Tissue-specific gene expression inference: application to human brain disorders.

Kevin Vervier[1], Leo Brueggeman[1] & Jacob Michaelson[1]

[1]Department of Psychiatry, The University of Iowa

Summary

machine learning training step

TiSAn

0

brain-SLINGER gene-level weights

	SNP2	SNP3
gene A	0	-1.8
gene B	1.2	0.5
gene C	-1.4	0

whole-blood expression

	gene A	gene B	gene C
Individual 1	2	0.1	1.2
Individual 2	2.5	0.2	1.3
Individual 3	3.5	1.2	0.2
Individual 4	0.1	1.5	1.1

	SNP1	SNP2	SNP3
Individual 1	0	1	0
Individual 2	0	1	0
Individual 3	1	2	0
Individual 4	1	1	2

genotype data

brain gene expression inference

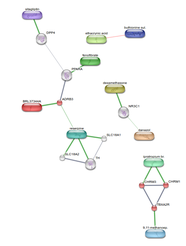
	SNP2	SNP3
case 1	2	2
case 2	2	1
control 1	1	2
control 2	0	2

new measured genotype

	gene A	gene B	gene C
case 1	0.1	1.6	**0.2**
case 2	0.6	1.4	**0.2**
control 1	0.1	1.2	**1.1**
control 2	0.1	0.8	**1.8**

differentially expressed genes

downregulated gene C in cases



candidate drug targets

Introduction

Gene expression varies with tissue context. Projects, like Gene-Tissue Expression (**GTEx**), have permitted quantification of such variation. Wheeler et al. (2016) estimated tissue-specific gene expression with the use of local heritability only, mentioning that current sample sizes in GTEx do not allow for integrating **distal heritability**. In this study, we propose an alternative framework, where we use more training observations from whole-blood samples, to predict gene expression in the human brain. In particular, we demonstrate that:

- filtering genotype based on tissue-specific functional score (**TiSAn**, [1]) improves inference quality,
- **transfer learning** can be used from **whole blood** (WB) to other tissues, such as **human brain**,
- brain models achieve higher brain enrichments compared to state of the art approaches: PrediXcan (Gamazon et al., 2015) and SLINGER [2].
- applying **brain-SLINGER** models to a large **bipolar disorder** cohort identified **103 candidates genes**, that we extensively validated through literature enrichment and **pharmacogenomics**.

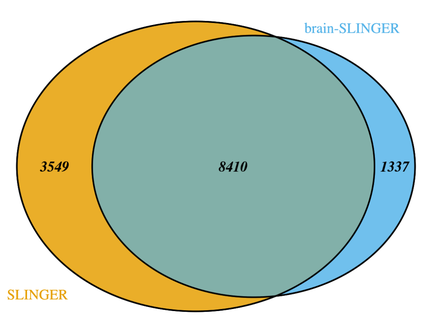
Transfer learning from whole blood to brain tissue

- Most of the local signal is shared across tissues [3] and distal interactions are enriched in tissue-specific features.
- We propose to combine two recently proposed approaches: SLINGER and TiSAn to capture brain-related signal from whole-blood expression data.
- For a given gene *g*, the model is a weighted combination of genotyped loci *s* related to *brain*:

$$\text{brain-SLINGER: } T_g = \sum_{s:\text{TiSAn}(s)>0} \gamma_{g,s} X_s$$

- **Elastic-net** penalization is used to limit the numbers of active SNPs in each gene model.
- We assume that the **brain-related enrichment in training support** helps infer accurate gene expression.

Validation and comparison with SLINGER



- We found 1,337 new genes only accessible through the brain enriched procedure.
- These genes are significantly enriched in brain-related functions: **neurological system process, signal transducer activity and receptor activity**.
- 3,549 genes are no longer accessible in brain-SLINGER, but no functional enrichment was found for those genes.
- using TiSAn filter **reduces spurious associations** in SLINGER models (*t*-test on R-squared $P = 6.3E\text{-}40$).
- brain models have a **significantly lower local enrichment** (KS-test $P < 2.2E\text{-}16$) compared to SLINGER models.

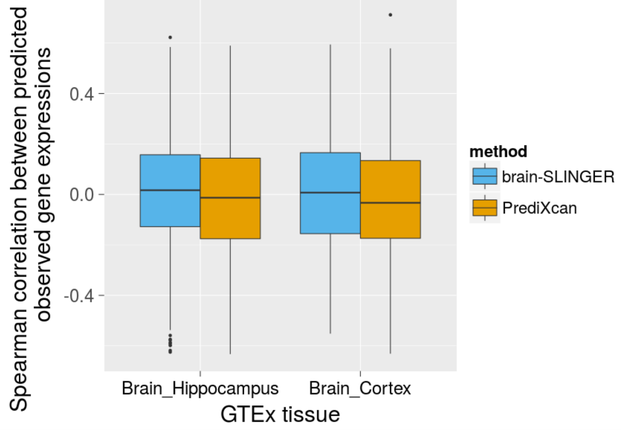
Correlation with GTEx brain tissues expression

- From GTEx genotypes, ~**2,000 genes expression** were inferred from SLINGER and brain-model.
- For each gene, find the best inference method, using the **correlation with measured expression**.
- Genes better predicted by brain models are **enriched** for Gene Ontology (GO) biological processes and molecular functions in **9 of 10 GTEx brain regions**.
- **No GO term enrichment** was found in the genes set better predicted by SLINGER.

GTEx tissue	top GO term	corrected *P*-value
Hypothalamus	positive regulation of **neurogenesis**	8.84E-06
Putamen basal ganglia	**calcium channel** activity	4.1E-04
Cerebellum	cell **migration**	1.15E-03
Cerebellar hemisphere	muscle contraction	4.56E-03
Anterior cingulate cortex	regulation of **axogenesis**	5.13E-03
Frontal cortex	cell **differentiation**	1.59E-02
Hippocampus	**neuron** differentiation	1.64E-02
Caudate basal ganglia	regulation of histone acetylation	1.77E-02
Cortex	**transmembrane** receptor	3.42E-02

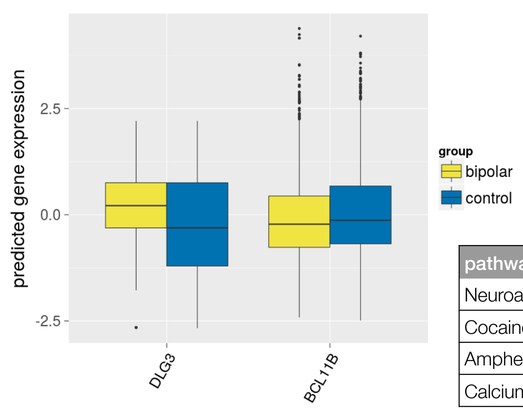
Comparison with tissue-specific PrediXcan

- PrediXcan models were trained based on GTEx expressions, using local interactions only.
- For each brain region, ~**2,000 genes expression** were inferred from PrediXcan and brain-model.
- For 8 brain regions, brain-SLINGER model trained on WB **performed as well** as the PrediXcan **trained on GTEx brain** data (paired Student test $P > 0.05$).
- For **Cortex and Hippocampus**, we found a significantly **higher correlation for brain-SLINGER**.
- It is better to combine **tissue-enriched loci and allow distal interactions** than train on actual tissue expression with local interactions only and smaller sample size.



Application to WTCCC bipolar disorder cohort

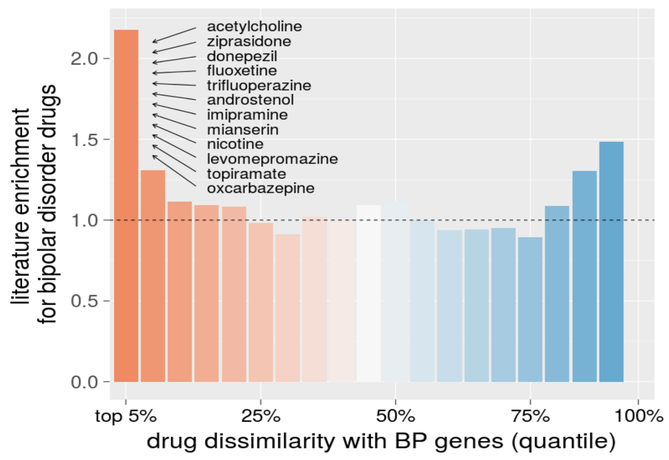
- WTCCC bipolar disorder (BP) cohort: **genotype data** on 1,868 cases and 2,938 controls.
- Expression for **9,747 genes** were inferred and tested for difference between cases and controls.
- **103 genes** were found to be significantly **different** in individuals with BP (after correction).
- Among them, 56 genes are up regulated (e.g., *DLG3*) and 47 down regulated (e.g., *BCL11B*).



KEGG pathway enrichment (genes + drugs in STITCH)

pathway name	count in gene set	false discovery rate
Neuroactive ligand-receptor intraction	5	0.0308
Cocaine addiction	3	0.0308
Amphetamine addiction	3	0.0308
Calcium signaling pathway	4	0.0316

- 2,837 **CMAP drug** profiles matched up/down regulated genes, and **rank** them by dissimilarity.
- Drugs with high dissimilarity might rescue affected genes expression → potential **treatments**.
- **Literature enrichment** for known BP drugs, first 5% might contain other drug candidates.



Future directions

- Reduce the number of tested genes, based on training performances
- Application to human heart tissue (available TiSAn score) and cardiovascular diseases.
- Multitask learning [4] on GTEx samples to gain power in detecting tissue-specific expression.
- Tissue-specific transcriptome-wide association strategy (lower multiple testing correction).

References

[1] Vervier K. & Michaelson J., *biorXiv*, 10.1101/141408 , 2017
[2] Vervier K. & Michaelson J., *Sci Rep*, 2016
[3] Liu et al., *AJHG*, 2017
[4] Vervier K. *et al.*, *Machine Learning and Knowledge Discovery in Databases*, 2014