Transcriptome-Wide Association Study (TWAS)

Lesson Outline

Lecture (2.30pm-3pm)

- Introduction to QTLs
- TWAS methods and considerations

Practical (3pm – 3.30pm)

- TWAS Hub

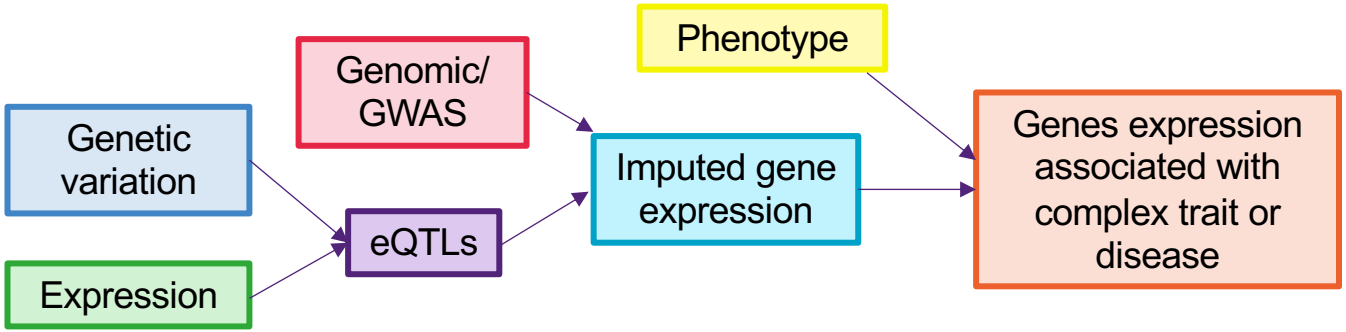
Li, B. & Ritchie, M.D. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Frontiers in Genetics* **12**(2021).
Wainberg, M., Sinnott-Armstrong, N., Mancuso, N. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* **51,** 592–599 (2019). https://doi.org/10.1038/s41588-019-0385-z

Transcriptome-Wide Association Study (TWAS)

Gene-based association approach that investigates associations between genetically regulated gene expression and complex diseases or traits.

Hypothesis: One or multiple eQTLs collectively regulate the transcriptional activities of a gene, and the genetically altered gene expression levels result in modulated disease risk.

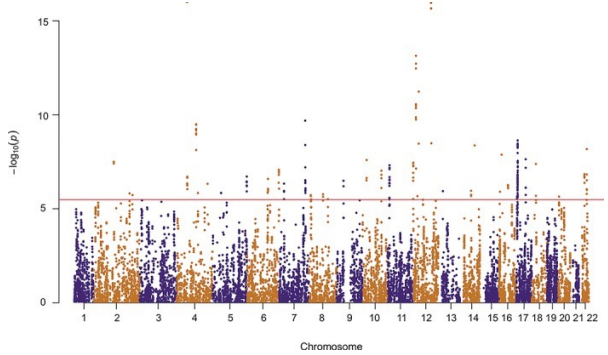
Relatively new approach ~2015

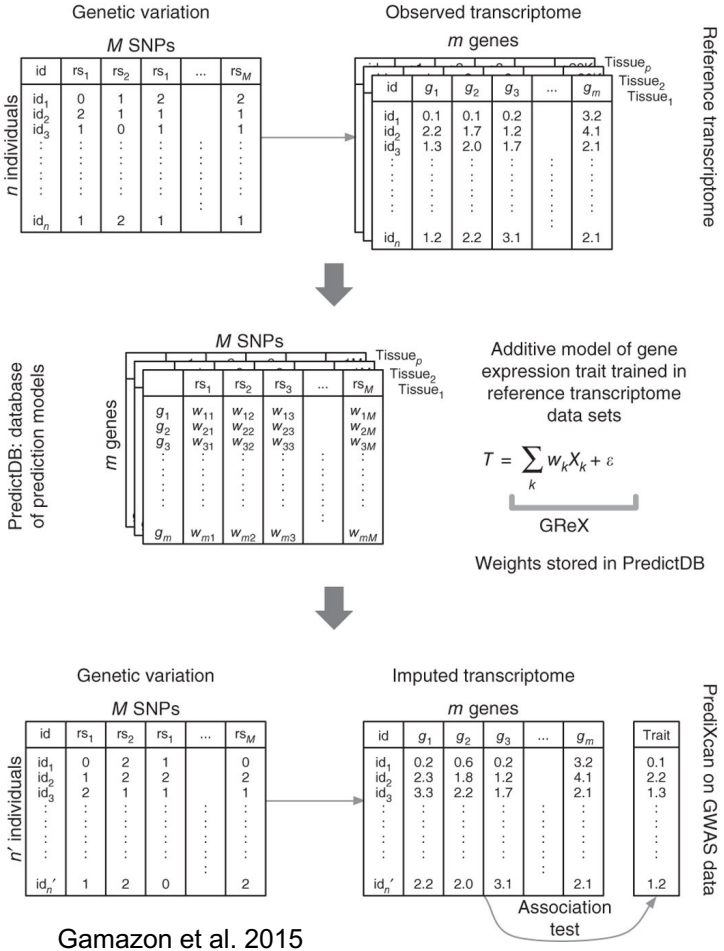


Two step analysis

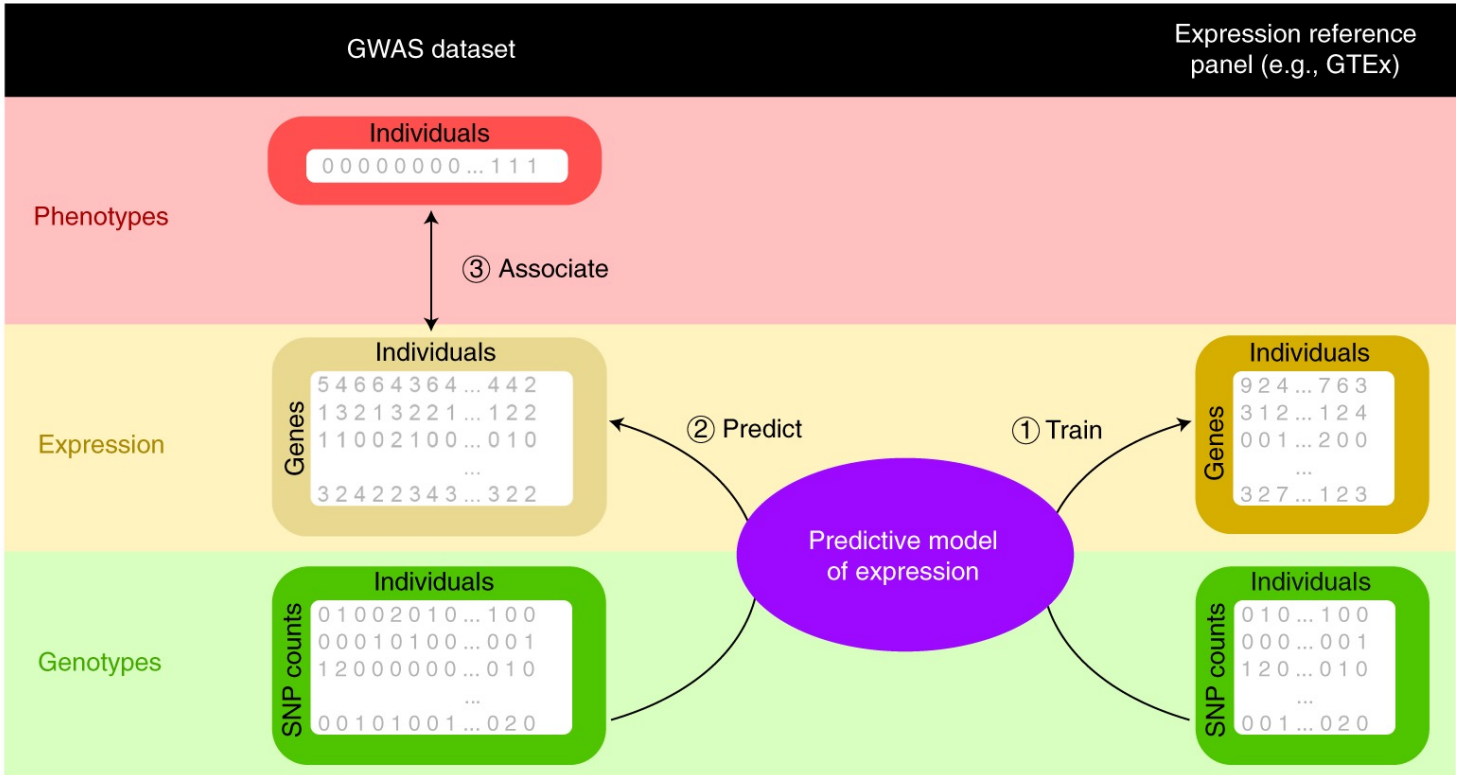
Step 1: Impute genetically regulated gene expression levels by combining regulatory effects of the eQTLs for a gene under an additive genetic model.

Step 2: Test association between imputed gene expression levels from step 1 with a disease phenotype of interest to estimate the statistical significance of each gene-disease association.



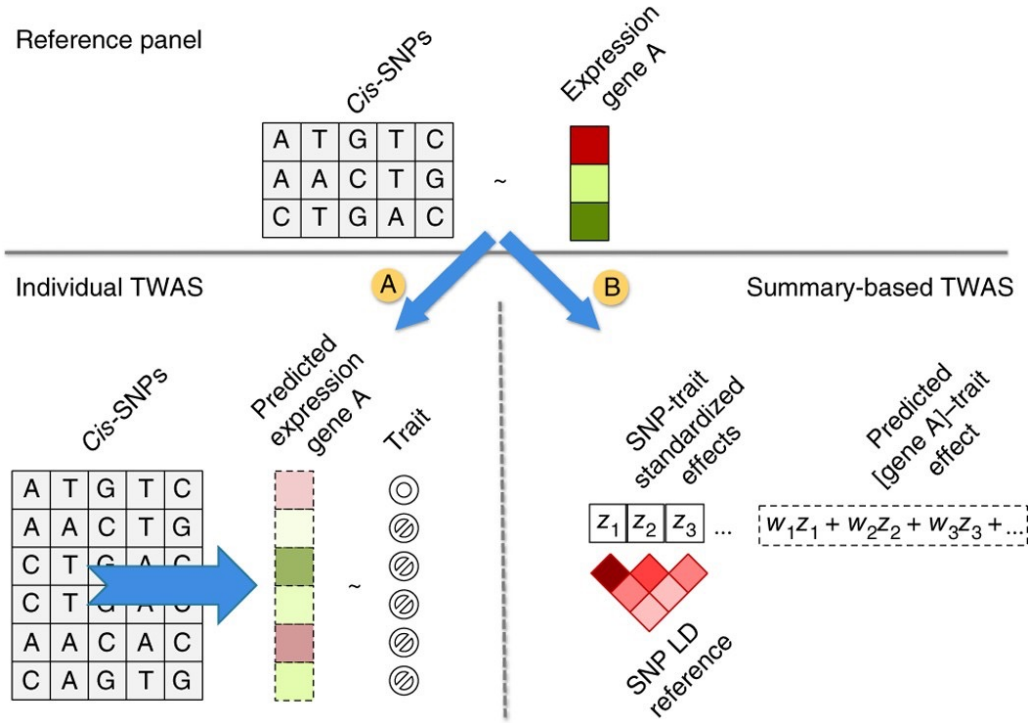


Gamazon et al. 2015

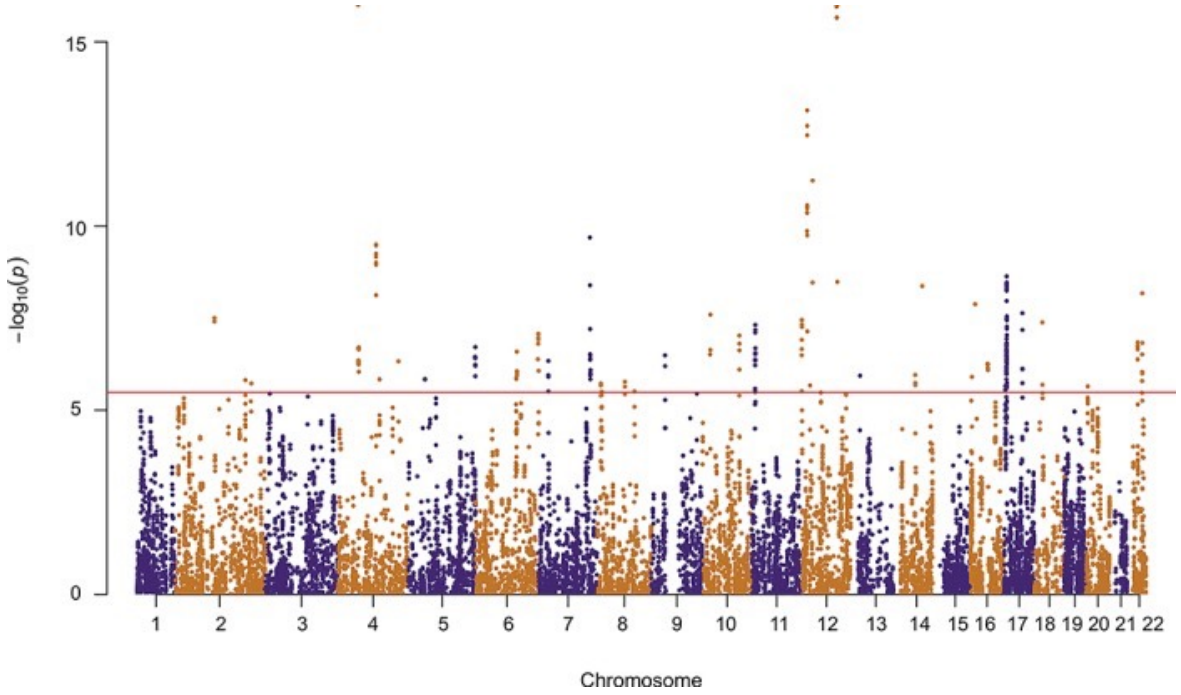


Wainberg et al. 2019

Individual-Level Data-Based TWAS Versus GWAS Summary Statistics-Based TWAS



Endometriosis TWAS



Endometrial gene expression and genotype data from the 206 samples was used to estimate the weighted effect of each SNP on each *cis*-gene

Combined with summary-level endometriosis GWAS data (17,045 endometriosis cases and 191,596 controls) to impute gene expression and test the effect of genetically regulated gene expression level on the endometriosis.

Identified 252 genes associated with endometriosis located at 39 independent loci.

Mortlock et al. 2020

Advantages

- Gene-based approach is easier to interpret functional disease mechanisms.
- The two steps in TWAS can be conducted independently i.e., you may have genetic and phenotype information for a large cohort, but not expression data, so you can use eQTLs generated in an independent dataset.
- You can perform the first step and apply this to multiple different phenotypes in step 2.
- The multiple testing burden is lower in TWAS compared to GWAS. Only need to adjust for the number of genes tested. (1000's of genes in TWAS vs millions of SNPs in GWAS)
- TWAS has the capability to predict tissue-specific genetically regulated gene expression levels and investigate gene-trait associations in disease-related or potentially pathological tissues.

Factors influencing results and interpretation

- The nature of input GWAS data (eg. individual-level versus GWAS summary statistics).
- The eQTL models used (power, tissue and cell type heterogeneity).
- The association method used to estimate gene-trait associations (tissue specific vs test-all-tissue).
- Correlated gene expression.

Individual Level

Eg. PrediXcan (Gamazon et al. 2015) & MultiXcan (Barbeira et al. 2019)

- Individual-level genotype data are not easily obtainable from published GWAS studies.
- Can directly estimate LD structure.
- More accurate estimates of gene-trait associations.
- Takes significant computational resources.

Summary Statistics

Eg. FUSION (Gusev et al. 2016), S-PrediXcan (Barbeira et al. 2018), S-MultiXcan (Barbeira et al. 2019), UPMOST (Hu et al. 2019)

- Can impute the regression statistics between the gene expression level of each gene and a trait directly from GWAS summary statistics.
- More computationally efficient and has the ability to analyse large GWAS.
- LD matrix derived from a reference set. Discrepancy between the reference LD matrix and the actual LD structure of a study cohort can introduce noise and may lead to false positive or false negative results.
- Can prioritize genes using only GWAS summary statistics to reduce multiple testing.
- Needs additional validation and careful interpretation.

eQTL Considerations

- Quality of the eQTL dataset - higher quality studies can identify more eQTLs and eQTLs with moderate to small effect sizes and improve the precision of eQTLs in complex gene regions.
- Power - power to detect eQTLs from transcriptome and genotype datasets is partially dependent on the sample size.
- Tissue - eQTLs can differ between tissues, cell types and cell states.
- Impacts the prediction accuracy of gene expression levels.

Quality eQTL data in more diverse tissues have been made publicly available thanks to several consortia (eg. GTEx, eQTLGen).

Association method

Tissue-specific (eg. PrediXcan, S-PrediXcan and FUSION)

- Identify tissue-specific disease mechanisms.
- Limited power if dataset relatively small.
- What if causal tissue unavailable?
- Exploration of multiple tissue can increase multiple testing burden.

Test-all-tissue approach (eg. MulTiXcan, s-MulTiXcan and UTMOST)

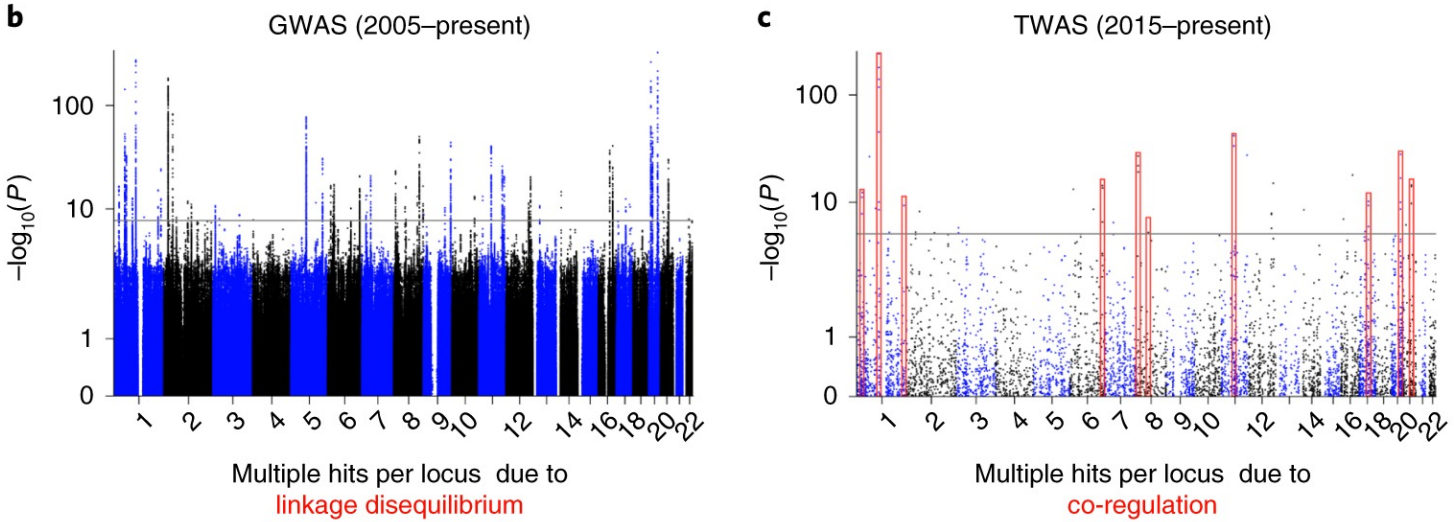
- Assumption that TWAS will only assign statistical significance to tissues that are biologically relevant to the complex trait of interest. This assumption, however, can be easily violated by eQTLs shared between tissues.
- The shared eQTL effects across tissues indicates that TWAS cannot distinguish disease-relevant tissues from irrelevant tissues that share similar gene expression levels from a statistical perspective
- Improved power but not tissue-specific and thus, cannot reveal tissue-specific genetic regulatory mechanisms.
- Computing resources and time required by cross-tissue TWAS methods are much higher.

Comparison	PrediXcan	S-PrediXcan	FUSION	MultiXcan	S-MultiXcan	UTMOST
Input GWAS data type	Individual-level genotype data	GWAS summary statistics	GWAS summary statistics	Individual-level genotype data	GWAS summary statistics	GWAS summary statistics
Statistical models for eQTL identifications	Elastic Net Fine-mapped MASHR-based models Joint-Tissue Imputation (JTI) models	Same as PrediXcan	Bayesian sparse linear mixed model (BSLMM)	Same as PrediXcan	Same as PrediXcan	Group LASSO with specialized regularization
Source reference panels	GTEx, MESA, CommonMind, StarNet, DGN, PsychENCODE	Same as PrediXcan	GTEx, TCGA	Same as PrediXcan	Same as PrediXcan	GTEx, StarNet, BLUEPRINT
eQTL Databases	http://predictdb.org/ https://zenodo.org/record/3842289#.YNVbJBOpGdY	http://predictdb.org/	http://gusevlab.org/projects/fusion/	http://predictdb.org/	http://predictdb.org/	https://github.com/Joker-Jerome/UTMOST
Current GTEx version[a]	GTEx v8	GTEx v8	GTEx v7	GTEx v8	GTEx v8	GTEx v6p
Gene-trait association methods	Linear or logistic regression	Dependent on GWAS method	Dependent on GWAS method	Principal component regression	Singular value decomposition (analogous to MultiXcan)	Generalized Berk-Jones test
Tissue-specificity	Tissue-specific	Tissue-specific	Tissue-specific	Cross-tissue	Cross-tissue	Cross-tissue
Output	Single-tissue gene-trait associations	Single-tissue gene-trait associations	Single-tissue gene-trait associations	Cross-tissue gene-trait associations	Cross-tissue gene-trait associations	Cross-tissue gene-trait associations
Pros	Up-to-date eQTL databases; Accurate representation of test cohort LD	Computationally efficient; Up-to-date eQTL databases;	Computationally efficient	Up-to-date eQTL databases;	Computationally efficient; Up-to-date eQTL databases;	Computationally efficient
Cons	Multiple testing burden; Computationally burdensome in comparison to summary-statistics based TWAS;	Reference LD matrix can introduce noises	Multiple testing burden; Reference LD matrix can introduce noises	Computationally burdensome;	Reference LD matrix can introduce noises	Reference LD matrix can introduce noises;
References using PMID	26258848, 32917697, 33020666	29739930	30926970	30668570	30668570	30804563

[a] *Dated August 2021.*

Limitations

- Prediction accuracy of gene expression levels is limited by the heritability (h^2) of each gene.
- Only using *cis*-eQTLs within a certain distance from genes. Trans-eQTLs may explain a large proportion of the heritability.
- Lack of eQTL data from different ancestry groups, diseases, medical conditions, sex, etc.
- TWAS power can be influenced by the quality of gene expression prediction (sample sizes, concordance between transcriptome reference population and testing populations, coverage of eQTLs in the test dataset, etc.), or genetic factors (e.g., genetic heritability of gene expression levels, heritability of the phenotype, sample size, MAF, etc.).
- When eQTL datasets have highly dissimilar sizes across tissues, the tissue with the most significant TWAS *P* value cannot necessarily be assumed to be causal, because reference-panel size affects the *P* value.
- Causal tissues or cell types are unclear in the majority of complex diseases or traits.
- Statistically significant TWAS results indicate only association, but not causation.
- TWAS prioritizes multiple genes, some likely to be non-causal.

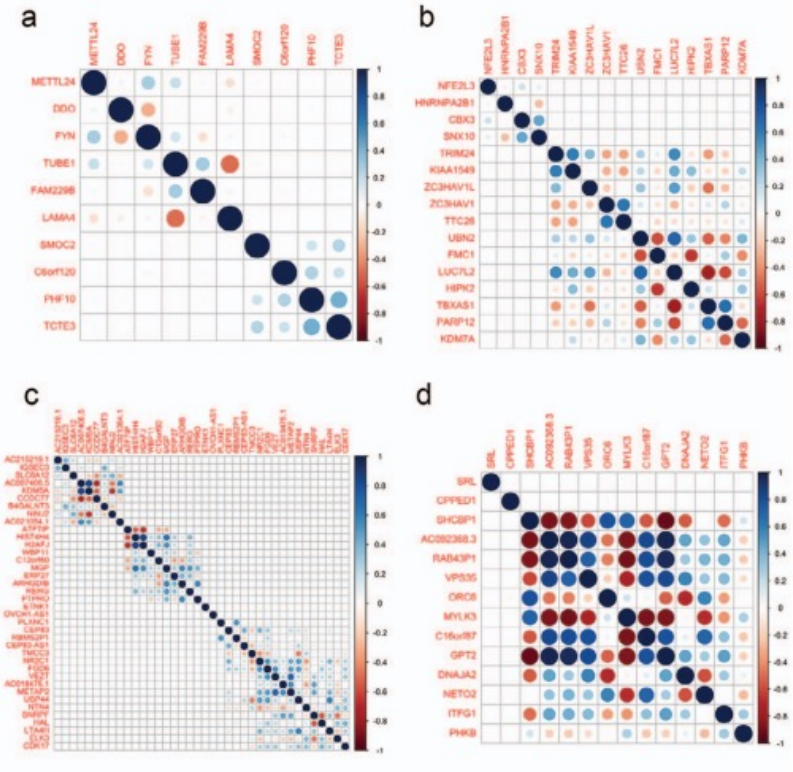


d

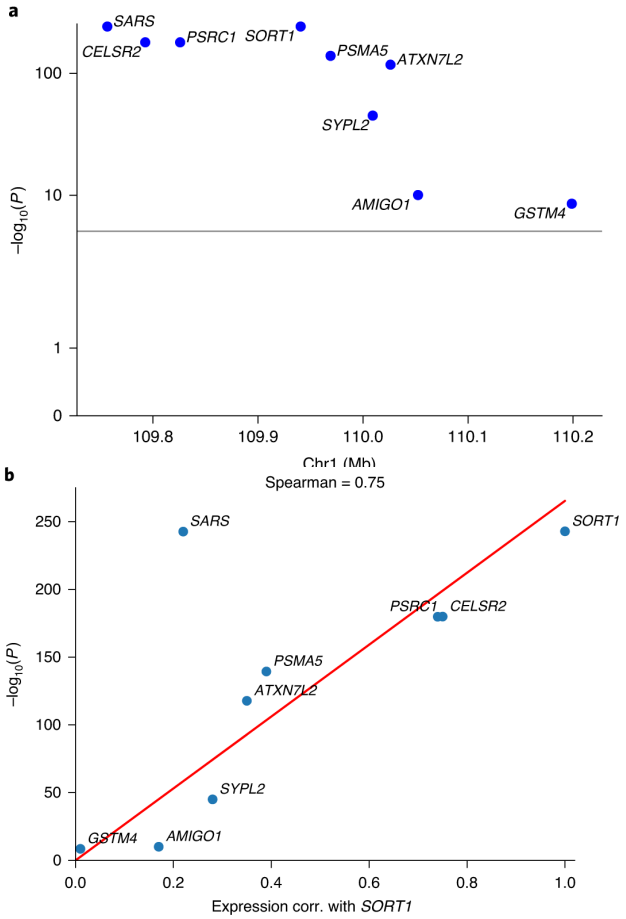
Scenario	Estimated percentage of non-causal hit genes	Case-study locus
Correlated expression	20% ($r^2 \geq 0.2$)	*SORT1* (LDL, liver)
Correlated predicted expression	75% ($r^2 \geq 0.2$)	*IRF2BP2* (LDL, liver)
Expression models share variants	69% (≥ 1 shared)	*NOD2* (Crohn's, whole blood)

Wainberg et al. 2019

Correlated expression



Mortlock et al. 2020



Wainberg et al. 2019

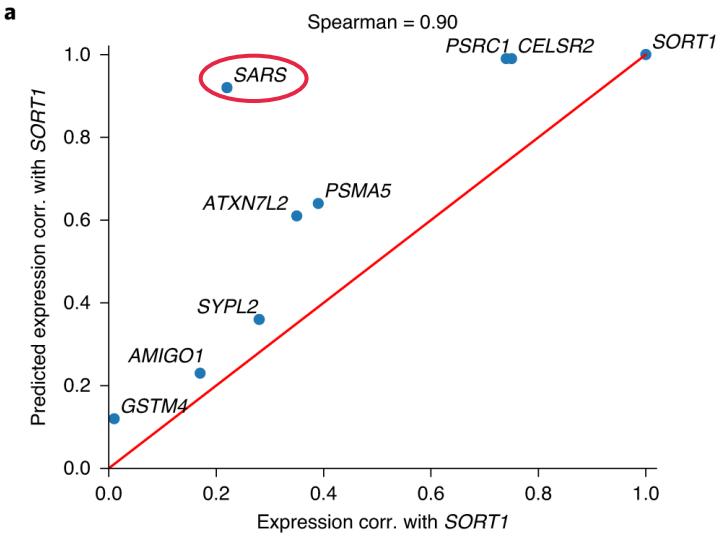
Correlated predicted expression

Total expression = genetic (cis-eQTLs, trans-eQTLs), environmental and technical components.

Predicted expression = genetic (common cis-eQTLs).

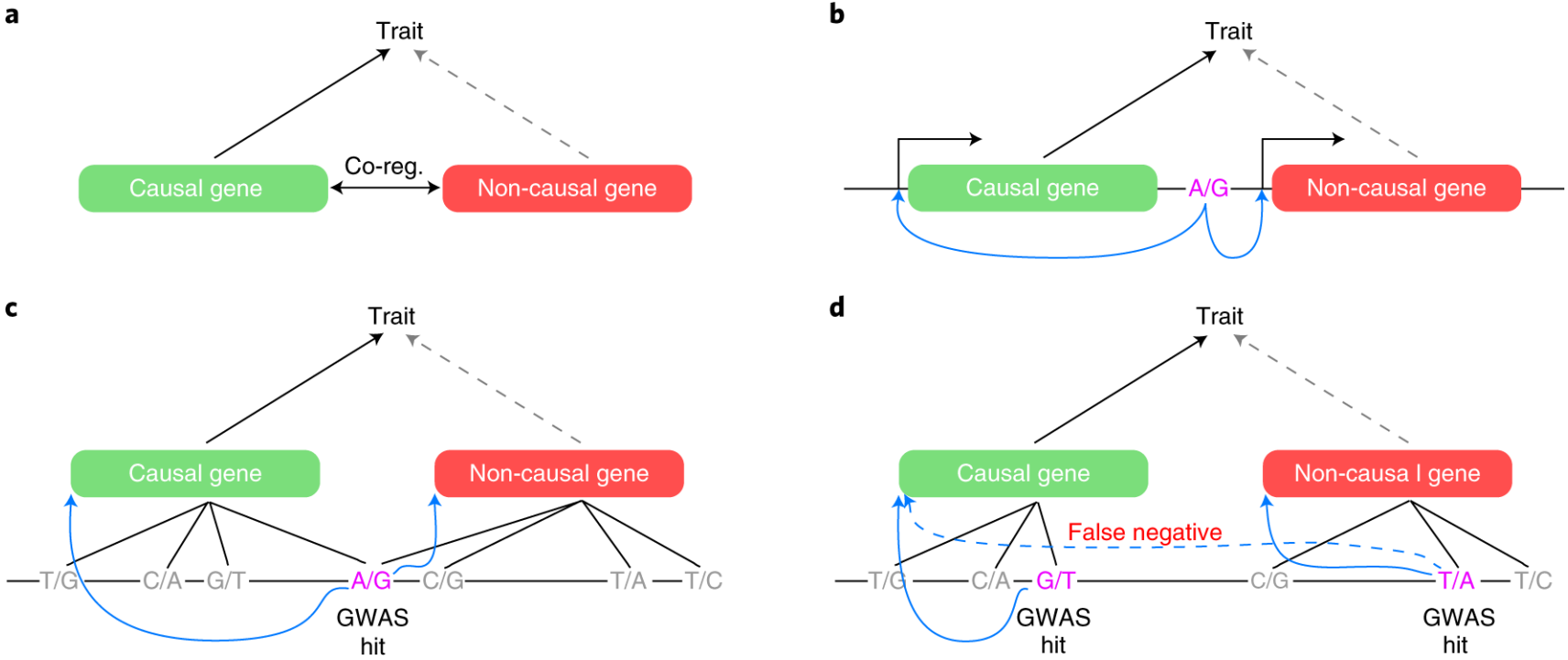
A gene pair can have correlated predicted expression if the same causal eQTL regulates both genes or if two causal eQTLs in LD each regulate one of the genes.

Correlated predicted expression can cause non-causal hits even in the absence of correlated total expression.



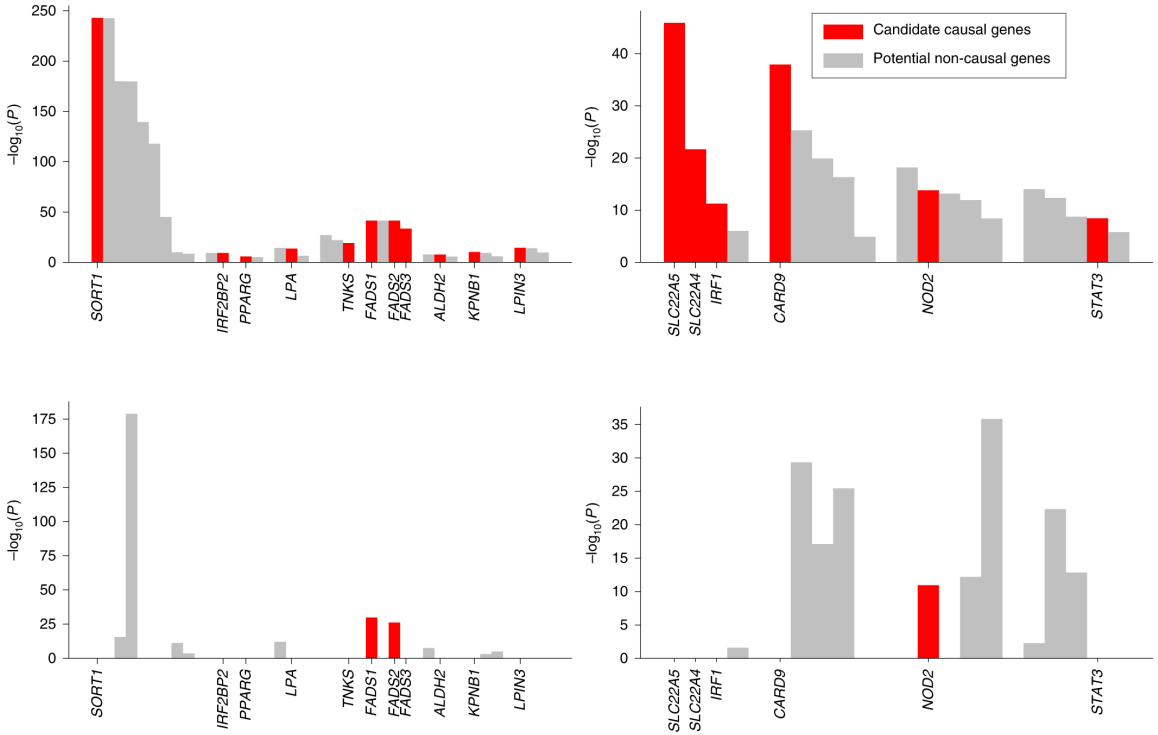
Wainberg et al. 2019

Scenarios in TWAS that may lead to non-causal hits



Wainberg et al. 2019

Bias with expression panels from non-trait-related tissues



Wainberg et al. 2019

Strategies to mitigate limitations

- Fine-mapping of causal gene sets (FOCUS)(Mancuso et al. 2019) directly models predicted expression correlations and uses them to assign genes posterior probabilities of causality.
- Use an eQTL dataset from only the most mechanistically related tissue available (balance between tissue bias and sample size).
- If no sufficiently large eQTL datasets from closely related tissues are available, we recommend aggregating information across all available tissues in a tissue-agnostic manner.
- eQTL dataset size affects the *P*-value. As such you should consider TWAS effect size in addition to *P*-value when investigating causal tissues for TWAS-associated genes.
- Test statistics can be inflated from by-chance QTL co-localization when the GWAS locus is highly significant and LD is extensive. Can test significance conditional on high GWAS effects (permutation test).

Practical – Exploring TWAS Hub

TWAS Hub

TWAS hub is an interactive browser of results from integrative analyses of GWAS and functional data for hundreds of traits and >100k expression models. The aim is facilitate the investigation of individual TWAS associations; pleiotropic disease/trait associations for a given gene of interest; predicted gene associations for a given disease/trait of interest with detailed per-locus statistics; and pleiotropic relationships between traits based on shared associated genes.

For each trait, a TWAS is carried out using the FUSION software (http://gusevlab.org/projects/fusion/). Gene models/weights were calculated from GTEX (45 tissues), METSIM (Adipose), NTR (Blood), ROSMAP (Brain), YFS (Blood), CommonMind (Brain) and TCGA (24 cancer tissues).

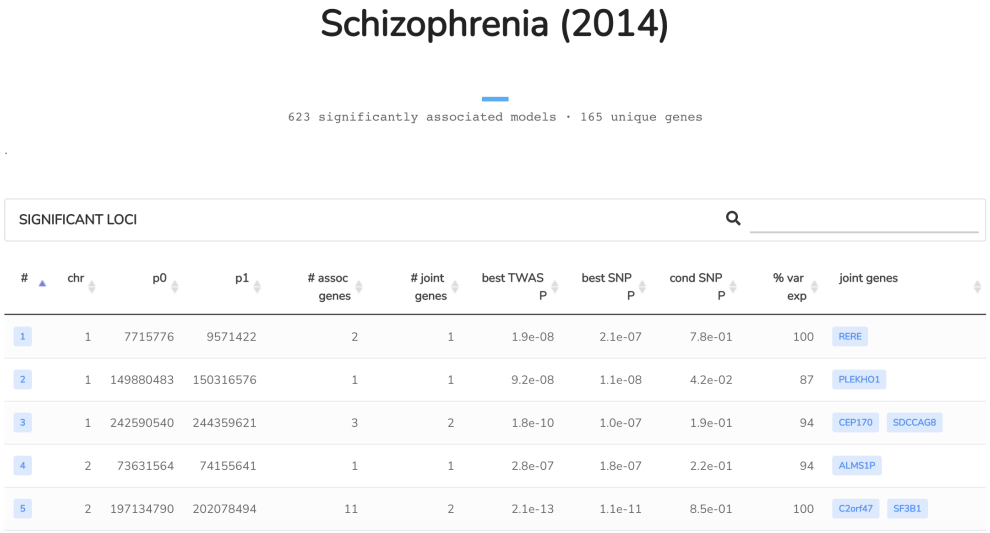
Genotypes are restricted to common, well-imputed HapMap3 SNPs. Typically, gene expression was analyzed with covariates for sex, age, genetic ancestry, and multiple gene expression PCs. For analyses of gene expression in tumors (from TCGA) local copy number alterations was also included as a covariate.

Open TWAS Hub in your browser http://twas-hub.org/

Trait View – Table 1

First, go to the traits tab and search for Schizophrenia.

The first table of the Trait View shows all of the transcriptome-wide significant associations for the given trait (after bonferroni correction for all models tested). Loci have been grouped into contiguous blocks and model selection run on each locus to identify the independently significant genes (which are reported in the right-most column).



#	chr	p0	p1	# assoc genes	# joint genes	best TWAS P	best SNP P	cond SNP P	% var exp	joint genes
1	1	7715776	9571422	2	1	1.9e-08	2.1e-07	7.8e-01	100	RERE
2	1	149880483	150316576	1	1	9.2e-08	1.1e-08	4.2e-02	87	PLEKHO1
3	1	242590540	244359621	3	2	1.8e-10	1.0e-07	1.9e-01	94	CEP170 SDCCAG8
4	2	73631564	74155641	1	1	2.8e-07	1.8e-07	2.2e-01	94	ALMS1P
5	2	197134790	202078494	11	2	2.1e-13	1.1e-11	8.5e-01	100	C2orf47 SF3B1

Trait View – Table 2

The second table shows all pleiotropic associations to other traits for any of the independently significant genes. The table is ordered by the “Chi^2 ratio” which is computed as the average Chi^2 statistic for the selected genes in the secondary trait, divided by the average statistic for all genes in the secondary trait. Ignoring issues of LD, this is an estimate of the heritability enrichment of the target genes relative to all genes and tends to provide reasonable results. For example, we can see that schizophrenia associated genes are also enriched for bipolar disorder, smoking, blood pressure, anxiety, nervous feelings, etc. The remaining columns list the number of significant genes in the target trait at Bonferroni correction [+] and at transcriptome-wide significance [++], the correlation of effect-sizes across the [+] genes, as well as links to each of the [+] genes.

PLEIOTROPIC ASSOCIATIONS

Trait	chisq ratio	# genes^{+}	# genes^{++}	% genes^{++}	corr	corr P	genes
Schizophrenia (2018)	14.7	48	30	66.7	1.00	9.5e-50	AC011816.1 AC103965.
Bipolar Disorder or Schizophrenia	14.6	47	33	73.3	0.99	4.3e-39	AC011816.1 AC103965.
Bipolar Disorder (2018)	5.7	8	3	6.7	0.99	5.6e-07	AC103965.1 CACNA2D4
Bipolar Disorder (2011)	5.1	3	0	0.0	0.00	1.0e+00	CACNA2D4 ITIH4-AS1
Pack years adult smoking proportion	4.9	2	1	2.2	0.00	1.0e+00	CHRNA5 KLC1
Pack years of smoking	4.4	1	1	2.2	0.00	1.0e+00	CHRNA5
Diastolic blood pressure, automated reading	4.1	10	8	17.8	-0.14	6.9e-01	ALMS1P CEP170 FE
Worrier / anxious feelings	3.9	7	5	11.1	0.39	3.9e-01	AS3MT C2orf47 CA

Trait View – Table 3

The third table shows the breakdown of associations by gene expression panel. These are ordered by the average TWAS Chi^2 statistic in the panel – an estimate of the average trait heritability explained by predictors from that expression study. The columns also report the # and % of significant associations from that study. In this case, we see no relevant tissue-specific enrichment for schizophrenia (see Prostate Cancer for an example of tissue specificity).

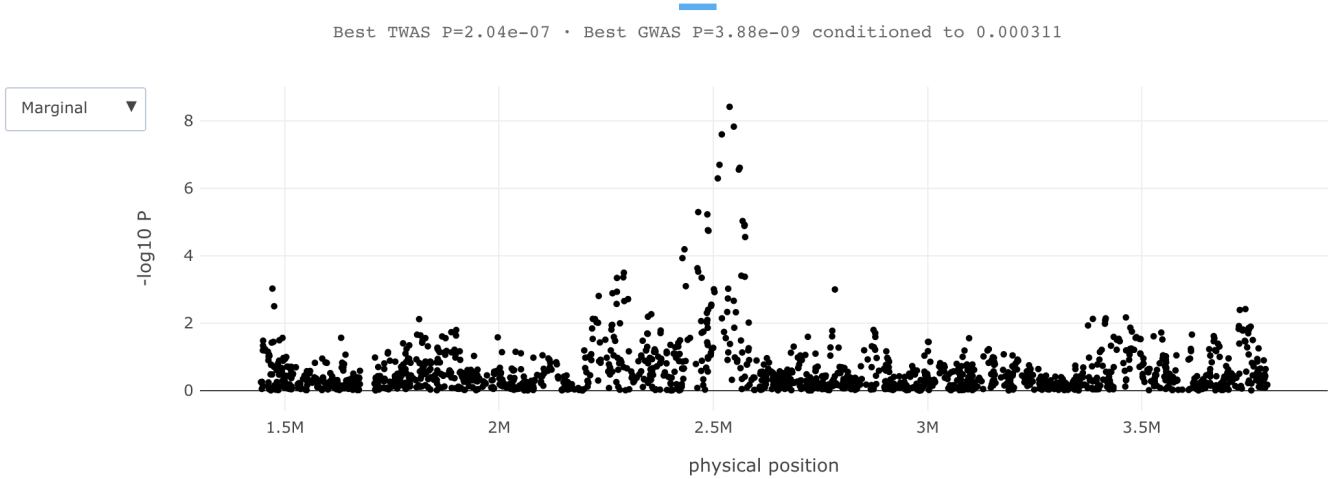
ASSOCIATIONS BY PANEL

study	tissue	# hits	% hits/tests	avg chisq
GTEx	Pancreas	7	0.44	2.6
GTEx	Prostate	4	0.49	2.6
GTEx	Small Intestine Terminal Ileum	3	0.68	2.6
GTEx	Brain Cerebellum	8	0.42	2.5
GTEx	Breast Mammary Tissue	13	0.68	2.5

Locus View

Click on locus #7 in the schizophrenia associations table to go to the Locus View for the *CNTN4* locus. The top panel shows a Manhattan plot of the GWAS association before and after conditioning on the predicted expression.

chr3:1,442,951-3,792,945



Locus View

The next panel shows all the significantly associated models, their model performance, correlation with the top index SNP, and *coloc* posterior probabilities (PP3 = two distinct causal variants; PP4 = a single shared causal variant). Here we see a single predictive model for *CNTN4* at this locus (from CommonMind brain) with a high PP4 and a much stronger TWAS vs eQTL Z-score, suggesting the TWAS is aggregating additional predictive signal - all good indicators of a pleiotropic effect. Since only one model is significant in the locus it is the “joint”ly selected model by default.

ASSOCIATED MODELS

#	Study	Tissue	Gene	h2	eQTL R2	model	# weights	model R2	model R2 P	eQTL GWAS Z	TWAS Z	TWAS P	Top SNP corr	PP3	PP4	joint
1	CommonMind	Brain Pre-frontal Cortex	CNTN4	0.15	0.03	enet	34	0.05	4.7e-07	3.5	5.2	2e-07	0.56	0.02	0.98	TRUE

Gene View

Click on CNTN4 to go to the Gene View. The top table shows all the predictive models that have been computed for this gene and their respective performance. Here we again see that for the model trained in brain, the best multivariate predictive model (in this case elastic net with cross-validation P=4.7e-07) far outperforms the best eQTL (P=2.3e-04), which provides further confidence that the TWAS predictor is capturing real additional signal and leading to a more significant disease association.

MODELS

#	panel	tissue	h2	h2 se	h2 P	eQTL R2	BLUP R2	ENET R2	BSLMM R2	LASSO R2	eQTL P	BLUP P	ENET P	BSLMM P	LASSO P
1	CommonMind	Brain Pre-frontal Cortex	0.149	0.034	1.7e-05	0.028	0.044	0.053	0.05	0.046	2.3e-04	4.0e-06	4.7e-07	9e-07	2.3e-06
2	GTEx	Adipose Visceral Omentum	0.241	0.084	8.7e-03	0.012	NA	0.023	NA	0.006	7.5e-02	NA	2.1e-02	NA	1.5e-01
3	GTEx	Thyroid	0.233	0.053	2.9e-05	0.209	NA	0.194	NA	0.195	6.6e-16	NA	9.0e-15	NA	6.8e-15
4	TCGA	Thyroid Carcinoma	0.192	0.040	3.6e-07	0.127	0.094	0.133	NA	0.136	2.5e-12	1.9e-09	7.5e-13	NA	3.7e-13

Gene View

The second table shows a heatmap of association for this gene between all traits (rows) and all models (columns 4-). We order the heatmap by the “avg Chi² ratio” column, which is computed as the average Chi² for the gene-disease pair (across all models) divided by the average Chi² for all genes in the listed disease (across all models). This normalization accounts for sample size and heritability differences between traits and emphasize associations that are stronger than expected by chance (without the normalization, highly heritable and polygenic traits like height, for example, would constantly be at the top of the list simply because they have so many detectable causal variants). The subsequent columns list the raw average Chi² statistic, maximum Chi² statistic across all models (to filter for model-specific associations), and then the individual Z-scores for each model. Here we see that schizophrenia is the second most enriched trait for *CNTN4* associations, followed by feelings-related measurements – potentially informing our understanding of how this gene fits into the cross-trait relationships. Sorting on column #1 shows that the brain model is only significantly associated with schizophrenia. Sorting on the “max chi2” column shows that no other models are strongly associated (with any trait).

TRAIT ASSOCIATIONS

Trait	Avg chi2 ratio	Avg chi2	Max chi2	1	2	3	4
Bipolar Disorder or Schizophrenia	10.1	24.6	38.6	6.2	1.7	4.4	6.1
Schizophrenia (2018)	8.8	18.2	31.2	5.6	2.5	3.6	4.8
Schizophrenia (2014)	7.6	16.8	27.0	5.2	1.1	3.7	5.1
Worry too long after embarrassment	5.4	9.1	15.1	-2.4	-0.9	-3.9	-3.8
Ever depressed for a whole week	5.2	6.5	10.6	2.7	-0.1	2.8	3.3
Bipolar Disorder (2018)	4.9	7.8	16.5	2.8	-0.3	2.6	4.1
Body Mass Index (BMI) (2010)	4.3	5.9	8.1	-2.1	-2.4	-2.8	-2.4
Depressed Affect (Nagel 2018)	4.1	8.6	14.7	-1.9	-1.7	-3.6	-3.8
Medication: Cholesterol lowering	4.1	6.3	10.5	-2.7	-0.1	-2.7	-3.2

Possible target gene

TWAS evidence suggests *CNTN4* is associated with schizophrenia and this is a brain-specific effect.

CNTN4 was recently implicated in schizophrenia and shown to change neurodevelopment in zebrafish by Fromer et al. 2016 Nat Neurosci.