# Novel Insight into the Aetiology of Autism Spectrum Disorder Gained by Integrating Expression Data with Genome-wide Association Statistics

## *Supplement 1*

## Supplementary Text

### Datasets

*PGC + iPSYCH ASD GWAS summary statistics*

The PGC + iPSYCH ASD GWAS summary statistics are publicly available and were downloaded from: https://www.med.unc.edu/pgc/results-and-downloads. The downloaded ASD GWAS summary statistics contained 7,583,402 variants. Additional quality control and processing was performed, including nomenclature and genome build synchronised to the 1000 genomes reference dataset, exclusion of insertion/deletions, and exclusion of strand ambiguous and variants with an INFO score of less than 0.8. A final dataset of 6,038,890 single nucleotide polymorphisms (SNPs) was carried forward for analysis.

*Gene expression datasets*

The 16 expression SNP-weight sets are listed in Table 1, including gene expression SNP-weights for fetal brain tissue, and brain, blood and adipose tissue in adults. Features were derived using data collected from the dorsolateral prefrontal cortex of 621 individuals collected by the CommonMind Consortium (CMC) (1), peripheral blood from 1,245 individuals from the Netherlands Twin Registry (NTR) (2), blood from 1,264 individuals within the Young Finns Study (YFS) (3, 4), adipose tissue from 563 individuals from the Metabolic Syndrome in Men study (METSIM) (5), 9 specific brain regions in 81-103 individuals in the Genotype-Tissue Expression project (GTEx) (6), and brain homogenates from 67 genetically-determined European fetuses aged 12-19 weeks post-conception which were collected through the Human Developmental Biology Resource (HDBR) (7). Features used in this TWAS primarily represent gene expression, but additional transcript specific expression features

are available in the O'Brien fetal brain dataset, and features representing splice events are available in the CMC dorsolateral prefrontal cortex dataset.

CMC, NTR, YFS, METSIM and GTEx SNP-weights were downloaded directly from the FUSION/TWAS website (see URLs). Information regarding the analysis of genotypes and gene expression from these datasets has been previously described previously: CMC (8), NTR, YFS, METSIM (9), GTEx (6), O'Brien fetal brain (7).

From the processed genotype and gene expression data, TWAS predictors were computed using FUSION software (see URLs). In brief, the FUSION software first estimates the cis-SNP-heritability of each feature based on SNPs +/-500kb from the feature boundary using the AI-REML algorithm in GCTA (10). For features that have nominally significant cis-SNP-heritability ($p < 0.01$), predictive models are generated using BLUP, elastic net, LASSO and BSLMM (CMC, YFS, NTR, METSIM only) models. Five-fold cross validation is then used to evaluate the out-of-sample variance explained by each model with the best model being used in the TWAS.

## TWAS

*Defining transcriptome-wide significance*

Many genes are available in multiple SNP-weight sets and have highly correlated predicted expression. Furthermore, genes near one another often have correlated expression and are therefore not independent. Therefore, a Bonferroni significance threshold for all features (gene/tissue combinations) would be highly conservative. We estimated transcriptome-wide significance using a permutation procedure, which accounts for the correlation between features, within and across SNP-weight sets. Initially, expression levels for all features from all SNP-weight sets (N features = 38,157) were imputed into the 1000 genomes reference dataset used by FUSION (N individuals = 489) using the FUSION protocol which involves PLINK (11) (see URLs). Then, for each permutation, a random normally distributed phenotype was generated, linear regression was performed to derive a *p*-value of association for each feature, and the minimum *p*-value was stored. This procedure was repeated 1,000 times. The

5% quantile of the minimum *p*–values is the transcriptome-wide significance threshold with the features used in this study. Based on these permutations the transcriptome-wide significance threshold was estimated at $p = 4.25\times10^{-6}$ (95% CI = $2.86\times10^{-6}$ – $5.53\times10^{-6}$). This approach for estimating transcriptome-wide significance is an adaptation of the permutation procedure used, in part, to estimate the genome-wide significance threshold (12).

*Colocalisation*

This method uses a Bayesian framework to estimate the posterior probability of five models: Model 0 = No association with either ASD or gene expression, Model 1 = Association with ASD only, Model 2 = Association with gene expression only, Model 3 = Association with ASD and gene expression, but from two independent SNPs, and Model 4 = Association with ASD and gene expression at a common SNP.

*Calculating proportion of SNP association explained by predicted expression*

The proportion of a SNP-level association accounted for by predicted expression in the TWAS was calculated as 1-($\chi$2 of conditioned GWAS association) / ($\chi$2 of unconditioned GWAS association). This is the same method used by TWAS-hub to calculate '% variance explained' (http://twas-hub.org).

*Similarities and differences between TWAS and MAGMA*

MAGMA and TWAS both aggregate SNP associations within gene regions. However, a key difference is that MAGMA aggregates the association of SNPs within gene regions without taking into account of SNP effects on gene expression. MAGMA is therefore considered functionally agnostic. In contrast, TWAS aggregates the association of SNPs within gene regions weighted by their effect on gene expression, so comparison of TWAS to MAGMA highlights the effect of considering SNP-effects on gene expression. Another important difference is that MAGMA includes any gene region for which SNPs are available in the GWAS and linkage disequilibrium (LD) reference, whereas TWAS only includes genes with significantly heritable gene expression (based on SNPs within a 500kb window).

## TWAS-based enrichment analysis

*Analytical procedure*

Competitive enrichment was tested for by performing a linear regression for each gene-set, whereby gene Z-scores were predicted by membership of each gene-set, including covariates for gene length and the number SNPs within the gene region. Given the functional consequence of each genes up- or down-regulation is unknown, Z-scores were calculated as probit transformed (1-*p*), resulting in an approximately normally distributed Z-score of non-zero association. To avoid potential bias due to outliers, Z-scores were truncated to be between -3 and 6. This regression approach for enrichment analysis can also be used to test for a correlation between TWAS associations and continuous gene annotations, termed gene property analysis, as is also implemented in MAGMA.

To avoid bias due to the correlation between genes we use lme4qtl (13) to fit a mixed model regression of TWAS Z-score on gene-set membership, accounting for the correlation in Z-scores between genes due to LD. The correlation matrix used was computed based on the same predicted gene expression values used when estimating the transcriptome wide-significance threshold. The correlation between genes that were more than 5Mb apart or on separate chromosomes were set to zero. Any gene-gene correlations with an R-squared less than 0.0001 were set to zero. The matrix was stored as a sparse matrix substantially reducing the memory requirements and duration of the analysis. The linear mixed model with the sparse matrix was performed using the lme4qtl package in R. The software used for this analysis (TWAS-GSEA) is publically available (see URLs).

We analysed TWAS association results from all 16 SNP-weight sets simultaneously to improve genome coverage and reduce the multiple testing burden. If multiple features represent the same gene, such as when a gene is captured in multiple SNP-weight sets, only the feature that gave the best prediction of expression (as measured by cross-validated R2) was retained.

For gene set analysis, the feature IDs were converted to entrez IDs using the biomaRt package in R, matching based on the ‘external_gene_name’ variable. Of the unique TWAS features, 11,470 had entrez IDs and could be included in the analysis. The MAGMA gene-set analysis was performed using the

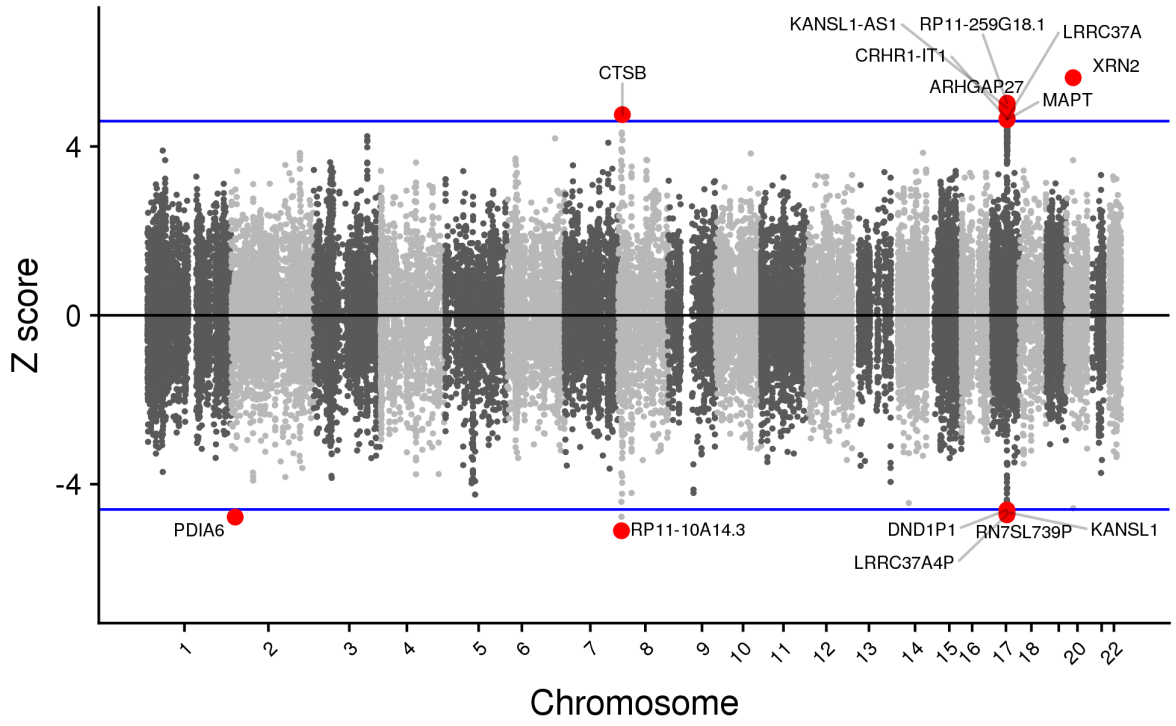
MAGMA derived gene-level associations and default settings. 11,424 genes in the MAGMA analysis had entrez IDs available and could be included in the gene-set analyses.

For gene property analysis, of the unique TWAS features, 8,699 were present in the BRAINSPAN preferential expression dataset. For comparative gene property analysis in MAGMA, 8,685 genes in the MAGMA gene analysis were present in the BRAINSPAN preferential expression dataset and were included in the analysis.

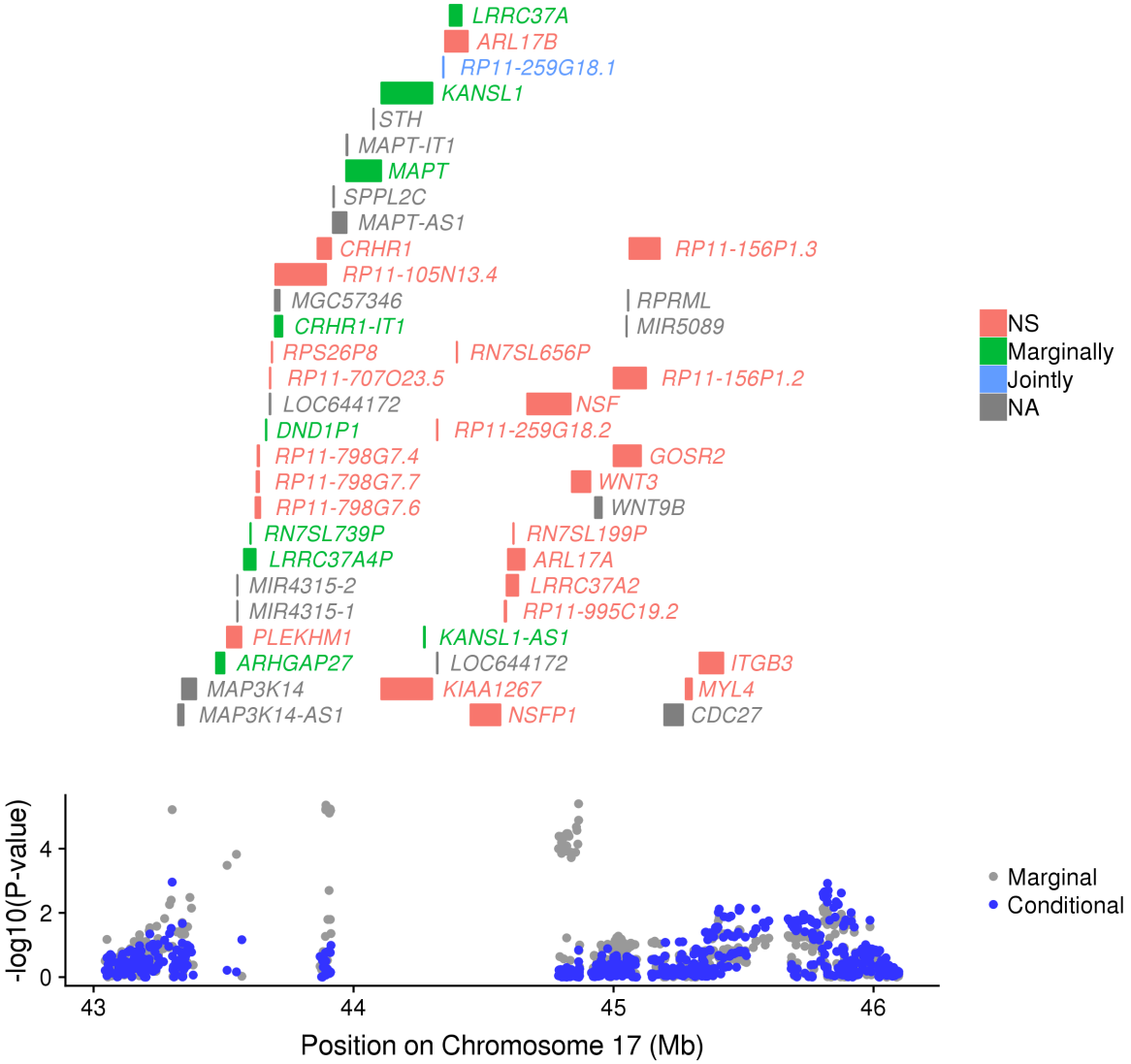
*Interpretation of S-LDSC TWAS-based heritability estimates*

Estimates of variance explained by each SNP-weight set should not be interpreted as a measure of enrichment as they are highly correlated with the sample size of the dataset used to derive the SNP-weights. SNP-weights are only available for a gene if the gene's expression has a statistically significant SNP-heritability. Larger samples often have greater power to detect significantly SNP-heritable expression, and therefore SNP-weight sets derived from larger samples typically include SNP-weights for more genes, which leads to an increased variance explained by the SNP-weight set.

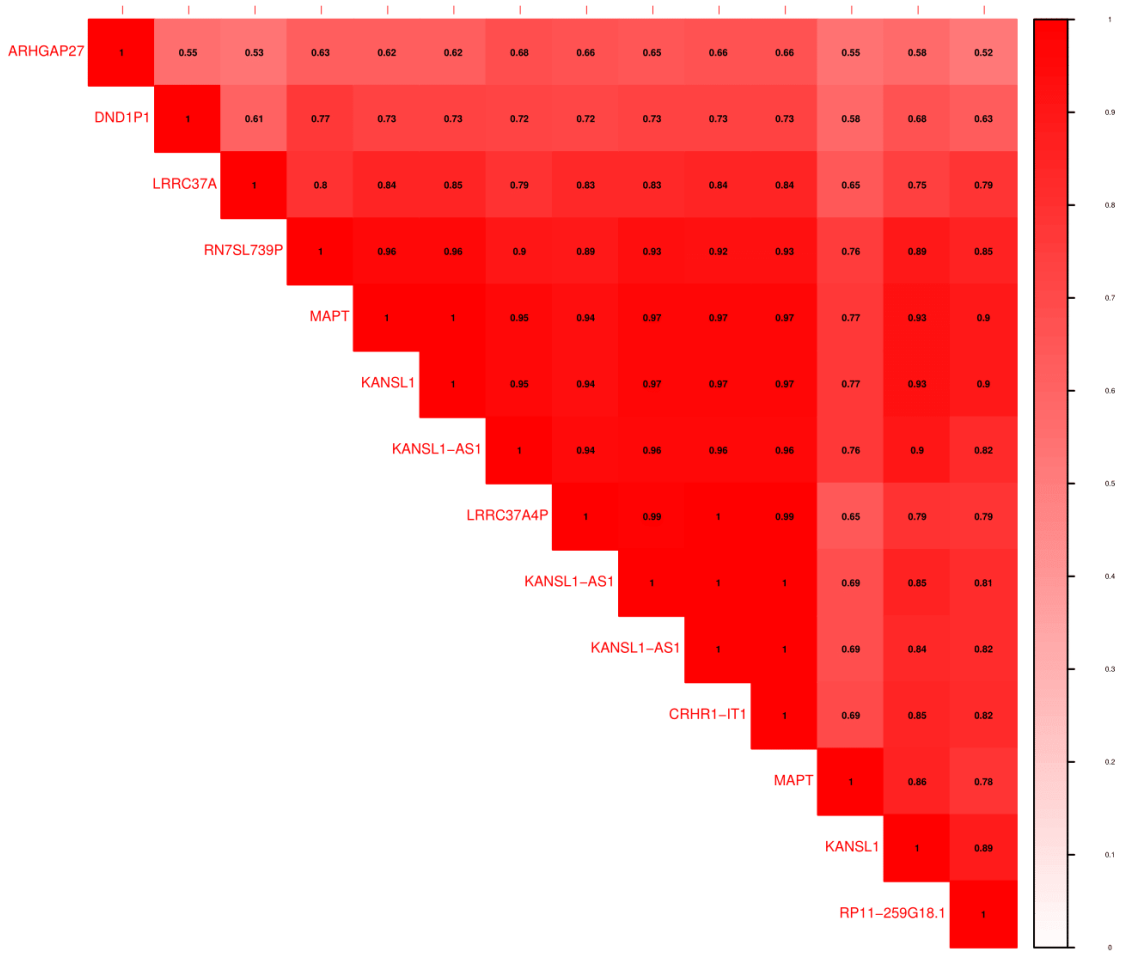
## Supplementary Figures



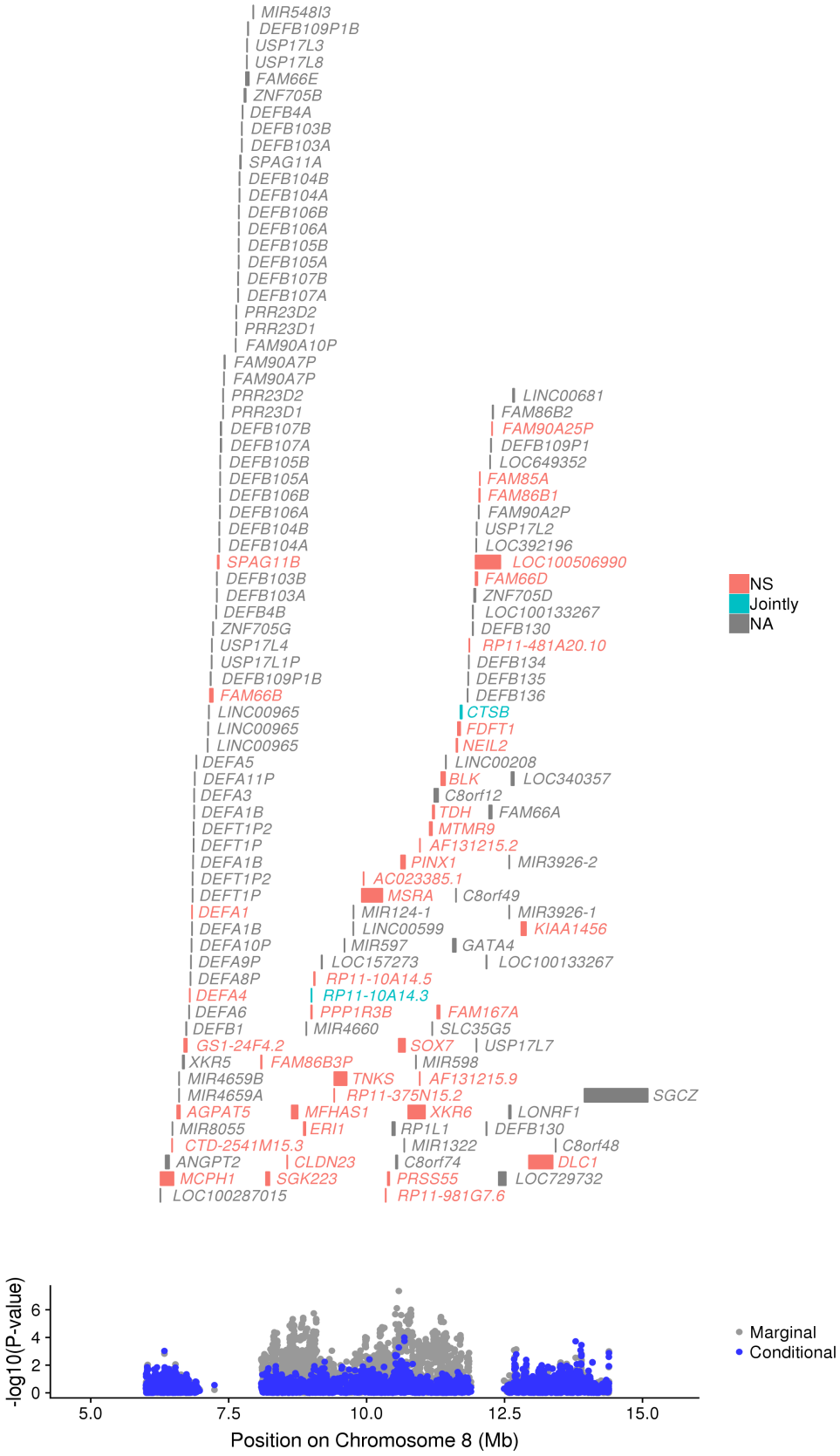

Supplementary Figure S1. Manhattan plot of all ASD TWAS associations. Each point represents a single gene tested, with physical position plotted on the x-axis and Z score of association between the gene and ASD plotted on the y-axis. Transcriptome-wide significant associations are highlighted as red points and are labelled with their ID. If more than one transcriptome-wide significant feature represents the same gene, only the most significant feature is highlighted in red and labelled. The blue horizontal line indicates transcriptome-wide significance ($p < 4.25 \times 10^{-6}$).



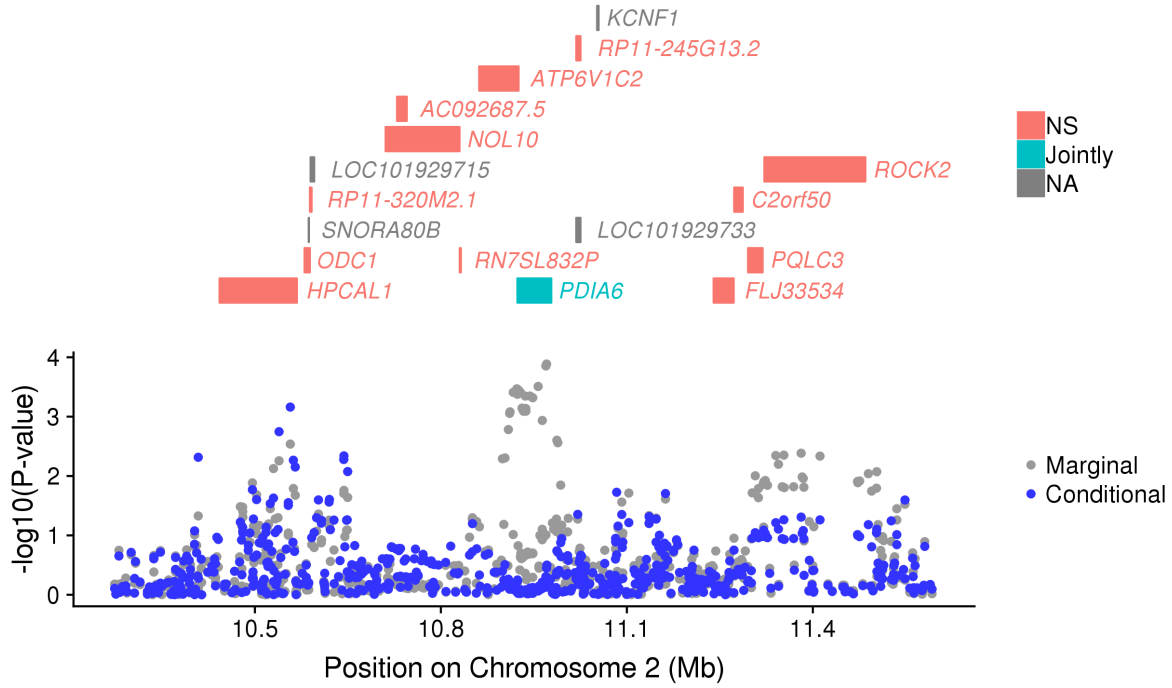

Supplementary Figure S2. Regional association plot. The top panel shows all of the genes in the locus. Marginally TWAS associated genes are highlighted in green, jointly significant genes are highlighted in blue, non-significant genes are in red, and genes that were not assessed in the TWAS are in grey. The bottom panel shows a Manhattan plot of the GWAS data before (grey) and after (blue) conditioning on the green genes.



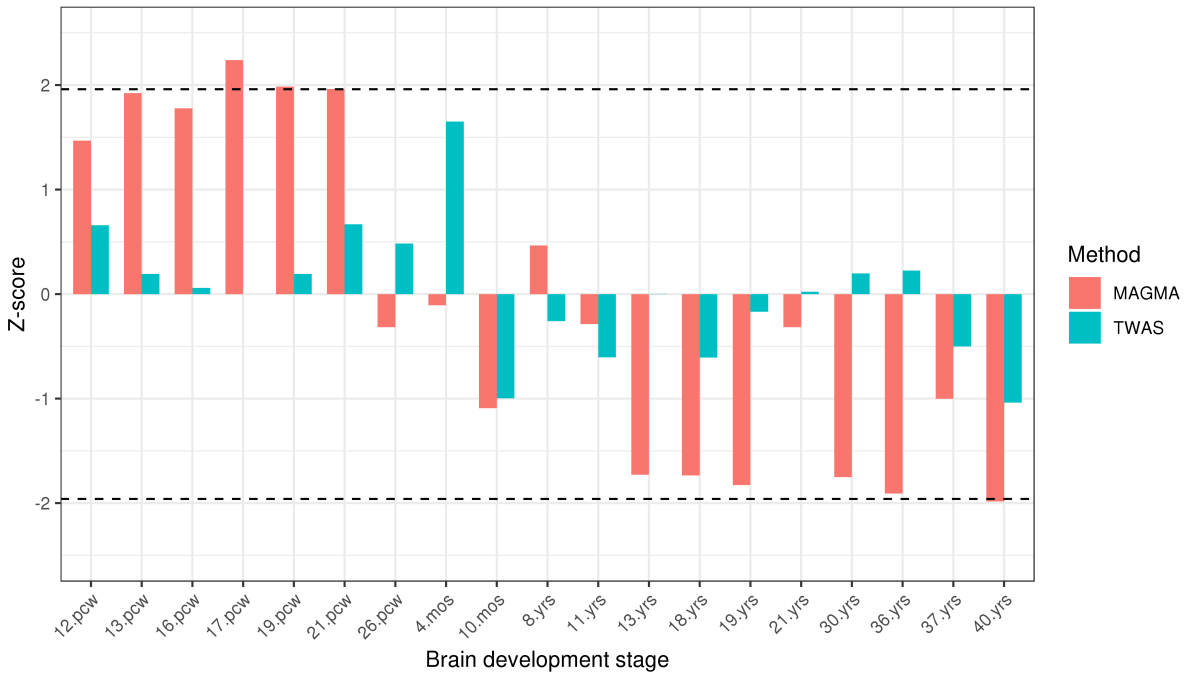

Supplementary Figure S3. Correlations between transcriptome-wide significant genes on chromosome 17.



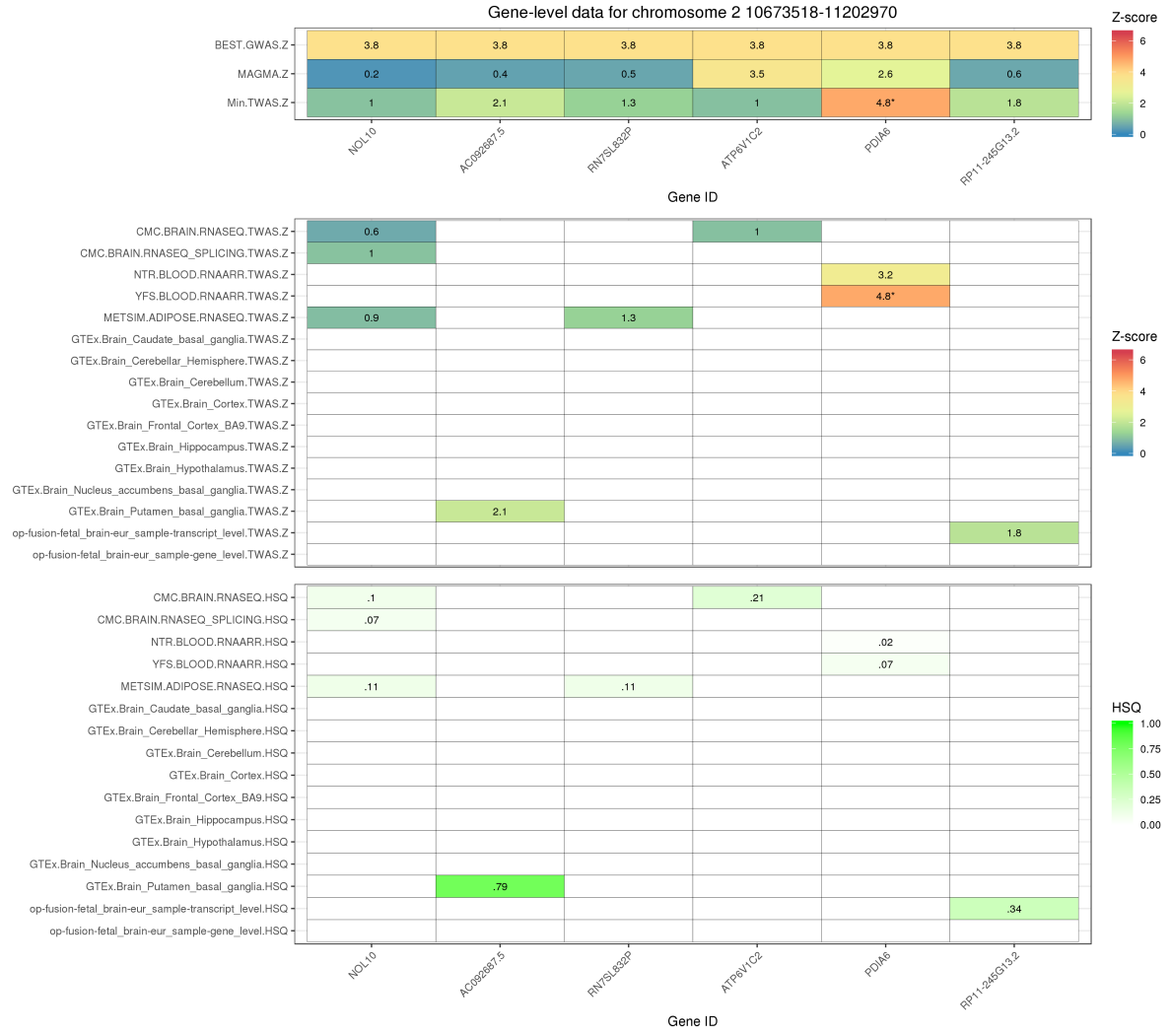

Supplementary Figure S4. Regional association plot. The top panel shows all of the genes in the locus. Marginally TWAS associated genes are highlighted in green, jointly significant genes are highlighted in blue, non-significant genes are in red, and genes that were not assessed in the TWAS are in grey. The bottom panel shows a Manhattan plot of the GWAS data before (grey) and after (blue) conditioning on the green genes.



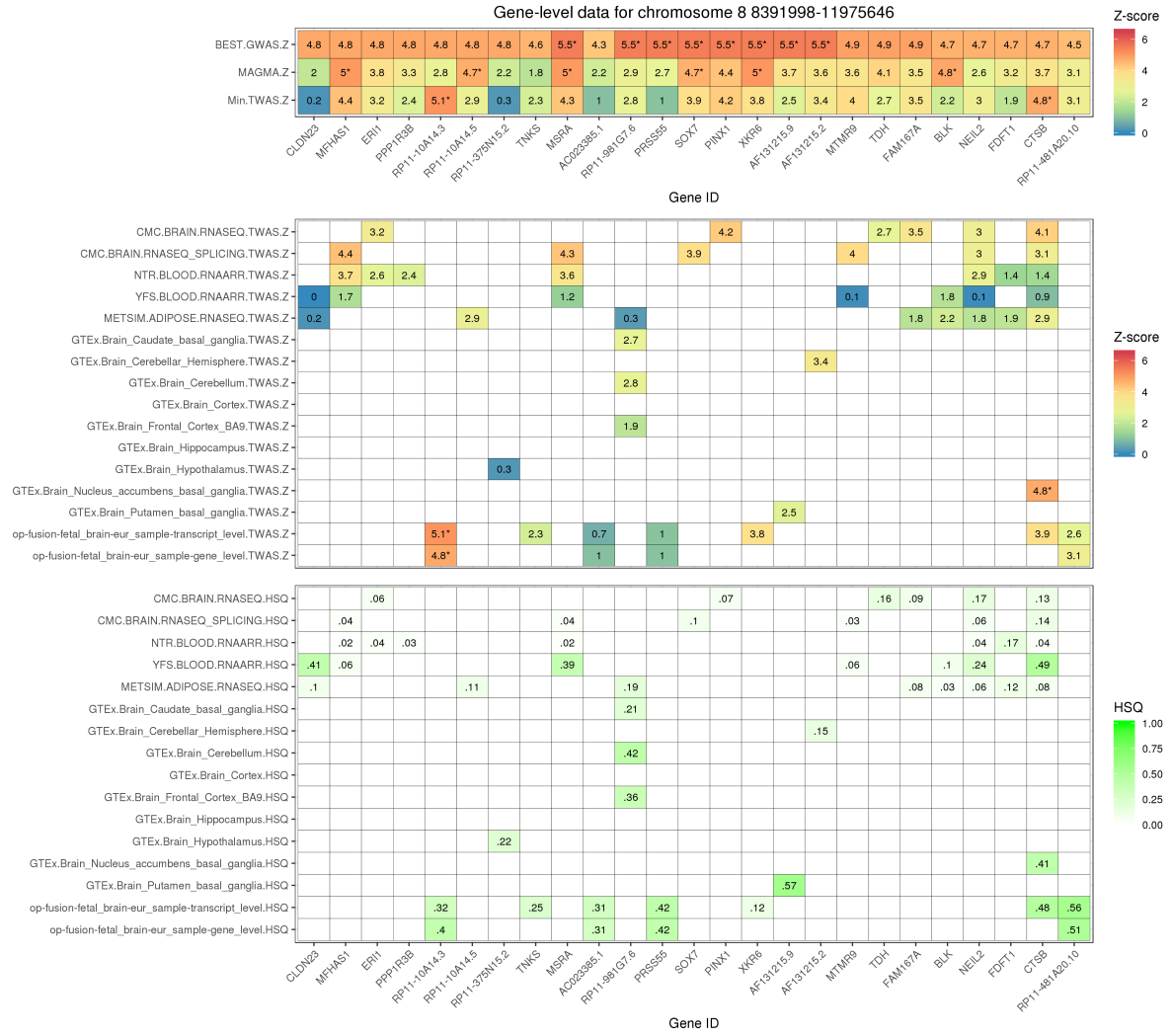

Supplementary Figure S5. Regional association plot. The top panel shows all of the genes in the locus. Marginally TWAS associated genes are highlighted in green, jointly significant genes are highlighted in blue, non-significant genes are in red, and genes that were not assessed in the TWAS are in grey. The bottom panel shows a Manhattan plot of the GWAS data before (grey) and after (blue) conditioning on the green genes.




Supplementary Figure S6. Gene property analysis of preferential gene expression across brain development. Results based on TWAS derived Z scores and MAGMA derived Z-scores are shown. The dashed lines indicate nominal significance (p=0.05).

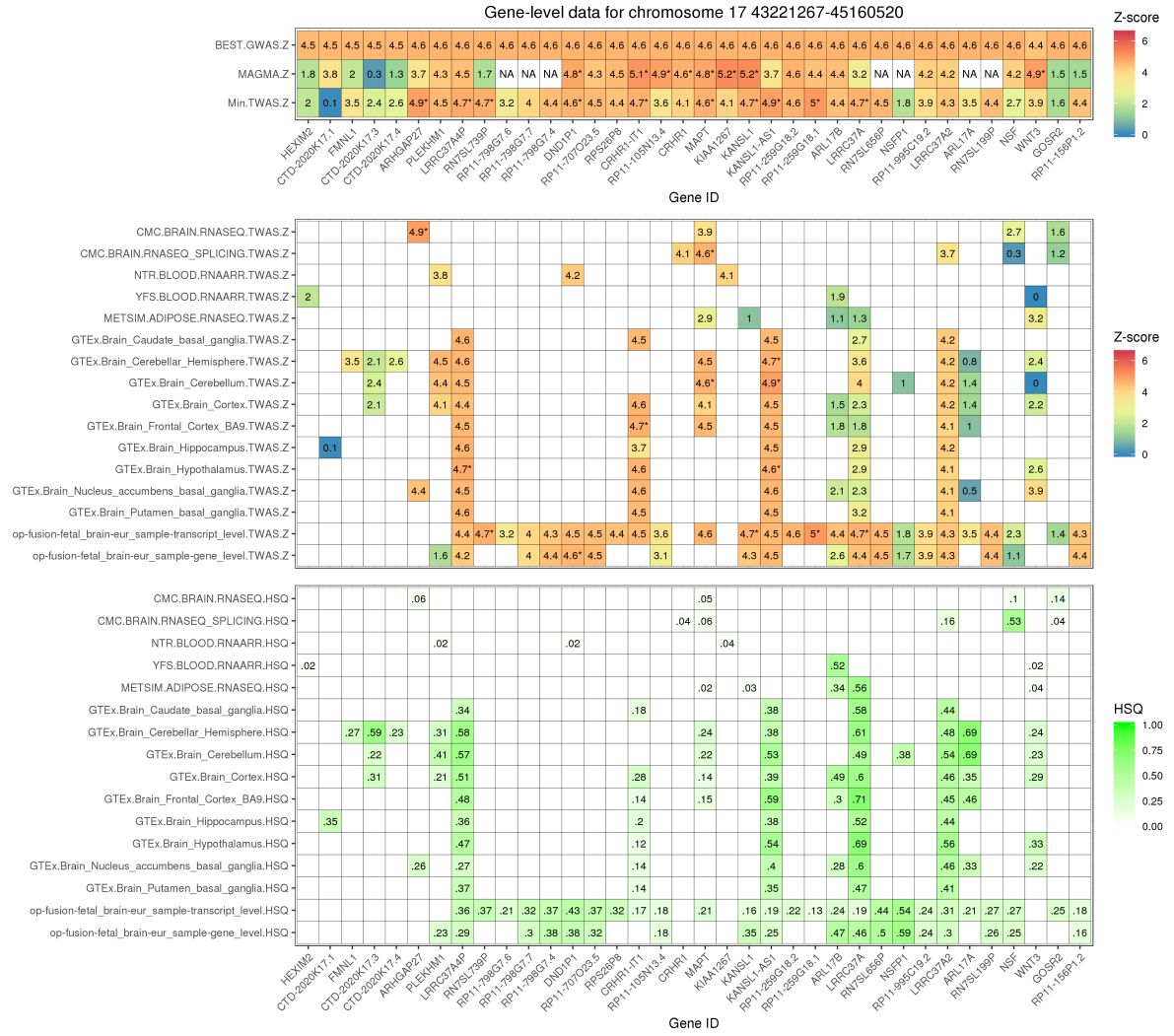



Supplementary Figure S7. Comparison of gene-level Z scores derived using MAGMA and TWAS, and SNP-level Z score in the corresponding GWAS, containing either a significant TWAS or MAGMA association. The absolute TWAS Z score is used here. Genes within 250kb of genes significant in the either the TWAS of MAGMA analysis are included. Empty cells indicate the gene was not available in the TWAS or MAGMA analysis. Asterisks indicate the Z score surpassed the corresponding significance threshold (TWAS = $p < 4.25x10^{-6}$; MAGMA = Bonferroni p-value </= 0.05, GWAS = $p <= 5x10^{-8}$).

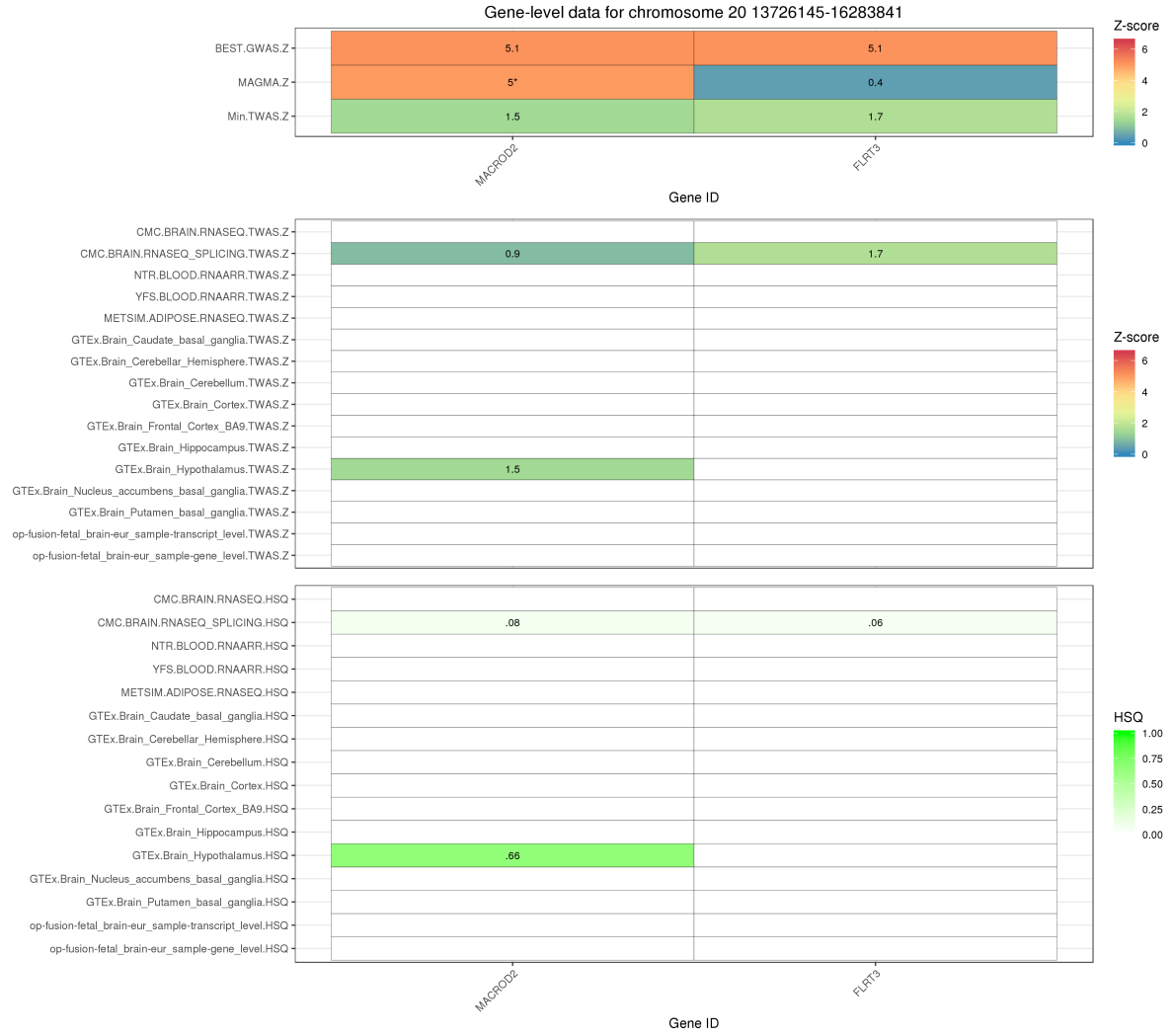



Supplementary Figure S8. Comparison of gene-level Z scores derived using MAGMA and TWAS, and SNP-level Z score in the corresponding GWAS, containing either a significant TWAS or MAGMA association. The absolute TWAS Z score is used here. Genes within 250kb of genes significant in the either the TWAS of MAGMA analysis are included. Empty cells indicate the gene was not available in the TWAS or MAGMA analysis. Asterisks indicate the Z score surpassed the corresponding significance threshold (TWAS = $p < 4.25x10^{-6}$; MAGMA = Bonferroni p-value </= 0.05, GWAS = $p <= 5x10^{-8}$).

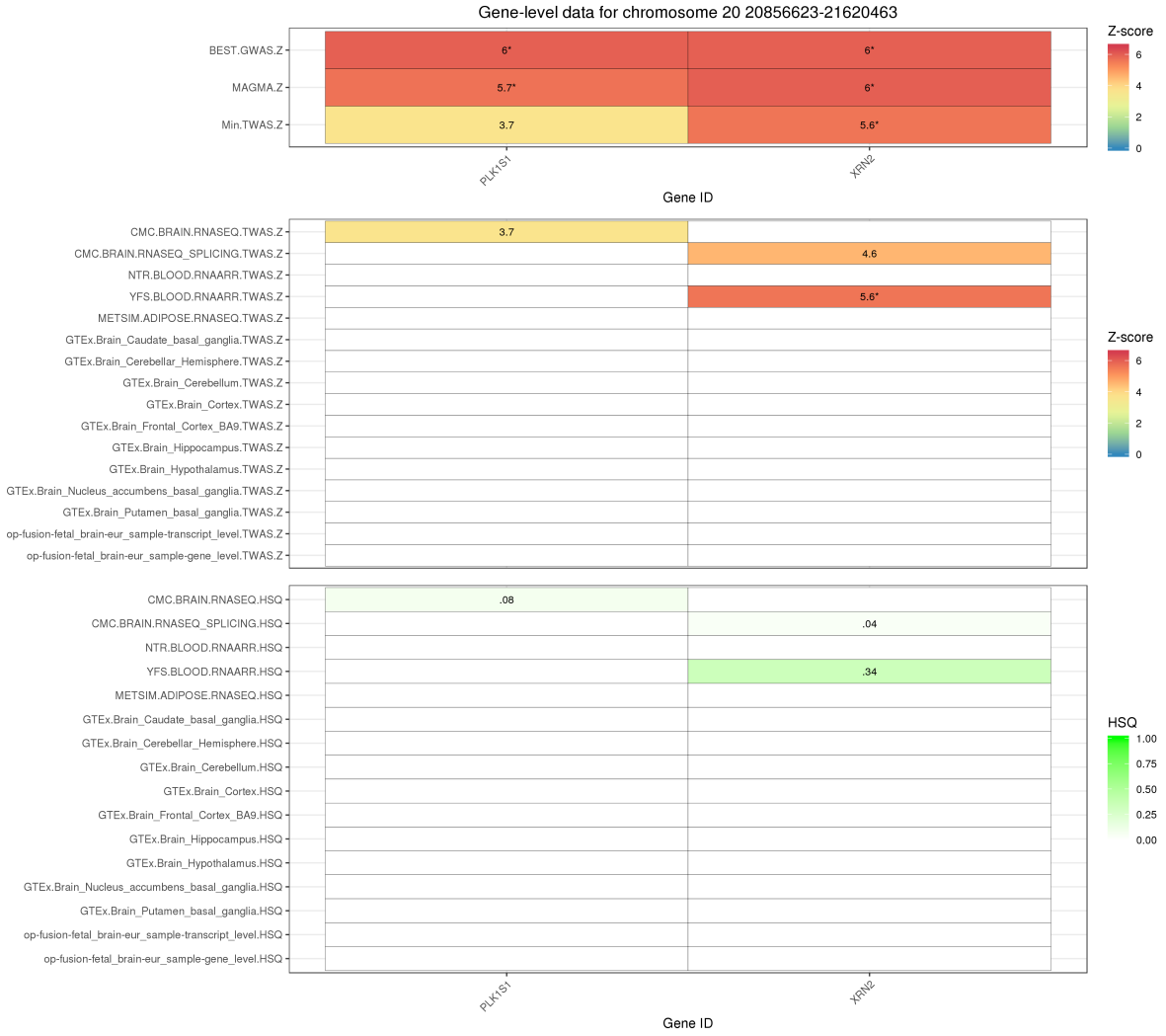

Supplementary Figure S9. Comparison of gene-level Z scores derived using MAGMA and TWAS, and SNP-level Z score in the corresponding GWAS, containing either a significant TWAS or MAGMA association. The absolute TWAS Z score is used here. Genes within 250kb of genes significant in the either the TWAS of MAGMA analysis are included. Empty cells indicate the gene was not available in the TWAS or MAGMA analysis. Asterisks indicate the Z score surpassed the corresponding significance threshold (TWAS = $p < 4.25x10^{-6}$; MAGMA = Bonferroni p-value </= 0.05, GWAS = $p <= 5x10^{-8}$).




Supplementary Figure S10. Comparison of gene-level Z scores derived using MAGMA and TWAS, and SNP-level Z score in the corresponding GWAS, containing either a significant TWAS or MAGMA association. The absolute TWAS Z score is used here. Genes within 250kb of genes significant in the either the TWAS of MAGMA analysis are included. Empty cells indicate the gene was not available in the TWAS or MAGMA analysis. Asterisks indicate the Z score surpassed the corresponding significance threshold (TWAS = $p < 4.25x10^{-6}$; MAGMA = Bonferroni p-value </= 0.05, GWAS = $p <= 5x10^{-8}$).




Supplementary Figure S11. Comparison of gene-level Z scores derived using MAGMA and TWAS, and SNP-level Z score in the corresponding GWAS, containing either a significant TWAS or MAGMA association. The absolute TWAS Z score is used here. Genes within 250kb of genes significant in the either the TWAS of MAGMA analysis are included. Empty cells indicate the gene was not available in the TWAS or MAGMA analysis. Asterisks indicate the Z score surpassed the corresponding significance threshold (TWAS = $p < 4.25x10^{-6}$; MAGMA = Bonferroni p-value </= 0.05, GWAS = $p <= 5x10^{-8}$).




Supplementary Figure S12. Comparison of gene-level Z scores derived using MAGMA and TWAS, and SNP-level Z score in the corresponding GWAS, containing either a significant TWAS or MAGMA association. The absolute TWAS Z score is used here. Genes within 250kb of genes significant in the either the TWAS of MAGMA analysis are included. Empty cells indicate the gene was not available in the TWAS or MAGMA analysis. Asterisks indicate the Z score surpassed the corresponding significance threshold (TWAS = $p < 4.25x10^{-6}$; MAGMA = Bonferroni p-value </= 0.05, GWAS = $p <= 5x10^{-8}$).

## Supplemental References

1. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, *et al.* (2016): Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci*. 19: 1442.

2. Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, *et al.* (2014): Heritability and genomics of gene expression in peripheral blood. *Nat Genet*. 46: 430.

3. Nuotio J, Oikonen M, Magnussen CG, Jokinen E, Laitinen T, Hutri-Kähönen N, *et al.* (2014): Cardiovascular risk factors in 2011 and secular trends since 2007: the Cardiovascular Risk in Young Finns Study. *Scand J Public Health*. 42: 563–571.

4. Raitakari OT, Juonala M, Rönnemaa T, Keltikangas-Järvinen L, Räsänen L, Pietikäinen M, *et al.* (2008): Cohort profile: the cardiovascular risk in Young Finns Study. *Int J Epidemiol*. 37: 1220–1226.

5. Laakso M, Kuusisto J, Stancakova A, Kuulasmaa T, Pajukanta P, Lusis AJ, *et al.* (2017): METabolic Syndrome In Men (METSIM) Study: a resource for studies of metabolic and cardiovascular diseases. *J Lipid Res*. jlr-O072629.

6. GTEx Consortium (2015): The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (80- )*. 348: 648–660.

7. O'Brien HE, Hannon E, Hill M., Toste C., Robertson MJ, Morgan JE, *et al.* (2018): Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biol*. 19: 194.

8. Gusev A, Mancuso N, Won H, Kousi M, Finucane HK, Reshef Y, *et al.* (2018): Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat Genet*. 50: 538.

9. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, *et al.* (2016): Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 48: 245.

10. Yang J, Lee SH, Goddard ME, Visscher PM (2011): GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 88: 76–82.

11. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015): Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 4: 1.

12. Dudbridge F, Gusnanto A (2008): Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol*. 32: 227–234.

13. Ziyatdinov A, Vázquez-Santiago M, Brunel H, Martinez-Perez A, Aschard H, Soria JM (2018): lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics*. 19: 68.