Developing Computational Representations of Disease-Relevant Molecules:

3 Cases Studies for AI in Biomedicine

Mark Gerstein, Yale

Slides freely downloadable from
Lectures.GersteinLab.org
& "tweetable" (via **@MarkGerstein**).
No Conflicts for this Talk. See last slide for more info.

Backgrounds

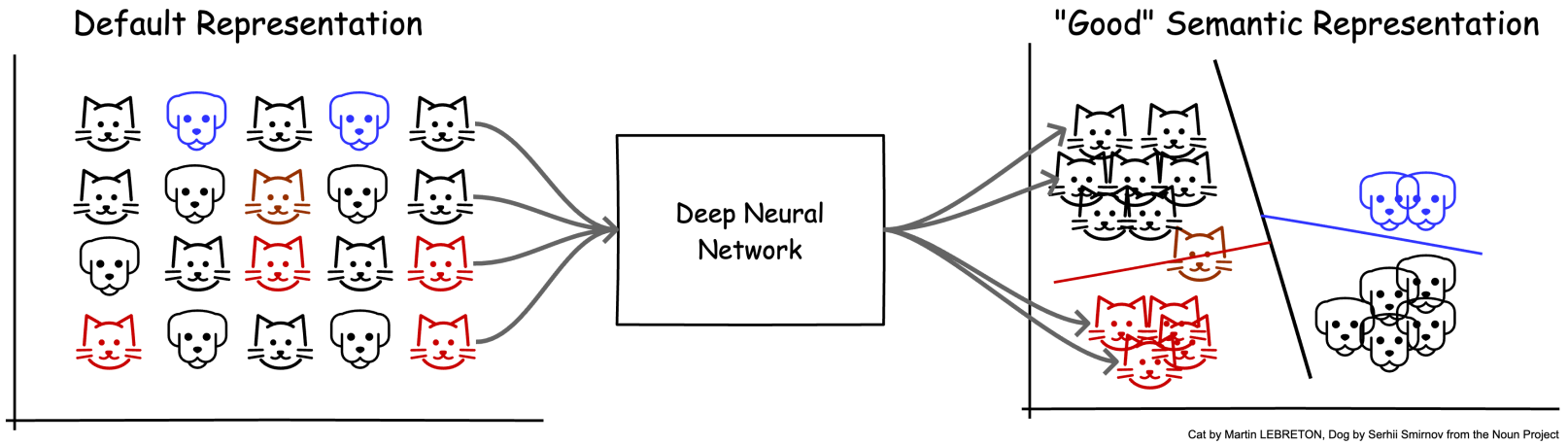
Learning meaningful representations
from large, complex biological data

Representation learning

- "*An AI must fundamentally understand the world around us, and we argue that this can only be achieved if it can learn to identify and disentangle the* ***underlying explanatory factors*** *hidden in the observed milieu of low-level sensory data.*"

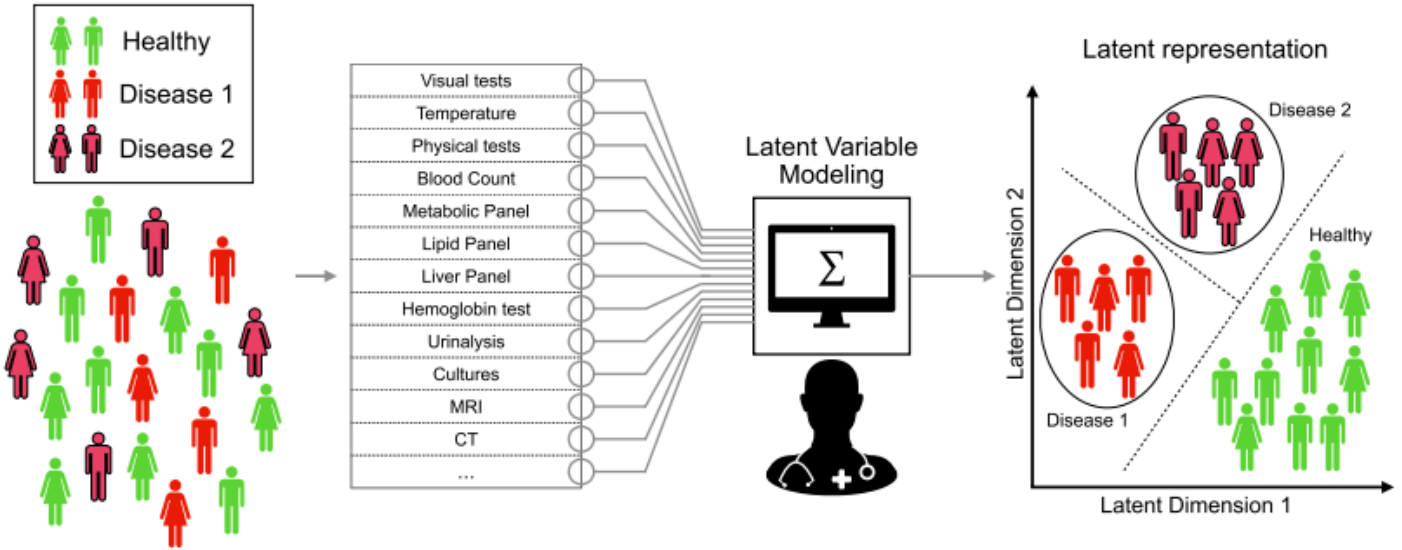
Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1798-1828.

Representation learning



https://blog.fastforwardlabs.com/2020/11/15/representation-learning-101-for-software-engineers.html

Representation learning



Kopf, Andreas, and Manfred Claassen. "Latent representation learning in biology and translational medicine." *Patterns* 2.3 (2021).

Representation learning

- Particularly well suited to molecular biomedical data because of
  - its scale and high-dimensionality
  - its difficulty for easy interpretation
- When it comes to scientific data:
  - We only have partial knowledge about their internal structures
  - Thus, if the learned representations could re-discover some known patterns in the data, they could can help us discover more potentially meaningful ones.

Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

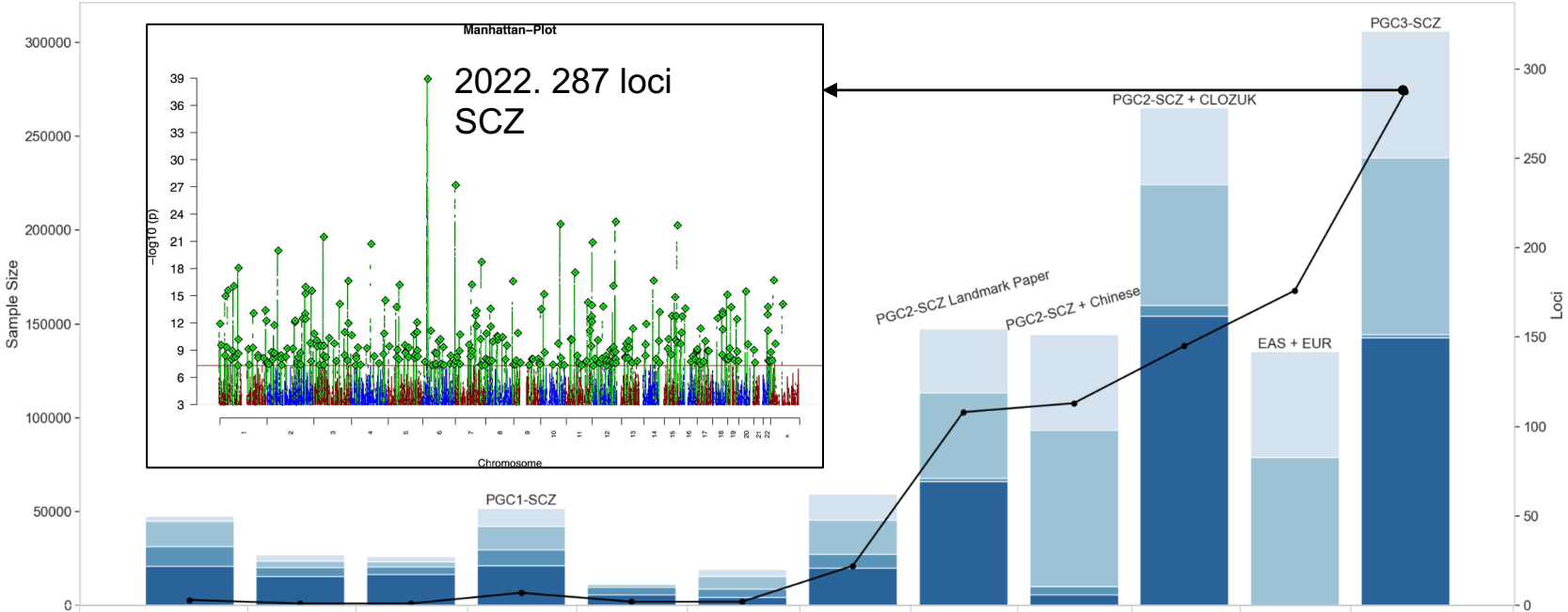
- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

Many psychiatric conditions are highly heritable in comparison to other disorders, but their mechanisms are unknown

Disease	Heritability	SNP-based Heritability	PMID	Sample Size (1000s)	Molecular **Mechanisms**
Schizophrenia	**79%**	**24%**	35396580	320	**(C4A)**
Bipolar disorder	60-85%	16-19%	34002096	414	**-**
Alzheimer's disease	60-80%	3%	34493870	1,127	APOE, Tau
Hypertension	30%	18%	38689001	1,029	Renin–angiotensin–aldosterone
Heart disease	50-60%	6%	26343387	184	Atherosclerosis, VCAM-1
Stroke	32%	17%-21%	33773637	262	Reactive oxygen species (ROS), Ischemia
Type-2 diabetes	26%	20%	30054458	659	Insulin resistance
Breast Cancer	31%	22%	38741014	40	BRCA, PTEN

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

Great Progress in Finding Variants Related to Brain Diseases: The history of reported schizophrenia GWAS



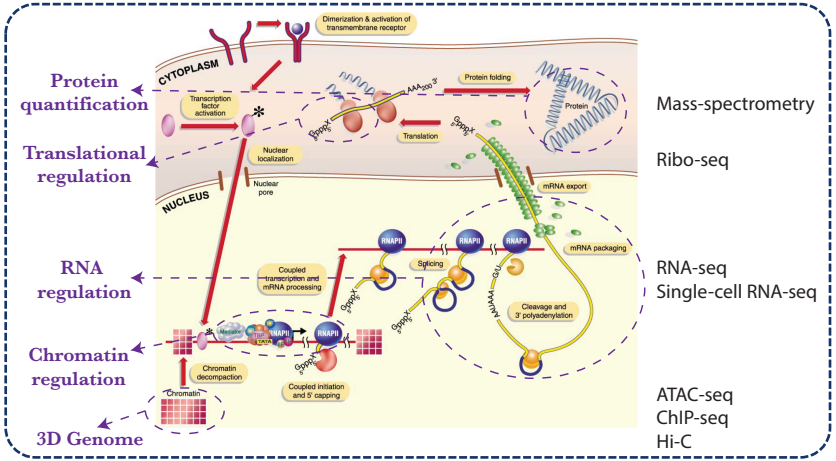
PMID	19571808	19571811	19571809	21926974	22037552	22037555	23974872	25056061	28991256	29483656	31740837	35396580
Replication-Control	21093	15490	16424	21397	5603	4539	19762	66236	5770	154224	0	142626
Replication-Case	10282	4686	4327	8442	4027	4383	7413	1513	4384	5662	0	1979
Discovery-Control	13498	3587	2653	12462	1599	6468	18310	45604	83493	64643	78818	94015
Discovery-Case	2663	3322	2681	9394	746	3750	13833	34241	50874	40675	56418	67390
loci	3	1	1	7	2	2	22	108	113	145	176	287
SNP-based heritability												0.24
Citation	Figures adopted and updated from Sophie E. Legge et al. Psychological Medicine. 2021; Vassily Trubetskoy et al. Nature. 2022											

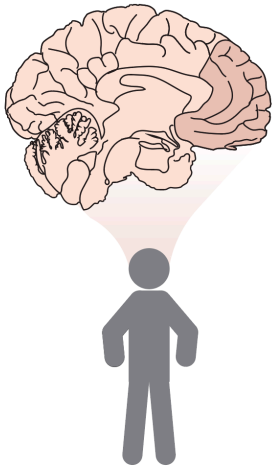
Assessing gene regulation to understand psychiatric disorders

Addressing the fact that molecular mechanisms are not known for most psychiatric disorders

Genotype → Intermediate Molecular Endophenotypes → Macrophenotype (Psychiatric Disorders)

?





Mechanistic Insight ← → Clinical Insight

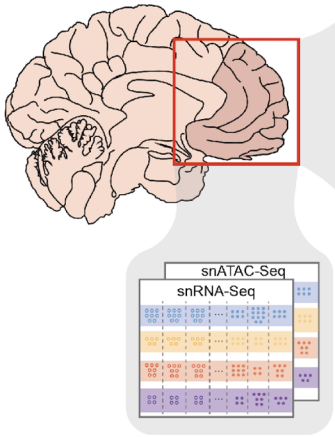
lectures.gersteinlab.org

Figure derived from Orphanides and Reinberg, *Cell* 2002

The PsychENCODE Consortium: Focusing on the PFC

PsychENCODE 200 researchers at 40 institutions

Main goal: *Understand the genetic, genomic and epigenomic etiologies of schizophrenia, bipolar disorder, autism spectrum disorder, and other neuropsychiatric disorders*



The prefrontal cortex (PFC) not only governs executive functions, but is also responsible for:

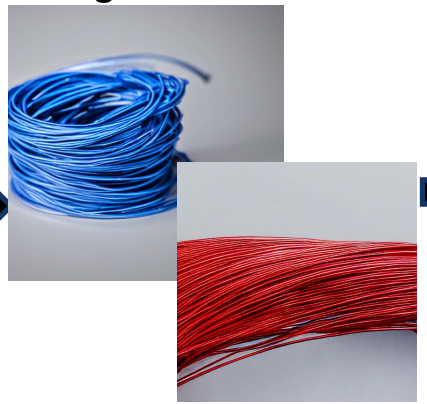
- behavioral regulation and mental health
- development and plasticity
- interplay with neurotransmitter systems

Elert, E. *Nature* **508**, S2–S3 (2014)

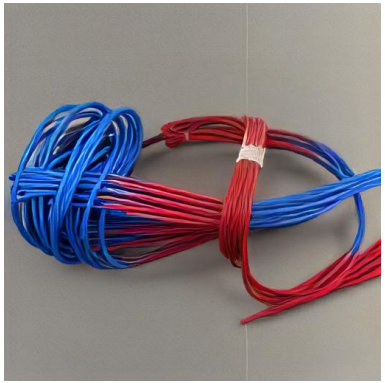
Advantages of single-cell resolution in the brain

Bulk Datasets

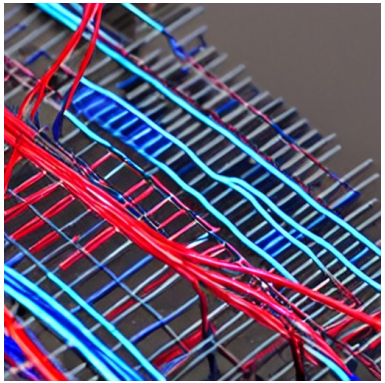
Single-cell Datasets



Spatial Analysis



Cell-to-cell Connectivity



Images generated using the DeepAI Image Generator tool

How PsychENCODE fits into the history of genome annotation

2000

2005

2010

2015

2020

2025

Worm Genome

modENCODE

1000 Genomes Pilot

GTEx

1000 Genomes Production

BICCN

HuBMAP

The Human Genome Project

ENCODE Pilot

ENCODE Production

Comparative ENCODE

TCGA

PsychENCODE

Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

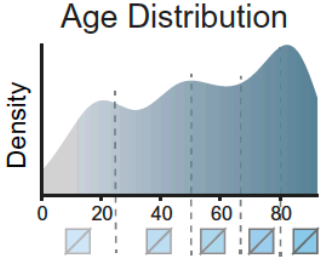
- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

Integrating multi-omics data for 388 adult brains

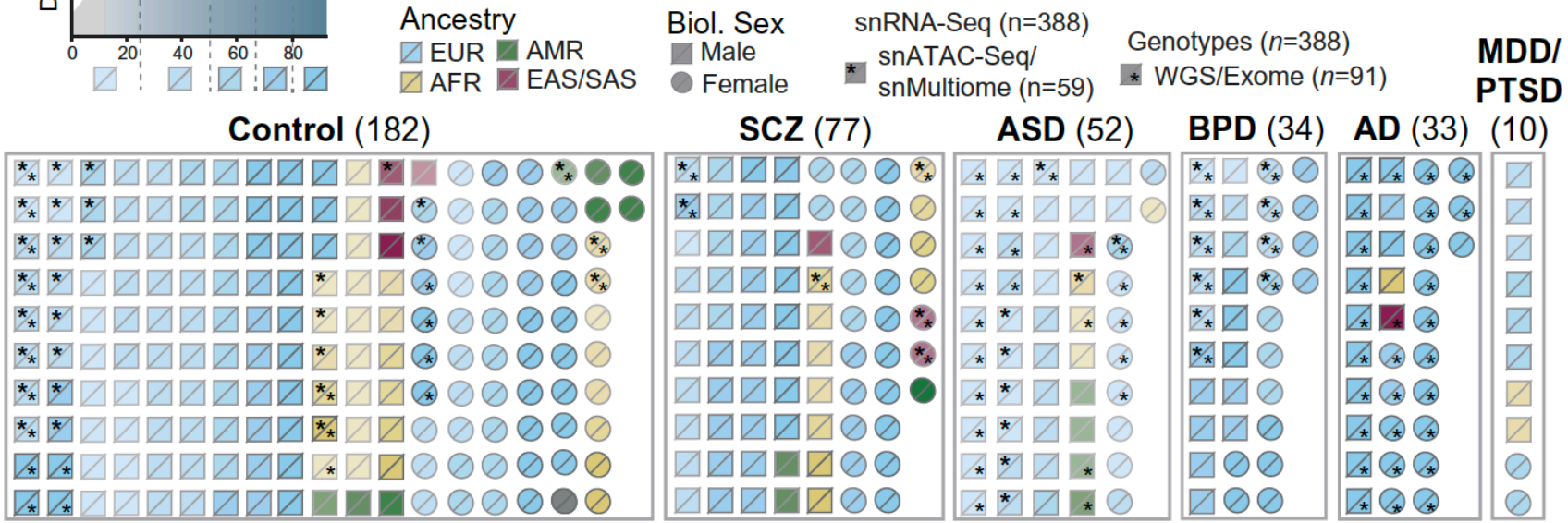
388 Samples

2.8M Nuclei



Single-cell data for 388 individuals (snRNA-seq, genotypes, ~60 snATAC-seq) – **one of the largest single-cell collections in the human brain**

Integrated study derived from 12 cohorts (PEC, AMP-AD, & other studies) for population and cross-disorder comparisons



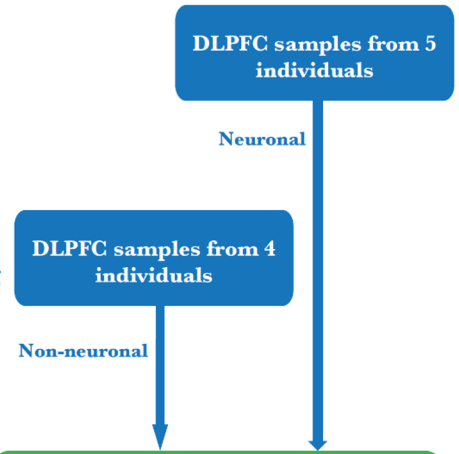
[Emani et al. ('24) Science]

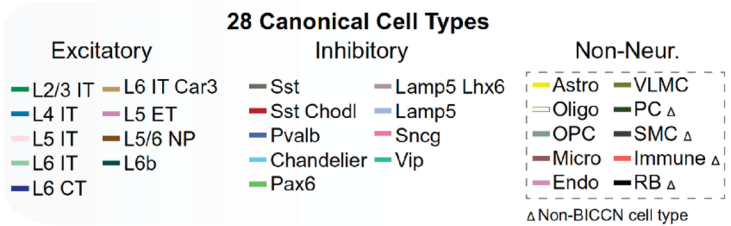
Objective. To synchronize activities across consortia for consistent DLPFC cell type definitions

BICCN:
Multi-modal Cross-brain-region Reference Atlas

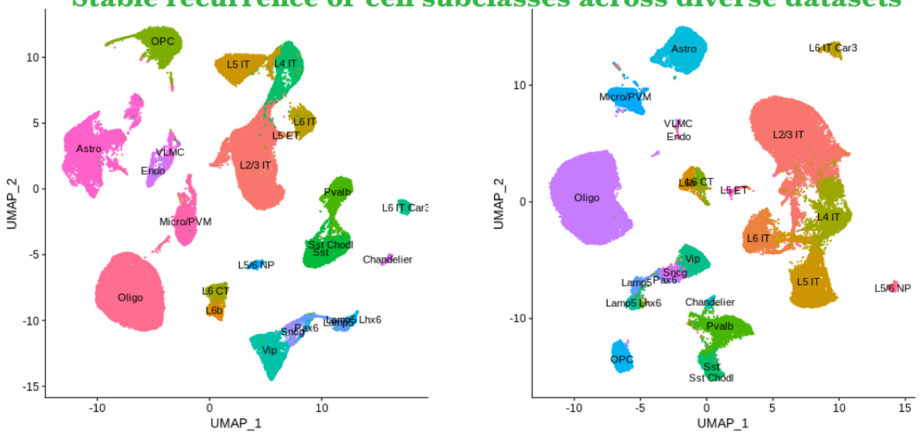
Ma-Sestan:
Focused PFC study with deeper sampling of non-neuronal cells

PsychENCODE:
Validation on psychiatric disorder snRNAseq datasets

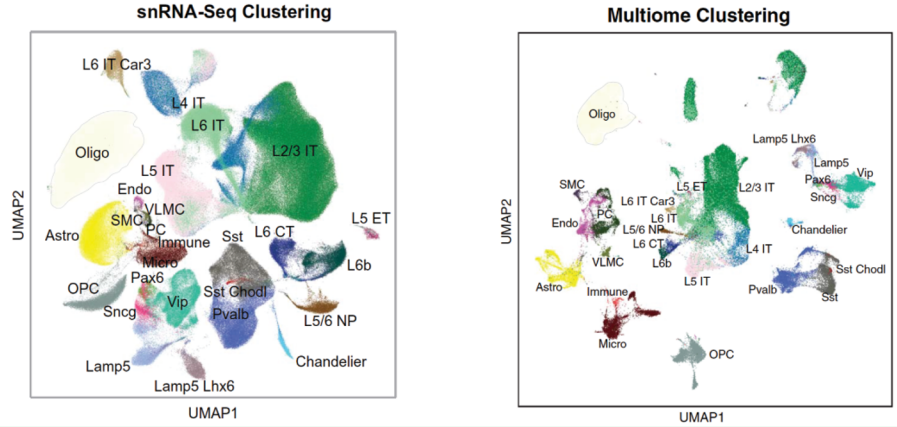




Stable recurrence of cell subclasses across diverse datasets



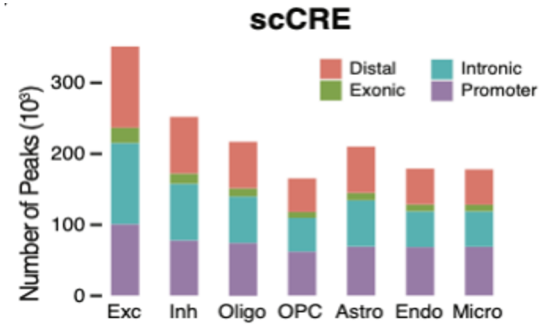
Stable recurrence of cell subclasses across different modalities

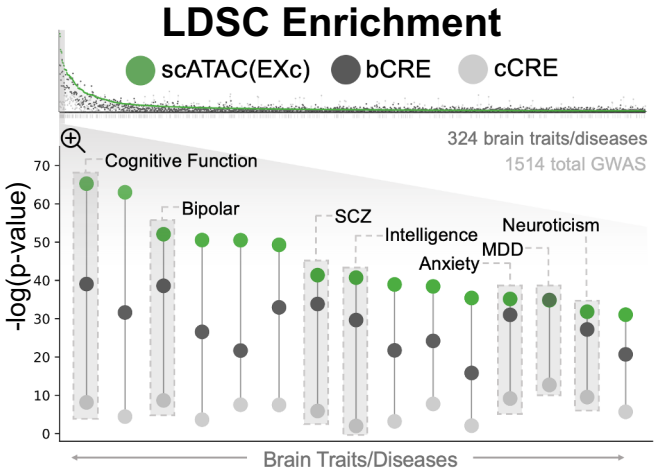


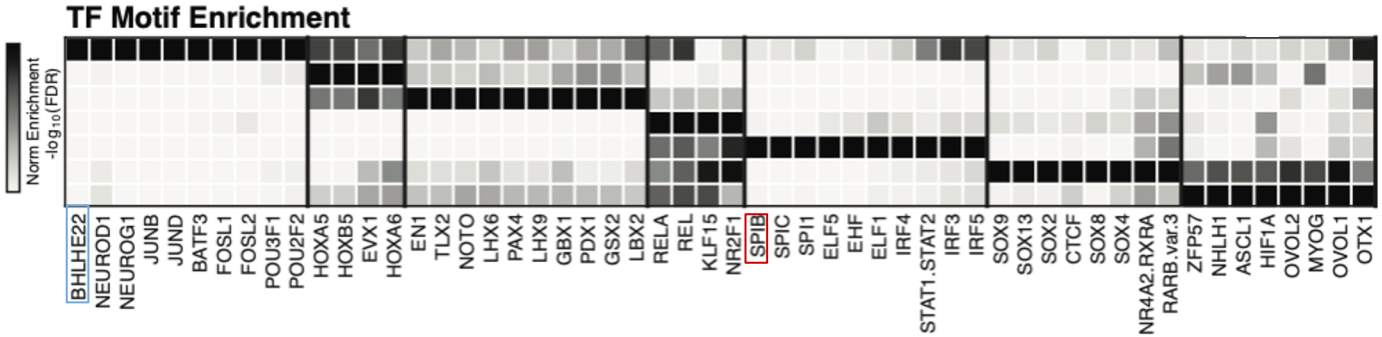
[Emani et al. ('24) Science]

scCREs show specific enrichment for TF motifs and GWAS signals

~560,000 single-cell cis-regulatory elements (scCREs) from ATAC peaks, more enriched for brain traits in GWAS than bulk cCREs







Cell-type-specific TF motif patterns - major brain cell types employ distinct groups of TFs.

[Emani et al. ('24) Science]

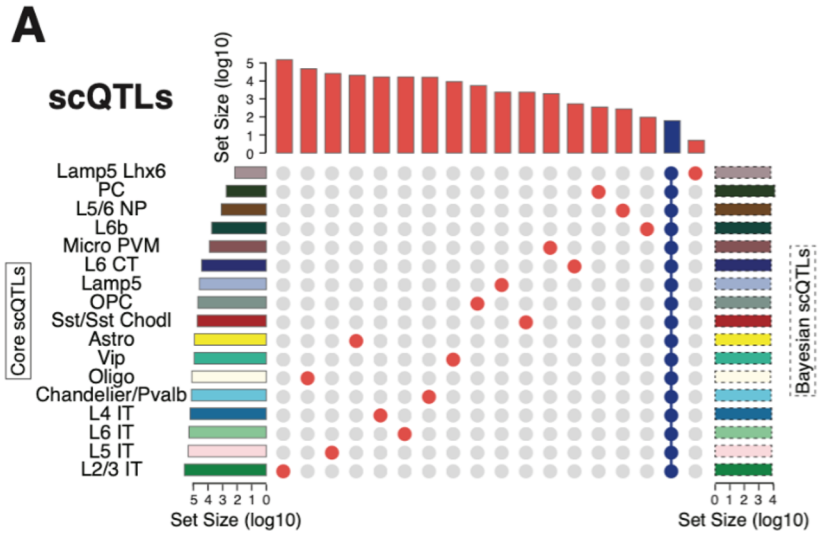
Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

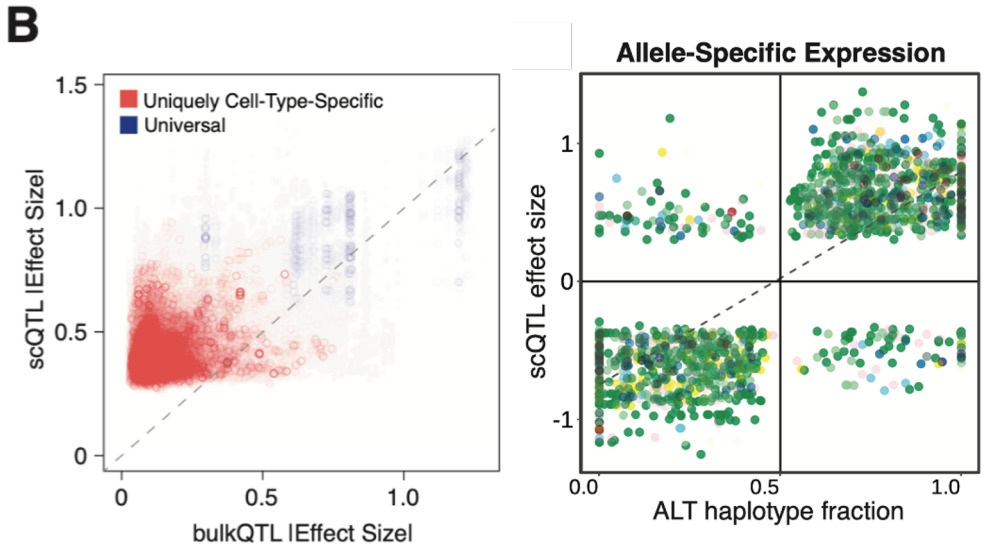
- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

Cell type-specific eQTLs (scQTLs)





Core set of 1.4M scQTLs based on standard GTEx QTL scheme

Overall, ~85K scQTLs & ~690 significant eGenes per cell type

~53% scQTLs cell-type-specific

Bayesian methods sharing information between cell types can identify more scQTLs in rarer cell types

Dilution Effect implies many scQTLs will not be seen in bulk

For scQTLs that overlap with bulk eQTLs, scQTL effect is larger & diluted out in bulk

Validation using single-cell ASE

Consistency betw. scQTL effect sizes and ALT haplotype fractions

[Emani et al. ('24) Science]

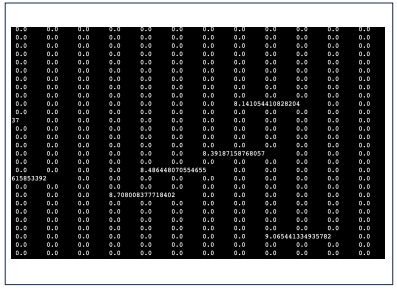
Challenges in Calling Cell type-specific eQTLs (scQTLs)

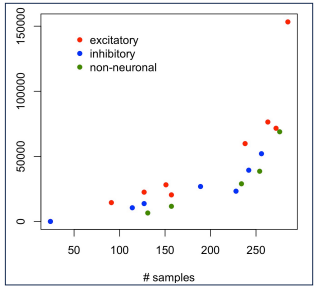
Matrix gymnastics: Cross-study data integration, filtering, and matrix synchronization

> Filtering:
- Genes
- Nuclei
- Variants
- Individuals

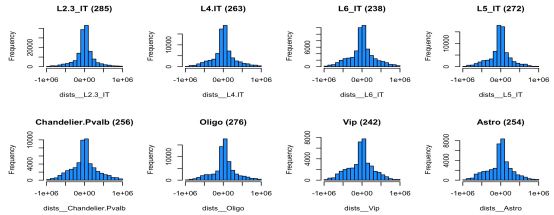
> Matrix synchronization

Data sparsity and limited statistical power in snRNA-seq contexts





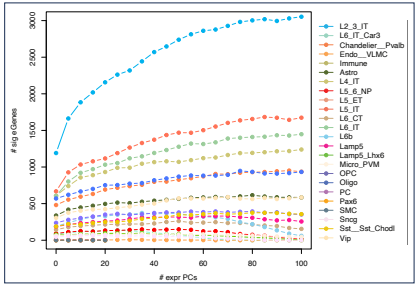
‘Semi-blind’ validation: Devising and performing quality checks and ‘validation’ without gold-standard reference dataset for comparisons

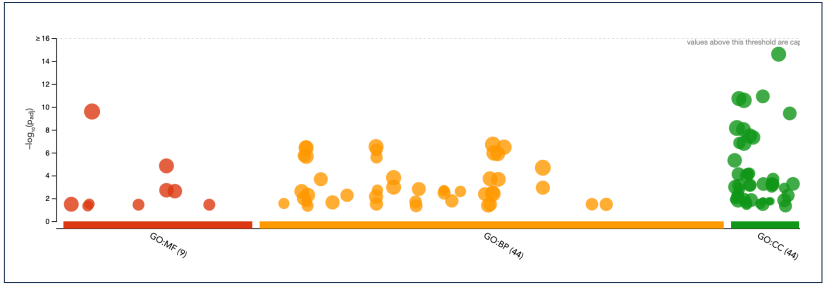


Slowly explore ‘decision space’ -- details re. pre-processing (e.g., expr normalization, etc)

- log TPM-normalization?
- log CPM-normalization?
- sc-transform normalization?
- TMM-normalization?
- Thresholds for # min nuclei & samples
- Stage to enforce MAF filters

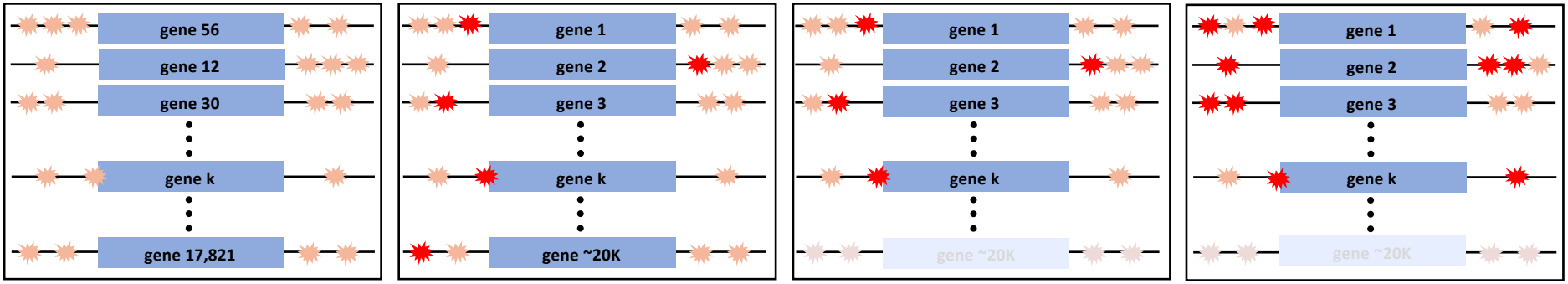
Batch effects (mult. cohorts): Optimizing calling setups (ex: selecting covariates and numbers of PCs to include)





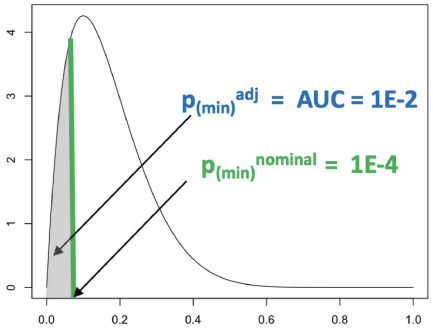
[Emani et al. (‘24) Science]

Multi-step (hierarchical) scheme to identify significant eGenes & their associated eSNPs (GTEx compatible approach)

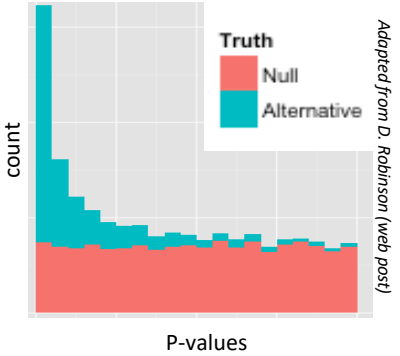


Genes are ordered by descending significance

Step 1: Identify the most significant eSNP per gene, and then correct p-values for multiple testing within each gene to derive adjusted *gene-level p-values*



Step 2: Multiple testing correction (BH to estimate FDR) is applied to the set of all ~20K adjusted gene-level p-values to yield the *significant eGenes* (FDR 0.05)



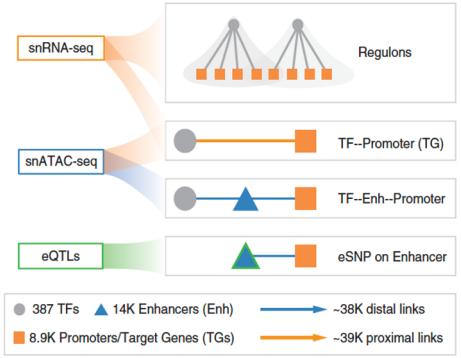
Step 3: Pull in all *significant eSNPs* associated with each significant eGene by using the scheme adopted by GTEx

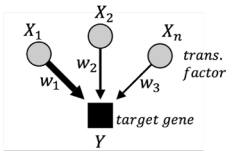
$$p_{nomimanl\_thresh}^{(eGene)} = F^{-1}(p_t)$$

$$F(p_{nomimanl\_thresh}^{(eGene)}) = p_t$$

[Emani et al. ('24) Science]

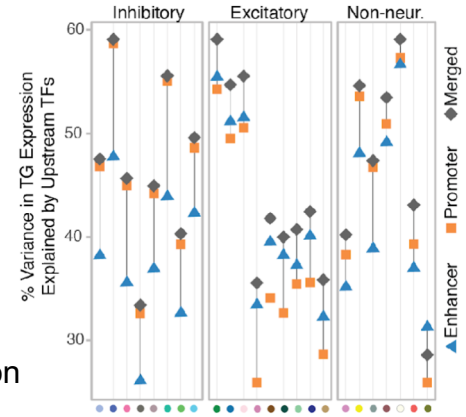
Constructing cell-type-specific gene regulatory networks

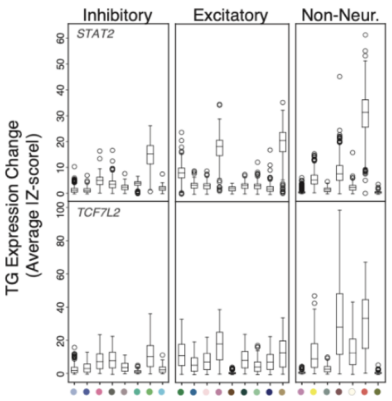


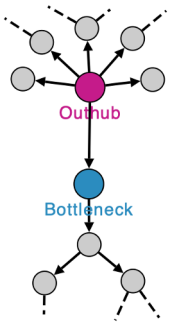


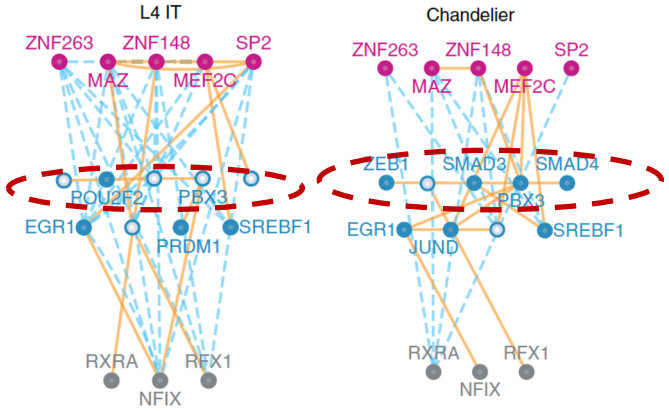
$$Y_i = \beta_0 + \sum_{j=1}^{n} \beta_j \cdot (X_j \cdot w_j) + \epsilon$$

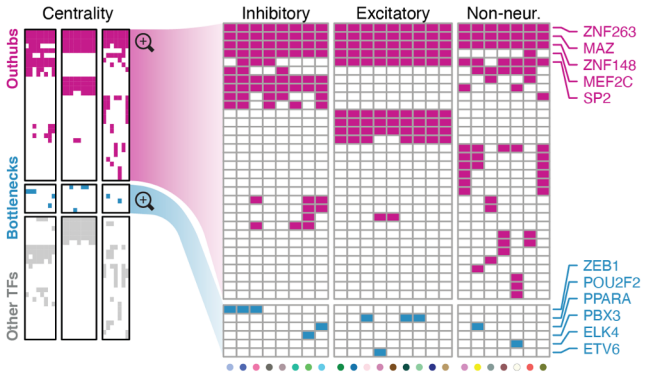
TF expression explains 52% of variation in target gene expression







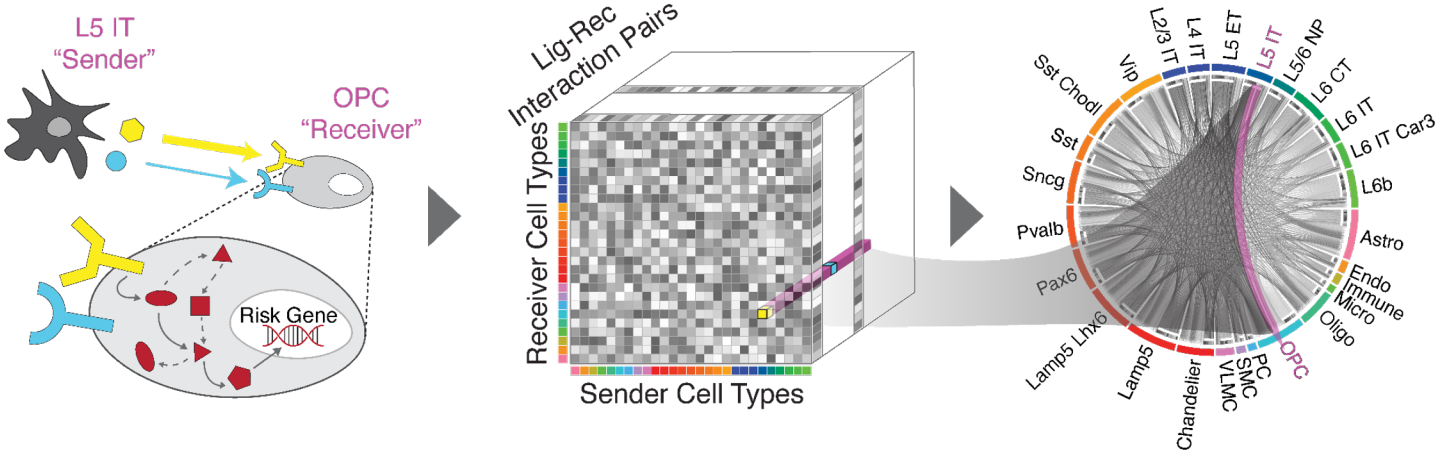




TFs show differential usage (i.e. out-hubs, bottlenecks) across cell-type GRNs

[Emani et al. ('24) Science]

Disease-specific alterations in cell-to-cell communication

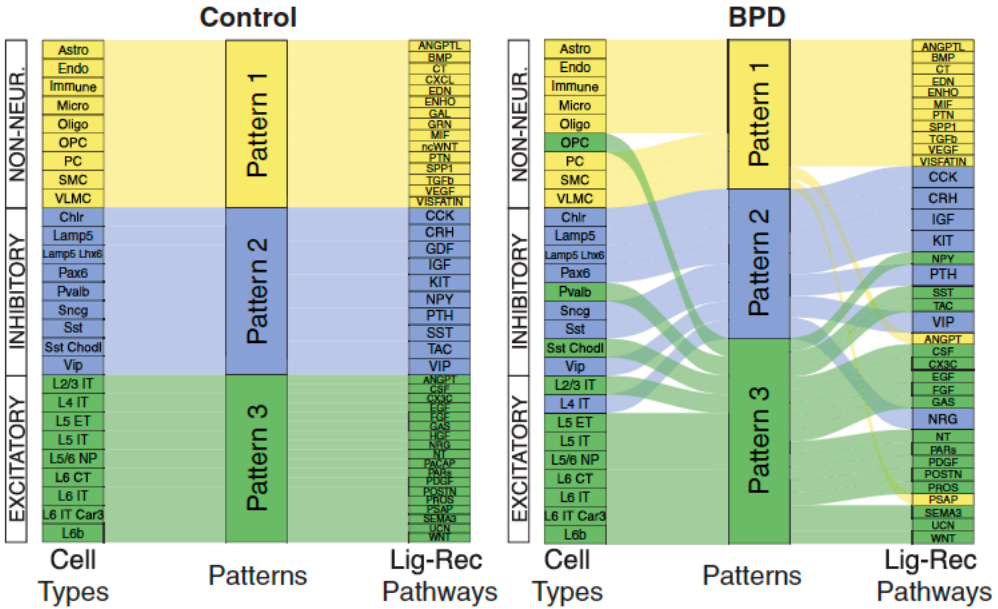


726-edge network
(CellChat)

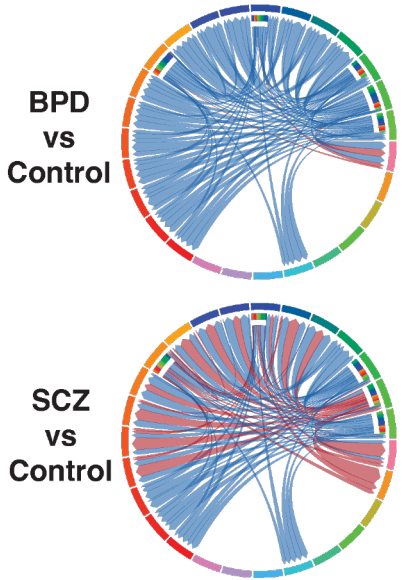
[Emani et al. ('24) Science]

Disease-specific alterations in cell-to-cell communication

Outgoing Communication Pattern Clustering



Differential Analysis of WNT Pathway



Large-scale changes in cell-cell communication patterns seen in individuals with neuropsychiatric disorders

[Emani et al. ('24) Science]

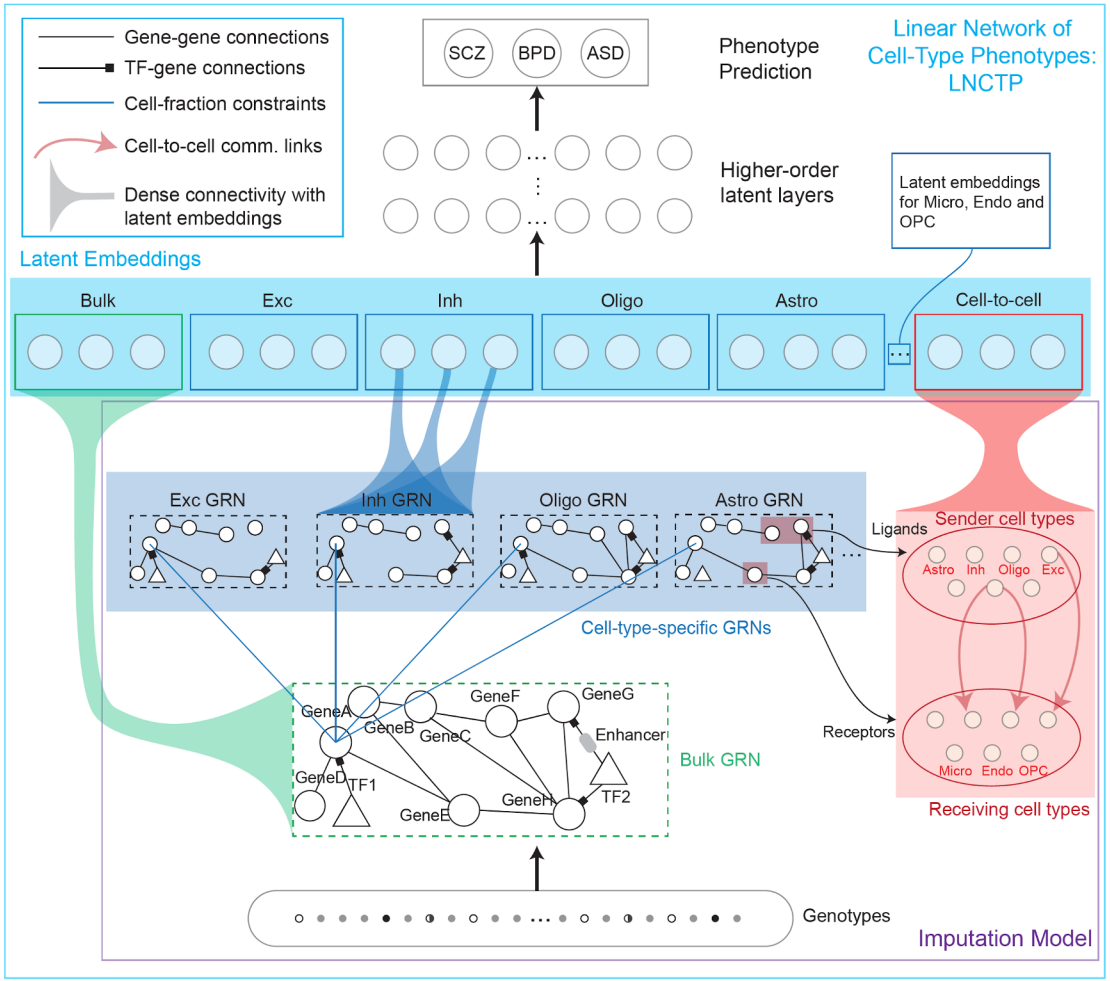
Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

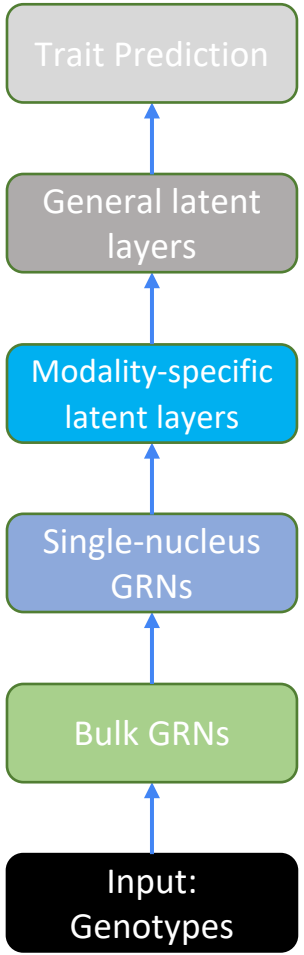
- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

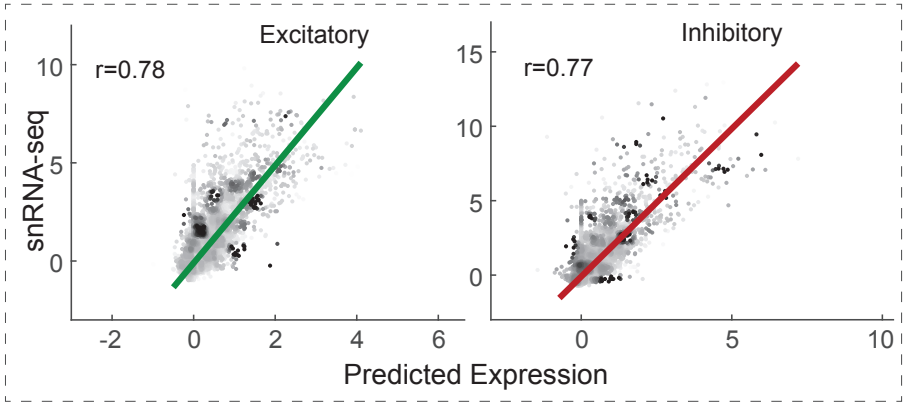
Linear Network of Cell-Type Phenotypes (LNCTP) model framework





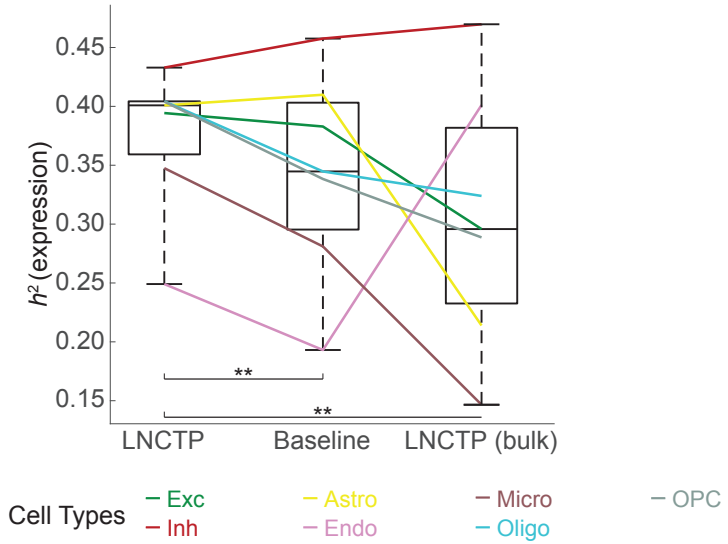
[Emani et al. ('24) Science]

LNCTP shows improved performance for imputing expression



-Prediction of single-cell expression in samples based only on genotypes

-Improves prediction of cell-type expression variance compared with other methods (i.e. baseline or bulk RNA models, PRS)



[Emani et al. ('24) Science]

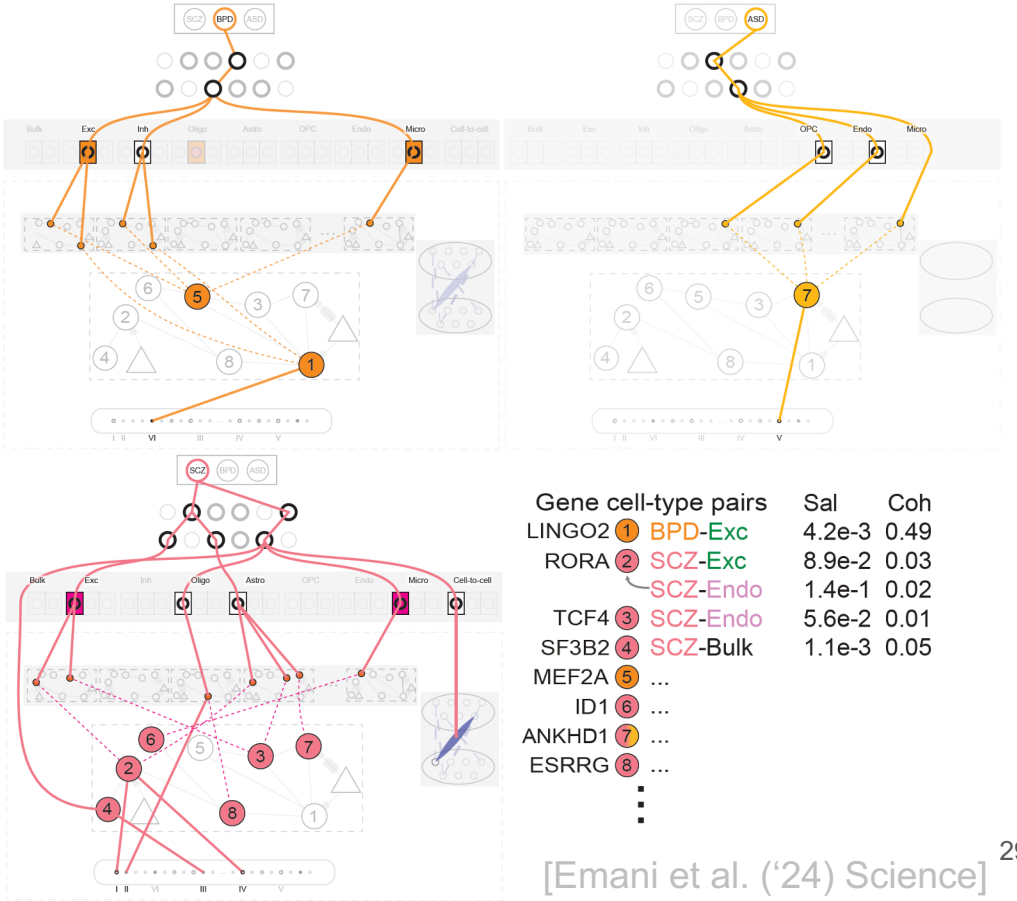
Using LNCTP to link genes, cell types, and phenotypes

Results from LNCTP allow the association of traits with genes in a cell-type-specific manner

By tracing the influence of genes through visible and latent layers, **cell-type-specific effects** towards disease can be identified.

~250 total gene + cell type pairs

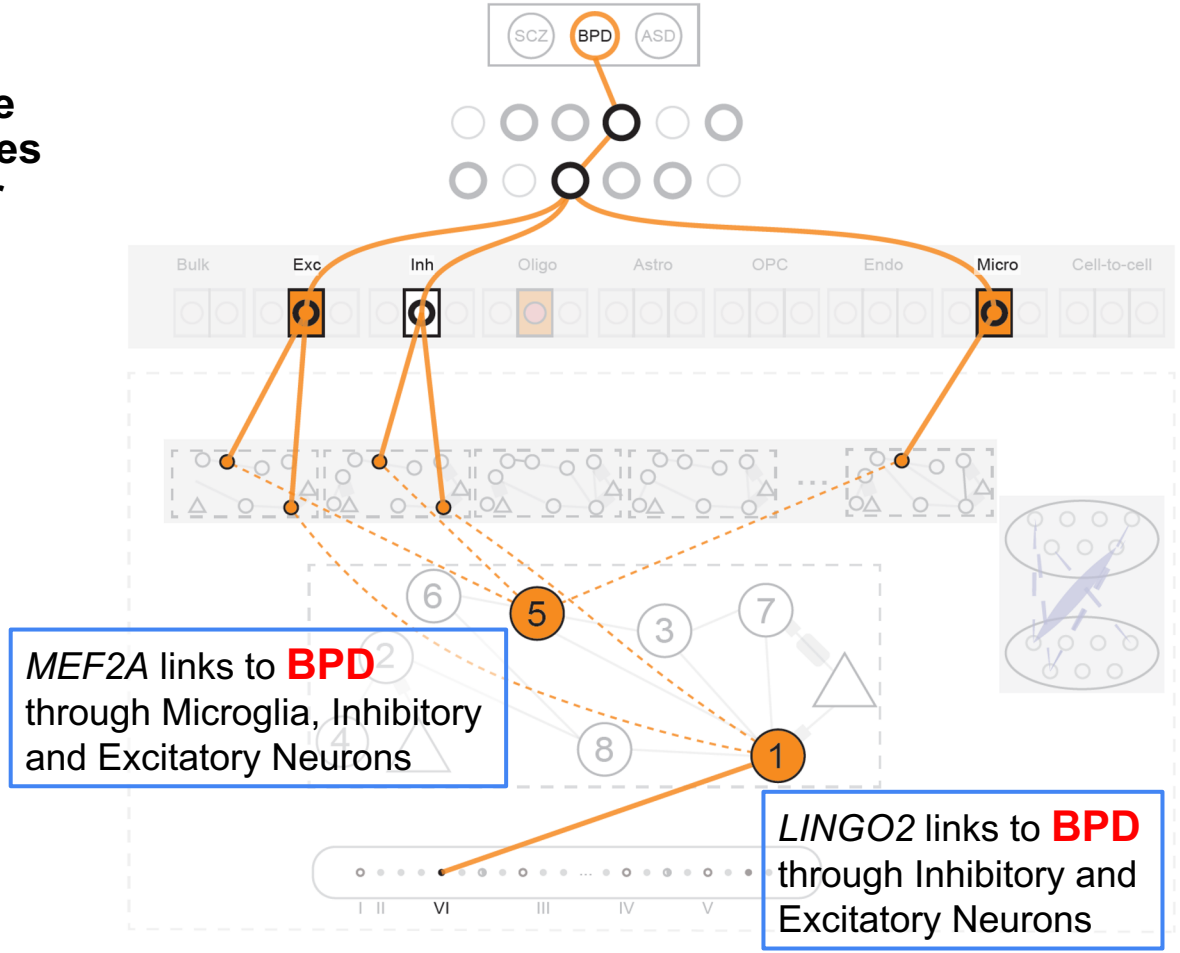
Salient pathways from genes through cell types to traits



[Emani et al. ('24) Science]

LNCTP examples: Prioritized cell types for BPD genes

Results from LNCTP allow the association of traits with genes in a cell-type-specific manner



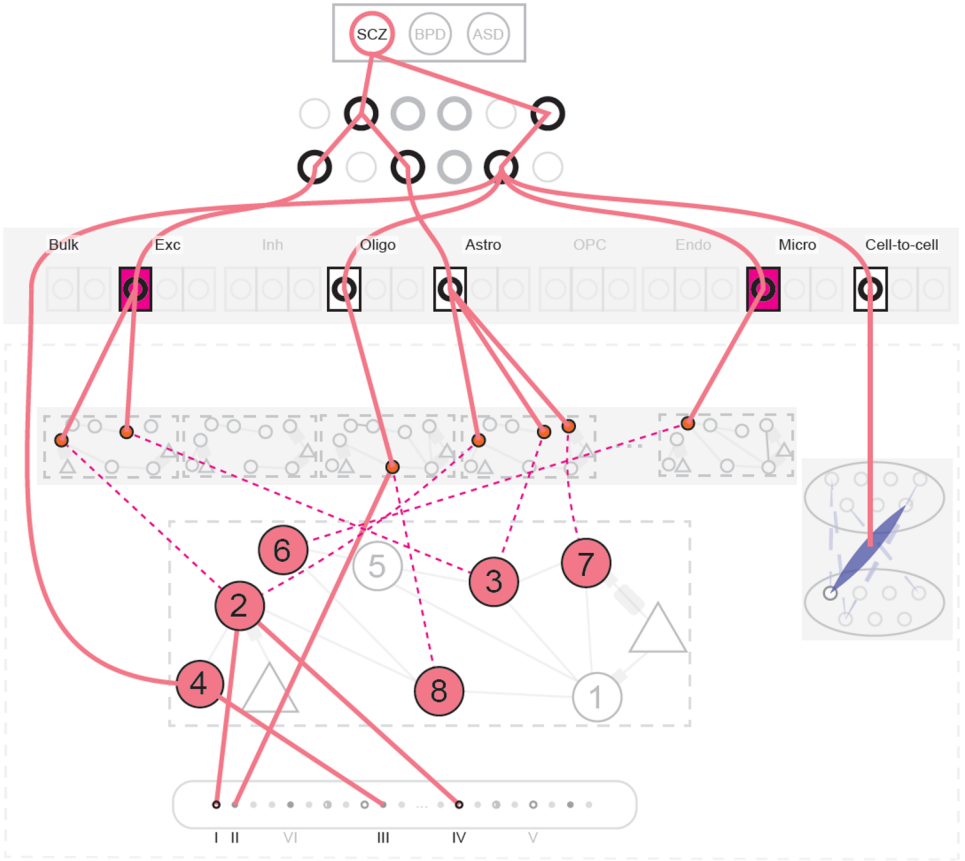
[Emani et al. (‘24) Science]

LNCTP examples: Prioritized cell types for SCZ genes

Results from LNCTP allow the association of traits with genes in a cell-type-specific manner

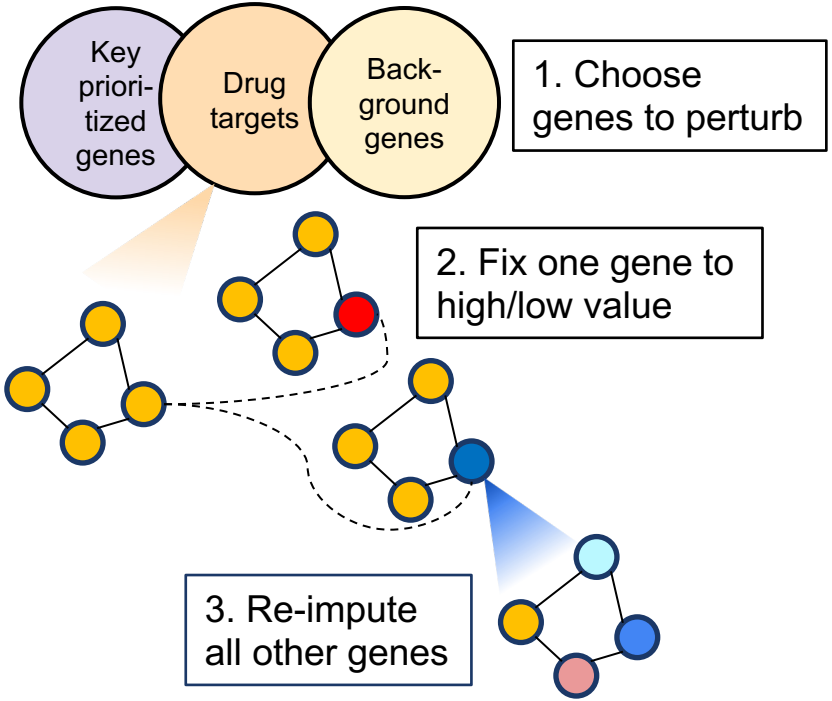
Highlights in **SCZ** include *TCF4*, *RORA,* & Micro-Exc linkage in cell-to-cell network

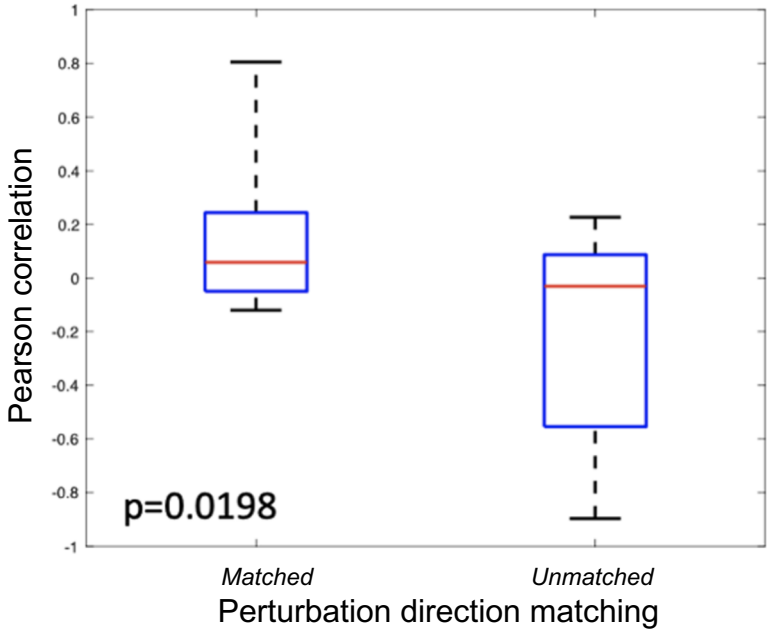
RORA	2	SCZ-Exc	8.9e-2	0.03
		SCZ-Endo	1.4e-1	0.02
TCF4	3	SCZ-Endo	5.6e-2	0.01
SF3B2	4	SCZ-Bulk	1.1e-3	0.05
MEF2A	5	...		
ID1	6	...		
ANKHD1	7	...		
ESRRG	8	...		



[Emani et al. ('24) Science]

Comparison of LNCTP predicted network to Glu neuron CRISPR experiments*





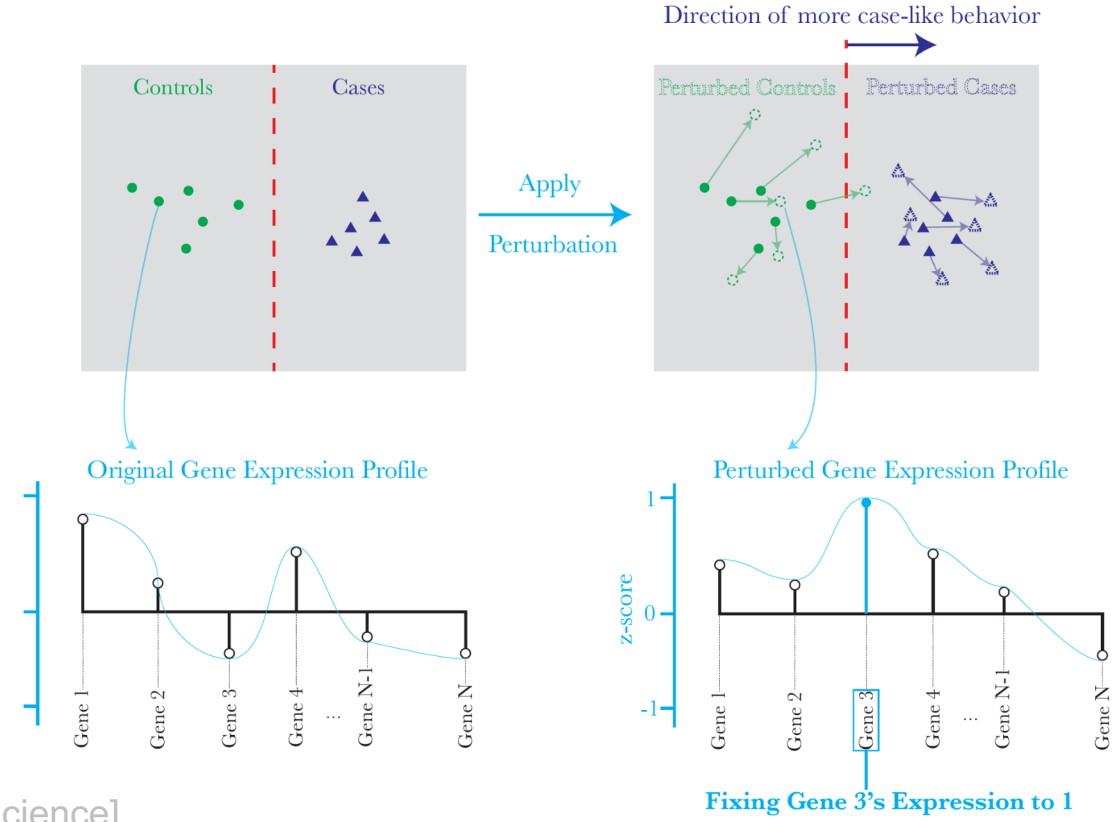
Comparing LNCTP and CRISPR perturbations

- Perturbations in excitatory neuron GRN
- Upper decile of genes according to LNCTP z-score changes
- Perturbation directions are **matched** or **unmatched**
  - **Unmatched** means LNCTP z-score changes correlate with CRISPR fold-change vectors for all genes **except** the perturbed gene

*Tian, R., Abarientos, A., Hong, J. *et al. Nat Neurosci* 24, 1020–1034 (2021).

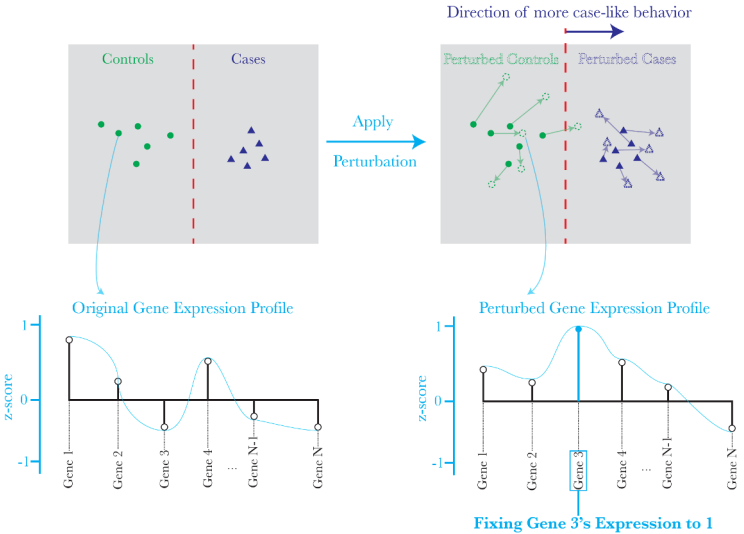
[Emani et al. (‘24) Science]

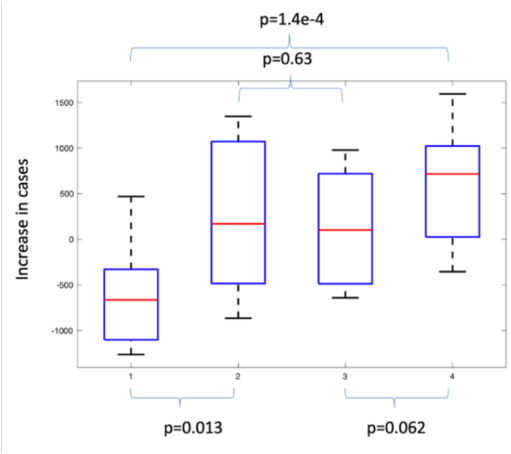
LNCTP model: Perturbation Analysis



[Emani et al. ('24) Science]

LNCTP model: Perturbation Analysis





Perturbations:

	Genes	Direction
1	Drugs/Key	reversed
2	Background	reversed
3	Background	forward
4	Drugs/Key	forward

LNCTP predicts effects of perturbations on case/control status

$$p_{LNCTP-perturbed}\left(y_i, h_i, x_{i,\neg(c^*,g^*)} | z_i, x_{ic^*g^*} = \{1,-1\}\right) \propto p_{GMRF}\left(x_{i,\neg(c^*,g^*)} | z_i, x_{ic^*g^*} = \{1,-1\}\right) \cdot p_{DNN}(y_i, h_i | x_i)$$

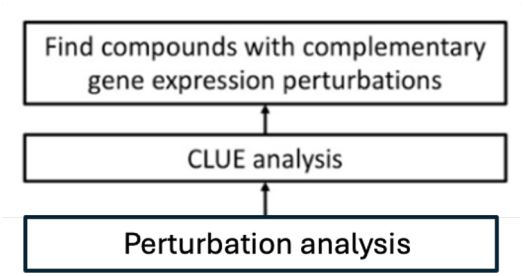
$$p_{GMRF}\left(x_{i,\neg(c^*,g^*)} | z_i, x_{ic^*g^*} = \{1,-1\}\right) \propto \exp\left(-E_{GMRF}\left(x_{i,\neg(c^*,g^*)} | z_i, x_{ic^*g^*} = \{1,-1\}\right)\right)$$

$$E_{GMRF}\left(x_{i,\neg(c^*,g^*)} | z_i, x_{ic^*g^*} = \{1,-1\}\right) = x_{i0}^T J x_{i0\cdot} + \sum_g x_{i0g}^T b\left(z, \beta_g\right) + \sum_c \left(x_{ic}^T J_c x_{ic\cdot} + x_{ic\cdot}^T b_c + x_{ic}^T J_c^{(c2c)} x_{i,LR,c}\right) +$$

$$\sum_{c_1,c_2} J_{c_1,c_2}^{(c2c)} x_{i,L,c_1}^{(c2c)} x_{i,R,c_2}^{(c2c)} + \lambda \sum_g \left(x_{i0g} - f(z)^T x_{i,1\ldots C,g}\right)^2 + \infty \cdot \delta(x_{ic^*g^*} \neq \{1,-1\})$$

[Emani et al. ('24) Science]

LNCTP model: Clue.io Analysis



Number of significant compounds in CLUE database

Gene	Bulk	Astro	Endo	Exc	Inh	Micro	Oligo	OPC
ANKHD1	797	234	15	405	375	13	1808	253
ESRRG	184	52	105	403	89	413	325	206
ID1	400	402	756	389	119	9073	594	1005
LINGO2	431	791	346	25	16	1688	492	1454
MEF2A	335	283	118	8	11	2670	661	349
RORA	242	196	301	156	79	1289	673	4605
SF3B2	176	100	1101	421	1014	4688	774	722
TCF4	489	185	7	253	96	1063	347	385

Well-known drugs used for neuropsychiatric disorders:

dopamine receptor antagonists,
dopamine receptor agonists,
glutamate receptor antagonists,
calcium channel blockers,
GABA receptor agonists,
MAP kinase inhibitors

Compounds with unknown effects:

Cytokine IL-1a: potential in reversing the expression changes of the ID1 gene in microglia
AKT inhibitor 10-DEBC: potential reversing the effects of TCF4, ID1, RORA, SF3B2
Consistent occurrence of bromodomain inhibitors for reversing effects of all eight genes

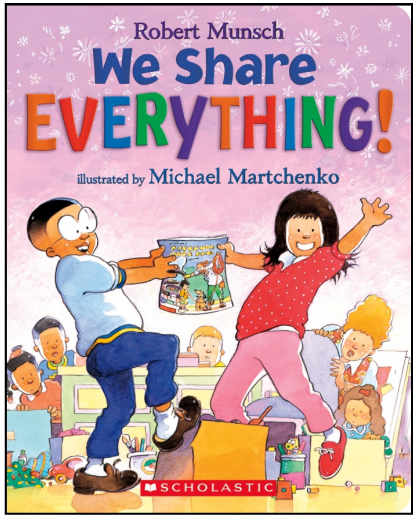
[Emani et al. ('24) Science]

Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

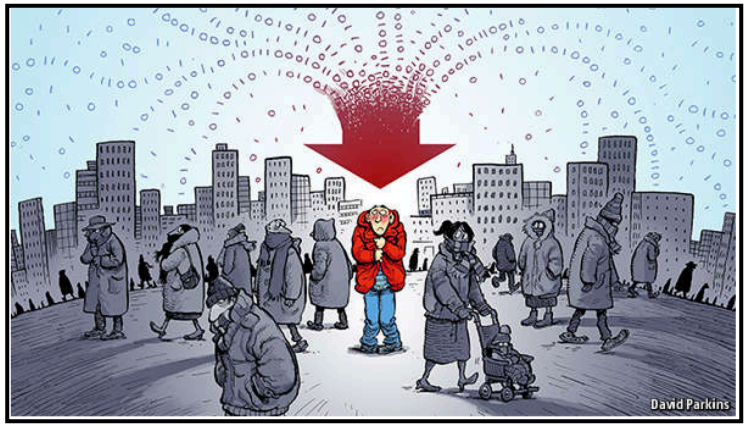


The Dilemma

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

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

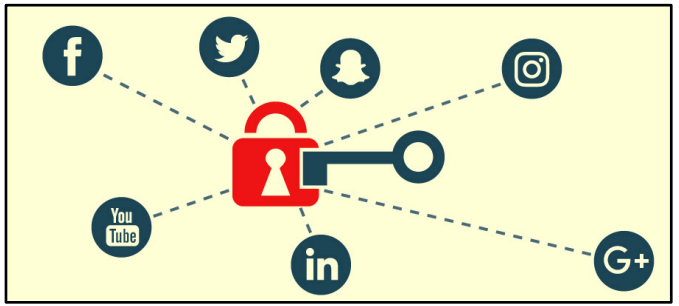
- Sharing helps **speed research**
  - Large-scale mining of this information is important for medical research
  - Statistical power
  - Privacy is cumbersome, particularly for big data



[Economist, 15 Aug '15]

[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)]

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]

Risk from environmental samples: sparse and noisy genotypes

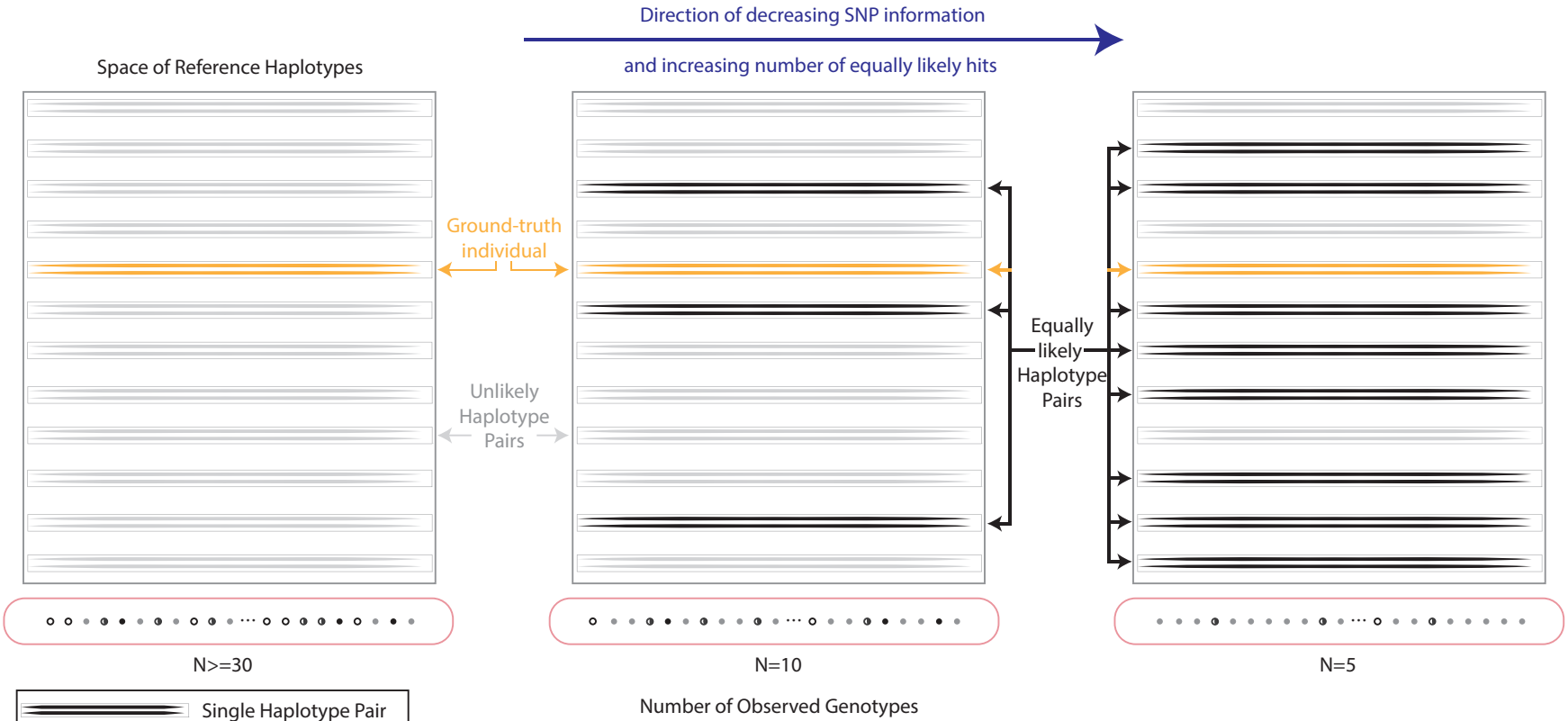
- Swabs of objects are easy to obtain
- How much information is contained in such DNA samples?
- Created a population-genetics-inspired approach* to quantify the risk:

PLIGHT = **P**rivacy **L**eakage through **I**nference across **G**enotypic **H**MM **T**rajectories

- Found risk of identification is high even with tens of noisy SNPs
- Identified a way of sanitizing data before publication

*Emani, P.S.; Geradi, M.N.; Gürsoy, G.; Grasty, M.R.; Miranker, A.; Gerstein, M.B. ``Assessing and mitigating privacy risks of sparse, noisy genotypes by local alignment to haplotype databases", *Genome Research* (2023), Vol. 33, Iss. 12, Pgs. 2156-2173

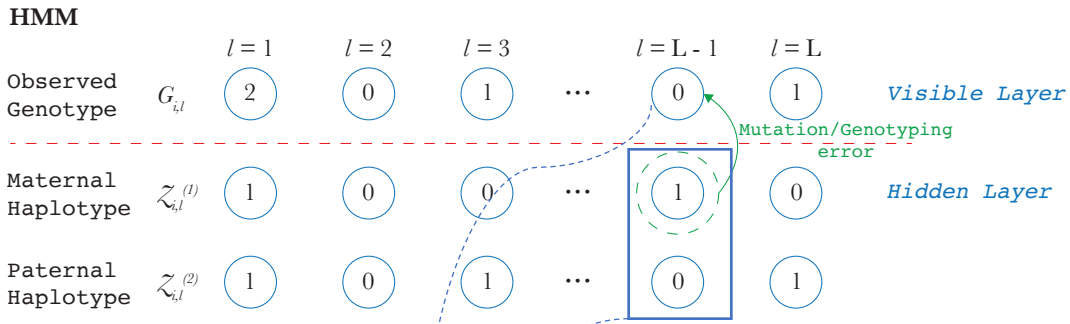
Quantifying identifying information in the limit of small SNP sets



The Li & Stephens HMM: Effect of mutation/genotyping error

(Li, N.; Stephens, M. *Genetics* **2003**, *165*, 2213–2233.)

When identifying individuals, can include the effects of genotyping error/noise using a population genetics approach:



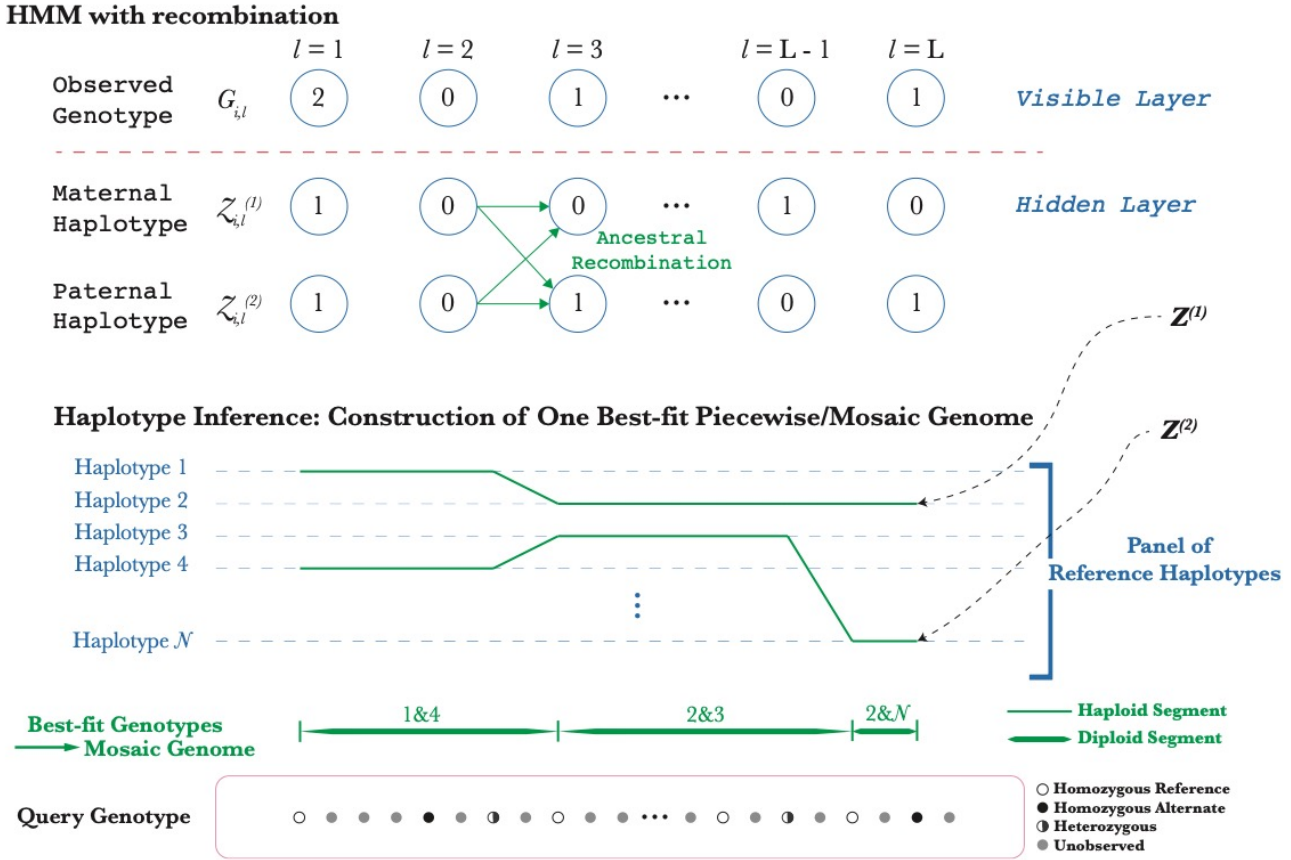
Error rate $\lambda = 0.05$

$Z^{(1)} + Z^{(2)}$	G = 0	1	2
0	0.9025	0.095	0.0025
1	0.0475	0.905	0.0475
2	0.0025	0.095	0.9025

Probability of observing a genotype G of 0, conditional on the sum of the haplotypes being 1

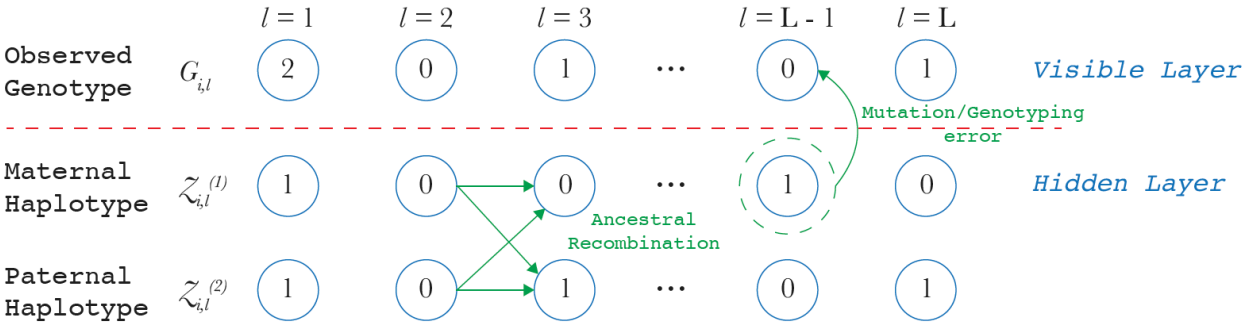
The Li & Stephens HMM: Effect of ancestral recombination

(Li, N.; Stephens, M. *Genetics* **2003**, *165*, 2213–2233.)



The Li & Stephens HMM

(Li, N.; Stephens, M. *Genetics* **2003**, *165*, 2213–2233.)



Addresses: What is the probability of observing a set of genotypes, based on underlying panel of haplotypes?

$$P(G_i|H) = \sum_{Z_j^{(1)},Z_k^{(2)}} P\left(G_i\middle|Z_j^{(1)},Z_k^{(2)}\right).P\left(Z_j^{(1)},Z_k^{(2)}\middle|H\right)$$

Encodes recombination

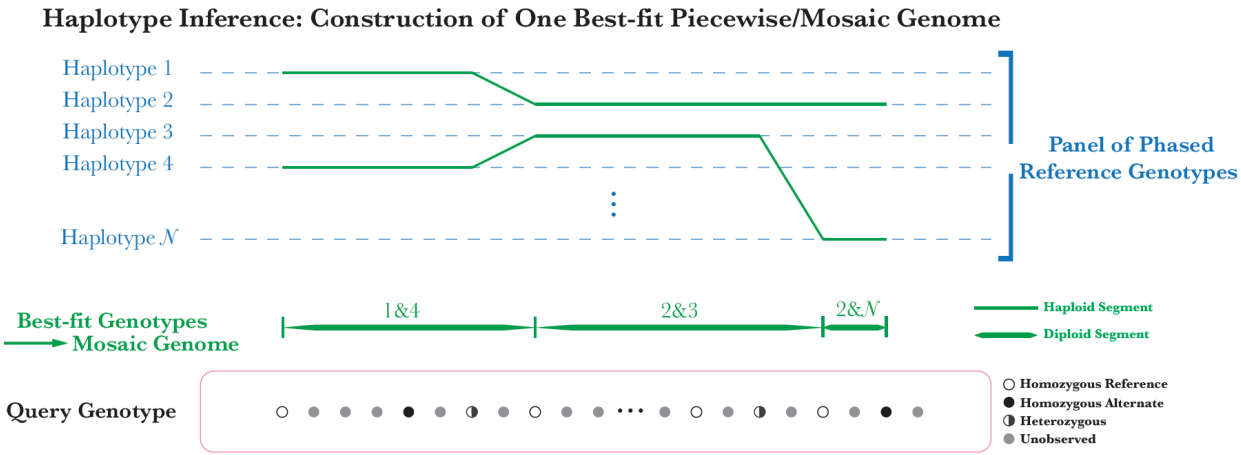
Encodes mutations or genotyping error

H = Set of all reference phased genotypes

$Z_j^{(\alpha)} = \left\{Z_{j(l),l}^{(\alpha)}\right\}_{l=1}^{L}$ = Set of all possible haplotypes at the observed loci l

PLIGHT

Viterbi algorithm for most-likely path



1. Search through the space of haplotype pairs (total dimension = $N \times N$) to match the diploid genotypes
2. Each of the two haplotypes can independently recombine with other haplotypes
3. Allow for mutations/genotyping error

4. **Result:** Piecewise matches of reference haplotypes to observed genotype

PLIGHT

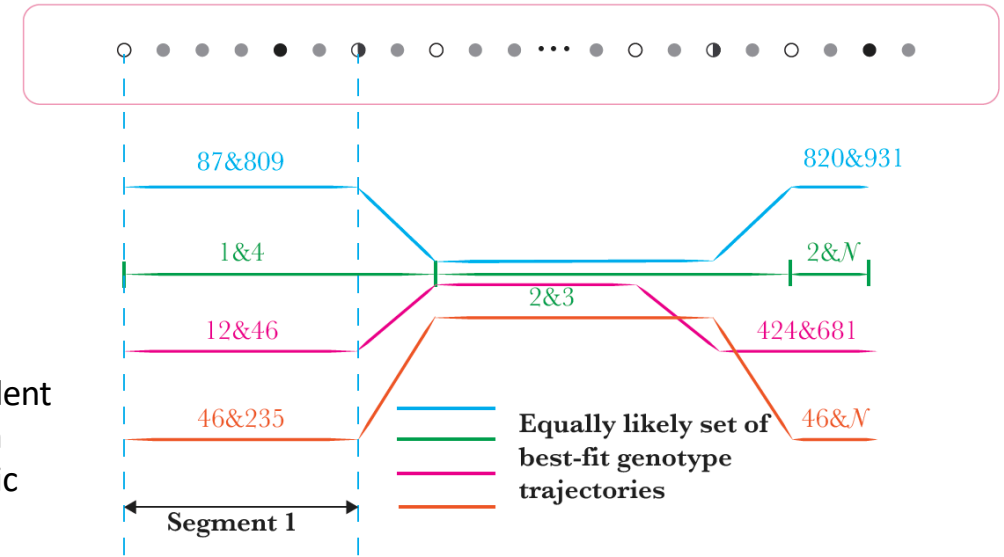
Viterbi algorithm for all equally likely "best-fit" paths

$$\underset{Z_j^{(1)},Z_k^{(2)} \in \text{Set of all possible haplotypes}}{\operatorname{argmax}} P\left(G_i \middle| Z_j^{(1)}, Z_k^{(2)}\right) . P\left(Z_j^{(1)}, Z_k^{(2)} \middle| H\right)$$

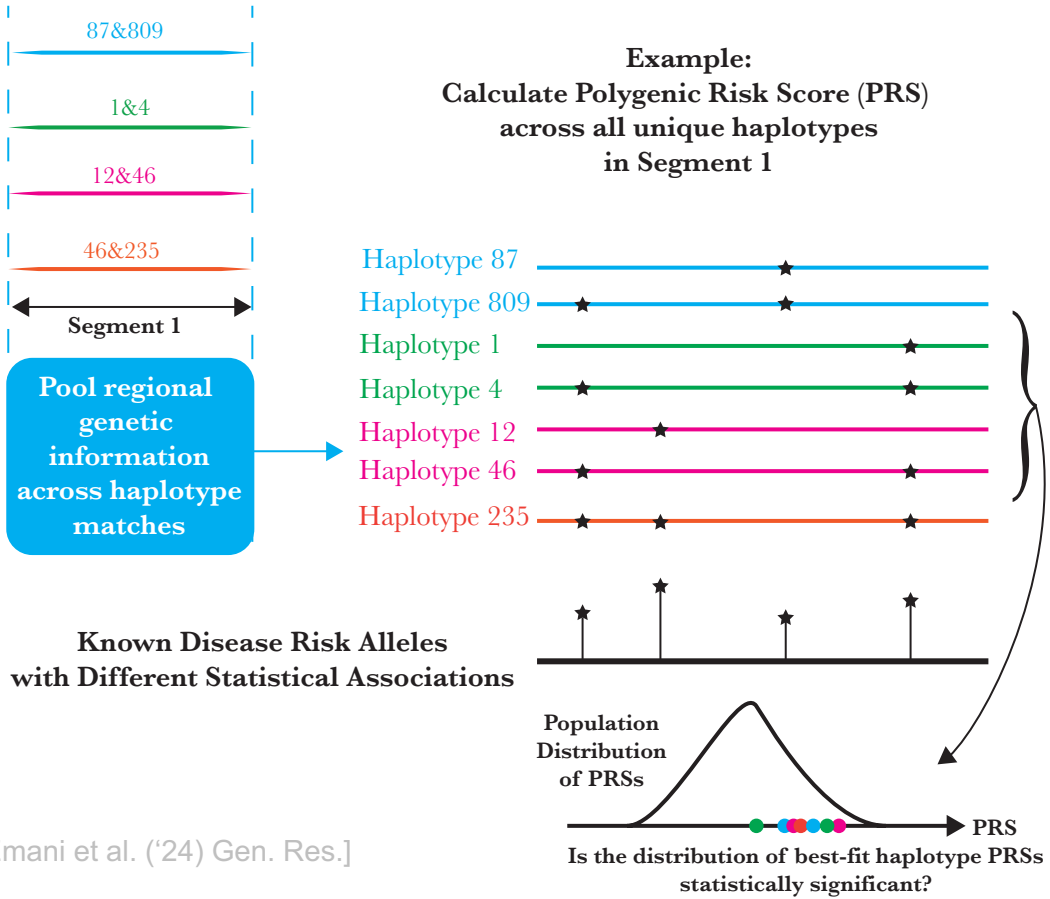
For very sparse data, several paths may be equally likely.

We term these paths 'Genotypic Trajectories': to signify the sequential exploration of haplotype space.

The number of independent trajectories in any region gives a sense of genotypic 'entropy'

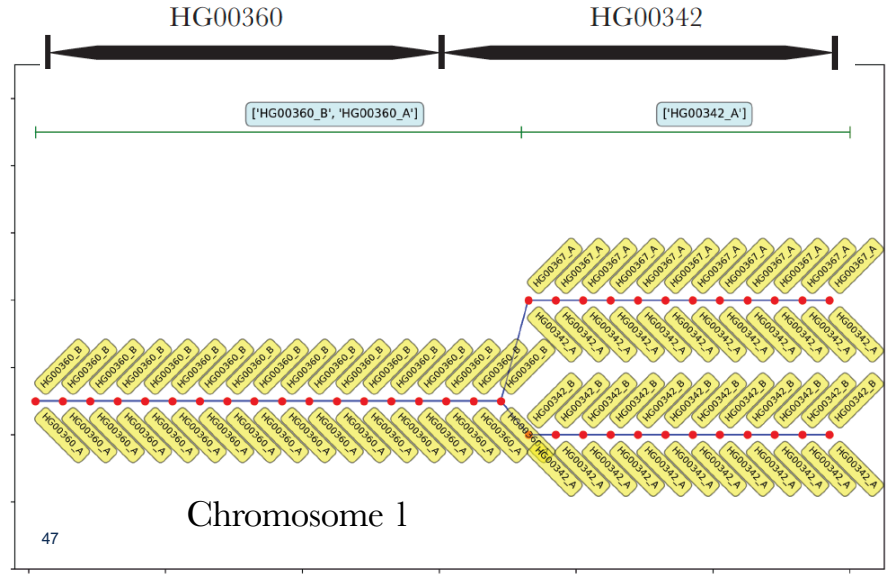


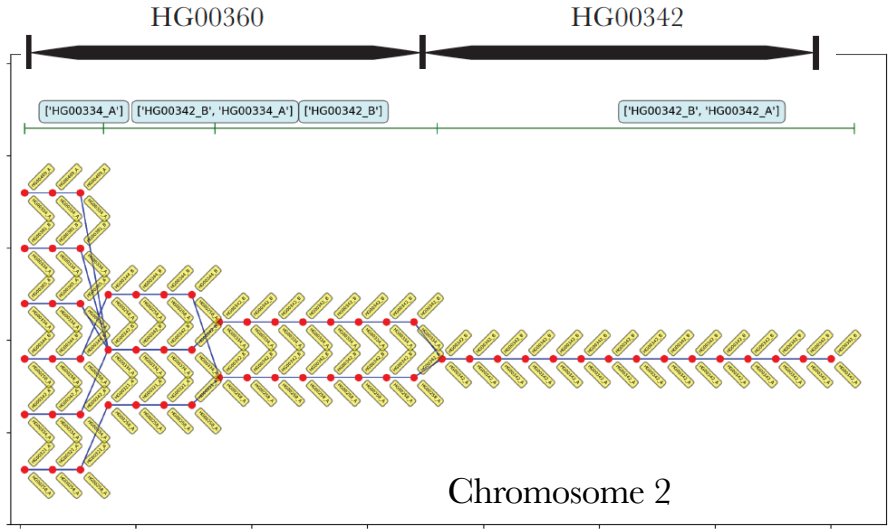
Privacy risk of partial/regional genotype matching



A few noisy SNPs can pose privacy risks

- Identified individuals with as little as 10 SNPs; robust to modest noise
- Found 1st-degree relatives (parents, siblings, children) with ~20-30 SNPs
- Environmental (saliva) swab SNPs still allow identification (with ~30 SNPs), in spite of noise
- Recommended sanitizing SNPs that specify identity with high degree of certainty





Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

Overall Problem: Finding Key Variants in Personal Genomes

CAN YOU FIND THE PANDA?

Millions of variants in a personal genome
Thousands, in a cancer genome
Different **contexts** for prioritization

In **rare disease**, only a few
high-impact variants are associated with disease

In **cancer**, a few positively selected drivers amongst many passengers

In **common disease**, more variants associated & each has weaker effect,
But one wants to find key ”functional” variant amongst many in LD

Overall Problem: Finding Key Variants in Personal Genomes

CAN YOU FIND THE PANDA?

Millions of variants in a personal genome
Thousands, in a cancer genome
Different **contexts** for prioritization

In **rare disease**, only a few
high-impact variants are associated with disease

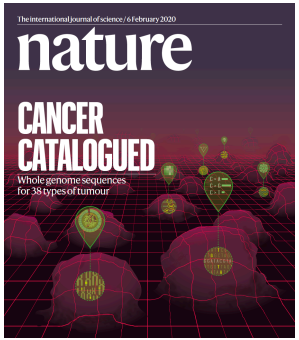
In **cancer**, a few positively selected drivers amongst many passengers

In **common disease**, more variants associated & each has weaker effect,
But one wants to find key ”functional” variant amongst many in LD

Thus: Need to find & prioritize high impact variants.
Particularly hard for non-coding regions.

Basic Science ➔ Medicine: Creation of Variant Catalogs of healthy & sick people

CLASSICS



INITIATIVES

AGTC

biobank uk

THE PRECISION MEDICINE INITIATIVE

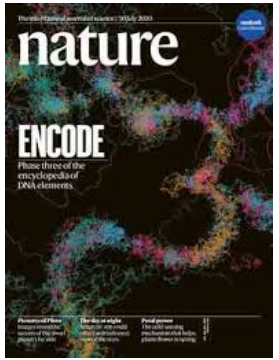
STARTUPS

23andMe

FOUNDATION MEDICINE

Resources for Variant Interpretation

More direct & clear interpretation through molecular endophenotypes (gene expression) rather than "macro phenotype" (disease diagnosis)



GTEx

iGVF

HUMAN CELL ATLAS

Medical Big Data: Promise and Challenges (Lee and Yoon , *Kidney Res. Clin. Pract.*, 2017)

Learning the DNA regulatory grammar

https://commons.wikimedia.org/wiki/File:Possible_Transcription_Factor_Binding_Sites_in_C12orf66_Promoter_Region.png

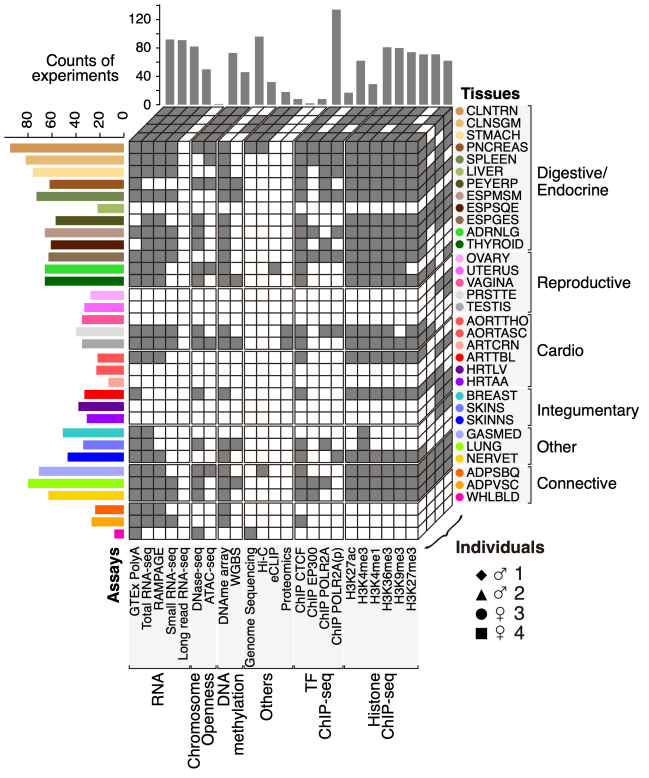
Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

The EN-TEx resource of multi-tissue functional genomics data



GTEx

Tissues

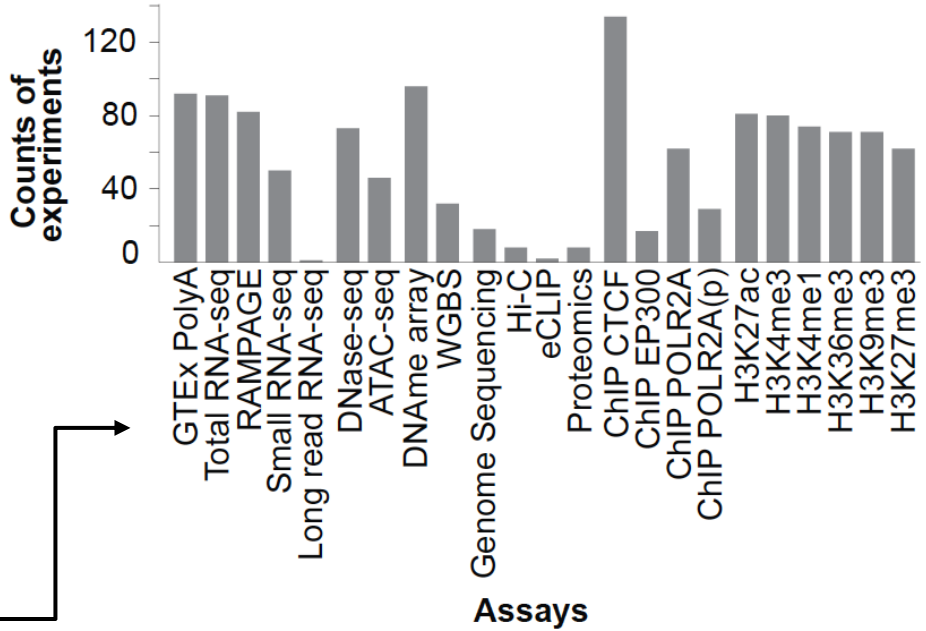
&

Individuals

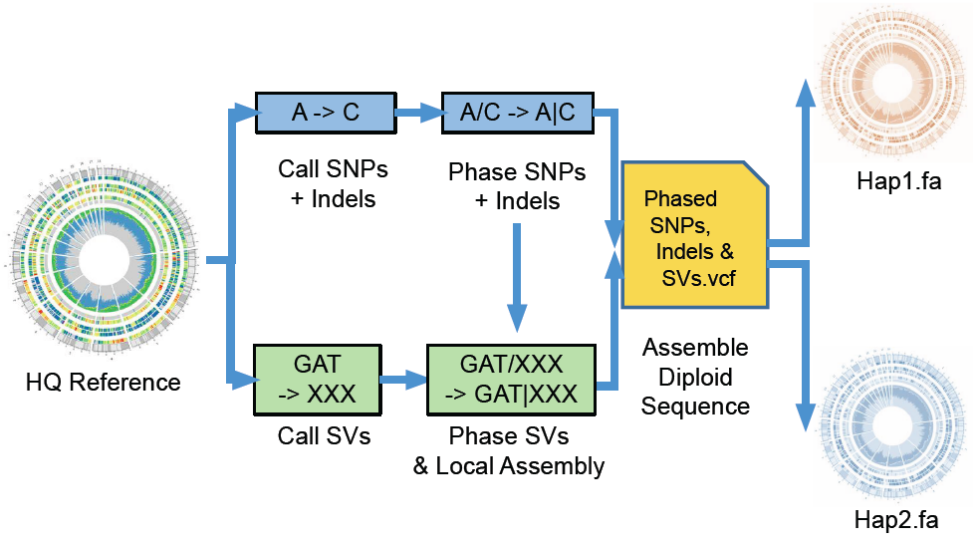
ENCODE

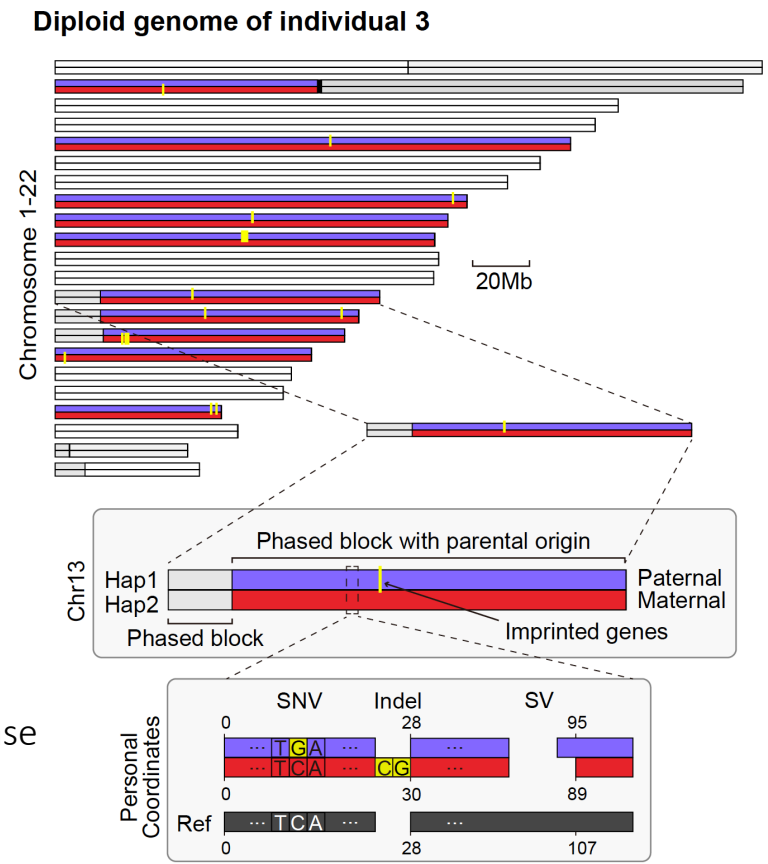
Assays

~1500 experiments



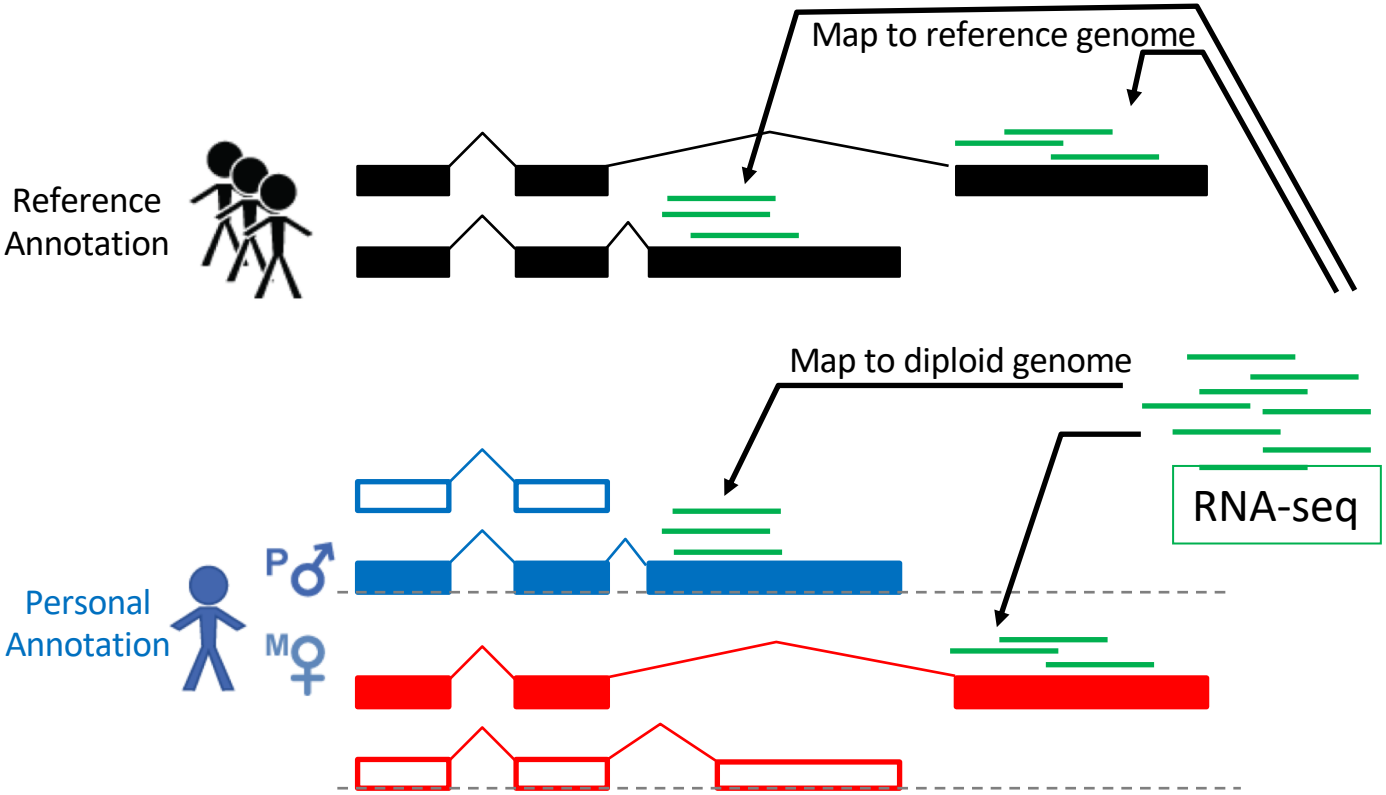
Construction of personal diploid genomes





- long-read (PacBio, 10x) and short-read (Illumina) DNA-seq
- >20K long structural variants (SVs) per person
- We integrate WGBS data to call imprinted regions and phase maternal and paternal haplotypes

How to map functional genomic data



[Rozowsk et al. Cell '23] – Lectures.gersteinlab.org

Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

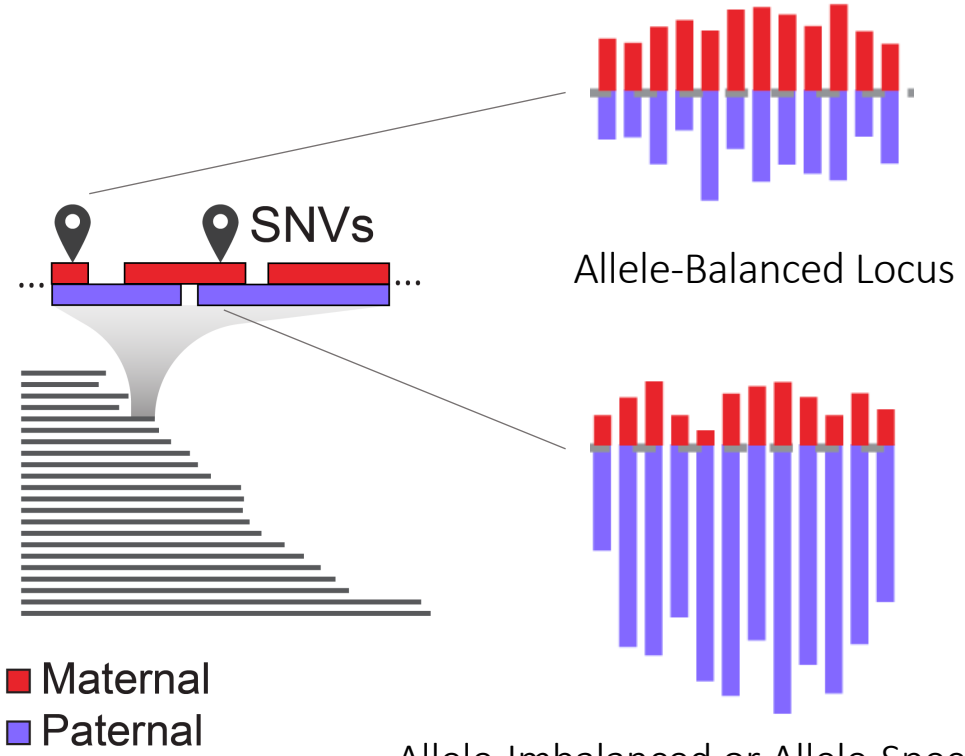
- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

Personal genomes and the detection of allele-specific events

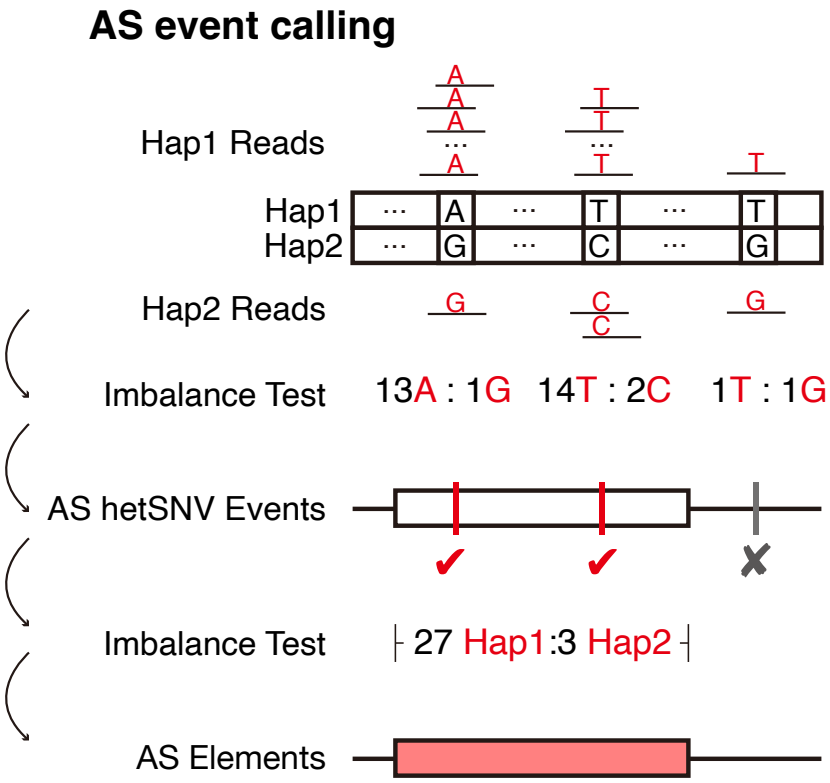
Personal genomes allow to study a broader set of SNVs, including individual-specific SNVs



Read Stack
(e.g., from RNA- or ChIP-seq)
has necessary info. to
determine correct haplotype
to map to
but tricky mapping issues must
be resolved

```
ACTTTGATAGCGTCAATG
 CTTTGATAGCGTCAATGC
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
    TGATAGCGTCAATGCACG
      ATAGCGTCAATGCACGTC
       TAGCGTCAATGCACGTCG
```

Personal genomes and the detection of allele-specific events



AlleleSeq pipeline

Identify accessible SNVs

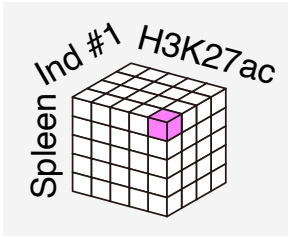
Perform beta binomial test

Identify heterozygous AS SNVs

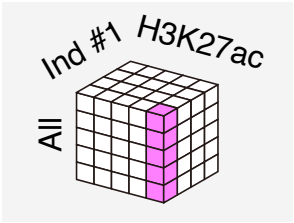
Merge reads within an element (genes, cCREs)
Perform beta binomial test

Identify AS elements (genes, cCREs)

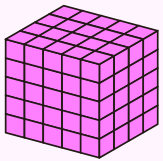
Aggregating AS Events into Catalog & their GWAS/eQTL Enrichment



EX:
H3K27ac, Ind. #1, Spleen
has ~2,600 AS SNVs

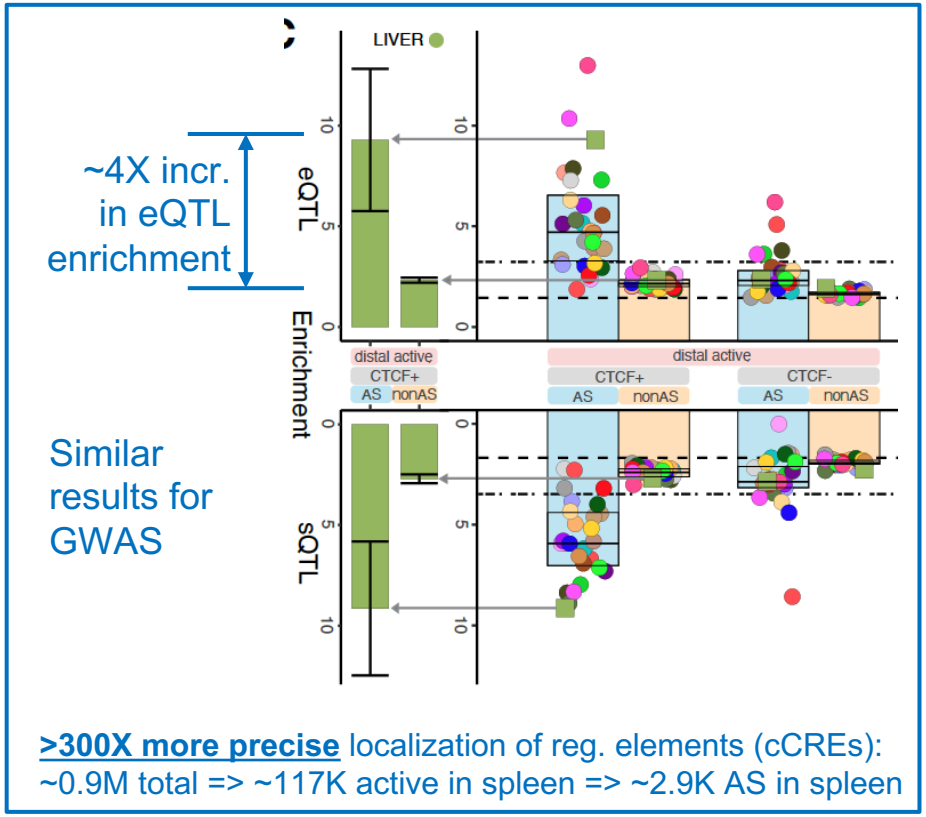


AS SNVs Across Tissues:
Union: 5,500 SNVs
Pooling: 27,000 SNVs
(Allows for **Joint Calling**)



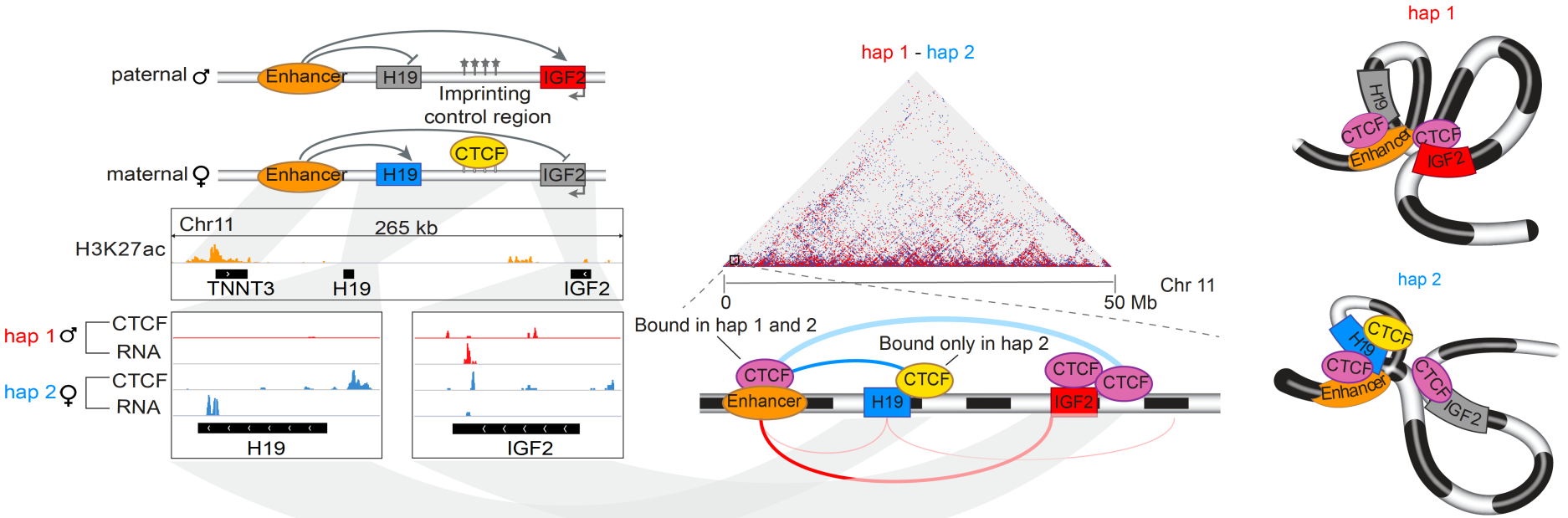
AS catalog
4 Individuals
12 Assays
31 Tissues

~1.3 M
SNVs
total



>300X more precise localization of reg. elements (cCREs):
~0.9M total => ~117K active in spleen => ~2.9K AS in spleen

The classic example of *H19* and *IGF2* allele-specific activity



Recapitulating a Classic Story

AS Hi-C analysis ==> Different M & P Chromatin

Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

Relating AS SNVs and TF binding motifs

Genome sequence

AS SNV non-AS SNV
(for a given assay)

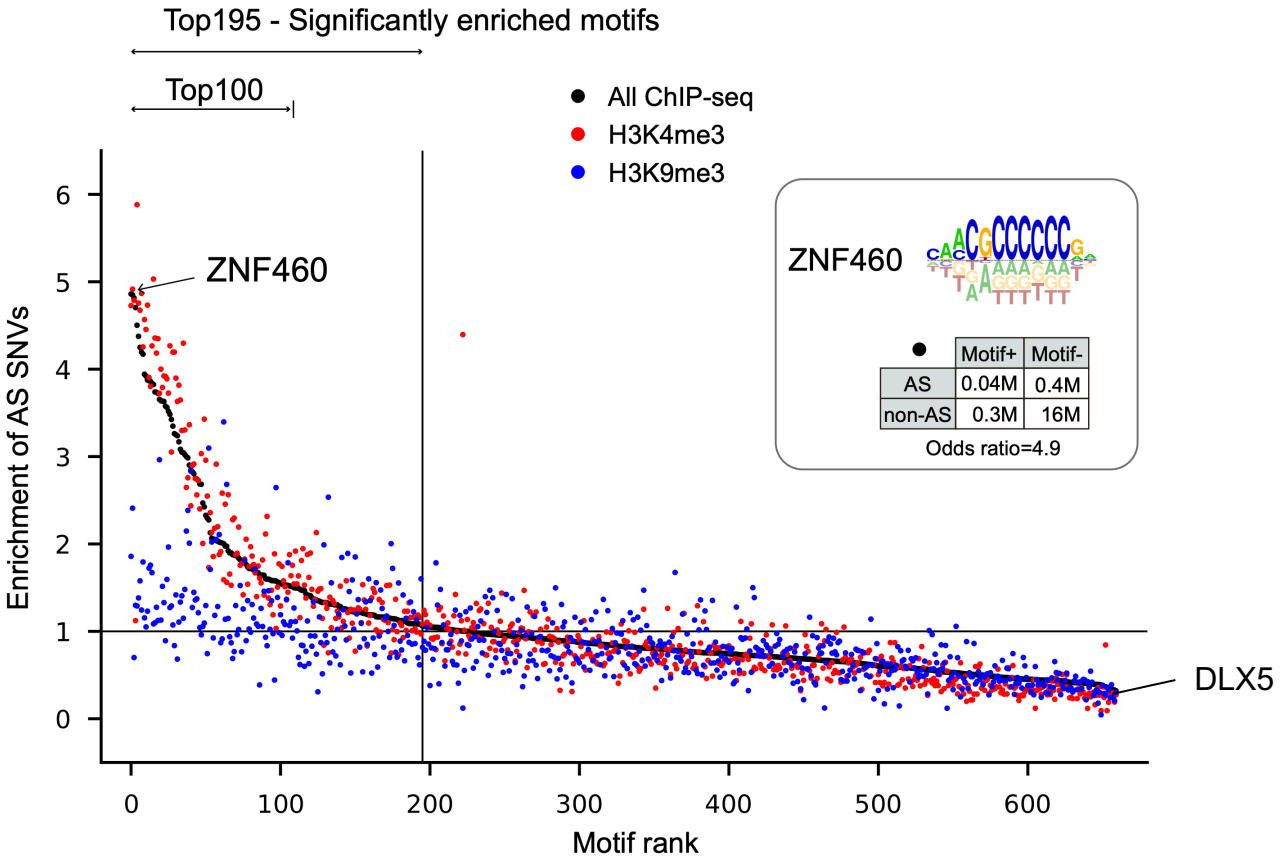
TF binding

There are many SNVs in the genome

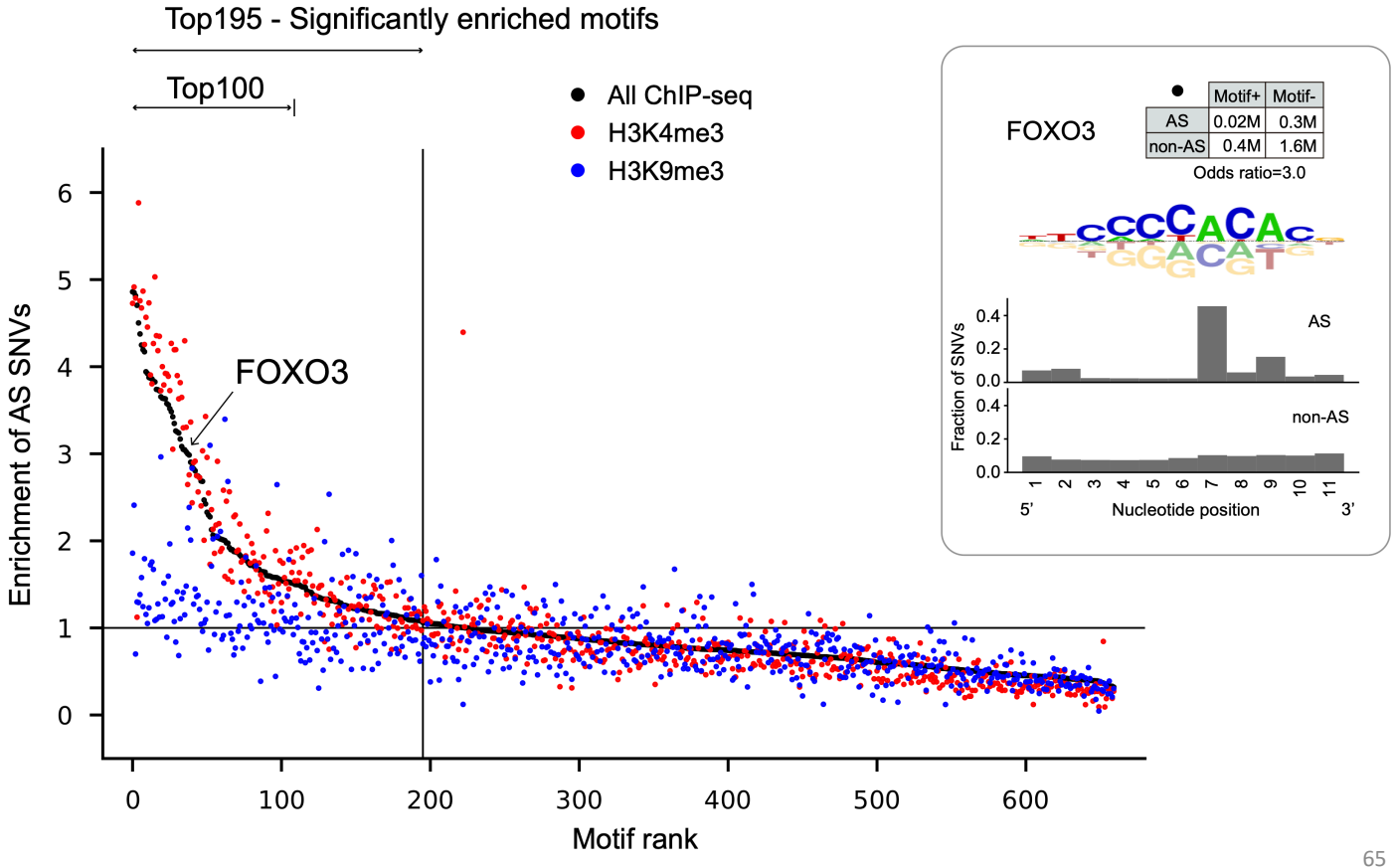
- Some can impact activity (AS SNVs)
- Some don't have any impact (non-AS SNVs)

We can identify TFs with motifs more sensitive to mutations, showing enrichment in AS events

Cross-referencing AS SNVs and TF binding motifs identified AS “sensitive” TFs



Cross-referencing AS SNVs and TF binding motifs identified AS “sensitive” TFs



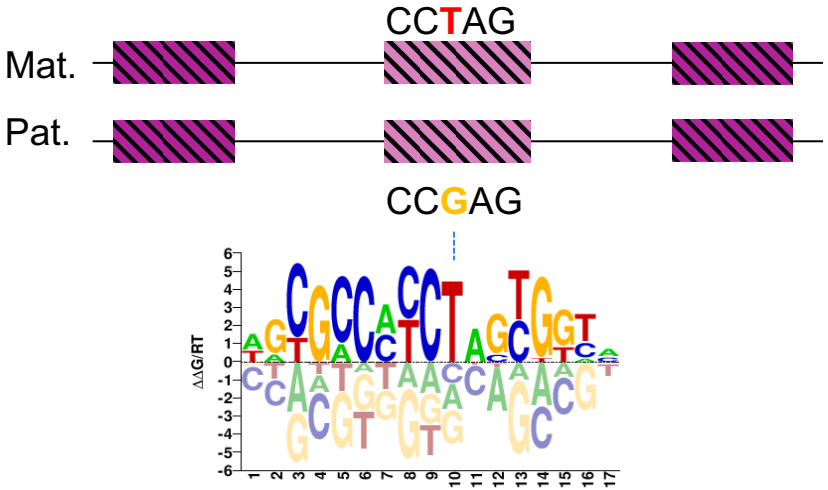
Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

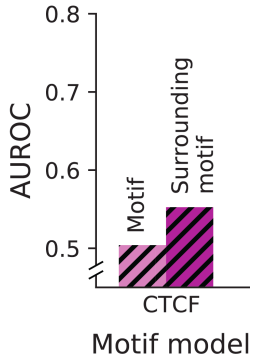
- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

Predicting AS activity just from nucleotide sequence (Simple)



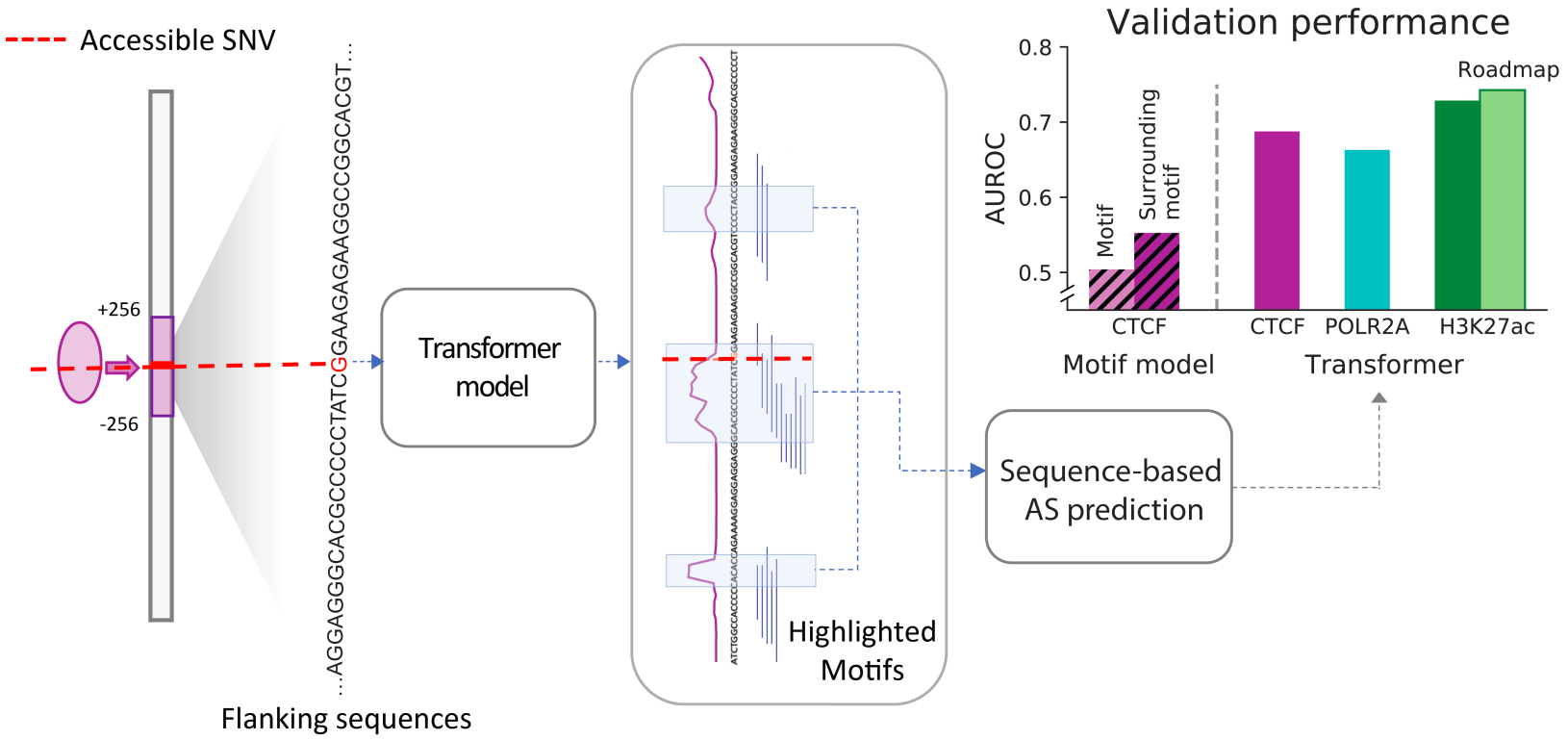


Simple logistic regression models based on overlapping & nearby motifs

- Each TF has a specific sequence that defines its motif.
  Can we predict AS events from those overlapping TF sequence motifs?
- EX: Predicting CTCF AS activity from its motif
- Intuition: Might think AS variant “knocking-out” a TF motif would give rise to differential AS binding

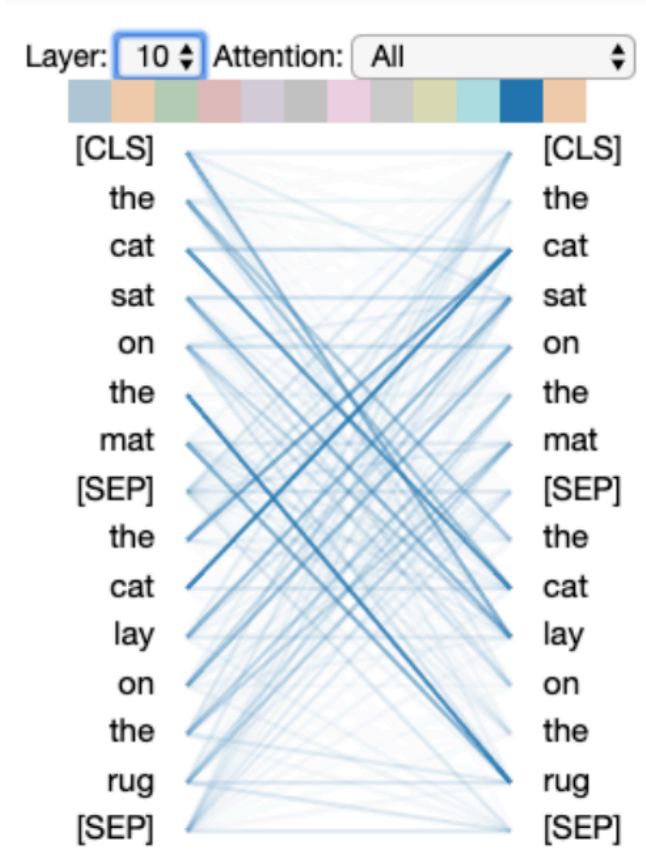
Predicting AS activity just from nucleotide sequence (Deep Learning)

Model taking into account a larger 250-bp window around the SNV yields better predictions



Contextual language model

- Attention
  - How “relevant” each part of the sentence is to the rest
  - $L \times L$ attention matrix for **every word** against **every word**
  - Calculated using the whole sentence

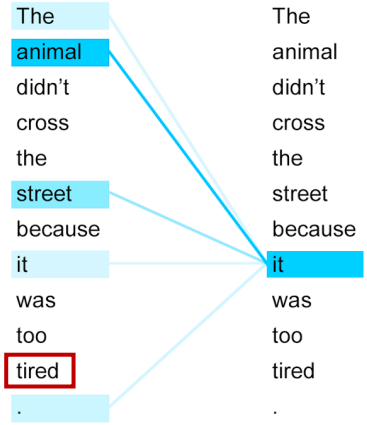


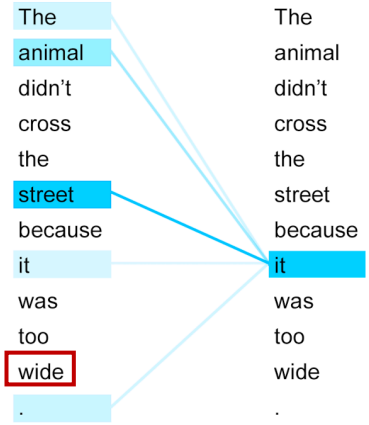
Vig, Jesse. "BertViz: A tool for visualizing multihead self-attention in the BERT model." *ICLR workshop: Debugging machine learning models*. Vol. 23. 2019.

Contextual language model

The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.

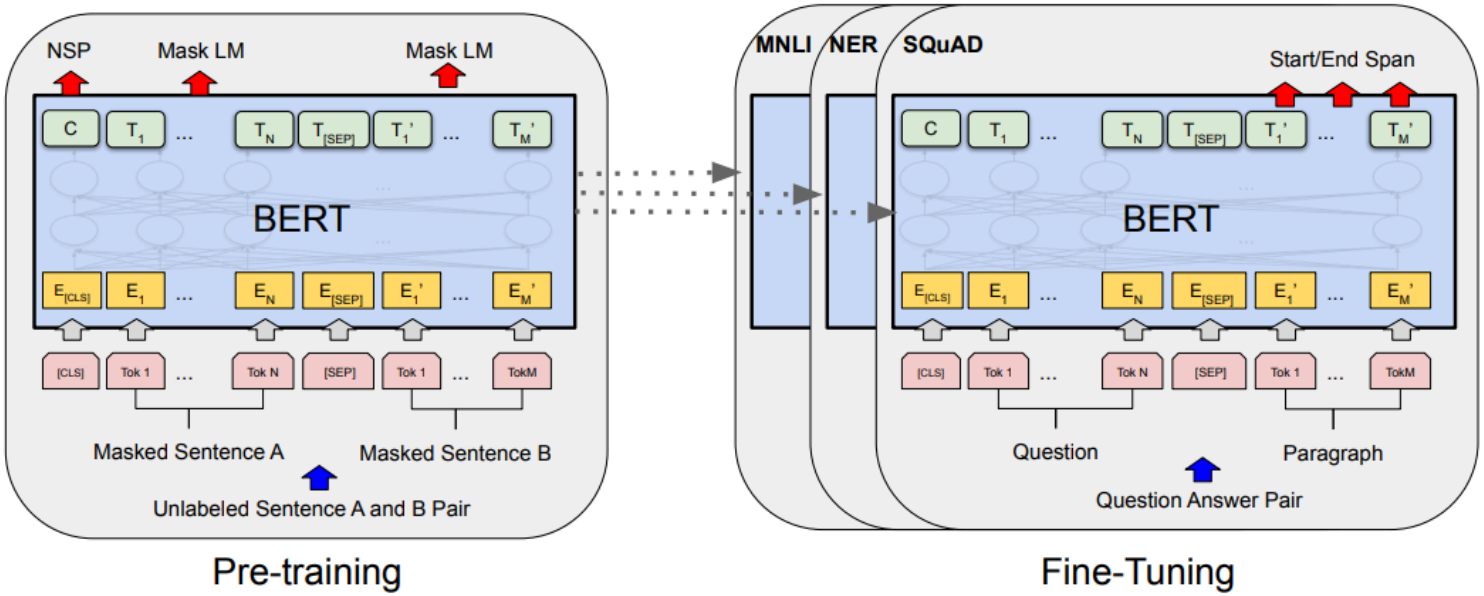




https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Transformer encoder & BERT

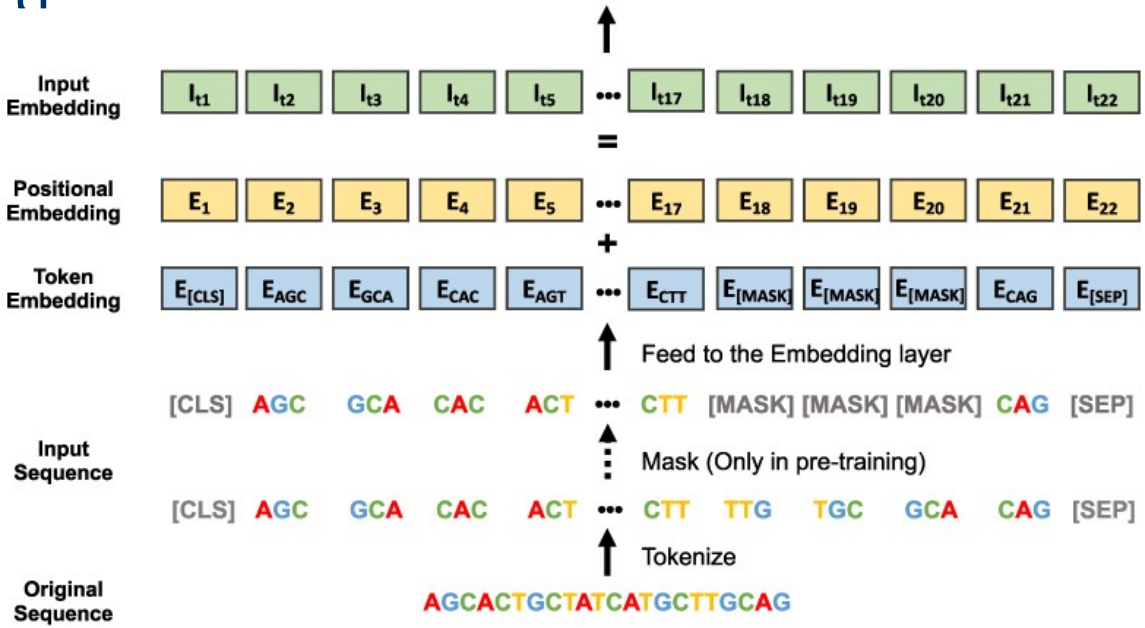
Attention layers



Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Transformer encoder & BERT

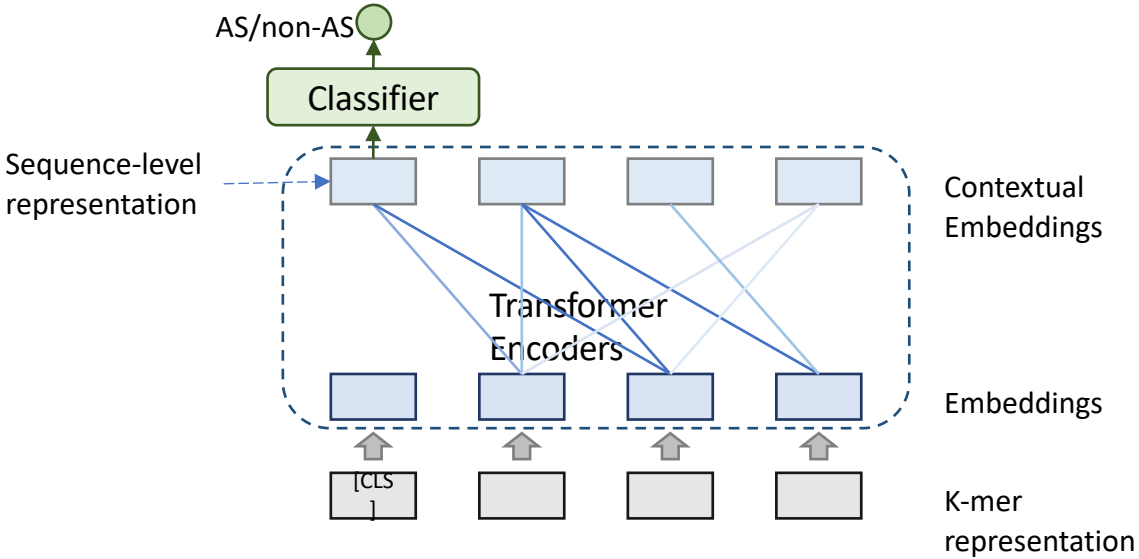
- DNABERT



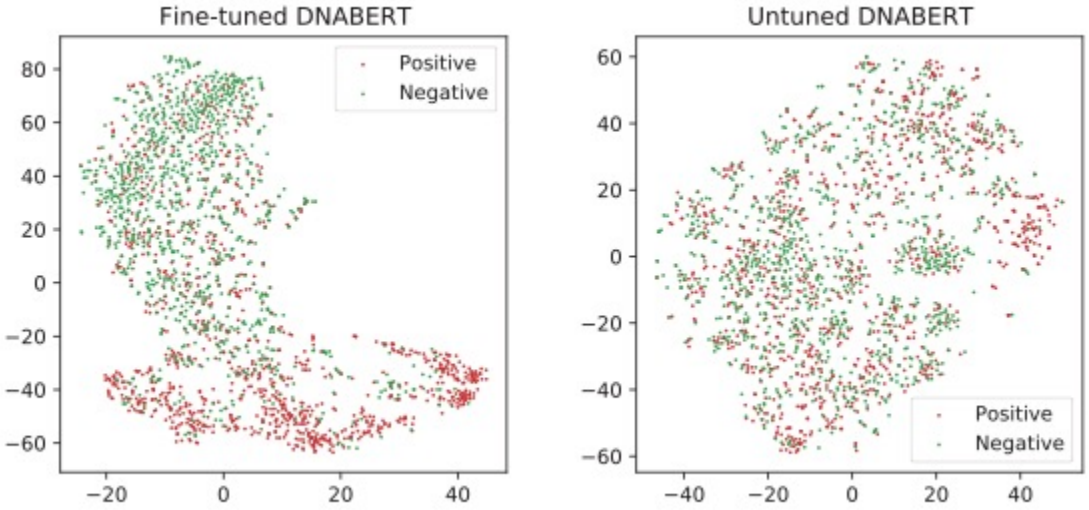
Ji, Yanrong, et al. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome." *Bioinformatics* 37.15 (2021): 2112-2120.

Allele specificity prediction

- Trained on the aggregated AS set of donor 3
  - ±128bp sequence context
  - Positive vs negative: AS vs non-AS heterozygous variants

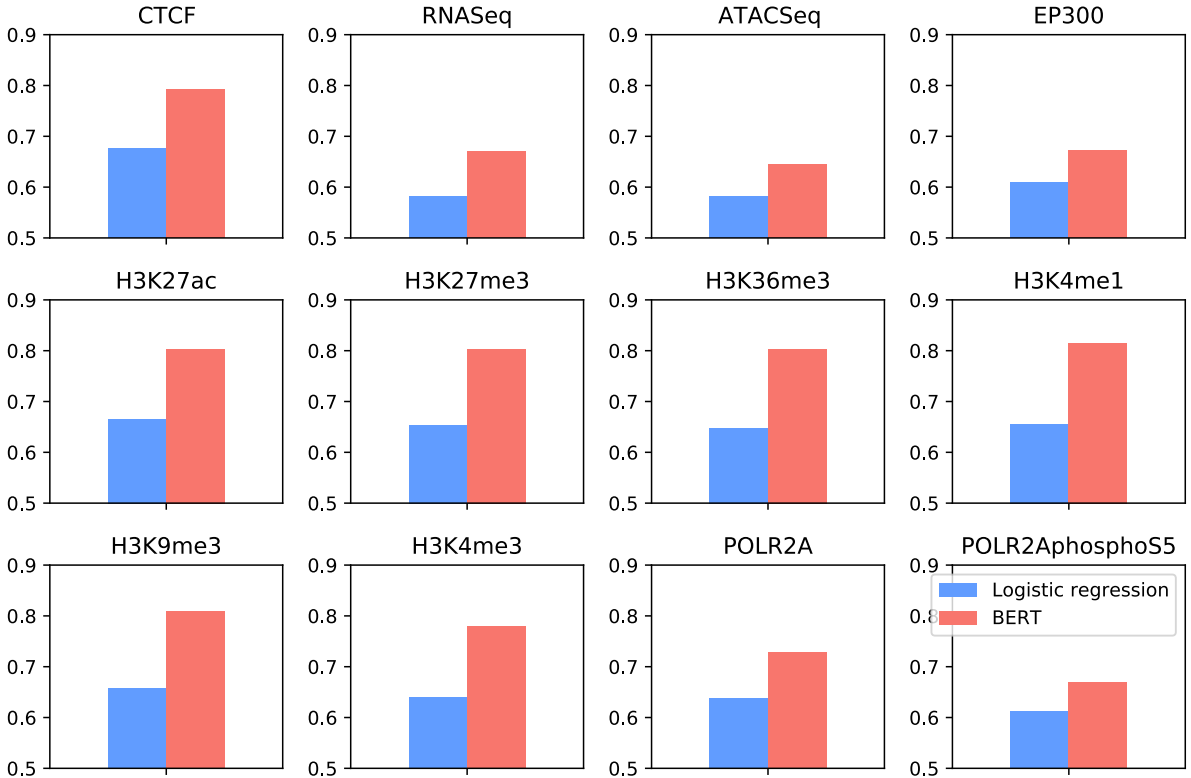


Sequence embedding

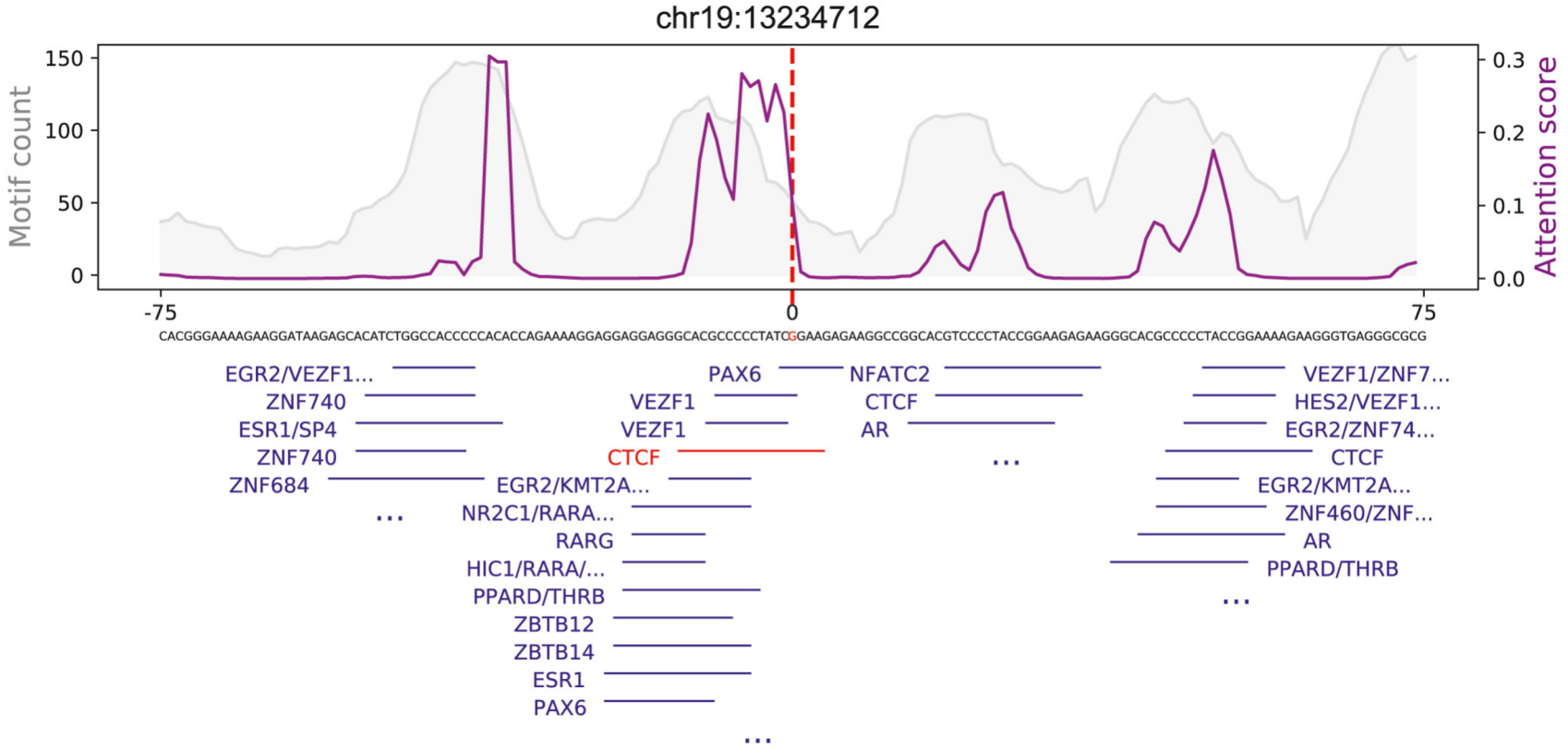


[Rozowsk et al. Cell '23] – Lectures.gersteinlab.org

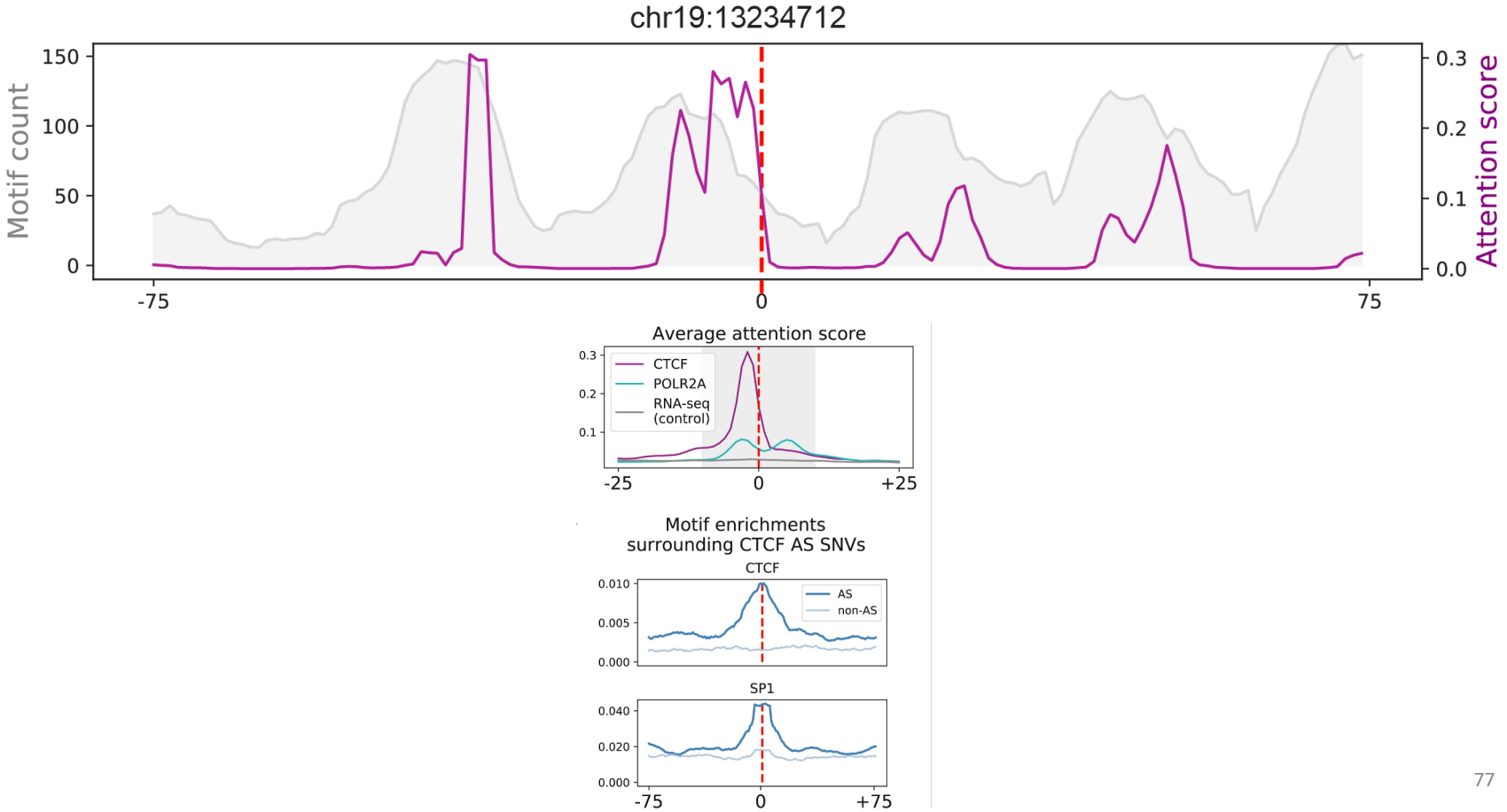
Allele specificity prediction



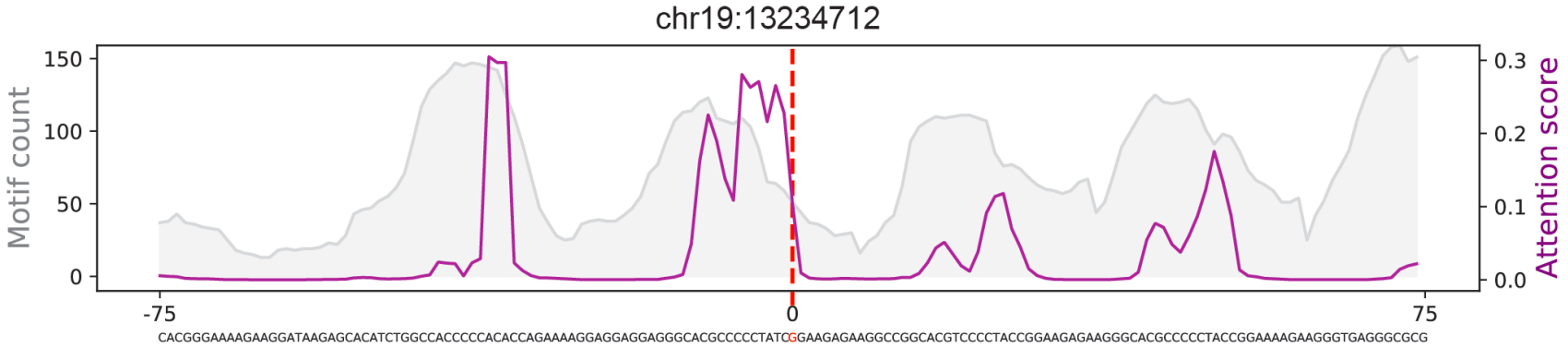
Predicting AS activity just from nucleotide sequence (Attention)



Predicting AS activity from just nucleotide sequence (Attention)

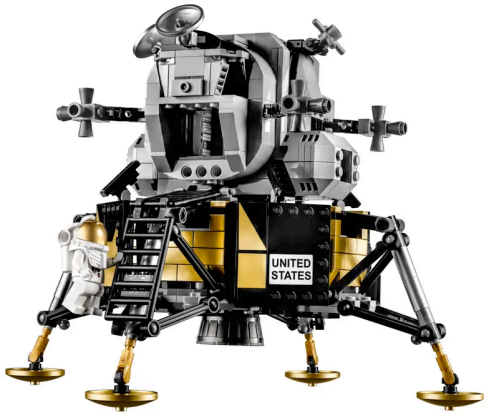


Predicting AS activity from just nucleotide sequence (Attention)



Both the motif centered at the SNV position & the surrounding sequence motifs of other TFs are relevant for AS behavior

- For instance, a mutated TF-binding site could be stabilized by other cofactors & show no AS behavior
- Similar to the legs of the Lunar Module:
  If one doesn't work, the three other legs can still anchor properly



Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

Developing Computational Representations of Disease-Relevant Molecules: 3 Cases Studies for AI in Biomedicine

- **Part 1: Multi-scale Modelling for Brain Disorders (Representing Molecular & Cellular Networks in a DL Framework)**
  - PsychENCODE consortium, leveraging high heritability of psychiatric diseases with functional genomics
- **Uniform Processing of Single Cell Data for 388 Brains**
  - 28 cell types, merging BICCN with a PFC-focused study
  - >500K scCREs
- **Creating Cellular Networks**
  - 1.4 M scQTLs from GTEx methods
  - Building cell-type-specific regulatory networks from scQTLs, scCREs & single-cell co-expression
  - Cell-to-cell communication networks, with changes in disease
- **Integrative Models Using These Networks**
  - Embedding regulatory networks & cell-to-cell communication networks in a deep-learning model to predict disease from genotype
  - Using this to prioritize specific pathways & genes.
  - Modelling perturbations & using these to suggest pot. drug targets

- **Part 2: Measuring Genomic Privacy Risk from a Few, Noisy SNPs (Understanding Information Leakage in terms of Constraints on Haplotype Trajectories)**
  - The Dilemma of Genomic Privacy: The genome as fundamental, inherited info that's private v. need for large-scale sharing & mining for med. research
  - 30 SNPs from "environmental" coffee cup sample sufficient for ID
  - Based on finding most likely haplotype "trajectories" in a genome DB
  - Single trajectory for a unique match for an ensemble of equivalent ones for near match
  - Calculating a PRS score over an ensemble

- **Part 3: Variant Impact for Precision Medicine (Learning Distributed Sequence Patterns via Transformer Attention)**
  - Need for variant catalogs & interpretation resources
- **EN-TEx: a Resource for Variant Interpretation**
  - >1500 functional experiments with diploid genomes of 4 individuals
  - Differential mapping to haplotypes
- **Development of AS Catalog**
  - >1M allele-specific events, over all samples from jt. calling
  - Useful biological interpretation: chrX, SVs, Igf2-H19
  - Association of AS events & eQTL/GWAS variants; provides a source of variant interpretation
  - Relating AS events to tissue-specificity & conservation
- **Transformer model to Predict AS Variants**
  - Identification of sensitive TF binding motifs
  - Model can successfully predict if a SNV will be AS purely from sequence; however, it requires extended context (~150 bp) around SNV
  - Interpretation in terms of anchoring co-factors

Acknowledgements

github.com/gersteinlab/PLIGHT

P Emani, Geradi, M.N.; Gürsoy, G.; Grasty, M.R.; Miranker, A

Also: JOBS.gersteinlab.org

Team work with >80 authors
13 Labs

P. S. Emani, J. J. Liu, D. Clarke, M. Jensen, J. Warrell, C. Gupta, R. Meng, C. Y. Lee, S. Xu, C. Dursun, S. Lou

Y. Chen, Z. Chu, T. Galeev, A. Hwang, Y. Li, P. Ni, X. Zhou

T. Chatterjee, Y. Dai, Zi. Duan, M. Gancz, D. Garrido-Martín, E. Henry, G. E. Hoffman, A. Huang, Y. Jiang, T. Jin, S. Khullar, J. Liu, S. Liu, J. Moore, E. Nguyen, N. Phalke, M. Pjanic, H. Pratt, A.S. Rajagopalan, T. R. Riesenmy, N. Shedd, G. Wang, Y. Xia, A. C. Yang, S. Zheng, D. Lee, Z. Weng, H. Won

T. E. Bakken, J. Bendl, L. Bicks, L. Cheng, Y. Cheng, M. Flaherty, J. F. Fullard, S. Gaynor-Gillett, J. Grundman, N. Hawken, N. L. Jorstad, R. Kawaguchi, J. Liu, S. Ma, M. Margolis, S. Mazariegos, J. R. Moran, D. Quintero, M. Shi, M. Spector, R. Terwilliger, K. J. Travaglini, B. Wamsley, S. Xiao, M. J. Gandal, E. S. Lein, P. Roussos, N. Sestan, K. P. White

M. J. Girgenti, J. Zhang, D. Wang, D. Geschwind, M. Gerstein

& PsychENCODE Consortium

National Institute of Mental Health

PsychENCODE Consortium

212 members
Labs from 37 universities/institutions

http://brainscope.psychencode.org/

brainSCOPE

Yale · The University of Wisconsin Madison · UCI · Allen Institute · Icahn School of Medicine at Mount Sinai · UCLA · TEMPUS · Penn University of Pennsylvania · Universitat de Barcelona

Cornell College · James J. Peters Department of Veterans Affairs Medical Center · University of Chinese Academy of Sciences · UMass Chan Medical School · Children's Hospital of Philadelphia · University of Washington · NUS National University of Singapore · The University of North Carolina at Chapel Hill · U.S. Department of Veterans Affairs

Team Science with >100 Authors, 20 Labs, 3 Consortia, 1 Funder

Joel Rozowsky[$], Jiahao Gao[$], Beatrice Borsari[$], Yucheng T Yang[$], Timur Galeev[$], Gamze Gursoy[$], Charles B Epstein[$], Kun Xiong[$], Jinrui Xu[$], Tianxiao Li[$], Jason Liu[$],

Keyang Yu[7†], Ana Berthel[2,3†], Zhanlin Chen[8†], Fabio Navarro[2,3†], Maxwell S Sun[2,3†], James Wright[9†], Justin Chang[2,3†], Christopher JF Cameron[2,3†], Noam Shoresh[6†], Elizabeth Gaskell[6†], Jorg Drenkow[10†],

Jessika Adrian[11], Sergey Aganezov[12], François Aguet[6], Gabriela Balderrama-Gutierrez[13], Samridhi Banskota[6], Guillermo Barreto Corona[6], Sora Chee[14], Surya B Chhetri[15], Gabriel Conte Cortez Martins[2,3], Cassidy Danyko[10], Carrie A Davis[10], Daniel Farid[2,3], Nina P Farrell[6], Idan Gabdank[11], Yoel Gofin[7], David U Gorkin[14], Mengting Gu[2,3], Vivian Hecht[6], Benjamin C Hitz[11], Robbyn Issner[6], Yunzhe Jiang[2,3], Melanie Kirsche[12], Xiangmeng Kong[2,3], Bonita R Lam[11], Shantao Li[2,3], Bian Li[2,3], Xiqi Li[7], Khine Zin Lin[11], Ruibang Luo[16], Mark Mackiewicz[15], Ran Meng[2,3], Jill E Moore[17], Jonathan Mudge[18], Nicholas Nelson[6], Chad Nusbaum[6], Ioann Popov[2,3], Henry E Pratt[17], Yunjiang Qiu[14], Srividya Ramakrishnan[12], Joe Raymond[6], Leonidas Salichos[2,3,19], Alexandra Scavelli[10], Jacob M Schreiber[20], Fritz J Sedlazeck[12,21,22], Lei Hoon See[10], Rachel M Sherman[12], Xu Shi[2,3], Minyi Shi[11], Cricket Alicia Sloan[11], J Seth Strattan[11], Zhen Tan[2,3], Forrest Y Tanaka[11], Anna Vlasova[4,23,24], Jun Wang[2,3], Jonathan Werner[10], Brian Williams[25], Min Xu[1,2], Chengfei Yan[2,3], Lu Yu[9], Christopher Zaleski[10], Jing Zhang[26],

Kristin Ardlie, Michael Cherry, Eric Mendenhall, William Noble, Zhiping Weng, Morgan Levine, Alexander Dobin, Barbara Wold, Ali Mortazavi, Bing Ren, Jesse Gillis, Richard Myers, Michael Snyder, Jyoti Choudhary, Aleksandar Milosavljevic, Michael Schatz[#], Bradley Bernstein[#], Roderic Guigó[#], Thomas Gingeras[#], Mark Gerstein[#]

ENCODE Labs, GTEx Labs, ROADMAP epigenomics PROJECT Labs

NIH NHGRI

Image Credit:
Susanna Liu (Yale)

entex.encodeproject.org

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2019.
  - Please read permissions statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  - Paper references in the talk were mostly from Papers.GersteinLab.org.

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt