

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

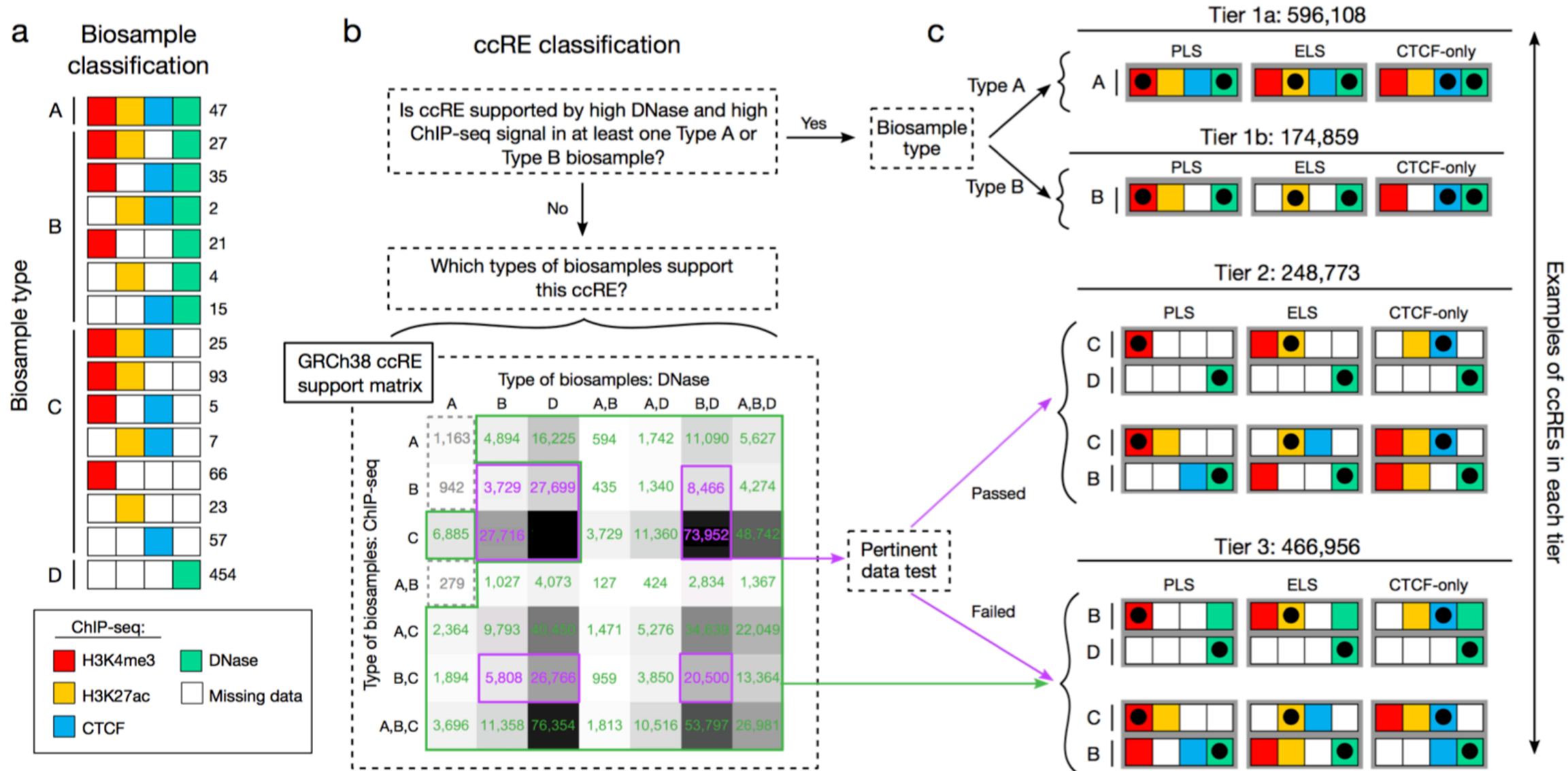
Updates to the Registry of ccREs: rDHSs for Anchoring ccREs

- 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.

Updates to the Registry of ccREs: New classification groups

		H3K4me3	H3K27ac	CTCF	DNase	TSS ccRE center ≤ 200 bp	Proximal ccRE center > 200 bp, ≤ 2 kb	Distal ccRE center > 2 kb
State	1	Red	Yellow	Blue	Green	PLS*	pPLS*	dPLS*
	2	Red	Yellow	Grey	Green	PLS	pPLS	dPLS
	3	Red	Grey	Blue	Green	PLS*	DNase-H3K4me3*	DNase-H3K4me3*
	4	Red	Grey	Grey	Green	PLS	DNase-H3K4me3	DNase-H3K4me3
	5	Grey	Yellow	Blue	Green	pELS*	pPLS*	dPLS*
	6	Grey	Yellow	Grey	Green	pELS	pELS	dPLS
	7	Grey	Grey	Blue	Green	CTCF-only	CTCF-only	CTCF-only

Updates to the Registry of ccREs: New tier system



Search Candidate cis-Regulatory Elements by ENCODE (SCREEN)

SCREEN: Search Candidate cis-Regulatory Elements by ENCODE

[Overview](#) [About](#) [Tutorials](#) [Downloads](#) [Versions](#)

SCREEN is a web interface for searching and visualizing the Registry of candidate cis-Regulatory Elements (ccREs) derived from [ENCODE data](#). The Registry contains 1.31M human ccREs in hg19 and 0.43M mouse ccREs in mm10, with orthologous ccREs cross-referenced. SCREEN presents the data that support biochemical activities of the ccREs and the expression of nearby genes in specific cell and tissue types.

You may launch SCREEN using the search box below or browse a curated list of SNPs from the NHGRI-EBI Genome Wide Association Study (GWAS) catalog to annotate genetic variants using ccREs.

[Browse GWAS](#)

Enter a gene name or alias, a SNP rsID, a ccRE accession, or a genomic region in the form chr:start-end. You may also enter a cell type name to filter results.
Examples: "K562 chr11:5226493-5403124", "SOX4", "rs4846913", "EH37E0204974"

[Search Human \(hg19\)](#) [Search Mouse \(mm10\)](#) [Search Human \(GRCh38\)](#)

SCREEN-Cistrome Interface (MM10 and GRCh38)

SCREEN mm10

ccRE Search Results

EM10E0072345 chr11:5,239,269-5,239,533 ★ P

Biosamples ⓘ

TSV Search:

cell type	tissue
<input type="radio"/> 129.DLCR liver male embryo (14.5 days)	liver
<input type="radio"/> 129 E14TG2a.4	ESC
<input type="radio"/> 129 ES-E14	ESC
<input type="radio"/> 129 G1E	blood
<input type="radio"/> 129 liver male embryo (14.5 days)	liver
<input type="radio"/> 129 ZHBTc4-mESC treated with doxycycline hyclate	ESC
<input type="radio"/> B10.H-2aH-4bp/Wts_CH12.LX	blood
<input type="radio"/> C57BL/6 3T3-L1	adipose
<input type="radio"/> C57BL/6 acute myeloid leukemia	blood
<input type="radio"/> C57BL/6 adipocyte	adipose

Total: 138

« ‹ 1 2 3 … 14 › »

intersecting cistrome TF exps Search:

factor	# of experiments that support TF binding	# experiments in total
RELA	4	29
OTX2	2	16
HNF1B	2	2
ZFP57	2	14
ASH2L	1	2
NR0B1	1	2
IRF4	1	56
POU5F1	1	59
TCF12	1	5
EP300	1	113

Total: 12

« ‹ 1 2 › »

intersecting cistrome histone mark exps Search:

mark	# of experiments that support histone modification	# experiments in total
H3K27ac	46	731
H3K4me1	31	565
H3K4me2	17	399
H3K4me3	13	1,007
H3K9ac, H3K14ac	3	17
H3K36me3	1	211
H3	1	119
H3K9ac	1	139

Total: 8



Cistrome DB Toolkit

<http://cistrome.org/db>

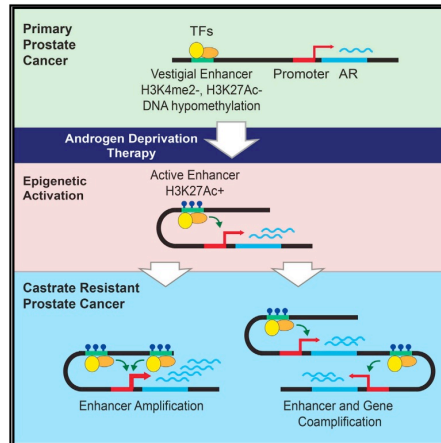
- ① Which factors bind to your genomic interval of interest?
- ② Which factors regulate your gene of interest?
- ③ Which cistromes in the Cistrome DB are similar to your cistrome of interest?

Which factors bind to your genomic interval?

Cell

A Somatically Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in Advanced Prostate Cancer

Graphical Abstract



Article

Authors

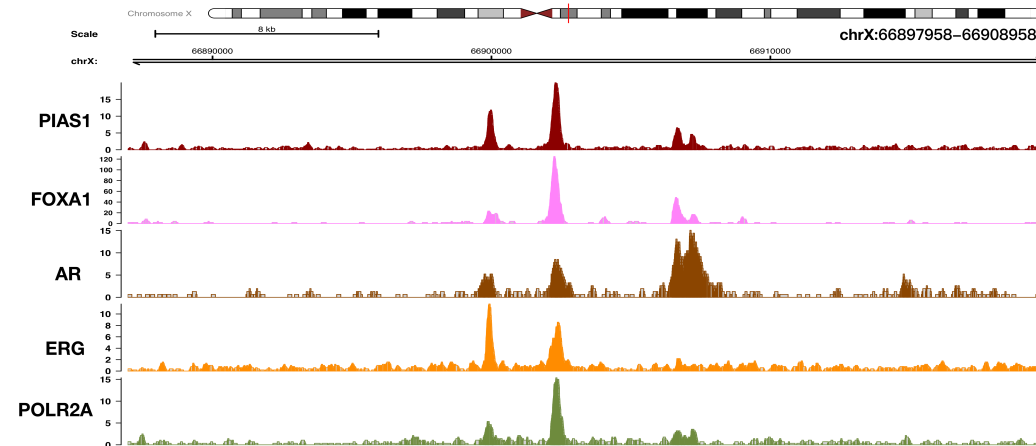
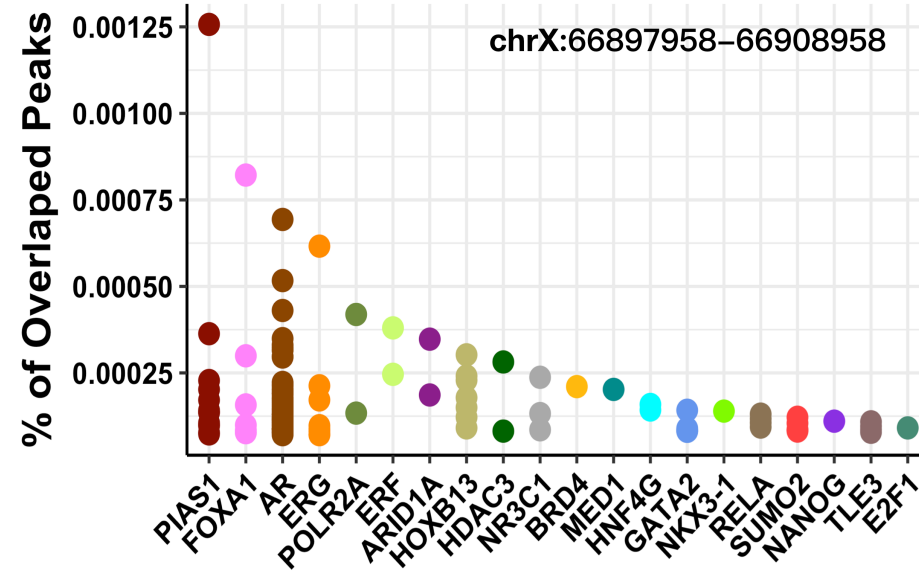
David Y. Takeda, Sándor Spisák, Ji-Heui Seo, ..., Mark M. Pomerantz, William C. Hahn, Matthew L. Freedman

Correspondence

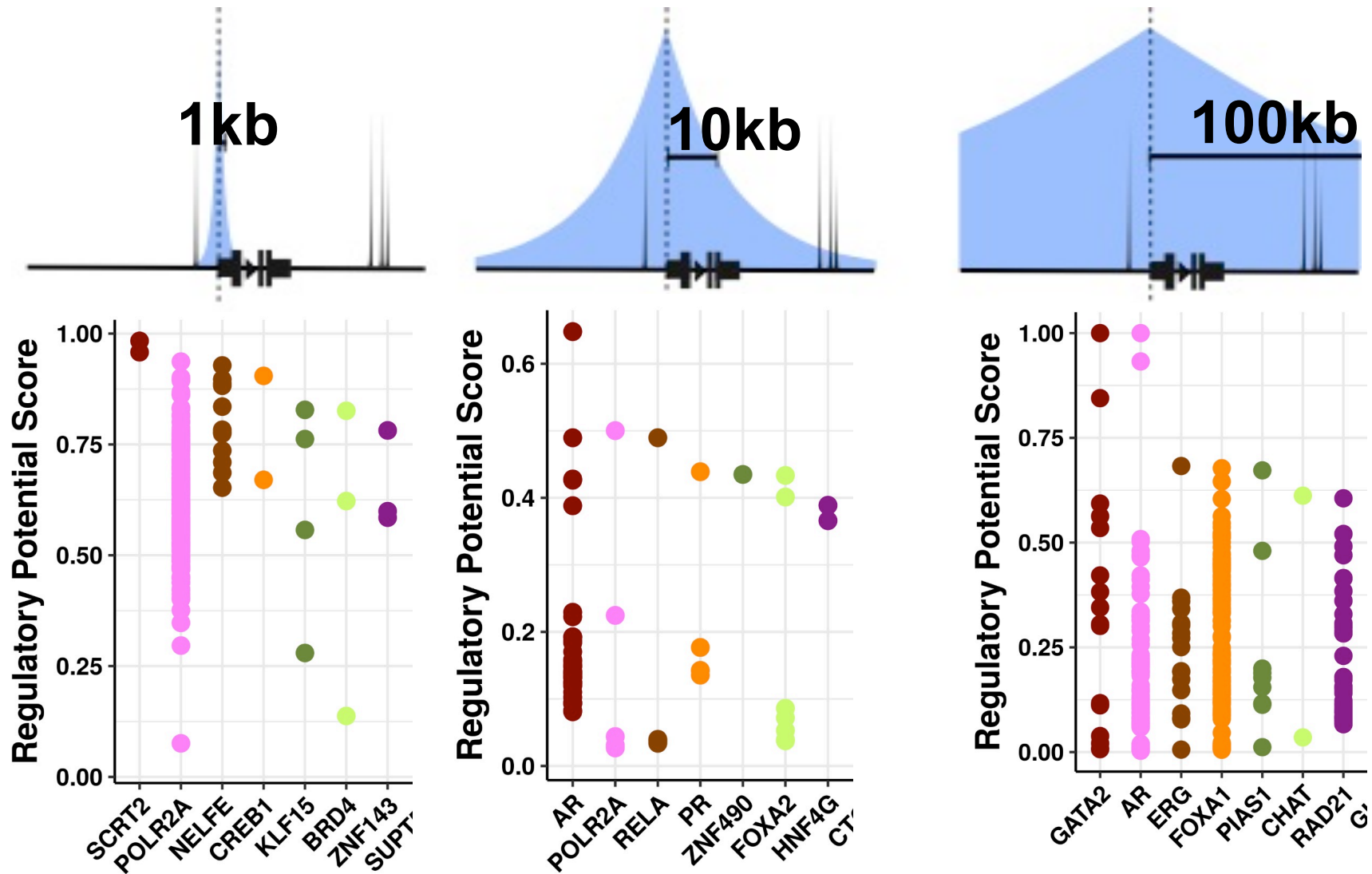
mfreedman@partners.org

In Brief

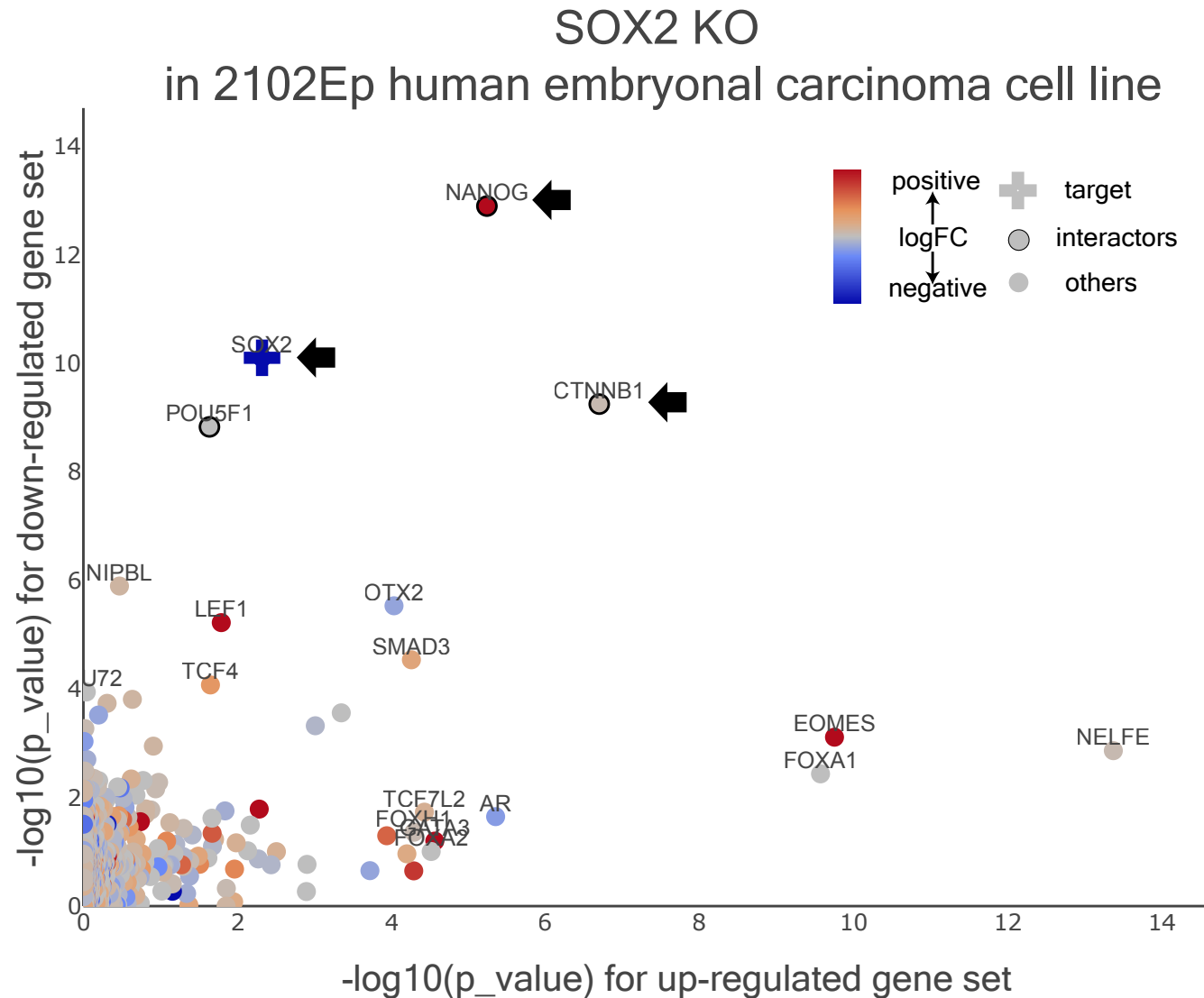
Activation and amplification of an enhancer upstream of the androgen receptor locus drives progression of metastatic castration-resistant prostate cancer.



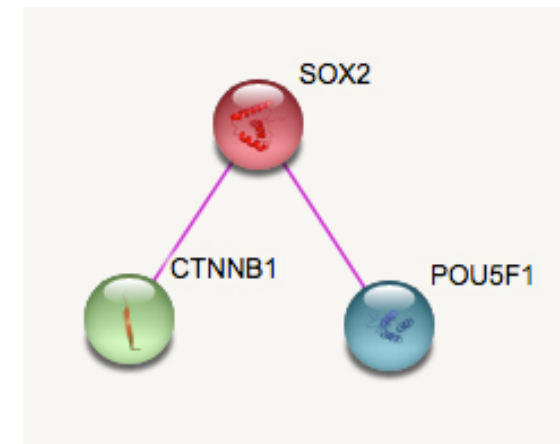
Which TFs regulates your gene?



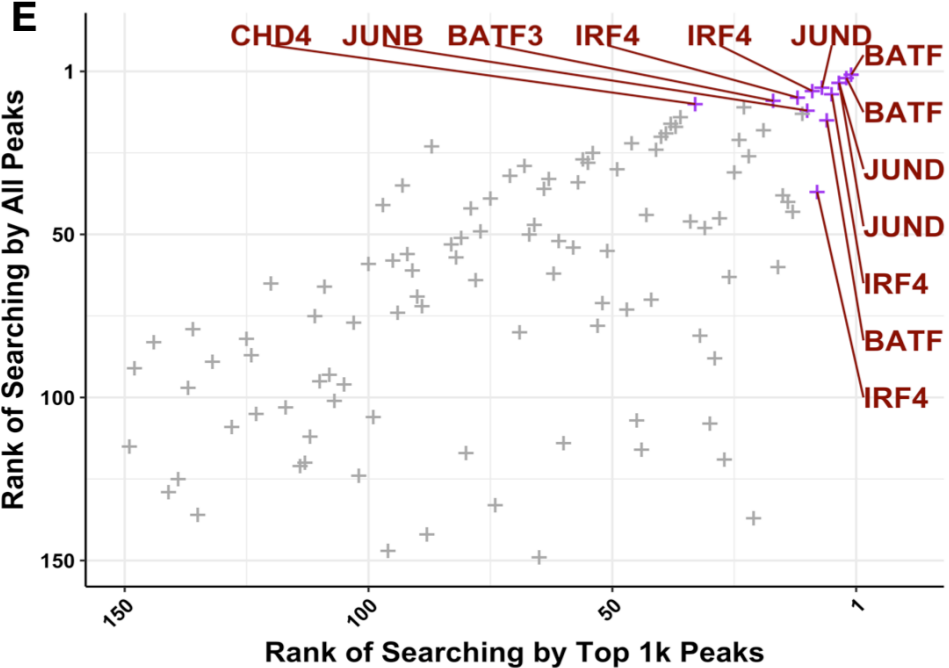
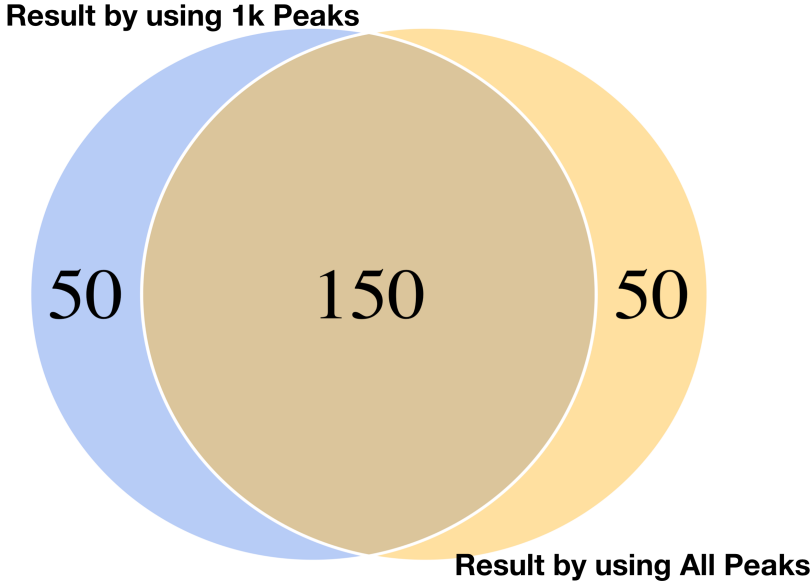
Co-regulatory TFs identified using Cistrome DB ChIP-seq data



STRING database:
Experimental evidence



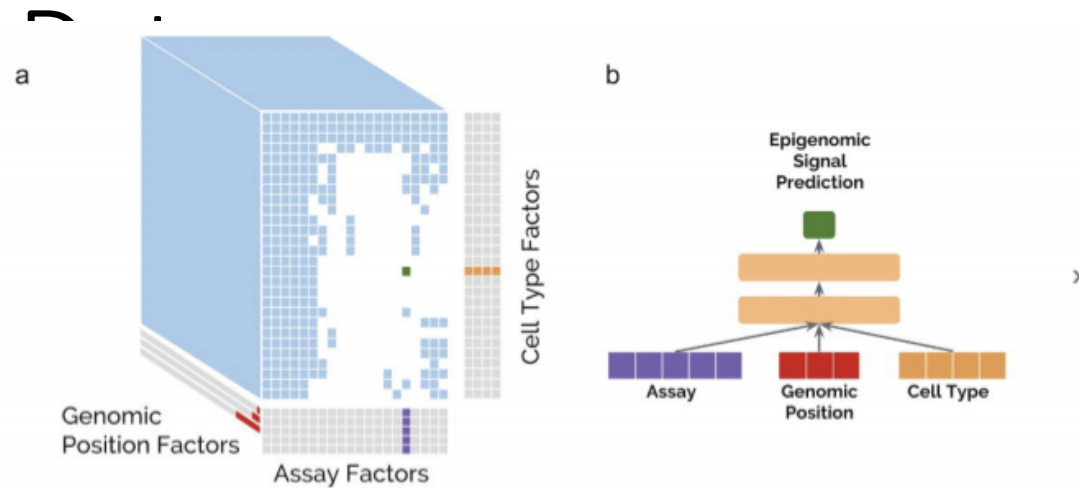
Which Cistrome DB cistromes are similar to your cistrome?



Chromatin & ChIP analysis

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

Avocado method for Imputation of Chromatin



bioRxiv
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 accessibility, 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,

UNIVERSITY OF WASHINGTON

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

Jacob Schreiber¹, Timothy Durham², 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 →

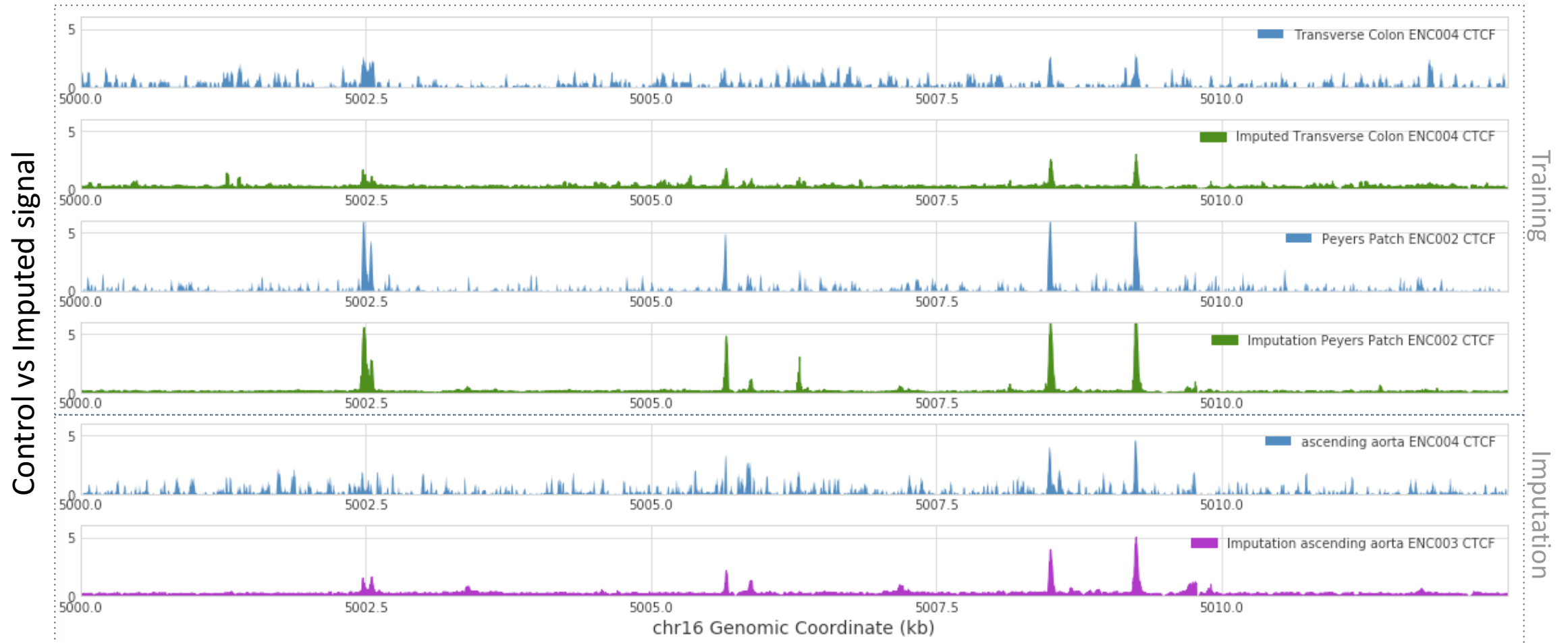
Read Supplement →

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

Jacob Schreiber, Bill Noble

Application of Avocado to the ENTEx Dataset

910 sample imputations from Avocado (Deep learning model)
Histone modifications + CTCF + RNAPol II



Synapse ID: syn17083203



DOI: 10.7303/syn17083203



Storage Location: Synapse Storage

Project Settings

Wiki Tools

Wiki

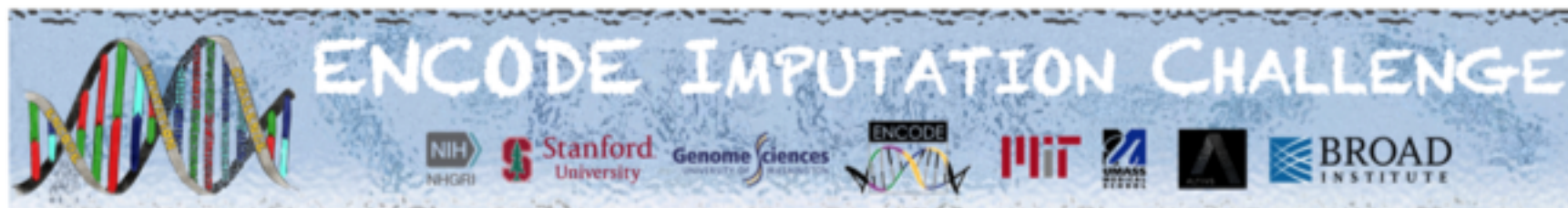
Discussion

ENCODE Imputation Challenge

1 - Challenge News and Updates

2 - Challenge Overview

3 - How to Participate



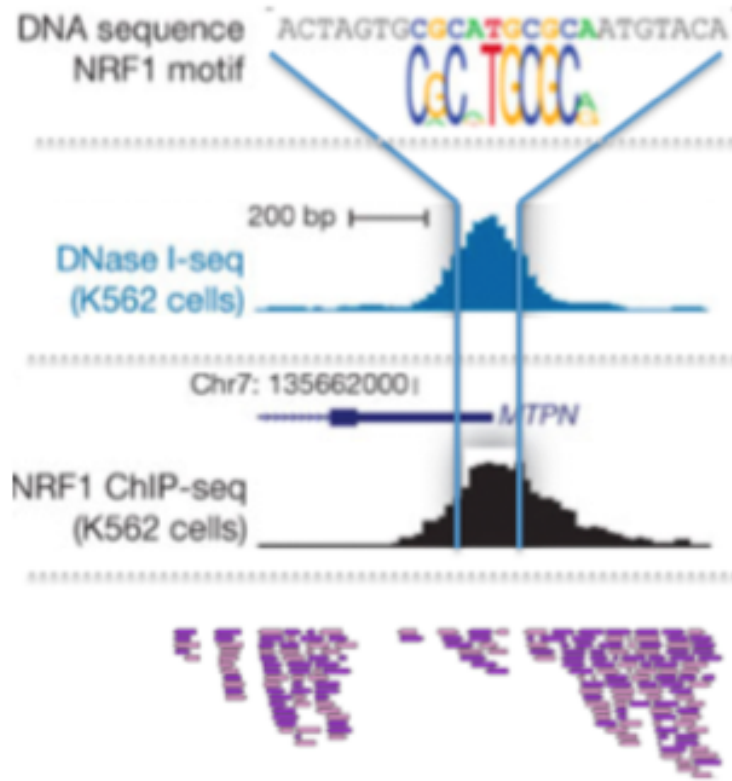
About the Challenge

How to Participate

News and Updates

https://www.encodeproject.org/encodeimpute

Participants will predict histone ChIP-seq

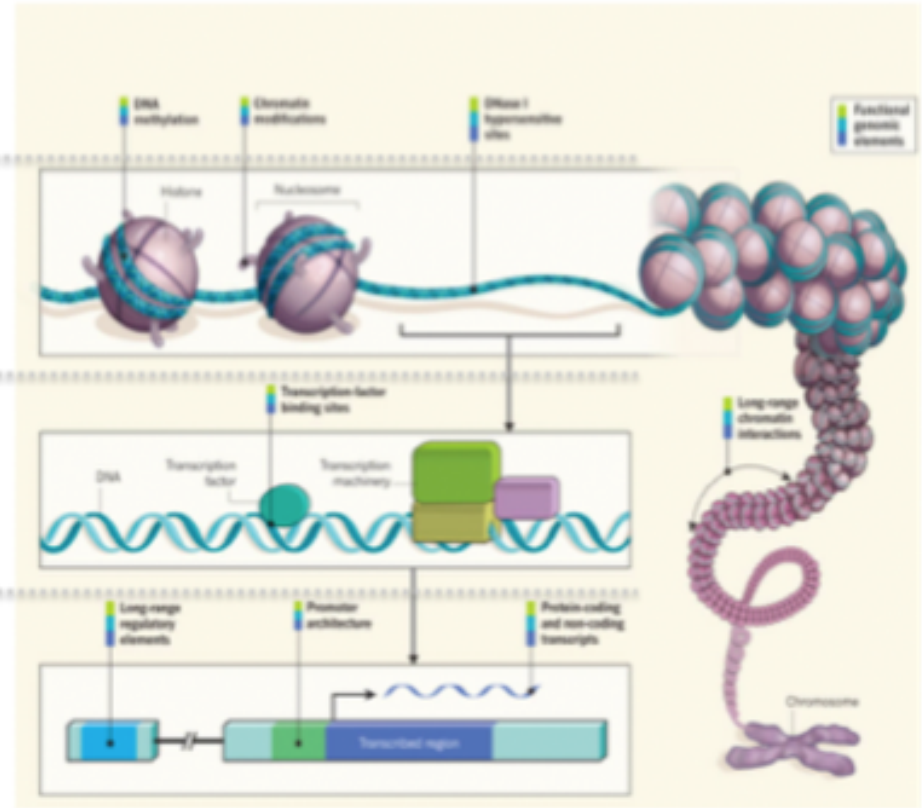


DNA-seq
Training

DNase-seq
Training

ChIP-seq
Training
Prediction

RNA-seq
Training



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

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

Decompose the variance

PCA/tSNE

cPCA

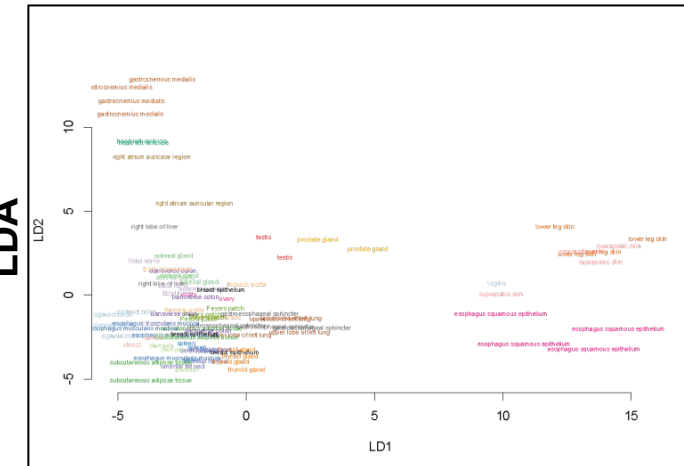
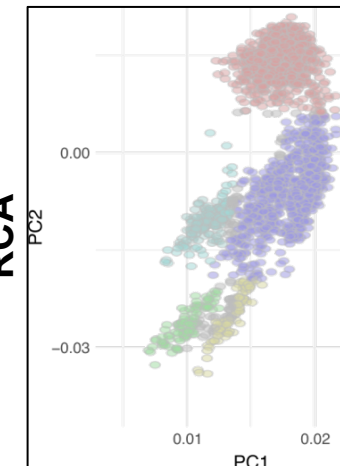
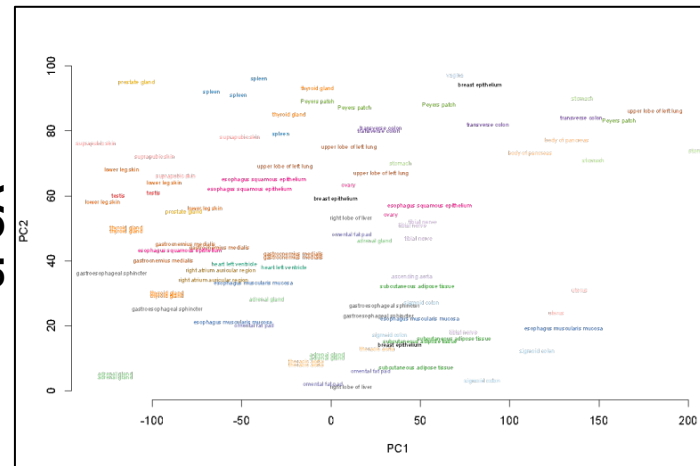
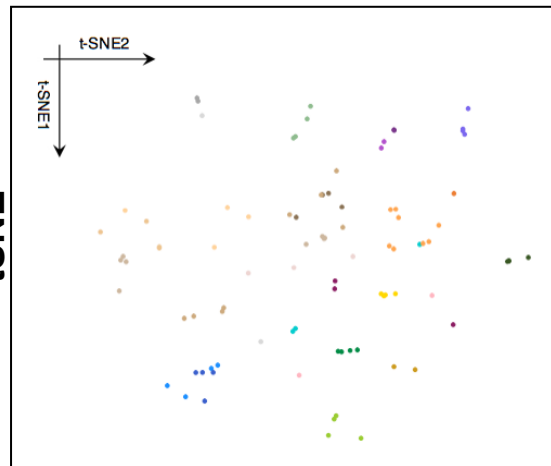
RCA

LDA

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

Unsupervised;
 no assumption on the structure

More supervised;
 with assumption on the structure



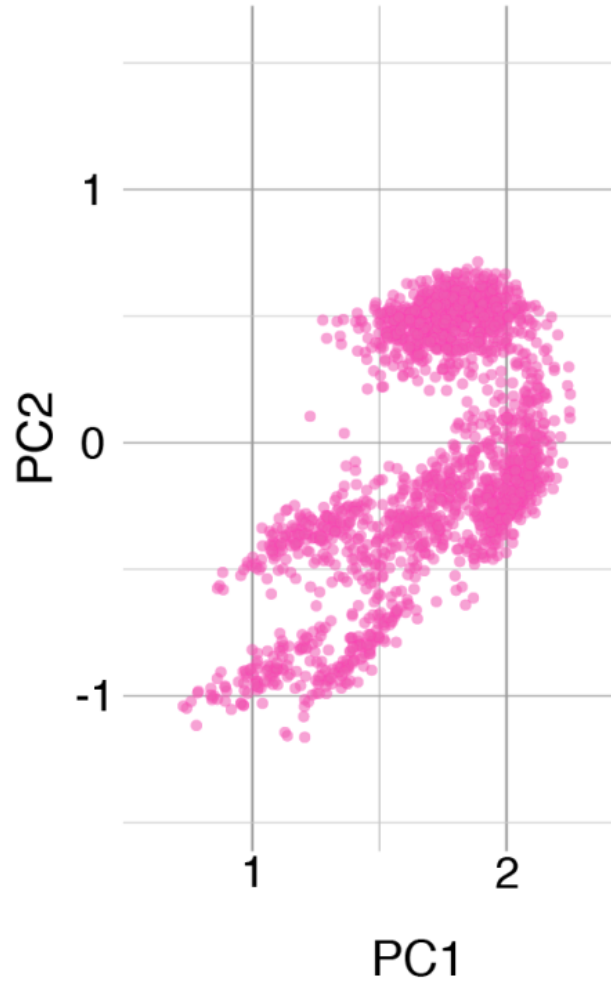
- For example in EN-TE_x, Observed variance = Individual variance + tissue-specific variance + ...

Reference Component Analysis (RCA)

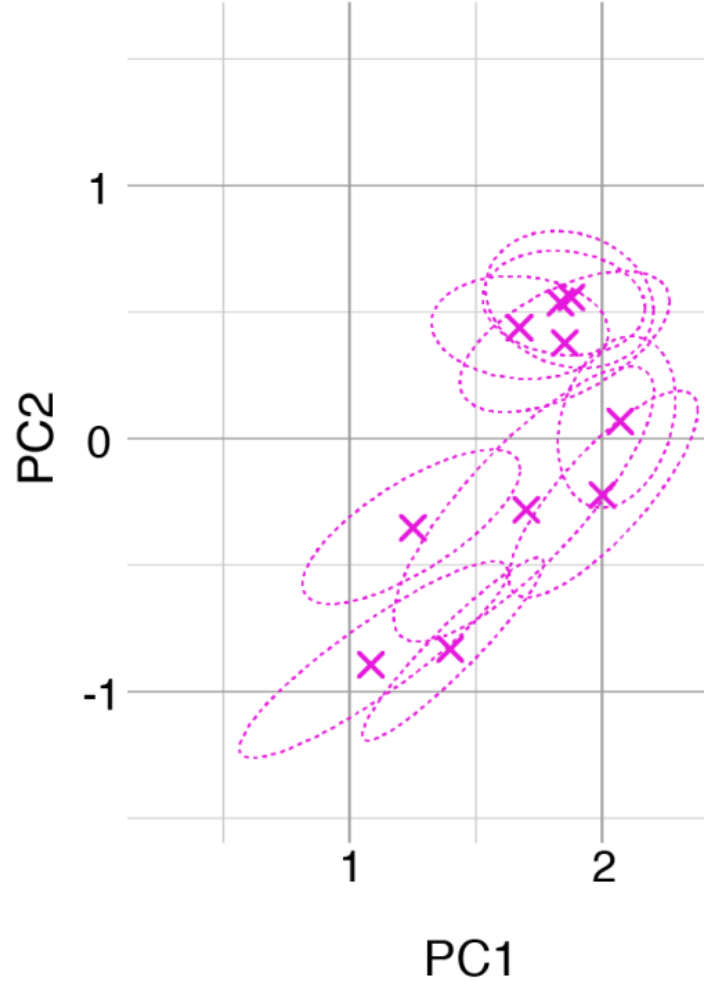
Applied to EN-TE_x data, giving consistent transcriptome v epigenome comparison

Projections
into EN-TE_x space

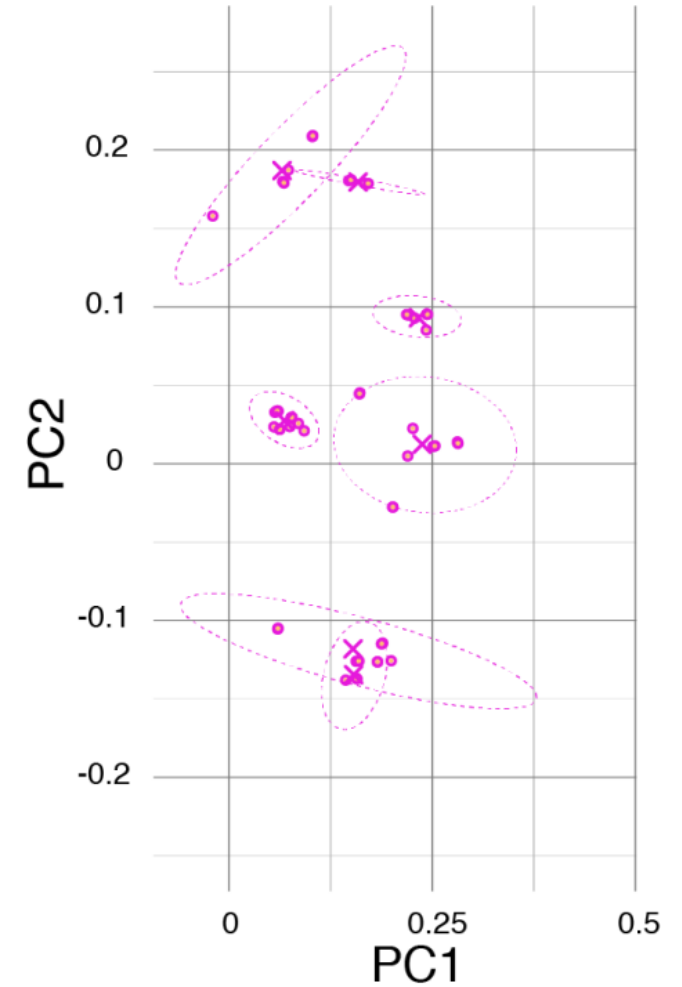
Transcriptome (GTEx)



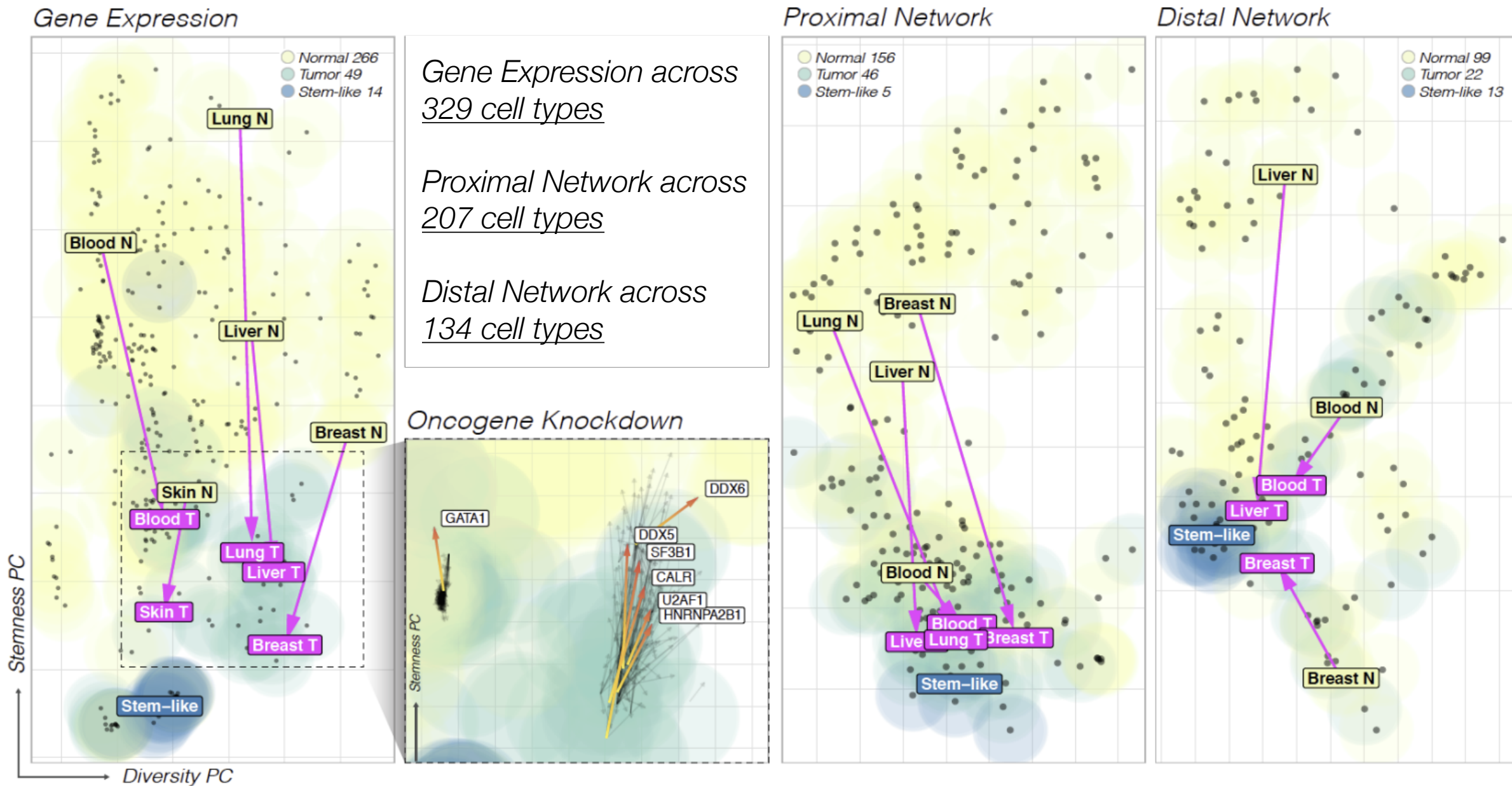
Transcriptome Grouped by tissue



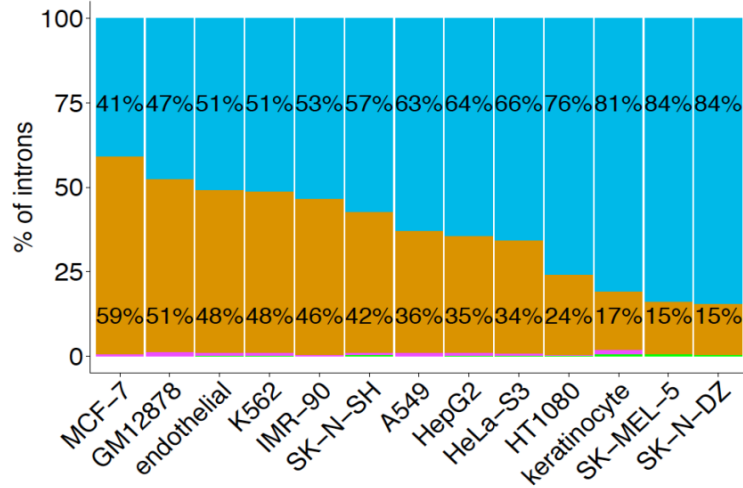
H3K27Ac



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



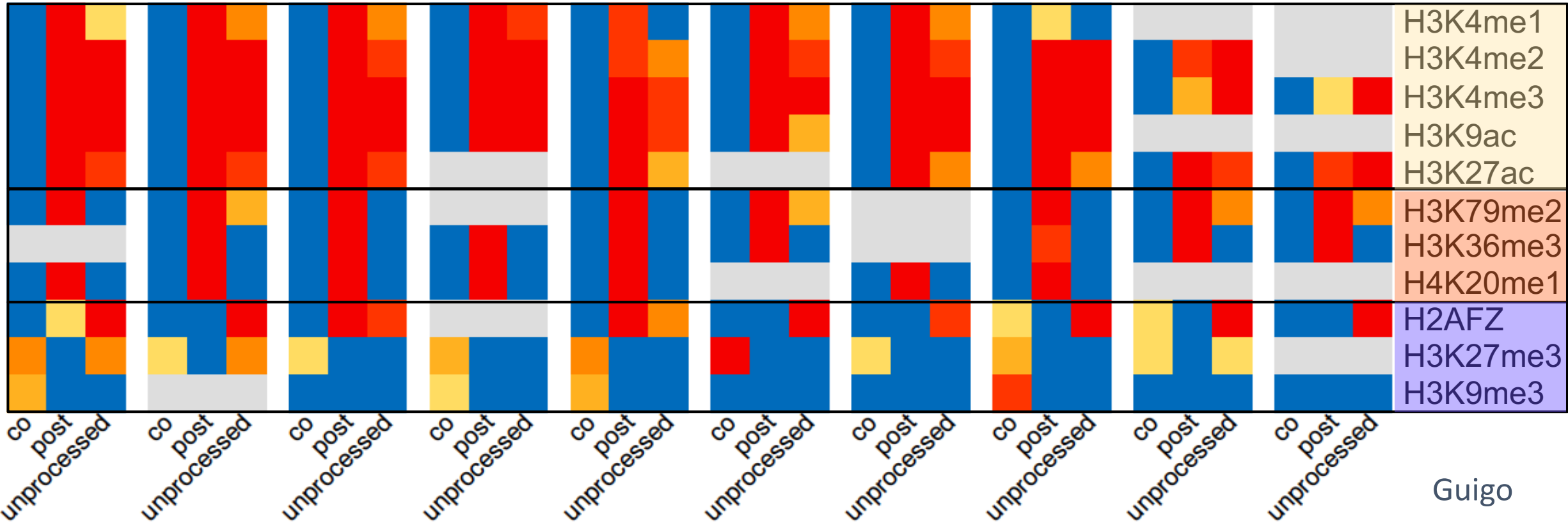
Epigenome & Transcriptome: HMs & Splicing



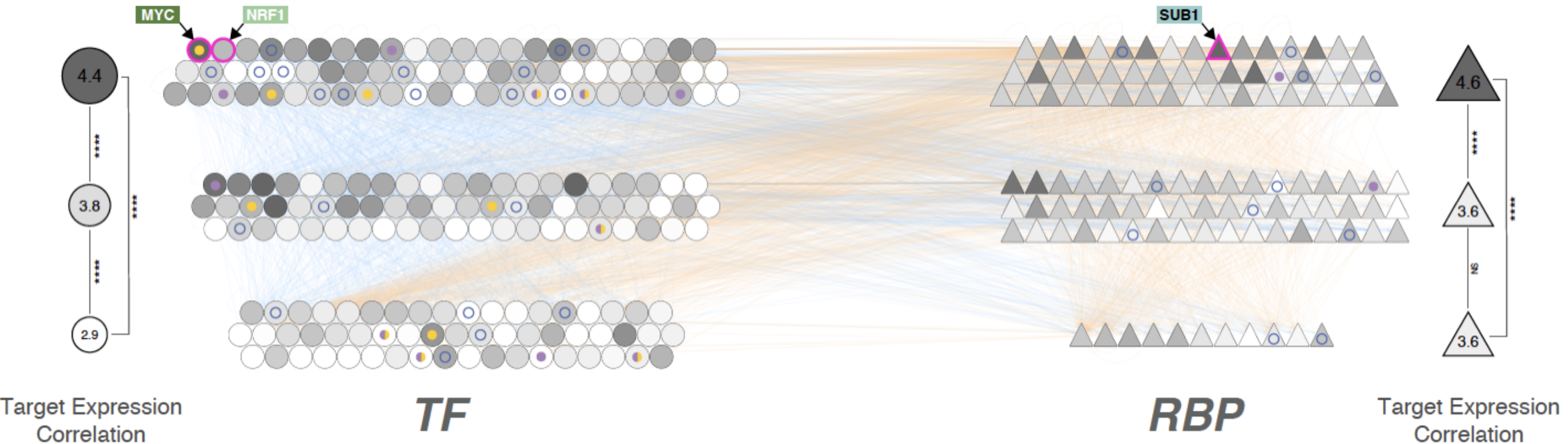
We observe enrichment of specific histone marks in the 3 distinct groups of introns (co-t. spliced, post-t. spliced, unprocessed)

- 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



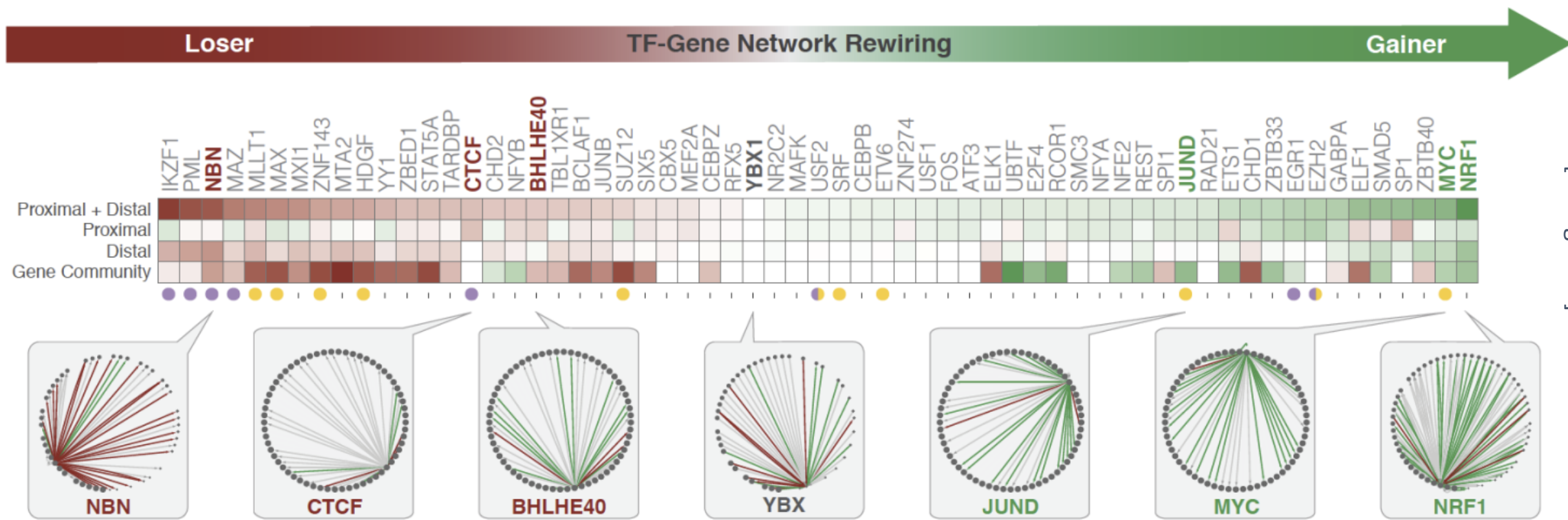
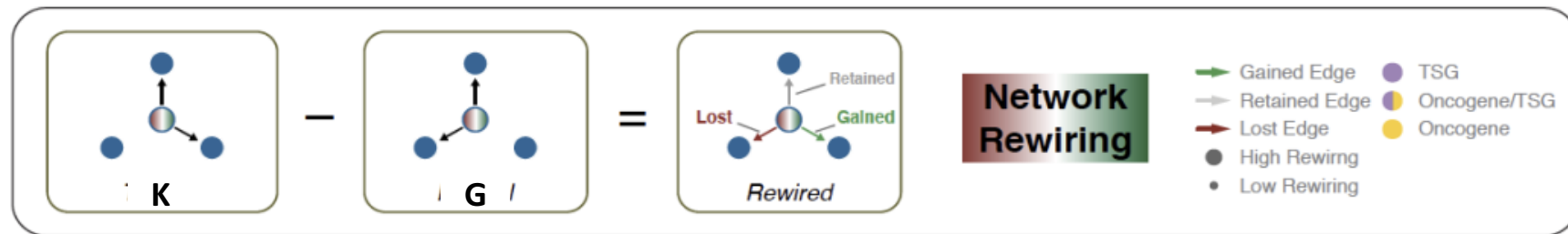
Recasting of ENCODE Annotations in terms of TF & RBP Regulatory Networks



TF to target, via promoter (or pot. enhancer); RBP to target;
& then TF \Leftrightarrow RBP interconnections

Developing Metrics for Network Change

(Rewiring Index for **K562 v GM12878**)



Variant Annotation

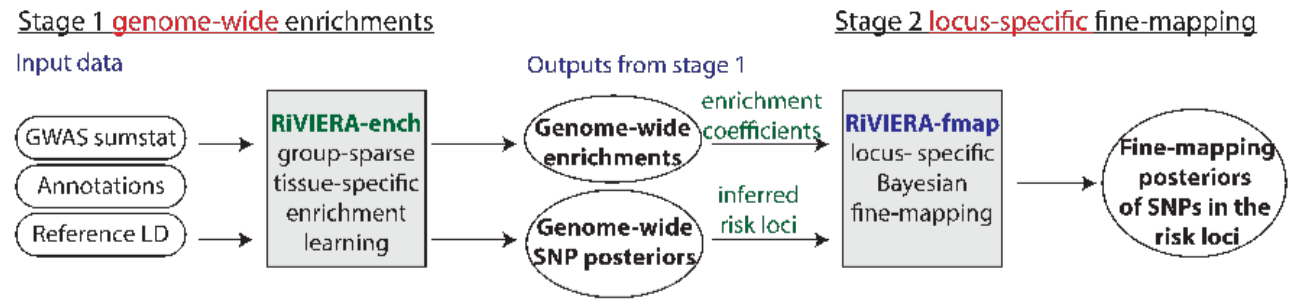
The DAC & other consortium members have developed tools for annotating variants with ENCODE annotations

Eg HaploReg, FunSeq, RegulomeDB

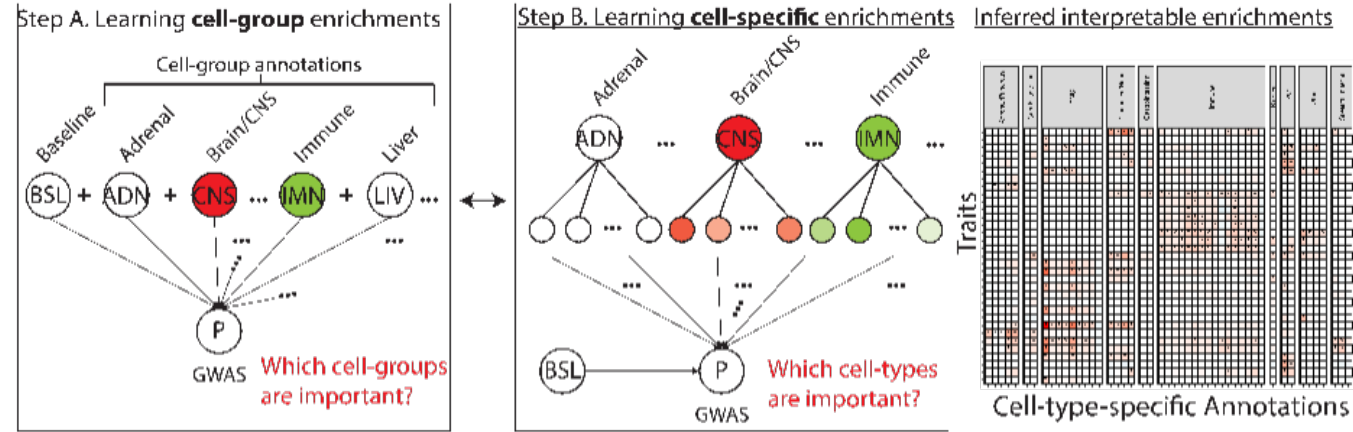
Work has continued on new tools

e.g. **Riviera**

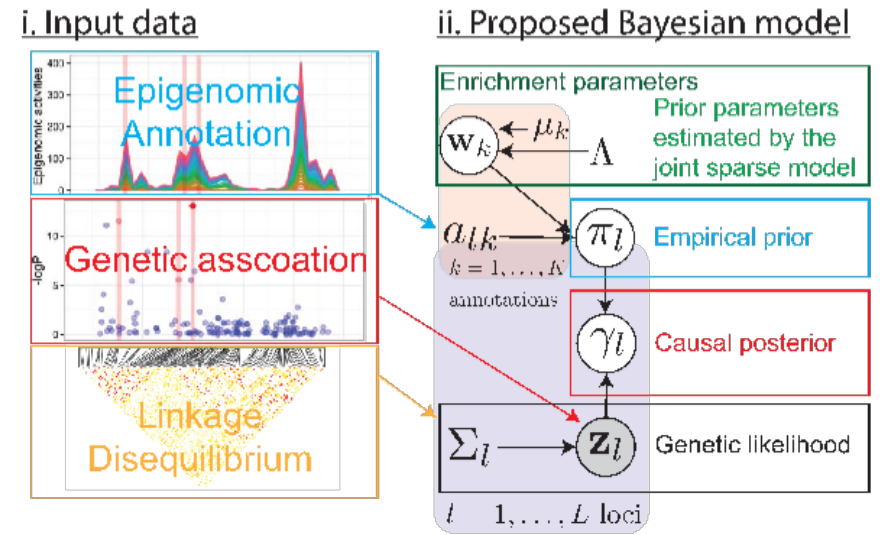
Risk Variant Inference using Epigenomic Reference Annotations is for inference of driver variants from summary statistics across multiple traits using hundreds of epigenomic annotations



b. RiVIERA-ench: cell-group-guided enrichments learning



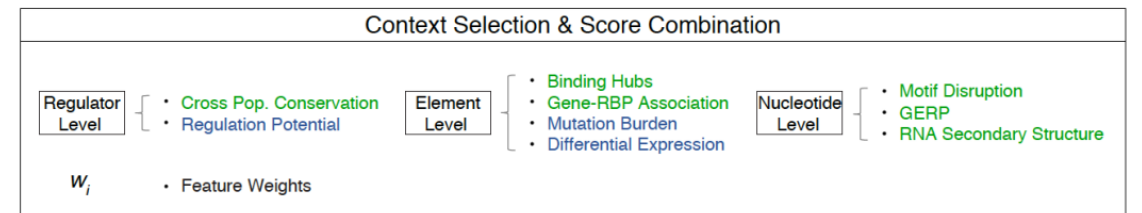
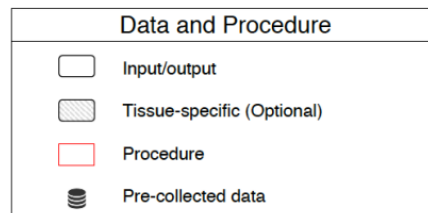
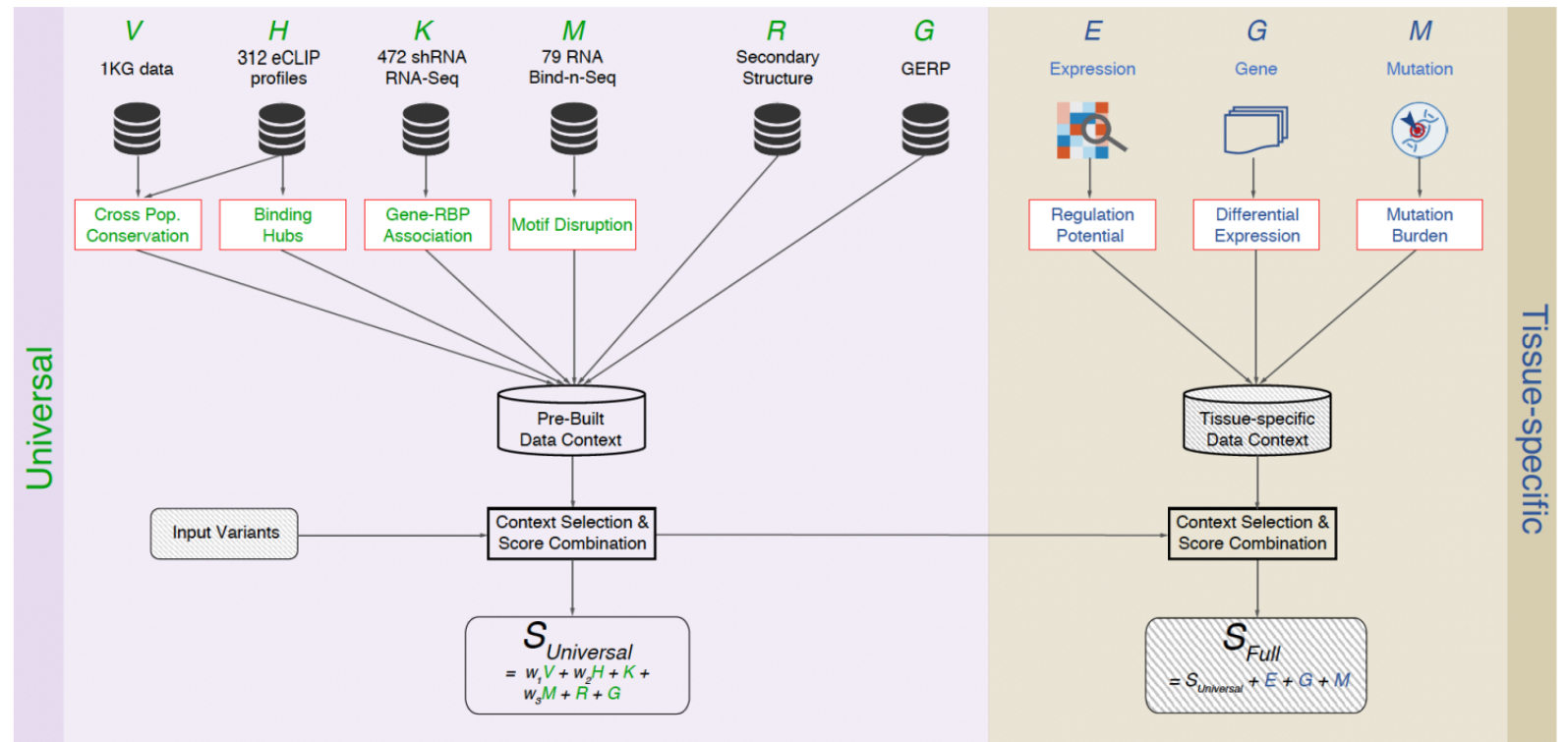
c. RiVIERA-fmap: proposed Bayesian fine-mapping model to infer causal variants



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