Tissue-specific functional annotation for genetic variation

Dr. Kévin Vervier, PhD

Department of Psychiatry
University of Iowa Hospitals and Clinics

May 24, 2017

1. Context
2. Collect information at the tissue level
3. Machine-learning for functional annotation
4. Tutorial
5. Applications

Context

Functional annotation for genetic variation

- Whole-genome sequencing to understand genetic trait architecture in large cohorts.
- Given thousands of candidate variants, prioritize candidate variants.
- Especially difficult for non-coding variants.
- Recent scores (e.g. CADD/DANN/Eigen) predict generic deleteriousness annotation.
- Idea: using tissue-related data to derive a functional score.

Strategy

We propose a new tool, **Ti**ssue **S**pecific **An**notation (TiSAn*) that:

- measures how likely a position is related to tissue functions,
- returns high score for tissue-related positions,
- can easily be adapted to many tissues,
- is a predictive model, based on machine learning,

$$\mathbb{P}(Y = \textit{tissue} \mid X) = f(X; w).$$

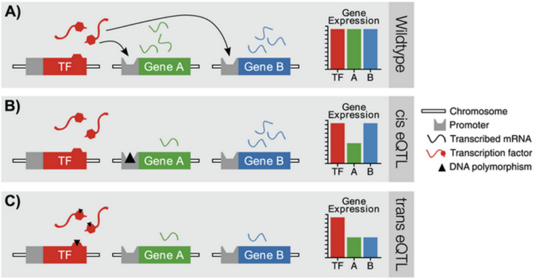
*also the french word for herbal tea

Collect information at the tissue level

Gene-Tissue Expression (GTEx)

GTExPortal

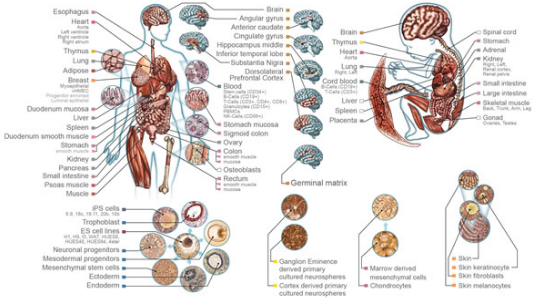
- study gene expression in $\sim$ 50 different tissues
- genotype available for most of the donors
- tissue-specific expression Quantitative Trait Loci (eQTL)



Source: Wolen and Miles, 2012

ENCODE/RoadMap Epigenomics (RME)

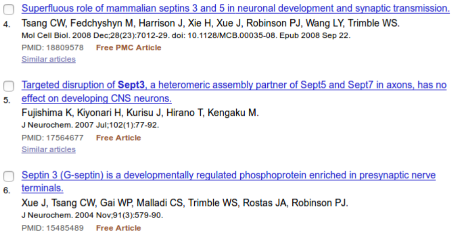
- map DNA methylation in more than 80 cell types
- tissue-specific regulation mechanisms
- differentially methylated regions



Source: RME Consortium, 2015

NCBI PubMed

- gene2pubmed: curated database for articles citing genes
- literature mining for tissue-related genes (NCBI API)
- query on tissue-gene co-citations (title+abstract)



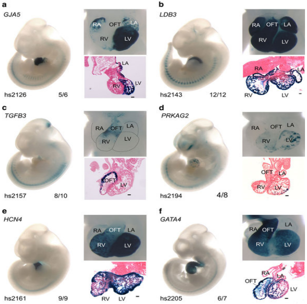
□ 4. Superfluous role of mammalian septins 3 and 5 in neuronal development and synaptic transmission.
Tsang CW, Fedchyshyn M, Harrison J, Xie H, Xue J, Robinson PJ, Wang LY, Trimble WS.
Mol Cell Biol. 2008 Dec;28(23):7012-29. doi: 10.1128/MCB.00035-08. Epub 2008 Sep 22.
PMID: 18809578 **Free PMC Article**
Similar articles

□ 5. Targeted disruption of **Sept3**, a heteromeric assembly partner of Sept5 and Sept7 in axons, has no effect on developing CNS neurons.
Fujishima K, Kiyonari H, Kurisu J, Hirano T, Kengaku M.
J Neurochem. 2007 Jul;102(1):77-92.
PMID: 17564677 **Free Article**
Similar articles

□ 6. Septin 3 (G-septin) is a developmentally regulated phosphoprotein enriched in presynaptic nerve terminals.
Xue J, Tsang CW, Gai WP, Malladi CS, Trimble WS, Rostas JA, Robinson PJ.
J Neurochem. 2004 Nov;91(3):579-90.
PMID: 15485489 **Free Article**

Projects dedicated to one tissue

- previous databases work for a large set of tissues
- we also integrate data from single-tissue projects
- developmental brain methylation (Spiers et al., 2015)
- fetal heart enhancers (Dickel et al., 2016)
- especially relevant for rare tissues



Source: Dickel, 2016

Machine-learning for functional annotation

Features space description

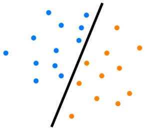
For each genomic position, we extract the following tissue-specific descriptors

- transcriptomics:
  - distance to the closest tissue eQTL (GTEx)
  - distance to the closest 'tissue gene' (PubMed)
- epigenomics:
  - distance to methylation regions (RME)
  - methylation level (RME)
- genomics:
  - *n*-nucleotides composition in 1kb neighborhood ($n = 1, 2, 3, 4$)
- single-tissue data:
  - fetal brain methylation (Spiers *et al.*, 2015)
  - heart enhancers (Dickel *et al.*, 2016)

Currently, ∼360 features are used in the data representation.

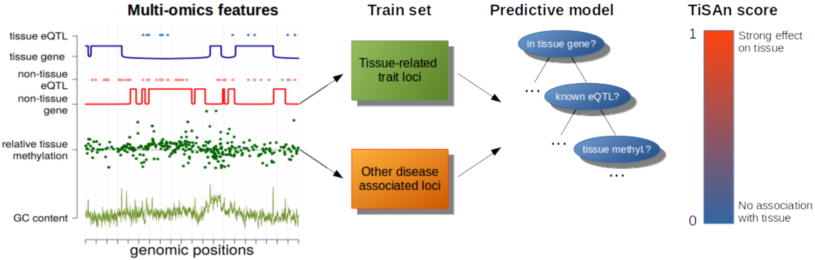
Training set definition

- Supervised machine learning consists in separating positive and negative examples
- Decision rule optimization and pattern detection



- Association between a location and a tissue can be found in disease-related loci.
- Online databases like GWAS Catalog, genotype array probes (e.g., PsychArray and MetaboChip).

Method overview



TiSAn framework

- Applied on two human tissues: brain and heart.

Tutorial

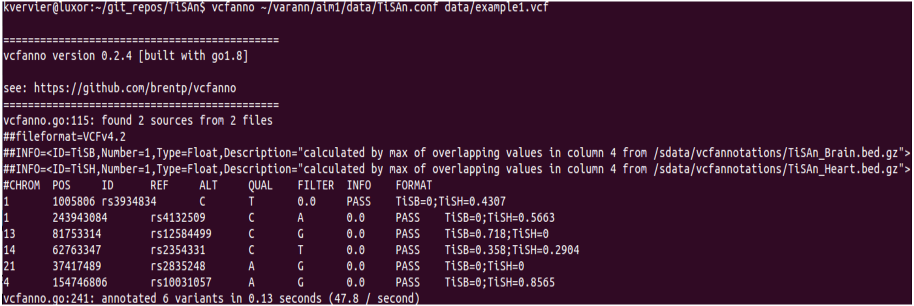
Tutorial: database + vignettes

- **R** vignettes are distributed as a package and provide guidelines for each model development step
- Genome-wide scores for heart and brain are available at `http://flamingo.psychiatry.uiowa.edu/TiSAn`
- .bed format makes TiSAn easy to integrate in most bioinformatics pipelines

Github: `http://github.com/kevinVervier/TiSAn`

Tutorial: get TiSAn scores for candidate loci

- Input: VCF file with loci of interest.
- Plug TiSAn databases in, for instance, *vcfanno* tool.



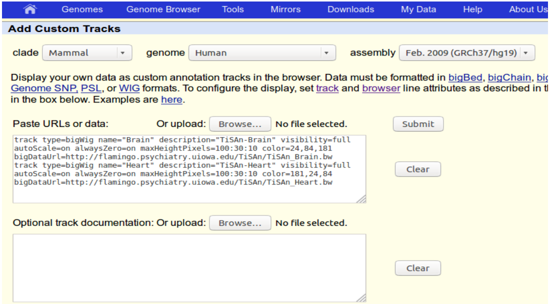
```
kvervier@luxor:~/git_repos/TiSAn$ vcfanno ~/varann/aim1/data/TiSAn.conf data/example1.vcf

=============================================
vcfanno version 0.2.4 [built with go1.8]

see: https://github.com/brentp/vcfanno
=============================================
vcfanno.go:115: found 2 sources from 2 files
##fileformat=VCFv4.2
##INFO=<ID=TiSB,Number=1,Type=Float,Description="calculated by max of overlapping values in column 4 from /sdata/vcfannotations/TiSAn_Brain.bed.gz">
##INFO=<ID=TiSH,Number=1,Type=Float,Description="calculated by max of overlapping values in column 4 from /sdata/vcfannotations/TiSAn_Heart.bed.gz">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT
1       1005806 rs3934834       C       T       0.0     PASS    TiSB=0;TiSH=0.4307
1       243943084       rs4132509       C       A       0.0     PASS    TiSB=0;TiSH=0.5663
13      81753314        rs12584499      C       G       0.0     PASS    TiSB=0.718;TiSH=0
14      62763347        rs2354331       C       T       0.0     PASS    TiSB=0.358;TiSH=0.2904
21      37417489        rs2835248       A       G       0.0     PASS    TiSB=0;TiSH=0
4       154746806       rs10031057      A       G       0.0     PASS    TiSB=0;TiSH=0.8565
vcfanno.go:241: annotated 6 variants in 0.13 seconds (47.8 / second)
```

Tutorial: visualization in UCSC Genome Browser

- Step 1: Access the UCSC Genome Browser custom track page
- For both TiSAn scores, paste instructions in "Paste URLs or data" box
- Submit your query.



Add Custom Tracks

clade Mammal genome Human assembly Feb. 2009 (GRCh37/hg19)

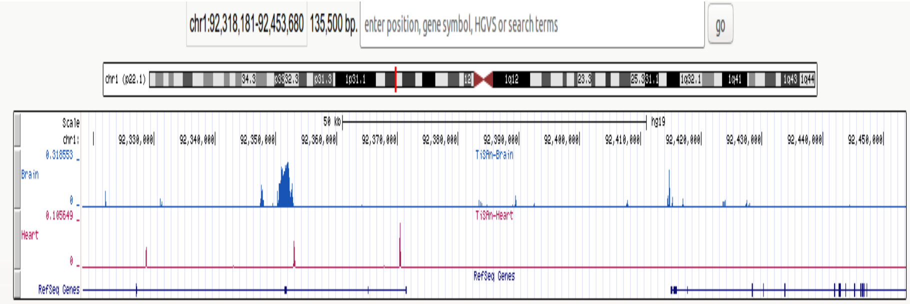
Display your own data as custom annotation tracks in the browser. Data must be formatted in bigBed, bigChain, bi
Genome SNP, PSL, or WIG formats. To configure the display, set track and browser line attributes as described in t
in the box below. Examples are here.

Paste URLs or data: Or upload: Browse... No file selected.

```
track type=bigWig name="Brain" description="TiSAn-Brain" visibility=full
autoScale=on alwaysZero=on maxHeightPixels=100:30:10 color=24,84,181
bigDataUrl=http://flamingo.psychiatry.uiowa.edu/TiSAn/TiSAn_Brain.bw
track type=bigWig name="Heart" description="TiSAn-Heart" visibility=full
autoScale=on alwaysZero=on maxHeightPixels=100:30:10 color=181,24,84
bigDataUrl=http://flamingo.psychiatry.uiowa.edu/TiSAn/TiSAn_Heart.bw
```

Optional track documentation: Or upload: Browse... No file selected.

Tutorial: visualization in UCSC Genome Browser

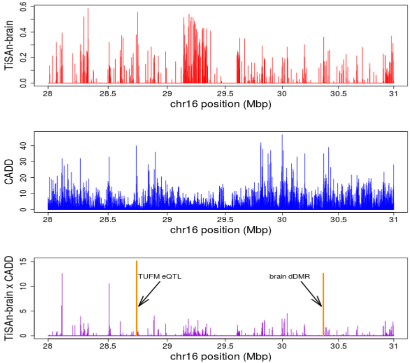


chr1:92,318,181-92,453,680 135,500 bp.

Applications

Application: region-based analysis: 16p11

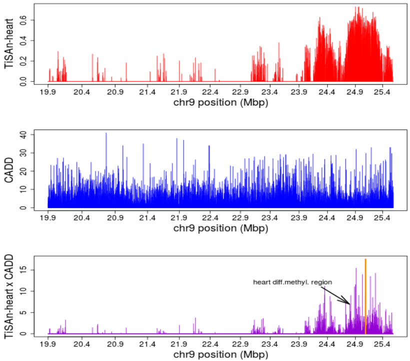
- Known to be related to autism and schizophrenia.
- On Chromosome 16, from $\sim$ 28.3Mb to 30.3Mb.
- TiSAn combined with pathogenicity scores, like CADD.



TiSAn-brain (top), CADD (mid), TiSAn × CADD (bot)

Application: region-based analysis: 9p21

- Known to be related to cardiovascular disease.
- On Chromosome 9, from ∼ 19.9Mb to 25.5Mb.
- TiSAn combined with pathogenicity scores, like CADD.



TiSAn-heart (top), CADD (mid), TiSAn × CADD (bot)

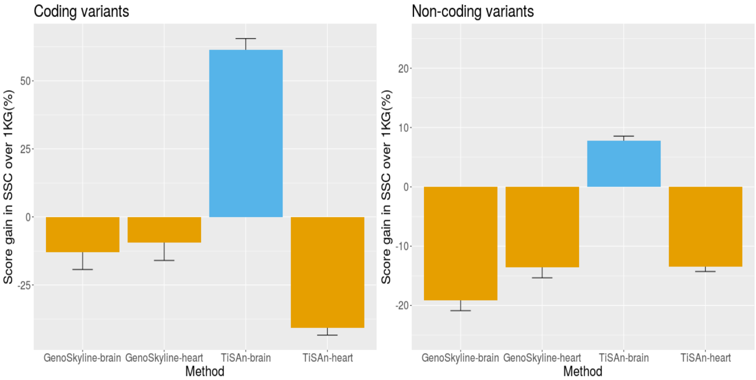
Autism genes set enrichment

SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

- Simons Simplex Collection (SSC): genetic repository of 2,600 simplex families with autistic proband
- 1,000 Genomes (1KG): genetic variation in unaffected population
- Expected enrichment in brain-related genetic burden in SSC cohort, even in unaffected family members.

Autism genes set enrichment

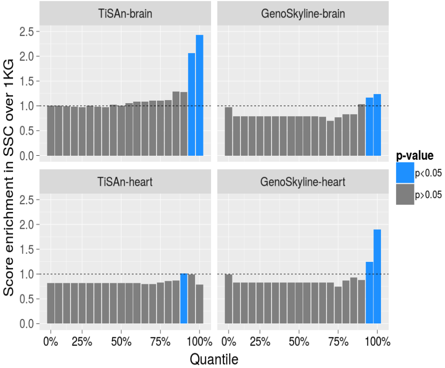
- Pathogen variants found around strong autism candidate genes
- Comparison in average score between SSC and 1KG variants
- Comparison with GenoSkyLine (Lu *et al.*, 2016)



Relative score gain between 1KG and SSC.

Autism genes set enrichment

- Mix SSC and 1KG variants and rank them based on their scores.
- For each quantile, compute relative enrichment in SSC over 1KG.



Quantile-wise enrichment in SSC variants over 1KG.

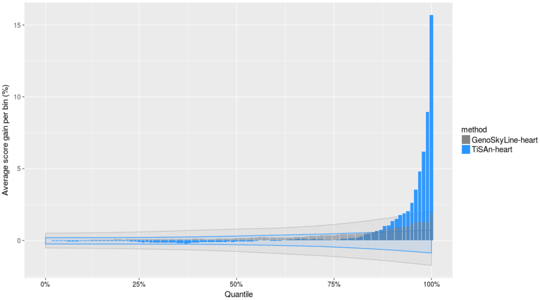
Genome-wide association for coronary artery disease

CARDIoGRAMplusC4D

- Combine multiple large scale genetic studies to identify risk loci for coronary artery disease
- Estimate trait association (*p*-value) for ∼8,000,000 SNPs.
- Hypothesis: TiSAn-heart score increases with association strength.

Genome-wide association for coronary artery disease

- Loci binned based on their *p*-values into percentile groups
- Score gain between top percentile and remaining groups



Cumulative quantile-wise functional score enrichment.

Conclusion

- General framework for tissue-related functional score
- Enrichment found for both brain and heart models in known loci
- Next steps:
  - Evaluate deep-learning based solutions
  - Discovery analysis in unpublished data (bipolar disorder, sudden death)
  - Combine with SLINGER for tissue-specific gene expression inference

Thank you for your attention

Fundings: NIH MH105527 and DC014489
Collaborators: Dr Jacob Michaelson lab (UI)

Brew your own TiSAn!
Github: `http://github.com/kevinVervier/TiSAn`