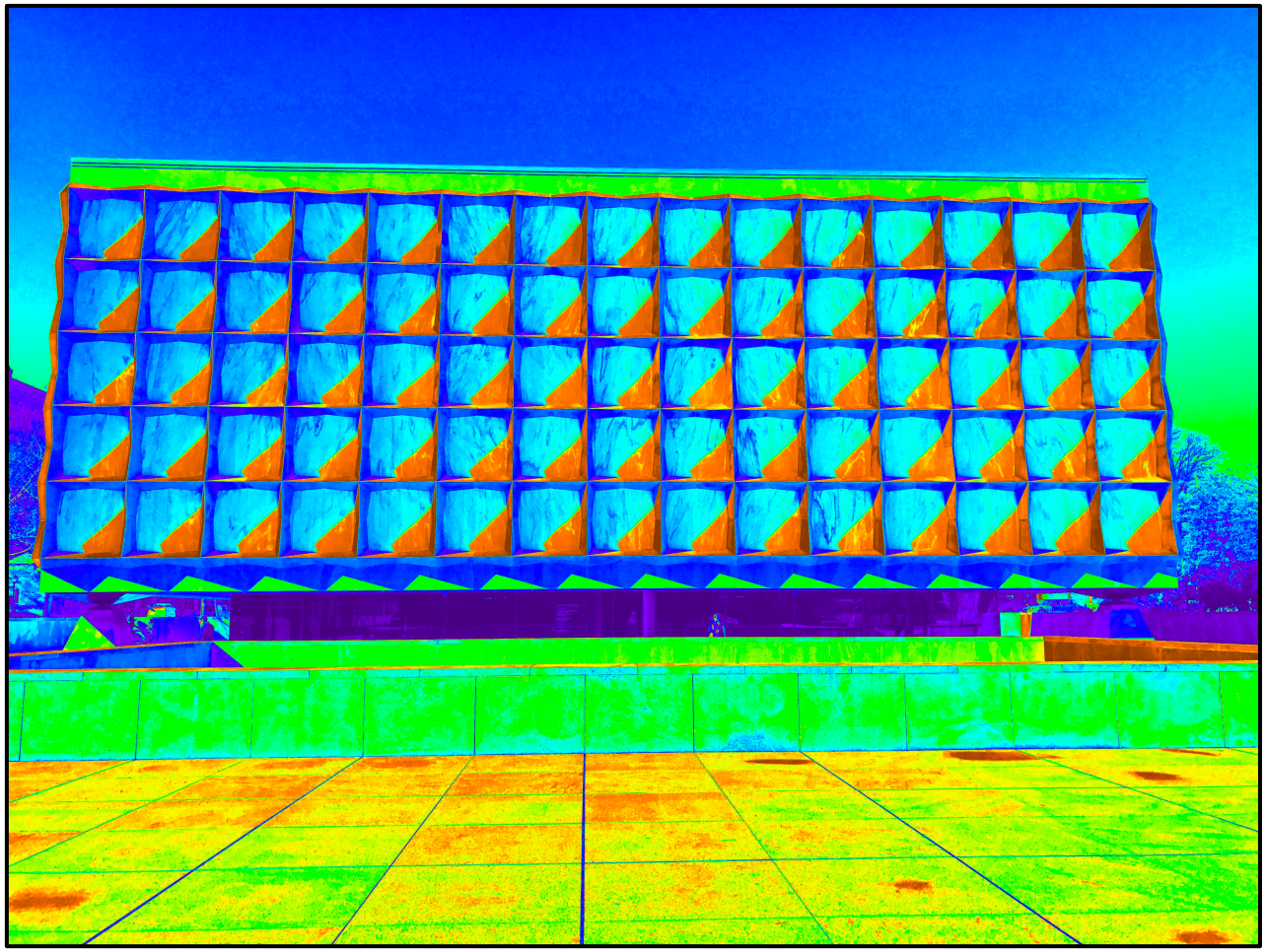
Personal Genomics & Data Science:

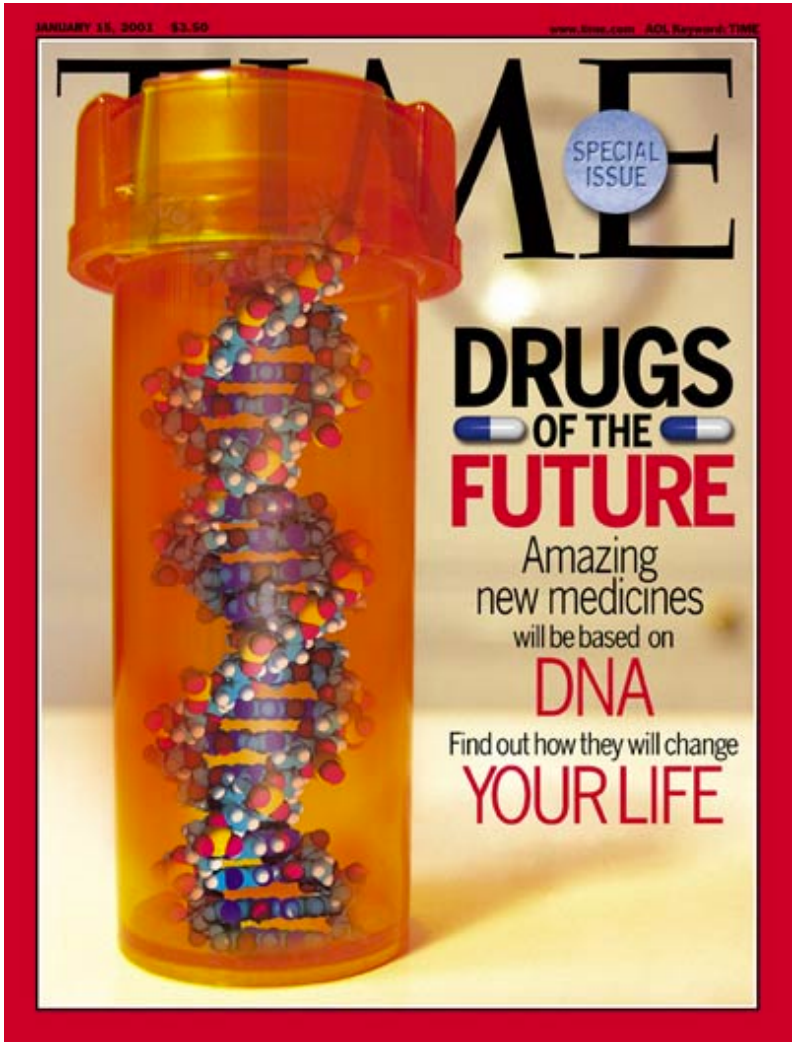
Using population-scale functional genomics to suggest potential neuropsychiatric drug targets & building a hybrid classifier to ascertain differential drug sensitivity



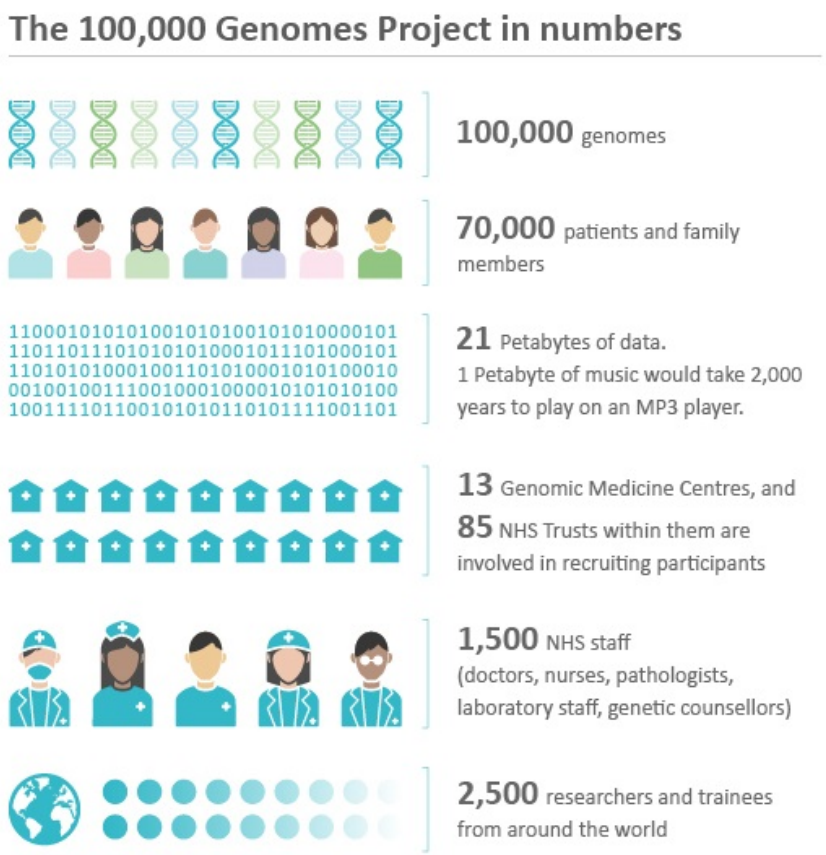
M Gerstein
Yale
(See last slide for more info.)

Slides freely downloadable
from **Lectures.GersteinLab.org**
& “tweetable” (via **@MarkGerstein**)

The Genomic Future



Many big projects.
Soon millions will be sequenced….



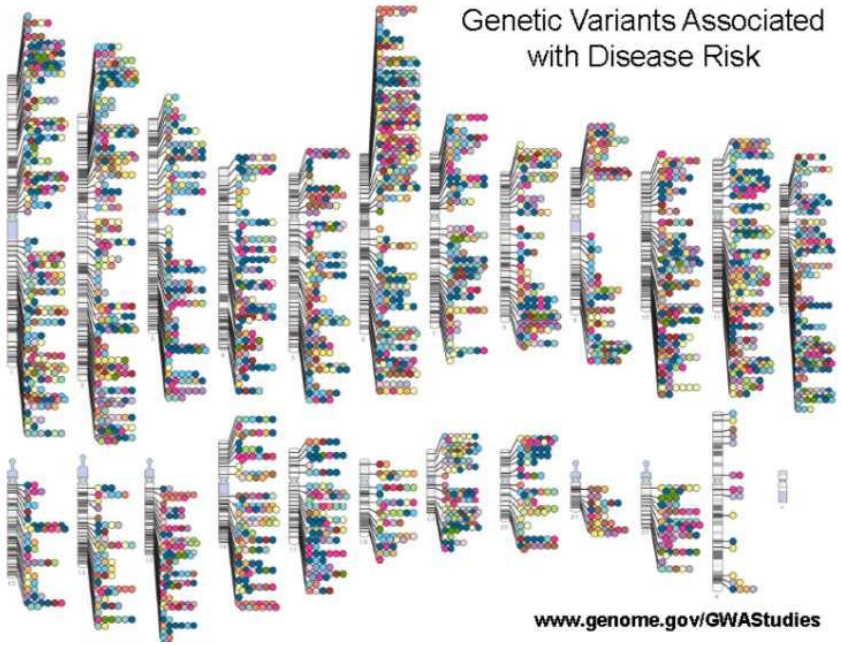
All of Us℠ The Precision Medicine Initiative®
THE FUTURE OF HEALTH BEGINS WITH YOU

NHLBI TOPMed

https://www.mongodb.com/press/genomics-england-uses-mongodb-to-power-the-data-science-behind-the-100000-genomes-project

What to do with these variants in relation to disease

- **Personalized risk prediction for many conditions**
- Precision oncology
- **Drug target identification via genetic associations**
- **Accounting for differential drug sensitivity**



Using population-scale functional genomics to suggest potential neuropsychiatric drug targets & building a hybrid classifier to ascertain differential drug sensitivity

- **PsychENCODE:** Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + full developmental time-course
  - Using the changing proportions of cell types (via **single-cell deconvolution**) to account for expression variation across a population, disorders & development
  - 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 + consistently comparing the brain to other organs
- **GenoDock:** Building a predictor for the sensitivity of drug binding to personal SNVs
  - Hybrid classifier connecting **physical modelling with statistical learning**
    - The modeling creates a pseudo gold-standard dataset, which is used to train the stat. classifier
  - **Classifier Results**
    - Independent validation on an expt. validation set
    - Gives higher disruption scores to cancer driver SNVs. Also, illustrates importance of different features (eg GERP).
    - Picks out certain drugs (eg imatinib) as being particularly sensitive to SNVs

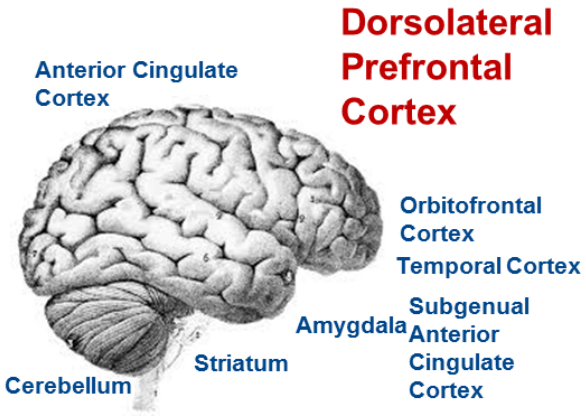
Using population-scale functional genomics to suggest potential neuropsychiatric drug targets & building a hybrid classifier to ascertain differential drug sensitivity

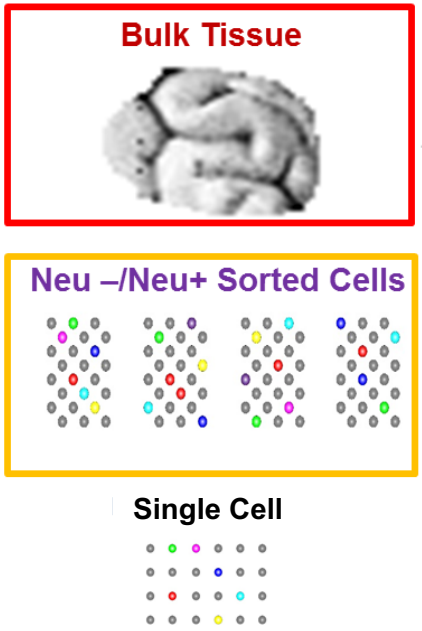
- **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + full developmental time-course
  - Using the changing proportions of cell types (via **single-cell deconvolution**) to account for expression variation across a population, disorders & development
  - 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 + consistently comparing the brain to other organs

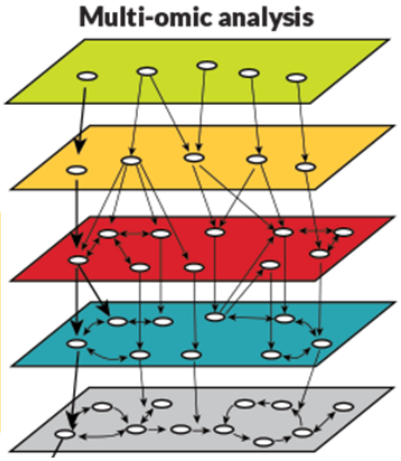
- **GenoDock**: Building a predictor for the sensitivity of drug binding to personal SNVs
  - Hybrid classifier connecting **physical modelling with statistical learning**
    - The modeling creates a pseudo gold-standard dataset, which is used to train the stat. classifier
  - **Classifier Results**
    - Independent validation on an expt. validation set
    - Gives higher disruption scores to cancer driver SNVs. Also, illustrates importance of different features (eg GERP).
    - Picks out certain drugs (eg imatinib) as being particularly sensitive to SNVs

Sample Sources: >2,500 brains

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





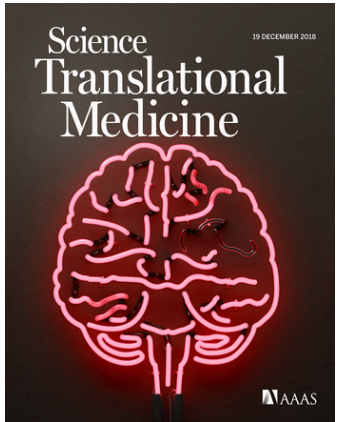


Genome:
WGS, genotype

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

Transcriptome:
RNA-seq, lncRNAseq,

Proteome:
MWP, LC-MS/MS



PsychENCODE

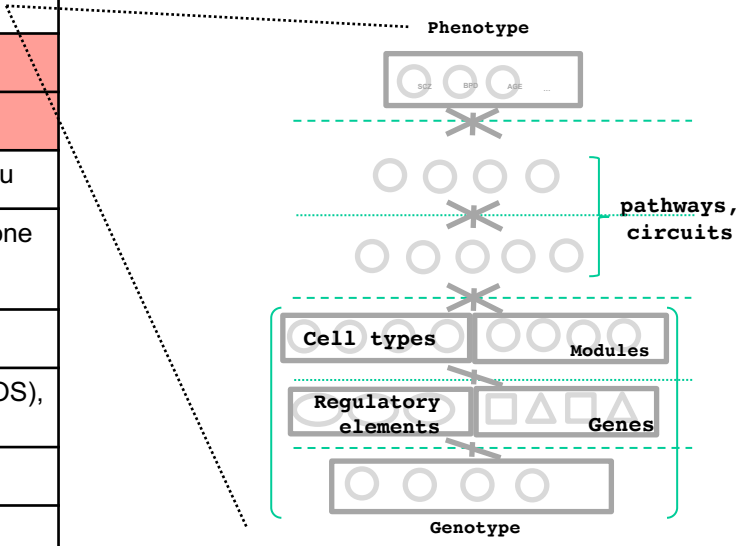
’18 rollout in Science

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



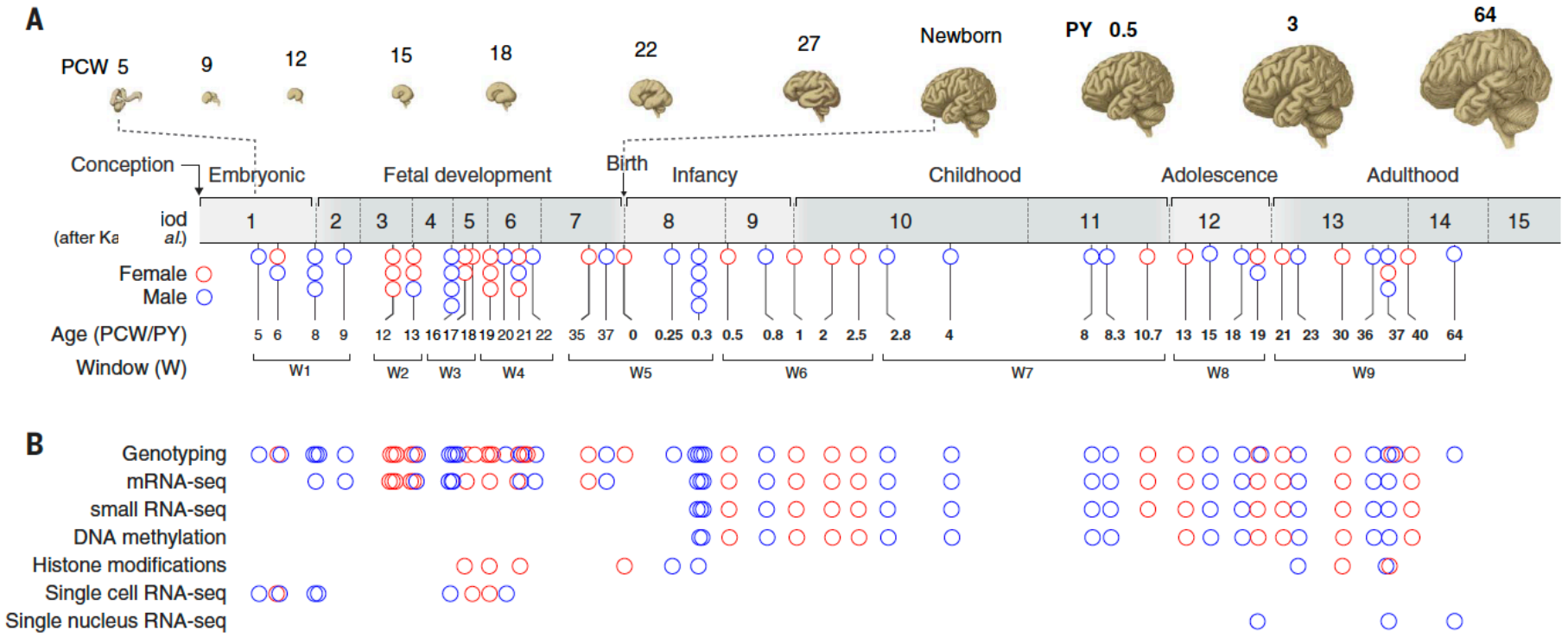
*https://www.snpedia.com/index.php/Heritability

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)

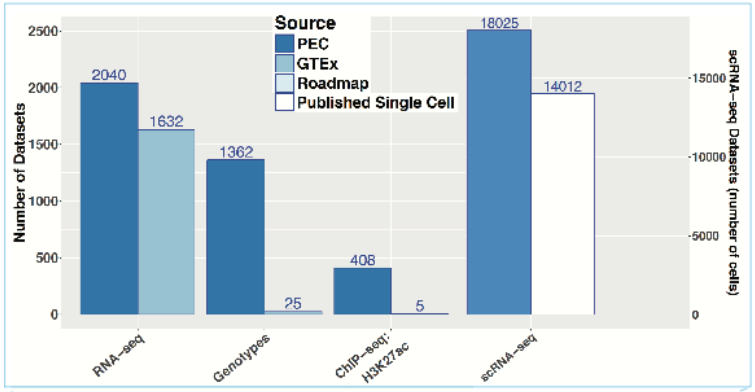
Thus, interested in developing predictive models of psychiatric traits which:
- Use observations at intermediate (molecular levels) levels to inform latent structure
- Use the predictive features of these "molecular endo phenotypes" to begin to suggest actors involved in mechanism

Developmental Capstone Data Set

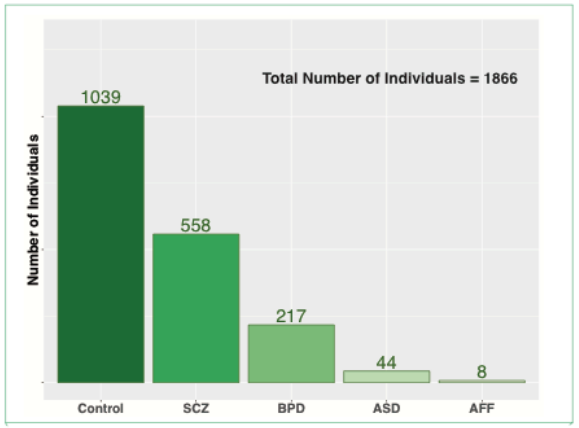


- 60 Individuals in total
- Ages from 5 PCW to 64 yrs.
- 16 brain regions for > 9 PCW

[Li et al. (‘18) Science]

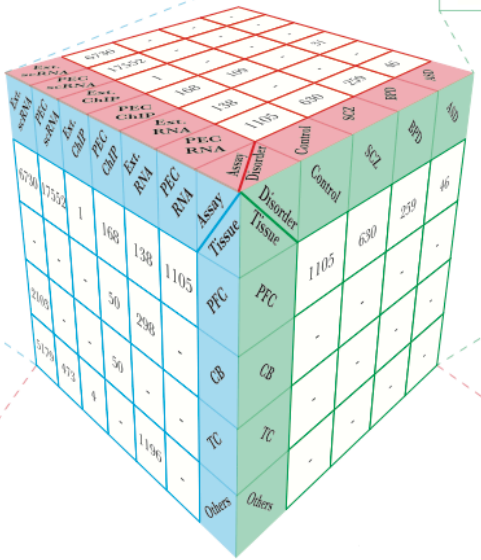


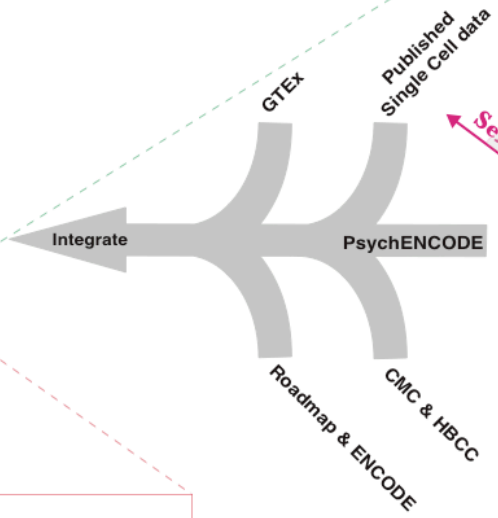
Assay

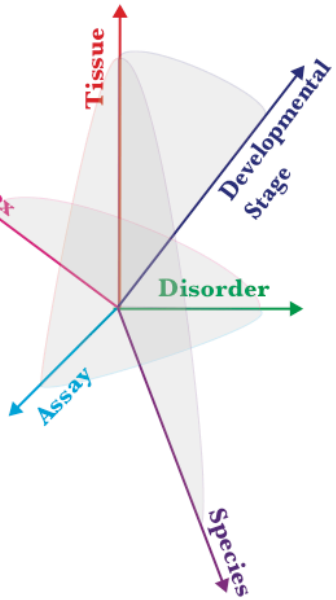


Disorder

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

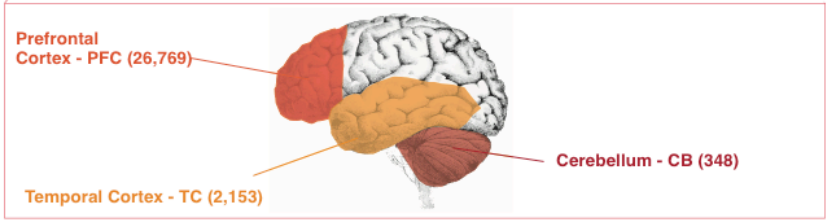






Collecting functional genomic datasets for the adult brain

from PsychENCODE, other large consortia & single cell studies

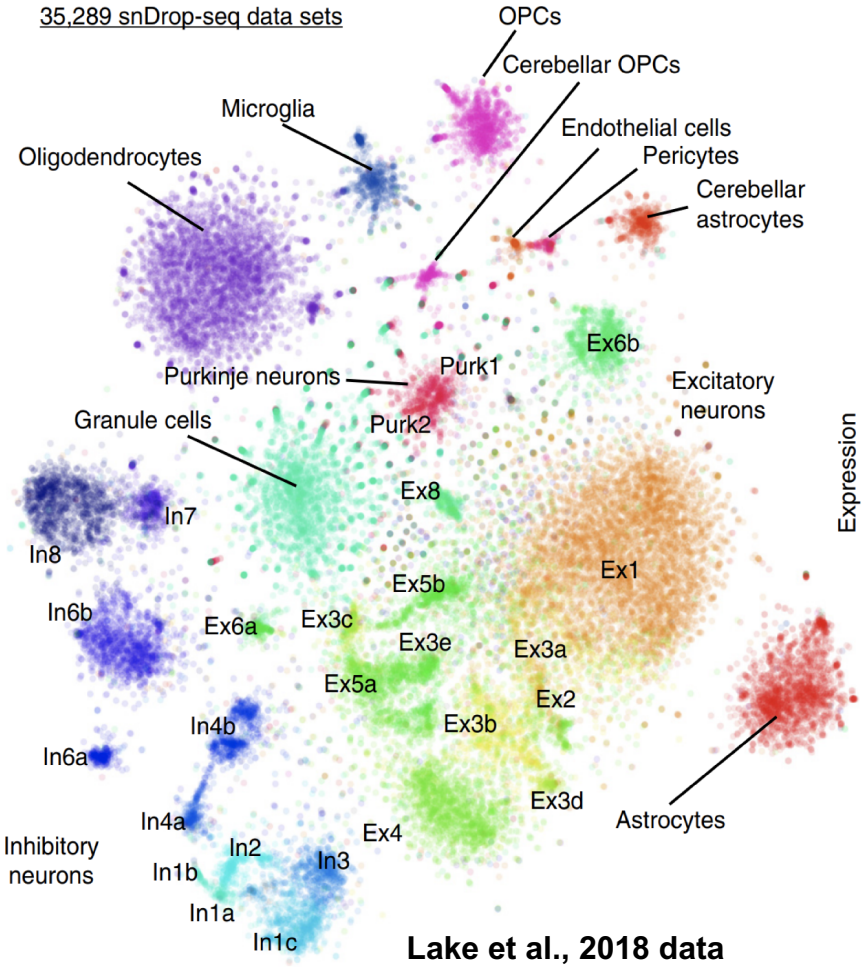


[Wang et al. (‘18) Science]

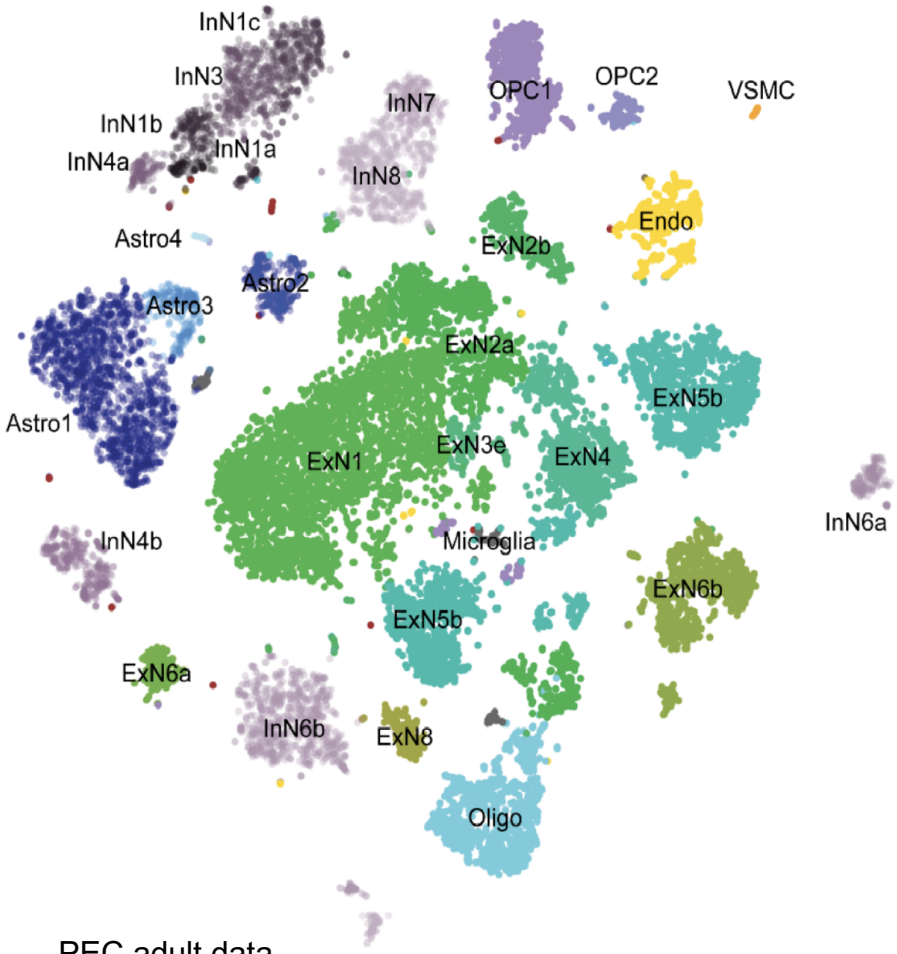
Merging & Clustering Single Cell Data Sets

Single cell signatures, from:

- ~14K cells
  (Lake et al.,'16 & '18)
- ~400 cells
  (Darmanis et al., PNAS, '15)
- ~18K cells (PsychENCODE)



Lake et al., 2018 data

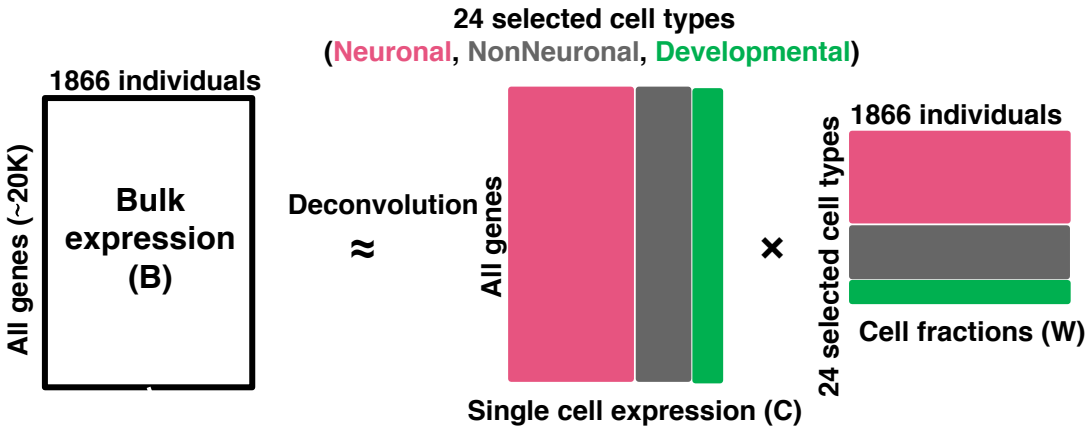


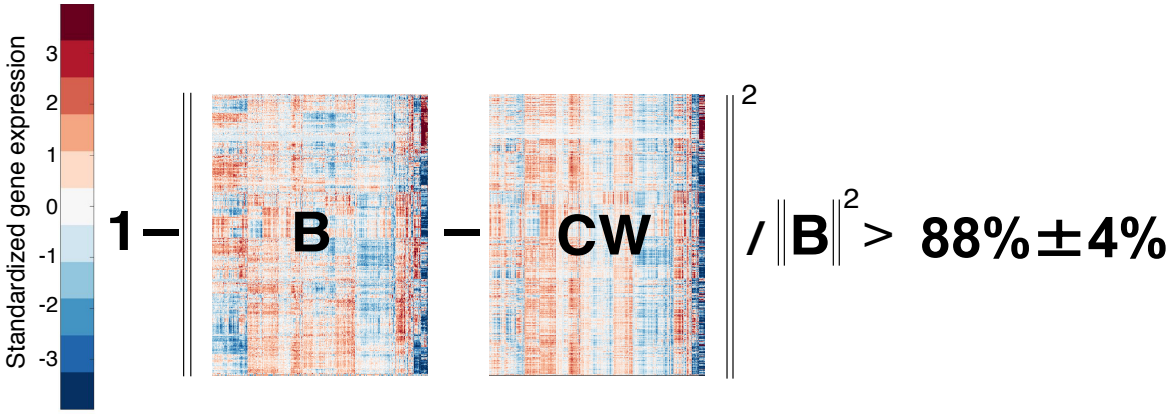
PEC adult data
[Li et al. ('18), Science. Wang et al. ('18). Science]

Single-cell deconvolution

Step 1:

Supervised learning to estimate cell fractions

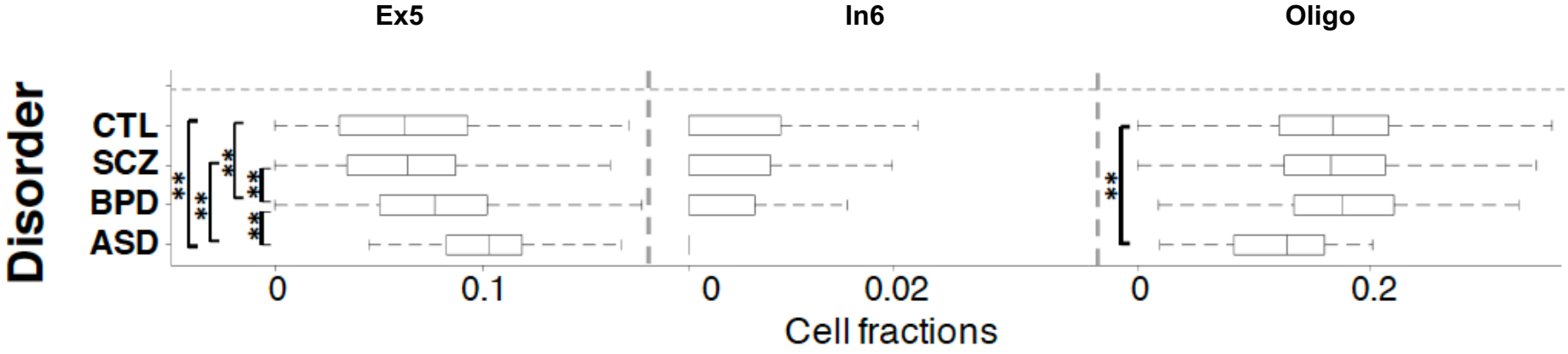




Individual and cross-population reconstruction accuracy via deconvolution

[Wang et al. ('18) Science]

Different neuronal & glial cell fractions across disorders

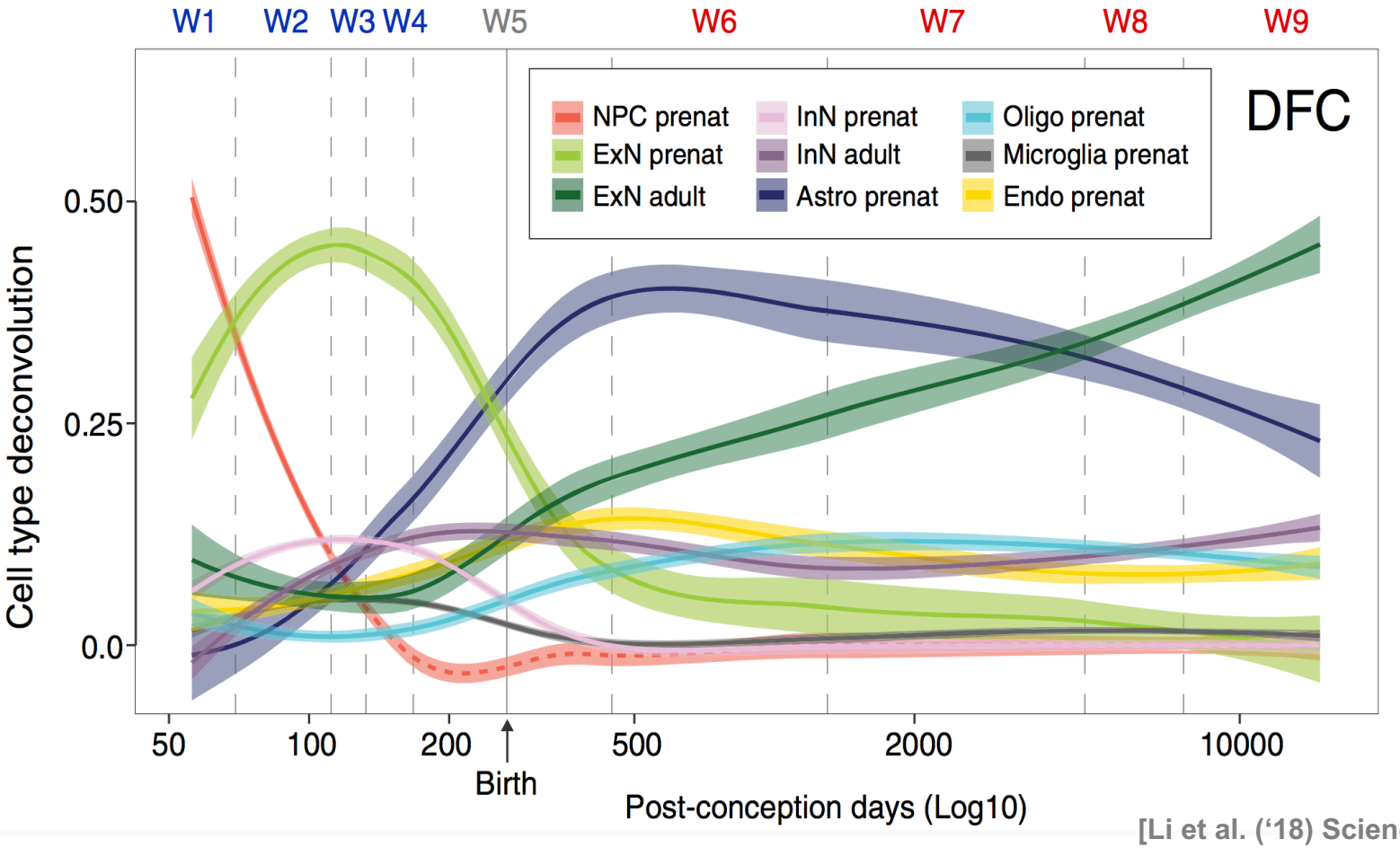


Excitatory to Inhibitory imbalance at neuronal subtype level for ASD*

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

[Wang et al. ('18) Science]

Different neuronal & glial cell fractions across ages



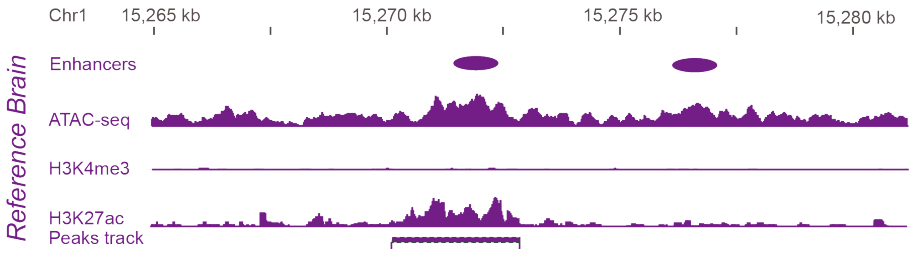
[Li et al. ('18) Science]

Using population-scale functional genomics to suggest potential neuropsychiatric drug targets & building a hybrid classifier to ascertain differential drug sensitivity

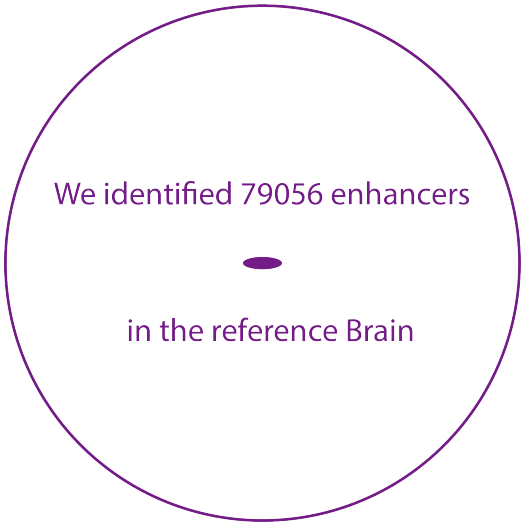
- **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + full developmental time-course
  - Using the changing proportions of cell types (via **single-cell deconvolution**) to account for expression variation across a population, disorders & development
  - 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 + consistently comparing the brain to other organs

- **GenoDock**: Building a predictor for the sensitivity of drug binding to personal SNVs
  - Hybrid classifier connecting **physical modelling with statistical learning**
    - The modeling creates a pseudo gold-standard dataset, which is used to train the stat. classifier
  - **Classifier Results**
    - Independent validation on an expt. validation set
    - Gives higher disruption scores to cancer driver SNVs. Also, illustrates importance of different features (eg GERP).
    - Picks out certain drugs (eg imatinib) as being particularly sensitive to SNVs

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

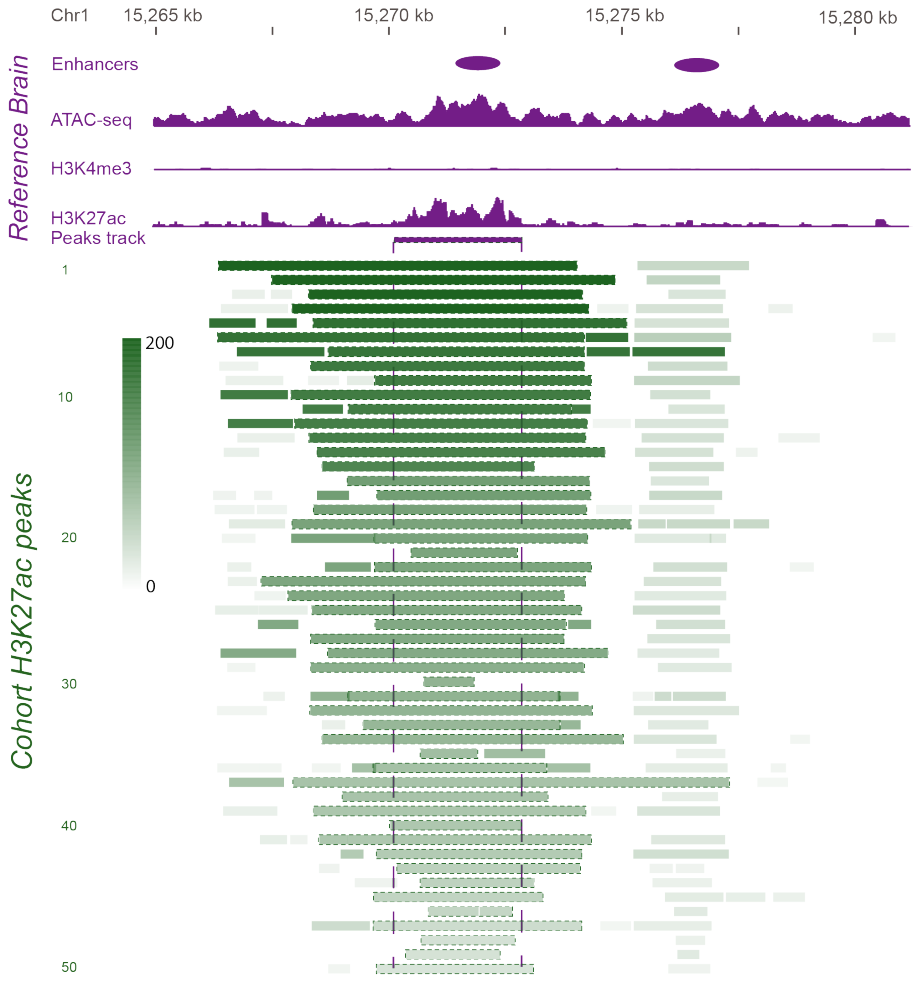


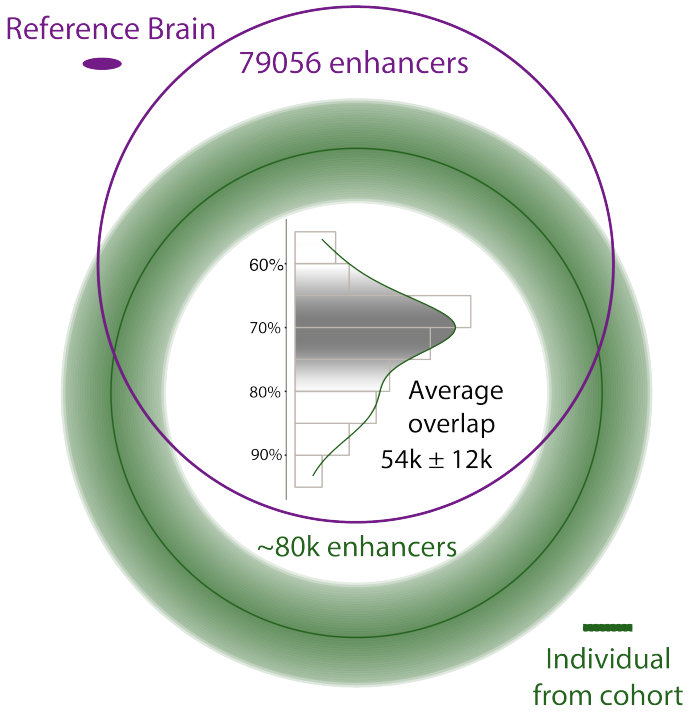
Consistent with ENCODE, active enhancers are identified as open chromatin regions enriched in H3K27ac and depleted in H3K4me3



[Wang et al. (‘18) Science]

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

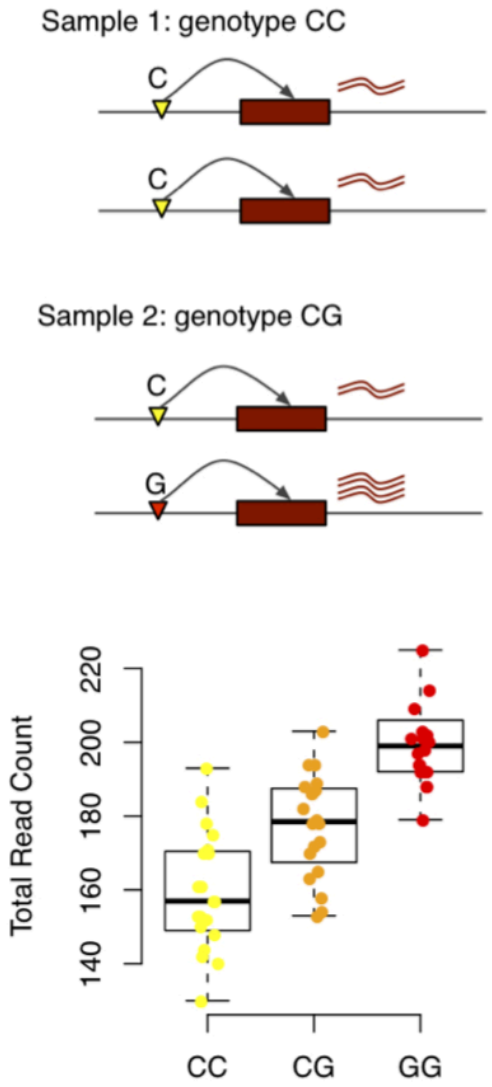




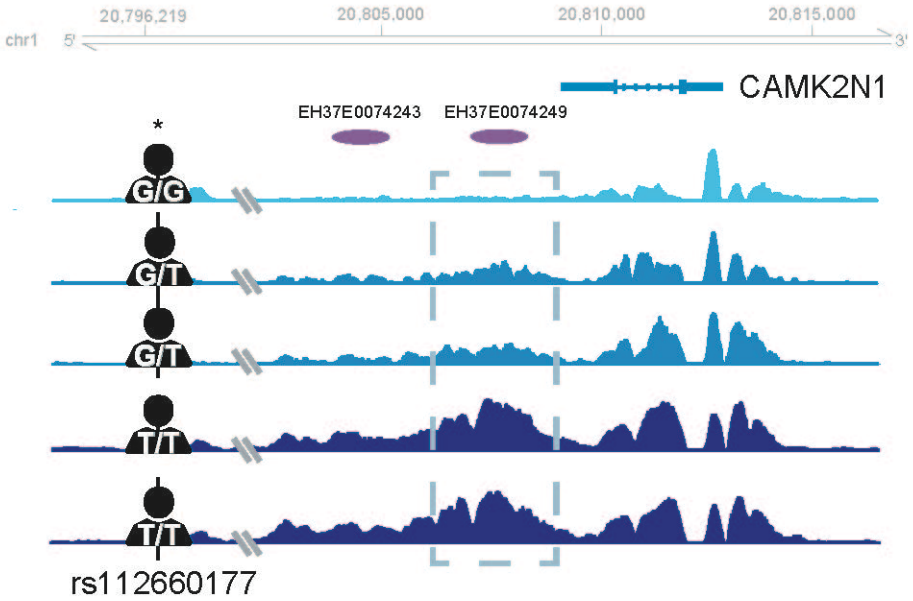
[Wang et al. ('18) Science]

Quantitaive Trait Loci (QTLs) associated with variation

Gene expression (eQTL)



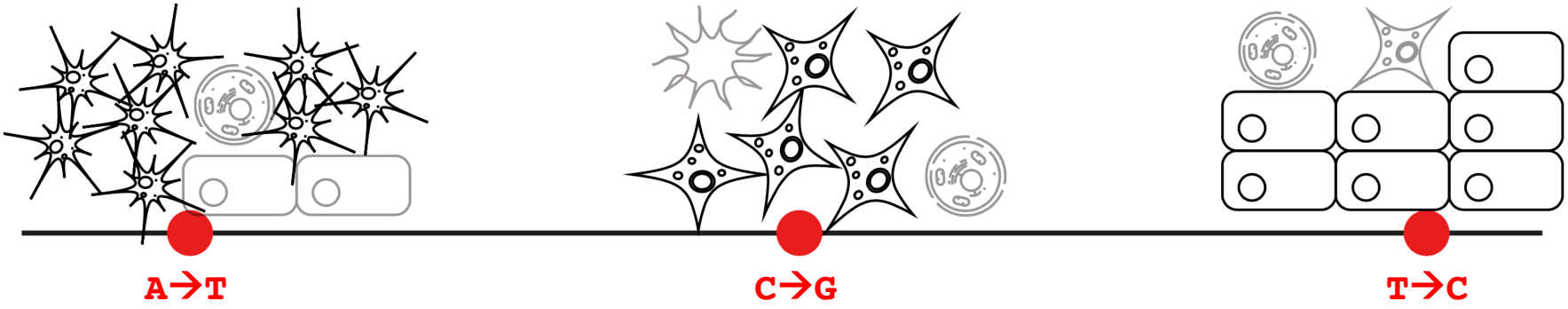
Chromatin (cQTL)

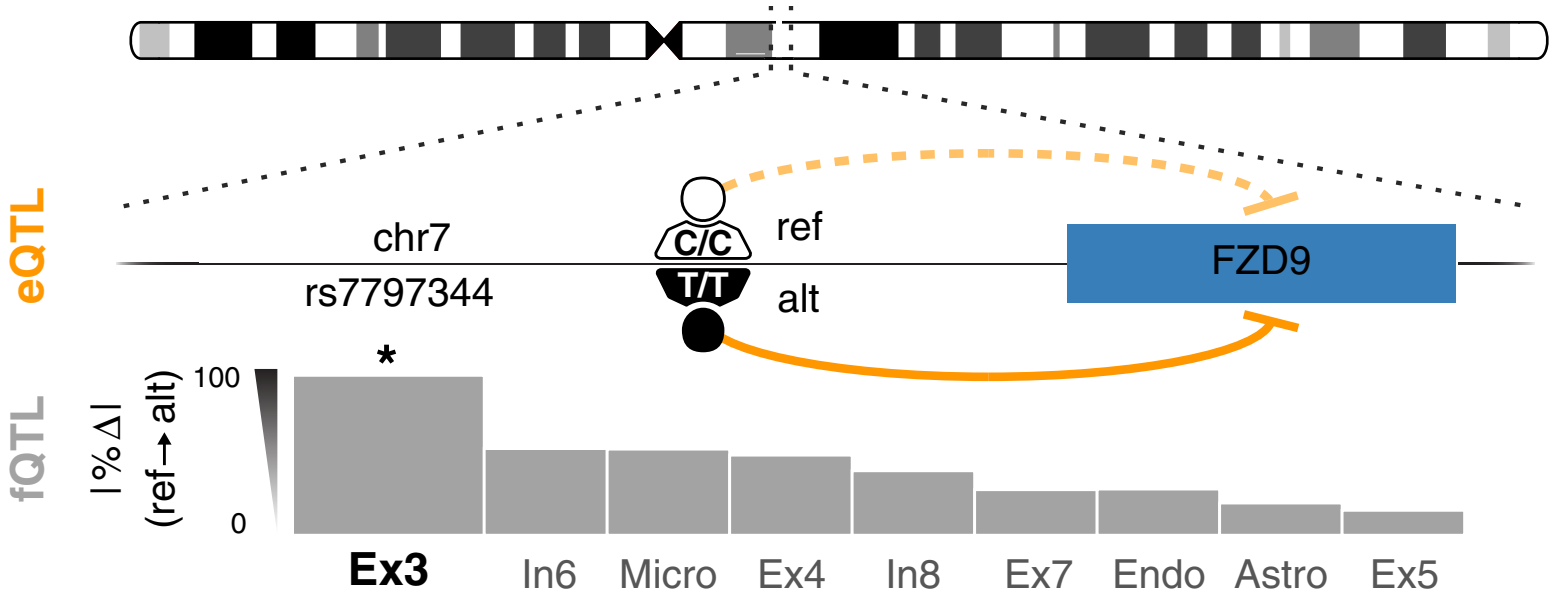


Sun, Wei, and Yijuan Hu. "eQTL mapping using RNA-seq data." *Statistics in biosciences* 5.1 (2013): 198-219.

[Wang et al. ('18) Science]

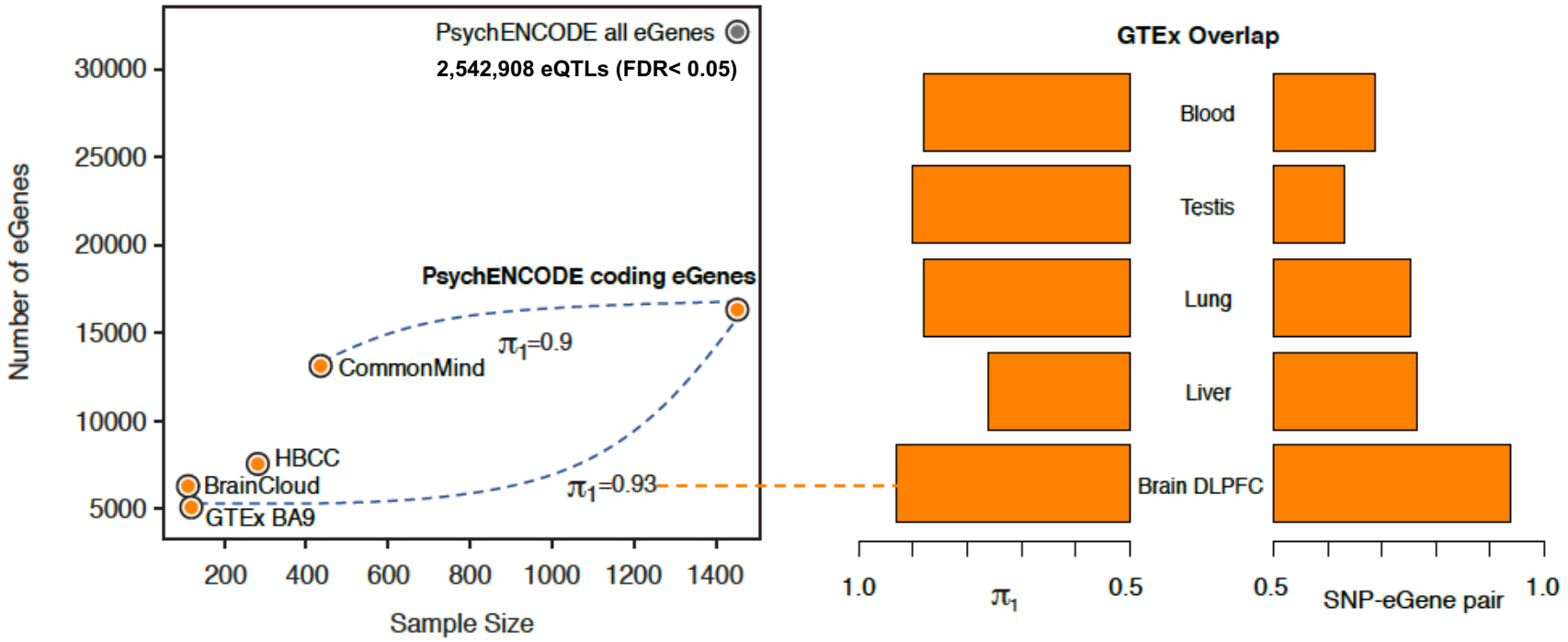
Cell fraction QTLs (fQTLs)





[Wang et al. ('18) Science]

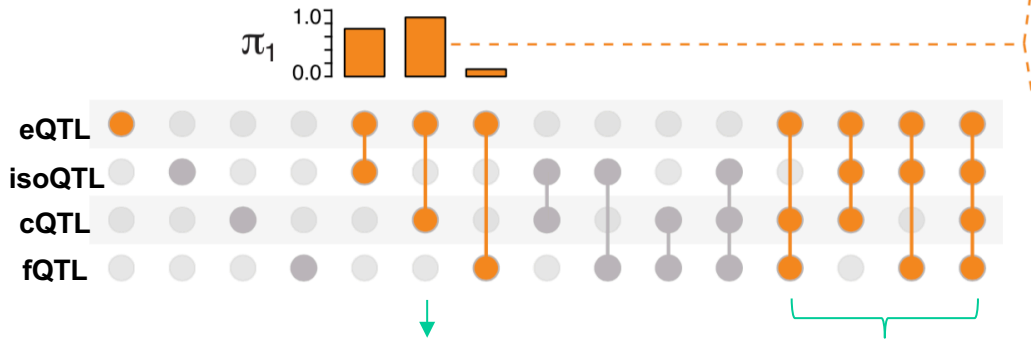
Larger brain eQTL sets than previous studies, but strong overlap with them



[Wang et al. ('18) Science]

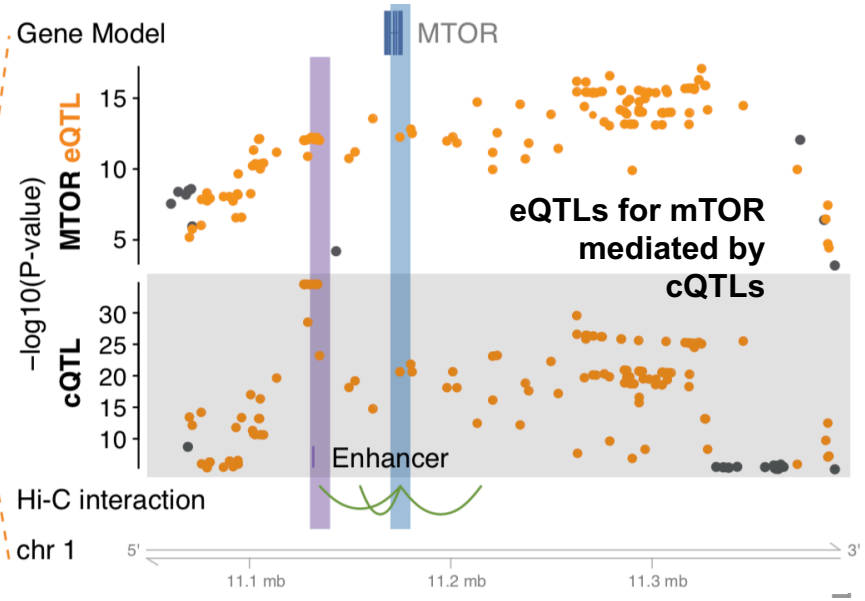
multi-QTLs from overlapping different types of QTLs: cQTL, fQTL, eQTL & isoQTL

	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 and cQTLs significantly overlap

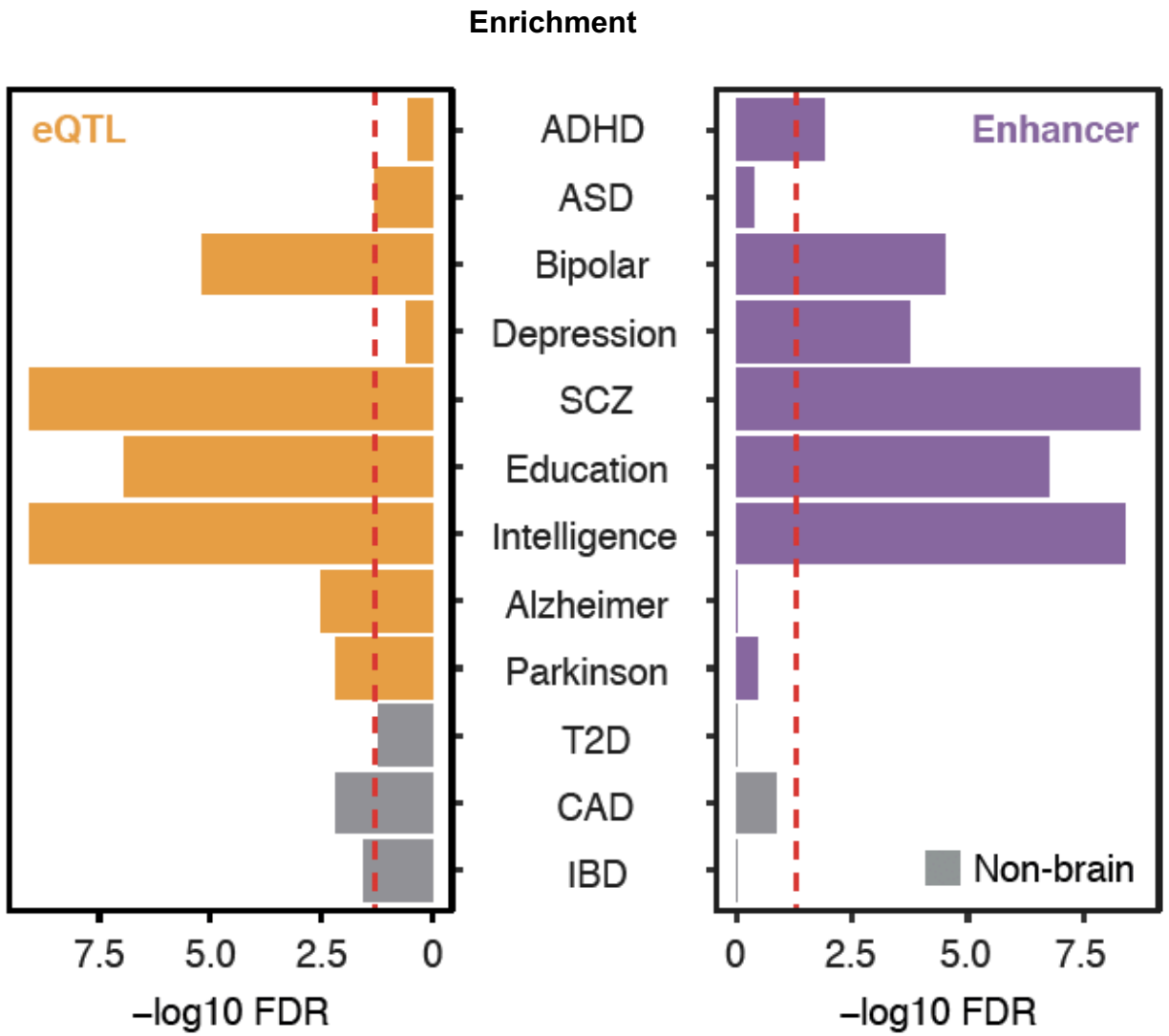
1391 SNPs (multi-QTLs) in at least three types among eQTLs, isoQTLs, cQTLs, fQTLs



eQTLs for mTOR mediated by cQTLs

[Wang et al. ('18) Science]

Brain eQTLs and enhancers enriched with GWAS SNPs for brain disorders



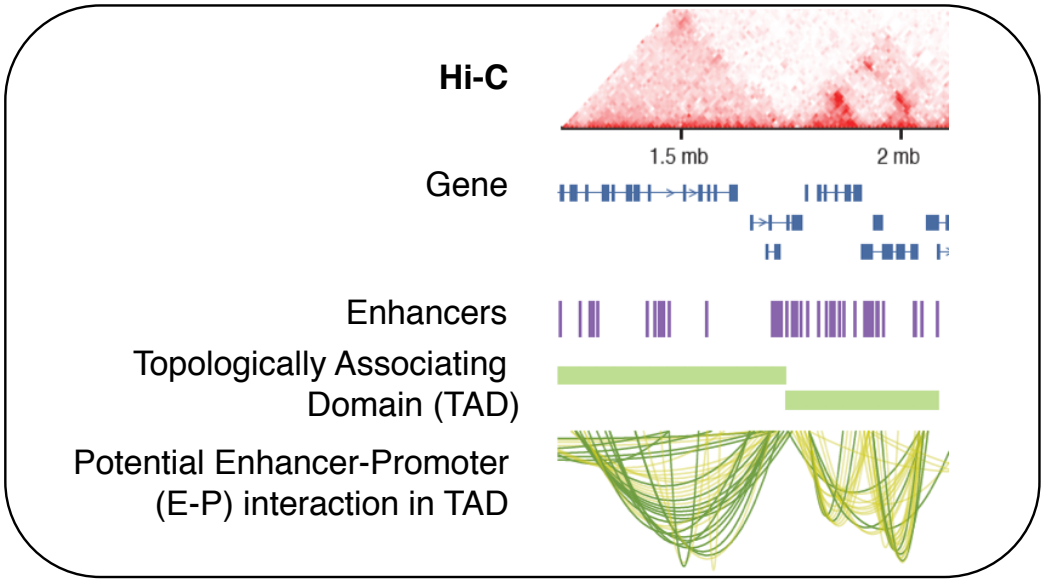
Wang, et al., *Science*, 2018

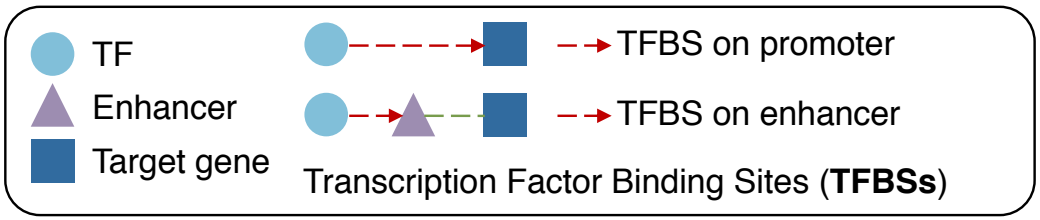
Using population-scale functional genomics to suggest potential neuropsychiatric drug targets & building a hybrid classifier to ascertain differential drug sensitivity

- **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + full developmental time-course
  - Using the changing proportions of cell types (via **single-cell deconvolution**) to account for expression variation across a population, disorders & development
  - 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 + consistently comparing the brain to other organs

- **GenoDock**: Building a predictor for the sensitivity of drug binding to personal SNVs
  - Hybrid classifier connecting **physical modelling with statistical learning**
    - The modeling creates a pseudo gold-standard dataset, which is used to train the stat. classifier
  - **Classifier Results**
    - Independent validation on an expt. validation set
    - Gives higher disruption scores to cancer driver SNVs. Also, illustrates importance of different features (eg GERP).
    - Picks out certain drugs (eg imatinib) as being particularly sensitive to SNVs

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

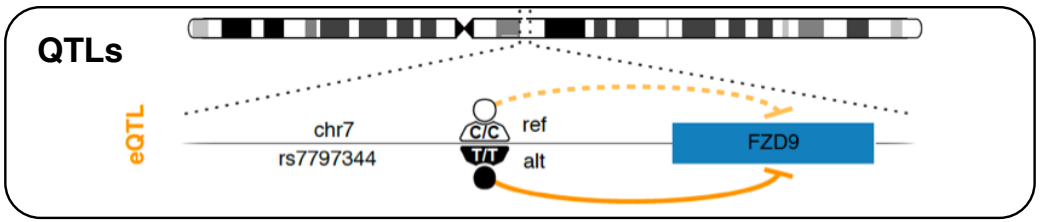




Expression activity relationship

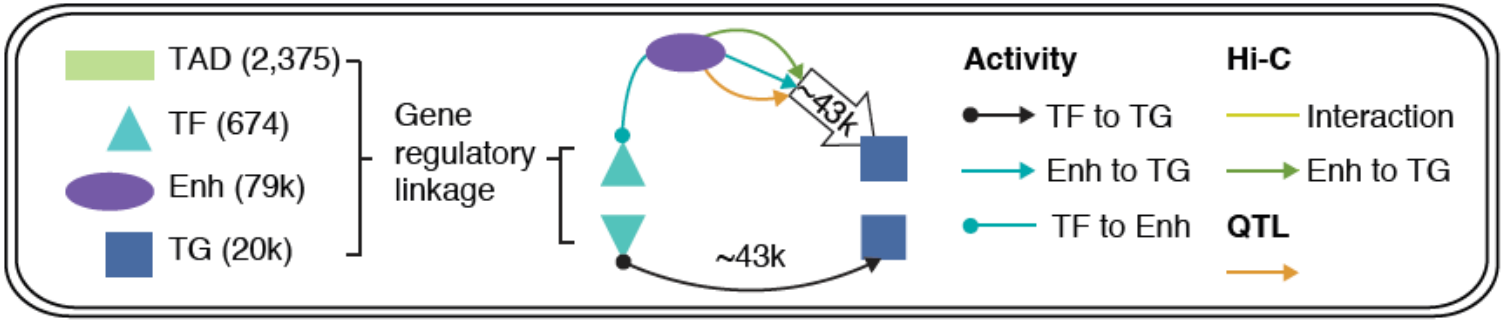
$C^* = argmin_C(\|Y - XC\|^2 + a\|C\|^2 + b\|C\|_{L1})$

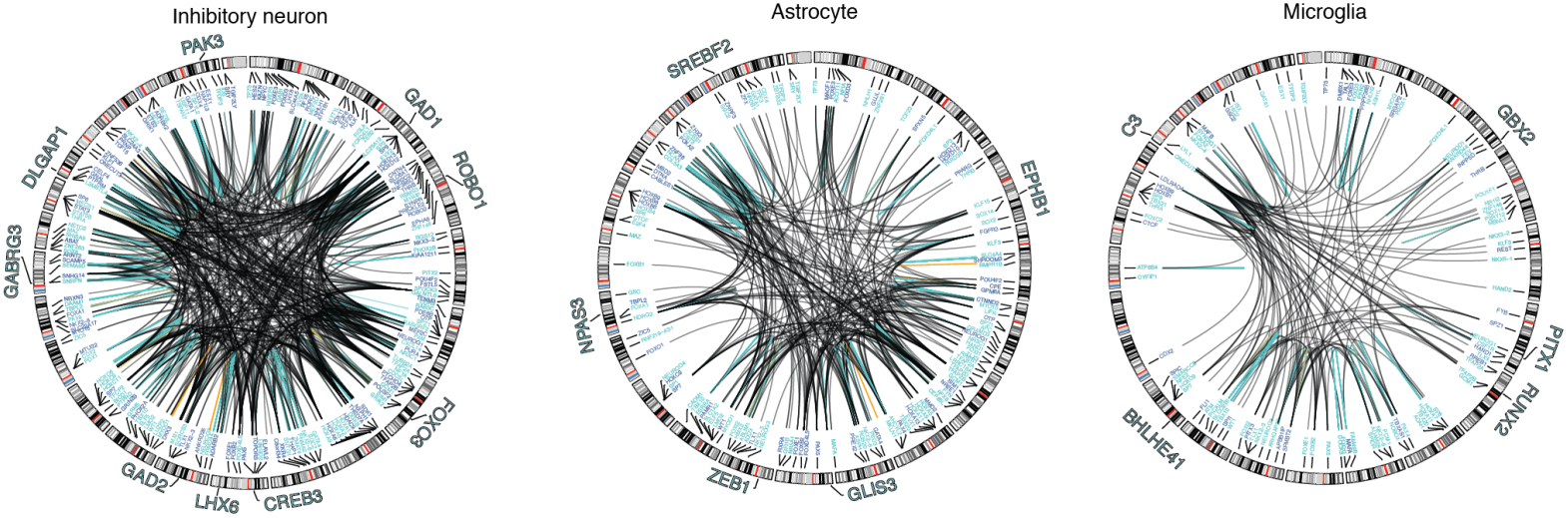
TF expression (X) to predict target gene expression (Y) using Elastic net regression



[Wang et al. ('18) Science]

Imputed gene regulatory network for the human brain

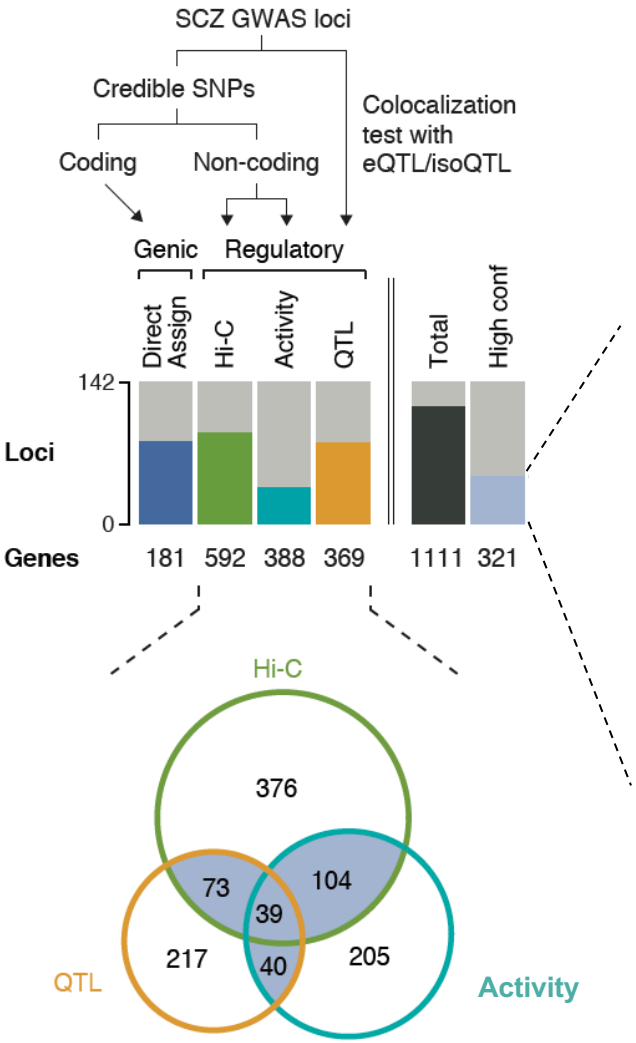


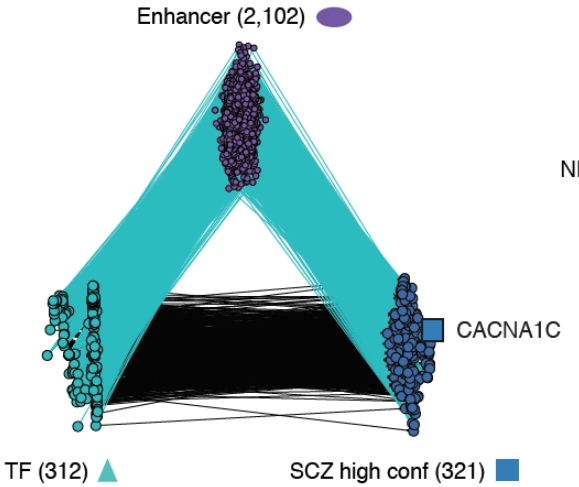


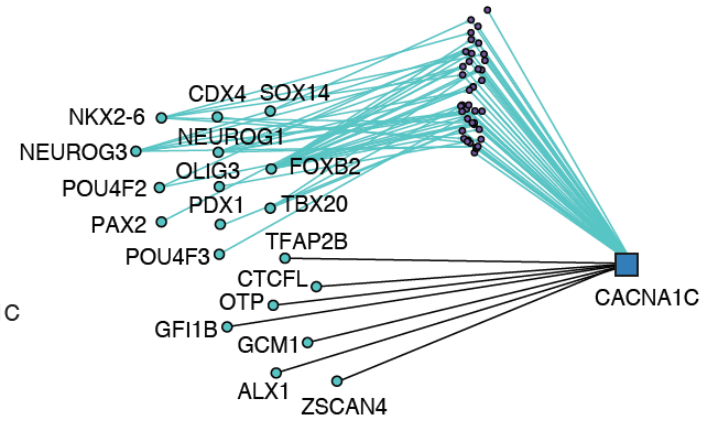
subnetworks targeting single cell marker genes

[Wang et al. ('18) Science]

Linking GWAS SNPs to disease genes using the regulatory network



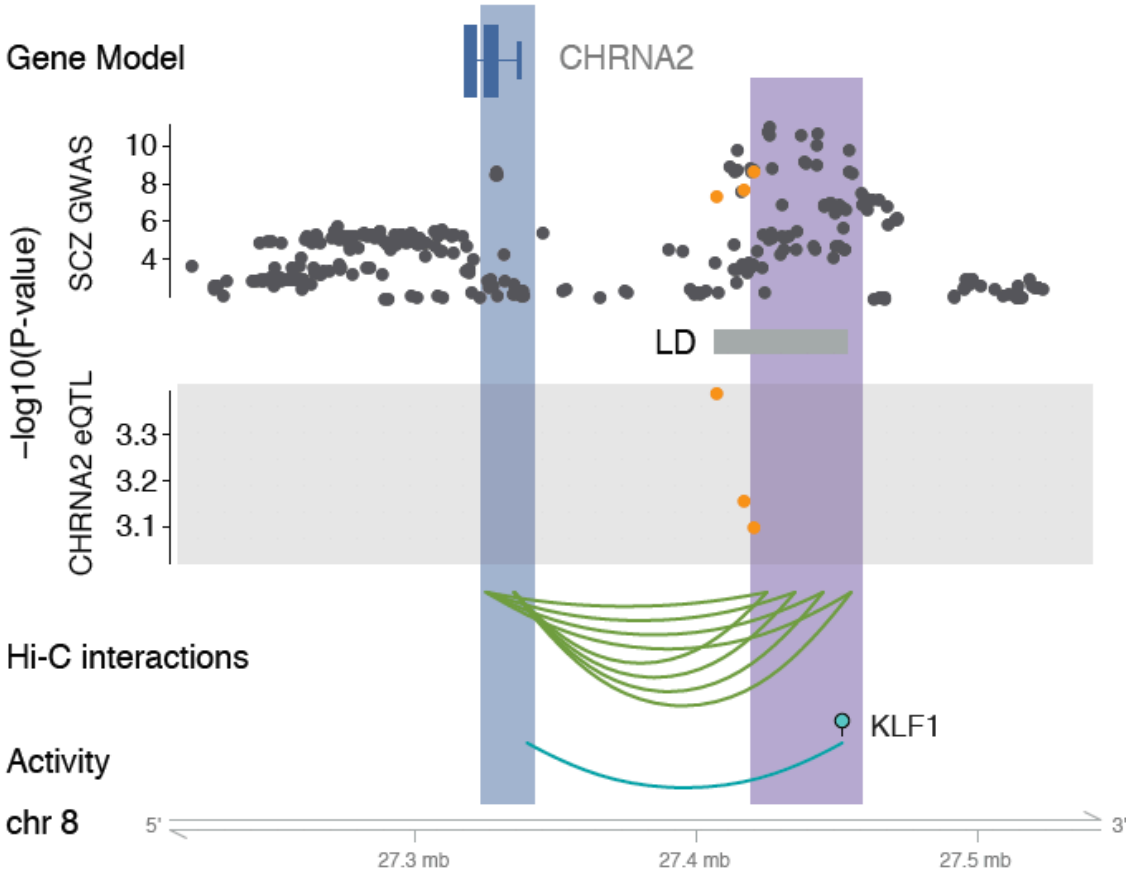


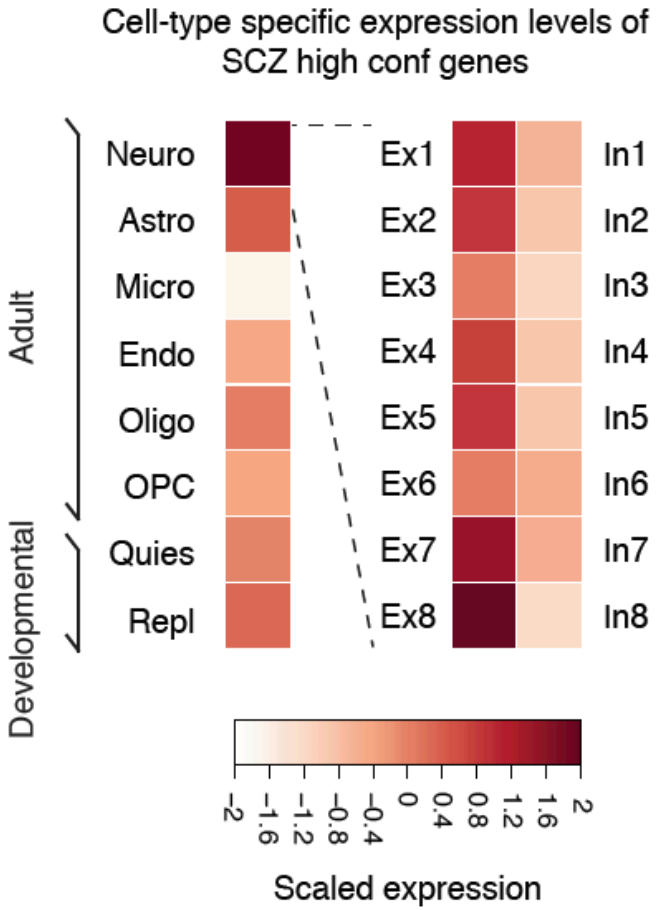


321
high-confident SCZ genes

[Wang et al. ('18) Science]

GWAS variants and single cell expression levels for SCZ genes





Wang, et al., *Science*, 2018

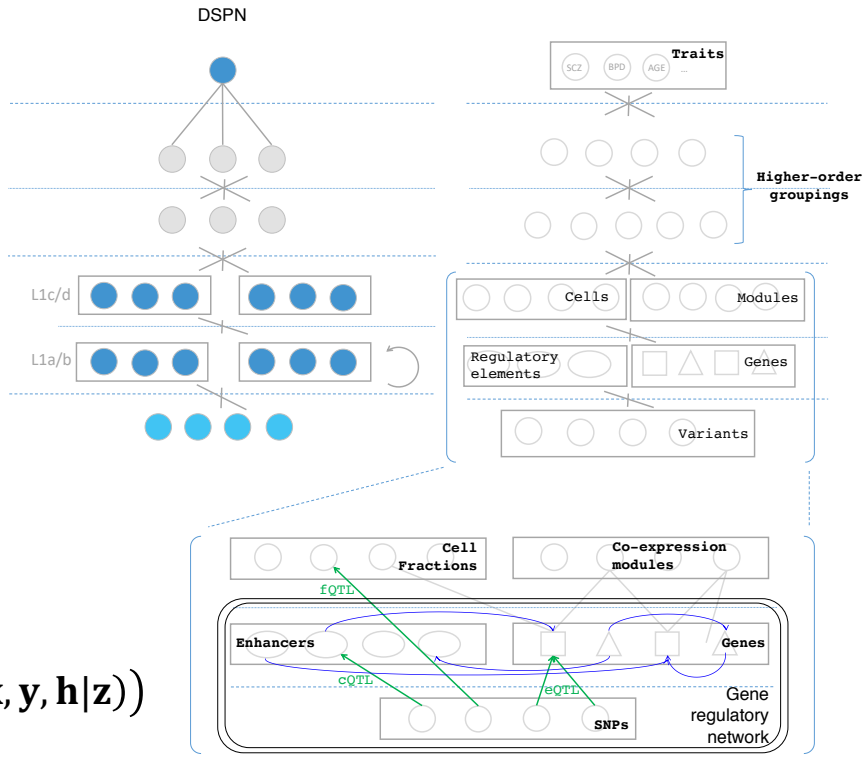
Using population-scale functional genomics to suggest potential neuropsychiatric drug targets & building a hybrid classifier to ascertain differential drug sensitivity

- **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + full developmental time-course
  - Using the changing proportions of cell types (via **single-cell deconvolution**) to account for expression variation across a population, disorders & development
  - 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 + consistently comparing the brain to other organs

- **GenoDock**: Building a predictor for the sensitivity of drug binding to personal SNVs
  - Hybrid classifier connecting **physical modelling with statistical learning**
    - The modeling creates a pseudo gold-standard dataset, which is used to train the stat. classifier
  - **Classifier Results**
    - Independent validation on an expt. validation set
    - Gives higher disruption scores to cancer driver SNVs. Also, illustrates importance of different features (eg GERP).
    - Picks out certain drugs (eg imatinib) as being particularly sensitive to SNVs

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)

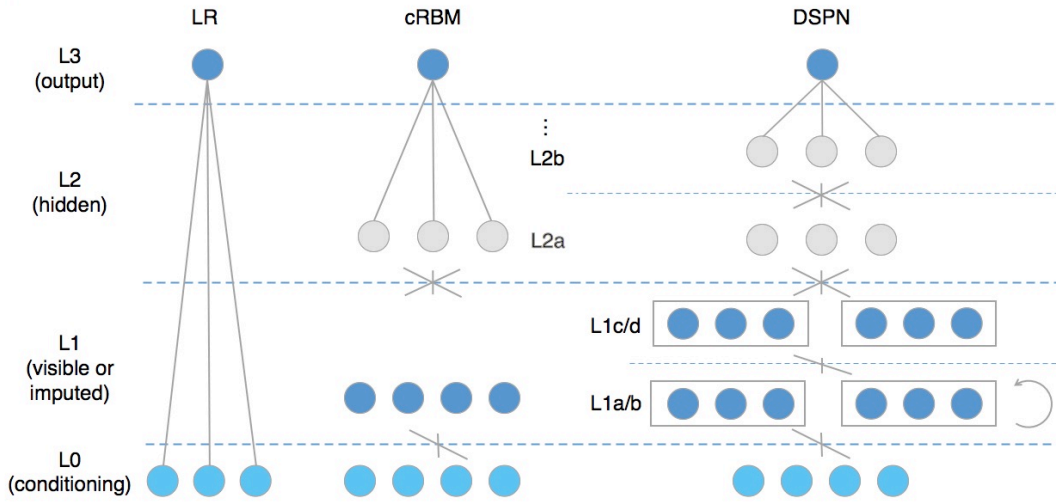
Energy model:

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

$$E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z}) = -\mathbf{z}^{\mathrm{T}}\mathbf{W}_1\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W}_2\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W}_3\mathbf{h} - \mathbf{h}^{\mathrm{T}}\mathbf{W}_4\mathbf{h} - \mathbf{h}^{\mathrm{T}}\mathbf{W}_5\mathbf{y} - \boldsymbol{Bias}$$

[Wang et al. ('18) Science]

DSPN improves brain disease prediction by adding deep layers



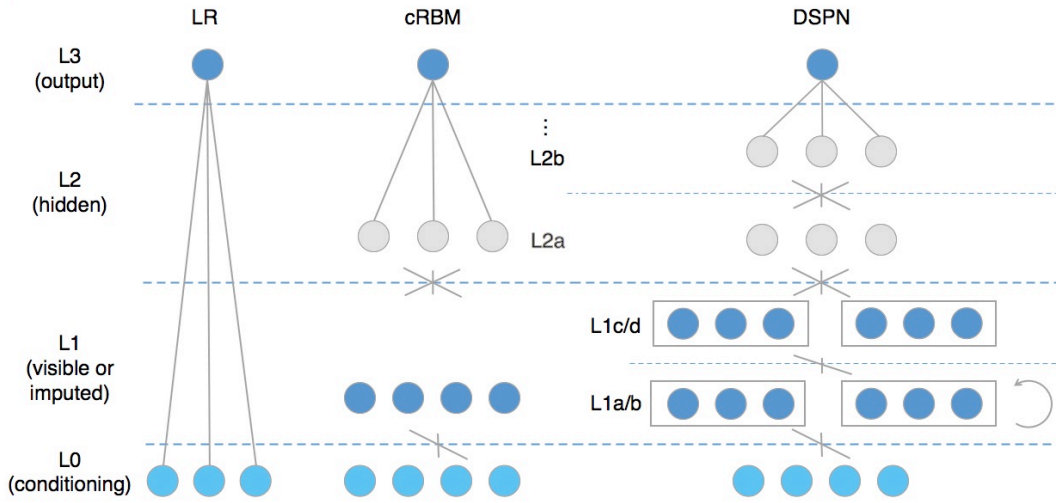
Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	**54.6%**	63.0%	70.0%	59.0%	**73.6%**
Bipolar Disorder	**56.7%**	63.3%	71.1%	67.2%	**76.7%**
Autism Spectrum Disorder	**50.0%**	51.7%	67.2%	62.5%	**68.3%**

X 6.0

Accuracy = chance to correctly predict disease/health

[Wang et al. ('18) Science]

DSPN improves brain disease prediction by adding deep layers



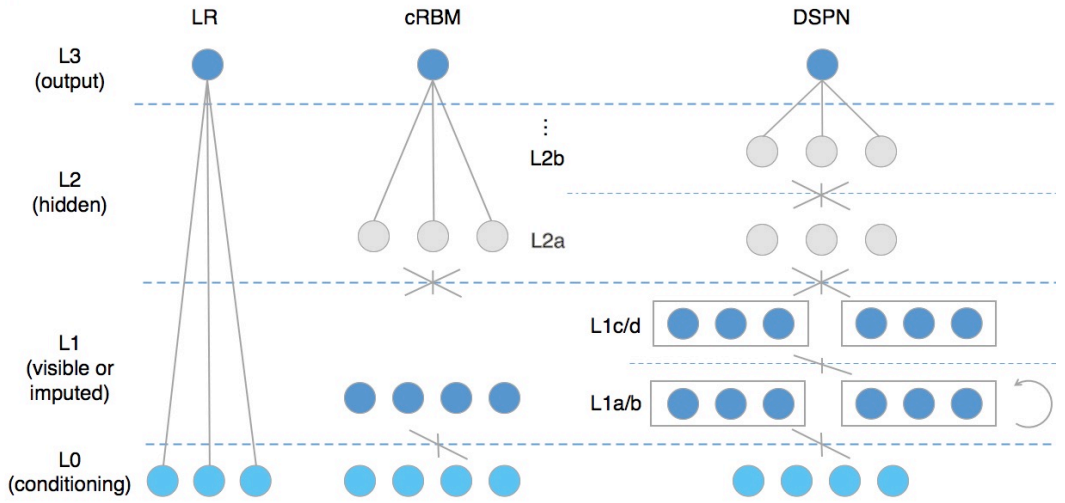
Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	54.6%	63.0%	70.0%	59.0%	73.6%
Bipolar Disorder	56.7%	63.3%	71.1%	67.2%	76.7%
Autism Spectrum Disorder	50.0%	51.7%	67.2%	62.5%	68.3%

X 2.5

Accuracy = chance to correctly predict disease/health

[Wang et al. ('18) Science]

DSPN improves brain disease prediction by adding deep layers



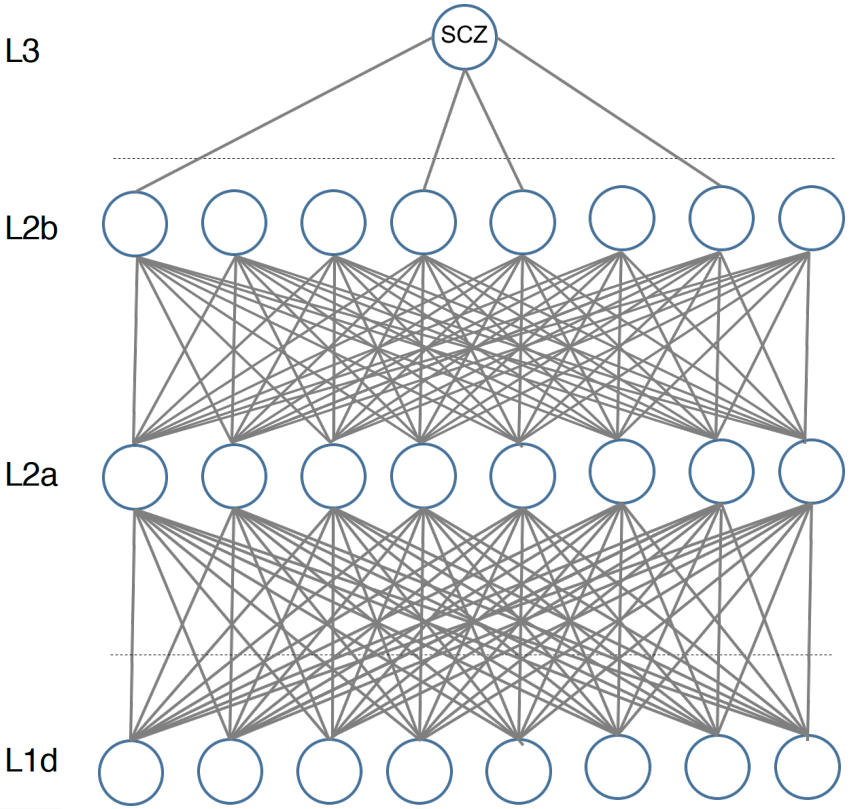
Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	**54.6%**	63.0%	70.0%	**59.0%**	73.6%
Bipolar Disorder	**56.7%**	63.3%	71.1%	**67.2%**	76.7%
Autism Spectrum Disorder	**50.0%**	51.7%	67.2%	**62.5%**	68.3%

X 3.1

Accuracy = chance to correctly predict disease/health

[Wang et al. ('18) Science]

Multilevel Network Interpretation

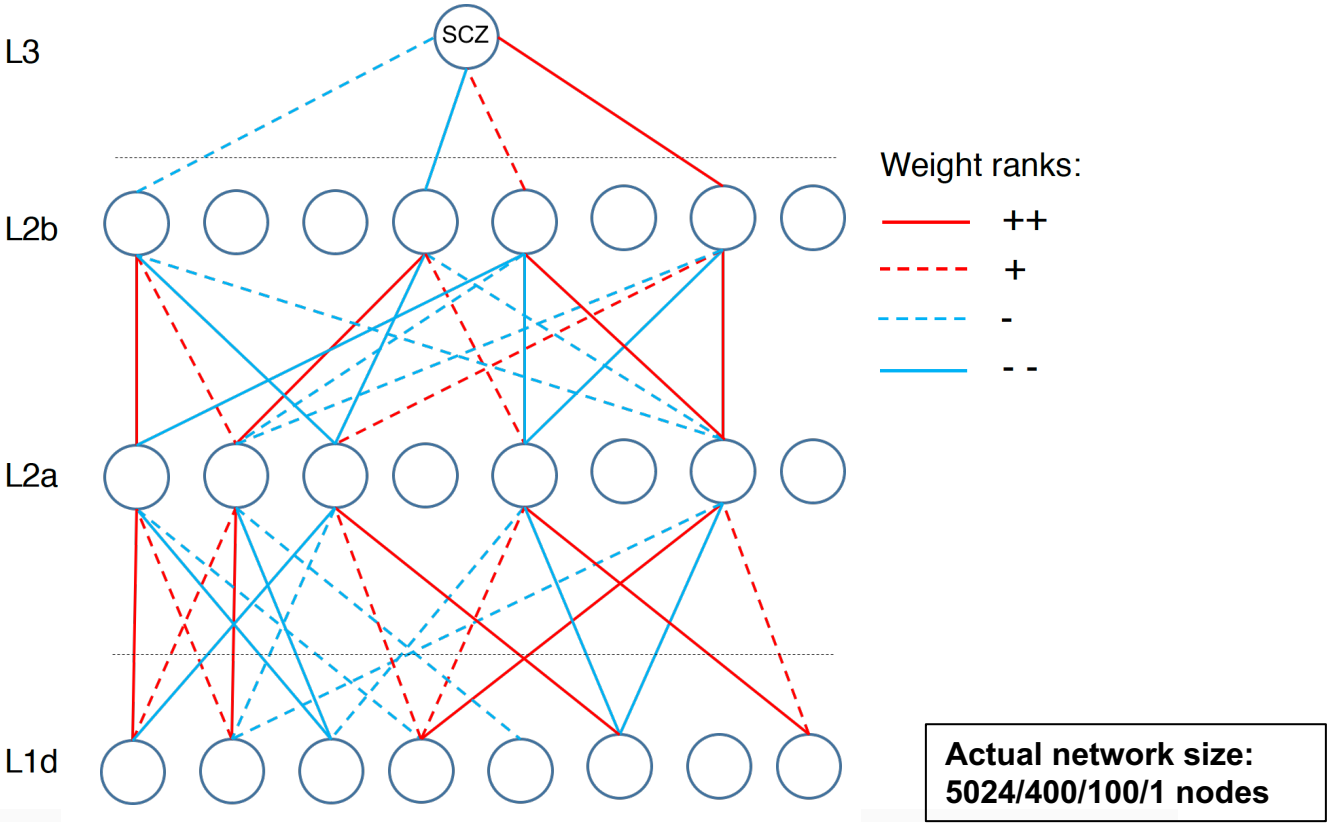


Actual network size:
5024/400/100/1 nodes

- Start with a fully connected trained network

[Wang et al. (‘18) Science]

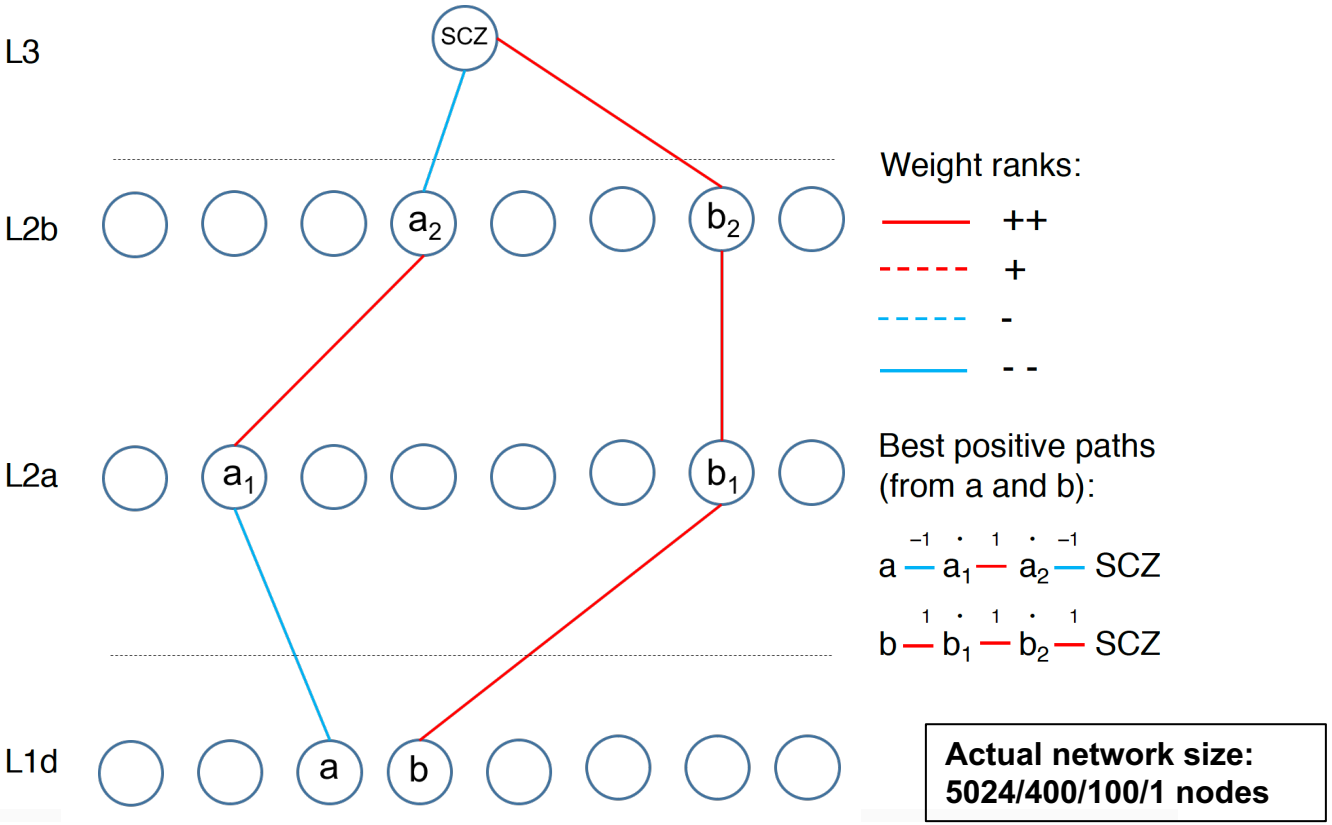
Multilevel Network Interpretation



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

[Wang et al. (‘18) Science]

Multilevel Network Interpretation

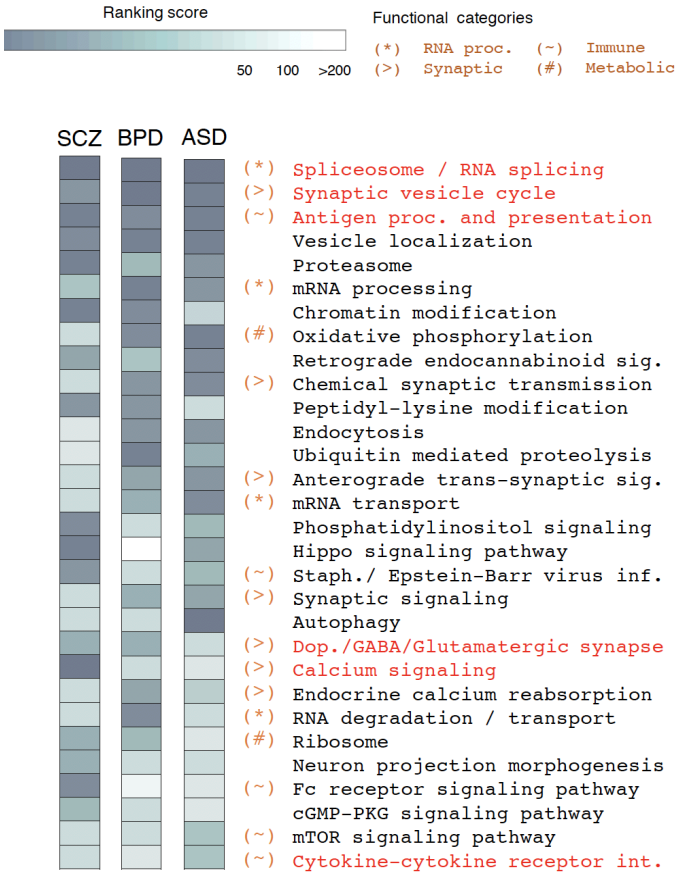


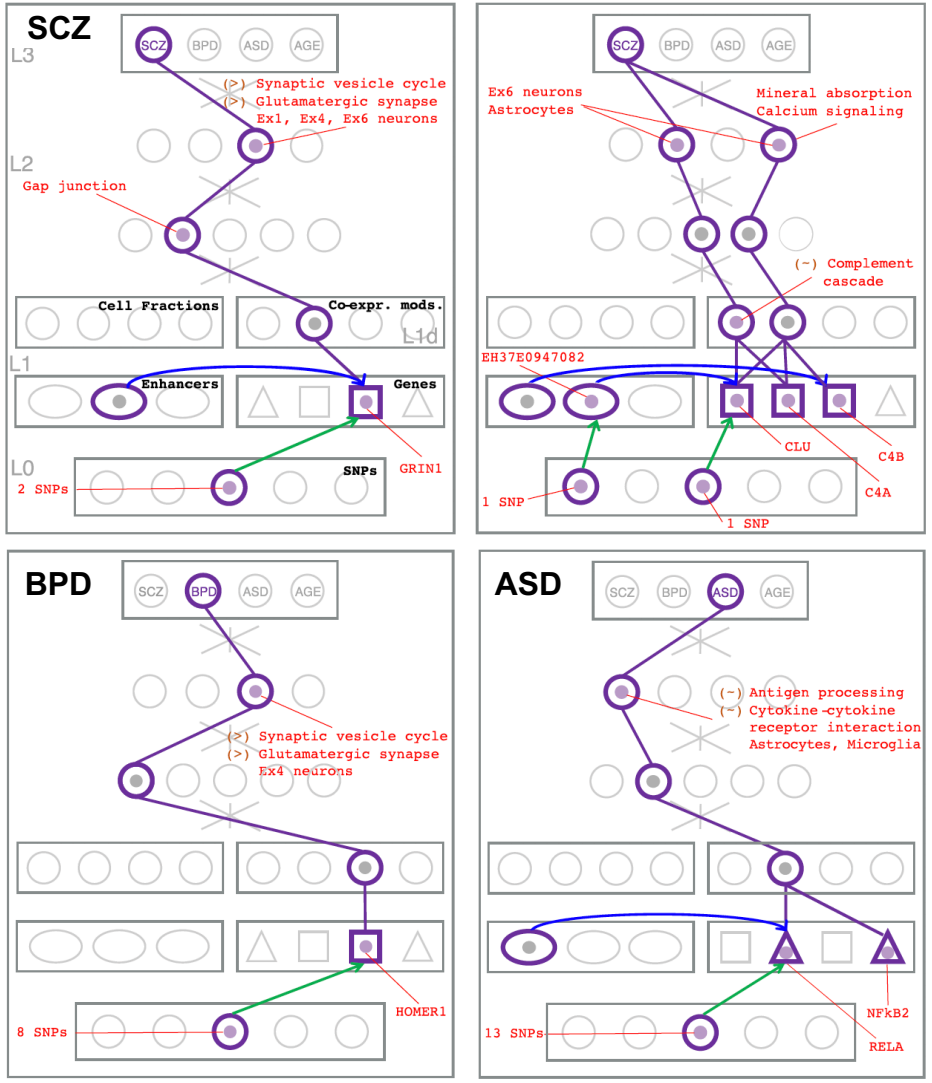
- Start with a fully connected trained network
- Sparsify network using edges with largest absolute weights (+/-)
- Extract 'best positive paths' to each prioritized module (e.g. a-a_1-a_2-SCZ) by summing weights and multiplying signs

[Wang et al. ('18) Science]

DSPN discovers enriched pathways and linkages to genetic variation

Cross-disorder MOD/HOG enrichment ranking





[Wang et al. ('18) Science]

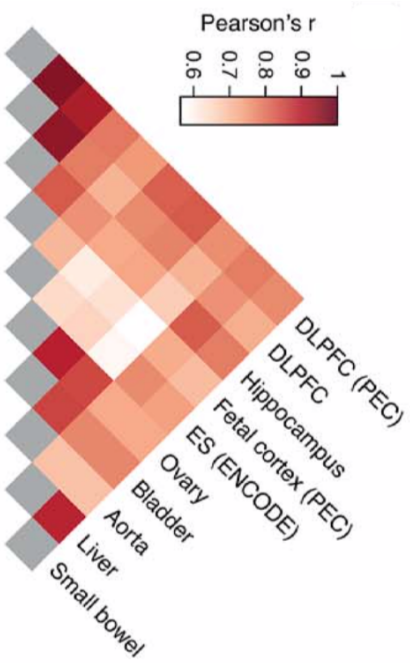
Using population-scale functional genomics to suggest potential neuropsychiatric drug targets & building a hybrid classifier to ascertain differential drug sensitivity

- **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + full developmental time-course
  - Using the changing proportions of cell types (via **single-cell deconvolution**) to account for expression variation across a population, disorders & development
  - 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 + consistently comparing the brain to other organs

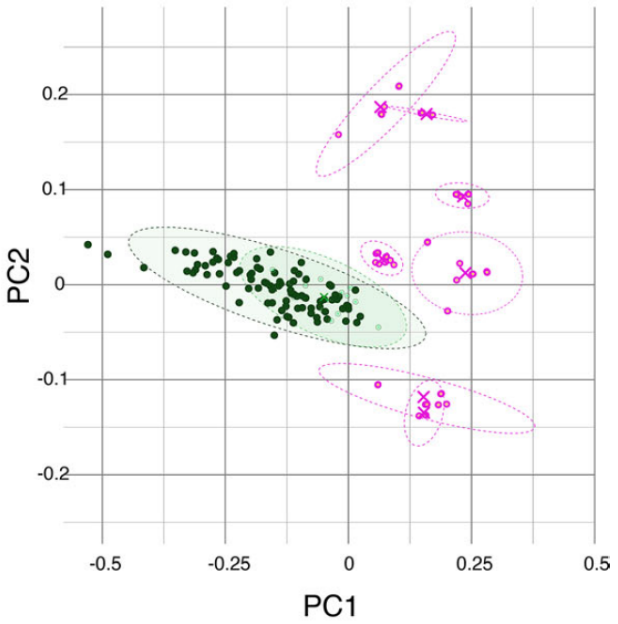
- **GenoDock**: Building a predictor for the sensitivity of drug binding to personal SNVs
  - Hybrid classifier connecting **physical modelling with statistical learning**
    - The modeling creates a pseudo gold-standard dataset, which is used to train the stat. classifier
  - **Classifier Results**
    - Independent validation on an expt. validation set
    - Gives higher disruption scores to cancer driver SNVs. Also, illustrates importance of different features (eg GERP).
    - Picks out certain drugs (eg imatinib) as being particularly sensitive to SNVs

Cross tissue variation in Chromatin & Expression

Placing the Brain in context of all other Body Tissues

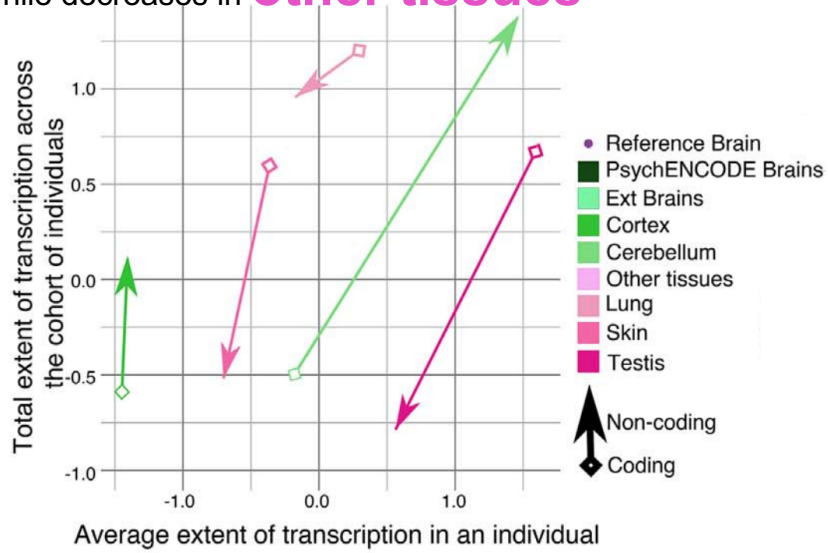


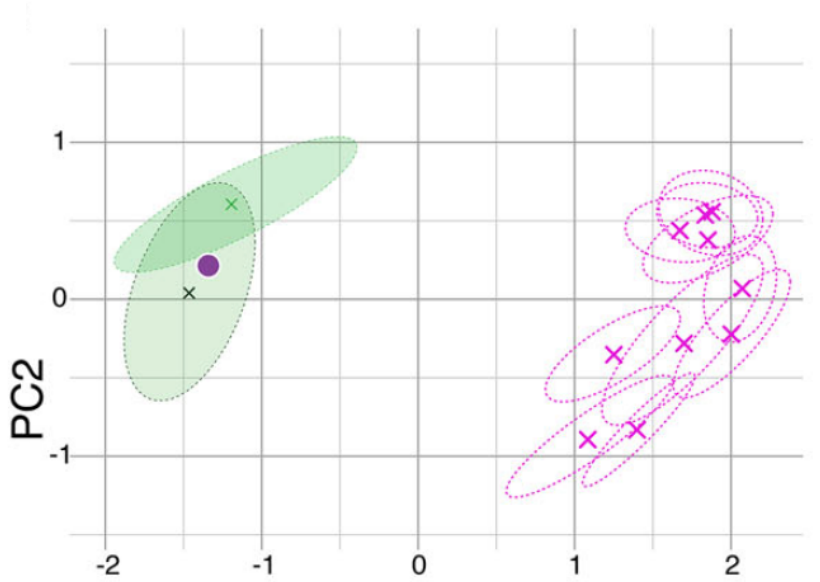
Hi-C



Chromatin

Transcriptome diversity increases in the non-coding portion of the **brain genome** while decreases in **other tissues**



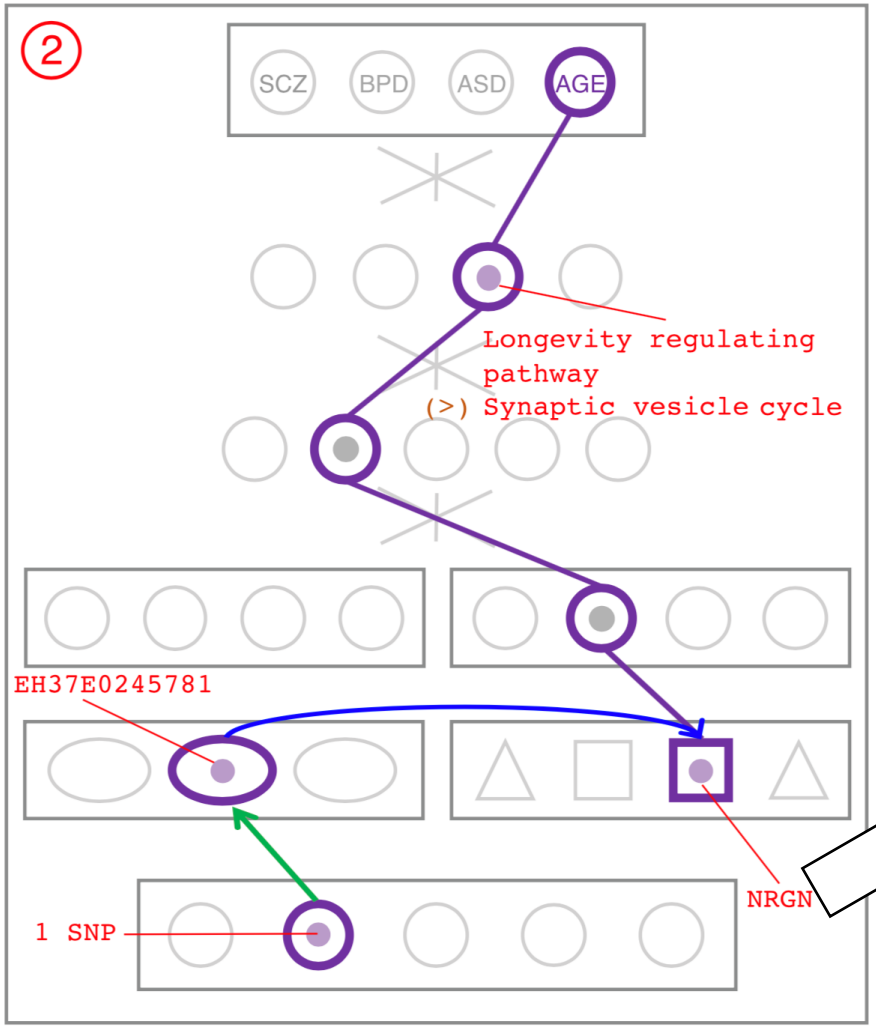


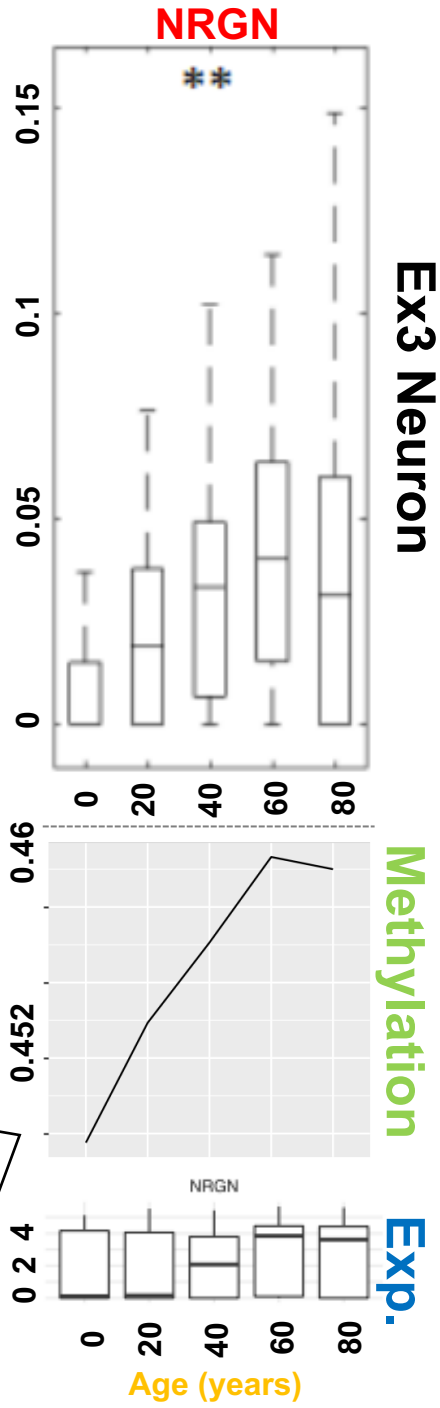
Expression

[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**



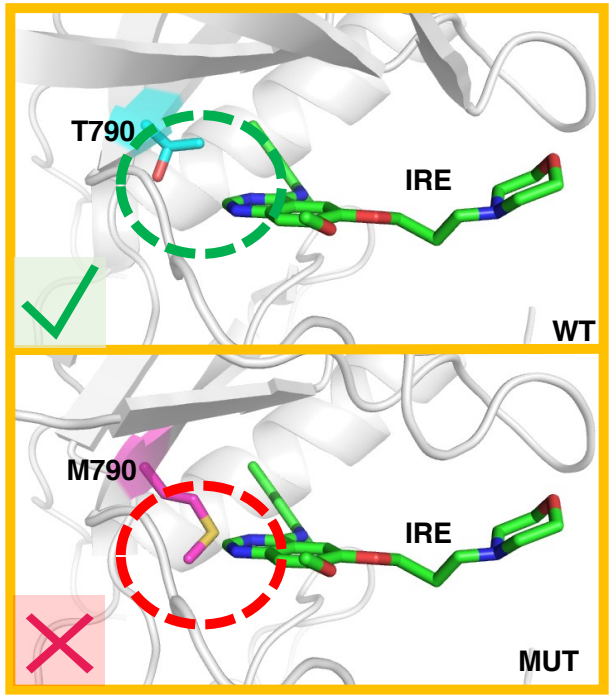


Using population-scale functional genomics to suggest potential neuropsychiatric drug targets & building a hybrid classifier to ascertain differential drug sensitivity

- **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + full developmental time-course
  - Using the changing proportions of cell types (via **single-cell deconvolution**) to account for expression variation across a population, disorders & development
  - 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 + consistently comparing the brain to other organs

- **GenoDock**: Building a predictor for the sensitivity of drug binding to personal SNVs
  - Hybrid classifier connecting **physical modelling with statistical learning**
    - The modeling creates a pseudo gold-standard dataset, which is used to train the stat. classifier
  - **Classifier Results**
    - Independent validation on an expt. validation set
    - Gives higher disruption scores to cancer driver SNVs. Also, illustrates importance of different features (eg GERP).
    - Picks out certain drugs (eg imatinib) as being particularly sensitive to SNVs

An Example of Binding Affinity Change between Protein & Drug Ligand under the Impact of Single Nucleotide Variants (SNV)



human EGFR & gefitinib (IRE)
PDB: 2ity, Chain A, amino acid 790
Modeling and Visualization: Modeller & PyMol

Epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors (EGFR-TKIs) are used in the treatments of non-small cell lung cancer (NSCLC)

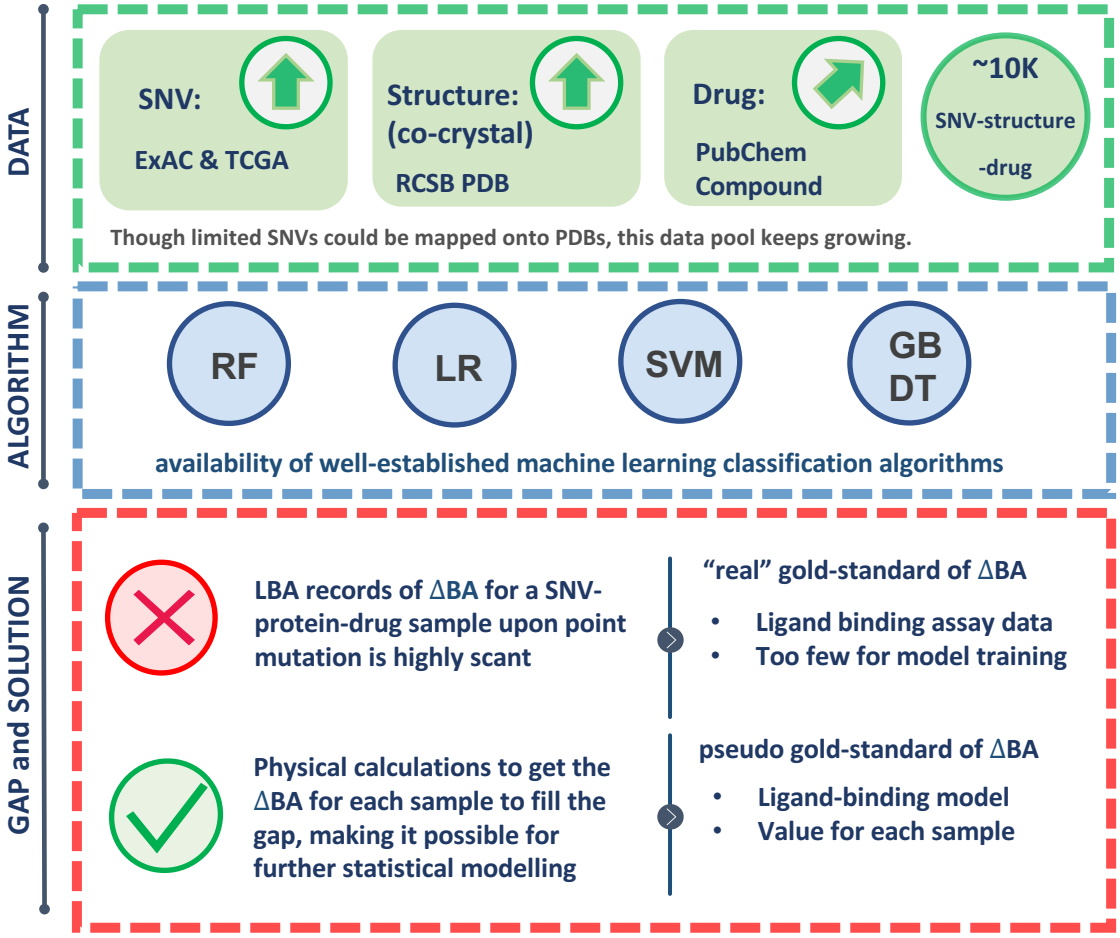
For protein-drug binding upon point mutation,

if ΔBA ≤ 0 → **non-disruptive SNV (ND)**

if ΔBA > 0 → **disruptive SNV (D)**

Is there any method that could predict the effects of SNVs to drug binding (D or ND)?

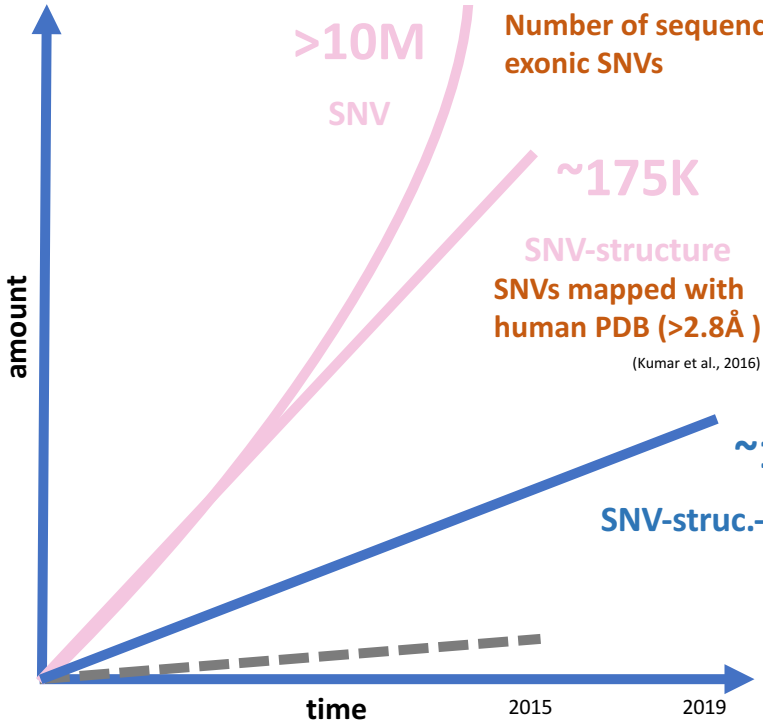
Assessment of feasibility to build a supervised-learning classifier for binding-disruptive SNVs



Wang et al. Structure, 2019

A Hot Topic in Machine Learning is "Hybrid" Model Integrating Physical & Statistical Calculations

The Major Hurdle:
Highly Scant Ligand Binding Assay Data for ΔBA



The Physically-based Data Augmentation Approach:
Leveraging Physical Calculations of ΔBA to Fill the Gap

(Reichstein *et al.*, **Nature**, 2019 & Xie et al., preprint, 2018)

- Expansion of the training dataset for under sampled domains
- Data augmentation is crucial to avoid overfitting

Physically-based Data Augmenting to expand the Δ BA set

ΔBA of each SNV-protein-drug tuple (*pseudo* gold-standard), for parameterizing statistical learning model

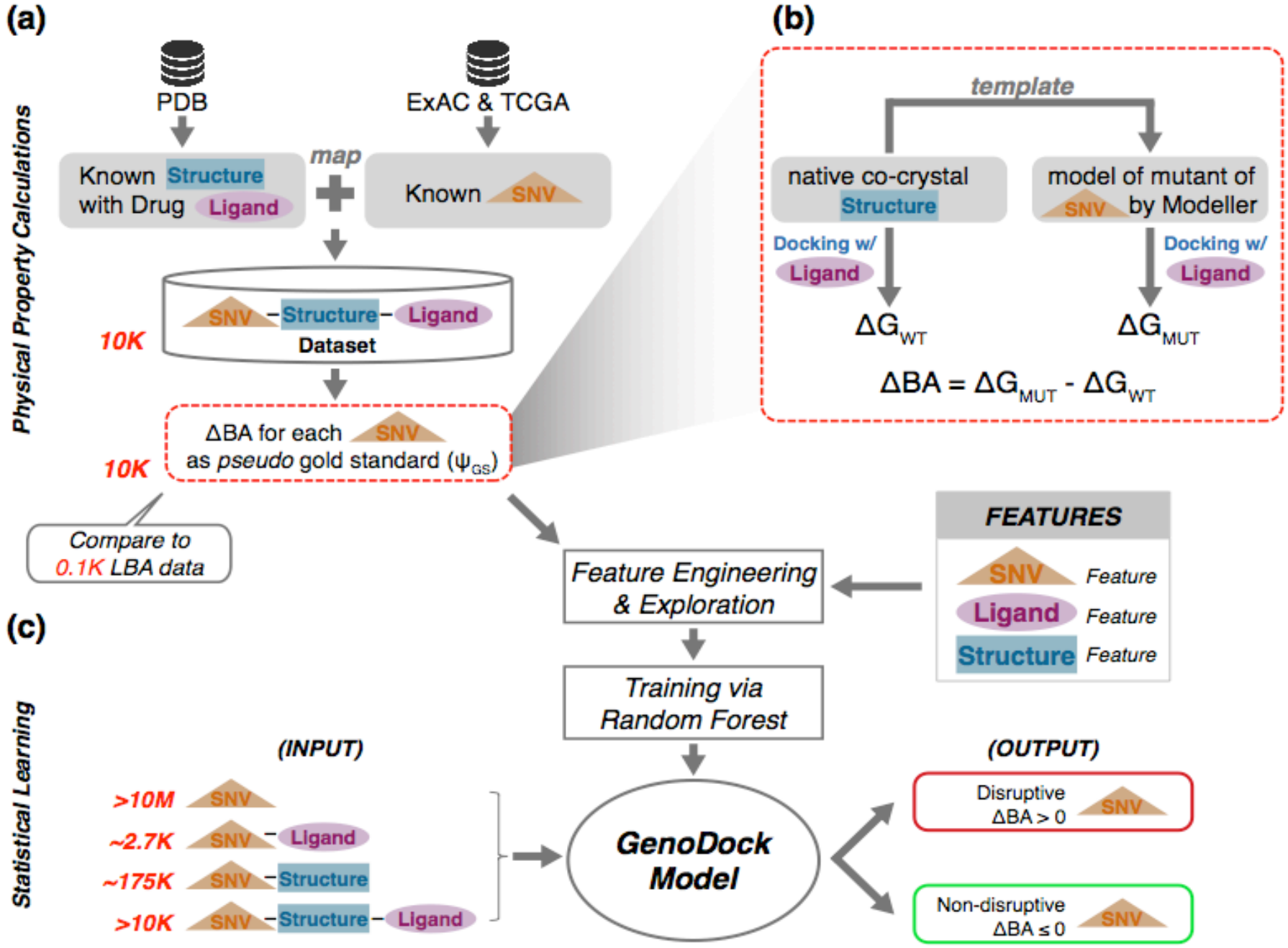
entries in GenoDock
(Wang et al., 2019)

~10K

Docking

experimental ΔBA of human protein from Platinum
(Pires et al., 2015)

Framework for GenoDock: from Dataset Preparation to Model Construction



3 Feature Groups as Predictor, with 4 Application Cases Based on Info Availability

What are features are effective for prioritization of disruptive SNVs?

3 groups of features as predictors

SNV Feature
- Allele Frequency
- SIFT
- PolyPhen-2
- GERP
- Germline/Somatic

Ligand Feature
- Molecular Weight
- H-bond donor
- H-bond acceptor
- Rotatable Bond #
- Polar Surface Area

Structure Feature
- Distance
- Binding Site
- Polarity Change
- Volume Change

Will SNV of interest disrupt protein-ligand binding

4 random forest model trained based on information available

- SNV + Structure + Ligand
- SNV + Structure
- SNV + Ligand
- SNV only

validate the "full feature" case

then, expand the model to 3 more "feature poor" cases

List of Models & Datasets in the Study

Model 1: statistical model (GenoDock)
Model 2: ligand binding model (to calculate ΔBA)

Model	Role	Parameterization	Validation	Description
1	Core Model	Statistical model from Ψ_{GS}	Platinum	Supervised learning model using the pseudo gold-standard set as target feature. The direct validation of this model is to apply the model to an independent, experiment-based validation dataset.
2	Auxillery Model	Physically based	–	A physical-based, previously published computational ligand-docking model to calculate binding affinity change for the pseudo gold standard set.

Dataset	Role	Size	Source	Description
Ψ_{GS}	Trains 1	~10k	Built from 2	Core dataset constructed for training the statistical model. Contains pseudo gold standard set as the target feature.
Platinum	Validates 1	86	Experiment	The human protein subset from Platinum. used as direct validation dataset of our statistical method.

KEY TAKE-AWAY

- **The statistical model and ligand binding model are the two models for this study;**
- **The validation of the statistical model and the assessment of rigor of the ligand binding model are two independent process.**

Wang et al. Structure, 2019

Using population-scale functional genomics to suggest potential neuropsychiatric drug targets & building a hybrid classifier to ascertain differential drug sensitivity

- **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + full developmental time-course
  - Using the changing proportions of cell types (via **single-cell deconvolution**) to account for expression variation across a population, disorders & development
  - 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 + consistently comparing the brain to other organs

- **GenoDock**: Building a predictor for the sensitivity of drug binding to personal SNVs
  - Hybrid classifier connecting **physical modelling with statistical learning**
    - The modeling creates a pseudo gold-standard dataset, which is used to train the stat. classifier
  - **Classifier Results**
    - Independent validation on an expt. validation set
    - Gives higher disruption scores to cancer driver SNVs. Also, illustrates importance of different features (eg GERP).
    - Picks out certain drugs (eg imatinib) as being particularly sensitive to SNVs

The *pseudo* Gold-Standard as Self-Constructed Prediction Target: Physical Calculations for Binding Affinity Score Change (ΔBA)

Dataset

ExAC: 8565 SNV-PDB native-mutant pairs
TCGA: 1718 SNV-PDB native-mutant pairs

Equation

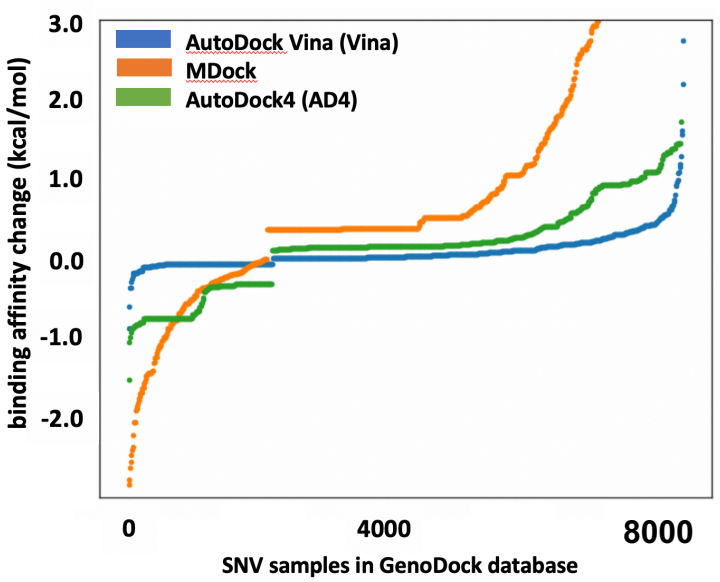
$$\Delta\Delta G(SNV) = \Delta G(SNV) - \Delta G(WT)$$

ΔG(WT): BA of WT protein-drug complex
ΔG(SNV): BA of point mutated protein-drug complex
ΔΔG(SNP): BA change

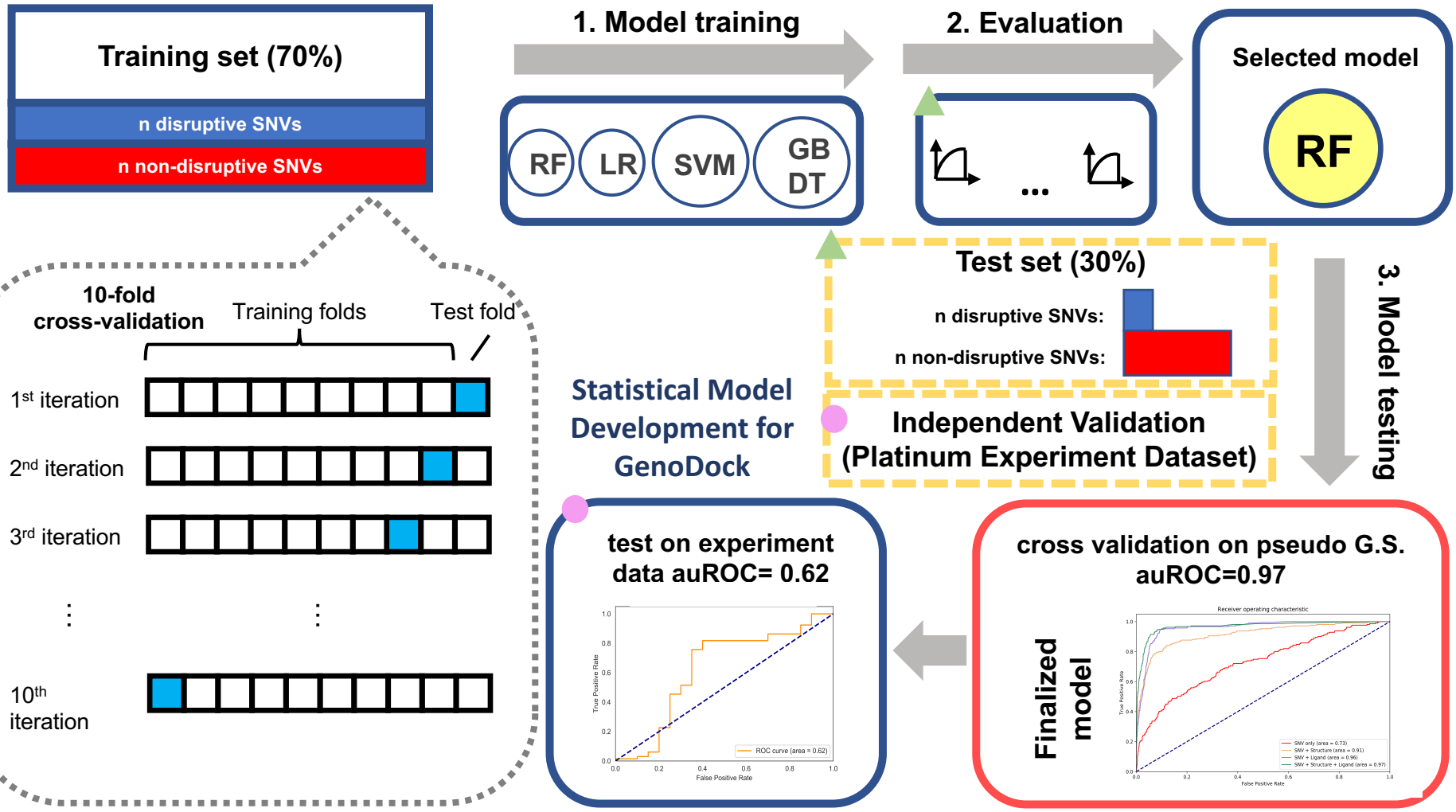
Approach

MDock | AutoDock Vina | AutoDock 4

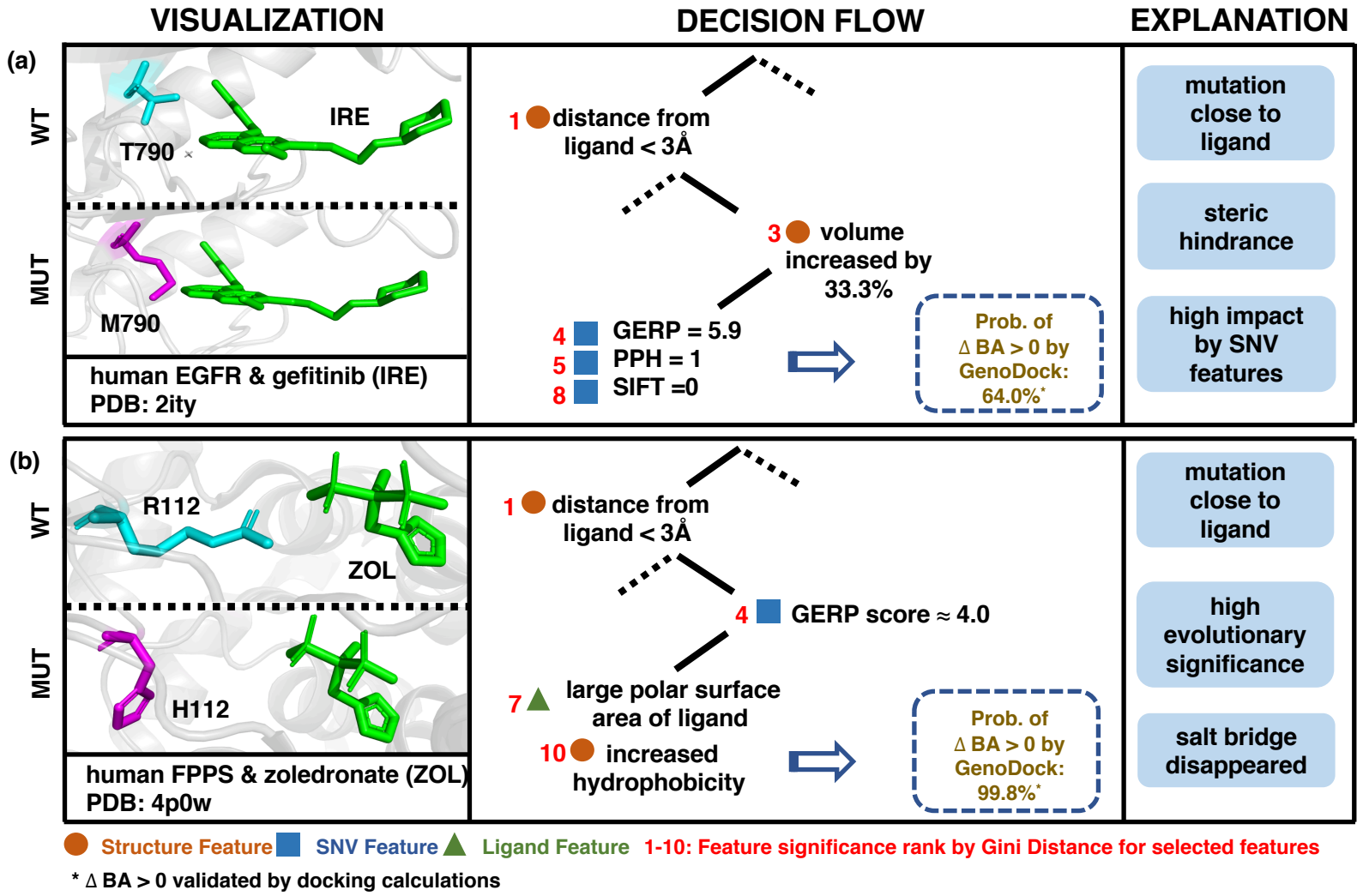
- Pearson Product-Moment Correlation (PMCC) reveals good consistency of different docking calculations
- PMCC (Vina & AD4) = 0.89
- PMCC (Vina & MDock) = 0.94



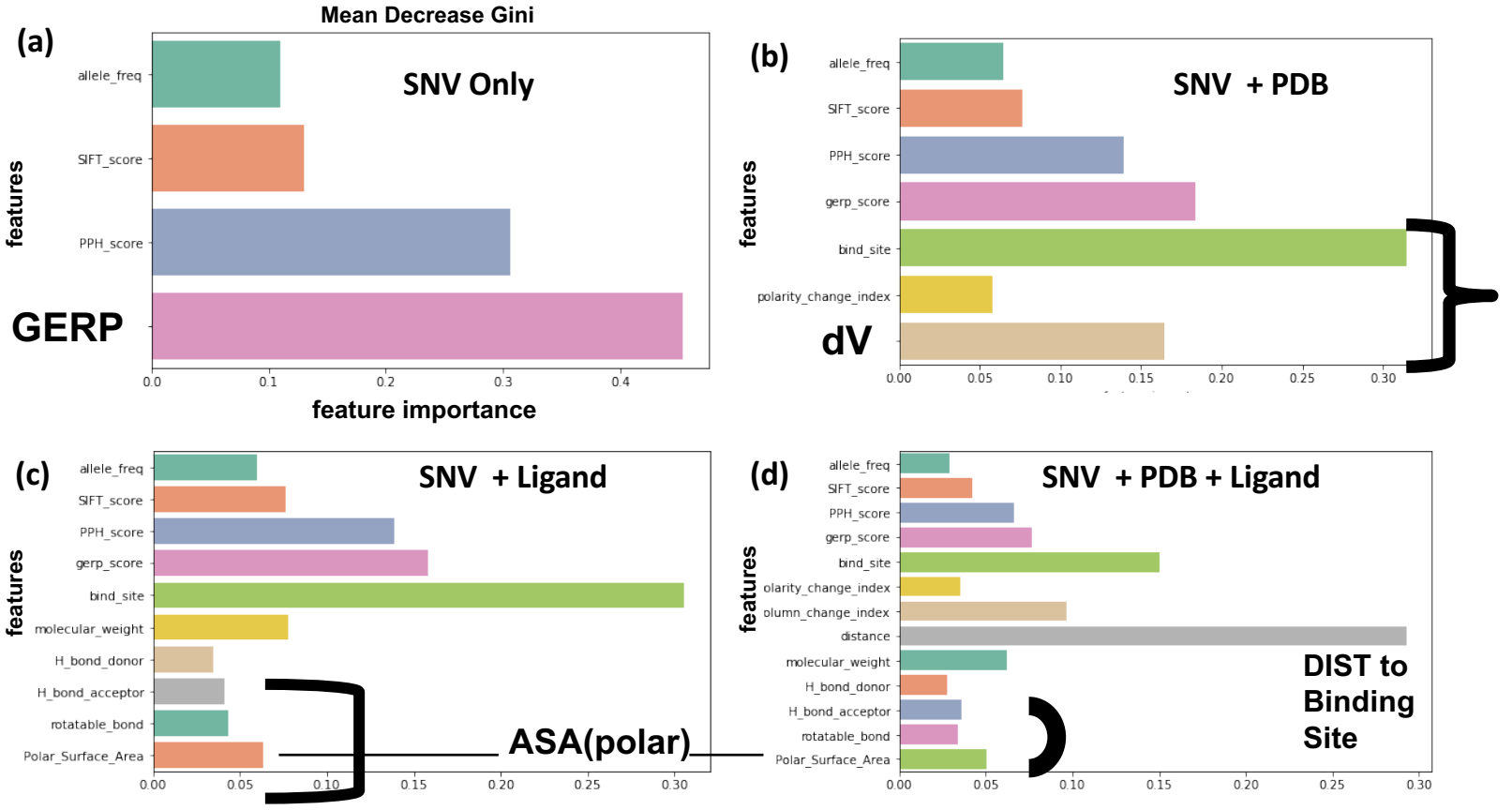
Given the pseudo Gold-Standard, the Workflow for Building the Statistical Model & its Performance in Cross-validation & Independent Testing



Example of the Output of the Classifier: GenoDock Helps Characterize Known & Unknown SNVs that Disrupt Protein-Ligand Binding

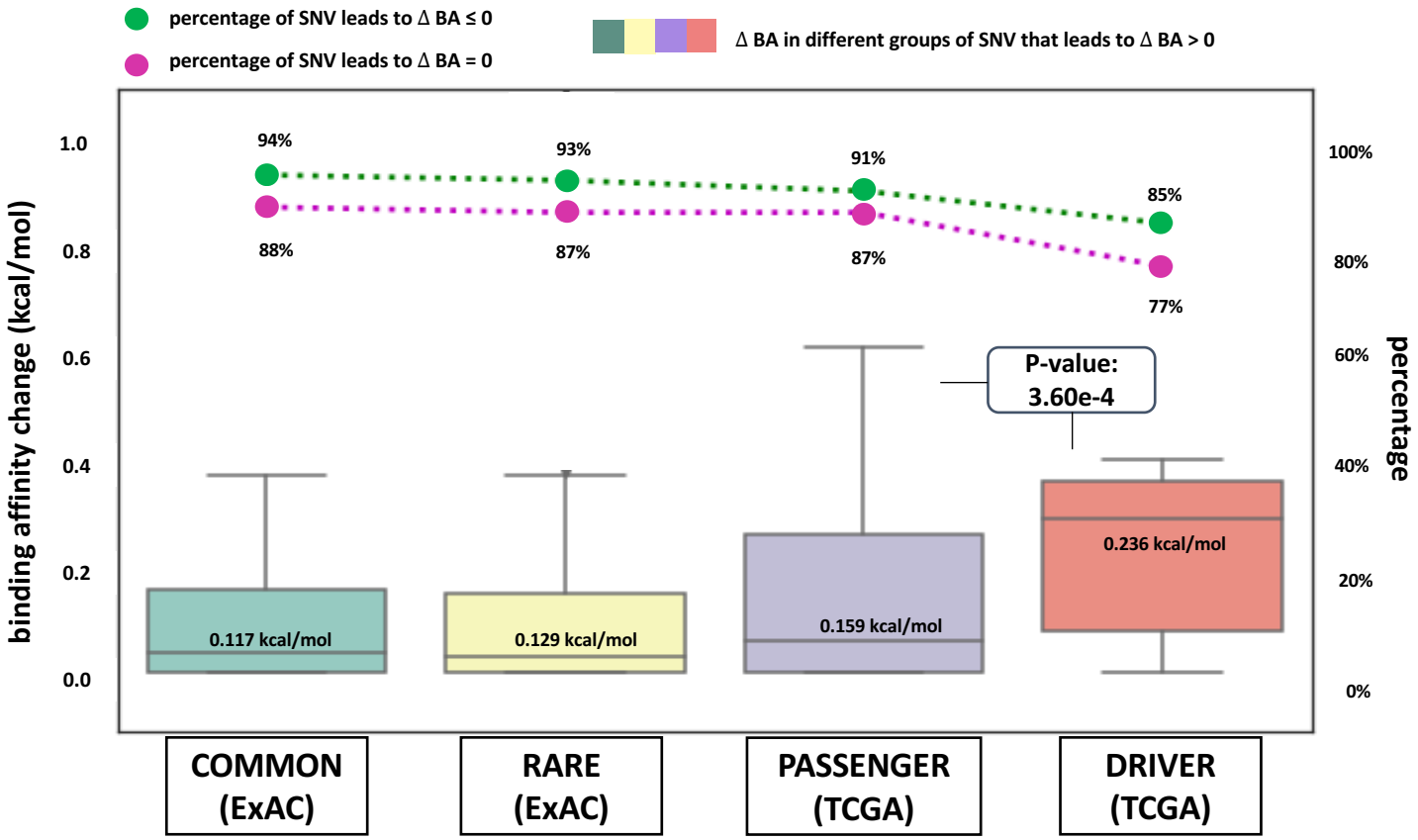


Gini Distance for Relative Feature Importance in 4 Models



Wang et al. Structure, 2019

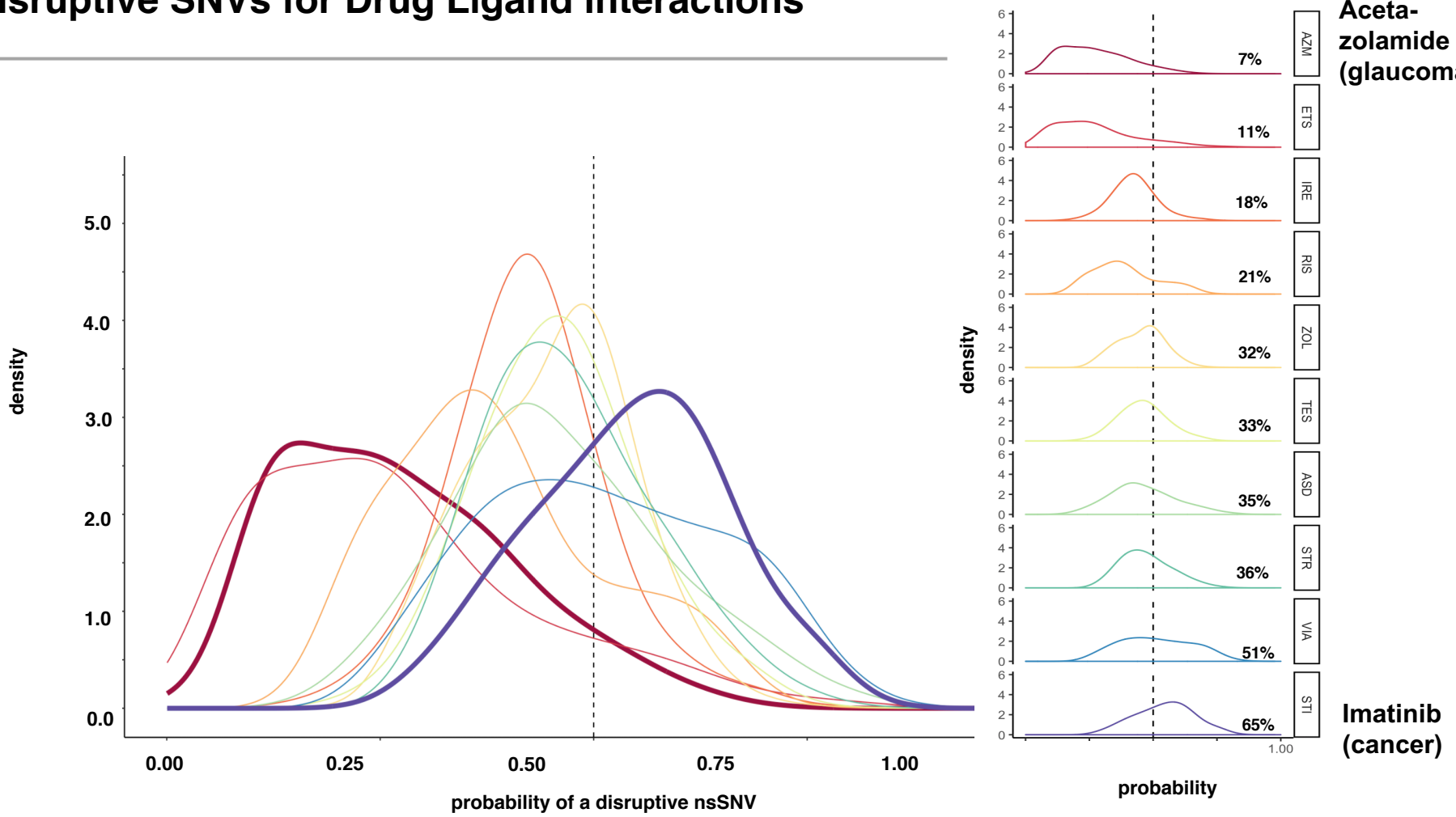
Boxplot of Overall Ligand Binding Affinity Changes for Different Types of SNVs in GenoDock



The more an SNV is considered disease-associated, the greater chance that this SNV would destabilize binding affinity of the protein and drug ligand.

Wang et al. Structure, 2019

Application of GenoDock to large-scale screening of disruptive SNVs for Drug Ligand interactions



Wang et al. Structure, 2019

Using population-scale functional genomics to suggest potential neuropsychiatric drug targets & building a hybrid classifier to ascertain differential drug sensitivity

- **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + full developmental time-course
  - Using the changing proportions of cell types (via **single-cell deconvolution**) to account for expression variation across a population, disorders & development
  - 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 + consistently comparing the brain to other organs

- **GenoDock**: Building a predictor for the sensitivity of drug binding to personal SNVs
  - Hybrid classifier connecting **physical modelling with statistical learning**
    - The modeling creates a pseudo gold-standard dataset, which is used to train the stat. classifier
  - **Classifier Results**
    - Independent validation on an expt. validation set
    - Gives higher disruption scores to cancer driver SNVs. Also, illustrates importance of different features (eg GERP).
    - Picks out certain drugs (eg imatinib) as being particularly sensitive to SNVs

Using population-scale functional genomics to suggest potential neuropsychiatric drug targets & building a hybrid classifier to ascertain differential drug sensitivity

- **PsychENCODE:** Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + full developmental time-course
  - Using the changing proportions of cell types (via **single-cell deconvolution**) to account for expression variation across a population, disorders & development
  - 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 + consistently comparing the brain to other organs
- **GenoDock:** Building a predictor for the sensitivity of drug binding to personal SNVs
  - Hybrid classifier connecting **physical modelling with statistical learning**
    - The modeling creates a pseudo gold-standard dataset, which is used to train the stat. classifier
  - **Classifier Results**
    - Independent validation on an expt. validation set
    - Gives higher disruption scores to cancer driver SNVs. Also, illustrates importance of different features (eg GERP).
    - Picks out certain drugs (eg imatinib) as being particularly sensitive to SNVs

PsychENCODE Acknowledgment

- Geetha Senthil
- Lora Bingaman
- David Panchision
- Alexander Arguello
- Thomas Lehner

"Adult Capstone" Team – 1 of 3 capstones

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

Dedicated to **Pamela Sklar**

Resource.psychencode.org

The PsychENCODE Consortium: Allison E Ashley-Koch, Duke University; Gregory E Crawford, Duke University; Melanie E Garrett, Duke University; Lingyun Song, Duke University; Alexias Safi, Duke University; Graham D Johnson, Duke University; Gregory A Wray, Duke University; Timothy E Reddy, Duke University; Fernando S Goes, Johns Hopkins University; Peter Zandi, Johns Hopkins University; Julien Bryois, Karolinska Institutet; Andrew E Jaffe, Lieber Institute for Brain Development; Amanda J Price, Lieber Institute for Brain Development; Nikolay A Ivanov, Lieber Institute for Brain Development; Leonardo Collado-Torres, Lieber Institute for Brain Development; Thomas M Hyde, Lieber Institute for Brain Development; Emily E Burke, Lieber Institute for Brain Development; Joel E Kleiman, Lieber Institute for Brain Development; Ran Tao, Lieber Institute for Brain Development; Joo Heon Shin, Lieber Institute for Brain Development; Schahram Akbarian, Icahn School of Medicine at Mount Sinai; Kiran Girdhar, Icahn School of Medicine at Mount Sinai; Yan Jiang, Icahn School of Medicine at Mount Sinai; Marija Kundakovic, Icahn School of Medicine at Mount Sinai; Leanne Brown, Icahn School of Medicine at Mount Sinai; Bibi S Kassim, Icahn School of Medicine at Mount Sinai; Royce B Park, Icahn School of Medicine at Mount Sinai; Jennifer R Wiseman, Icahn School of Medicine at Mount Sinai; Elizabeth Zharovsky, Icahn School of Medicine at Mount Sinai; Rivka Jacobov, Icahn School of Medicine at Mount Sinai; Olivia Devillers, Icahn School of Medicine at Mount Sinai; Elie Flatow, Icahn School of Medicine at Mount Sinai; Gabriel E Hoffman, Icahn School of Medicine at Mount Sinai; Barbara K Lipska, Human Brain Collection Core, National Institutes of Health, Bethesda, MD; David A Lewis, University of Pittsburgh; Vahram Haroutunian, Icahn School of Medicine at Mount Sinai and James J Peters VA Medical Center; Chang-Gyu Hahn, University of Pennsylvania; Alexander W Charney, Mount Sinai; Stella Dracheva, Mount Sinai; Alexey Kozlenkov, Mount Sinai; Judson Belmont, Icahn School of Medicine at Mount Sinai; Diane DelValle, Icahn School of Medicine at Mount Sinai; Nancy Francoeur, Icahn School of Medicine at Mount Sinai; Evi Hadjimichael, Icahn School of Medicine at Mount Sinai; Dalila Pinto, Icahn School of Medicine at Mount Sinai; Harm van Bakel, Icahn School of Medicine at Mount Sinai; Panos Roussos, Mount Sinai; John F Fullard, Mount Sinai; Jaroslav Bendl, Mount Sinai; Mads E Hauberg, Mount Sinai; Lara M Mangravite, Sage Bionetworks; Mette A Peters, Sage Bionetworks; Yooree Chae, Sage Bionetworks; Junmin Peng, St. Jude Children's Hospital; Mingming Niu, St. Jude Children's Hospital; Xusheng Wang, St. Jude Children's Hospital; Maree J Webster, Stanley Medical Research Institute; Thomas G Beach, Banner Sun Health Research Institute; Chao Chen, Central South University; Yi Jiang, Central South University; Rujia Dai, Central South University; Annie W Shieh, SUNY Upstate Medical University; Chunyu Liu, SUNY Upstate Medical University; Kay S. Grennan, SUNY Upstate Medical University; Yan Xia, SUNY Upstate Medical University/Central South University; Ramu Vadukapuram, SUNY Upstate Medical University; Yongjun Wang, Central South University; Dominic Fitzgerald, The University of Chicago; Lijun Cheng, The University of Chicago; Miguel Brown, The University of Chicago; Mimi Brown, The University of Chicago; Tonya Brunetti, The University of Chicago; Thomas Goodman, The University of Chicago; Majd Alsayed, The University of Chicago; Michael J Gandal, University of California, Los Angeles; Daniel H Geschwind, University of California, Los Angeles; Hyejung Won, University of California, Los Angeles; Damon Polioudakis, University of California, Los Angeles; Brie Wamsley, University of California, Los Angeles; Jiani Yin, University of California, Los Angeles; Tarik Hadzic, University of California, Los Angeles; Luis De La Torre Ubieta, UCLA; Vivek Swarup, University of California, Los Angeles; Stephan J Sanders, University of California, San Francisco; Matthew W State, University of California, San Francisco; Donna M Werling, University of California, San Francisco; Joon-Yong An, University of California, San Francisco; Brooke Sheppard, University of California, San Francisco; A Jeremy Willsey, University of California, San Francisco; Kevin P White, The University of Chicago; Mohana Ray, The University of Chicago; Gina Giase, SUNY Upstate Medical University; Amira Kefi, University of Illinois at Chicago; Eugenio Mattei, University of Massachusetts Medical School; Michael Purcaro, University of Massachusetts Medical School; Zhiping Weng, University of Massachusetts Medical School; Jill Moore, University of Massachusetts Medical School; Henry Pratt, University of Massachusetts Medical School; Jack Huey, University of Massachusetts Medical School; Tyler Borrman, University of Massachusetts Medical School; Patrick F Sullivan, University of North Carolina - Chapel Hill; Paola Giusti-Rodriguez, University of North Carolina - Chapel Hill; Yunjung Kim, University of North Carolina - Chapel Hill; Patrick Sullivan, University of North Carolina - Chapel Hill; Jin Szatkiewicz, University of North Carolina - Chapel Hill; Suhn Kyong Rhie, University of Southern California; Christoper Armoskus, University of Southern California; Adrian Camarena, University of Southern California; Peggy J Farnham, University of Southern California; Valeria N Spitsyna, University of Southern California; Heather Witt, University of Southern California; Shannon Schreiner, University of Southern California; Oleg V Evgrafov, SUNY Downstate Medical Center; James A Knowles, SUNY Downstate Medical Center; Mark Gerstein, Yale University; Shuang Liu, Yale University; Daifeng Wang, Stony Brook University; Fabio C. P. Navarro, Yale University; Jonathan Warrell, Yale University; Declan Clarke, Yale University; Prashant S. Emani, Yale University; Mengting Gu, Yale University; Xu Shi, Yale University; Min Xu, Yale University; Yucheng T. Yang, Yale University; Robert R. Kitchen, Yale University; Gamze Gürsoy, Yale University; Jing Zhang, Yale University; Becky C Carlyle, Yale University; Angus C Nairn, Yale University; Mingfeng Li, Yale University; Sirisha Pochareddy, Yale University; Nenad Sestan, Yale University; Mario Skarica, Yale University; Zhen Li, Yale University; Andre M.M. Sousa, Yale University; Gabriel Santpere, Yale University; Jinmyung Choi, Yale University; Ying Zhu, Yale University; Tianliuyun Gao, Yale University; Daniel J Miller, Yale University; Adriana Cherskov, Yale University; Mo Yang, Yale University; Anahita Amiri, Yale University; Gianfilippo Coppola, Yale University; Jessica Mariani, Yale University; Soraya Scuderi, Yale University; Anna Szekely, Yale University; Flora M Vaccarino, Yale University; Feinan Wu, Yale University; Sherman Weissman, Yale University; Tanmoy Roychowdhury, Mayo Clinic Rochester; Alexej Abyzov, Mayo Clinic Rochester;.

Developmental Capstone

- **M Li, G Santpere, Y Imamura Kawasawa, OV Evgrafov, FO Gulden, S Pochareddy, SM Sunkin, Z Li, Y Shin,**

Y Zhu, AMM Sousa, DM Werling, RR Kitchen, HJ Kang, M Pletikos, J Choi, S Muchnik, X Xu, D Wang, B Lorente-Galdos, S Liu, P Giusti-Rodriguez, H Won, CA de Leeuw, AF Pardinas, BrainSpan Consortium,
PsychENCODE Consortium, PsychENCODE Developmental Subgroup,
M Hu, F Jin, Y Li, MJ Owen, MC O'Donovan, JTR Walters, D Posthuma, MA Reimers, P Levitt, DR Weinberger, TM Hyde, JE Kleinman, DH Geschwind, MJ Hawrylycz, MW State, SJ Sanders, PF Sullivan,

ES Lein, JA Knowles, N Sestan

psychencode.org

See

JOBS.gersteinlab.org

Hiring Postdocs

GenoDock.molmovdb.org

B Wang, C Yan,

S Lou, P Emani, B Li, M Xu, X Kong, W Meyerson, Y Yang, D Lee

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