

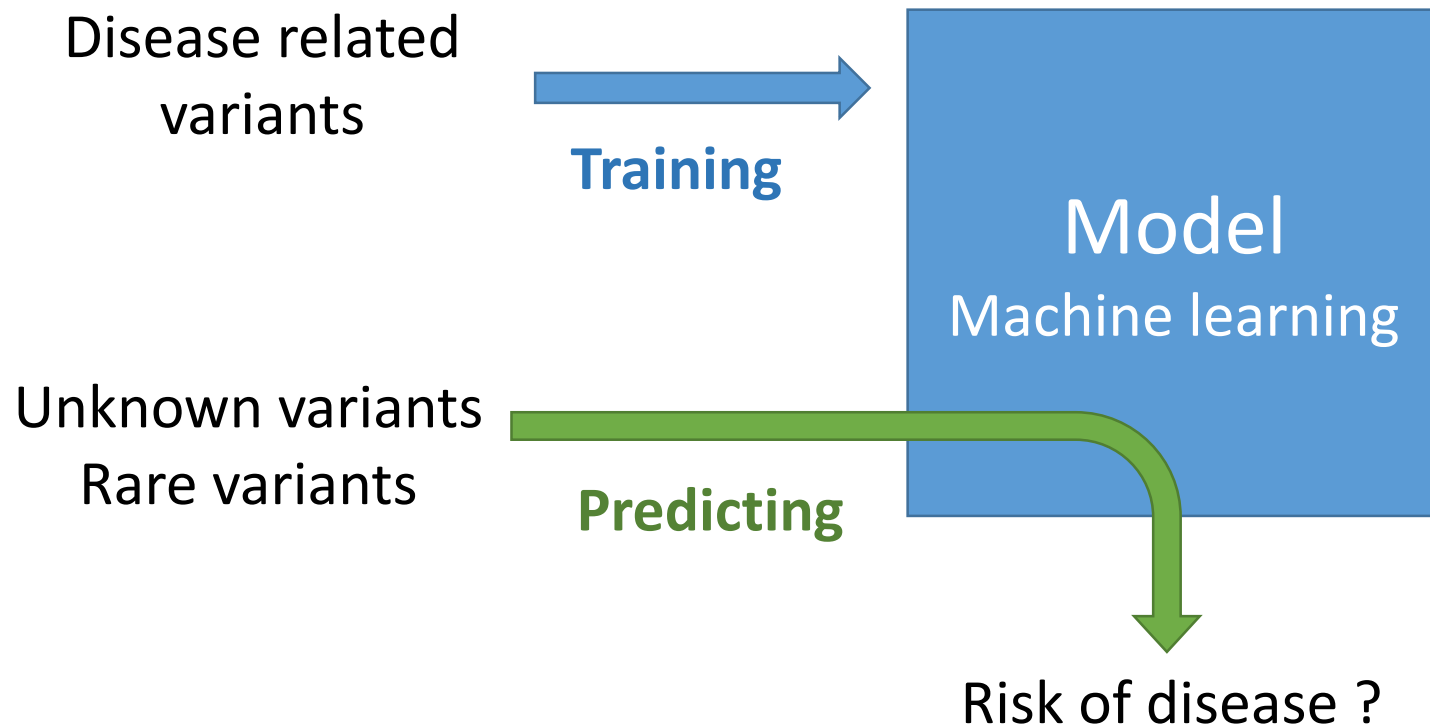
Deep learning model for genetic variants and omics data

Woojin Yang
PostDoc, KOBIC

August 31, 2018

The 16th KJC Bioinformatics Symposium, Hayama, Japan

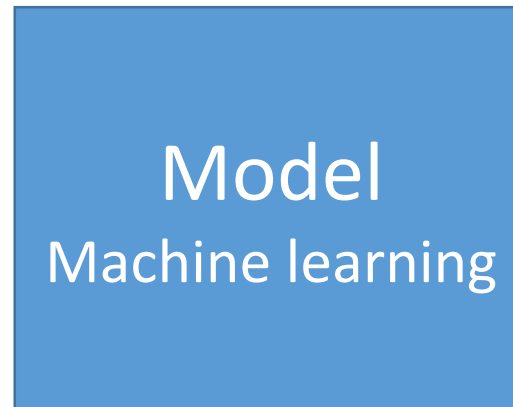
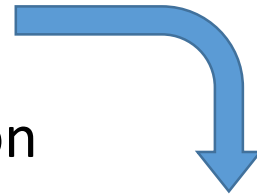
Goal: Disease = $f(\text{Variant})$



Machine learning for causal variants

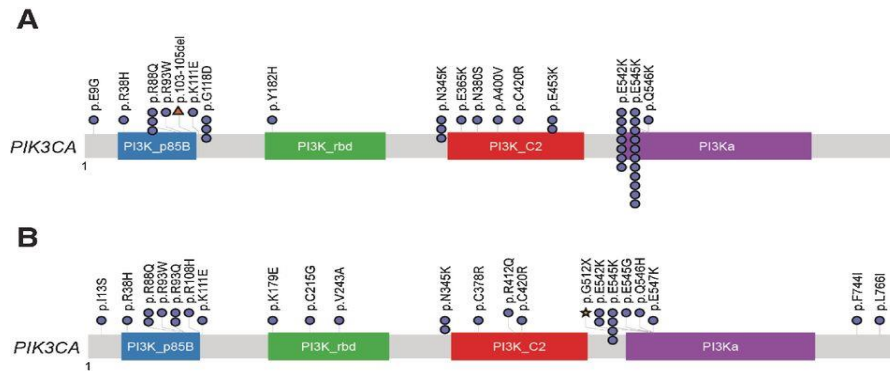
1. What to train
 1. True variants
 2. False variants
 3. Feature extraction
2. How to train
 1. Model
 2. Hyper-parameter optimization
3. Validation
 1. Cross-validation
 2. Biologically

Training



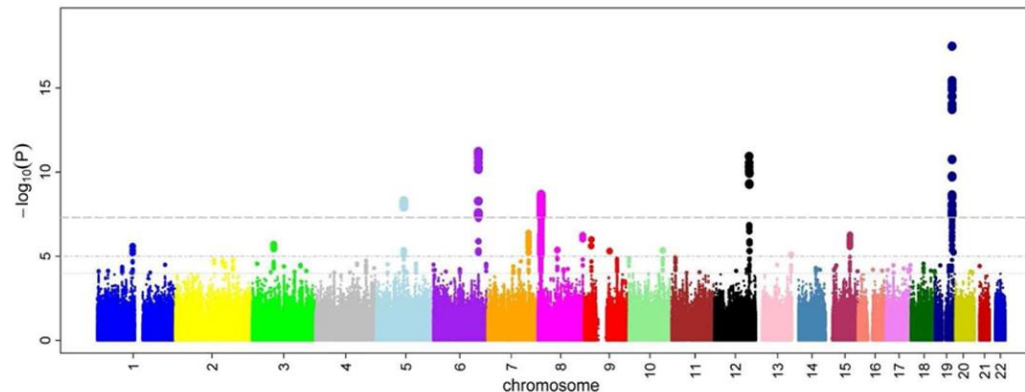
The true set: Disease related variants

Cancer recurrent mutation → cancer



LI, Xiangchun, et al. Distinct subtypes of gastric cancer defined by molecular characterization include novel mutational signatures with prognostic capability. *Cancer research*, 2016.

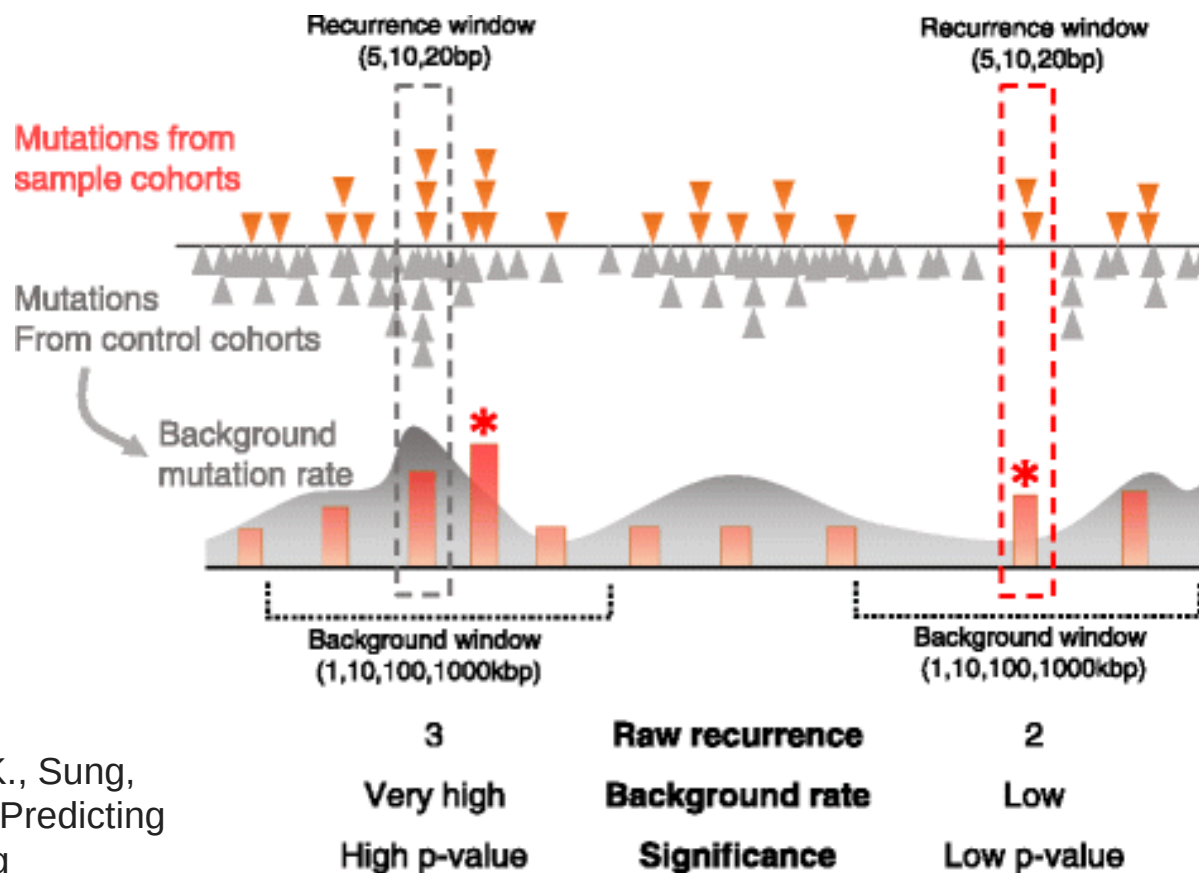
GWAS snps → common diseases



IKRAM, M. Kamran, et al. Four novel Loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo. *PLoS genetics*, 2010, 6.10: e1001184.

Recurrent mutations: Cause of cancer?

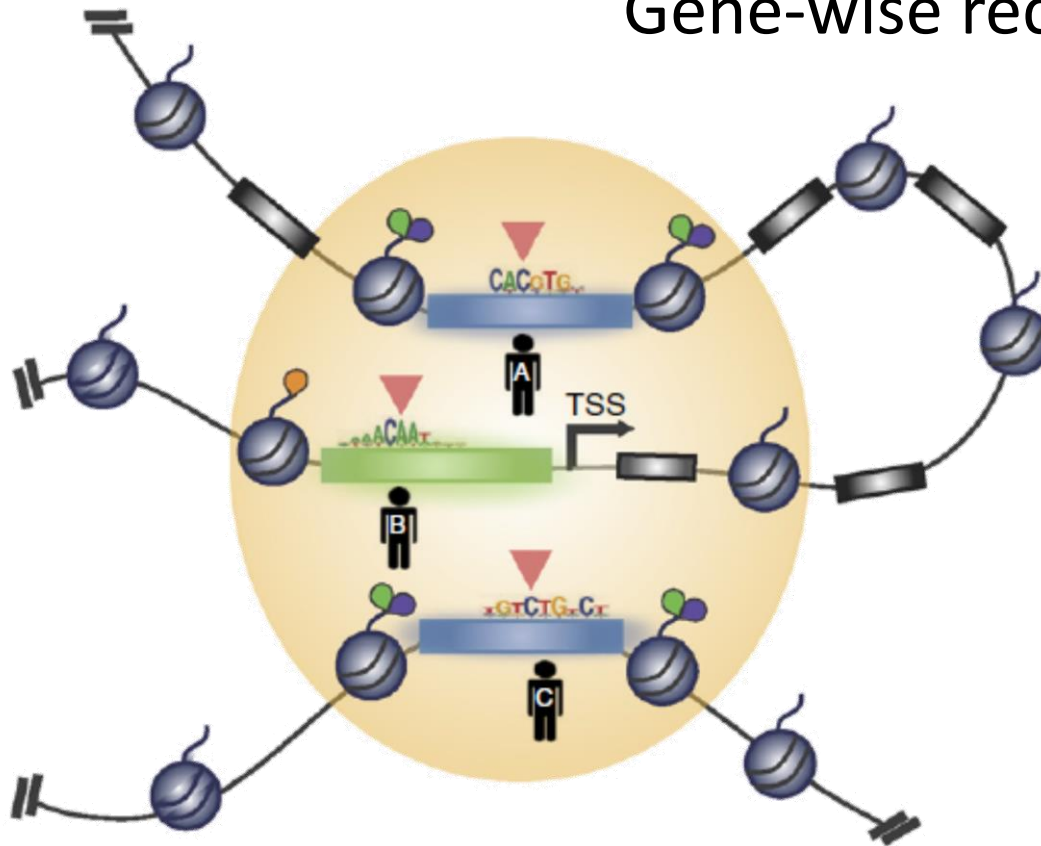
Site-specific recurrent mutations



Yang, W., Bang, H., Jang, K., Sung, M. K., & Choi, J. K. (2016). Predicting the recurrence of noncoding regulatory mutations in cancer. *BMC bioinformatics*, 17(1), 492.

Recurrent mutations: Cause of cancer?

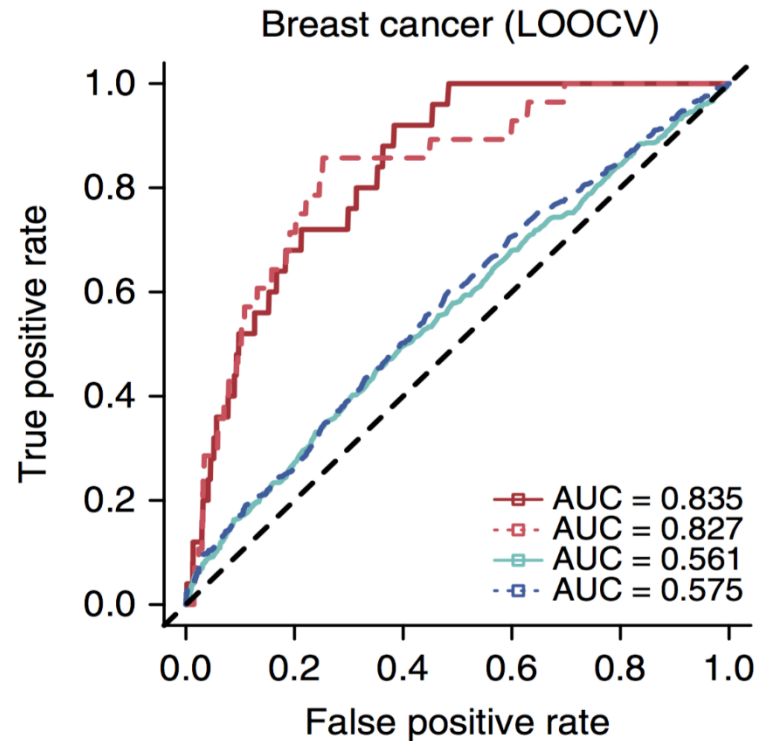
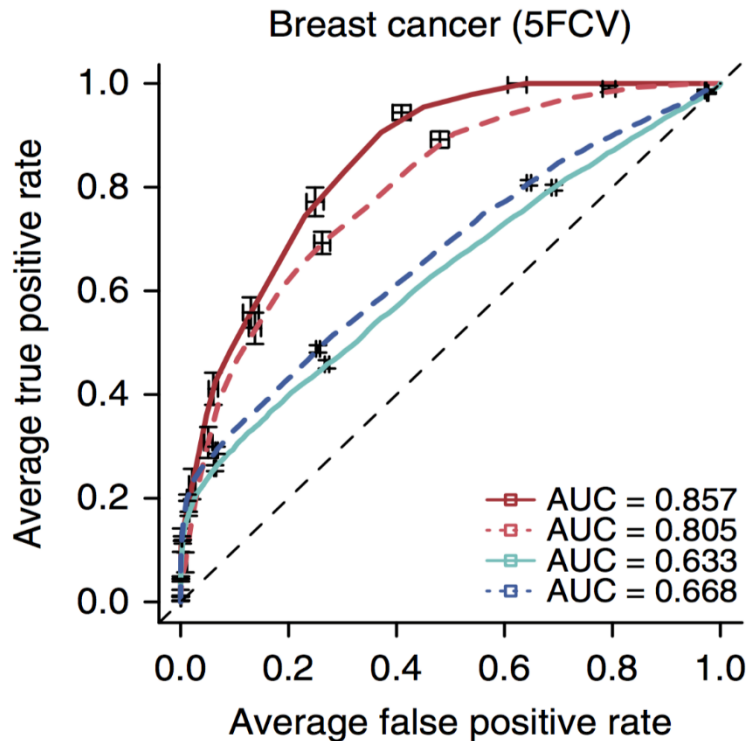
Gene-wise recurrent mutations



Kim, K., Jang, K., Yang, W., Choi, E. Y., Park, S. M., Bae, M., ... & Choi, J. K. (2016). Chromatin structure-based prediction of recurrent noncoding mutations in cancer. *Nature genetics*, 48(11), 1321.

Recurrent mutations may be causal

Performance of model

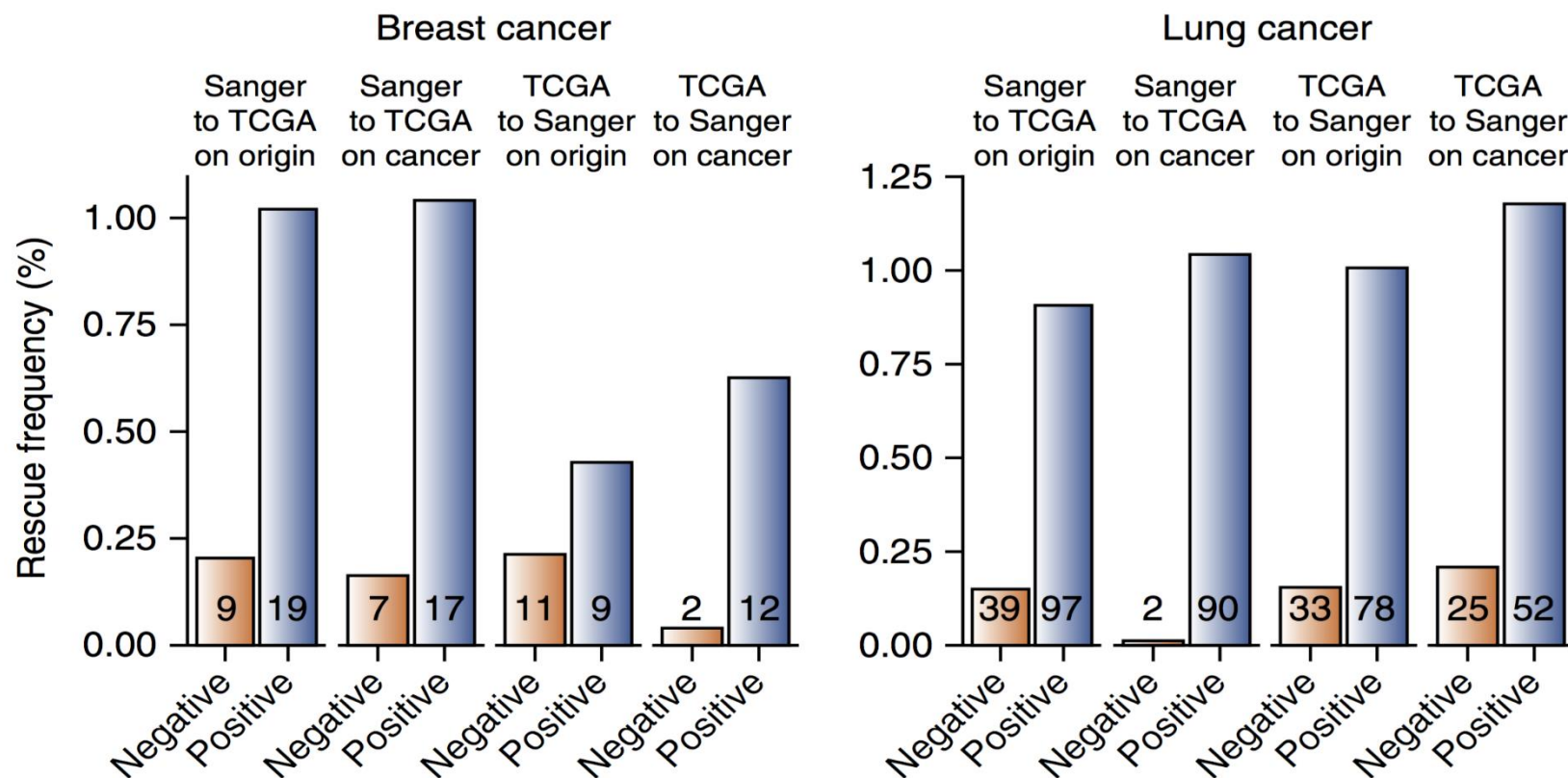


—■— Gene-level recurrence with cancer epigenome
- - ■ - Gene-level recurrence with cell-of-origin epigenome

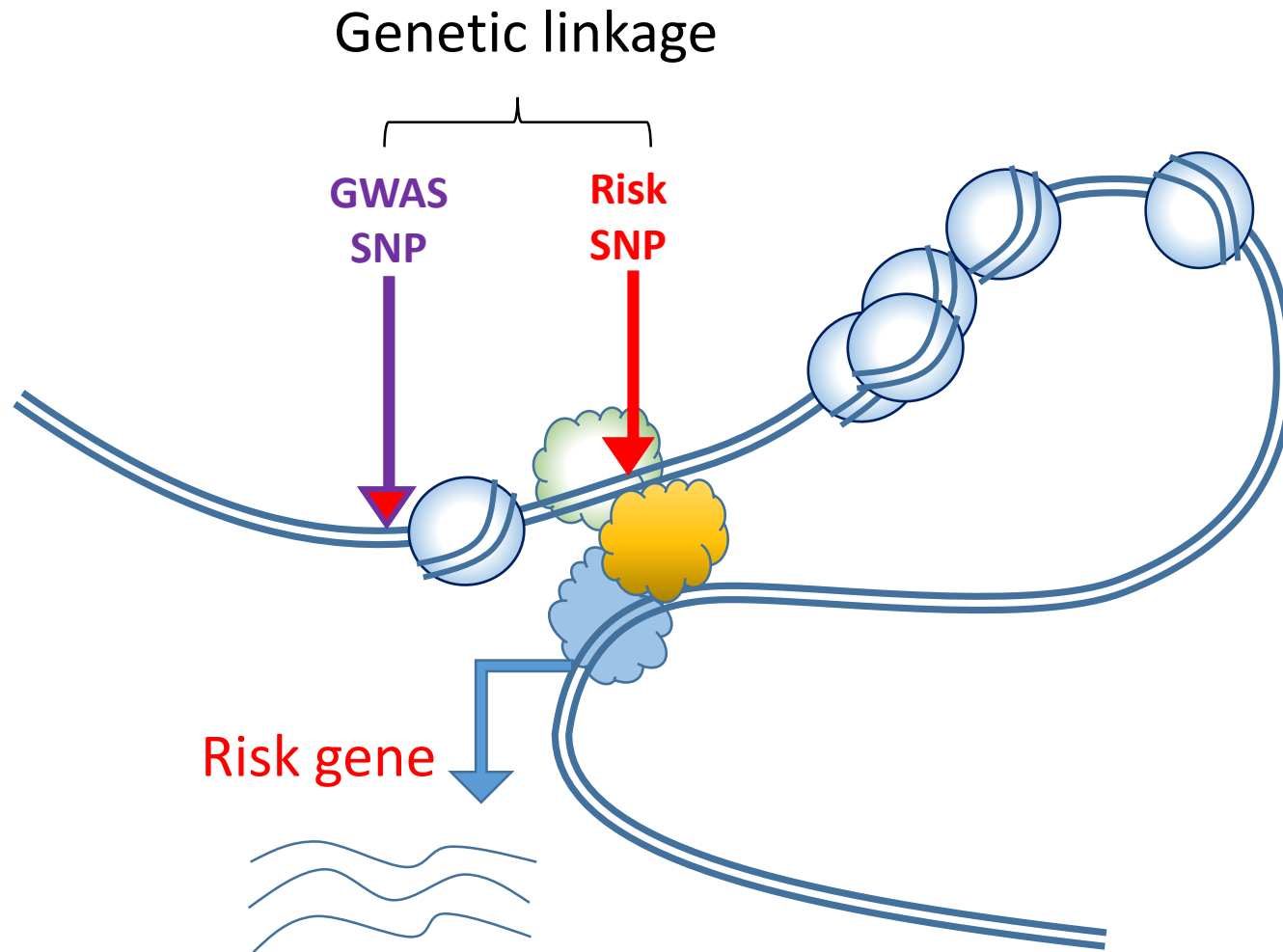
—■— Site-specific recurrence with cancer epigenome
- - ■ - Site-specific recurrence with cell-of-origin epigenome

Verification in new data

Prediction of model

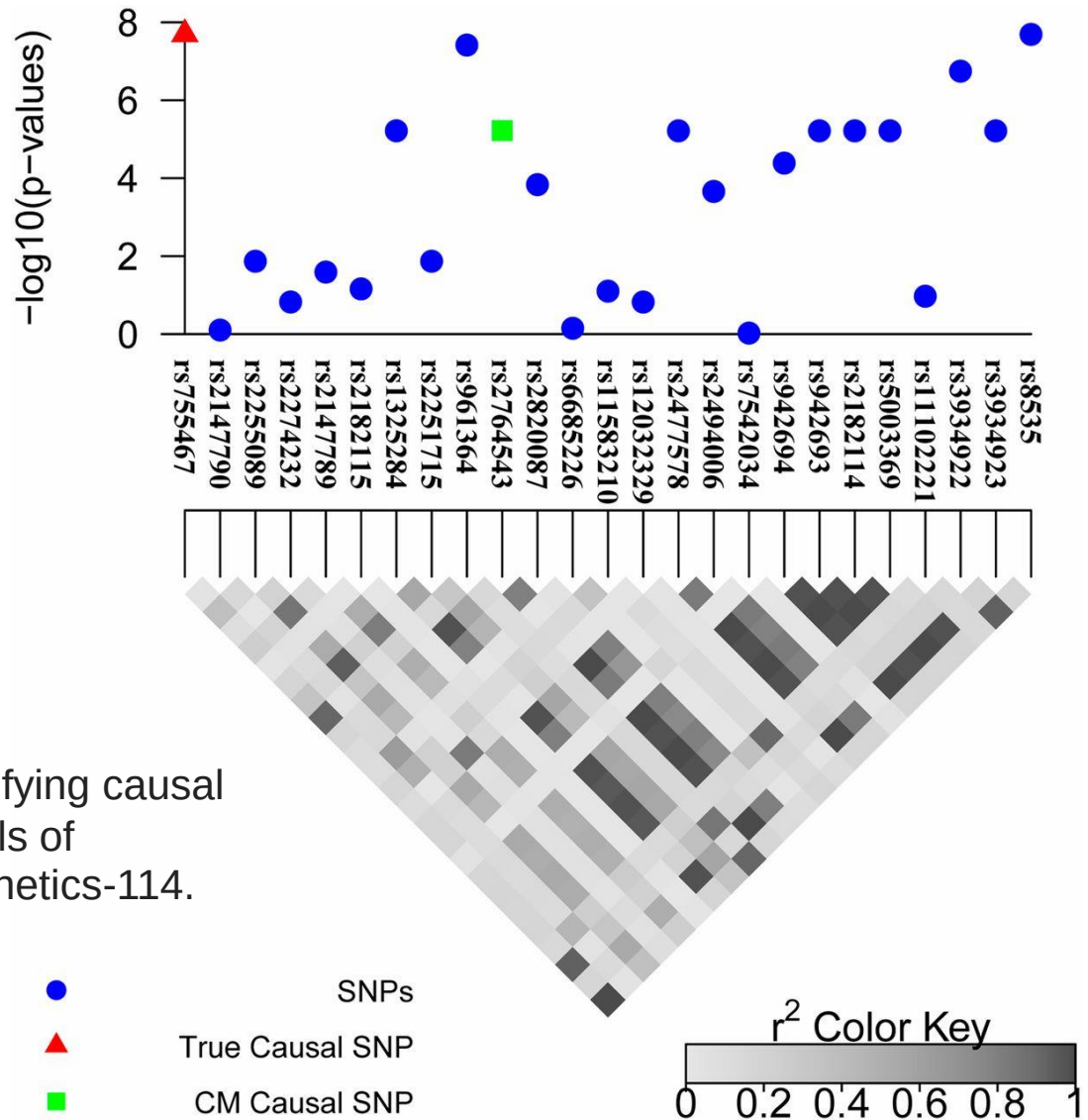


GWAS SNPs: Cause of disease?



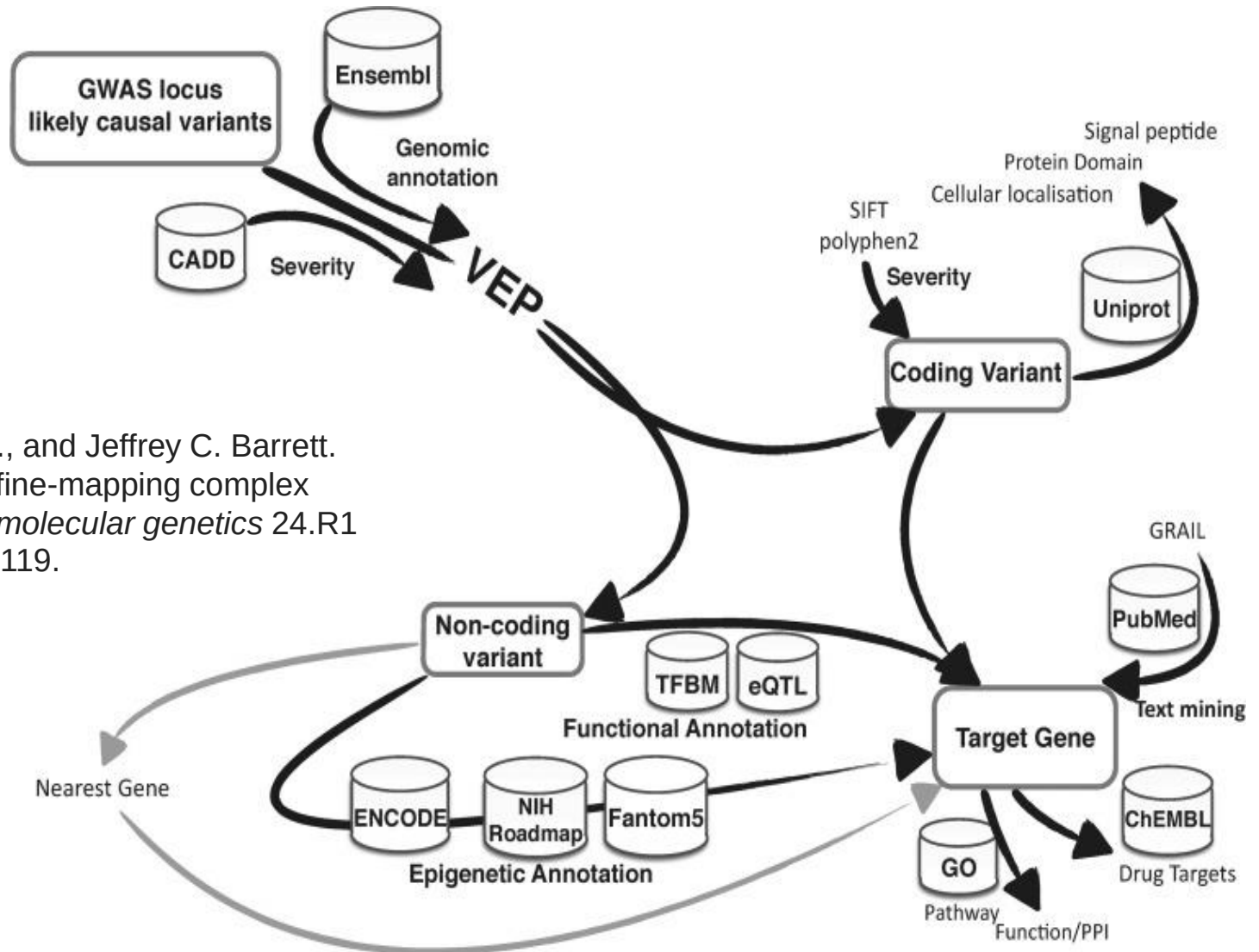
GWAS fine mapping

Statistical



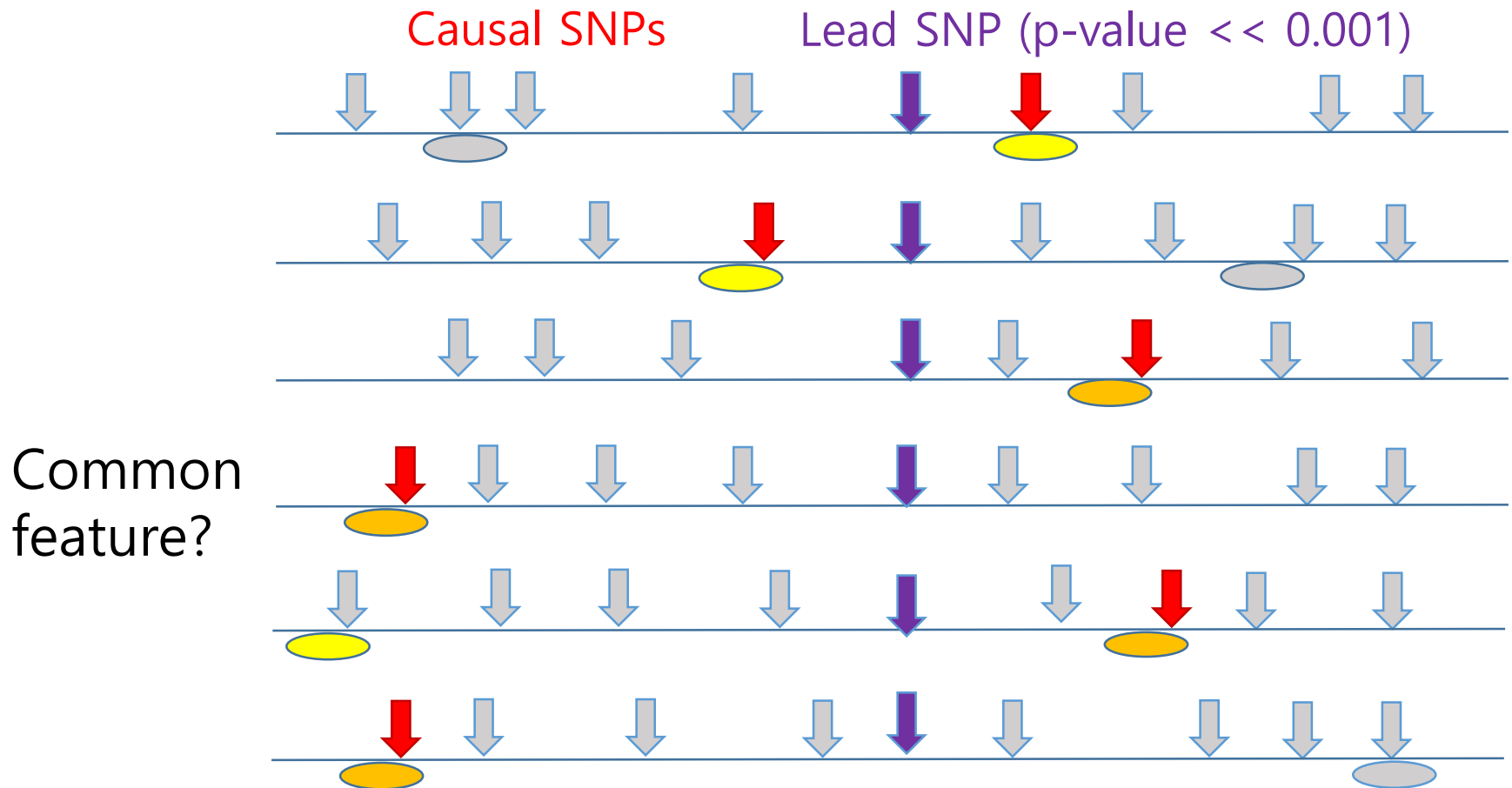
Hormozdiari, Farhad, et al. "Identifying causal variants at loci with multiple signals of association." *Genetics* (2014): genetics-114.

Functional fine mapping



Spain, Sarah L., and Jeffrey C. Barrett.
"Strategies for fine-mapping complex
traits." *Human molecular genetics* 24.R1
(2015): R111-R119.

Causal Variant Prediction by CNN-based Fine Mapping



What to train: The features

Disease related variants



Features

Gene	PROTEIN	PROTEIN	PROTEIN	PROTEIN	PROTEIN	PROTEIN	PROTEIN	PROTEIN	PROTEIN
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Training

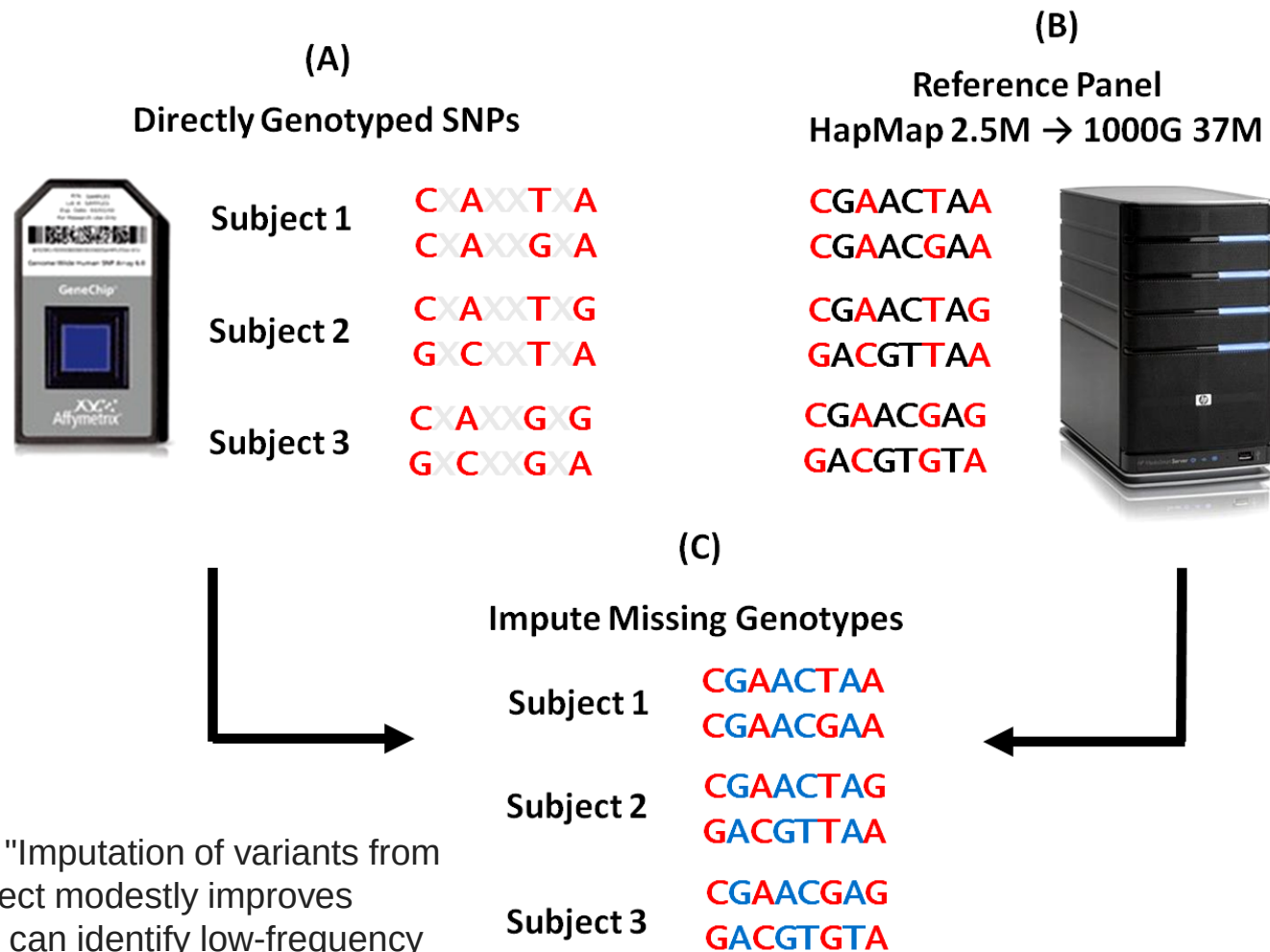


Model
Machine learning

SNPs to features

	Number of associations		
Disease	ADHD	ASD	BPD
Significant tag SNPs ($P < 5 \times 10^{-4}$)	642	943	1,424
Disease	MDD	SCZ	RA
Significant tag SNPs ($P < 5 \times 10^{-4}$)	832	601	435
Disease	SLE	CD	UC
Significant tag SNPs ($P < 5 \times 10^{-4}$)	849	431	383

Imputation with 1000 genomes



Wood, Andrew R., et al. "Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation." *PLoS One* 8.5 (2013): e64343.

LD proxy calculation

Home LDassoc LDhap LDmatrix LDpair **LDproxy** SNPchip SNPclip Help

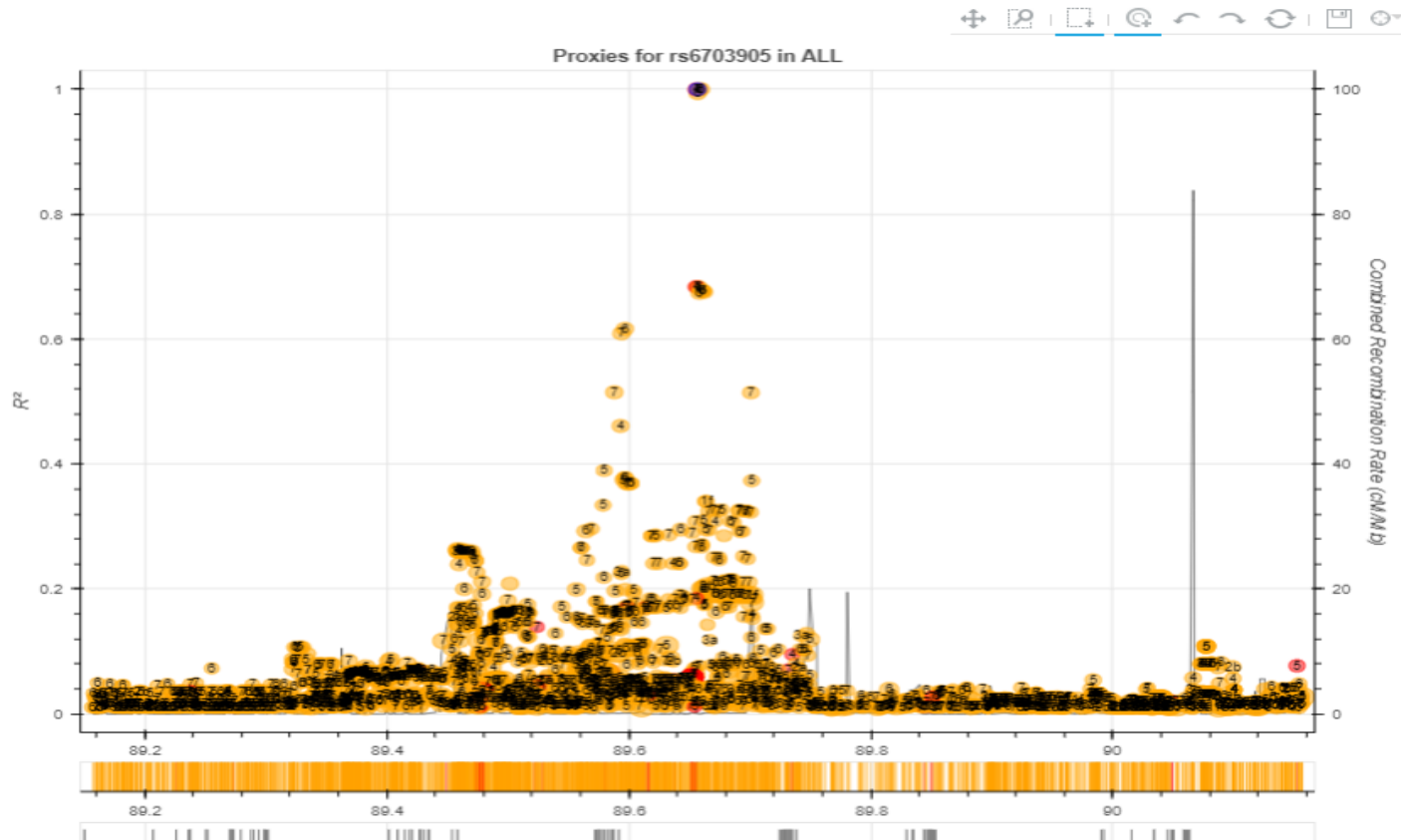
rs6703905

All Populations ▾

R^2

D'

Calculate



LD proxy calculation

Proxy Variants

Show entries

Search:

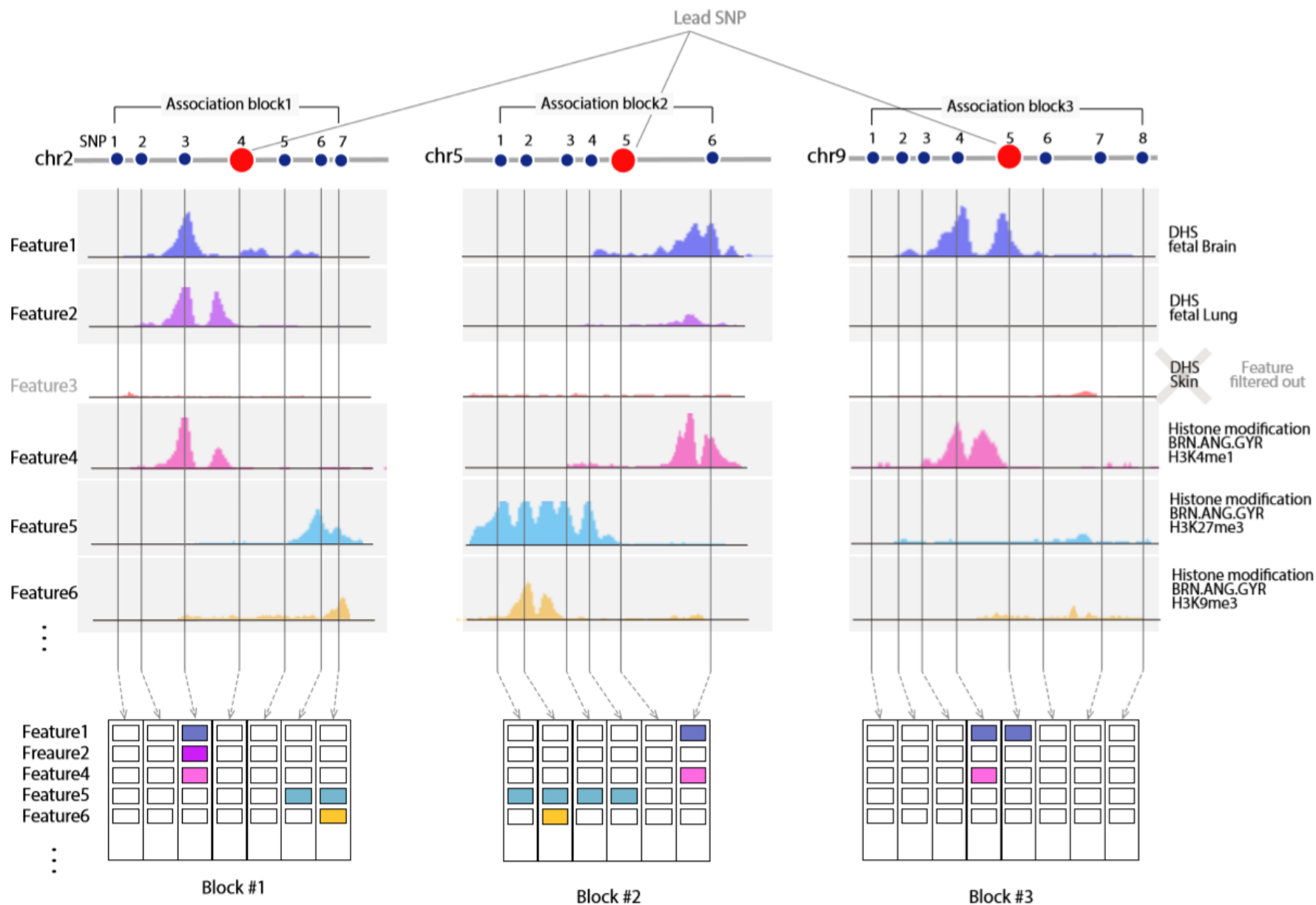
RS Number	Chr	Position (GRCh37)	Alleles	MAF	Distance	D'	R ²	Correlated Alleles	RegulomeDB	HaploReg	Functional Class
rs12077011	1	89659652	(G/A)	0.1072	2903	1.0	1.0	C=G,T=A	6		NA
rs12069631	1	89657754	(A/C)	0.1072	1005	1.0	1.0	C=A,T=C	6		NA
rs1410871	1	89655994	(C/T)	0.1072	755	1.0	1.0	C=C,T=T	4		synonymous
rs6667970	1	89656861	(A/G)	0.1078	112	1.0	0.9938	C=A,T=G	6		NA
rs1831240	1	89655783	(T/C)	0.0759	966	1.0	0.6836	C=T,T=C	4		missense
rs12084701	1	89657679	(C/A)	0.0759	930	1.0	0.6836	C=C,T=A	6		NA
rs1324333	1	89656534	(C/T)	0.0759	215	1.0	0.6836	C=C,T=T	7		NA
rs1324331	1	89661163	(A/G)	0.0761	4414	0.9941	0.6775	C=A,T=G	5		NA
rs1324332	1	89660833	(T/A)	0.0761	4084	0.9941	0.6775	C=T,T=A	6		NA
rs7537509	1	89660405	(T/A)	0.0761	3656	0.9941	0.6775	C=T,T=A	6		NA
RS Results	Chr	Position (GRCh37)	Alleles	MAF	Distance	D'	R ²	Correlated Alleles	RegulomeDB	HaploReg	Functional Class

Showing 1 to 10 of 3,033 entries

Previous 2 3 4 5 ... 304 Next

[Download all proxy variants](#)

The diagram illustrates the architecture of the DeepPrio model for prioritizing causal SNPs. It shows three parallel processing paths for different chromosomes: chr2, chr5, and chr9. Each path is divided into three association blocks (Block #1, Block #2, Block #3). The Lead SNP is identified at the top of each path. Below the association blocks, the model processes various genomic features (Feature1 to Feature6) represented by colored bars. The features are derived from various genomic annotations, including DHS (fetal Brain, fetal Lung, Skin), Histone modifications (BRN, ANG, GYR, H3K4me1, H3K27me3, H3K9me3), and other genomic features. The model uses a hierarchical structure where the Lead SNP is identified at the top, and the association blocks are processed in parallel. The final output is a prioritized list of SNPs across the three blocks.



Epigenetic data

ENCODE

Data

Encyclopedia

Materials & Methods

Help

Search...

Q

Organism

Homo sapiens10909

Mus musculus1864

Drosophila melanogaster1464

Caenorhabditis elegans1009

Drosophila pseudoobscura12

+ See more...

Biosample type

cell line5327

tissue2910

primary cell1505

in vitro differentiated cells630

stem cell400

+ See more...

Organ

blood2538

bodily fluid2538

liver1036

lung850

epithelium640

+ See more...

Project

ENCODE7727

Roadmap2761

GGR418

community3

BIOSAMPLE

10909 results

Clear Filters

cell line

K562	697	10	276	19	191	12	3	12	1	8	1	1	2	1	6	9	77	2	9
HepG2	370	3	257	11	164	7	3	6	2	3	2		2		6	6		1	2
A549	384	14		27		2	2		1	9	1		2		2	5	3		
GM12878	249	3		14		8	3	4	2	6	2	1	2	2	6			2	2
HEK293	271					1	2		2				2						

+ See 166 more...

tissue

stomach	57	20		10			3	5	1	4	6	5	1		2				
adrenal gland	46	11		8	4	2	6	4	1	2	5	4	1	2	3				
liver	73	2		5			1	2	1	1	3	2	3					3	
transverse colon	41	4					4	4		4		4	4		4				
sigmoid colon	54	4		4			4	4		4	2	4			4				

+ See 130 more...

primary cell

foreskin keratinocyte	37	2		6				3			1	13	1	3	13			3	
common myeloid progenitor, CD34-positive	44	12		1		12			8		1								
endothelial cell of umbilical vein	36	2		5		2	1			1			1		6		5		
neurosphere	34			8							4				8			4	
mammary epithelial cell	34	3		3		2	3	1	1	1	2		3	2			1		1

ASSAY

ChIP-seq

DNase-seq

siRNA RNA-seq

polyA RNA-seq

eCLIP

RNA microarray

DNAse array

total RNA-seq

RRBS

small RNA-seq

WGBS

RAMPAGE

genotyping array

microRNA-seq

Repli-seq

ATAC-seq

CAGE

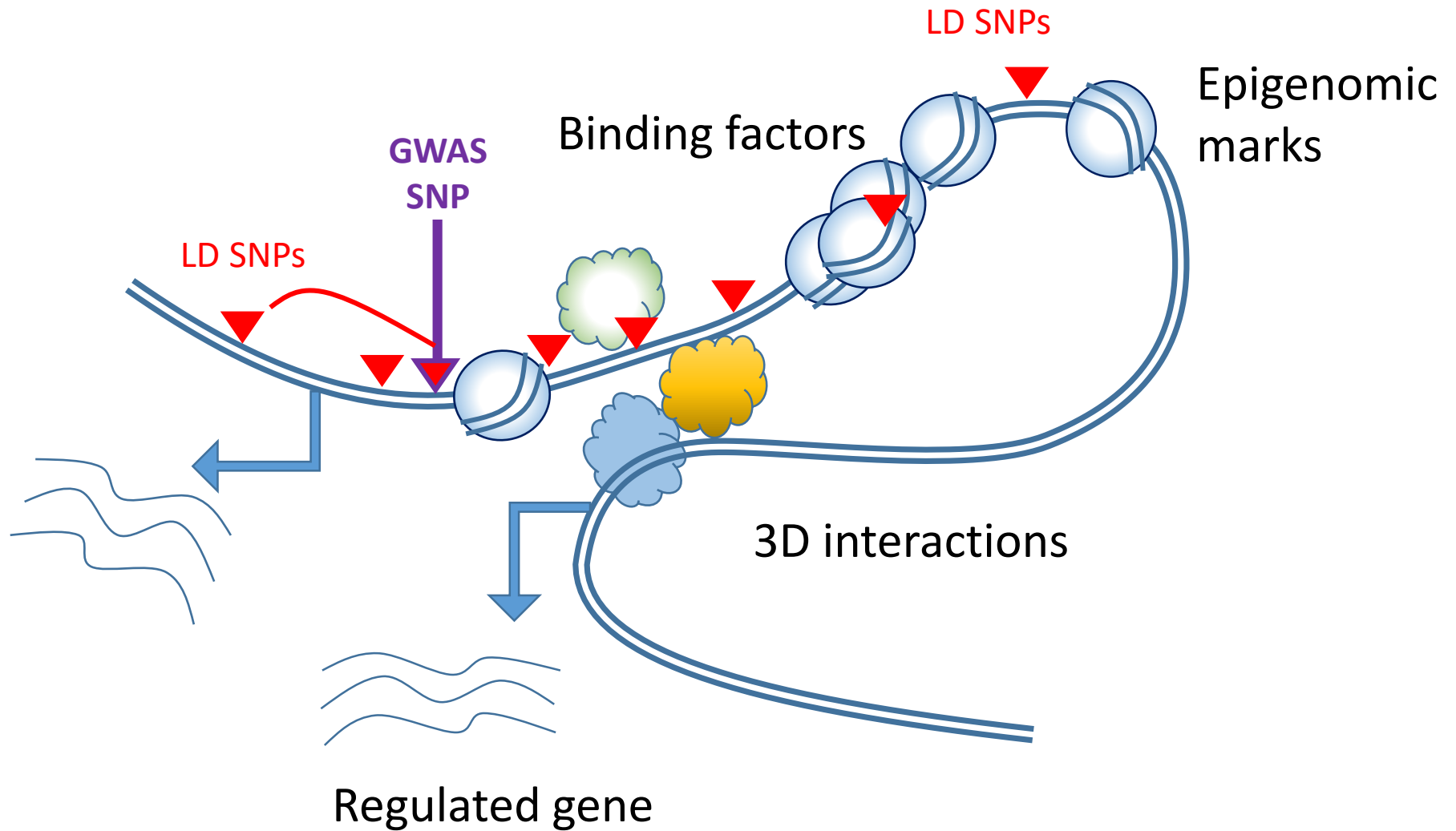
CRISPRi RNA-seq

MRE-seq

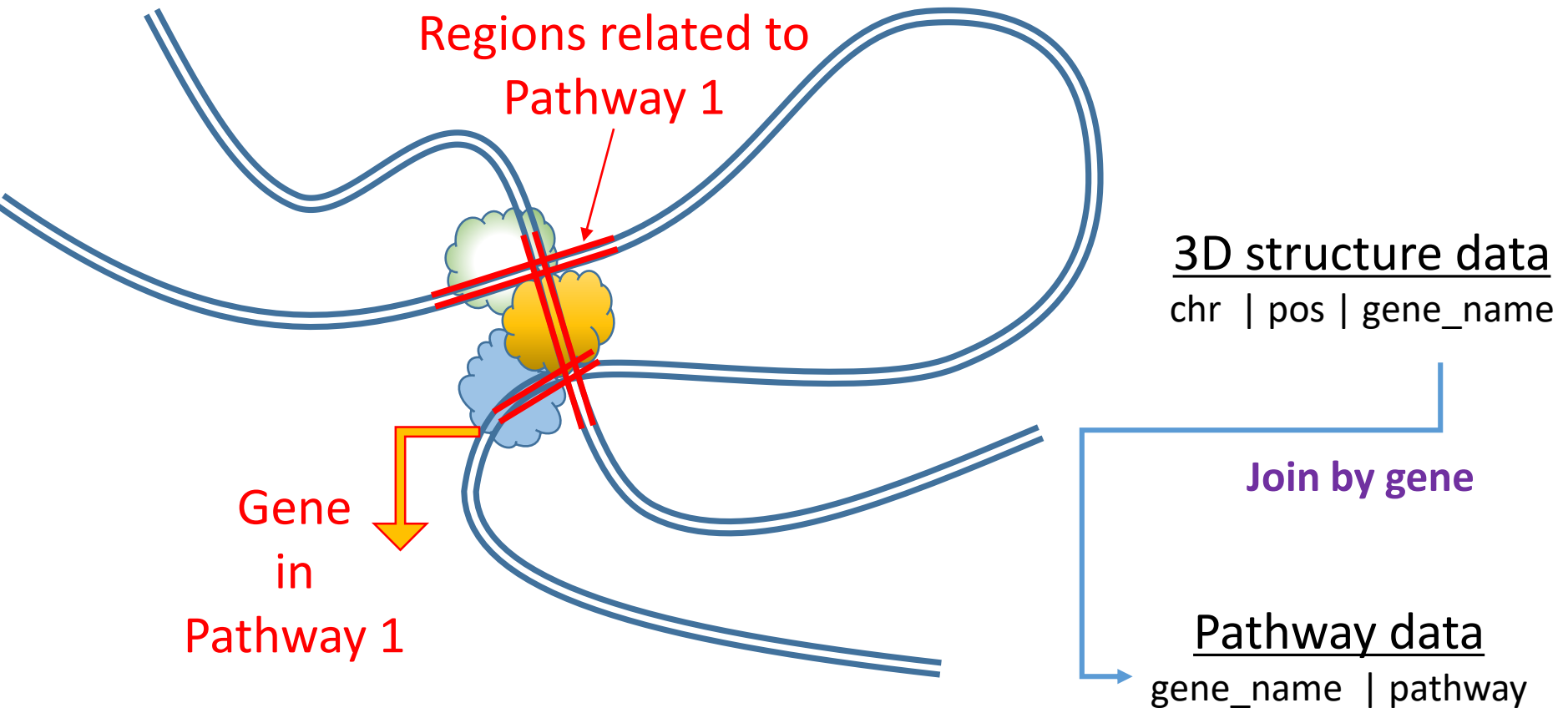
ChIA-PET

...and 23 more

Regulatory effect of variants

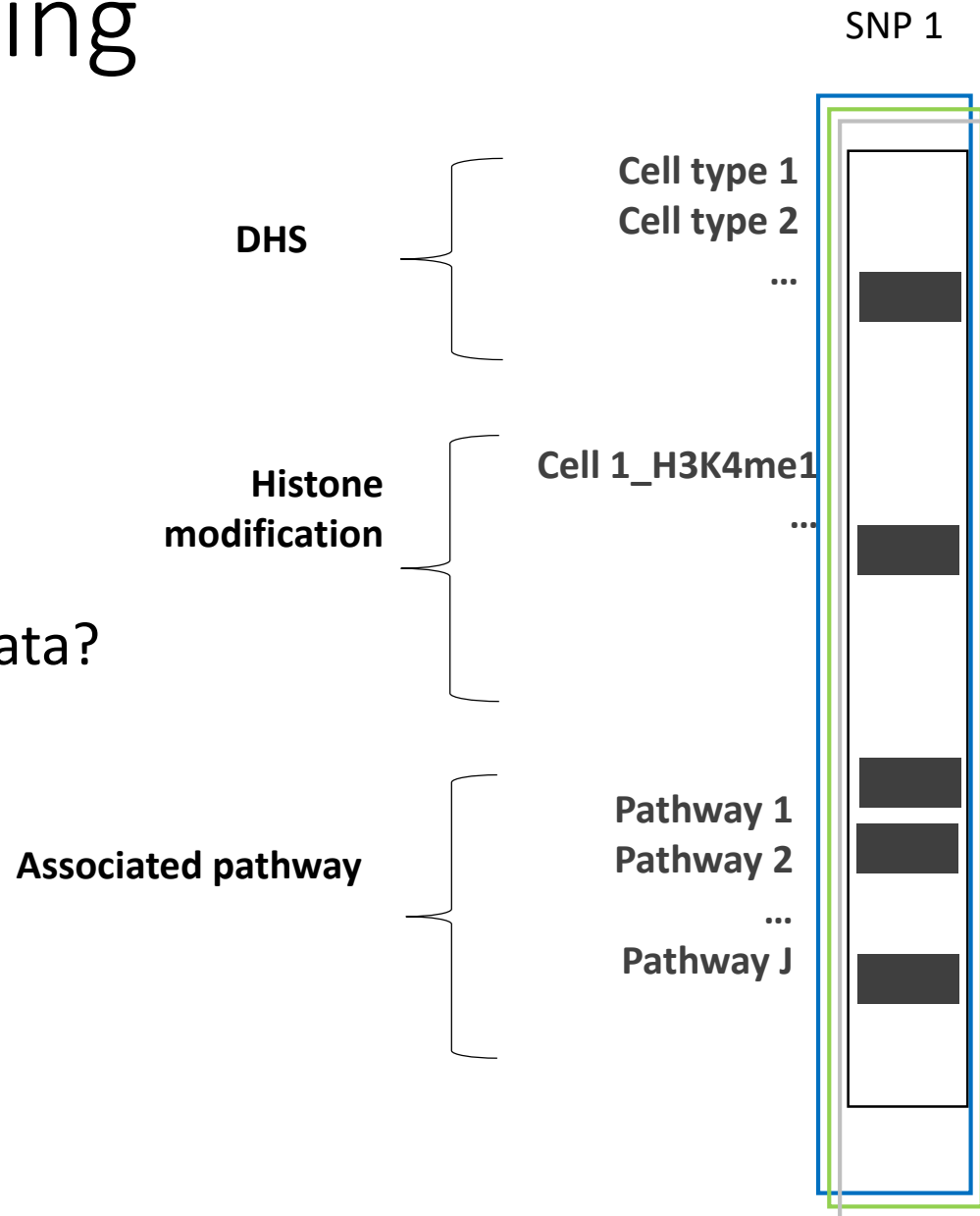


Pathway data

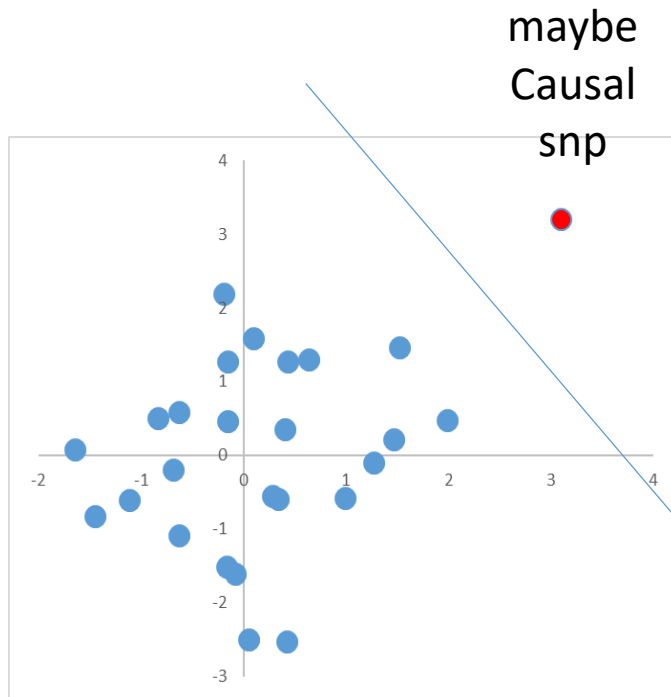


Data merging

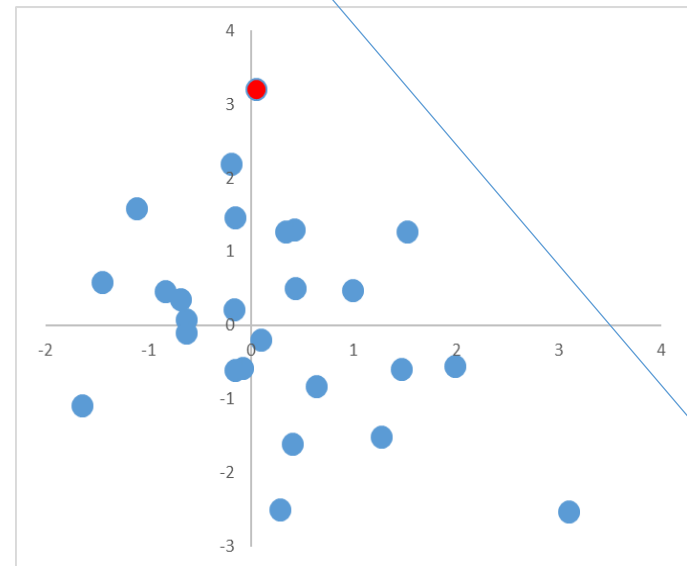
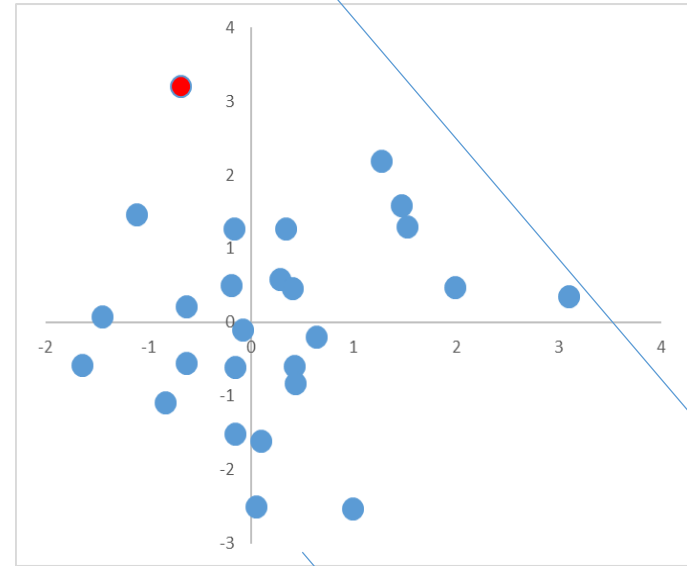
Binding factor data?



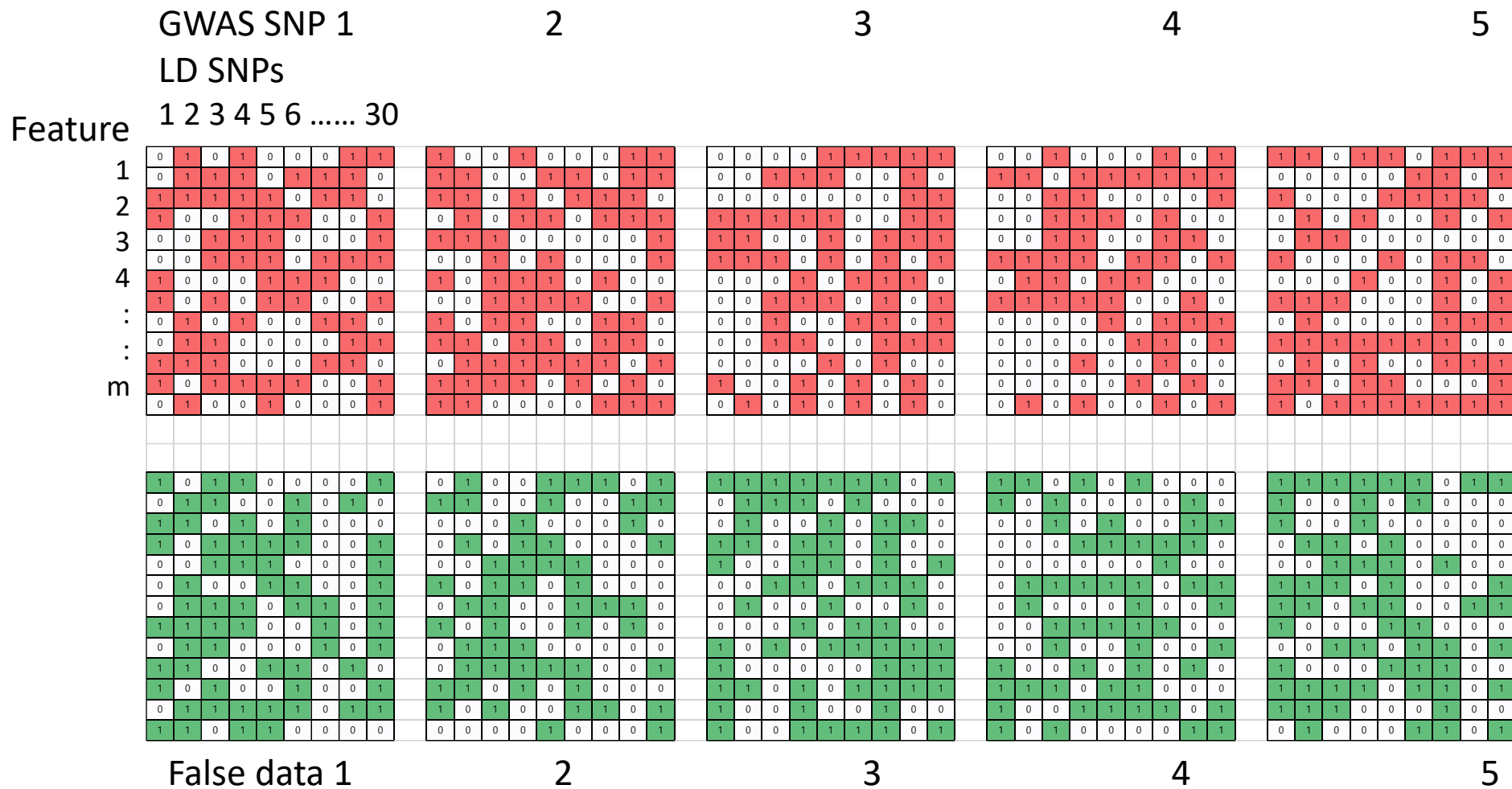
Make false set



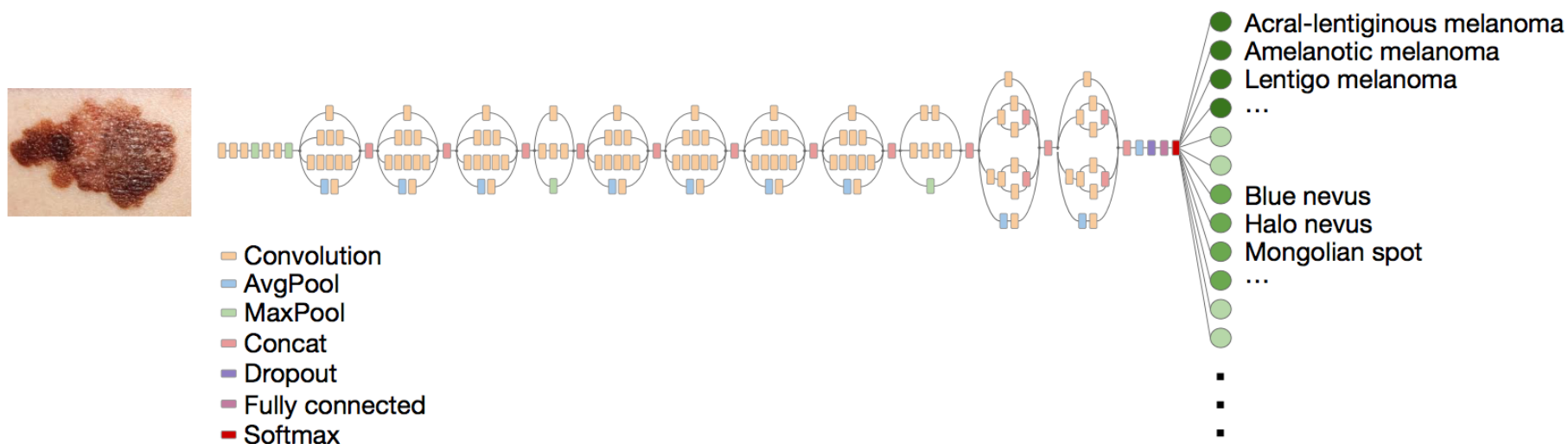
Shuffle
x and y



Datasets to train

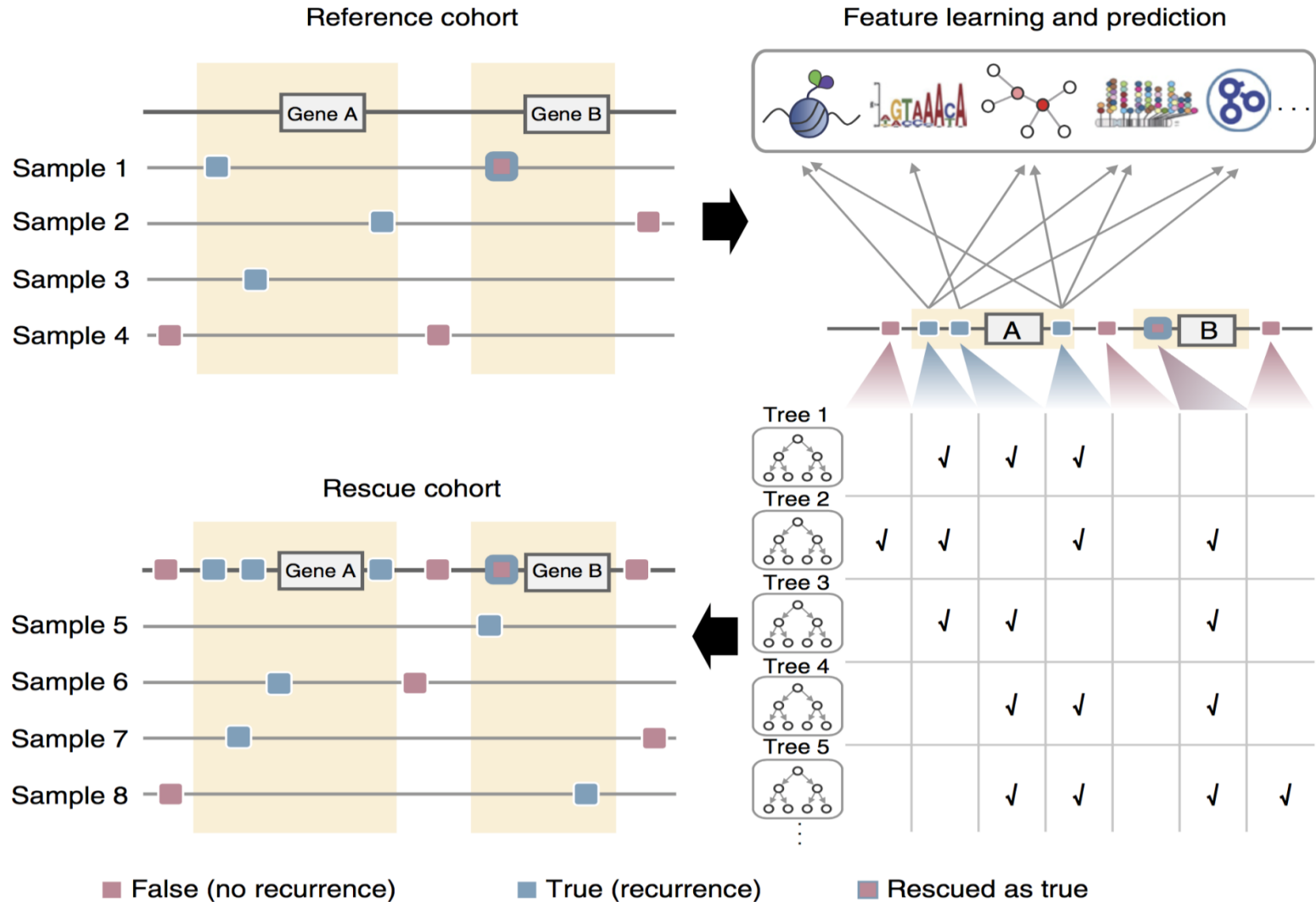


How to train: The model

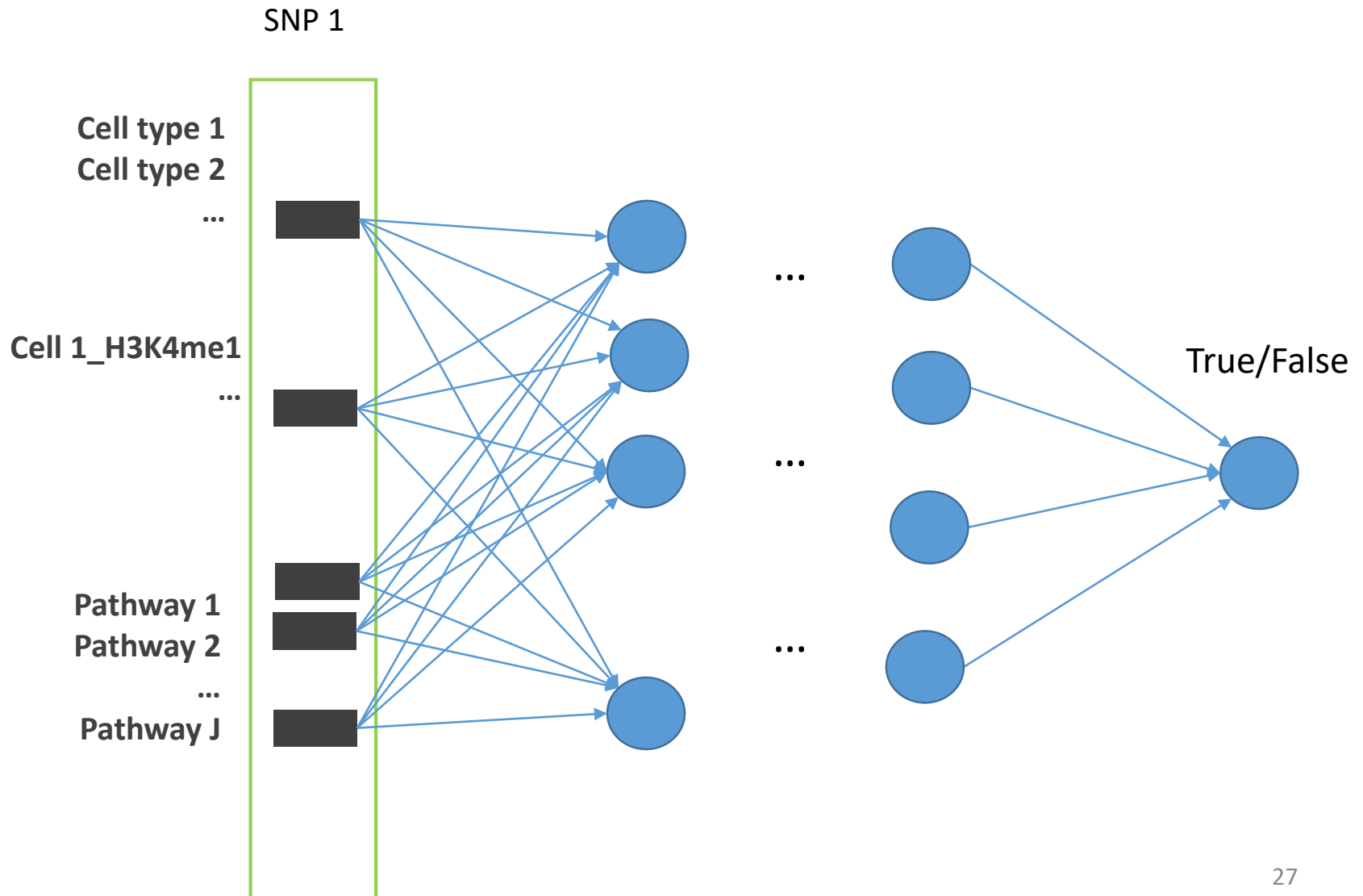


Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* 542.7639 (2017): 115.

Machine Learning for Recurrent Mutations

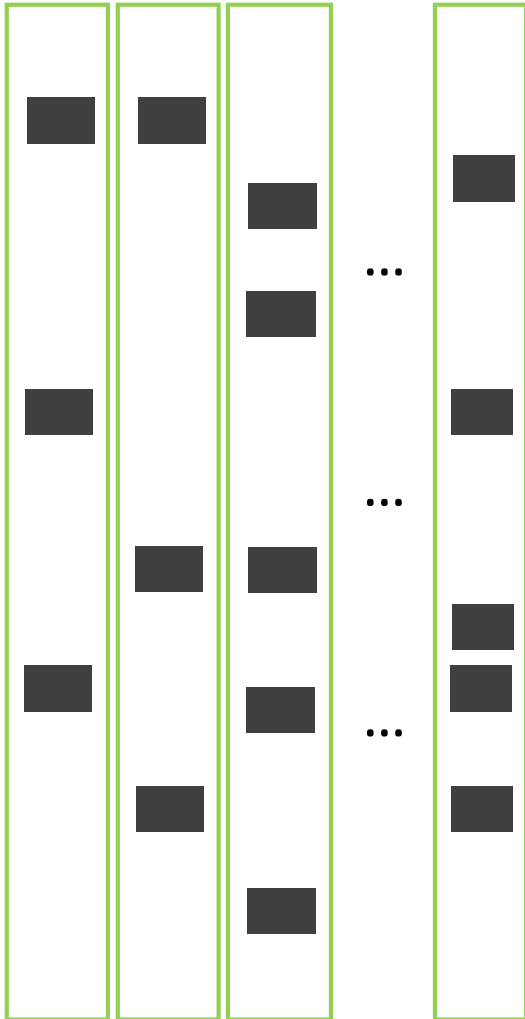


Classify single SNP

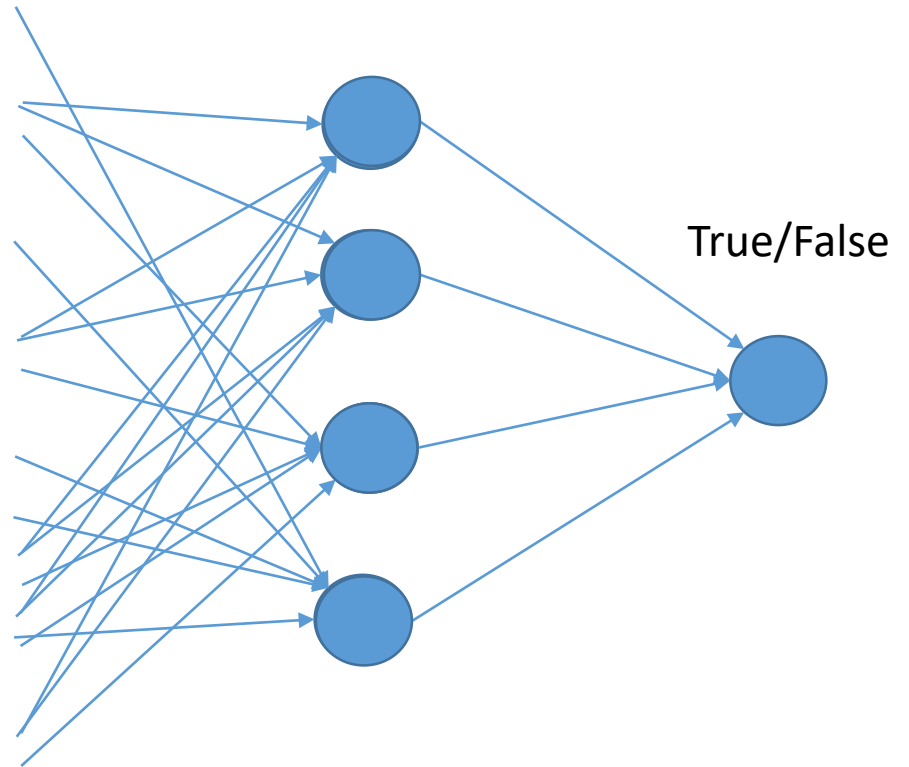


Classify group of SNP

SNP 1 2 3 ... SNP 30

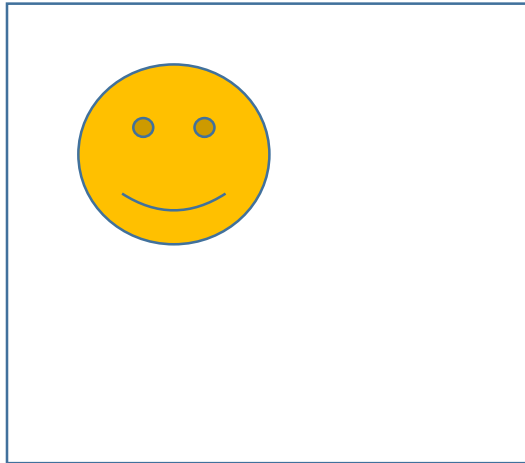


?



Use of convolutional neural network

smile



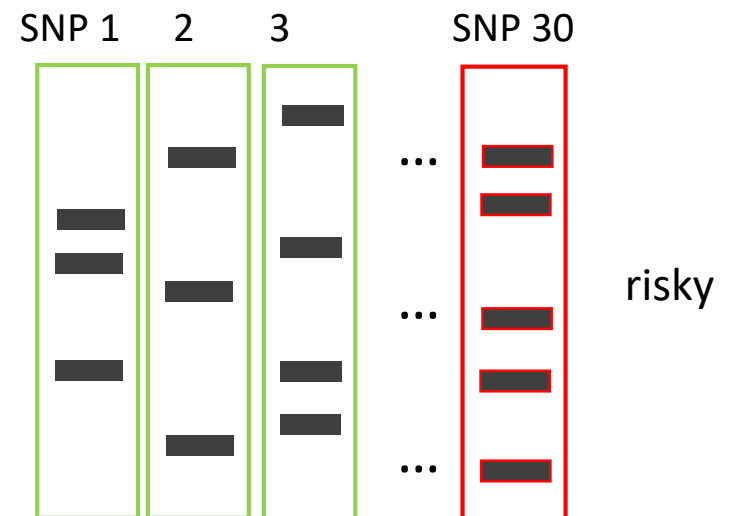
Find same
pattern of
image



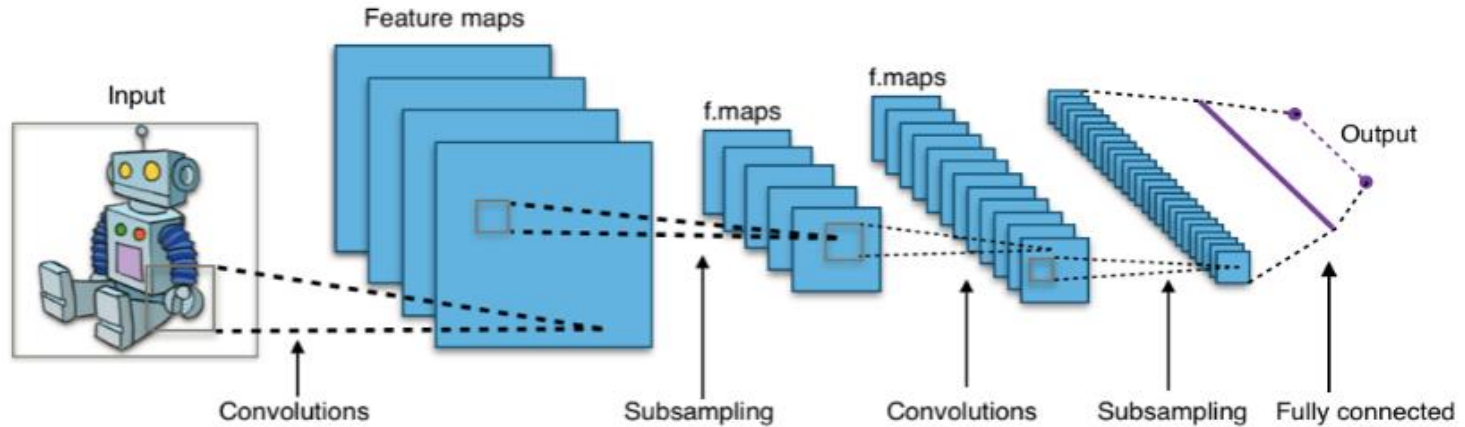
smile



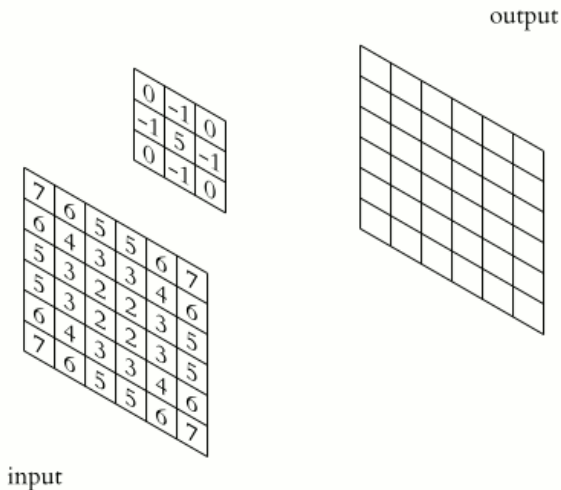
Not
sensitive
to position



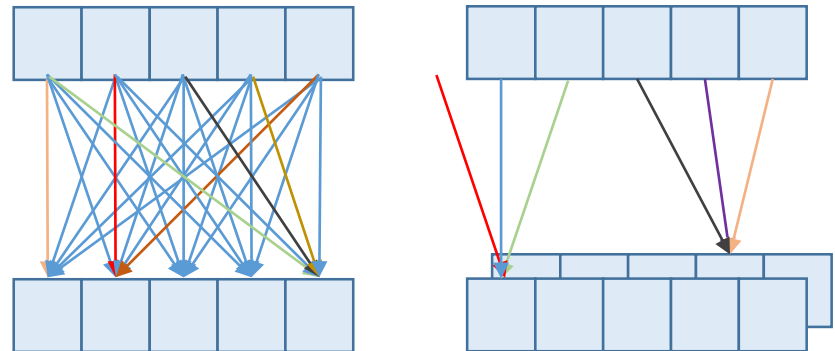
Convolutional neural network



Convolution



Fully-connected vs convolutional layer



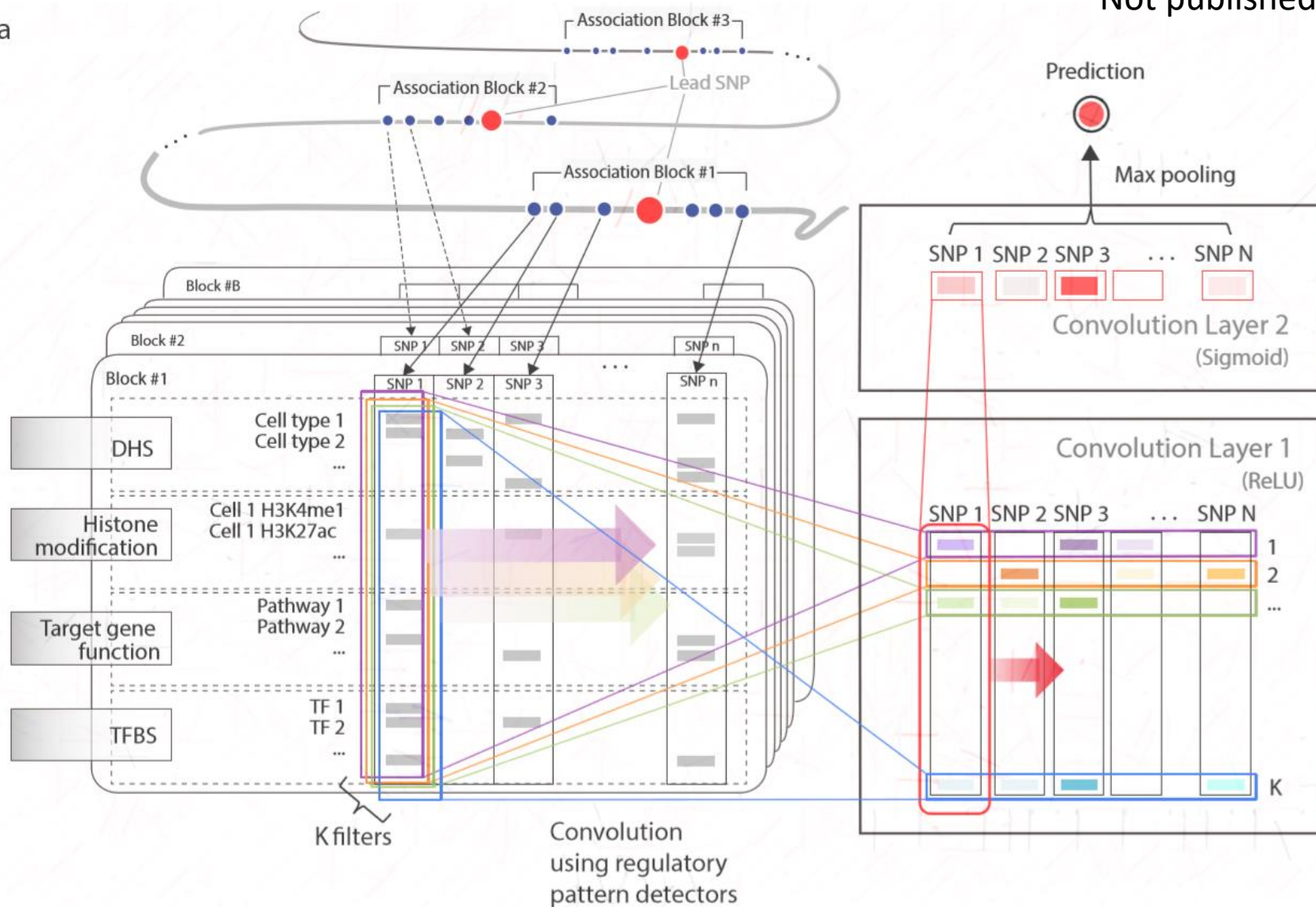
Weight of all links are different
(parameters to learn)

Num of weights
= num of kernels
* size of kernel

Model using CNN

Not published

a



Feature reduction

Small number of samples ~ 500

Too many and
sparse features

0	1	0	1	0	0	0	1	1	
0	1	1	1	0	1	1	1	0	
1	1	1	1	1	0	1	1	0	
1	0	0	1	1	1	0	0	1	
0	0	1	1	1	0	0	0	1	
0	0	1	1	1	0	1	1	1	
1	0	0	0	1	1	1	0	0	
1	0	1	0	1	1	0	0	1	
0	1	0	1	0	0	1	1	0	
0	1	1	0	0	0	0	1	1	

Remove sparse
features < 1%

1	0	1	1	0	0	0	0	1	
0	1	1	0	0	1	0	1	0	
1	1	0	1	0	1	0	0	0	
1	0	1	1	1	1	0	0	1	
0	0	1	1	1	0	0	0	1	
0	1	0	0	1	1	0	0	1	
0	1	1	1	0	1	1	0	1	
1	1	1	1	0	0	1	0	1	
0	1	1	0	0	1	1	0	1	
1	0	1	0	0	1	1	0	1	

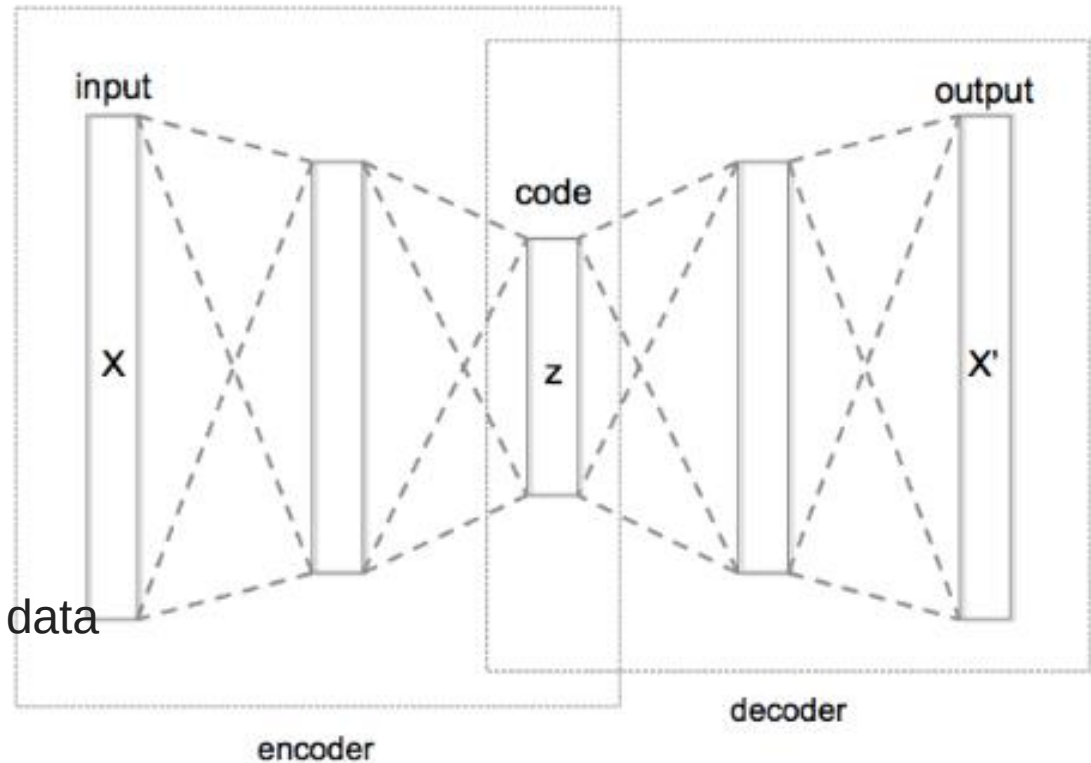
Binding features are removed

Autoencoder

Unsupervised learning

ANN

For efficient encoding of input data



Good tool for dimension reduction

Feature extraction

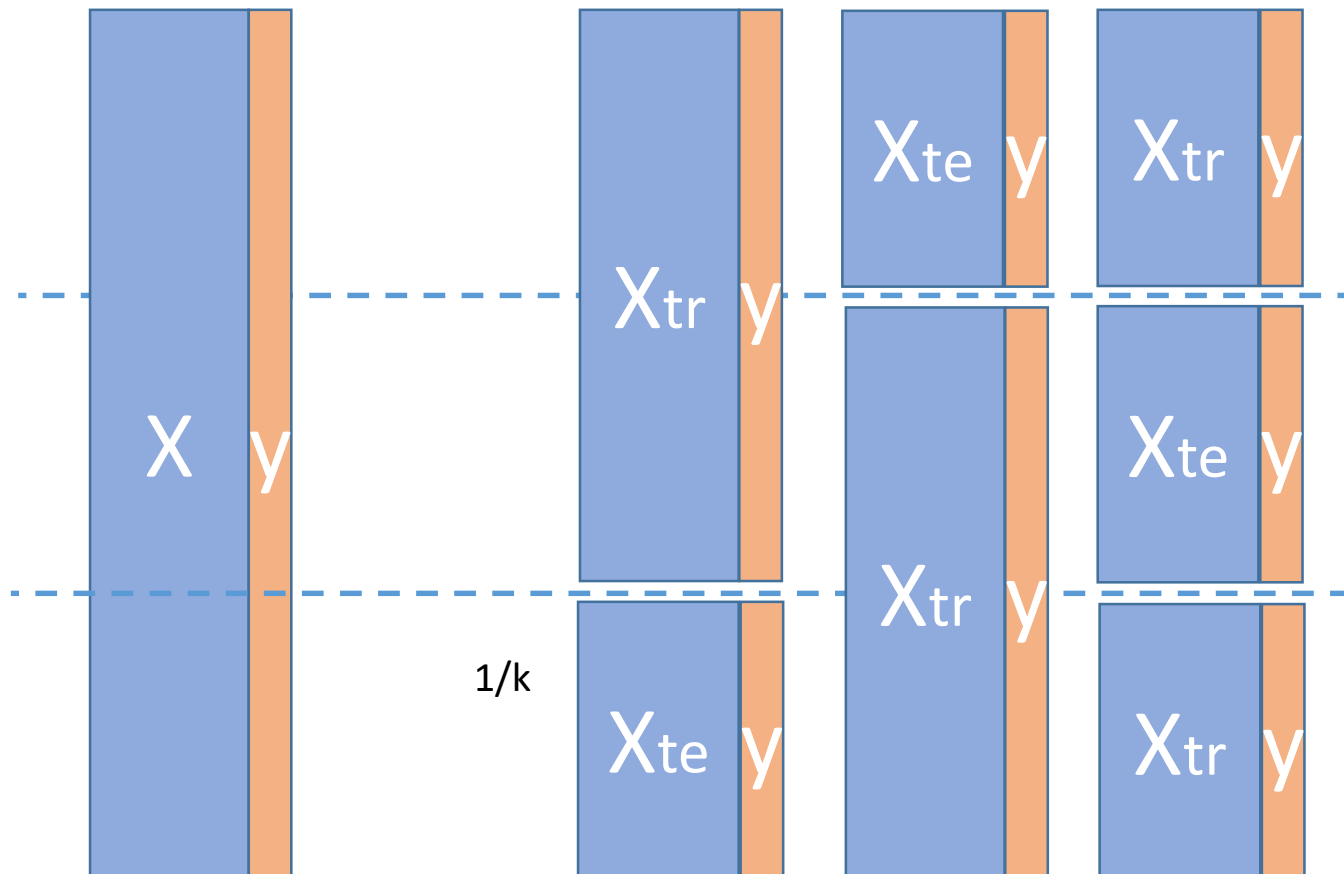
Non-linear (if 2 or more encoding layers are used)

Cf. PCA

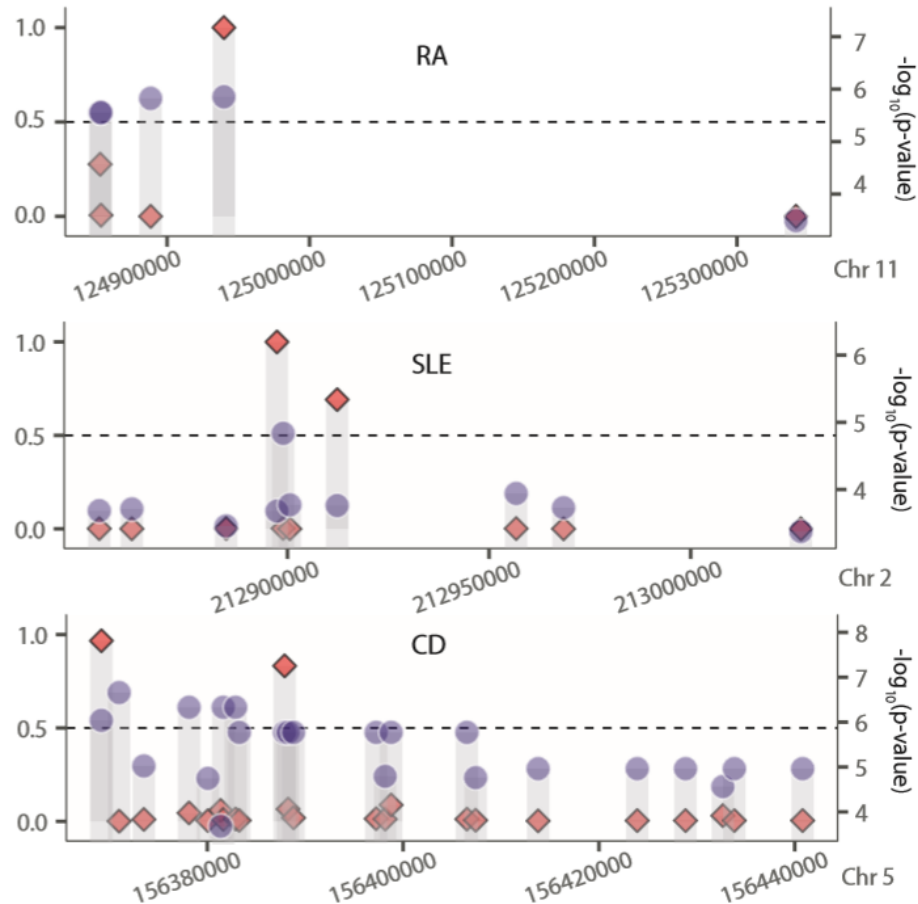
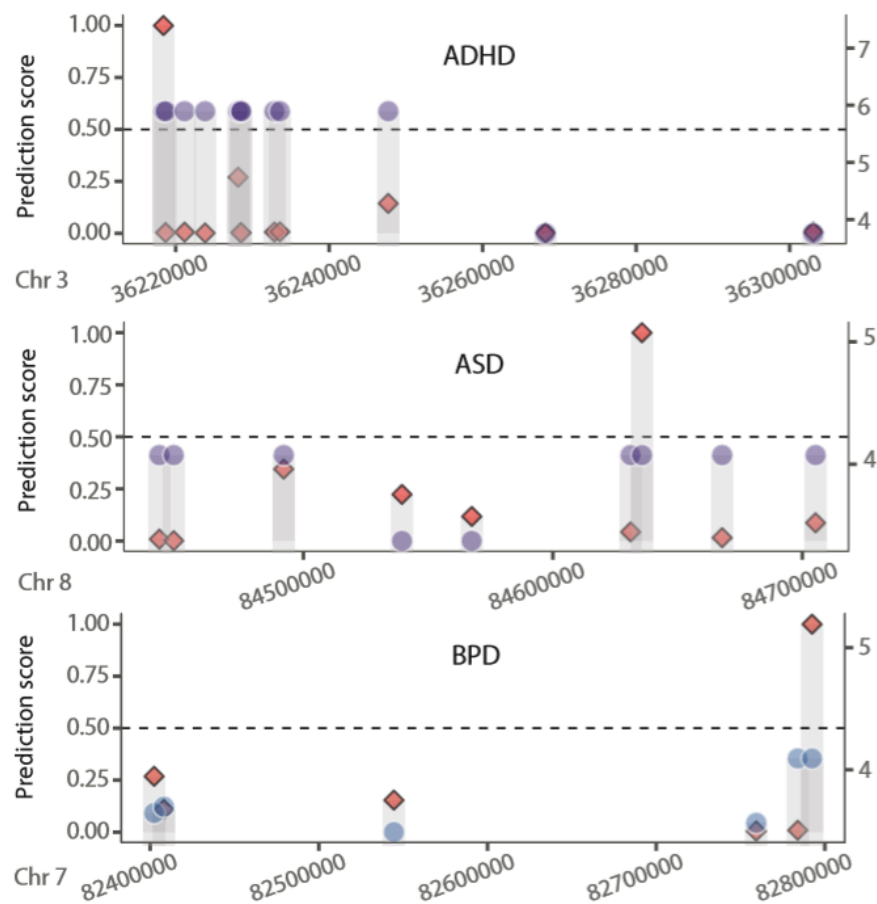
PCA is linear and greedy (1st PC explain more than 2nd)

Result and validation

k-fold cross-validation



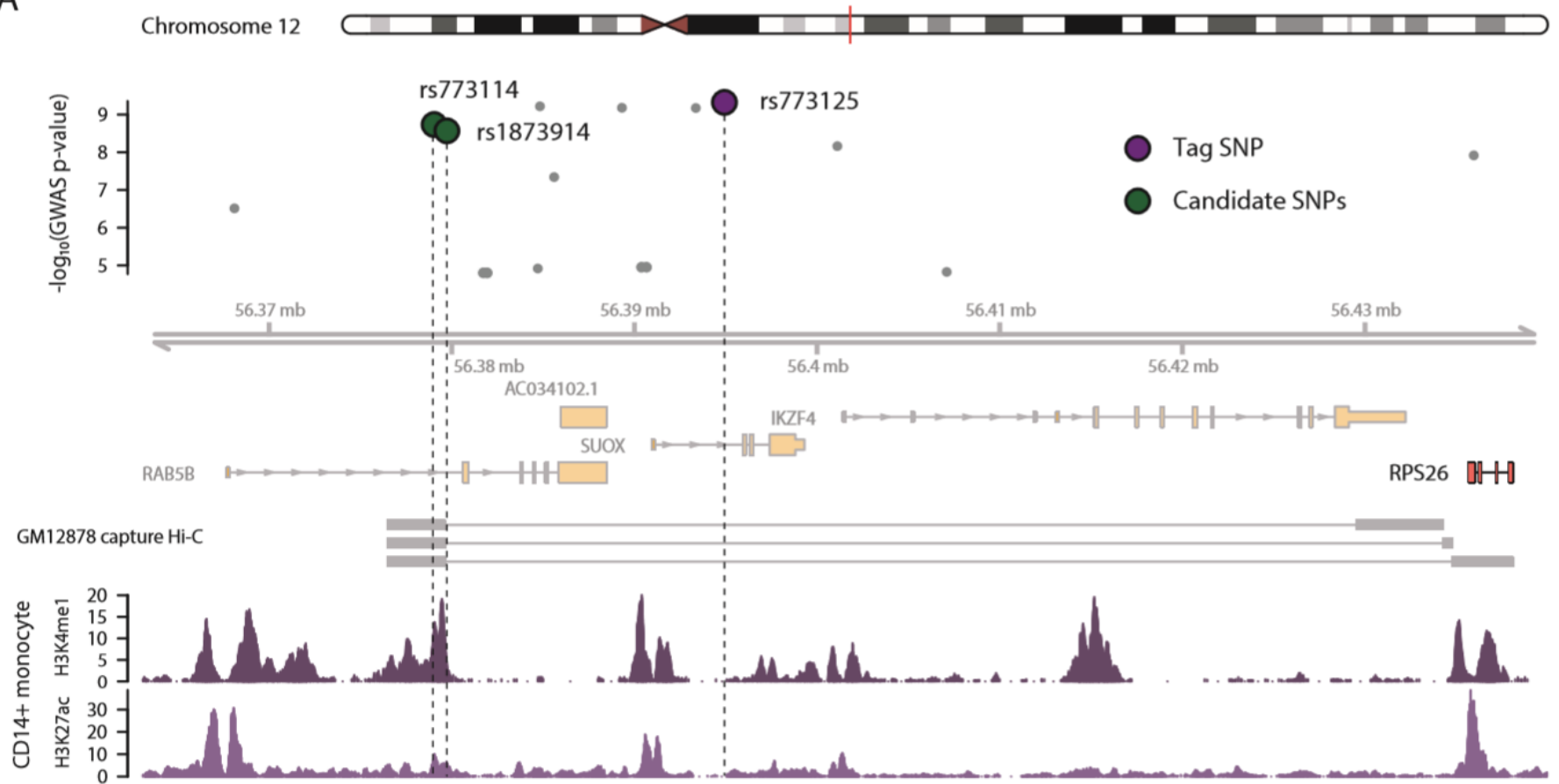
Distinguished SNPs



Not published

Biological validation

A

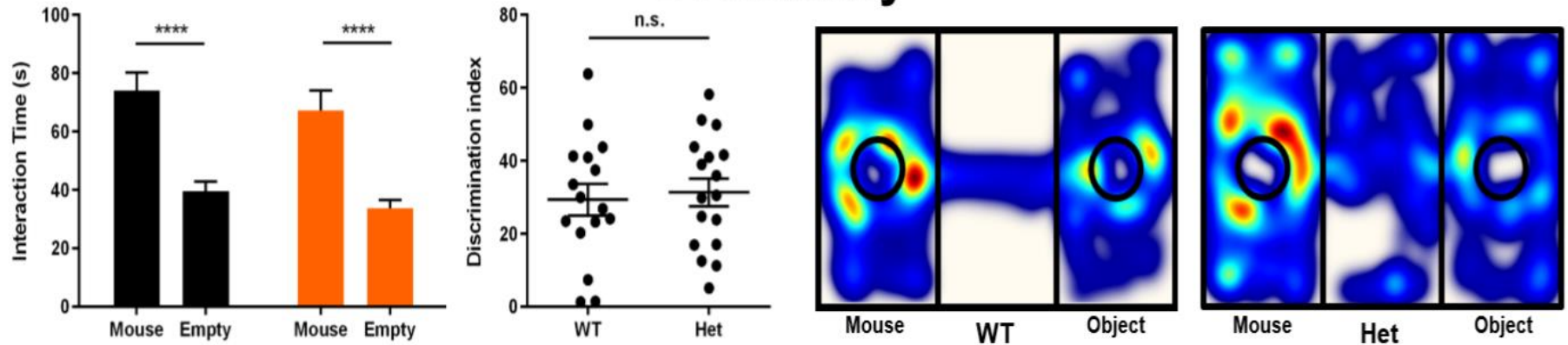


Not published

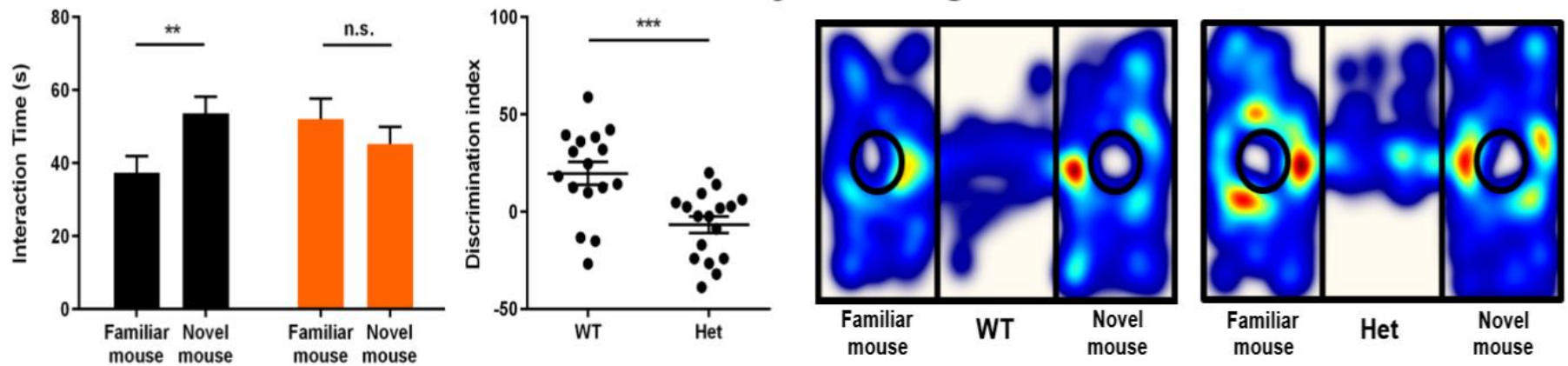
Knockout mice

Not published

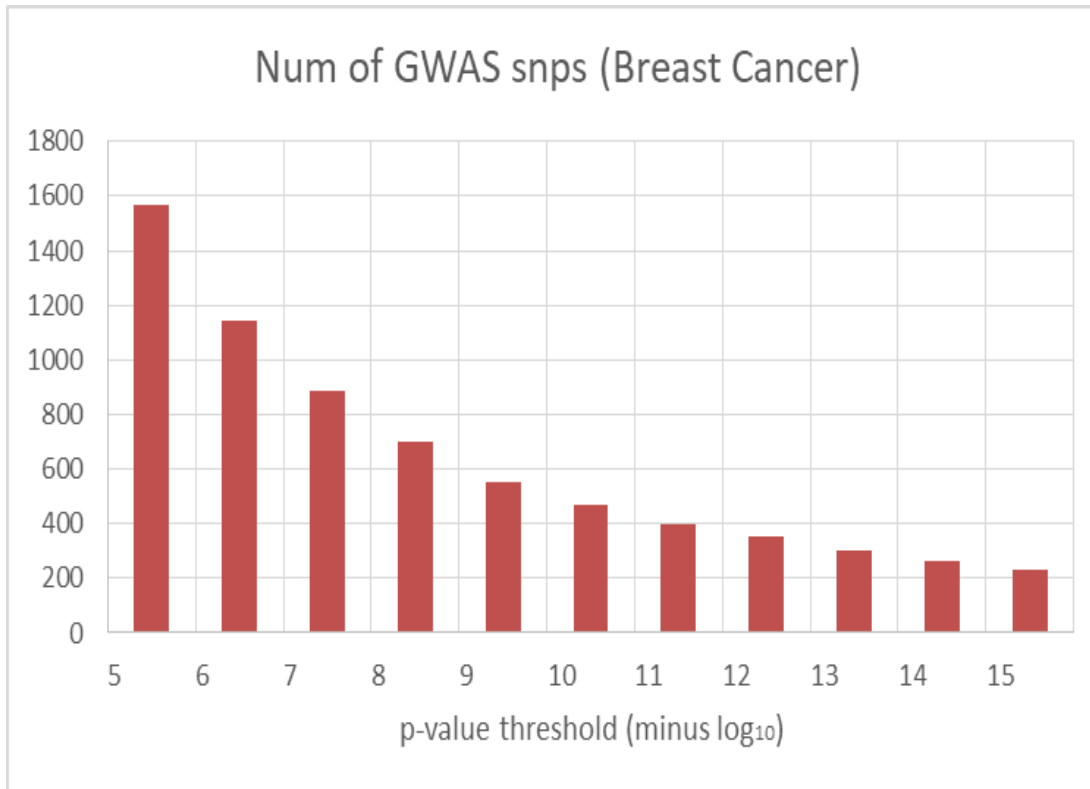
Sociability



Social Novelty Recognition



! Lack of samples !



Cf.
ImageNet
~ 14m images

MNIST
~ 60k hand-written

Recent trend of deep learning

Type ◆	Classifier ◆	Distortion ◆	Preprocessing ◆	Error rate (%) ▲
Convolutional neural network	Committee of 5 CNNs, 6-layer 784-50-100-500-1000-10-10	None	Expansion of the training data	0.21 ^[17]
Convolutional neural network	Committee of 35 CNNs, 1-20-P-40-P-150-10	elastic distortions	Width normalizations	0.23 ^[8]
Convolutional neural network	6-layer 784-50-100-500-1000-10-10	None	Expansion of the training data	0.27 ^[24]
Convolutional neural network	6-layer 784-40-80-500-1000-2000-10	None	Expansion of the training data	0.31 ^[15]
Deep neural network	6-layer 784-2500-2000-1500-1000-500-10	elastic distortions	None	0.35 ^[23]
K-Nearest Neighbors	K-NN with non-linear deformation (P2DHMDM)	None	Shiftable edges	0.52 ^[19]
Support vector machine	Virtual SVM, deg-9 poly, 2-pixel jittered	None	Deskewing	0.56 ^[21]

Data augmentation

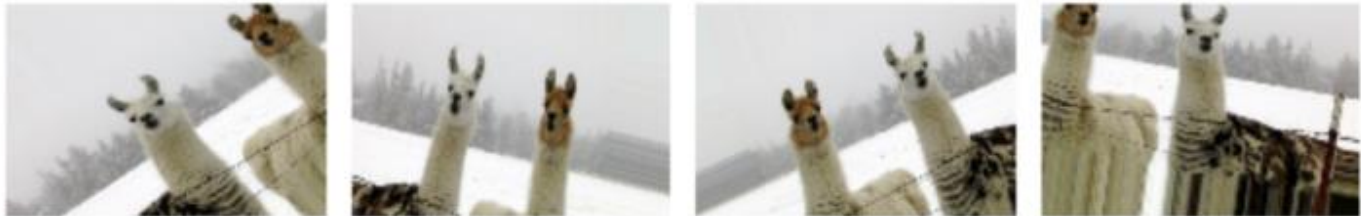


Figure I: Traditional Transformations



Figure II: Style Transformations via GANs

Perez, Luis, and Jason Wang. "The effectiveness of data augmentation in image classification using deep learning." *arXiv preprint arXiv:1712.04621* (2017).

Augmentation for multi-dimensional data

SNP	1	2	3	4	5	6	7	8
DHS	1.0							
CTCF								
H2AZ		1.0						
H3K27ac								
H3K27me3		1.0	1.0		1.0			
H3K36me3	1.0							
H3K4me1	1.0					1.0		
H3K4me3		1.0						
H3K9ac	1.0		1.0		1.0			1.0
H3K9me3				1.0	1.0		1.0	1.0
H4K20me1			1.0					

Traditional



SNP	1	2	3	4	5	6	7	8
DHS					1.0			
CTCF			1.0	1.0				
H2AZ		1.0		1.0				
H3K27ac						1.0		
H3K27me3	1.0	1.0				1.0		
H3K36me3			1.0					
H3K4me1				1.0			1.0	
H3K4me3		1.0	1.0	1.0				
H3K9ac				1.0				
H3K9me3						1.0	1.0	
H4K20me1						1.0		

Style transform



SNP	1	2	3	4	5	6	7	8
DHS								
CTCF	0.3							
H2AZ								
H3K27ac	0.3	0.3	0.3					
H3K27me3	0.3	0.3						
H3K36me3	0.5	0.4						
H3K4me1	0.5	0.3						
H3K4me3	0.5	0.4						
H3K9ac	0.3	0.3	0.3	0.4	0.3	0.3	0.3	0.5
H3K9me3		0.3	0.3	0.6	0.3	0.3	0.3	0.5
H4K20me1			0.3	0.5	0.3	0.3	0.3	0.5

It is not 2D image

Augmentation for multi-dimensional data

SNP	1	2	3	4	5	6	7	8
DHS	1.0							
CTCF								
H2AZ		1.0						
H3K27ac								
H3K27me3		1.0	1.0		1.0			
H3K36me3	1.0							
H3K4me1	1.0					1.0		
H3K4me3		1.0						
H3K9ac	1.0		1.0		1.0			1.0
H3K9me3				1.0	1.0		1.0	1.0
H4K20me1			1.0					

Add
random
features



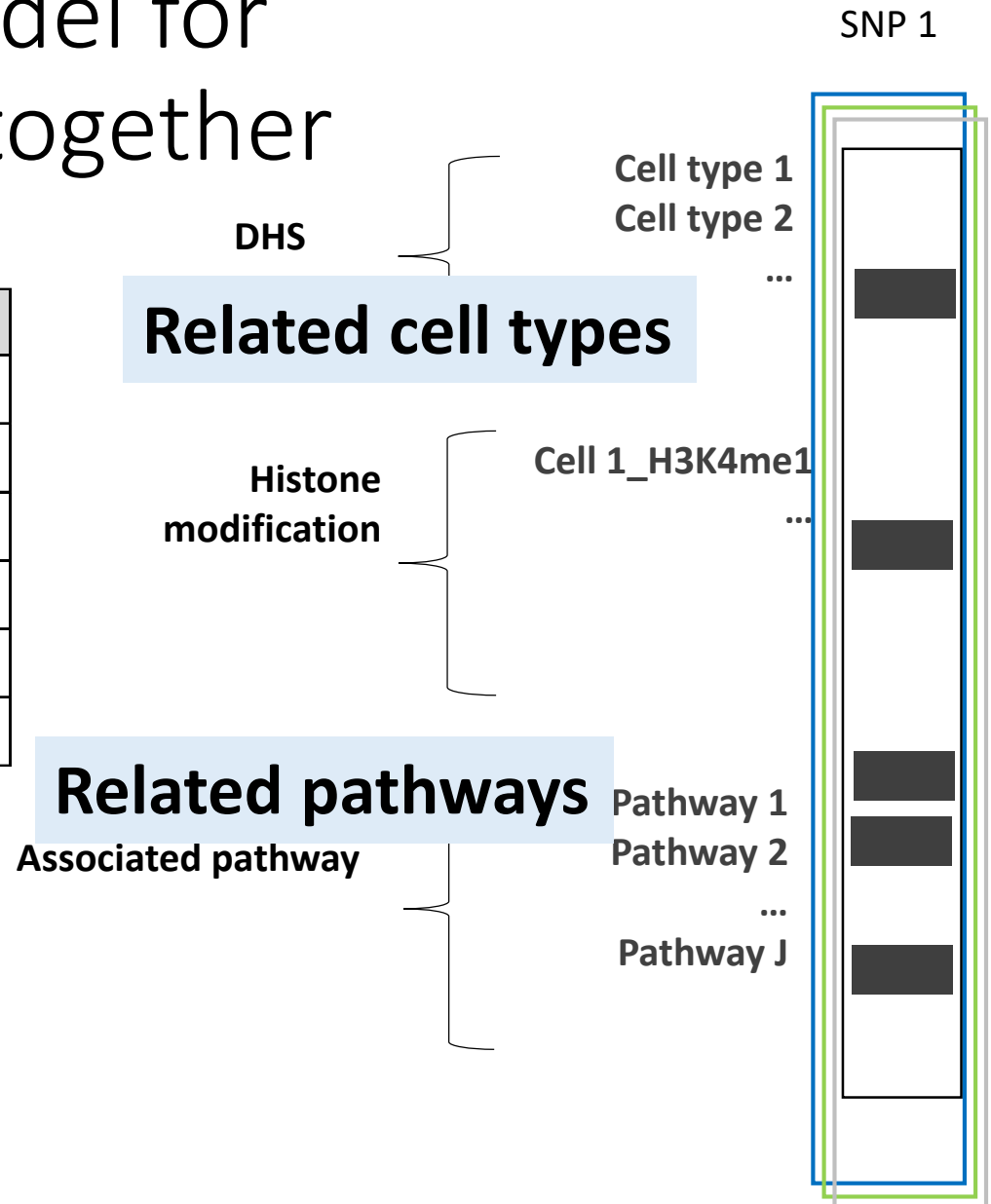
X 10

SNP	1	2	3	4	5	6	7	8
DHS	1.0							
CTCF								
H2AZ		1.0						
H3K27ac								
H3K27me3		1.0	1.0		1.0			
H3K36me3	1.0							
H3K4me1	1.0					1.0		
H3K4me3		1.0						
H3K9ac	1.0		1.0		1.0			1.0
H3K9me3				1.0	1.0		1.0	1.0
H4K20me1			1.0					

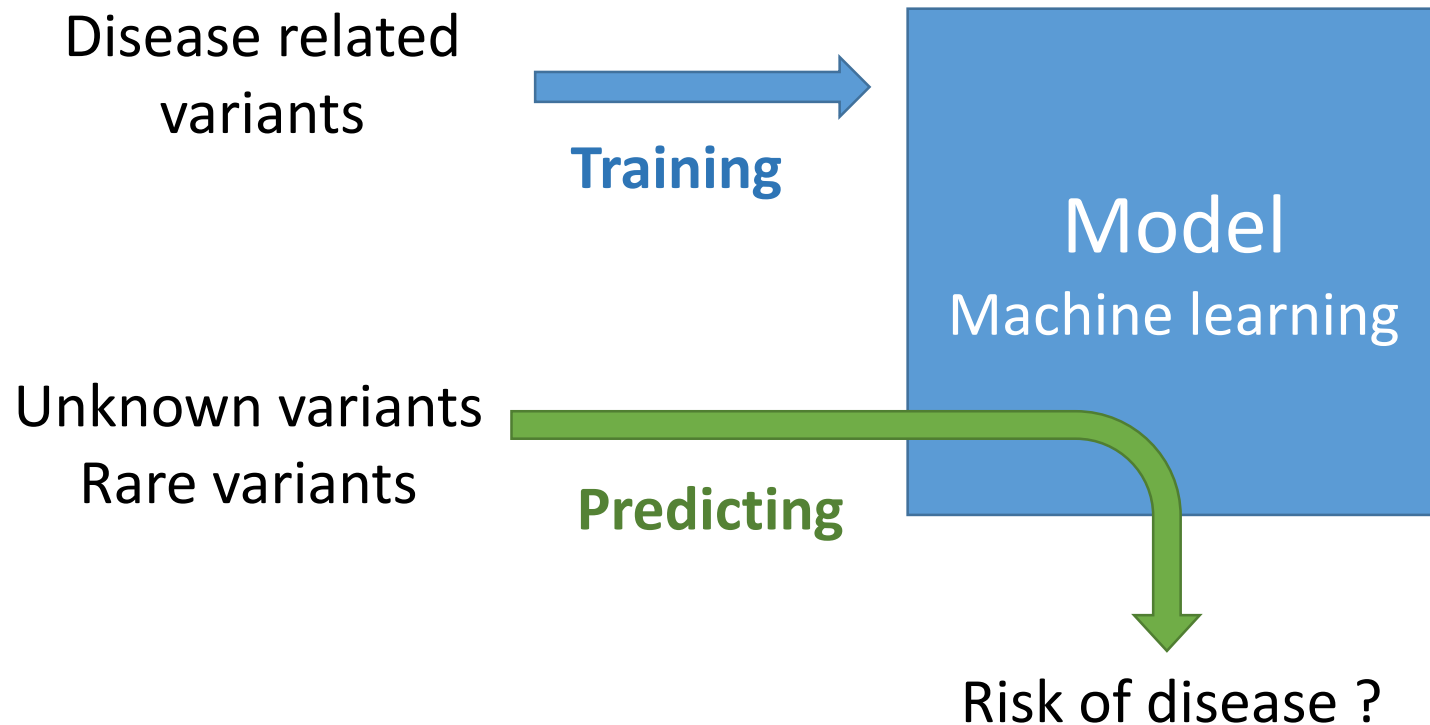
SNP	1	2	3	4	5	6	7	8
DHS	1.0							
CTCF								
H2AZ		1.0						
H3K27ac								
H3K27me3		1.0	1.0		1.0			
H3K36me3	1.0							
H3K4me1	1.0					1.0		
H3K4me3		1.0						
H3K9ac	1.0		1.0		1.0			1.0
H3K9me3				1.0	1.0		1.0	1.0
H4K20me1			1.0					

Generalized model for all GWAS SNPs together

Number of associations		
ADHD	ASD	BPD
642	943	1,424
MDD	SCZ	RA
832	601	435
SLE	CD	UC
849	431	383



Summary

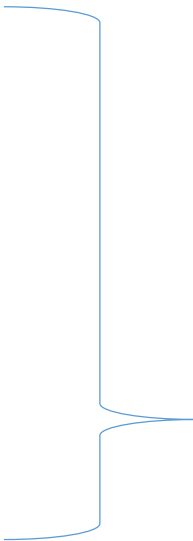


History

1943	MCP artificial neuron
1958	Perceptron (machine learning)
1969	Multi-layer perceptron (ANN)
1986	Backpropagation
1997	LSTM
1998	LeNet (CNN)
2012	...
2016	AlphaGo
2025	My work?

Tools

2006	CUDA
2008	Python3
2010	Theano
2015	Tensorflow Keras



Thank you