**Novel Bayesian approach for transcriptome-wide association study by leveraging genome-wide eQTL information through summary statistics**

Justin M. Luningham1, Junyu Chen5, Shizhen Tang2,5, Philip L. De Jager3, David A. Bennett4, Aron S. Buchman4, Jingjing Yang5\*

1. Department of Population Health Sciences, Georgia State University School of Public Health, Atlanta, GA, 30303, USA
2. Department of Biostatistics and Bioinformatics, Emory University School of Public Health, Atlanta, GA, 30322, USA
3. Medical Center Neurological Institute, Columbia University, New York, NY, 10032, USA
4. Rush Alzheimer’s Disease Center, Rush University Medical Center, Chicago, IL, 60612, USA
5. Center for Computational and Quantitative Genetics, Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, 30322, USA

\* Correspondence author: [jingjing.yang@emory.edu](mailto:jingjing.yang@emory.edu)

**Abstract**

Transcriptome-wide association study (TWAS) has been widely used to integrate gene expression and genetic data to conduct gene-based association studies. Existing TWAS methods only use genetic data of nearby SNPs (e.g., cis-SNPs within 1MB region) of the target gene as predictors to fit a regression model for gene expression levels, where the predicted values are referred as genetically regulated gene expression (GReX) that are tested for association with the phenotype of interest. Since trans-SNPs (e.g., outside of the 1MB region) of the target gene also explain a significant amount of variation for most expression quantitative traits, taking both cis- and trans- SNPs as predictors is expected to increase the prediction accuracy of GReX and then increase TWAS power. However, this requires enormous computation power for fitting prediction models of genome-wide genes. Here, we propose a novel TWAS approach that accounts for both cis- and trans- genotype data in a Bayesian variable selection regression model and uses summary statistics of eQTL analyses and EM-MCMC algorithm to enable practical usage. Both our simulation and real application studies have shown that our Bayesian approach achieved higher TWAS power by considering genome-wide genotype data for predicting GReX levels. Especially, we identified a novel risk gene *ZC3H12* to be associated with both clinical diagnosis of AD (p-value=) and global brain pathology index of AD (p-value=), and gene *KCTD12* to be associated with brain pathology of -amyloid load (p-value=). A free software for implementing our proposed Bayesian TWAS approach is available on Github.

**Introduction**

Genome-wide association studies (GWAS) have successfully identified thousands of variants associated with complex diseases and traits over the past 10-15 years1-5. However, most of these GWAS hits are located within noncoding regions and the underlying biological mechanisms by which these variants impact a phenotype are still largely unknown6; 7. Recent studies have shown that GWAS associations were enriched for regulatory elements such as expression of quantitative trait loci (eQTL)8-10, suggesting that integrating both genetic and transcriptomic data might help identify key biological mechanisms of complex traits.

One of such integrative methods is Transcriptome-wide association study (TWAS)11-13. Specifically, TWAS first fits a prediction regression model for the expression quantitative trait of the target gene using nearby genetic data (e.g., cis-SNPs within 1MB region) as predictors, and then tests the association between the predicted genetically regulated gene expression (GReX) levels and the phenotype of interest. Generally, the GReX prediction models can be trained using reference data sets that have both profiled transcriptomic and genetic data, such as the Genotype-Tissue Expression (GTEx) project with transcriptomic data for >44 human tissues8, Genetic European Variation in Health and Disease (GEUVADIS) for lymphoblastoid cell lines14, and the North American Brain Expression Consortium (NABEC) for cortex tissues15. Then TWAS can be conducted with both individual-level and summary-level GWAS data.

Essentially, TWAS is equivalent to a burden type gene-based test taking “cis-eQTL effect sizes” that are coefficients of cis-SNPs from the gene expression prediction regression model as their corresponding burden weights11-13. Because of weighting genetic variants by cis-eQTL effect sizes, the phenotypic effects of the risk genes identified by TWAS are potentially mediated through transcriptome variations. For example, TWAS has been shown to prioritize known GWAS risk genes as well as identify novel risk genes for a wide range of complex phenotypes, including schizophrenia, breast cancer, and Alzheimer’s disease (AD)16-20.

However, existing TWAS methods only use genetic data of cis-SNPs of the target gene as predictors to fit gene expression prediction regression models11-13. Since trans-SNPs (e.g., outside of the 1MB region) of the target gene not only explain a significant amount of variation for most expression quantitative traits but also contain significant trans-eQTL with important biological interpretations21; 22, taking both cis- and trans- SNPs as predictors is expected to increase the prediction accuracy of GReX and then increase TWAS power. The biggest challenge would be requiring enormous computation power to fit GReX prediction models for ~20K genome-wide genes with ~10M genome-wide SNPs.

Here, we propose a novel TWAS approach that accounts for both cis- and trans- genotype data in a Bayesian variable selection regression (BVSR) model23 for predicting GReX and uses summary statistics of eQTL analyses and EM-MCMC algorithm24 to enable efficient computation in practice. We demonstrated the viability of this Bayesian approach in simulation studies with varying proportion of true causal cis- and trans-eQTL for gene expression levels. Particularly, we compared our approach to the previously proposed PrediXcan method that fits GReX prediction models using only cis-SNPs in an Elastic-net penalized regression11. We further applied our approach to conduct TWAS for the clinical diagnosis of AD as well as quantitative pathology indices of AD including neurofibrillary tangle density (tangles), -amyloid load (amyloid), and global pathology (gpath, a combination of tangles and amyloid).

Both our simulation and real application studies have shown that our Bayesian approach achieved higher TWAS power by considering the genotype data of both cis- and trans- SNPs for predicting GReX levels. Especially, we identified a novel risk gene *ZC3H12* to be associated with both clinical diagnosis of AD (p-value=) and global brain pathology index of AD (p-value=), and gene *KCTD12* to be associated with -amyloid load (p-value=). Additionally, a free software for implementing our proposed Bayesian TWAS approach is available on Github.

**Method**

TWAS Procedure

The first step in TWAS is to train a prediction model for observed gene expression levels using genotype data as predictors on a per-gene basis11-13. The general linear prediction model can be written as

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where denotes gene-expression levels for a given target gene, centered and adjusted for non-genetic covariates; denotes centered genotype data; denotes the corresponding eQTL effect sizes for the target gene; and is an error term following a ) distribution. With estimated from the training cohort that have both transcriptomic and genetic data profiled and genotype data of the test cohort, TWAS will test the association between the phenotype of interest and the predicted genetically regulated gene expression (GReX) levels that can be estimated by

Bayesian Variable Selection Regression Model

Generally, SNPs within 1MB of the flanking 5’ and 3’ ends (cis-SNPs) are included in the gene expression prediction model (Equation (1)) and the corresponding non-zero estimates will be used for TWAS by existing methods [ref]. In order to leverage the additional information provided by trans-eQTL that are located outside the 1MB flanking of the target gene, we propose to utilize the Bayesian variable selection regression (BVSR)23 model to account for both cis- and trans- SNPs in the gene expression prediction model as

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

The BVSR model assumes a spike-and-slab prior distribution to the effect sizes , that is, the prior on is a mixture distribution of a normal distribution with zero mean and a point-mass density function at 0. Basically, we assume the following respective priors for the coefficients of cis- and trans- SNPs,

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

where denotes the probability that the coefficient is continuous and normally distributed, and is the point mass density function. Further, the following conjugate hyper prior distributions are assumed for the cis- and trans- specific parameters,

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

where *IG* indicates the Inverse Gamma distribution and hyper parameters will be chosen to enable non-informative hyper prior distributions.

To facilitate computation, a latent indicator is assumed such that, if , and follows a normal distribution if . Then the expected value of this indicator, , represents the *posterior inclusion probability* ( for each individual SNP to have non-zero effect sizes. Particularly, the BVSR model enables a novel Bayesian approach to estimate the GReX for test samples that can account for the uncertainty for each SNP to be a broad sense of eQTL with a non-zero effect size:

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

where represents the genotype data of variant for test samples and (, ) denote estimates by BVSR model (Equations (2-4)). This Bayesian GReX estimate can then be used to conduct TWAS by testing the association between phenotype of interest and this Bayesian GReX (Equation (5)).

Efficient Computation Techniques

Although the estimates of eQTL effect sizes and corresponding posterior inclusion probabilities (, ) can be obtained by standard Markov Chain Monte Carlo (MCMC)25 method in theory, the computation burden for modeling genome-wide genotype data is nearly impossible because of enormous memory usage and slow convergence rate. Therefore, we employ novel computational techniques to enable computation efficiency such that this Bayesian TWAS method is practical for accounting for both cis- and trans- eQTL information. Specifically, we employed the previously developed scalable expectation-maximization Markov Chain Monte Carlo (EM-MCMC) algorithm24 to estimate per gene. Unlike the original EM-MCMC algorithm requiring individual-level genotype data, we reduced the computation time by enabling the EM-MCMC algorithm to utilize only summary statistics, including the pre-calculated score statistics from single variant eQTL analyses and LD coefficients. To further reduce the computation burden, we will pre-select a subset of approximately independent genome regions before applying the BVSR model. Particularly, we only consider approximately independent genome regions that contain at least one SNP with p-value for testing based on a single variant regression model, .

ROS/MAP and Mayo Clinic GWAS data of Alzheimer’s Disease

We will apply our proposed Bayesian TWAS method for studying AD, by using the transcriptomic and GWAS data from Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) 26-28 as well as GWAS data from Mayo Clinic Alzheimer’s Disease Genetics Studies (MCADGS)29-31. Specifically, ROS and MAP are prospective cohort studies of aging and dementia, which recruited senior adults without known dementia at enrollment who underwent annual clinical evaluation. Brain autopsy was done at the time of death for each participant. All participants signed an informed consent and Anatomic Gift Act, and the studies were approved by the Institutional Review Board of Rush University Medical Center, Chicago, IL. The Mayo Clinic late-onsite Alzheimer’s Disease GWAS study contains samples from two clinical AD Case-Control series (Mayo Clinic Jacksonville and Mayo Clinic Rochester) as well as a neuropathological series of autopsy-confirmed subjects from the Mayo Clinic Brain Bank. Microarray genotype data generated for 2,093 European-decent from ROS/MAP32 and 2,099 European-decent samples from MCADGS, were further imputed to the 1000 Genome Project Phase 333 in our analysis.

The post-mortem brain samples (gray matter of the dorsolateral prefrontal cortex) from ~30% these ROS/MAP participants with assayed genotype data were profiled for transcriptomic data by next-generation RNA-seqencing34, which were used as reference data to train GReX prediction models. In this paper, we conducted TWAS for clinical diagnosis of late-onsite AD (available for both ROS/MAP and MCADGS) as well as pathology indices of AD (available only for ROS/MAP cohort) quantified with -antibody specific immunostains, including neurofibrillary tangle density (tangles), -amyloid load (amyloid), and global pathology (a combination of tangles and amyloid)26-28. The neurofibrillary tangle density quantifies the average Tau tangle density within two or more 20µm sections from eight brain regions –– hippocampus, entorhinal cortex, midfrontal cortex, inferior temporal, angular gyrus, calcarine cortex, anterior cingulate cortex, and superior frontal cortex. The -amyloid load quantifies the average percent area of cortex occupied by -amyloid protein in adjacent sections from the same eight brain regions.

Simulation Study Design

We conducted simulation studies to validate the performance of our proposed novel Bayesian approach using both cis- and trans- eQTL information for TWAS and compare with the existing PrediXcan method using only cis-eQTL information. Specifically, we used samples with both profiled gene expression and genotype data from the ROS/MAP study as our training data to estimate eQTL effect sizes and the corresponding posterior inclusion probabilities (N=499). The remaining samples from the ROS/MAP study with profiled genotype data only were used as test data. We arbitrarily selected five approximately independent genome blocks, including one “cis-” genotype block and four “trans-” genotype blocks (variants were filtered with minor allele frequency (MAF) > 5% and Hardy-Weinberg *p*-value ). With genotype matrix of the randomly selected causal eQTL, we generated effect-sizes from and then re-scaled the effect-sizes to ensure the targeted gene expression heritability . Gene expression levels were generated by , with . Then we simulated phenotypes by , where was selected with respect to phenotype heritability and .

Particularly, to mimic the complex real genomic architecture of gene expression, we considered two scenarios, one with 5 true causal eQTL representing a very sparse scenario with relatively large effect sizes and the other one with 22 true causal eQTL (~0.1% of all variants considered in the simulation studies) representing a less sparse scenario with relatively small effect sizes. For the scenario with 5 true causal eQTL, we considered three sub-scenarios with respect to how these true causal eQTL distributed over genome blocks, i) all causal eQTL are from the cis-block; ii) two causal eQTL are from the cis-block explaining 70% of the specified while the other three causal eQTL are from the trans-blocks explaining the other 30% of ; iii) all causal QTL are from the trans-block. Similarly, For the scenario with 22 true causal eQTL, we considered three scenarios where the proportions of causal eQTL were 30%, 50%, and 70% from *cis-* genome blocks. We also varied the total expression trait heritability and phenotype heritability in both scenarios, i.e., for the scenario with 5 true causal eQTL and for the scenario with 22 true causal eQTL. Here, different levels of phenotype heritability were chosen to achieve similar levels of power across all scenarios.

In each simulation, we first fitted the BVSR model with both *cis*- and *trans*- genome blocks and the Elastic-Net model used by PrediXcan with only *cis*- genome block using training data. Then we conducted TWAS with predicted *GReX* by both methods. The performance of our Bayesian approach is compared to PrediXcan in terms of of the predicted GReX and power of association studies with test samples. The association study power was calculated as the proportion of 1,000 repeated simulations of each scenario with *p*-value (genome-wide significance threshold for gene-based association studies).

**Results**

Simulation Results

For the scenario with 5 true causal eQTL and various true expression heritability, our simulation studies show that the Bayesian approach based on BVSR model lead to higher test for GReX and TWAS power than the PrediXcan method when a portion of the true causal eQTL are distributed over trans- genome blocks (Figure 1(A, B)). This is because our Bayesian approach leverages both cis- and trans- eQTL information while the PrediXcan method dose not account for trans- eQTL information. As expected, the performance of our Bayesian approach is similar to PrediXcan when all causal eQTL are from the cis- genome block, while the PrediXcan method failed with nearly zero test when all causal eQTL are from trans- genome blocks.

For the scenario with 22 mixed cis- and trans- eQTL, the performance comparison between our Bayesian approach and PrediXcan method became more complicated with respect various true expression heritability (Figure 1(C, D)). Particularly, when , both methods have difficulties to accurately estimate the eQTL effect sizes and the test for GReX is nearly zero. As expression heritability increases, the advantage of modeling both cis- and trans- genotype data by BVSR model arose, leading to higher test for GReX and higher TWAS power. When and 70% of the true causal eQTL are cis-*,* our Bayesian approach had relatively worse performance than the PrediXcan method, which is because the Elastic-Net regression model used by PrediXcan is more suitable than the BVSR model when the true causal eQTL are relatively less sparse with small effect sizes.

Overall, our Bayesian approach based on BVSR model is preferred for the scenario with sparse true eQTL signals that have relatively large effect sizes and are distributed not only within cis- genome region of the target gene. This is due to the properties of BVSR model that is generally preferred for sparse true causal signals and the advantage of leveraging both cis- and trans- genome data to predict the gene expression levels of the target gene.

TWAS of Alzheimer’s Disease Related Phenotypes

Here, we applied our Bayesian approach to conduct TWAS of AD using the GWAS data of ROS/MAP28; 35 and MCADGS29. First, we trained GReX prediction model by applying the BVSR model to samples (N=499) from the ROS/MAP cohort that possess both profiled genotype data and transcriptomic data of postmortem prefrontal cortex tissue. All quantitative gene expression traits were normalized and corrected for Sex, Age at Death, PMI, Study (ROS or MAP), Batch Effects, RIN Scores, and Cell type proportions ("oligodendrocytes", "astrocytes", "microglia", "neurons"). We obtained GReX prediction models for a total of 14,156 genes by our Bayesian approach, while only 6,011 valid prediction models (with at least one cis-SNP with nonzero effect size on gene expression) by PrediXcan. Across all of these 6,011 genes with valid prediction models by PrediXcan, our Bayesian approach had a smaller train value for quantitative gene expression traits than the PrediXcan method for only 855 genes (Supplementary Figure 4).

Second, we imputed GReX values for all remain samples that only possess profiled genotype data in ROS/MAP and MCADGS. Last, we conducted TWAS by testing the association between the GReX values (fitted values of training samples and imputed values for remain samples) and phenotypes of AD clinical diagnosis, quantitative pathology indices of AD including neurofibrillary tangle density (tangles), -amyloid load (amyloid), and global pathology (gpath, a combination of tangles and amyloid).

For the dichotomous phenotype of AD clinical diagnosis, the case/control status was determined by different rules and the available confounding variables are different for ROS/MAP and MCADGS. Specifically, samples from the ROS/MAP cohort have their AD status determined by cognitive consensus diagnosis at the time of death. Individuals were determined to control with no cognitive impairment (control), or case with mild cognitive impairment or AD dementia. Confounding variables that were adjusted for during TWAS with the ROS/MAP cohort include age at death, sex, smoker status, ROS or MAP study, education level, and top three principal components derived from genome-wide genotype data. Whereas, for the Mayo Clinic GWAS samples, cases were determined for samples with a medical history of late onsite AD diagnosis, and available confounding variables only include age, sex, and top three principal components derived from genome-wide genotype data. Therefore, we meta-analyzed these two cohorts for AD clinical diagnosis by applying the inverse-variance weighting (Cooper, Hedges, & Valentine, 2009) method to summary statistics generated by TWAS within each cohort. We also compared the meta-TWAS results by our Bayesian approach to the ones generated by PrediXcan that uses only *cis*-eQTL information for predicting GReX.

Our Bayesian approach identified one significantly associated gene *ZC3H12B* (located on the X chromosome) for AD clinical diagnosis with effect size and p-value . Both within-cohort TWAS obtained positive effect sizes () for ROS/MAP and Mayo Clinic cohorts, with respective p-value , . On the other hand, this gene was missed by the PrediXcan method, because zero effect sizes were estimated by PrediXcan for all cis-SNPs of gene *ZC3H12B* and therefore the corresponding GReX values were not imputed for TWAS. Manhattan plots for our Bayesian approach and PrediXcan are presented in Figure 2A and Supplementary Figure 1A, respectively.

Because the phenotypes of quantitative pathology indexes (tangles, amyloid, and gpath) were only profiled by the ROS/MAP cohort, we only conducted TWAS of these phenotypes using the ROS/MAP samples. Specifically, TWAS were conducted for a total of 1,121, 1,114, and 1,139 ROS/MAP samples with GWAS genotype data and respective phenotype data for tangles, amyloid, and gpath. Confounding variables including age at death, sex, smoker status, ROS or MAP study, education level, and top three principal components derived from genome-wide genotype data were adjusted for in TWAS of these phenotypes.

Interestingly, by our Bayesian approach, the *ZC3H12B* gene was also identified to be significantly associated with the quantitative global pathology index (gpath) with genome-wide significant p-value (Figure 2B), and marginally significantly associated with neurofibrillary tangle density (tangles) with p-value . Gene *KCTD12* was identified to be significantly associated with -amyloid load (amyloid) with p-value . Additionally, gene *RPAP2* is marginally associated with gpath with p-value . Q-Q plots of all TWAS are presented in Supplementary Figure 3.

Particularly, the *ZC3H12B* gene was found to regulate proinflammatory activation of macrophages36 and expressed higher in brain, spinal cord and thymus tissue types comparing to others37. Protein RPAP2 was found as a hub protein in the RNA polymerase-associated module that might contribute to impaired transcription in the AD entorhinal cortex region38.

Insights about eQTL Genetic Architecture

In addition to imputing the GReX values (Equation 5), the posterior inclusion probabilities also provide insights into the genetic architecture of eQTL (i.e., how eQTL are distributed across the whole genome). Note that the posterior inclusion probability obtained from the BVSR model (Equations 2-4) is essentially the expected probability for a SNP to be an eQTL. Therefore, the sum of these posterior inclusion probabilities within cis- or trans- genome regions will represent the expected number of cis- or trans- eQTL. For a total of 14,156 genes with fitted BVSR models, after excluding XX outlier genes with >99 expected eQTL, we obtained the number of expected eQTL as 2.44 (SD = 5.70) across genome-wide regions, 0.25 (SD = 1.24) within cis- genome regions, and 2.48 (SD = 5.49) within trans- genome regions. That is, on overage, 88% of eQTL with non-zero estimated effect sizes by the BVSR model were from *trans-* genome regions with respect to the target gene.

Interestingly, genes with train R2 < 0.05 tend to have the sum of genome-wide posterior inclusion probabilities >5 (Table 2). This is consistent with our simulation studies where the BVSR model dose not out-perform the PrediXcan when there exists 22 true causal eQTL, the true expression heritability <0.1 (test R2<0.05), and >50% of true causal eQTL are from cis- regions (Figure 1C, Table 3). These results are like due to the nature of the BVSR model that is preferred for the scenario with sparse true causal signals (5 as in our simulation studies) and relatively large effect sizes. This also show that TWAS results by our Bayesian method would be more reliable for genes with train R2 > 0.05. Note that a general rule applied by previous studies is to exclude genes with train R2 < 0.05 from TWAS, which is also applied to our application studies of Alzheimer’s Disease related phenotypes.

**Discussion**

The current paper utilized a Bayesian variable selection model to obtain predicted values of genetically regulated gene expression using genome-wide genotypes for application to TWAS. This expands on previous TWAS methods that utilize a small window of *cis* genotypes only to build the *GReX* prediction model. We adapted a BVSR program for genetic analysis that employed an efficient EM-MCMC estimation algorithm (Yang et al., 2017). Our method further utilized single-variant summary statistics to prune the genome-wide genotypes that were considered in predicting *GReX* in order to speed up the algorithm. Results indicated that variants outside of the *cis* region are indeed eQTL that play a large role in predicting *GReX.* Future developments of TWAS approaches stand to benefit from considering genome-wide variants in gene expression training models.

This investigation also indicated that the BVSR may be better-suited for TWAS methods than the Elastic Net. In general, the BVSR method found a small number of *effective* eQTL, as seen in the small sums of the *PPs*. Both the BVSR and Elastic Net tended to keep a fairly large number of variants in the final prediction model, but the BVSR applies weights that reflect selection uncertainty. For many genes, it is expected that there are a small number eQTL with large effect, rather than many eQTL with diffuse effects. The BVSR is better suited for the former case, whereas Elastic Net is better suited for the latter (reference?). Therefore, the BVSR approach may be preferred for a TWAS. This is somewhat speculative, however, as it is not clear from our simulation study how much of the BVSR advantage is due to considering *trans* variants which are true eQTL, and how much is gained by a different estimation method.

The application of our method to Alzheimer’s diagnosis and related phenotypes also highlighted a couple of genes that warrant future investigation. The ZC3H12B gene was strongly associated with AD status in our meta-analysis of the ROS/MAP and Mayo samples, even after false discovery rate correction. This gene was also associated with Global AD Pathology and increased neurofibrillary tangles. The KCTD12 gene (associated with amyloid plaques) and the RPAP2 gene (suggestive association with GPATH) should be the subject of future AD studies.

The current study is not without limitations. Primarily, a massive amount of computing resources is needed to build training models for all available genes using genome-wide markers. Even with the pruning method we applied, a single gene required approximately 8 hours of serial CPU time. However, our algorithm is relatively memory-efficient (3 to 4 GB of virtual memory per gene, on average). If one has access to a few hundred compute threads, training models can be built in a couple of weeks. Much more time would be required if more limited resources were available. Our simulation study also did not consider a case where the Elastic Net could detect genome-wide variants to compare the two prediction models. However, in practice, the Elastic Net is not suited for genome-wide analysis with a large number of variants (reference?). We also detected some inflation in the TWAS when using the BVSR method in our analysis of Alzheimer’s Disease and related phenotypes. Further investigation of the cause of this inflation is warranted.

In conclusion, genome-wide analysis for training *GReX* models is an important next step for TWAS. The BVSR method presented herein provides a framework for considering both *cis* and *trans* variants as potential eQTL. Our study indicates that this consideration can lead to better prediction of *GReX*, thus increasing statistical power in TWAS.

**Table and Figure Legends**

**Supplemental Data**

Supplementary Text and Figures

**Declaration of Interests**

The authors declare no competing interests.

**Acknowledgement**

J.Y. was supported by the startup funding from Department of Human Genetics at Emory University School of Medicine. ROS/MAP study data were provided by the Rush Alzheimer’s Disease Center, Rush University Medical Center, Chicago, IL. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute. The Mayo Clinic GWAS data were provided by the following sources: The Mayo Clinic Alzheimer's Disease Genetic Studies, led by Dr. Nilüfer Ertekin-Taner and Dr. Steven G. Younkin, Mayo Clinic, Jacksonville, FL using samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's Disease Research Center, and the Mayo Clinic Brain Bank. Mayo Clinic GWAS data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation.

**References**

1. Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nature reviews Genetics 6, 95-108.

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

3. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews Genetics 9, 356-369.

4. Nikpay, M., Goel, A., Won, H.H., Hall, L.M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C.P., Hopewell, J.C., et al. (2015). A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nature genetics 47, 1121-1130.

5. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. American journal of human genetics 101, 5-22.

6. Huang, Q. (2015). Genetic study of complex diseases in the post-GWAS era. J Genet Genomics 42, 87-98.

7. Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The Post-GWAS Era: From Association to Function. American journal of human genetics 102, 717-730.

8. GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. Nature 550, 204-213.

9. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS genetics 6, e1000888.

10. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464, 768-772.

11. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nature genetics 47, 1091-1098.

12. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics 48, 245-252.

13. Nagpal, S., Meng, X., Epstein, M.P., Tsoi, L.C., Patrick, M., Gibson, G., De Jager, P.L., Bennett, D.A., Wingo, A.P., Wingo, T.S., et al. (2019). TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. American journal of human genetics 105, 258-266.

14. Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506-511.

15. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J., et al. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS genetics 6, e1000952.

16. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. American journal of human genetics 100, 473-487.

17. Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Schizophrenia Working Group of the Psychiatric Genomics, C., McCarroll, S., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nature genetics 50, 538-548.

18. Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M.K., Shu, X.O., Lu, Y., Cai, Q., et al. (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nature genetics 50, 968-978.

19. Raj, T., Li, Y.I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.C., Ng, B., Gupta, I., Haroutunian, V., et al. (2018). Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. Nature genetics 50, 1584-1592.

20. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nature genetics 51, 592-599.

21. Lloyd-Jones, L.R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M., et al. (2017). The Genetic Architecture of Gene Expression in Peripheral Blood. American journal of human genetics 100, 371.

22. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. 447367.

23. Guan, Y.T., and Stephens, M. (2011). Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems. Annals of Applied Statistics 5, 1780-1815.

24. Yang, J., Fritsche, L.G., Zhou, X., Abecasis, G., and International Age-Related Macular Degeneration Genomics, C. (2017). A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies. American journal of human genetics 101, 404-416.

25. Casella, G. (2001). Empirical Bayes Gibbs sampling. Biostatistics 2, 485-500.

26. Bennett, D.A., Schneider, J.A., Arvanitakis, Z., and Wilson, R.S. (2012). Overview and findings from the religious orders study. Curr Alzheimer Res 9, 628-645.

27. Bennett, D.A., Schneider, J.A., Buchman, A.S., Barnes, L.L., Boyle, P.A., and Wilson, R.S. (2012). Overview and findings from the rush Memory and Aging Project. Curr Alzheimer Res 9, 646-663.

28. Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., and Schneider, J.A. (2018). Religious Orders Study and Rush Memory and Aging Project. J Alzheimers Dis 64, S161-S189.

29. Carrasquillo, M.M., Zou, F., Pankratz, V.S., Wilcox, S.L., Ma, L., Walker, L.P., Younkin, S.G., Younkin, C.S., Younkin, L.H., Bisceglio, G.D., et al. (2009). Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nature genetics 41, 192-198.

30. Zou, F., Chai, H.S., Younkin, C.S., Allen, M., Crook, J., Pankratz, V.S., Carrasquillo, M.M., Rowley, C.N., Nair, A.A., Middha, S., et al. (2012). Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. PLoS genetics 8, e1002707.

31. Allen, M., Carrasquillo, M.M., Funk, C., Heavner, B.D., Zou, F., Younkin, C.S., Burgess, J.D., Chai, H.S., Crook, J., Eddy, J.A., et al. (2016). Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci Data 3, 160089.

32. De Jager, P.L., Shulman, J.M., Chibnik, L.B., Keenan, B.T., Raj, T., Wilson, R.S., Yu, L., Leurgans, S.E., Tran, D., Aubin, C., et al. (2012). A genome-wide scan for common variants affecting the rate of age-related cognitive decline. Neurobiol Aging 33, 1017 e1011-1015.

33. 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.

34. De Jager, P.L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L.C., Yu, L., Eaton, M.L., Keenan, B.T., Ernst, J., McCabe, C., et al. (2014). Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci 17, 1156-1163.

35. De Jager, P.L., Ma, Y., McCabe, C., Xu, J., Vardarajan, B.N., Felsky, D., Klein, H.U., White, C.C., Peters, M.A., Lodgson, B., et al. (2018). A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci Data 5, 180142.

36. Liang, J., Wang, J., Azfer, A., Song, W., Tromp, G., Kolattukudy, P.E., and Fu, M. (2008). A novel CCCH-zinc finger protein family regulates proinflammatory activation of macrophages. J Biol Chem 283, 6337-6346.

37. Fagerberg, L., Hallstrom, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., et al. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics 13, 397-406.

38. Kikuchi, M., Ogishima, S., Miyamoto, T., Miyashita, A., Kuwano, R., Nakaya, J., and Tanaka, H. (2013). Identification of unstable network modules reveals disease modules associated with the progression of Alzheimer's disease. PloS one 8, e76162.