Data supplement for Seligowski et al., Leveraging Large-Scale Genetics of PTSD and Cardiovascular Disease to Demonstrate Robust Shared Risk and Improve Risk Prediction Accuracy. Am J Psychiatry (doi: 10.1176/appi.ajp.21111113)

Phenotype data

Diagnostic data was extracted from the MGBB and ICD-10 codes and curated disease populations were used to determine diagnosis.[31,32] Curated Disease Populations are collected by a validated phenotype algorithm with a positive predictive value of 0.90. These algorithms rely on coded diagnosis as well as natural language processing terms extracted from clinical narratives. Predictive features for each characterization were identified using an automated feature extraction protocol, which identifies comorbidities, symptoms, and medications. Concepts were screened based on frequency in patient clinical notes. Billing diagnosis and prescriptions were also included in algorithms.[31,32] Patients with lifetime history of congestive heart failure, coronary artery disease, hypertension, ischemic stroke, and depression as determined by Curated Disease Population status were considered positive for these traits in our dataset. Curated Disease Populations were not available for PTSD or myocardial infarction.

Curated Disease Population Algorithms

Model definitions with features and betas: Below are the feature weights of the final hypertension phenotype algorithms. The weights below are used to derive a predicted probability of hypertension or no hypertension for every Biobank participant.

Feature_ID	Beta (weight)	Feature Description
(Intercept)	0.247	Model Intercept (beta 0)
HTN_COD_DX_Hypertension	0.793	Count of coded diagnosis of hypertension
HTN_COD_MED_Antihypertensives	0.627	Count of prescriptions for an antihypertensive

HTN_COD_MED_Aceinhibitor	0.476	Count of prescriptions for an ACE inhibitor
max_bmi_new	-0.323	Maximum BMI (if missing, impute to mean sample BMI of 30)
patient_dxenct	-0.586	Total number of visits with a coded diagnosis

Below are the feature weights of the final CAD phenotype algorithms. The weights below can be used to derive a predicted probability of CAD or no CAD for any Biobank participant.

Feature_ID		Beta (weight)	Feature Description
(Intercept)		5.288948213	Model Intercept (beta 0)
patient_dxenct		-3.121364389	Number of encounters with ICD-9 codes
CAD_COD_DX_IschemicHeartDisease		1.934743408	Coded mentions of a Ischemic Heart Disease diagnosis
CAD_NLP_alcohol		0.16107732	Hitex non-negated NLP mentions of UMLS CUI C0001962: alcohol
CAD_NLP_angioplasty		0.022473675	Hitex non-negated NLP mentions of UMLS CUI C0162577: angioplasty
CAD_NLP_antiplateletagents		0.503346441	Hitex non-negated NLP mentions of UMLS CUI C0085826: antiplatelet agents
CAD_NLP_coronaryarterybypassgrafting		1.054659295	Hitex non-negated NLP mentions of UMLS CUI C0010055: coronary artery bypass grafting
CAD_NLP_coronaryatherosclerosis		-1.194440371	Hitex non-negated NLP mentions of UMLS CUI C0010054: coronary atherosclerosis
CAD_NLP_coronaryheartdisease		2.588093909	Hitex non-negated NLP mentions of UMLS CUI C0010068: coronary heart disease

CAD_NLP_creatinine		-0.342506235	Hitex non-negated NLP mentions of UMLS CUI C0010294: creatinine
CAD_NLP_electrocardiogram		-1.62020437	Hitex non-negated NLP mentions of UMLS CUI C0013798: electrocardiogram
CAD_NLP_ischemia		1.037838709	Hitex non-negated NLP mentions of UMLS CUI C0022116: ischemia
CAD_NLP_ischemiccardiomyopathy		-0.106484107	Hitex non-negated NLP mentions of UMLS CUI C0349782: ischemic cardiomyopathy
CAD_NLP_myocardialinfarction		0.230624677	Hitex non-negated NLP mentions of UMLS CUI C0027051: myocardial infarction
CAD_NLP_nitroglycerin		1.769966347	Hitex non-negated NLP mentions of UMLS CUI C0017887: nitroglycerin
CAD_NLP_plateletaggregationinhibitors		0.433529361	Hitex non-negated NLP mentions of UMLS CUI C0032177: platelet aggregation inhibitors

Below are the feature weights of the final depression phenotype algorithms. The weights below are used to derive a predicted probability of depression or no depression for every Biobank participant.

Feature_ID	Beta (weight)	Feature Description
(Intercept)	-2.539	Model Intercept (beta 0)
Depression_COD_DX_Depression	1.973	Count of coded diagnosis of depression or MDD

Depression_COD_DX_Mentalhealthdisorders	0.175	Count of coded diagnosis of any mental health disorder
Depression_COD_DX_Bipolardisorder	-0.813	Count of coded diagnosis of bipolar disorder
Depression_COD_MED_Antidepressants	1.096	Count of prescriptions for an antidepressant
Depression_COD_MED_Antipsychotics	-0.395	Count of prescriptions for an antipsychotic
Depression_COD_MED_Anticonvulsants	-0.323	Count of prescriptions for an anticonvulsant
patient_dxenct	-0.515	Total number of visits with a coded diagnosis

Below are the feature weights of the final CHF phenotype algorithms. The weights below can be used to derive a predicted probability of CHF or no CHF for any Biobank participant.

Feature_ID	Beta (weight)	Feature Description
(Intercept)	-0.85309914	Model Intercept (beta 0)
patient_dxenct	-0.790872064	Number of encounters with an ICD-9 code
CHF_COD_DX_CHF	0.299480034	Count of CHF diagnoses
CHF_NLP_heartfailure	0.36842715	Hitex non-negated NLP mentions of UMLS CUI C0018802: heart failure
CHF_NLP_ischemiccardiomyopathy	0.39455684	Hitex non-negated NLP mentions of UMLS CUI C0349782: ischemic cardiomyopathy
CHF_NLP_loopdiuretics	0.650788804	Hitex non-negated NLP mentions of UMLS CUI C0354100: loop diuretics

Below are the feature weights of the final Ischemic stroke phenotype algorithms. The weights below are used to derive a predicted probability of ISTR for every Biobank participant.

Feature_ID	Beta (weight)	Feature Description
(Intercept)	2.640	Model Intercept (beta 0)
Stroke_COD_DX_IschemicStroke	1.205	Count of coded diagnosis of ischemic stroke
Stroke_COD_DX_HemorrhagicStroke	-0.507	Count of coded diagnosis of hemorrhagic stroke
Stroke_NLP_brainattack	0.828	Count of non-negated NLP mentions of brain attack
Stroke_NLP_transientischaemicattack	-0.137	Count of non-negated NLP mentions of transient ischemic attack
patient_dxenct	-1.480	Total number of visits with a coded diagnosis

Genetic data

De-identified genetic data was requested from the MGBB. Whole blood samples are collected from participants during research draws or at the time of clinical draws. Samples are genotyped in batches using three versions of Illumina SNP array. Imputed genotype data from Multi-Ethnic Genotyping Array (MEGA), Expanded Multi-Ethnic Genotyping Array (MEGA Ex), and Multi-Ethnic Global (MEG) BeadChip were included in this analysis. Imputation was performed using the Michigan Imputation Server with a Minimac3. Only variants on the 22 autosomal chromosomes were considered. Data were obtained in 8 batches/sub-cohorts: Batch 1: MEG_A1_A (*n*=4780), Batch 2: MEG_A1_B (*n*=5020), Batch 3: MEG_C (*n*=5492), Batch 4: MEG_D (*n*=5146), Batch 5: MEG_E (*n*=4850), Batch 6: MEG_X1 (*n*=866), Batch 7: MEGA (*n*=4924), Batch 8: MEGAEX (*n*=5344).

Preprocessing and QC of genetic data

QC was performed on each batch individually and merged together, resulting in a total of 36,422 individuals and 9,036,179 variants after imputation. Unique ID's (rsID's) of variants were inserted based on chromosomal locations from the dbSNP database. Variants were filtered with two more criteria: minor allele frequency (MAF>0.01) and Hardy-Weinberg equilibrium (HWE p>1E-6). This resulted in 6,001,335 SNPs.

Mendelian randomization analyses

Genetic correlations and phenotypic comorbidity between PTSD/MDD and hypertension (and cardiovascular illnesses in general) have been repeatedly reported. In the current manuscript we demonstrated significant genetic correlations between the two psychiatric disorders and cardiovascular illnesses with a robust statistical evidence. These correlations observed at various levels may stem from one of the following four scenarios/mechanisms.

- Scenario 1: Onset of MDD/PTSD causes higher predisposition to hypertension and other cardiovascular illnesses. This could possibly be mediated by lifestyle changes (such as smoking and less exercise/mobility) and subsequent decline in anthropometric measurements (such as higher BMI).
- Scenario 2: Existence of cardiovascular illnesses leads to elevated predisposition to MDD/PTSD.
- Scenario 3: Both MDD/PTSD and cardiovascular illness are caused by shared genetic variants. This shared genetic underpinning can possibly influence distinct pathways/processed in the two groups of disorders. This is commonly known as horizontal pleiotropy.

- Scenario 4: MDD/PTSD and hypertension are caused by separate genetic variants that happen to be in linkage disequilibrium with each other.

It is important to note previous publications suggesting positive causal effect of MDD on cardiovascular diseases (Scenario 1 above) [PMID: 33032663], [PMID: 33372528]. Another interesting observation is the significant difference in age distribution between the two group of disorders. In the MGBB data, the median age for PTSD-positive and MDD-positive patients are 51 and 62; while the median age for hypertension and CAD positive patients are 70 and 75, respectively. Hence, those diagnosed with cardiovascular illnesses are, on average, significantly older than those diagnosed with the psychiatric disorders. Even if this is not observed on a longitudinal follow-up observation, it is suggestive of the sequence of which disorder is likely to be diagnosed first.

Here, we hypothesize that there is a causal link from MDD/PTSD to hypertension. The gold standard for inferring a causal link is Randomized control trial (RCT). However, in most situations, RCT is not feasible and requires a long timeframe to implement. This holds true in our current dataset as well. A feasible alternative that is cost-effective and implementable using our current study data is Mendelian Randomization (MR) [PMID: 34698778]. MR can be thought of as a 'randomized trial' where randomization is done at birth, where predisposing genetic variants are assorted randomly from parents to offspring [PMID: 19509388]. In other words, the predisposing genetic variants can be thought of as proxies for randomized intervention (assuming random mating).

MR is a special case of what is commonly known as instrumental variable analysis where genetic variation is used as the instrumental variable [PMID: 26282889]. For a variable to be

used as an instrumental variable (genetic variants in MR case), three main assumptions should hold true: (1) Relevance: the genetic variant is associated with the exposure, (2) Exclusion restriction: the genetic variant does not affect the outcome directly, and (3) Random assignment: the genetic variant does not affect the outcome through confounding variables. The main goal is to infer a causal link from the exposure variable to the outcome variable.

To test the causal link from MDD/PTSD to hypertension using MR, MDD/PTSD is our exposure variable and hypertension is the outcome variable. There are two possible choices for genetic variants as instrumental variables. The first and more common approach is to select few individual variants that are significantly associated with the exposure. MR analysis is conducted on each genetic variant as the instrumental variable and the results are combined afterwards. The second approach is to use polygenic risk scores as instrumental variables. The use of polygenic scores instead of using multiple individual variants (by far the most common approach) has been shown to mitigate instrumental variable bias (PMID:24062299).

Here, we use polygenic scores computed in the current manuscript to conduct the MR procedure. We implemented a procedure commonly known as two stage least square regression (2SLS) [PMID: 24114802]. In the first stage, the exposure (binary MDD label) is regressed against MDD polygenic risk score and along with confounders we want to control for (age and the first five principal components). Using this regression model, predicted value of the exposure is computed (MDD_hat). In the second stage, the outcome variable (hypertension binary label) is regressed against the predicted exposure variable (MDD_hat) and the potential confounding variables (age and the first five principal components). Then, (epidemiological) causality is inferred based on the significance of the coefficient of the predicted exposure variable (MDD_hat). The result of the statistics of the coefficient is shown below.

Estimate	Std. Error	p	Significance
0.428	0.095	5.97E-06	***

Therefore, we can infer that MDD onset is linked to hypertension diagnosis. We implemented a similar MR procedure for three other models. The results are shown in the table below.

Model	beta	Std. Error	p
MDD --> hypertension	0.428	0.095	5.97e-06
MDD --> CAD	0.466	0.141	9.22e-04
PTSD --> hypertension	0.201	0.044	5.97e-06
PTSD --> CAD	0.219	0.066	9.22e-04

Model	beta	Std. Error	p
hypertension --> MDD	0.054	0.048	0.258
hypertension --> PTSD	0.144	0.068	0.0345
CAD --> MDD	0.157	0.075	0.0373
CAD --> PTSD	0.310	0.108	0.0041

As shown, our analysis provides support for a causal link from psychiatric disorders to cardiovascular illnesses, but not the other way around.

Note that all analysis is only for EUR sub-population of the MGBB dataset. In the future, similar analysis on sub-populations (based on ancestry, gender and age) needs to be conducted. All analysis is done on R statistical software.

CRP analyses

We investigated genetic level correlations between the two groups of phenotypes and the largest publicly available CRP GWAS summary statistics (PMID: 33462484). As expected, CRP level in the blood is genetically correlated with both group of phenotypes. The main result is summarized below.

Summary statistics	r_g	p
Hypertension_meta	0.3183	5.98E-16
UKBB_Essential_Hypertension	0.3206	6.06E-16
CAD_meta	0.2254	1.04E-07
UKBB_broad_Depression	0.2326	4.63E-07
PGC_MDD2018_ex23andMe	0.1691	2.70E-05
MDD_meta	0.1622	5.46E-05
PTSD_meta	0.2211	0.000138
pts_eur_freeze2	0.1859	0.001913

Supplementary Tables and Figures

****** *Figures S5–S22 are provided in a separate file.* ******

		estimated				
		EUR	AFR	EAS	SAS	Total
self-identified	White	30296	107	13	318	30734
	Black	31	1759	0	19	1809
	Asian	17	10	487	274	788
	Other	327	222	12	541	1102
	Unknown	1087	315	34	552	1988
	Total	31758	2413	546	1704	36421

TABLE S1. Self-identified and genetically predicted ancestry composition. Genetic analysis was done with those participants predicted to have EUR ancestry.

TABLE S2. List of publicly available summary statistics used in the current study

Study name	Phenotype	Type of variable	Cases/controls or n	Sample description	No. of variants	SNP heritability (SE)	Reference
ptsd_eur_freeze2	PTSD	binary	23,212/151,447	PGC freez-2 EUR	9.77M	0.063 (0.011)	(PMID:31594949)
PGC_MDD2018_ex23andMe	MDD	binary	59,851/113,154	PGC	13.5M	0.076 (0.005)	(PMID:29700475)
UKBB_probable_MDD	Probable MDD	binary	30,603/143,916	UKBB	7.67M	0.022 (0.003)	(PMID:29662059)
UKBB_ICD_MDD	MDD	binary	8,276/209,308	UKBB	7.67M	0.019 (0.002)	(PMID:29662059)
UKBB_broad_Depression	Depression	binary	113,769/208,811	UKBB	7.67M	0.063 (0.003)	(PMID:29662059)
UKBB_Essential_Hypertension	Hypertension	binary	77,723/330,366	UKBB	28.3M	0.074 (0.004)	(PMID:32589924)
UKBB_doctor_highBP	High blood pressure	binary	144,793/313,761	UKBB	5.26M	0.127 (0.006)	(PMID:30940143)
Hypertension_diverse	Hypertension	binary	27,123/22,018	Ancestrally diverse cohort	28.3M	0.042 (0.009)	(PMID:31217584)
CAD_Nikpay	CAD	binary	60,801/123,504	mostly EUR	6.7M	0.076 (0.008)	(PMID:26343387)
UKBB_ICBP_SBP	SBP	quantitative	757K	UKBB & ICBP*	7.4M	0.140 (0.006)	(PMID:30224653)
UKBB_ICBP_DBP	DBP	quantitative	757K	UKBB & ICBP*	7.4M	0.136 (0.006)	(PMID:30224653)
UKBB_ICBP_PP	pulse pressure (PP=SBP-DBP)	quantitative	757K	UKBB & ICBP*	7.4M	0.123 (0.005)	(PMID:30224653)
UKBB_RHR	HR resting	quantitative	460K	UKBB	5.26M	0.152 (0.012)	(PMID:30940143)
HR50	HR recovery	quantitative	58.8K	EUR	14.7M	0.030 (0.005)	(PMID:29497042)
HRinc	HR increase	quantitative	58.8K	EUR	14.7M	0.03 (0.003)	(PMID:29497042)

Note. PTSD = posttraumatic stress disorder; CVD = cardiovascular disease; MDD = major depressive disorder; CAD = coronary artery disease; SBP = systolic blood pressure; DBP = diastolic blood pressure; HR = heart rate; SNP = single nucleotide polymorphism; SE = standard error; UKBB & ICBP: UKBB (n=458) + International Consortium for BP GWAS (n=150K+149K).

TABLE S3. Genetic correlations among PTSD/MDD and CVD (hypertension/CAD)

	PGC_MDD2018	**UKBB_broad_Depression**	**UKBB_Essential_Hypertension**	**Nikpay_CAD**
	rG (S.E.), p	rG (S.E.), p	rG (S.E.), p	rG (S.E.), p
ptsd_eur_freeze2	0.78 (0.086), 1.99E-19	0.67 (0.074), 1.93E-19	0.34 (0.061), 1.91E-08	0.27 (0.082), 8.83E-4
PGC_MDD2018_ex23andMe		0.92 (0.029), 1.65E-225	0.28 (0.029), 6.74E-22	0.18 (0.046), 1.34E-4
UKBB_broad_Depression			0.22 (0.028), 3.07E-15	0.19 (0.039), 1.01E-06
UKBB_Essential_Hypertension				0.55 (0.038), 6.01E-47

Note. Correlation results using publicly available summary statistics.

TABLE 4. Mendelian randomization results suggest causal pathway from PTSD and MDD to hypertension

Model	Estimate	Std. Error	p	
MDD --> hypertension	0.3970	9.31E-02	2.04E-05	***
hypertension --> MDD	0.0539	4.76E-02	0.257656	
PTSD --> hypertension	0.3818	0.104408	0.000255	***
hypertension --> PTSD	0.0539	4.76E-02	0.257656	

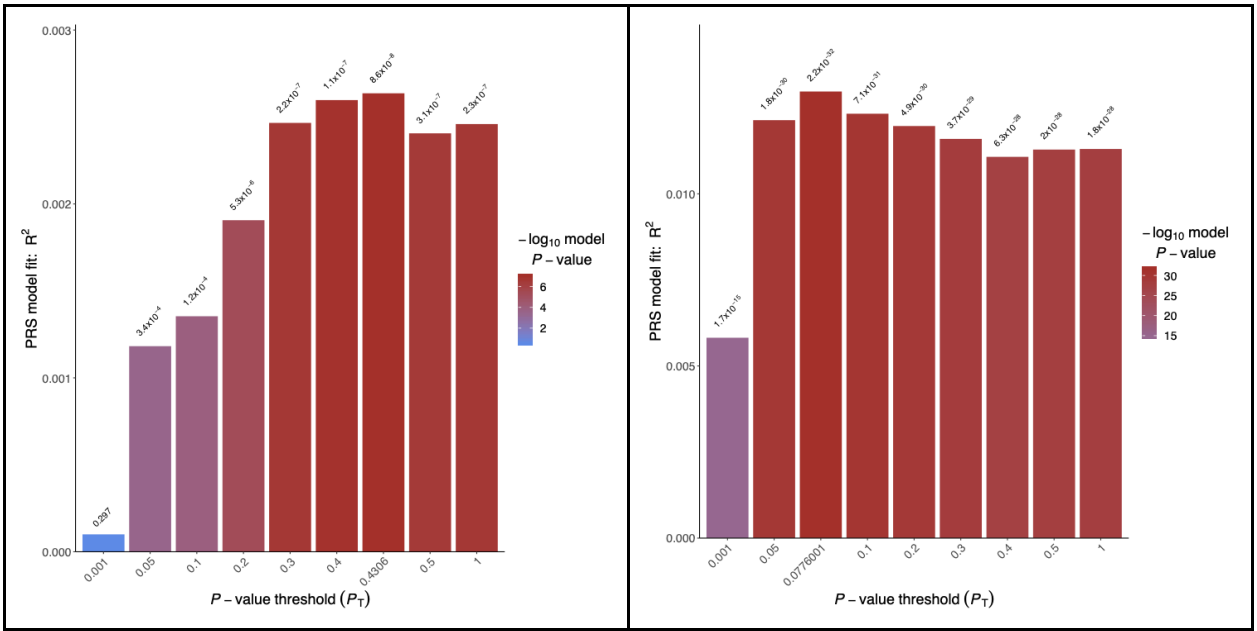

FIGURE S1. Thresholding step with MTAG summary statistics for PTSD polygenic risk score (PRS)

Base summary stats	Target phenotype	PRS R^2	PRS P-value	P-value threshold	Num_SNP
CAD-Hypertension	CAD	0.0104702	9.99E-42	0.1862	22,138
CAD-Hypertension-PTSD-MDD	CAD	0.0106199	2.89E-42	0.14535	18,242
Hypertension-CAD	hypertension	0.0169782	7.26E-114	0.0265001	6,743
Hypertension-CAD-PTSD-MDD	hypertension	0.0172907	6.72E-116	0.0420001	8,981

FIGURE S2. Improvement in the prediction of coronary artery disease (CAD) and hypertension using PTSD and MDD summary statistics

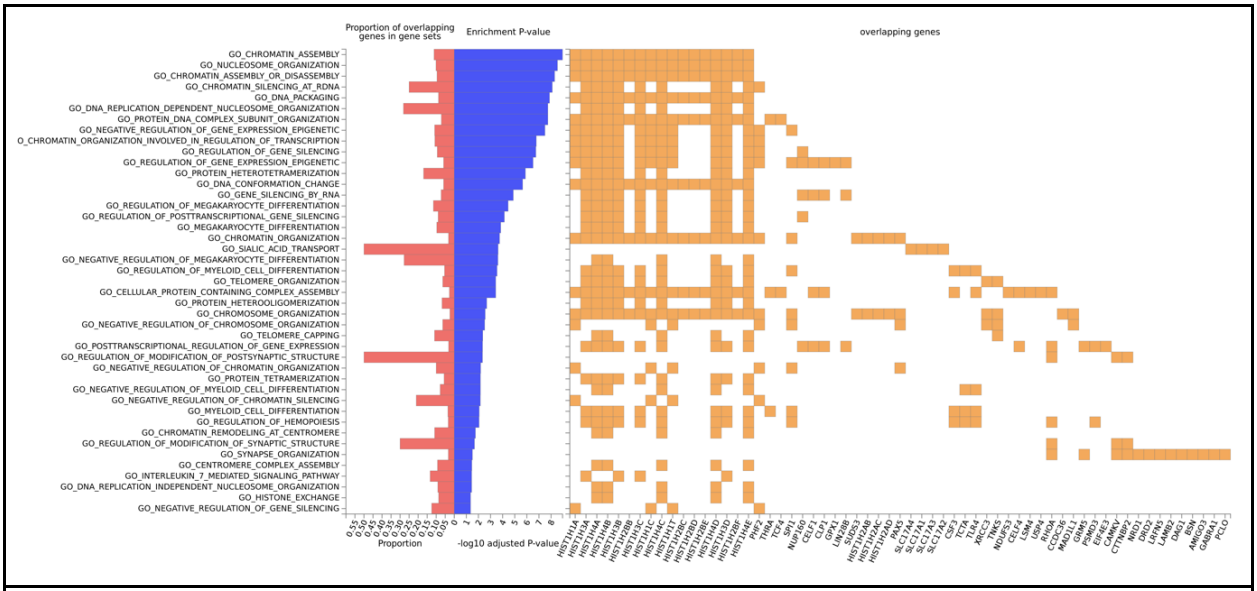

FIGURE S3. Genomic risk loci included in pathway analysis

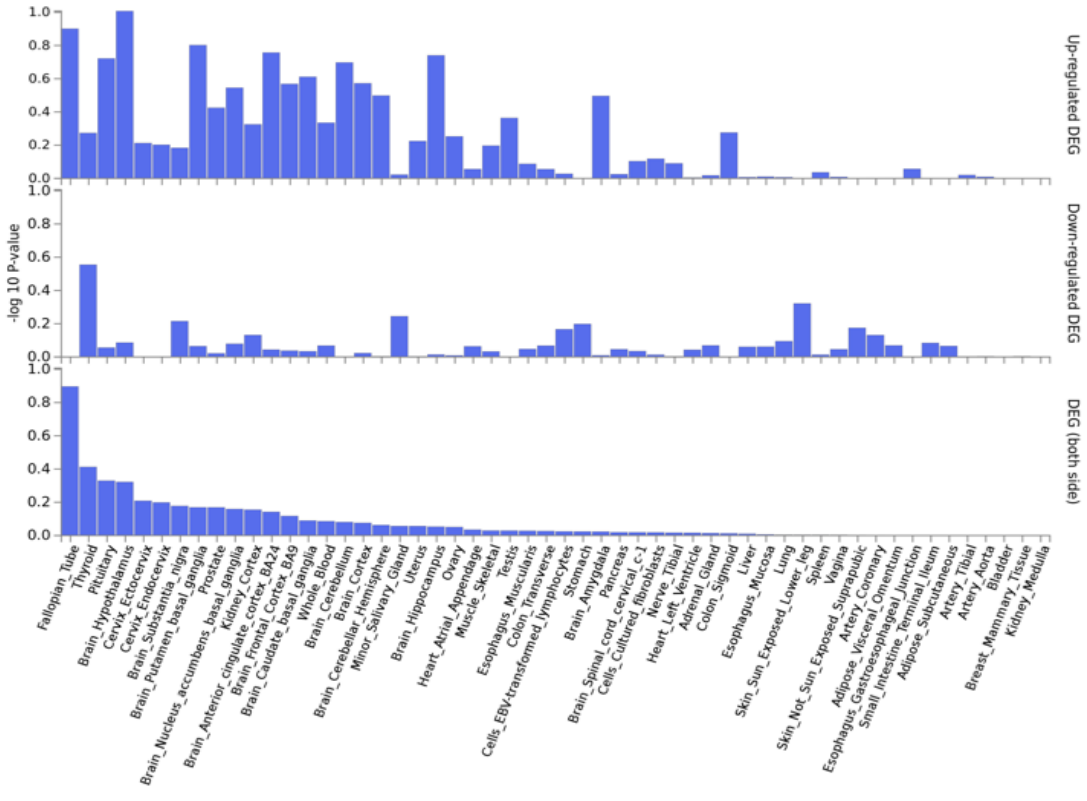

FIGURE S4. Tissue specificity of pathway analysis