BIRD: Big data Regression for Predicting DNase I Hypersensitivity

Hongkai Ji
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Email: hji@jhu.edu

Big Data Prediction and Regression

X_{px1} → Y_{qx1}

$p = 10^4 - 10^9$

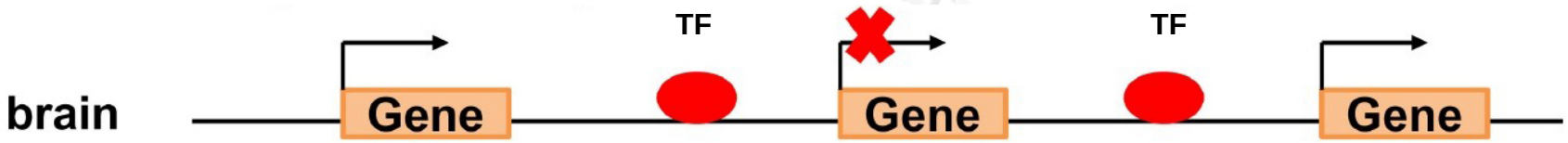
$q = 10^4 - 10^9$

- **Genotype**
- **Gene Expression**
- **Histone modification**
- **DNA methylation**

→

- **Chromatin states**
- **Transcription factor binding sites**
- **3D chromatin interaction**

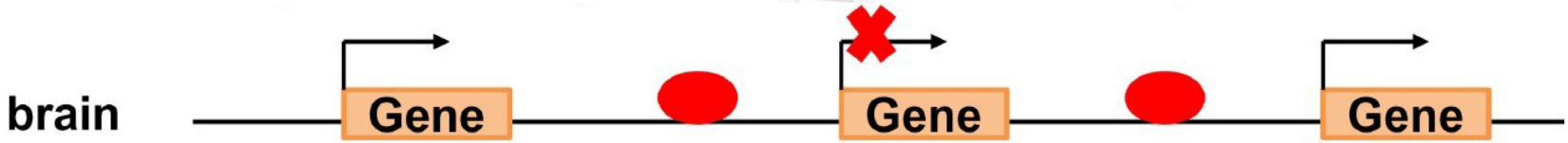
Gene Regulation, Transcription Factor (TF), Cis-regulatory Elements



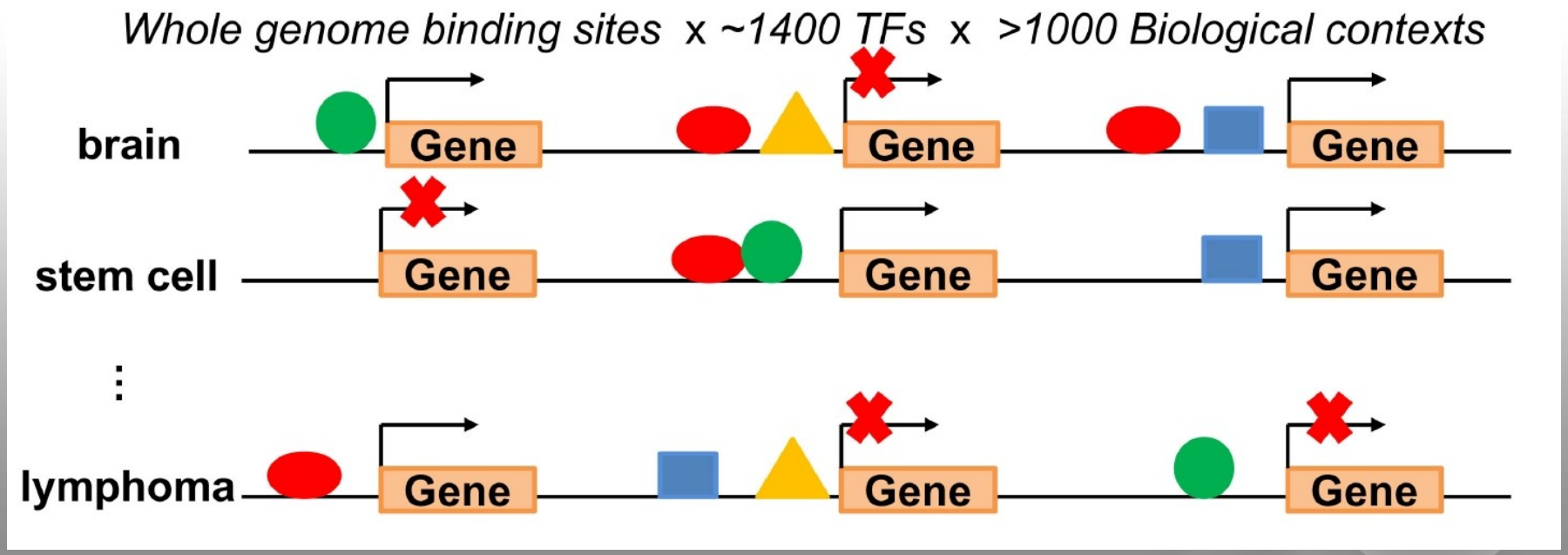
ChIP-seq/ChIP-chip

Problems with ChIP-seq/ChIP-chip

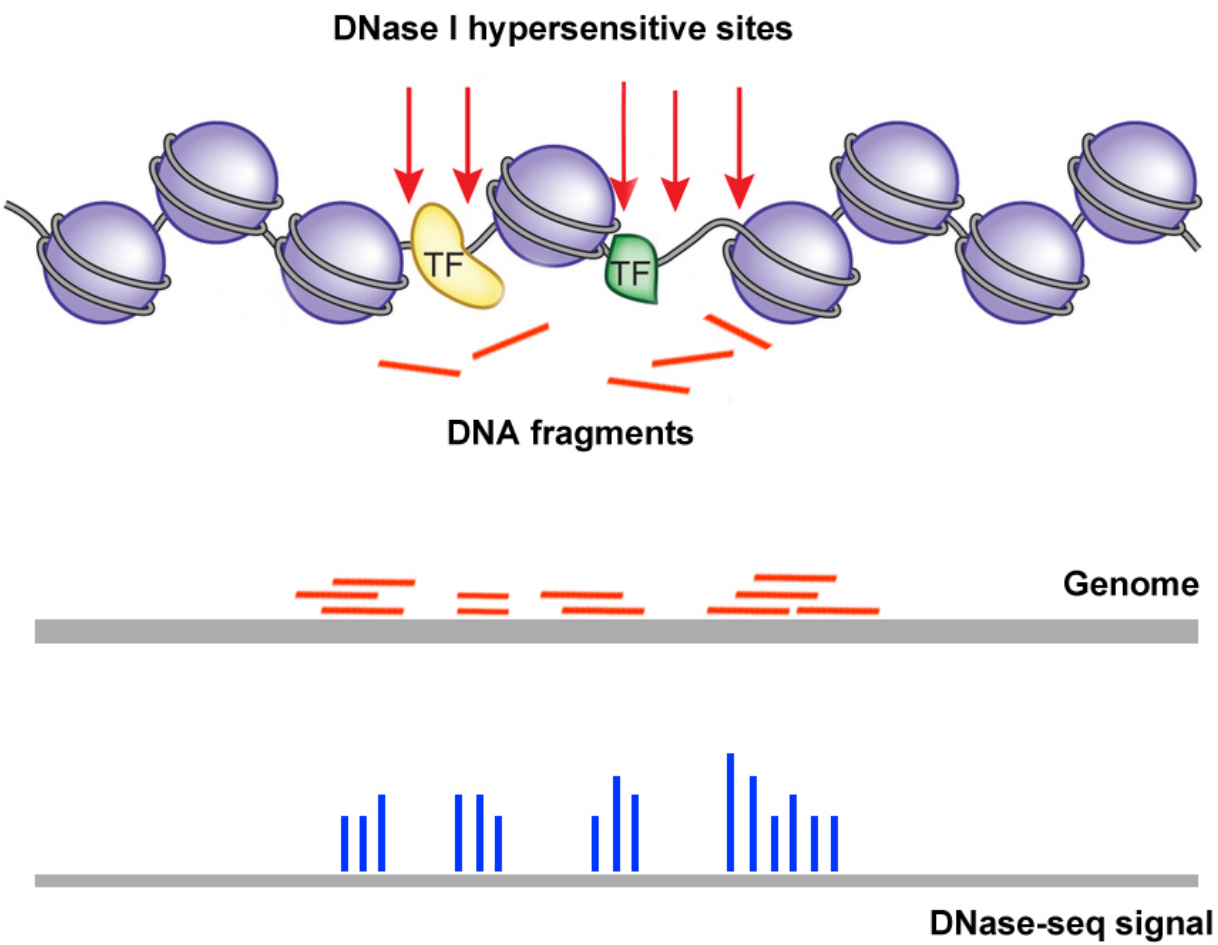
What ChIP-seq/ChIP-chip can do:



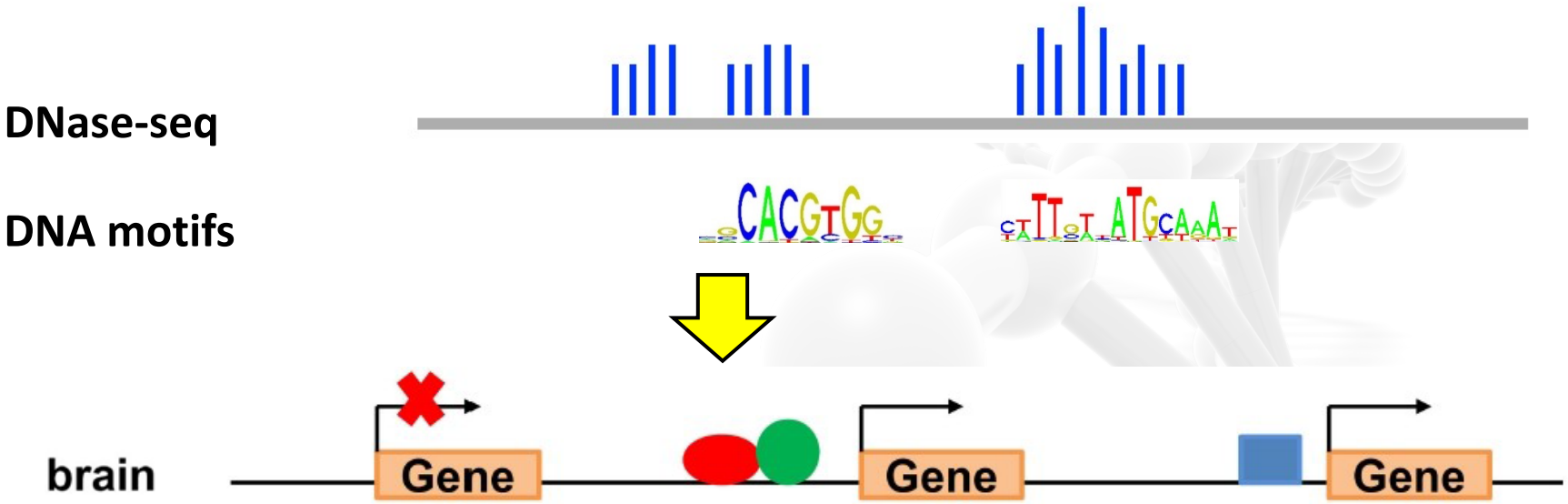
What we want to have:



DNase I Hypersensitivity (DHS) & DNase-seq

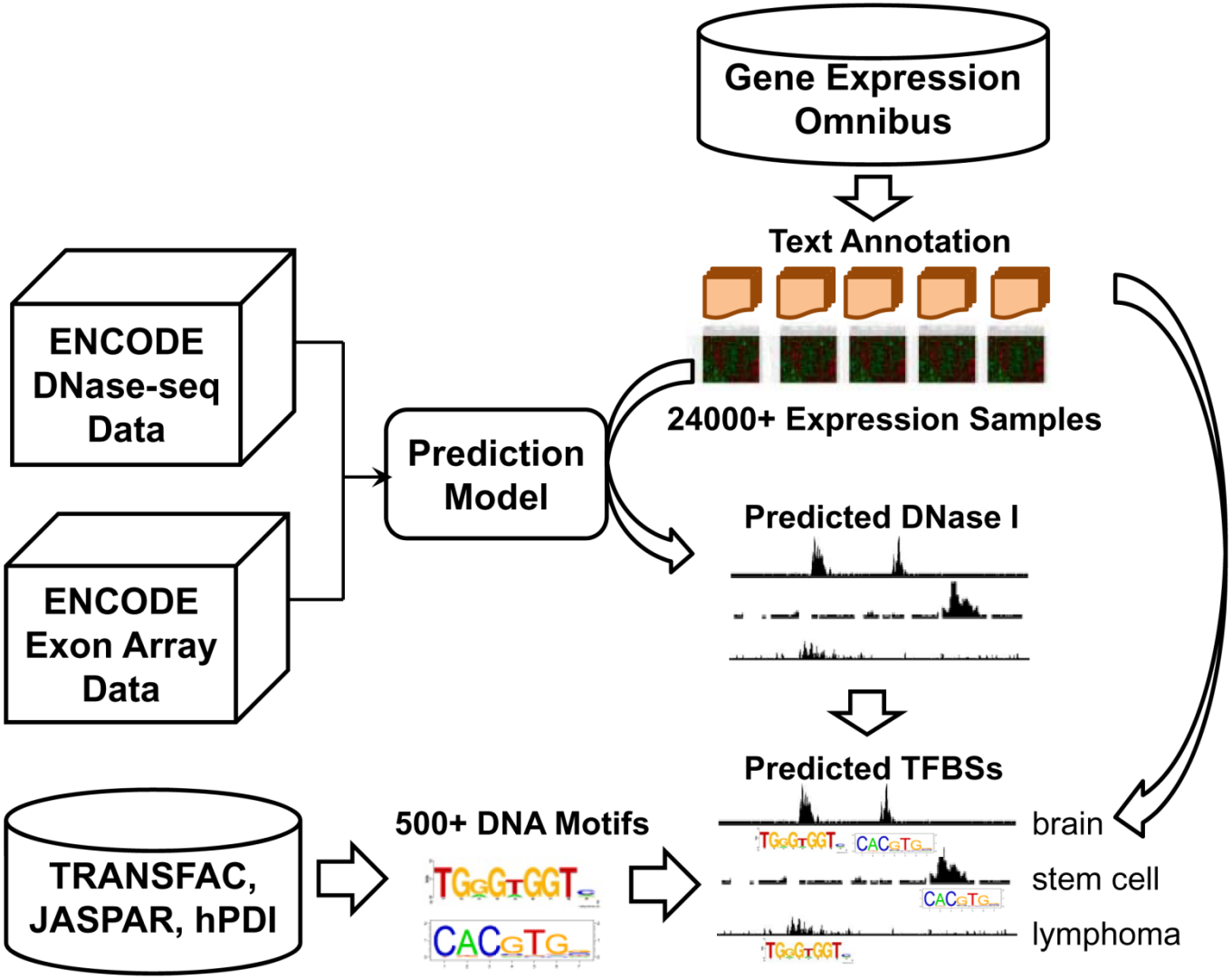


Predict Transcription Factor Binding Sites (TFBS) Based on DNase-seq



Pique-Regi et al. Genome Res. 2011. 21:447-455

Our Approach: A Solution Based on Big Data



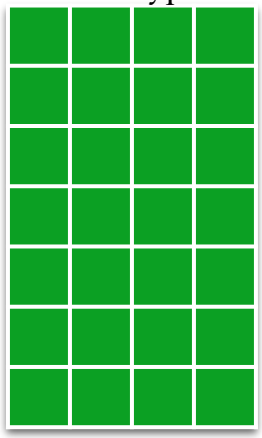
Training Data

X: Gene Expression

Cell type

Gene

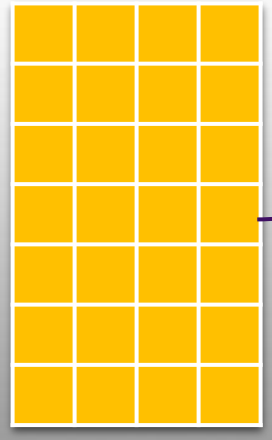
$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{bmatrix}$$



Y: DHS

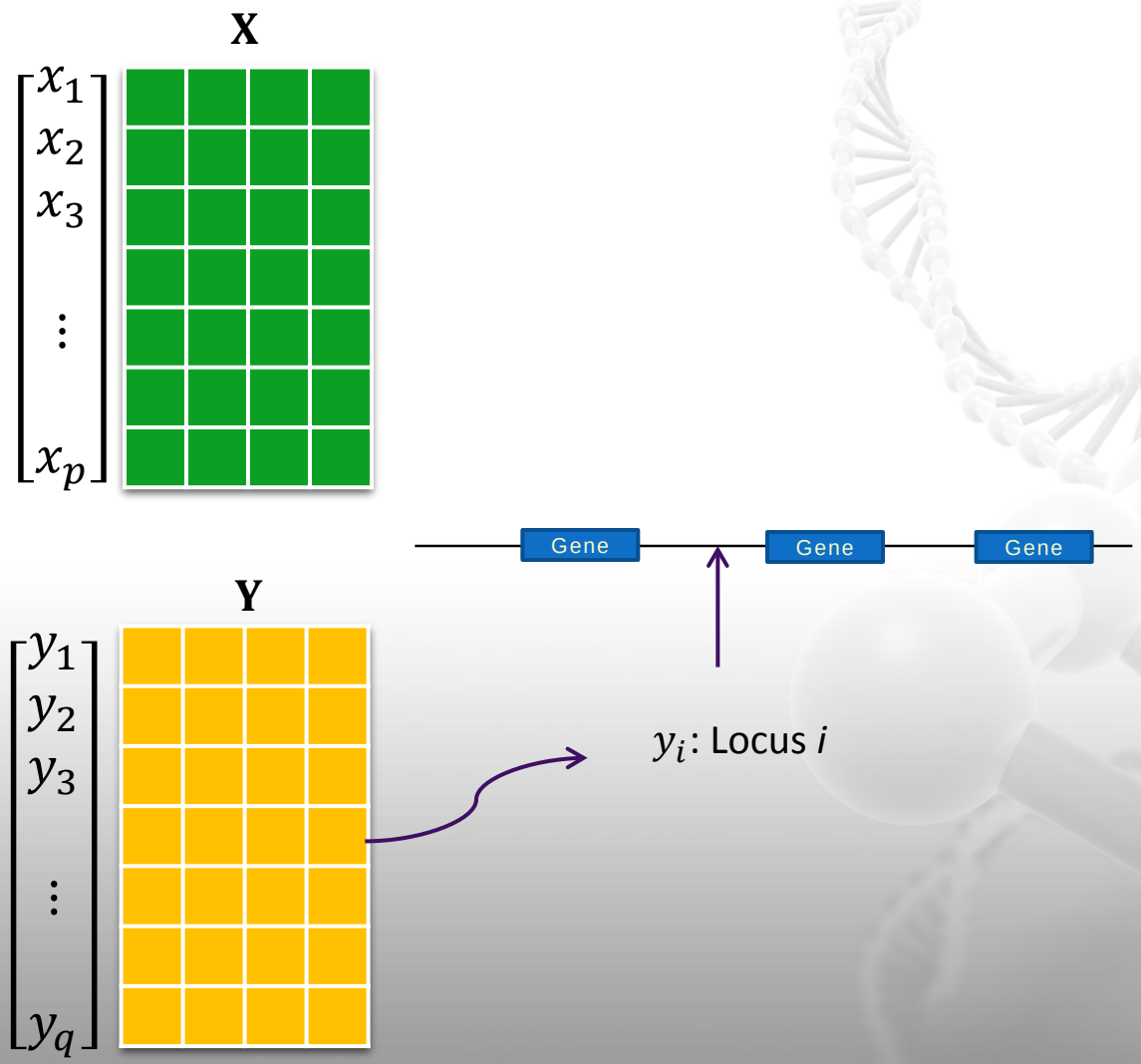
Locus

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_q \end{bmatrix}$$

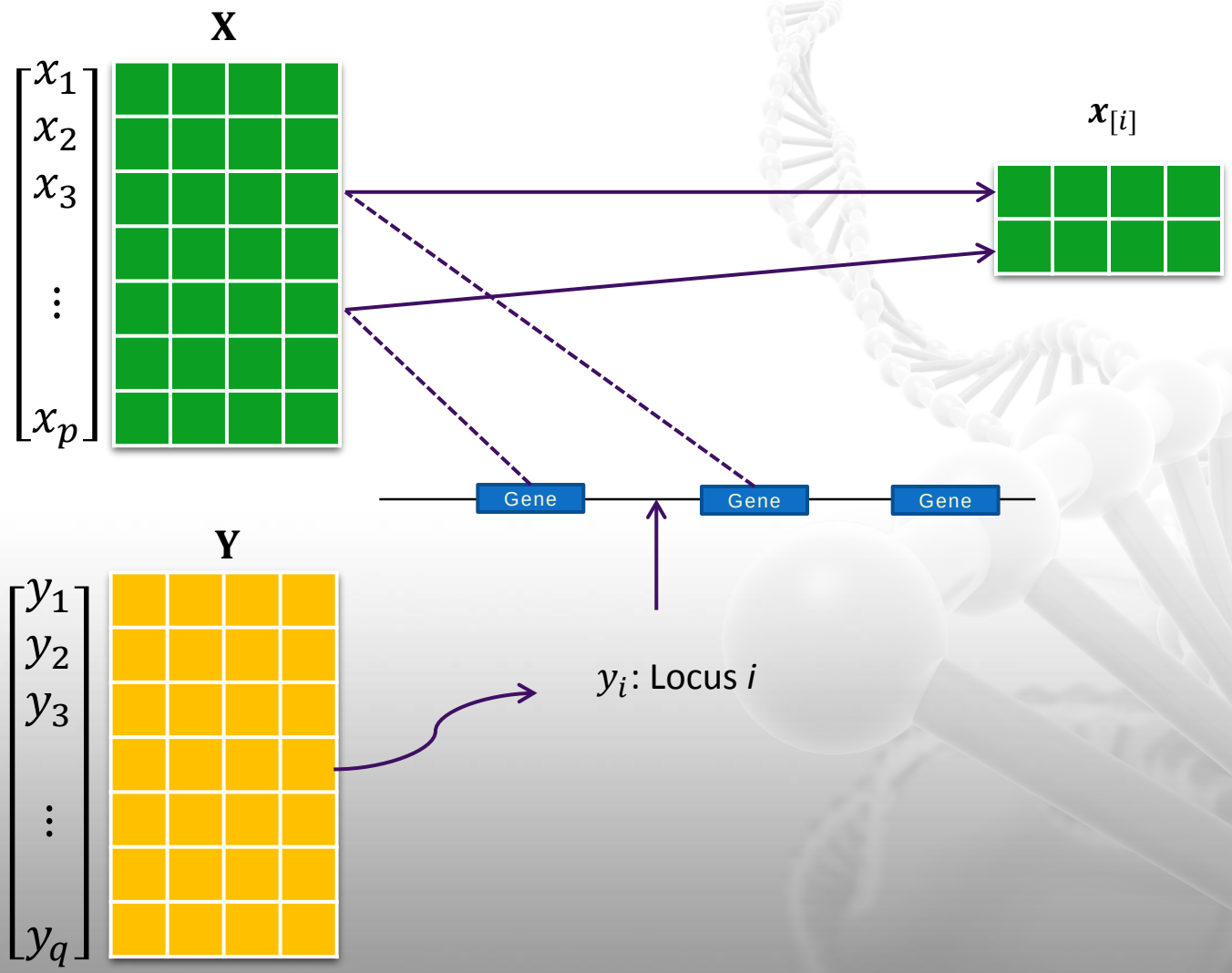


Locus *i*: $y_i = f_i(\mathbf{X}) + e$

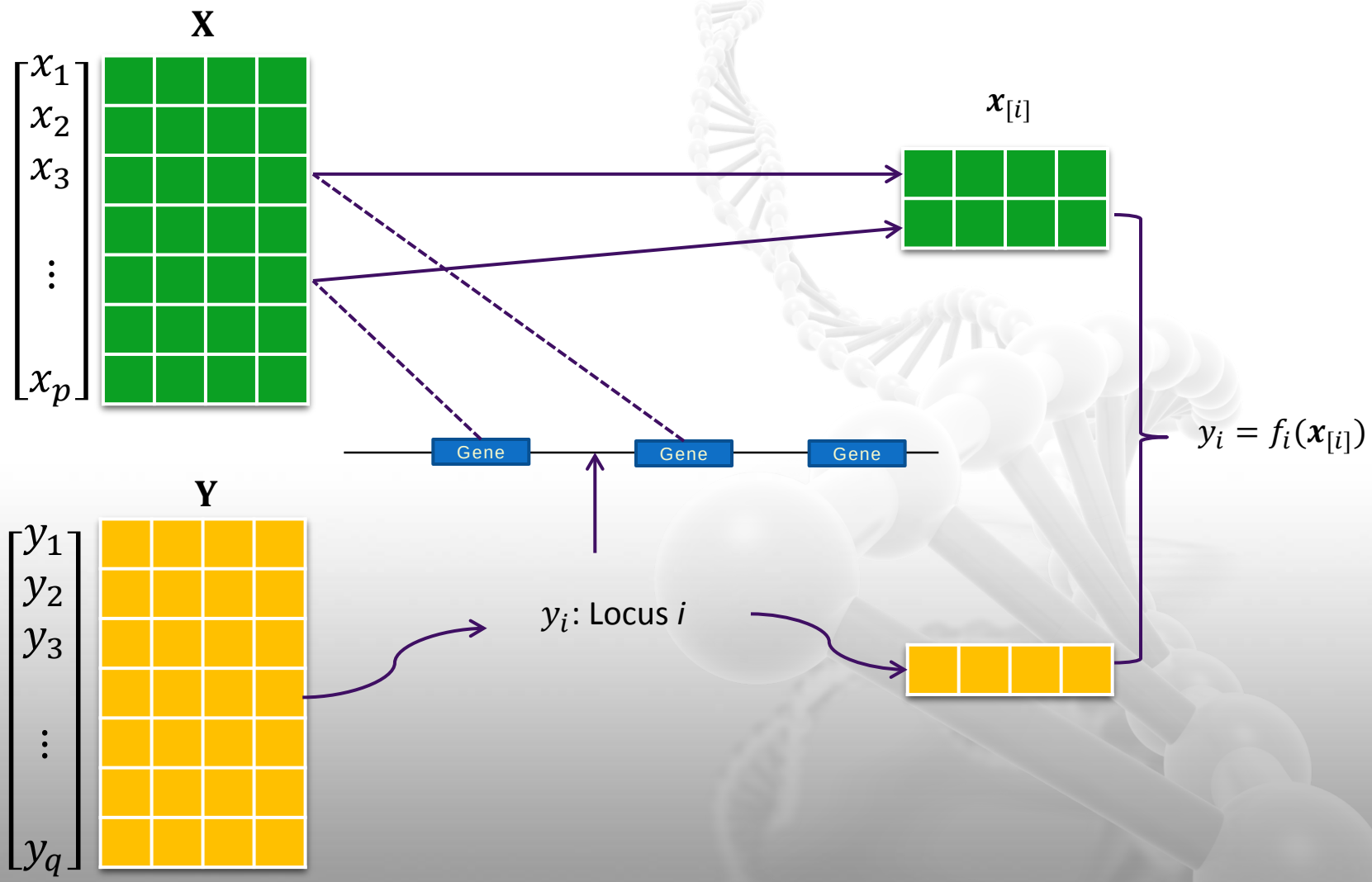
Neighboring Gene Approach



Neighboring Gene Approach

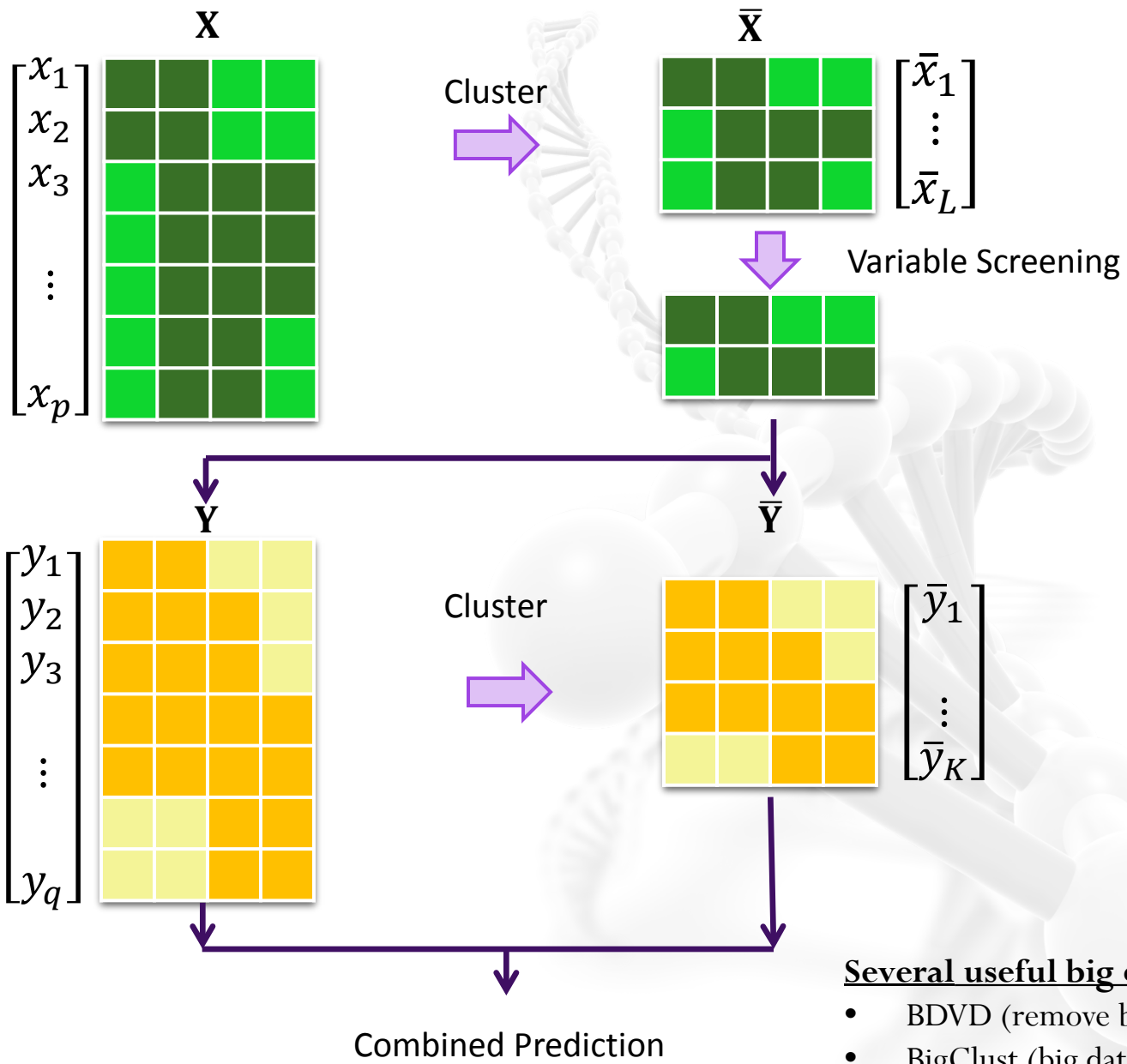


Neighboring Gene Approach



Problem: not all information is contained in the neighbors.

BIRD: Big Data Regression for Predicting DNase I Hypersensitivity

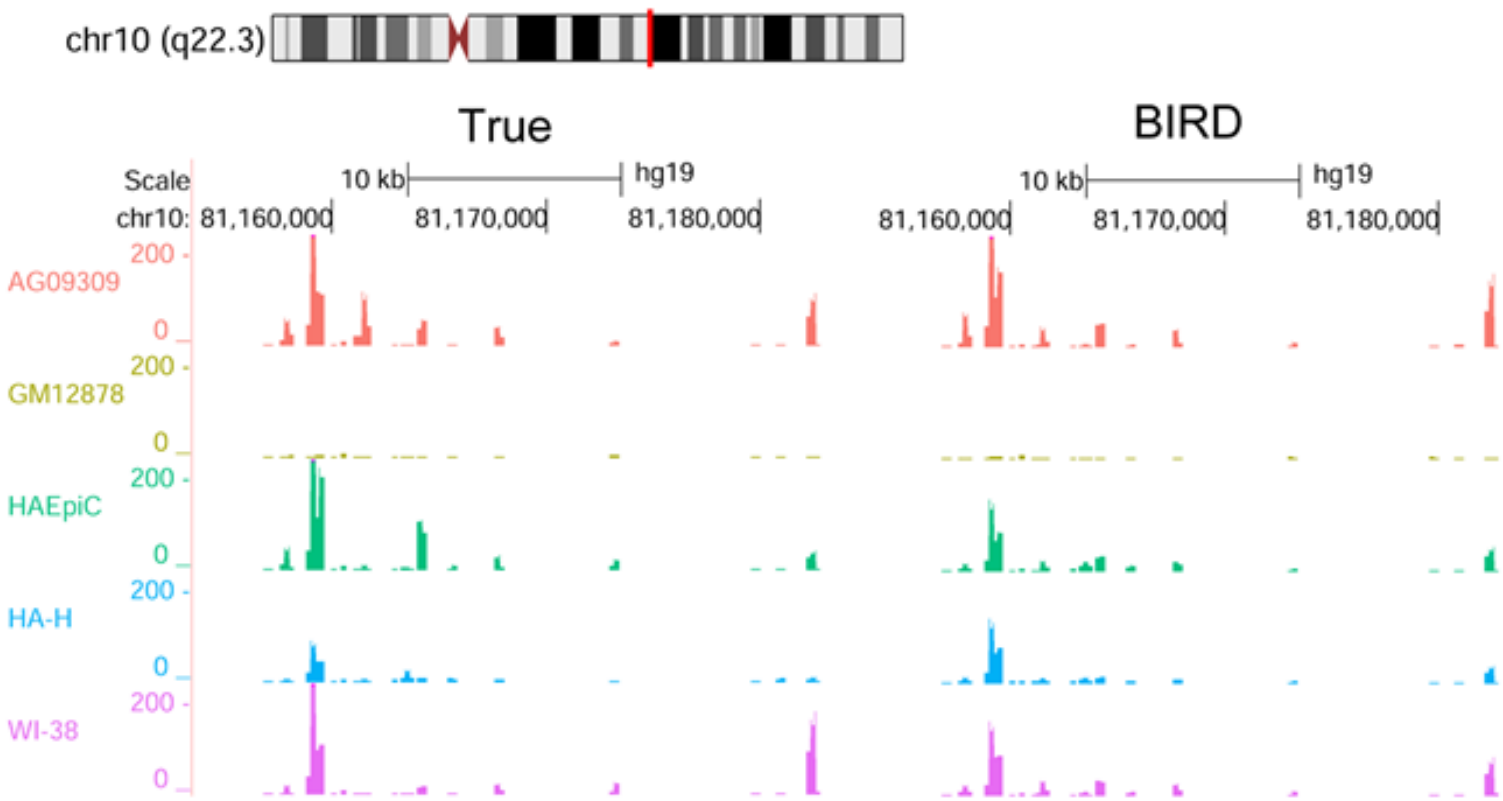


Several useful big data tools:

- BDVD (remove batch effects)
- BigClust (big data clustering)

Evaluation Based on ENCODE Data

- 57 distinct human cell lines with DNase-seq and exon array
- 40 cell types as training dataset
- 17 cell types as test dataset

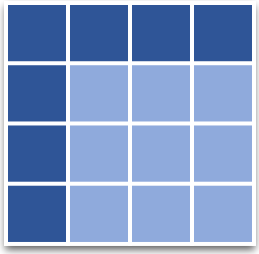


DHS Prediction Performance

Truth (Y)

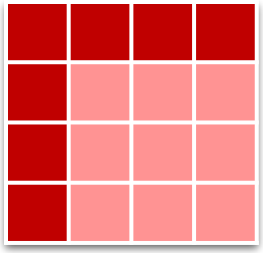
Cell type

Locus

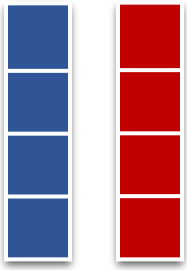


vs

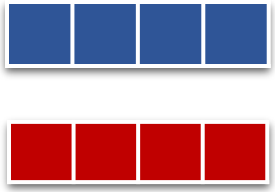
Prediction ($\hat{Y}$)



(1) Cross-Locus Correlation

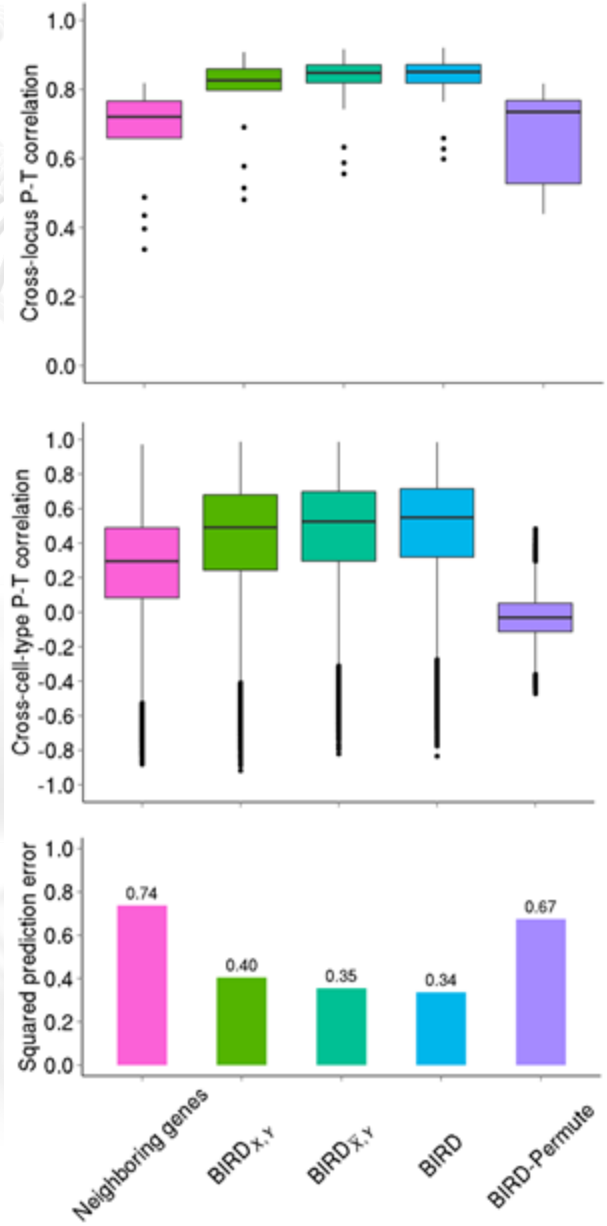


(2) Cross-Cell-Type Correlation

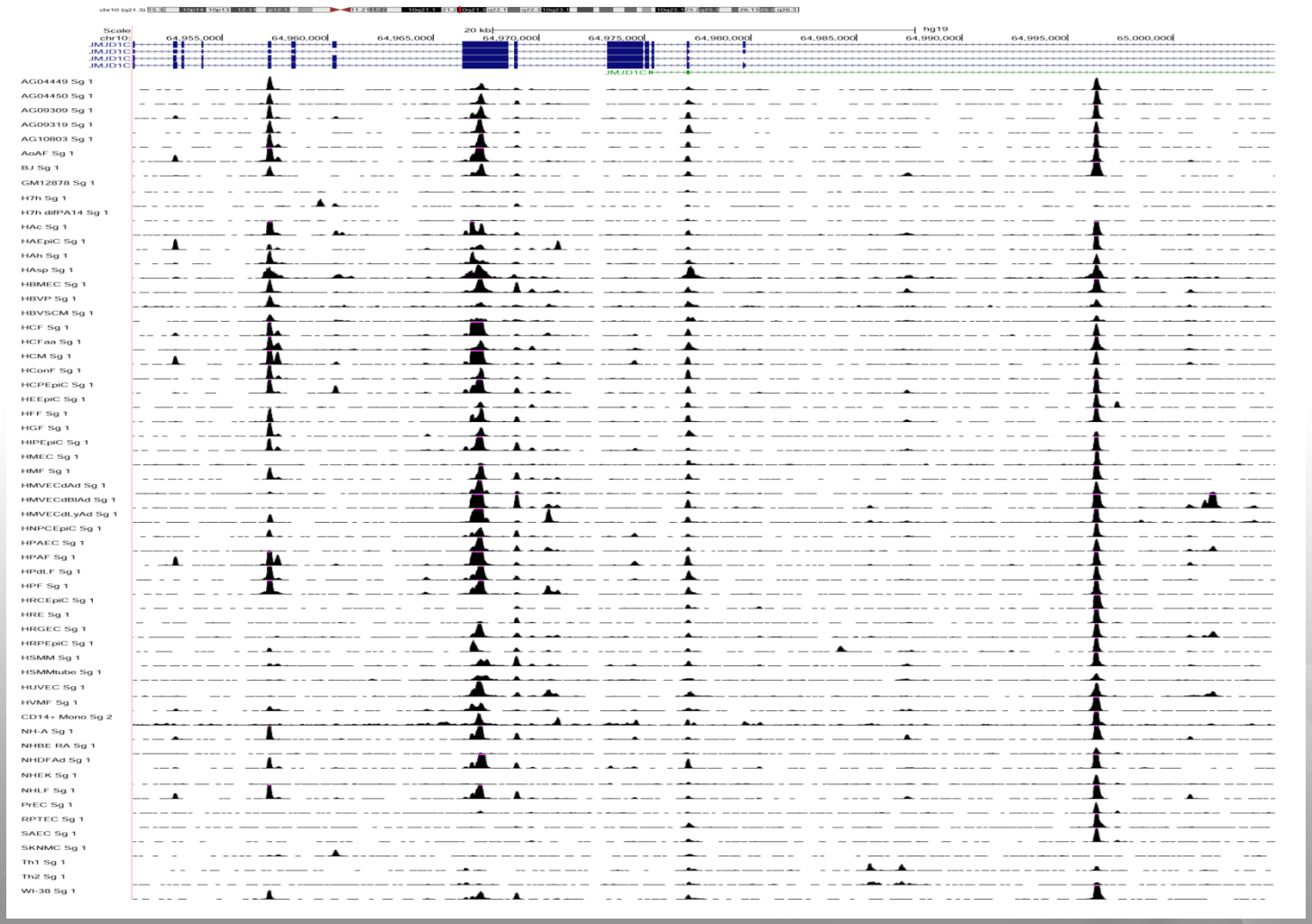


(3) Squared Pred. Err.

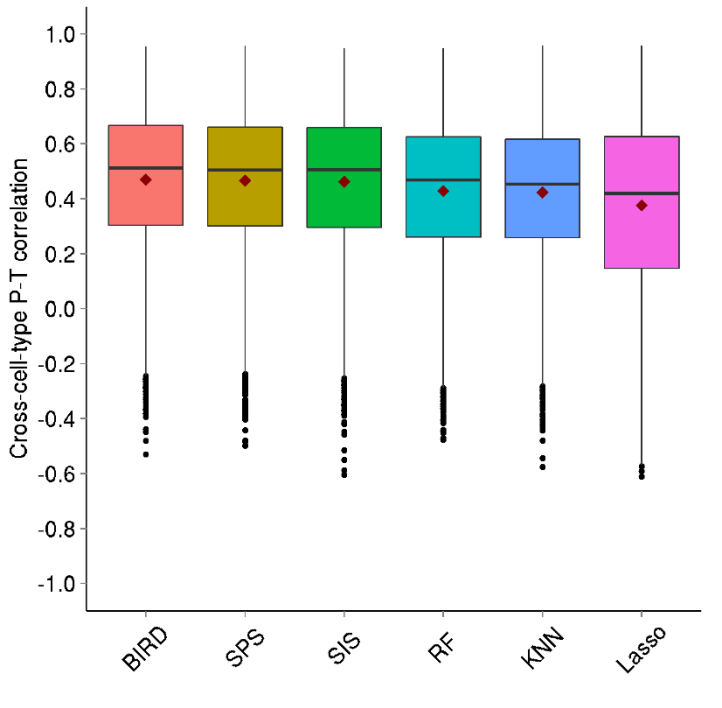
$$\tau = \frac{\sum_l \sum_m (y_{lm} - \hat{y}_{lm})^2}{\sum_l \sum_m (y_{lm} - \bar{y})^2}$$

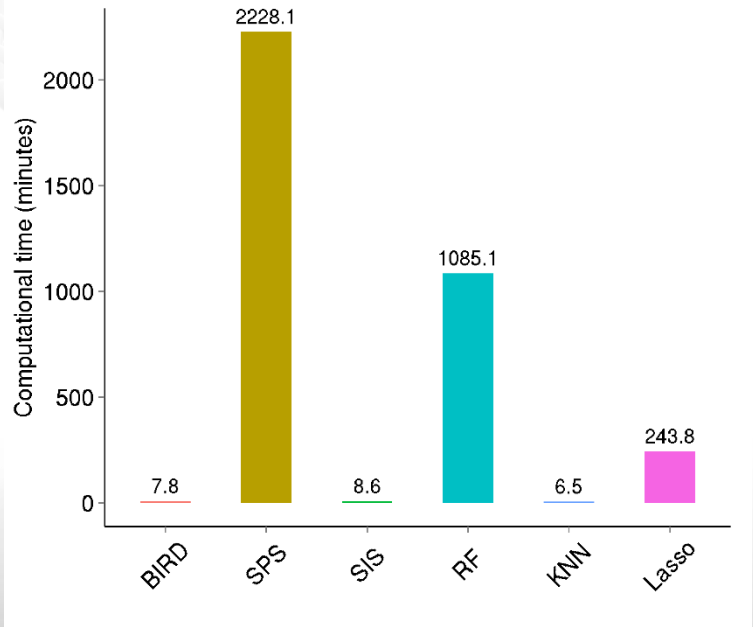


Locus Effects



BIRD vs. alternative methods

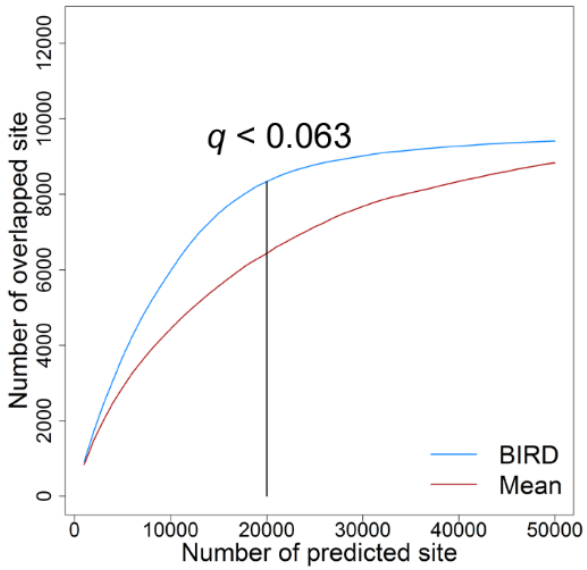


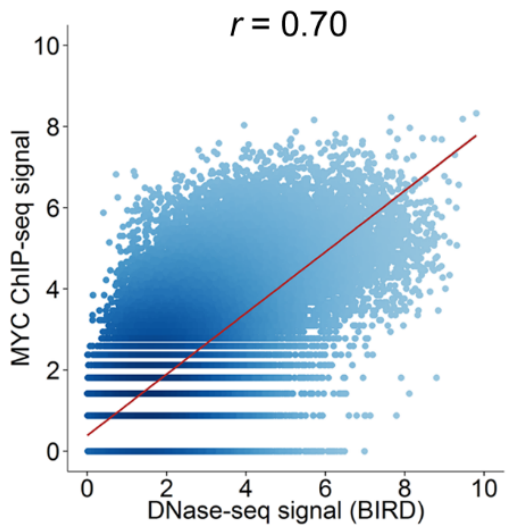


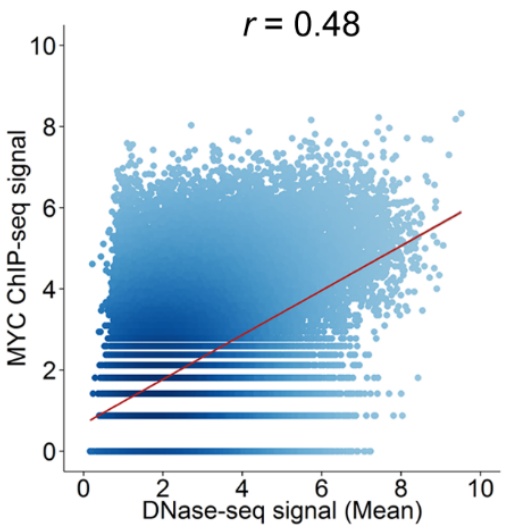
Method	BIRD	SPS	SIS	RF	KNN	Lasso
Mean r_c	0.4703	0.4667	0.4625	0.4289	0.4241	0.3757
Runtime (minute)	7.8	2228.1	8.6	1085.1	6.5	243.8

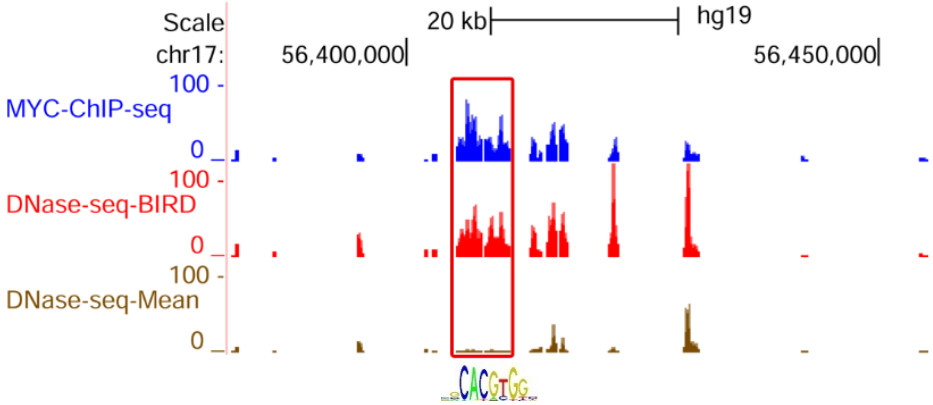
Transcription Factor Binding Site Prediction

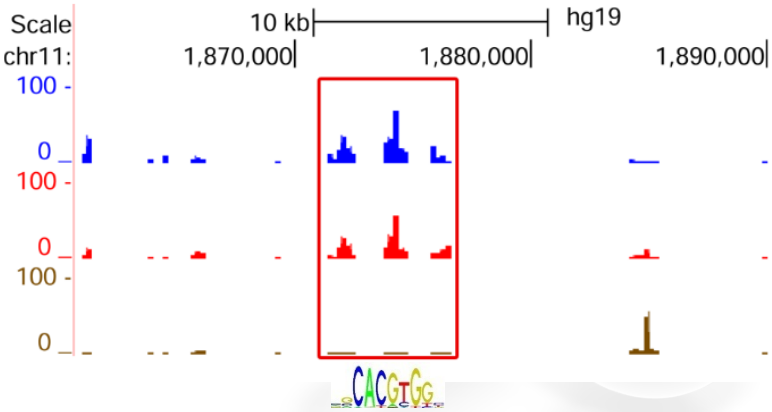
MYC binding in P493-6 B-cell lymphoma





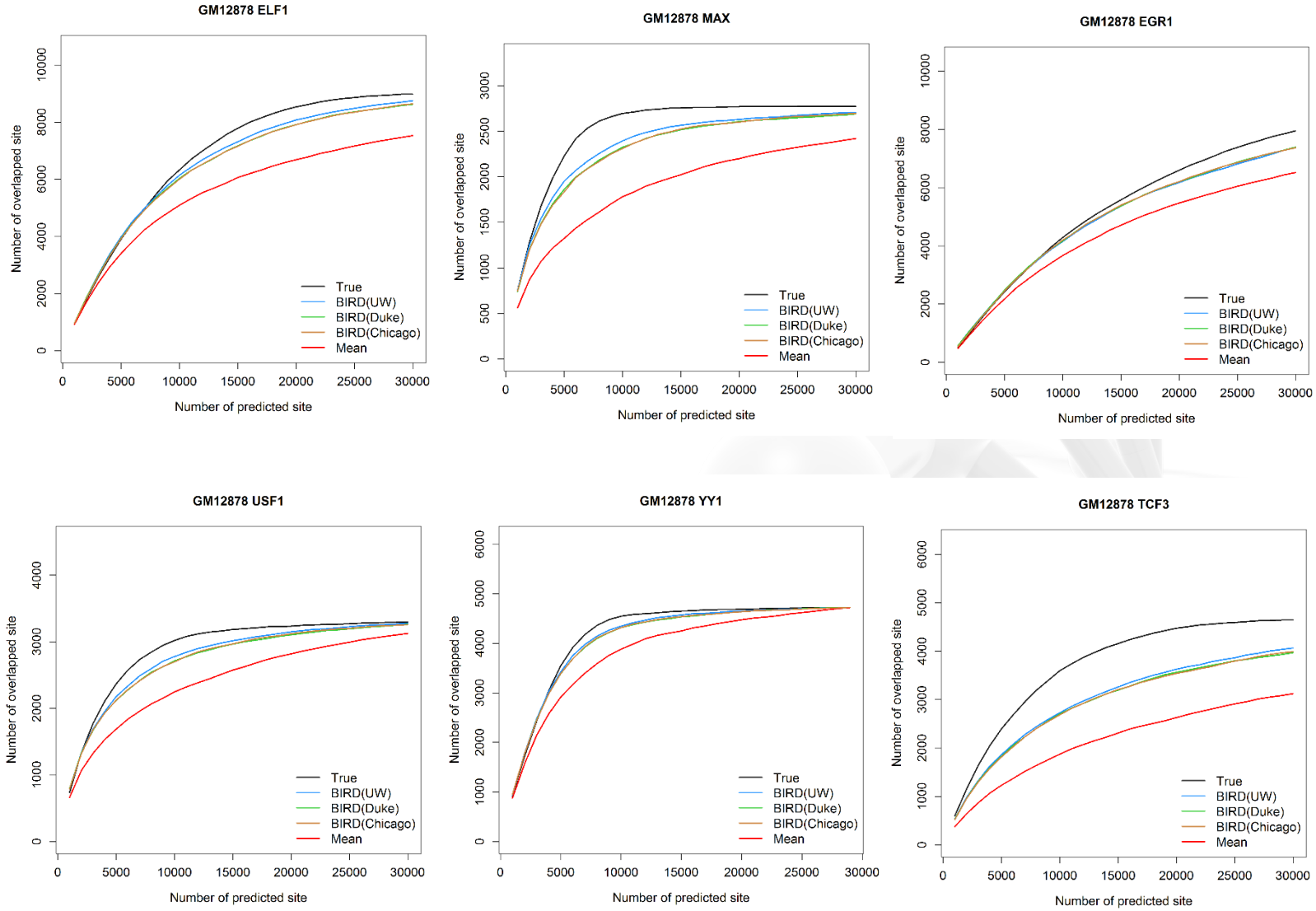






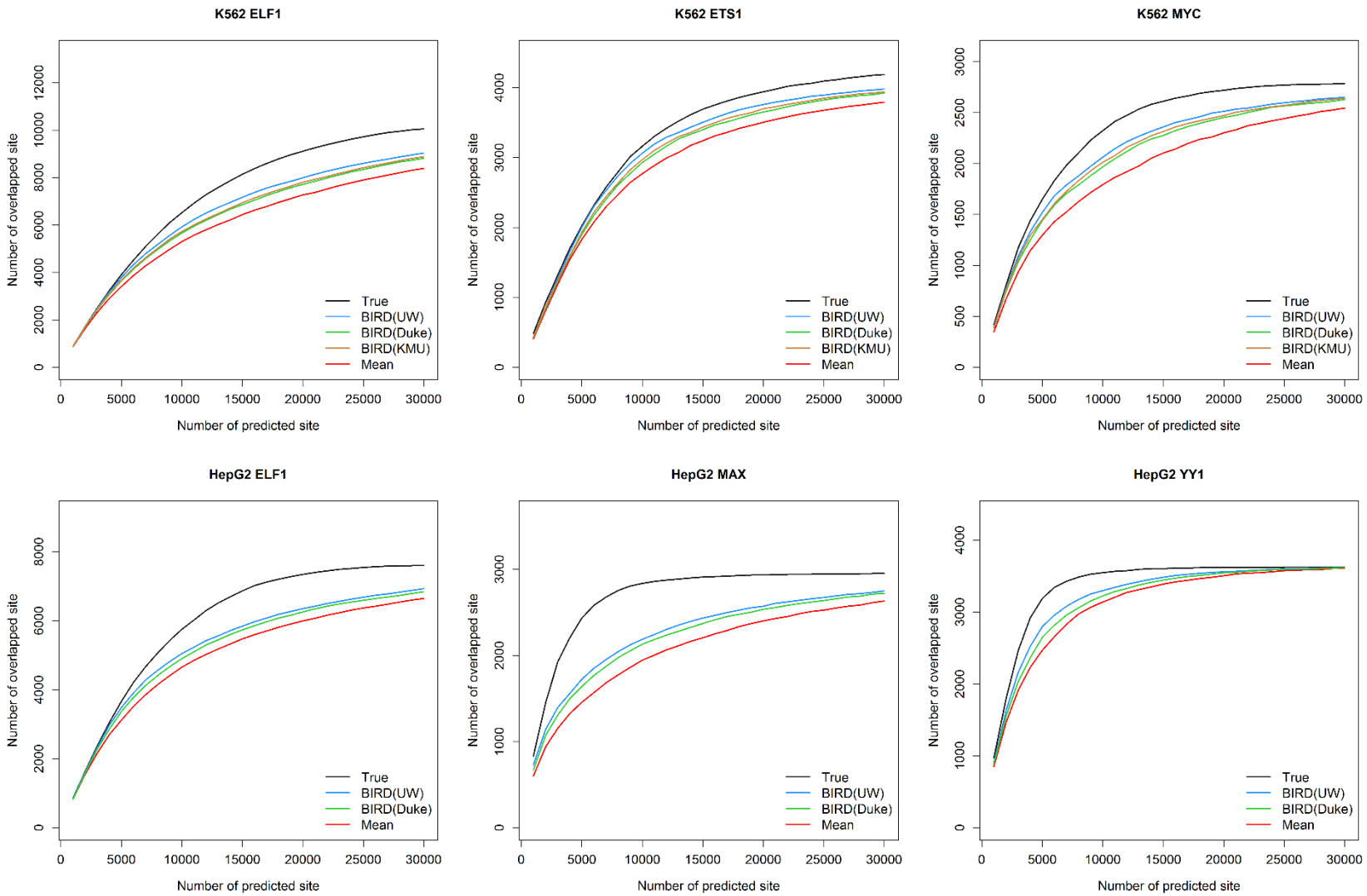
Transcription Factor Binding Site Prediction

GM12878



Transcription Factor Binding Site Prediction

K562 and HepG2



Global Prediction of DHS Using GEO

- **Apply BIRD to 2000 exon array samples in GEO**
- **Web database resource**

Upload tab delimited BED file (chr start_base_par end_base_pair):

Choose File No file chosen

File was not uploaded. Check if it is formatted correctly

Or use textfield

Example:

chr1 10000 20000

chr2 20000 50000

… … …

Use tab to separate chromosome, start base pair and end base pair

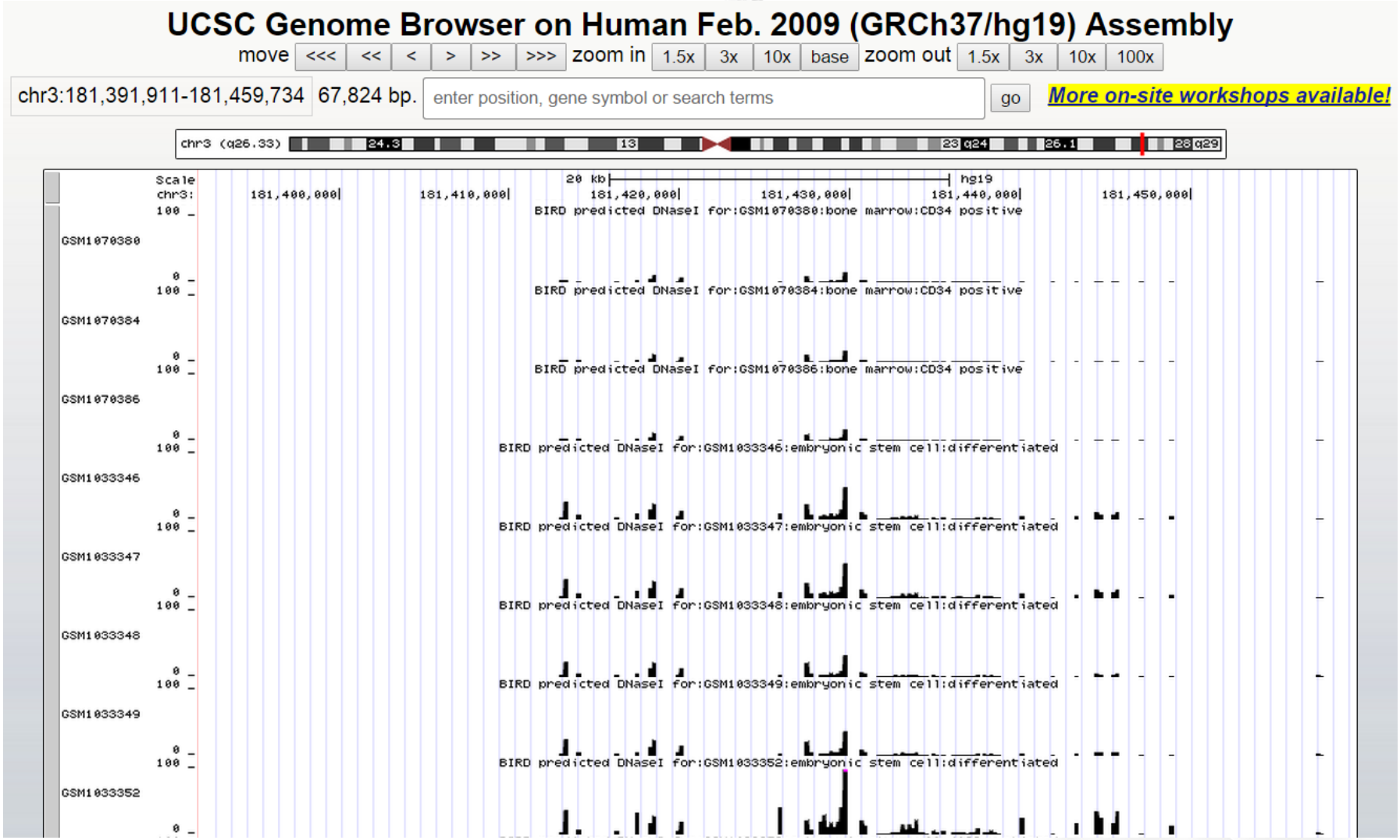
Search | Download DNase-Seq | Create Heat Map | Differential Analysis | Download Annotation D

Visualization of Predicted DNase-Seq data in UCSC Browser

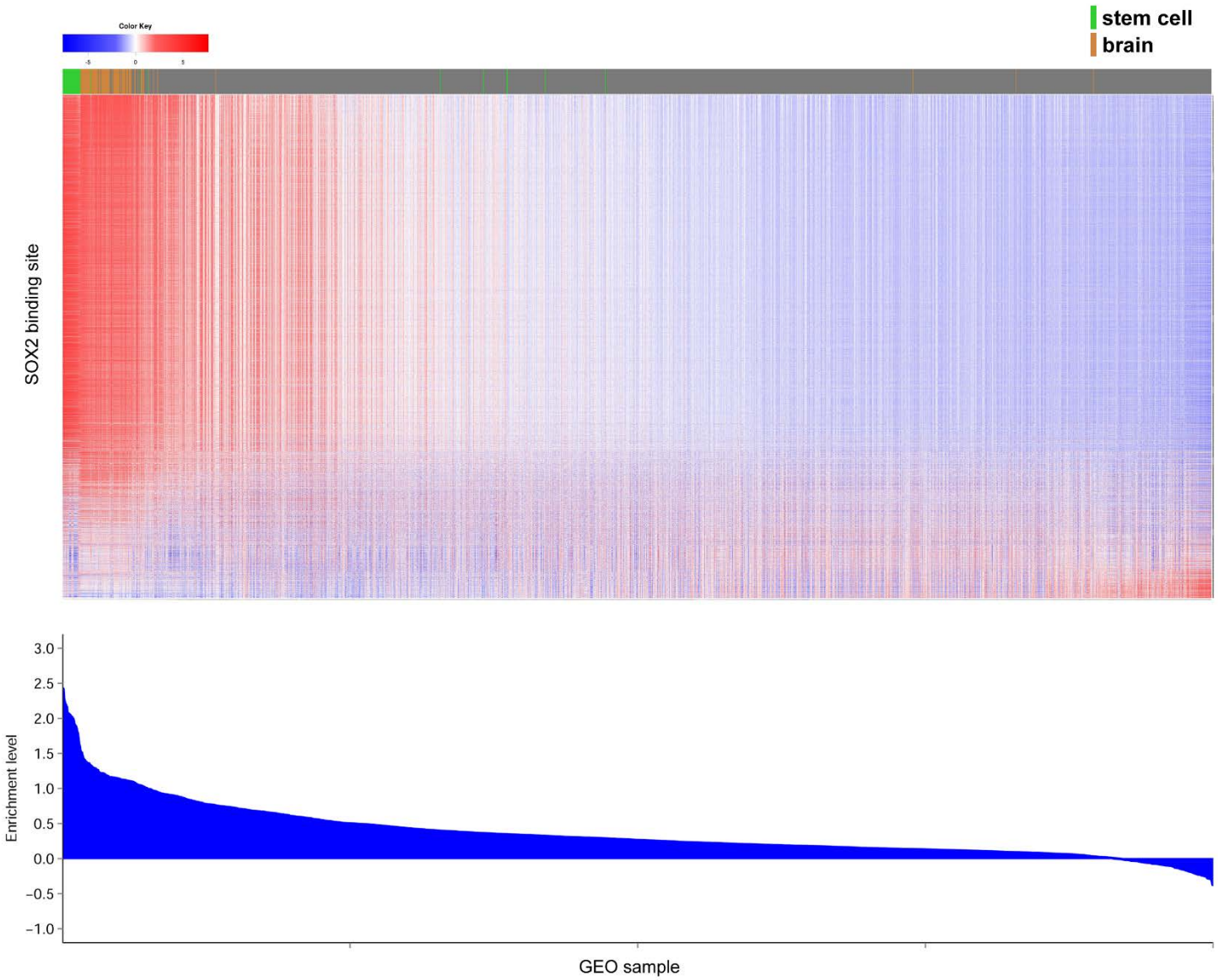
GSE: all ×	GSM: all ×	Cell Type: stem cell ×	Cell Status: all ×	Sex: all ×	Other: all ×	Chromosome: all ×
GSE19090	GSM1033346	embryonic stem cell	differentiated	NA	cell line: H7-hESC; diffProtA_5d	
GSE19090	GSM1033347	embryonic stem cell	differentiated	NA	cell line: H7-hESC; diffProtA_5d	
GSE19090	GSM1033348	embryonic stem cell	differentiated	NA	cell line: H7-hESC; diffProtA_9d	
GSE19090	GSM1033349	embryonic stem cell	differentiated	NA	cell line: H7-hESC; diffProtA_9d	
GSE19090	GSM1033350	embryonic stem cell	differentiated	NA	cell line: H7-hESC; diffProtA_14d	
GSE19090	GSM1033351	embryonic stem cell	differentiated	NA	cell line: H7-hESC; diffProtA_14d	

Global Prediction of DHS Using GEO

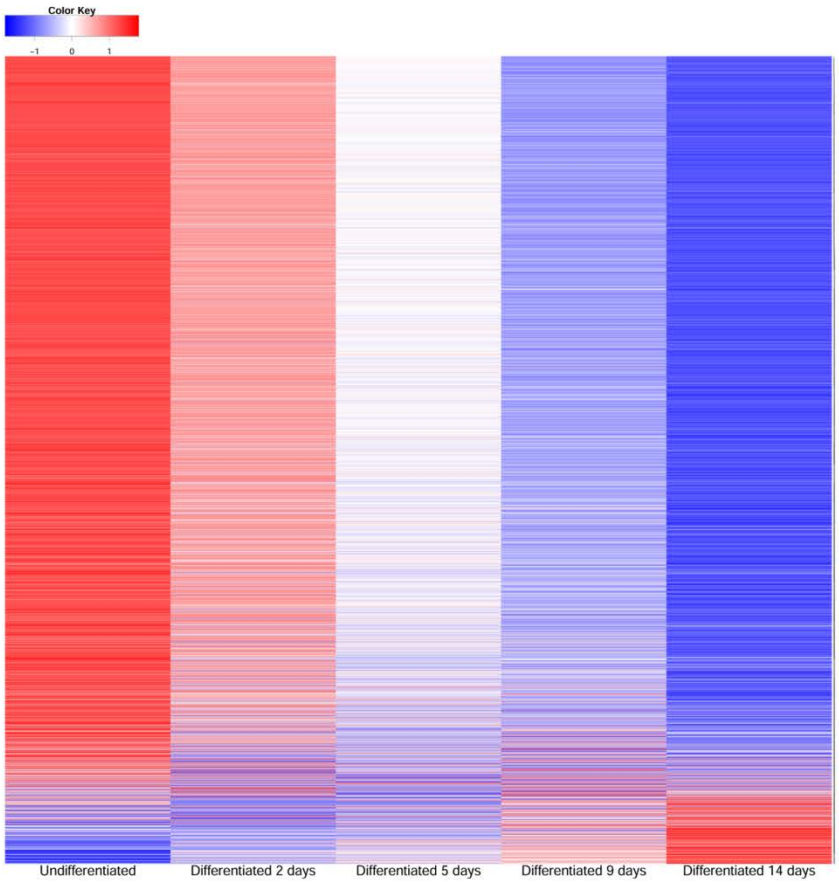
Visualization in UCSC genome browser

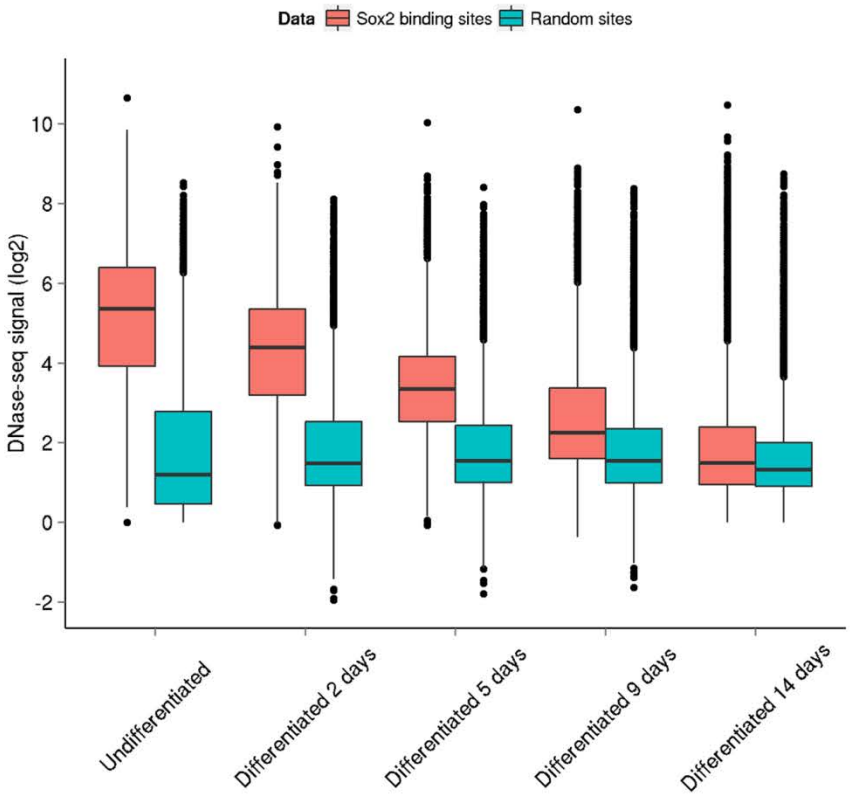


Example: SOX2 Binding Dynamics



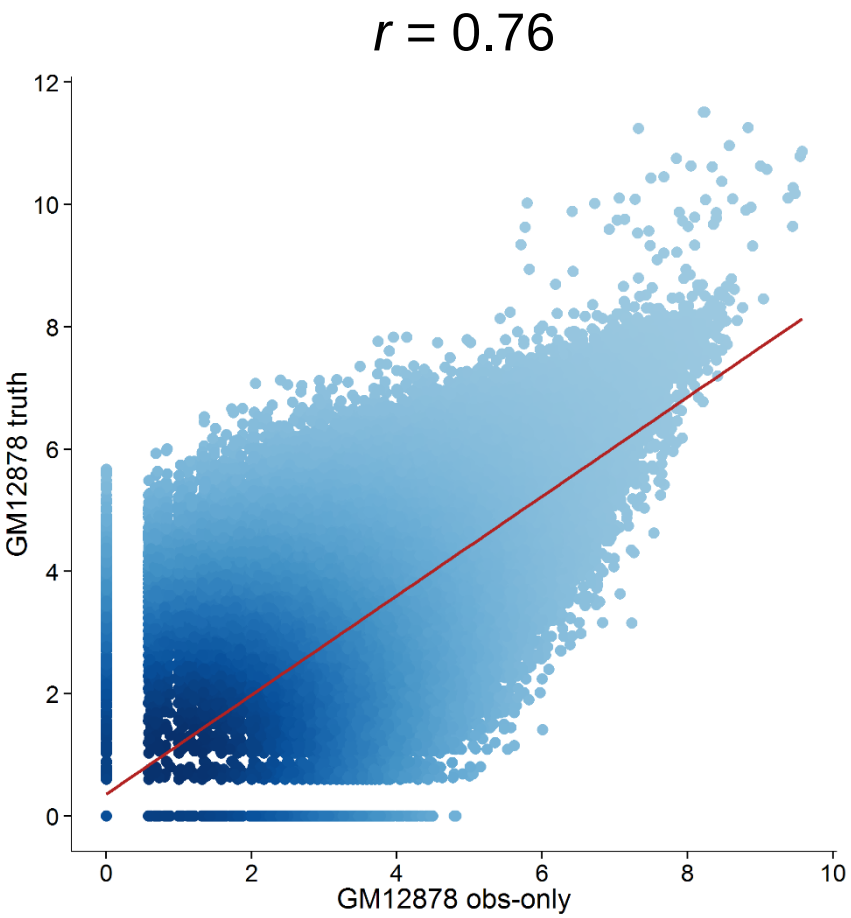
Dynamic SOX2 activity during stem cell differentiation

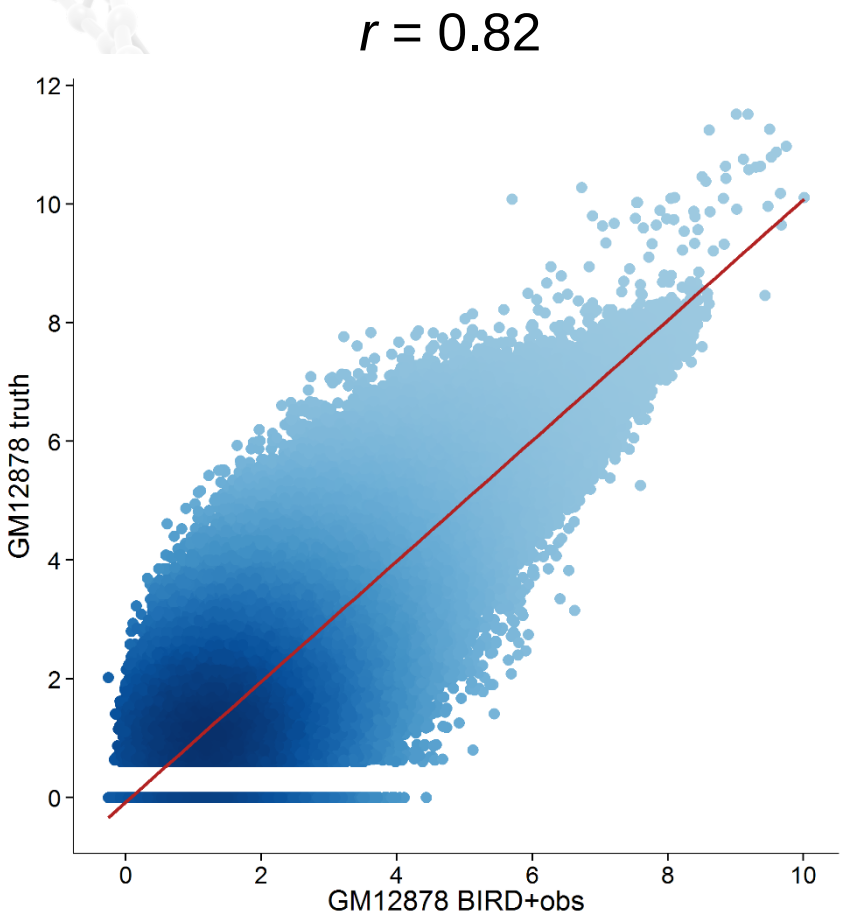




Improve Data Analysis Using Predicted DHS

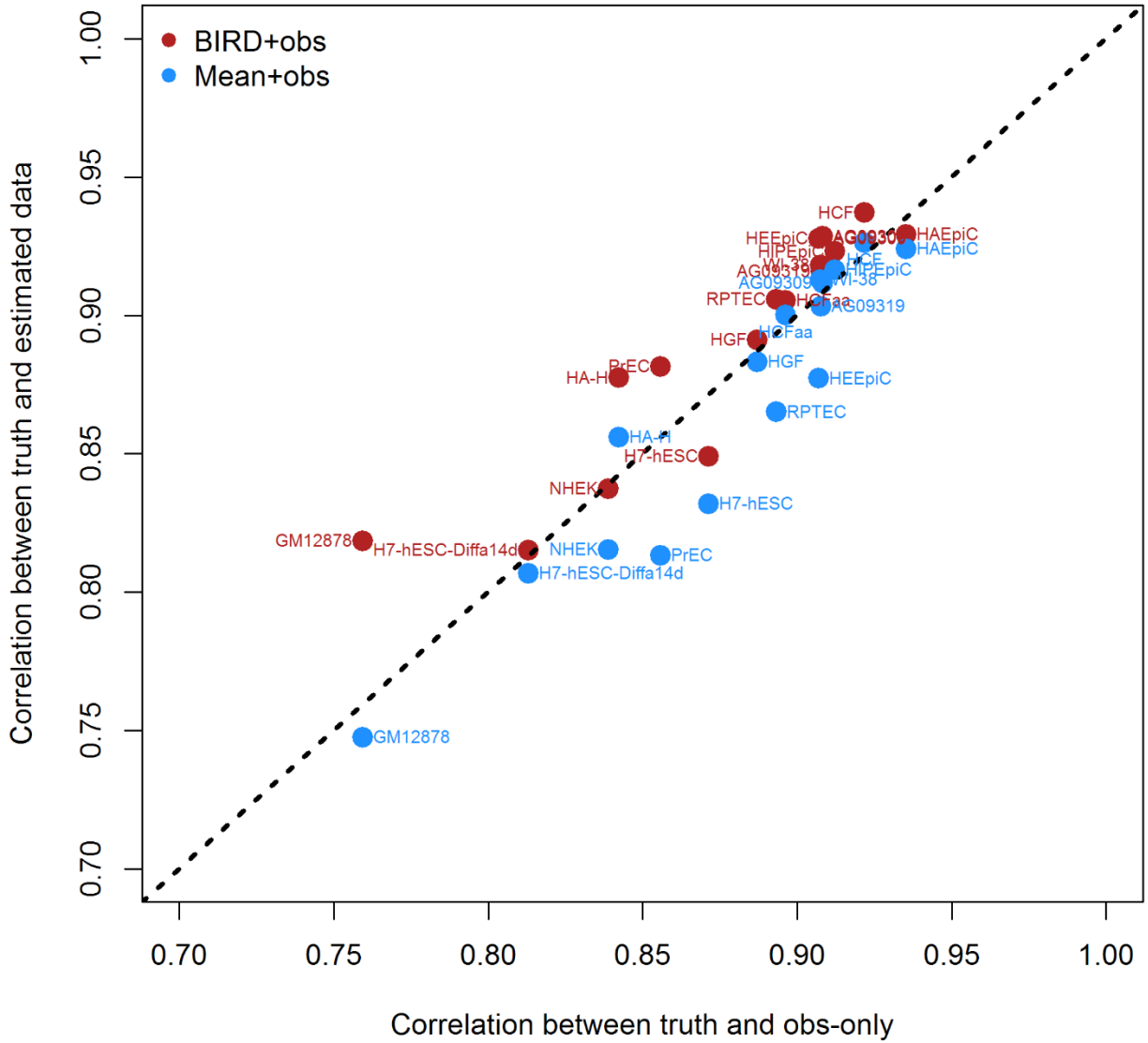
GM12878 DNase-seq





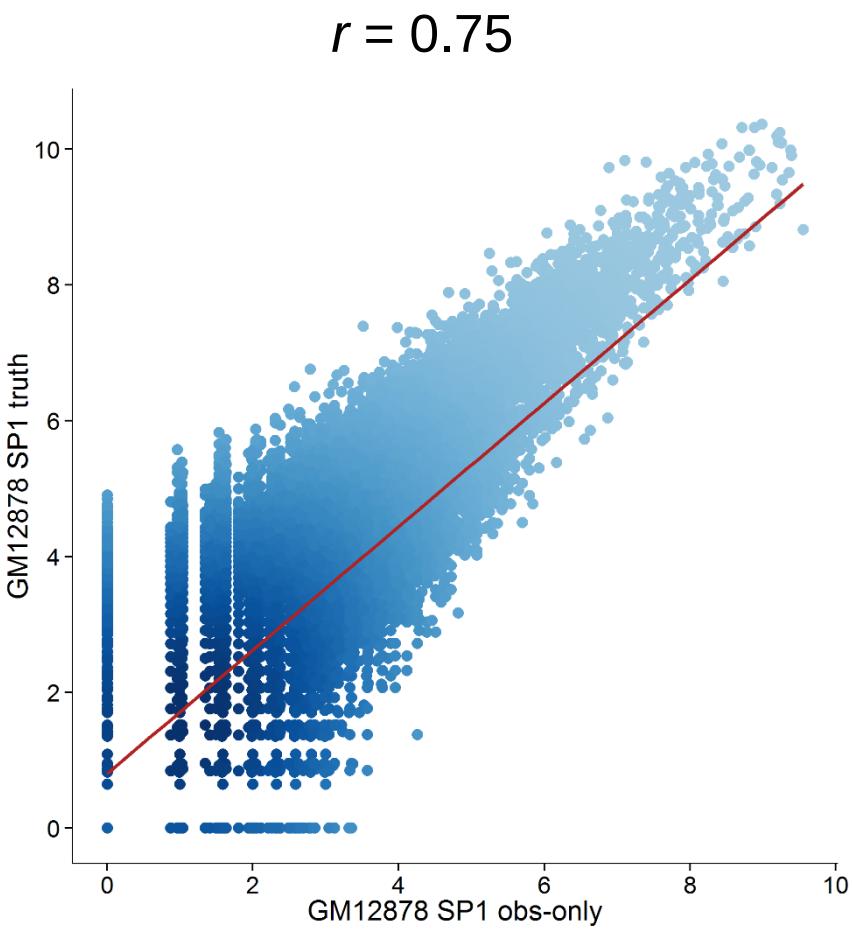
Improve Data Analysis Using Predicted DHS

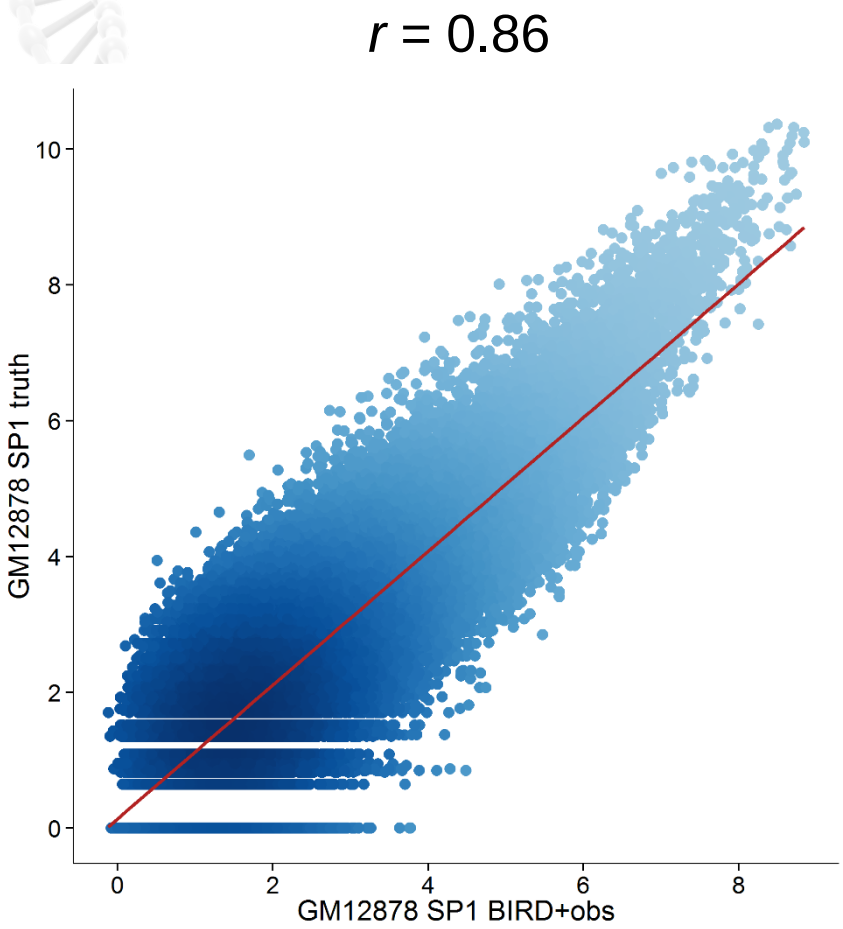
16 test cells



Improve Data Analysis Using Predicted DHS

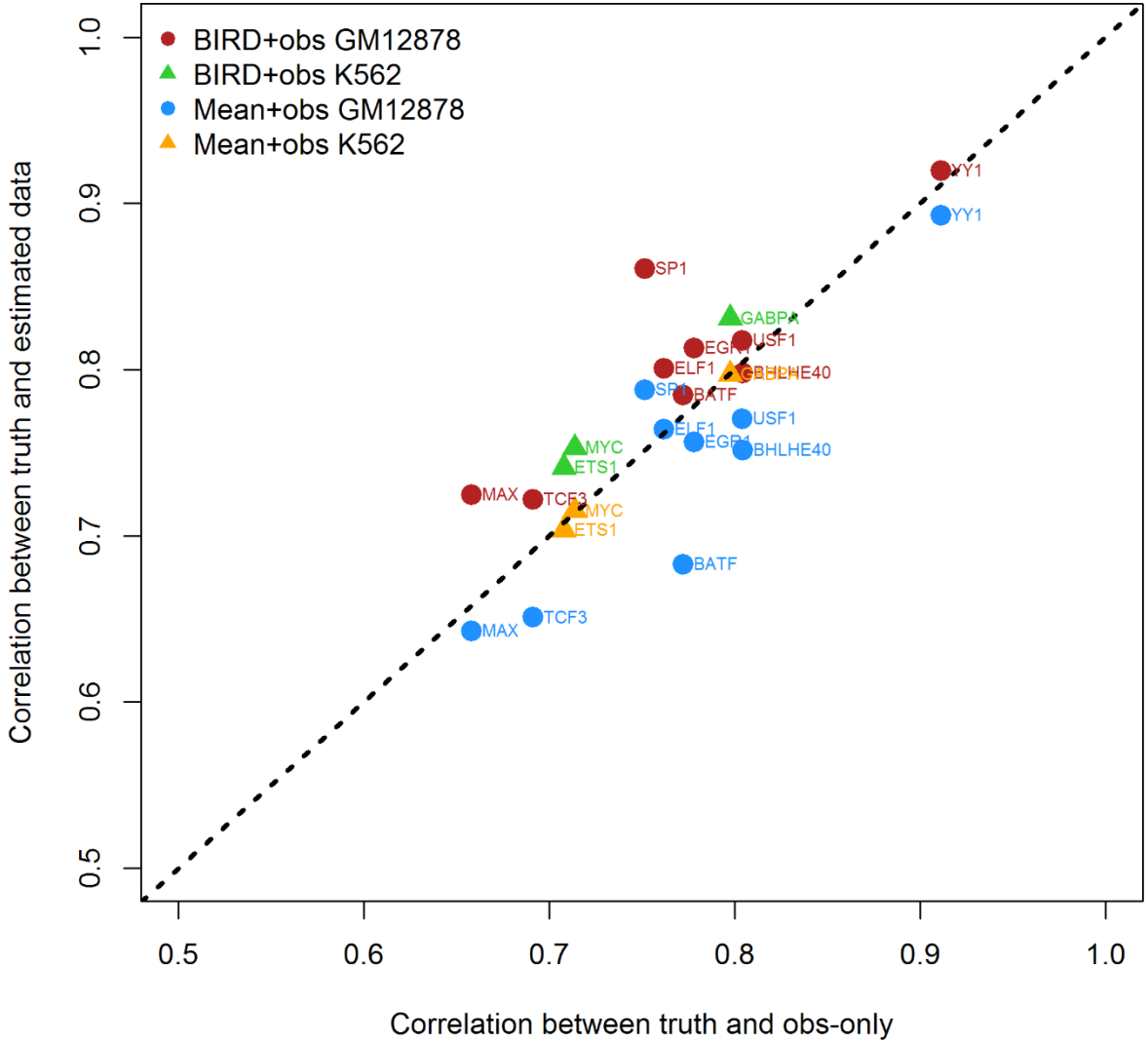
GM12878 ChIP-seq for SP1





Improve Data Analysis Using Predicted DHS

GM12878: 9 TFs, K562: 3 TFs

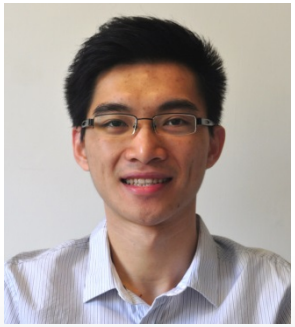


Summary

- Public data is a powerful tool to make discoveries
- BIRD: big data regression and prediction
- Regulatory element activities may be predicted using gene expression
- Prediction provides a new way to integrate two different data types

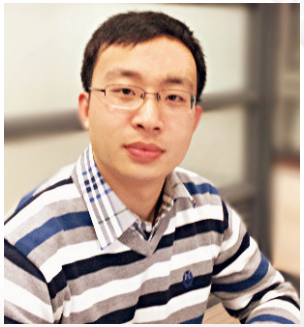
Acknowledgment

Group Members:



Weiqiang Zhou

Ben Sherwood



Zhicheng Ji

Fang Du

Jiawei Bai

Funding:

NIH R01HG006841, R01HG006282
Maryland Stem Cell Research Fund 2012-MSCRFE-0135-00