The ENCODE Data Analysis Center (DAC)

MPI

Zhiping Weng
UMMS

Mark Gerstein
Yale

Manager

Jill Moore
UMMS

Co-I

Roderic Guigo
CRG

Rafa Irizarry
Harvard

Manolis Kellis
MIT

Anshul Kundaje
Stanford

Shirley Liu
Harvard

Bill Noble
UW

Registry of candidate cis-Regulatory Elements (ccREs)

- Collection of putative regulatory regions defined using DNase-seq and H3K4me3, H3K27ac, and CTCF ChIP-seq

- V0 – Currently available at portal and SCREEN: 1.31 M ccREs in human (hg19) and 432 k in mouse (mm10)

- V1 – soon to be released, featured in Encyclopedia manuscript: 1.46 M ccREs in human (hg38) and 499 k in mouse (mm10)
    - 300+ more experiments than V1
    - Updated methodology to achieve finer resolution and classification of elements

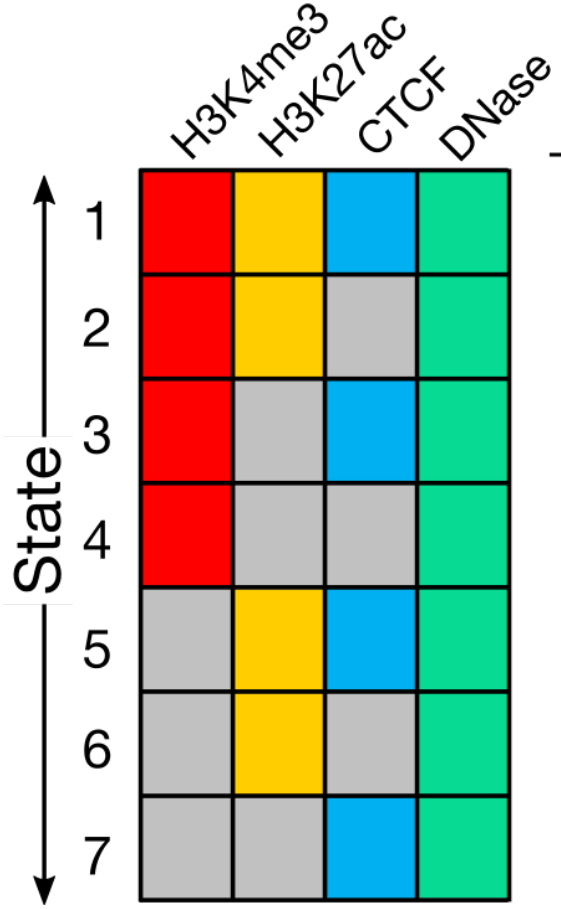
Jill Moore, Henry Pratt, Zhiping Weng

Updates to the Registry of ccREs: rDHSs for Anchoring ccREs

- Worked with the Stam lab to select a list of biosamples with DNase-seq data to build representative DHSs (rDHSs).
- Developed a new method of calling DNase hotspots in individual samples to account for differences in sequencing depth and then compile these DNase hotspots in the list of biosamples into rDHSs.
- Worked with the Stam lab to ensure that the rDHSs were congruent with the Stam lab's consensus DHSs (cDHSs).
- The final rDHSs form the collection from which ccREs are selected based on DNase, H3K4me3, H3K27ac, and CTCF signal levels.

Jill Moore, Henry Pratt, Zhiping Weng

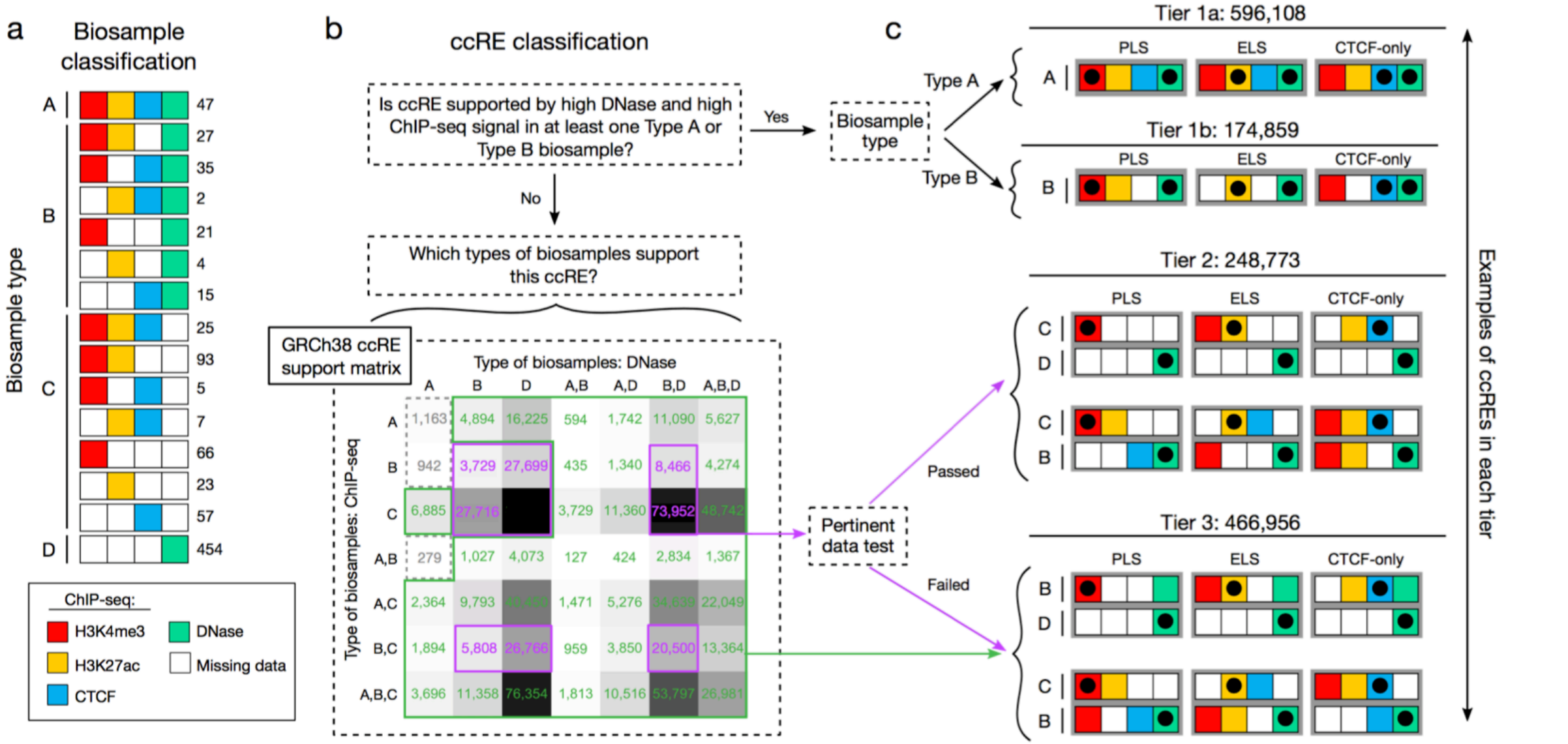
Updates to the Registry of ccREs: New classification groups



TSS ccRE center ≤ 200 bp	Proximal ccRE center > 200 bp, ≤ 2 kb	Distal ccRE center > 2 kb
PLS*	**pPLS***	**dPLS***
PLS	**pPLS**	**dPLS**
PLS*	**DNase-H3K4me3***	**DNase-H3K4me3***
PLS	**DNase-H3K4me3**	**DNase-H3K4me3**
pELS*	**pPLS***	**dPLS***
pELS	**pELS**	**dPLS**
CTCF-only	**CTCF-only**	**CTCF-only**

Jill Moore, Zhiping Weng
ccRE group (Wold, Hardison, Berstein, Stam, Snyder, Gerstein, Kundaje)

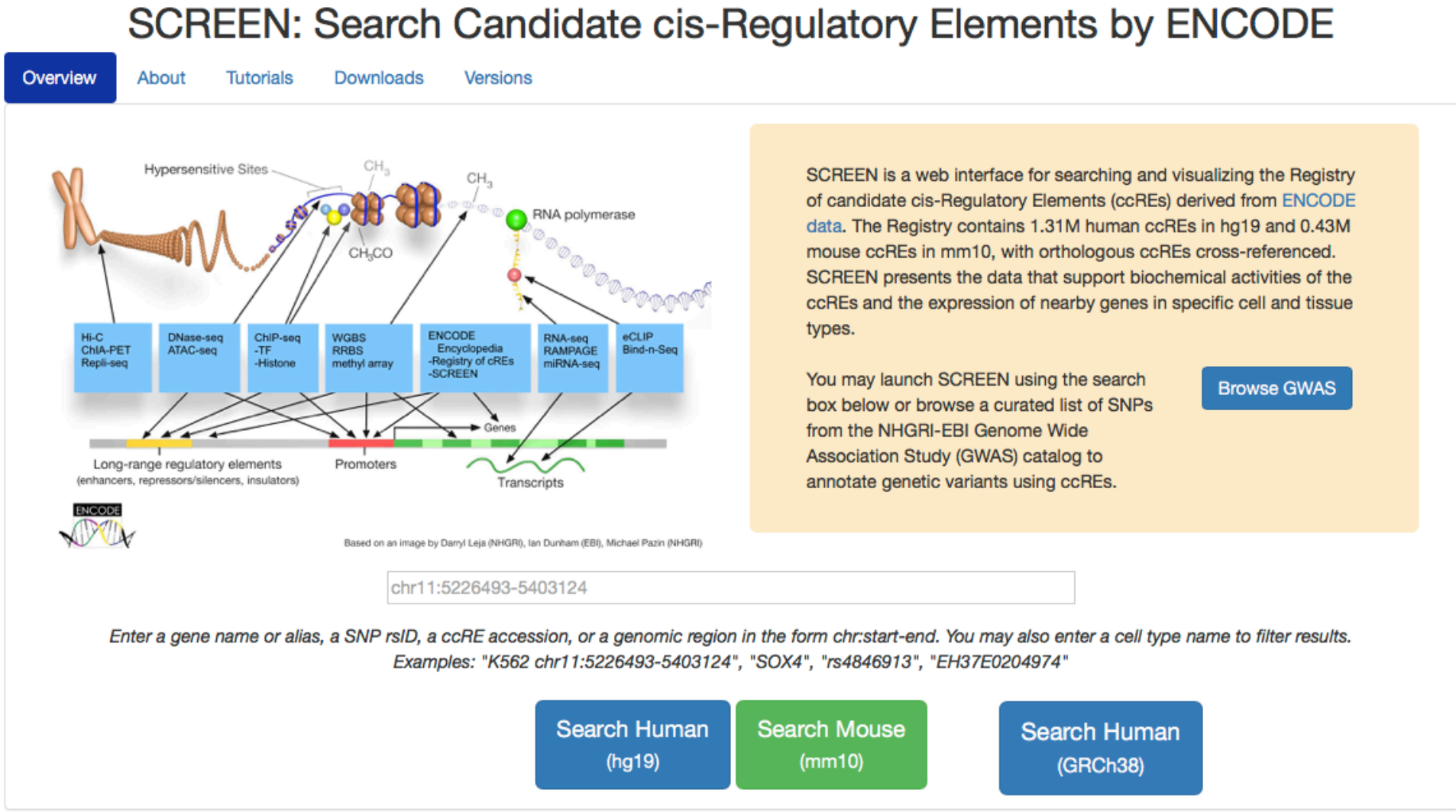
Updates to the Registry of ccREs: New tier system



Jill Moore, Zhiping Weng

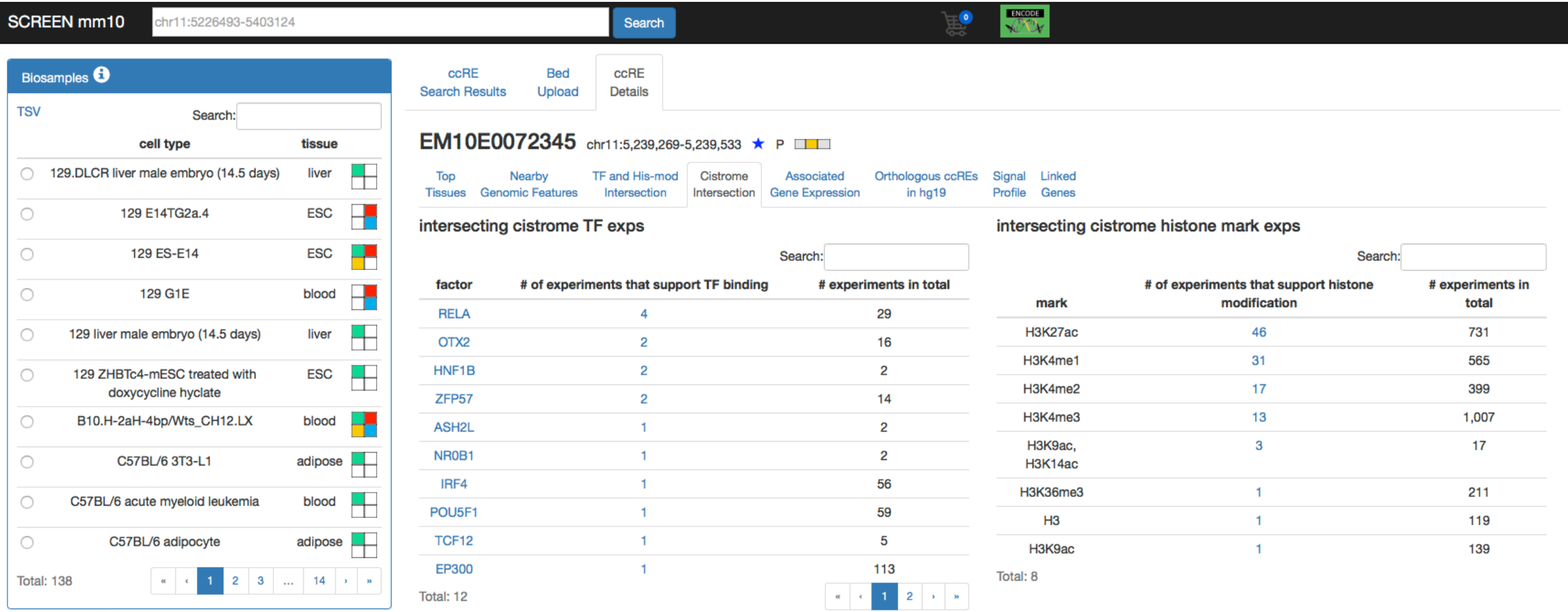
ccRE group (Wold, Hardison, Berstein, Stam, Snyder, Gerstein, Kundaje)

Search Candidate cis-Regulatory Elements by ENCODE (SCREEN)



Henry Pratt, Jill Moore, Zhiping Weng

SCREEN-Cistrome Interface (MM10 and GRCh38)



Henry Pratt, Zhiping Weng, Shirley Liu

Cistrome DB Toolkit http://cistrome.org/db

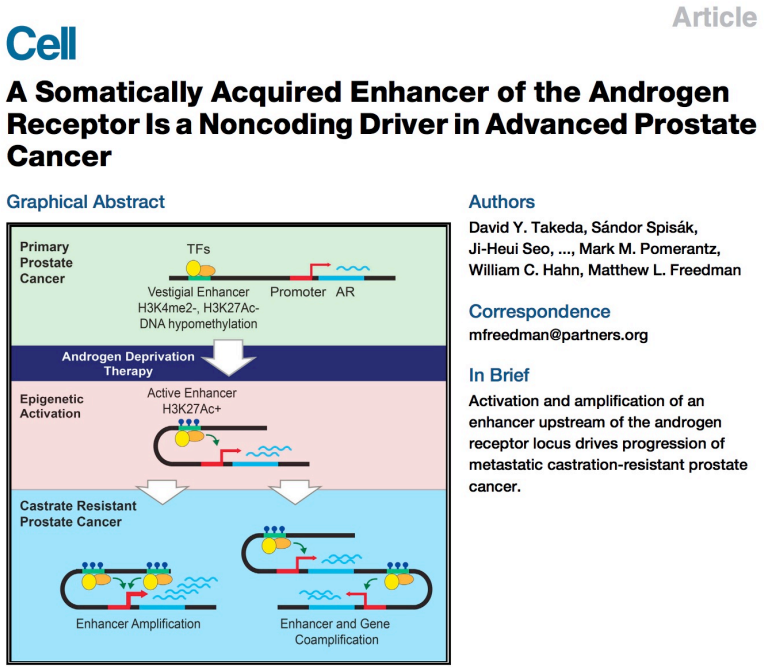
① Which factors bind to your genomic interval of interest?

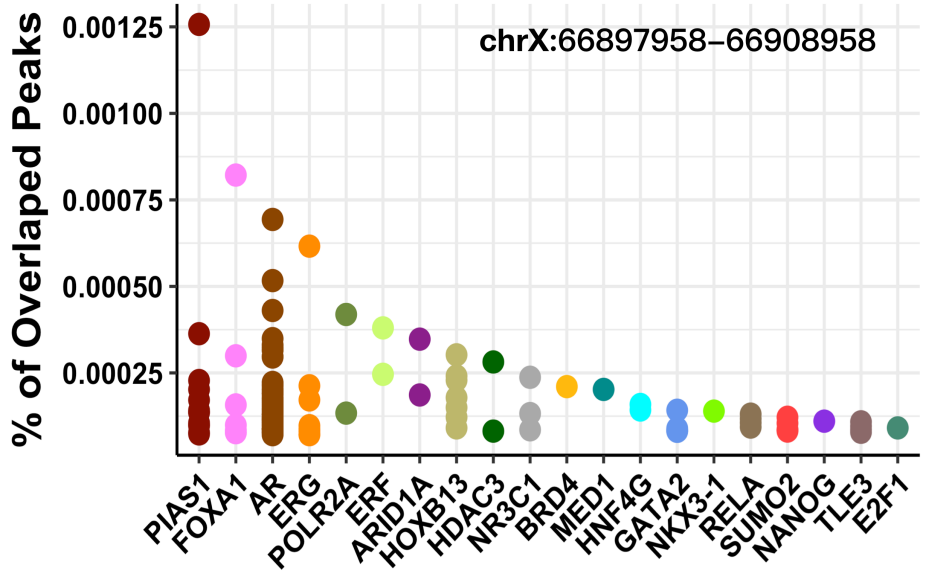
② Which factors regulate your gene of interest?

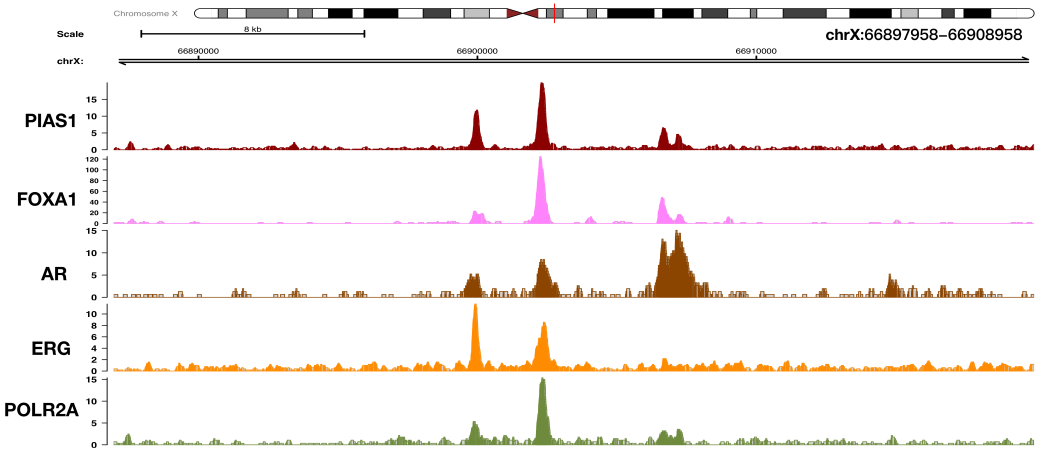
③ Which cistromes in the Cistrome DB are similar to your cistrome of interest?

Rongbin Zheng, Cliff Meyer, Shirley Liu

Which factors bind to your genomic interval?

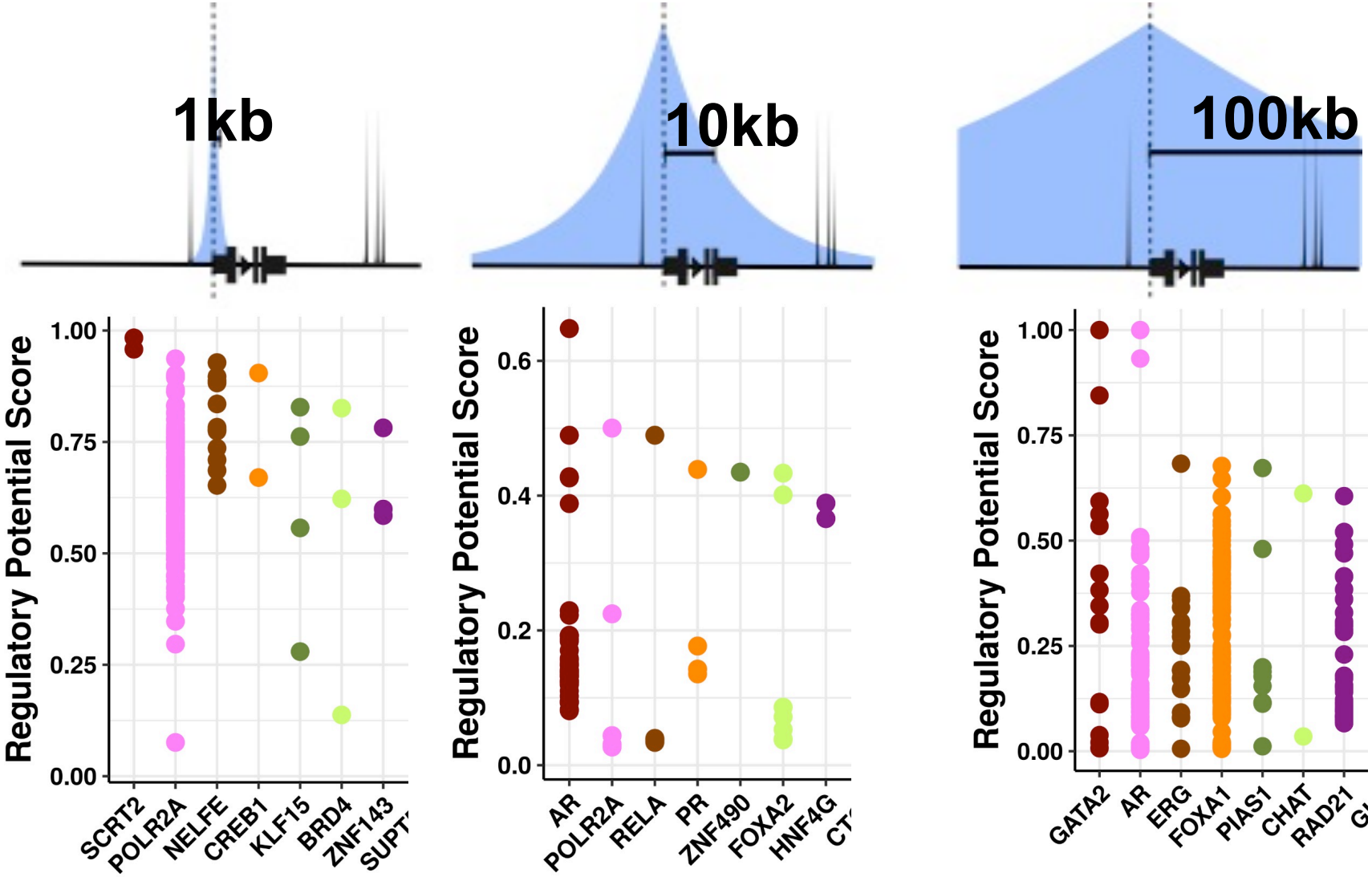






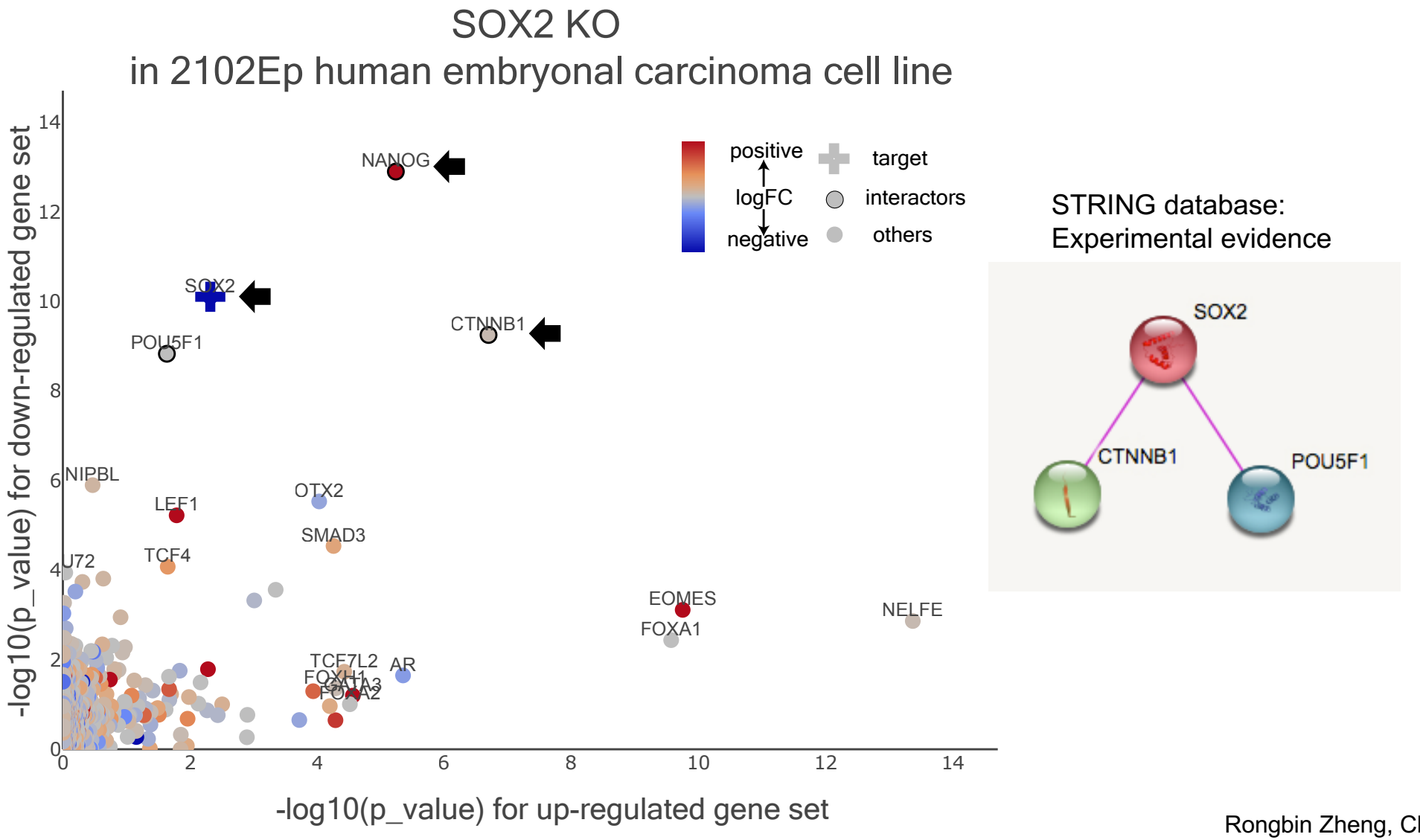
Rongbin Zheng, Cliff Meyer, Shirley Liu

Which TFs regulates your gene?



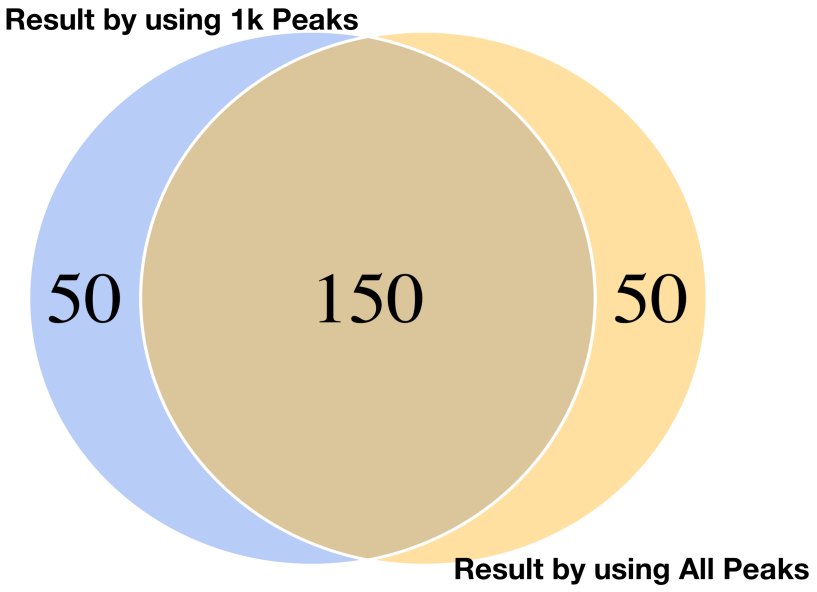
Rongbin Zheng, Cliff Meyer, Shirley Liu

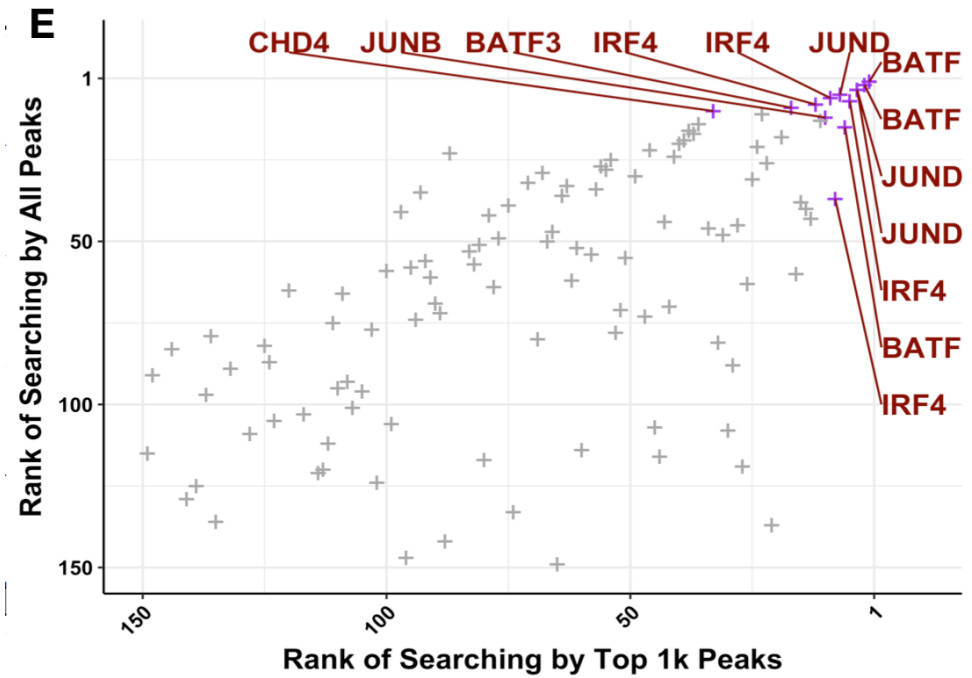
Co-regulatory TFs identified using Cistrome DB ChIP-seq data



Rongbin Zheng, Cliff Meyer, Shirley Liu

Which Cistrome DB cistromes are similar to your cistrome?





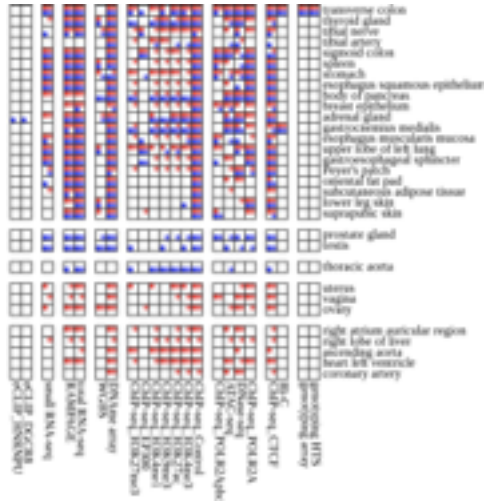
Rongbin Zheng, Cliff Meyer, Shirley Liu

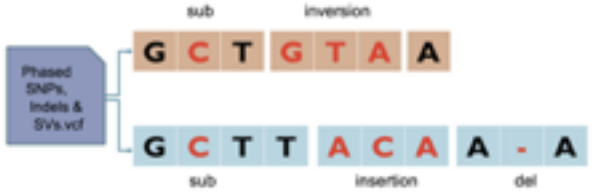
EN-TEx effort & the DAC

Involves Gerstein, Guigo & Noble labs

Cataloguing genomic elements in personal genomes with matched ENCODE datasets & the full breadth of assays on many tissues

1. Collect matched WGS & functional genomic data across ~20 tissues for 4 individuals.
2. Catalogue & compare genomic elements between individuals and tissues.
3. Study utility of phased personal diploid genomes

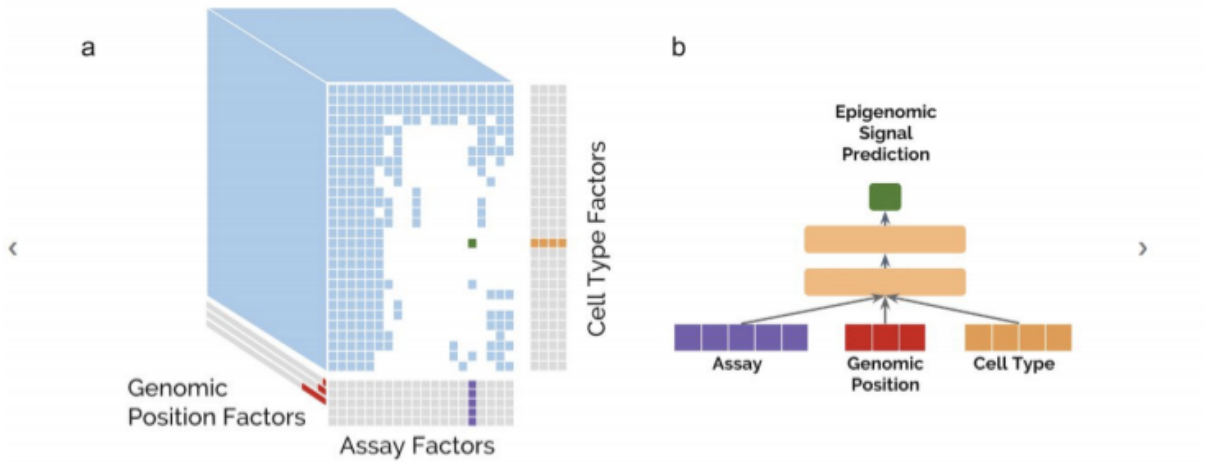




Chromatin & ChIP analysis

The DAC has continued to develop methods for the large-scale analysis of ENCODE ChIP-seq data, many applied to EN-TEx

Avocado method for Imputation of Chromatin



UNIVERSITY OF WASHINGTON

Avocado: Multi-scale Deep Tensor Factorization Learns a Latent Representation of the Human Epigenome

Jacob Schreiber[1], Timothy Durham[2], Jeffrey Bilmes[1, 3], and William Noble[1, 2]

1. Paul G. Allen School of Computer Science and Engineering, University of Washington
2. Department of Genome Science, University of Washington
3. Department of Electrical Engineering, University of Washington

Download Model → Download Imputations → Read Paper → Read Supplement →

https://noble.gs.washington.edu/proj/avocado

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | AB | CHANNEL

Search

New Results

Multi-scale deep tensor factorization learns a latent representation of the human epigenome

Jacob Schreiber, Timothy J Durham, Jeffrey Bilmes, William Stafford Noble

doi: https://doi.org/10.1101/364976

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract | Info/History | Metrics | Supplementary material | Preview PDF

Abstract

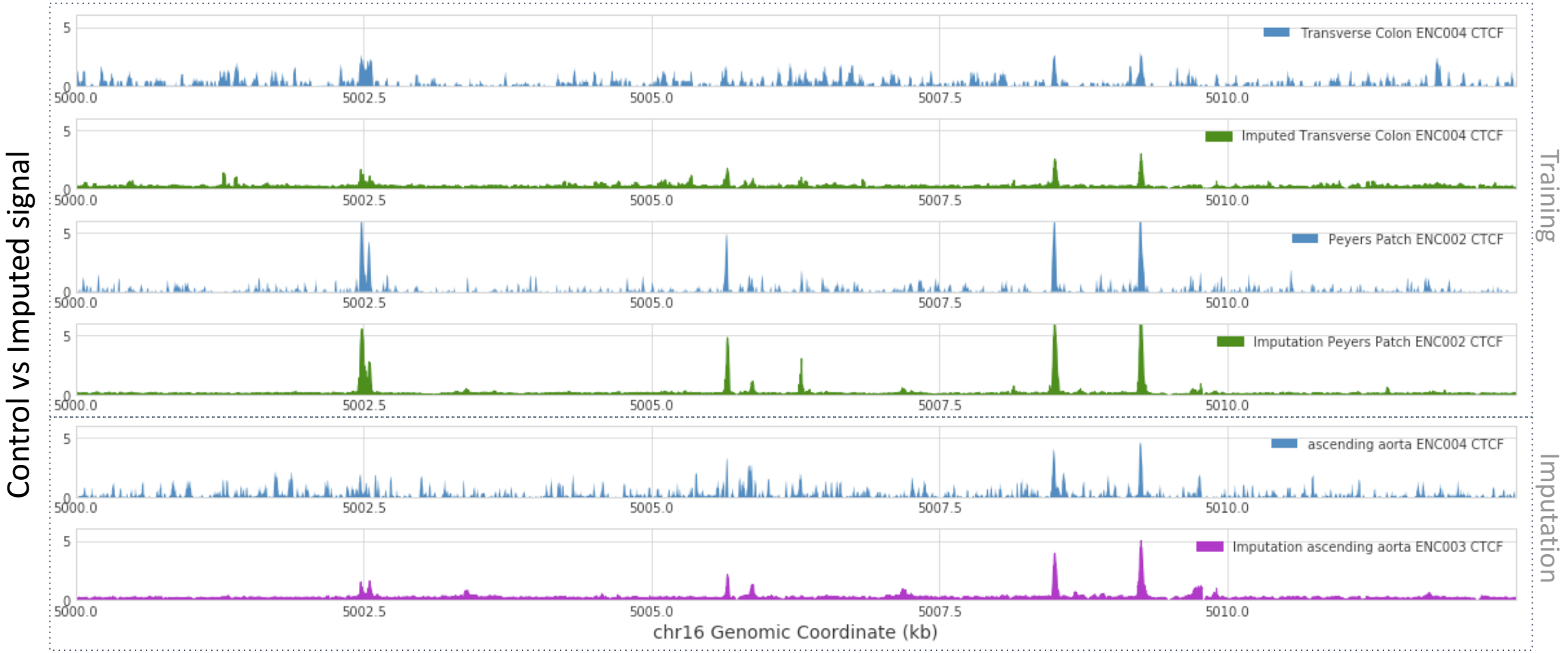
The human epigenome has been experimentally characterized by measurements of protein binding, chromatin acessibility, methylation, and histone modification in hundreds of cell types. The result is a huge compendium of data, consisting of thousands of measurements for every basepair in the human genome. These data are difficult to make sense of, not only for humans,

Jacob Schreiber, Bill Noble

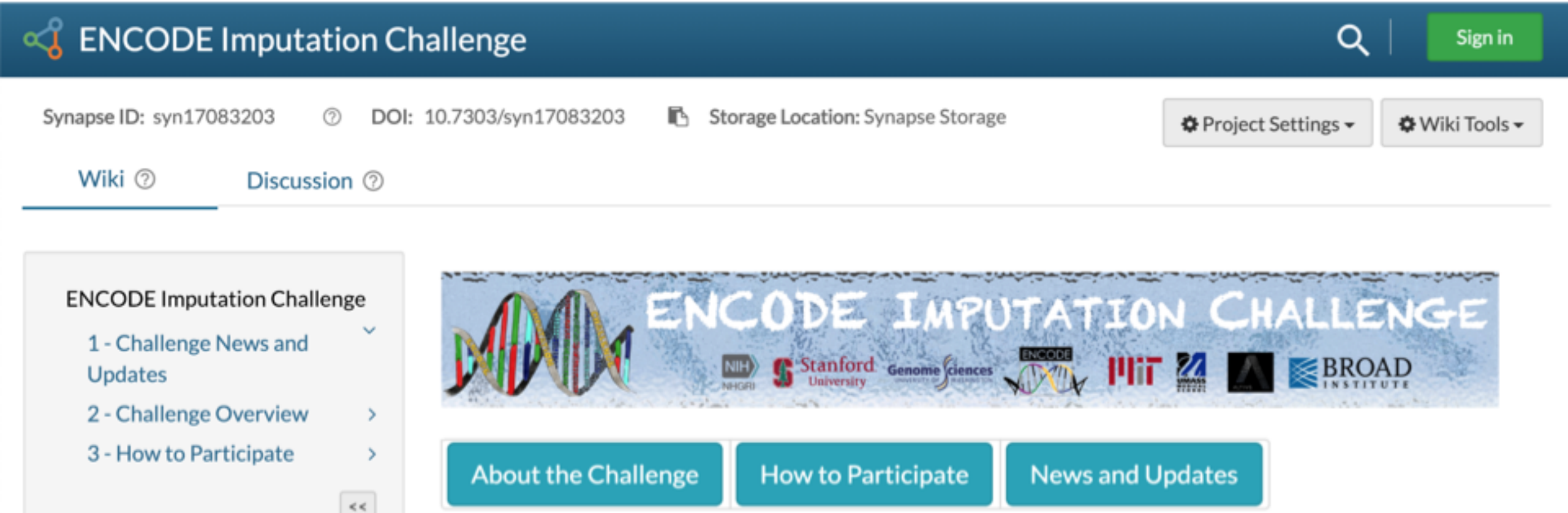
Application of Avocado to the ENTEx Dataset

910 sample imputations from Avocado (Deep learning model)

Histone modifications + CTCF + RNApol II



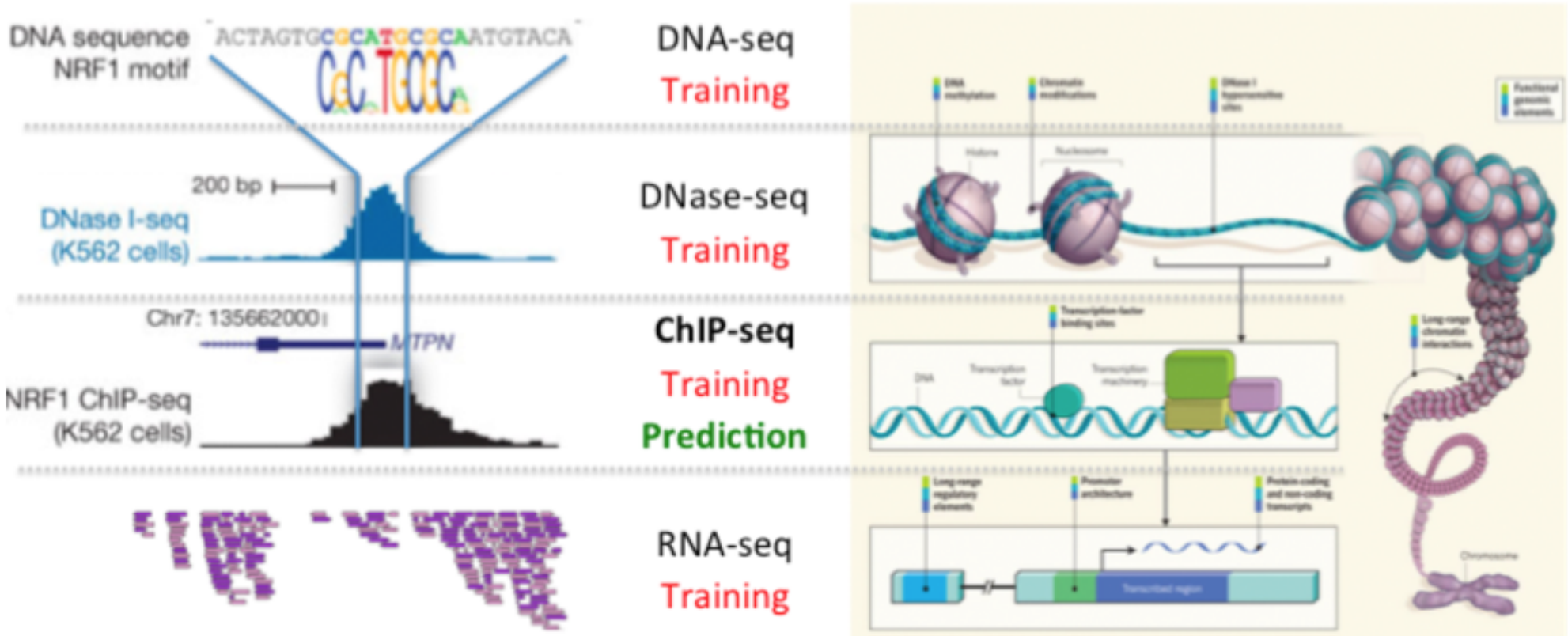
Jacob Schreiber, Bill Noble



https://www.encodeproject.org/encodeimpute

Jacob Schreiber, Bill Noble

Participants will predict histone ChIP-seq



Jacob Schreiber, Bill Noble

Dimensionality reduction & consistent visualization of transcriptome & epigenetic data across cell types

PCA: Principal Component Analysis
cPCA: Contrastive PCA
RCA: Reference Component Analysis
LDA: Linear Discriminant Analysis

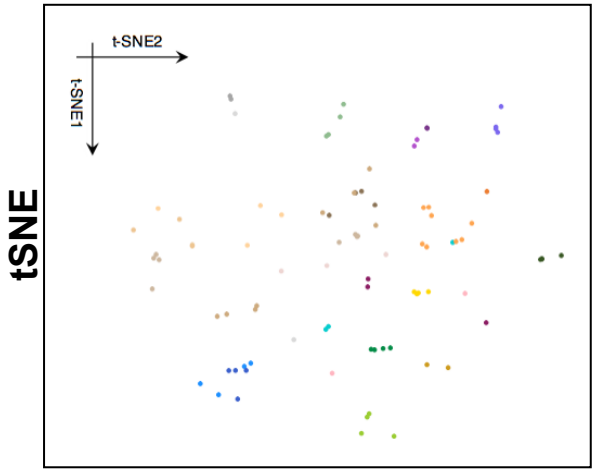
Decompose the variance

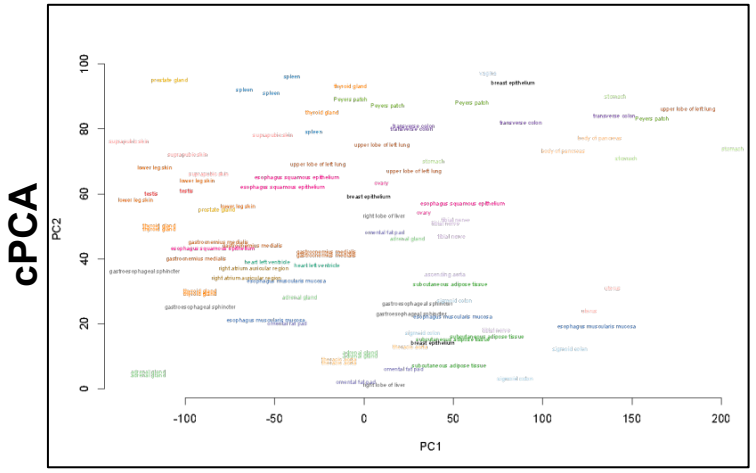
PCA/tSNE — **cPCA** — **RCA** — **LDA**

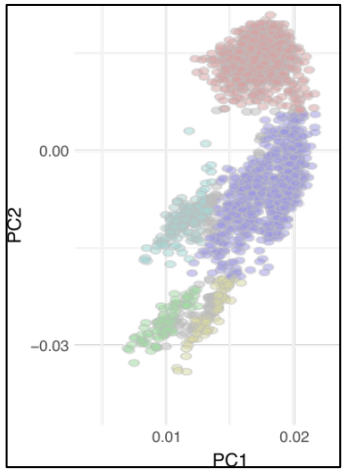
Traditional PCA is fully unsupervised and operates on an individual sample. Can we make the models aware of the structure?

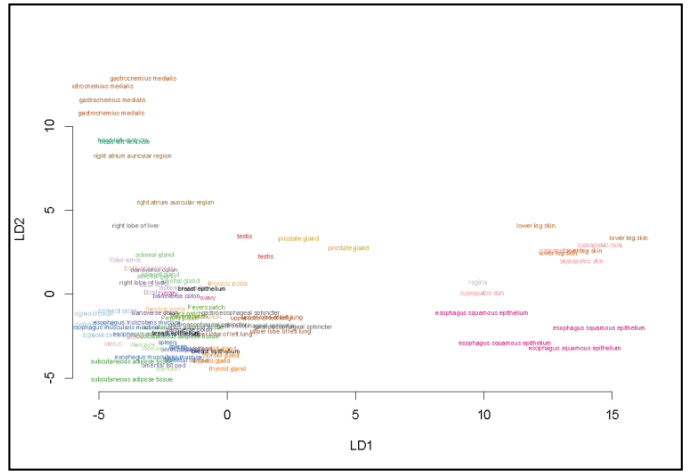
Unsupervised;
no assumption on the structure

More supervised;
with assumption on the structure









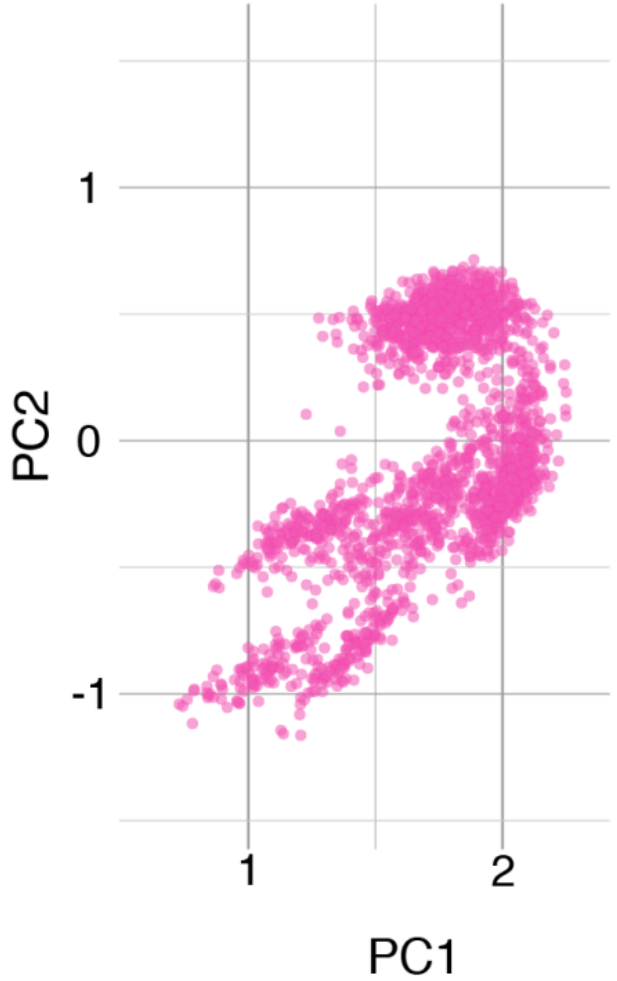
- For example in EN-TEx, Observed variance = Individual variance + tissue-specific variance + …

Reference Component Analysis (RCA)

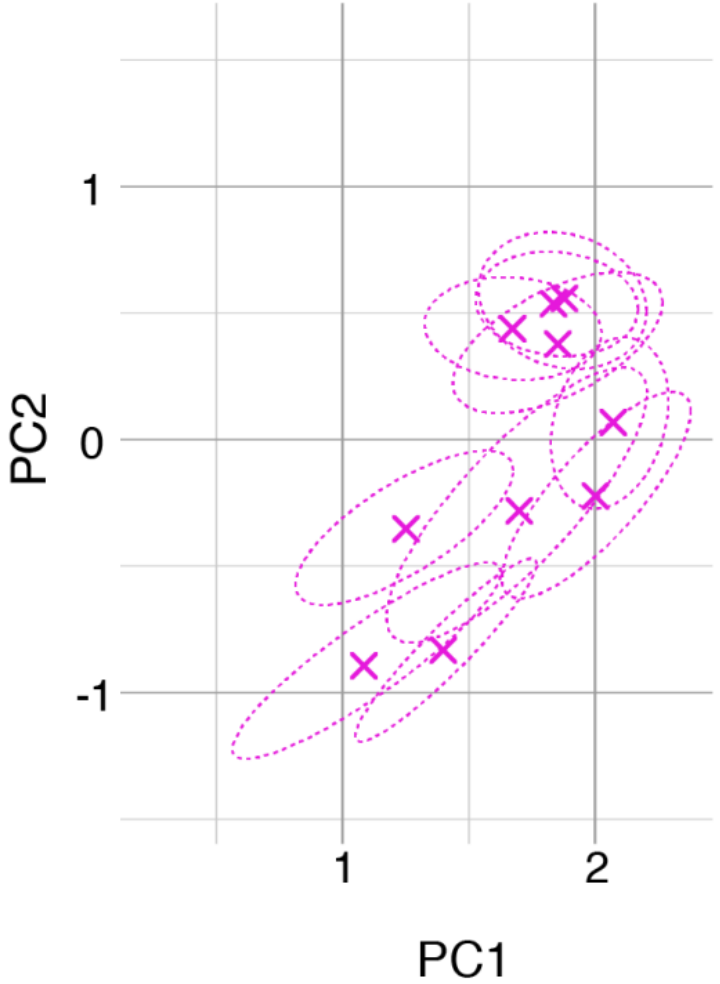
Applied to EN-TEx data, giving consistent transcriptome v epigenome comparison

Projections into EN-TEx space

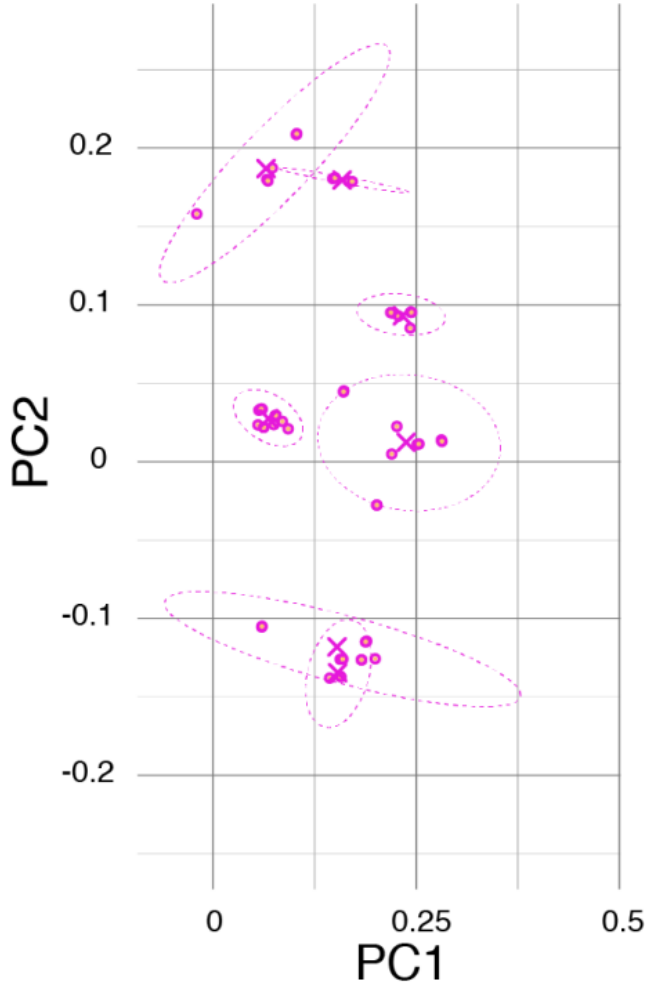
Transcriptome (GTEx)



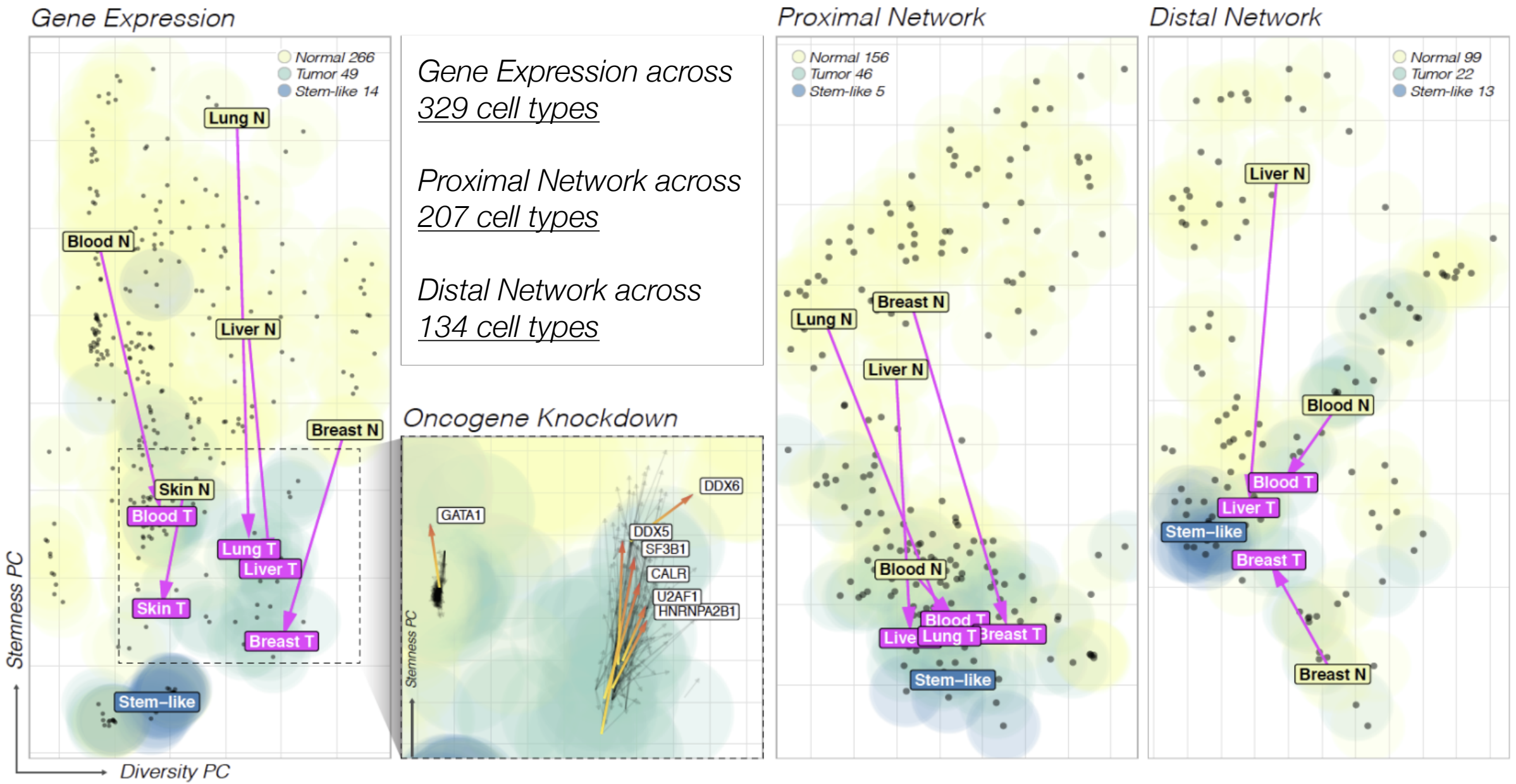
Transcriptome
Grouped by tissue



H3K27Ac

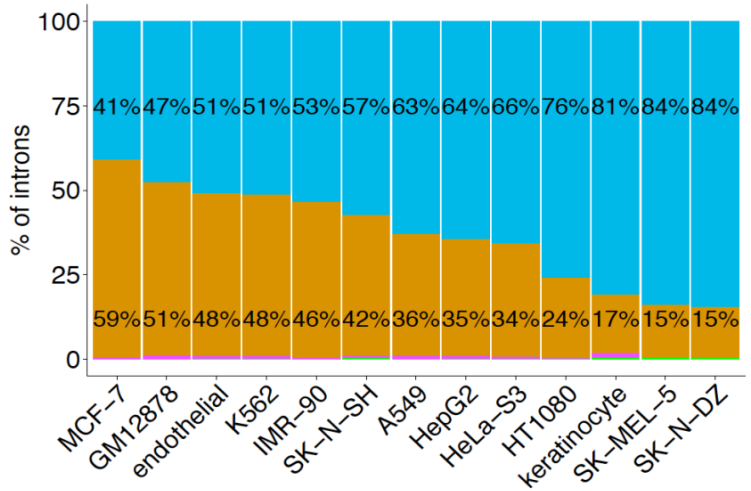


Applying RCA to many ENCODE cell types, focusing on Tumor-Normal Comparison



[Zhang et al.]

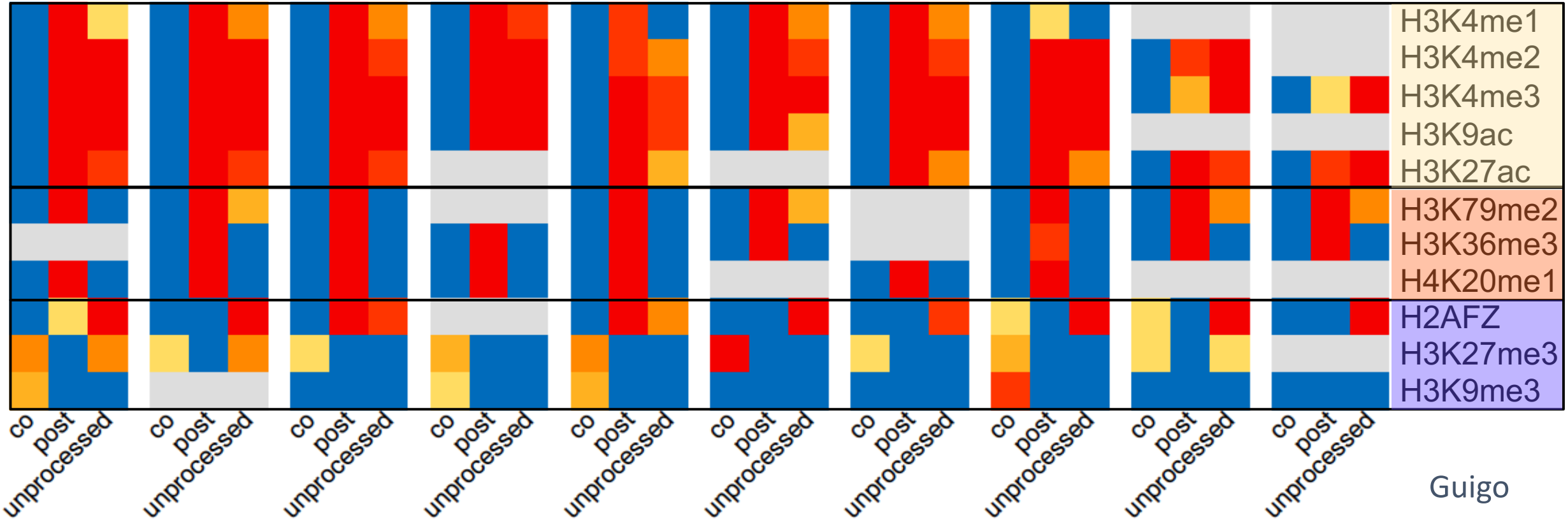
Epigenome & Transcriptome: HMs & Splicing



We observe enrichment of specific histone marks in the 3 distinct groups of introns (co-t. spliced, post-t. spliced, unprocessed)

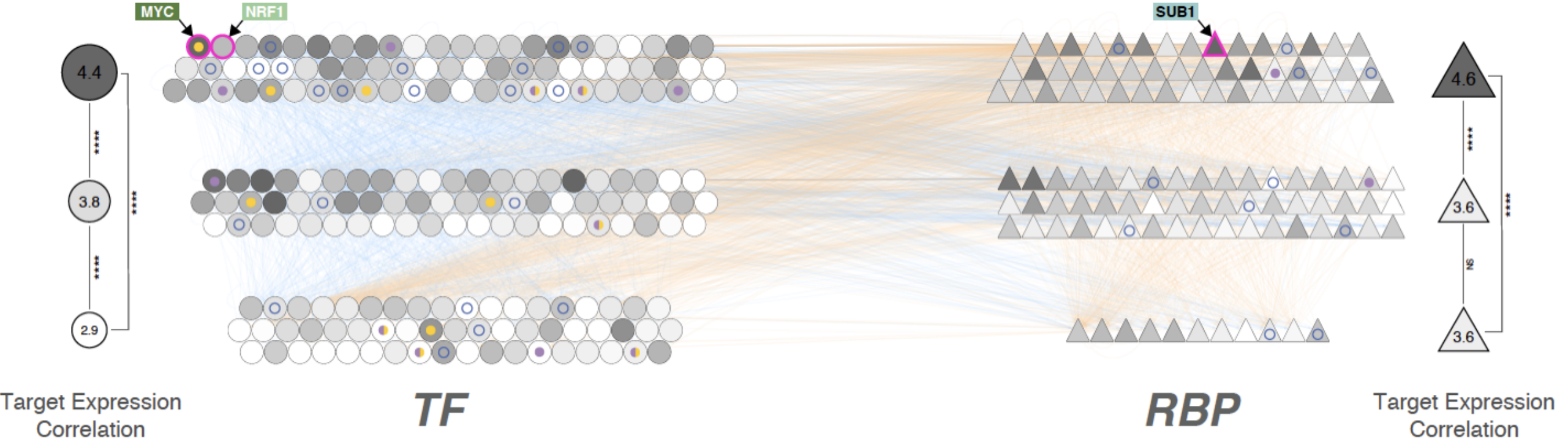
- **narrow and active histone marks → post-t. spliced & unprocessed**
- **broad and active histone marks → post-t. Spliced**
- **broad and repressive histone marks → co-t. spliced (& unprocessed)**

A549 GM12878 HeLa-S3 HepG2 IMR-90 K562 MCF-7 SK-N-SH endoth. keratinocyte



Guigo

Recasting of ENCODE Annotations in terms of TF & RBP Regulatory Networks

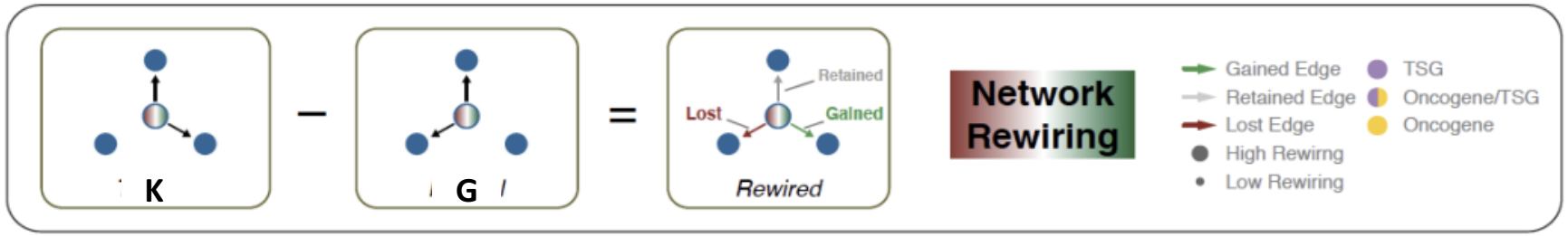


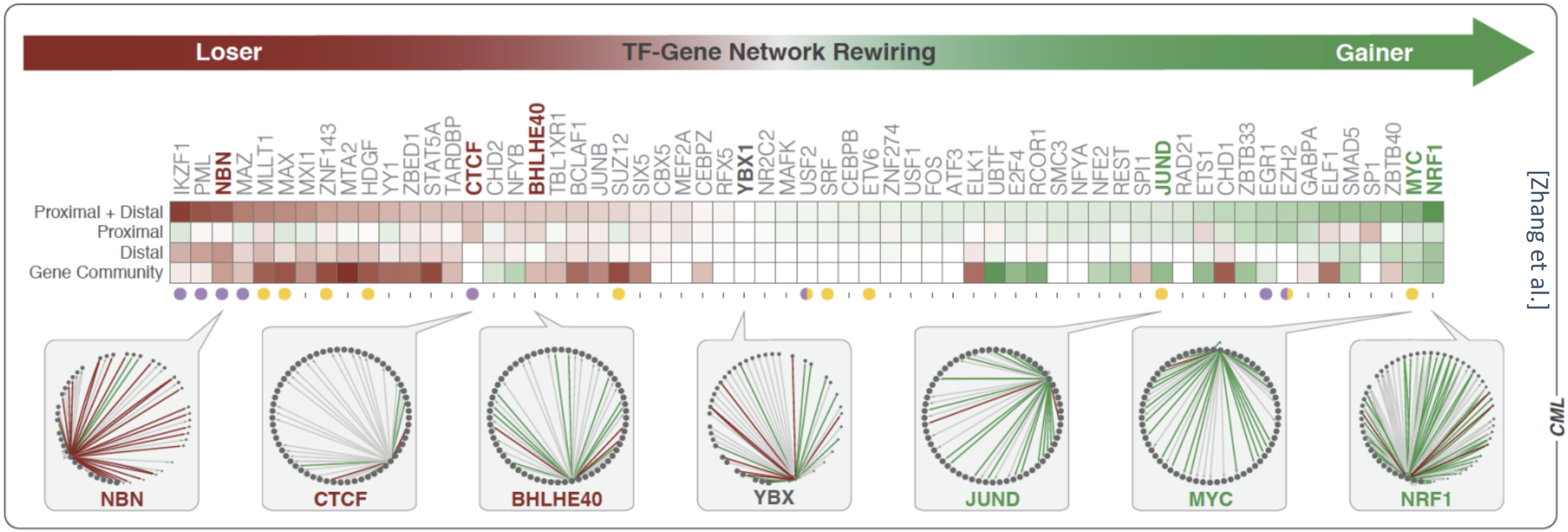
TF to target, via promoter (or pot. enhancer); RBP to target;
& then TF <=> RBP interconnections

[Zhang et al.]

Developing Metrics for Network Change

(Rewiring Index for **K**562 v **G**M12878)





Variant Annotation

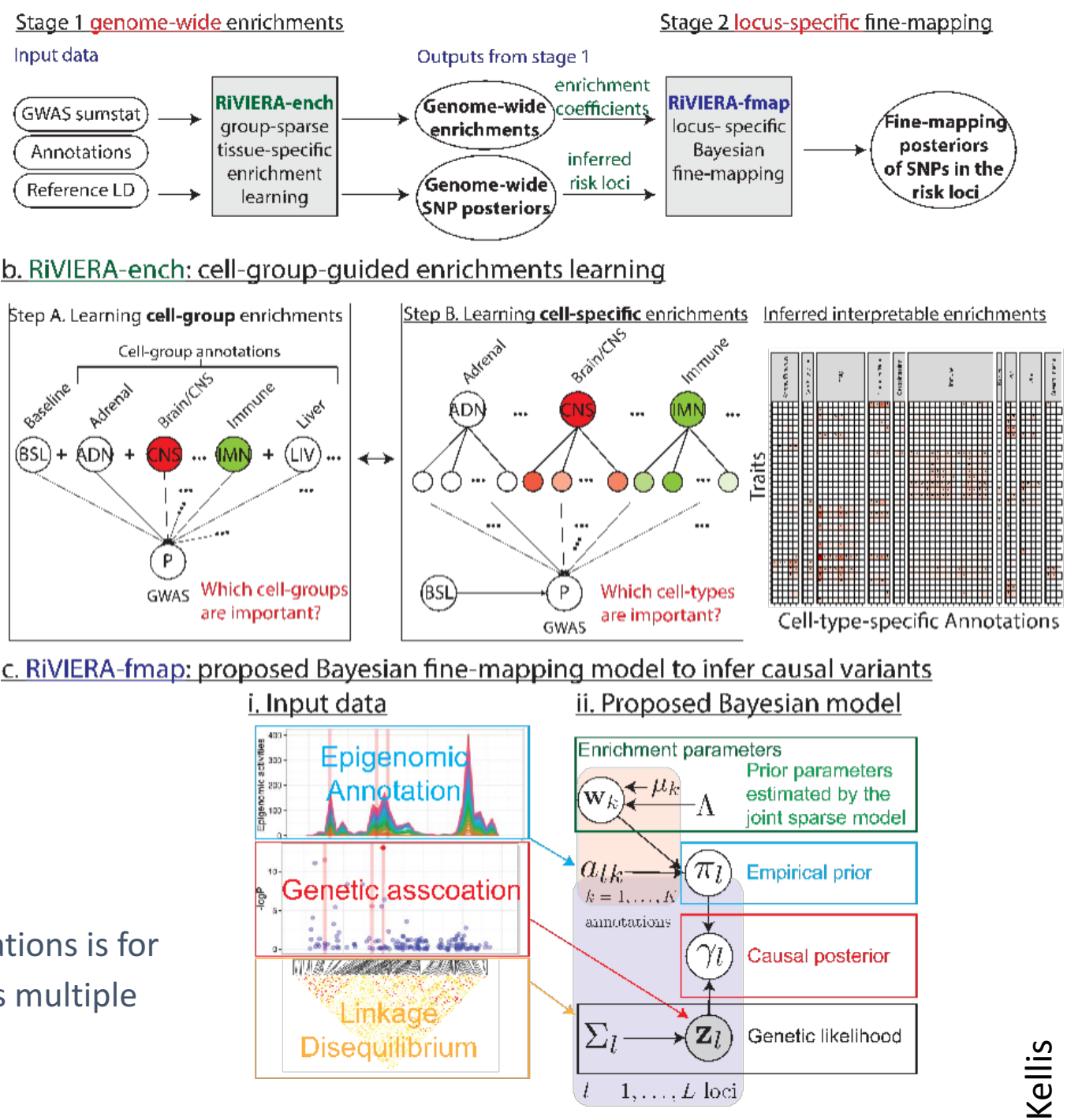
The DAC & other consortium members have developed tools for annotating variants with ENCODE annotations

Eg HaploReg, FunSeq, RegulomeDB

Work has continued on new tools

e.g. **Riviera**

Risk Variant Inference using Epigenomic Reference Annotations is for inference of driver variants from summary statistics across multiple traits using hundreds of epigenomic annotations

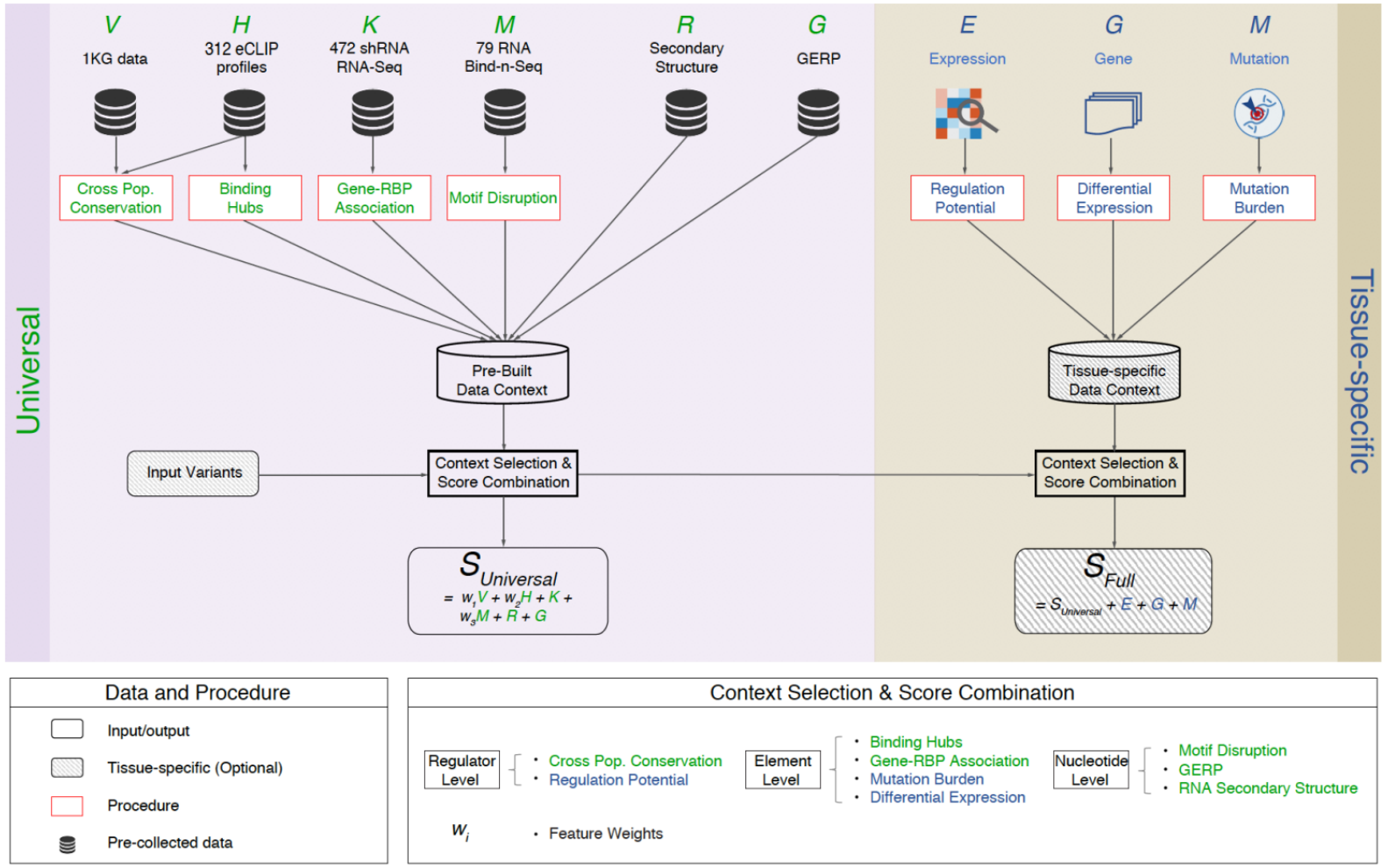


Kellis

Variant Annotation #2

Another new tool:

RADAR annotates the effect of variants in relation to RBP binding sites



Conclusions

ENCODE DAC priorities

- Building version 1.0 of the Encyclopedia
- Building SCREEN 1.0
- Developing pipelines & QC metrics
- Participating in working groups (EN-TEx, NAWG, RNA, etc.)
- Development of Chromatin Analysis Tools & Approaches

Thanks to:

- DAC members, especially trainees
- Other ENCODE members