

Machine learning interprets genomics

- *Understanding functional genomics in human brain*

CS540 Introduction to Artificial Intelligence, Fall 2020

Daifeng Wang, Ph.D.

Assistant Professor

Department of Biostatistics and Medical Informatics

Department of Computer Sciences (affiliate)

Investigator, Waisman Center

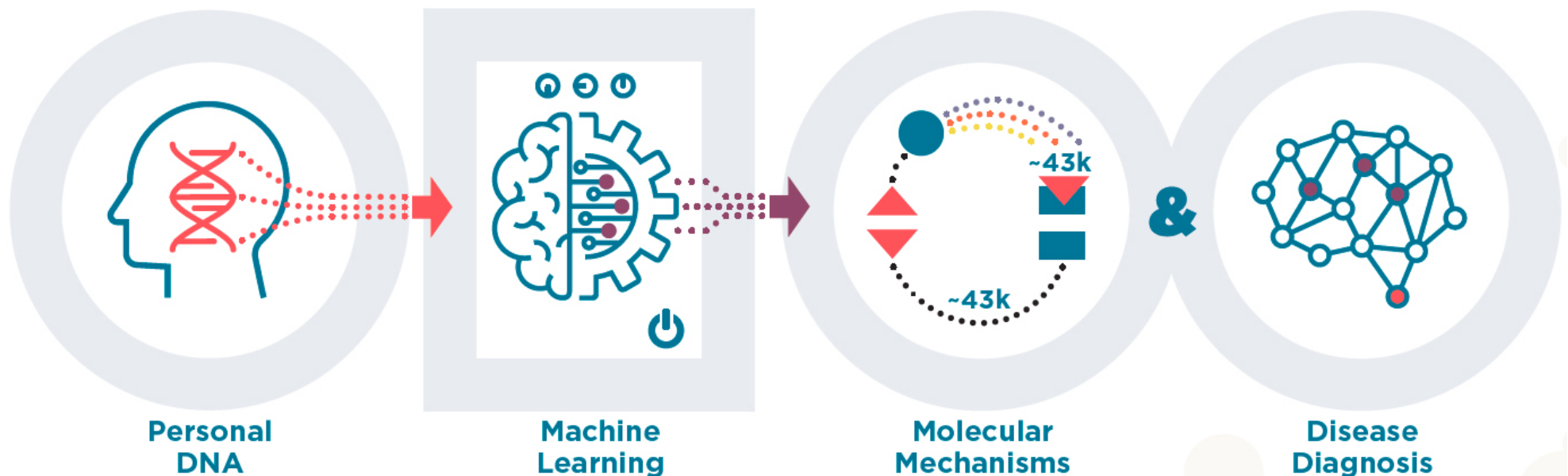
University of Wisconsin-Madison

daifeng.wang@wisc.edu

Research in my lab

- Goal
 - *Advance biological knowledge on genomics in brain diseases*
- Approach
 - *Interpretable computational approaches; e.g., machine learning*

Decoding Genomic Information to Better Understand Molecular Mechanisms and Improve Disease Diagnosis

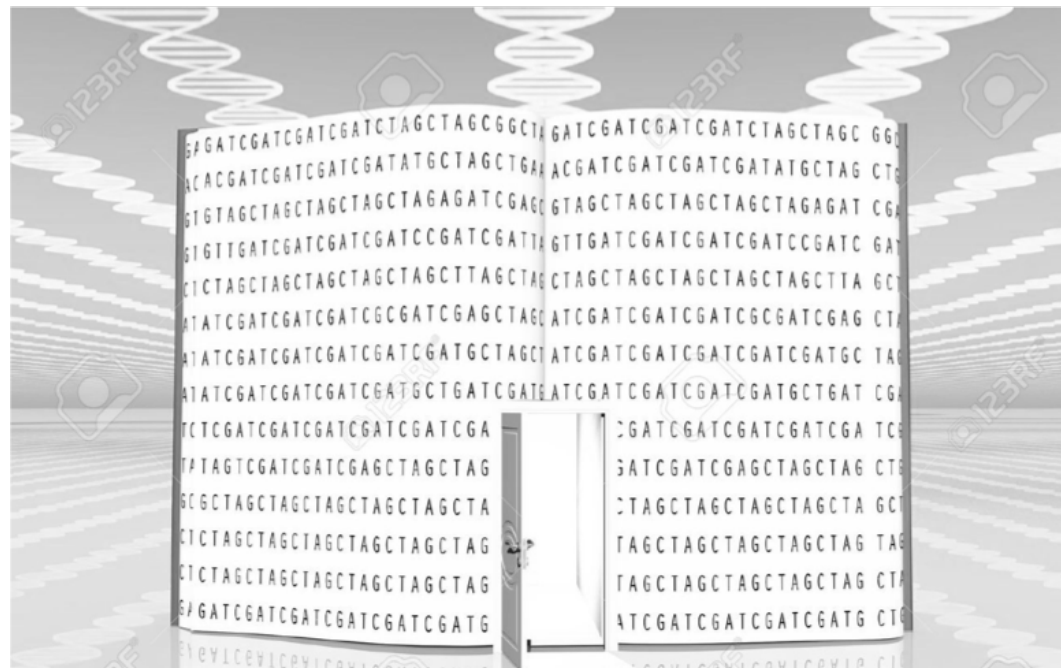


Your genome is your genetic code book

Book	Genome
Chapters	Chromosomes
Sentences	Genes
Words	Elements
Letters	Bases

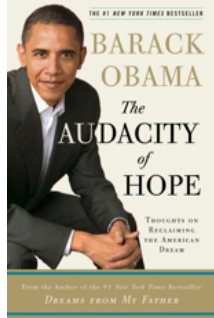
Human

- 46 chromosomes
- ~ 20,000 – 25,000 genes
- ~ Millions elements
- 4 unique bases (A, T, C, G), ~3 billion in total



<https://goo.gl/images/vMaz4T>

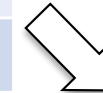
How to read sentences/genes for understanding book/genome?



Chapter One
Republicans and Democrats



Book	Genome
Chapters	Chromosomes
Sentences	Genes
Words	Elements
Letters	Bases

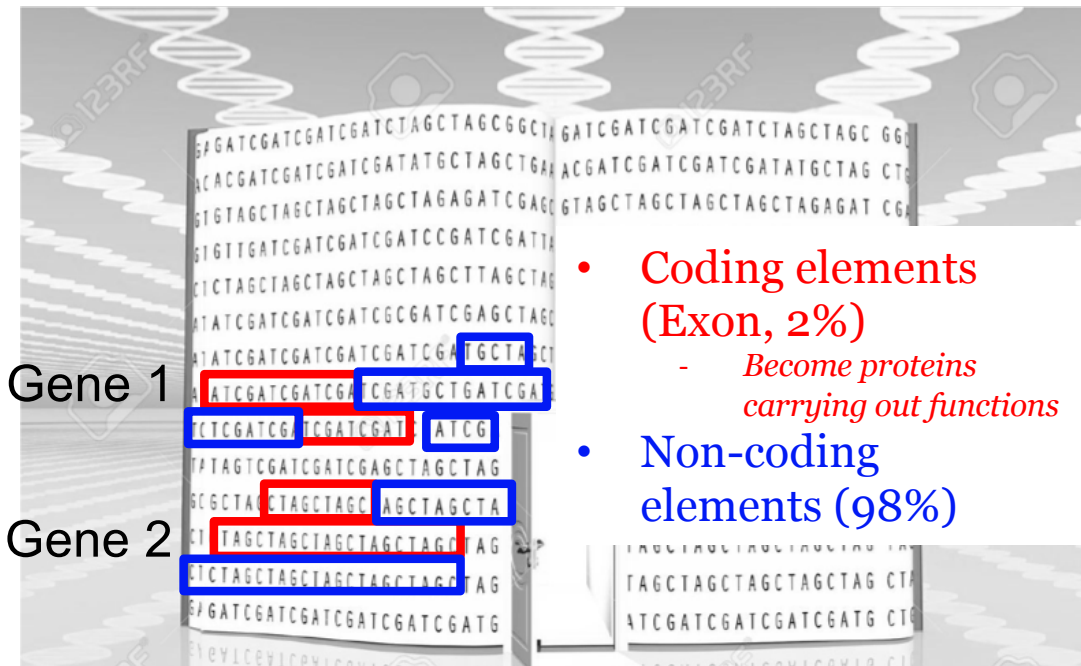


“On most days, I enter the Capitol through the basement. A small subway train carries me from the Hart Building, where ...”

- Key words
- Non-key words

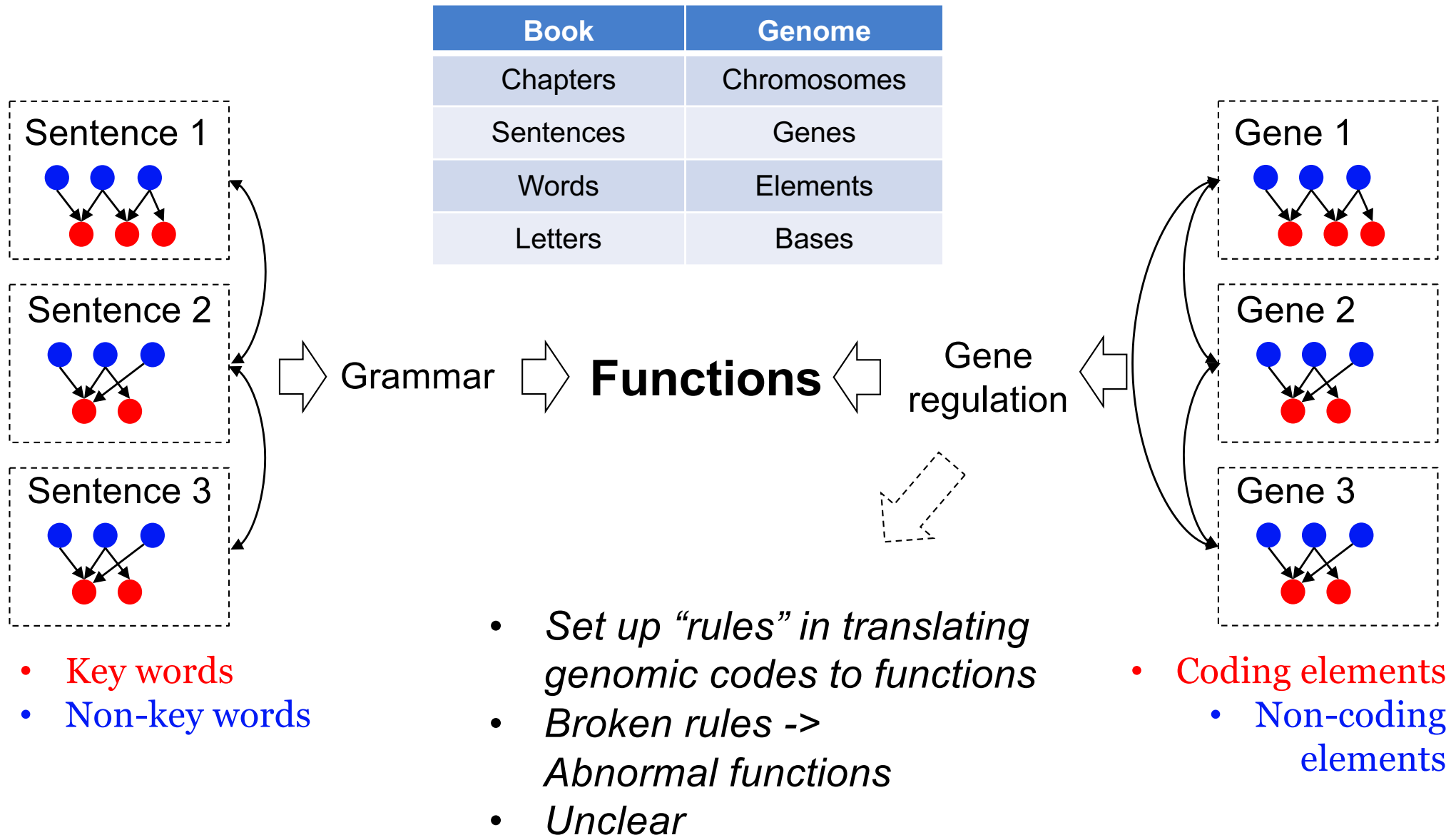
Overhead, the ceiling forms a creamy white oval, with an American eagle etched in its center. Above the visitors' gallery, the busts of the nation's first twenty vice presidents sit in solemn repose.

And in gentle steps, one hundred mahogany desks rise from the well of the Senate in four horseshoe-shaped rows. Some of these desks date back to 1819, and atop each desk is a tidy receptacle for inkwells and quills. Open the drawer of any desk, and you will find within the names of the senators who once used it—Taft and Long, Stennis and Kennedy—scratched or penned in the senator's own hand. Sometimes, standing there in



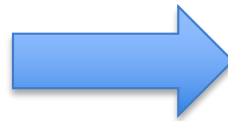
<https://goo.gl/images/vMaz4T>

Grammar for book is clear but not for genome



Genes to Functions

- Genes and elements



- Connections (“rules”)

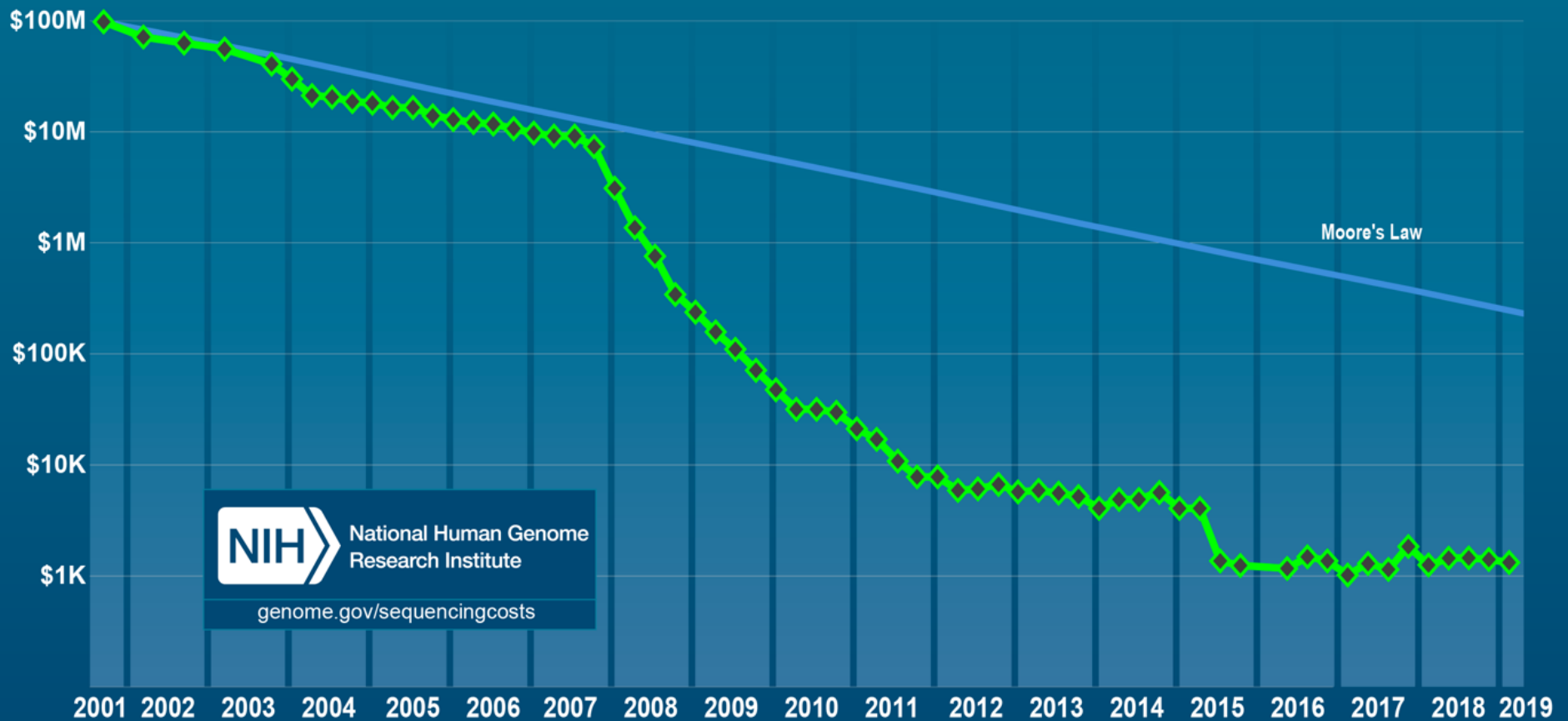


- Functions



Low sequencing cost enables reading our whole genome

Cost per Genome



After reading our genomes, we find differences: DNA mutations (i.e., genomic variants)

Single Nucleotide Polymorphisms (SNPs) normally happen ~1% on individual human genome.

Individual 1
 Chr 2 ...CGATA!
 copy1 ...GCTAT!
 Chr 2 ...CGATATTCCATCGAATG!
 copy2 ...GCTATAAGGGTAGCTTAC.

Individual 2
 Chr 2 ...CGATATTCCATCGAATG!
 copy1 ...GCTATAAGGGTAGCTTAC.
 Chr 2 ...CGATATTCCATCGAATG!
 copy2 ...GCTATAAGGGTAGCTTAC.

Individual 3
 Chr 2 ...CGATATTCCATCGAATG!
 copy1 ...GCTATAAGGATAGCTTAC.
 Chr 2 ...CGATATTCCATCGAATG!
 copy2 ...GCTATAAGGATAGCTTACAG...

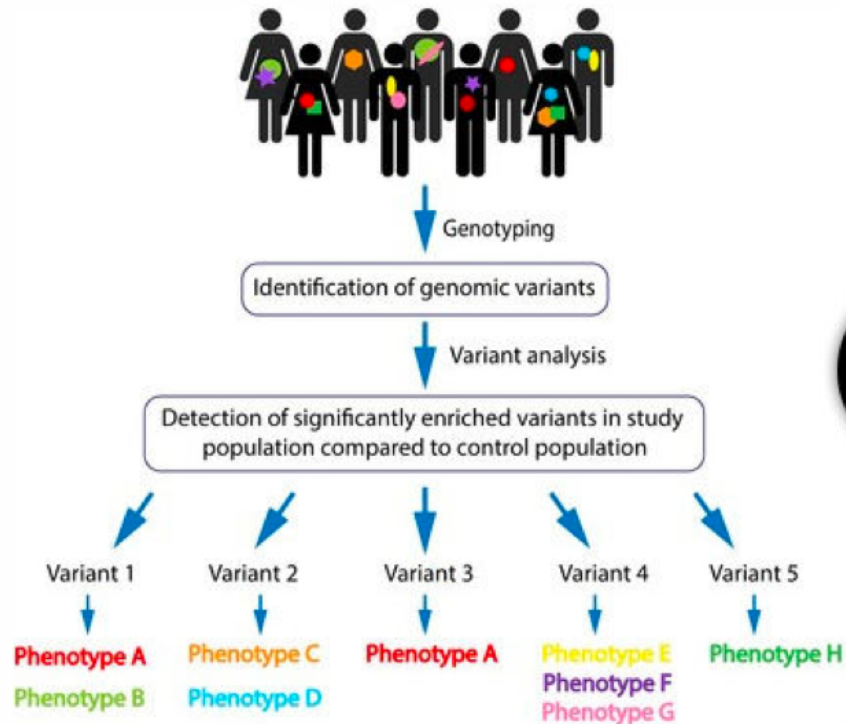
Most SNPs are harmless but some break “rules”



Single Nucleotide Polymorphism (SNPs)



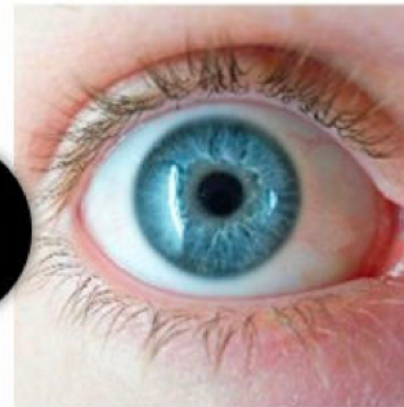
Genotype to Phenotype



VS

Phenotype= Blue Eyes

Phenotype=Brown Eyes



Genotype= bb
Recessive= b

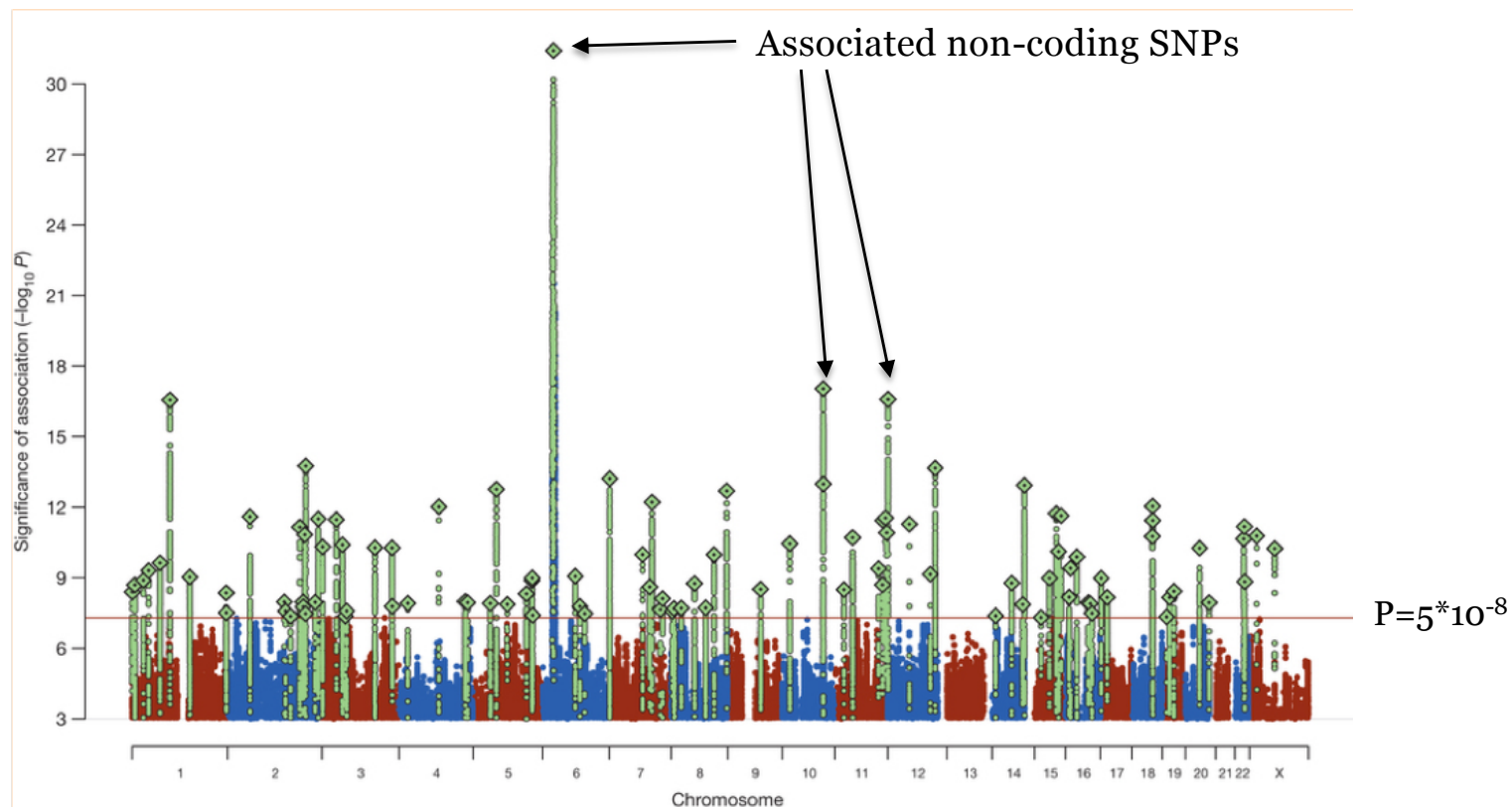
Genotype = Bb or BB
Dominant = B



Genotype vs. Phenotype

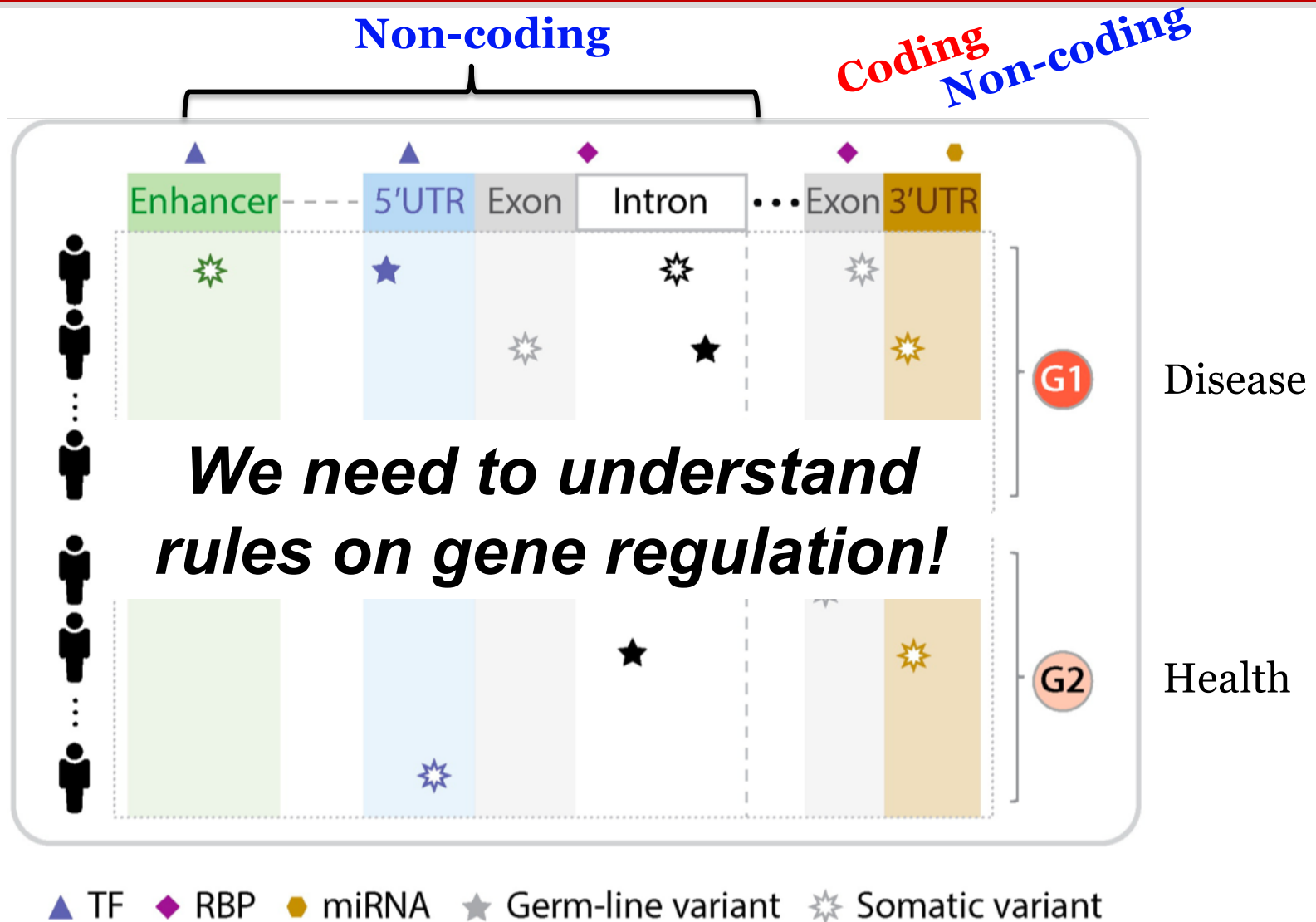
Example: Genome-Wide Association Study (GWAS) identifies disease associated noncoding variants

36,989 schizophrenia cases and 113,075 controls
in Psychiatric Genomics Consortium



However, association can't tell "rules" in genome

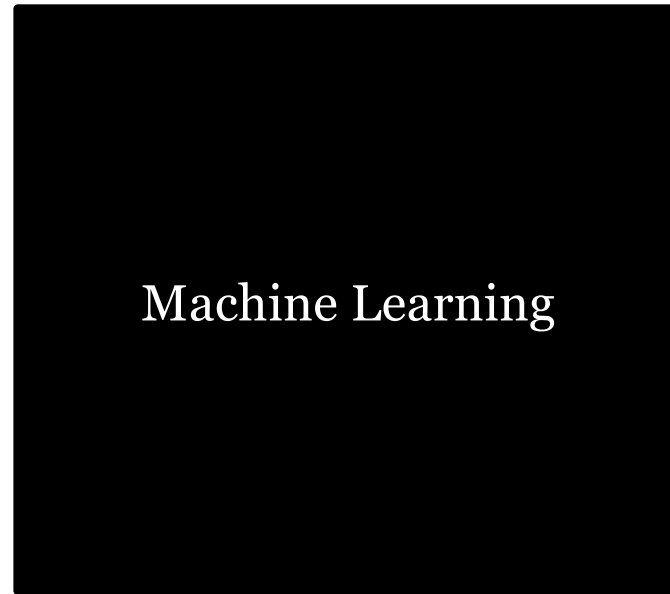
Genotype to Phenotype is a complex process



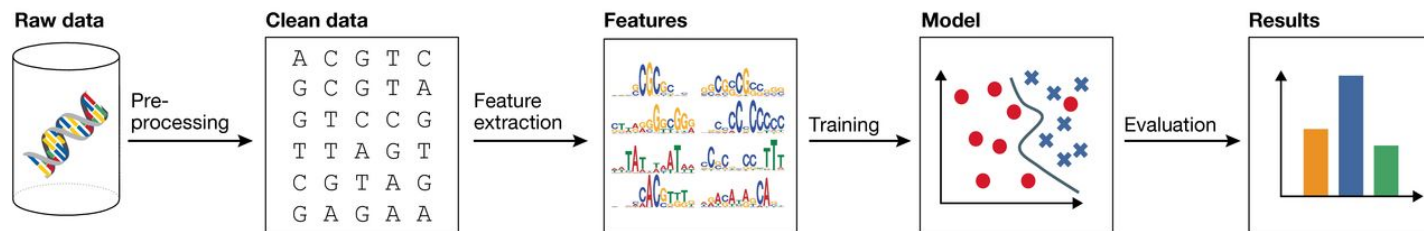
Machine learning deciphers “rules” for disease prediction



Big genomic data →

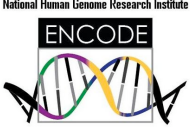


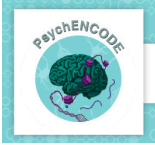
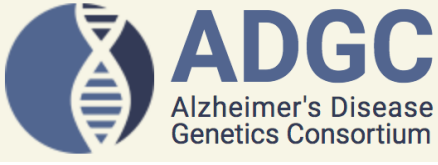


→ Genetic rules
→ Disease diagnosis
...



Machine Learning approaches

Big genomic data enable learning rules on gene regulation

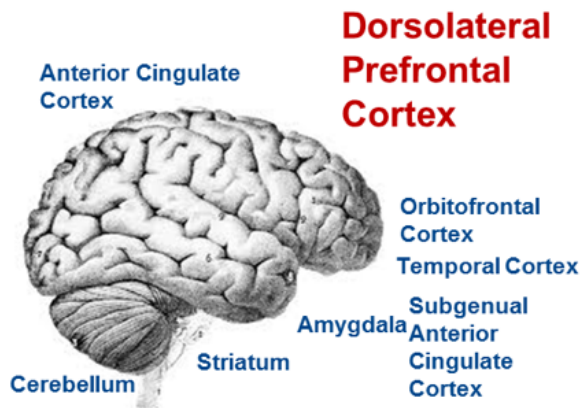
Human	20,000 genes (2% genome) Other genomic elements: non-coding RNAs, gene regulatory regions, repeats, and so on... (98% genome)	
Cell lines	ENCODE (Encyclopedia of DNA Elements) Consortium (> 300 cell types) 	
Tissues	 Genotype-Tissue Expression (GTEx) (> 40 tissues)	
Cancers	THE CANCER GENOME ATLAS National Cancer Institute National Human Genome Research Institute The Cancer Genome Atlas (TCGA) (> 40 cancer types)	
Development	 (13 developmental stages, 16 brain regions)	
Psychiatric disorders	 PsychENCODE Consortium (~2,000 tissues incl. health, Schizophrenia, Autism, Bipolar)	
Neurodegenerative diseases	 Religious Orders Study International Parkinson's and Memory and Aging Disease Genomics Project (ROSMAP) Consortium (IPDGC)	

Example: PsychENCODE (PEC) consortium

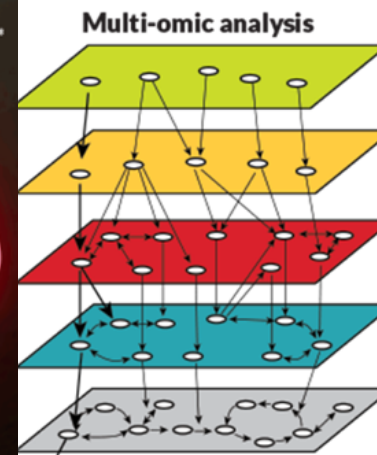


Sample Sources: >2,500 brains

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



Wang, et al., *Science*, 2018



Genome:
WGS, genotype

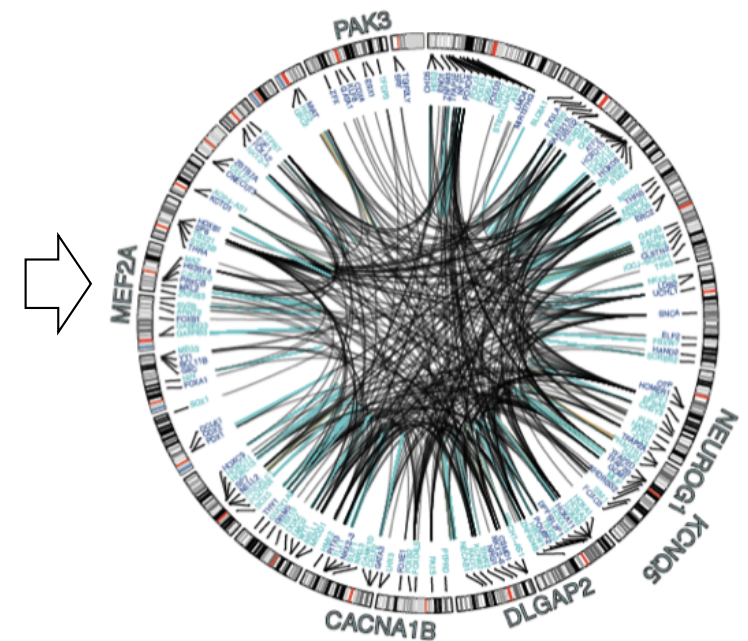
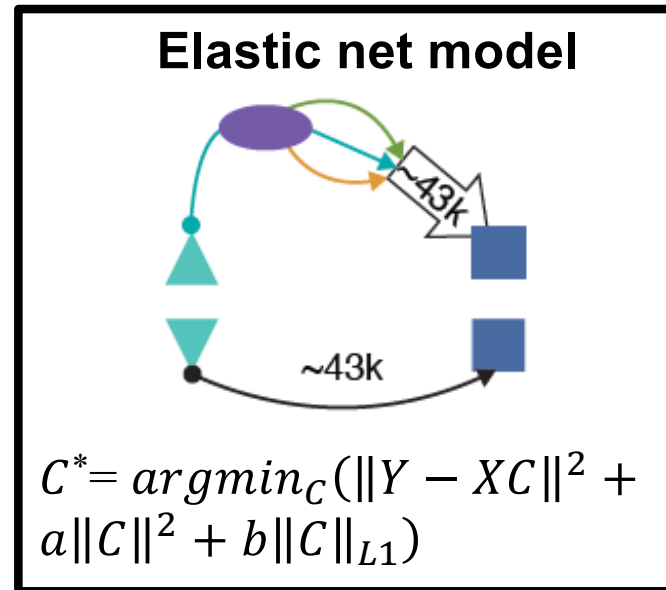
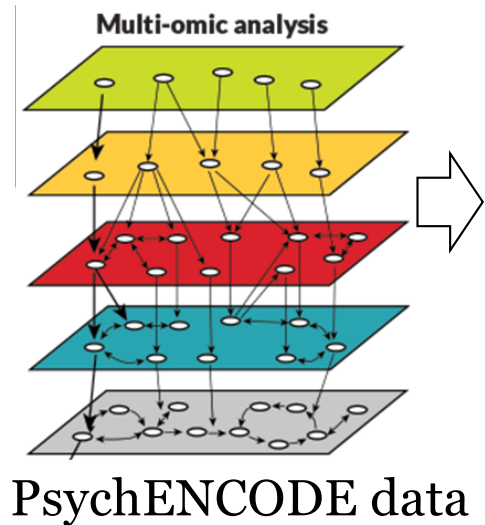
Epigenome:
ChIP-seq, ATAC-seq, HiC, ERRBS, Array Methylation, NOMeSeq

Transcriptome:
RNA-seq, lncRNAseq,

Proteome:
MWP, LC-MS/MS

Big genome data of human brain for the first time!

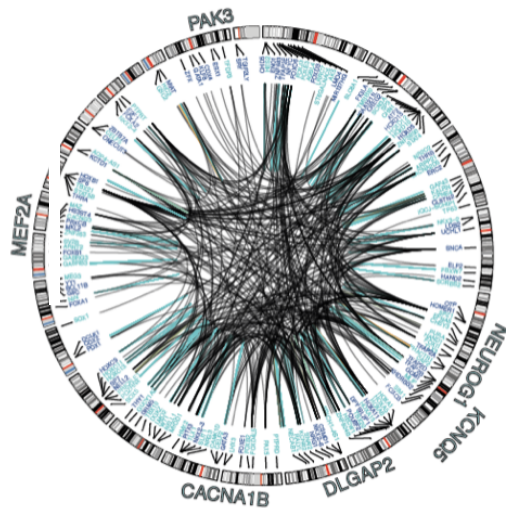
Identifying gene regulatory network in the human brain using PsychENCODE data



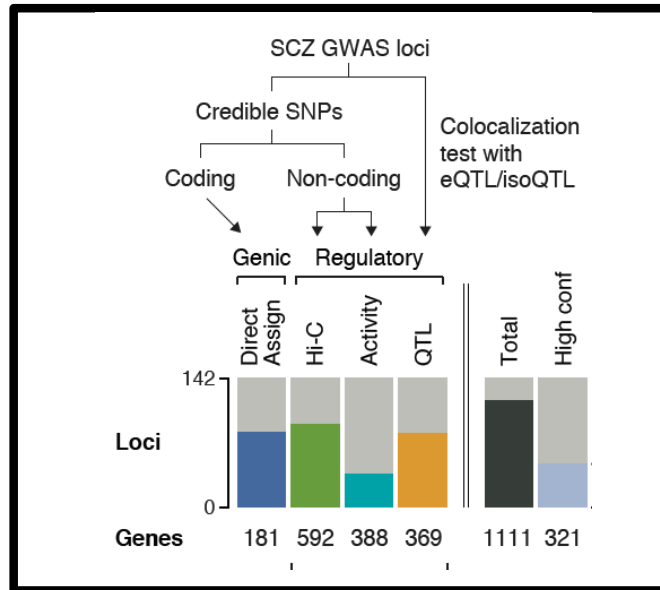
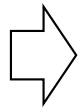
Gene regulatory network -
“rules” in human brain genome

Linking novel disease genes using learned “rules”

GWAS data



Gene regulatory network -
“rules” in human brain genome



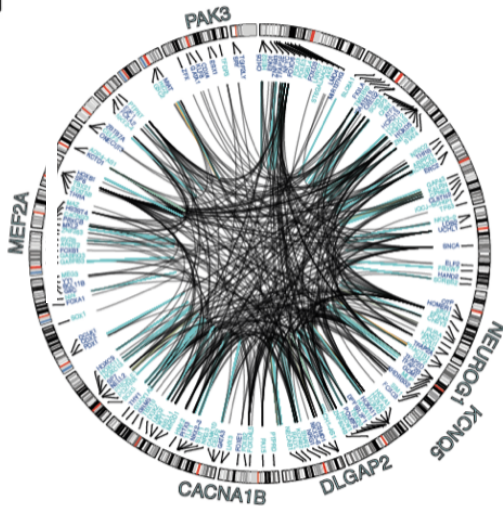
- BOLA1
- SV2A
- KAT5
- GFOD2
- FXR1
- SF3B4
- INPP4B
- PRIM1
- RNASEH2C
- AP5B1
- OVOL1
- CFL1
- SNX32
- MTMR11
- MUS81
- MAD1L1
- EFEMP2
- ...

321
 schizophrenia
 genes

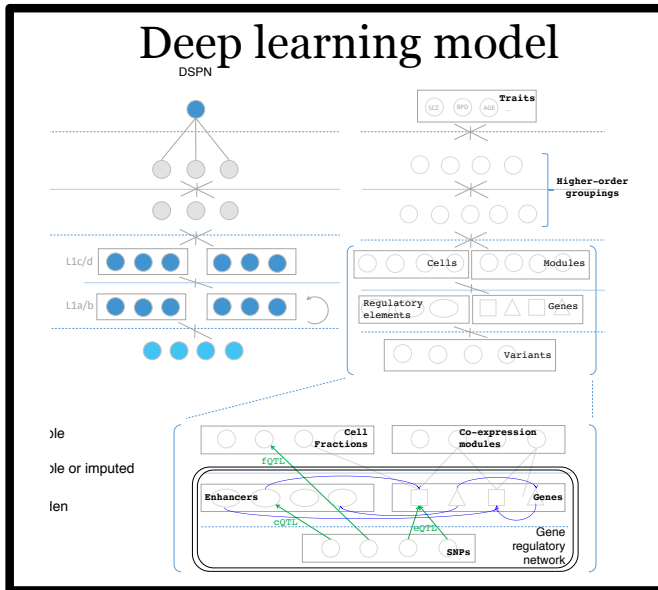
Improving brain disease prediction by applying learned “rules”



Population genotype data



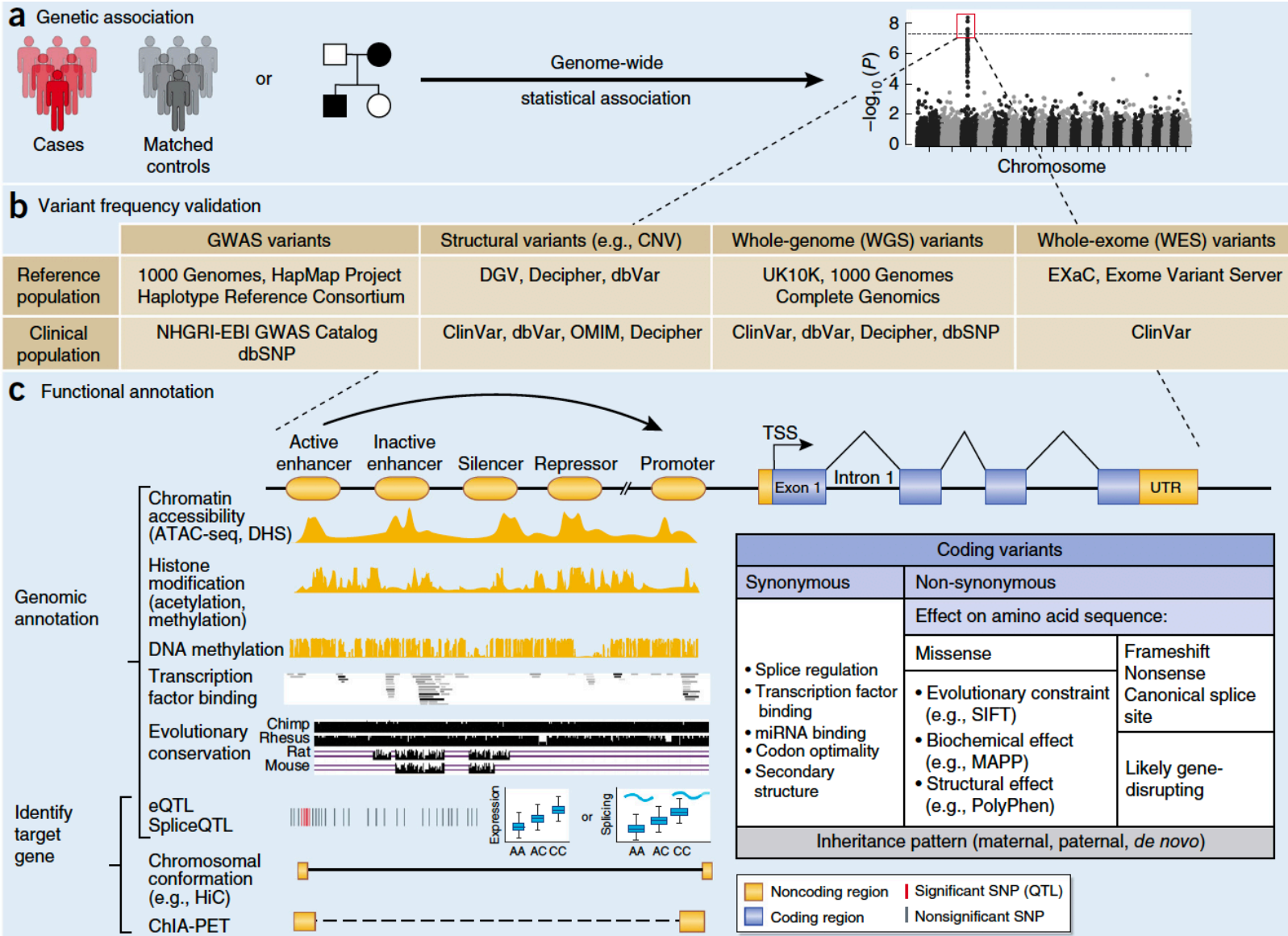
Gene regulatory network - “rules” in human brain genome



0.93
0.25
...
0.17

Schizophrenia diagnosis (probability)

The human genome is more complex than a book. Many unknown “rules” (i.e., biological mechanisms)!



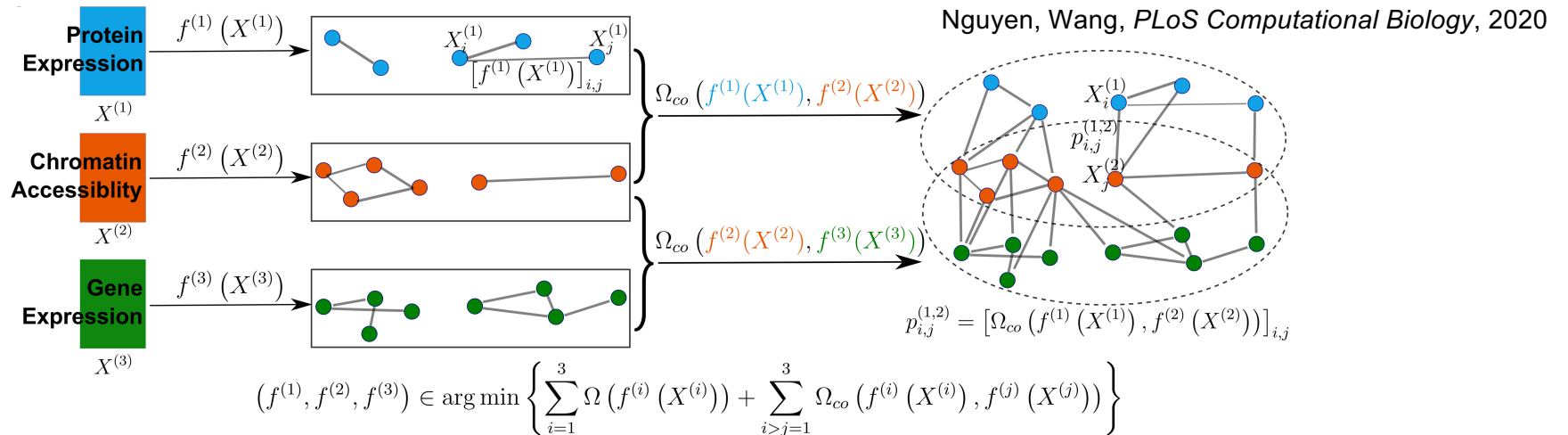
Disease-associated genomic variants



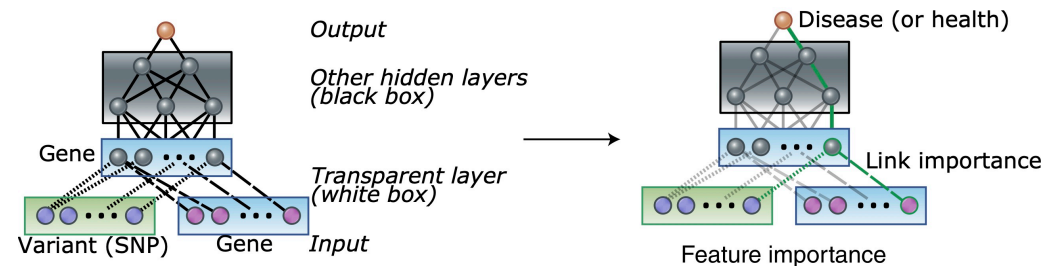
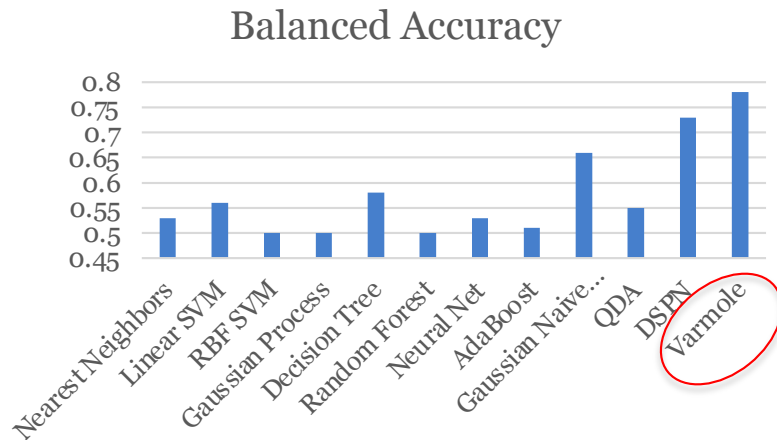
How do variants function?

Multi-view learning application in functional genomics

- A multi-view learning framework for understanding multi-omics



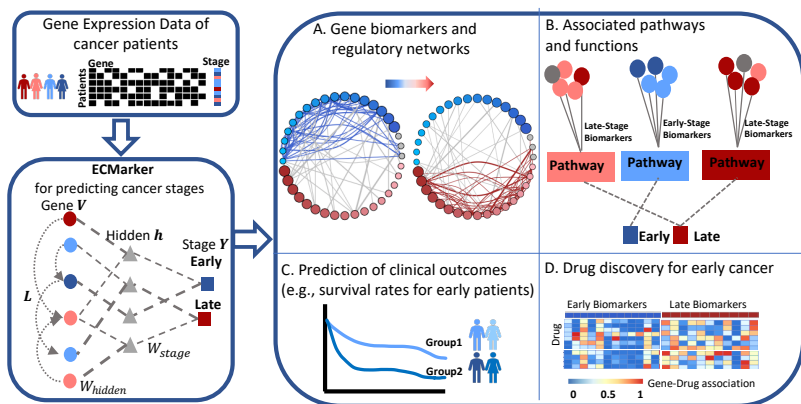
- Interpretable deep neural network model prioritizes disease variants and genes via drop-connect



Nguyen, Jin, Wang, *Bioinformatics*, 2020

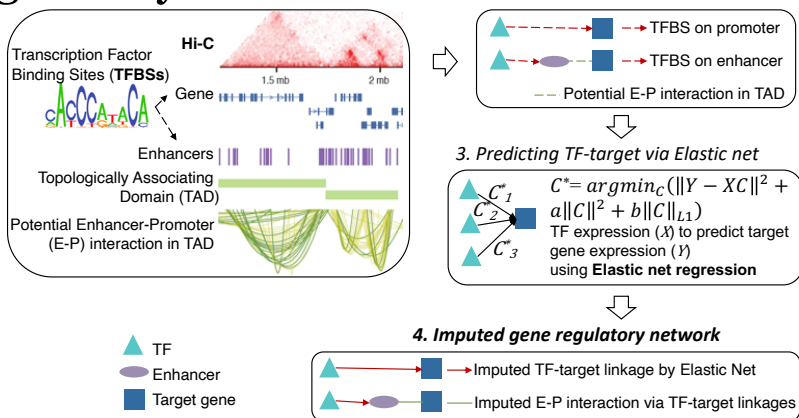
Select ongoing applications

- Genomic biomarkers in early disease stages (e.g., cancer, neurodegeneration)



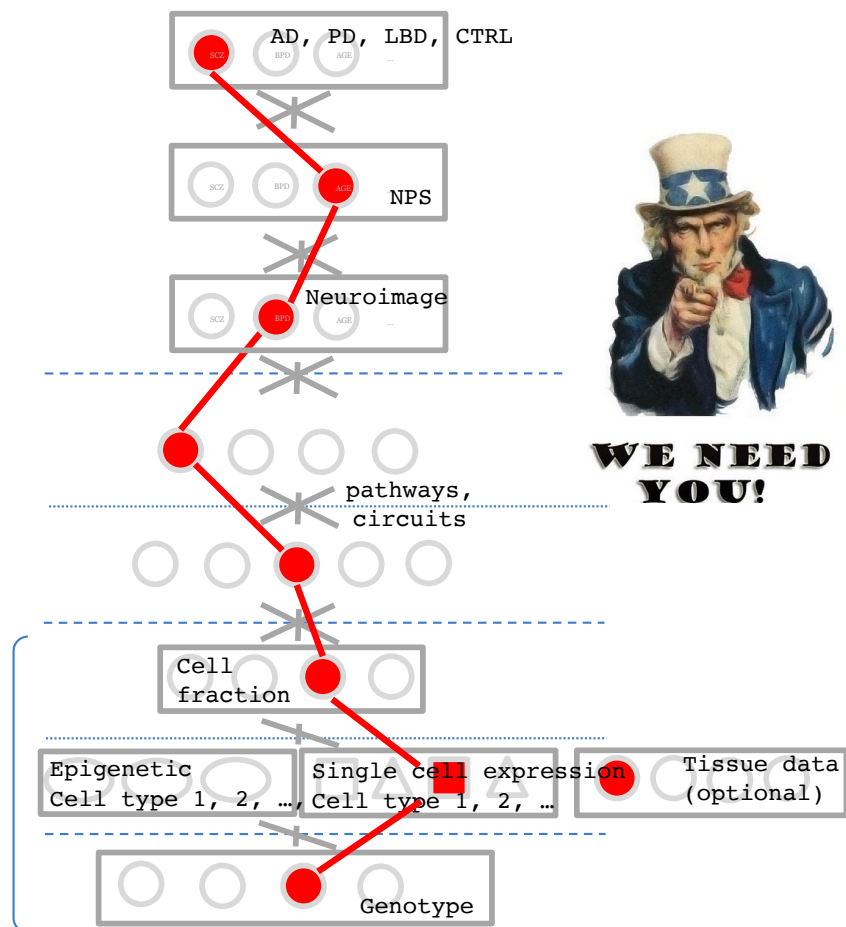
Jin, Nguyen, Talos, Wang, 2020

- Disease & Cell-type specific genes and regulatory networks



Ying, Rehani, Liu, Roussos, Wang, in revision

- Deep learning for deep phenotypes (e.g., symptom, imaging, cross-disease)



Thank you!

Ph.D. positions available
Please contact daifeng.wang@wisc.edu
Website: <https://daifengwanglab.org/>

