Interpretable deep learning model, embedding the gene regulatory network, for understanding functional genomics in neuropsychiatric disorders

M Gerstein
Yale
(See last slide for more info.)

Sample Sources: >2,500 brains

Genome:
WGS, genotype

Cross-disorder: ASD, SCZ, BP, Neurodevelopmental, Neurotypical

Bulk Tissue

Multi-omic analysis
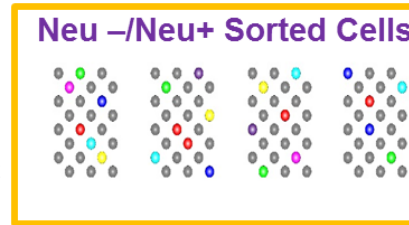
Epigenome:
ChIP-seq, ATAC-seq, HiC, ERRBS, Array Methylation, NOMeSeq

Dorsolateral Prefrontal Cortex

Anterior Cingulate Cortex

Orbitofrontal Cortex
Temporal Cortex

Neu –/Neu+ Sorted Cells

Transcriptome:
RNA-seq, lncRNAseq,

Amygdala
Subgenual Anterior Cingulate Cortex

Striatum

Cerebellum

Single Cell

Proteome:
MWP, LC-MS/MS
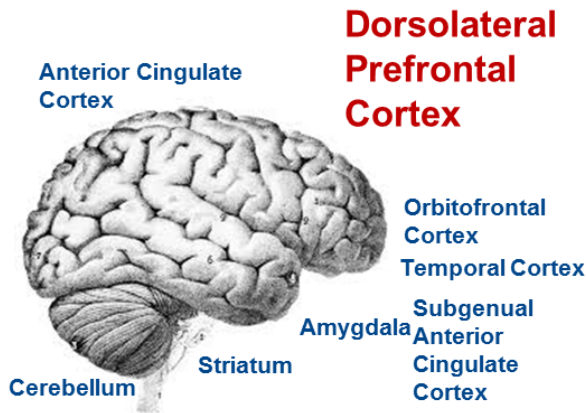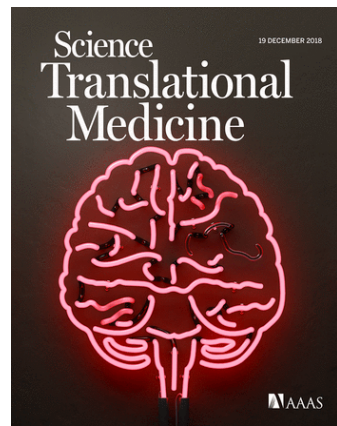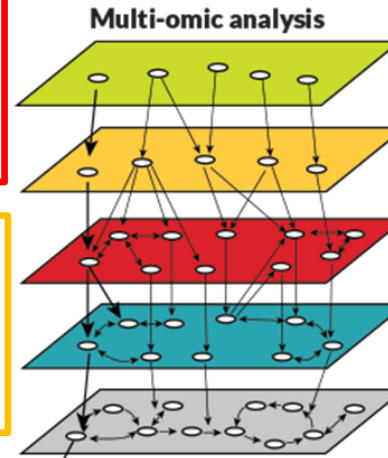
# PsychENCODE

'18 rollout in Science

11 papers in total.
Major material in the 3 capstones:

Wang et al. ('18), Li et al. ('18), Gandal et al. ('18)

A core issue addressed by PsychENCODE:
Using functional genomics to reveal molecular mechanisms between genotype and phenotype in brain disorders

Disease	Heritability*	Molecular **Mechanisms**
Schizophrenia	**81%**	(C4A)
Bipolar disorder	70%	-
Alzheimer's disease	58 - 79%	Apolipoprotein E (APOE), Tau
Hypertension	30%	Renin–angiotensin–aldosterone
Heart disease	34-53%	Atherosclerosis, VCAM-1
Stroke	32%	Reactive oxygen species (ROS), Ischemia
Type-2 diabetes	26%	Insulin resistance
Breast Cancer	25-56%	BRCA, PTEN

Phenotype

pathways, circuits

Cell types · Modules

Regulatory elements · Genes

Genotype
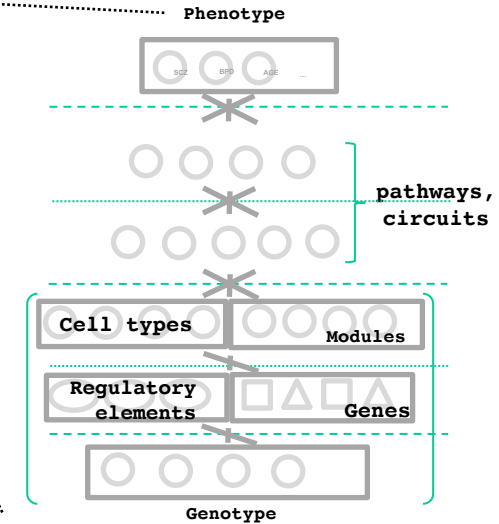
Many psychiatric conditions are highly heritable
  Schizophrenia: up to 80%
But we don't understand basic molecular mechanisms underpinning this association
  (in contrast to many other diseases such as cancer & heart disease)
Thus, interested in developing predictive models of psychiatric traits which:
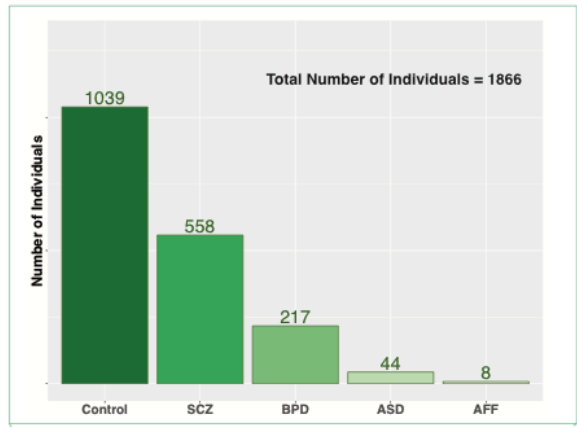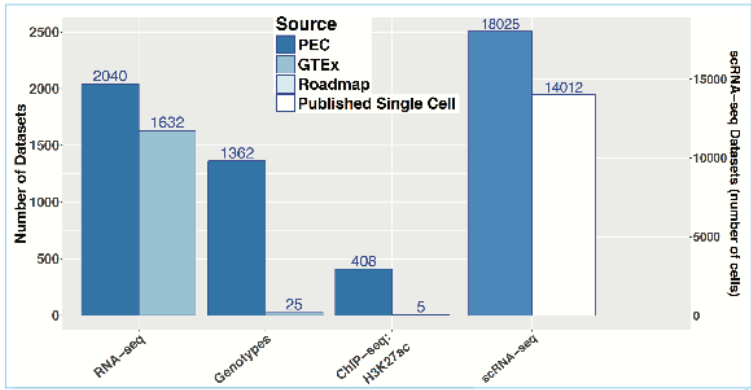 Use observations at intermediate (molecular levels) levels to inform latent structure
  Use the predictive features of these "molecular endo phenotypes" to begin to suggest actors involved in mechanism

Interpretable deep learning model, embedding the gene regulatory network, for understanding functional genomics in neuropsychiatric disorders

- Construction of an adult brain resource with 1866 individuals
- Using the changing proportions of cell types
  (via **single-cell deconvolution**)
  to account for expression variation across a population
- Large-scale processing defines ~79K PFC
  enhancers & creates a comprehensive QTL resource
  (~2.5M eQTLs + cQTLs & fQTLs)
- Connecting QTLs, enhancer activity relationships & Hi-C contacts into a
  brain regulatory network
  & using this to link SCZ GWAS SNPs to genes
- Embedding the reg. network in a
  deep-learning model to predict psychiatric disease from genotype &
  transcriptome. Using this to suggest specific pathways & genes, as
  potential drug targets.
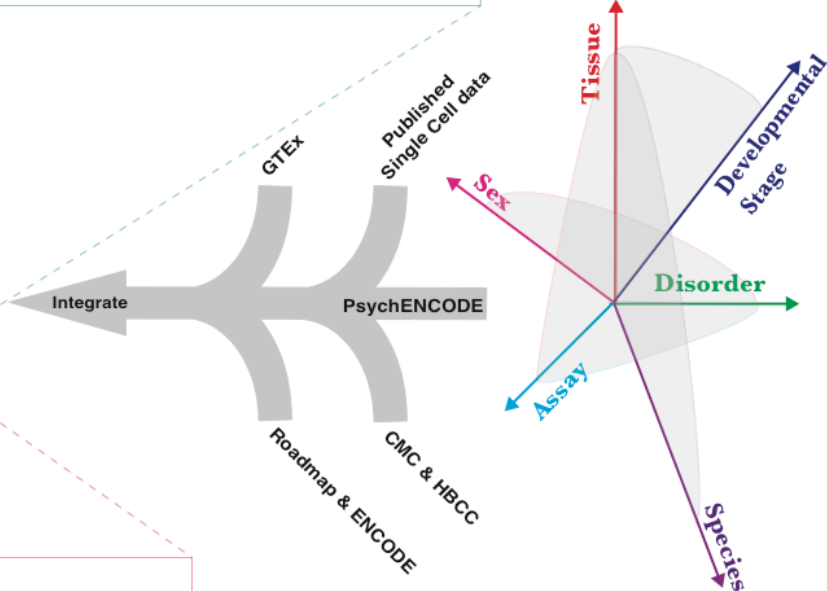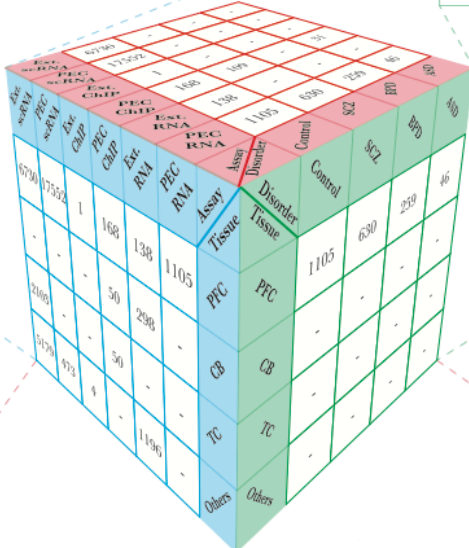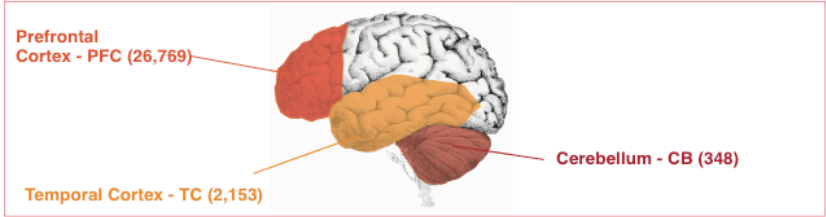- Other resource uses: highlighting aging related genes

Interpretable deep learning model, embedding the gene regulatory network, for understanding functional genomics in neuropsychiatric disorders

- Construction of an adult brain resource with 1866 individuals
- Using the changing proportions of cell types
  (via **single-cell deconvolution**)
  to account for expression variation across a population
- Large-scale processing defines ~79K PFC
  enhancers & creates a comprehensive QTL resource
  (~2.5M eQTLs + cQTLs & fQTLs)
- Connecting QTLs, enhancer activity relationships & Hi-C contacts into a
  brain regulatory network
  & using this to link SCZ GWAS SNPs to genes
- Embedding the reg. network in a
  deep-learning model to predict psychiatric disease from genotype &
  transcriptome. Using this to suggest specific pathways & genes, as
  potential drug targets.
- Other resource uses: highlighting aging related genes

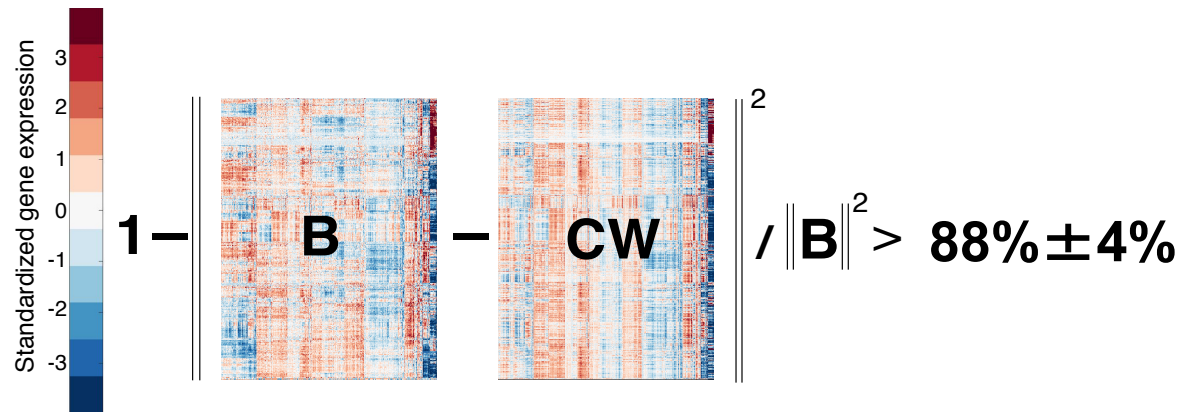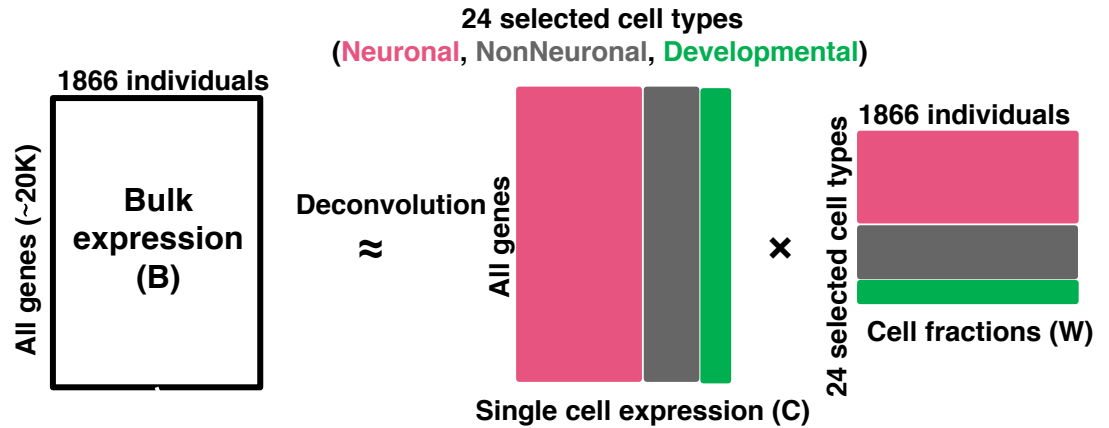**Collecting functional genomic datasets for the adult brain**

from PsychENCODE, other large consortia & single cell studies

1866
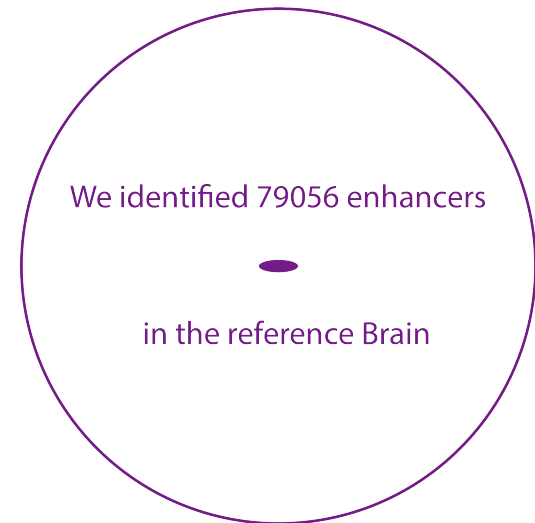Individuals
~3.7K bulk RNA-seq
~32K single-cells

[Wang et al. ('18) Science]

Single-cell deconvolution :

Supervised learning to estimate cell fractions

1866 individuals

24 selected cell types
(**Neuronal**, NonNeuronal, **Developmental**)

All genes (~20K)

Bulk expression (B)

Deconvolution

$\approx$

All genes

$\times$

1866 individuals

24 selected cell types

Cell fractions (W)

Single cell expression (C)

Standardized gene expression

$$1 - \left\| B - CW \right\|^2 / \left\| B \right\|^2 > 88\% \pm 4\%$$

Individual and cross-population reconstruction accuracy via deconvolution

Different neuronal & glial cell fractions across disorders

Ex5   In6   Oligo
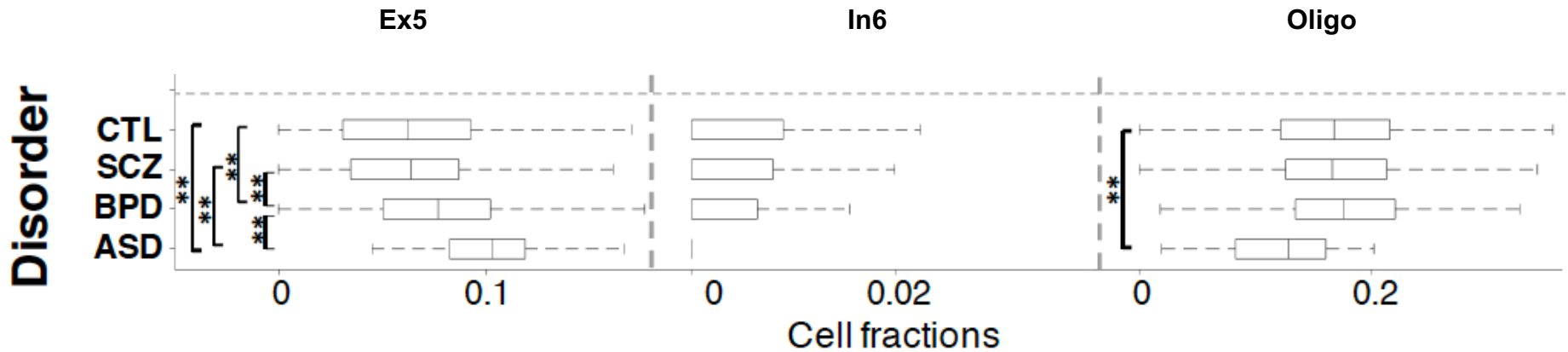
Excitatory to Inhibitory imbalance at
neuronal subtype level for ASD*

* Rubenstein et al., Model of autism: increased ratio of excitation/inhibition in key neural systems, Genes Brain Behav. 2003

[Wang et al. ('18) Science]

Interpretable deep learning model, embedding the gene regulatory network, for understanding functional genomics in neuropsychiatric disorders

- Construction of an adult brain resource with 1866 individuals
- Using the changing proportions of cell types
  (via **single-cell deconvolution**)
  to account for expression variation across a population
- Large-scale processing defines ~79K PFC
  enhancers & creates a comprehensive QTL resource
  (~2.5M eQTLs + cQTLs & fQTLs)
- Connecting QTLs, enhancer activity relationships & Hi-C contacts into a
  brain regulatory network
  & using this to link SCZ GWAS SNPs to genes
- Embedding the reg. network in a
  deep-learning model to predict psychiatric disease from genotype &
  transcriptome. Using this to suggest specific pathways & genes, as
  potential drug targets.
- Other resource uses: highlighting aging related genes

Developing a Reference Set of ~79K PFC Enhancers & Studying Their Population Variation
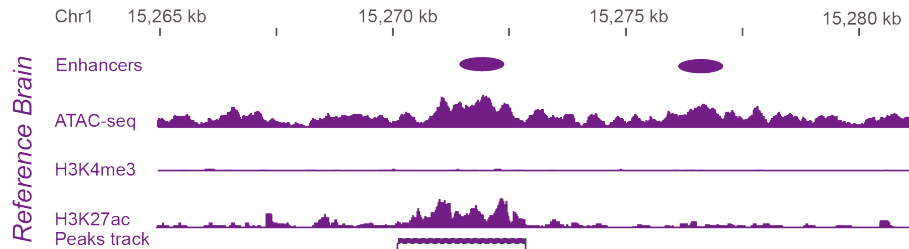

Consistent with ENCODE, active enhancers are identified as open chromatin regions enriched in H3K27ac and depleted in H3K4me3

We identified 79056 enhancers

in the reference Brain

[Wang et al. ('18) Science]

Developing a Reference Set of ~79K PFC Enhancers & Studying Their Population Variation

[Wang et al. ('18) Science]
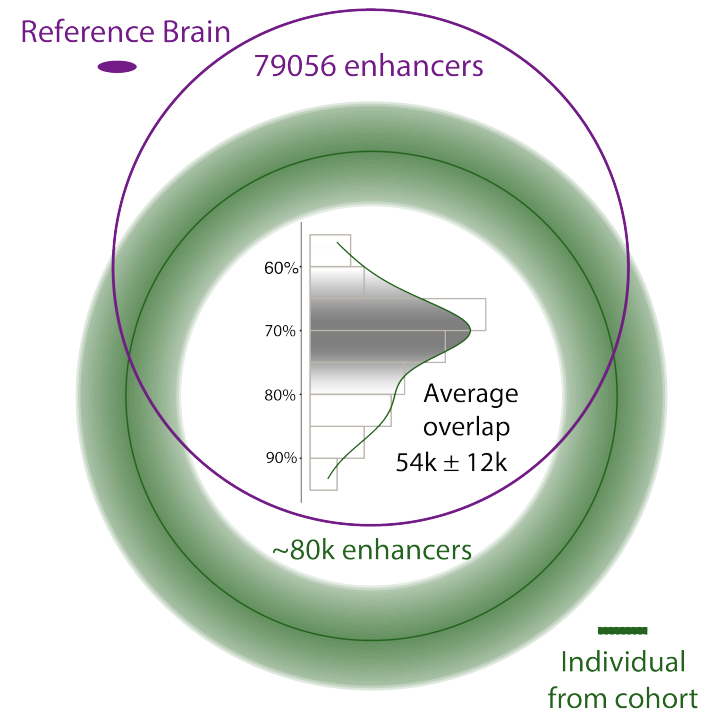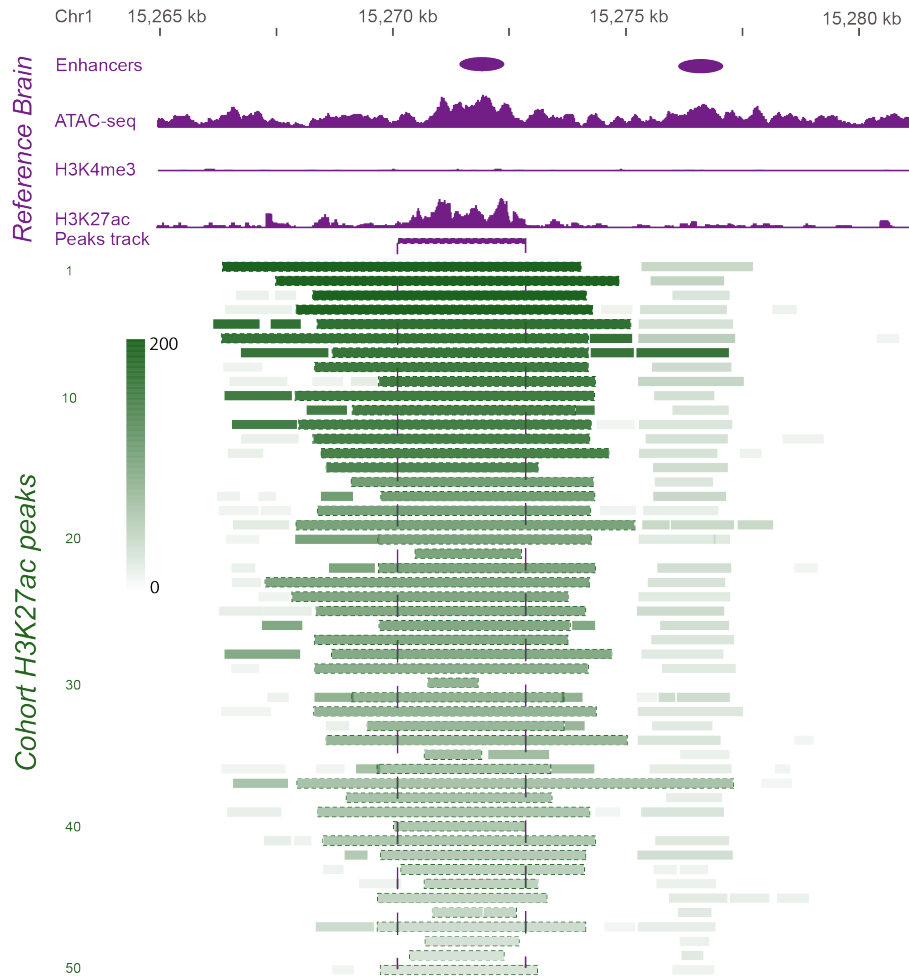
Larger brain eQTL sets than previous studies, but strong overlap with them
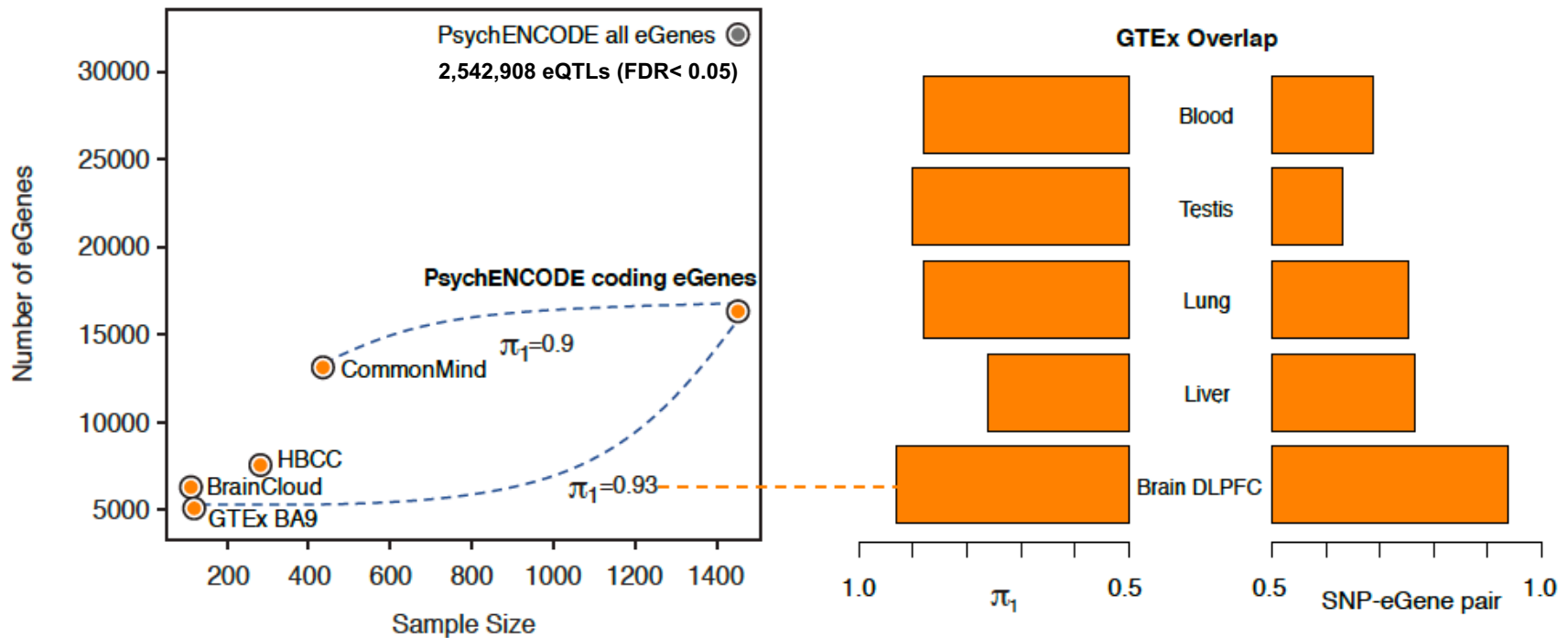
multi-QTLs from overlapping different types of QTLs: cQTL, fQTL, eQTL & isoQTL

	Numbers of QTLs	eGenes Enhancers Cell types	SNPs
eQTL	2,542,908	32,944	1,341,182
isoQTL	2,628,259	19,790	1,052,939
cQTL*	8,464	8,484	7,983
fQTL	4,199	9	1,672

eQTLs for mTOR mediated by cQTLs

eQTLs and cQTLs significantly overlap

1391 SNPs (multi-QTLs) in at least three types among eQTLs, isoQTLs, cQTLs, fQTLs

[Wang et al. ('18) Science]

13

Interpretable deep learning model, embedding the gene regulatory network, for understanding functional genomics in neuropsychiatric disorders

- Construction of an adult brain resource with 1866 individuals
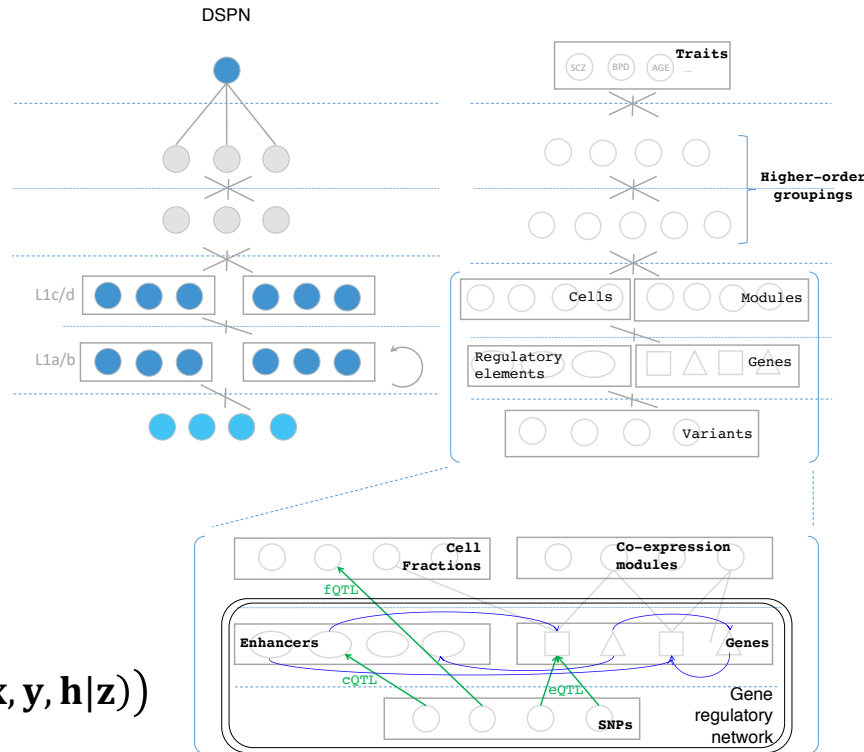- Using the changing proportions of cell types
(via **single-cell deconvolution**)
to account for expression variation across a population
- Large-scale processing defines ~79K PFC
enhancers & creates a comprehensive QTL resource
(~2.5M eQTLs + cQTLs & fQTLs)
- Connecting QTLs, enhancer activity relationships & Hi-C contacts into a
brain regulatory network
& using this to link SCZ GWAS SNPs to genes
- Embedding the reg. network in a
deep-learning model to predict psychiatric disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as potential drug targets.
- Other resource uses: highlighting aging related genes

Hi-C

1.5 mb   2 mb

Gene

Enhancers

Topologically Associating Domain (TAD)

Potential Enhancer-Promoter (E-P) interaction in TAD

TF

Enhancer

Target gene

TFBS on promoter

TFBS on enhancer

Transcription Factor Binding Sites (**TFBSs**)

Expression activity relationship

$C^*= argmin_C (\|Y - XC\|^2 + a\|C\|^2 + b\|C\|_{L1})$

TF expression (X) to predict target gene expression (Y) using Elastic net regression

C^*_i
C^*_j
C^*_k

QTLs

eQTL

chr7
rs7797344

C/C ref
T/T alt

FZD9

Gene regulatory network inference from Hi-C, QTLs & Activity Correlations

[Wang et al. ('18) Science]

Imputed gene regulatory network for the human brain
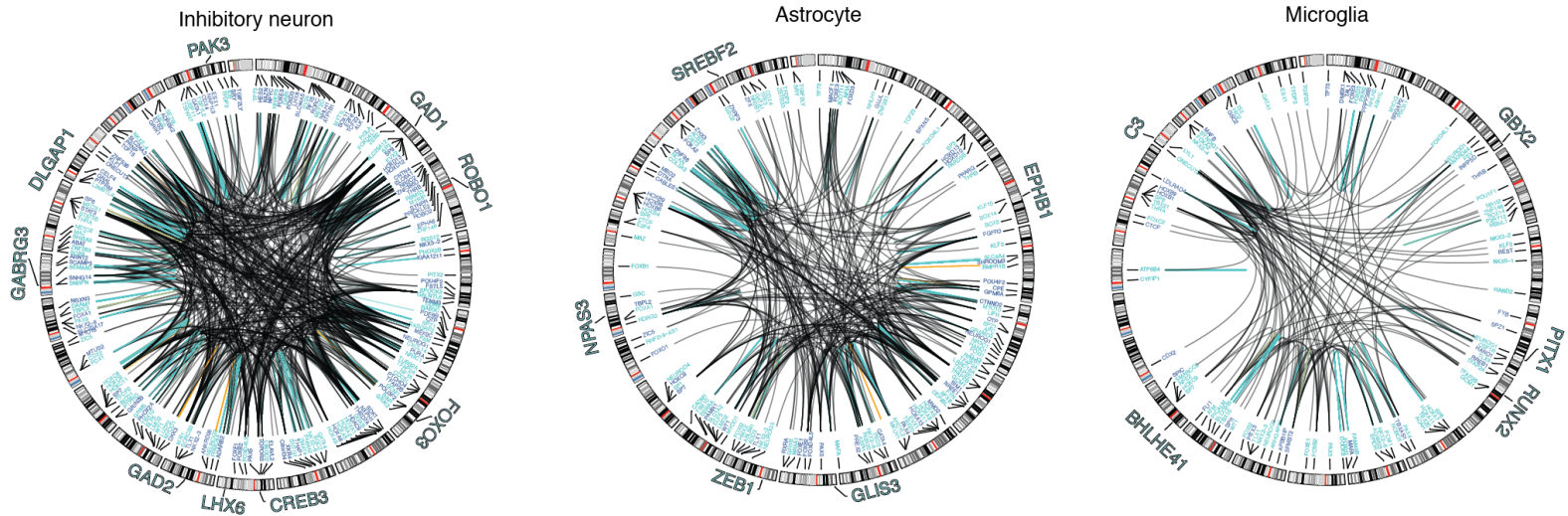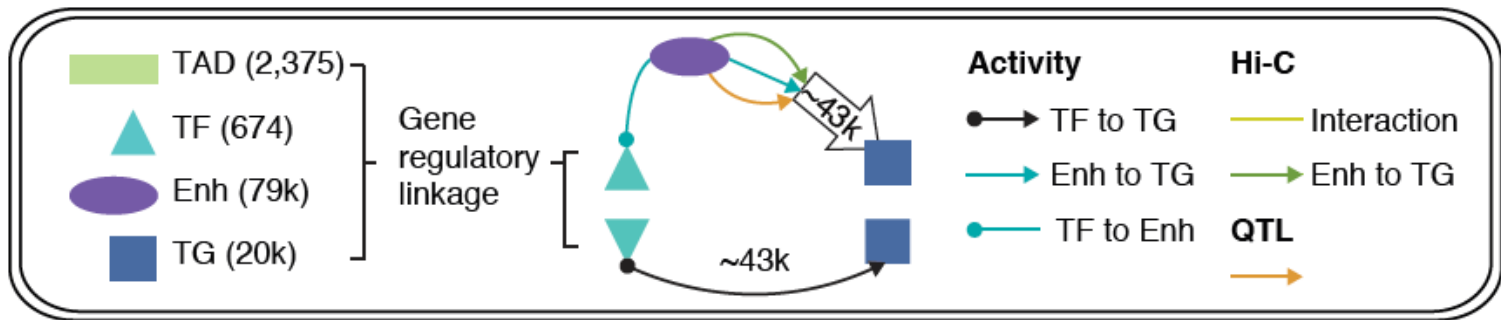
subnetworks targeting single cell marker genes

GWAS variants for SCZ genes
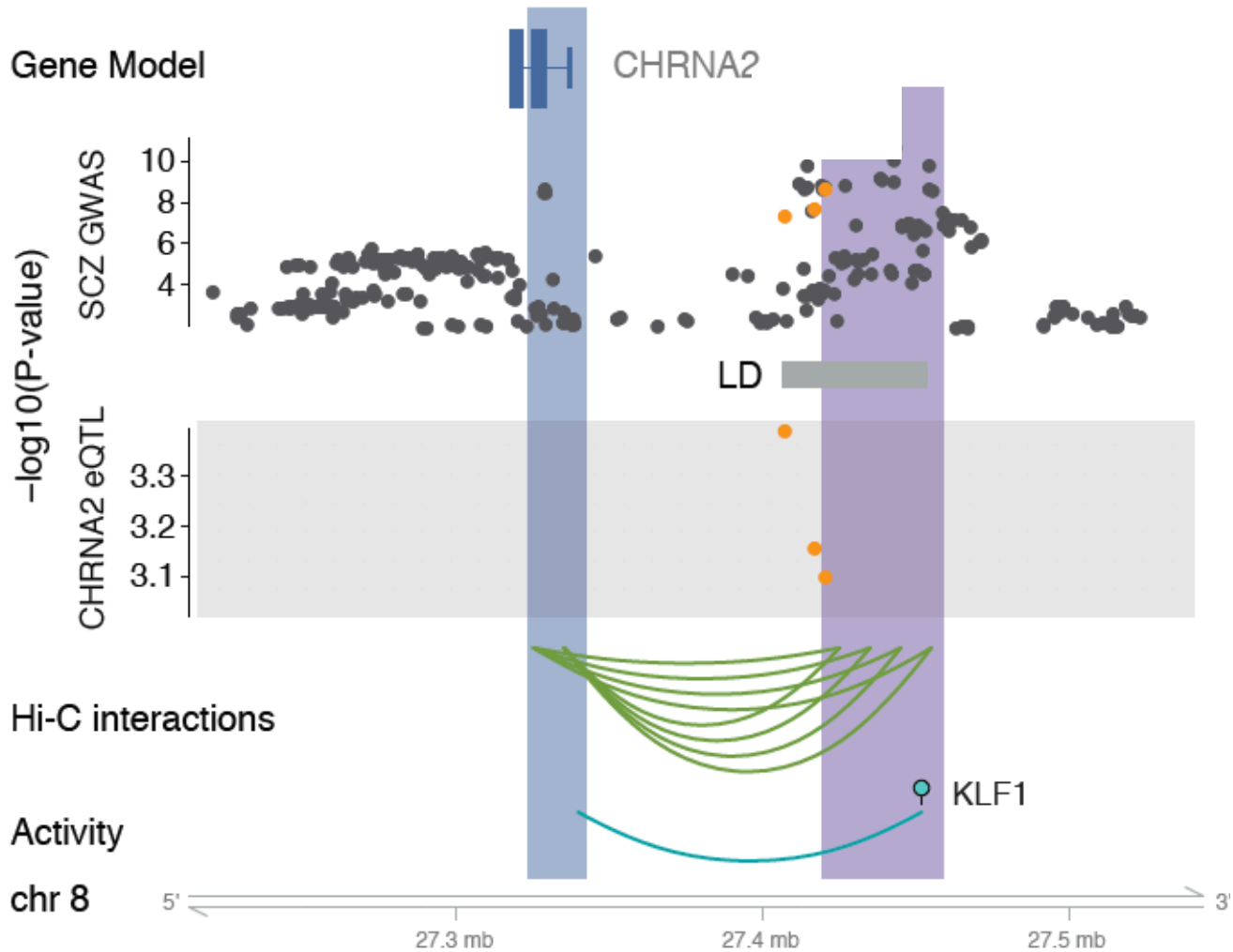
Linking GWAS SNPs to disease genes using the regulatory network

142
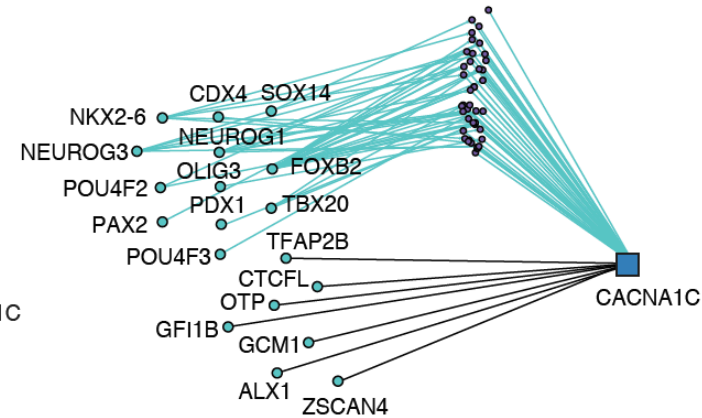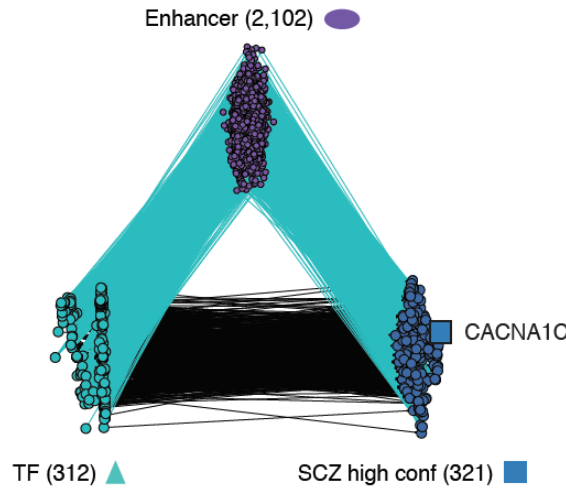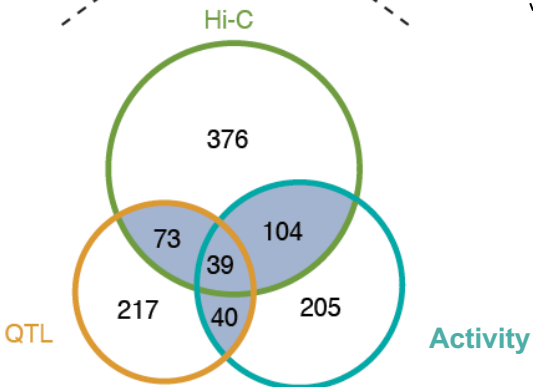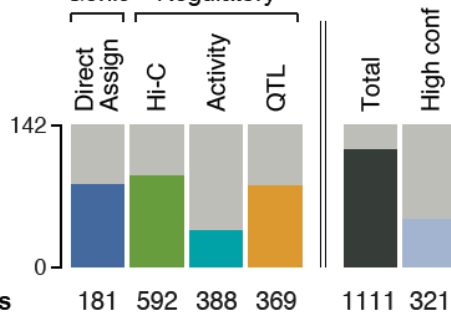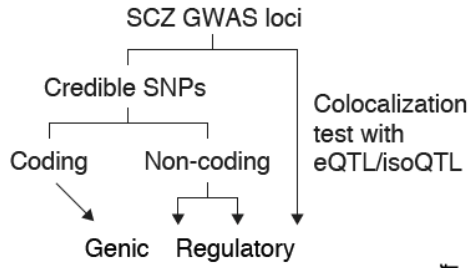
321
**high-confident
SCZ genes**

[Wang et al. ('18) Science]

Interpretable deep learning model, embedding the gene regulatory network, for understanding functional genomics in neuropsychiatric disorders

- Construction of an adult brain resource with 1866 individuals
- Using the changing proportions of cell types
  (via **single-cell deconvolution**)
  to account for expression variation across a population
- Large-scale processing defines ~79K PFC
  enhancers & creates a comprehensive QTL resource
  (~2.5M eQTLs + cQTLs & fQTLs)
- Connecting QTLs, enhancer activity relationships & Hi-C contacts into a
  brain regulatory network
  & using this to link SCZ GWAS SNPs to genes
- Embedding the reg. network in a
  deep-learning model to predict psychiatric disease from genotype &
  transcriptome. Using this to suggest specific pathways & genes, as
  potential drug targets.
- Other resource uses: highlighting aging related genes

Deep Structured Phenotype Network (DSPN)

Gene regulatory network builds skeleton

Boltzmann machine

y: phenotypes

h: hidden units (e.g., circuits)

x: intermediate phenotypes (e.g., genes, enhancers)

z: genotypes (e.g., SNPs)

W: weights (e.g., regulatory network)

Energy model:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z}) \propto \exp\big(-E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z})\big)$$

$$E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z}) = -\mathbf{z}^{\mathrm{T}}\mathbf{W_1}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W_2}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W_3}\mathbf{h} - \mathbf{h}^{\mathrm{T}}\mathbf{W_4}\mathbf{h} - \mathbf{h}^{\mathrm{T}}\mathbf{W_5}\mathbf{y} - Bias$$
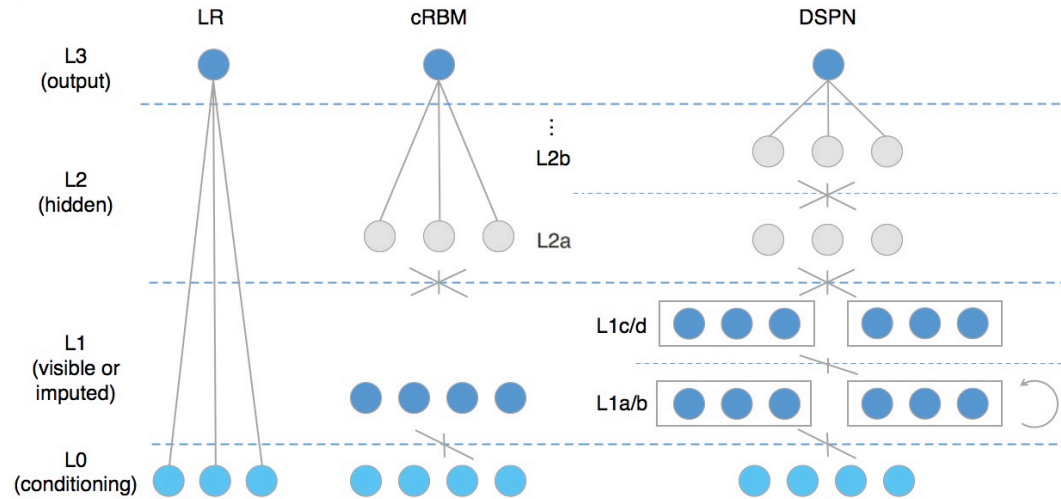
[Wang et al. ('18) Science]

DSPN improves brain disease prediction by adding deep layers

Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	**54.6%**	63.0%	70.0%	59.0%	**73.6%**
Bipolar Disorder	**56.7%**	63.3%	71.1%	67.2%	**76.7%**
Autism Spectrum Disorder	**50.0%**	51.7%	67.2%	62.5%	**68.3%**

X 6.0

Accuracy = chance to correctly predict disease/health

DSPN improves brain disease prediction by adding deep layers
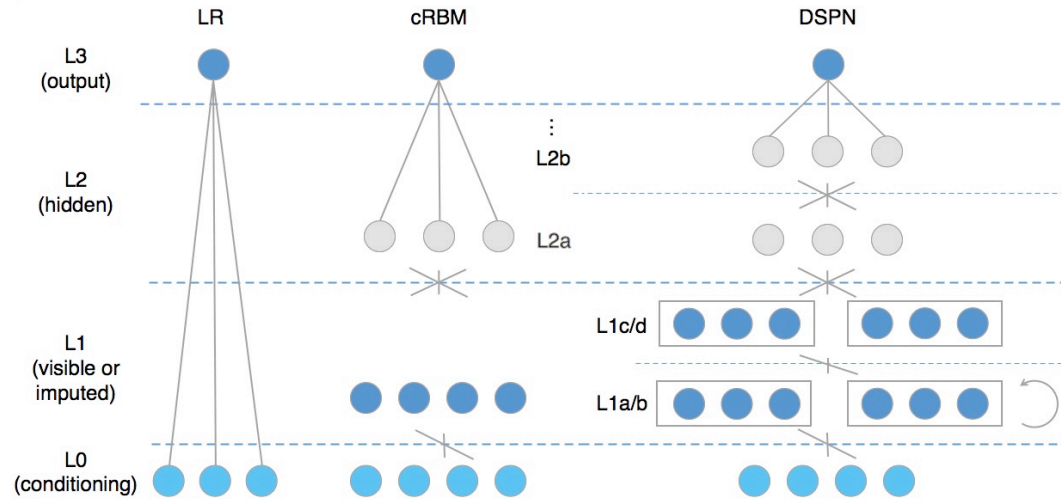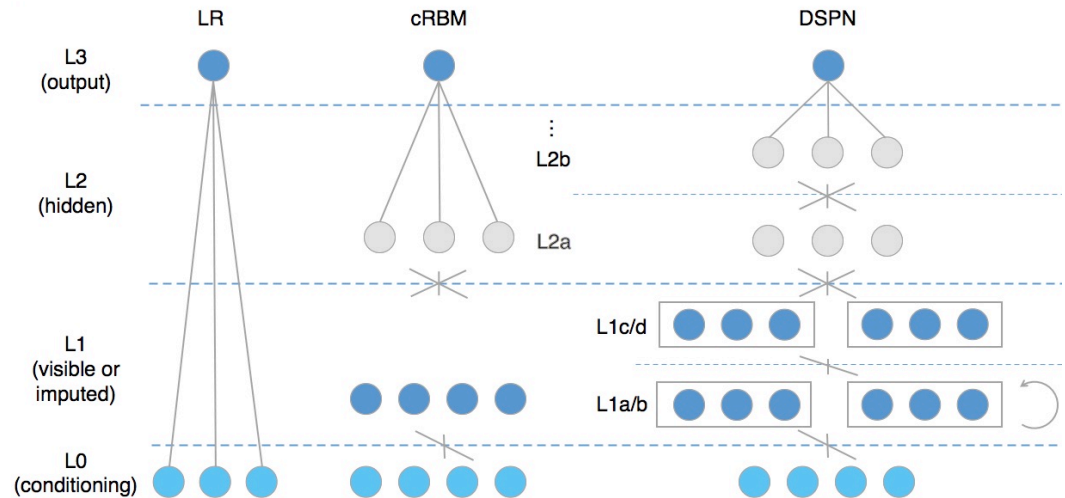
Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	54.6%	**63.0%**	70.0%	59.0%	**73.6%**
Bipolar Disorder	56.7%	**63.3%**	71.1%	67.2%	**76.7%**
Autism Spectrum Disorder	50.0%	**51.7%**	67.2%	62.5%	**68.3%**

X 2.5

Accuracy = chance to correctly predict disease/health

[Wang et al. ('18) Science]

DSPN improves brain disease prediction by adding deep layers

Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	**54.6%**	63.0%	70.0%	**59.0%**	73.6%
Bipolar Disorder	**56.7%**	63.3%	71.1%	**67.2%**	76.7%
Autism Spectrum Disorder	**50.0%**	51.7%	67.2%	**62.5%**	68.3%

X 3.1

Accuracy = chance to correctly predict disease/health

Interpretable deep learning model, embedding the gene regulatory network, for understanding functional genomics in neuropsychiatric disorders

- Construction of an adult brain resource with 1866 individuals
- Using the changing proportions of cell types
  (via **single-cell deconvolution**)
  to account for expression variation across a population
- Large-scale processing defines ~79K PFC
  enhancers & creates a comprehensive QTL resource
  (~2.5M eQTLs + cQTLs & fQTLs)
- Connecting QTLs, enhancer activity relationships & Hi-C contacts into a
  brain regulatory network
  & using this to link SCZ GWAS SNPs to genes
- Embedding the reg. network in a
  deep-learning model to predict psychiatric disease from genotype &
  transcriptome. Using this to suggest specific pathways & genes, as
  potential drug targets.
- Other resource uses: highlighting aging related genes

Multilevel Network Interpretation

L3

L2b

L2a

L1d

**Actual network size:
5024/400/100/1 nodes**
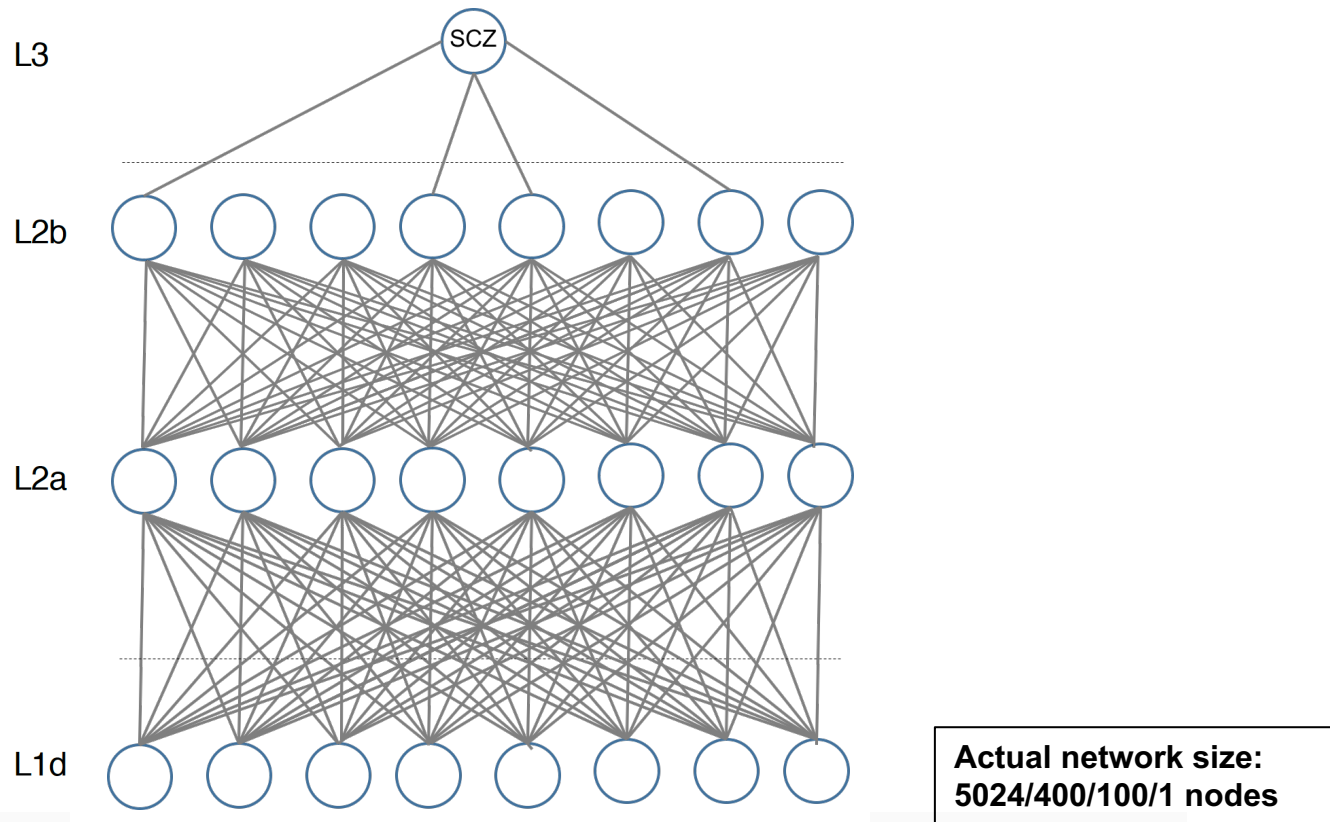
- Start with a fully connected trained network

Multilevel Network Interpretation
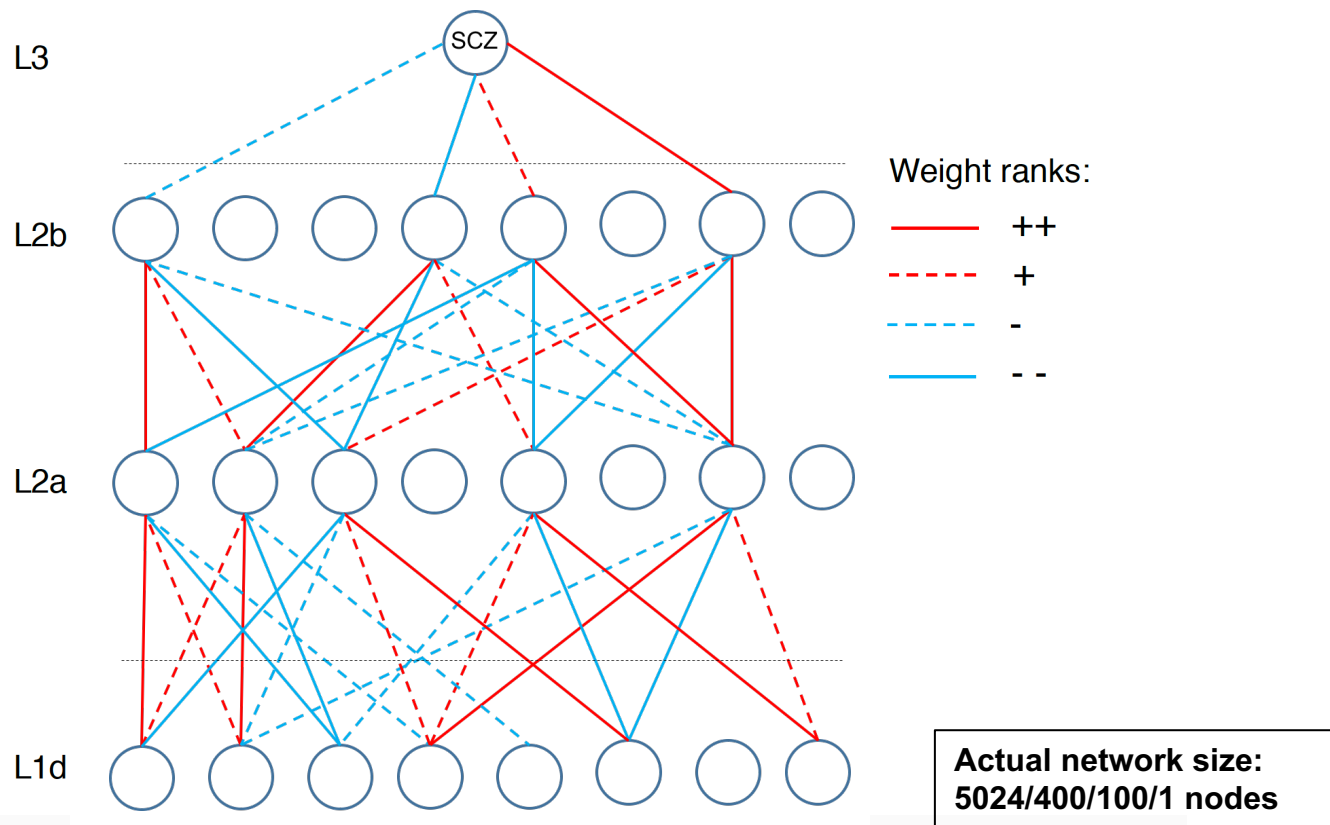
- Start with a fully connected trained network
- Sparsify network using edges with largest absolute weights (+/-)

Multilevel Network Interpretation
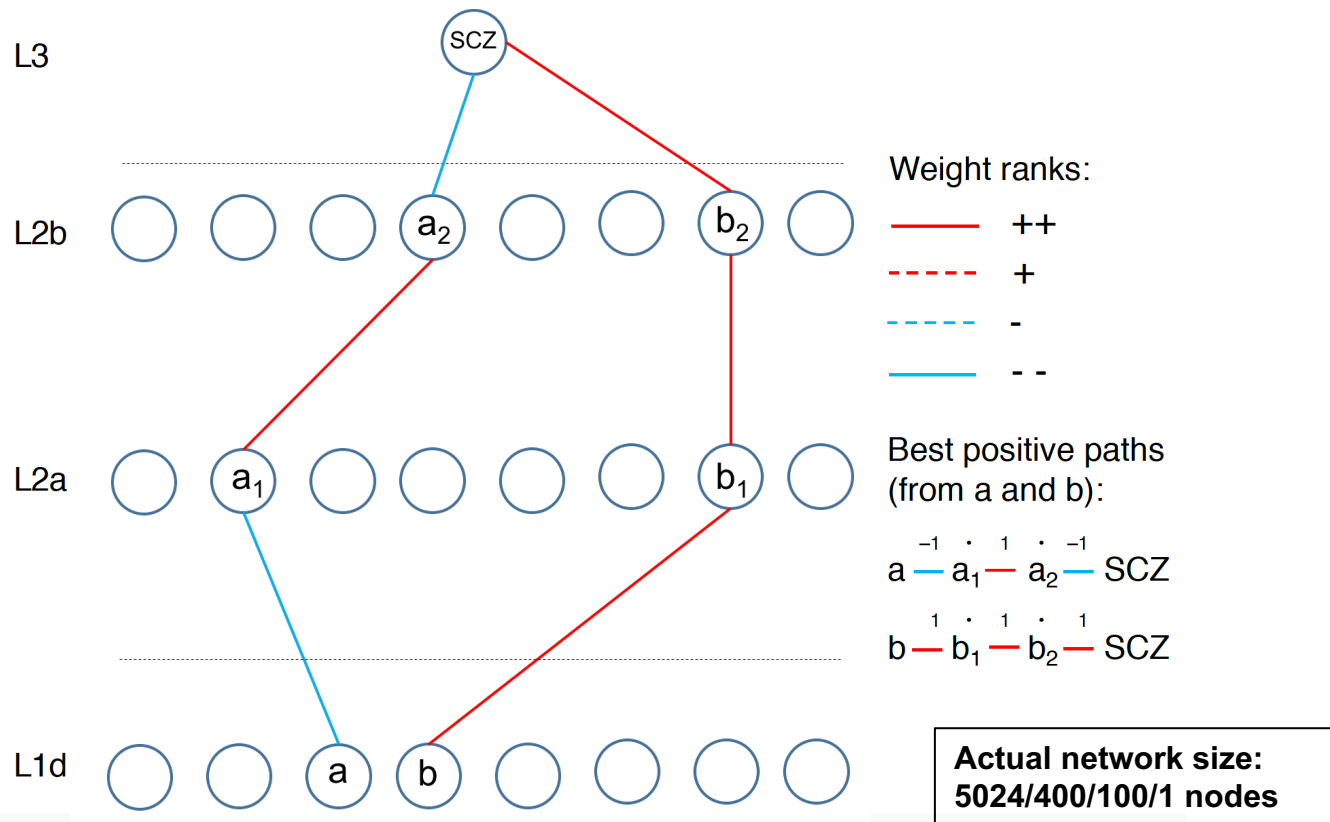
- Start with a fully connected trained network
- Sparsify network using edges with largest absolute weights (+/-)
- Extract 'best positive paths' to each prioritized module (e.g. a-a_1-a_2-SCZ) by summing weights and multiplying signs

DSPN discovers enriched pathways and linkages to genetic variation

Cross-disorder MOD/HOG enrichment ranking

Ranking score

50 100 >200

Functional categories
(*) RNA proc.   (~) Immune
(>) Synaptic   (#) Metabolic
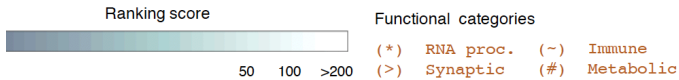
SCZ BPD ASD
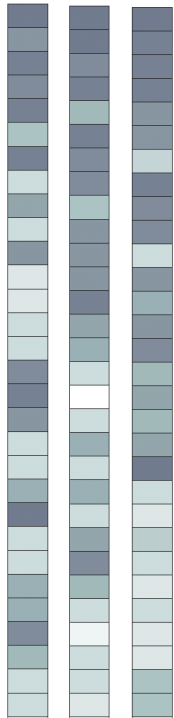
(*) Spliceosome / RNA splicing
(>) Synaptic vesicle cycle
(~) Antigen proc. and presentation
    Vesicle localization
    Proteasome
(*) mRNA processing
    Chromatin modification
(#) Oxidative phosphorylation
    Retrograde endocannabinoid sig.
(>) Chemical synaptic transmission
    Peptidyl-lysine modification
    Endocytosis
    Ubiquitin mediated proteolysis
(>) Anterograde trans-synaptic sig.
(*) mRNA transport
    Phosphatidylinositol signaling
    Hippo signaling pathway
(~) Staph./ Epstein-Barr virus inf.
(>) Synaptic signaling
    Autophagy
(>) Dop./GABA/Glutamatergic synapse
(>) Calcium signaling
(>) Endocrine calcium reabsorption
(*) RNA degradation / transport
(#) Ribosome
    Neuron projection morphogenesis
(~) Fc receptor signaling pathway
    cGMP-PKG signaling pathway
(~) mTOR signaling pathway
(~) Cytokine-cytokine receptor int.

[Wang et al. ('18) Science]
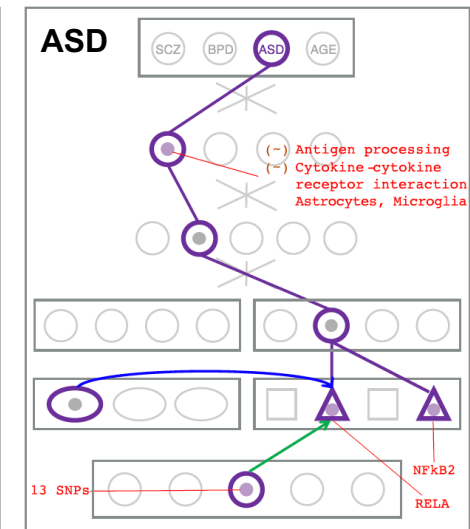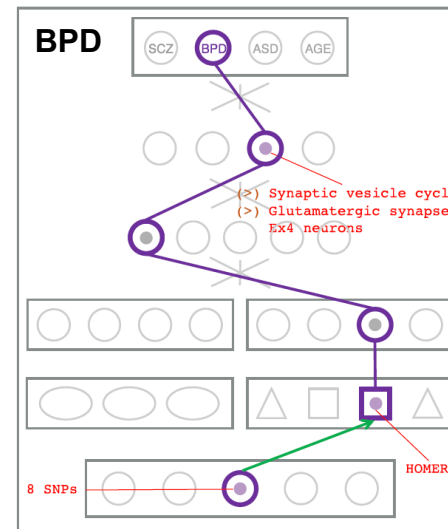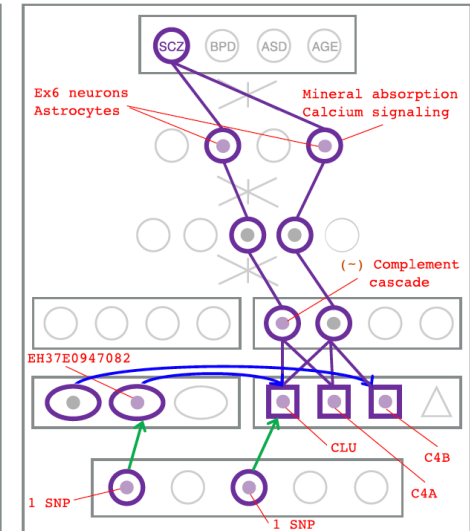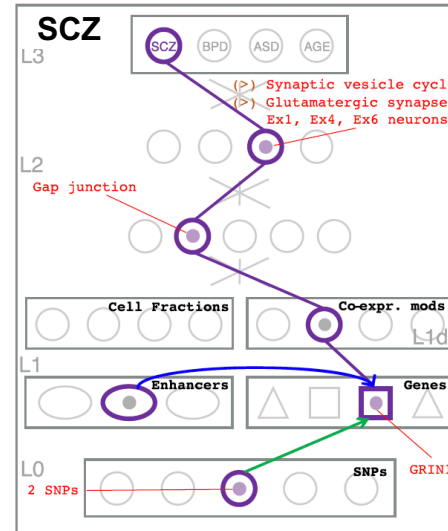
NRGN has variable expression over age and is in Synaptic vesicle cycle pathway is enriched in SCZ, BPD, ASD

NGRN is a gene associated with the **Synaptic vesicle pathway** and NGRN **expression** and **methylation** is correlated with **Age**

② SCZ BPD ASD AGE

Longevity regulating pathway
(>) Synaptic vesicle cycle

EH37E0245781

NRGN

1 SNP

NRGN

Ex3 Neuron

**

Methylation

Exp.

NRGN

Age (years)

Interpretable deep learning model, embedding the gene regulatory network, for understanding functional genomics in neuropsychiatric disorders

- Construction of an adult brain resource with 1866 individuals
- Using the changing proportions of cell types
  (via **single-cell deconvolution**)
  to account for expression variation across a population
- Large-scale processing defines ~79K PFC
  enhancers & creates a comprehensive QTL resource
  (~2.5M eQTLs + cQTLs & fQTLs)
- Connecting QTLs, enhancer activity relationships & Hi-C contacts into a
  brain regulatory network
  & using this to link SCZ GWAS SNPs to genes
- Embedding the reg. network in a
  deep-learning model to predict psychiatric disease from genotype &
  transcriptome. Using this to suggest specific pathways & genes, as
  potential drug targets.
- Other resource uses: highlighting aging related genes

- Construction of an adult brain resource with 1866 individuals
- Using the changing proportions of cell types
(via **single-cell deconvolution**)
to account for expression variation across a population
- Large-scale processing defines ~79K PFC
enhancers & creates a comprehensive QTL resource
(~2.5M eQTLs + cQTLs & fQTLs)
- Connecting QTLs, enhancer activity relationships & Hi-C contacts into a
brain regulatory network
& using this to link SCZ GWAS SNPs to genes
- Embedding the reg. network in a
deep-learning model to predict psychiatric disease from genotype &
transcriptome. Using this to suggest specific pathways & genes, as
potential drug targets.
- Other resource uses: highlighting aging related genes

Acknowledgments

Dedicated to
Pamela Sklar

NIMH:
Geetha Senthil
Lora Bingaman
David Panchision
Alexander Arguello
Thomas Lehner

Daifeng Wang, Shuang Liu, Jonathan Warrell, Hyejung Won, Xu Shi, Fabio Navarro, Declan Clarke, Mengting Gu, Prashant Emani, Yucheng T. Yang, Min Xu, Michael Gandal, Shaoke Lou, Jing Zhang, Jonathan J. Park, Chengfei Yan, Suhn Kyong Rhie, Kasidet Manakongtreecheep, Holly Zhou, Aparna Nathan, Mette Peters, Eugenio Mattei, Dominic Fitzgerald, Tonya Brunetti, Jill Moore, Yan Jiang, Kiran Girdhar, Gabriel Hoffman, Selim Kalayci, Zeynep Hulya Gumus, Greg Crawford, **PsychENCODE Consortium**, **Panos Roussos, Schahram Akbarian, Andrew E. Jaffe, Kevin White, Zhiping Weng, Nenad Sestan,**
Daniel H. Geschwind, James A. Knowles

Extra

Info about content in this slide pack

- General PERMISSIONS

  - This Presentation is copyright Mark Gerstein, Yale University, 2019.

  - Please read permissions statement at www.**gersteinlab.org/misc/permissions.html** .

  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).

  - Paper references in the talk were mostly from Papers.GersteinLab.org.

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .

  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt