JEPEGMIX2-P: a novel transcriptomic pathway method that greatly enhances detection of the molecular underpinnings for complex traits

Chris Chatzinakos[1*], Donghyung Lee[2], Cai Na[3], Vladimir I. Vladimirov[1], Anna Docherty[1], Bradley T. Webb[1], Brien P. Riley[1], Jonathan Flint[4], Kenneth S. Kendler[1] and Silviu-Alin Bacanu[1]

[1]Department of Psychiatry, Virginia Commonwealth University, Richmond Virginia, United States of America

[2]The Jackson Laboratory for Genomic Medicine, Farmington Connecticut, United States of America

[3]Welcome Trust Centre for Human Genetics, Oxford, United Kingdom

[4]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, United States of America

* chris.chatzinakos@vcuhealth.org

ABSTRACT

Genetic signal detection in genome-wide association studies (GWAS) can be enhanced by pooling small signals from multiple Single Nucleotide Polymorphism (SNP), e.g. across genes and pathways. Because genes are believed to influence traits via gene expression, it is of interest to combine information from expression Quantitative Trait Loci (eQTLs) in a gene or genes in the same pathway. Such methods, widely referred as transcriptomic, already exists for gene analysis, e.g. our group's **J**oint **E**ffect on **P**henotype of **e**QTLs associated with a **G**ene in **Mix**ed cohorts (JEPEGMIX/JEPEGMIX2). However, due to the its quadratic (in the number of SNPs) computational burden for computing linkage disequilibrium (LD) across large regions, transcriptomic methods are not yet available for arbitrarily large pathways/gene sets. To overcome this obstacle, we propose

JEPEGMIX2-pathways (JEPEGMIX-P), which implements a novel transcriptomic pathway method having a desirable linear computational burden. It 1) automatically estimates the ethnic composition (weights) of the cohort using a very large and diverse reference panel (33K subjects, including ~11K Han Chinese), 2) uses these weights and the reference panel to estimate the LD between gene transcriptomic statistics and 3) uses the estimated LD values along with GWAS summary statistics to rapidly test for the association between trait and the expression of genes even in the largest pathways. To underline its potential for increasing the power to uncover genetic signals over the state-of-the-art and commonly used non-transcriptomics (agnostic) methods, e.g. MAGMA, we applied JEPEGMIX2-P to summary statistics of several large meta-analyses from Psychiatric Genetics Consortium (PGC). Surprisingly, most of these significant pathways do not seem to be directly involved in the activity of the central nervous system. While our work is just the first step on the road toward the end goal of clinical translation, PGC anorexia results suggest possible avenues for (personalized) treatment.

Author summary

By using summary statistics from genetic studies to infer the association between the biologically relevant measure of gene expression and traits, transcriptomics methods are a promising avenue for uncovering risk genes and pathway of genes for complex human diseases. While numerous such transcriptomic methods were used to uncover a large number of gene level signals, due to the extreme computational burden, none of these methods was successfully extended for detecting signals at the, probably even more biologically relevant, pathway of genes level. In this paper we propose a novel

transcriptomic pathway method that has a close to minimally attainable computation burden and is applicable "as-is" to ethnically diverse studies. The proposed method adequately controls the false positive rates. Its application to psychiatric disorder studies unveils numerous new signals that were not detected by state-of-the art non-transcriptomic (agnostic) methods.

Introduction

Genome-wide association studies (GWAS) have been very successful for identifying diseases loci using single-marker based association tests [1]. Unfortunately, these methods have had limited power to identify causal genes or pathways [2]. For most complex traits, genetic risks are likely the result of the joint effect of multiple genes located in causal pathways [3]. Consequently, pooling information across genes in a pathway is likely to greatly improve signal detection.

Given that gene expression (GE), is widely posited to be the critical causal mechanism linking variant to phenotype [4], the paradigm for pooling of information should be informed by this mediating factor. GE based methods, widely denoted as transcriptomic, exist for gene-level inference [5-7]. They combine summary statistics at expression Quantitative Traits Loci (eQTL) known to best predict GE to infer the association between trait and GE for gene under investigation. The variance of the linear combination [8] is assessed using the estimated linkage disequilibrium (LD) matrix for all eQTLs which, for m variants, requires an $O(m^2)$ computational burden. This quadratic running time makes them unsuitable for transcriptomic pathway methods, due to the possibility of a very large

number of variants being analyzed in large regions of a chromosome. For instance, to estimate the correlation between statistics of genes in a pathway, we might have to combine information over tens of thousands of Single Nucleotide Polymorphisms (SNPs) in Major Histocompatibility (MHC) region from chromosome 6p (~25-35Mbp). Due to its irregular LD patterns, LD between any two SNPs in this region cannot be assumed to be negligible.

Currently, pathway analysis methods are non-transcriptomic, i.e. at a minimum they do not use the LD between transcriptomic gene statistics. Most of them just search for "agnostic" (i.e. not GE mediated) signal enrichment in a pathway/gene set. Among existing pathway methods we mention ALIGATOR [9], GSEA [10], DAPPPLE [11], as MAGENTA [12], INRICH [13] and MAGMA [14], as well as online tools: GeneGo/MetaCore (www.genego.com), Ingenuity Pathway Analysis (www.ingenuity.com), PANTHER (www.pantherdb.org), WebGestalt (bioinfo.vanderbilt.edu/webgestalt), DAVID (david.abcc.ncifcrf.gov) and Pathway Painter (pathway.painter.gsa-online.de. While not designed for pathway analyses, LDpred [15, 16] can also be adapted to test whether pathways are enriched above the polygenic background while adjusting for genomic covariates. Although all these tools were shown to be very powerful, a transcriptomic based pathway analyses can greatly complement the "agnostic" findings of all these tools.

To extend transcriptomic methods to pathway-level inference that models the LD between transcriptomic gene statistics, we propose a novel method, called JEPEGMIX2 Pathway

(JEPEGMIX2-P). It i) uses a very large and diverse reference panel consisting of 33K subjects (including around 11K Han Chinese), ii) automatically estimates ethnic composition of cohort, iii) uses these weight to compute LD for gene statistics via a linear running time procedure, iv) uses LD and GWAS summary statistics to rapidly test for the association between trait and expression of genes even in even the largest pathways, v) avoids an accumulation of just averagely enriched polygenic information by adjusting gene statistics for coding regions enrichment, and vi), to avoid the large signal in a gene inducing significant signals in all small pathways that include it, provides the option of a conditional analysis that eliminates the effect of SNPs with significant signals.

Results

JEPEGMIX2-P using our proposed automatic weight detection procedure (see Methods), controlled the false positive rates at or below the nominal threshold, even when this threshold was 10^{-6} , under both null (H_0) and "polygenic null" (H_p - enrichment in association signals is uniform over the entire genome) scenarios. When the method used "precise" prespecified subpopulation weights (e.g. using the closest subpopulations from the reference sample, i.e. as derived from the study description), the false positives rates were increased, especially for lower nominal rates, by up to ~220-450 (Text S1, Figs S1-S5 in SI). However, JEPEGMIX2-P with pre-estimated weights based on super populations (i.e. European, East Asian, African etc) had a much lower inflation of false positive rates; only for 10^{-6} threshold the false positive rate was increased by ~2-4 times, under both H_0 and H_p scenarios (Text S1, Fig S6 in SI).

For high-LD pathways, e.g. those defined by single chromosome bands in MSigDB [17-19], the behavior of automatically estimated weights is similar to the one for all pathways. However, false positive rates increase by ~300-1200 for the precise prespecified subpopulation weights (Text S1, Figs S7-S11 in SI), while when using super population-based weights, it remained practically unchanged from the 2-4X increase derived for all pathways (Text S1, Fig S12 in SI).

For the "nullified" data sets, i.e. those obtained from real data sets by substituting the study Z-scores by their expected quantile under H_0, JEPEGMIX2-P with automatic weights adequately controlled the size of the test. However, for the commonly used MAGMA false positive rates were up to one order of magnitude higher than the nominal ones, especially for the lower nominal thresholds (Fig 1).

Using the FDR p-value adjustment, for both unconditional and conditional JEPEG2-P analyses, we uncovered numerous significant pathway signals for Psychiatric Genetics Consortium (PGC) traits (Table 1). For the most significant we present heatmaps (Fig. 2-3, Text S2, Fig. S13-S17 in SI) while extended tables (Supplementary Excel file) include all significant signals. On the other hand, MAGMA (applied to the same GWAS summary statistics), most likely due to not employing the transcriptomic information, found fewer signals and only for SCZ (Table 2). JEPEGMIX2-P running time for a gene and pathway transcriptomic analysis of PGC data was less than 5 days on a single core of a cluster node with 4x Intel Xeon 6 core 2.67 GHz. MAGMA's running time was less than 3 days.

Table 1. Description of GWAS studies and traits that were analyses.

Trait	Trait Abbreviation	Dataset Description
Schizophrenia	SCZ	PGC2 SCZ [20]
Attention Deficit Hyperactivity Disorder	ADHD	PGC ADHD [21]
Autism	AUT	PGC AUT [22]
Bipolar	BIP	PGC BIP [23]
Eating Disorders	EAT	PGC EAT [24]
Major depression disorder	MDD	PGC MDD [25]

Table 2. Numbers of signals found by JEPEGMIX2-P and MAGMA.

Trait	**JEPEGMIX2-P without conditional analysis**	**JEPEGMIX2-P condition analysis**	**MAGMA**
ADHD	1	1	-
AUT	2	2	-
EAT	268	-	-
MDD	607	7	-
SCZ	825	27	5

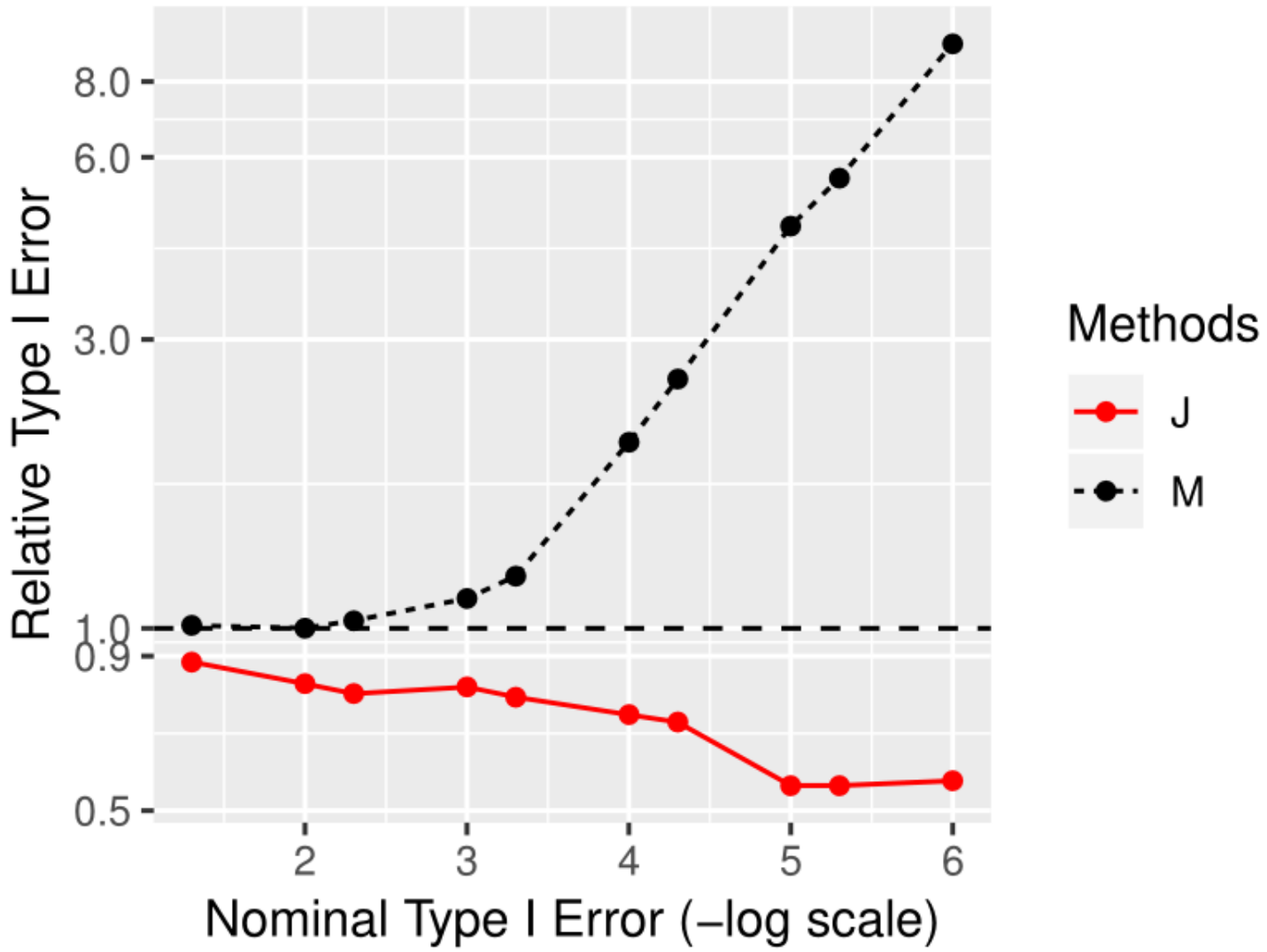

Fig 1. Relative size of the test (the quotient of empirical false positive rate and nominal type I error), for all pathways in the analysis of 20 nullified GWAS. In legend, Methods denotes whether the statistics had estimate from JEPEGMIX2-P (J) or from MAGMA (M).

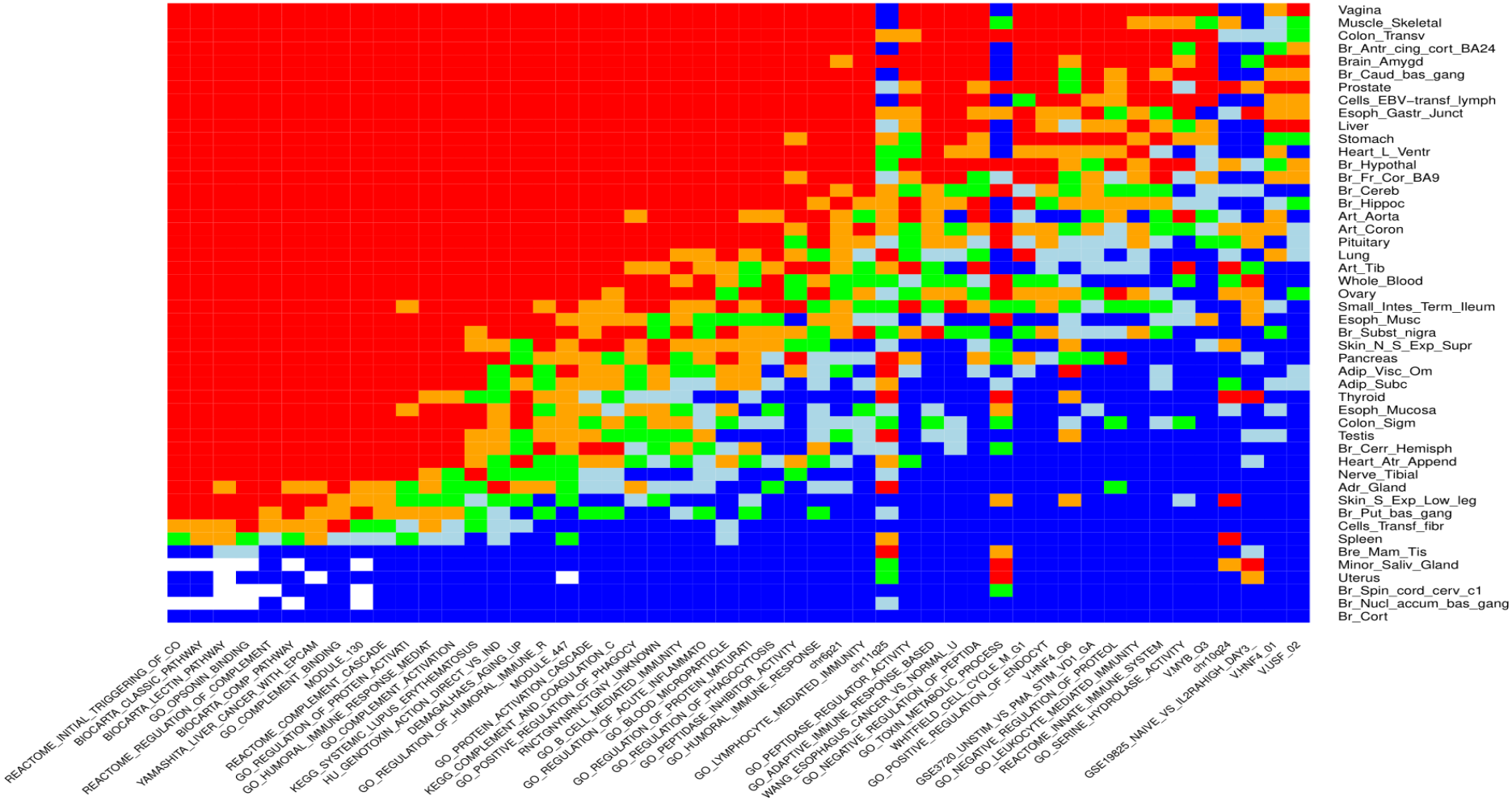

Fig 2. Top 50 SCZ pathway signals heatmap. The pathways and tissues are ordered in the decreasing order of the overall sum of $-log10(p$-values) for all tissues and pathways with at least two significant signals. Where red color denotes $q < 0.001$, orange $0.001 < q < 0.01$, green $0.01 < q < 0.05$, light blue $0.05 < q < 0.16$ and blue $0.16 < q < 1$. (See Excel Supplementary file for Abbreviations and the list of the signals not plotted above.)

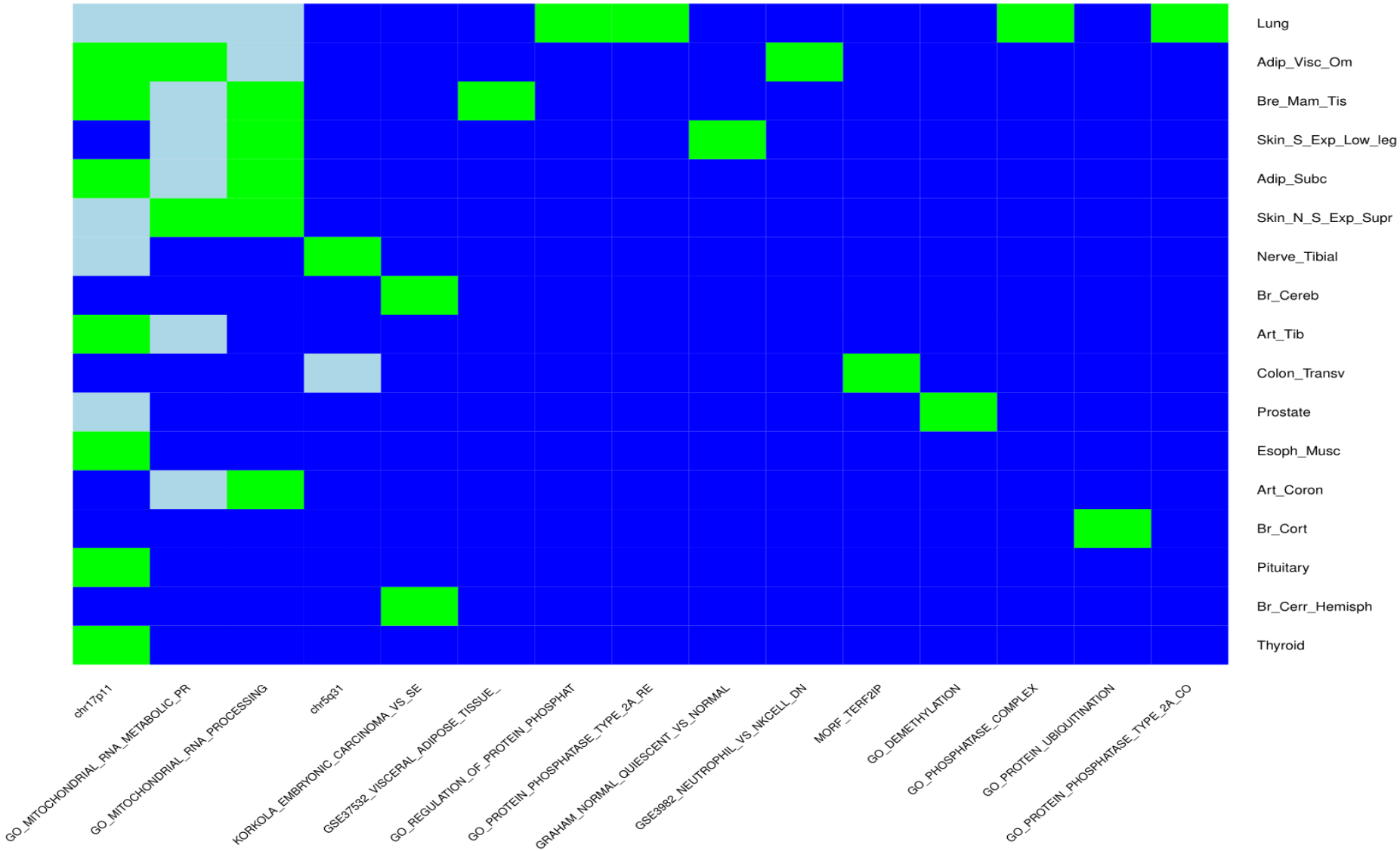

Fig 3. SCZ heatmap for pathway signals after conditioning on all significant SNP signals. See Fig 2. for background.

Method evaluation and comparison

To estimate the false positive rates of JEPEGMIX2-P, for five different cosmopolitan studies scenarios, we simulated (under H_0) 100 cosmopolitan cohorts of 10,000 subjects for Ilumina 1M autosomal SNPs using 1KG haplotype patterns [26] (Text S1, Table 1 of SI). The subject phenotypes were simulated independent of genotypes as a random Gaussian sample. SNP phenotype-genotype association summary statistics were computed from a correlation test. For each cohort, we obtained JEPEGMIX2-P statistics, for the two "null" enrichment scenarios i) under null (H_0), i.e. no enrichment, and ii) polygenic null (H_p), i.e. when enrichment is uniform over the entire genome regardless of functionality of individual genomic regions. For the JEPEGMIX2-P analyses of the resulting data we used i) prespecified (PRE) and ii) automatically estimated ethnic weights (EST). The prespecified (PRE) weights were assigned assuming information from the studies about subpopulations involved were available. As PRE-weights, we assigned "study-published" weights for the closest subpopulations from our new reference panel. Given that i) subjects were re-assigned to subpopulations in the new panel and ii) the populations labels in the new panel do not correspond to the ones from 100 Genomes, this induced possible mismatches that might result in increased false positive rates. To avoid this, a second version of the PRE-approach provides the published weights to continental superpopulations, i.e. European [EUR], East Asian [ASN], South Asian [SAS], African [AFR] and America native [AMR].

During our initial simulations we observed that pathways with name lengths ≤ 8, e.g. chr3p21, ch6p21 etc., have increased false positive rates due to having numerous genes

in high LD due to their proximity. For that reason, we also estimated the size of the test for all cohort scenarios, only for these high LD pathways.

Recently, due to its strength, MAGMA is one of the most used pathway analysis methods. Consequently, we compare the results obtained from our method with those obtained by this state-of-the-art method. However, to compare JEPEGMIX2-P with MAGMA, given that 1) the simulated cohorts might not reflect real data and 2) these data sets are for cosmopolitan cohorts (MAGMA software does not provide reference panel for these cohorts), we used real data to create “nullified” data sets. These nullified data sets are based on 20-real GWAS SCZ, ADHD, AUT, MDD and a further (preponderantly European) 16 data sets that are not yet publicly available. This approximation for null data is obtained by substituting the expected quantile of the Gaussian distribution for the (ordered) Z-score (see also Text S3 of the SI) after eliminating SNPs with significant association p-values in the original GWAS. However, one side effect of this approach consists of statistics within/near the peak signals in original GWASs might be too concentrated into the tails of the distribution to be a perfect “null data”. This can result in a slight increase in false positive rate, especially when applied to the nullified version of a highly enriched GWAS (e.g. PGC2 SCZ). However, most of the data sets used in “nullification” were not highly enriched in association signals.

Practical Applications. We applied JEPEGMIX2-P and MAGMA to summary statistics coming from Psychiatric Genetics Consortium (PGC- http://www.med.unc.edu/pgc/) datasets, i.e. SCZ, ADHD, AUT, BIP, EAT, MDD. To limit the increase in Type I error

rates of JEPEGMIX2-P, we deem as significantly associated only those pathways that yield an FDR-adjusted p-value (q-value)< 0.05. Due to *C4* explaining most of Major Histocompatibility (MHC) (chr6:25-33 Mb [27], gene/signals for SCZ, for this trait, we omit non-*C4* genes in this region. Moreover, due to the high correlation between SNPs in MHC (chr6:25-33 Mb), we also omit genes in this region for MDD, which also showed MHC signals.

Discussion

The discovery of biological pathways implicated in diseases is the target for any genetic analysis. Despite the numerous methods available for pathways analysis, none of these methods relies solely on eQTLs to infer the association between expression of genes in pathway and trait, which is widely posited to be the critical causal mechanism. To overcome these two main factors, we propose JEPEGMIX2-P for testing the association between pathway expression and trait. Even for uniformly enriched GWAS and high LD pathways, JEPEGMIX2-P with the automatic weights fully controls the false positive rates at or below nominal levels.

Narrowly assigning the ethnic weights to the subpopulations perceived as the “closest” to the ones in the studies is not advisable due to the possibility of great mismatch between the cohort and the “re-arranged” subpopulations from ourreference panel, which can result in greatly increased false positive rates (see Methods). Consequetly, users should use the automatic detection of cohort composition, regardless whether cohort AFs are available or not.

Applying JEPEGMIX2-P to psychiatric phenotypes, we discovered numerous pathways that were deemed significant for SCZ, ADHD, AUT, EAT and MDD. We mention that while the original SCZ paper did not report any genome-wide significant pathways and MAGMA reports only five, JEPEGMIX2 detected hundreds of them. Even more, these signals are not very likely to be false discoveries due to our method i) adjusting both SNP and gene statistics for polygenic/gene enrichment background, ii) accurately estimating of ethnic weights and iii) excellent control of Type I error rates.

Interpreting and validating all pathway signals require substantially more work. It is always “the last mile” that is the most laborious. However, JEPEGMIX2-P provides carefully vetted targets for wet-lab validation. Nonetheless, our findings sometimes allow for reasonably informed inferences. For instance, in anorexia results, as one of the uncovered signals, the pathway GEISS_RESPONSE_TO_DSRNA_UP (Supplementary Excel file) is a pathway that is involved in response to virus infections. This finding suggests a possible avenue of treatment: patients with active virus infections might benefit from being treated with anti-viral medication. However, the responders to such treatment are likely to form only a minority of the anorexia patients, i.e. fraction of those with active viral infections.

While the method is a welcome addition to our pathway tools, it still has limitations when used for assigning “causal” tissues/cell. First, due to the rather small sample sizes of existing GE experiments, GE in different tissues is often correlated and greatly incomplete due to ~80% of genes not having good GE prediction from eQTL SNPs. Consequently,

the capacity of inferring causal tissues/cell types will be greatly enhanced by future updates that i) use larger GE studies and ii) impute the statistics of most genes that do not have reliable eQTL predictions by using a) statistics from genes with good GE predictions and ii) the empirical correlations between their gene expressions (e.g. GTEx derived).

Methods

Naïve application of many analysis methods for genes/pathways with numerous SNPs/genes might yield large signals merely by accumulating “average” polygenic signals from well-powered studies. To avoid such an accumulation of average polygenic information, we competitively adjust SNP and gene level χ^2 statistics for the background enrichment of genome wide SNPs and transcriptomic gene statistics, respectively. This is achieved simply by adjusting gene statistics for average non-centrality (Text S4, S5 of Supplementary Information-SI). Subsequently, as detailed below, we use the GWAS summary statistics i) to estimate the ethnic composition of the study cohort and ii) use the estimated ethnic weights to build a pathway statistic that has a highly desirable $O(m)$ computational burden.

GWAS summary data comprise of a large range of effect sizes and it is unclear whether the estimated pathway statistics are related to the whole range, including SNPs with very small effects, or just SNPs with large effects. To avoid a very large signal in a gene inducing a significant signal in all smaller pathways including the gene, we also offer the option to eliminate the effect size of big SNPs, by applying a novel conditional analysis

procedure (Text S6 in SI) to summary statistics before their use in our transcriptomic pathway tool.

Automatic detection of the ethnic composition for the cohort. The LD between markers can vary widely between human populations. Thus, to compute the LD, which is necessary for internal imputation and variance estimation for gene statistics, we need to estimate the ethnic composition of the cohort. Our group has previously described in DISTMIX paper [26] a method of using the reference panel to estimate the ethnic composition when the cohort allele frequencies (AF) are available. However, lately consortia do not provide such summary measure; they often might provide just the Caucasians AF. *Consequently, there is a need for a method to estimate the ethnic composition of the cohort even when no AFs are provided*. Below is the theoretical outline of such method, which uses only the SNP Z-scores summary statistics.

Assume that the cohort genotype is a mixture of genotypes from k ethnic subpopulation from a large and diverse reference panel. If the i-th subject at the j-th SNP has genotype G_{ij} and belongs to the l-th group, let $p_j^{(l)}$ be the frequency of the reference allele frequency for this SNP in the l-th group. Let $q_j^{(l)} = 1 - p_j^{(l)}$ and $G'_{ij} = \frac{G_{ij} - 2\,p_j^{(l)}}{\sqrt{2p_j^{(l)}\,q_j^{(l)}}}$ be the normalized genotype, i.e. the transformation to a variable with zero mean and unit variance. Near H_0, SNP Z-score statics $Z_j s$ have the approximately the same correlation structure as the genotypes used to construct it, G_{*j}'s, and, thus, the same correlation structure as its transformation, G'_{*j}'s. However, given that both $G'_{*j}{}'s$ and $Z_j s$ have unit variance, it follows that the two have the same covariance (i.e. not only the same correlation) structure. Therefore, for any $s \geq 1$

$E(Z_j\ Z_{j+s}) = E(G'_{*j}\ G'_{*\,j+s})$, which, assuming that $w^{(l)}$ is the expected fraction of subjects from the entire cohort that belong to the l-th subpopulation from the reference panel, becomes

$$E(Z_j\ Z_{j+s}) = \sum_{l=1}^{k} w^{(l)}\ E\left[{G'_{*j}}^{(l)}\ {G'_{*(j+s)}}^{(l)}\right] = \sum_{l=1}^{k} w^{(l)}\ Cov\left({G'_{*j}}^{(l)}, {G'_{*(j+s)}}^{(l)}\right) =$$

$$\sum_{l=1}^{k} w^{(l)}\ Cor({G'_{*j}}^{(l)}, {G'_{*(j+s)}}^{(l)}) \quad (1).$$

Henceforth, we will simply denote the $\boldsymbol{w}$ vector as weights.

While $Cor({G'_{*j}}^{(l)}, {G'_{*(j+s)}}^{(l)})$ is unknown, it can be easily estimated using their reference panel counterparts with appropriate ethnic weights. Thus, the weights, $w^{(l)}$, can be simply estimated by simply regressing the product of Z-scores of reasonably close SNP Z-scores, $Z_j\ Z_{j+s}$, on correlations between normalized genotypes at the same SNP pairs for all subpopulations in the reference panel. Because some GWAS might have numerous large signals, e.g. latest height meta-analysis [28], a more accurate estimation of the weights in equation (1) is very likely to be obtained by substituting the expected Gaussian quantiles for Z_j (see text S3 in SI).

Due to the strong LD among SNPs, the estimation of the correlation using all SNPs in a genome simultaneously might lead to a poor regression estimate in (1). To avoid this, we sequentially split GWAS SNPs into 1000 non-overlapping SNP sets, e.g. first set consists of the 1st, 1001st, 2001st, etc. map ordered SNPs in the study. The large distances between SNPs in the same set make them quasi-independent which, thus, improves the accuracy of the estimated correlation. $W = (w^{(l)})$ is subsequently estimated as the

average of the weights obtained from the 1000 SNP sets. Finally, we set to zero the negatives weights and normalize the remaining weights to sum to 1 [29].

While approximate continental (EUR, ASN, SAS, AFR and AMR) ethnic distribution of subjects can be easily estimated from study info, it is not always clear how these weights should be apportioned among continental subpopulations. This further apportioning is likely to be important when the GWAS cohorts contain a large number of admixed populations, e.g.

African Americans and American native populations, which in the making of the reference panel, had many subjects re-assigned to related subpopulations. Consequently, when continental proportions are provided by the users, we can use the above described automatic detection to ditribute these weights to the most likely subpopulations in the reference panel.

$O(m)$ LD estimation procedure. It is very computationally challenging [$O(m^2)$] for m genetic variants] to estimate the large correlation matrices needed to compute transcriptomic pathway statistics (substantially more so for the upcoming larger reference panels). The same heavy computational burden occurs in fine-mapping when there is a desire to output correlation between statistics of genes and pathways with suggestive/significant signals. Thus, for computational feasibility, we need to find an approach that avoids computing correlation matrices.

For the theoretical justification of such an approach we use the mathematical notation from the automatic weight estimation, where $G'_{*j} = \frac{G_{ij} - 2\,p_j^{(l)}}{\sqrt{2\,p_j^{(l)} q_j^{(l)}}}$ is the normalized version of G_{*j}, i.e. with means 0 and variance 1, like the Z-scores. The Z-score transcriptomics statistic per gene or pathway is a linear combination of the Z-scores from expression Quantitative Trait Loci (eQTL) SNPs [6]: $Z = \frac{\sum_{j=1}^{m} b_j\, Z_j}{SD(\sum_{j=1}^{m} b_j\, Z_j)}$,

where the $SD(\sum_{j=1}^{m} b_j\, Z_j)$ is not known and should be estimated reasonably fast. Thus, in general we are interested in computing the covariance between two very large pathway scores (or the variance of a large one), i.e. linear combinations of Z-scores: $Cov(\sum_{j=1}^{m} a_j\, Z_j, \sum_{j=1}^{m} b_j\, Z_j)$. As stated above, working "by SNP" and computing the correlation is $O(m^2)$ an, thus, highly untenable for very large combinations of SNP statistics. However, it is possible to work by "mimicking" the higher order entity (gene, pathways) statistics by observing that, under the null hypothesis, $\sum_{j=1}^{m} a_j\, Z_j$ and $\sum_{j=1}^{m} b_j\, Z_j$ have, due to normalization of G'_{*j}, a distribution that is identical to the distribution of $\sum_{j=1}^{m} a_j\, G'_{*j}$ and $\sum_{j=1}^{m} b_j G'_{*j}$, respectively.

Thus, $Cov(\sum_{j=1}^{m} a_j\, Z_j, \sum_{j=1}^{m} b_j\, Z_j) = Cov(\sum_{j=1}^{m} a_j\, G'_{*j}, \sum_{j=1}^{m} b_j\, G'_{*j})$, which is easily estimated from a reference sample without computing correlation matrices, by using just a highly desirable linear [$O(m)$] running time procedure. For the correlation between two pathway statistics, then:

$$Cor(\sum_{j=1}^{m} a_j Z_j, \sum_{j=1}^{m} b_j Z_j) = \frac{Cov(\sum_{j=1}^{m} a_j G'_{*j}, \sum_{j=1}^{m} b_j G'_{*j})}{\sqrt{Var(\sum_{j=1}^{m} a_j G'_{*j})}\sqrt{Var(\sum_{j=1}^{m} b_j G'_{*j}, \sum_{j=1}^{m} b_j G'_{*j})}} \quad (2)$$

Within JEPEGMIX-P, the covariances and correlations of the statistics are transparently computed using subject weights reflecting the fraction in the study cohort for each ethnic group from the reference panel. Thus, computing the correlations reduces to simply applying linear combinations to normalized genotype vectors in reference panels followed by very simple estimations of weighted covariance and variance matrices for the two vectors. We need to underscore again that besides the huge memory savings, the proposed method has linear running time while estimating the correlation matrix has a quadratic (in the number of SNPs) running time.

Computation of pathway statistic. Generic transcriptomic methods output Z-score statistics by gene. Thus, if the correlation between gene statistics is available, e.g. by using the $O(m)$ method described above, these statistics can be combined using a Mahalanobis χ^2 statistics with the number of degrees of freedom (df) equal to the number of genes. Unfortunately, this can become quickly very involved if we need to compute the LD between statistics of all ~20,000 genes. However, given that the genotypes of variants in different chromosome arms are practically independent, if follows that Z-scores for genes on different chromosome arms are independent. Thus, the Mahalanobis type statistics can be computed more easily by chromosome arm and the pathway χ^2 statistics are computed simply as the sum of chromosome arm statistics (Fig 4). Similarly, pathway statistic df equals the sum of the dfs for chromosome arm statistics.

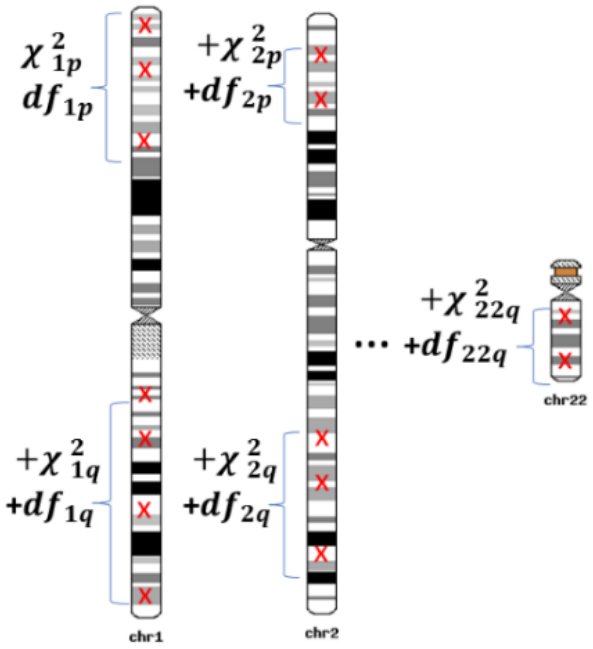

Fig 4. Computation of pathway

Annotation and reference panel. As pathway database we used MSigDB [17-19], which is well maintained and widely used by researchers. To facilitate user-specific input for new pathways along with future extensions, the annotation file now includes an R-like formula for the expression of each gene as a function of its eQTL genotypes and of the content for each pathway as a function of the names its constituting genes. The updated annotation file includes cis-eQTL for all tissues available in the v0.7 version of PredictDB (http://predictdb.hakyimlab.org/). To avoid making inference about genes poorly predicted by SNPs, for the 48 available tissues (Text S7, Table S2 of SI), we retain only genes for which the expression is reasonably accurately predicted (q-value < 0.05) from its eQTLs. The current version uses the $32{,}953$ subjects (33K) as the reference panel. It consists of $20{,}281$ Europeans, $10{,}800$ East Asians (from CONVERGE study, Text S8 of SI), 522 South Asians, 817 Africans and 533 Native of Americas (Text S9 Table S3 of SI).

Software and data availability

JEPEGMIX2-P is freely available for *academic use* at https://github.com/Chatzinakos/JEPEGMIX-P. The JEPEGMIX2-P executable requires only the GWAS summary statistics from the user. The reference panel and the annotation files are also available at the same repo.

Supporting information

SI. Text and Figures.
(PDF)

SE. Table with the significant signals for the real applications.
(EXCEL)

Reference

1. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLOS Computational Biology. 2012;8(12):e1002822.
2. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet. 2010;11(12):843-54.
3. Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends Genet. 2012;28(7):323-32.
4. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, et al. Genetics of gene expression and its effect on disease. Nature. 2008;452(7186):423-8.
5. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091-8.
6. Chatzinakos C, Lee D, Webb BT, Vladimirov VI, Kendler KS, Bacanu SA. JEPEGMIX2: improved gene-level joint analysis of eQTLs in cosmopolitan cohorts. Bioinformatics. 2017.
7. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48(3):245-52.
8. Jin L, Zuo XY, Su WY, Zhao XL, Yuan MQ, Han LZ, et al. Pathway-based analysis tools for complex diseases: a review. Genomics Proteomics Bioinformatics. 2014;12(5):210-20.
9. Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, et al. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet. 2009;85(1):13-24.
10. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50.

11. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. 2011;7(1):e1001273.
12. Segre AV, Groop L, Mootha VK, Daly MJ, Altshuler D, Consortium D, et al. Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. Plos Genetics. 2010;6(8).
13. Lee PH, O'Dushlaine C, Thomas B, Purcell SM. INRICH: interval-based enrichment analysis for genome-wide association studies. Bioinformatics. 2012;28(13):1797-9.
14. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput Biol. 2015;11(4):e1004219.
15. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291-5.
16. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47(11):1228-35.
17. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-25.
18. Liberzon A. A description of the Molecular Signatures Database (MSigDB) Web site. Methods Mol Biol. 2014;1150:153-60.
19. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739-40.
20. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013;45(10):1150-9.
21. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, et al. Discovery Of The First Genome-Wide Significant Risk Loci For ADHD. bioRxiv. 2017.
22. Buxbaum JD, Daly MJ, Devlin B, Lehner T, Roeder K, MW S. The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. Neuron. 2012;76(6):1052-6.
23. Psychiatric GCBDWG. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet. 2011;43(10):977-83.
24. Duncan L, Yilmaz Z, Gaspar H, Walters R, Goldstein J, Anttila V, et al. Significant Locus and Metabolic Genetic Correlations Revealed in Genome-Wide Association Study of Anorexia Nervosa. Am J Psychiat. 2017;174(9):850-8.
25. Major Depressive Disorder Working Group of the Psychiatric GC, Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, et al. A mega-analysis of genome-wide association studies for major depressive disorder. Mol Psychiatry. 2013;18(4):497-511.
26. Lee D, Bigdeli TB, Williamson VS, Vladimirov VI, Riley BP, Fanous AH, et al. DISTMIX: Direct imputation of summary statistics for unmeasured SNPs from mixed ethnicity cohorts. Bioinformatics. 2015.
27. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1279-83.
28. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46(11):1173-86.
29. Chatzinakos C, Lee D, Webb BT, Vladimirov VI, Kendler KS, Bacanu S-A. JEPEGMIX2: improved gene-level joint analysis of eQTLs in cosmopolitan cohorts. Bioinformatics. 2017:btx509-btx.