Supplementary Material for "TiSAn: Estimating Tissue Specific Effects of Coding and Noncoding Variants"

Kévin Vervier ,Jacob J Michaelson

March 29, 2018

TiSAn enrichment in tissue-specific transcriptome from GTEx consortium

We consider 44 tissues characterized by RNA-seq, from the GTEx project, including 2 heart tissues (atrial appendage and left ventricle), and 10 brain tissues. For each gene, we compute the average TiSAn score profile using both brain and heart models. We also derive an average TiSAn score for each gene flanking region (+/-10kb). Figure S10 shows the correlation between the measured gene expression (in RPKM) and the corresponding TiSAn average score across all tissues.

TiSAn-brain annotates 9 out of 10 brain regions among the highest positive correlations, when considering gene region (Fig. S11A), but also when considering flanking regions only (Fig. S11B), providing support that TiSAn scores are informative for non-genic regions. TiSAn-heart also shows a strong heart-related enrichment for both genic (Fig. S11C) and flanking regions (Fig. S11D). Interestingly, Liver and Pancreas also show high positive correlations with TISAn score, which is reasonable given their known implication in hypercholesterolemia or diabetes, conditions that have a direct bearing on cardiovascular health.

Tissue-specific signal in transcription factor binding sites

Transcription factor (TF) binding sites (TFBS) are associated with observed differences in gene expression across tissues [4]. We hypothesized that computing TiSAn score profiles in TFBS could provide insight about the tissue-related action of specific TFs. The ENCODE project provides TFBS detection in 80 different cell types for more than 50 TFs. TiSAn scores were predicted for millions of loci using a 1,000bp window centered on TFBS. Average TiSAn profiles for each TF allow us to identify the sites showing an overall enrichment across cell types. This enrichment is measured by comparing TiSAn score at the TFBS location (center), and on

the flanking regions. Statistical tests were performed by comparing the values observed on the central region ($[334:666]$bp) to flanking regions ($[1:333] \bigcup [667:1000]$bp). For instance, strong heart-related signal was found among TFBS for BHLHE40, CEBPB, FOXA1, GATA1, HNF4A, JUN, MAFK, MAX, MYC, POU2F2, STAT1, and TAL1. Notably, CEBPB TFBS are significantly enriched for TiSAn-heart score in 6 cell types (t-test Bonferoni corrected P-values < 0.05), including two related to smooth muscles (A549 and IMR90) and one related to liver (HepG2) (Supplementary Figure 17a), and CEBPB has been associated with cardiac hypertrophy [1] and fatty liver disease [2].

Functional enrichment patterns were also found for the critical brain transcription factor REST in 10 different cell types (Supplementary Figure 17b). 6 of those 10 cell types are significantly enriched in the central region of the TFBS (t-test Bonferoni corrected P-value < 0.05). Among these significantly enriched cell lines, 3 were brain cancer cell lines (U87, SK-N-SH, and PFSK-1), which support recent findings on the importance of REST in neuroblastoma drug sensitivity(29).

References

[1] Redondo-Angulo, I., et al. (2016) C/EBPbeta is required in pregnancy-induced cardiac hypertrophy, *International journal of cardiology*, **202**, 819-828.

[2] Sookoian, S., et al. (2017) Genetic variation in long noncoding RNAs and the risk of nonalcoholic fatty liver disease, *Oncotarget*, **8**, 22917-22926.

[3] Spiers, H., et al. (2015) Methylomic trajectories across human fetal brain development, *Genome Res*, **25**, 338-352.

[4] Zhong, S., He, X. and Bar-Joseph, Z. (2013) Predicting tissue specific transcription factor binding sites, *BMC genomics*, **14**, 796.

Tissue	Category	Count
Brain	Positive	10,097 (5,535 + 4,562)
Brain	Negative	19,699 (12,305 + 7,394)
Heart	Positive	21,248 (7,476 + 13,772)
Heart	Negative	28,743 (9,760 + 18,713)

Table S1: Training set composition. For both heart and brain tissues, we report the count of positive and negative examples used to train TiSAn models. The counts are divided in two parts: the first number corresponds to variants found in large intergenic non-coding RNAs database LincSNP, and the second number to genotype array probesets (PsychArray or MetaboChip).

Features	Database	Summary
n-nucleotide frequencies	Hg19 genome fasta	$\pm$500bp neighborhood, $n \in (1, 2, 3, 4)$
distance to the closest tissue eQTL	GTEx v6	Weibull distance
is it a tissue eQTL?	GTEx v6	binary value
distance to the closest eQTL	GTEx v6	Weibull distance
is it a eQTL?	GTEx v6	binary value
distance to the closest tissue gene	PubMed gene2ID	Weibull distance
is it in a tissue gene?	PubMed gene2ID	binary value
distance to the closest gene	PubMed gene2ID	Weibull distance
is it in a gene?	PubMed gene2ID	binary value
distance to the closest methylated region	RoadMap Epigenomics	Weibull distance
methylation level in tissue cell lines	RoadMap Epigenomics	if the position falls in a methylated region
methylation level in other cell lines	RoadMap Epigenomics	if the position falls in a methylated region
distance to the closest dDMR	Fetal brain from [3]	Weibull distance (brain model only)
is it in a dDMR?	Fetal brain from [3]	binary value
distance to the closest dDMP	Fetal brain from [3]	Weibull distance (brain model only)
is it in a dDMP?	Fetal brain from [3]	binary value (brain model only)
distance to the closest heart enhancer	Heart Enhancer Compendium	Weibull distance (heart model only)
is it in an heart enhancer?	Heart Enhancer Compendium	binary value (heart model only)

Table S2: Feature space description. For each variable used to train predictive models, we report the corresponding public database, and a brief summary of its content. eQTL stands for Expression quantitative trait loci. dDMR stands for developmentally differentially methylated region, and dDMP for developmentally differentially methylated position.

Tissue	compositional	GTEx eQTL	RME methylation	literature
Adipose	yes	yes	yes	yes
Adrenal gland	yes	yes	yes	yes
Bone	yes	no	yes	yes
Brain	yes	yes	yes	yes
Breast	yes	yes	yes	yes
Colon	yes	yes	yes	yes
Esophagus	yes	yes	yes	yes
Heart	yes	yes	yes	yes
Intestine	yes	yes	yes	yes
Kidney	yes	no	yes	yes
Liver	yes	yes	yes	yes
Lung	yes	yes	yes	yes
Muscle	yes	yes	yes	yes
Nerve	yes	yes	no	yes
Ovary	yes	yes	yes	yes
Pancreas	yes	yes	yes	yes
Placenta	yes	no	yes	yes
Prostate	yes	yes	no	yes
Rectum	yes	no	yes	yes
Skin	yes	yes	yes	yes
Spleen	yes	yes	yes	yes
Stomach	yes	yes	yes	yes
Testis	yes	yes	no	yes
Thymus	yes	no	yes	yes
Thyroid	yes	yes	no	yes
Uterus	yes	yes	yes	yes
Vagina	yes	yes	no	yes
Whole Blood	yes	yes	yes	yes

Table S3: Feature availability for different tissues. For each tissue, we report if the features used to train either the brain or heart model could be derived for other tissues using the TiSAn-train tool. GTEx: Gene-Tissue Expression. RME: RoadMap Epigenomics. literature: genes found to be co-cited with the tissue of interest on PubMed.

Distribution	brain gene	non-brain gene	brain eQTL	non-brain eQTL	methylated region
Weibull	8597 (**1**)	26726 (2)	20939 (**1**)	95059 (2)	123156 (3)
Beta	8145 (3)	24885 (3)	20486 (3)	91568 (3)	125018 (**1**)
Log-Normal	7645	27003 (**1**)	20933 (2)	96032 (**1**)	123960 (2)
Exponential	8335 (2)	21491 (4)	9698 (4)	40844 (4)	28393 (4)

Table S4: Distance to annotations for different distributions. Estimations were done using 1,000 random genomic positions. Fit quality is reported as the Log-Likelihood, and for each annotation, the distributions were ranked. Bold numbers correspond to the best distribution for a given annotation.

Method	Cross-validated AUC
Random Forest	0.795
Logistic Regression	0.635
Support Vector Machine	0.584

Table S5: 10-folds cross-validated performances obtained during TiSAn-brain model training, for different classification strategies. AUC: Area under ROC curve.

Gene	Citations	Gene	Citations
NRXN1	61	*GABRB3*	30
CNTNAP2	49	*SCN2A*	30
SHANK3	49	*FOXP2*	28
PTEN	39	*RBFOX1*	28
CACNA1C	35	*SYNGAP1*	28
OXTR	34	*AUTS2*	27
RELN	34	*GRIN2B*	25
MET	32	*DPP6*	22
DISC1	31	*MBD5*	22
SCN1A	31	*SLC6A4*	22

Table S6: List of SFARI autism-related genes, supported by literature.

Tissue ID	Interaction	Stability
Brain	GC & proximity with brain eQTL	1
Brain	proximity with fetal brain dDMP & proximity with non-brain eQTL	1
Brain	proximity with brain gene & proximity with non-brain eQTL	0.95
Heart	proximity with fetal heart enhancer & proximity with heart eQTL	0.95
Heart	proximity with heart eQTL & proximity with non-heart eQTL	0.95

Table S7: Examples of features interactions in TiSAn models. Iterative Random Forest approach was used to identify combinations of features frequently occuring in decision trees. Stability values are estimated using 20 bootstrap samples on an optimized model by 10 iterations. *dDMP* stands for developmentally differentially methylated position.

LincSNP ID	Chr	Start	End	Associated disorder
LSLNC096364	12	2907933	2909631	Suicide attempts in bipolar disorder
LSLNC141017	12	2901746	2904044	Autism with low IQ
LSLNC169768	12	2901745	2904044	Autism with low IQ
LSLNC169511	12	1936729	1940267	Tourette's syndrome
LSLNC027117	12	1936728	1940267	Tourette's syndrome

Table S8: Non-coding RNAs found in linkage disequilibrium with neurodevelopmental and psychiatric disorders

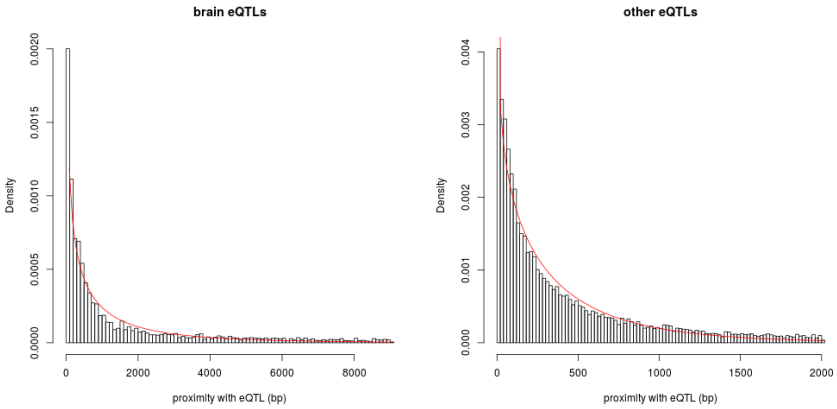

Figure S1: Distribution of distance to the closest GTEx expression quantitative trait locus (eQTL) for brain (left) and non-brain (right) tissues. The red lines correspond to a Weibull distribution fit. Estimated parameters for left (resp. right) figure are: shape = 0.351 (resp. 0.315) and scale = 21,888 (resp. 9,111). 10,000 random genomic loci were used to estimate the distributions.

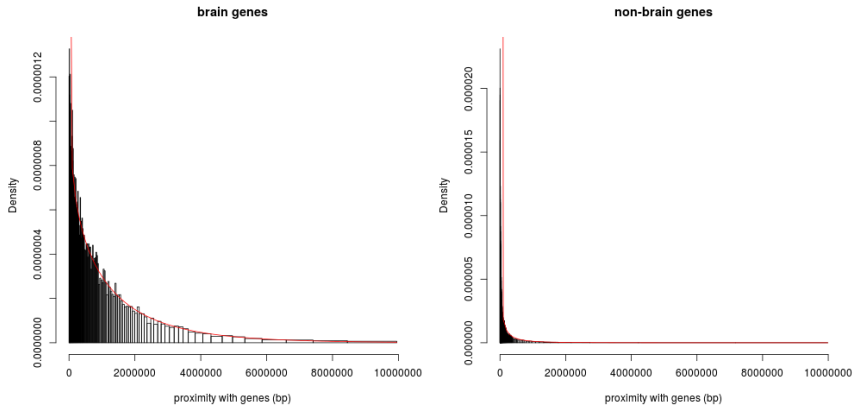

Figure S2: Distribution of distance to the closest gene for brain (left) and non-brain (right) tissues. The red lines correspond to a Weibull distribution fit. Estimated parameters for left (resp. right) figure are: shape = 0.852 (resp. 0.529) and scale = 1,453,217 (resp. 201,985). 10,000 random genomic loci were used to estimate the distributions.

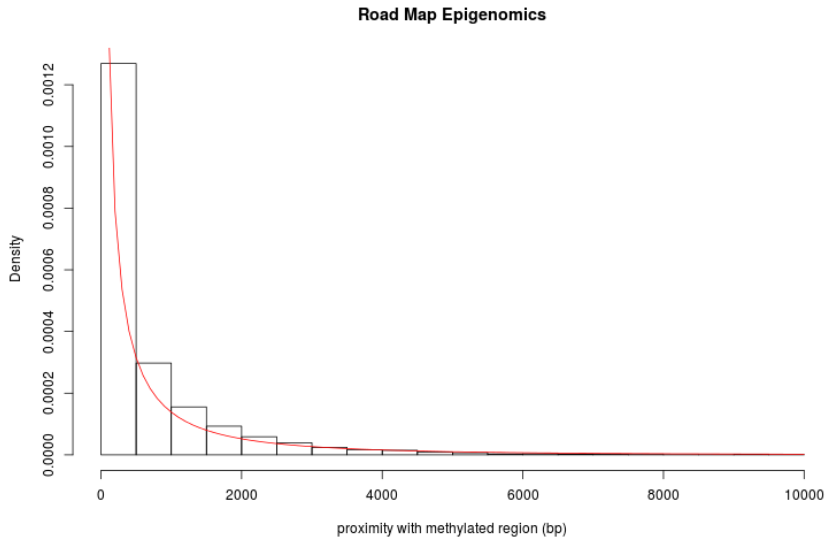

Figure S3: Distribution of distance to the closest methylated region found in RoadMap Epigenomics database. The red line corresponds to a Weibull distribution fit. Estimated parameters are: shape = 0.746 and scale = 590.3. 10,000 random genomic loci were used to estimate the distributions.

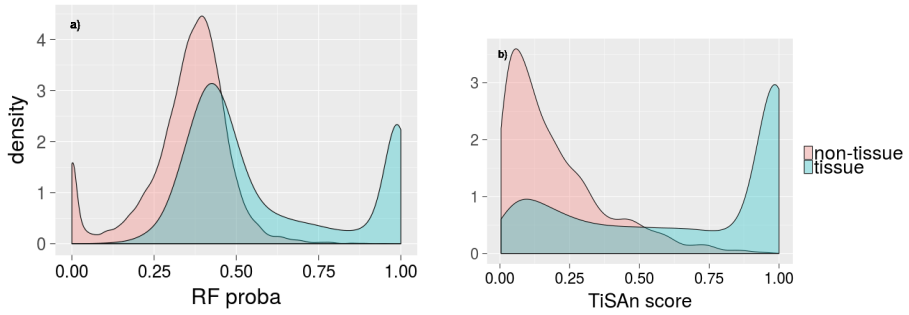

Figure S4: TiSAn-brain cross-validation performances. (a) Random forest raw output distribution. (b) TiSAn score obtained after rescaling odd-ratios. For clarity purpose, only strictly positive odd-ratios values are shown. See Supplementary Figure 8 for the distribution with zero scores.

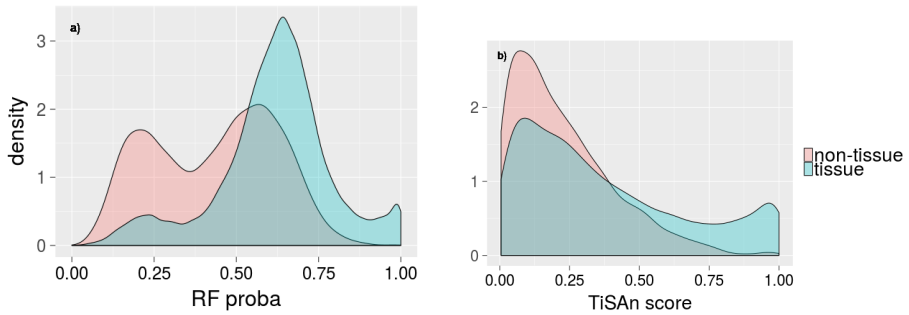

Figure S5: TiSAn-heart cross-validation performances. (a) Random forest raw output distribution. (b) TiSAn score obtained after rescaling odd-ratios. For clarity purpose, only strictly positive odd-ratios values are shown.

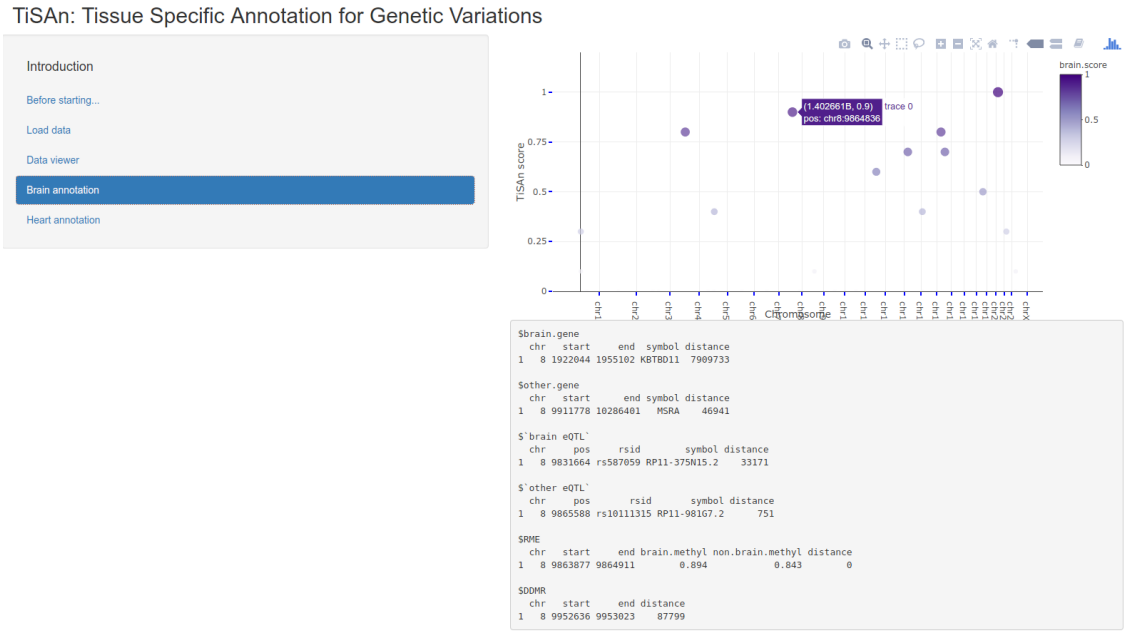

Figure S6: TiSAn-view application. In this Shiny-based tool, users can upload a short list of variants, and get detailed annotations on which features were used to score them.

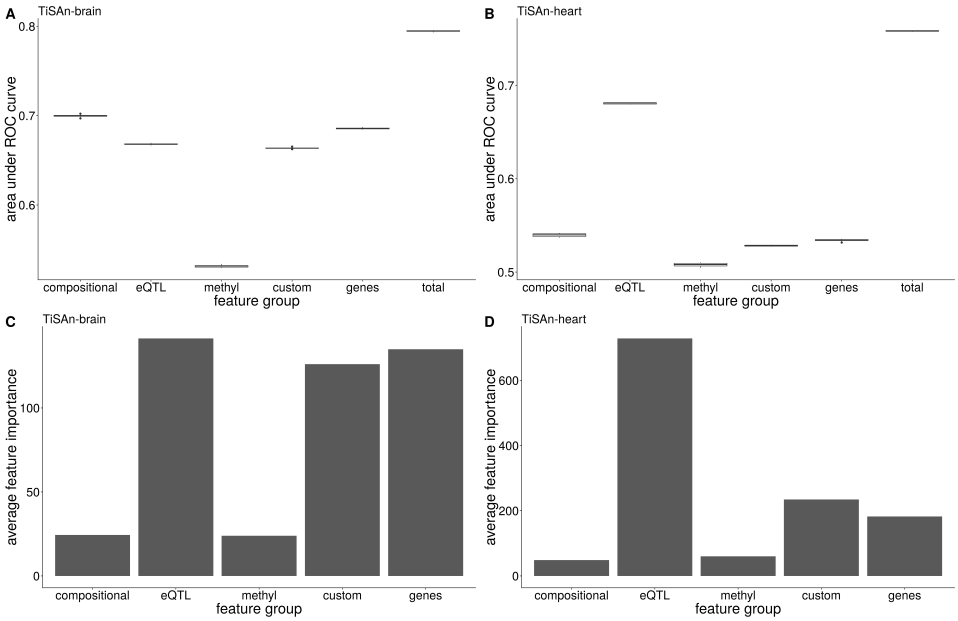

Figure S7: Features group importance in model predictive power. **(A)** TiSAn-brain model performances trained on different features groups. **(B)** TiSAn-heart model performances trained on different features groups. For each set of features, cross-validated area under the ROC curve were computed. **(C)** TiSAn-brain model features importance. **(D)** TiSAn-heart model features importance. For each set of features, average mean decrease in Gini index was computed. *eQTL*: expression quantitative trait loci refers to both tissue and non-tissue eQTL features (distance and binary). *methyl* refers to DNA methylation data from the RoadMap consortium (binary , distance and methylation level). *custom* refers to tissue-specific databases (brain: fetal brain development, heart: known enhancers). *genes* refers to both tissue and non-tissue genes (binary features and distance). *total* refers to the model trained using all the features.

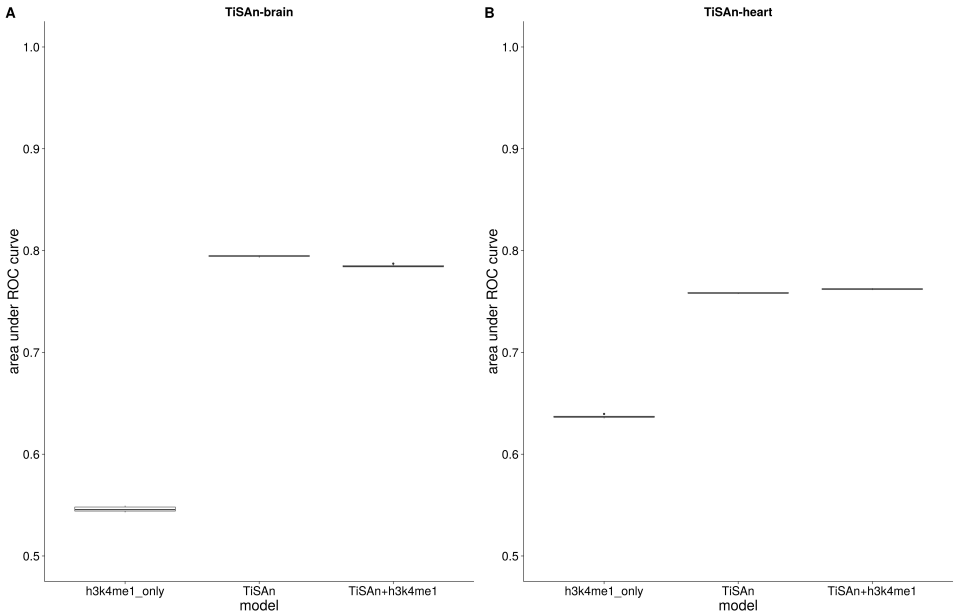

Figure S8: Model performances on additional epigenome data. **(A)** Brain models performances using different feature sets. **(B)**Heart models performances using different feature sets. For each set of features, 5-folds cross-validated area under the ROC curve were computed (10 repeats). *h3k4me1 only*: model only using binary and region methylation found in tissue and non-tissue samples. *TiSAn* refers to the model presented in the study. *TiSAn-h3k4me1* refers to a model trained on TiSAn features and the h3k4me1 epigenome data.

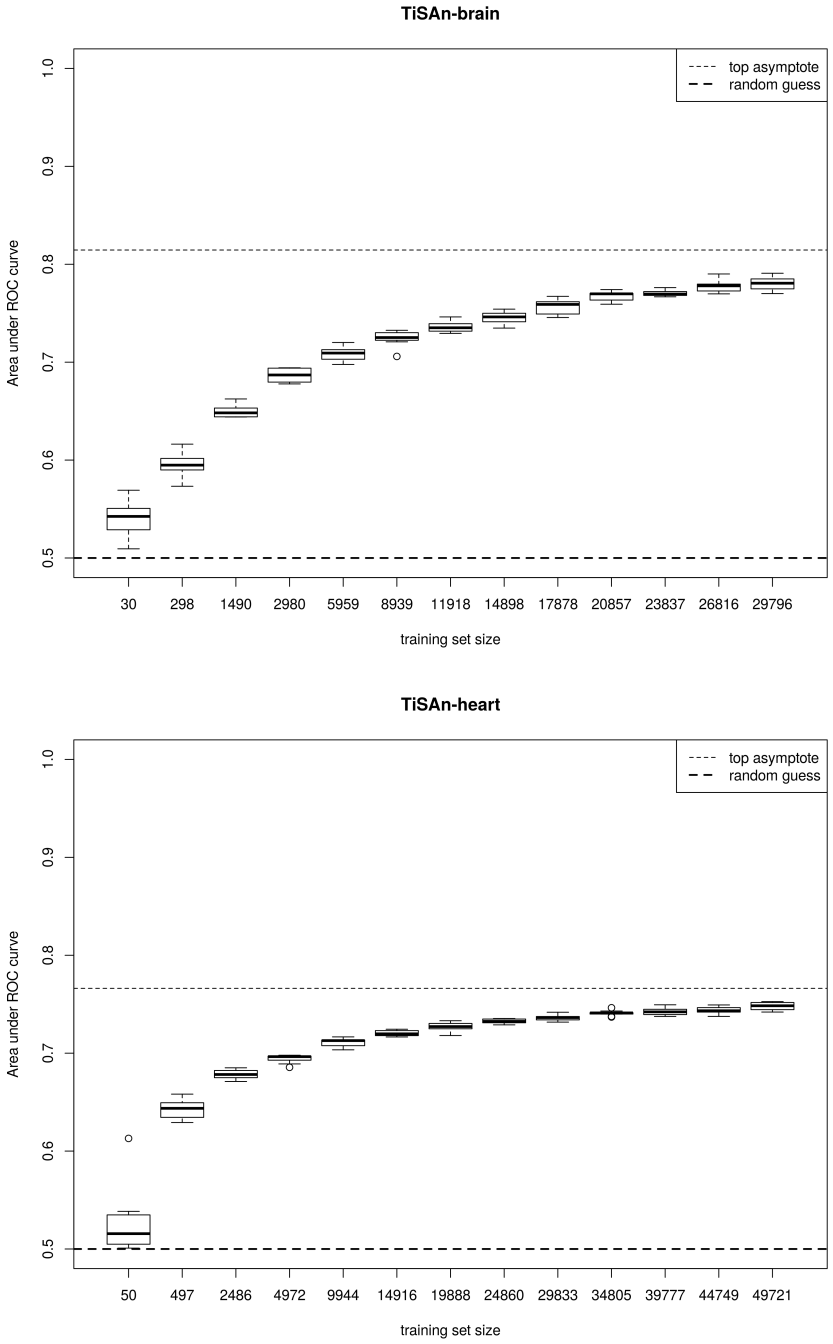

Figure S9: Learning curve and traing set size impact on model performances. (Top) Brain model. (Bottom) Heart model. For different subsets of the total training set, cross-validated area under the ROC curve were computed. Top asymptotes were obtained by fitting an exponential growth model based on the observed median values.

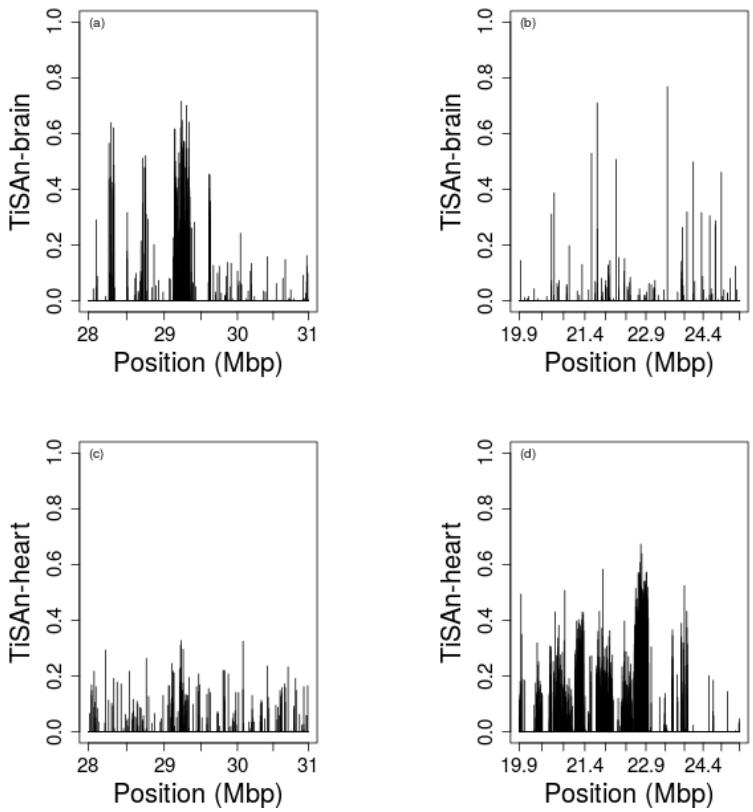

Figure S10: Region based analysis. Two disease-associated genomic regions (left: 16p11, right: 9p21) were held out during the model training. We annotated one locus every kilobase with both TiSAn-brain (top) and TiSAn-Heart (bottom) and represent the score distribution along those regions.

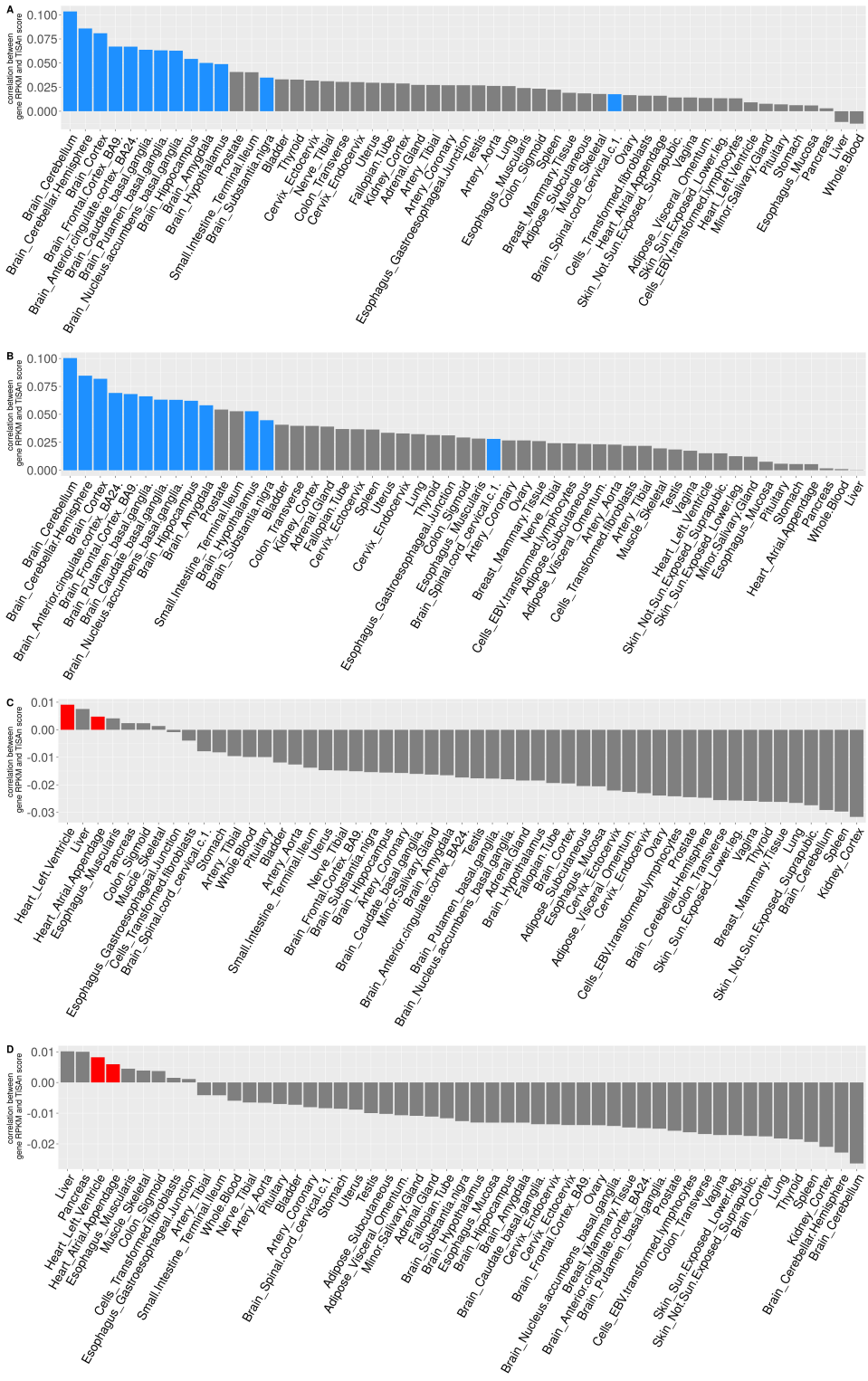

Figure S11: **Correlation between tissue gene expression and TiSAn scores**. Across 44 tissues transcriptomes and 17,803 genes, we compute the average TiSAn score at each location and report the Spearman correlation between level of expression and TiSAn score. (**A**) TiSAn-brain score enrichment computed over the full gene body. (**B**) TiSAn-brain score enrichment computed over flanking regions of genes (±10kb). (**C**) TiSAn-heart score enrichment computed over the full gene body. (**D**) TiSAn-heart score enrichment computed over flanking regions of genes (±10kb). Bars colored in blue (resp. in red) are tagged in GTEX as brain (resp. heart) tissues.

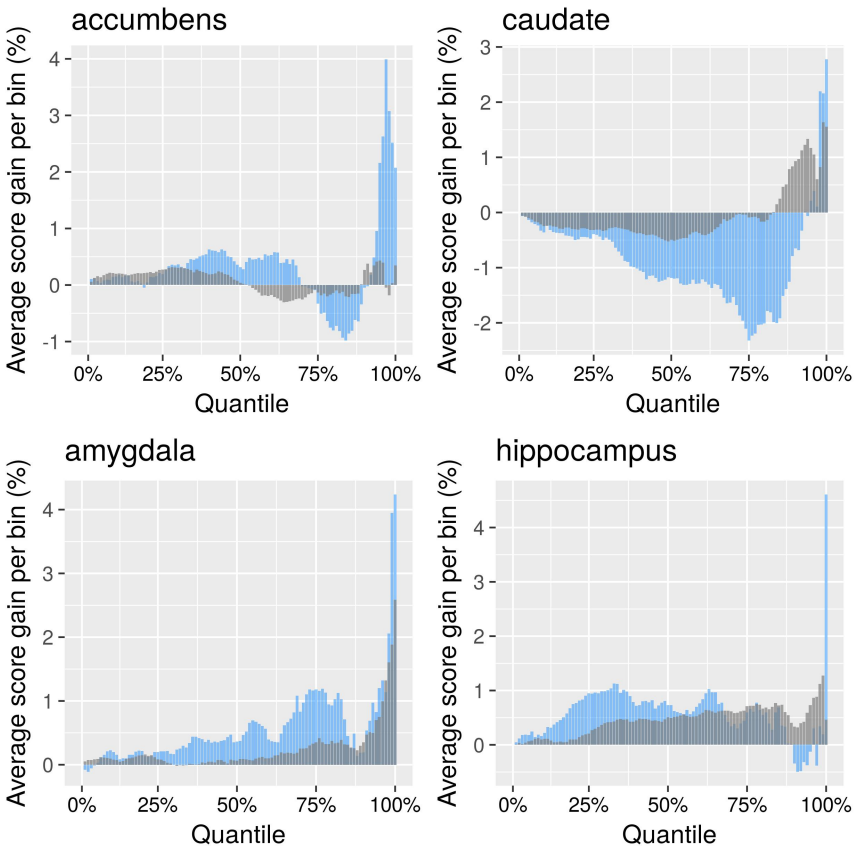

Figure S12: **TiSAn annotation for both brain (blue) and heart (grey) scores of intergenic loci associated with brain volume**. Each of 974,045 ENIGMA loci was binned, into 100 percentile groups, based on the statistical association strength with brain volume. For each bin, the average TiSAn score enrichment is computed with respect to the average score across the entire set of loci. The right part of each panel corresponds to loci with the stronger association with the brain region volume.

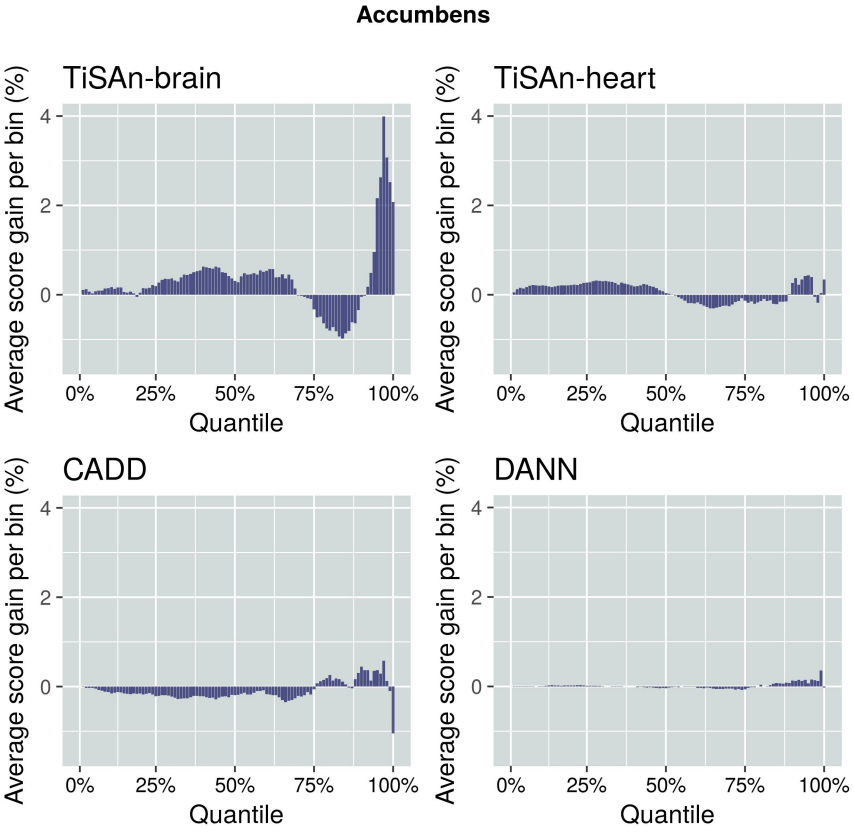

Figure S13: **Functional annotation for intergenic loci associated with nucleus accumbens volume**. Each of 974,045 ENIGMA loci was binned, into 100 percentile groups, based on the statistical association strength with nucleus accumbens volume. For each bin, average functional score enrichment is computed with respect to the average score across the entire set of loci. The right part of each panel corresponds to loci with the stronger association with the nucleus accumbens volume. Top left: TiSAn-brain, Top right: TiSAn-heart, Bottom left: CADD, Bottom right: DANN.

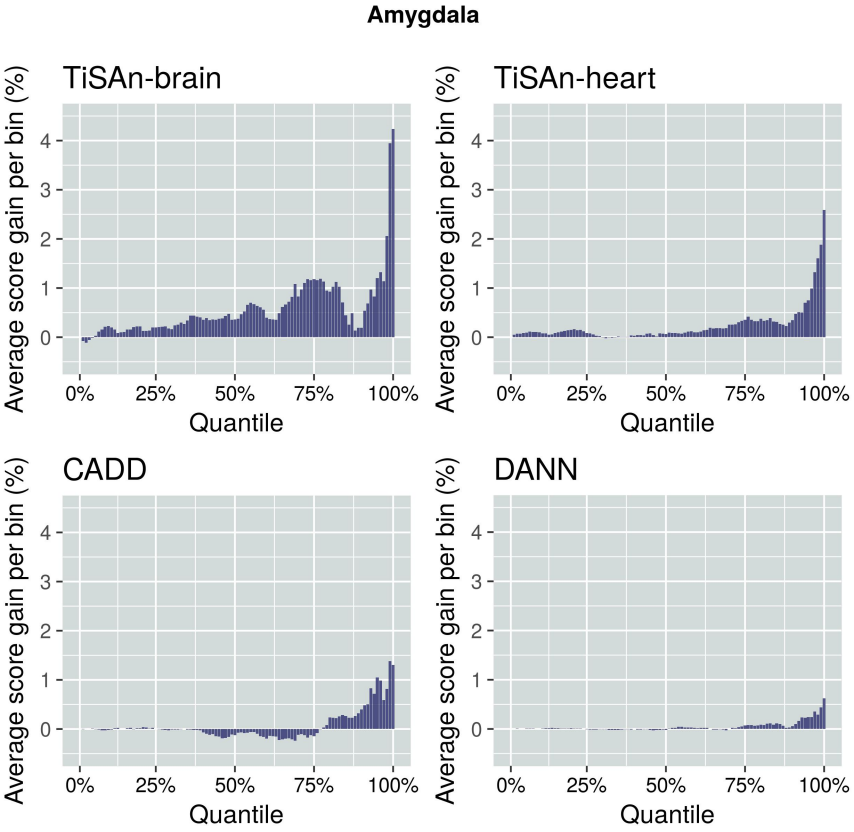

Figure S14: **Functional annotation for intergenic loci associated with amygdala volume**. Each of 974,045 ENIGMA loci was binned, into 100 percentile groups, based on the statistical association strength with amygdala volume. For each bin, average functional score enrichment is computed with respect to the average score across the entire set of loci. The right part of each panel corresponds to loci with the stronger association with the amygdala volume. Top left: TiSAn-brain, Top right: TiSAn-heart, Bottom left: CADD, Bottom right: DANN.

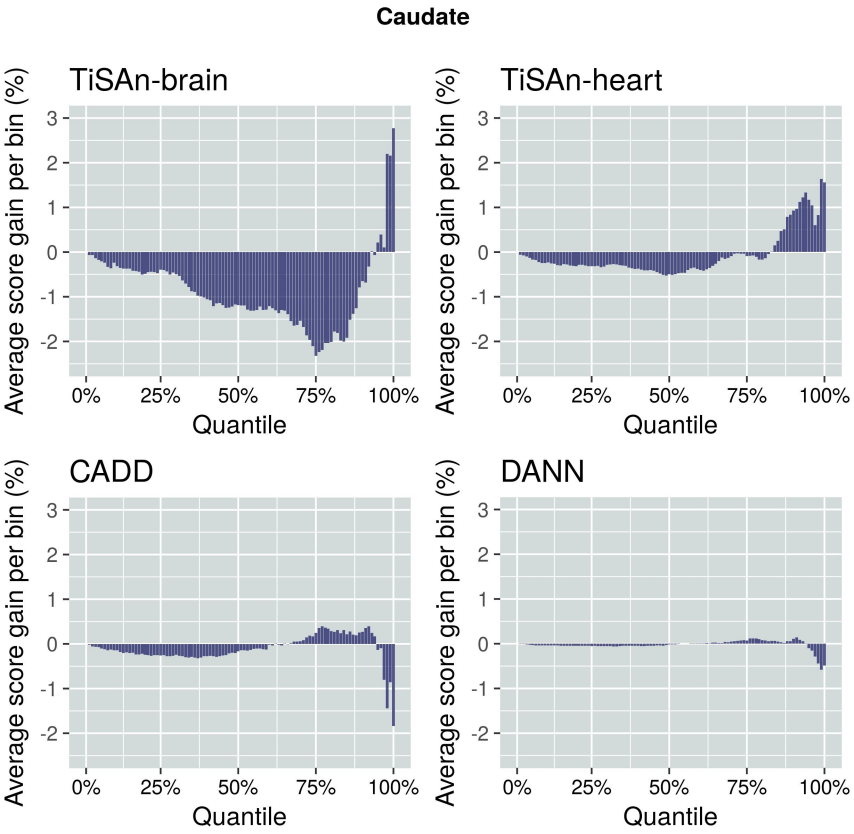

Figure S15: **Functional annotation for intergenic loci associated with caudate volume**. Each of 974,045 ENIGMA loci was binned, into 100 percentile groups, based on the statistical association strength with caudate volume. For each bin, average functional score enrichment is computed with respect to the average score across the entire set of loci. The right part of each panel corresponds to loci with the stronger association with the caudate volume. Top left: TiSAn-brain, Top right: TiSAn-heart, Bottom left: CADD, Bottom right: DANN.

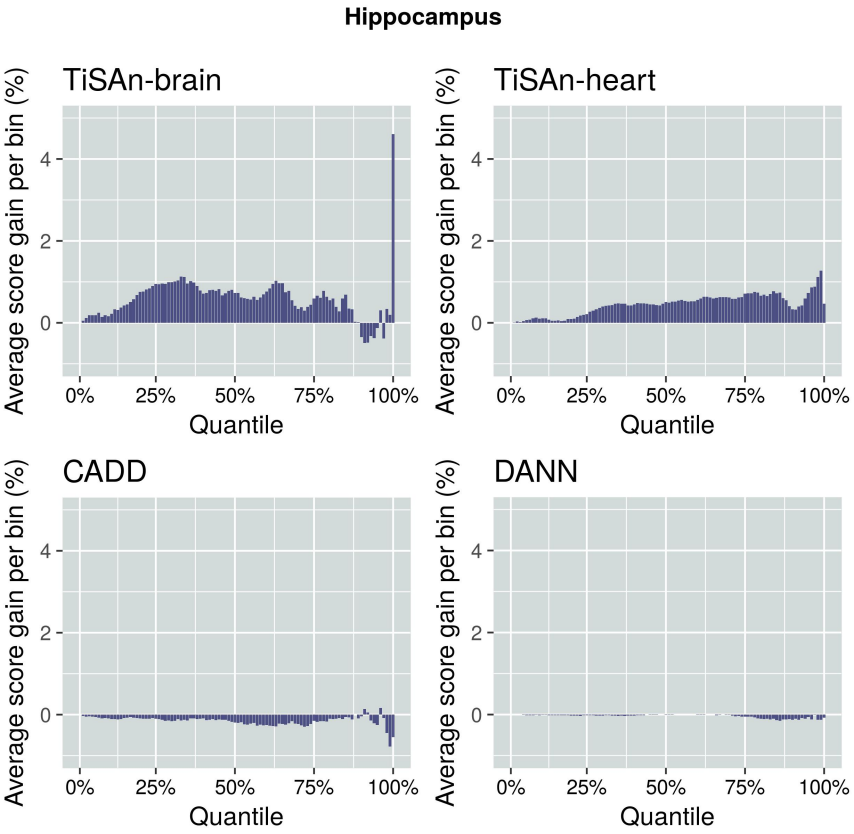

Figure S16: **Functional annotation for intergenic loci associated with hippocampus volume**. Each of 974,045 ENIGMA loci was binned, into 100 percentile groups, based on the statistical association strength with hippocampus volume. For each bin, average functional score enrichment is computed with respect to the average score across the entire set of loci. The right part of each panel corresponds to loci with the stronger association with the hippocampus volume. Top left: TiSAn-brain, Top right: TiSAn-heart, Bottom left: CADD, Bottom right: DANN.

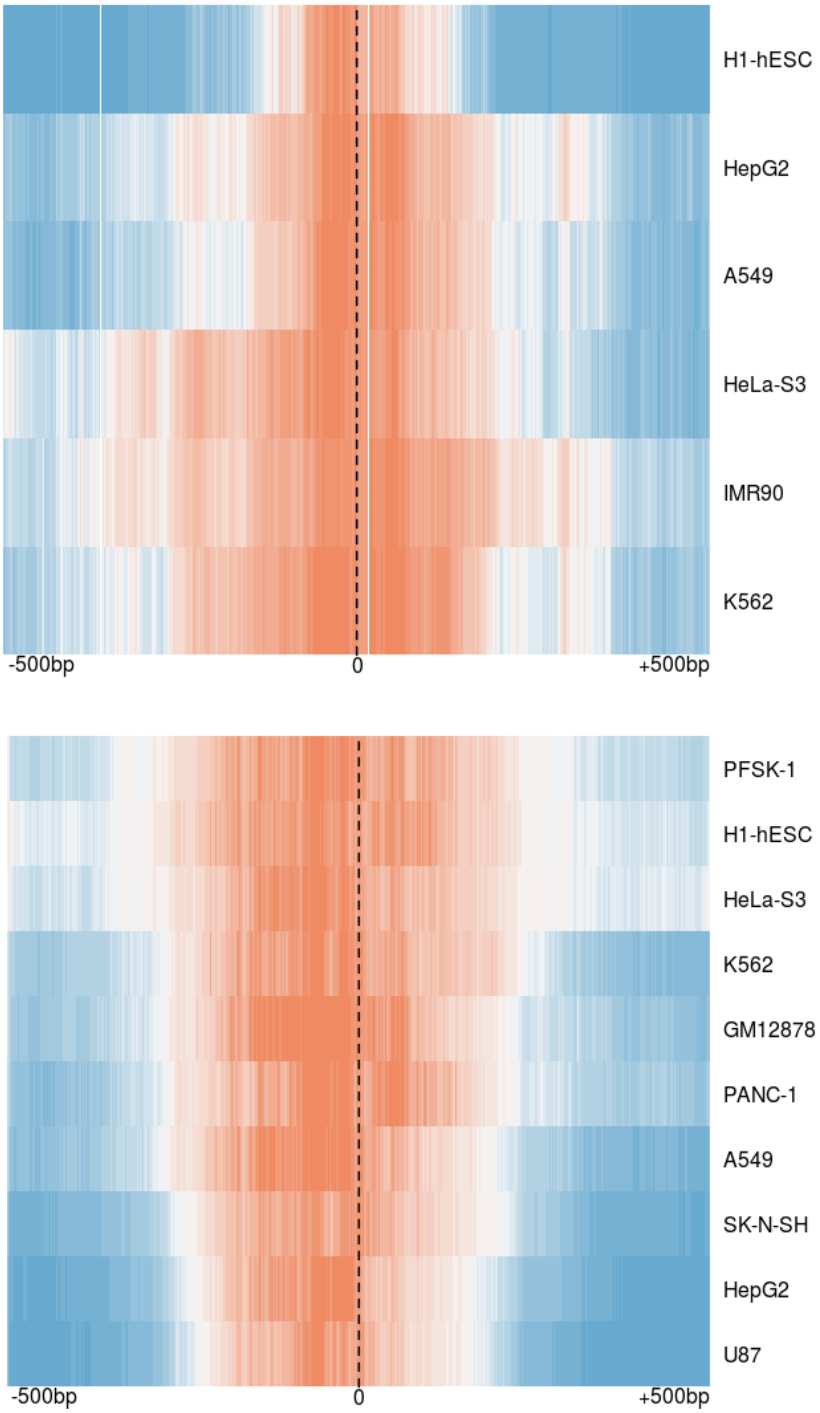

Figure S17: Genome-wide tissue-specific transcription factor binding sites (TFBS) characterization. Functional score profiles were obtained using a 1,000bp window centered on the TFBS (dash line). Positive enrichment (orange) and negative enrichment (blue) are reported for each different cell type in row. (**Top**)TiSAn-heart enrichment in CEBPB TFBS. Locations for ENCODE TFBS were found in 6 different cell types. (**Bottom**) TiSAn-brain enrichment in REST TFBS. Locations for ENCODE TFBS were found in 10 different cell types.

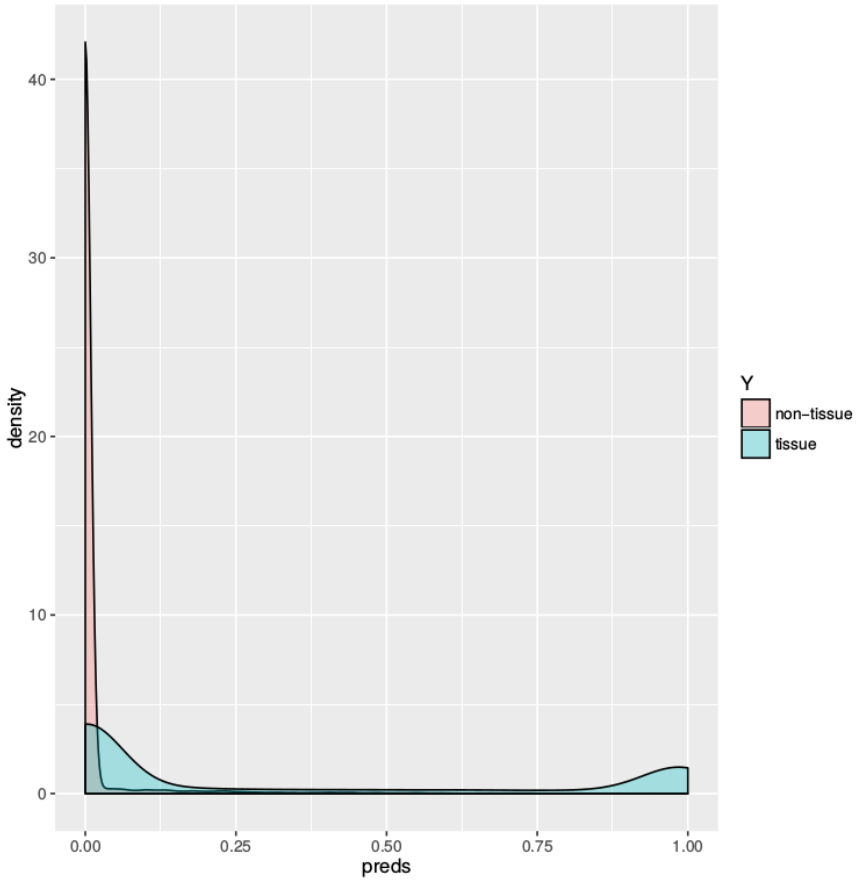

Figure S18: TiSAn-brain rescaled scores. This distribution is using the same data than Supplementary Figure 4B, but also included the loci with a TiSAn score equal to zero.

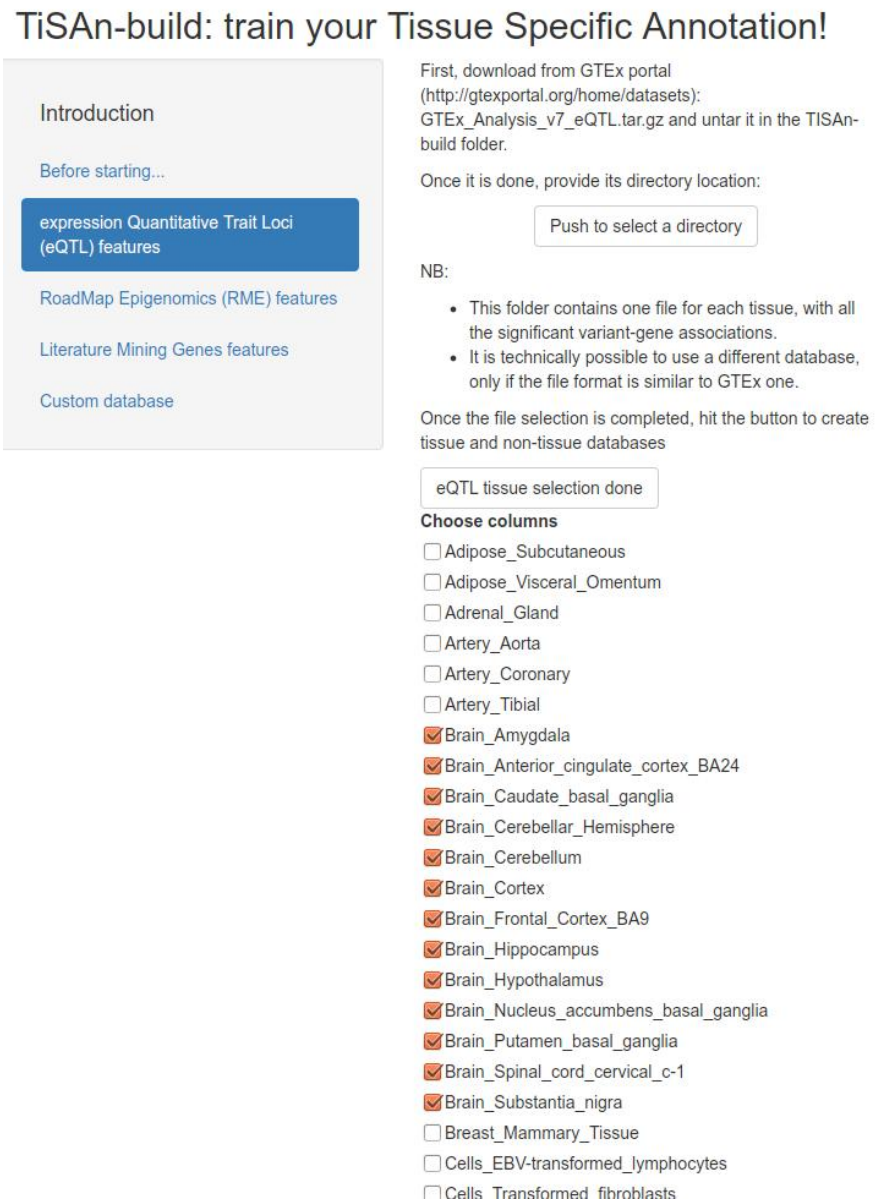

Figure S19: TiSAn-build application (GTEx). In this Shiny-based tool, users can extract tissue-specific signal from public databases. This screenshot shows the panel related to the extraction of expression quantitative trait loci from the Gene-Tissue Expression (GTEx) consortium. The user can select the tissues of interest by clicking in checkboxes. After pressing 'eQTL tissue selection done', the program will separate eQTL data into a tissue-specific and a non-tissue specific databases that could be used as references during the model training.

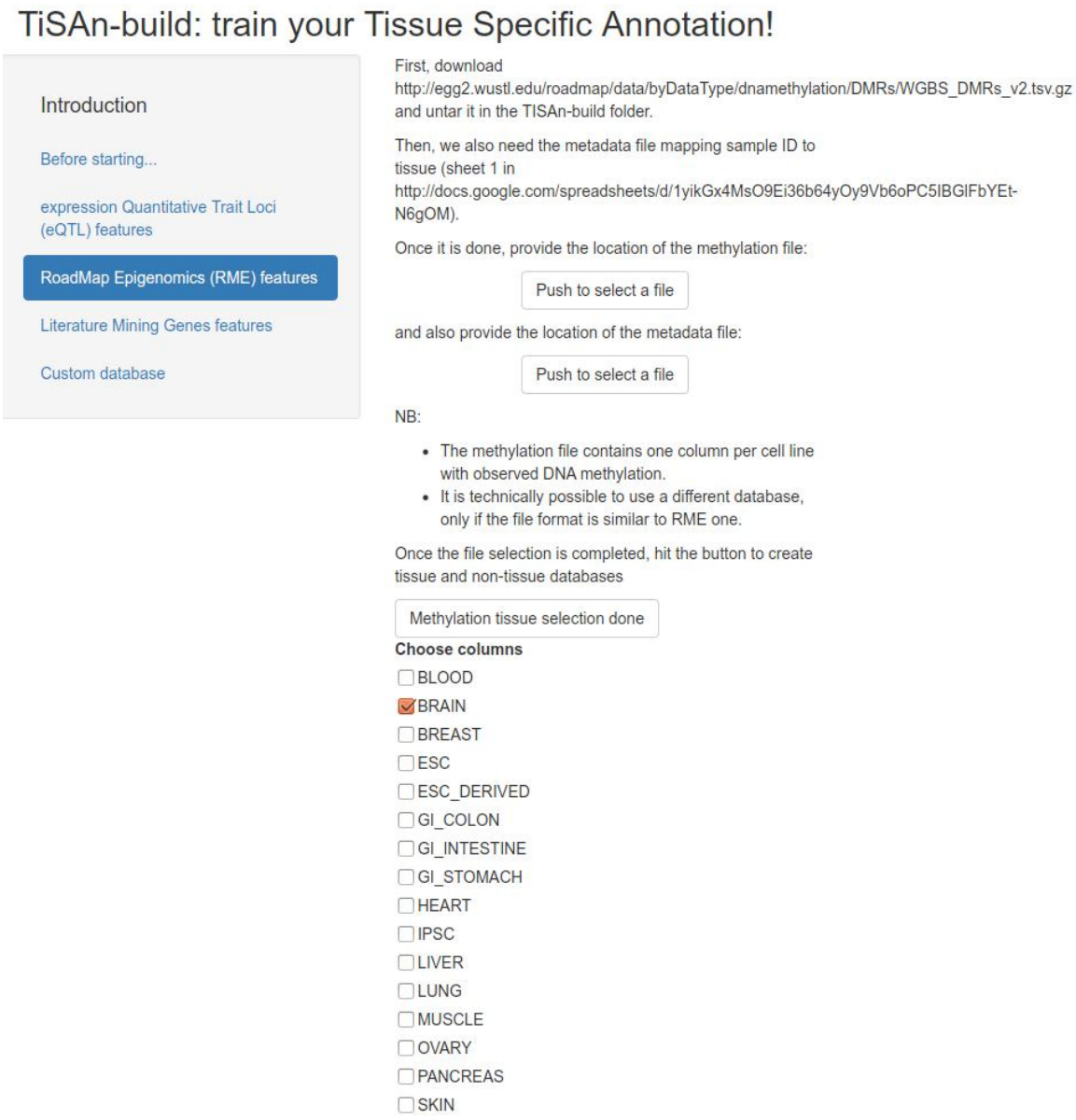

Figure S20: TiSAn-build application (RME). In this Shiny-based tool, users can extract tissue-specific signal from public databases. This screenshot shows the panel related to the extraction of DNA methylation from the RoadMap Epigenomics (RME) consortium. The user can select the tissues of interest by clicking in checkboxes. After pressing 'Methylation tissue selection done', the program will compute average methylation level for tissue-specific and non-tissue specific cell lines that could be used as references during the model training.

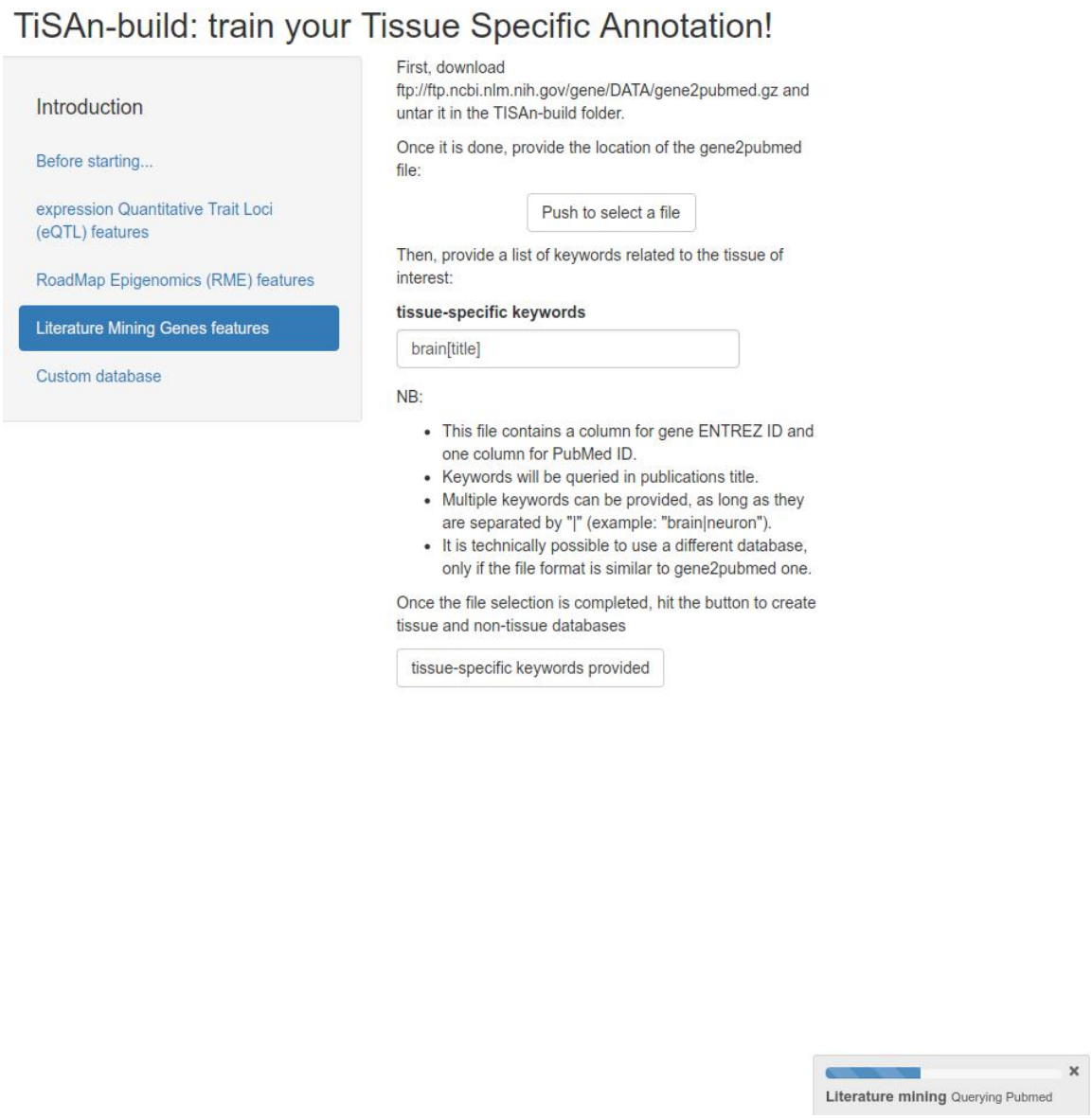

Figure S21: TiSAn-build application (PubMed). In this Shiny-based tool, users can extract tissue-specific signal from public databases.This screenshot shows the panel related to the extraction of tissue-related genes from PubMed. The user can enter keywords describing the tissue of interest. After pressing 'tissue-specific keywords provided', the program will separate genes into a tissue-specific and a non-tissue specific databases that could be used as references during the model training.

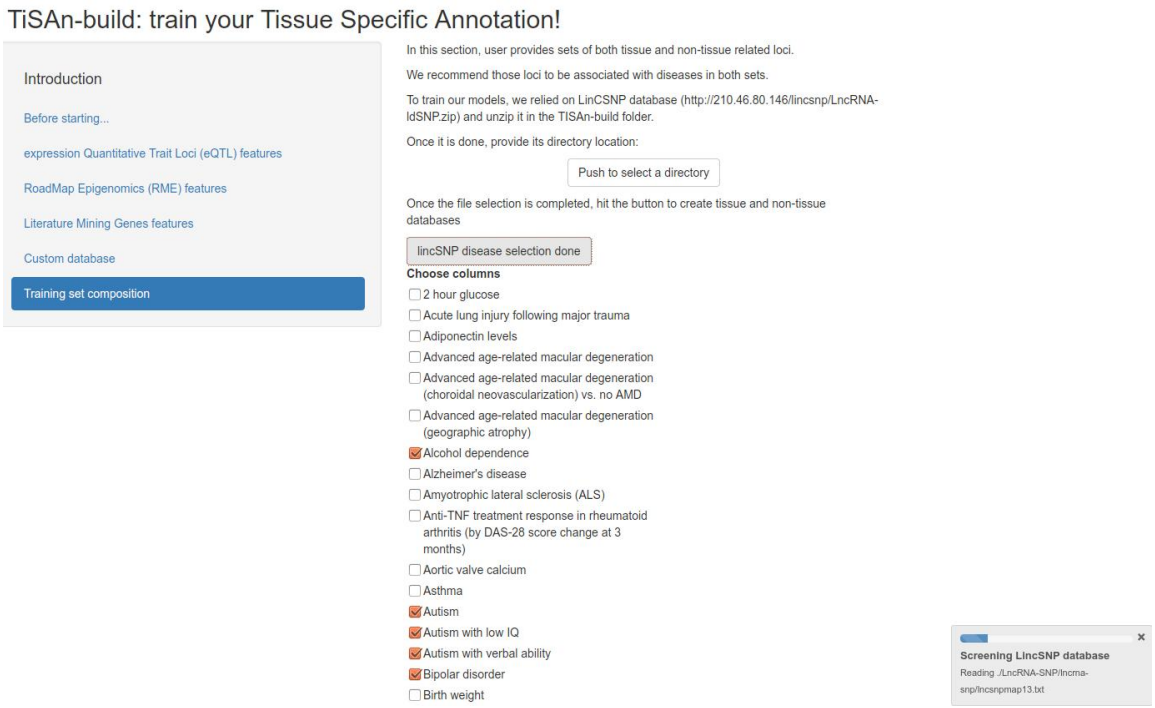

Figure S22: TiSAn-build application (training set from lincSNP). In this Shiny-based tool, users can extract training set positions for both tissue and non-tissue related diseases.This screenshot shows the panel related to the selection of disease-related loci from the LincSNP database. The user can select tissue-related terms among a long list of diseases and traits. After pressing 'lincSNP disease selection done', the program will separate SNPs into a tissue-specific and a non-tissue specific sets that could be used as references during the model training.