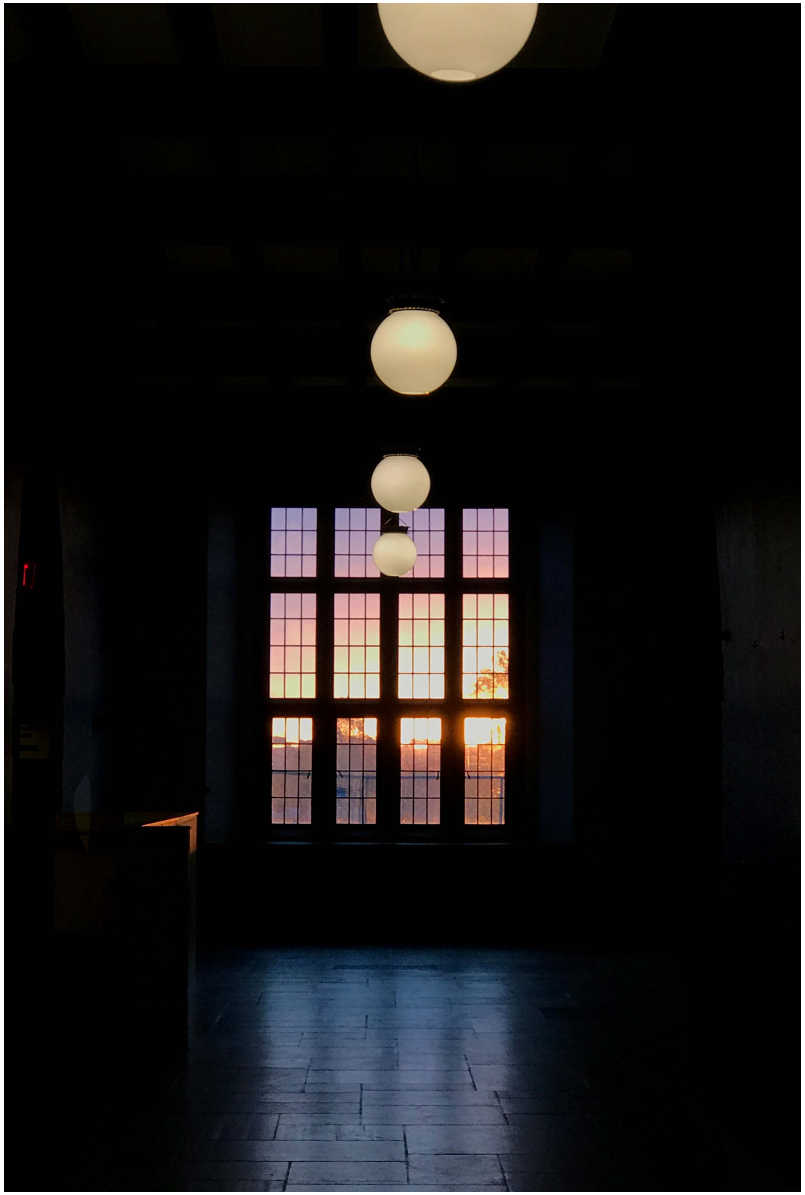


Personal Genomics

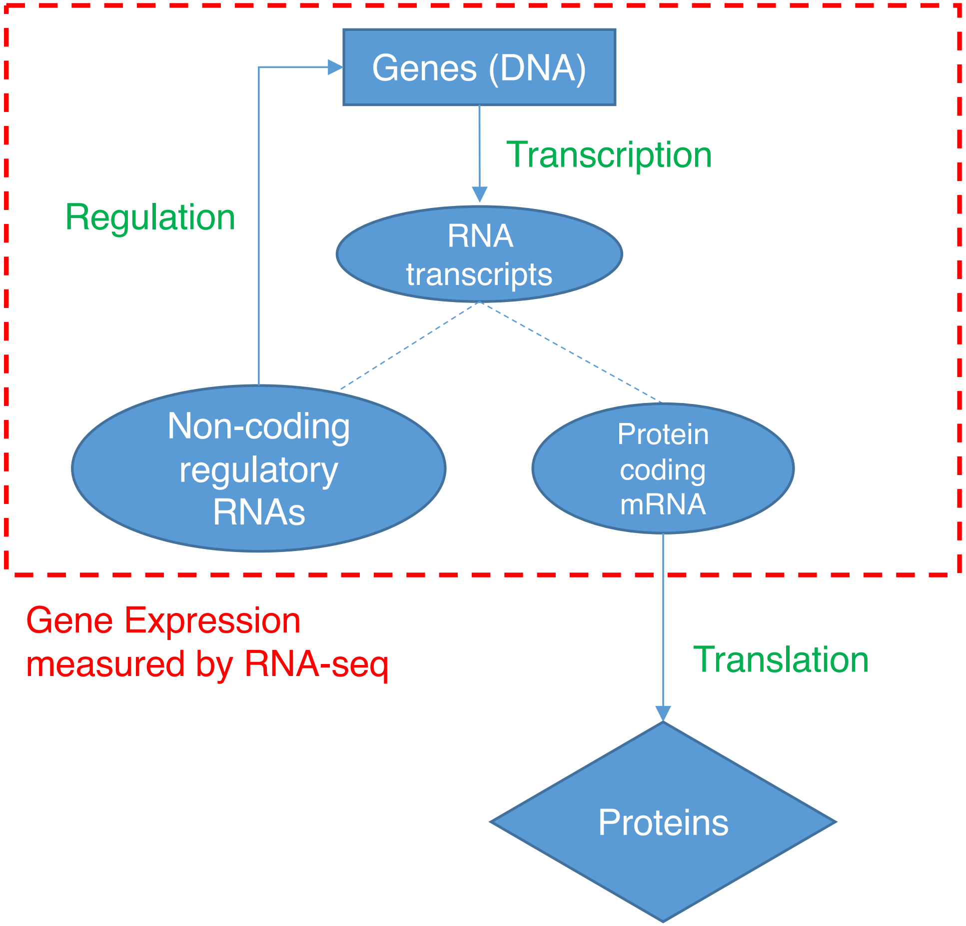
& Data Science:

Using population-scale functional genomics to understand neuropsychiatric disease & interpreting the data exhaust from this activity

Mark Gerstein
Yale

Slides freely downloadable from
Lectures.GersteinLab.org
& “tweetable” (via **@markgerstein**).
See last slide for more info.

Transcriptome = Gene Activity of All Genes in the Genome, usually quantified by RNA-seq



Genes (DNA)

Transcription

Regulation

RNA transcripts

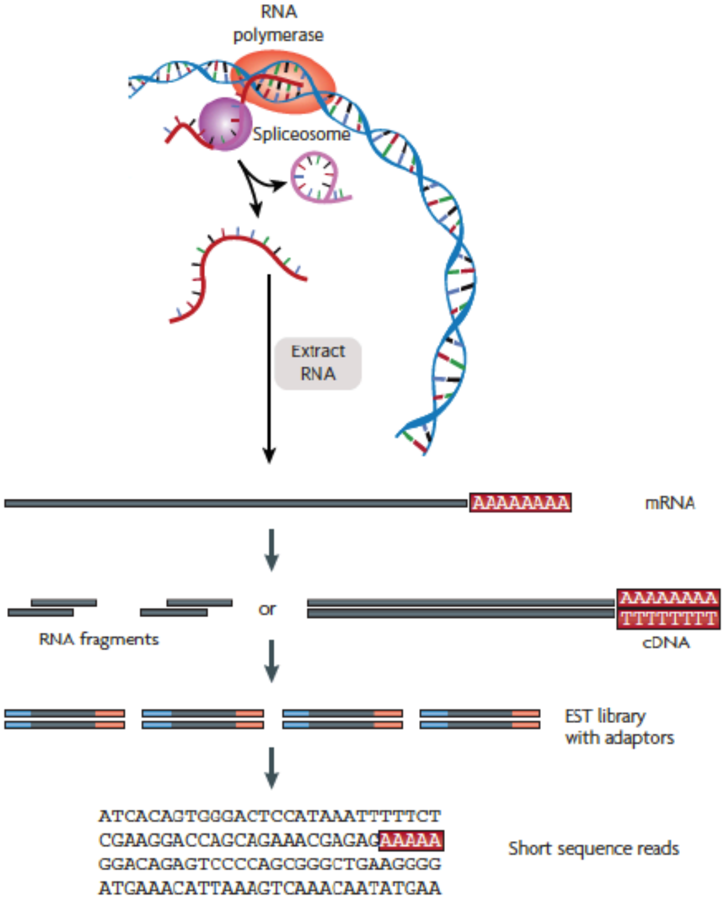
Non-coding regulatory RNAs

Protein coding mRNA

Translation

Proteins

Gene Expression measured by RNA-seq



[NATURE 459: 927; NAT. REV. GEN. 10: 57]

Expression of genes is quantified by transcription:
RNA-Seq measures mRNA transcript amounts

RNA-Seq Overview

Successive steps of Data Reduction

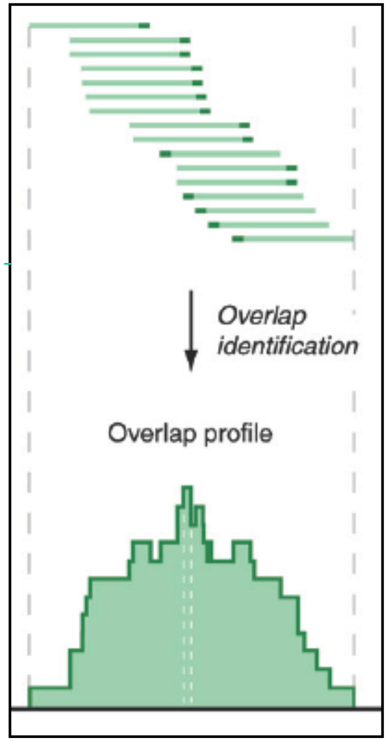
```
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
TACCGATCGAGTCGACTGTAAACGTAGGCATA
ATTCTGACTGGTGTCATGCTGATGTACTTAAA
```

Fastq sequence files
~5-10 GB

Index-building + Alignment to reference genome

BAM files
~1-2-fold reduction

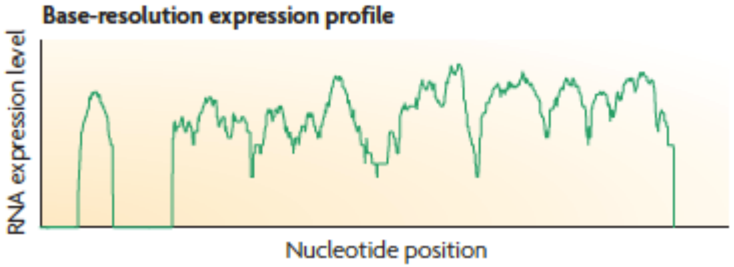
Conversion to signal track by overlapping reads



Reads => Signal

BigWig files
~25-fold reduction

Mapping to genes



Quantitative information from RNA-seq signal:
average signals at exon level (RPKMs)

Gene/Transcript expression matrix
~20-fold reduction

[*NAT. REV.* 10: 57; *PLOS CB* 4:e1000158; *PNAS* 4:107: 5254]

Activity Patterns

- RNA Seq. gives rise to activity patterns of genes & regions in the genome

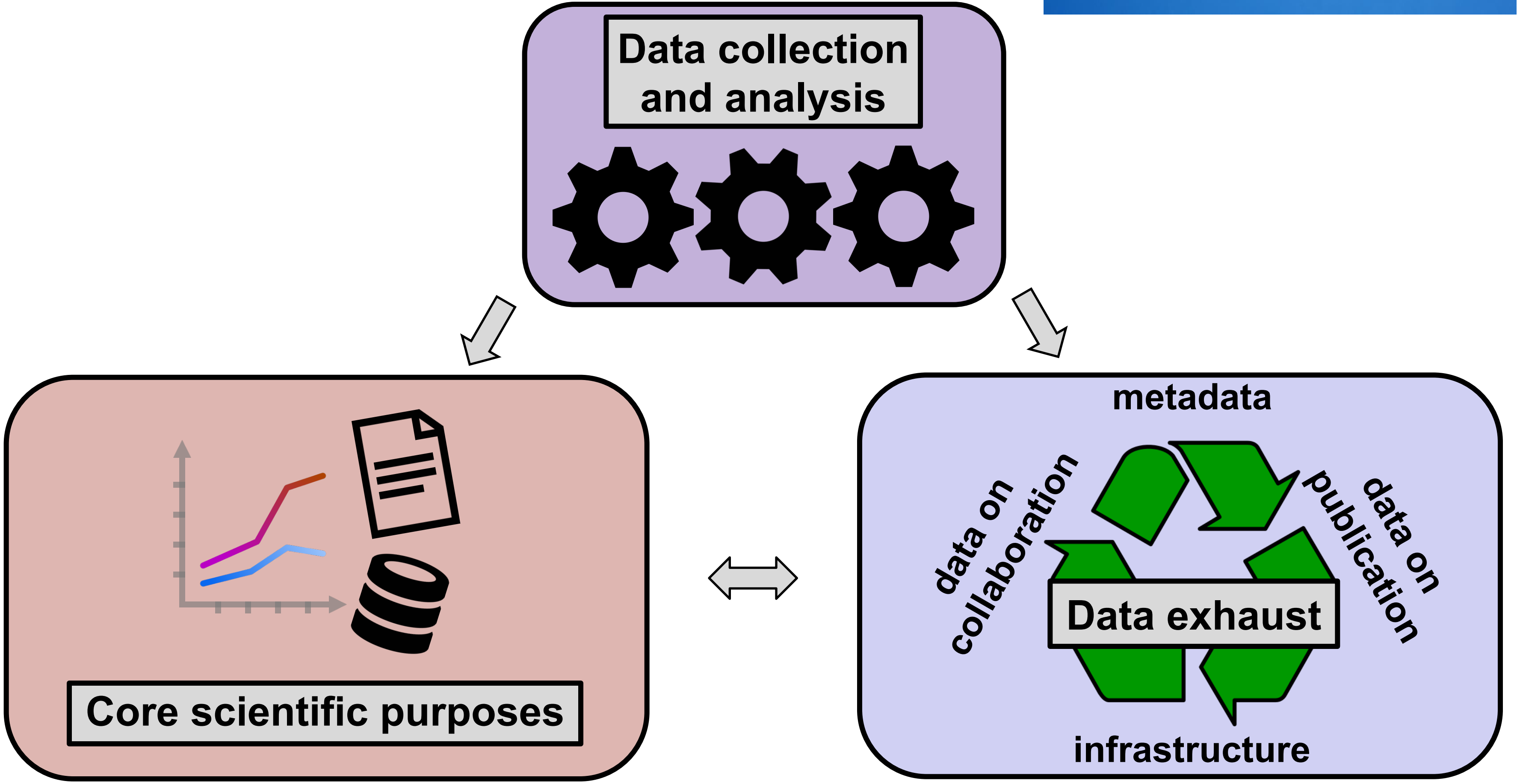
Some Core Science Qs Addressed by RNA-seq

- Gene activity as a function of:
  - **Developmental** stage: basic patterns of co-active genes across development
  - **Cell-type** & Tissue: relationship to specialized functions
  - **Evolutionary** relationships: behavior preserved across a wide range of organisms; patterns in model organisms in relation to those in humans
  - **Individual**, across the human population
  - **Disease** phenotypes: disruption of patterns in disease

- Some overarching Qs:
  Are there core patterns of gene activity ?
  How do they vary across individual ?
  Are they disrupted by disease?

Data Exhaust

- Creative use of data is key to data science!
- Data exhaust = exploitable byproducts of big data collection and analysis


Data collection and analysis

Core scientific purposes

Data exhaust

[photos: wikipedia/wikimedia]

Using population-scale functional genomics to understand neuropsychiatric disease & interpreting the data exhaust from this activity

- *[Core]* **PsychENCODE:** Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + dev. 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 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 disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as targets.

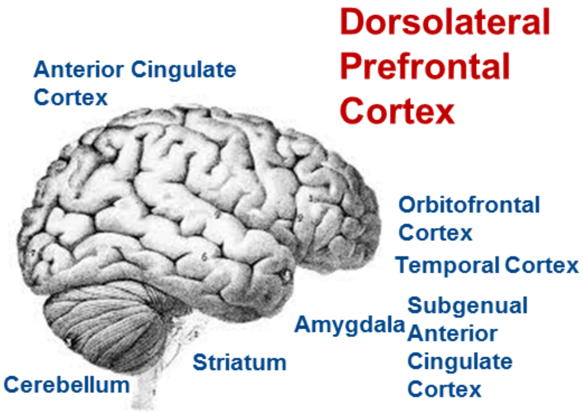
- *[Exhaust]* **Other uses** for the resource
  - Highlighting aging related genes + consistently comparing the brain to other organs
- *[Exhaust]* **Genomic Privacy**
  - The **Dilemma**
    - The genome as fundamental, inherited info that's very private v. need for large-scale mining for med. research
    - 2-sided nature of RNA-seq presents tricky privacy issues
  - **eQTLs**: Quantifying & removing variant info from expression levels with ICI & predictability. Instantiating a practical linking attack with noisy quasi-identifiers
  - **Signal Profiles**: Manifest appreciable leakage from large & small deletions. Linking attacks possible but additional complication of SV discovery in addition to genotyping

Using population-scale functional genomics to understand neuropsychiatric disease & interpreting the data exhaust from this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + dev. 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 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 disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as targets.
- *[Exhaust]* **Other uses** for the resource
  - Highlighting aging related genes + consistently comparing the brain to other organs
- *[Exhaust]* **Genomic Privacy**
  - The **Dilemma**
    - The genome as fundamental, inherited info that's very private v. need for large-scale mining for med. research
    - 2-sided nature of RNA-seq presents tricky privacy issues
  - **eQTLs**: Quantifying & removing variant info from expression levels with ICI & predictability. Instantiating a practical linking attack with noisy quasi-identifiers
  - **Signal Profiles**: Manifest appreciable leakage from large & small deletions. Linking attacks possible but additional complication of SV discovery in addition to genotyping

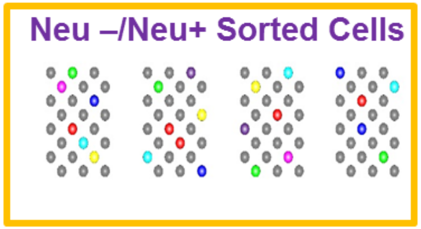
Sample Sources: >2,500 brains

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



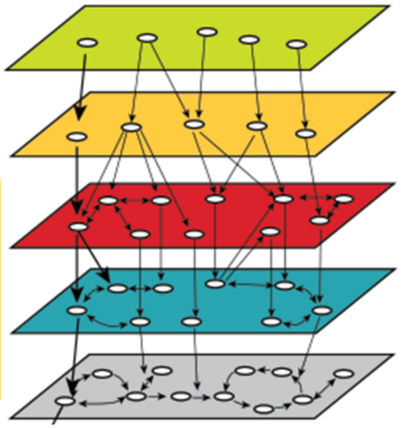
Bulk Tissue

Neu –/Neu+ Sorted Cells



Single Cell

Multi-omic analysis

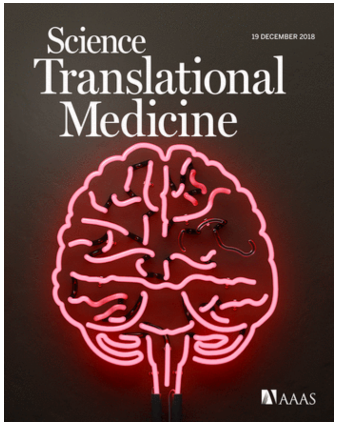


Genome:
WGS, genotype

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

Transcriptome:
RNA-seq, lncRNAseq,

Proteome:
MWP, LC-MS/MS



PsychENCODE

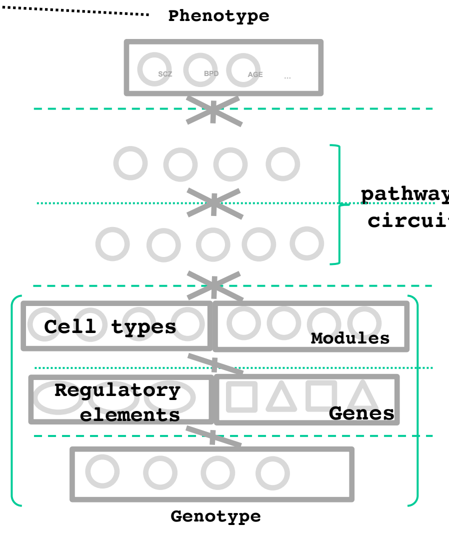
'18 rollout in Science

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

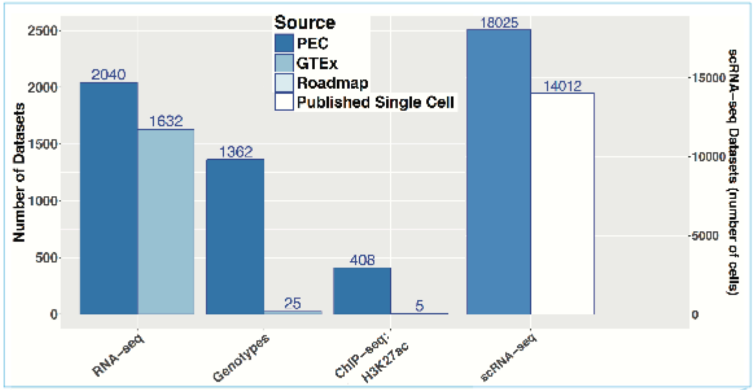
Many psychiatric conditions are highly heritable
- Schizophrenia: up to 80%

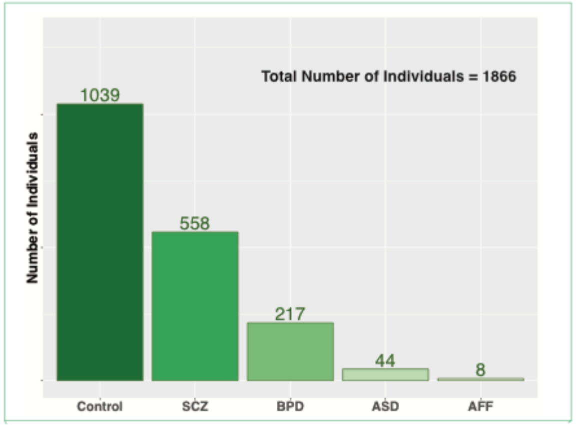
But we don’t understand basic molecular mechanisms underpinning this association
- (in contrast to many other diseases such as cancer & heart disease)

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

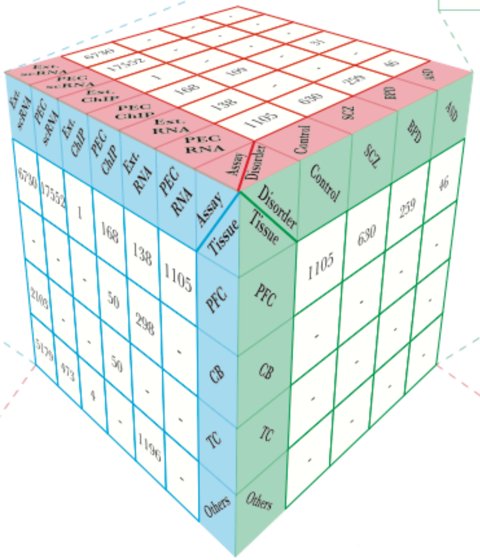
Collecting functional genomic datasets for the adult brain

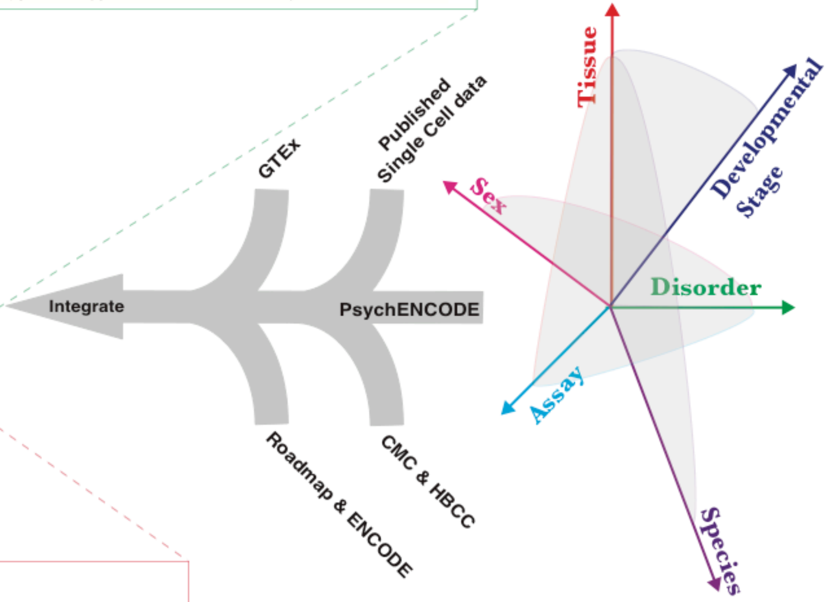
from PsychENCODE, other large consortia & single cell studies

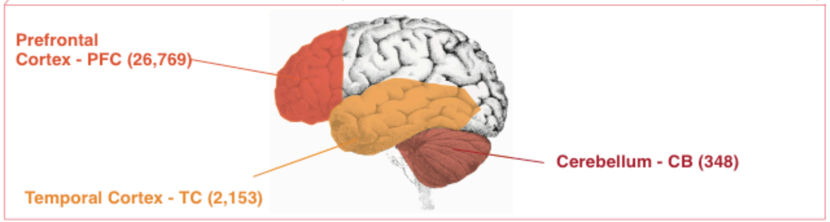




1866 Individuals
~3.7K bulk RNA-seq
~32K single-cells



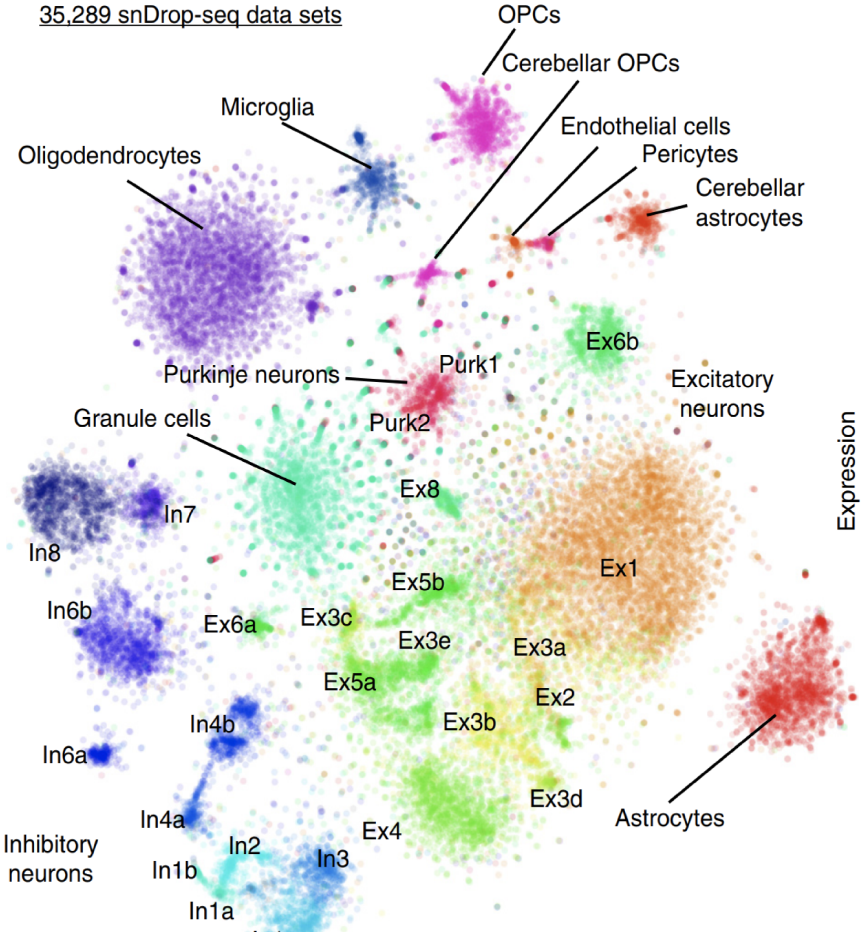




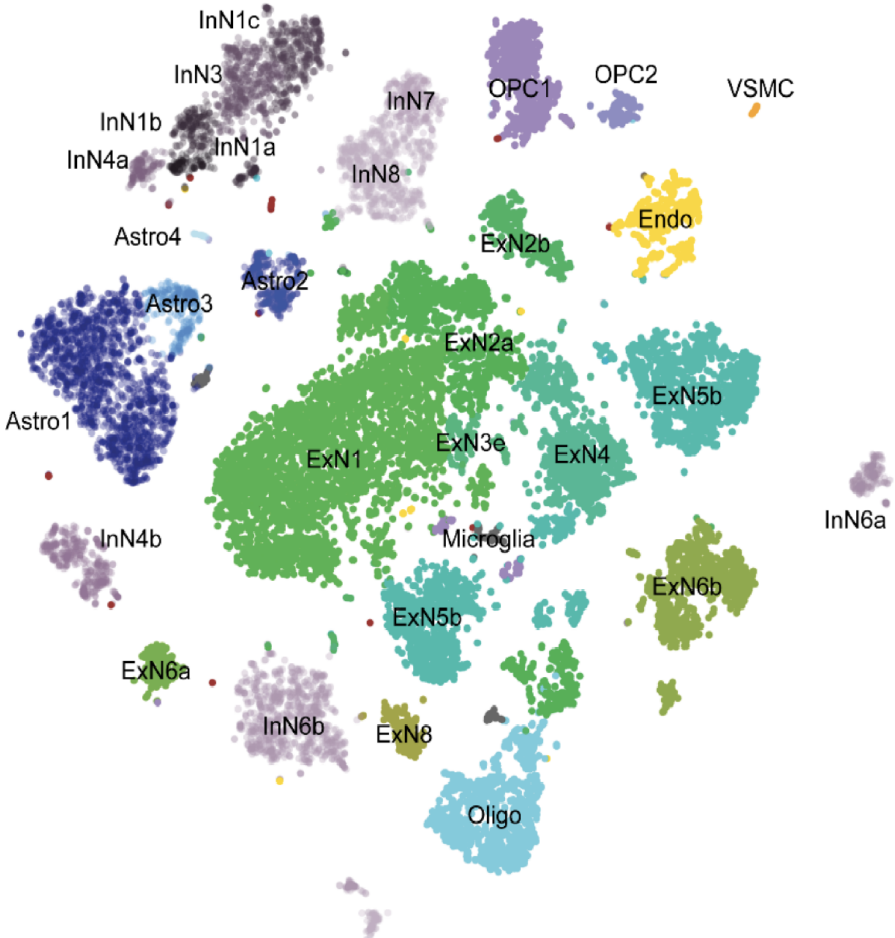
Merging & Clustering Single Cell Data Sets

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

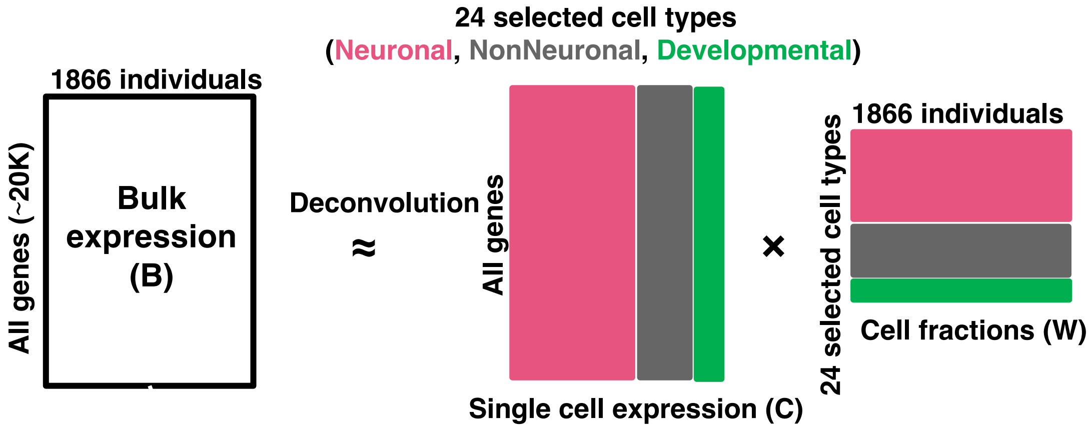


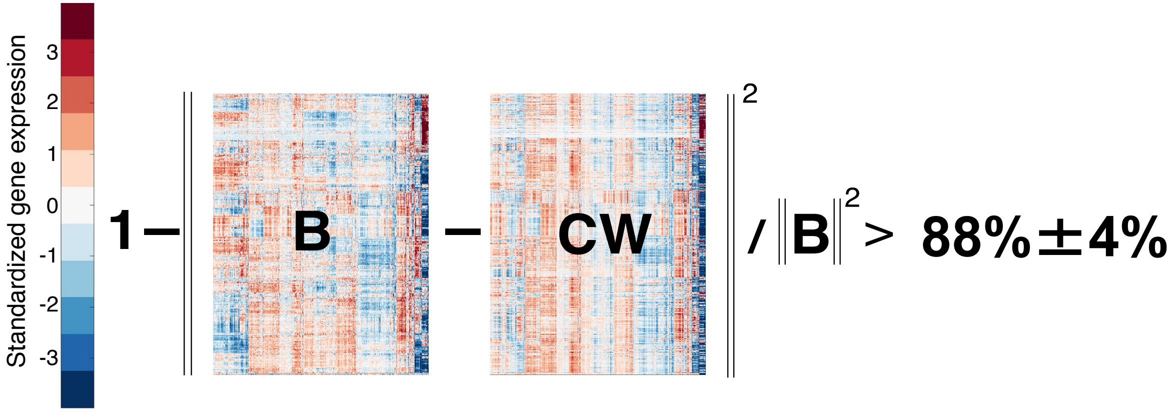
PEC adult data
[Li et al. (‘18), Science. Wang et al. (‘18). Science]

Single-cell deconvolution

Step 1:

Supervised learning to estimate cell fractions



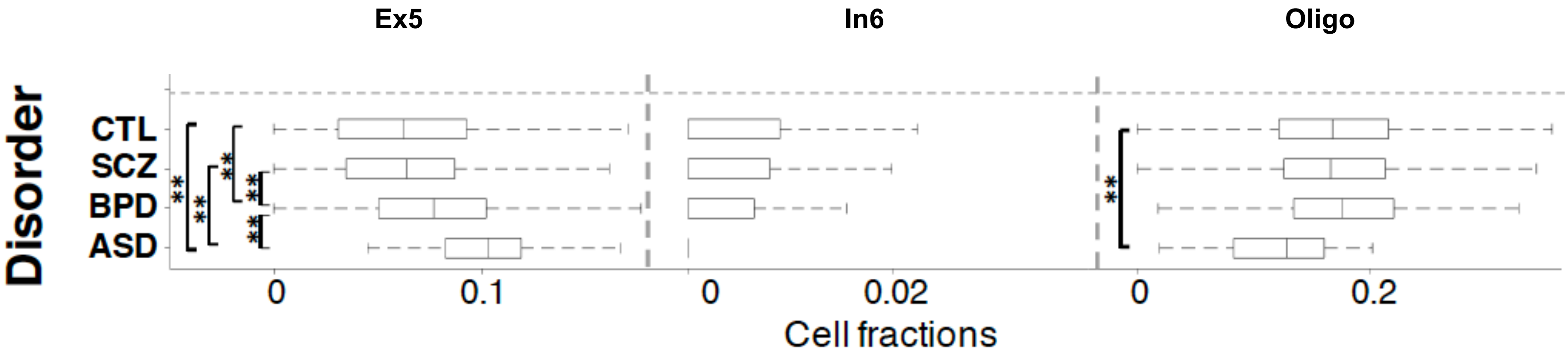


$$1 - \left\| B - CW \right\|^2 / \left\| B \right\|^2 > 88\% \pm 4\%$$

Individual and cross-population reconstruction accuracy via deconvolution

[Wang et al. ('18) Science]

Different neuronal & glial cell fractions across disorders

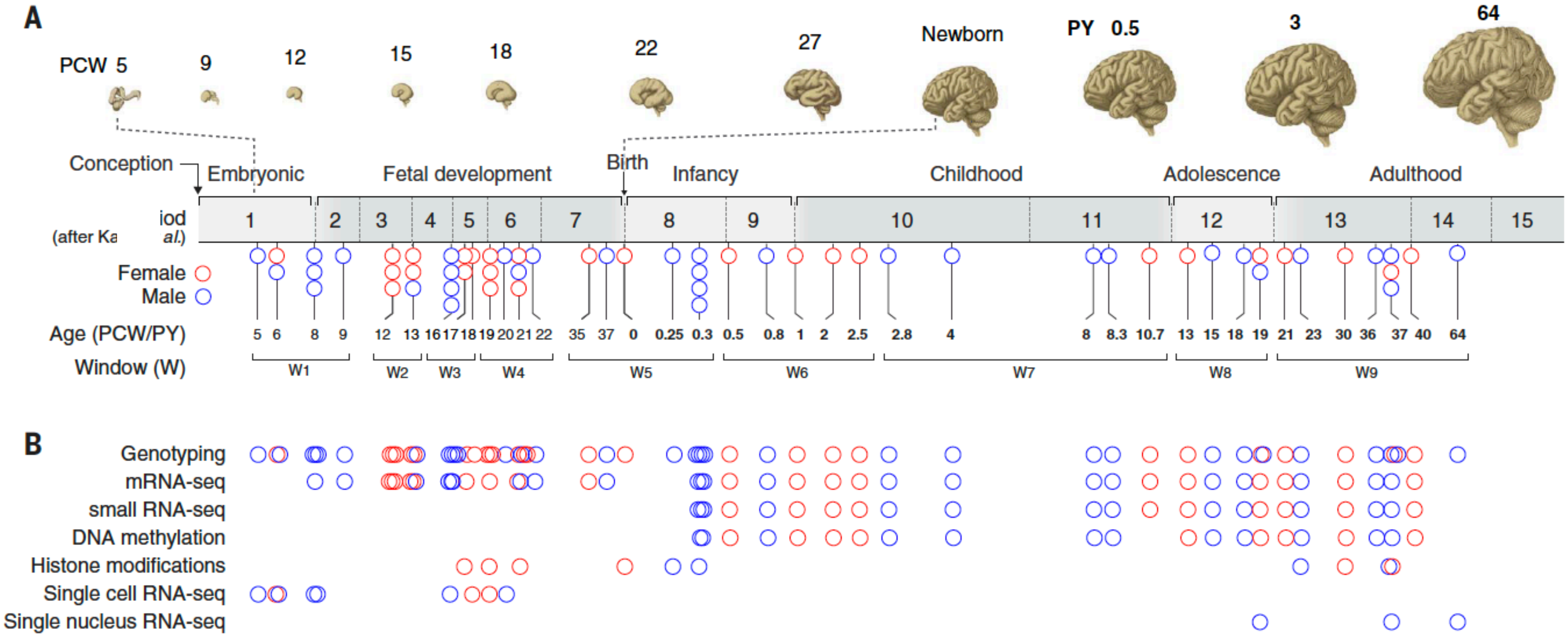


Excitatory to Inhibitory imbalance at neuronal subtype level for ASD*

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

[Wang et al. ('18) Science]

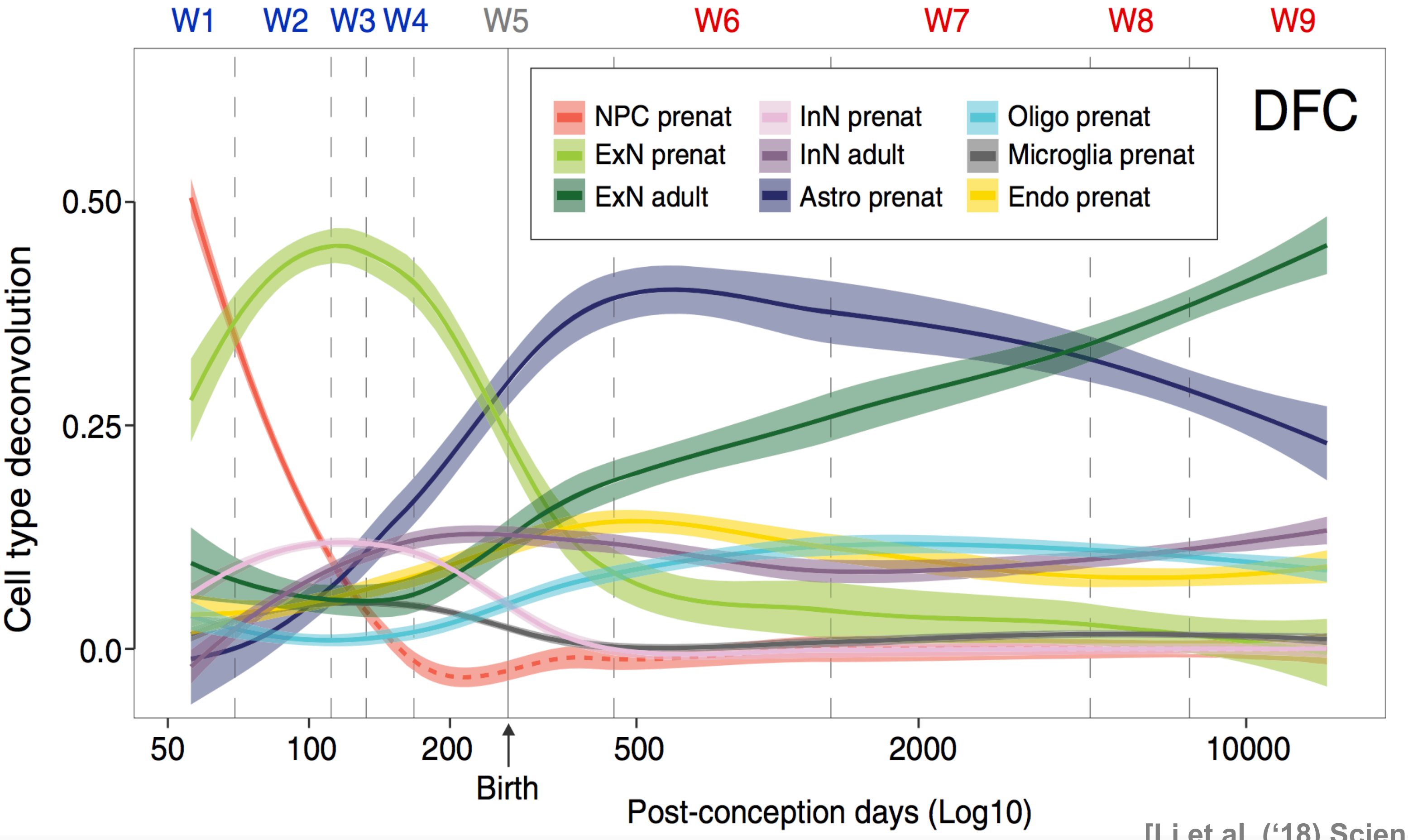
Developmental Capstone Data Set



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

[Li et al. ('18) Science]

Different neuronal & glial cell fractions across ages



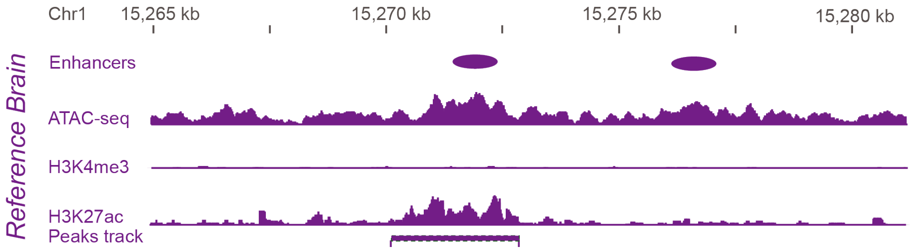
[Li et al. ('18) Science]

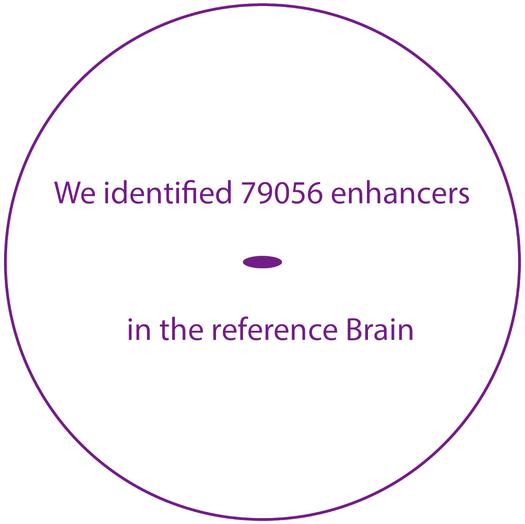
Using population-scale functional genomics to understand neuropsychiatric disease & interpreting the data exhaust from this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + dev. 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 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 disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as targets.

- *[Exhaust]* **Other uses** for the resource
  - Highlighting aging related genes + consistently comparing the brain to other organs
- *[Exhaust]* **Genomic Privacy**
  - The **Dilemma**
    - The genome as fundamental, inherited info that's very private v. need for large-scale mining for med. research
    - 2-sided nature of RNA-seq presents tricky privacy issues
  - **eQTLs**: Quantifying & removing variant info from expression levels with ICI & predictability. Instantiating a practical linking attack with noisy quasi-identifiers
  - **Signal Profiles**: Manifest appreciable leakage from large & small deletions. Linking attacks possible but additional complication of SV discovery in addition to genotyping

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





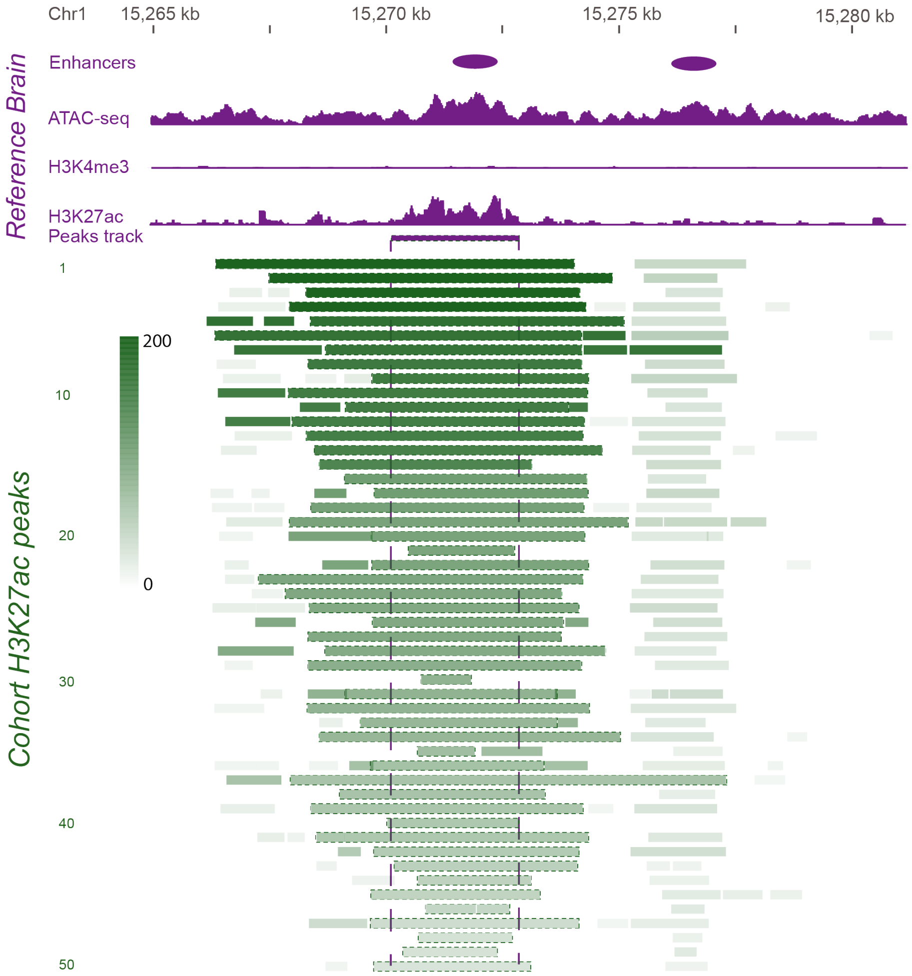
We identified 79056 enhancers

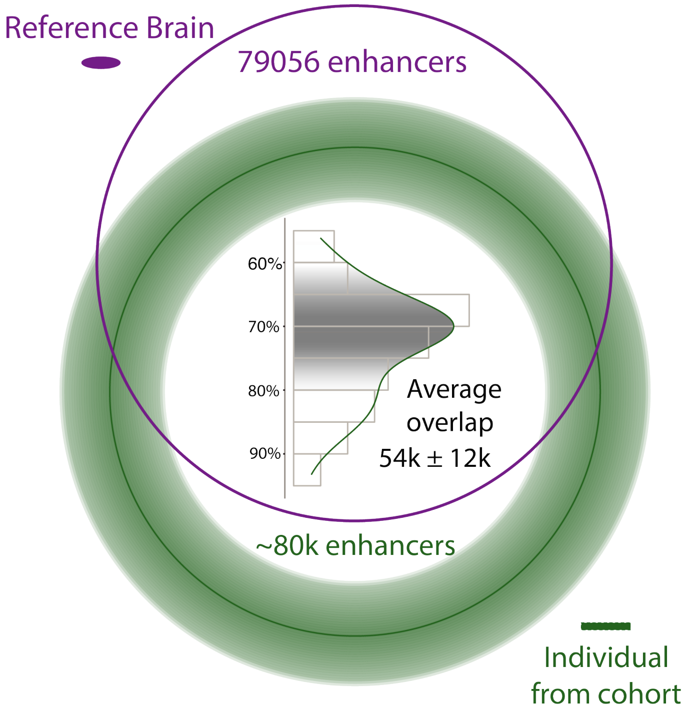
in the reference Brain

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

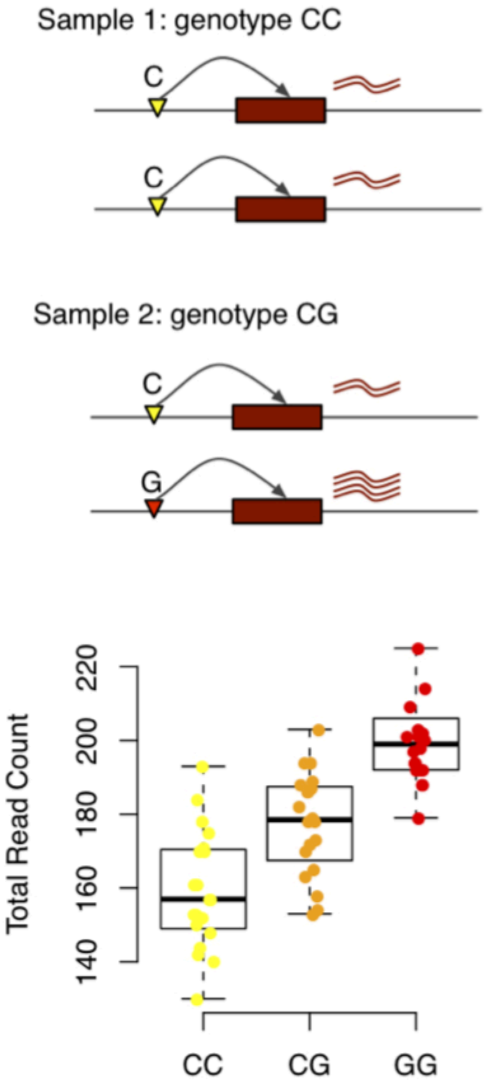




[Wang et al. ('18) Science]

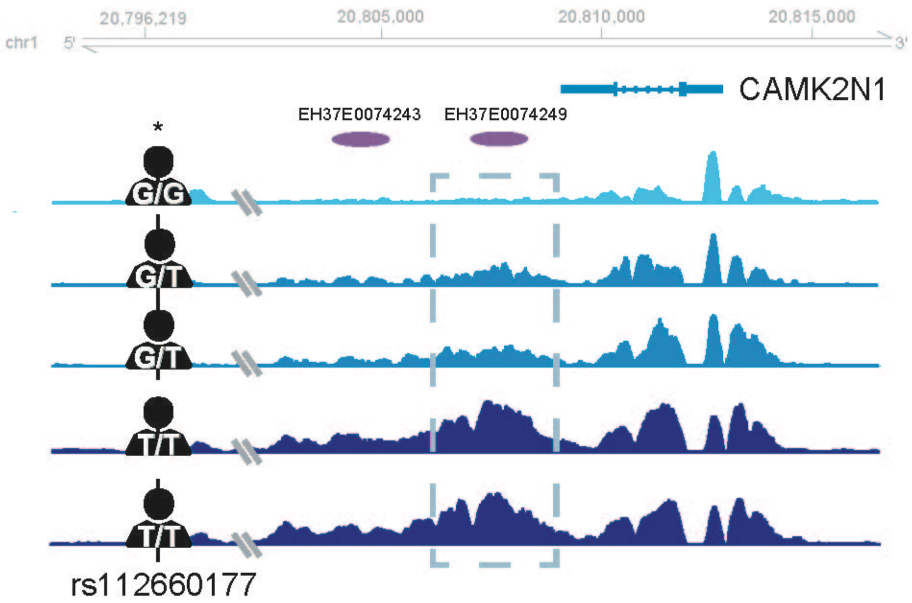
Quantitaive Trait Loci (QTLs) associated with variation

Gene expression (eQTL)



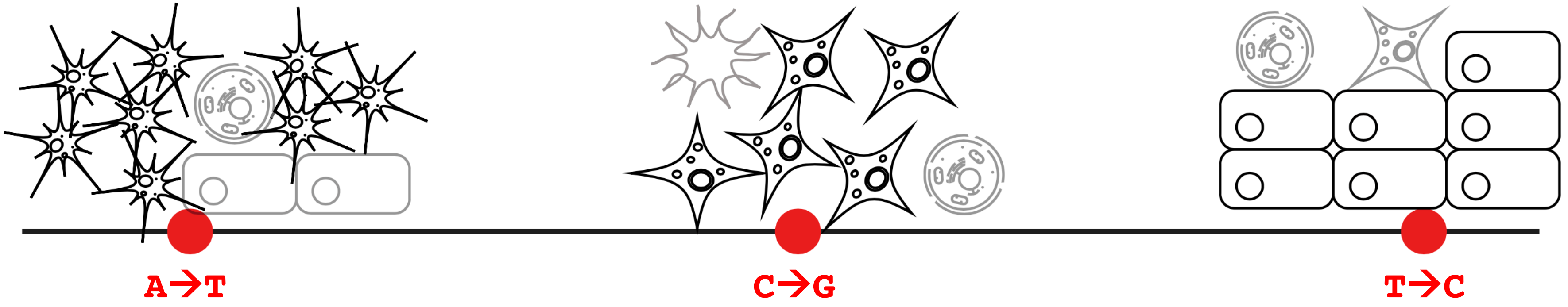
Sun, Wei, and Yijuan Hu. "eQTL mapping using RNA-seq data." Statistics in biosciences 5.1 (2013): 198-219.

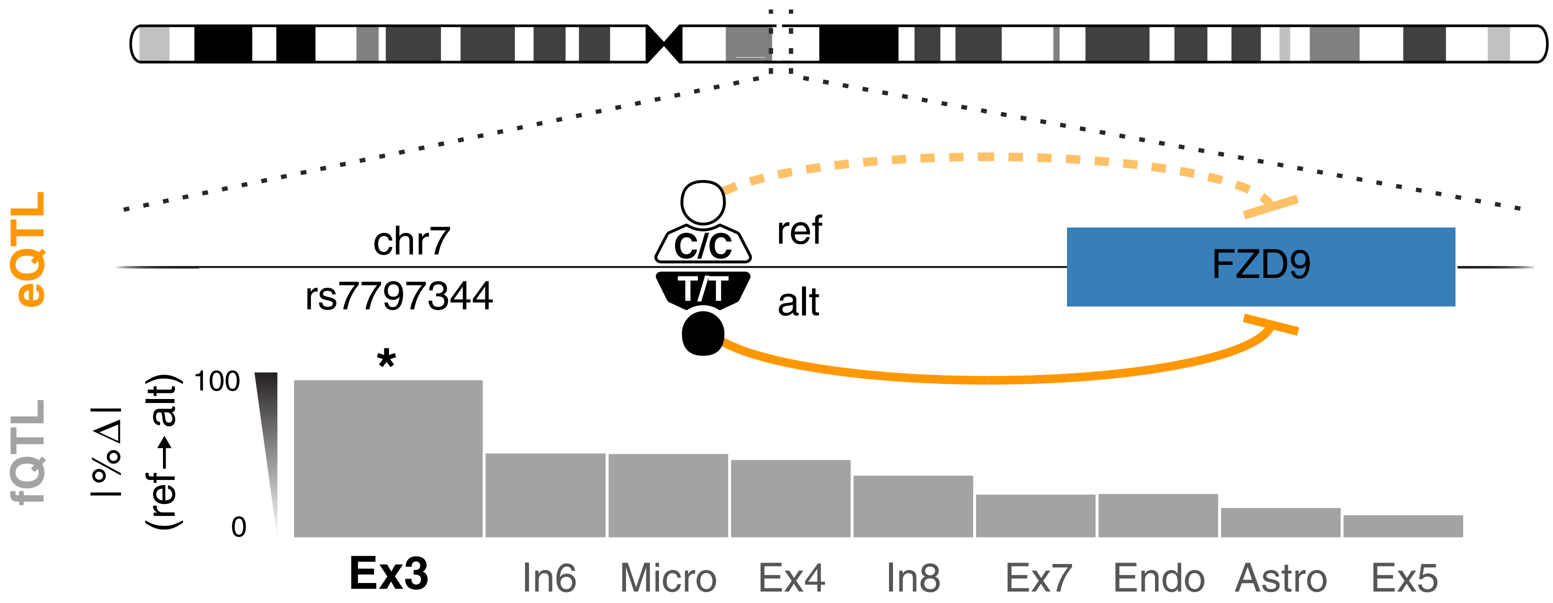
Chromatin (cQTL)



[Wang et al. ('18) Science]

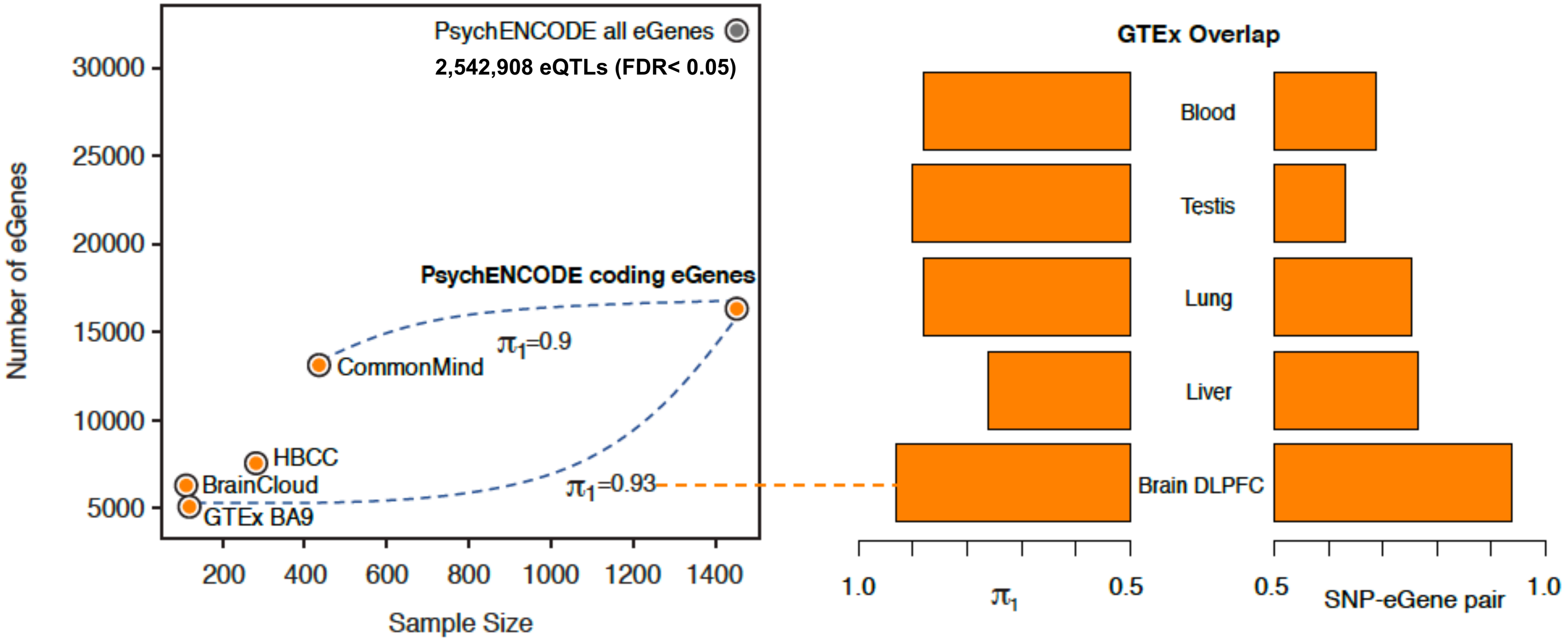
Cell fraction QTLs (fQTLs)





[Wang et al. ('18) Science]

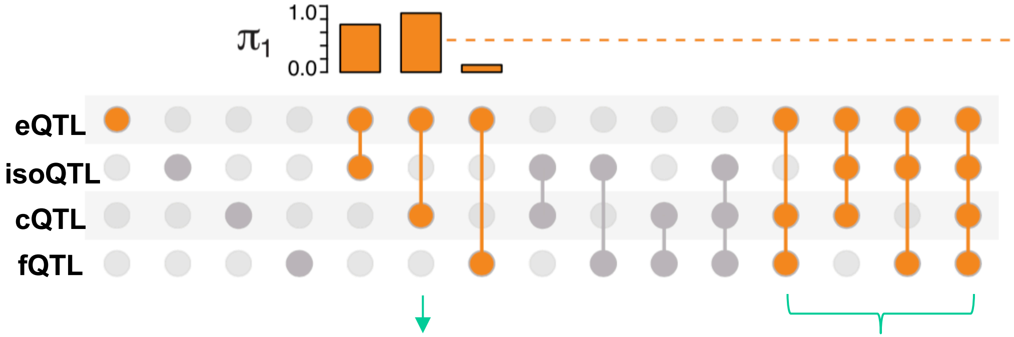
Larger brain eQTL sets than previous studies, but strong overlap with them



[Wang et al. ('18) Science]

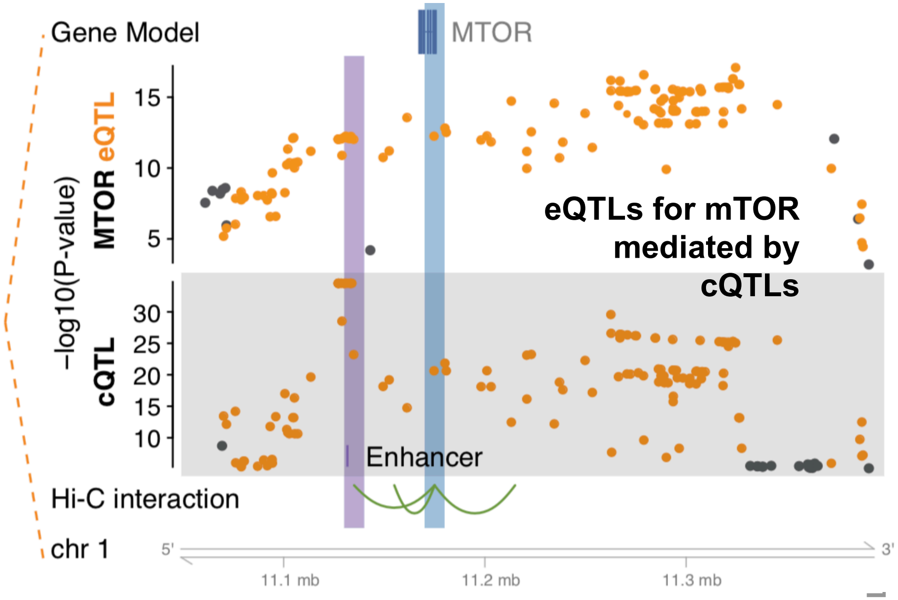
multi-QTLs from overlapping different types of QTLs: cQTL, fQTL, eQTL & isoQTL

	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

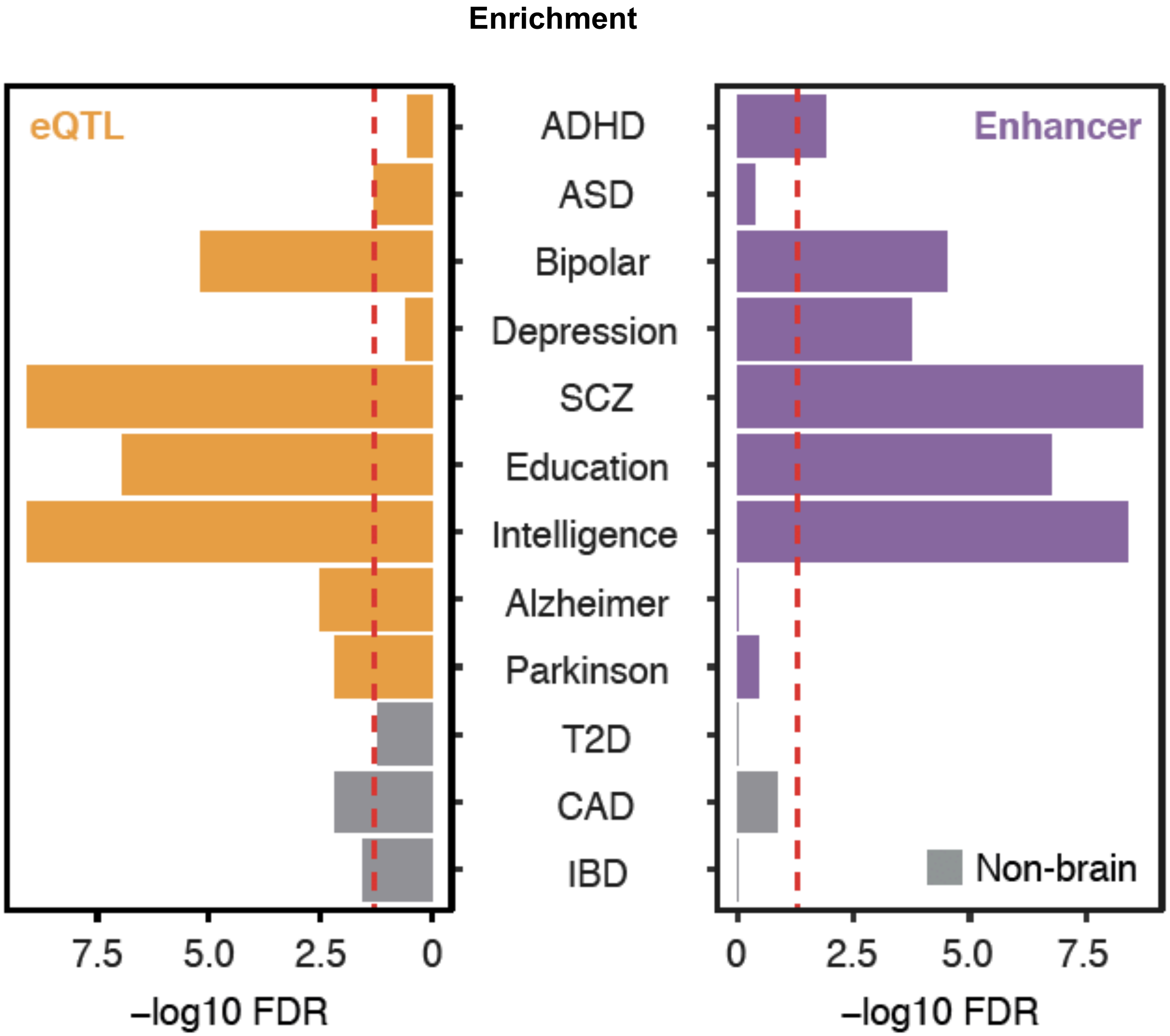
1391 SNPs (multi-QTLs) in at least three types among eQTLs, isoQTLs, cQTLs, fQTLs



eQTLs for mTOR mediated by cQTLs

[Wang et al. ('18) Science]

Brain eQTLs and enhancers enriched with GWAS SNPs for brain disorders



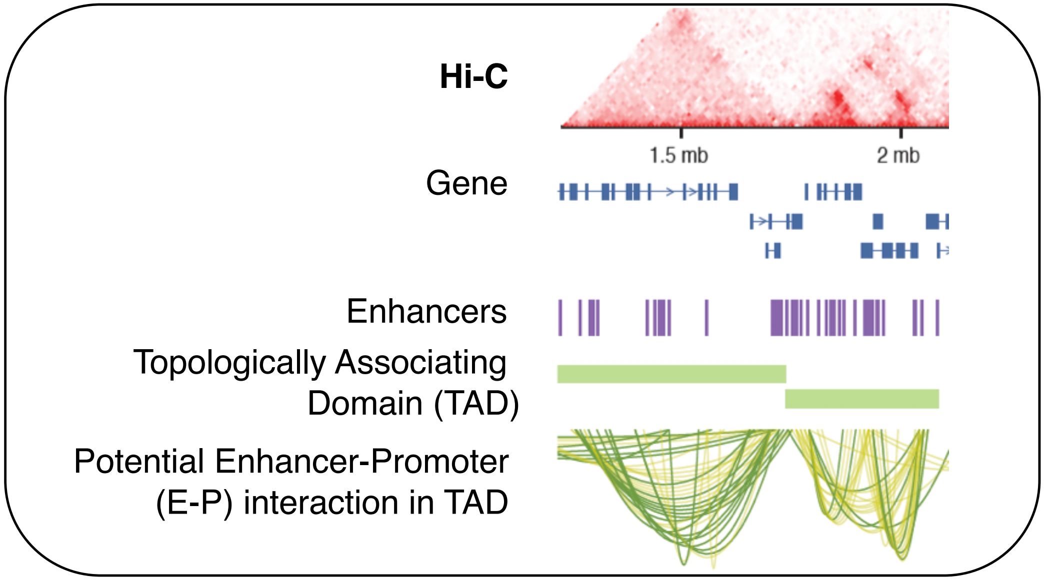
Wang, et al., *Science*, 2018

Using population-scale functional genomics to understand neuropsychiatric disease & interpreting the data exhaust from this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + dev. 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 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 disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as targets.

- *[Exhaust]* **Other uses** for the resource
  - Highlighting aging related genes + consistently comparing the brain to other organs
- *[Exhaust]* **Genomic Privacy**
  - The **Dilemma**
    - The genome as fundamental, inherited info that's very private v. need for large-scale mining for med. research
    - 2-sided nature of RNA-seq presents tricky privacy issues
  - **eQTLs**: Quantifying & removing variant info from expression levels with ICI & predictability. Instantiating a practical linking attack with noisy quasi-identifiers
  - **Signal Profiles**: Manifest appreciable leakage from large & small deletions. Linking attacks possible but additional complication of SV discovery in addition to genotyping

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

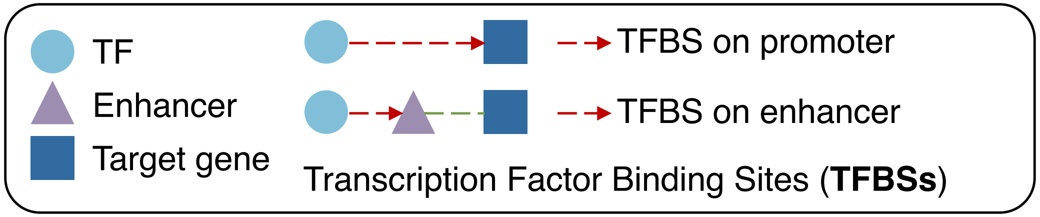

Hi-C

Gene

Enhancers

Topologically Associating Domain (TAD)

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


TF

Enhancer

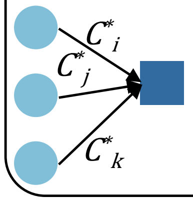
Target gene

TFBS on promoter

TFBS on enhancer

Transcription Factor Binding Sites (**TFBSs**)

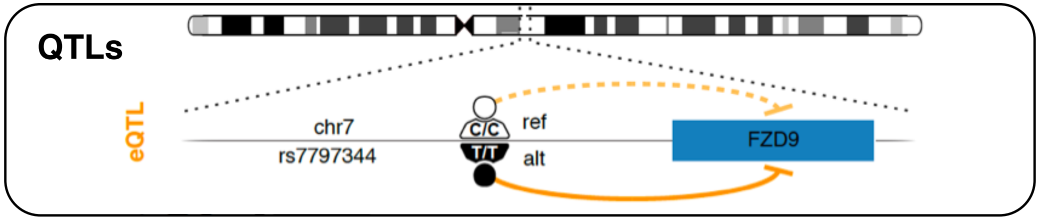
Expression activity relationship



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

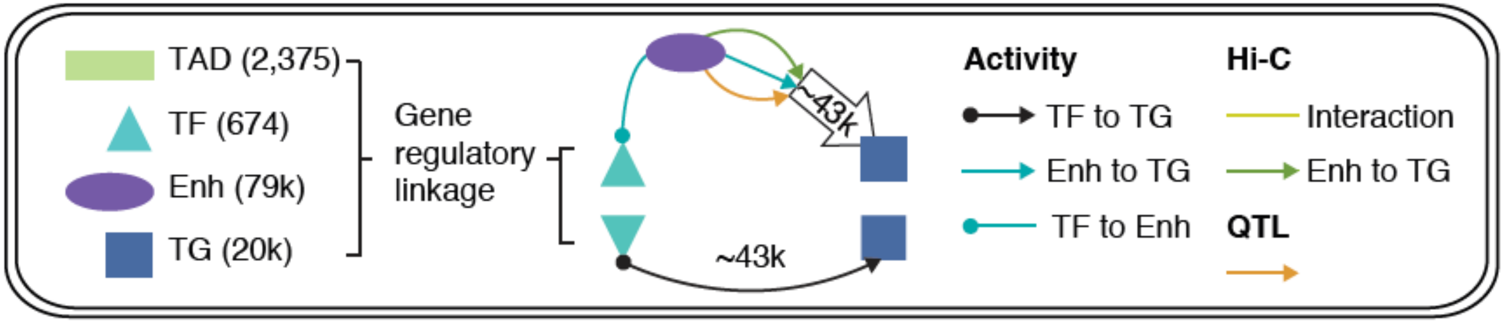
TF expression (X) to predict target gene expression (Y) using Elastic net regression

QTLs



[Wang et al. ('18) Science]

Imputed gene regulatory network for the human brain



TAD (2,375)

TF (674)

Enh (79k)

TG (20k)

Gene regulatory linkage

~43k

~43k

Activity

TF to TG

Enh to TG

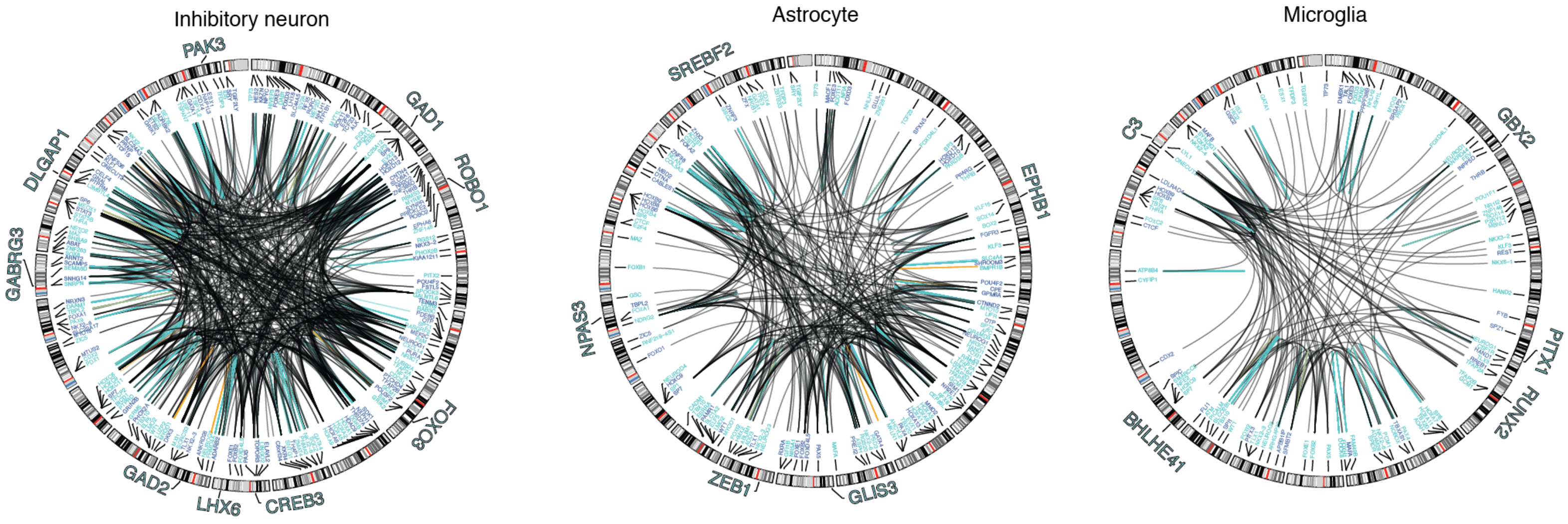
TF to Enh

Hi-C

Interaction

Enh to TG

QTL



Inhibitory neuron

Astrocyte

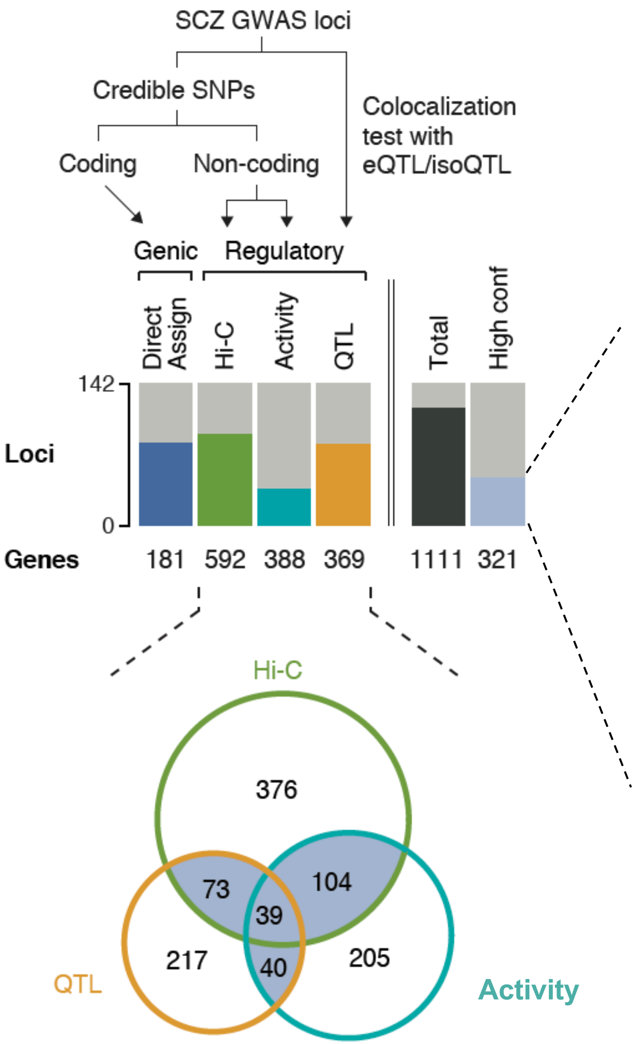
Microglia

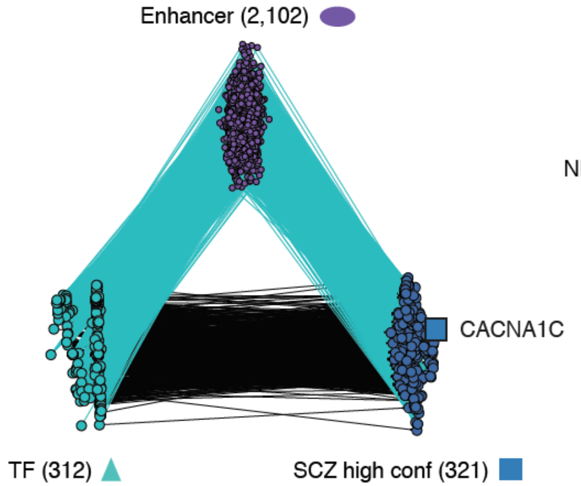
subnetworks targeting single cell marker genes

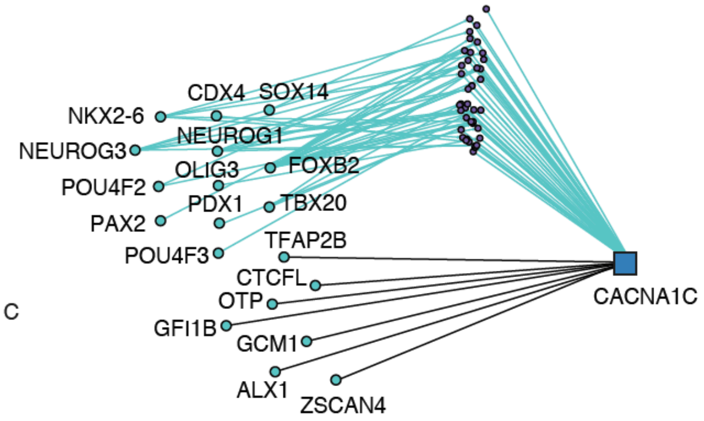
[Wang et al. ('18) Science]

Linking GWAS SNPs to disease genes using the regulatory network

142



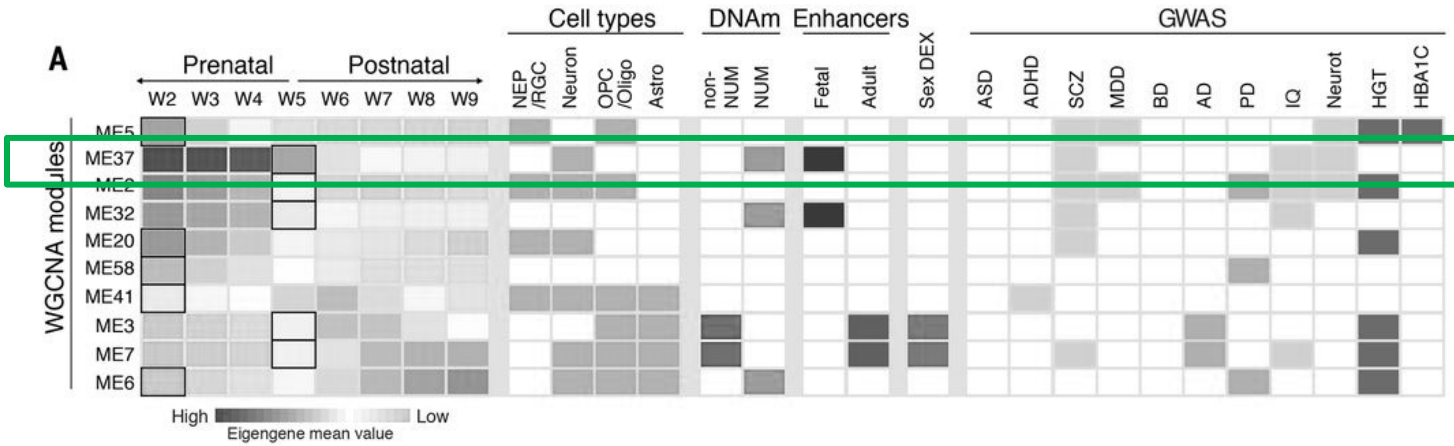




321
high-confident SCZ genes

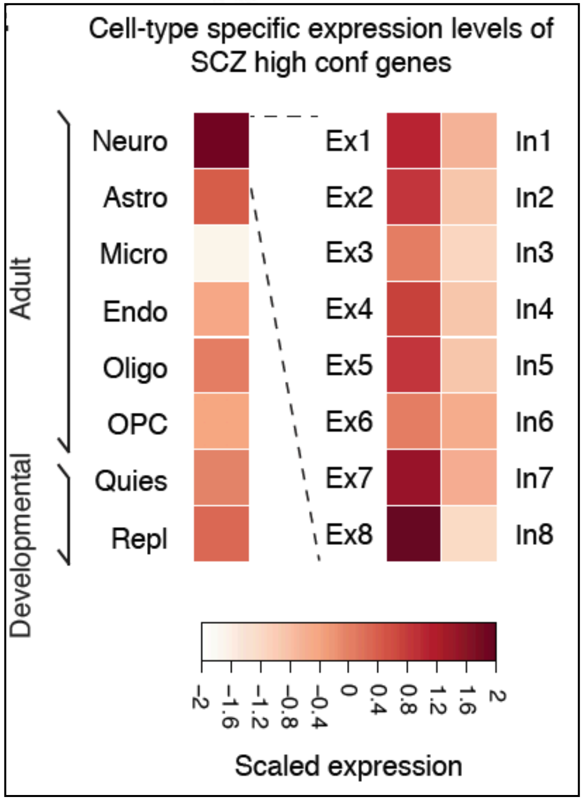
[Wang et al. ('18) Science]

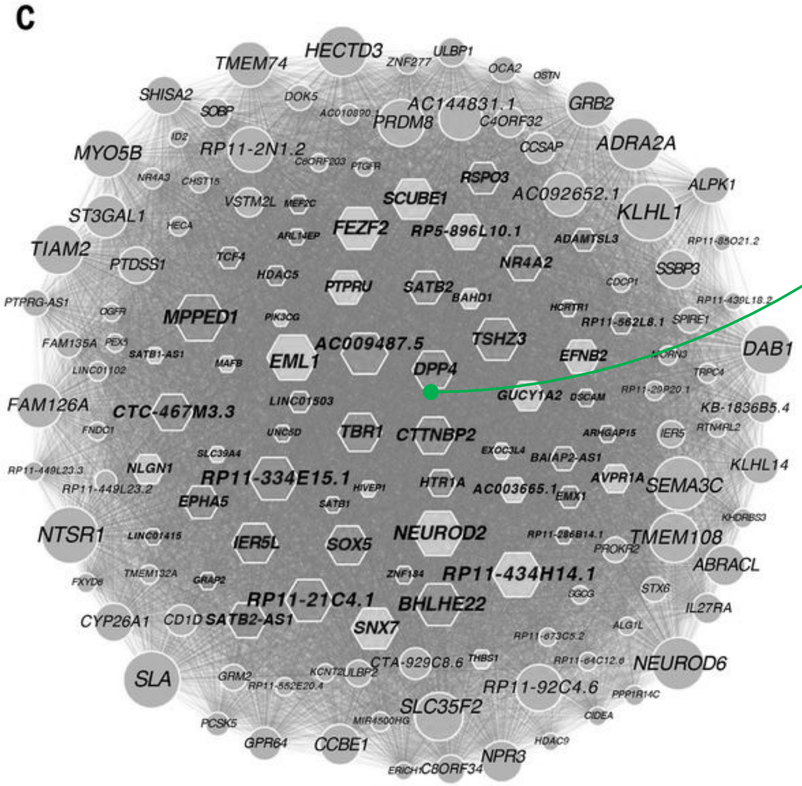
Genes associated with SCZ enriched in specific neuronal cell types & co-expression modules, active prenatally



ME37

Gene co-expression network





Mingfeng Li et al.
Science 2018

Wang, et al.,
Science, 2018

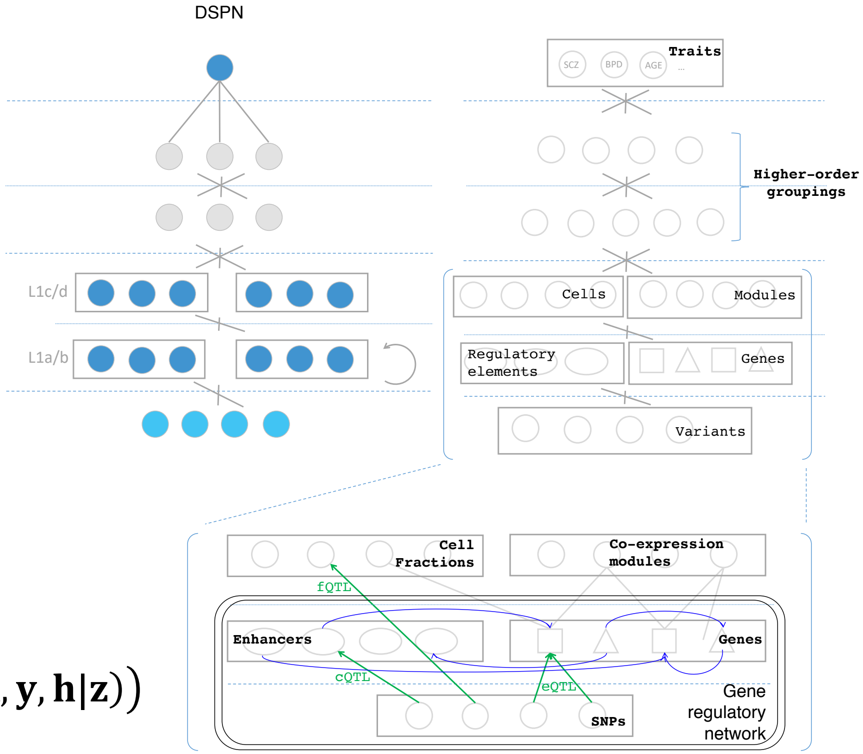
Using population-scale functional genomics to understand neuropsychiatric disease & interpreting the data exhaust from this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + dev. 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 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 disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as targets.

- *[Exhaust]* **Other uses** for the resource
  - Highlighting aging related genes + consistently comparing the brain to other organs
- *[Exhaust]* **Genomic Privacy**
  - The **Dilemma**
    - The genome as fundamental, inherited info that's very private v. need for large-scale mining for med. research
    - 2-sided nature of RNA-seq presents tricky privacy issues
  - **eQTLs**: Quantifying & removing variant info from expression levels with ICI & predictability. Instantiating a practical linking attack with noisy quasi-identifiers
  - **Signal Profiles**: Manifest appreciable leakage from large & small deletions. Linking attacks possible but additional complication of SV discovery in addition to genotyping

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)

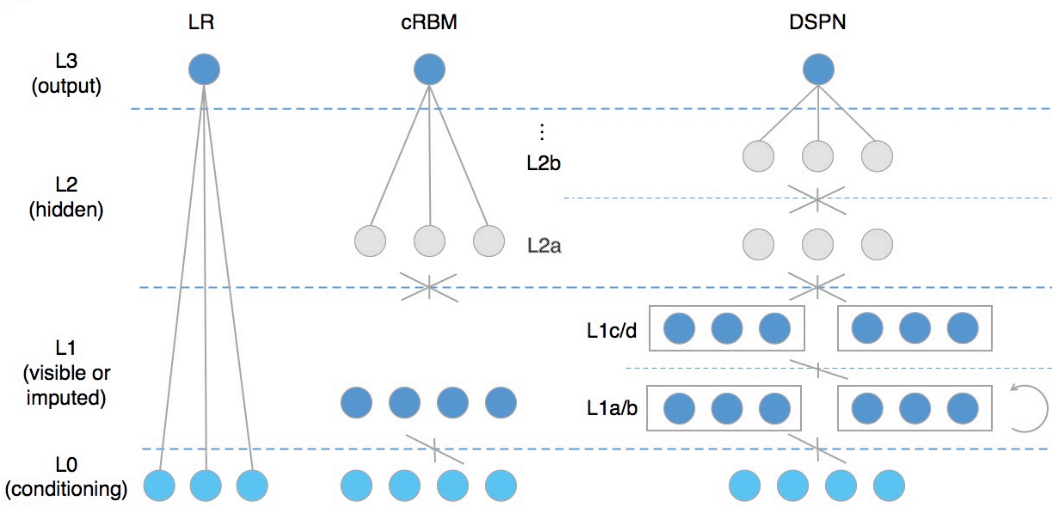
Energy model:

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

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

[Wang et al. ('18) Science]

DSPN improves brain disease prediction by adding deep layers



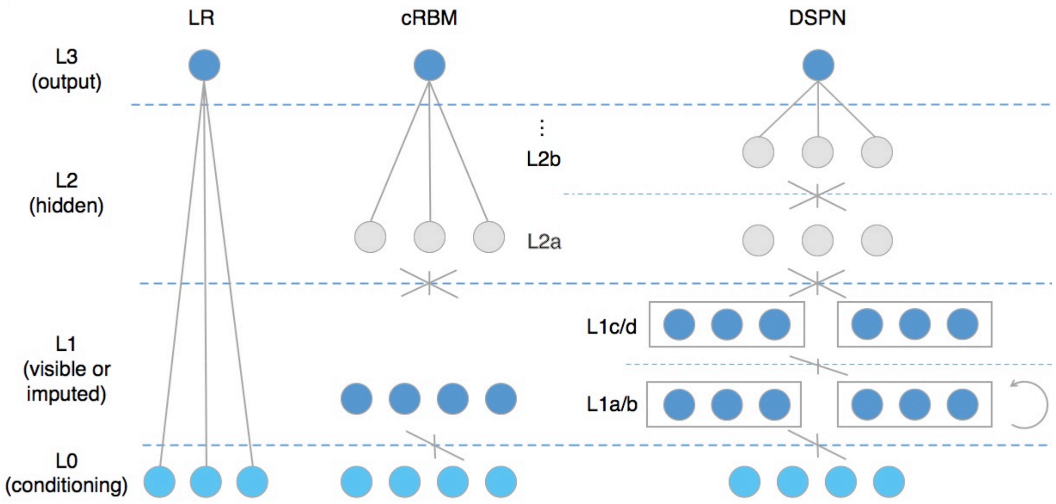
Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	**54.6%**	63.0%	70.0%	59.0%	**73.6%**
Bipolar Disorder	**56.7%**	63.3%	71.1%	67.2%	**76.7%**
Autism Spectrum Disorder	**50.0%**	51.7%	67.2%	62.5%	**68.3%**

X 6.0

Accuracy = chance to correctly predict disease/health

[Wang et al. ('18) Science]

DSPN improves brain disease prediction by adding deep layers



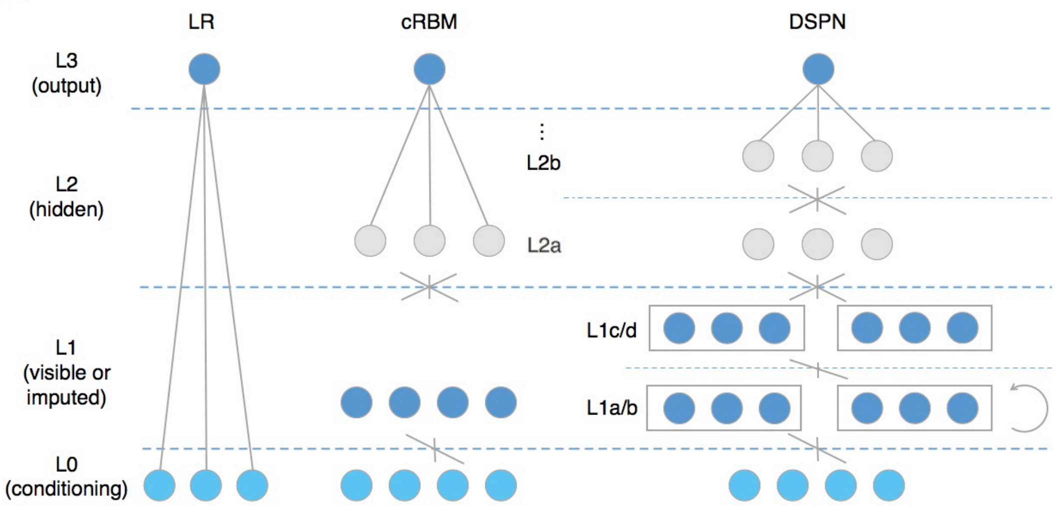
Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	54.6%	**63.0%**	70.0%	59.0%	**73.6%**
Bipolar Disorder	56.7%	**63.3%**	71.1%	67.2%	**76.7%**
Autism Spectrum Disorder	50.0%	**51.7%**	67.2%	62.5%	**68.3%**

X 2.5

Accuracy = chance to correctly predict disease/health

[Wang et al. ('18) Science]

DSPN improves brain disease prediction by adding deep layers



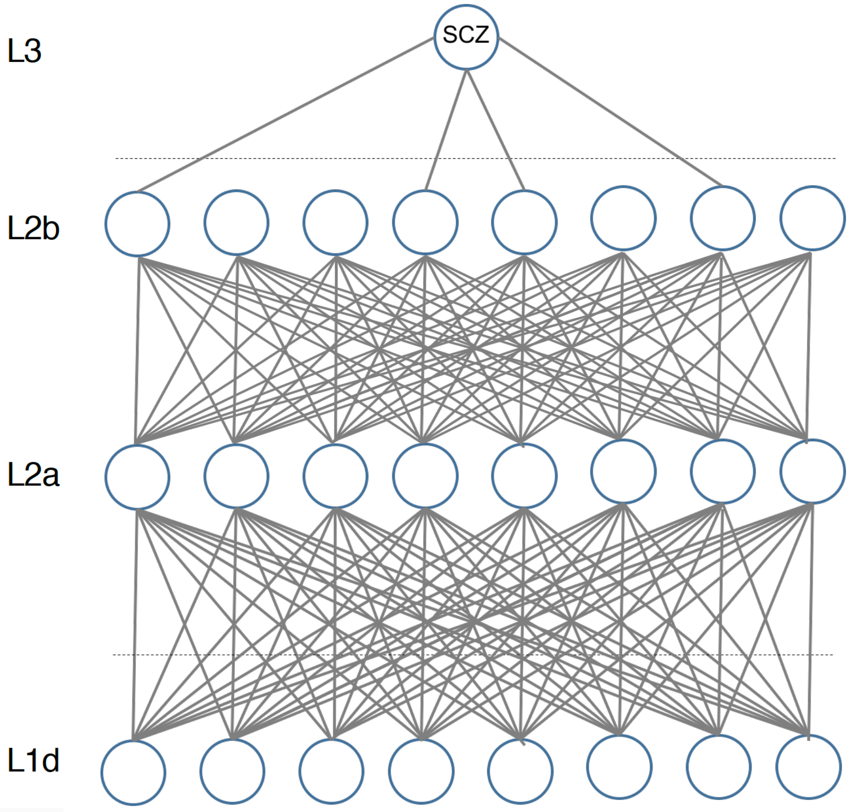
Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	**54.6%**	63.0%	70.0%	**59.0%**	73.6%
Bipolar Disorder	**56.7%**	63.3%	71.1%	**67.2%**	76.7%
Autism Spectrum Disorder	**50.0%**	51.7%	67.2%	**62.5%**	68.3%

X 3.1

Accuracy = chance to correctly predict disease/health

[Wang et al. ('18) Science]

Multilevel Network Interpretation

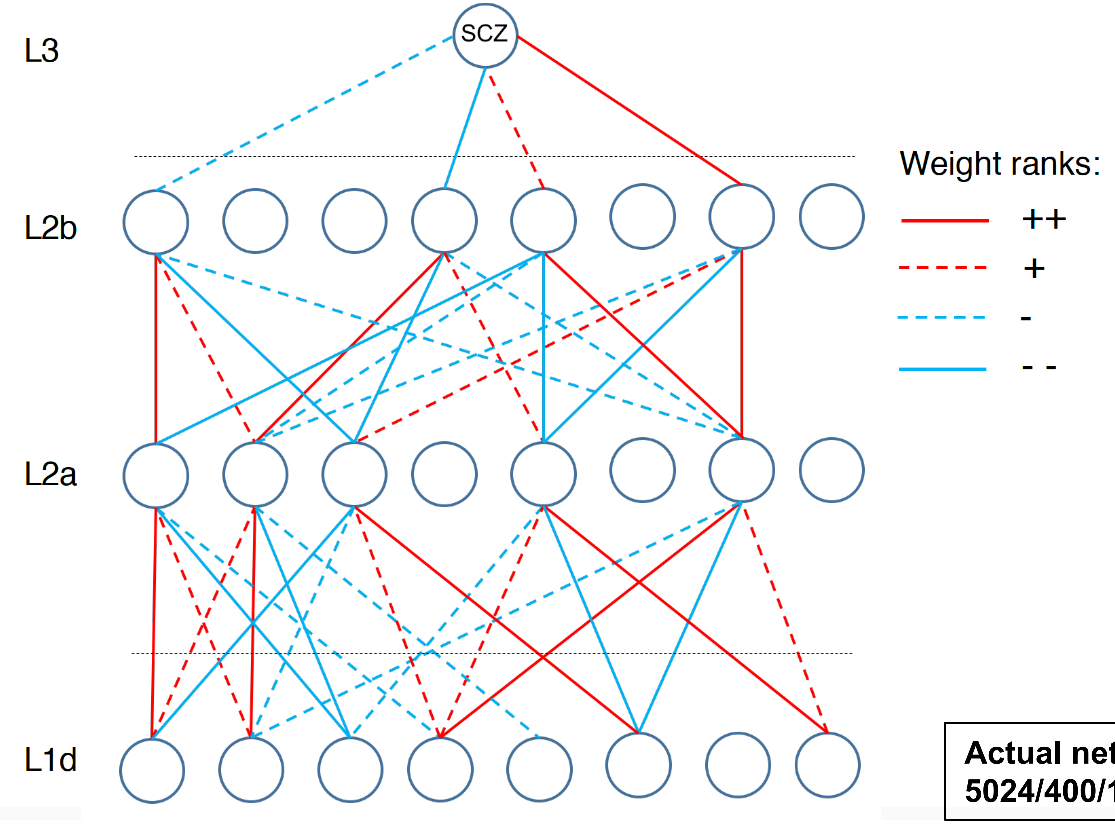


Actual network size:
5024/400/100/1 nodes

- Start with a fully connected trained network

[Wang et al. ('18) Science]

Multilevel Network Interpretation

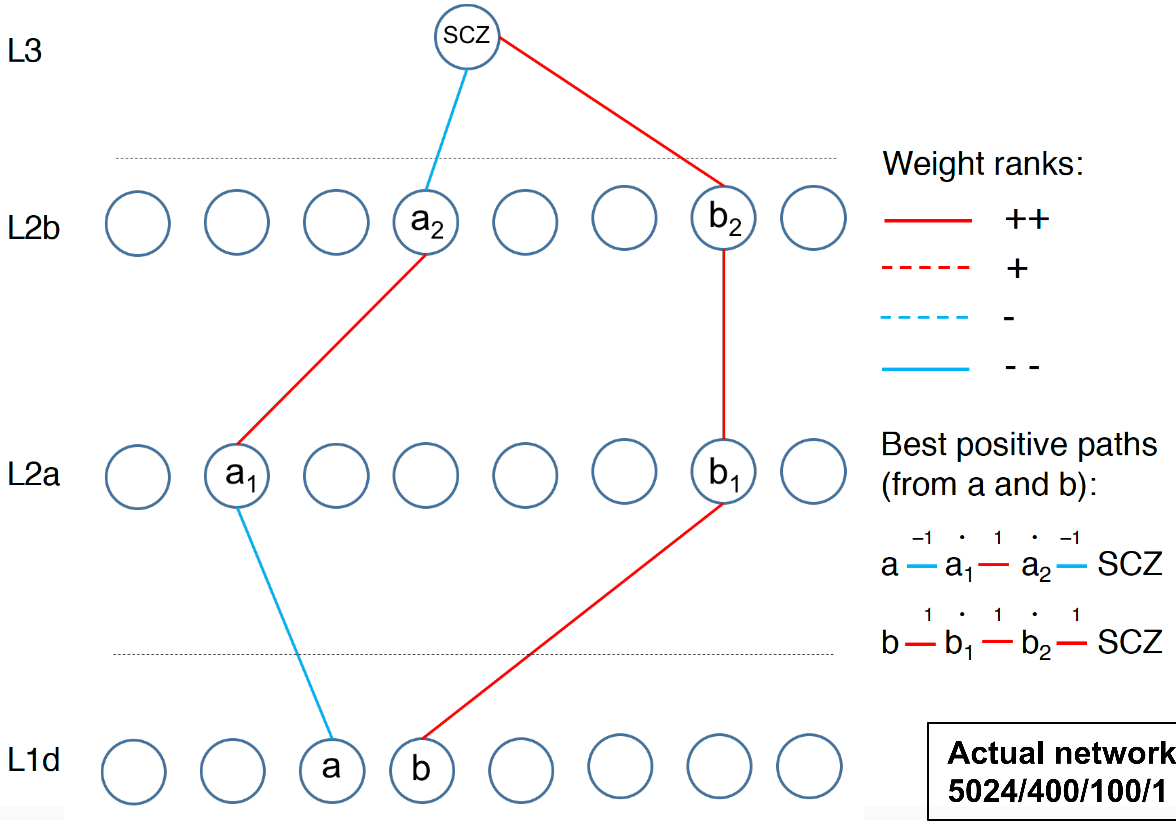


Actual network size:
5024/400/100/1 nodes

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

[Wang et al. ('18) Science]

Multilevel Network Interpretation



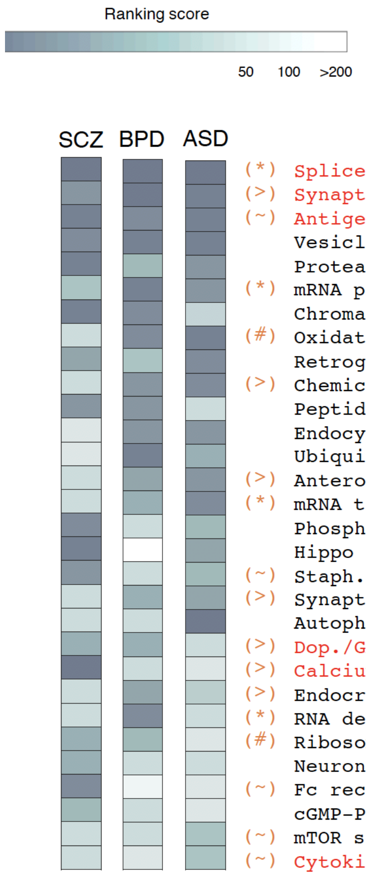
Actual network size: 5024/400/100/1 nodes

- 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

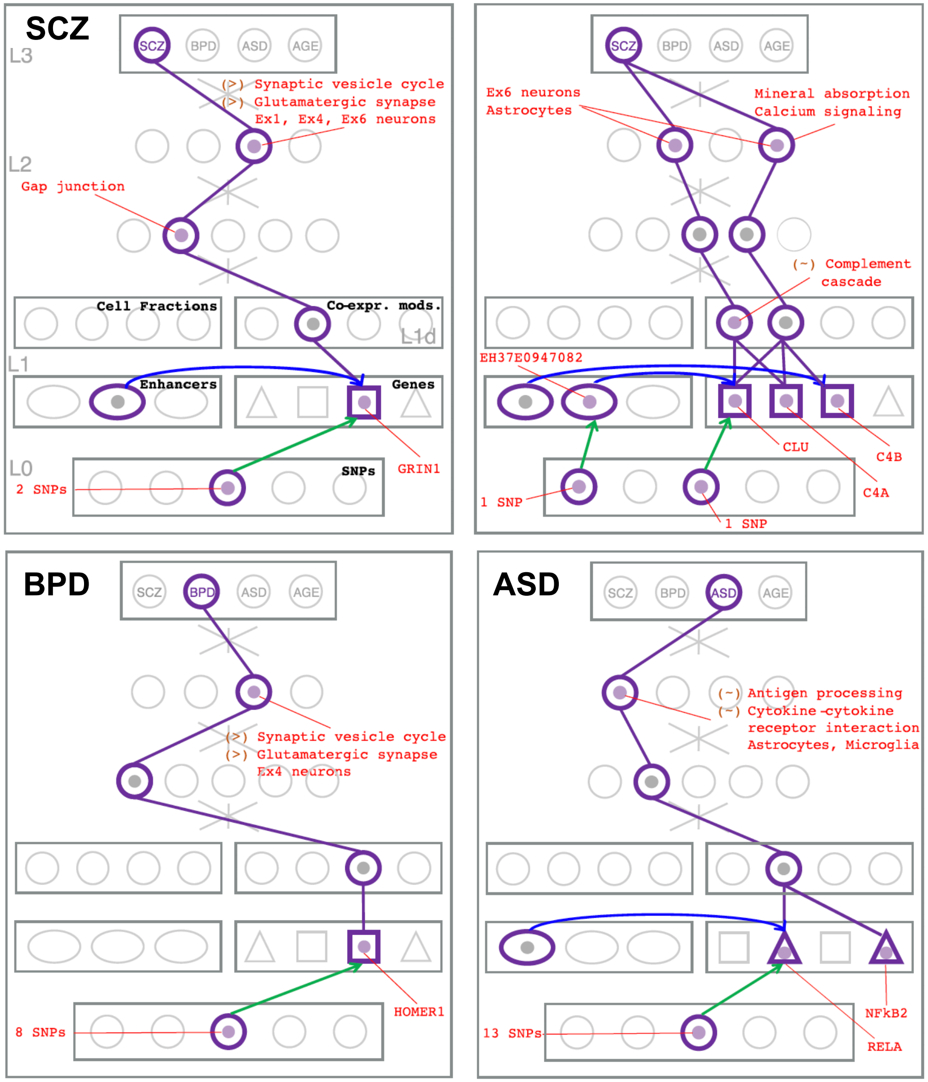
Cross-disorder MOD/HOG enrichment ranking



Functional categories

(*)	RNA proc.	(~)	Immune
(>)	Synaptic	(#)	Metabolic

(*)	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.



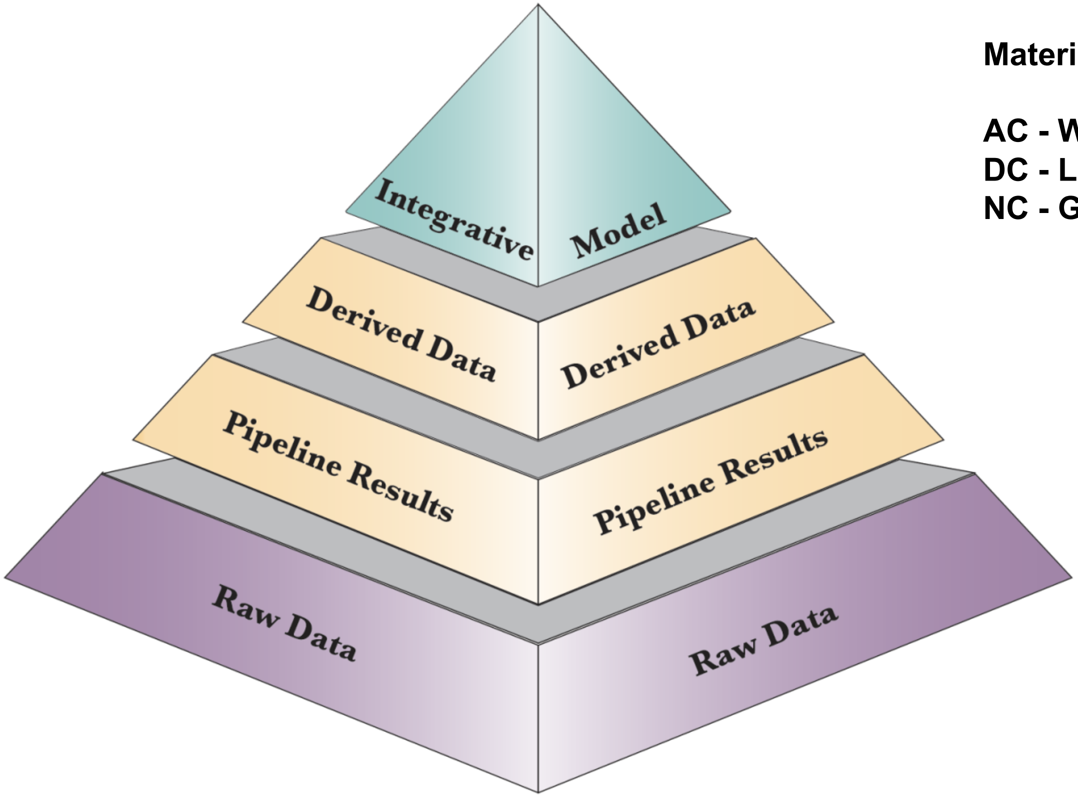
[Wang et al. ('18) Science]

Using population-scale functional genomics to understand neuropsychiatric disease & interpreting the data exhaust from this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + dev. 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 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 disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as targets.

- *[Exhaust]* **Other uses** for the resource
  - Highlighting aging related genes + consistently comparing the brain to other organs
- *[Exhaust]* **Genomic Privacy**
  - The **Dilemma**
    - The genome as fundamental, inherited info that's very private v. need for large-scale mining for med. research
    - 2-sided nature of RNA-seq presents tricky privacy issues
  - **eQTLs**: Quantifying & removing variant info from expression levels with ICI & predictability. Instantiating a practical linking attack with noisy quasi-identifiers
  - **Signal Profiles**: Manifest appreciable leakage from large & small deletions. Linking attacks possible but additional complication of SV discovery in addition to genotyping

Phase 1 PsychENCODE **capstone resource:** Layers of distributed information



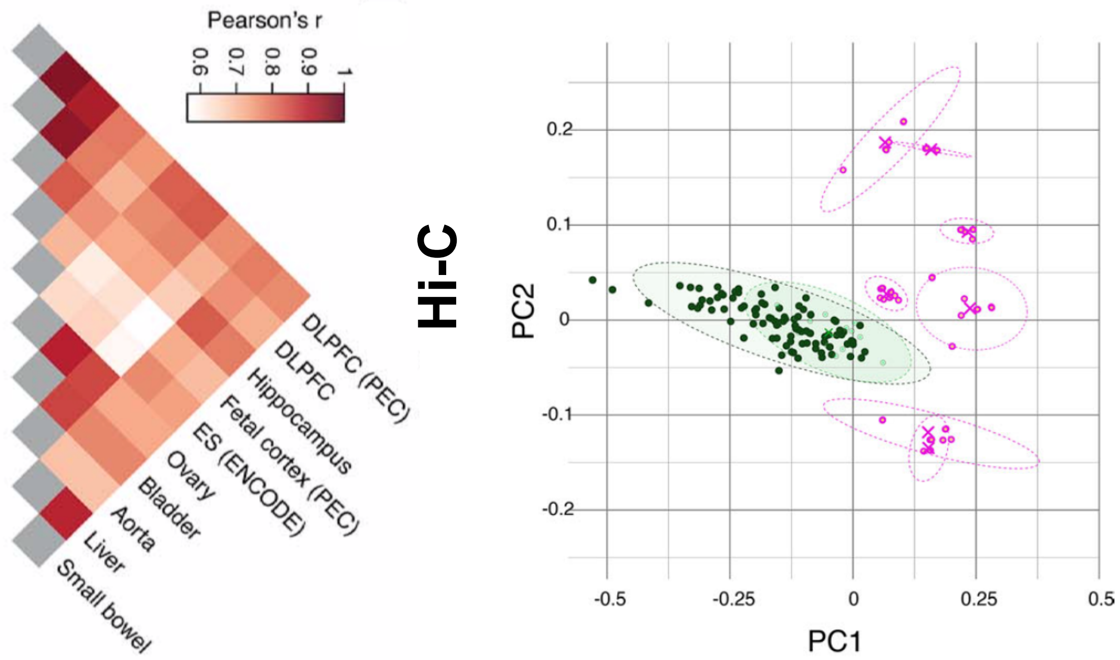
Material in the 3 capstones:

AC - Wang et al. ('18)
DC - Li et al. ('18)
NC - Gandal et al. ('18)

Resource.psychencode.org
Development.psychencode.org

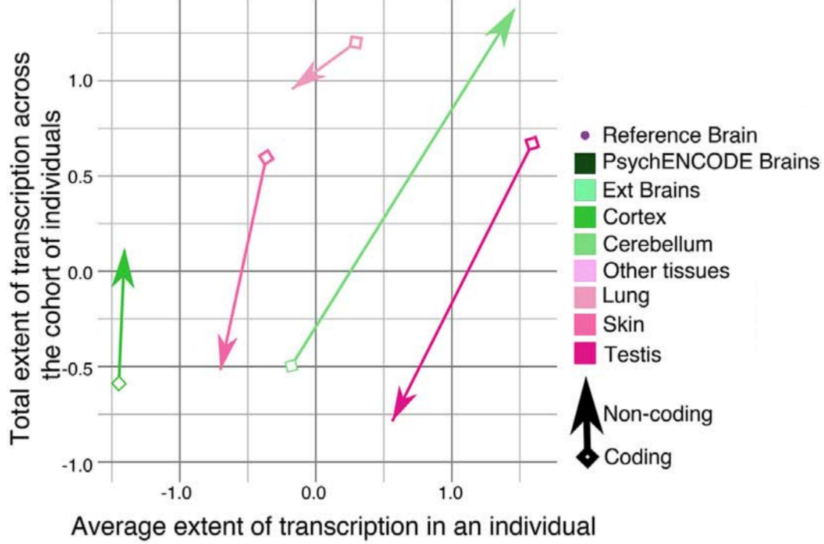
Cross tissue variation in Chromatin & Expression

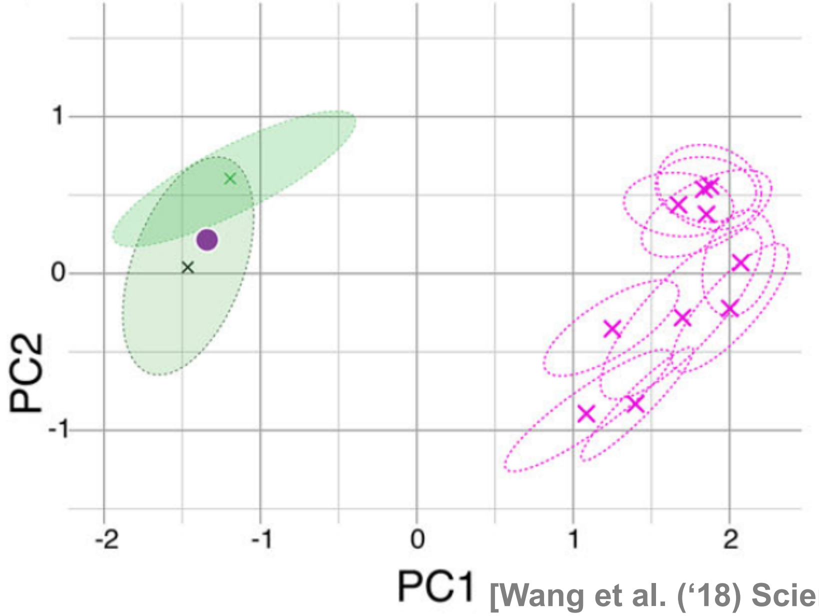
Placing the Brain in context of all other Body Tissues



Chromatin

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



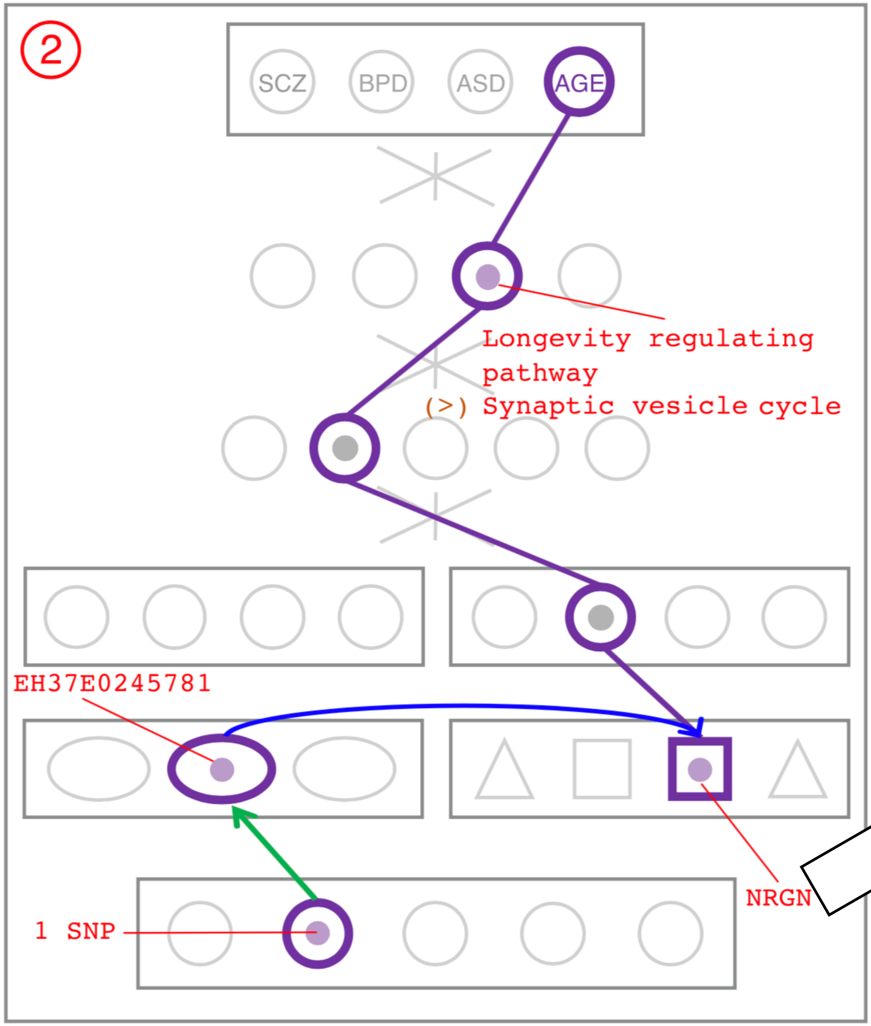


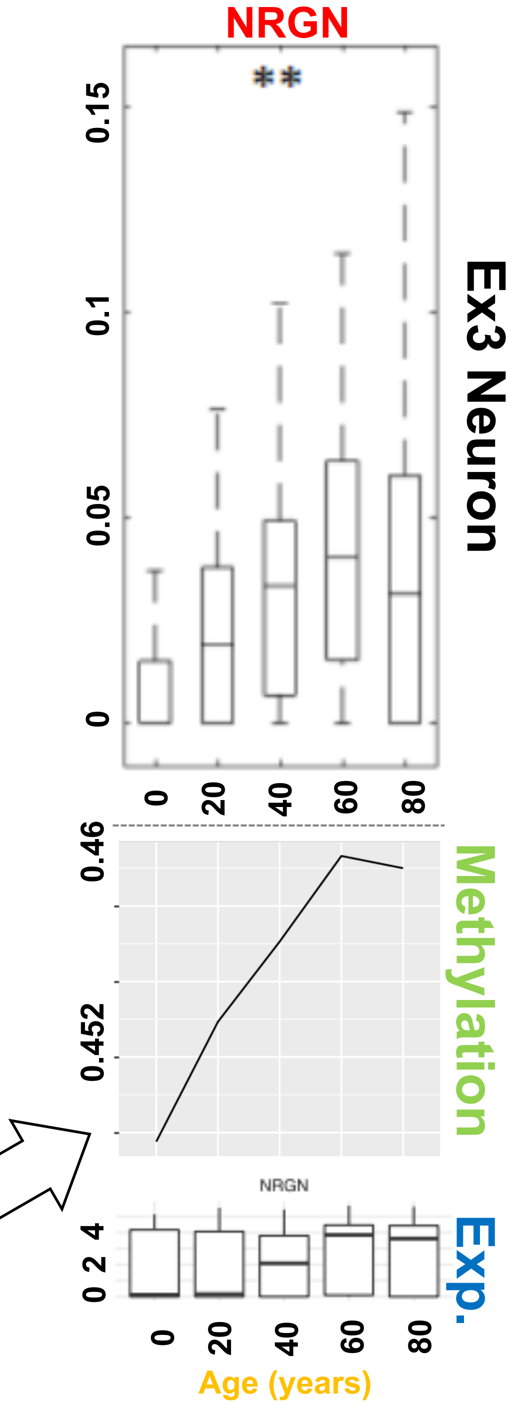
Expression

[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 understand neuropsychiatric disease & interpreting the data exhaust from this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + dev. 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 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 disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as targets.
- *[Exhaust]* **Other uses** for the resource
  - Highlighting aging related genes + consistently comparing the brain to other organs
- *[Exhaust]* **Genomic Privacy**
  - The **Dilemma**
    - The genome as fundamental, inherited info that's very private v. need for large-scale mining for med. research
    - 2-sided nature of RNA-seq presents tricky privacy issues
  - **eQTLs**: Quantifying & removing variant info from expression levels with ICI & predictability. Instantiating a practical linking attack with noisy quasi-identifiers
  - **Signal Profiles**: Manifest appreciable leakage from large & small deletions. Linking attacks possible but additional complication of SV discovery in addition to genotyping

2-sided nature of functional genomics data: Analysis can be very General/Public or Individual/Private

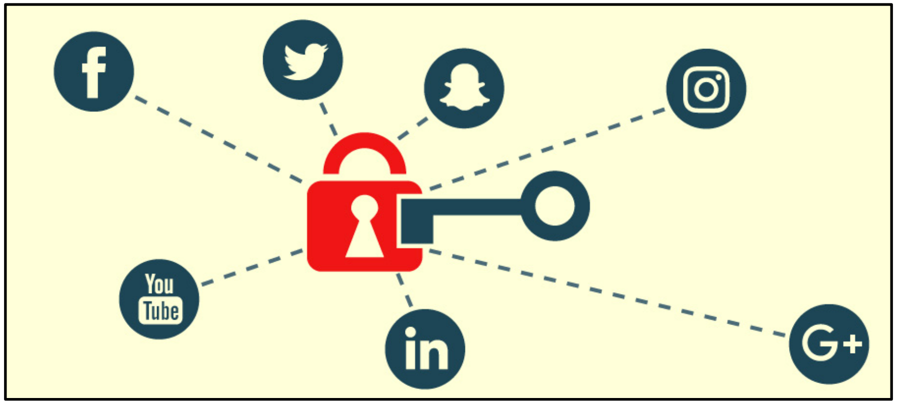
- **General quantifications** related to overall aspects of a condition – ie gene activity as a function of:
  - Developmental stage, Evolutionary relationships, Cell-type, Disease

- **Above are not tied to an individual's genotype. However, data is derived from individuals & tagged with their genotypes**

- (Note, a few calculations aim to use explicitly genotype to derive general relations related to sequence variation & gene expression - eg allelic activity)

Privacy: Does Genomics has similar "Big Data" Dilemma as in the Rest of Society?

- We confront privacy risks every day we access the internet (e.g., social media, e-commerce).
- Sharing & "peer-production" is central to success of many new ventures, with analogous risks to genomics
  - **EG web search**: Large-scale mining essential



Genetic Exceptionalism :

The Genome is very fundamental data, potentially very revealing about one's identity & characteristics

Personal Genomic info. essentially meaningless currently but will it be in 20 yrs? 50 yrs?

Genomic sequence very revealing about one's children. Is true consent possible?

Once put on the web it can't be taken back

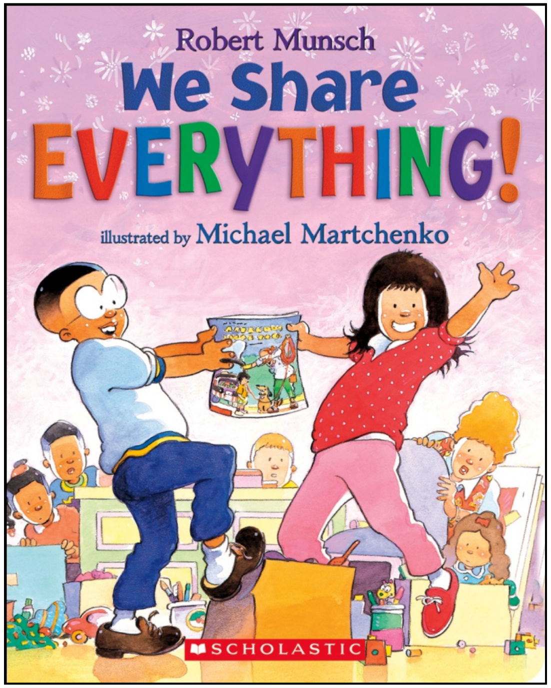
Ethically challenged history of genetics

Ownership of the data & what consent means (Hela)

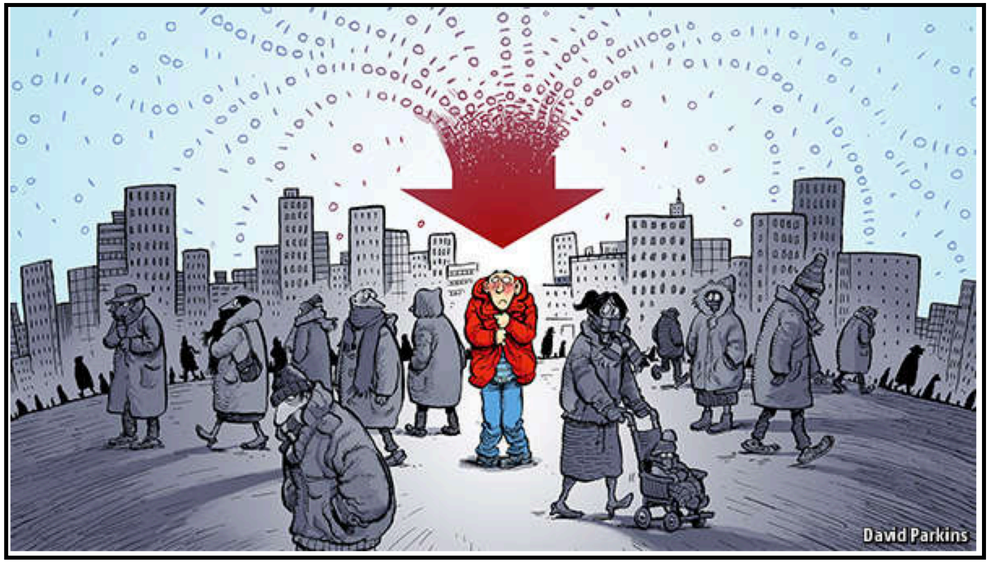
Could your genetic data give rise to a product line?

[Seringhaus & Gerstein ('09), *Hart. Courant* (Jun 5); Greenbaum & Gerstein ('11), *NY Times* (6 Oct), D Greenbaum & M Gerstein ('08). Am J. Bioethics; D Greenbaum & M Gerstein, Hartford Courant, 10 Jul. '08 ; SF Chronicle, 2 Nov. '08; Greenbaum et al. *PLOS CB* ('11) ; Greenbaum & Gerstein ('13), The Scientist; Photos from NY Times, it.wisc.edu]

The Other Side of the Coin for Genomics: Why we should share



- Sharing helps **speed research**
  - Large-scale mining of this information is important for medical research
  - Statistical power
  - Privacy is cumbersome, particularly for big data



[Economist, 15 Aug '15]

The Dilemma

- The individual (harmed?) v the collective (benefits)
  - But do sick patients care about their privacy?
- How to balance risks v rewards
  - Quantification

[Yale Law Roundtable ('10). Comp. in Sci. & Eng. 12:8; D Greenbaum & M Gerstein ('09). Am. J. Bioethics; D Greenbaum & M Gerstein ('10). SF Chronicle, May 2, Page E-4; Greenbaum et al. *PLOS CB* ('11)]

Current Social & Technical Solutions: The quandary where are now

- **Closed Data** Approach
  - Consents
  - “Protected” distribution via dbGAP
  - Local computes on secure computer
- Issues with Closed Data
  - Non-uniformity of consents & paperwork
    - Different, confusing int’l norms
  - Computer security is burdensome
  - Many schemes get “hacked” .
  - **Tricky aspects of high-dimensional data** (leakage & ease of creating quasi-identifiers)
- **Open Data**
  - Genomic "test pilots” (ala PGP)?
    - Sports stars & celebrities?
  - Some public data & data donation is helpful but is this a realistic solution for an unbiased sample of ~1M

[Greenbuam et al ('04), Nat. Biotech; Greenbaum & Gerstein ('13), The Scientist; Photo: futurism.com]

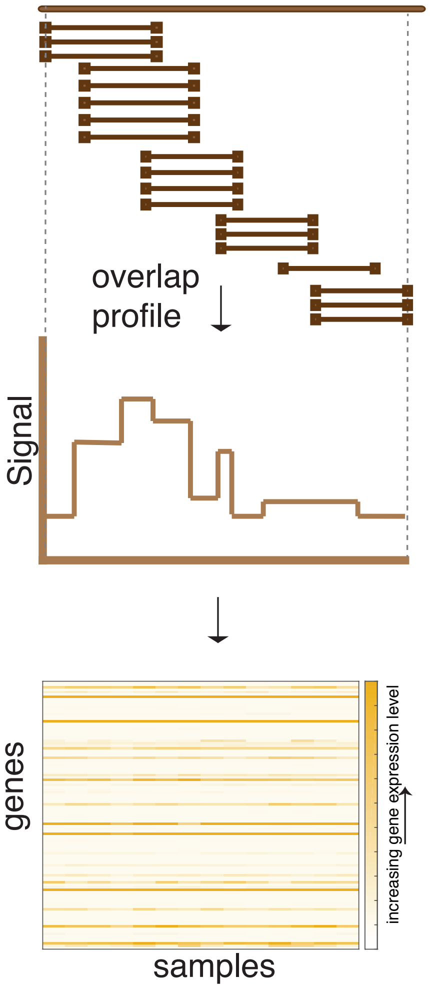
Strawman Hybrid Social & Tech Proposed Solution?

- Fundamentally, researchers have to keep genetic secrets.
  - **Need for an (international) legal framework**
  - Genetic Licensure & training for individuals (similar to medical license, drivers license)
- Technology to make things easier
  - Cloud computing & enclaves (eg solution of Genomics England)
- Technological barriers shouldn't create a social incentive for "hacking"

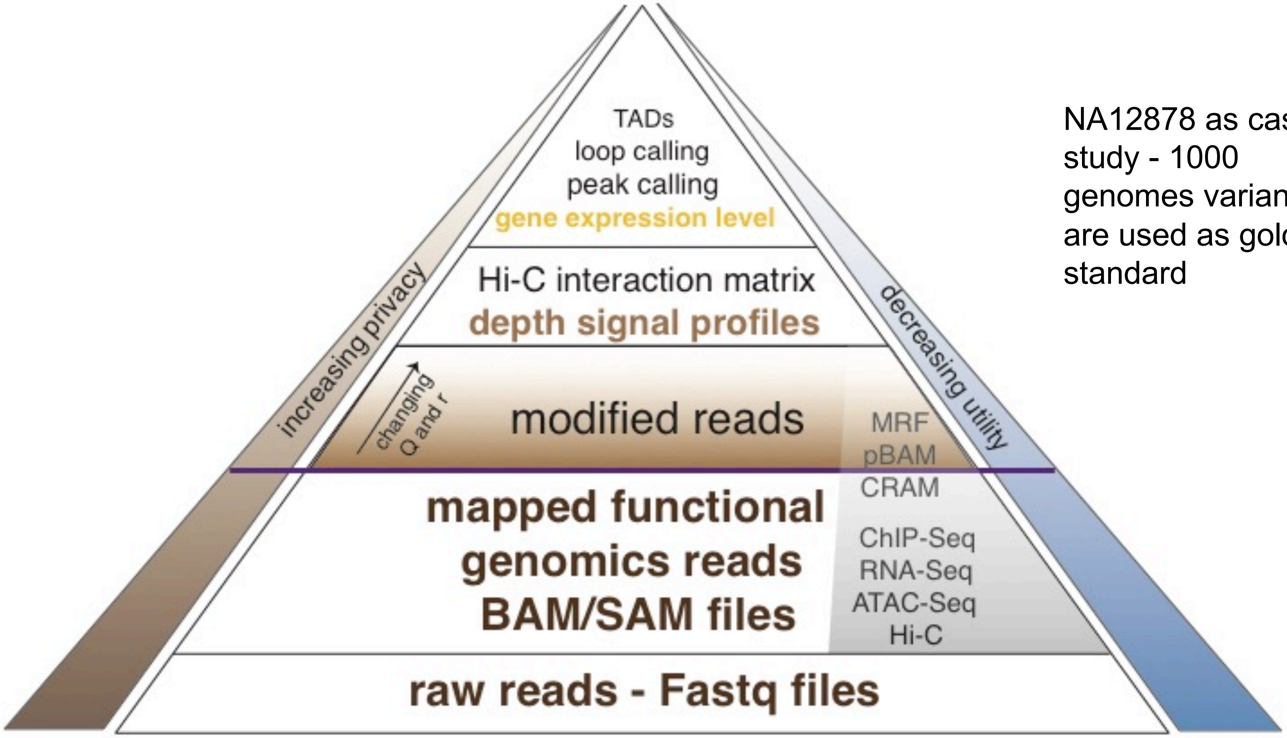
- **Quantifying Leakage & allowing a small amounts of it**
- **Careful separation & coupling of private & public data**
  - **Lightweight, freely accessible secondary datasets coupled to underlying variants**
  - Selection of stub & "test pilot" datasets for benchmarking
  - Develop programs on public stubs on your laptop, then move the program to the cloud for private production run

[D Greenbaum, M Gerstein ('11). Am J Bioeth 11:39. Greenbaum & Gerstein, The Scientist ('13)]

Functional genomics data comes with a great deal of sequencing; We can quantify amount of leakage at every step of the data summarization process.



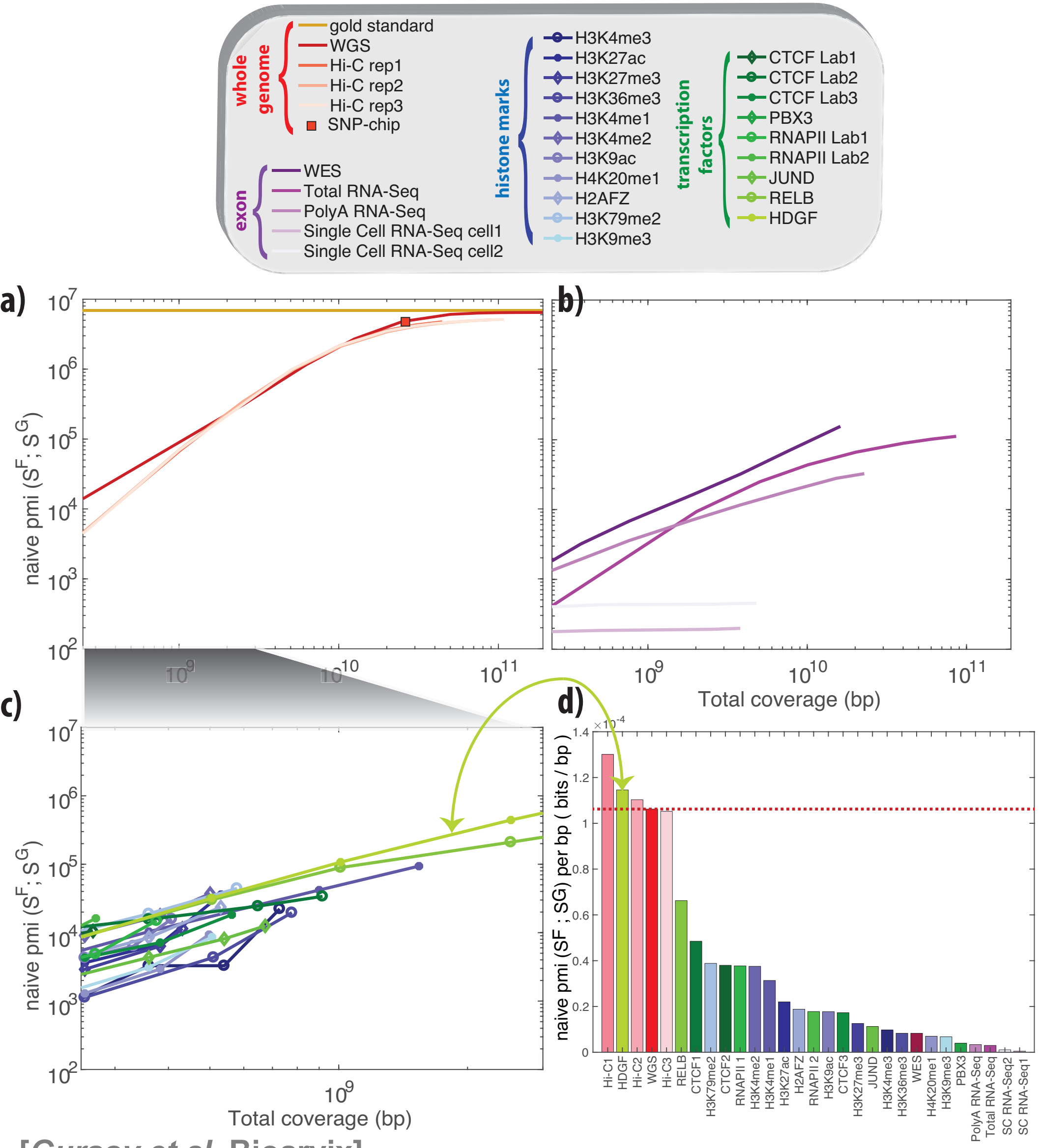
[*Gursoy et al,* Bioarxiv]



NA12878 as case study - 1000 genomes variants are used as gold standard

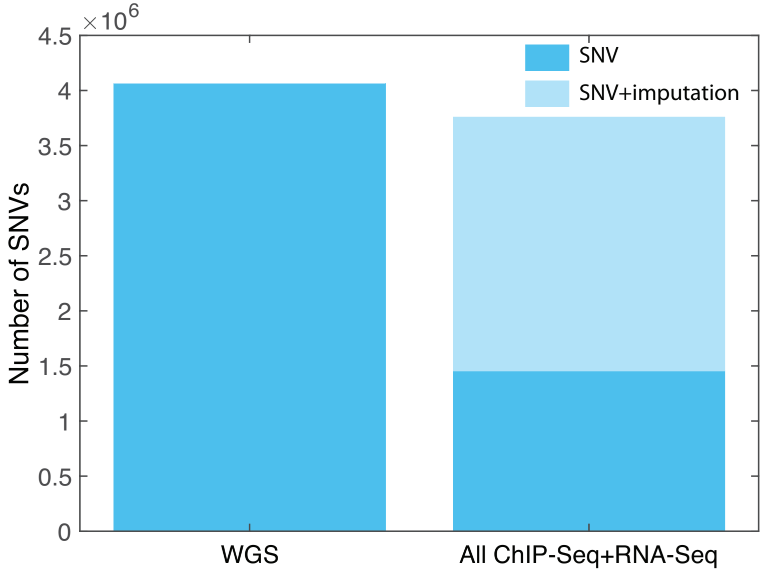
Leakage Source	Leaking Variants	# of potential variants	Average leakage per variant (bits)	Maximum leakage per variant (bits)	# of accessible variants	Total leakage (bits)
Raw reads	Exonic variants	2,682,417	0.10 ± 0.28	9.88 ± 2.12	246,893	24,689
Modified reads Q = {indels}	Exonic SNVs	2,607,969	0.09 ± 0.27	9.95 ± 2.02	231,031	207,92
Modified reads Q = {mismatches}	Exonic indels	51,408	0.33 ± 0.47	7.64 ± 2.42	15,862	5234
Signal profiles	Exonic deletions	48,019	0.29 ± 0.45	7.97 ± 2.42	1,067	298
Gene expression quantification	eQTLs	3,175	1.19 ± 0.36	4.00 ± 1.92	158	188

- **How much information, for example, do RNA-Seq reads (or ChIP-Seq) reads contain? Does that information enough to identify individuals?**



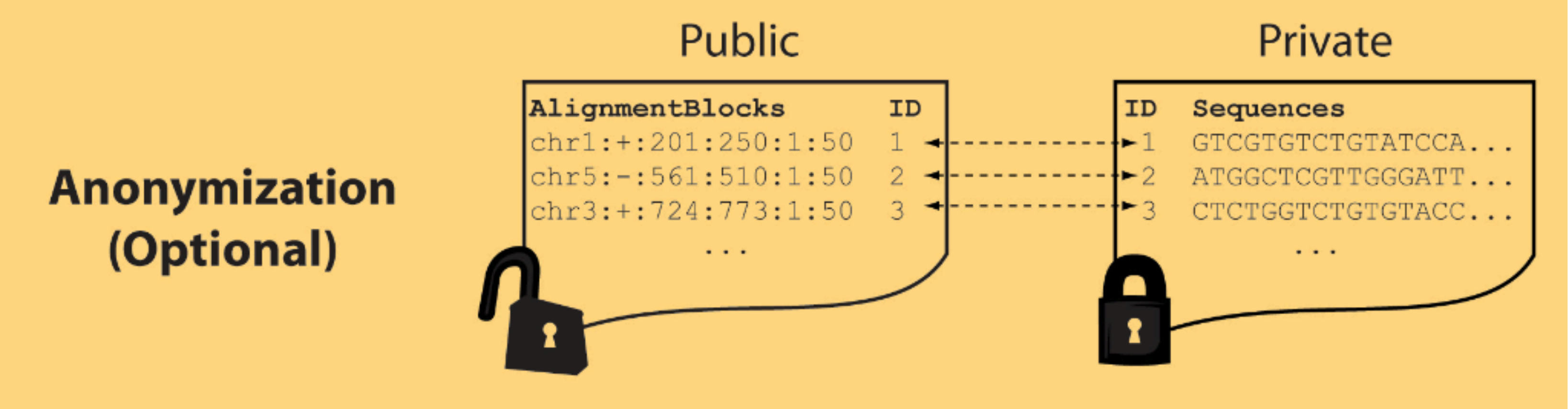
[*Gursoy et al,* Bioarvix]

- **It might seem like we don't infer much information from single ChIP-Seq and RNA-Seq experiments compared to WGS**
  - **However putting 10 different ChIP-Seq experiments and RNA-Seq together with imputation provides a great deal of information about the individual**



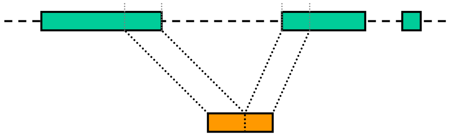
Light-weight formats to Hide Most of the Read Data (Signal Tracks)

- Some lightweight format clearly separate public & private info., aiding exchange
- Files become much smaller. Similar to CRAM
- Distinction between formats to compute on and those to archive with – become sharper with big data



AlignmentBlocks	ID
chr1:+:201:250:1:50	1
chr5:-:561:510:1:50	2
chr3:+:724:773:1:50	3
...	

ID	Sequences
1	GTCGTGTCTGTATCCA...
2	ATGGCTCGTTGGGATT...
3	CTCTGGTCTGTGTACC...
	...



Mapping coordinates without variants (MRF)

Reads (linked via ID, 10X larger than mapping coord.)

[*Bioinformatics* 27: 281]

Using population-scale functional genomics to understand neuropsychiatric disease & interpreting the data exhaust from this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + dev. 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 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 disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as targets.

- *[Exhaust]* **Other uses** for the resource
  - Highlighting aging related genes + consistently comparing the brain to other organs
- *[Exhaust]* **Genomic Privacy**
  - The **Dilemma**
    - The genome as fundamental, inherited info that's very private v. need for large-scale mining for med. research
    - 2-sided nature of RNA-seq presents tricky privacy issues
  - **eQTLs**: Quantifying & removing variant info from expression levels with ICI & predictability. Instantiating a practical linking attack with noisy quasi-identifiers
  - **Signal Profiles**: Manifest appreciable leakage from large & small deletions. Linking attacks possible but additional complication of SV discovery in addition to genotyping

Representative Functional Genomics, Genotype, eQTL Datasets

- Genotypes are available from the 1000 Genomes Project
- mRNA sequencing for 462 individuals from gEUVADIS and ENCODE
  - Publicly available quantification for protein coding genes
- Functional genomics data (ChIP-Seq, RNA-Seq, Hi-C) available from ENCODE
- Approximately 3,000 cis-eQTL (FDR<0.05)

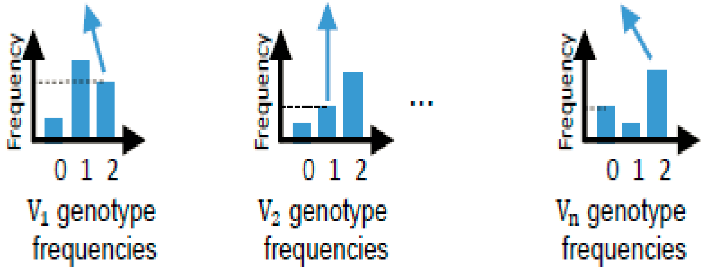
1000 Genomes
A Deep Catalog of Human Genetic Variation

GEUVADIS
CONSORTIUM

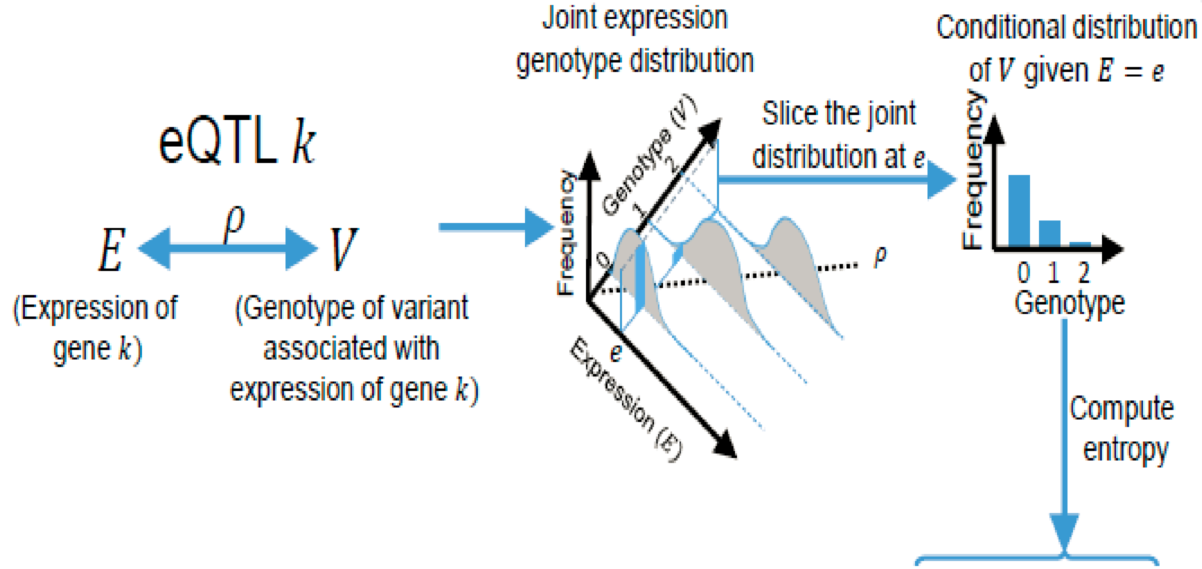
ENCODE

Information Content and Predictability

$$ICI\left(\begin{array}{c}\text{Individual has variant}\\ \text{genotypes } g_1, g_2, \ldots, g_n\\ \text{for variants } V_1, V_1, \ldots, V_n\end{array}\right) = \log\left(\frac{1}{\begin{array}{c}\text{Frequency of}\\ V_1 \text{ genotype}\\ g_1 = 2\end{array}}\right) + \log\left(\frac{1}{\begin{array}{c}\text{Frequency of}\\ V_2 \text{ genotype}\\ g_2 = 1\end{array}}\right) + \ldots + \log\left(\frac{1}{\begin{array}{c}\text{Frequency of}\\ V_n \text{ genotype}\\ g_n = 2\end{array}}\right)$$



- **Naive measure of information (no LD, distant correlations, pop. struc., &c)**
- **Higher frequency: Lower ICI**
- **Additive for multiple variants**

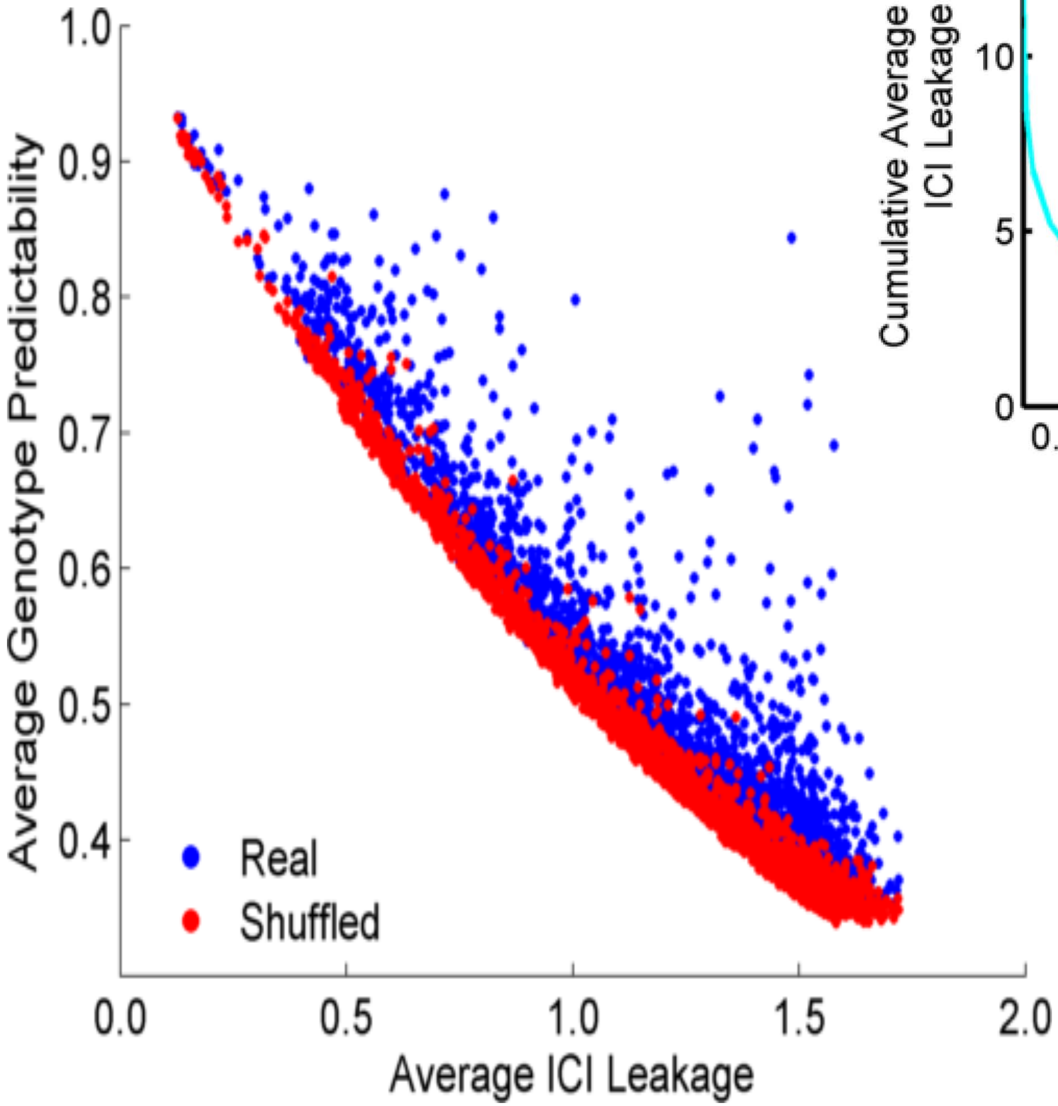


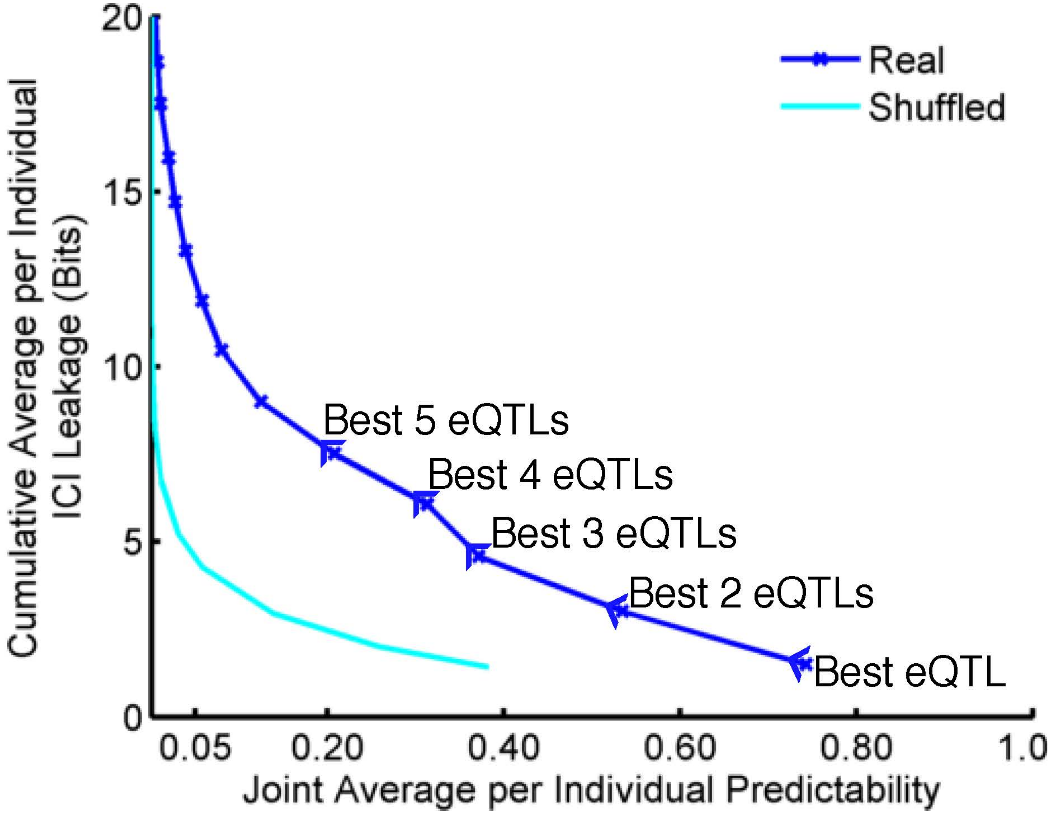
$$\pi\left(V \middle| E = e\right) = e^{-H\left(V \middle| E = e\right)}$$

- **Condition specific entropy**
- **Higher cond. entropy: Lower predictability**
- **Additive for multiple eQTLs**

[Harmanci et al. Nat. Meth. 2016]

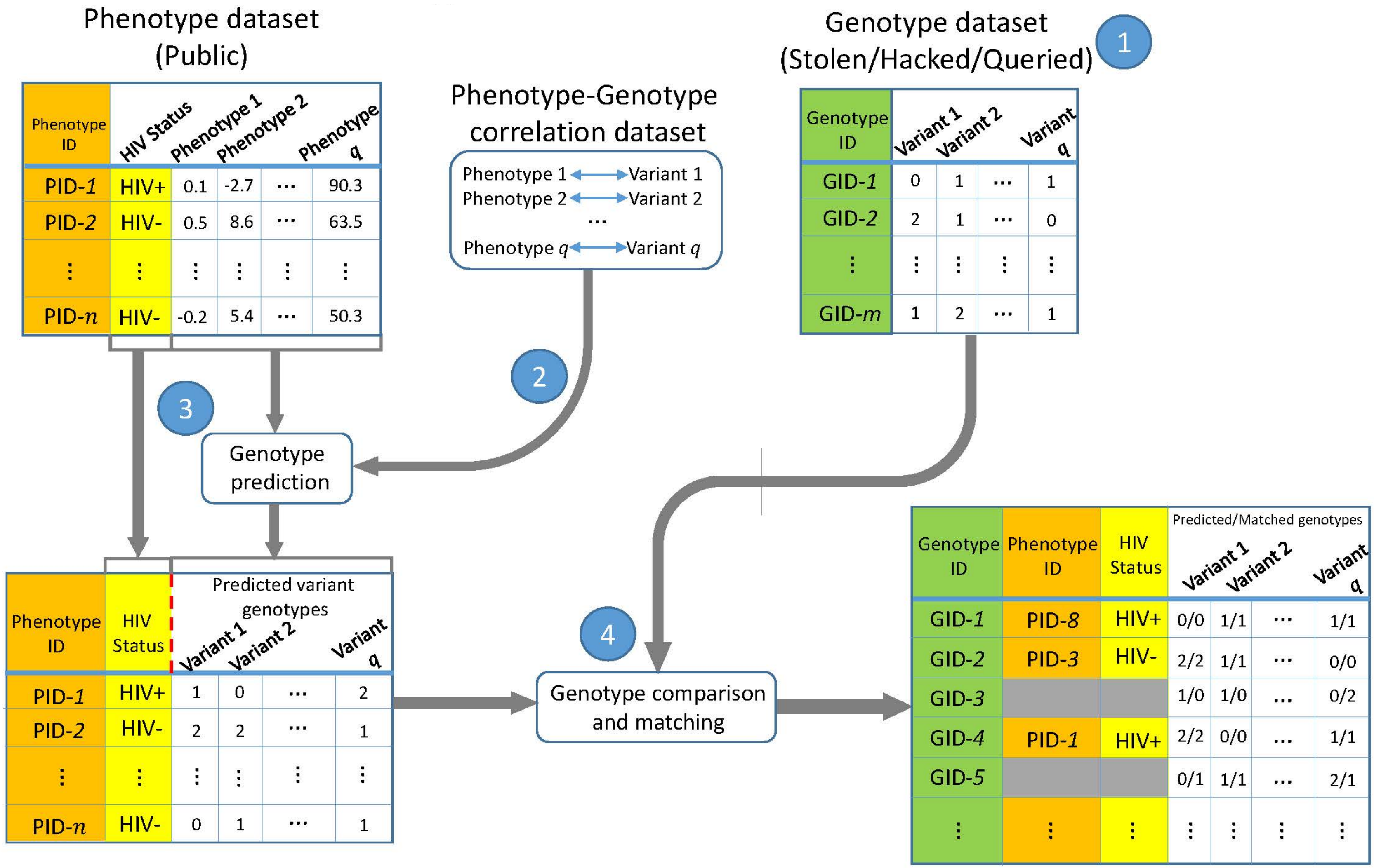
[Harmanciet al. Nat. Meth. ('16]





ICI Leakage versus Genotype Predictability

Linking Attack Scenario



Phenotype dataset
(Public)

Phenotype ID	HIV Status	Phenotype 1	Phenotype 2		Phenotype *q*
PID-*1*	HIV+	0.1	-2.7	···	90.3
PID-*2*	HIV-	0.5	8.6	···	63.5
⋮	⋮	⋮	⋮	⋮	⋮
PID-*n*	HIV-	-0.2	5.4	···	50.3

Phenotype-Genotype
correlation dataset

Phenotype 1 ⟷ Variant 1
Phenotype 2 ⟷ Variant 2
···
Phenotype *q* ⟷ Variant *q*

Genotype dataset
(Stolen/Hacked/Queried)

Genotype ID	Variant 1	Variant 2		Variant *q*
GID-*1*	0	1	···	1
GID-*2*	2	1	···	0
⋮	⋮	⋮	⋮	⋮
GID-*m*	1	2	···	1

Genotype
prediction

Phenotype ID	HIV Status	Predicted variant genotypes			
		Variant 1	Variant 2		Variant *q*
PID-*1*	HIV+	1	0	···	2
PID-*2*	HIV-	2	2	···	1
⋮	⋮	⋮	⋮	⋮	⋮
PID-*n*	HIV-	0	1	···	1

Genotype comparison
and matching

Genotype ID	Phenotype ID	HIV Status	Predicted/Matched genotypes			
			Variant 1	Variant 2		Variant *q*
GID-*1*	PID-*8*	HIV+	0/0	1/1	···	1/1
GID-*2*	PID-*3*	HIV-	2/2	1/1	···	0/0
GID-*3*			1/0	1/0	···	0/2
GID-*4*	PID-*1*	HIV+	2/2	0/0	···	1/1
GID-*5*			0/1	1/1	···	2/1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

[Harmanciet al. Nat. Meth. (in revision)]

Linking Attacks: Case of Netflix Prize

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
…	…	…	…
…	…	…	…

Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
…	…	…	…
…	…	…	…
…	…	…	…

- **Many users are shared**
- **The grades of same users are correlated**
- **A user grades one movie around the same date in two databases**

Anonymized Netflix Prize Training Dataset
made available to contestants

Linking Attacks: Case of Netflix Prize

NETFLIX ⟷ IMDb

Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
…	…	…	…
…	…	…	…

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
…	…	…	…
…	…	…	…
…	…	…	…

- **Many users are shared**
- **The grades of same users are correlated**
- **A user grades one movie around the same date in two databases**

- **IMDB users are public**

- **NetFLIX and IMdB moves are public**

Linking Attacks: Case of Netflix Prize

NETFLIX

IMDb

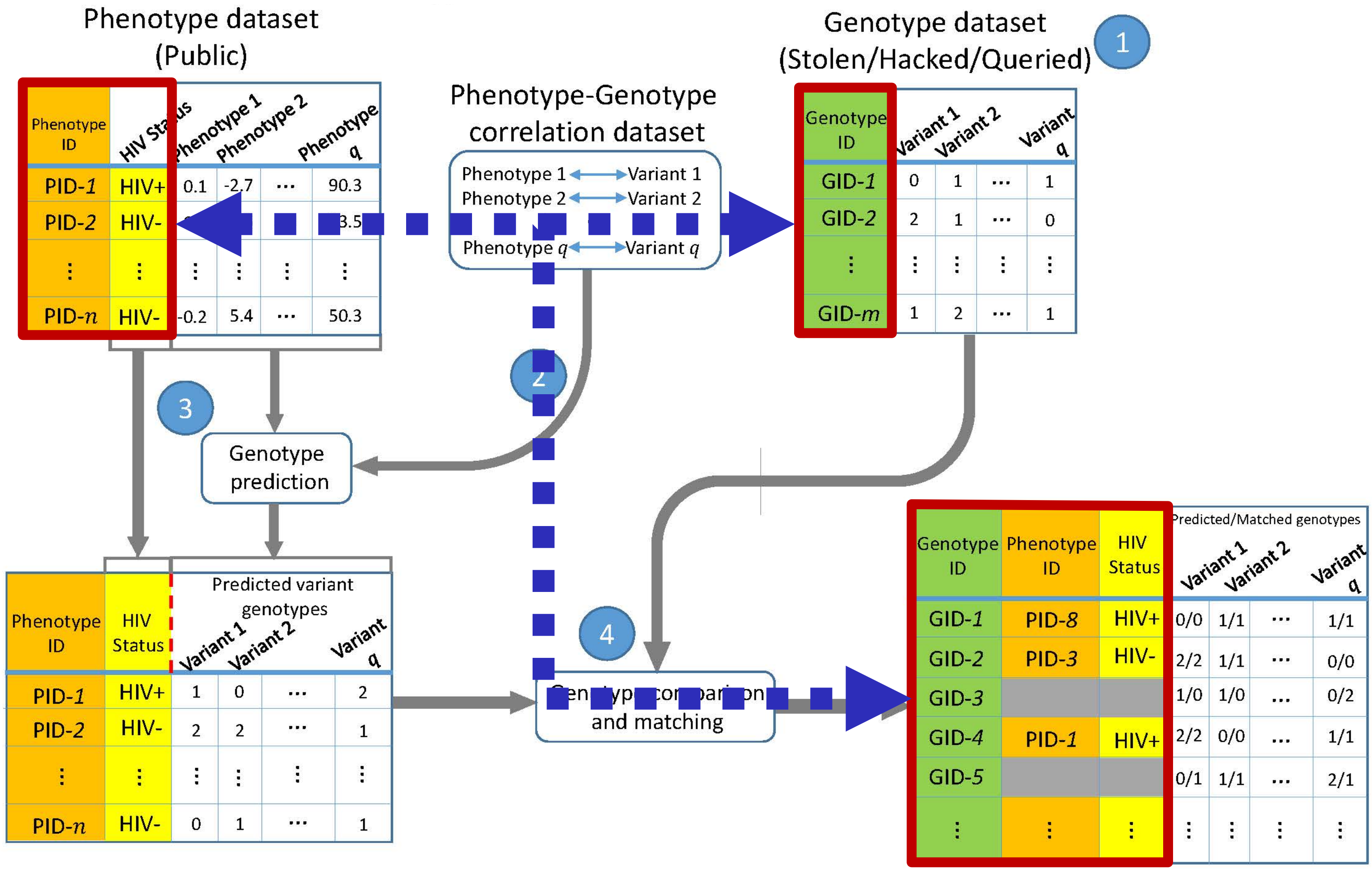
Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	**NTFLX-666**	6/6/2016	5
…	…	…	…
…	…	…	…

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
…	…	…	…
…	…	…	…
…	…	…	…

- **Many users are shared**
- **The grades of same users are correlated**
- **A user grades one movie around the same date in two databases**

Linking Attack Scenario



Phenotype dataset (Public)

Phenotype ID	HIV Status	Phenotype 1	Phenotype 2	···	Phenotype *q*
PID-*1*	HIV+	0.1	-2.7	···	90.3
PID-*2*	HIV-	[illegible]	[illegible]	[illegible]	3.5
⋮	⋮	⋮	⋮	⋮	⋮
PID-*n*	HIV-	-0.2	5.4	···	50.3

Phenotype-Genotype correlation dataset

Phenotype 1 ↔ Variant 1
Phenotype 2 ↔ Variant 2
Phenotype *q* ↔ Variant *q*

Genotype dataset (Stolen/Hacked/Queried)

Genotype ID	Variant 1	Variant 2	···	Variant *q*
GID-*1*	0	1	···	1
GID-*2*	2	1	···	0
⋮	⋮	⋮	⋮	⋮
GID-*m*	1	2	···	1

Genotype prediction

Phenotype ID	HIV Status	Predicted variant genotypes Variant 1	Variant 2	···	Variant *q*
PID-*1*	HIV+	1	0	···	2
PID-*2*	HIV-	2	2	···	1
⋮	⋮	⋮	⋮	⋮	⋮
PID-*n*	HIV-	0	1	···	1

Genotype comparison and matching

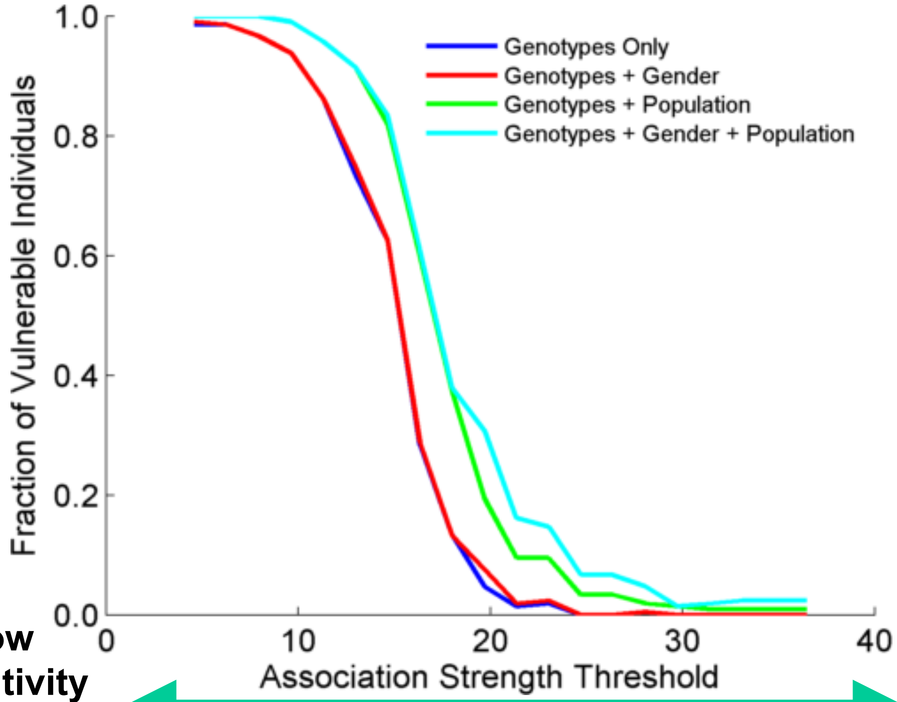
Genotype ID	Phenotype ID	HIV Status	Predicted/Matched genotypes Variant 1	Variant 2	···	Variant *q*
GID-*1*	PID-*8*	HIV+	0/0	1/1	···	1/1
GID-*2*	PID-*3*	HIV-	2/2	1/1	···	0/0
GID-*3*			1/0	1/0	···	0/2
GID-*4*	PID-*1*	HIV+	2/2	0/0	···	1/1
GID-*5*			0/1	1/1	···	2/1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

[Harmanciet al. Nat. Meth. (in revision)]

Success in Linking Attack with Extremity based Genotype Prediction

200 individuals eQTL Discovery
200 individuals in Linking Attack

High Sensitivity



Low Sensitivity

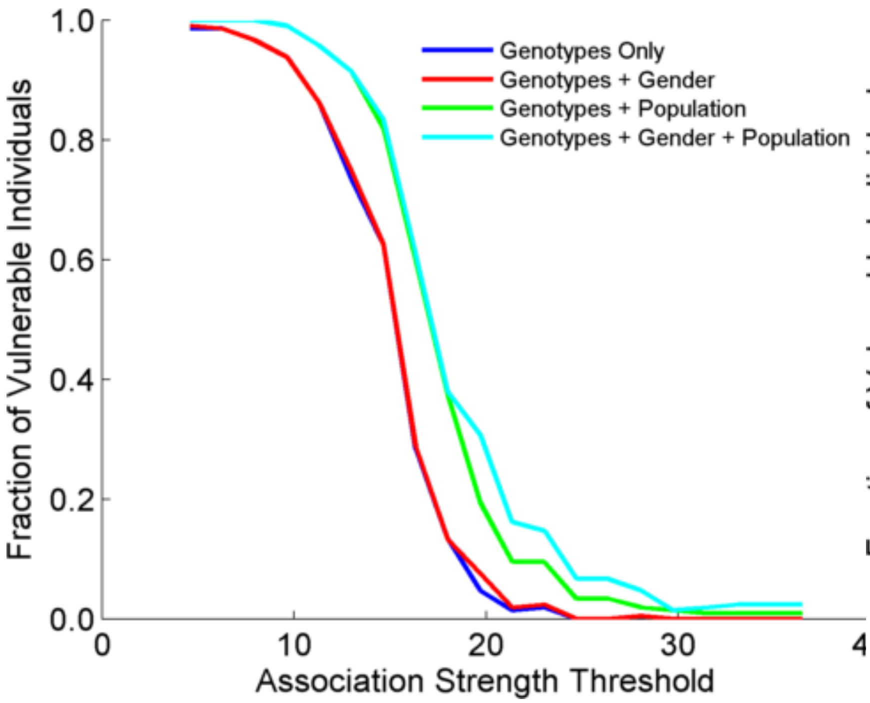
High Number Of eQTLs

Low Number Of eQTLs

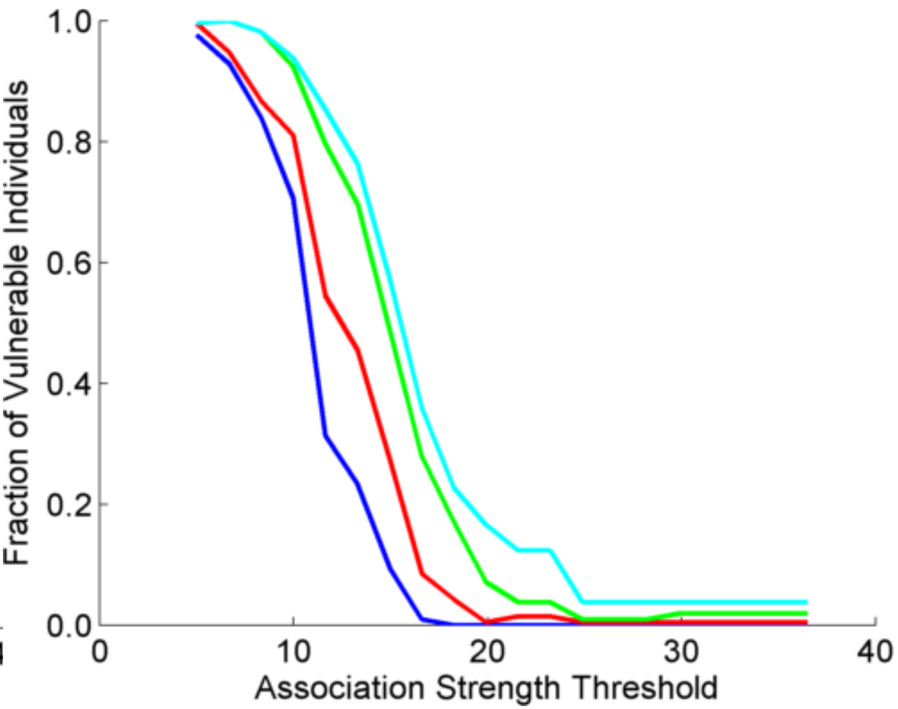
[Harmanci et al. *Nat. Meth.* (16)]

Success in Linking Attack with Extremity based Genotype Prediction

200 individuals eQTL Discovery
200 individuals in Linking Attack



200 individuals eQTL Discovery
100,200 individuals in Linking Attack



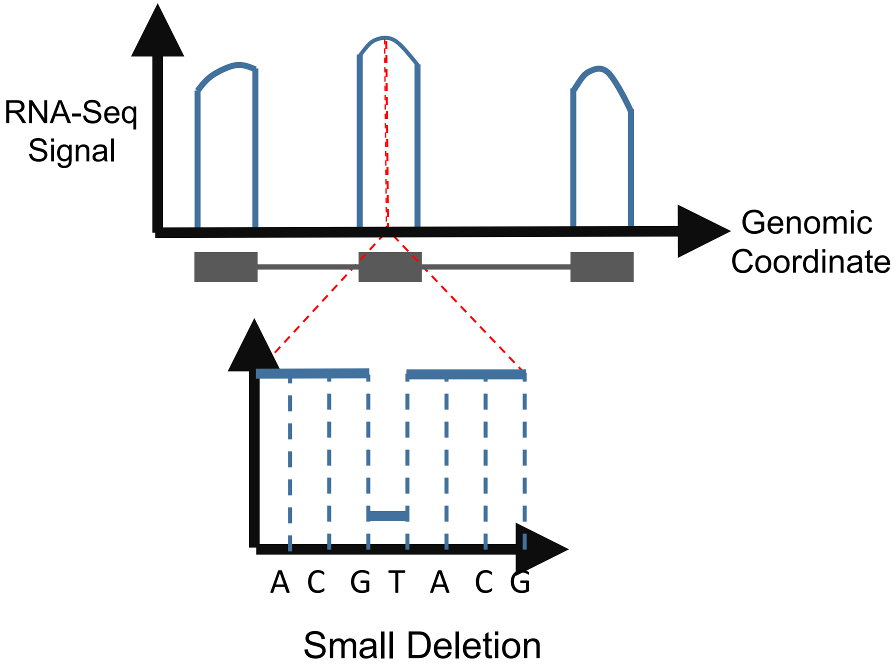
[Harmanci et al. *Nat. Meth.* (16)]

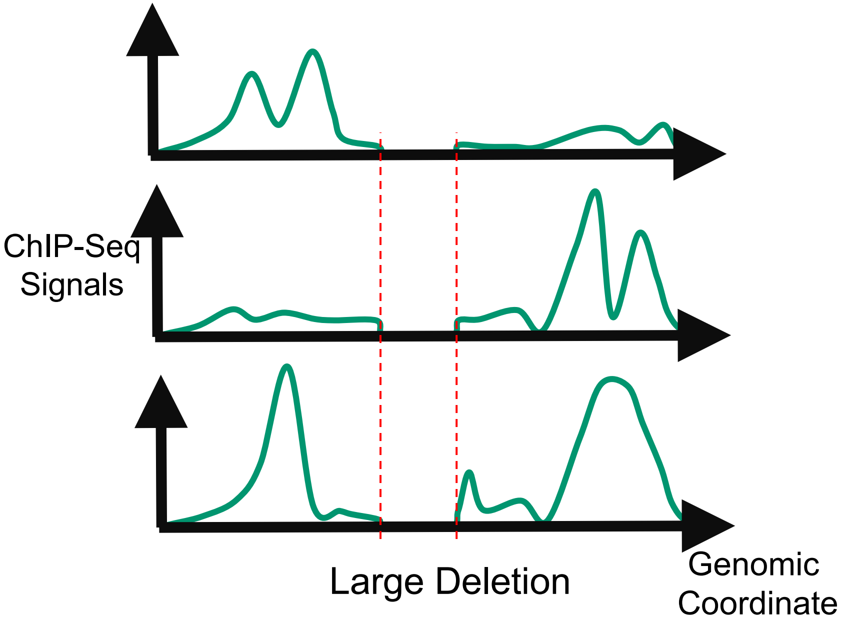
Using population-scale functional genomics to understand neuropsychiatric disease & interpreting the data exhaust from this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + dev. 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 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 disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as targets.

- *[Exhaust]* **Other uses** for the resource
  - Highlighting aging related genes + consistently comparing the brain to other organs
- *[Exhaust]* **Genomic Privacy**
  - The **Dilemma**
    - The genome as fundamental, inherited info that's very private v. need for large-scale mining for med. research
    - 2-sided nature of RNA-seq presents tricky privacy issues
  - **eQTLs**: Quantifying & removing variant info from expression levels with ICI & predictability. Instantiating a practical linking attack with noisy quasi-identifiers
  - **Signal Profiles**: Manifest appreciable leakage from large & small deletions. Linking attacks possible but additional complication of SV discovery in addition to genotyping

Detection & Genotyping of small & large SV deletions from signal profiles

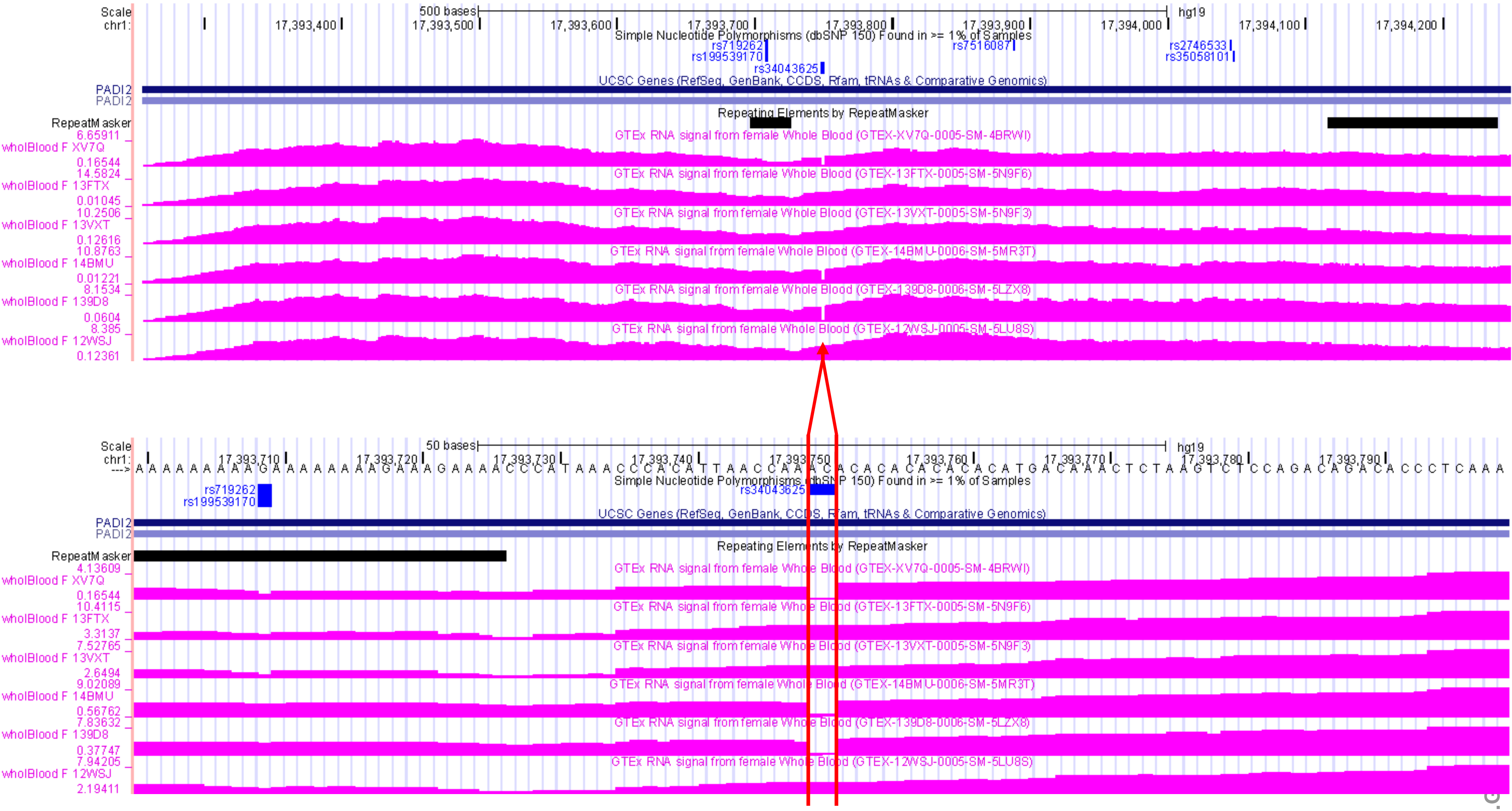




RNA-seq also shows large deletions

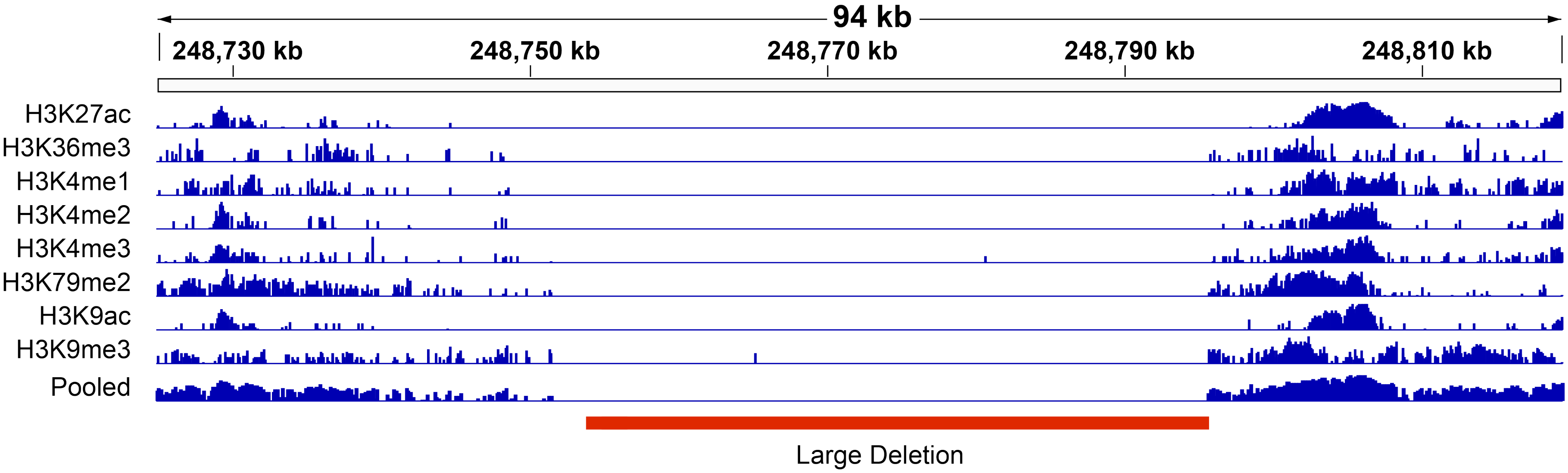
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Example of Small Deletion Evident in Signal Profile



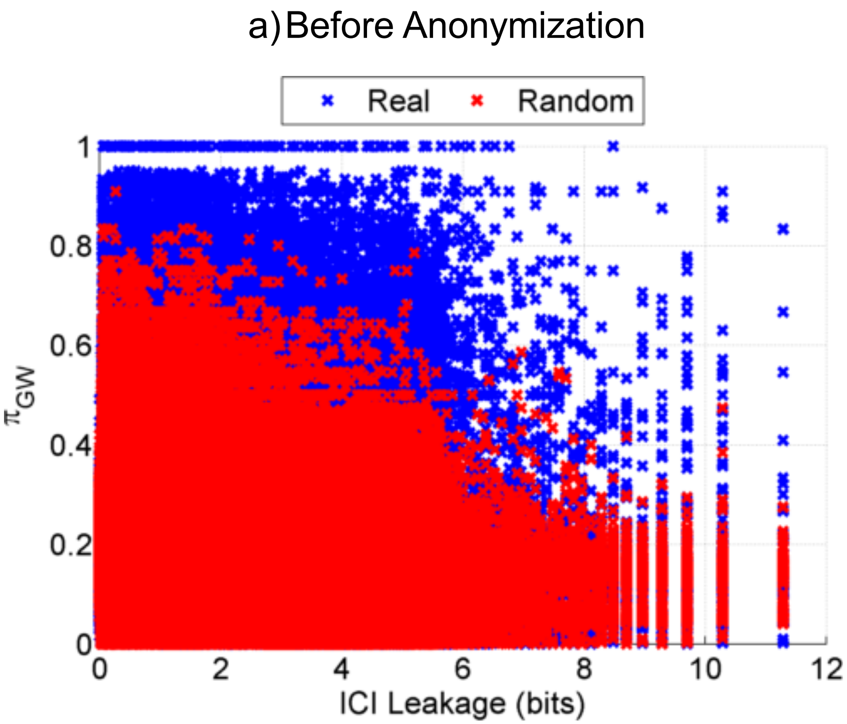
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

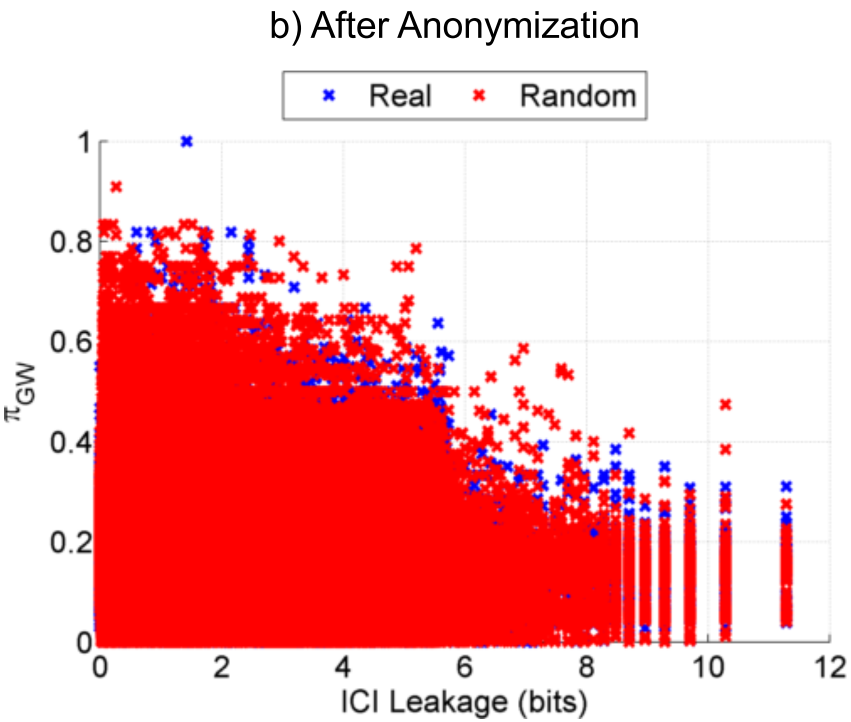
Example of Large Deletion Evident in Signal Profile



[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Information Leakage from SV Deletions

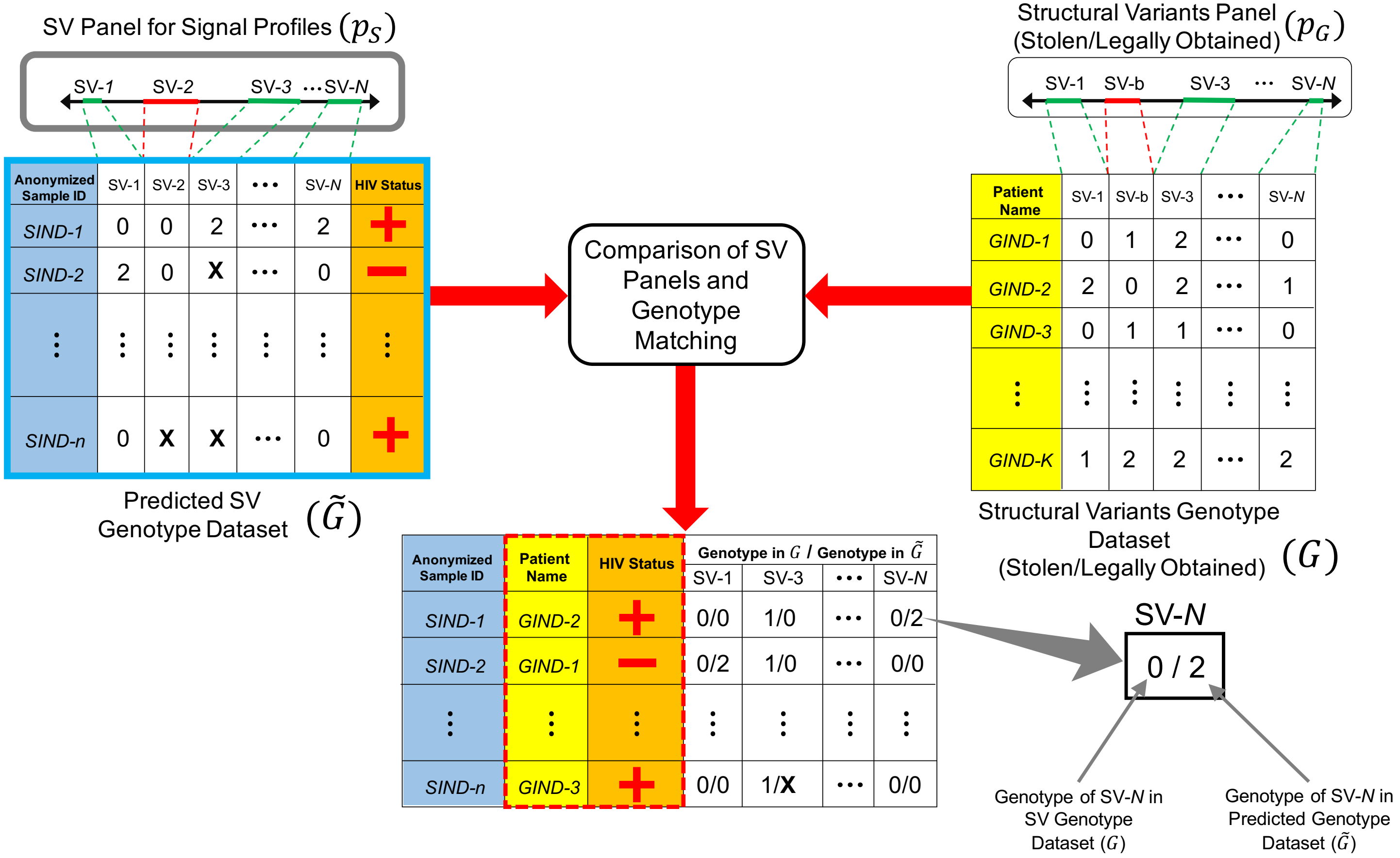

a) Before Anonymization


b) After Anonymization

Simple anonymization procedure (filling in deletion by value at endpoints) has dramatic effect

[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Another type of Linking Attack: Linking based on SV Genotyping



Anonymized Sample ID	SV-1	SV-2	SV-3	···	SV-N	HIV Status
SIND-1	0	0	2	···	2	+
SIND-2	2	0	X	···	0	−
⋮	⋮	⋮	⋮	⋮	⋮	⋮
SIND-n	0	X	X	···	0	+

Predicted SV Genotype Dataset $(\tilde{G})$

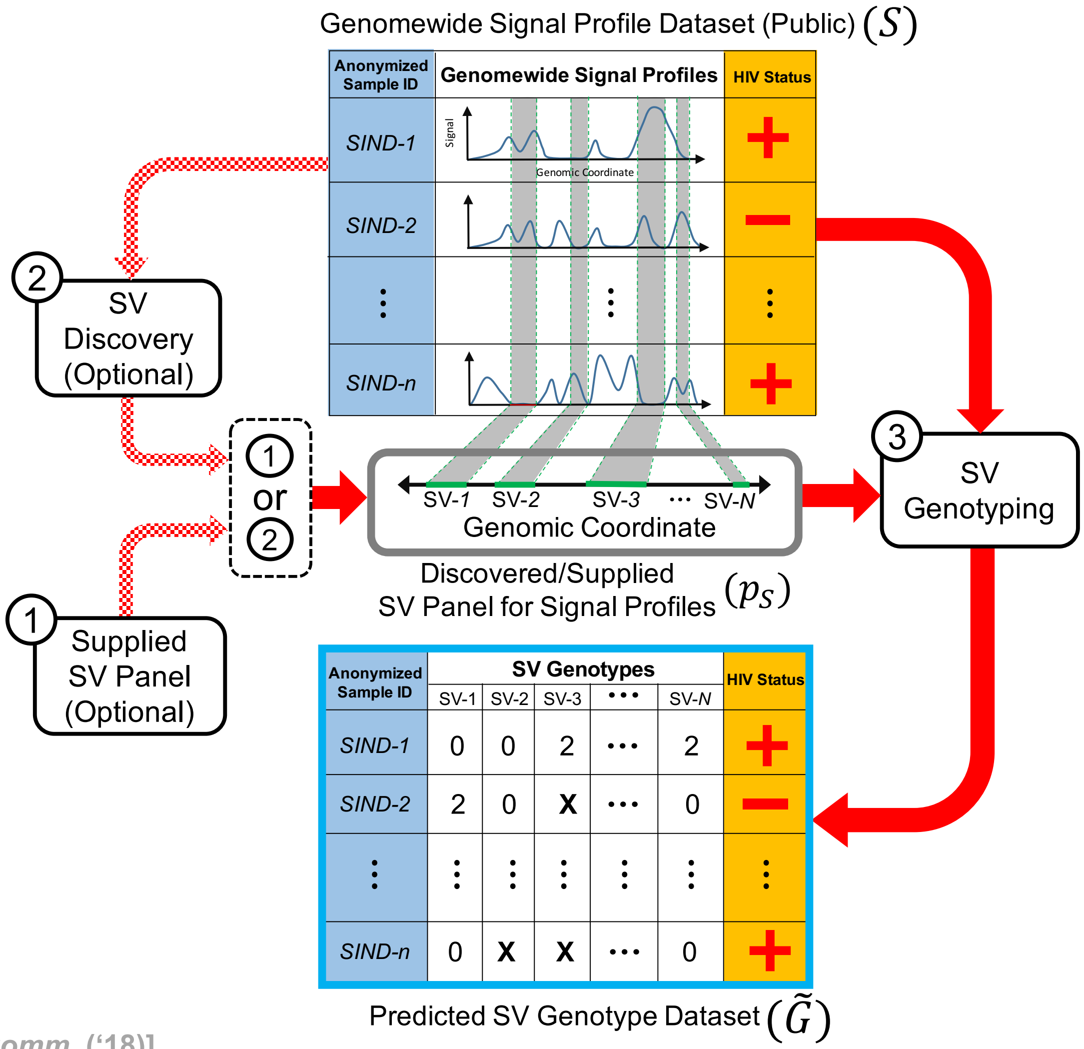
Patient Name	SV-1	SV-b	SV-3	···	SV-N
GIND-1	0	1	2	···	0
GIND-2	2	0	2	···	1
GIND-3	0	1	1	···	0
⋮	⋮	⋮	⋮	⋮	⋮
GIND-K	1	2	2	···	2

Structural Variants Genotype Dataset (Stolen/Legally Obtained) (G)

Anonymized Sample ID	Patient Name	HIV Status	Genotype in G / Genotype in $\tilde{G}$			
			SV-1	SV-3	···	SV-N
SIND-1	GIND-2	+	0/0	1/0	···	0/2
SIND-2	GIND-1	−	0/2	1/0	···	0/0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
SIND-n	GIND-3	+	0/0	1/X	···	0/0

[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Another type of Linking Attack:
First Doing SV Genotyping



Genomewide Signal Profile Dataset (Public) (S)

Anonymized Sample ID	Genomewide Signal Profiles	HIV Status
SIND-1		+
SIND-2		−
⋮	⋮	⋮
SIND-n		+

Discovered/Supplied SV Panel for Signal Profiles (p_S)

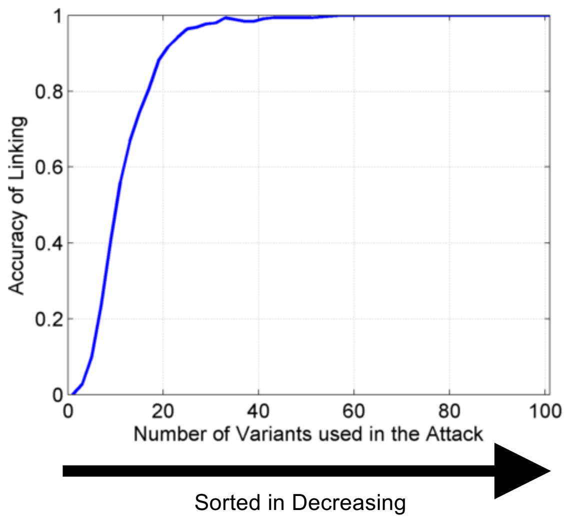
Anonymized Sample ID	SV Genotypes					HIV Status
	SV-1	SV-2	SV-3	•••	SV-*N*	
SIND-1	0	0	2	•••	2	+
SIND-2	2	0	X	•••	0	−
⋮	⋮	⋮	⋮	⋮	⋮	⋮
SIND-n	0	X	X	•••	0	+

Predicted SV Genotype Dataset $(\tilde{G})$

[Harmanci & Gerstein, *Nat. Comm.* ('18)]

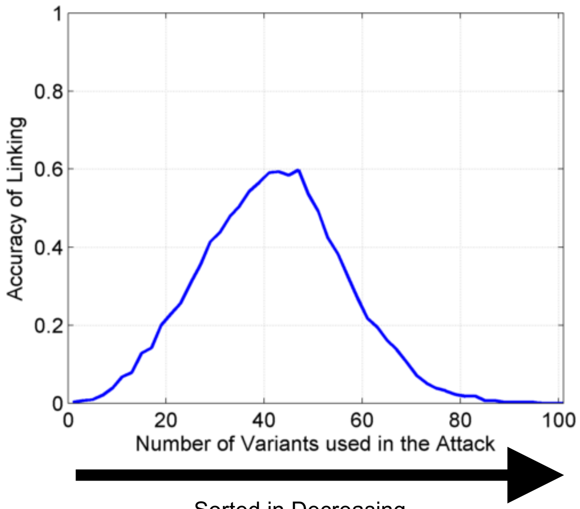
Linking Attack Based on SV Deletions in gEUVADIS Dataset

c) Genotyping
(1kG MAF>0.01)



Sorted in Decreasing
Predictability

d) Discovery + Genotyping



Sorted in Decreasing
Predictability

[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Using population-scale functional genomics to understand neuropsychiatric disease & interpreting the data exhaust from this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + dev. 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 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 disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as targets.

- *[Exhaust]* **Other uses** for the resource
  - Highlighting aging related genes + consistently comparing the brain to other organs
- *[Exhaust]* **Genomic Privacy**
  - The **Dilemma**
    - The genome as fundamental, inherited info that's very private v. need for large-scale mining for med. research
    - 2-sided nature of RNA-seq presents tricky privacy issues
  - **eQTLs**: Quantifying & removing variant info from expression levels with ICI & predictability. Instantiating a practical linking attack with noisy quasi-identifiers
  - **Signal Profiles**: Manifest appreciable leakage from large & small deletions. Linking attacks possible but additional complication of SV discovery in addition to genotyping

Using population-scale functional genomics to understand neuropsychiatric disease & interpreting the data exhaust from this activity

- *[Core]* **PsychENCODE:** Population-level analysis of functional genomics data related to neuropsychiatric disease
  - Construction of an adult brain resource with 1866 individuals + dev. 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 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 disease from genotype & transcriptome. Using this to suggest specific pathways & genes, as targets.
- *[Exhaust]* **Other uses** for the resource
  - Highlighting aging related genes + consistently comparing the brain to other organs
- *[Exhaust]* **Genomic Privacy**
  - The **Dilemma**
    - The genome as fundamental, inherited info that’s very private v. need for large-scale mining for med. research
    - 2-sided nature of RNA-seq presents tricky privacy issues
  - **eQTLs**: Quantifying & removing variant info from expression levels with ICI & predictability. Instantiating a practical linking attack with noisy quasi-identifiers
  - **Signal Profiles**: Manifest appreciable leakage from large & small deletions. Linking attacks possible but additional complication of SV discovery in addition to genotyping

PsychENCODE Acknowledgment

National Institute
of Mental Health

- 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

Acknowledgements

papers.gersteinlab.org/subject/**privacy**

PrivaSig.gersteinlab.org
PrivaSeq.gersteinlab.org

A **Harmanci**,
D **Greenbaum**, G **Gürsoy**

Hiring Postdocs. See
JOBS.gersteinlab.org !

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2017.
  - 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