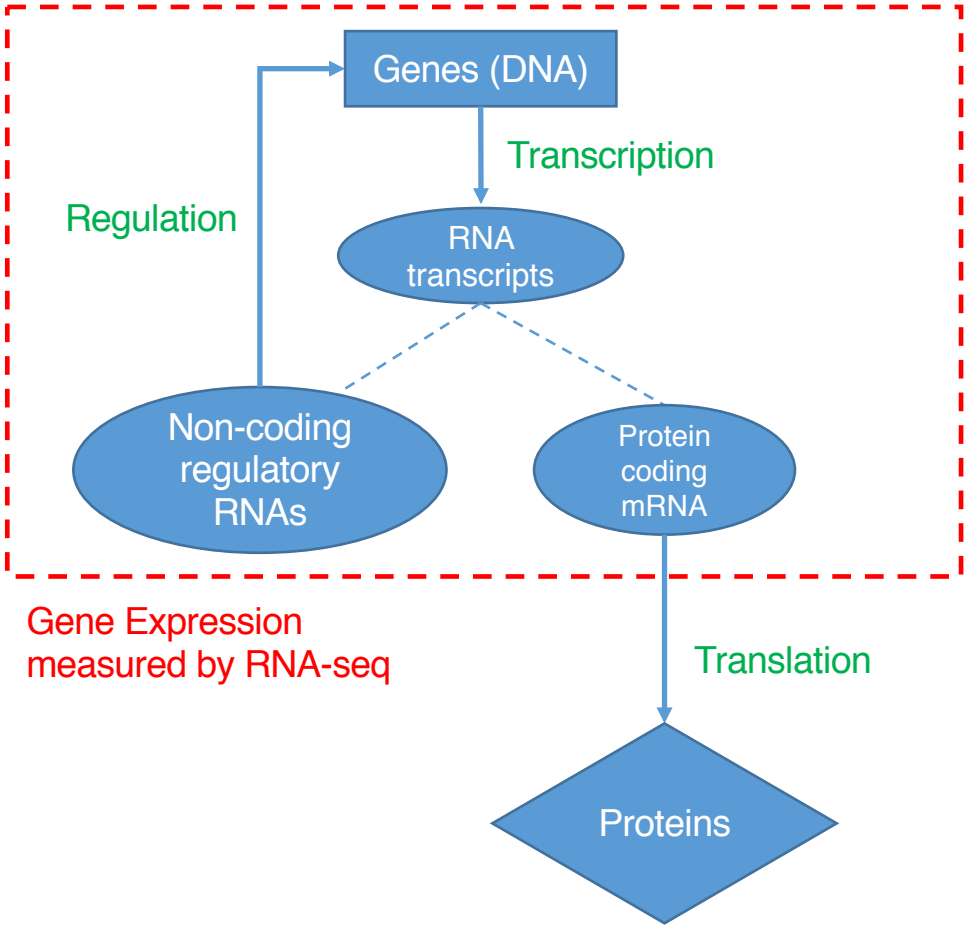
Transcriptome Mining:

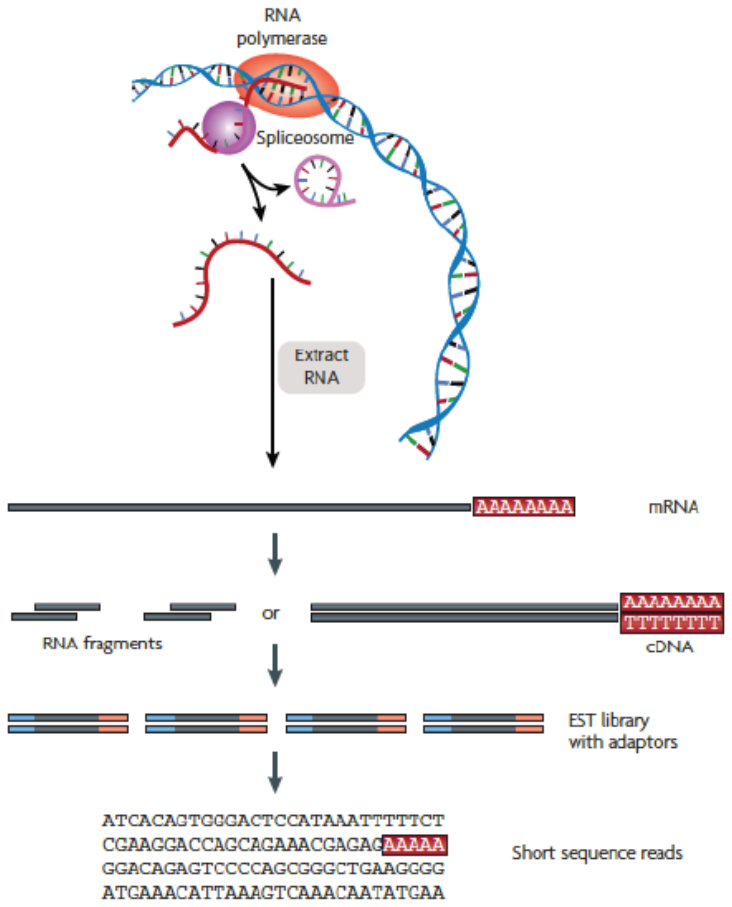
Population-scale genomic analysis
to better understand mental disease
& the subtle privacy risks
of this activity

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





[*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

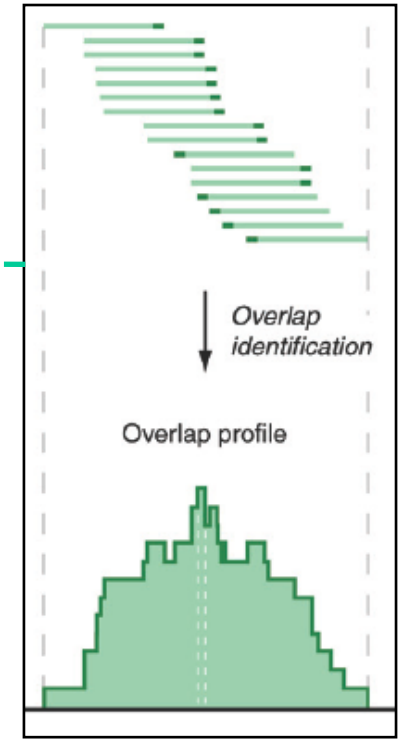
Fastq sequence files
~5-10 GB

```
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
TACCGATCGAGTCGACTGTAAACGTAGGCATA
ATTCTGACTGGTGTCATGCTGATGTACTTAAA
```

Index-building + Alignment to reference genome

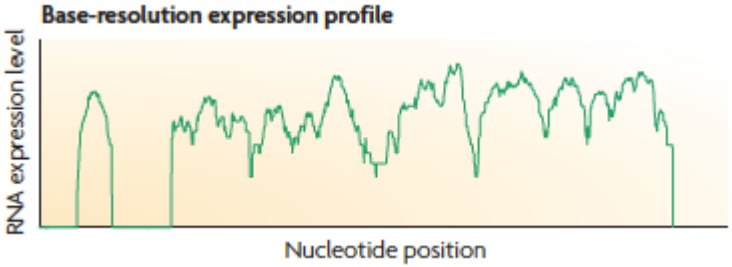
BAM files
~1-2-fold reduction

Conversion to signal track by overlapping reads



Reads => Signal

BigWig files
~25-fold reduction



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

Mapping to genes

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

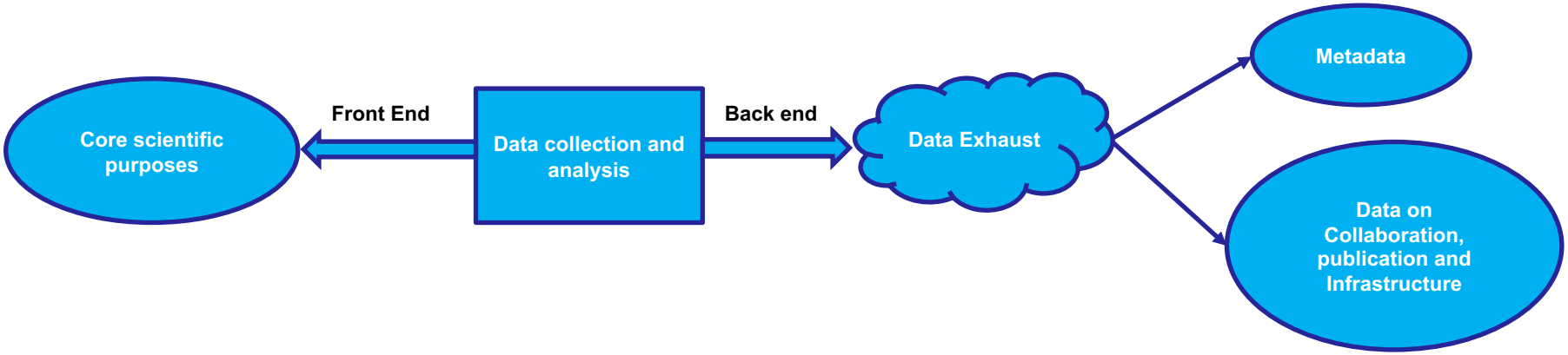
Some Core Science Qs Addressed by RNA-seq

- 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?

Studying large-scale transcriptome data also produces

Data Exhaust



- Data Exhaust = Exploitable byproducts of big data collection and analysis
- Creative use of Data is key to Data Science !

[PHOTO: RELAXNEWS; from http://www.lapresse.ca]

Transcriptome Mining: Population-scale genomic analysis to better understand mental disease & the subtle privacy risks of this activity

- *[Core]* **PsychENCODE:** Population-level analysis of functional genomics data related to mental disease
  - Consortium intro & construction of an adult brain resource w/ 1866 individuals
  - Explanation of across-population variation via changing proportions of cell types (using single-cell deconvolution)
  - Generation of a large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C & activity relationships. Using this to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype & transcriptome
- *[Exhaust]* **Genomic Privacy** & RNA-seq
  - Introduction to 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 particularly tricky privacy issues
  - Using file formats to remove obvious variants
  - eQTLs: Quantifying & removing further variant info from expression levels w/ ICI & predictability. Instantiating a practical linking attack w/ 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
- *[Exhaust]* **Publication Patterns** from data producing consortia
  - Co-authorship network stats relate to publication rollouts & show gradual adoption by community
  - Key role of brokers in data dissemination

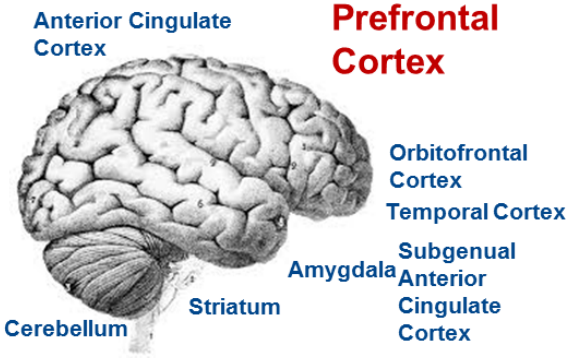
Transcriptome Mining: Population-scale genomic analysis to better understand mental disease & the subtle privacy risks of this activity

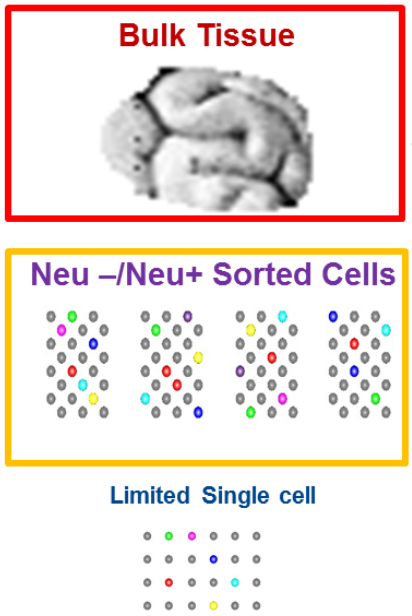
- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to mental disease
  - Consortium intro & construction of an adult brain resource w/ 1866 individuals
  - Explanation of across-population variation via changing proportions of cell types (using single-cell deconvolution)
  - Generation of a large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C & activity relationships. Using this to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype & transcriptome
- *[Exhaust]* **Genomic Privacy** & RNA-seq
  - Introduction to 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 particularly tricky privacy issues
  - Using file formats to remove obvious variants
  - eQTLs: Quantifying & removing further variant info from expression levels w/ ICI & predictability. Instantiating a practical linking attack w/ 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
- *[Exhaust]* **Publication Patterns** from data producing consortia
  - Co-authorship network stats relate to publication rollouts & show gradual adoption by community
  - Key role of brokers in data dissemination

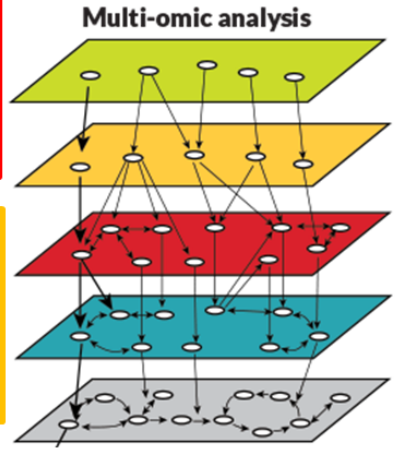
Sample Sources: >2,500 brains

Genome:
WGS, genotype

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







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

Transcriptome:
RNA-seq, lncRNAseq,

Proteome:
MWP, LC-MS/MS

Data Coordination/Analysis Center - Uniformly processed data across disorders and developmental time periods!

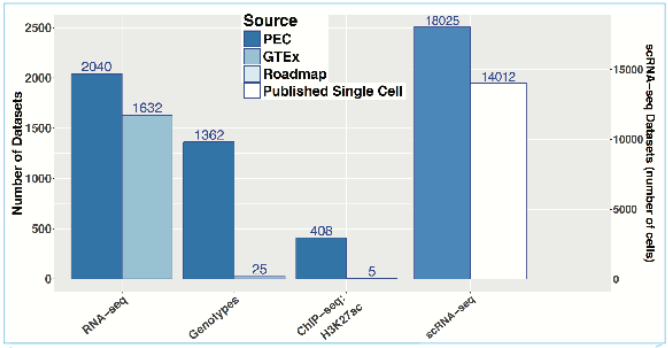
The PsychENCODE consortium

NIH National Institute of Mental Health

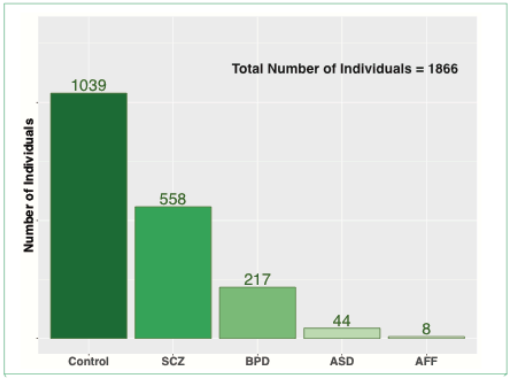
Some of the Qs addressed by PsychENCODE

- **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)
- **Moreover**, current models substantially underestimate heritability using genetic data
  - Schizophrenia : ~25%
- **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

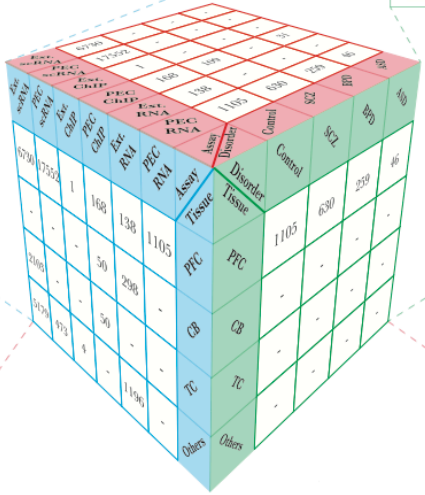
Collecting functional genomic datasets for the adult human brain from PsychENCODE, other large consortia & single cell studies

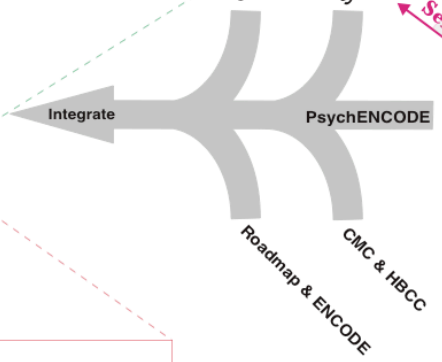


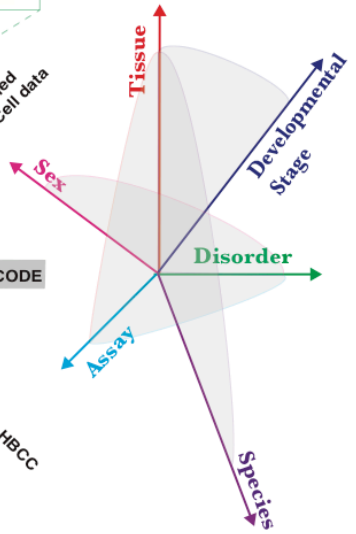
Assay



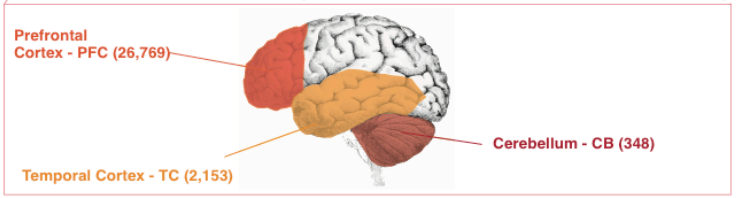
Disorder



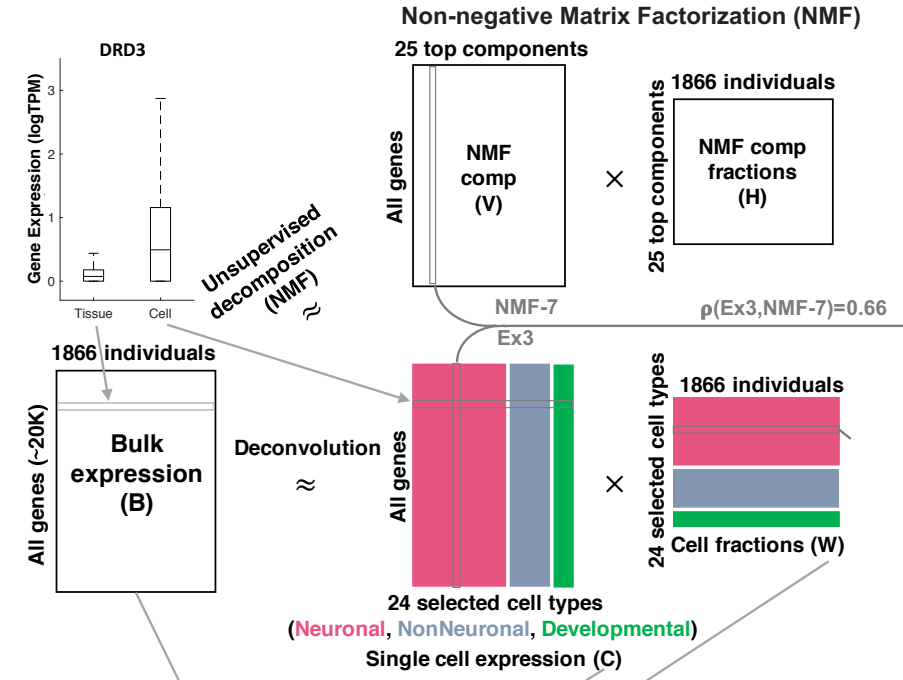


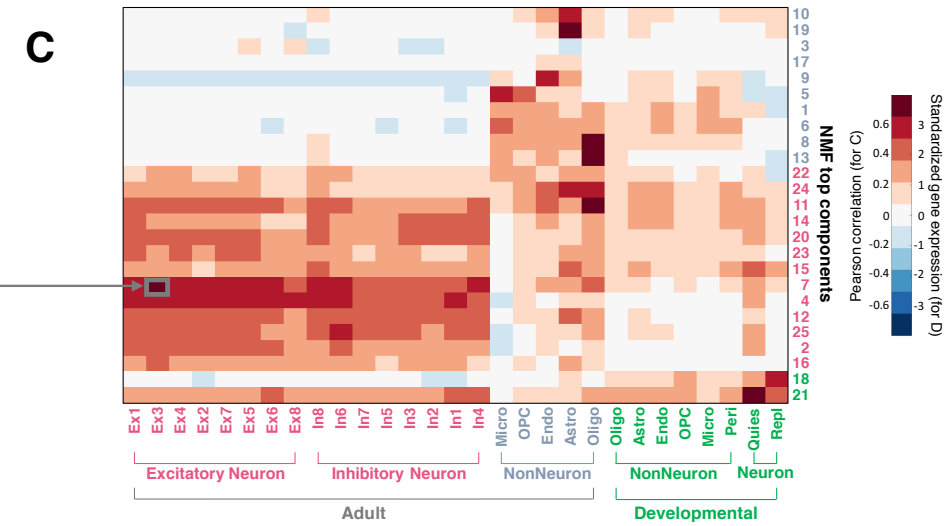


Tissue

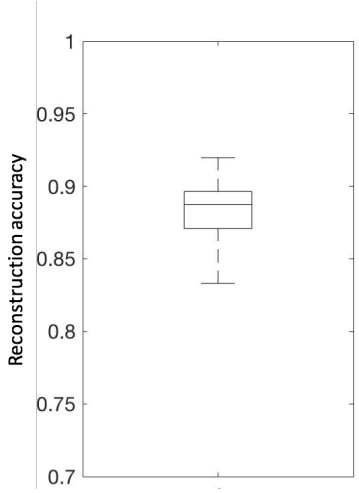


De-convolving bulk gene expression variation in brain PFC across a population





$$1 - \frac{\left\| B - CW \right\|^2}{\left\| B \right\|^2} > 85\%$$

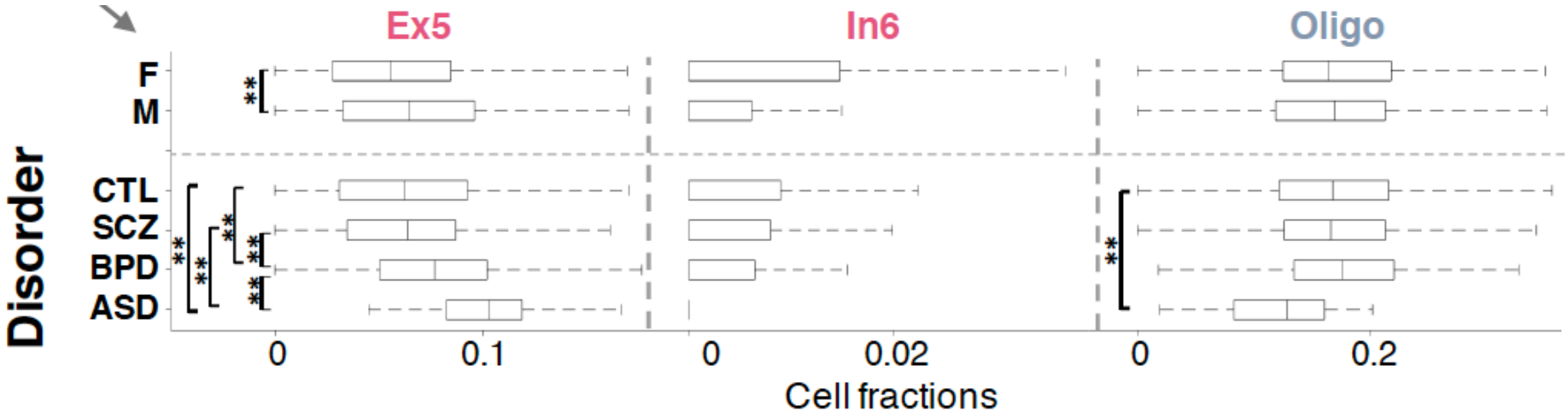


Consistent w/ **V** = unbiased bulk-derived signatures from NMF

Explain most of the variation in the bulk data expression matrix across 1866 individuals (**B**)

C = Expression signatures of basic cell types from single cell data

Neuronal & glial cell fraction change across gender & disorders



NB - Excitatory to Inhibitory imbalance at neuronal subtype level for ASD

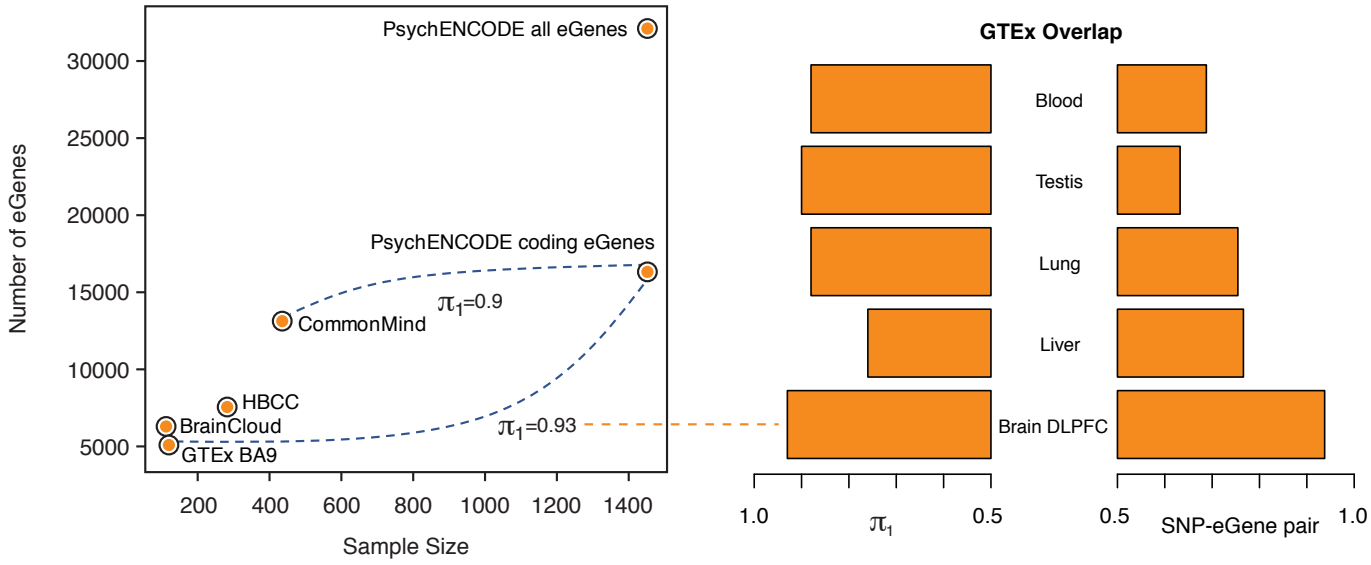
Rubenstein et al., Model of autism: increased ratio of excitation/inhibition in key neural systems, Genes Brain Behav. 2003

Transcriptome Mining: Population-scale genomic analysis to better understand mental disease & the subtle privacy risks of this activity

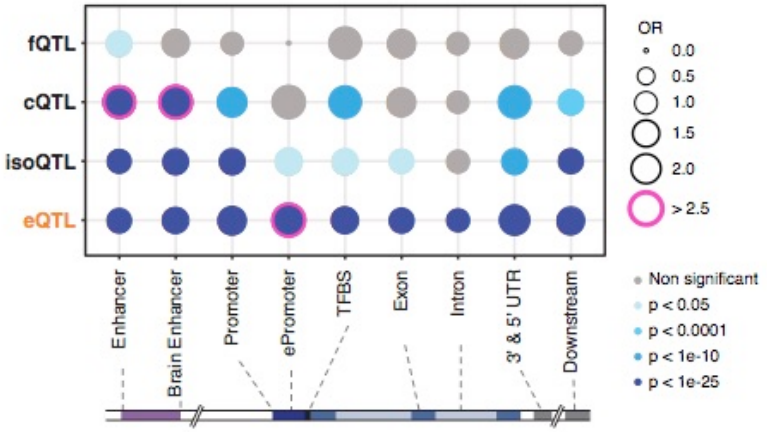
- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to mental disease
  - Consortium intro & construction of an adult brain resource w/ 1866 individuals
  - Explanation of across-population variation via changing proportions of cell types (using single-cell deconvolution)
  - Generation of a large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C & activity relationships. Using this to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype & transcriptome
- *[Exhaust]* **Genomic Privacy** & RNA-seq
  - Introduction to 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 particularly tricky privacy issues
  - Using file formats to remove obvious variants
  - eQTLs: Quantifying & removing further variant info from expression levels w/ ICI & predictability. Instantiating a practical linking attack w/ 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
- *[Exhaust]* **Publication Patterns** from data producing consortia
  - Co-authorship network stats relate to publication rollouts & show gradual adoption by community
  - Key role of brokers in data dissemination

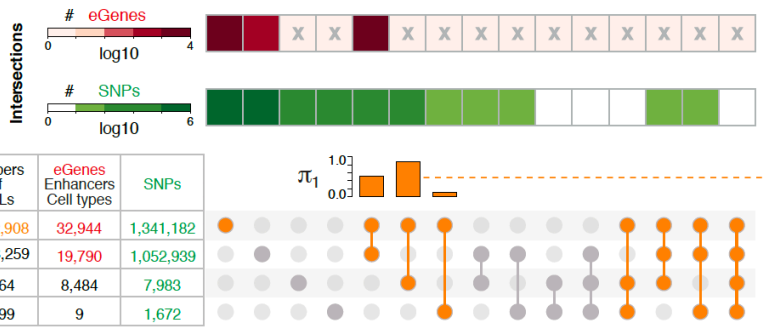
Brain eQTL sets larger than previous studies

- Brain eQTLs (FDR< 0.05)
  - 32944(75%) eGenes
  - 2,542,908 eQTLs
  - 1,341,182 unique cis-eSNPs
    (~238K independent SNPs after linkage-disequilibrium (LD) pruning)
- Large overlap with GTEx brain eQTLs



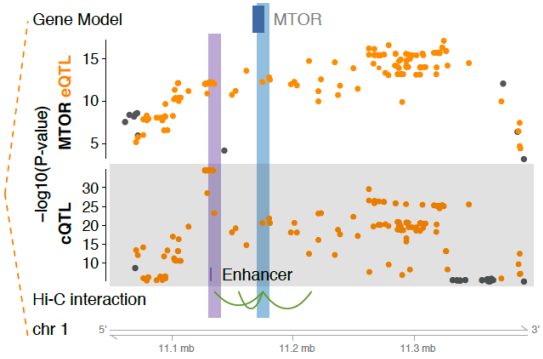
Many multi-QTLs (ie SNP acting as both cQTL & eQTL)



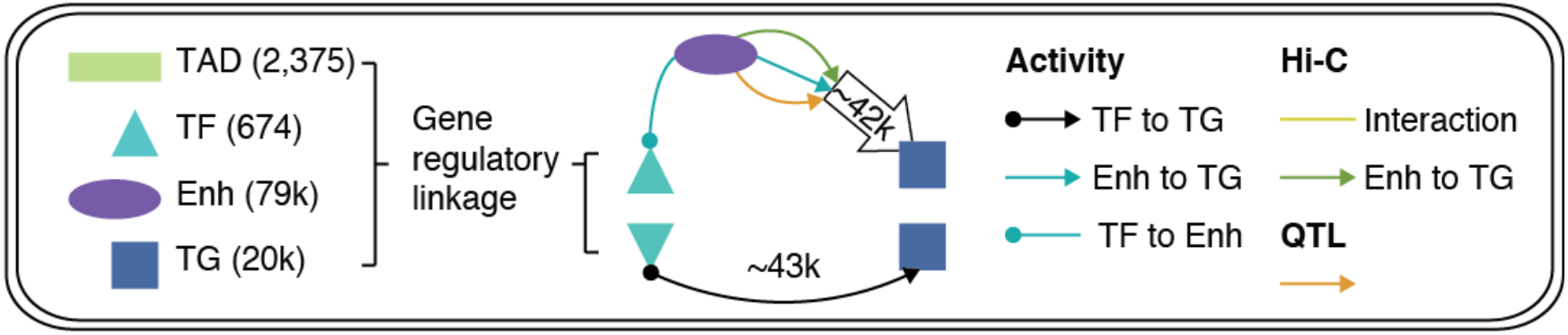


	Numbers of QTLs	eGenes Enhancers Cell types	SNPs
eQTL	2,542,908	32,944	1,341,182
isoQTL	2,628,259	19,790	1,052,939
cQTL*	8,464	8,484	7,983
fQTL	4,199	9	1,672

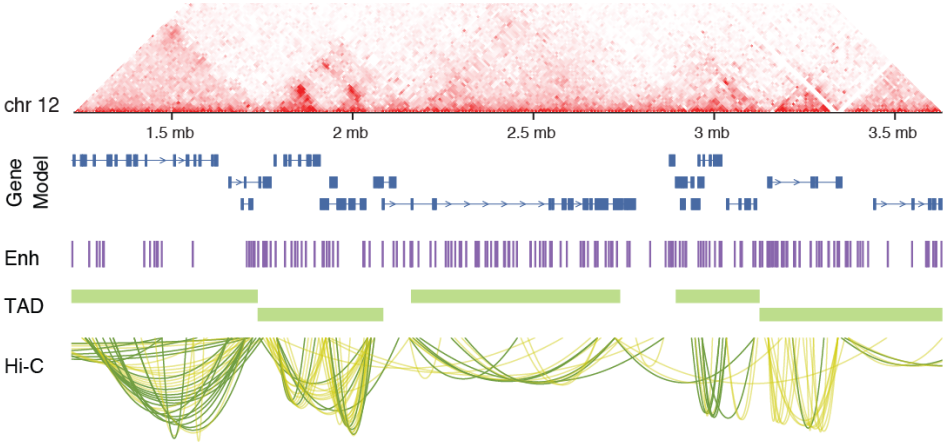
eQTLs and cQTLs significantly overlap

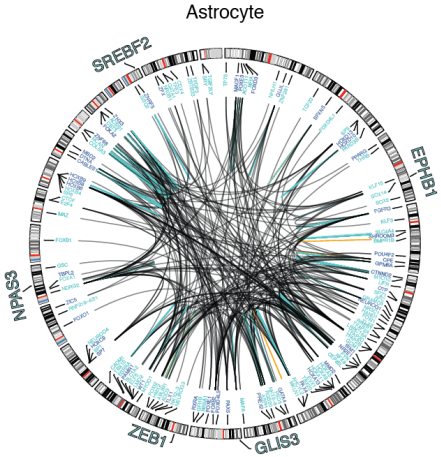


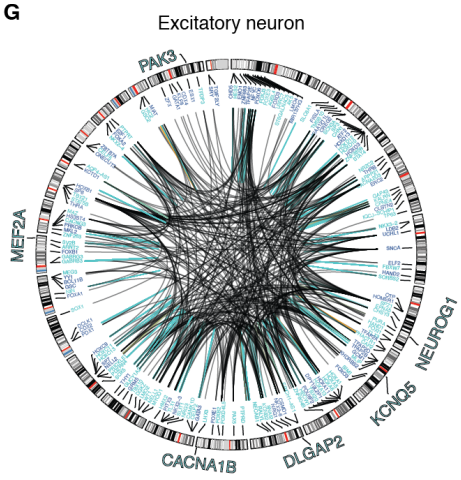
eQTLs for aging gene, mammalian target of rapamycin (mTOR) potentially mediated by cQTLs



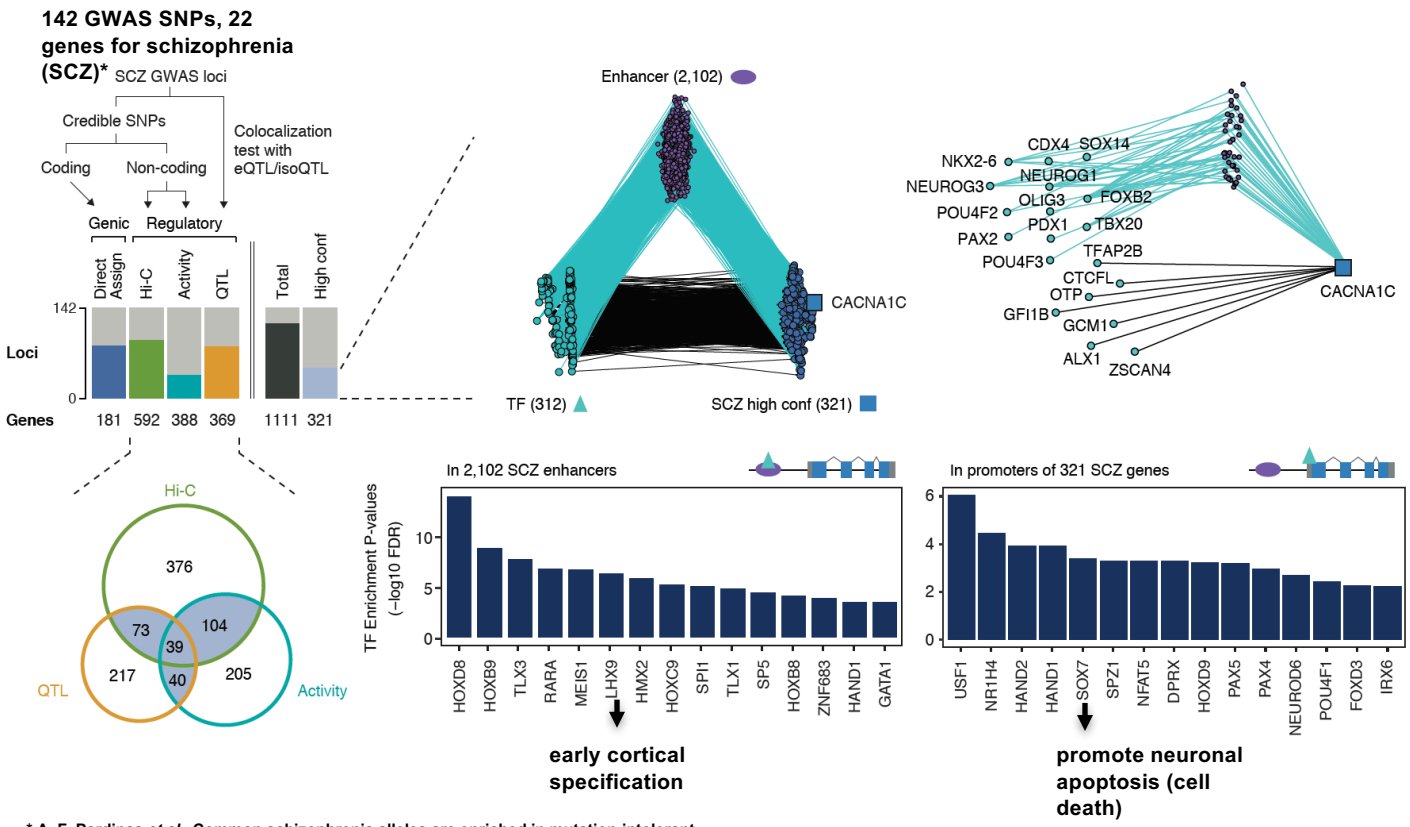
Gene regulatory network construction







Linking GWAS non-coding SNPs to new disease genes using gene regulatory network



* A. F. Pardinas *et al.*, Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* 50, 381-389 (2018)

Integrative modeling of brain phenotype data

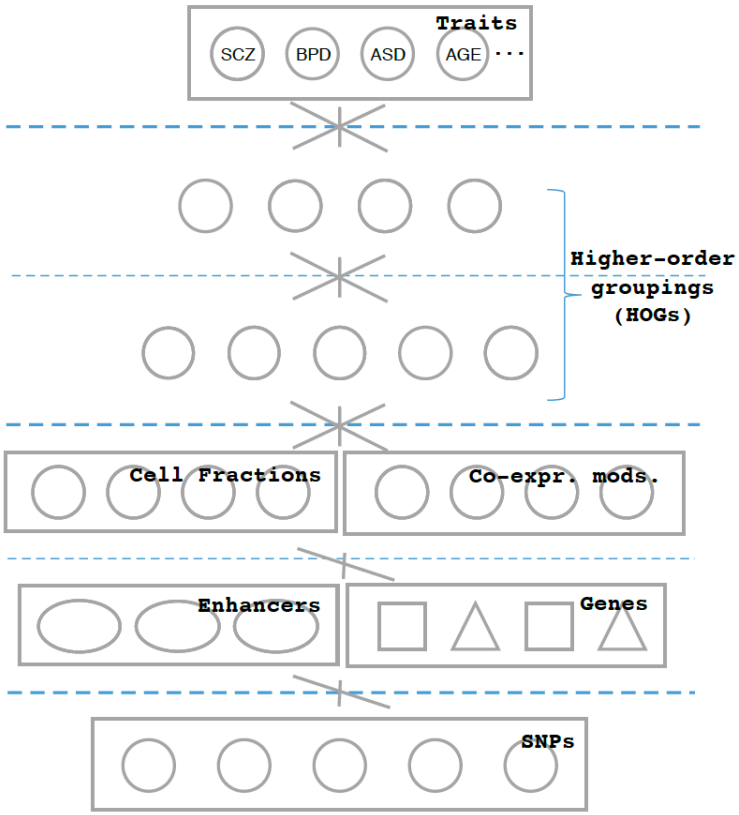
- We use the framework of Boltzmann machines to integrate phenotypes at multiple levels, while conditioning on genotype
- Evaluate joint Energy model of conditional distribution
- Inference and training
  - Prediction by minimizing free energy
  - 'Persistent' MCMC for training
- Boltzmann machine variables
  - **x**: visible units
  - **h**: hidden units
  - **z**: conditioning units
  - W: weights

$$p(\mathbf{x},\mathbf{h}|\mathbf{z}) = \exp\left(-E(\mathbf{x},\mathbf{h}|\mathbf{z})\right)/Z(\mathbf{z})$$

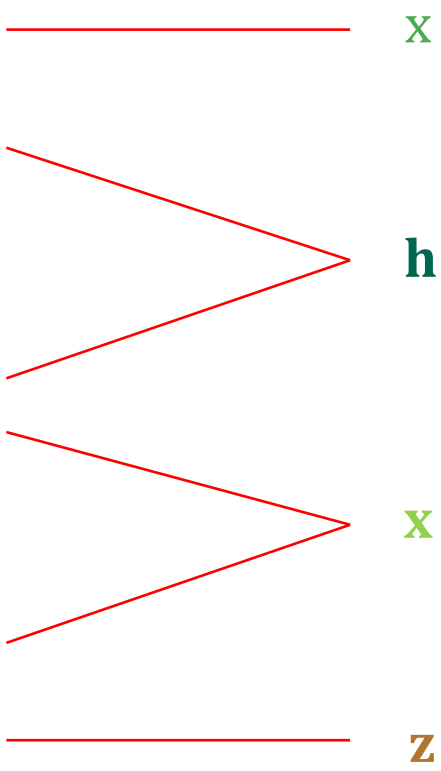
$$Z(\mathbf{z}) = \sum_{\mathbf{x},\mathbf{h}} \exp\left(-E(\mathbf{x},\mathbf{h}|\mathbf{z})\right)$$

$$E(\mathbf{x},\mathbf{h}|\mathbf{z}) = -\mathbf{z}^{\mathrm{T}}\mathbf{W_1}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W_2}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W_3}\mathbf{h} - \mathbf{h}^{\mathrm{T}}\mathbf{W_4}\mathbf{h} - \boldsymbol{Bias}$$

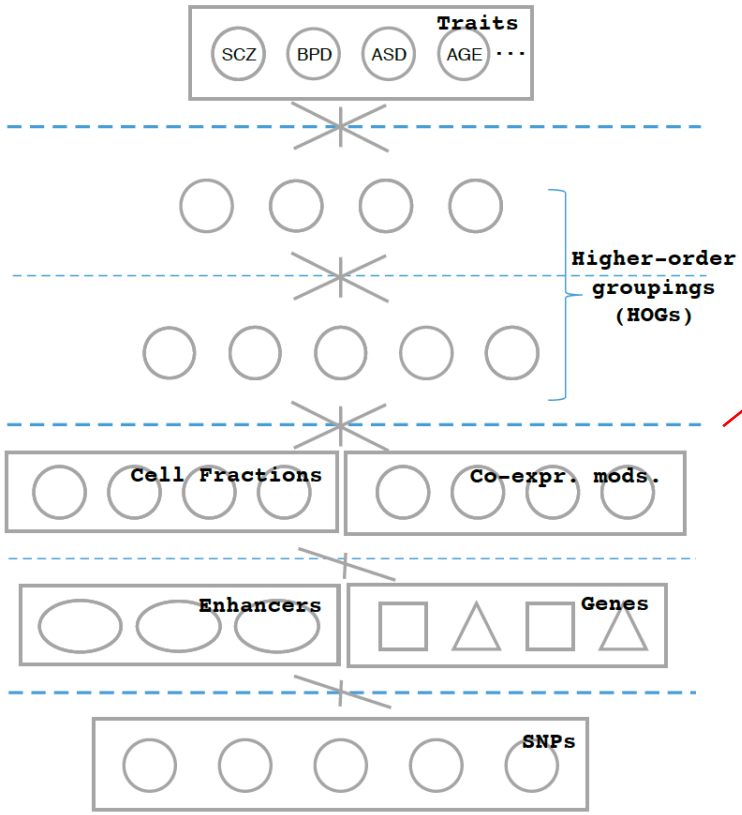
Deep Structured Phenotype Network (DSPN)



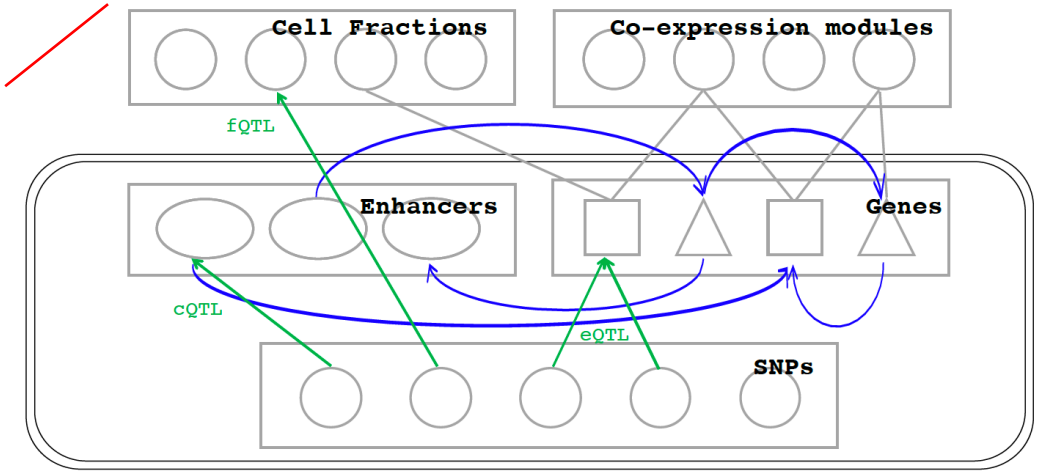
Boltzmann machine variables



Deep Structured Phenotype Network (DSPN)



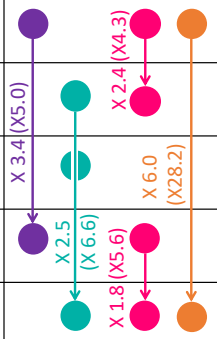
Gene network structure



DSPN improves brain disease prediction

Accuracy = chance to correctly predict disease/health

Method	SCZ	BPD	ASD	AVG (SCZ+BPD+ASD)
LR-gene	54.6% (0.5%)	56.7% (2.5%)	50.0% (0.0%)	53.8% (1.0%)
LR-trans	63.0% (4.8%)	63.3% (6.3%)	51.7% (1.8%)	59.3% (4.3%)
cRBM	70.0% (31.0%)	71.1% (22.6%)	56.7% (3.8%)	65.9% (19.1%)
DSPN-imput	59.0% (1.8%)	67.2% (10.7%)	62.5% (2.6%)	62.9% (5.0%)
DSPN-full	73.6% (32.8%)	76.7% (37.4%)	68.3% (14.4%)	72.9% (28.2%)

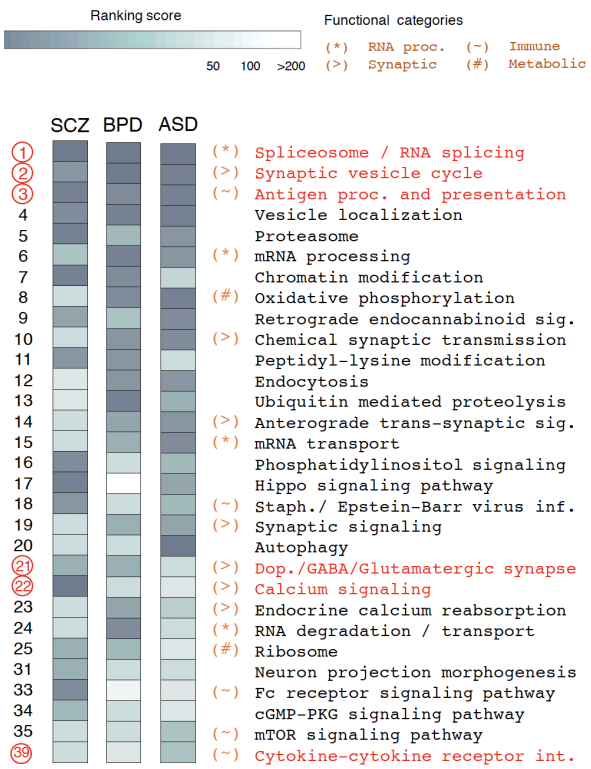


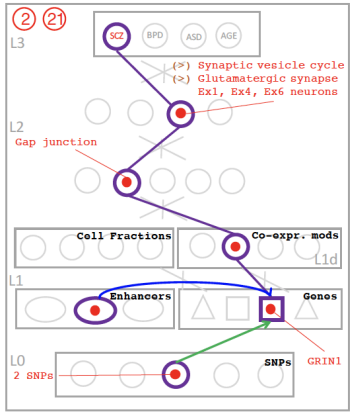
Model complexity	increasing	increasing	constant	increasing
Predictors	genotype	transcriptome	genotype->transcriptome	genotype->transcriptome

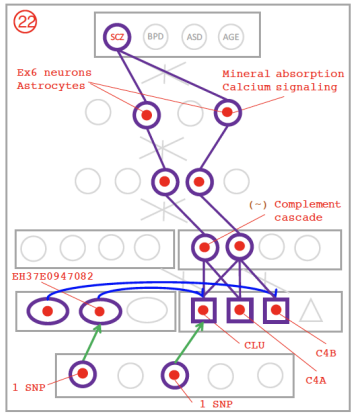
Unbracketed figures show test-set performance accuracy, with chance at 50%; bracketed figures show variance explained on liability scale

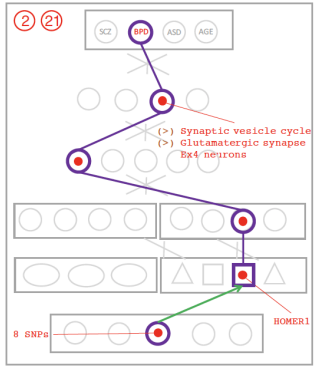
*** Brainstorm consortium (~1.2 million individuals, *Science*, 2018) used linear predictive model to find that common SNPs explain 25.6%, 20.5%, and 19% of the genetic variance for SCZ, BPD and ASD**

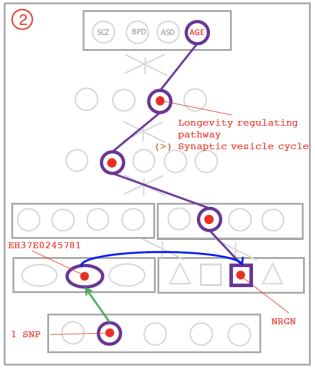
DSPN discovers molecular pathways from genotype to phenotype











Transcriptome Mining: Population-scale genomic analysis to better understand mental disease & the subtle privacy risks of this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to mental disease
  - Consortium intro & construction of an adult brain resource w/ 1866 individuals
  - Explanation of across-population variation via changing proportions of cell types (using single-cell deconvolution)
  - Generation of a large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C & activity relationships. Using this to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype & transcriptome
- *[Exhaust]* **Genomic Privacy** & RNA-seq
  - Introduction to 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 particularly tricky privacy issues
  - Using file formats to remove obvious variants
  - eQTLs: Quantifying & removing further variant info from expression levels w/ ICI & predictability. Instantiating a practical linking attack w/ 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
- *[Exhaust]* **Publication Patterns** from data producing consortia
  - Co-authorship network stats relate to publication rollouts & show gradual adoption by community
  - Key role of brokers in data dissemination

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)

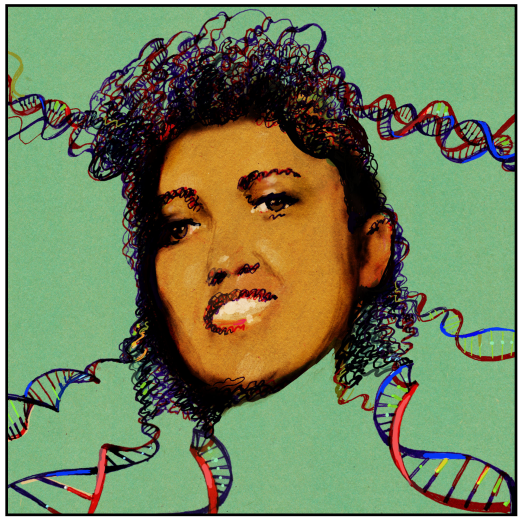
Genomics has similar "Big Data" Dilemma in the Rest of Society

- Sharing & "peer-production" is central to success of many new ventures, with the same risks as in genomics
  - **EG web search**: Large-scale mining essential
- We confront privacy risks every day we access the internet

[Seringhaus & Gerstein ('09), *Hart. Courant* (Jun 5); Greenbaum & Gerstein ('11), *NY Times* (6 Oct)]

Tricky Privacy Considerations in Personal Genomics

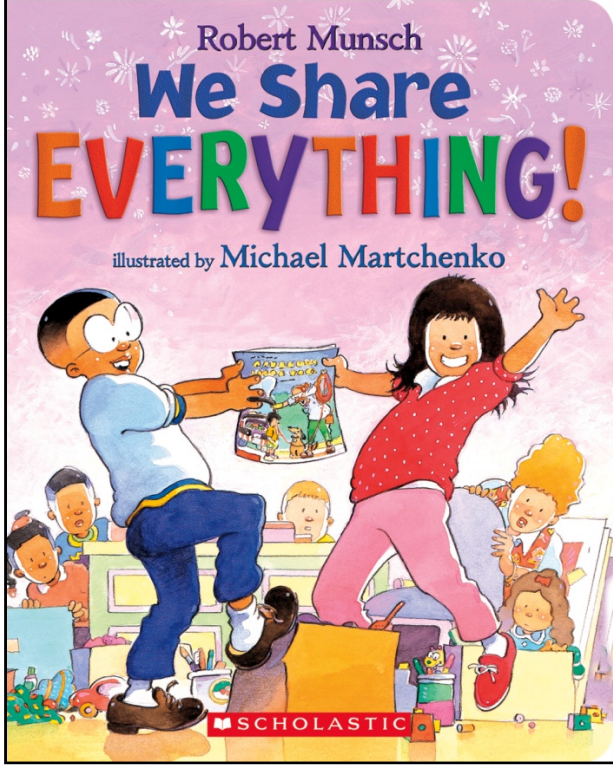
- **Genetic Exceptionalism :** The Genome is very fundamental data, potentially very revealing about one's identity & characteristics
- **Personal Genomic info. essentially meaningless currently but will it be in 20 yrs? 50 yrs?**
  - Genomic sequence very revealing about one's children. Is true consent possible?
  - Once put on the web it can't be taken back
- **Culture Clash:** Genomics historically has been a proponent of "open data" but not clear personal genomics fits this.
  - Clinical Medline has a very different culture.
- **Ethically challenged** history of genetics
  - Ownership of the data & what consent means (Hela)
    - Could your genetic data give rise to a product line?



[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; Photo from NY Times]

The Other Side of the Coin: Why we should share

- Sharing helps **speed research**
  - Large-scale mining of this information is important for medical research
  - Privacy is cumbersome, particularly for big data
- Sharing is important for **reproducible research**
- Sharing is useful for **education**
  - More fun to study a known person's genome
    - Eg Zimmer's Game of Genomes in STAT



[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)]

[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
  - What is acceptable risk?
    Can we quantify leakage?
    - Ex: photos of eye color
  - Cost Benefit Analysis

Current Social & Technical Solutions

- **Closed Data** Approach
  - Consents
  - "Protected" distribution via dbGAP
  - Local computes on secure computer
- Issues with Closed Data
  - Non-uniformity of consents & paperwork
    - Different international norms, leading to confusion
  - Encryption & computer security creates burdensome requirements on data sharing & large scale analysis
  - Many schemes get "hacked"

- **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]

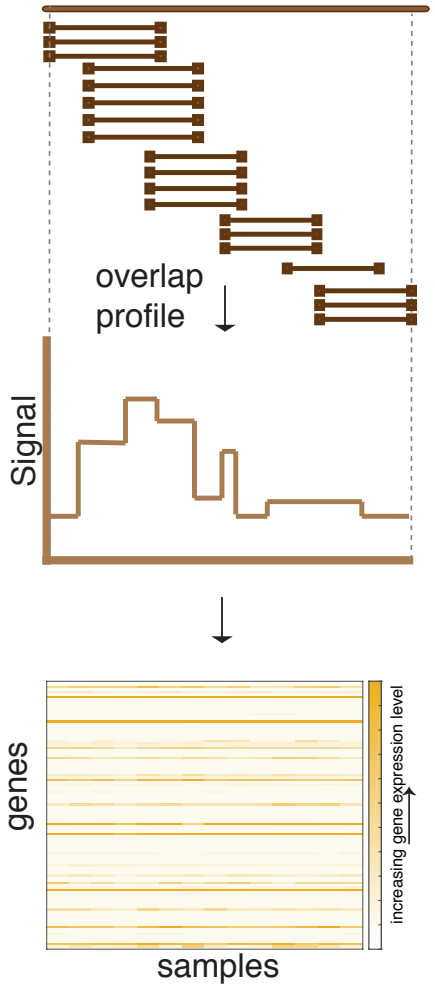
Strawman Hybrid Social & Tech Proposed Solution?

- 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”

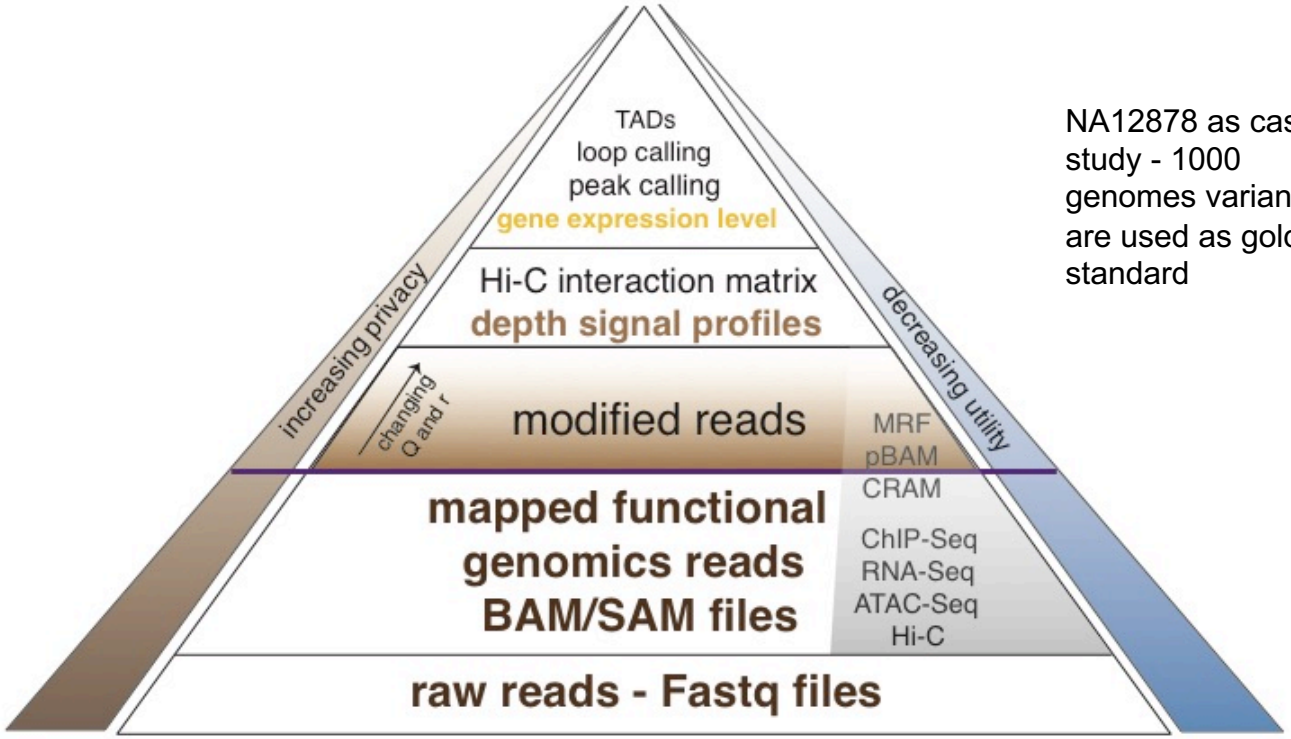
- **Quantifying Leakage & allowing a small amounts of it**
- **Careful separation & coupling of private & public data**
  - **Lightweight, freely accessible secondary datasets coupled to underlying variants**
  - Selection of stub & "test pilot" datasets for benchmarking
  - Develop programs on public 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.



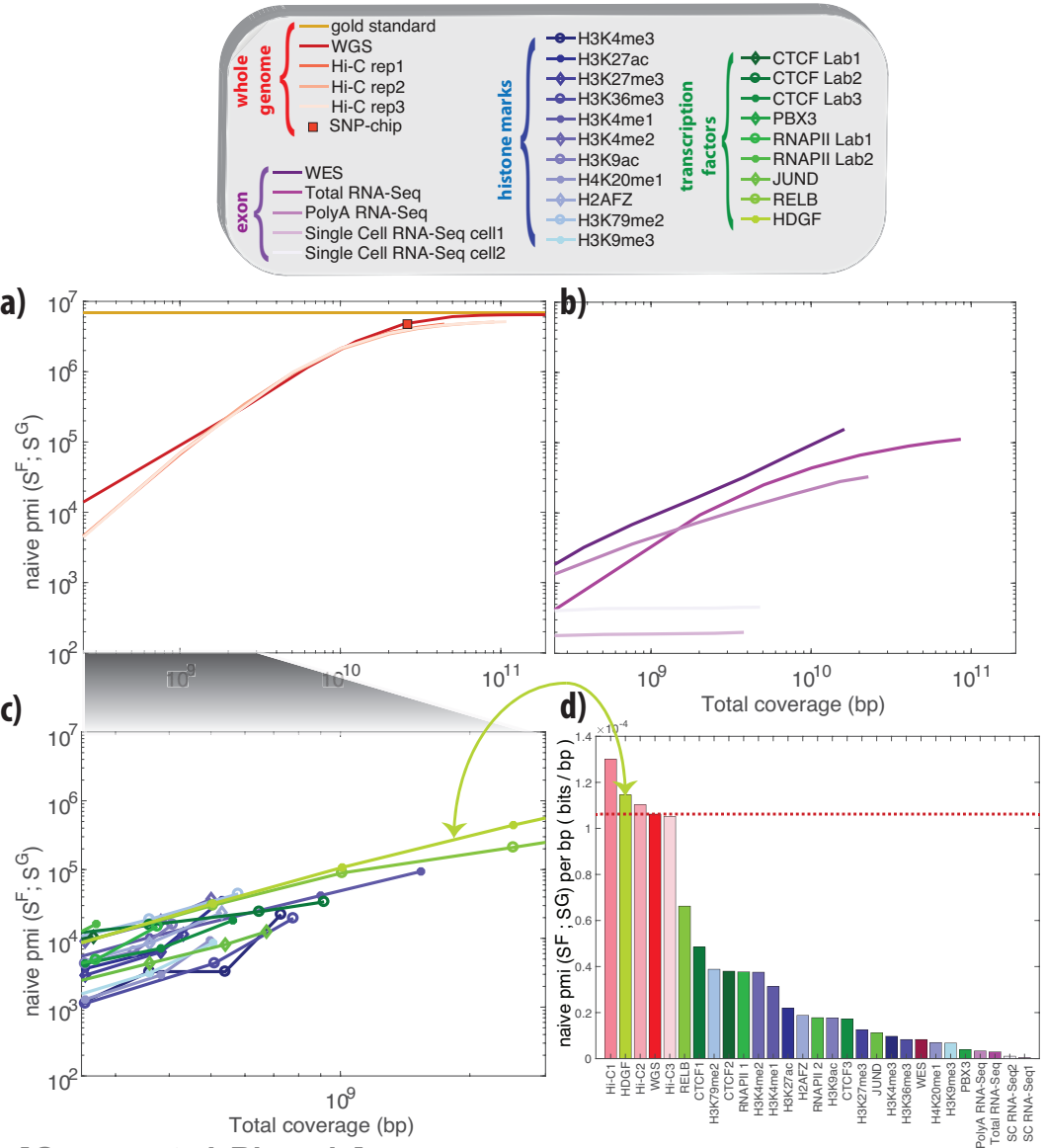
[*Gursoy et al*, Bioarvix]



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

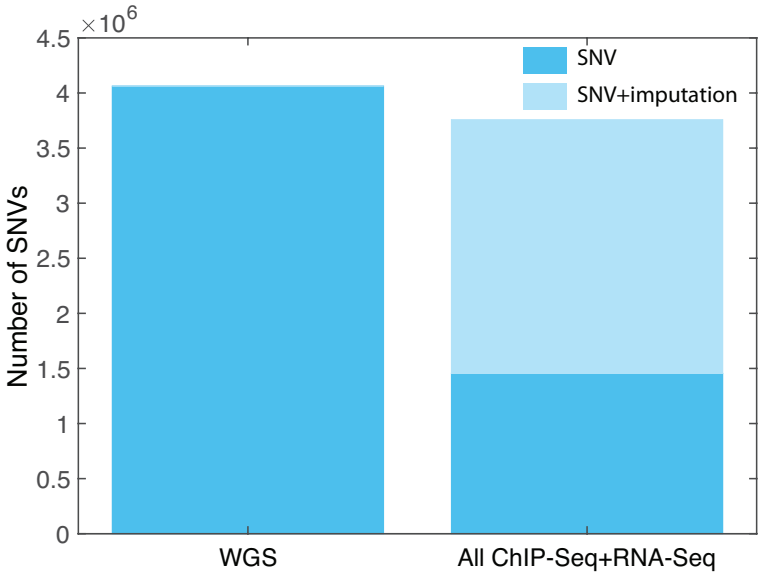
Leakage Source	Leaking Variants	# of potential variants	Average leakage per variant (bits)	Maximum leakage per variant (bits)	# of accessible variants	Total leakage (bits)
Raw reads	Exonic variants	2,682,417	0.10 ± 0.28	9.88 ± 2.12	246,893	24,689
Modified reads Q = {indels}	Exonic SNVs	2,607,969	0.09 ± 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?**



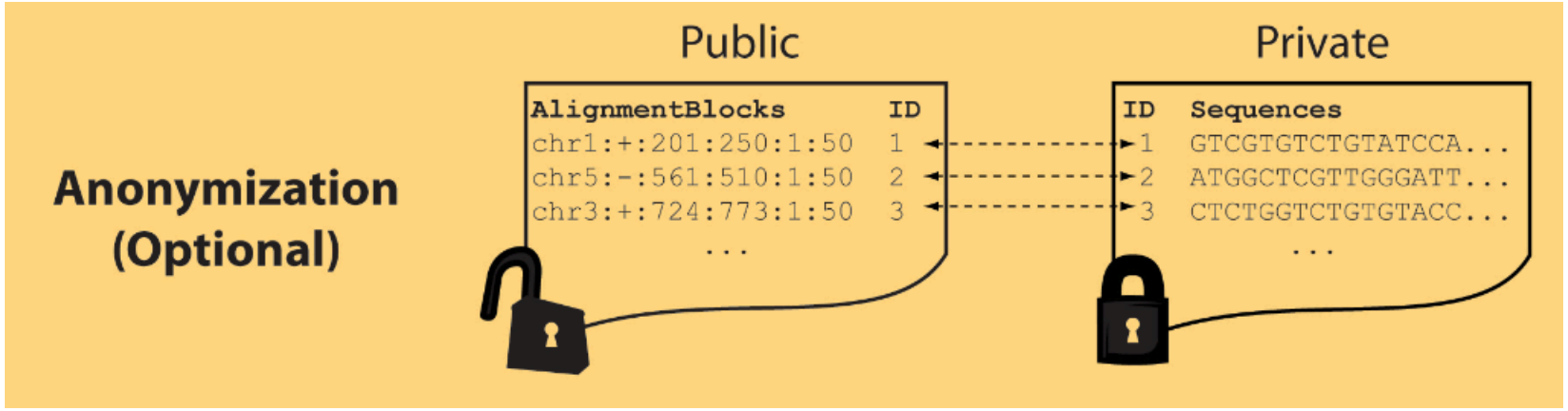
[*Gursoy et al*, Bioarvix]

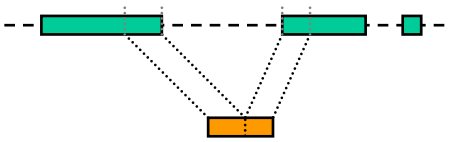
- **It might seem like we don't infer much information from single ChIP-Seq and RNA-Seq experiments compared to WGS**
  - **However putting 10 different ChIP-Seq experiments and RNA-Seq together with imputation provides a great deal of information about the individual**



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
- Distinction between formats to compute on and those to archive with – become sharper with big data



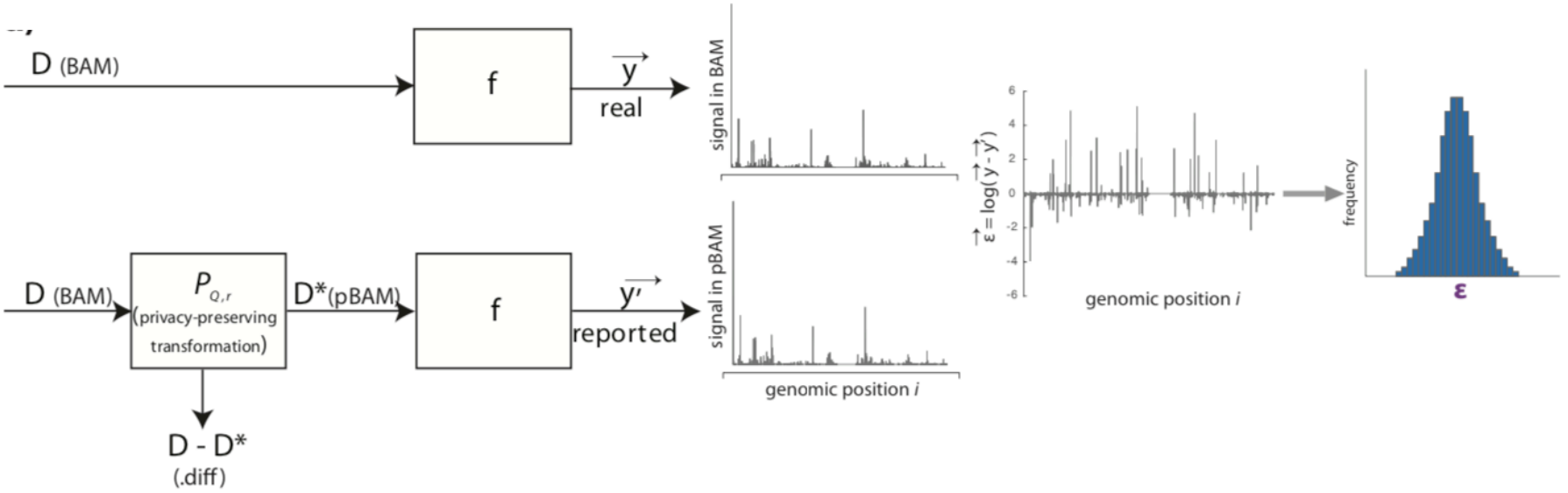


Mapping coordinates without variants (MRF)

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

[*Bioinformatics* 27: 281]

Privacy-aware Binary Alignment Mapping (pBAM)

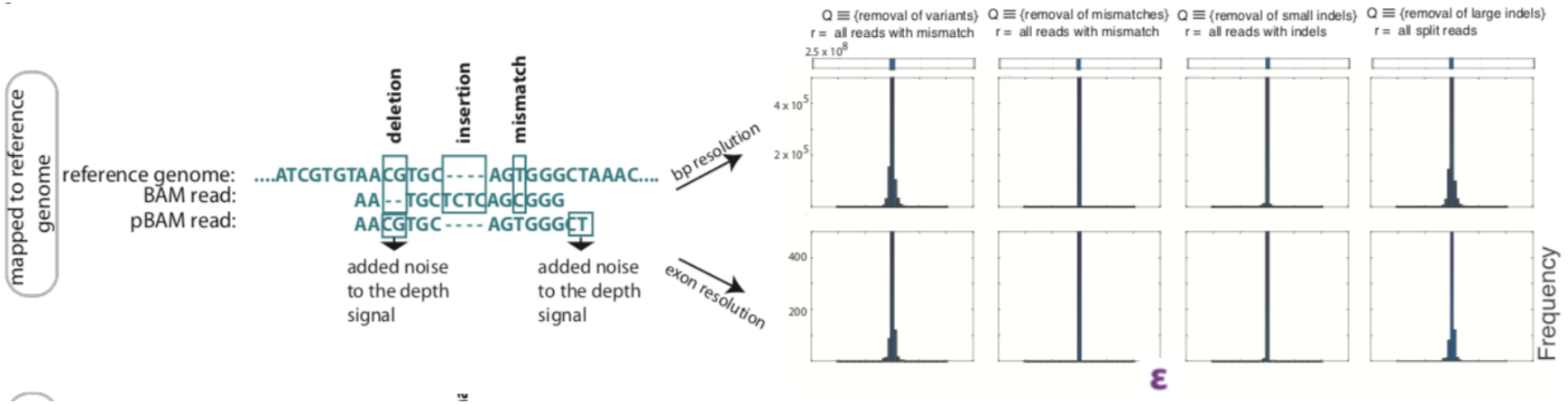


- A manipulation on Binary Alignment Files (BAM)
  - Suppression: replace sequence and quality string with (*)
  - Generalization: convert cigar, alignment score and MD tag into perfectly matched strings

- Works with majority of functions of SAMtools.

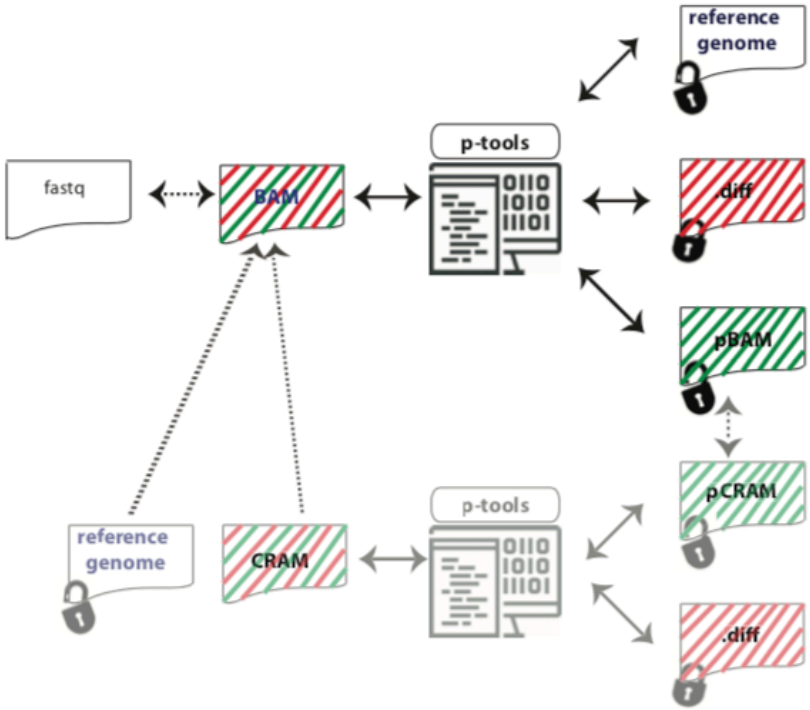
[*Gursoy et al,* Bioarvix]

pBAMs are high in utility and can be converted BAM



- Works well with many functional genomics pipelines, including STAR signal tracks, RSEM gene expression and quantification and MACS2 for ChIP-Seq peak calling.

- The original BAM does not need be stored. Rather, a smaller file called .diff can be safely stored for sensitive information in the BAM file.



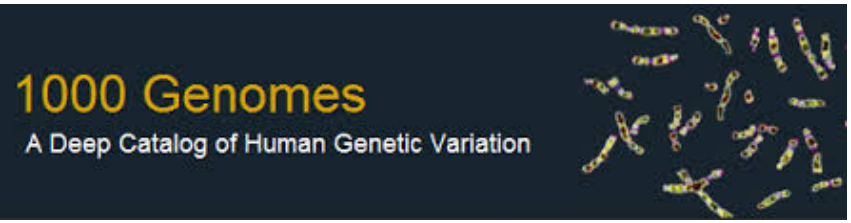
[*Gursoy et al*, Bioarvix]

Transcriptome Mining: Population-scale genomic analysis to better understand mental disease & the subtle privacy risks of this activity

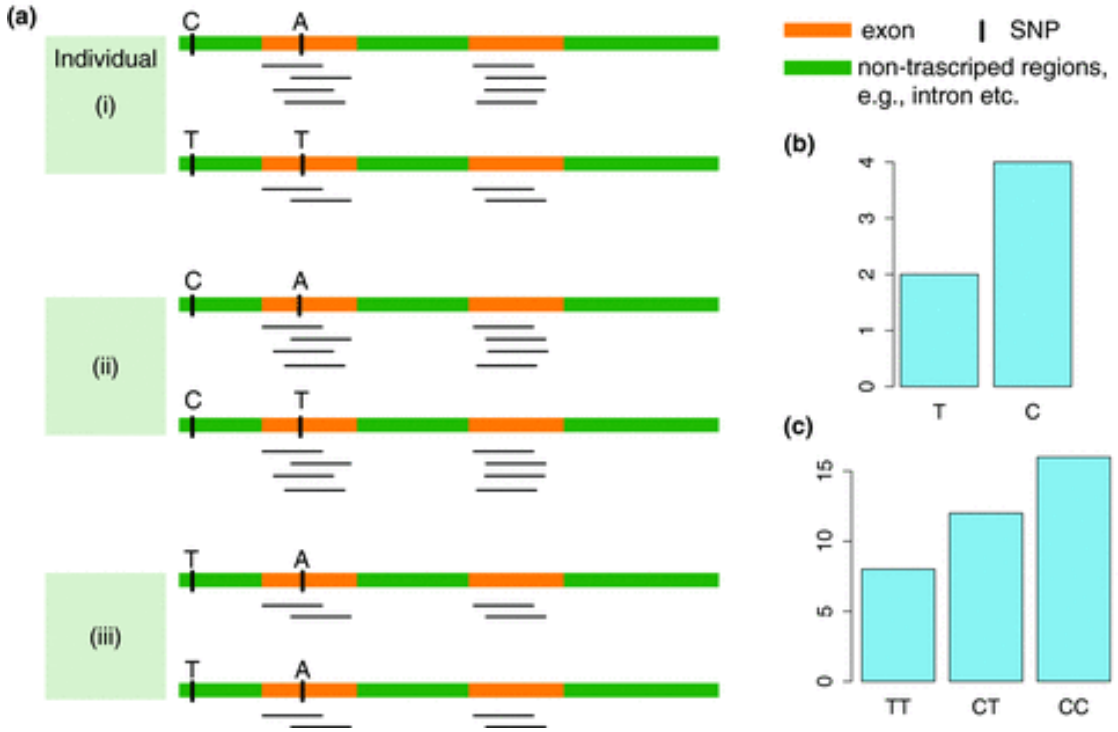
- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to mental disease
  - Consortium intro & construction of an adult brain resource w/ 1866 individuals
  - Explanation of across-population variation via changing proportions of cell types (using single-cell deconvolution)
  - Generation of a large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C & activity relationships. Using this to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype & transcriptome
- *[Exhaust]* **Genomic Privacy** & RNA-seq
  - Introduction to 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 particularly tricky privacy issues
  - Using file formats to remove obvious variants
  - eQTLs: Quantifying & removing further variant info from expression levels w/ ICI & predictability. Instantiating a practical linking attack w/ 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
- *[Exhaust]* **Publication Patterns** from data producing consortia
  - Co-authorship network stats relate to publication rollouts & show gradual adoption by community
  - Key role of brokers in data dissemination

Representative Functional Genomics, Genotype, eQTL Datasets

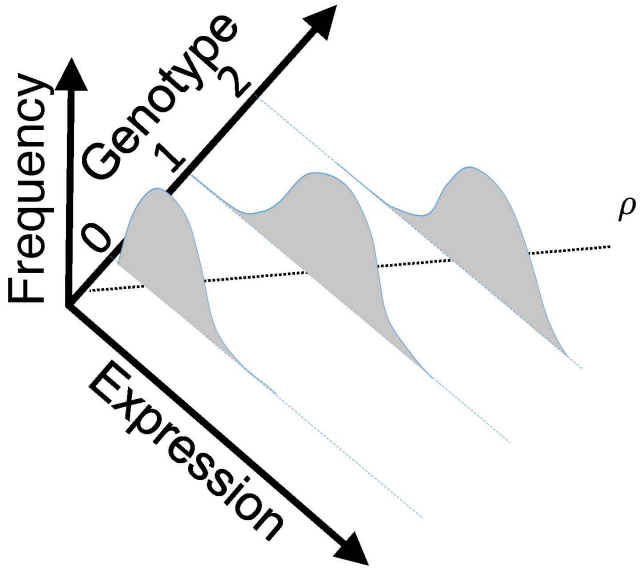
- Genotypes are available from the 1000 Genomes Project
- mRNA sequencing for 462 individuals from gEUVADIS and ENCODE
  - Publicly available quantification for protein coding genes
- Functional genomics data (ChIP-Seq, RNA-Seq, Hi-C) available from ENCODE
- Approximately 3,000 cis-eQTL (FDR<0.05)



eQTL Mapping Using RNA-Seq Data



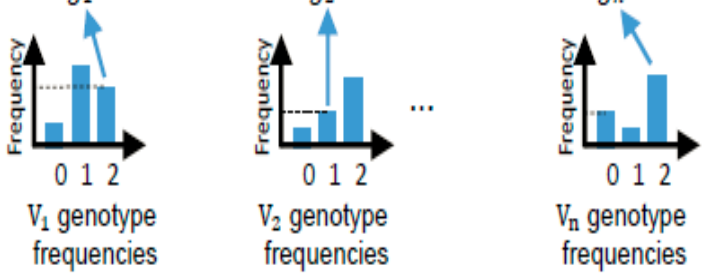
[*Biometrics 68(1) 1–11*]



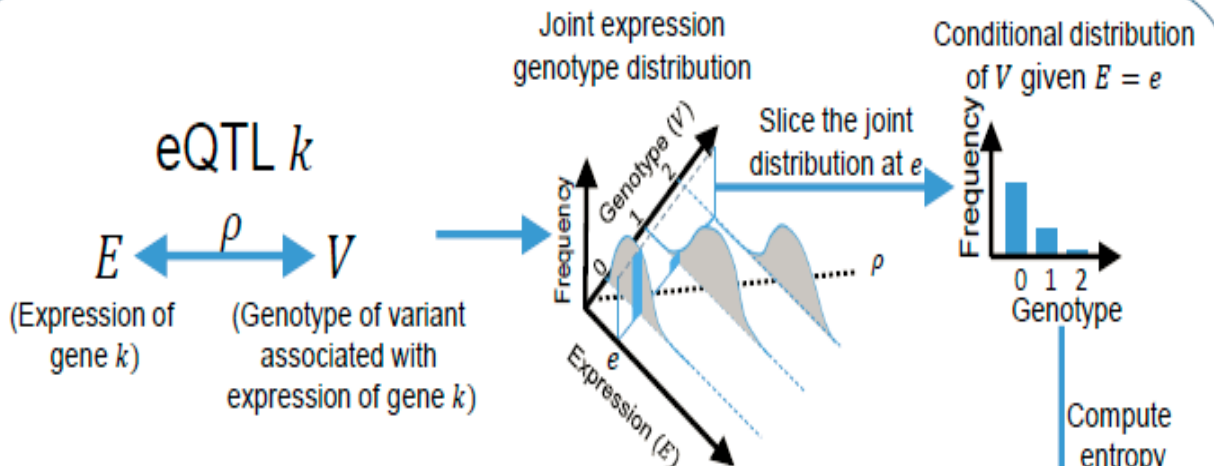
- eQTLs are genomic loci that contribute to variation in mRNA expression levels
- eQTLs provide insights on transcription regulation, and the molecular basis of phenotypic outcomes
- eQTL mapping can be done with RNA-Seq data

Information Content and Predictability

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



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

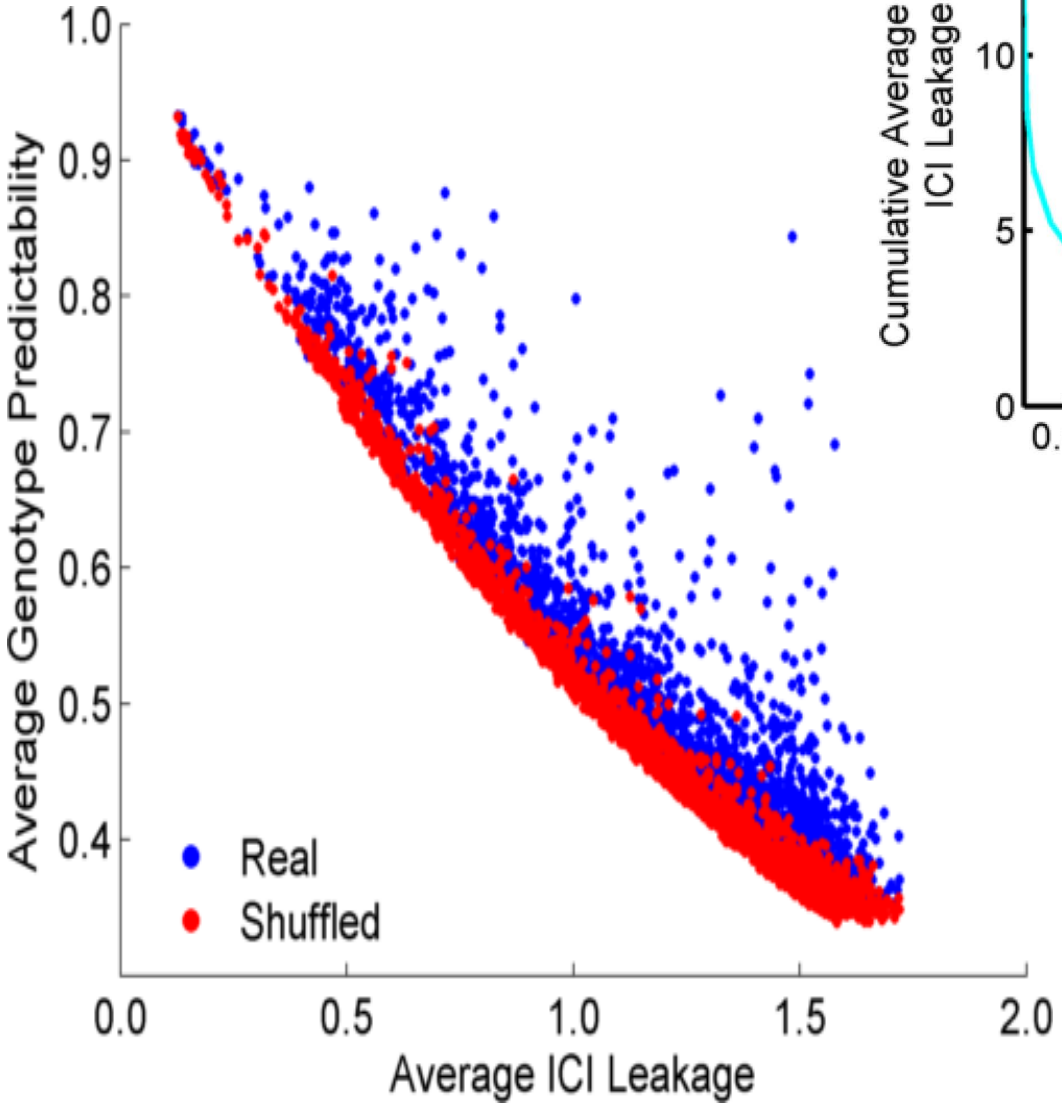


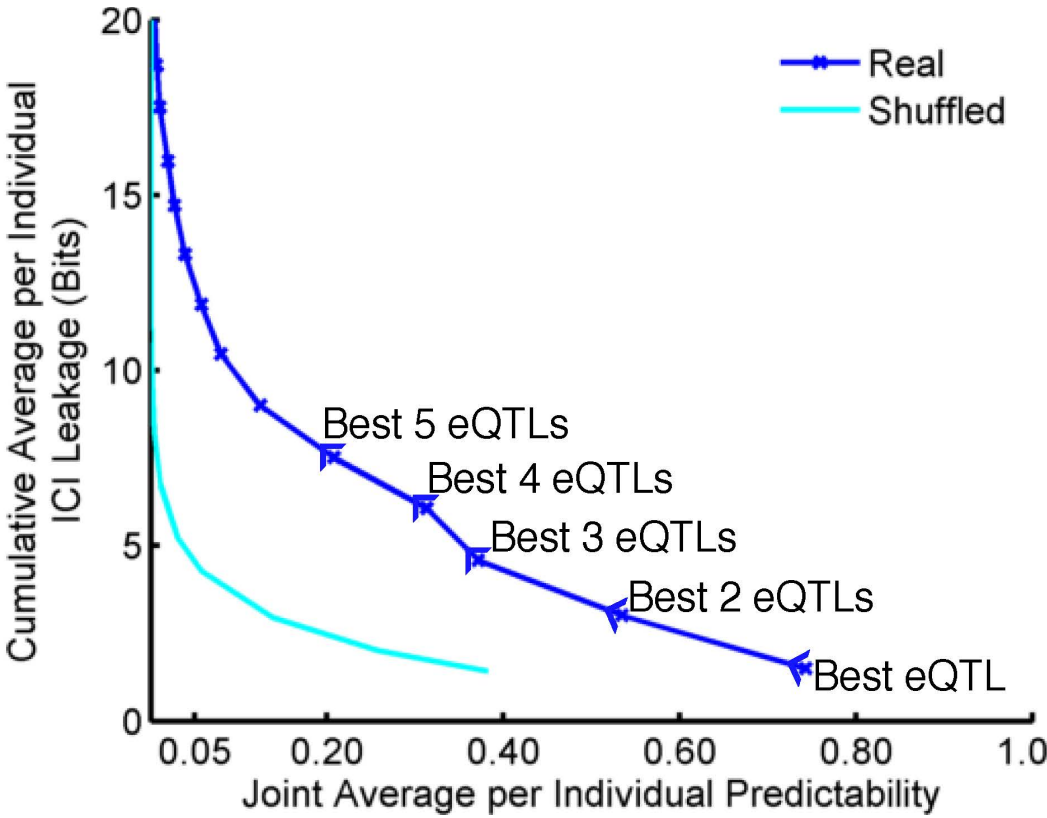
$$\pi(V \mid E = e) = e^{-H(V \mid E = e)}$$

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

[Harmanci et al. Nat. Meth. 2016]

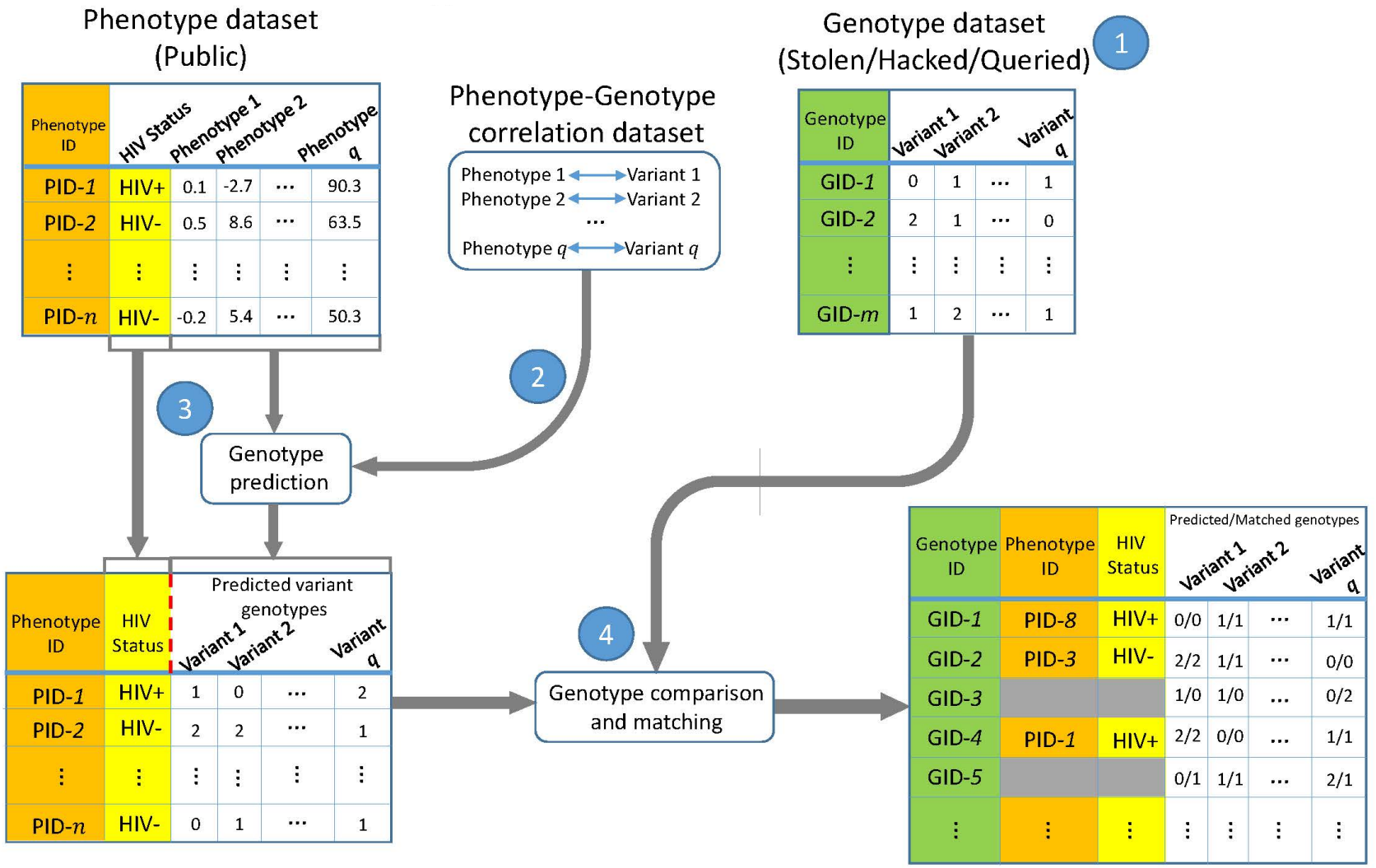
[Harmanciet al. Nat. Meth. (‘16]





ICI Leakage versus Genotype Predictability

Linking Attack Scenario



[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

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

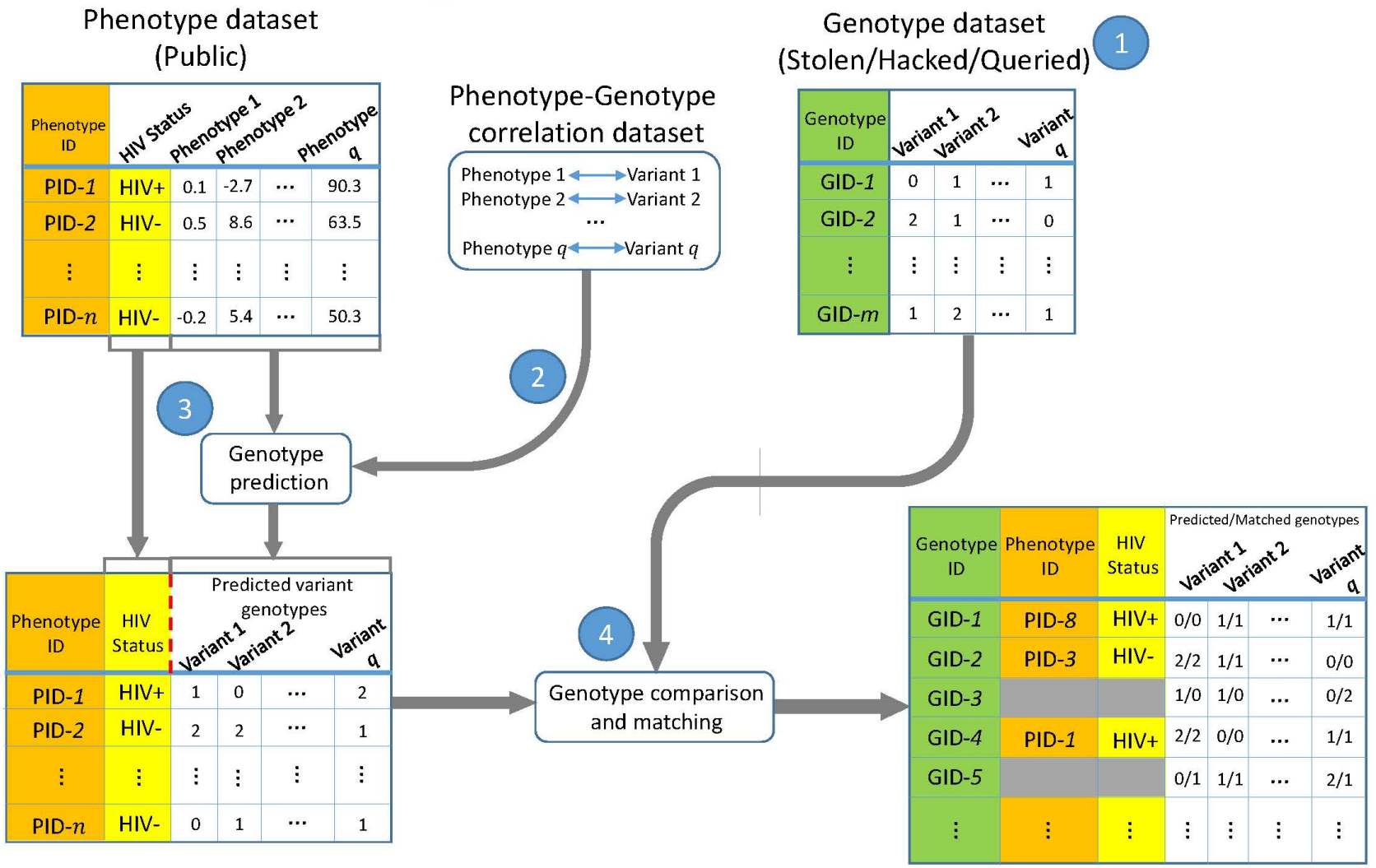
Names available for many users!

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	-
…	…	…	…
…	…	…	…
…	…	…	…

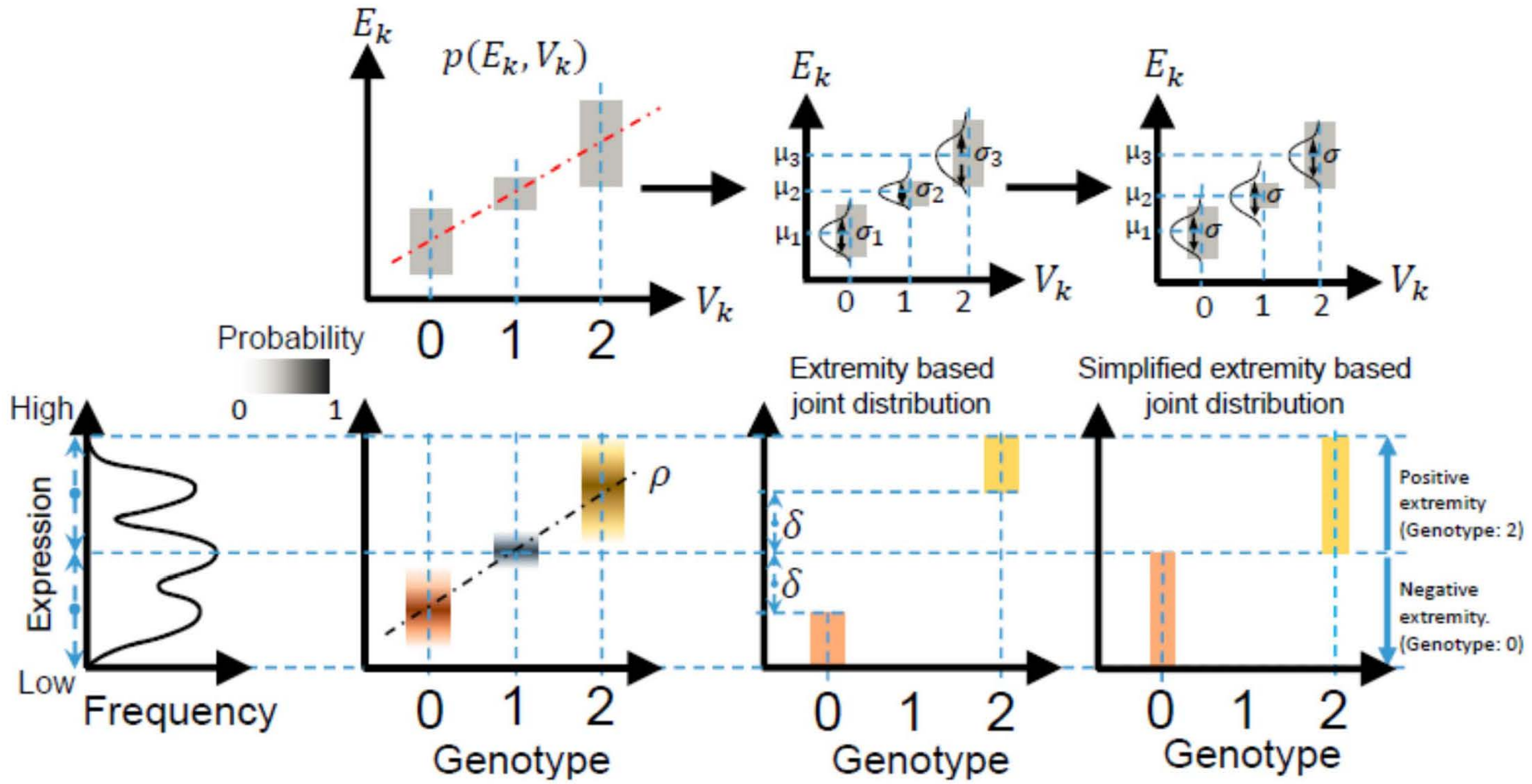
- **Many users are shared**
- **The grades of same users are correlated**
- **A user grades one movie around the same date in two databases**

Linking Attack Scenario



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

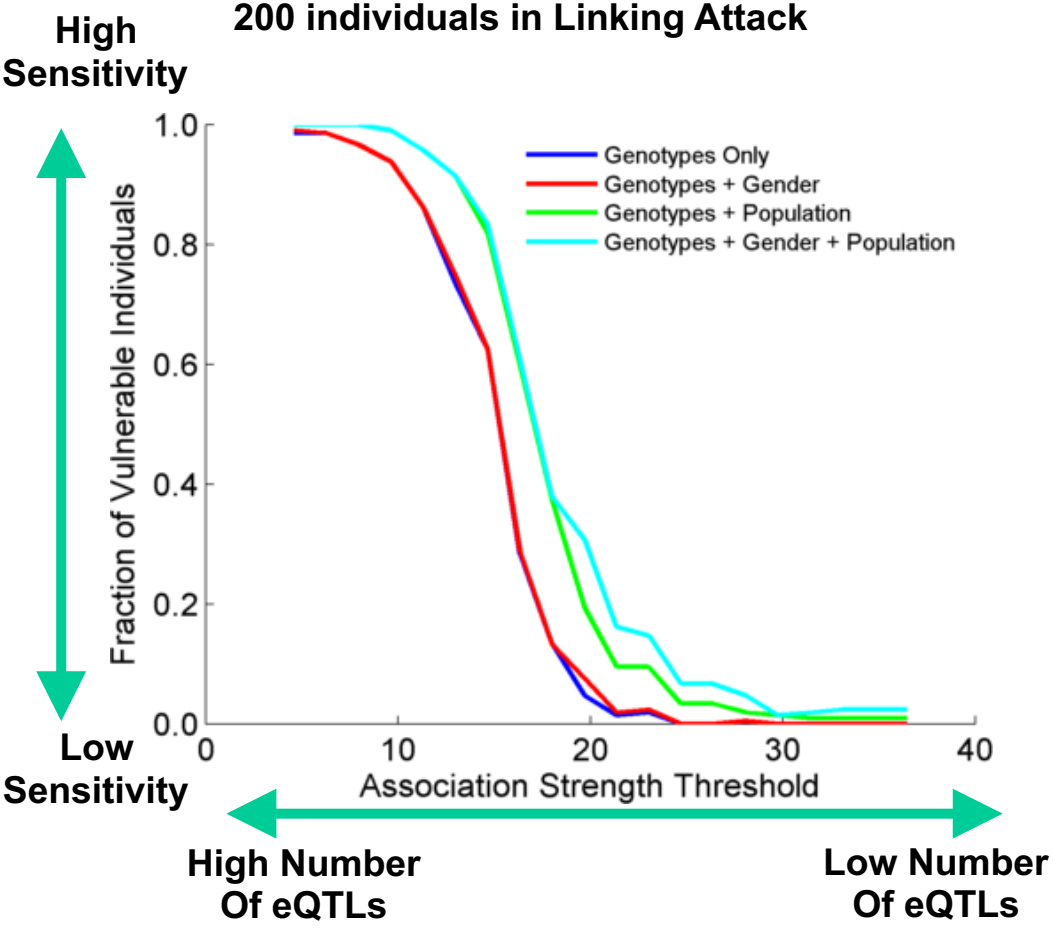
Levels of Expression-Genotype Model Simplifications for Genotype Prediction



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

Success in Linking Attack with Extremity based Genotype Prediction

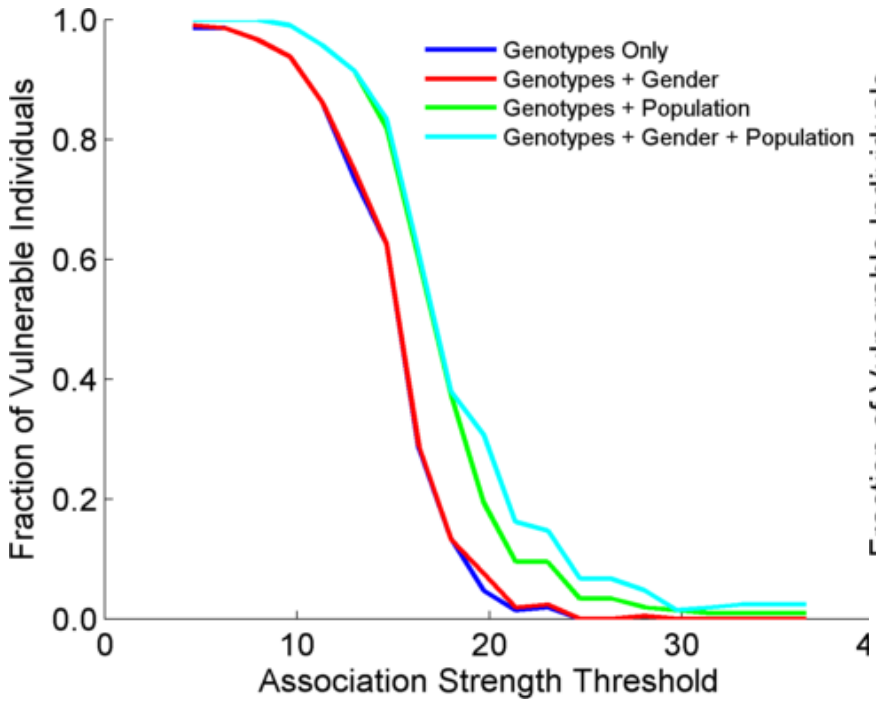
200 individuals eQTL Discovery
200 individuals in Linking Attack



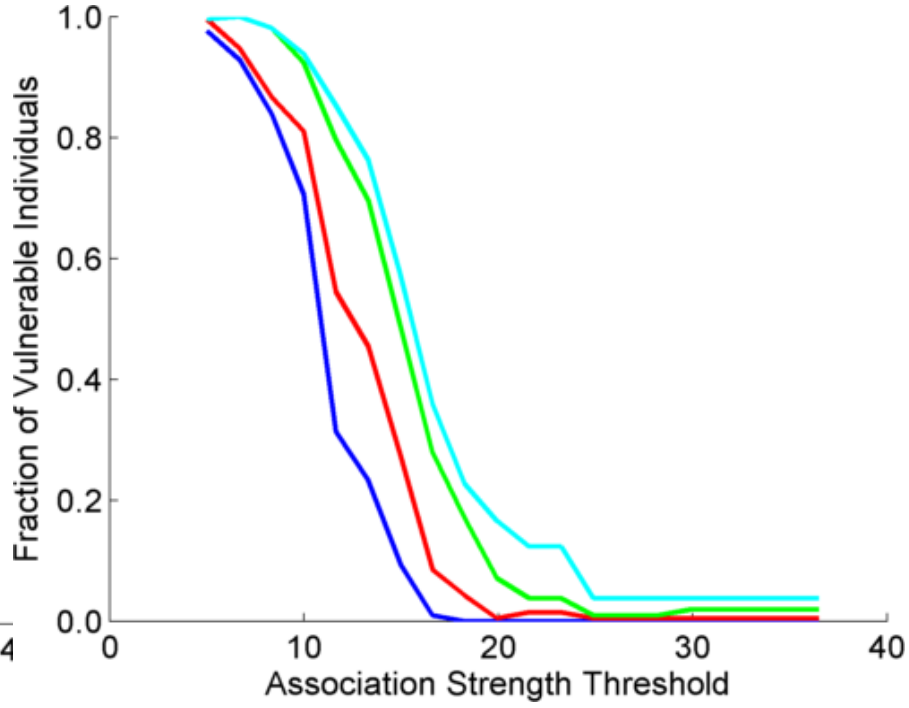
[Harmanci et al. *Nat. Meth.* (16)]

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

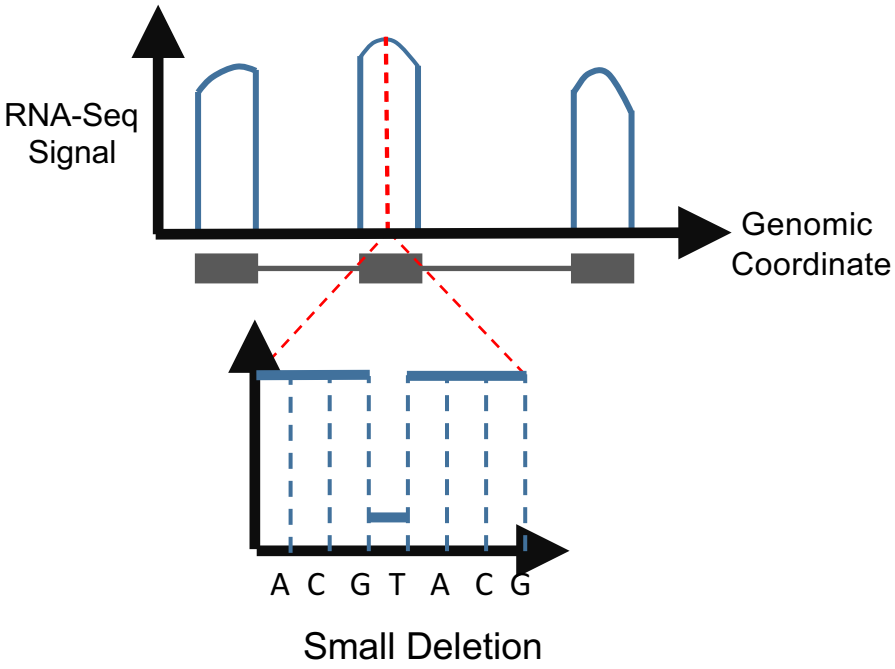


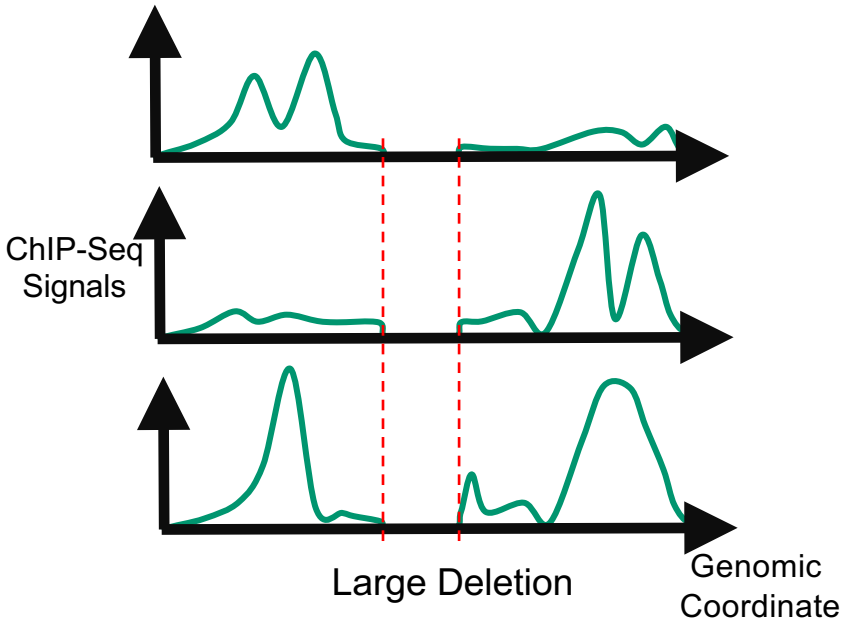
[Harmanci et al. *Nat. Meth.* (16)]

Transcriptome Mining: Population-scale genomic analysis to better understand mental disease & the subtle privacy risks of this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to mental disease
  - Consortium intro & construction of an adult brain resource w/ 1866 individuals
  - Explanation of across-population variation via changing proportions of cell types (using single-cell deconvolution)
  - Generation of a large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C & activity relationships. Using this to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype & transcriptome
- *[Exhaust]* **Genomic Privacy** & RNA-seq
  - Introduction to 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 particularly tricky privacy issues
  - Using file formats to remove obvious variants
  - eQTLs: Quantifying & removing further variant info from expression levels w/ ICI & predictability. Instantiating a practical linking attack w/ 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
- *[Exhaust]* **Publication Patterns** from data producing consortia
  - Co-authorship network stats relate to publication rollouts & show gradual adoption by community
  - Key role of brokers in data dissemination

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

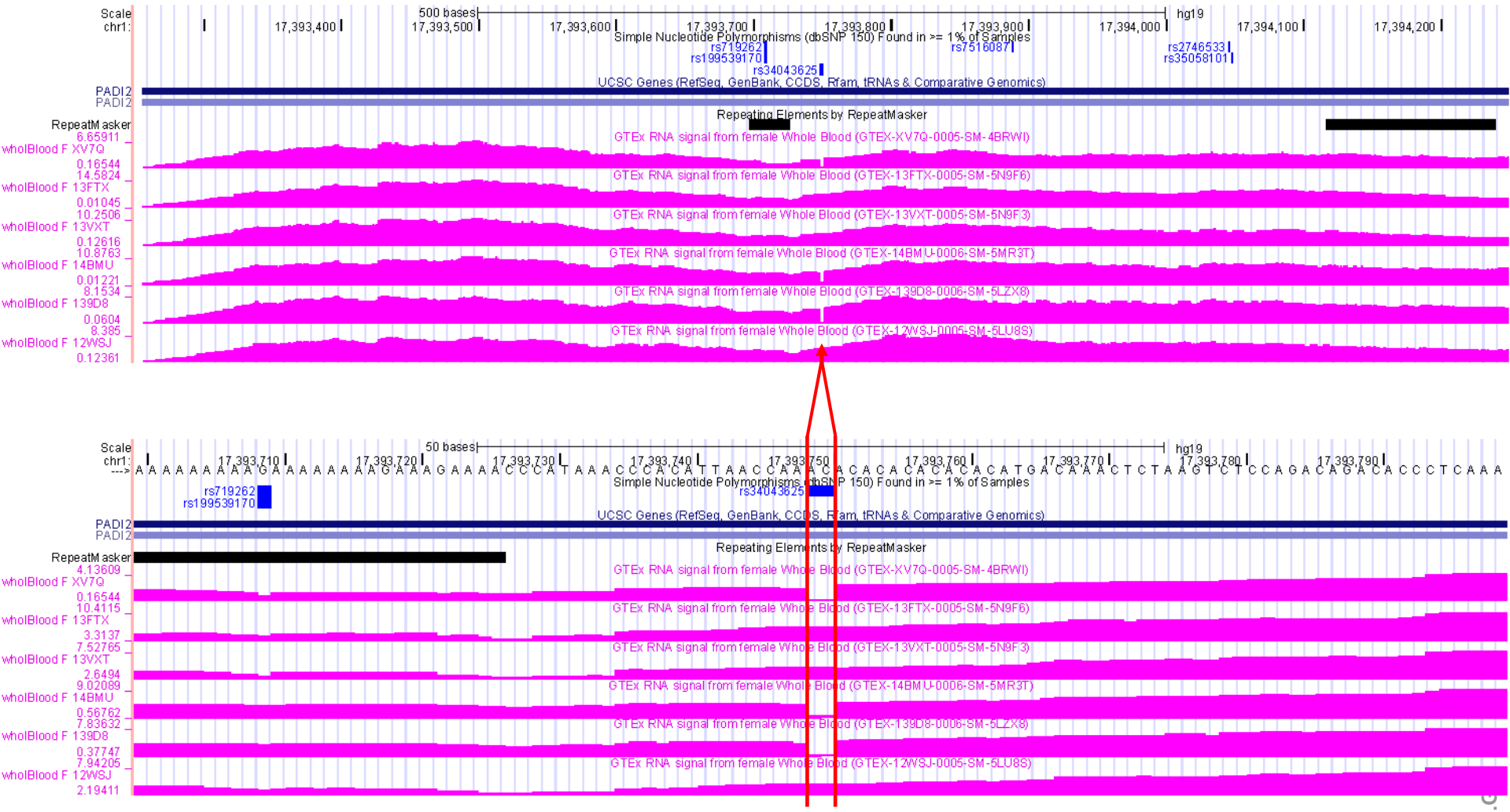




RNA-seq also shows large deletions

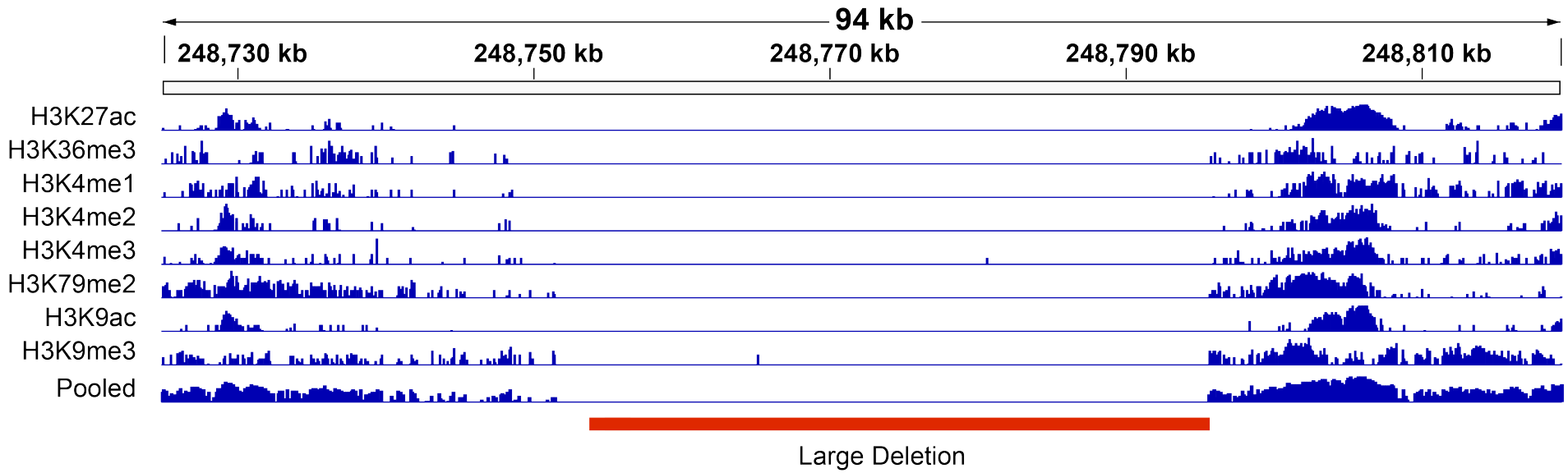
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Example of Small Deletion Evident in Signal Profile



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

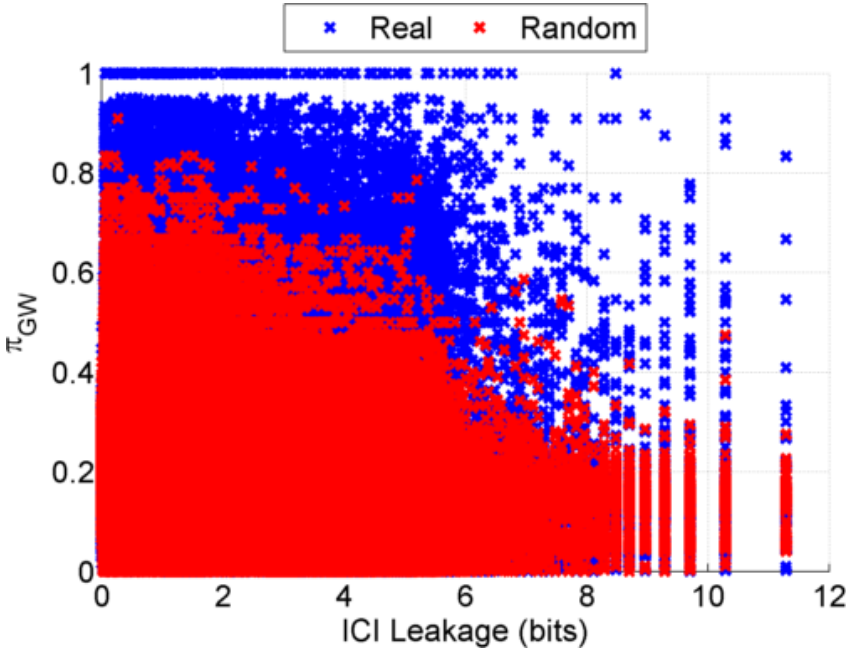
Example of Large Deletion Evident in Signal Profile



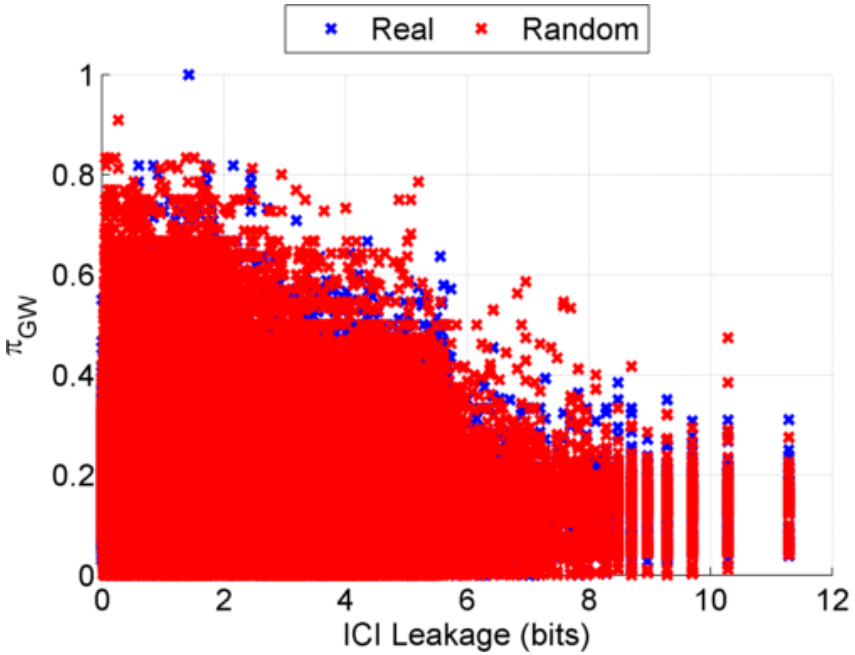
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Information Leakage from SV Deletions

a) Before Anonymization



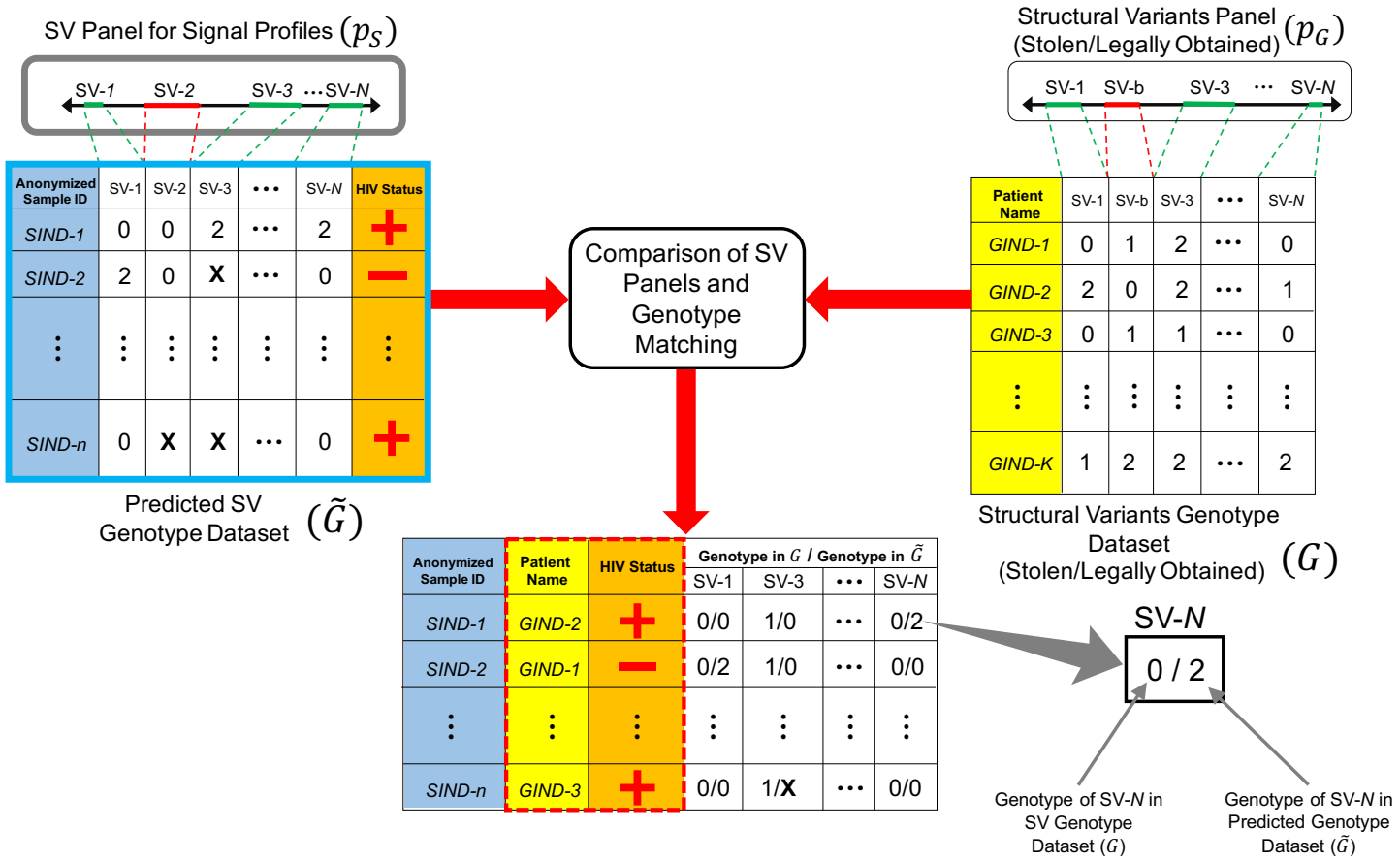
b) After Anonymization



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

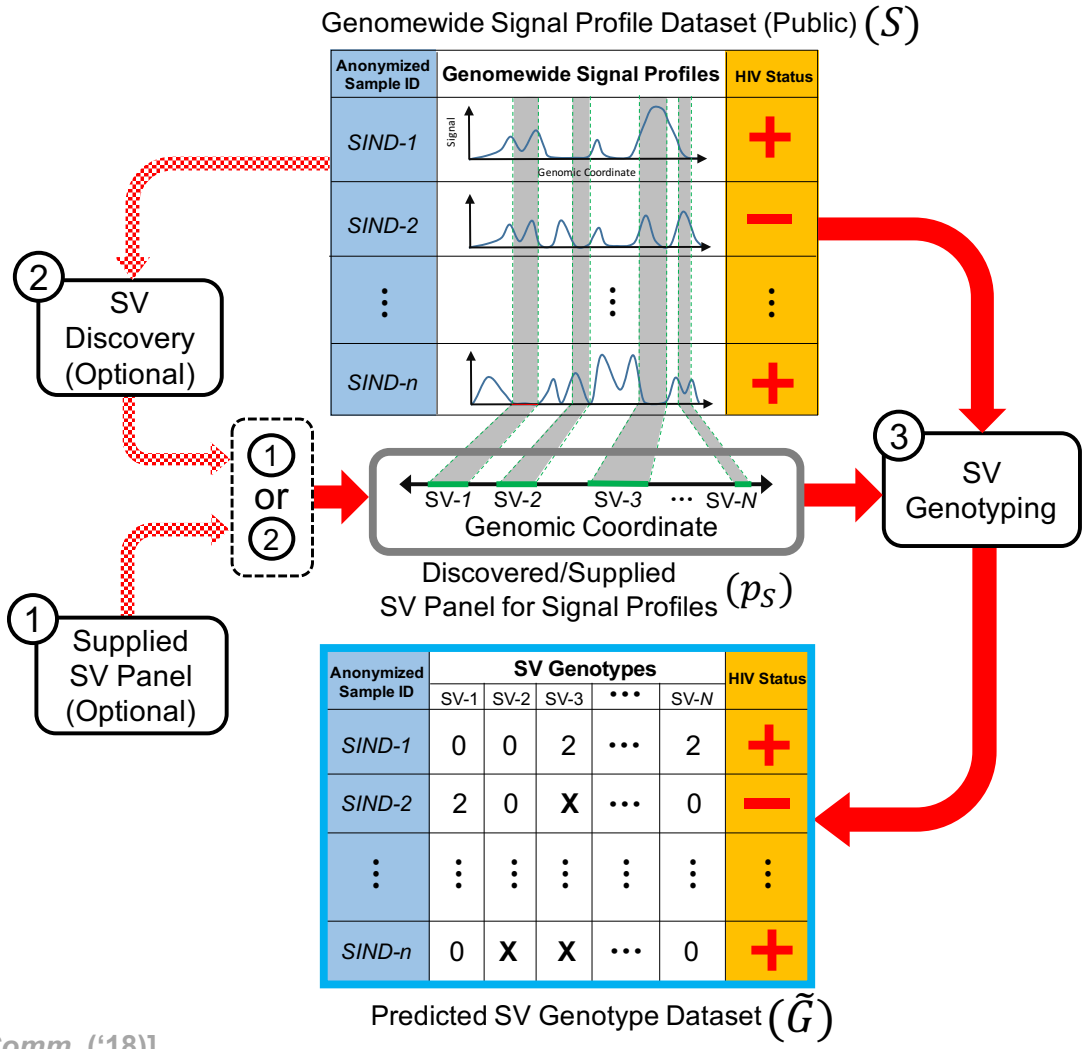
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Another type of Linking Attack: Linking based on SV Genotyping



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

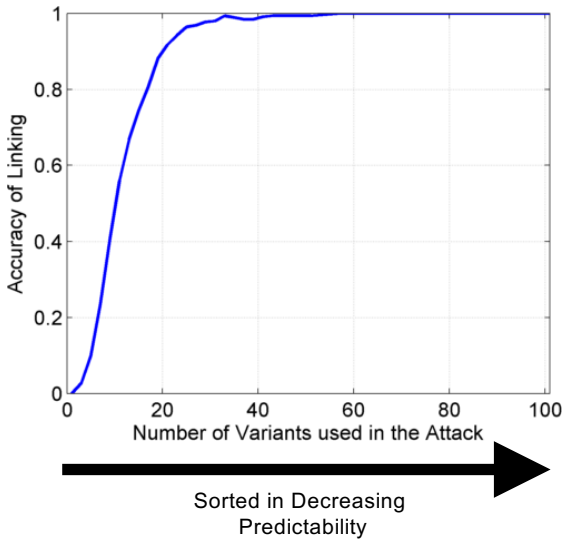
Another type of Linking Attack: First Doing SV Genotyping



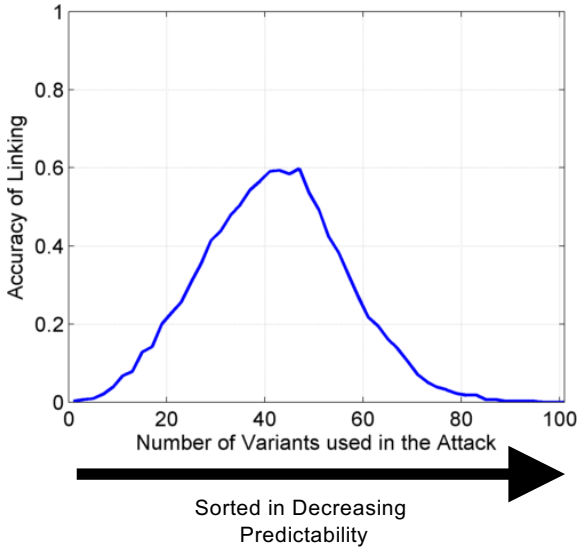
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Linking Attack Based on SV Deletions in gEUVADIS Dataset

c) Genotyping
(1kG MAF>0.01)



d) Discovery + Genotyping

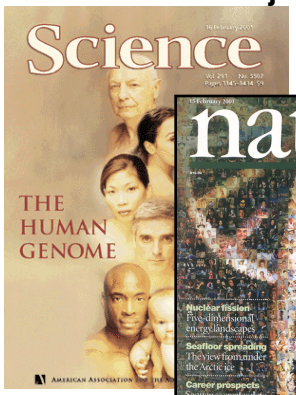


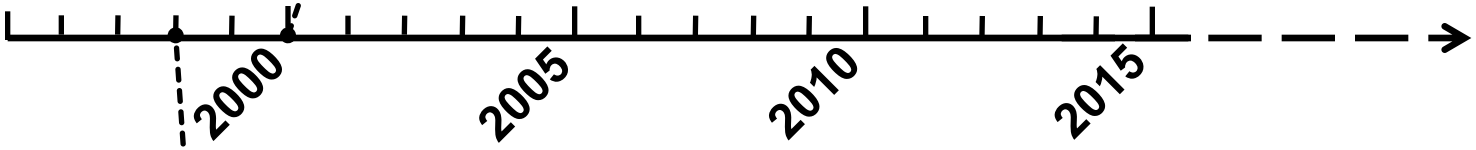
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Transcriptome Mining: Population-scale genomic analysis to better understand mental disease & the subtle privacy risks of this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to mental disease
  - Consortium intro & construction of an adult brain resource w/ 1866 individuals
  - Explanation of across-population variation via changing proportions of cell types (using single-cell deconvolution)
  - Generation of a large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C & activity relationships. Using this to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype & transcriptome
- *[Exhaust]* **Genomic Privacy** & RNA-seq
  - Introduction to 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 particularly tricky privacy issues
  - Using file formats to remove obvious variants
  - eQTLs: Quantifying & removing further variant info from expression levels w/ ICI & predictability. Instantiating a practical linking attack w/ 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
- *[Exhaust]* **Publication Patterns** from data producing consortia
  - Co-authorship network stats relate to publication rollouts & show gradual adoption by community
  - Key role of brokers in data dissemination

The Human Genome Project



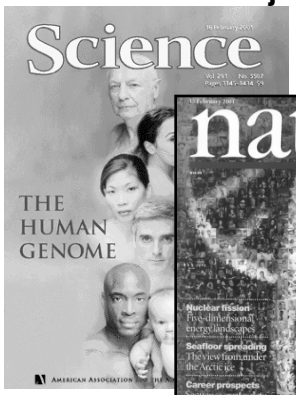


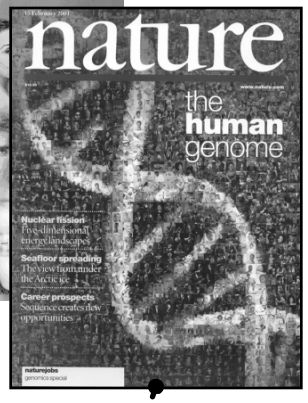
Worm Genome

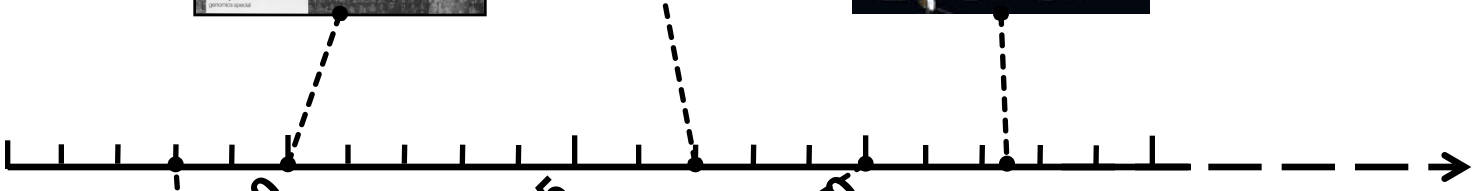
The Human Genome Project

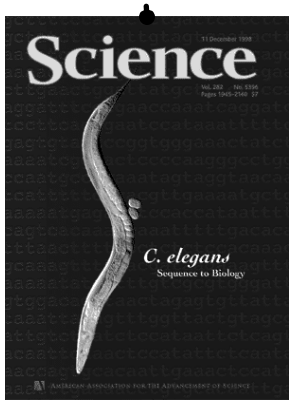
ENCODE Pilot

ENCODE Production









Worm Genome

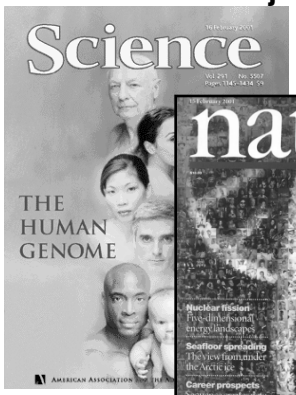
modENCODE

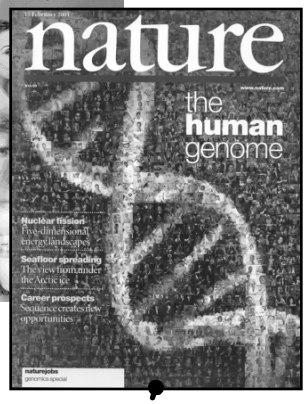
The Human Genome Project

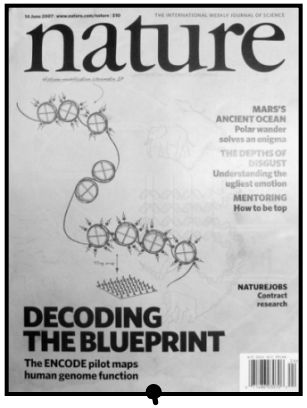
ENCODE Pilot

ENCODE Production

Comparative ENCODE





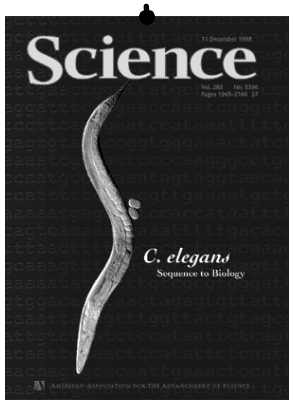


2000

2005

2010

2015



Worm Genome

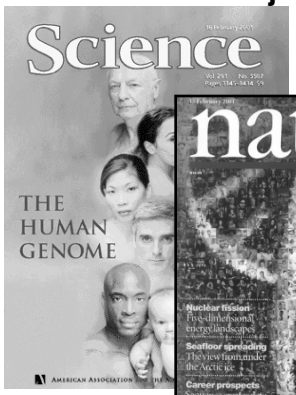
modENCODE

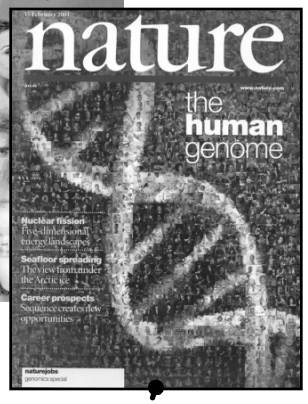
The Human Genome Project

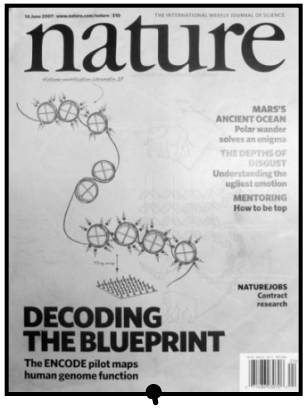
ENCODE Pilot

ENCODE Production

Comparative ENCODE





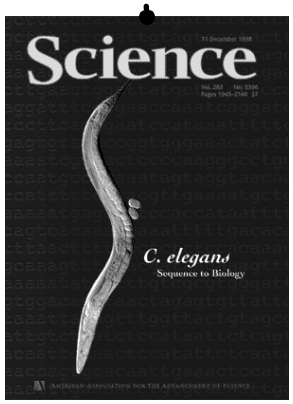


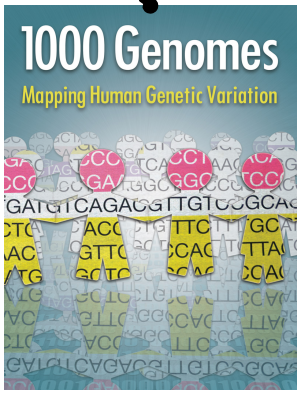
2000

2005

2010

2015





Worm Genome

modENCODE

1000 Genomes Pilot

1000 Genomes Production

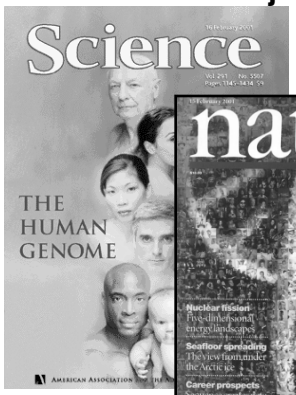
The Human Genome Project

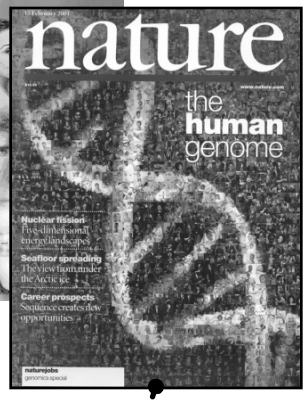
ENCODE Pilot

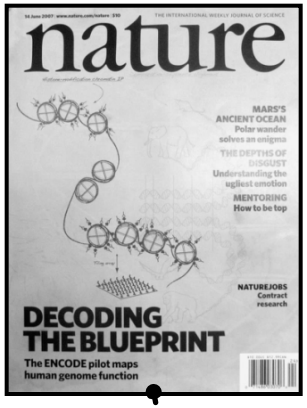
ENCODE Production

Comparative ENCODE

Epigenome Roadmap





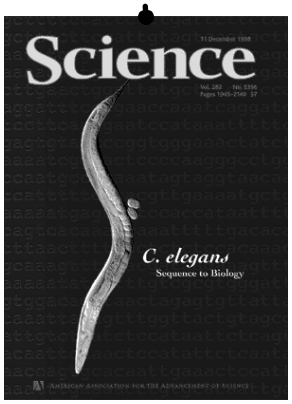


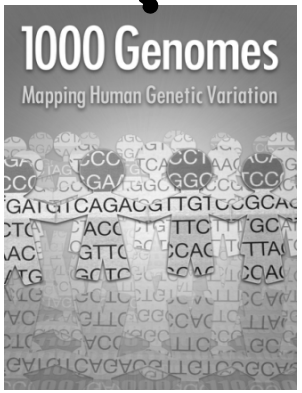
2000

2005

2010

2015





Worm Genome

modENCODE

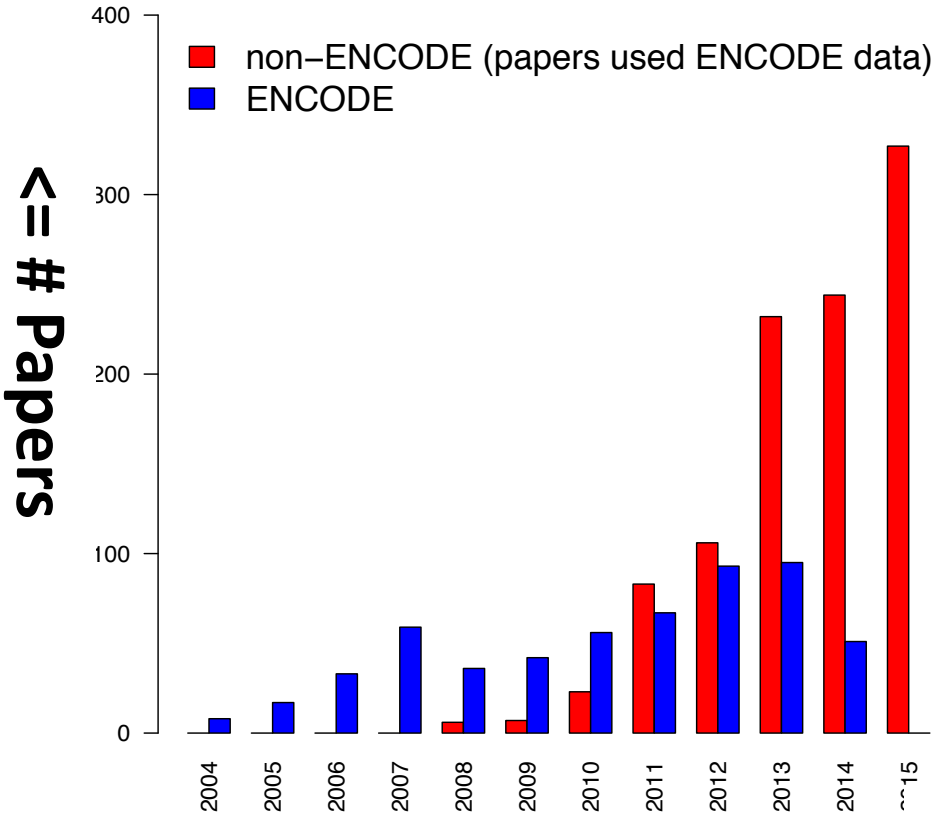
1000 Genomes Pilot

1000 Genomes Production

GTEx

With help of M Pazin at NHGRI, identified: 702 community papers that used ENCODE data but were not supported by ENCODE funding &
558 consortium papers supported by ENCODE funding
(https://www.encodeproject.org/search/?type=Publication for up-to-date query)
Then identified 1,786 ENCODE members & 8,263 non-members .

non–ENCODE (papers used ENCODE data) ENCODE



 # Authors

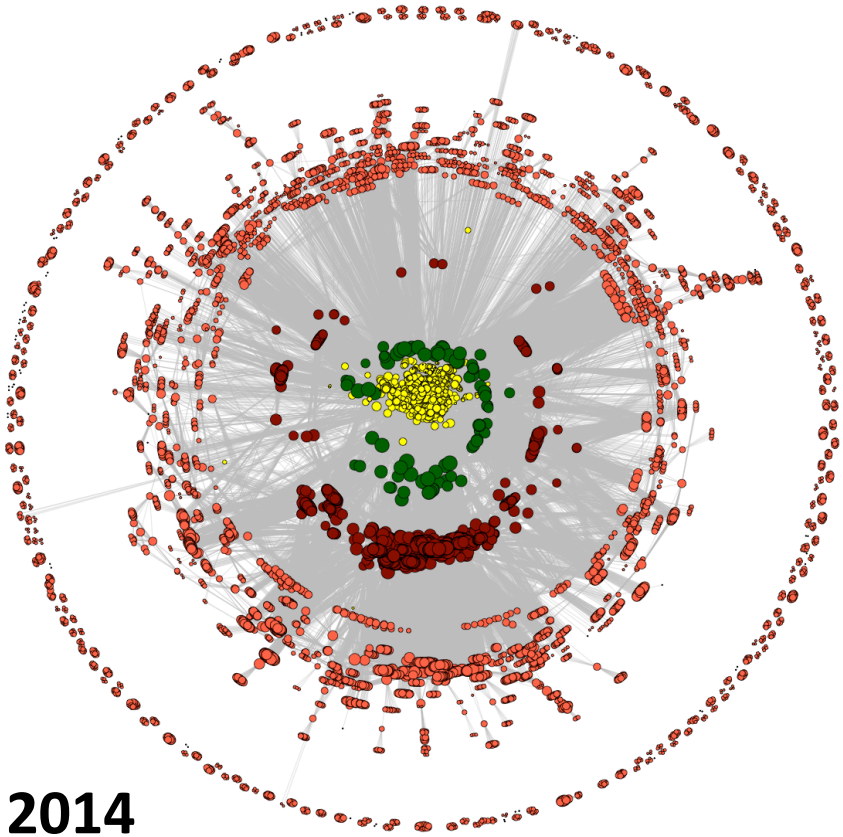
[Wang et al., TIG ('16)]

Yr. ('04 to '15)

Co-authorship Network of ENCODE members & Data Users

- ENCODE member
- non-member
- ENCODE member broker
- non-member broker

co-authorship



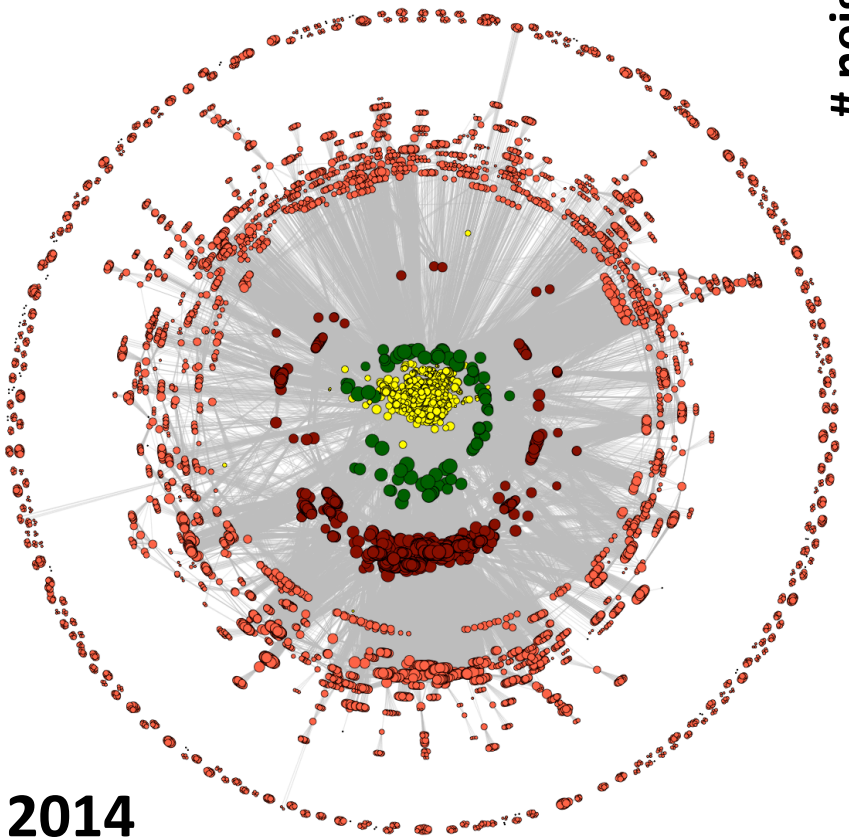
2014

[Wang et al., TIG ('16)]

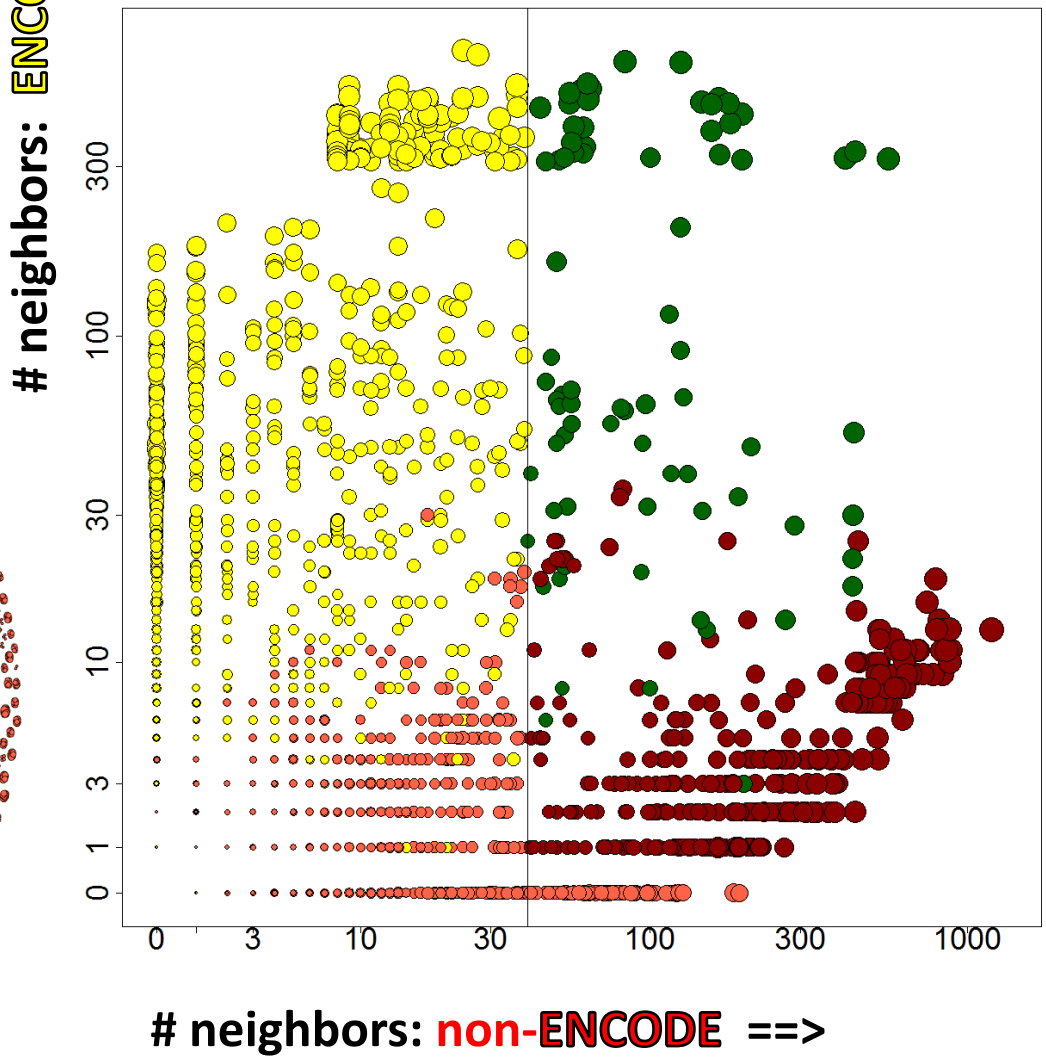
Co-authorship Network of ENCODE members & Data Users

- ENCODE member
- non-member
- ENCODE member broker
- non-member broker

co-authorship



2014

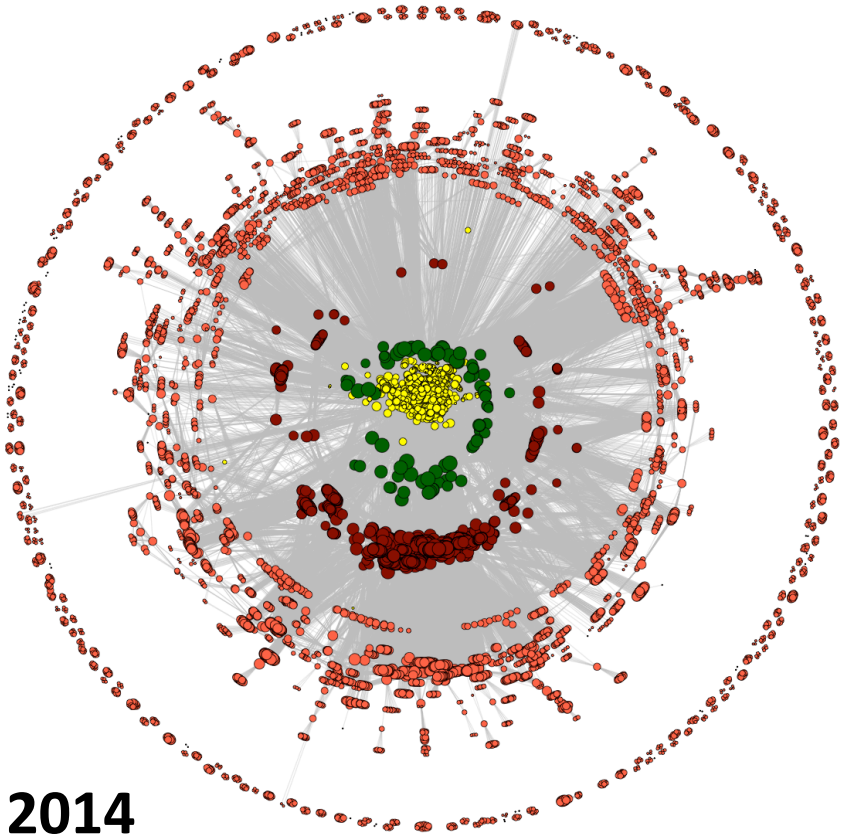


[Wang et al., TIG ('16)]

Co-authorship Network of ENCODE members & Data Users

- ENCODE member
- non-member
- ENCODE member broker
- non-member broker

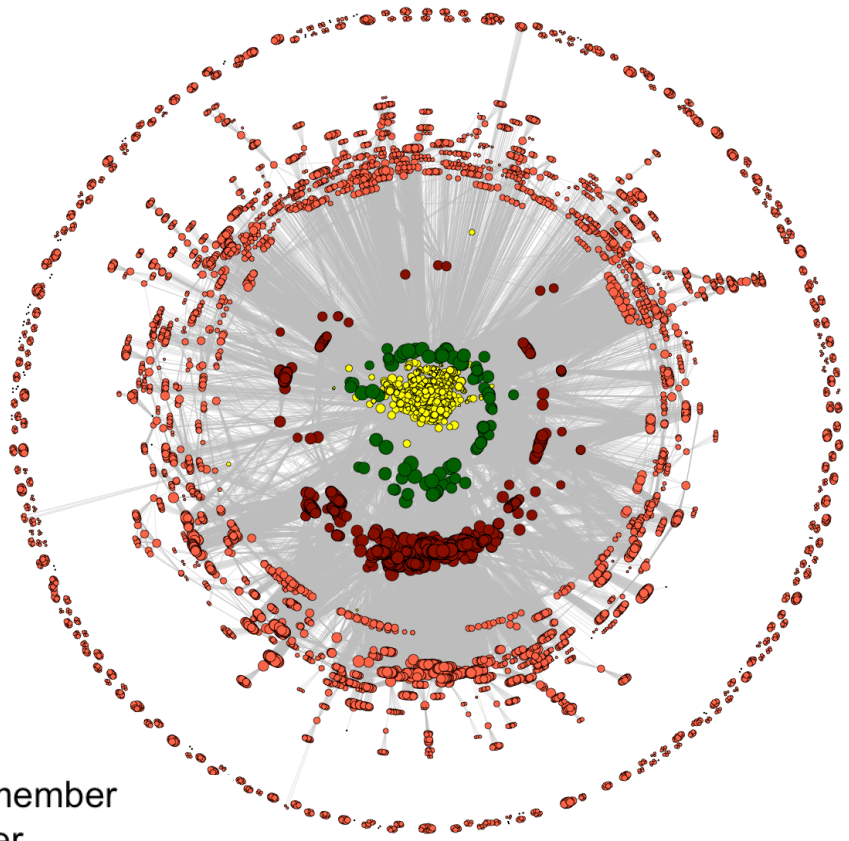
co-authorship



2014

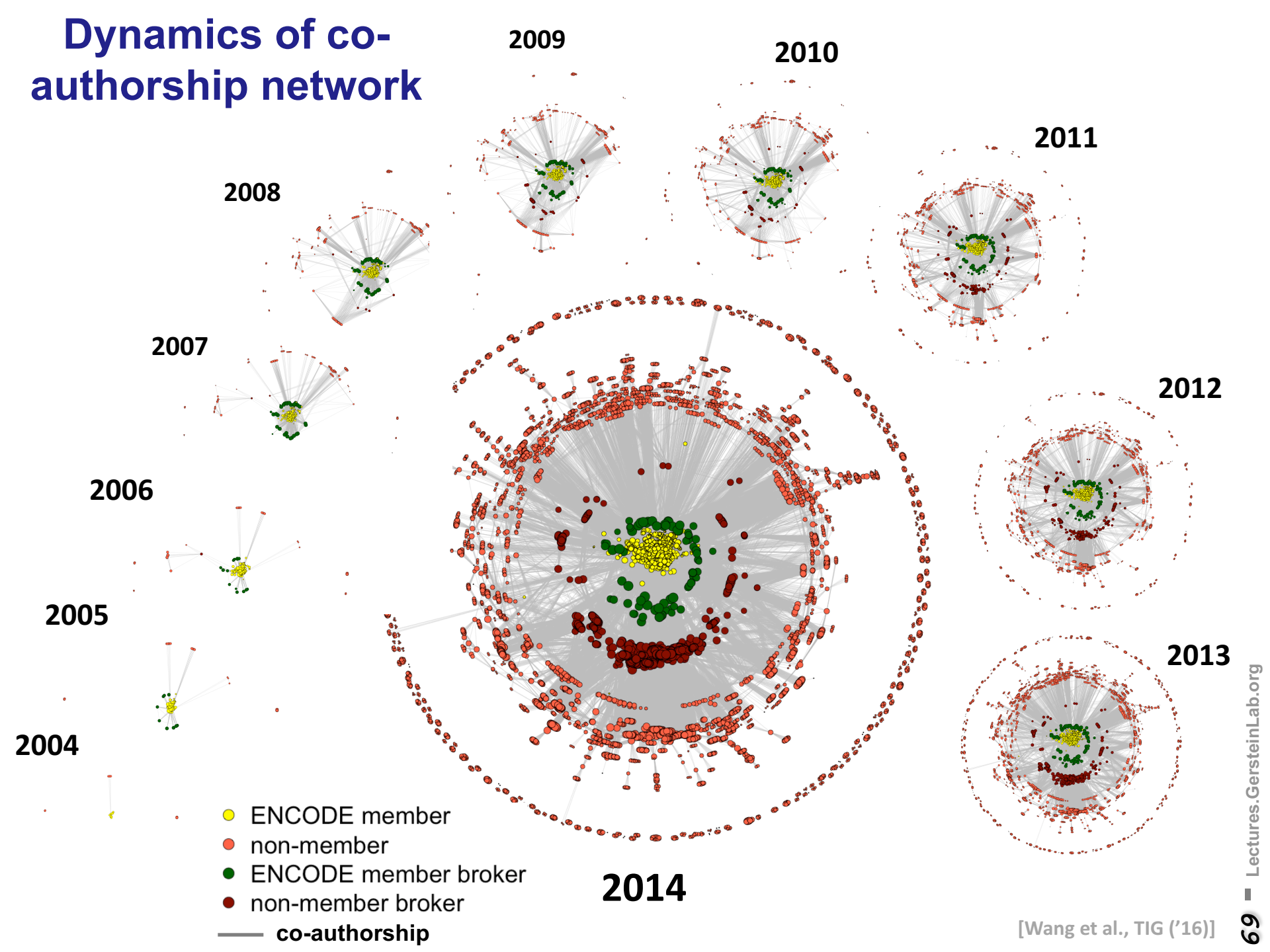
[Wang et al., TIG ('16)]

Dynamics of co-authorship network



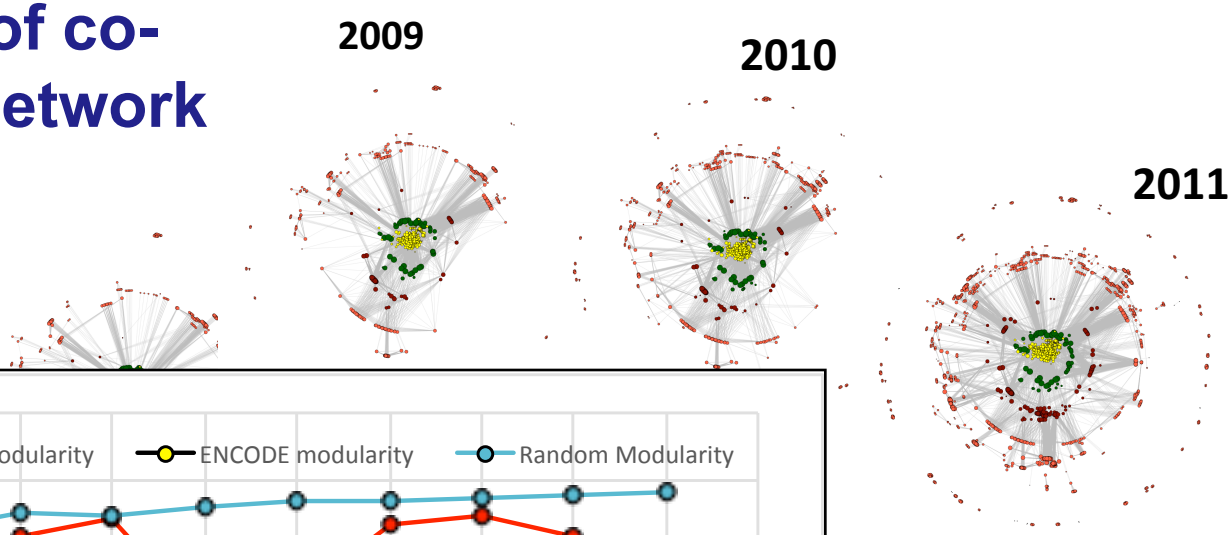
[Wang et al., TIG ('16)]

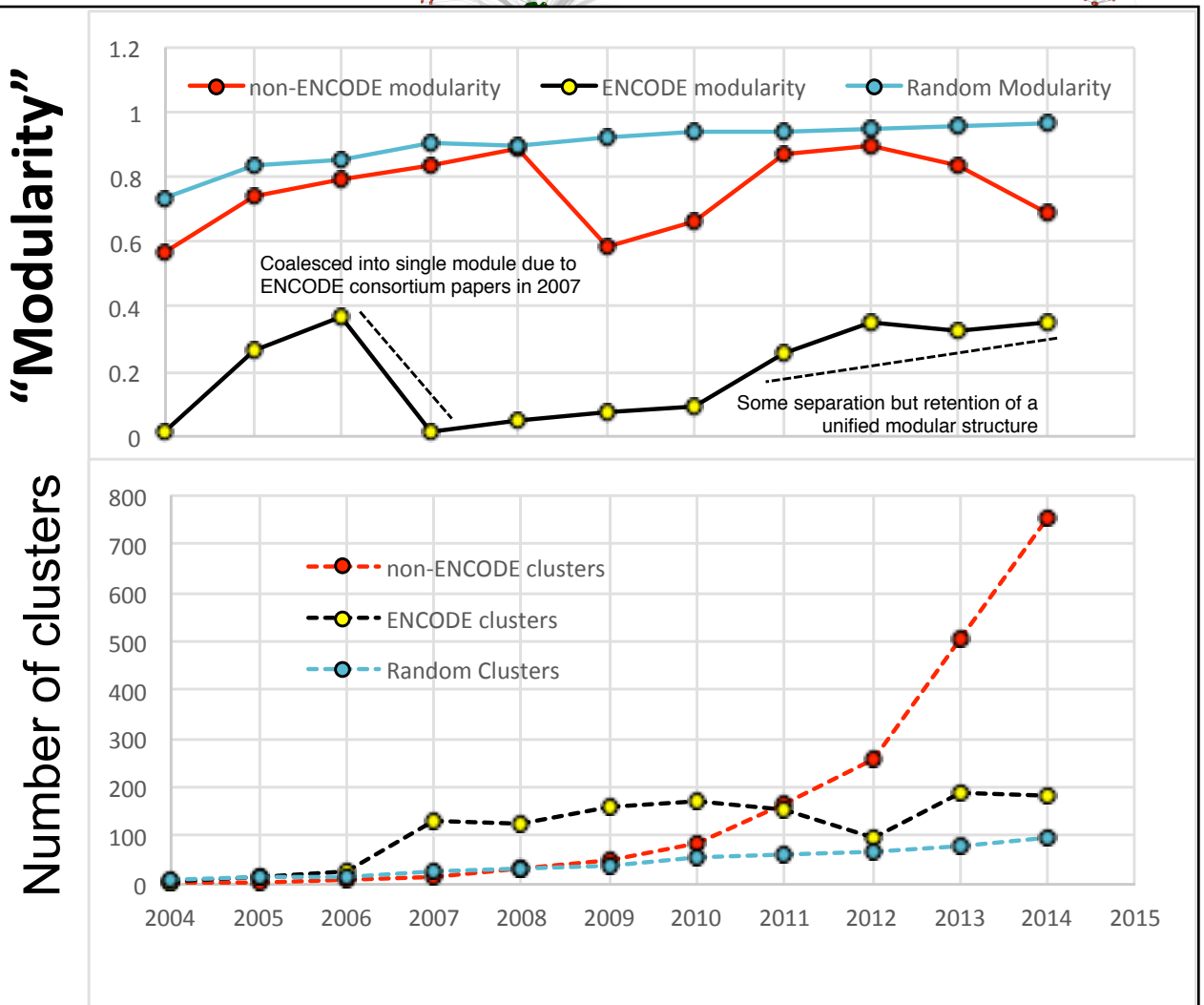
Dynamics of co-authorship network



[Wang et al., TIG ('16)]

Dynamics of co-authorship network





[Wang et al., TIG ('16)]

Transcriptome Mining: Population-scale genomic analysis to better understand mental disease & the subtle privacy risks of this activity

- *[Core]* **PsychENCODE**: Population-level analysis of functional genomics data related to mental disease
  - Consortium intro & construction of an adult brain resource w/ 1866 individuals
  - Explanation of across-population variation via changing proportions of cell types (using single-cell deconvolution)
  - Generation of a large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C & activity relationships. Using this to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype & transcriptome
- *[Exhaust]* **Genomic Privacy** & RNA-seq
  - Introduction to 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 particularly tricky privacy issues
  - Using file formats to remove obvious variants
  - eQTLs: Quantifying & removing further variant info from expression levels w/ ICI & predictability. Instantiating a practical linking attack w/ 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
- *[Exhaust]* **Publication Patterns** from data producing consortia
  - Co-authorship network stats relate to publication rollouts & show gradual adoption by community
  - Key role of brokers in data dissemination

Transcriptome Mining: Population-scale genomic analysis to better understand mental disease & the subtle privacy risks of this activity

- *[Core]* **PsychENCODE:** Population-level analysis of functional genomics data related to mental disease
  - Consortium intro & construction of an adult brain resource w/ 1866 individuals
  - Explanation of across-population variation via changing proportions of cell types (using single-cell deconvolution)
  - Generation of a large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C & activity relationships. Using this to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype & transcriptome
- *[Exhaust]* **Genomic Privacy** & RNA-seq
  - Introduction to 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 particularly tricky privacy issues
  - Using file formats to remove obvious variants
  - eQTLs: Quantifying & removing further variant info from expression levels w/ ICI & predictability. Instantiating a practical linking attack w/ 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
- *[Exhaust]* **Publication Patterns** from data producing consortia
  - Co-authorship network stats relate to publication rollouts & show gradual adoption by community
  - Key role of brokers in data dissemination

PsychENCODE Acknowledgment

NIH 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

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

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;.

Acknowledgements

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

PrivaSig.gersteinlab.org

PrivaSeq.gersteinlab.org

A **Harmanci**,
D **Greenbaum**
G **Gürsoy**,
F Navarro, M Green

Publication patterns [“encode authors”]

D **Wang**, KK Yan, J Rozowsky, E Pan

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