TiSAn: tissue specific variant annotation

Kévin Vervier[1] and Jacob J Michaelson[1,2]

University of Iowa, Carver College of Medicine, [1]Department of Psychiatry

[2]Corresponding Author: Jacob J Michaelson, 501 Newton Road, Iowa City, IA 52242. 319-335-8066

Contact: Jacob-Michaelson@uiowa.edu

Abstract:

The impact of genetic variation on molecular functions, such as gene expression, varies across tissues and cell types, complicating the interpretation of genetic variants. We introduce a functional Tissue-Specific Annotation (TiSAn) tool that predicts how related a genomic position is to a given tissue (http://github.com/kevinVervier/TiSAn). We demonstrate the accuracy and versatility of TiSAn by introducing predictive models for human heart and human brain, and detecting tissue-relevant variations in large cohorts for autism spectrum disorder and coronary artery disease.

Whole genome sequencing (WGS) is assuming its role as the technology of choice for an increasing number of genetic studies. A vast majority of the information yielded by WGS resides in non-coding and less well-characterized regions of the genome. Recent work in the annotation of non-coding variation has shown that multiple levels of information, integrated using machine learning algorithms, are required to capture the diverse regulatory potentials in these regions[1-4]. However, current state-of-the-art variant annotation methods predict generic pathogenicity, and largely sidestep the question of which tissues, organs, and systems are likely to be most susceptible to a particular genetic variation. Projects such as the Genotype-Tissue Expression (GTEx)[5] repository and the NIH Roadmap Epigenomics Mapping Consortium (RME)[6], provide clear evidence that a variant will not necessarily have the same impact on gene expression in different tissues or cell types. Recently proposed approaches, such as GenoSkyline[7], have employed cross-tissue methylation levels to annotate genetic variations. However, such methods have limitations because they were trained only using data that was uniformly collected across a wide variety of tissues, leaving out potentially informative features derived from one-off databases for specific tissues. This results in emphasizing performance over many tissues rather than optimizing for a specific tissue.

In this work, we introduce Tissue Specific Annotation (TiSAn), which combines the power of supervised machine learning with tissue-specific annotations, including genomics, transcriptomics and epigenomics (Supplementary Software 1 and http://github.com/kevinVervier/TiSAn for the latest version). We describe a general statistical learning framework in which researchers can derive a nucleotide-resolution score for the tissues they focus on. As a proof of principle, we apply our methodology to two human tissues, namely brain and heart.

The design of TiSAn models is outlined in Figure 1a (details in Online Methods). Taking advantage of publically available datasets[5,6], we extracted more than 350 different genome-wide

variables which were used to describe two large sets of disease-related loci. Training a supervised machine learning model requires positive and negative examples: here, positive examples were nucleotide positions that had been previously linked to a tissue-specific disease, and negative examples were variants that had no established link to the tissue-specific disease in question. Predictive models were trained on the labeled datasets and optimized to achieve high discrimination of tissue-specific loci (Sup. Fig. 4-5). Here, a position with a score equal to 1 can be considered strongly associated with the tissue, whereas a score of 0 means no association at all, and such a position is usually discarded in subsequent analysis.

In the following, we demonstrate TiSAn performance in three different settings, i) performing tissue-specific enrichment in case-control cohort, ii) enhancing results from a genome-wide association study (GWAS), iii) extracting genome-wide tissue-specific transcription factors. We also consider a recently proposed tool, GenoSkyline[7] that provides a genome partition in terms of functional segments, using methylation data only. Our approach aims to provide functional prediction at the single nucleotide resolution, because variants found in large predicted functional blocks (as is the case in GenoSkyline) may in fact have different functional effects.

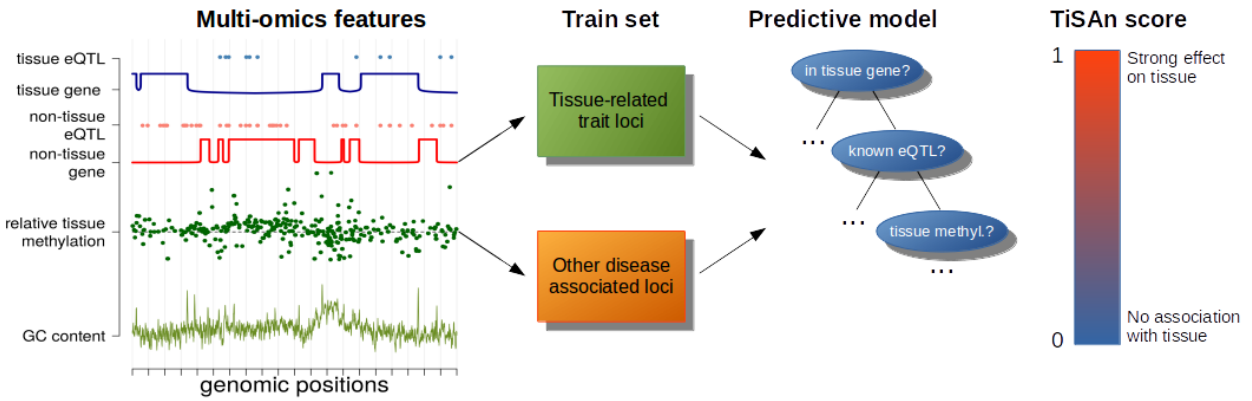


Figure 1: *TiSAn framework overview. Each single position in the genome is annotated with multiple levels of omics information, such as nucleotide frequency, methylation level, proximity with genes. Especially, information is extracted for training sets, made of disease-associated loci with or without an association with the tissue of interest. Using supervised machine learning techniques, a predictive model combines each feature with respect to its importance in predicting if a position will be functionally enriched for the tissue of interest. Model output consists in a tissue-specific functional score spanning between no-association and strong effect on the tissue.*

The Simons Simplex Collection (SSC) provides whole genome sequencing for one of the largest autism spectrum disorder (ASD) cohorts currently available. We hypothesized that deleterious genetic variation found in the vicinity of ASD-related genes would show higher enrichment in terms of brain-related functional consequences (as measured by the TiSAn-brain and GenoSkyline-brain scores) in the SSC compared to the 1000 Genomes (1KG). We further

assessed enrichment using the respective heart-specific scores as a form of negative control. In this analysis, the TiSAn-brain score shows the only positive tissue-specific enrichment, over 50% for coding variants (Fig. 2a) and around 10% for non-coding variants (Fig. 2b). Notably, there is a significant difference between TiSAn brain and heart scores (Wilcoxon signed-rank test, $P < 2x10^{-16}$), suggesting effective tissue specificity, whereas this was not observed for GenoSkyline models (Wilcoxon signed-rank test, $P = 0.351$). Genes *RNPS1*, *TNIP2*, and *TSC2* were the 3 genes with the highest TiSAn-brain score enrichment in SSC data, suggesting that deleterious variants in these regions in are more likely to affect brain functions.

Next, we ranked and binned variants according to their tissue-specific scores (i.e., TiSAn or GenoSkyline) and calculated the enrichment of SSC deleterious variants in each bin, with respect to deleterious 1KG variants. Because the SSC is a neurodevelopmental cohort, we expect to see over-representation of SSC variants in the most confidently called brain-related genomic regions. Indeed, significant enrichment of SSC variants was observed in the top quantiles for TiSAn-brain but also for both GenoSkyline models (Fig. 2c). Surprisingly, the GenoSkyline-heart model reports a more pronounced enrichment than the corresponding brain model, suggesting a potential lack of tissue specificity for GenoSkyline. TiSAn-brain achieves the highest enrichment by ranking 2.5 times more SSC variants in the top 5% than 1KG variants.

Current approaches to GWAS analysis rely mostly on association strength (e.g., P-value) to prioritize candidate regions. These variants often belong to large linkage-disequilibrium (LD) blocks, making it difficult to decipher the actual causal genetic mechanism. Here, we apply TiSAn to the Coronary Artery Disease (CAD) CARDIoGRAM consortium GWAS meta-analysis[8], and we demonstrate that the TiSAn-heart score is significantly higher among the most associated variants (Fig. 3d, (Wilcoxon signed-rank test, $P < 2x10^{-16}$)). Furthermore, the top 100 SNPs (according to their *P*-value) with a non-zero TiSAn were all found in LD with genomic regions strongly associated with coronary artery disease, demonstrating TiSAn high sensitivity. In this analysis, no significant enrichment was observed for GenoSkyline-heart (Wilcoxon signed-rank test, $P = 0.12$) or brain models (Supplementary Fig. 8).

Transcription factor (TF) binding sites (TFBS) are associated with observed differences in gene expression across tissues[9]. 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. 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 enriched for TiSAn-heart score in 6 cell types, including two related to smooth muscles (A549 and IMR90) and one related to liver (HepG2) (Supplementary Fig. 7), and has been associated with cardiac hypertrophy[10] and fatty liver disease[11].

Brain-specific binding patterns in mouse have been observed for the major regulator of chromatin state CTCF[12], and our analysis demonstrates a consistent signature with TiSAn-Brain enrichment at the CTCF binding locations across 70 different cell types (Supplementary Fig. 8), suggesting the importance of chromatin conformation in brain development and function. Functional enrichment patterns were also found for the critical brain transcription factor REST in 10 different cell types (Supplementary Fig. 9). Among these were 3 brain cancer cell lines (U87,

SK-N-SH, and PFSK-1), which may support recent findings on the importance of REST in neuroblastoma drug sensitivity[13].

We have demonstrated the effectiveness of TiSAn in real world use cases and in comparison to state of the art competing methods. These kinds of approaches hold great promise for helping genomics researchers narrow down massive lists of variants to focus on those that are most relevant to the tissue or disease at hand. Researchers interested in other tissues beyond brain or heart can derive their own functional annotation for a selected tissue of interest (Online Methods), and we have provided thorough documentation, including tutorials, on how to use TiSAn in genome informatics workflows.

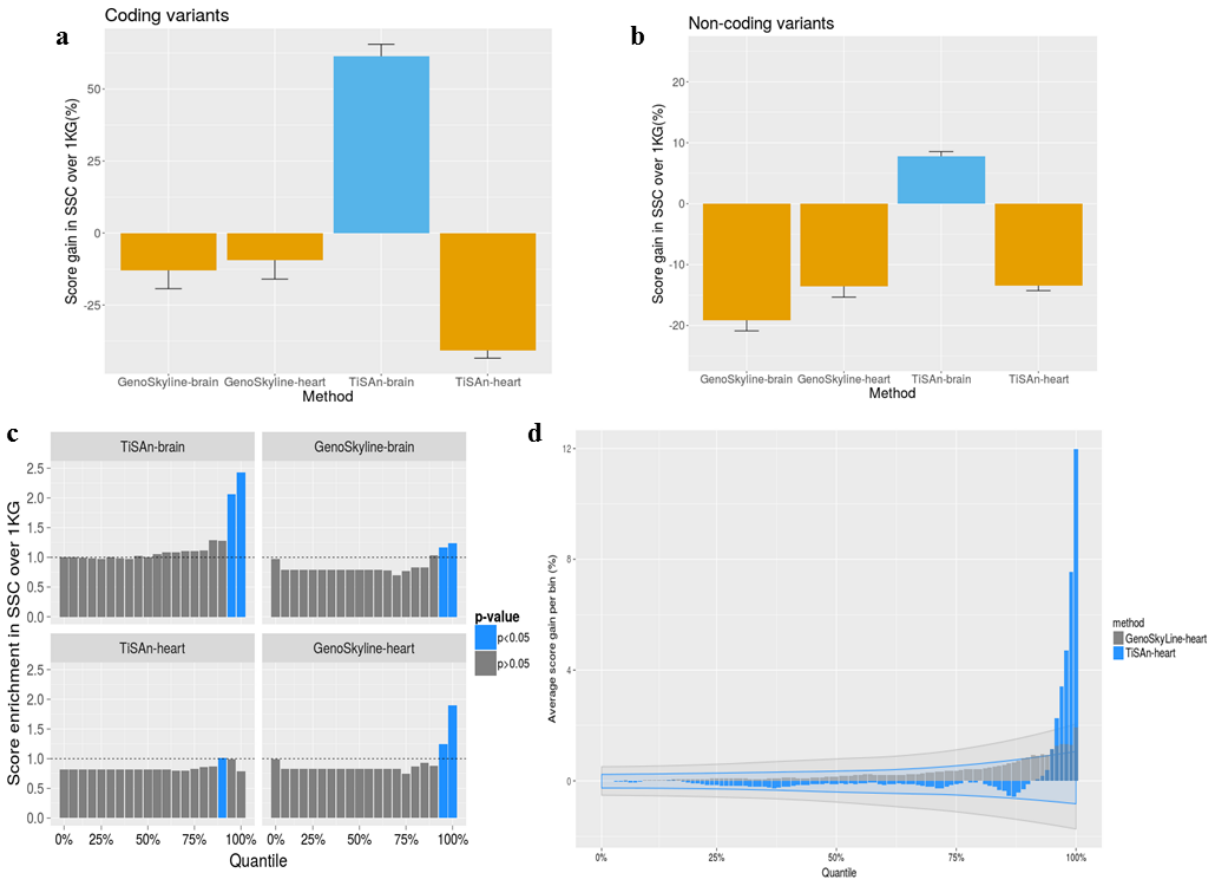


Figure 2: ***Tissue-related functional enrichment in a case-control cohort setting.*** *Comparison of Simons Simplex Collection (SSC) variants with 1000Genomes (1KG) variants. Coding variants (a) and non-coding variants (b). Both brain and heart models for TiSAn and GenoSkyline were evaluated.* ***(c) Functional score enrichment in SSC variants compared to 1KG variants****. After sorting SSC and 1KG variants based on their score, we compute cumulative enrichment for each 5%-ile. Blue bars correspond to significant difference between SSC and 1KG, using the Chi-squared test (p-value <0.05).* ***(d) CAD-GWAS signal prioritization using heart-related models.*** *Genetic*

variants were binned by percentiles, based on their association P-values. In each of those bins, we reported average functional scores (blue: TiSAn-heart, grey: GenoSkyline-heart). Shaded areas represent confidence interval for the corresponding method, after GWAS P-values random permutations.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by NIH grants MH105527 and DC014489.

Data on autism spectrum disorder variation have been contributed by Simons Simplex Collection investigators and have been downloaded from http://sfari.org/resources/sfari-base.

Data on coronary artery disease have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from www.cardiogramplusc4d.org.

AUTHOR CONTRIBUTIONS

K.V. conceived and carried out the analysis. J.J.M. provided supplemental assistance in the analysis and figures. K.V. and J.J.M. wrote the manuscript. All authors reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

References

1. Kellis M, Wold B, Snyder MP, et al. Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences of the United States of America 2014;111:6131-8.
2. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome research 2005;15:1051-60.
3. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics 2014;46:310-5.
4. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature genetics 2016;48:214-20.
5. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348:648-60.
6. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nature biotechnology 2010;28:1045-8.
7. Lu Q, Powles RL, Wang Q, He BJ, Zhao H. Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. PLoS genetics 2016;12:e1005947.

8. Nikpay M, Goel A, Won HH, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nature genetics 2015;47:1121-30.
9. Zhong S, He X, Bar-Joseph Z. Predicting tissue specific transcription factor binding sites. BMC genomics 2013;14:796.
10. Redondo-Angulo I, Mas-Stachurska A, Sitges M, Giralt M, Villarroya F, Planavila A. C/EBPbeta is required in pregnancy-induced cardiac hypertrophy. International journal of cardiology 2016;202:819-28.
11. Sookoian S, Rohr C, Salatino A, et al. Genetic variation in long noncoding RNAs and the risk of nonalcoholic fatty liver disease. Oncotarget 2017;8:22917-26.
12. Prickett AR, Barkas N, McCole RB, et al. Genome-wide and parental allele-specific analysis of CTCF and cohesin DNA binding in mouse brain reveals a tissue-specific binding pattern and an association with imprinted differentially methylated regions. Genome research 2013;23:1624-35.
13. Liang J, Tong P, Zhao W, et al. The REST gene signature predicts drug sensitivity in neuroblastoma cell lines and is significantly associated with neuroblastoma tumor stage. International journal of molecular sciences 2014;15:11220-33.
14. Voight BF, Kang HM, Ding J, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS genetics 2012;8:e1002793.
15. Ning S, Zhao Z, Ye J, et al. LincSNP: a database of linking disease-associated SNPs to human large intergenic non-coding RNAs. BMC bioinformatics 2014;15:152.
16. Cherkasov A, Ho Sui SJ, Brunham RC, Jones SJ. Structural characterization of genomes by large scale sequence-structure threading: application of reliability analysis in structural genomics. BMC bioinformatics 2004;5:101.
17. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nature methods 2015;12:931-4.
18. Wheeler HE, Shah KP, Brenner J, et al. Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. PLoS genetics 2016;12:e1006423.
19. Liu Y, Liang Y, Wishart D. PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic acids research 2015;43:W535-42.
20. Miller CL, Pjanic M, Wang T, et al. Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. Nature communications 2016;7:12092.
21. Spiers H, Hannon E, Schalkwyk LC, et al. Methylomic trajectories across human fetal brain development. Genome research 2015;25:338-52.
22. Dickel DE, Barozzi I, Zhu Y, et al. Genome-wide compendium and functional assessment of in vivo heart enhancers. Nature communications 2016;7:12923.
23. Lischke H, Loffler TJ, Fischlin A. Aggregation of individual trees and patches in forest succession models: capturing variability with height structured, random, spatial distributions. Theoretical population biology 1998;54:213-26.
24. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research 2016;44:D733-45.

ONLINE METHODS

Training set definition. We identify multiple sources for training examples with respect to a given tissue *T.* One way to build such a training set is to look for positions known to be causal of a disease in tissue *T.* This way, we reduce the risk of training a pathogenicity score and make

sure we extract a signal orthogonal to deleteriousness. For deriving genome-wide predictor, the training set needs to cover both coding and non-coding loci, but also loci related to *T* (positive examples) and unrelated (negative examples). Two types of public databases were used to derive training sets:

Genotype array loci: Disease-related loci could be found in Consortium developed arrays, designed for targeting specific disorders, such as the MetaboChip[14] for type2 diabetes and cardiovascular diseases, or Illumina Infinium PsychArray Beadchip for psychiatric disorders. Those probe sets usually contain tissue-related variants (positive examples), but also backbone/control loci which we consider as negative examples.

Large intergenic non-coding RNAs: Usually, non-coding variants are less functionally characterized than coding ones. Large intergenic non-coding RNAs (lincRNAs) represent a well-study group of non-coding elements. Databases, such as LincSNP[15], contain disease-related variants that occur in lincRNA loci. After defining a list of tissue-related disorders, we propose to divide this database in two subsets: one related to tissue T (positive examples) and one containing background variants (negative examples). This way, we enrich the training set with non-coding loci.

Exact number of training examples used for TiSAn brain and heart models can be found in Supplementary Table 1.

Weibull distribution. Following Cherkasov et al.[16], we model the distance between a given locus $\boldsymbol{x}$ and a known annotation, following the Weibull distribution and its Extreme Value Theory application. Therefore, in the following paragraphs, the distance is measured as:

$$\boldsymbol{d(x, anno)} = \left(\frac{\boldsymbol{\beta}}{\boldsymbol{\alpha}}\right) \times \left(\frac{|\boldsymbol{x-anno}|}{\boldsymbol{\alpha}}\right)^{\boldsymbol{\beta-1}} \times \boldsymbol{exp}\left\{-\left(\frac{|\boldsymbol{x-anno}|}{\boldsymbol{\alpha}}\right)^{\boldsymbol{\beta}}\right\},$$

where anno refers to a known annotation position, α is a scale factor, and β is a shape parameter. Parameters fitting was performed separately for each annotation, using *MASS* R package.

Features extraction. We represent each genomic position in a functional space made of hundreds of different annotations. In the following, we describe how such signal can be extracted using publically available data sets. More details can be found in Github vignettes (http://github.com/kevinVervier/TiSAn/tree/master/vignettes).

Nucleotides frequencies are linked to overall regulatory activity (G/C content), and patterns in n-nucleotides chains are at the core of transcription factor binding sites detection[17]. Recently, specific patterns have been identified to be tissue-specific[9], and we incorporate this information by computing frequencies for all $\boldsymbol{n}$-nucleotides ($\boldsymbol{n} \in (\boldsymbol{1}, \boldsymbol{2}, \boldsymbol{3}, \boldsymbol{4})$), found in a +/- 500 basepairs neighborhood around a locus $\boldsymbol{x}$.

Links between disease traits and tissue-specific gene expression have been reported in studies using the rich GTEx dataset[18]. For each genomic location, we extract features based on how close $\boldsymbol{x}$ is to known eQTLs for tissue *T*, and for other tissues. Weibull distribution was used for modeling the minimal distance to a GTEx eQTL (Supplementary Fig. 1). We also derive Boolean features for whether or not the genomic position $\boldsymbol{x}$ is at the exact location of a GTEx eQTL, which puts more weight for being a known locus. Although some genes expression shows variation across tissues, comprehensive resources and exhaustive list of tissue-specific genes are limited. It has already been shown that text mining techniques may help to extract

relationship between genes and disease traits[19]. Therefore, we propose to adapt such methods to identify, in the Pubmed database (May 2016 gene2ID database), genes reported to be associated with tissue *T.* Only genes with at least 3 citations were kept. We observed, when training the brain model, that around 1,000 tissue-related genes represent enough genome coverage to derive a feature based on the proximity between a locus $\boldsymbol{x}$ and a gene, under Weibull distribution assumption (Supplementary Fig. 2).

Epigenomics and in particular, methylation profiles have been integrated to explain tissue specific regulatory mechanisms[20]. Weibull distribution for modeling how the minimal distance to a methylated region found in RME database is compared to distance to all the other methylated regions (Supplementary Fig. 3). If it happens that the considered position $\boldsymbol{x}$ belongs to a methylated region characterized in RME project, we also get the average methylation level for *T* and all the other tissues.

Compared to other approaches mostly relying on RME and/or GTEx, we also considered tissue-specific data sets made available by the research works focusing on one single tissue. For the brain model, we integrate developmentally differentially methylated positions (dDMPs)[21] found in fetal brain. For the heart model, Heart Enhancer Compendium database[22] for heart development candidates was used.

Supervised machine learning model training. Considering the aforementioned training sets, we fit several machine learning approaches and compare them based on their 10-folds cross-validated performances (here, area under the ROC curve, AUC), and selected random forest[23] algorithm to train the final model (Supplementary Table 2, AUC = 0.8). Using cross-validation, we optimize both the number of trees and the number of variables to consider at each node in the *randomForest* R package.

From class probability to rescaled odd-ratio. Current approaches often consider the raw class probability as their functional score, requiring additional tuning step from the user. Here, we propose to rescale the classifier output, into a ready-to-use score. First, we define an optimal cutoff value on the probability (Supplementary Fig. 4a and 5a), as the smallest value which reaches a false discovery rate of 10%. For instance, this threshold is equal to 0.48 for the brain model and to 0.67 for the heart model. Then, we rescale the filtered probability to a score between 0 and 1, using the formula:

$$\boldsymbol{max}\left(\mathbf{0}, \mathbf{1} - \boldsymbol{thresh} - \frac{\mathbb{P}(\boldsymbol{x} \notin \boldsymbol{tissue})}{\mathbb{P}(\boldsymbol{x} \in \boldsymbol{tissue})} \times \boldsymbol{thresh}\right).$$

The main advantage of this step is to standardize predictive models, and push loci not tissue-related to a score strictly equal to 0 (Supplementary Fig. 4b and 5b).

Reference genome: Analyses done in this study used hg19 reference genome.

Evaluation framework. In all analyses presented in this study, we carefully removed positions that were both found in the TiSAn training and the validation sets, avoiding over-optimistic performances.

GWAS prioritization in coronary artery disease (CAD) cohort. CARDIoGRAM consortium GWAS meta-analysis summary statistics for 8,443,810 SNPs were downloaded at http://www.cardiogramplusc4d.org/media/cardiogramplusc4d-consortium/data-downloads/cad.additive.Oct2015.pub.zip. Correlation between functional score and association

strength (Fig. 3c) was obtained by binning in 100 percentile bins on reported association P-value. Then, relative score enrichment is computed for top1% variants and iteratively, until merging all the data. We derived confidence interval for both TiSAn and GenoSkyline by random permutations on the GWAS p-values. On the purpose of ranking variants, we filtered variants not predicted as functional by either TiSAn (zero score), or by GenoSkyline (score < 0.15).

Variants enrichment in vicinity of ASD genes. Variants found in 960 Simons Simplex Collection (SSC) individuals, including probands and parents were filtered based on their pathogenicity using CADD score. We estimated two different threshold values for coding (>15) and non-coding (>10.7) variants. Those values correspond to the first 10%-ile found in the 1000 Genomes data. We also focused the analysis on variants found in a +/-50,000bp windows around well-supported ASD genes, with more than 20 citations in the June 2016 SFARI gene list at http://gene.sfari.org/autdb/HG_Home.do (Supplementary Table 3). The same filters were applied to variants found in 1000 Genomes (1KG) European ancestry population (Phase 3). Coding and non-coding variants were separated based on their RefSeq[24] function annotation. The relative gain in average score (Fig. 2a and b) is calculated by doing the difference between average score in SSC and in 1KG, divided by the score in 1KG. Cumulative score enrichment for SSC over 1KG variants (Fig. 2c) is obtained by binning all variants from the two datasets based on their score, in 5%-ile groups. Then, average score ratio between the two groups is computed in each bin, and summed in a cumulative way, from the top 5% to all the data.

Transcription factor binding sites (TFBS) enrichment in tissue and cell type. ENCODE project provides a large repository for TFBS location in various cell type contexts. Here, we put together two databases, both available as UCSC Genome Browser tracks, *factorbookMotif*, which contains the location of more than 2 million TFBS across the genome, and *EncodeRegTfbsClustered*, which provides information regarding the cell types where TFBS were observed. Overlapping the two databases results in 1,514,086 unique TFBS found in 53 families. For each of those TFBS, we expanded their location using a 1,000 base pairs window, and TiSAn heart and brain scores were extracted. Scores were centered and scaled around the center value and show the actual score enrichment along the window. An average profile was computed for all TFS categories and cell types.

Software availability.

- Genome-wide TiSAn score databases are available in bed format (with index) at:

  - http://flamingo.psychiatry.uiowa.edu/TiSAn/TiSAn_Brain.bed.gz
  - http://flamingo.psychiatry.uiowa.edu/TiSAn/TiSAn_Heart.bed.gz
  - http://flamingo.psychiatry.uiowa.edu/TiSAn/TiSAn_Brain.bed.gz.tbi
  - http://flamingo.psychiatry.uiowa.edu/TiSAn/TiSAn_Heart.bed.gz.tbi
  - Tutorial and vignettes are also available at http://github.com/kevinVervier/TiSAn.

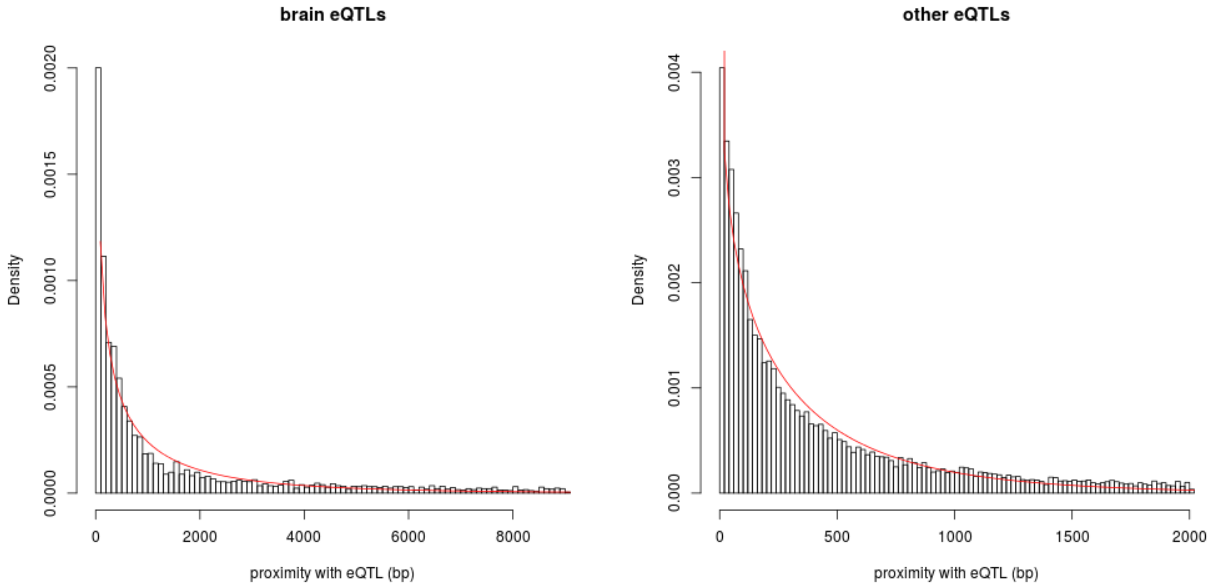
- GenoSkyline approach: we downloaded brain and heart models on the tool website (http://genocanyon.med.yale.edu/GenoSkyline), in November 2016.

- Combined Annotation Dependent Depletion (CADD): Deleteriousness annotation were performed using the CADD v1.0 (published version) at http://krishna.gs.washington.edu/download/CADD/v1.0/whole_genome_SNVs.tsv.gz

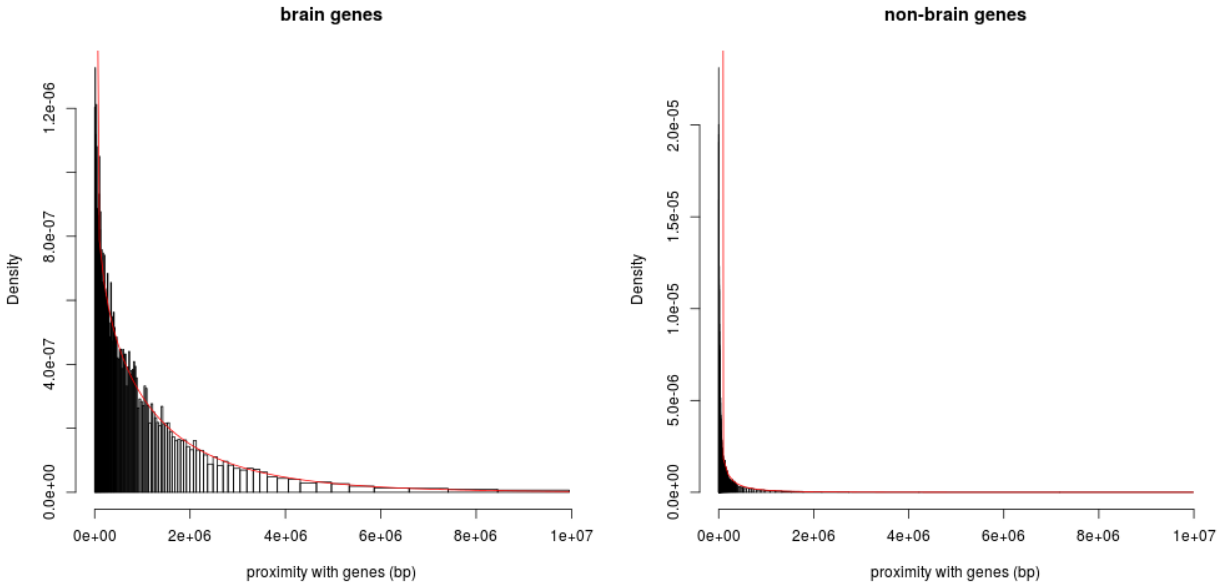
Supplementary Materials:

Supplementary Table 1: 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, corresponding to variants found in large intergenic non-coding RNAs database, or in* genotype array probesets*.*

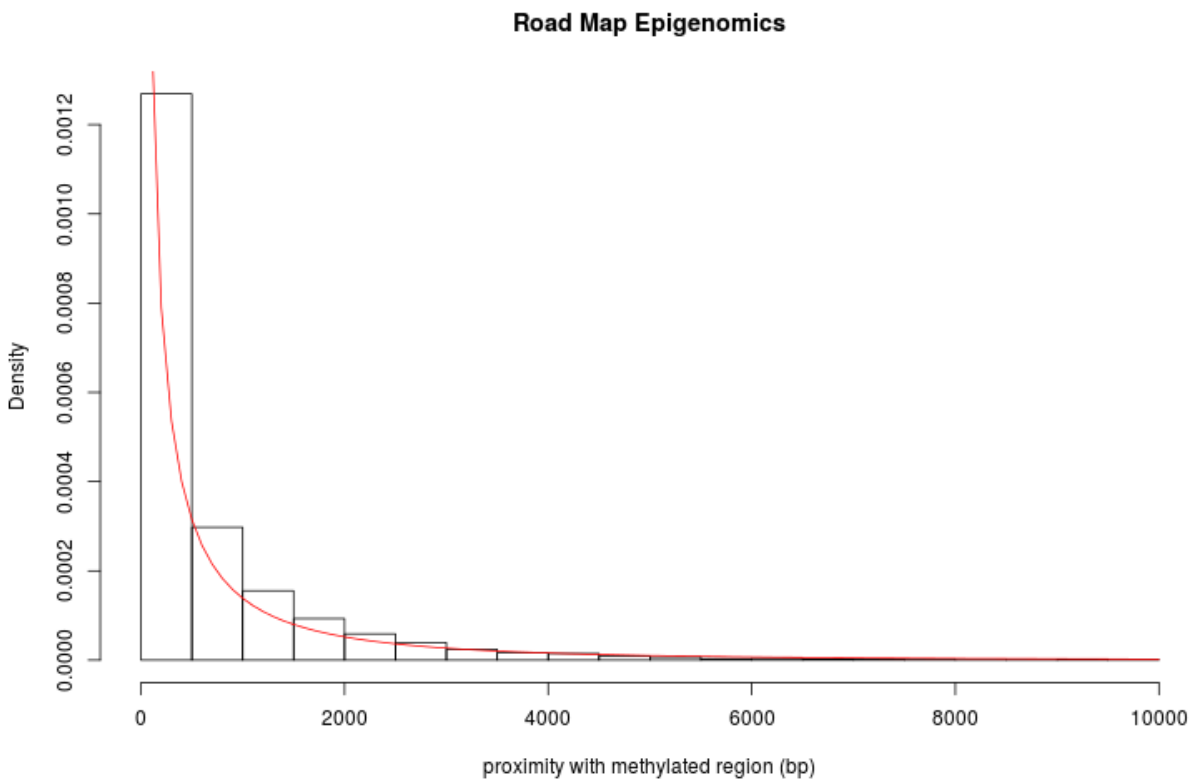
Tissue	Category	Count
Brain	Positive	10,715 (5,535 + 5,180)
Brain	Negative	22,811 (13,861 + 8,950)
Heart	Positive	28,473 (9,760 + 18,713)
Heart	Negative	62,438 (22,476 + 39,962)



Supplementary Figure 1*: 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).*



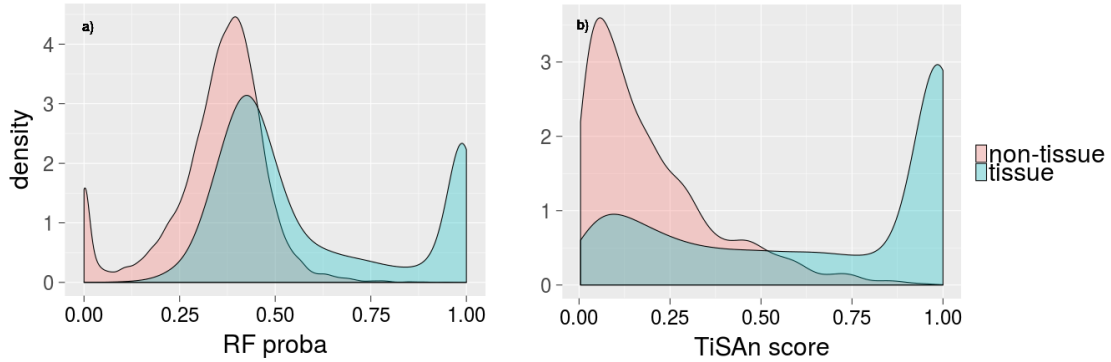
Supplementary Figure 2*: Distribution of distance to the closest gene for brain (left) and non-brain (right) tissues. The red lines correspond to a Weibull distribution fit. Estimated parameters for left (resp. right) figure are: shape = 0.852 (resp. 0.529) and scale = 1,453,217 (resp. 201,985).*



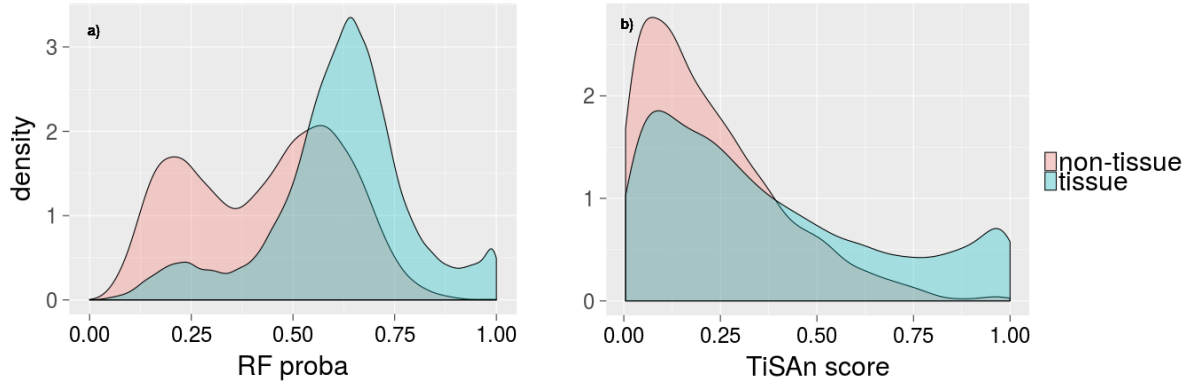
Supplementary Figure 3*: Distribution of distance to the closest methylated region found in RoadMap Epigenomics database. The red line corresponds to a Weibull distribution fit. Estimated parameters are: shape = 0.746 and scale = 590.3.*

Supplementary Table 2*: 10-folds cross-validated performances obtained during TiSAn-brain model training, for different classification strategies. AUC: Area under ROC curve.*

Methods	Cross-validated AUC
Random Forest	0.795
Logistic regression	0.635
Linear SVM	0.584



Supplementary Figure 4: TiSAn-brain *cross-validation performances. (a) Random forest raw output distribution. (b) TiSAn score obtained after rescaling odd-ratios.*

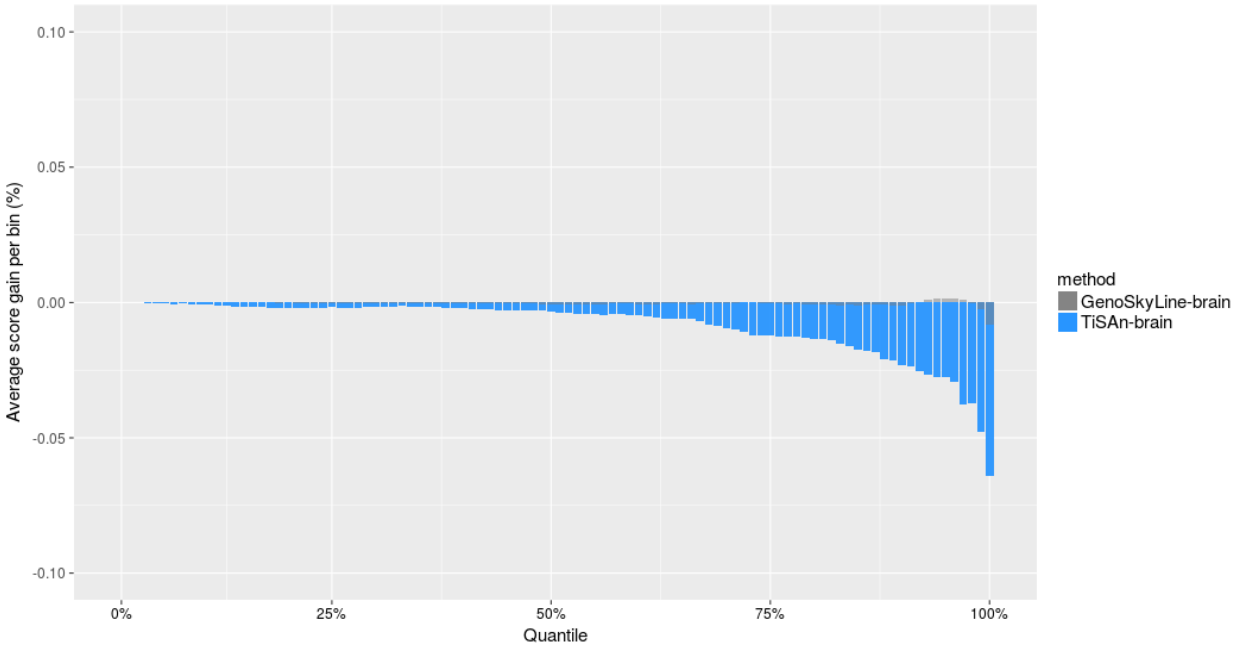


Supplementary Figure 5: TiSAn-heart *cross-validation performances. (a) Random forest raw output distribution. (b) TiSAn score obtained after rescaling odd-ratios.*

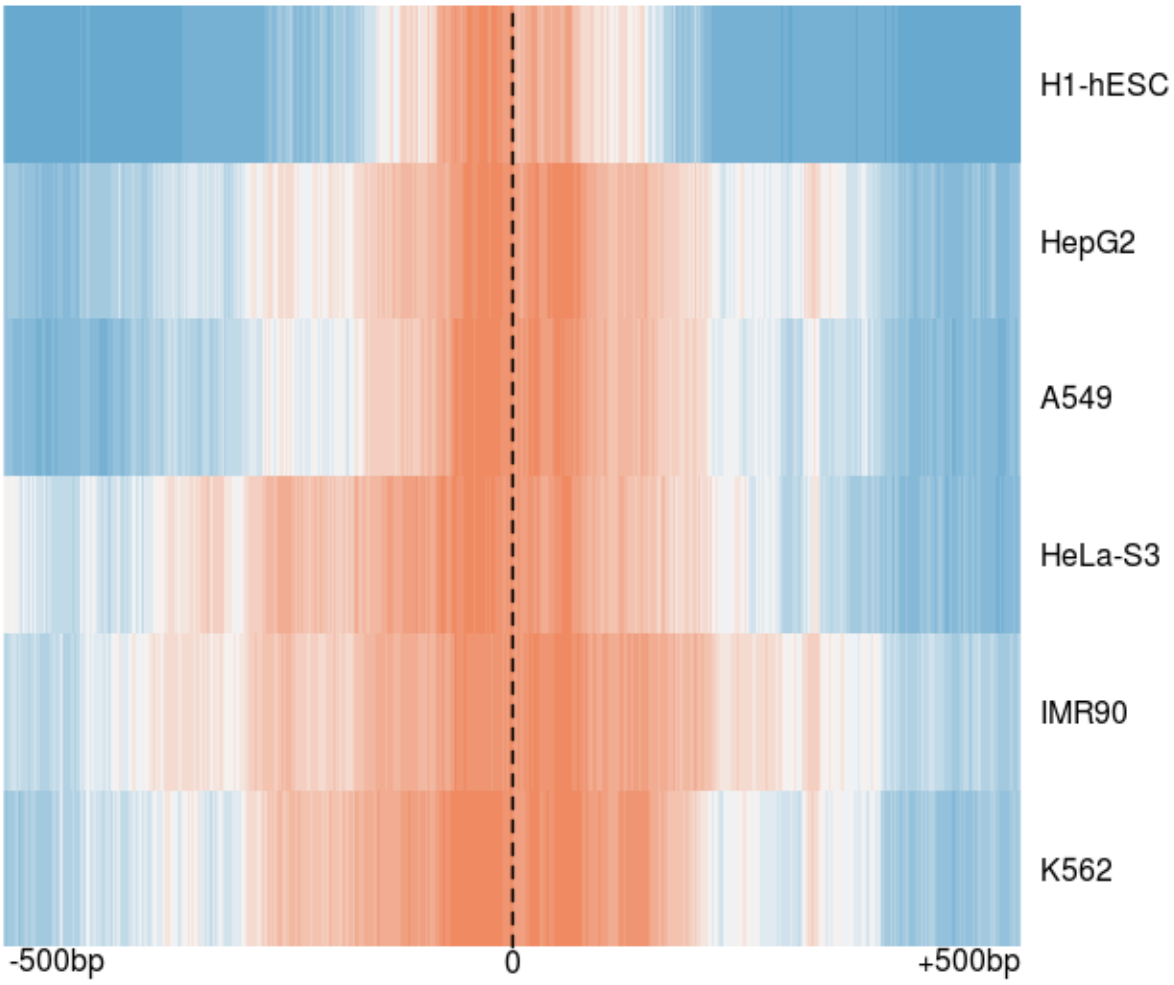
Supplementary Table 3*: List of SFARI autism-related genes, supported by literature.*

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

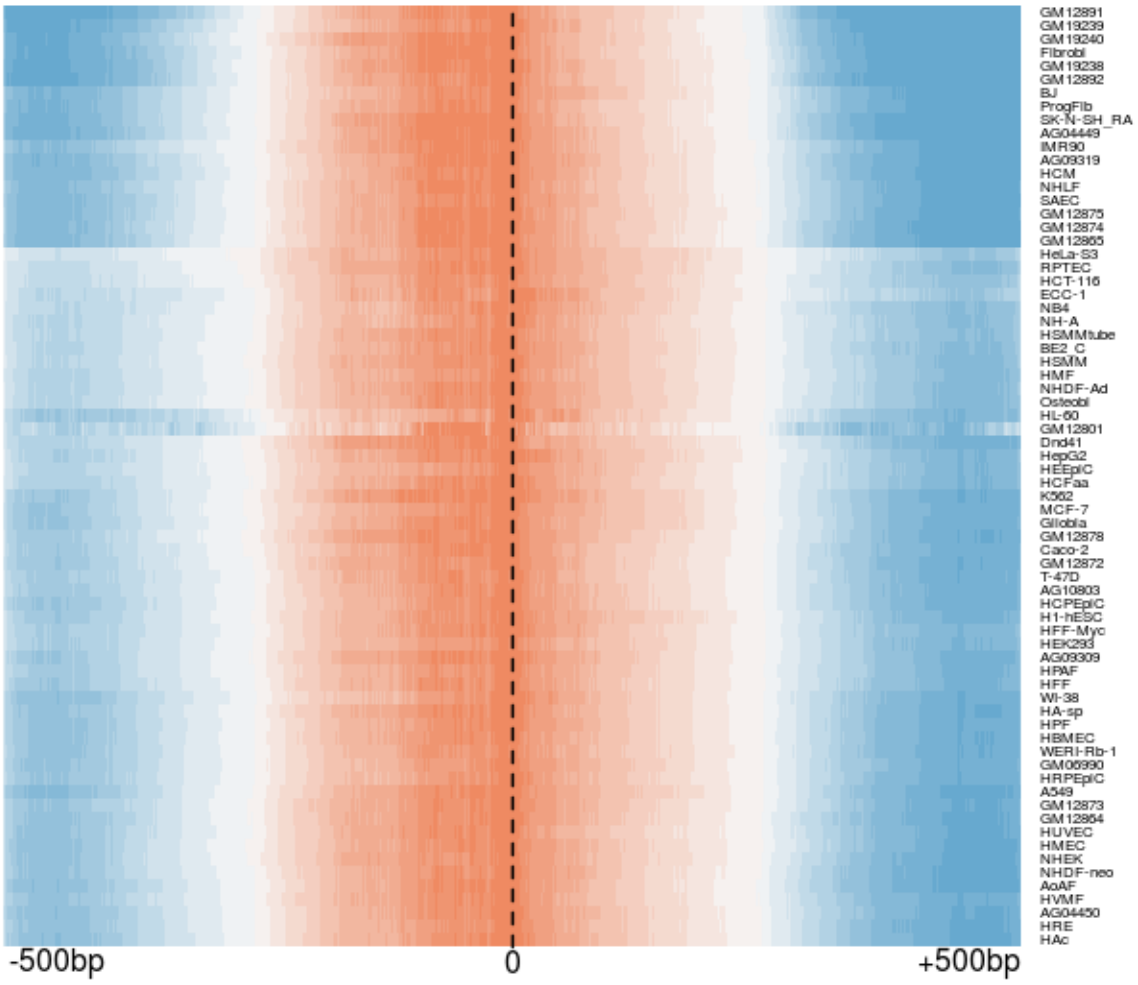
SCN1A	*31*	*SLC6A4*	22



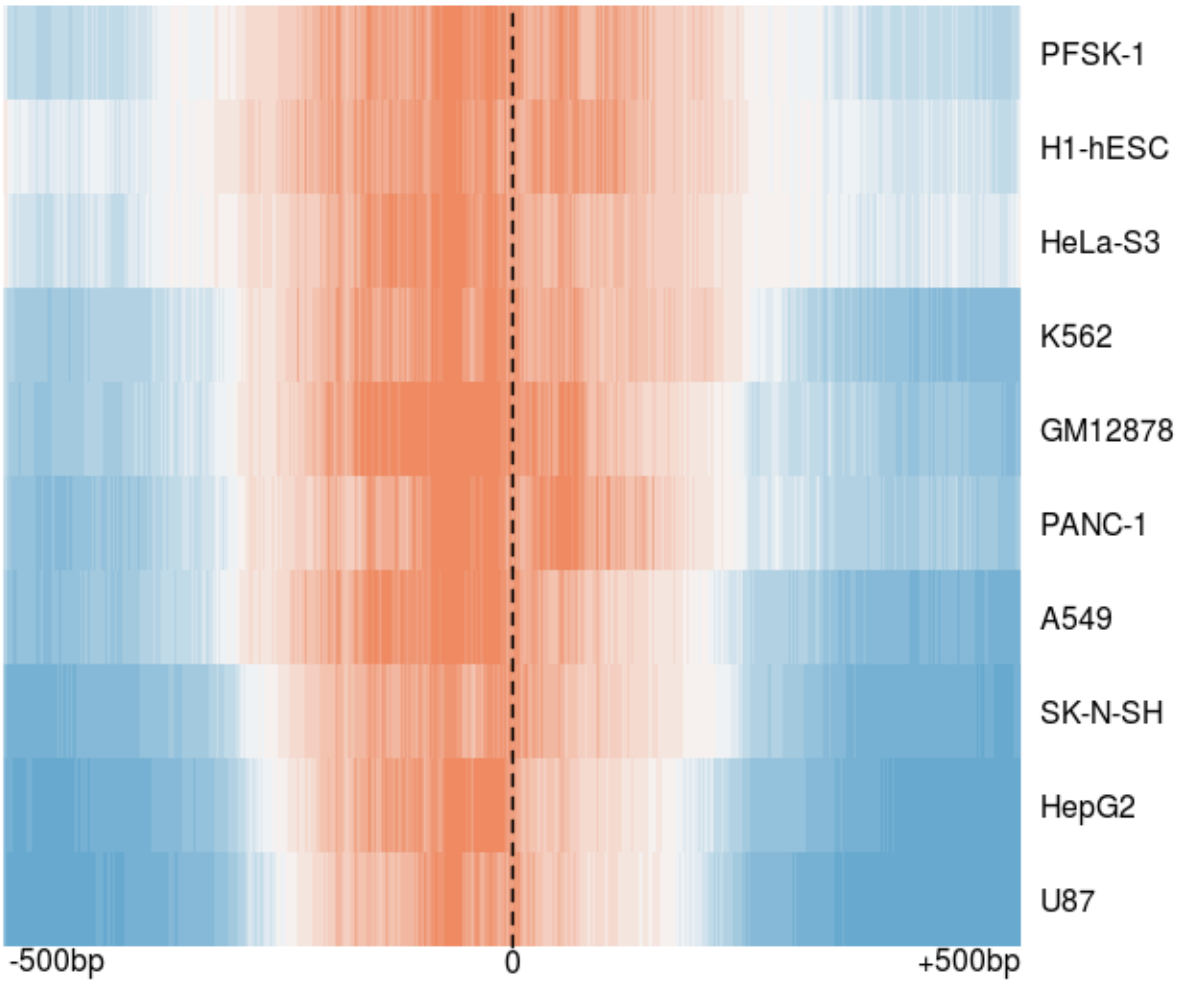
Supplementary Figure 6*: CAD-GWAS signal prioritization using brain-related models. Genetic variants were binned by percentiles, based on their association p-values. In each of those bins, we reported average functional scores (blue: TiSAn-brain, grey: GenoSkyline-brain)*



Supplementary Figure 7*: TiSAn-heart enrichment in CEBPB transcription factor binding sites (TFBS). Locations for ENCODE TFBS were found in 6 different cell types. 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 column.*



Supplementary Figure 8*: TiSAn-brain enrichment in CTCF transcription factor binding sites (TFBS). Locations for ENCODE TFBS were found in 70 different cell types. 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 column.*



Supplementary Figure 9*: TiSAn-brain enrichment in REST transcription factor binding sites (TFBS). Locations for ENCODE TFBS were found in 10 different cell types. 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 column.*