Integrative machine learning for regulatory genomics

Alexis Battle

NHGRI Machine Learning in Genomics Workshop, 2021

Outline

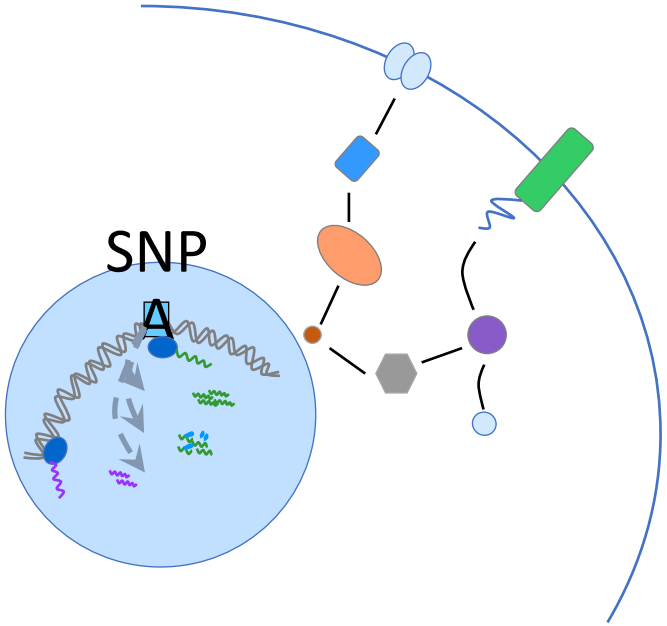
- Introduction and background
- Deep dive: The effects of rare genetic variation on gene expression
- Parting thoughts on ML in genomics

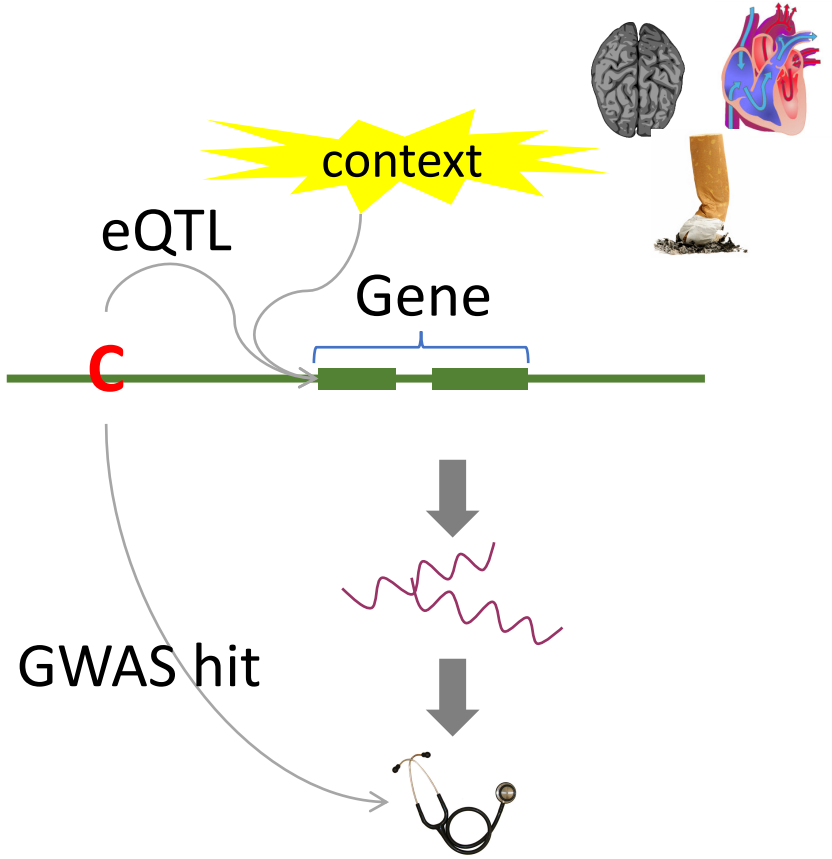
Introduction: integrative approaches for understanding the genetics of gene expression

Understanding regulatory variation

Identify the effects of regulatory genetic variation on gene expression and high-level phenotype

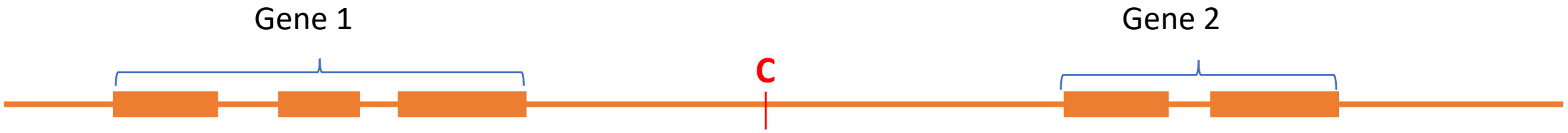
- Computational methods development





Non-coding genetic variation

Most variants, and most disease-associated variants, are non-coding



Challenges:

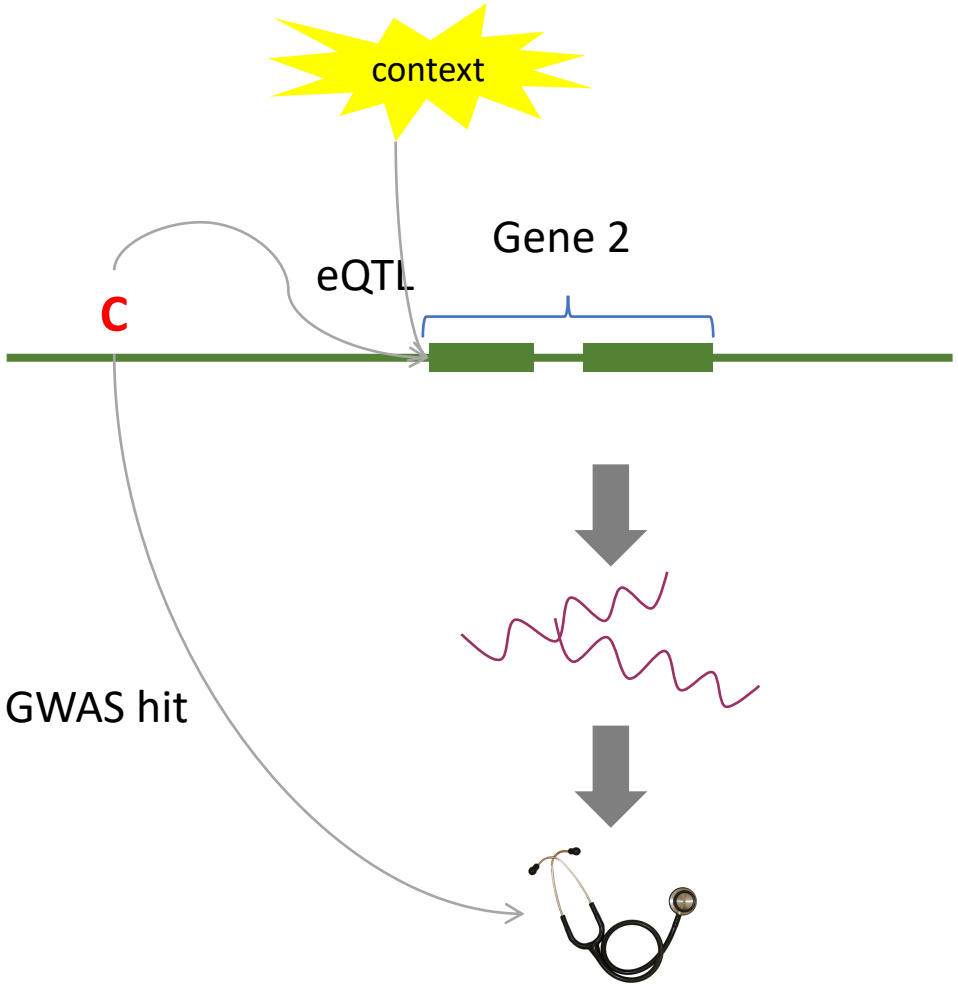
- Hard to predict effects of non-coding variants from sequence
- For disease-associated non-coding variants, difficult to interpret functional mechanism or design interventions

Context specificity

- Many factors *modulate* regulatory genetic effects

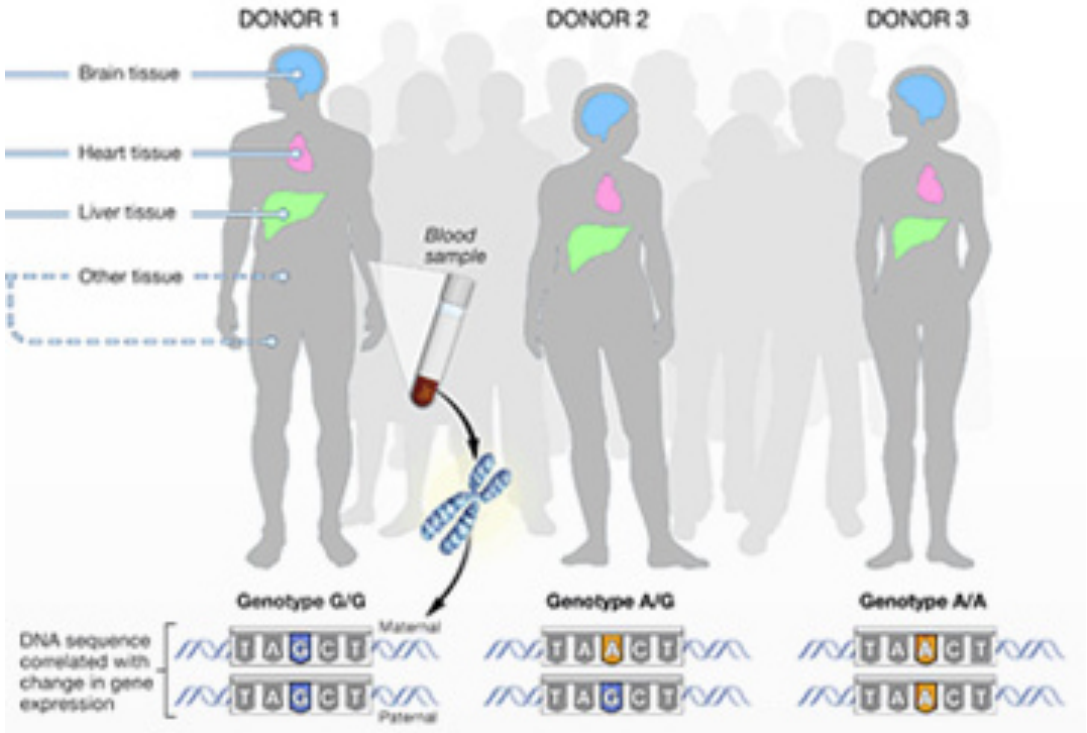
- Disease etiology may involve specific cell types, developmental stages, or environmental responses

- Need for tailored data and methods…



The GTEx project: tissue-specific gene expression

- 948 donors with WGS
- RNA-seq in 54 tissues
- Cis- and trans-eQTLs in each tissue
- Huge catalog of eQTLs reveals regulatory biology
- Intersection with genetics of disease
- https://www.sciencemag.org/collections/genetic-variation



GTEx enabled dozens of creative projects beyond eQTL study

GTEx Consortium Publications

2020

The GTEx Consortium atlas of genetic regulatory effects across human tissues
The GTEx Consortium.
Science. 369 (1318-1330), 10 Sep 2020. doi:10.1126/science.aaz1776

Cell type specific genetic regulation of gene expression across human tissues
Kim-Hellmuth* S, Aguet* F, Oliva M, Muñoz-Aguirre M, Kasela S, *et al*.
Science. 369 (eaaz8528), 10 Sep 2020. doi:10.1126/science.aaz8528

Transcriptomic signatures across human tissues identify functional rare genetic variation
Ferraro* NM, Strober* BJ, Einson J, Abell NS, Aguet F, *et al*.
Science. 369 (aaz5900), 10 Sep 2020. doi:10.1126/science.aaz5900

Determinants of telomere length across human tissues
Demanelis K, Jasmine F, Chen LS, Chernoff M, Tong L, *et al*.
Science. 369 (aaz6876), 10 Sep 2020. doi:10.1126/science.aaz6876

The impact of sex on gene expression across human tissues
Oliva* M, Muñoz-Aguirre* M, Kim-Hellmuth* S, Wucher V, Gewirtz ADH, *et al*.
Science. 369 (aba3066), 10 Sep 2020. doi:10.1126/science.aba3066

Tissue-specific genetic features inform prediction of drug side effects in clinical trials
Duffy A, Verbanck M, Dobbyn A, Won H-H, Rein JL, *et al*.
Science Advances. 6(37), eabb6242, 10 Sep 2020. doi:10.1126/sciadv.abb6242

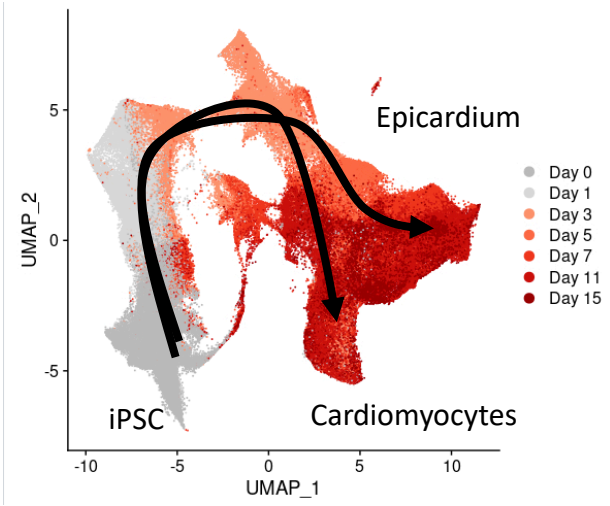
PhenomeXcan: Mapping the genome to the phenome through the transcriptome
Rividori M, Rajagopal PS, Barbeira A, Liang Y, Melia O, *et al*.

True for other large-scale datasets as well:

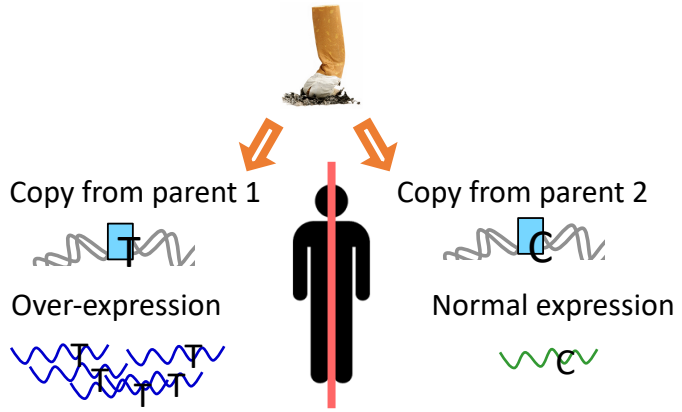
Depression Genes and Networks
ENCODE
Roadmap Epigenomics
UKBB
HapMap and 1000 Genomes
GEUVADIS
etc

Lab projects: ML + diverse transcriptomic data

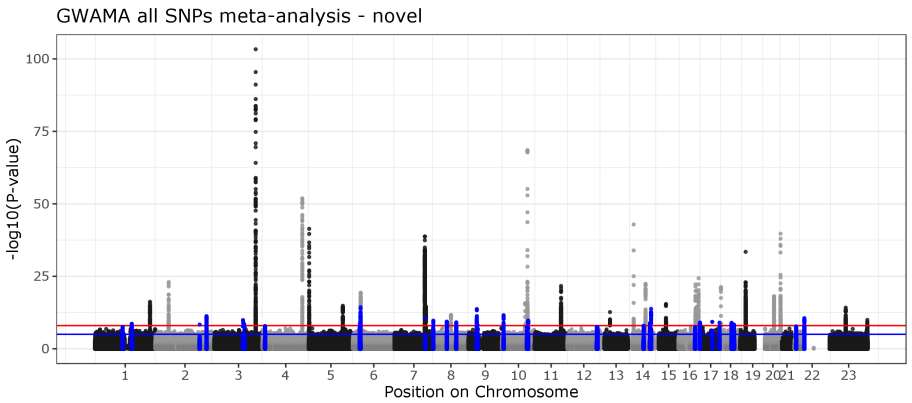
Single-cell and dynamic eQTL models



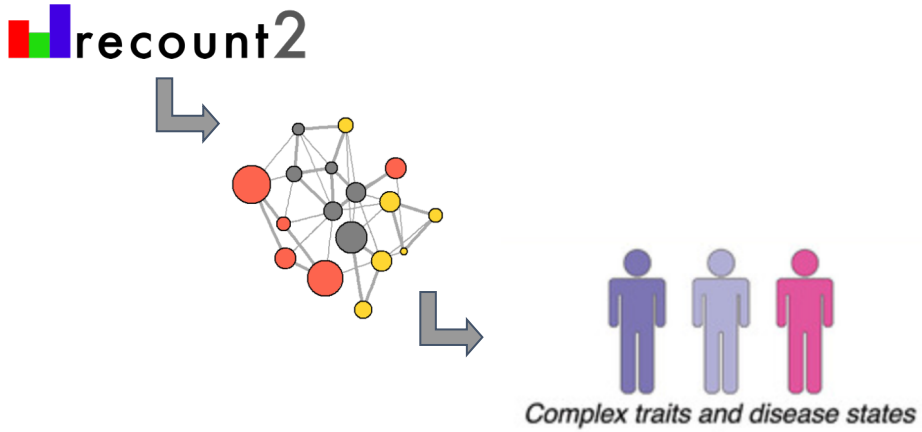
Context specificity



Disease and multi-omic integrative analysis



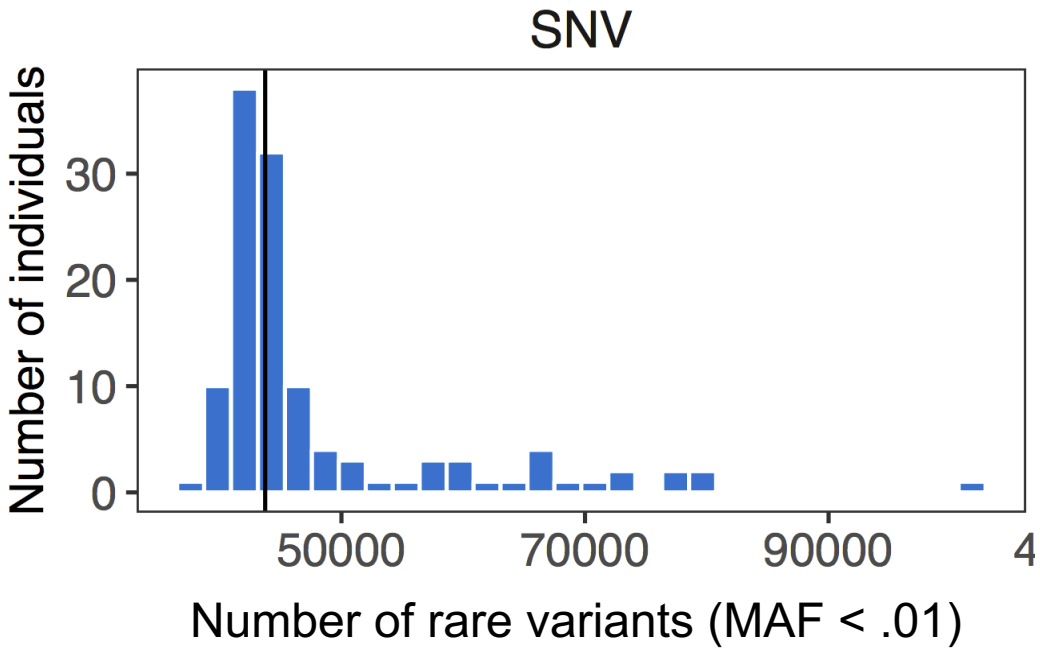
Large-scale network inference and integration



ML for rare genetic variation

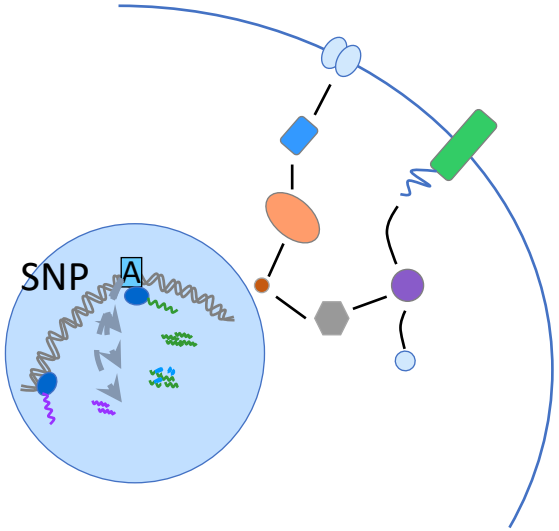
Motivation – Rare variation is abundant and mostly uncharacterized

- Individual genomes have a median of approximately 50,000 rare variants with MAF below 0.01
- Rare variants are enriched for deleterious properties, contribute to rare and complex disease
- Evaluating rare variants from whole genome sequencing remains very challenging
- Approximately half of rare disease patients go undiagnosed with current approaches



Li et al. Nature 2017

Project goals

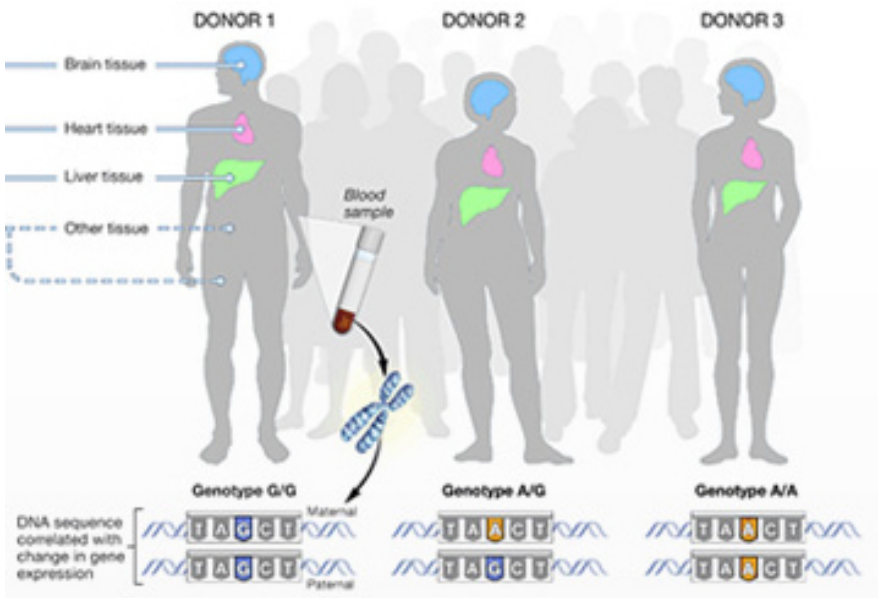


- Explore the impact of *rare*, regulatory variation
- Identify complex effects of rare variants from RNA-seq
- **Integrative model to prioritize rare regulatory variants from personal genomes supplemented with RNA-seq**

Analysis of rare variation in GTEx data

GTEx Project v8 rare variant analysis

- 714 individuals of European ancestry:
  - RNA-seq across multiple tissues
  - Whole genome sequencing



Using RNA-seq to help prioritize functional rare variants

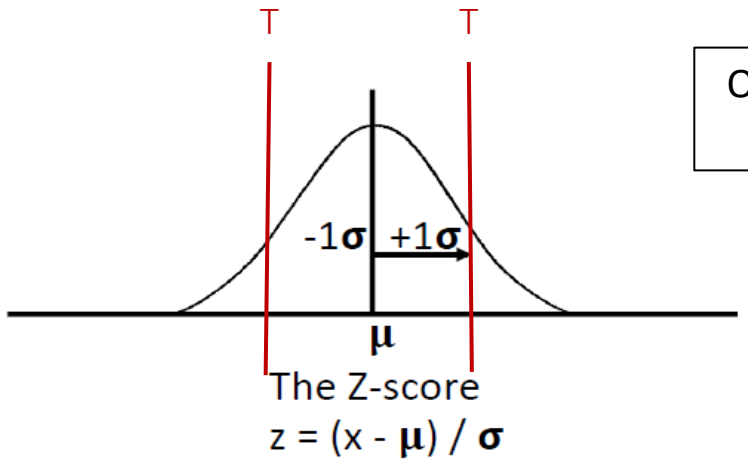
Hypothesis:

- Functional variants cause disruption at a **cellular level**
- Rare regulatory variation will result in **unusual** expression of nearby genes

Simple approach:

- Identify individuals whose gene expression is far from the population average

Total Expression Outliers



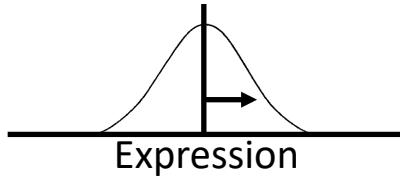
Outlier criterion
$|z| > T$

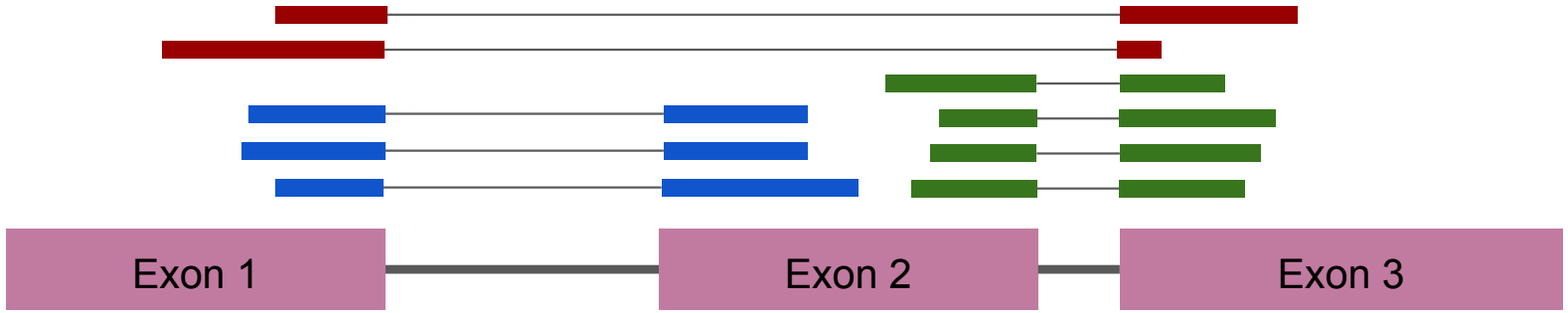
Li *et al.*, Nature 2017
Li *et al.*, *AJHG* 2014
Zeng *et al.*,*PLoS Genet* 2015
Zhao *et al.*, *AJHG* 2016
Cummings *et al.*, STM, 2017

Alternative splicing outliers

- Both rare and common genetic variants affecting splicing have been implicated in disease

- Abnormal *total* gene expression simply goes up or down compared to normal





- So how to define an outlier over a multi-dimensional space of possible splice junctions?

Cummings et al, STM, 201

Kremer et al, Nature 201

SPOT (SPlicing Outlier deTection)

J

N [X]

Junction counts

Estimate parameters of Dirichlet multinomial (DM) distribution from N samples

$X_1,\ldots X_j \sim \text{Multinomial}(p_1,\ldots p_j)$

$p_1,\ldots p_j \sim \text{Dirichlet}(\alpha_1,\ldots \alpha_j)$

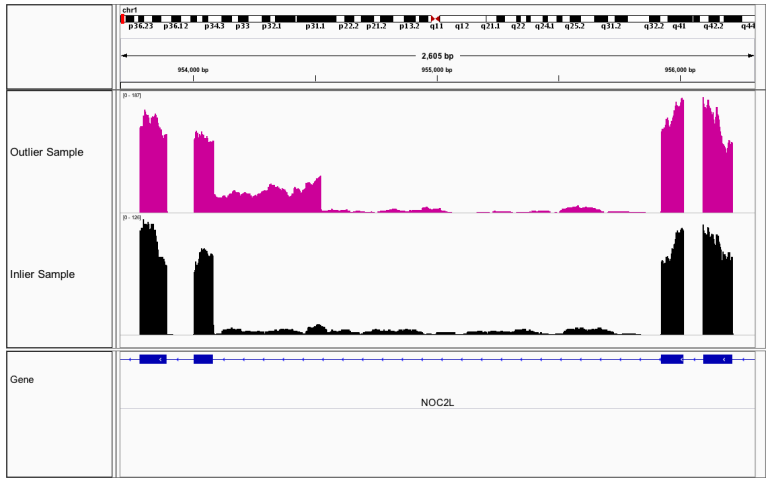
Compute Mahalanobis distance (MD) for each sample based on estimated DM mean and covariance

$$MD(X) = \sqrt{(X-\mu)^T \Sigma^{-1} (X-\mu)}$$

N – Number of individuals

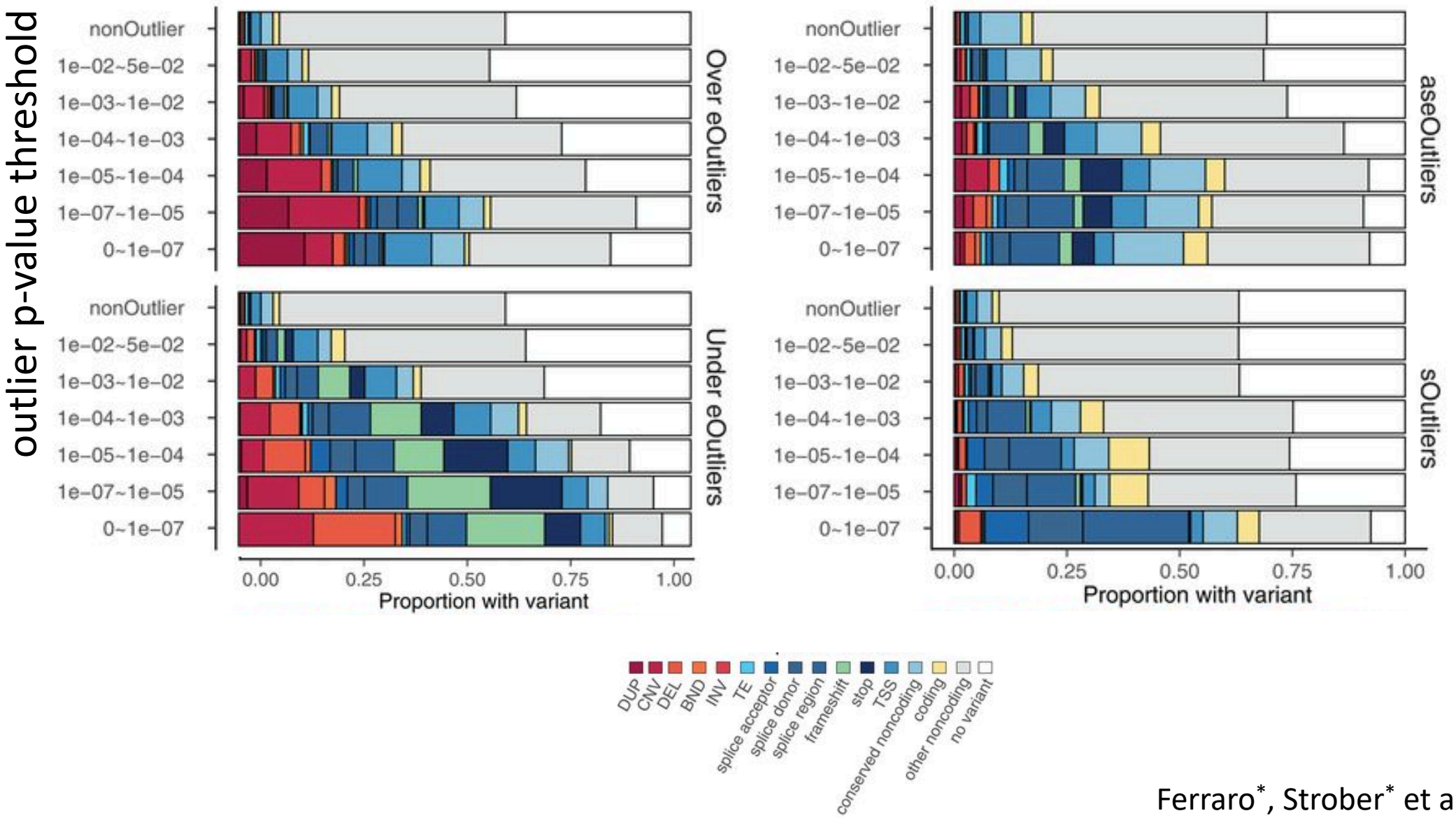
J – Number of observed junctions

LeafCutter quantification



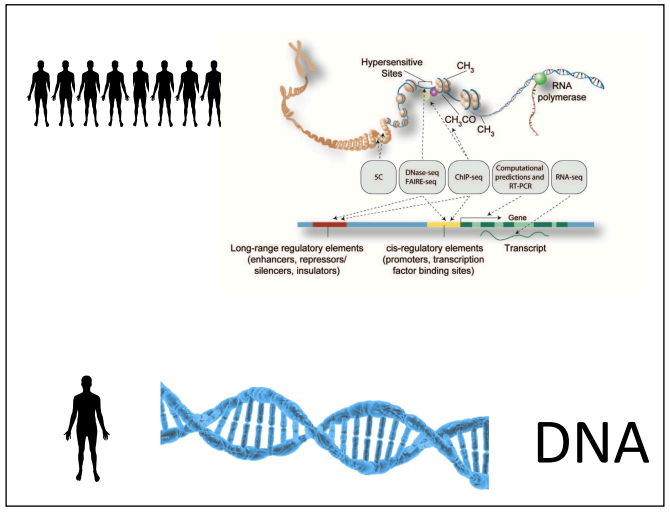
Ferraro*, Strober* et al, Science 2020

Outliers are enriched for distinct functional classes of rare variants

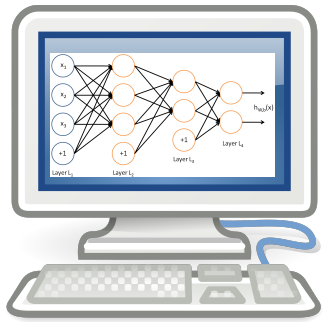


Ferraro*, Strober* et al, Science 2020

Machine learning for personal genomics



- CADD (Kircher et al, Nature, 2014)
- GWAVA (Ritchie et al, Nature Methods, 2014)
- BASSET (Kelley et al, Genome Research, 2016)
- DeepSEA (Zhou et al, Nature Methods, 2015)



Prediction function Y = f(X; θ)

Personal genomic predictions

Likely functional

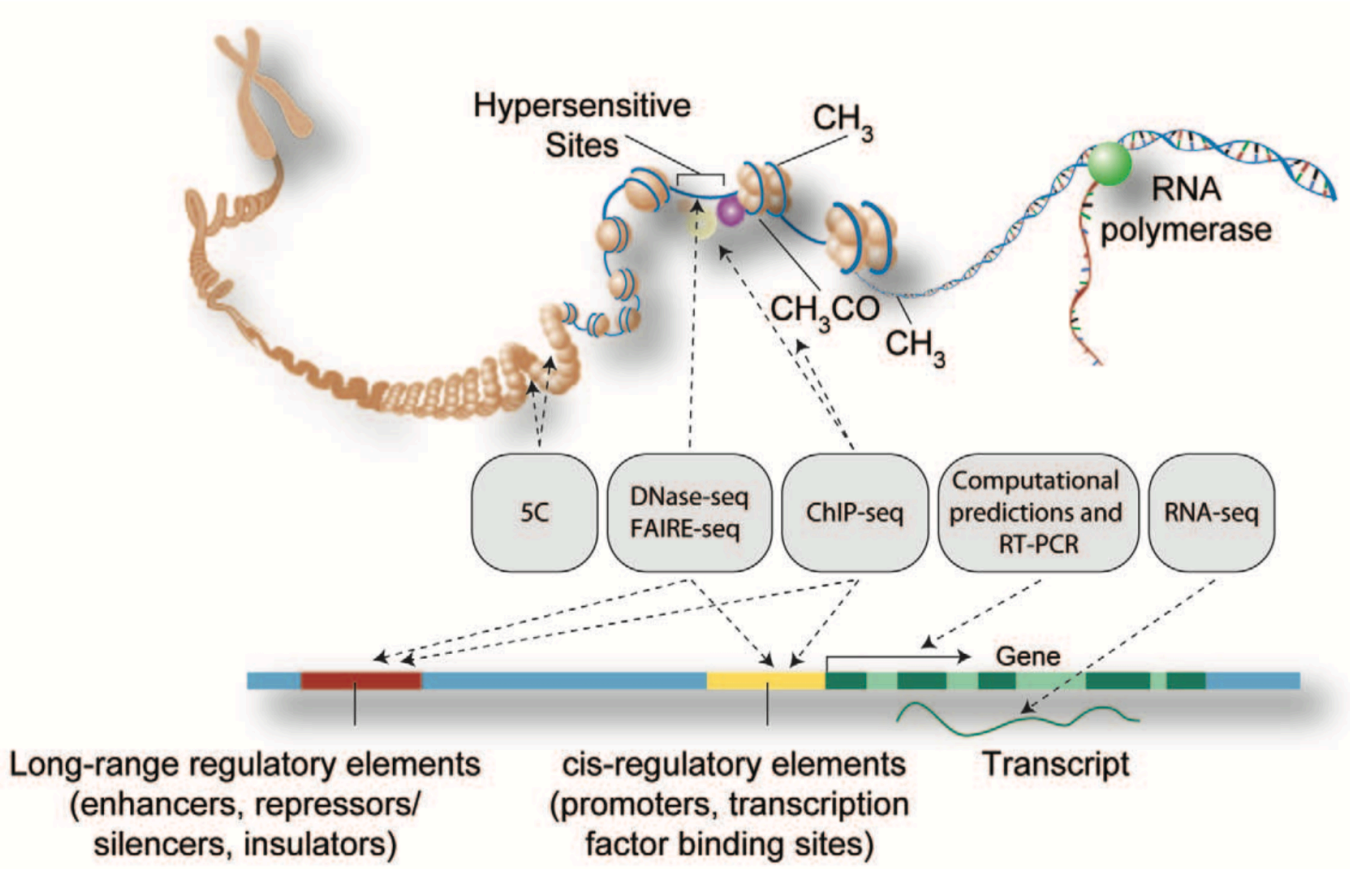
Chr2: AAC**T**TA
Chr16: TGC**A**TC

benign

Chr7: AAA**G**TC
Chr16: GCG**A**CC
..
Chr21: GGC**A**AT

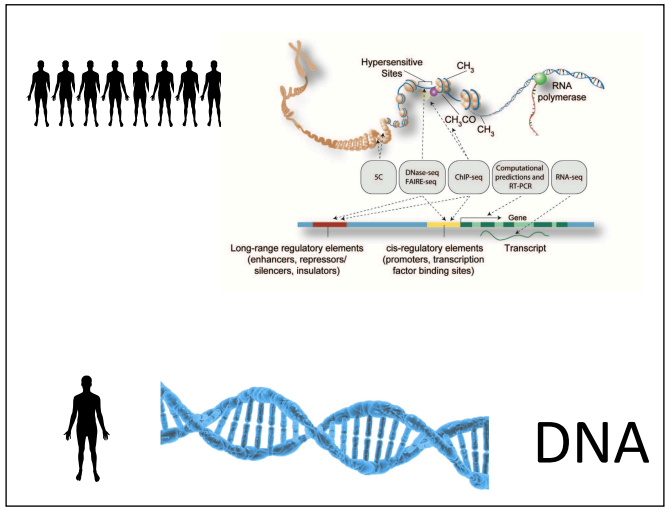
Diverse genomic feature data available

ENCODE Project Consortium. Plos Biology 2011.

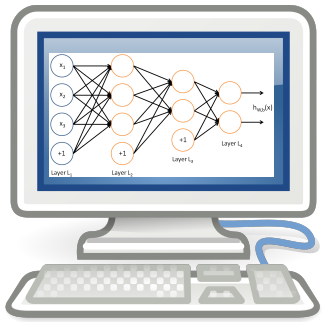


- Regulatory elements from Roadmap, ENCODE
- Conservation scores
- Transcription factor binding sites
- CpG sites
- Summary scores from existing WGS models

Machine learning for personal genomics



Hypothesis: a rare variant that is impacting health will also have a molecular signature in the affected person



Prediction function Y = f(X; θ)

Personal genomic predictions

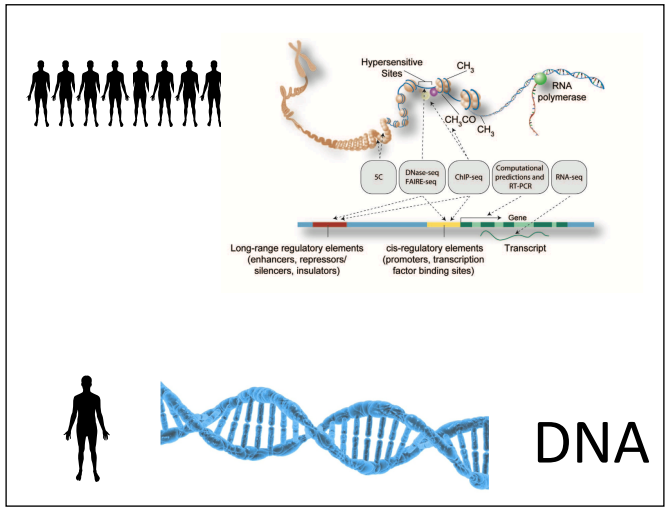
Likely functional

```
Chr2:  AACTTA
Chr16: TGCATC
```

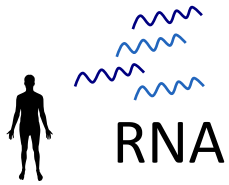
benign

```
Chr7:  AAAGTC
Chr16: GCGACC
..
Chr21: GGCAAT
```

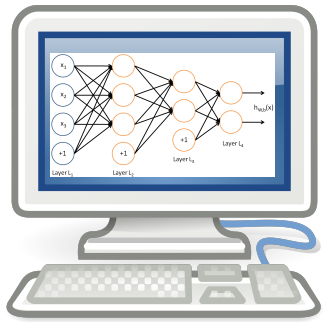
Machine learning for personal genomics



Personal transcriptomic data



Hypothesis: a rare variant that is impacting health will also have a molecular signature in the affected person



Prediction function Y = f(X; θ)

Personal genomic predictions

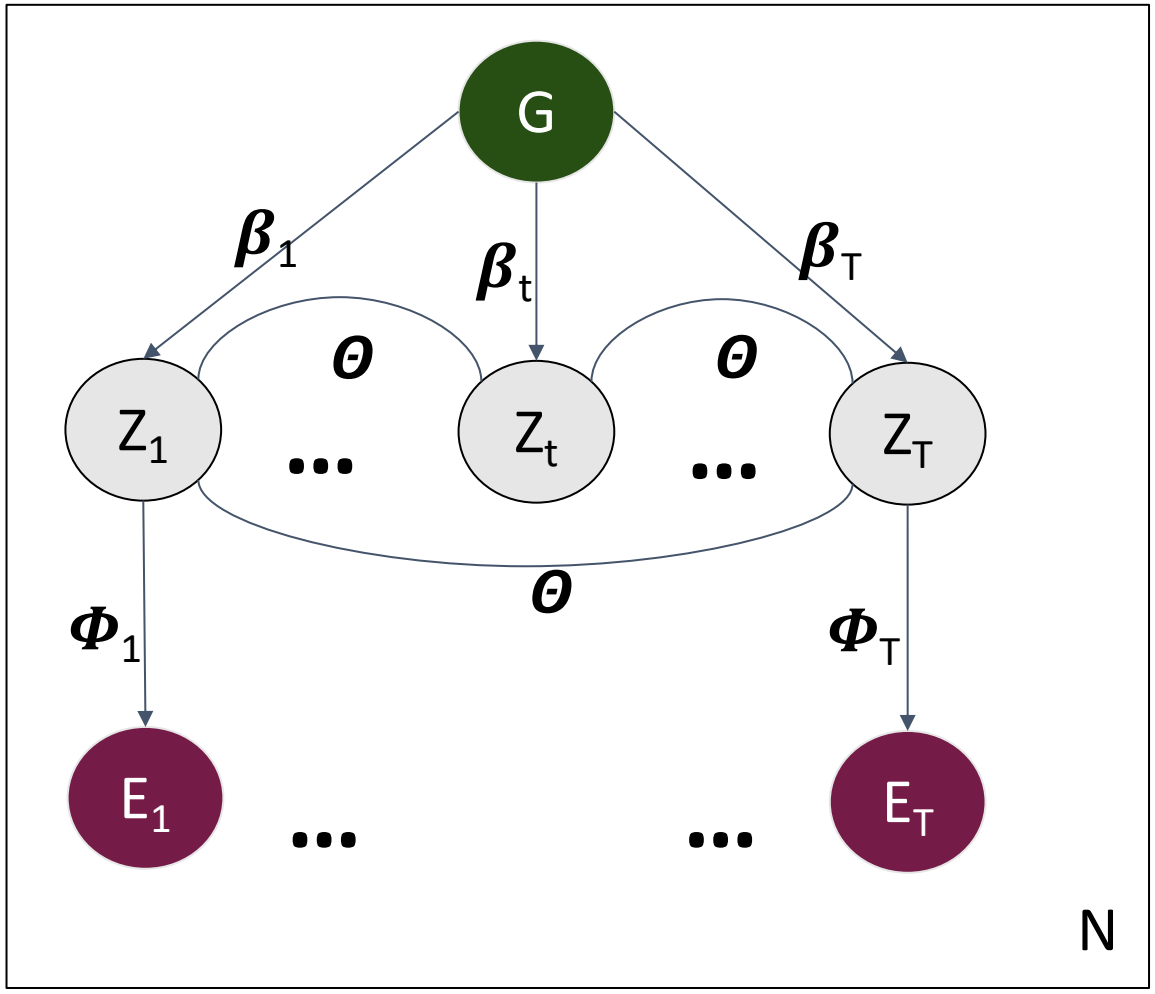
Likely functional

```
Chr2:  AACTTA
Chr16: TGCATC
```

benign

```
Chr7:  AAAGTC
Chr16: GCGACC
..
Chr21: GGCAAT
```

Watershed model integrates multiple molecular signals



Model each gene, in each individual (N instances)

G Genomic features of rare variants from whole genome sequence

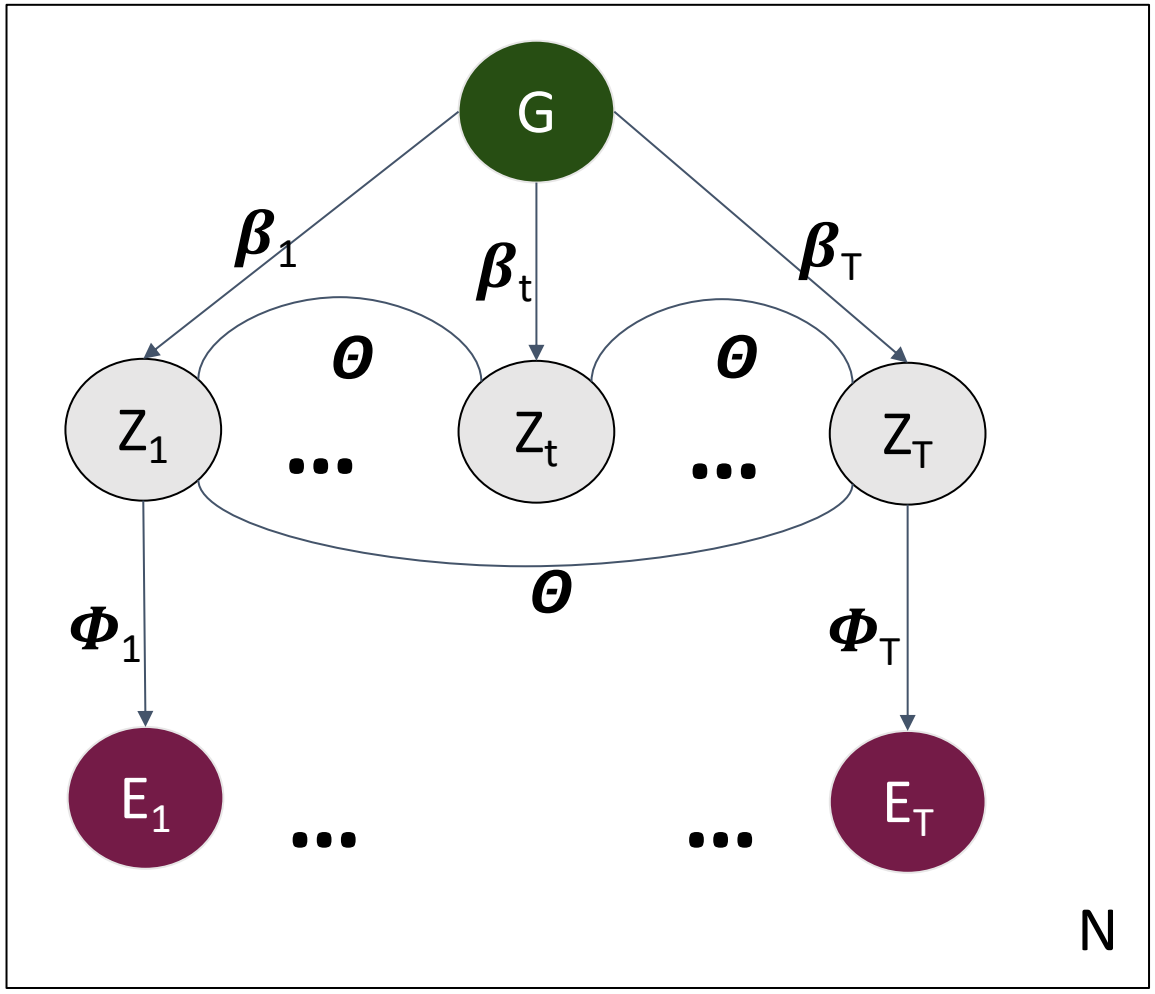
Z_t Latent variable of whether this rare variant has a regulatory effect on molecular phenotype t, ising model

E_t Signal from molecular phenotype t (outlier status) – ASE, splicing, total expression

Extensible to any molecular signal or data type

Ferraro*, Strober* et al, Science 2020

Watershed model integrates multiple molecular signals



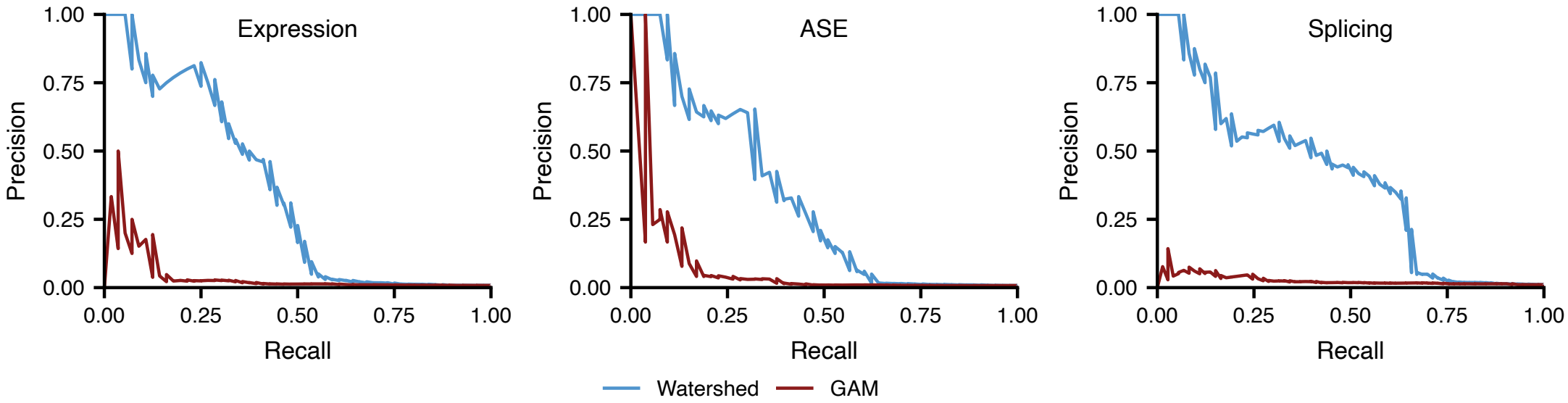
Unsupervised: Z unobserved, does not require labeled training data

Efficient: optimize model parameters using EM with approximate inference

Provides posterior probability of impact for each rare variant (Z_t) given any observed WGS and RNA-seq data in a new patient

Ferraro*, Strober* et al, Science 2020

Watershed: RNA improves prediction over WGS

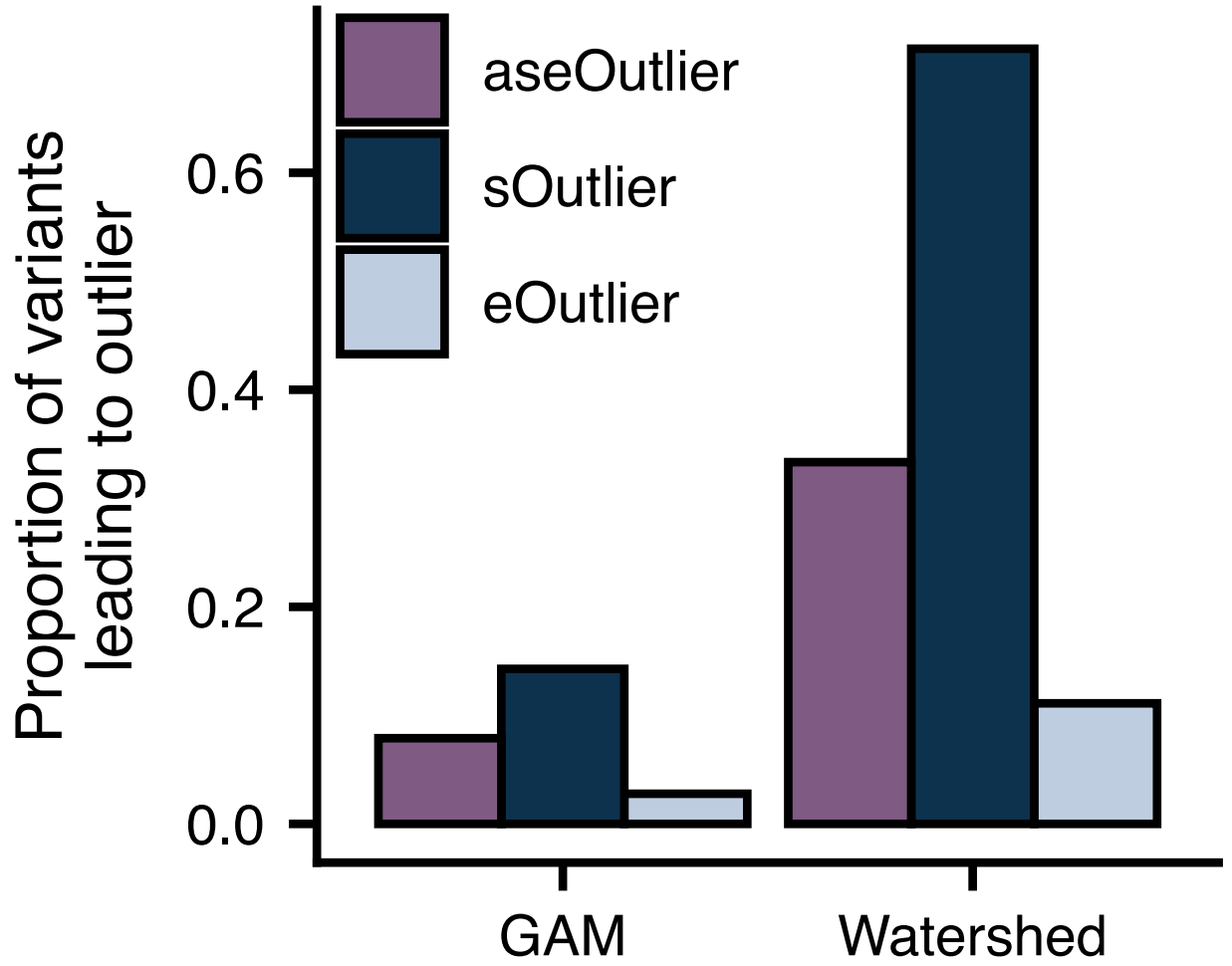


- Predicting variant effects in held out individuals (N=2 analysis)
- Watershed, utilizing RNA-seq, offers large improvements over WGS alone ("GAM")

- Replicated in independent data, and several variants validated with CRISPR-Cas9

Ferraro*, Strober* et al, Science 2020

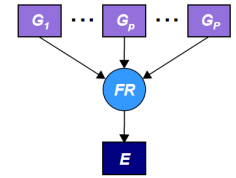
Watershed dramatically improves identification of rare variants with high risk of functional impact



Ferraro*, Strober* et al, Science 2020

Conclusions

- Rare genetic variants coincide with large transcriptomic changes
- Integrative model Watershed uses diverse signals from RNA, providing improvements in rare variant prioritization over only WGS
  - Extensible to multi-omic and other data types



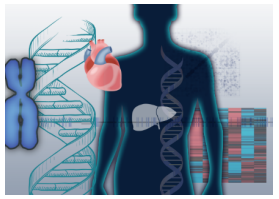
- https://science.sciencemag.org/content/369/6509/eaaz5900
- https://github.com/BennyStrobes/SPOT, https://github.com/BennyStrobes/Watershed

Parting thoughts on enabling ML in genomics

- Key resources and opportunities:
  - Large, accessible datasets enable diverse creative applications
  - Diverse data types
  - Flexible computational resources (increasing interest in cloud)
  - Tools and software for powerful ML frameworks (deep learning, probabilistic models, traditional ML)

- Challenges:
  - Confounders and technical artifacts (extensive metadata!)
  - Training researchers for highly interdisciplinary work
  - Vetting and maintaining computational tools in academia
  - Reproducibility
  - Interpretability

Thank you



Battle Lab members

Marios Arvanitis
Boris Brenerman
Jessica Bonnie
Diptavo Dutta
Surya Chhetri
Matthew Figdore
Seraj Grimes
Yuan He
Rebecca Keener
April Kim
Taibo Li
Bohan Ni
Ashton Omdahl
Princy Parsana
Joshua Popp
Guanghao Qi
Prashanthi Ravichandran
Ashis Saha
Benj Shapiro
Ben Strober
Karl Tayeb
Victor Wang

Collaborators

GTEx Consortium
Tuuli Lappalainen
Kristen Ardlie
Francois Aguet
Hae Kyung Im
Stephen Montgomery
Christopher Brown
Barbara Engelhardt
Pejman Mohammadi
Johan Einson
Nicole Ferraro
Xin Li
Yoav Gilad
Reem Elorbany
Katie Rhodes
David Knowles
Jeff Leek
Kasper Hansen
Maja Bucan