The American Journal of Human Genetics, Volume *100*

Supplemental Data

Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits

Nicholas Mancuso, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc

Figure S1. Estimates of $\rho_{g,local}$ between gene expression and trait are unbiased in simulations. Starting from real genotype data, we simulated gene expression at independent loci. We then simulated complex trait as a linear function of predicted expression at these loci. We performed a GWAS using complex trait and subsequent TWAS at each gene (using GBLUP weights) which was used as input to estimate $\rho_{g,local}$. A) Results for 2,500 simulations where the causal SNPs driving gene expression were typed in the data. B) Results for 2,500 simulates where causal SNPs driving gene expression were untyped.

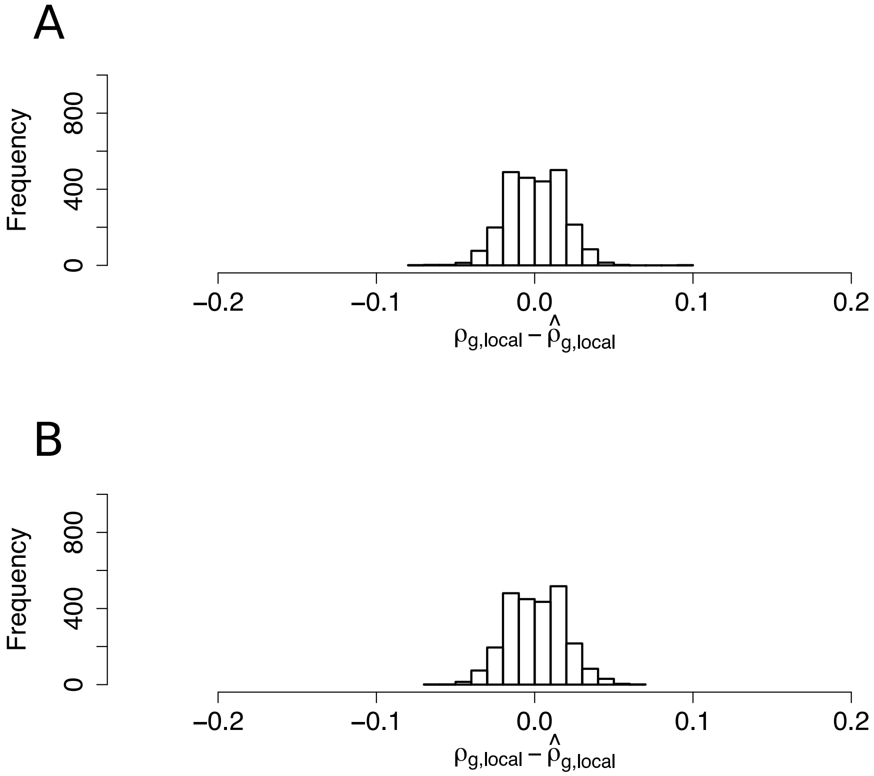


Figure S2. QQ-plot of null distribution in simulations measuring ρ_{GE}. The red line represents the identity line and the gray area is the 95% confidence interval of the null.

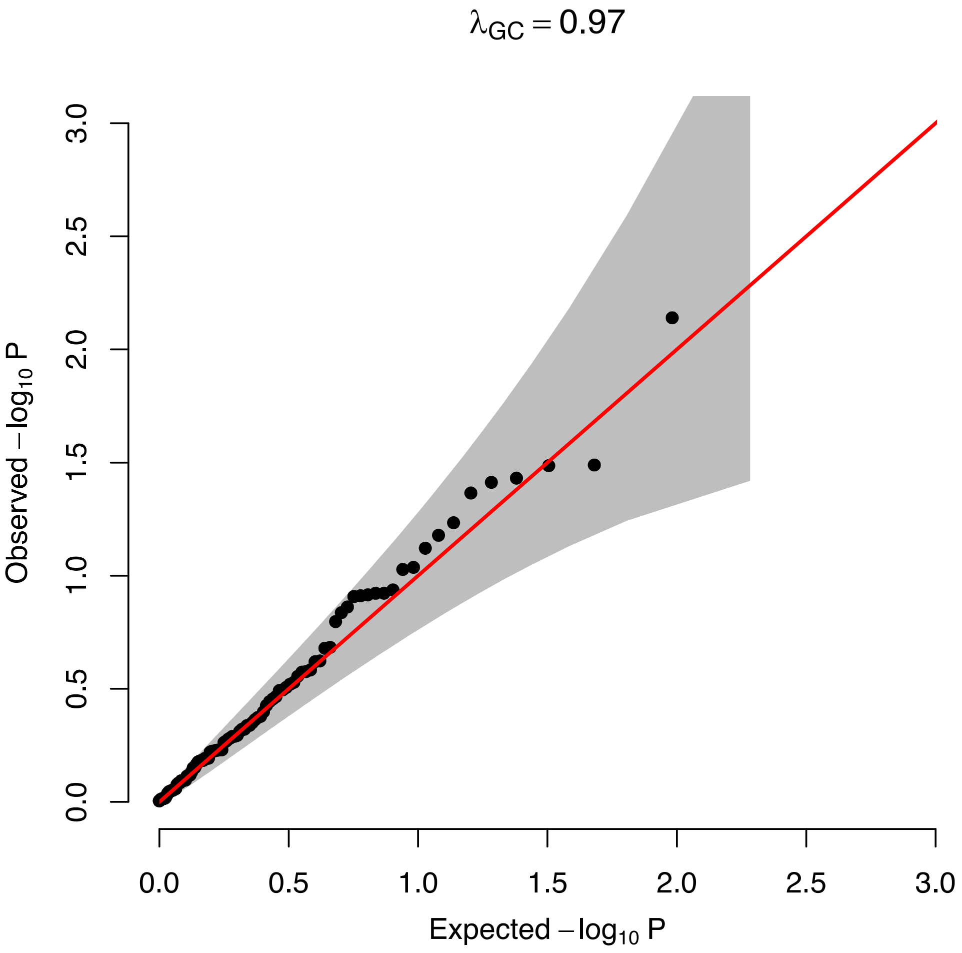


Figure S3. Estimating $h^2_{g,local}$ using all SNPs in a locus compared to the top eQTL. A) Estimates of $h^2_{g,local}$ using the described joint estimator versus the top eQTL. Results in the top row are obtained with causal SNPs typed in the data. Results in the bottom row have causal SNPs untyped/pruned from the genotype data. B) Joint estimation of $h^2_{g,local}$ results in better estimates for ρ_{GE}. The dotted line is the identity line. Each point represents the mean estimated ρ_{GE} over 100 simulations. Error bars capture the 95% confidence interval.

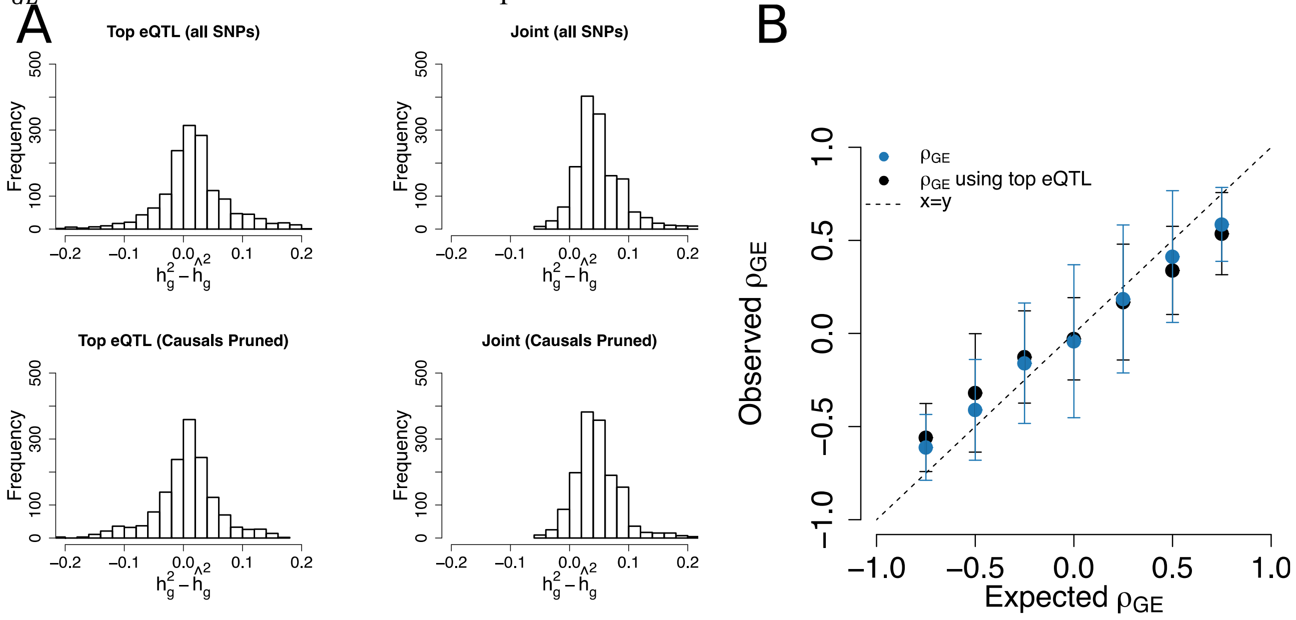


Figure S4. SNP association P-values at reported susceptibility genes not proximal to a genome-wide significant SNP tend to be suggestive. The red dotted line represents genome-wide significance.

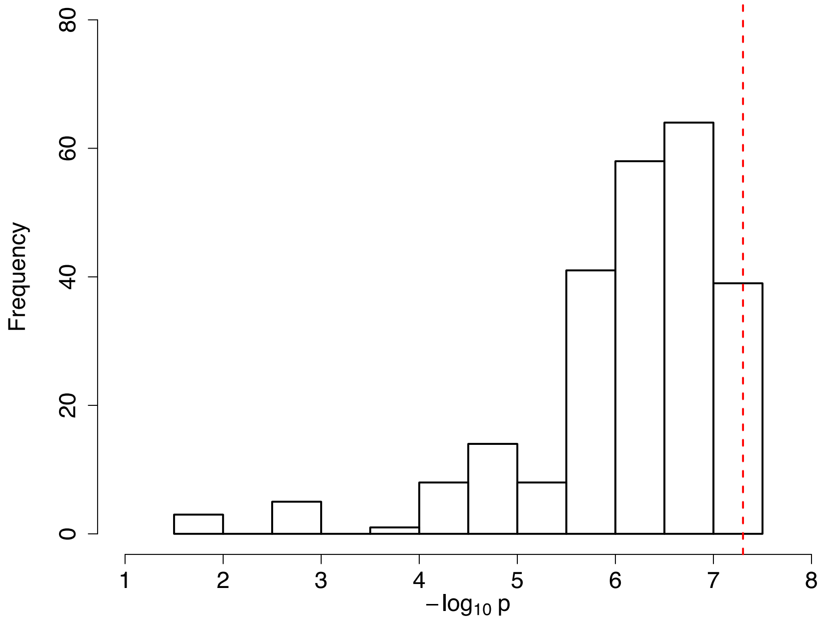


Figure S5. Distance from index SNP to gene transcription start site. Overlap was determined my selecting all SNPs within a 1 Mb flanking region around the gene's TSS. The red curve represents the estimated density assuming a Laplacian distribution of distances from the TSS.

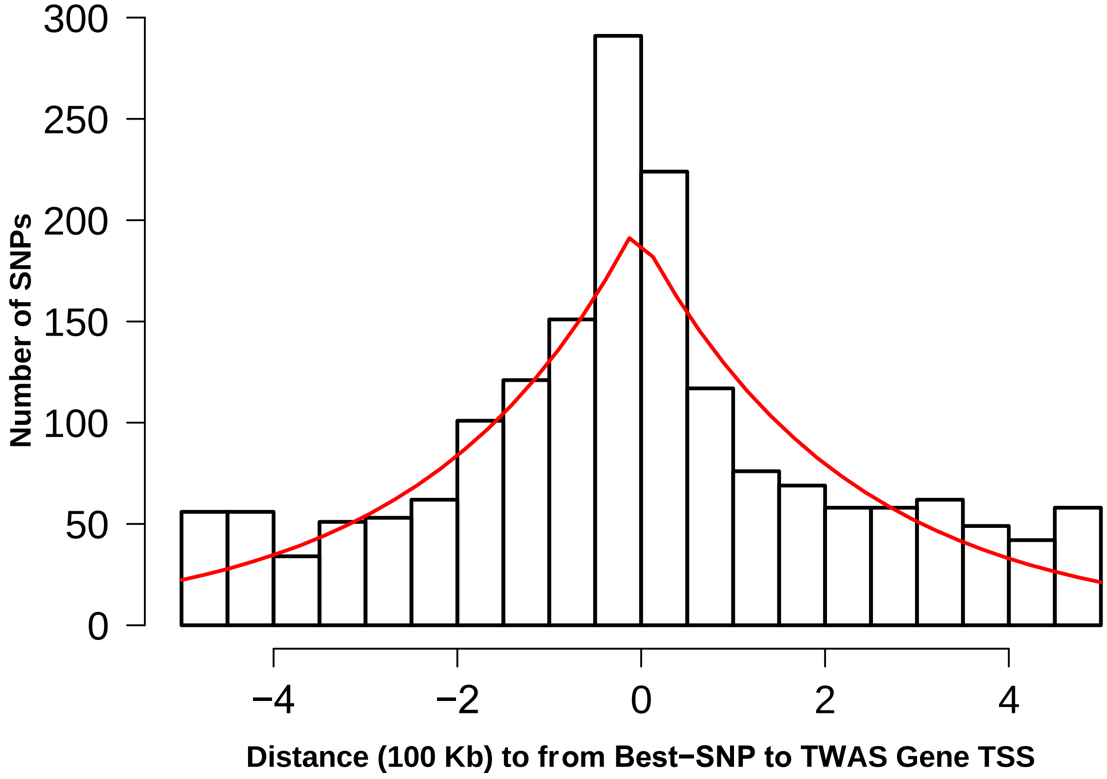


Figure S6. Distribution of association statistics for genes closest to index SNPs versus the top gene. The difference in means was significant under a Welch's t-test. Error-bars capture the lower and upper quartiles, with outliers represented as points.

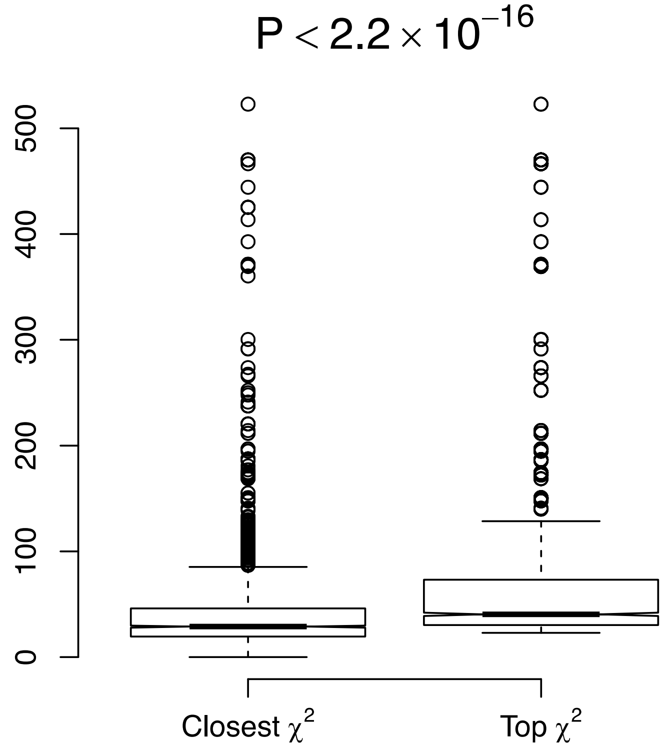


Figure S7. Number of genes associated with multiple traits.

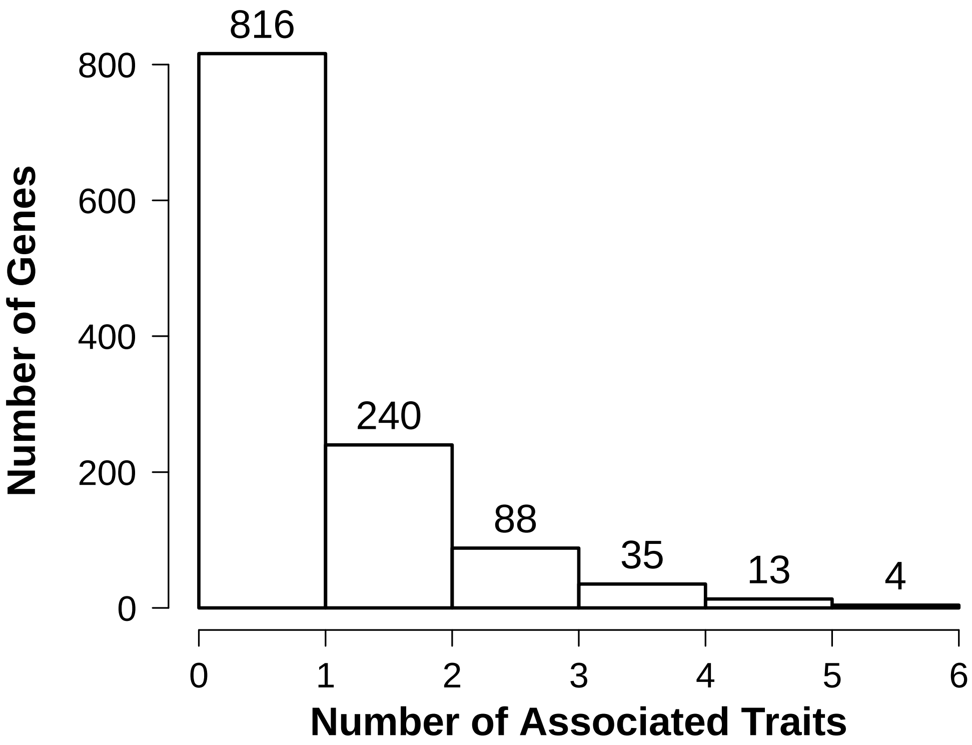


Figure S8. Comparing ρ_{GE} estimates computed using the top eQTL versus the entire locus. Estimates of ρ_{GE} in real data using the top eQTL are highly consistent with original estimates. The blue line represents the regression line fitted to the data.

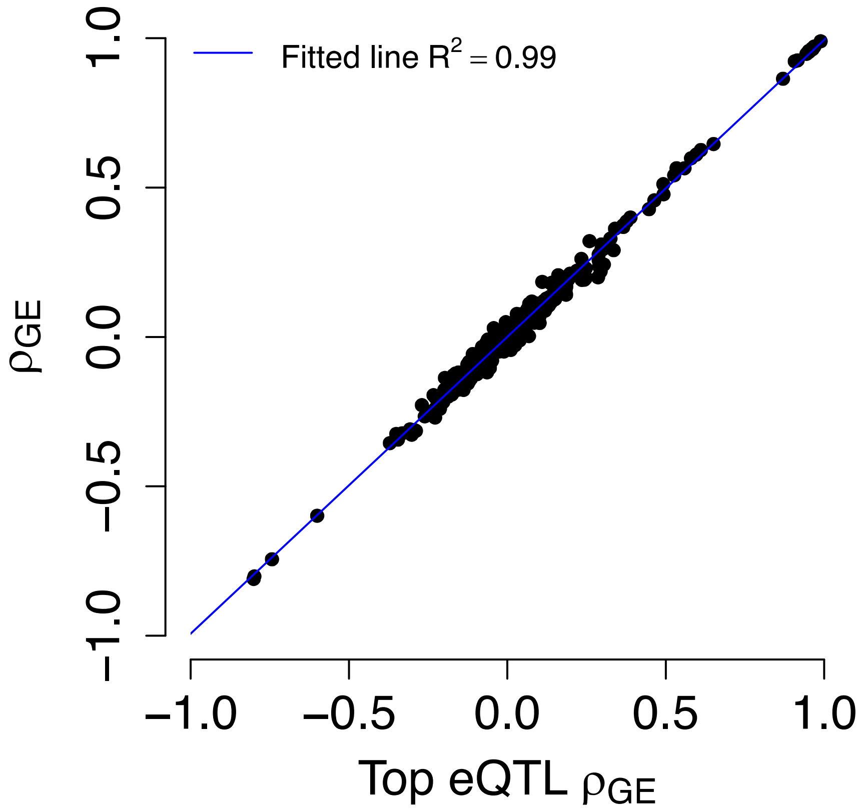


Figure S9. Molecular function analysis of TWAS genes for all traits. We only list functions that are significant ($P < 0.05$) after Bonferroni correction.

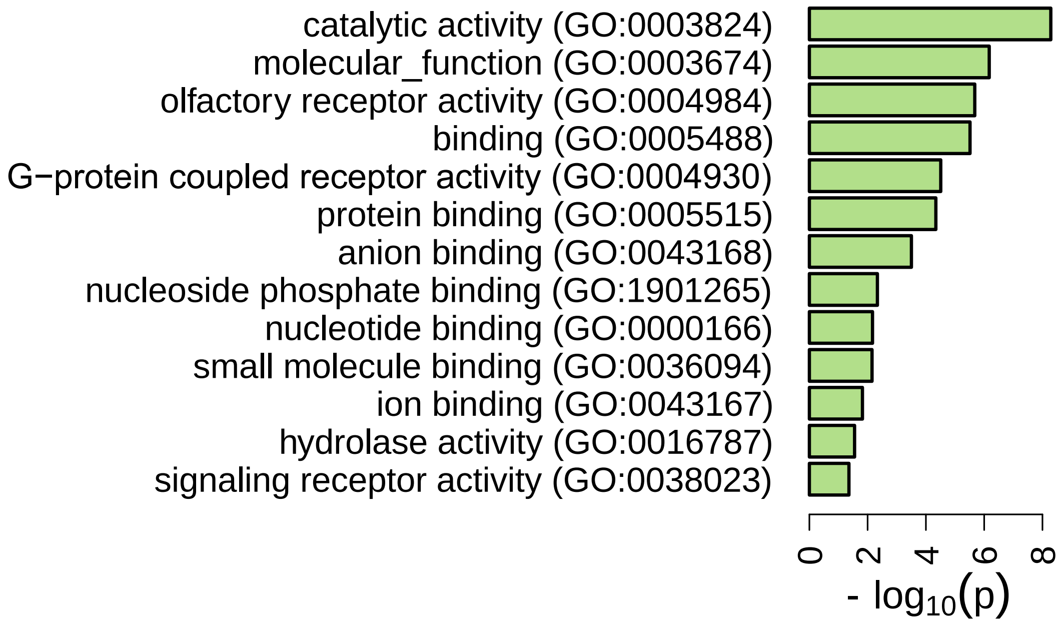
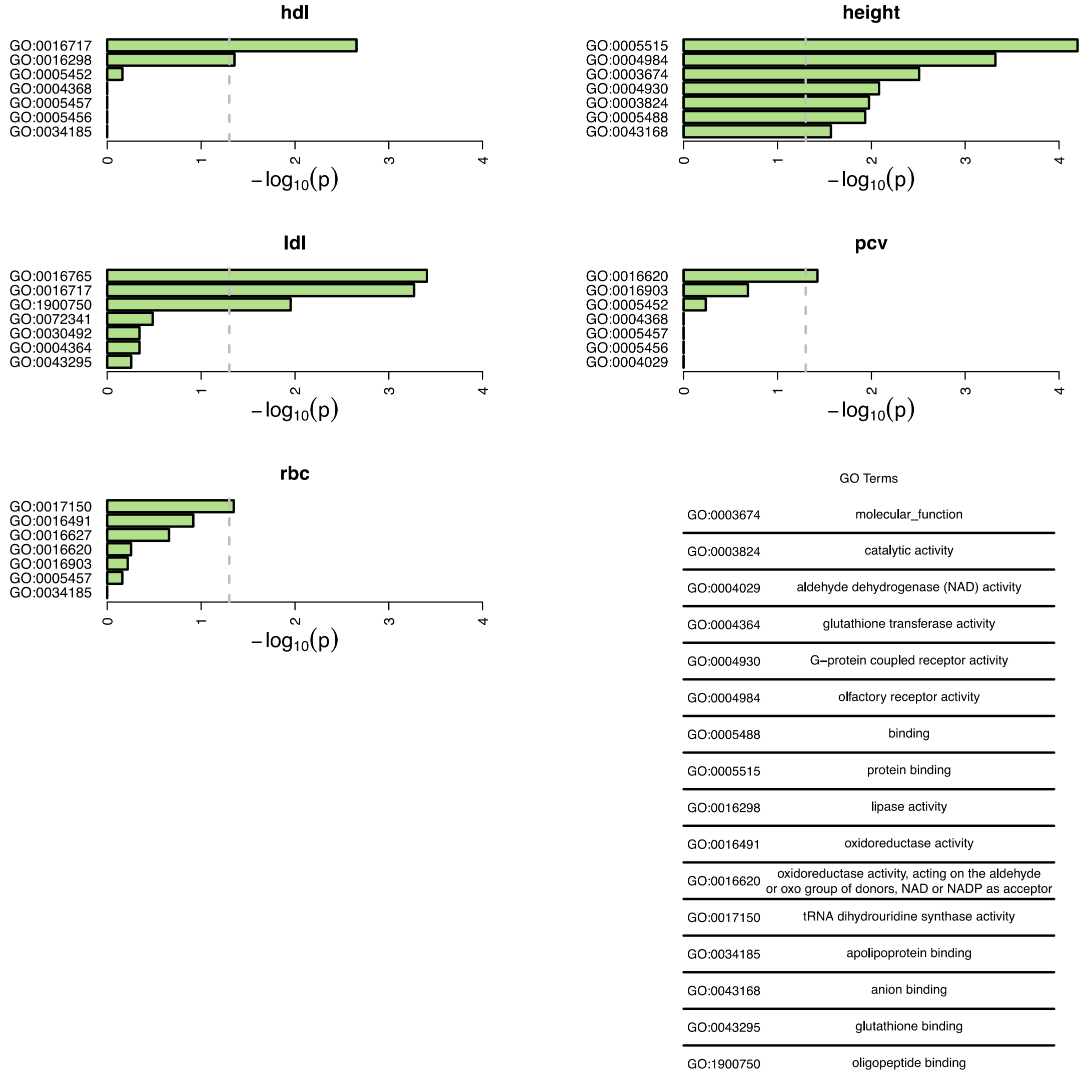


Figure S10. Molecular function analysis of TWAS genes. We only list functions that are significant ($P < 0.05$) after Bonferroni correction.



GO Terms

GO:0003674	molecular_function
GO:0003824	catalytic activity
GO:0004029	aldehyde dehydrogenase (NAD) activity
GO:0004364	glutathione transferase activity
GO:0004930	G−protein coupled receptor activity
GO:0004984	olfactory receptor activity
GO:0005488	binding
GO:0005515	protein binding
GO:0016298	lipase activity
GO:0016491	oxidoreductase activity
GO:0016620	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor
GO:0017150	tRNA dihydrouridine synthase activity
GO:0034185	apolipoprotein binding
GO:0043168	anion binding
GO:0043295	glutathione binding
GO:1900750	oligopeptide binding

Figure S11. Biological process analysis of TWAS genes for all traits. We only list functions that are significant ($P < 0.05$) after Bonferroni correction.

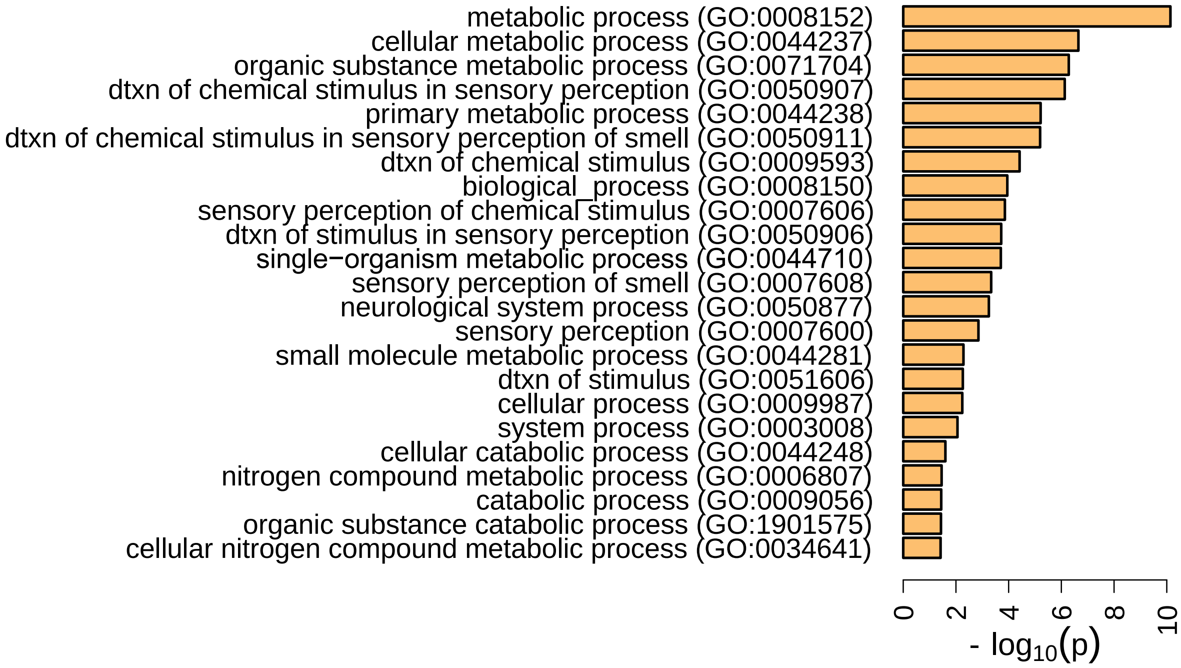
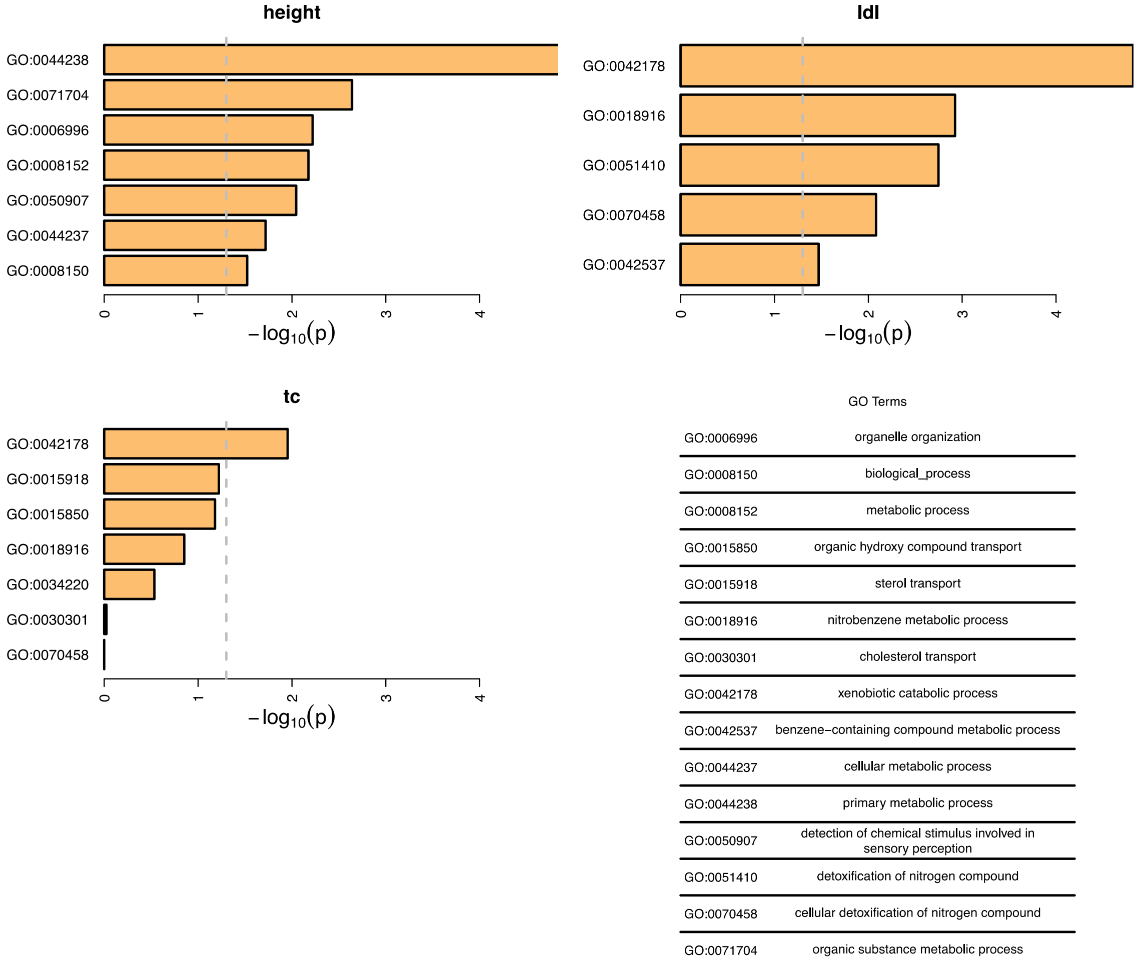


Figure S12. Biological process analysis of TWAS genes for height, LDL, and total cholesterol. We only list functions that are significant ($P < 0.05$) after Bonferroni correction.



GO Terms

GO:0006996	organelle organization
GO:0008150	biological_process
GO:0008152	metabolic process
GO:0015850	organic hydroxy compound transport
GO:0015918	sterol transport
GO:0018916	nitrobenzene metabolic process
GO:0030301	cholesterol transport
GO:0042178	xenobiotic catabolic process
GO:0042537	benzene–containing compound metabolic process
GO:0044237	cellular metabolic process
GO:0044238	primary metabolic process
GO:0050907	detection of chemical stimulus involved in sensory perception
GO:0051410	detoxification of nitrogen compound
GO:0070458	cellular detoxification of nitrogen compound
GO:0071704	organic substance metabolic process

Trait	Short Name	Sample Size	Number of SNPs	Trait measurement	Trait Group
Age at Menarche	AM	132989	1821879	Quantitative	Metabolic
Body Mass Index	BMI	226814	1859666	Quantitative	Anthropometric
College	COL	126559	1792881	Dichotomous	Social
Crohn's Disease	CD	51874	4822932	Dichotomous	Immune-related
Education Years	EY	126559	1788888	Quantitative	Social
Fasting Glucose	FG	46186	1824182	Quantitative	Metabolic
Fasting Insulin	FI	46186	1822388	Quantitative	Metabolic
Femoral Neck BMD	FN	53236	4637340	Quantitative	Anthropometric
Forearm BMD	FA	53236	4725343	Quantitative	Anthropometric
Hemoglobin	HB	51496	1894024	Quantitative	Hematopoietic
HBA1C	HBA1C	46368	1870395	Quantitative	Hematopoietic
Height	HEIGHT	241286	1854761	Quantitative	Anthropometric
High Density Lipoprotein	HDL	95572	1805617	Quantitative	Metabolic
HOMA-B	HOMA-B	46186	1820938	Quantitative	Metabolic
HOMA-IR	HOMA-IR	46186	1821061	Quantitative	Metabolic
Inflammatory Bowel Disease	IBD	65643	4823603	Dichotomous	Immune-related
Low Density Lipoprotein	LDL	90811	1803637	Quantitative	Metabolic
Lumbar Spine	LS	53236	4636561	Quantitative	Anthropometric
MCH Concentration	MCHC	47157	1893281	Quantitative	Hematopoietic
Mean Cell Hemoglobin	MCH	43733	1892019	Quantitative	Hematopoietic
Mean Cell Volume	MCV	48689	1893769	Quantitative	Hematopoietic
Number of Platelets	PLT	66867	1954590	Quantitative	Hematopoietic
Packed Cell Volume	PCV	45125	1893412	Quantitative	Hematopoietic
Red Blood Cell Count	RBC	45500	1892553	Quantitative	Hematopoietic
Rheumatoid Arthritis	RA	58284	4265540	Dichotomous	Immune-related
Schizophrenia	SCZ	74626	4772186	Dichotomous	Neurological
Total Cholesterol	TC	95802	1805676	Quantitative	Metabolic
Triglycerides	TG	92007	1803908	Quantitative	Metabolic
Type 2 Diabetes	T2D	61857	1806359	Dichotomous	Metabolic
Ulcerative Colitis	UC	47746	4823578	Dichotomous	Immune-related

Table S1. Summary of the 30 GWAS data.

Expression Weights	**Causal variants**	$\boldsymbol{h^2_{GE}}$	**SE**
GBLUP	Typed	0.30	0.01
GBLUP	Untyped	0.27	0.01
True	Typed	0.50	0.01

Table S3. Simulation results for $\boldsymbol{h^2_{GE}}$ estimates. We simulated 100 complex traits as a linear function of gene expression at 50 loci (see Material and Methods). We re-ran simulations with the causal variants for expression untyped in the genotyping data. We present the mean h^2_{GE}estimate along with the standard error across all simulation runs.

Gene	Chr	Tx Start	Tx End	Current Index SNP	SNP P	New Index SNP	BP	SNP P
SDCCAG8	chr1	243419306	243663393	rs12080886	5.73E-07	rs2992632	243503764	3.245E-11
ABCB9 *MPHOSPH9*	chr12	123405497	123451056	rs7980687	1.59E-06	rs10773002	123746961	7.742E-18
STK24	chr13	99102452	99174379	rs17574378	1.52E-07	rs9556958	99100046	1.208E-11
EIF3CL *SULT1A1* *RP11-1348G14.4* *TUFM* *MIR4721* *SH2B1* *NFATC2IP*	chr16	28390902	28415206	rs8049439	1.52E-07	rs8049439	28837515	6.992E-11

Table S7. TWAS predicted susceptibility loci for Education Years. Reported TWAS susceptibility loci for Education Years that did not overlap a genome-wide significant SNP within ±0.5Mb of transcription start-site in the Rietveld et al. Science 2013 study ($N = 126,559$) were proximal to genome-wide significant SNPs found in the much larger Okbay et al. Nature 2016 study ($N = 293,723$).