TiSAn: Estimating Tissue Specific Effects of Genetic Variants

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 non-coding genetic variation on molecular functions, such as gene expression, varies across tissues and cell types. Consequently, whether or not a genetic variant is deleterious depends on its tissue-specific context. 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). These predictions can then be used to contextualize and filter variants of interest in whole genome sequencing or genome wide association studies (GWAS). 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.

Author Summary:

Genome-wide association studies (GWAS) provide insights into the mechanisms underlying diseases and other traits, yet interpreting the effect of non-coding genetic variation (which represents the majority of GWAS hits) is an ongoing and unresolved challenge. This is evident because while we as a field are sequencing more and more whole genomes, we often retreat to the exome portion whenever we want to propose a clear interpretation of the variants we discover. Current annotation approaches are effective at predicting whether a given variant is damaging, and potentially disease-associated. However, these methods do not provide information on which organs or tissues are most susceptible to the variant's effect. The work presented here addresses these challenges. It describes the function and implementation of a machine learning approach (TiSAn) to annotate genetic variants, especially non-coding variants, according to their function in a specific tissue. Such annotation is critical in prioritizing genetic variations and making inferences about which tissues, organs, and systems they will most likely impact. We detail several use cases, including interpreting variants from whole genome sequencing studies of autism, as well as GWAS hits from studies of cardiovascular disease. Further analyses demonstrating the power of tissue-specific variant annotation are included in the supplementary materials. Thorough documentation for TiSAn, including tutorials, can be found at http://github.com/kevinVervier/TiSAn.

Introduction:

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.

Results:

Machine learning for predicting tissue-specific functional annotation

The design of TiSAn models is outlined in Figure 1 (details in Materials & Methods section). 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.

Brain-specific variant prioritization in a sample with familial risk for autism

Genome-wide enrichment for brain-related variations in affected individuals

The Simons Simplex Collection (SSC, http://base.sfari.org) 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)[8]. 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, $P = 0.351$).

Case-control variants filtering with brain-specific annotation

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.

Autism and calcium channel genes

An autism-related calcium voltage-gated channel gene, *CACNA1C*[9], was the gene with the highest TiSAn-brain score enrichment in SSC data, suggesting that deleterious variants at this locus are more likely to affect brain function. In particular, we identified 116 non-coding deleterious variants (see methods for CADD thresholds) in CTCF[10] transcription factor binding sites around *CACNA1C* location in SSC data, while none were found in unaffected population. Mutations in the same region also hit non-coding RNAs (ncRNAs) more frequently in SSC population than in control population (Fisher's exact test, $P < 2x10^{-16}$). Interestingly, 5 of these ncRNAs (Supplementary Table 4) were found in linkage disequilibrium with loci associated with autism and Tourette's syndrome according to the LincSNP database[11].

Heart-related signal prioritization in coronary artery disease

Genome-wide association strength and annotation score

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[12], and we demonstrate that the TiSAn-heart score is significantly higher among the most associated variants (Fig. 3, (Student t-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).

Reduction of multiple hypothesis burden in GWAS

We filtered tissue-relevant genotyped variants before the GWAS analysis, so that only heart-related variants would be considered in the correction for multiple testing. In the case of CAD-GWAS, this reduced the number of SNPs considered by 75%. This resulted in significant loci being narrowed by 20% on average (paired Student t-test, $P = 0.019$). Furthermore, the overall enrichment in transcription factor binding sites (TFBS) among significant loci is conserved between the original set of and the TiSAn-filtered one (Chi-squared test, $P = 0.51$), suggesting that the regulatory content is preserved after the filtering step (a further analysis of TFBS, provided in the supplementary information, suggests that TiSAn can reveal which TFs have important functional roles in specific tissues). Reducing the number of tested variants directly recalibrates the multiple-testing correction threshold used to determine significant loci from $5x10^{-8}$ to $1.6x10^{-7}$. Here, 91 new loci were found significantly associated with coronary artery disease, and show a significant enrichment in EBF1 TFBS (Fisher's Exact test, $P = 3.2x10^{-6}$), which has been previously linked to obesity, diabetes, and cardiovascular disease[13].

Discussion

Integrative approaches like TiSAn 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. However, few such tools currently exist, with most development efforts focusing on improving estimators of general (and not tissue-specific) deleteriousness[14]. GenoSkyLine, a recently developed tool that utilizes genome-scale tissue-specific epigenetic data, allowed us to benchmark TiSAn and demonstrate its effectiveness in prioritizing genetic variants that are most likely to play a role in the tissue-specific disease processes under consideration.

Specifically, we showed that individuals with elevated risk for autism (i.e., probands and their family members) had more deleterious WGS variants that were predicted to be brain relevant (by TiSAn-Brain) than controls. At the same time, we were unable to show any such differences in regions identified as brain-relevant by GenoSkyLine. Specificity was demonstrated in this

analysis by the absence of a similar case/control difference in deleterious variants predicted (by TiSAn-Heart) to be relevant to cardiovascular tissue. Further, we showed that strongly associated GWAS hits in a study of coronary artery disease have a significantly higher TiSAn-Heart signal than non-associated SNPs, supporting our method's ability to correctly prioritize tissue-specific variants. Again, we were unable to observe this difference using the GenoSkyLine score for cardiovascular tissue. We also demonstrated the practical advantages of reducing GWAS multiple testing burden by pre-filtering SNPs on the basis of their estimated tissue relevance. In each of these analyses, TiSAn showed an ability to correctly prioritize variants according to tissue specific action, while we were unable to do so with GenoSkyLine, the current state of the art for this application. TiSAn thus represents an important development towards leveraging the massive amount of underutilized information (i.e., non-coding variation) coming from whole genome sequencing studies.

Several technical points related to the development of TiSAn are worth mentioning. First, comparison between multiple machine learning algorithms (Supplementary Table 2) led us to use random forests, known to better handle non-linearity and correlation between variables. Recently, deep learning has been evaluated in the context of variation effects on chromatin[15], and future analyses will investigate the impact of using such an algorithmic framework. Another issue is that supervised learning requires genomic positions with accurate class labels, in this case, known to be either associated with a given tissue or not. However, for most of the available data, such a label does not exist, especially for positive association with a tissue-specific trait. Imbalance-aware machine learning[16] could be a solution to efficiently train predictive models in the case of underrepresented classes.

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.

Materials & 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 a 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). A complete list of the positions used for training brain and heart models is provided in Github vignettes (http://github.com/kevinVervier/TiSAn/tree/master/vignettes). 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[17] for cardiovascular diseases, or Illumina Infinium PsychArray Beadchip for psychiatric disorders. These probe sets contain tissue-related variants (positive examples), but also backbone/non-psychiatric variants which we consider as negative examples if they meet a minimal CADD threshold described below.

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[11], 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), i.e., deleterious randomly sampled variants not related to the tissue at hand. 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. [18], 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:

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

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[15]. Recently, specific patterns have been identified to be tissue-specific[19], and we incorporate this information by computing frequencies for all $\boldsymbol{n}$-nucleotides ($\boldsymbol{n} \in (\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4})$), found in a +/- 500 base pair 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[20]. 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[21]. 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 the Weibull distribution assumption (Supplementary Fig. 2).

Epigenomics and in particular, methylation profiles have been integrated to explain tissue specific regulatory mechanisms[22]. 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 projects focusing on a single tissue. For the brain model, we integrate developmentally differentially methylated positions (dDMPs) [23] found in fetal brain. For the heart model, Heart Enhancer Compendium database[24] 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[25] 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:

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

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 performed in this study used the 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).

Variant 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 top 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[26] 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. 3c) 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 site (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

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. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, et al. (2014) Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A 111: 6131-6138.
2. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, et al. (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res 15: 1051-1060.
3. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46: 310-315.
4. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 48: 214-220.
5. Consortium GT (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648-660.
6. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, et al. (2010) The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28: 1045-1048.
7. Lu Q, Powles RL, Wang Q, He BJ, Zhao H (2016) 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 Genet 12: e1005947.
8. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.
9. Kabir ZD, Che A, Fischer DK, Rice RC, Rizzo BK, et al. (2017) Rescue of impaired sociability and anxiety-like behavior in adult cacna1c-deficient mice by pharmacologically targeting eIF2alpha. Mol Psychiatry.
10. Prickett AR, Barkas N, McCole RB, Hughes S, Amante SM, et al. (2013) 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 Res 23: 1624-1635.
11. Ning S, Zhao Z, Ye J, Wang P, Zhi H, et al. (2014) LincSNP: a database of linking disease-associated SNPs to human large intergenic non-coding RNAs. BMC Bioinformatics 15: 152.
12. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, et al. (2015) A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet 47: 1121-1130.
13. Singh A, Babyak MA, Nolan DK, Brummett BH, Jiang R, et al. (2015) Gene by stress genome-wide interaction analysis and path analysis identify EBF1 as a cardiovascular and metabolic risk gene. Eur J Hum Genet 23: 854-862.
14. Capriotti E, Fariselli P (2017) PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. Nucleic Acids Res.
15. Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 12: 931-934.
16. Schubach M, Re M, Robinson PN, Valentini G (2017) Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. Sci Rep 7: 2959.
17. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, et al. (2012) The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet 8: e1002793.
18. Cherkasov A, Ho Sui SJ, Brunham RC, Jones SJ (2004) Structural characterization of genomes by large scale sequence-structure threading: application of reliability analysis in structural genomics. BMC Bioinformatics 5: 101.
19. Zhong S, He X, Bar-Joseph Z (2013) Predicting tissue specific transcription factor binding sites. BMC Genomics 14: 796.
20. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, et al. (2016) Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. PLoS Genet 12: e1006423.
21. Liu Y, Liang Y, Wishart D (2015) PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic Acids Res 43: W535-542.
22. Miller CL, Pjanic M, Wang T, Nguyen T, Cohain A, et al. (2016) Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. Nat Commun 7: 12092.
23. Spiers H, Hannon E, Schalkwyk LC, Smith R, Wong CC, et al. (2015) Methylomic trajectories across human fetal brain development. Genome Res 25: 338-352.
24. Dickel DE, Barozzi I, Zhu Y, Fukuda-Yuzawa Y, Osterwalder M, et al. (2016) Genome-wide compendium and functional assessment of in vivo heart enhancers. Nat Commun 7: 12923.
25. Lischke H, Loffler TJ, Fischlin A (1998) Aggregation of individual trees and patches in forest succession models: capturing variability with height structured, random, spatial distributions. Theor Popul Biol 54: 213-226.
26. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44: D733-745.

Figure Legends:

Figure 1: TiSAn framework overview*. Each nucleotide position in the genome is annotated with multiple levels of omics information, such as sequence content,*

methylation level, proximity to genes, etc. (see Methods). This information is extracted for training sets, comprised of deleterious variants with or without an association with the tissue of interest. Using supervised machine learning, specifically a Random Forest (RF), a predictive model combines each feature with respect to its ability to predict whether a position will be functionally associated with the tissue of interest. Model output consists in a tissue-specific functional score spanning between no functional relevance to the tissue (0) and strong functional relevance to the tissue (1). This score can then be used, for instance, to filter down large lists of candidate variants for further investigation, or to isolate the contribution of different tissues to a complex trait.

Figure 2: Brain-related functional enrichment in a case-control 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).*

Figure 3: 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 report average functional scores (blue: TiSAn-heart, grey: GenoSkyline-heart). Shaded areas represent confidence interval for the corresponding method, after GWAS P-value random permutations.*

Supporting Information Legends:

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 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 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: 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. **(A)TiSAn-heart enrichment in CEBPB TFBS**. Locations for ENCODE TFBS were found in 6 different cell types. **(B) TiSAn-brain enrichment in REST TFBS**. Locations for ENCODE TFBS were found in 10 different cell types.*

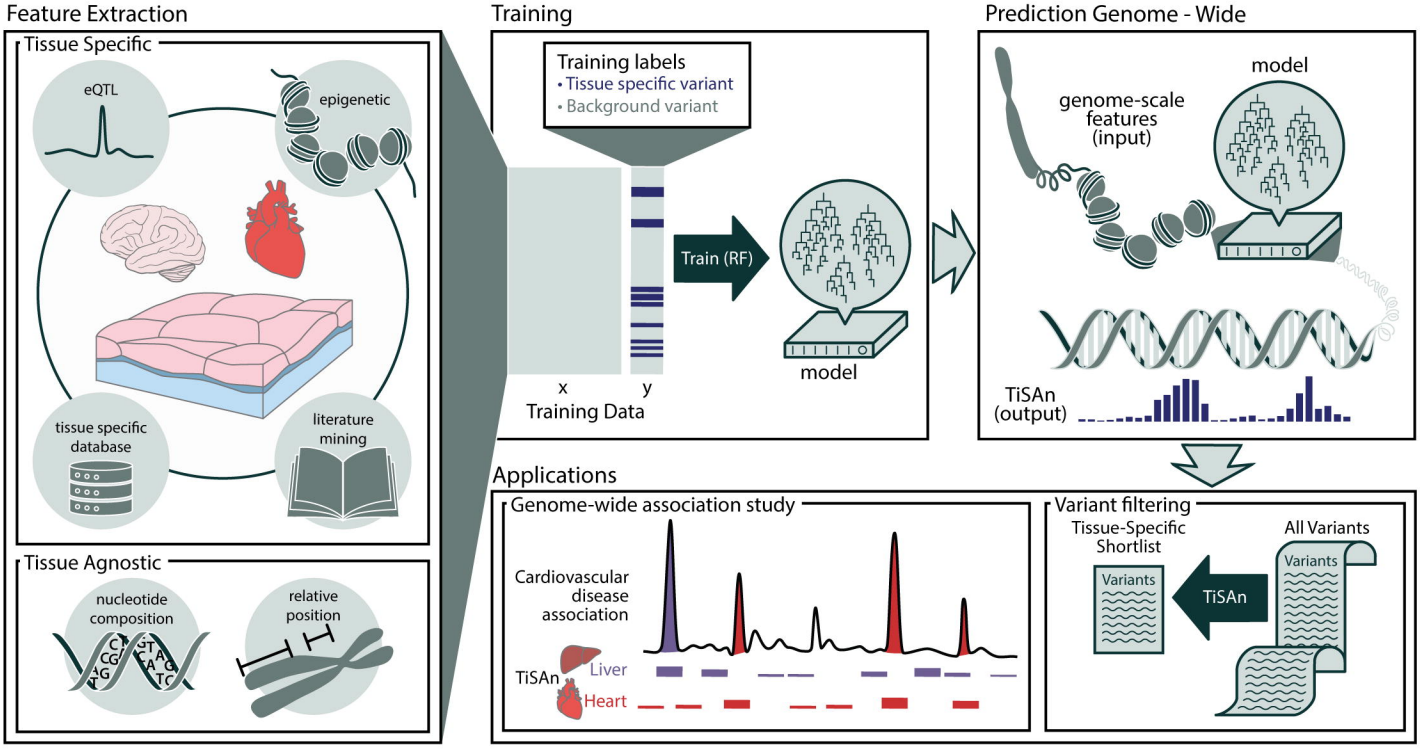
***Supplementary Figure 8**: TiSAn-brain enrichment in CTCF transcription factor binding sites (TFBS). Locations for ENCODE TFBS were found in 70 different cell types. Functional score profiles were obtained using a 1,000bp window centered on the TFBS (dash line). Positive enrichment (orange) and negative enrichment (blue) are reported for each different cell type in row.*

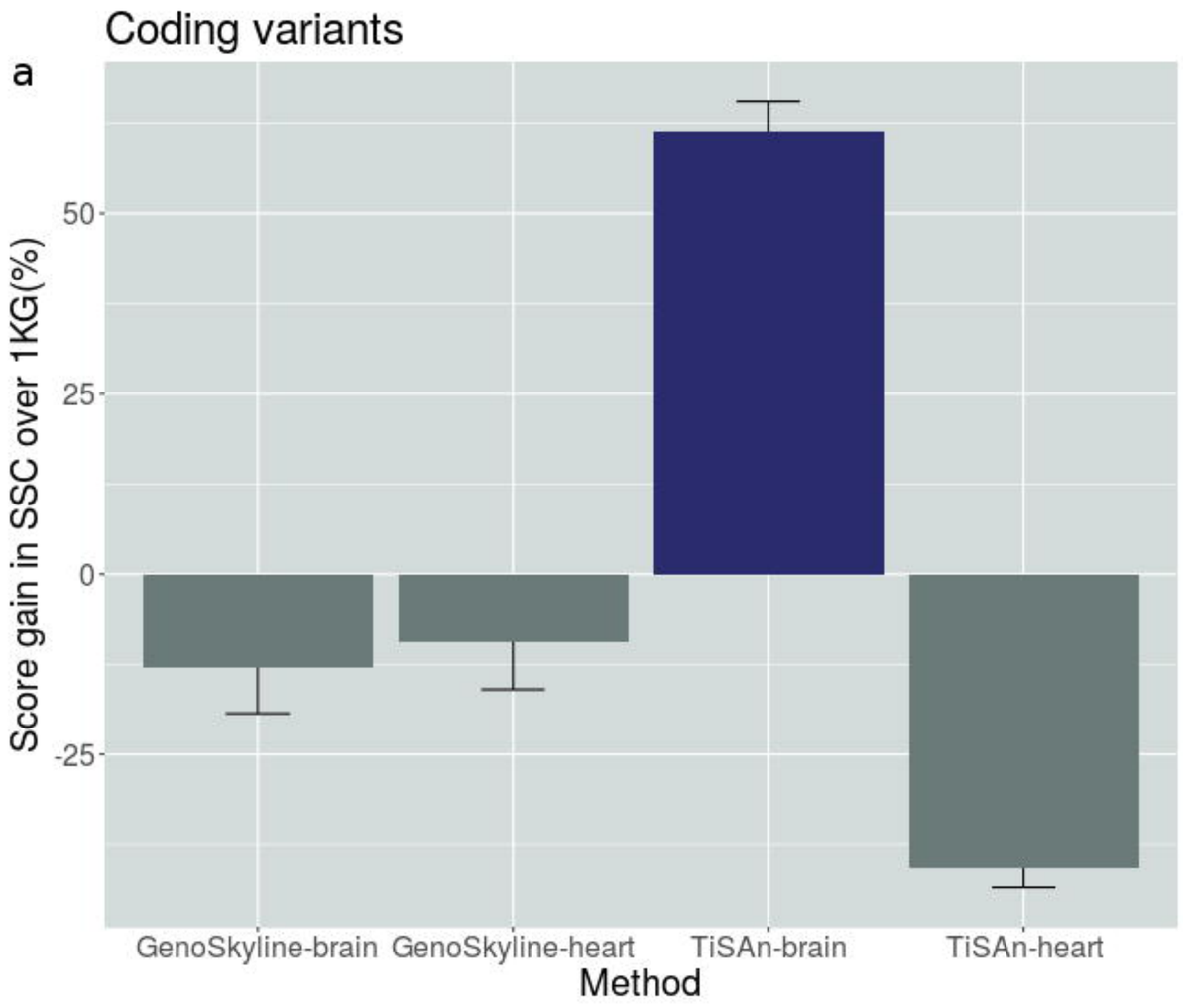
***Supplementary Table 1: Training set composition.** For both heart and brain tissues, we report the count of positive and negative examples used to train TiSAn models. The counts are divided in two parts, corresponding to variants found in large intergenic non-coding RNAs database, or in* genotype array probesets*.*

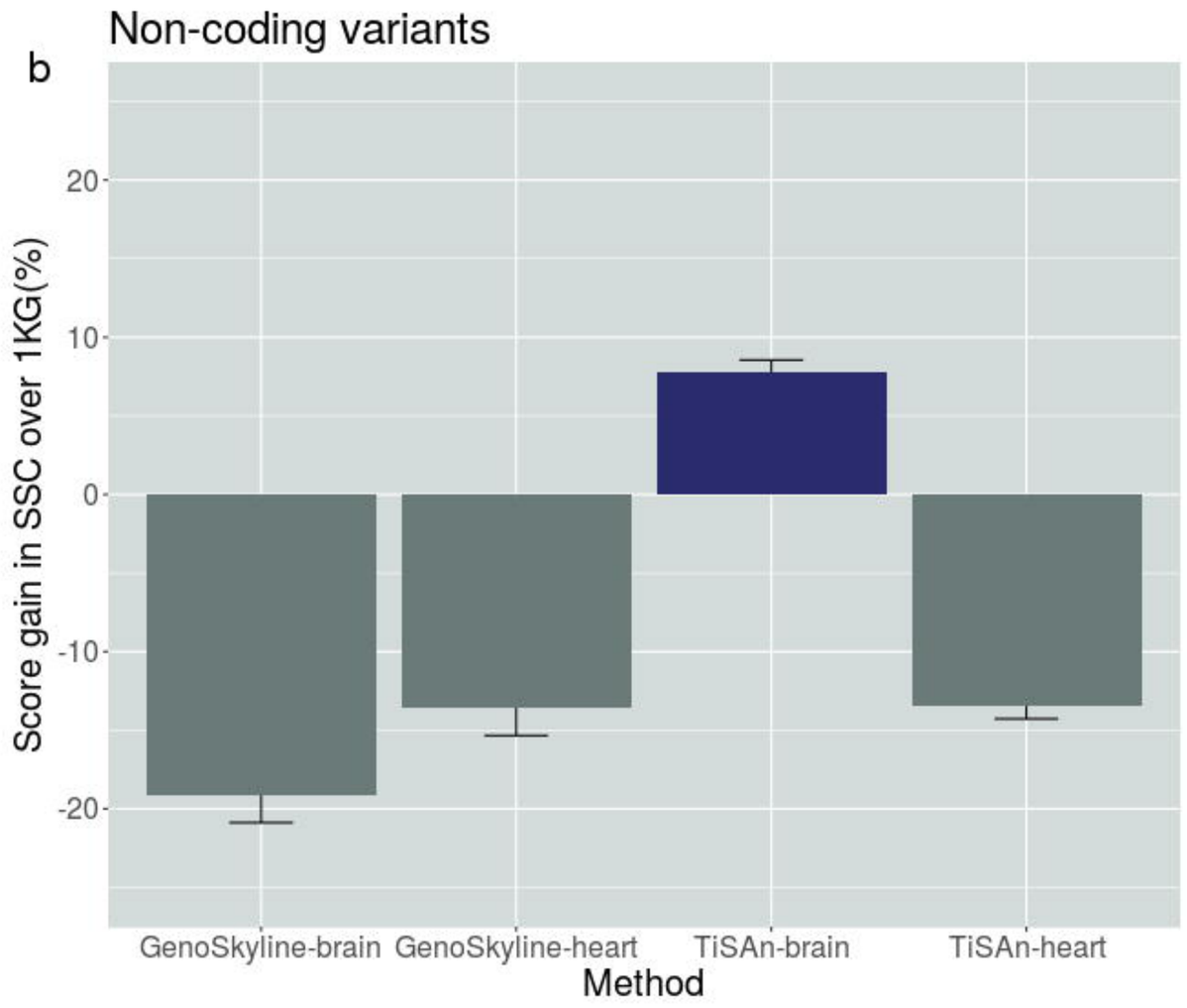
***Supplementary Table 2**: 10-folds cross-validated performances obtained during TiSAn-brain model training, for different classification strategies. AUC: Area under ROC curve.*

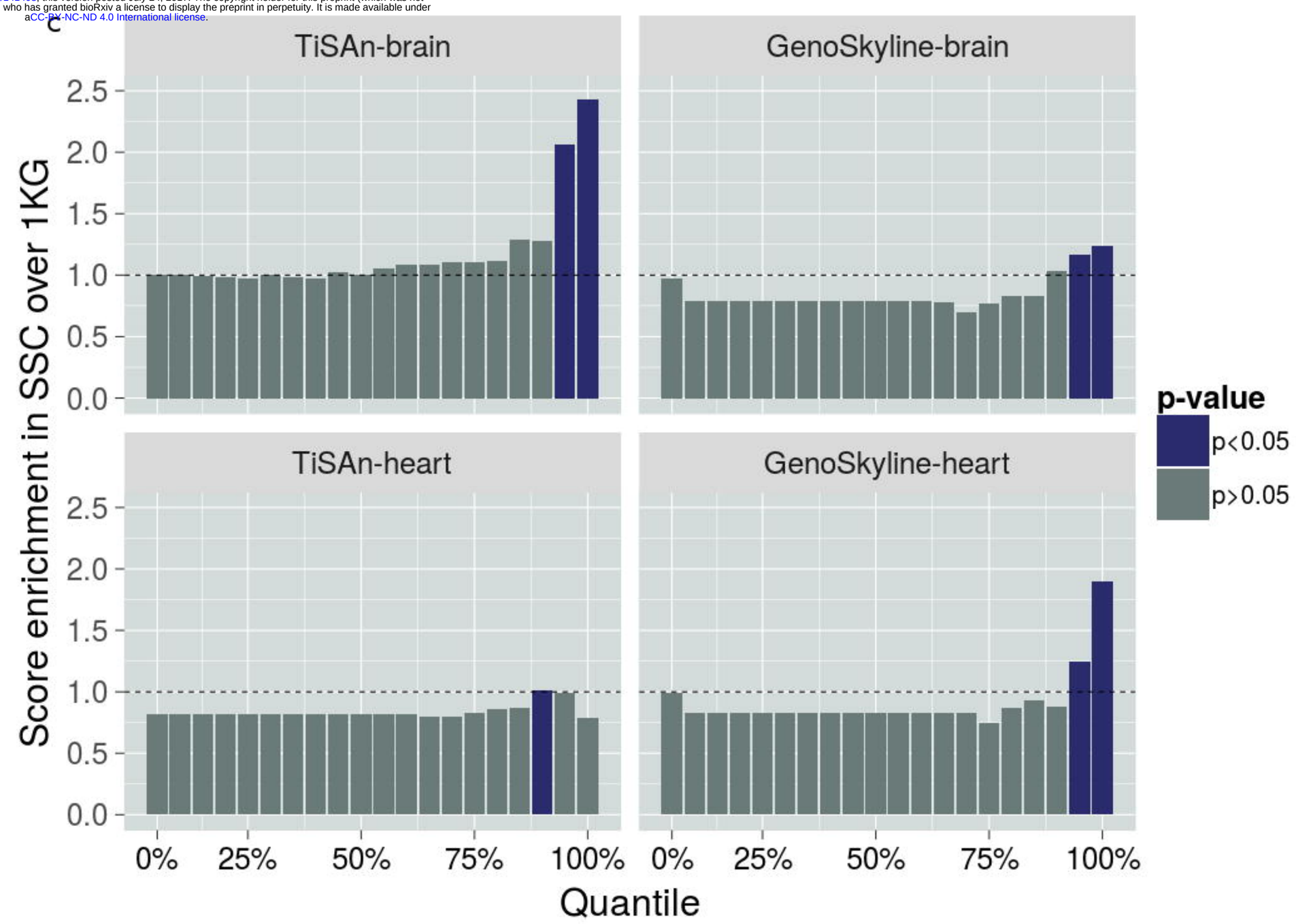
***Supplementary Table 3**: List of SFARI autism-related genes, supported by literature.*

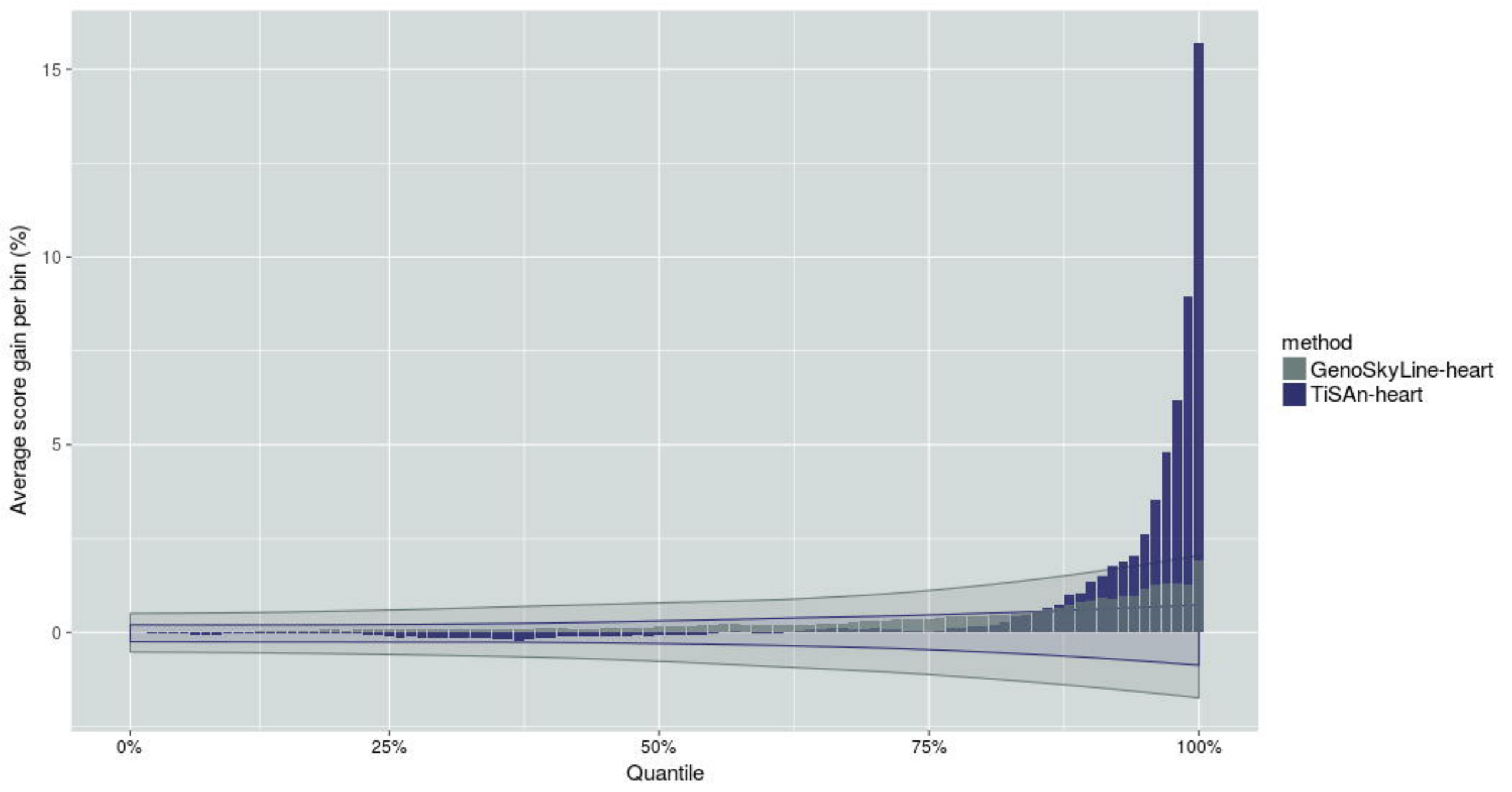
***Supplementary Table 4**: Non-coding RNAs found in linkage disequilibrium with neurodevelopmental and psychiatric disorders*


Feature Extraction
Tissue Specific
eQTL
epigenetic
tissue specific database
literature mining
Tissue Agnostic
nucleotide composition
relative position
Training
Training labels
• Tissue specific variant
• Background variant
Train (RF)
model
x
y
Training Data
Prediction Genome - Wide
model
genome-scale features (input)
TiSAn (output)
Applications
Genome-wide association study
Cardiovascular disease association
TiSAn
Liver
Heart
Variant filtering
Tissue-Specific Shortlist
All Variants
Variants
Variants
TiSAn








15
10
5
0
Average score gain per bin (%)
0%
25%
50%
75%
100%
Quantile
method
GenoSkyLine-heart
TiSAn-heart