*Genome analysis*

# JEPEGMIX2-P: Novel pathway transcriptomic method greatly increases detection of molecular pathways in cosmopolitan cohorts

Chris Chatzinakos[1*], Donghyung Lee[2], Cai Na[3], Anna Docherty[1], Vladimir I. Vladimirov[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 23298, USA

[2]The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut 06032, USA

[3]Welcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK

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

## ABSTRACT

**Motivation**: 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 trait 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, denoted as transcriptomic, already exists for gene analysis, e.g. our **G**ene-level **J**oint **A**nalysis of functional SNPs in **C**osmopolitan **C**ohorts (JEPEGMIX/JEPEGMIX2).
However, due to the $O(m^2)$ computational burden for computing linkage disequilibrium (LD) between numerous (m) SNPs across large regions, transcriptomic methods are not yet applicable to large pathways/gene sets. To overcome this obstacle, we propose JEPEGMIX2 pathways (JEPEGMIX-P), a novel $O(m)$ transcriptomic pathway method. It 1) automatically estimates the ethnic composition of the cohort along with a very large reference panel for a) imputing unmeasured eQTLs and b) accurately estimating LD for gene statistics and 2) uses estimated LD and GWAS summary statistics to rapidly test for the association between trait and expression of genes even in the largest pathways.

**Results:** We applied JEPEGMIX2-P to summary statistics from several large meta-analyses from Psychiatric Genetics Consortium Schizophrenia (PGC). Its application to PGC anorexia seem to suggest interesting avenues for treatment.

**Contact:** chris.chatzinakos@vcuhealth.org
Software: https://github.com/Chatzinakos/JEPEGMIX2-P
**Supplementary information:** Supplementary material is available.

## 1 INTRODUCTION

Genome-wide association studies (GWAS) have been very successful for identifying diseases loci using single-marker based association tests (Bush and Moore, 2012). Unfortunately, these methods have had limited power to identify disease genes (Wang, et al., 2010). For most complex traits, genetic risks are likely the result of the joint effect of multiple genes located in causal pathways (Ramanan, et al., 2012). 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 (Emilsson, et al., 2008), the pooling of information should be informed by this mediating factor. GE based methods, also denoted as transcriptomic, exist for gene-level inference (Chatzinakos, et al., 2017; Gamazon, et al., 2015; Gusev, et al., 2016). They combine summary statistics at expression Quantitative Traits Loci (eQTL) known to best predict GE to infer the association between GE and trait for gene under investigation. The variance of the linear combination (Jin, et al., 2014) is assessed using the estimated linkage disequilibrium (LD) matrix for all eQTLs which, for m variants, requires $O(m^2)$ computational burden. This quadratic running time makes them unsuitable for transcriptomic pathway methods, due to possibly very large number of variants needed to be analyzed in 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) 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 (Holmans, et al., 2009), GSEA (Subramanian, et al.,

2005), DAPPPLE (Rossin, et al., 2011), as MAGENTA (Segre, et al., 2010), , INRICH (Lee, et al., 2012) and MAGMA (de Leeuw, et al., 2015) 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 (Bulik-Sullivan, et al., 2015; Finucane, et al., 2015) can also be adapted to test if pathways are enriched above the polygenic background while adjusting for genomic covariates. While all these tools were shown to be very powerful, a transcriptomic based pathway analyses can greatly complement their "agnostic" findings.

To extend the transcriptomic method to pathways in a parametric manner that models the LD between transcriptomic gene statistics, we propose JEPEGMIX2 Pathway (JEPEGMIX2-P). It is a non-trivial extension of our gene-level JEPEGMIX2 transcriptomic method. JEPEGMIX2-P i) uses a large reference panel with 33K subjects, ii) automatically estimates ethnic composition of cohort, iii) computes LD for gene statistics using these weights in an $O(m)$ procedure, where $m$ is the number of SNPs, iv) uses LD and GWAS summary statistics to rapidly test for the association between trait and expression of genes even in the largest pathways, v) avoid 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 small pathways that include it, provides the option of a conditional analysis that eliminates the effect of SNPs with significant signals.

## 2 METHOD

Naïve application of many analyses methods for genes/pathways with a large number of 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 $\chi^2$ statistics for background enrichment of genome wide SNPs and gene statistics, respectively. This is achieved simply by adjusting these statistics for average non-centrality (Text S1, S2 of Supplementary Data-SD). Subsequently, as detailed below, we use the GWAS summary statistics i) to estimate the ethnic composition of the study cohort and ii), along the estimated ethnic weights, we build a pathway statistic that has a highly desirable $O(m)$ computational burden.

GWAS summary data comprise 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 S3 of SD) to summary statistics before inputting them in our transcriptomic pathway tool.

### 2.1 Automatic detection of the ethnic composition for the cohort

The LD between markers is ethnicity specific. 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 (Lee, et al., 2015) a method to estimate the ethnic composition when the cohort allele frequencies (AF) are available. However, lately consortia do not provide such 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.

Assume that the cohort genotype is a mixture of genotypes from $k$ ethnic groups from the 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 $G'_{ij} = \frac{G_{ij} - 2\,p_j^{(l)}}{\sqrt{2p_j^{(l)}\,(1-p_j^{(l)})}}$ 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, due independence of genotypes in different ethnic groups becomes

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

where $w^{(l)}$ is the expected fraction of subjects from the entire cohort that belong to the $l$-th group.

While $Cor(G'^{(l)}_{*j}, G'^{(l)}_{*(j+s)})$ is unknown, it can be easily approximated using their reference panel counterparts with appropriate ethnic weights. Thus, the weights, $w^{(l)}$, can be simply estimated by simply regressing the product of product 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. To increase discrimination power, in practice the parameter $s$ is chosen to maximize the variance of the within panel ethnic group correlations while keeping $j+s$-th SNP no more than 25Kb away from $j$ −th SNP. Because some GWAS might have numerous large signals, e.g. latest height meta-analysis (Wood, et al., 2014), a more accurate estimation of the weights is very likely to be obtained by substituting expected gaussian quantiles for $Z'_j$ (see text S4 in SD).

Due to the strong LD among SNPs, the estimation of the correlation using all SNPs in a genome 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 (Chatzinakos, et al., 2017).

While approximate continental (European [EUR], East Asian [ASN], South Asian [SAS], African [AFR] and America native[AMR]) ethnic distribution of subjects can be easily estimated from study info, it is not always clear how these proportions 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.

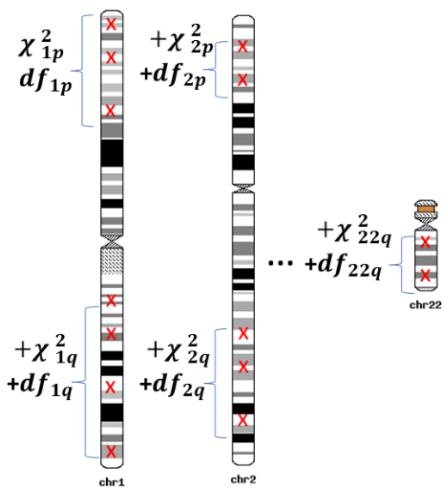


***Fig. 1 Computation of pathway statistics***

## 2.2 $O(m)$algorithm

It is very computationally challenging [$O(m^2)$ for $m$ genetic variants] to estimate the large correlation matrices needed for transcriptomic pathway statistics (substantially more so for 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.

In more details, let $G_{ij}, p_j$ and $Z_j$ , $i = 1, \dots, n$, $j = 1, \dots, m$ the matrix of genotypes for the $n$ subjects at $m$ SNPs, the reference allele frequencies at the $n$ SNPs their corresponding vector of Z-scores of testing the association of SNPs with trait. Let $q_j = 1 - p_j$ and vectors $G'_{*j} = \frac{G_{ij} - 2\,p_j}{\sqrt{2\,p_j\,q_j}}$ be 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 expression Quantitative Trait Loci (eQTL) SNPs (Chatzinakos, et al., 2017):

$$Z = \frac{\sum_{j=1}^{m} b_j\, Z_j}{SD(\sum_{j=1}^{m} b_j\, Z_j)},$$

where the $SD(\sum_{j=1}^{m} b_j\, Z_j)$ is not known and has to 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 linear combinations. However, it is possible to work by "mimicking" the higher 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 $G'_{*j}$, identical distribution to $\sum_{j=1}^{m} a_j\, G'_{*j}$ and $\sum_{j=1}^{m} b_j G'_{*j}$ . 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 an $O(m)$ procedure. For the correlation between two pathway statistics, then:

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

Within JEPEGMIX2 pathway, the covariances and correlations of the statistics are computed transparently 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 is linear in time in $m$ (the number of SNPs) while estimating the correlation matrix has a quadratic running time in $m$.

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 $\chi^2$ 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 $\chi^2$ statistics are computed simply as the sum of chromosome arm statistics (Fig. 1). Similarly, pathway statistic df equals the sum of the dfs for chromosome arm statistics.

## 2.3 Annotation and reference panel

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 its gene names. 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 S5, Table 1. of SD), we retain only genes for which the expression is predicted with q-value $< 0.05$ from its eQTLs. The current version uses the 33,000 subjects (33K) as the reference panel. It consists of 20,281 Europeans, 10,800 East Asians (Converge, Text S6 of SD), 522 South Asians, 817 Africans and 533 Native of Americas (Text S7 Table 2. of SD).

# 3 SIMULATIONS

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 (Lee, et al., 2015) (Text S8 Table 3. of SD). 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 scenario, we obtained

JEPEGMIX2-P statistics, for the two enrichment cases i) under null ( $H_0$) and ii) null polygenic ($H_p$) with prespecified ethnic weights (PRE) and with automatic ethnic weights (EST) for JEPEGMIX2-P. The prespecified weights were assigned under two scenarios, both assuming information from the studies about subpopulations involved were available. In the first scenario, we assigned the cohort weights for the closes subpopulations in 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. Consequently, in the second scenario we provided weights to continental superpopulations, i.e. AFR, AMR, ASN, EUR and SAS.

During our initial simulations we observed that pathways with character length names less or equals to 8, e.g. chr15q25, ch6p21 etc., have increased false positive rates due to having numerous genes in high LD due to their close proximity. For that reason, we also estimated the size of the test for all cohort scenarios, only for these high LD pathways.

### 3.1 Practical Applications

We applied JEPEGMIX2-P to psychiatric summary datasets available for download from Psychiatric Genetics Consortium (PGC-http://www.med.unc.edu/pgc/) , i.e Schizophrenia (SCZ) (Ripke, et al., 2013), Attention Deficit Hyperactivity Disorder (ADHD) (Demontis, et al., 2017), Anxiety (ANX) (Otowa, et al., 2016), Autism (AUT) (Buxbaum, et al., 2012), Bipolar (BIP) (Psychiatric, 2011), Eating Disorders (EAT) (Duncan, et al., 2017), Education (EDU), Major Depression Disorder (MDD) (Wray, et al., 2018) (Text S9, Table 4. in SD). To limit the increase in Type I error rates of JEPEGMIX2-P, we deem as significantly associated only pathways with an FDR-adjusted p-value (q-value)$< 0.05$. Due to C4 explaining most of Major Histocompatibility (MHC) (chr6:25-33 Mb (McCarthy, et al., 2016), gene/signals for schizophrenia (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.

## 4 RESULTS

JEPEGMIX2-P with automatic weights, controls the false positive rates at or below the nominal threshold, even at $10^{-6}$ , under both $H_0$ and $H_p$ scenarios, contrary the false positives rates were increased, especially for lower nominal rates, up to ~~220-450 at the version with the "precise" prespecified subpopulation weights (Text S8, Figure S1-S5 in SD). While, JEPEGMIX2-P, with pre-estimated weights according to the super populations, controls the false positive rates close to the nominal threshold, only at $10^{-6}$ the size of the test increases around 2 to 4, under both $H_0$ and $H_p$ scenarios (Text S8, Figure S6 in SD).

For high-LD pathways, JEPGMIX2-P with automatic weights, controls the false positive rates at or below the nominal threshold, even at $10^{-6}$ , for both $H_0$ and $H_p$ scenarios, while the false positives rates were increased, for all nominal rates, up to ~300-1200 at the version with the precise prespecified subpopulation weights (Text S8, Figure S7-S11 in SD). JEPEGMIX2-P, with pre-estimated weights according to the super populations, controls the false positive rates close to the nominal threshold, only at $10^{-6}$ the size of the test increases around 2 to 4, under both $H_0$ and $H_p$ scenarios (Text S8, Figure S12 in SD).

Using the FDR p-value adjustment, for JEPEGMIX2-P, for both cases (with and without elimination), for the PGC datasets, significant pathway signals were found, for which we present heatmaps (Text S9, Figure S13-S25 in SD). Each analysis ran in less than 5 days on a cluster node with 4x Intel Xeon 6 core 2.67 GHz.

## 5 CONCLUSIONS

The discovery of pathways implicated in the difference between cases and controls is a powerful approach for genetic analysis. Despite the different methods available for pathways, 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 enriched GWAS, JEPEGMIX2-P with the automatic weights fully controls the false positive rates at or below nominal levels, even for high high-LD pathways.

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 reference panel subpopulations, which can result in greatly increased false positive rates. To eliminate unforseen aretfacts, we recommend to the users to provide continental super-population weights when allele frequencies (AFs) are not available. Subsequently, our automatic ethnic detection algorith can reapportion the continental weights to the most likely subpopulations. When cohort AFs are available there is no need to provide suh continental weights.

Applying JEPEGMIX2-P to psychiatric phenotypes, we discovered significant interactions in SCZ, ADHD, ANX (factor scores only), AUT, EAT, EDU and MDD. We mention that the original SCZ paper did not report any genome-wide significant pathways, JEPEGMIX2 found hundreds of them. Even more, they are believable as we avoid false discoveries by i) adjusting both SNP and gene statistics for polygenic/gene enrichment background and ii) accurately estimating of ethnic weights. The Anorexia results are interesting, as they suggest a possible treatment for a part of the patient population. The most significant pathway is Geiss response to double stranded RNA, i.e. the genes that are positively regulated in glia when viruses are detected. Consequently, patients with active virus infections might be treated with anti-viral medication. However, the responders are likely to be only a minority of the patients with active such infections.

While the method is a welcome addition to pathway tools is still has limitations when used for assigning "causal" tissues. First, gene expression (GE) in different tissues is often correlated and greatly incomplete due to ~80% of genes not having good GE prediction from eQTL SNPs due to the rather small sample sizes of existing GE experiments. Consequently, the capacity of discriminating causal tissues will be greatly enhanced by future updates that i) use larger GE studies and ii) imputing the statistics of most

genes without good predictions by using a) statistics from genes with good GE predictions and ii) the empirical correlations between their gene expressions.

## REFERENCES

Bulik-Sullivan, B.K., *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet* 2015;47(3):291-295.

Bush, W.S. and Moore, J.H. Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology* 2012;8(12):e1002822.

Buxbaum, J.D., *et al.* The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron* 2012;76(6):1052-1056.

Chatzinakos, C., *et al.* JEPEGMIX2: improved gene-level joint analysis of eQTLs in cosmopolitan cohorts. *Bioinformatics* 2017.

de Leeuw, C.A., *et al.* MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS. Comput. Biol* 2015;11(4):e1004219.

Demontis, D., *et al.* Discovery Of The First Genome-Wide Significant Risk Loci For ADHD. *bioRxiv* 2017.

Duncan, L., *et al.* Significant Locus and Metabolic Genetic Correlations Revealed in Genome-Wide Association Study of Anorexia Nervosa. *Am J Psychiat* 2017;174(9):850-858.

Emilsson, V., *et al.* Genetics of gene expression and its effect on disease. *Nature* 2008;452(7186):423-428.

Finucane, H.K., *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet* 2015;47(11):1228-1235.

Gamazon, E.R., *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 2015;47(9):1091-1098.

Gusev, A., *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 2016;48(3):245-252.

Holmans, 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.

Jin, L., *et al.* Pathway-based analysis tools for complex diseases: a review. *Genomics Proteomics Bioinformatics* 2014;12(5):210-220.

Lee, D., *et al.* DISTMIX: Direct imputation of summary statistics for unmeasured SNPs from mixed ethnicity cohorts. *Bioinformatics* 2015.

Lee, P.H., *et al.* INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* 2012;28(13):1797-1799.

McCarthy, S., *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48(10):1279-1283.

Otowa, T., *et al.* Meta-analysis of genome-wide association studies of anxiety disorders. *Mol Psychiatry* 2016;21(10):1485.

Psychiatric, G.C.B.D.W.G. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* 2011;43(10):977-983.

Ramanan, V.K., *et al.* Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet* 2012;28(7):323-332.

Ripke, S., *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* 2013;45(10):1150-1159.

Rossin, E.J., *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* 2011;7(1):e1001273.

Segre, A.V., *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).

Subramanian, A., *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):15545-15550.

Wang, K., Li, M. and Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* 2010;11(12):843-854.

Wood, A.R., *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014;46(11):1173-1186.

Wray, N.R., *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* 2018;50(5):668-681.