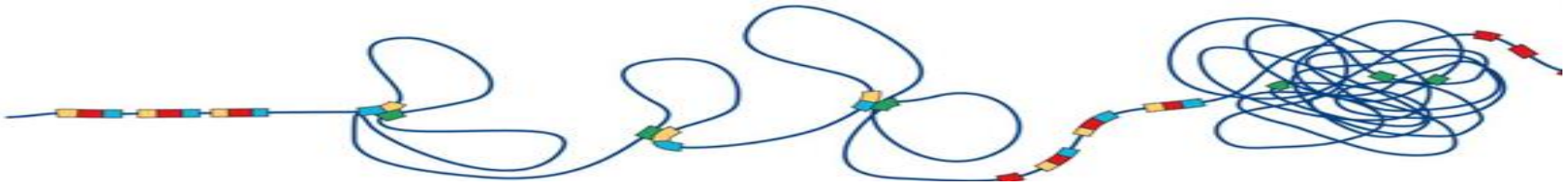
STANFORD SCHOOL OF MEDICINE | Genetics

STANFORD COMPUTER SCIENCE



# Dynamic enhancer-gene associations across diverse human cell types and tissues

Jianrong Wang


**Anshul Kundaje**

Assistant Professor

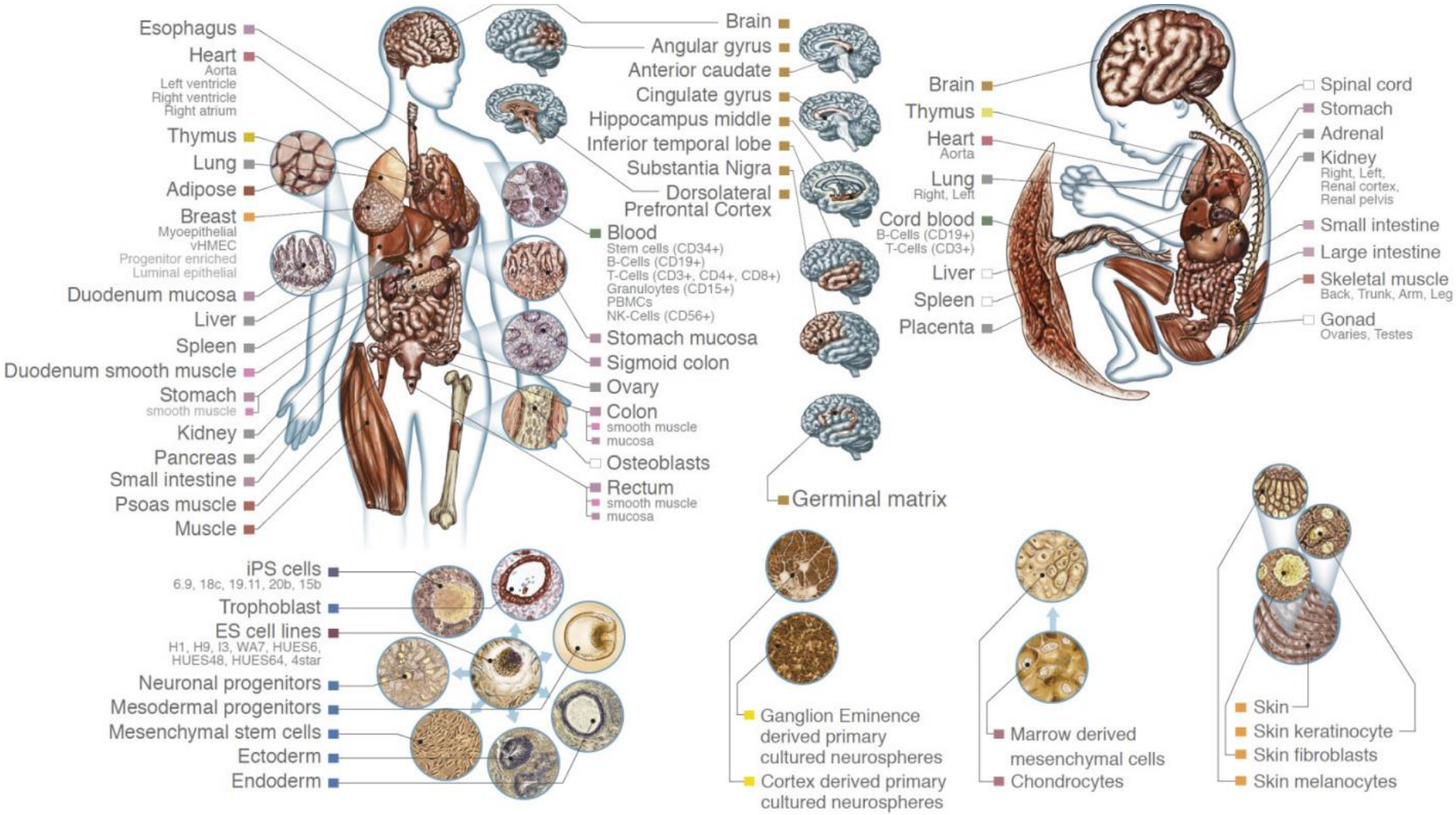
Dept. of Genetics

Dept. of Computer Science

http://anshul.kundaje.net

akundaje@stanford.edu

# Epigenomes and transcriptomes of 127 human tissues/cell-types

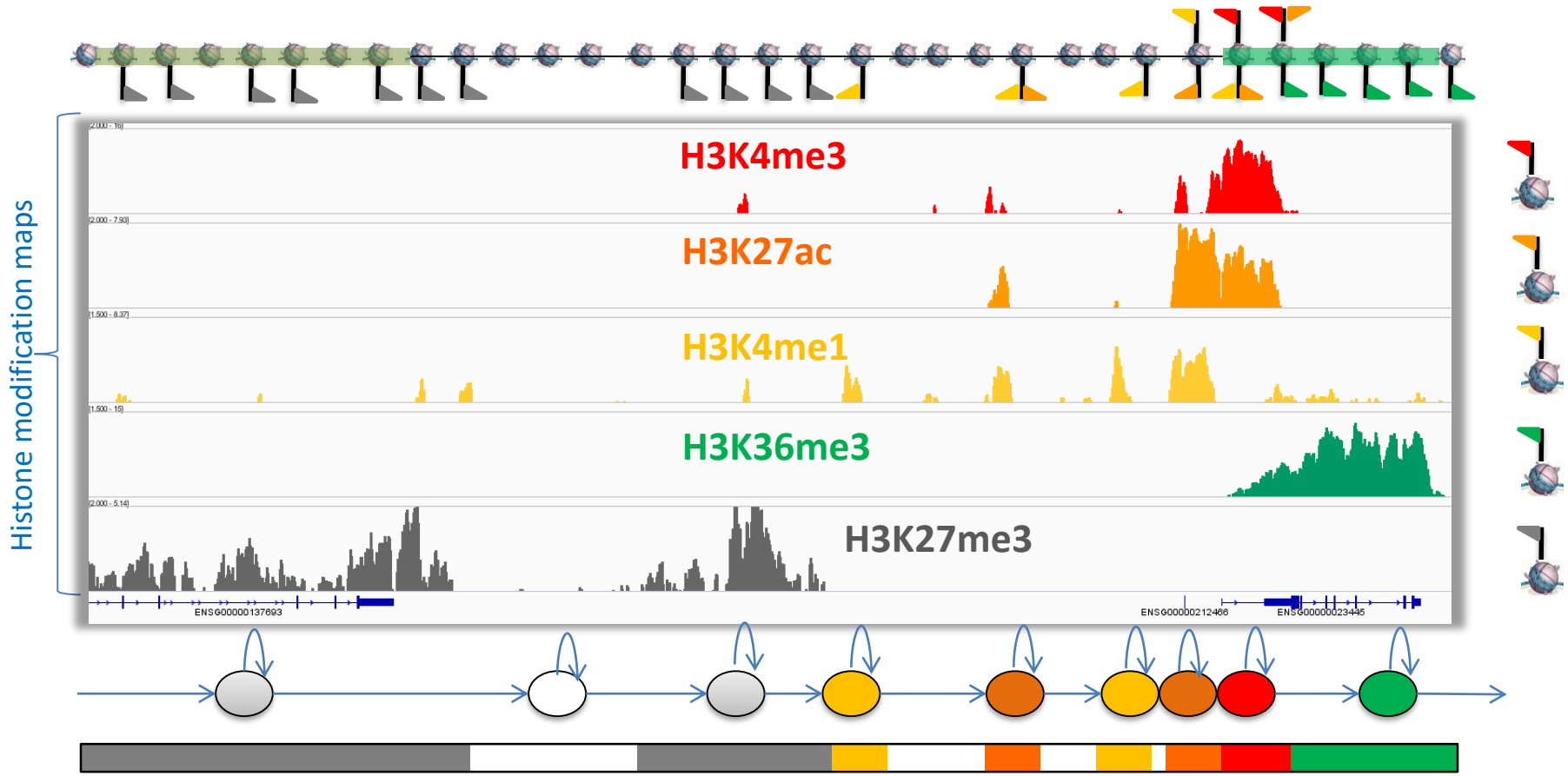



- 6+ key histone marks (Histone ChIP-seq)
- Open chromatin (DNase-seq)
- DNA methylation
- Gene expression (RNA-seq)

ENCODE + ROADMAP epigenomics PROJECT

http://roadmapepigenomics.org

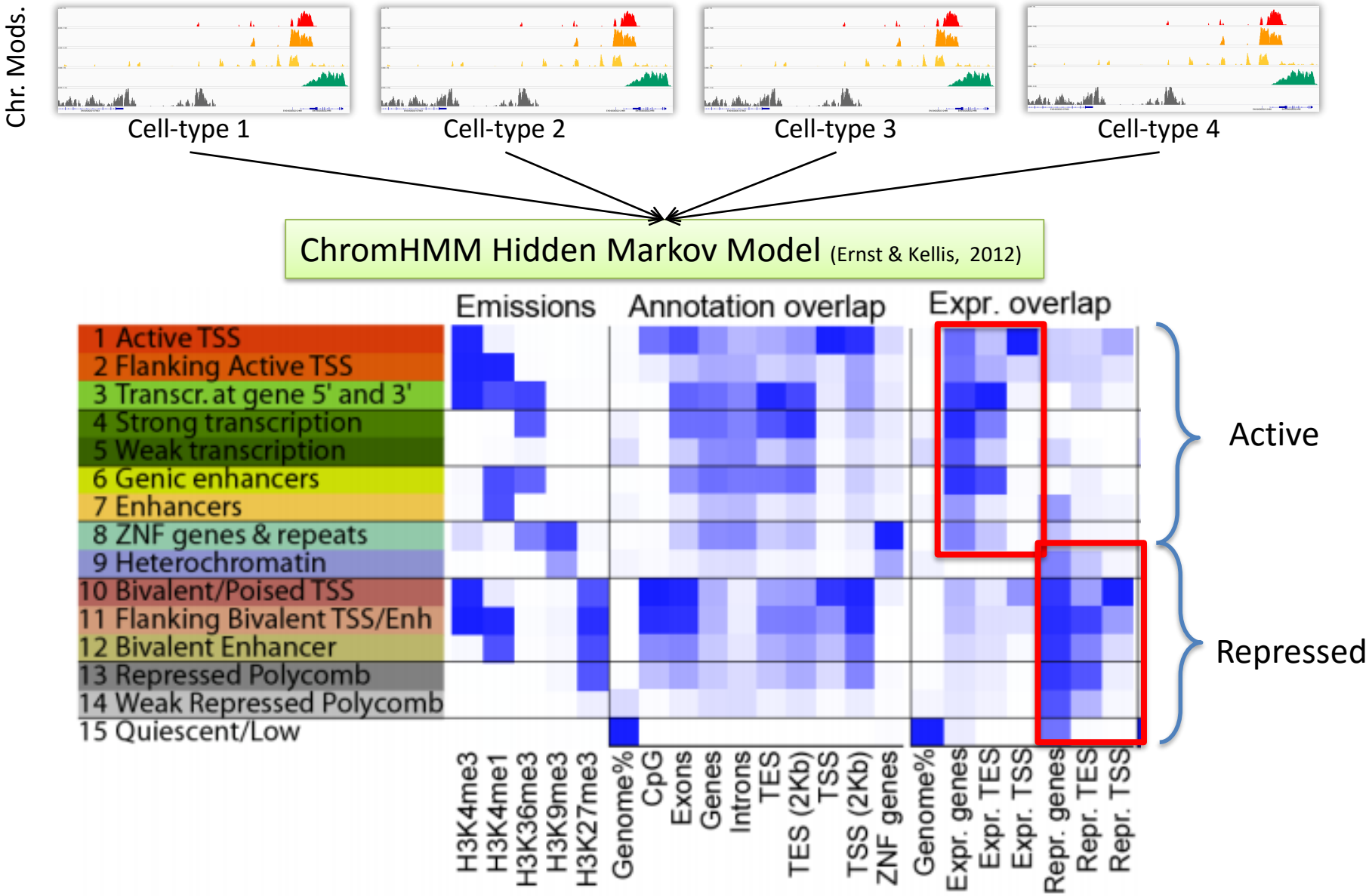
# Combinatorial chromatin states define regulatory elements



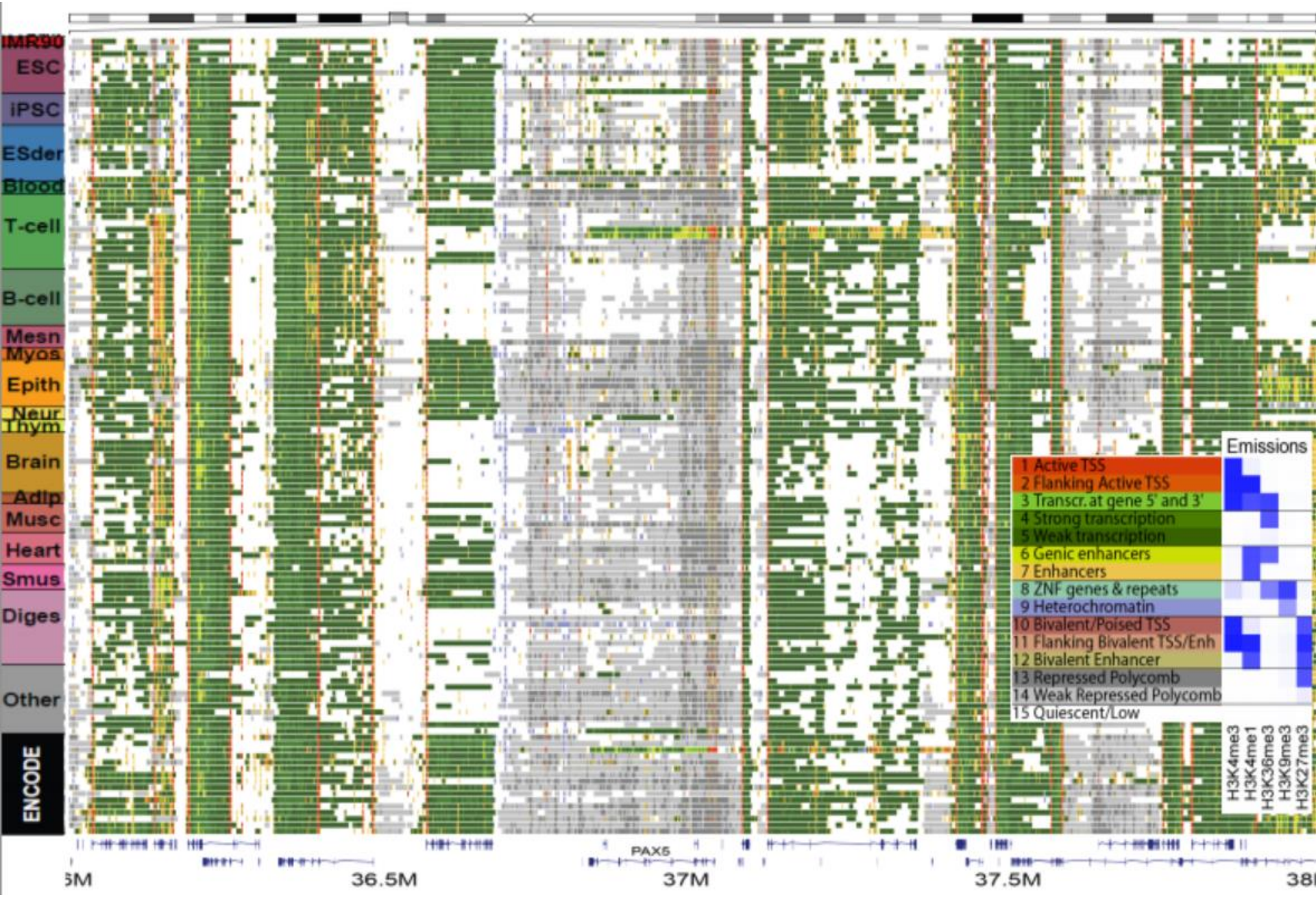

Identify hidden states using hidden Markov models

# Joint learning of regulatory chromatin states

**Joint training (virtual concatenation) across cell-types**

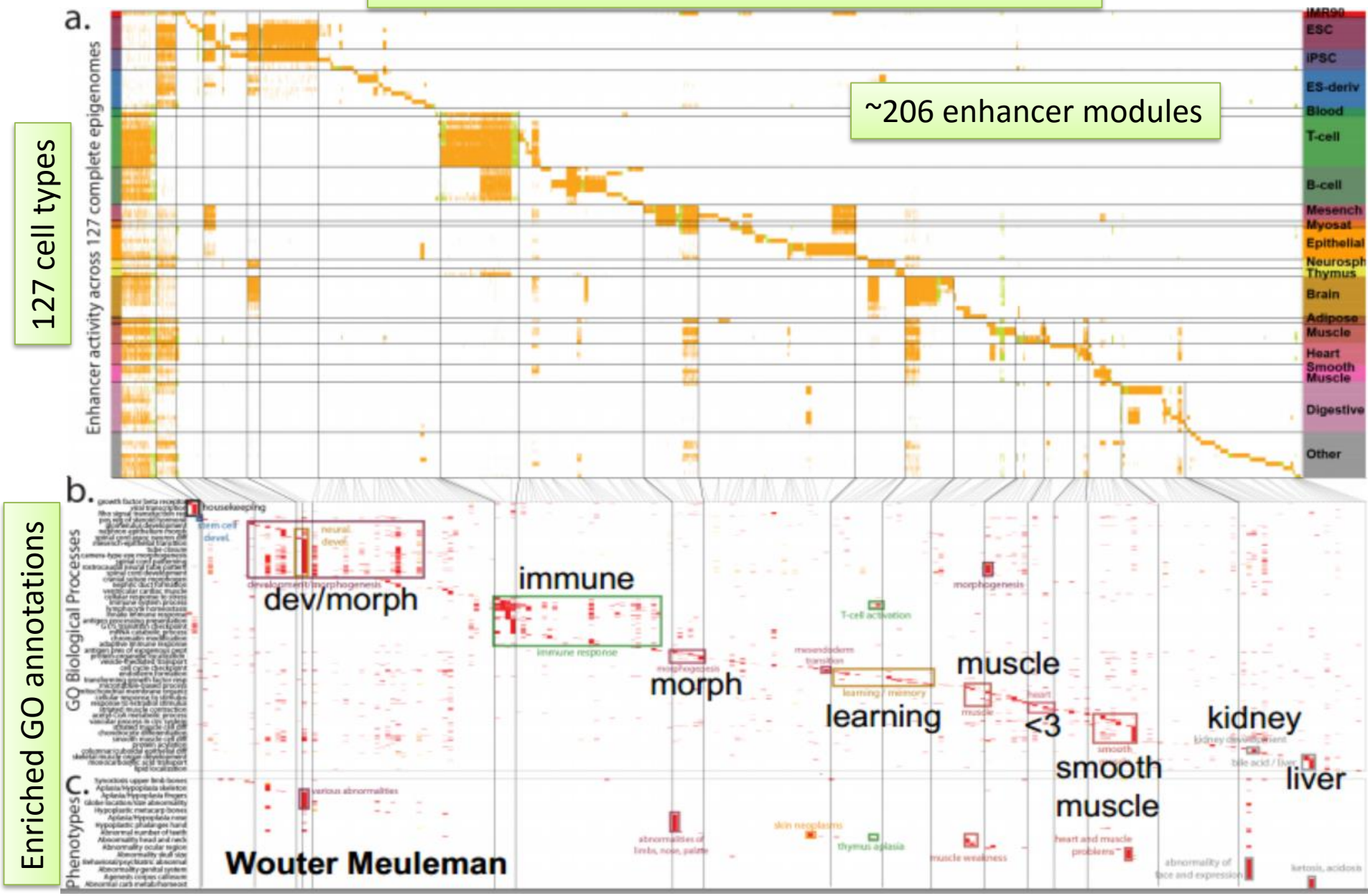


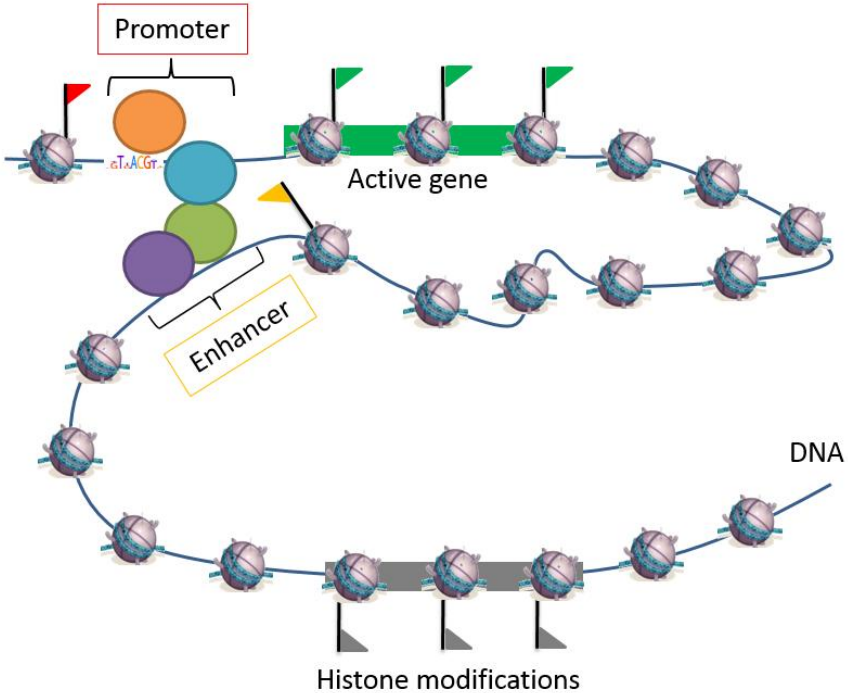
# Chromatin state dynamics across cell-types/tissues



# Modular chromatin activity dynamics of 2 million enhancers

~2M chromatin accessible sites in enhancer states

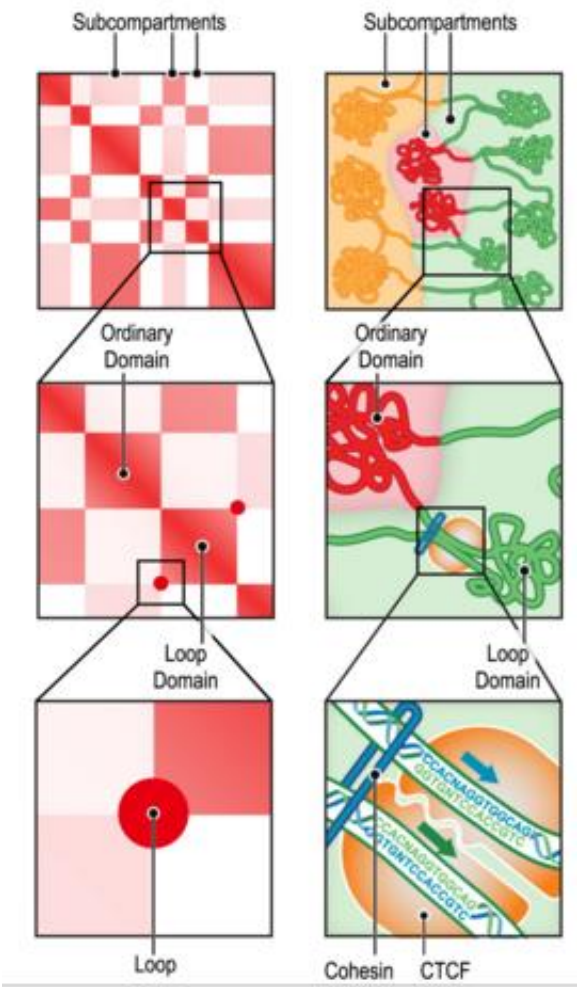




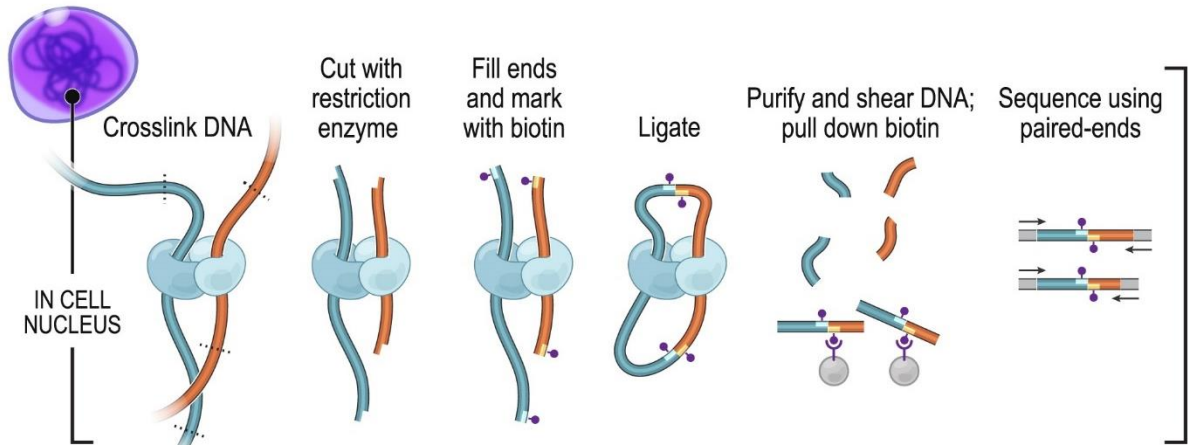

# Which genes do these distal enhancers regulate in different cell types?

With Jianrong Wang, MIT

# Chromatin contact mapping experiments



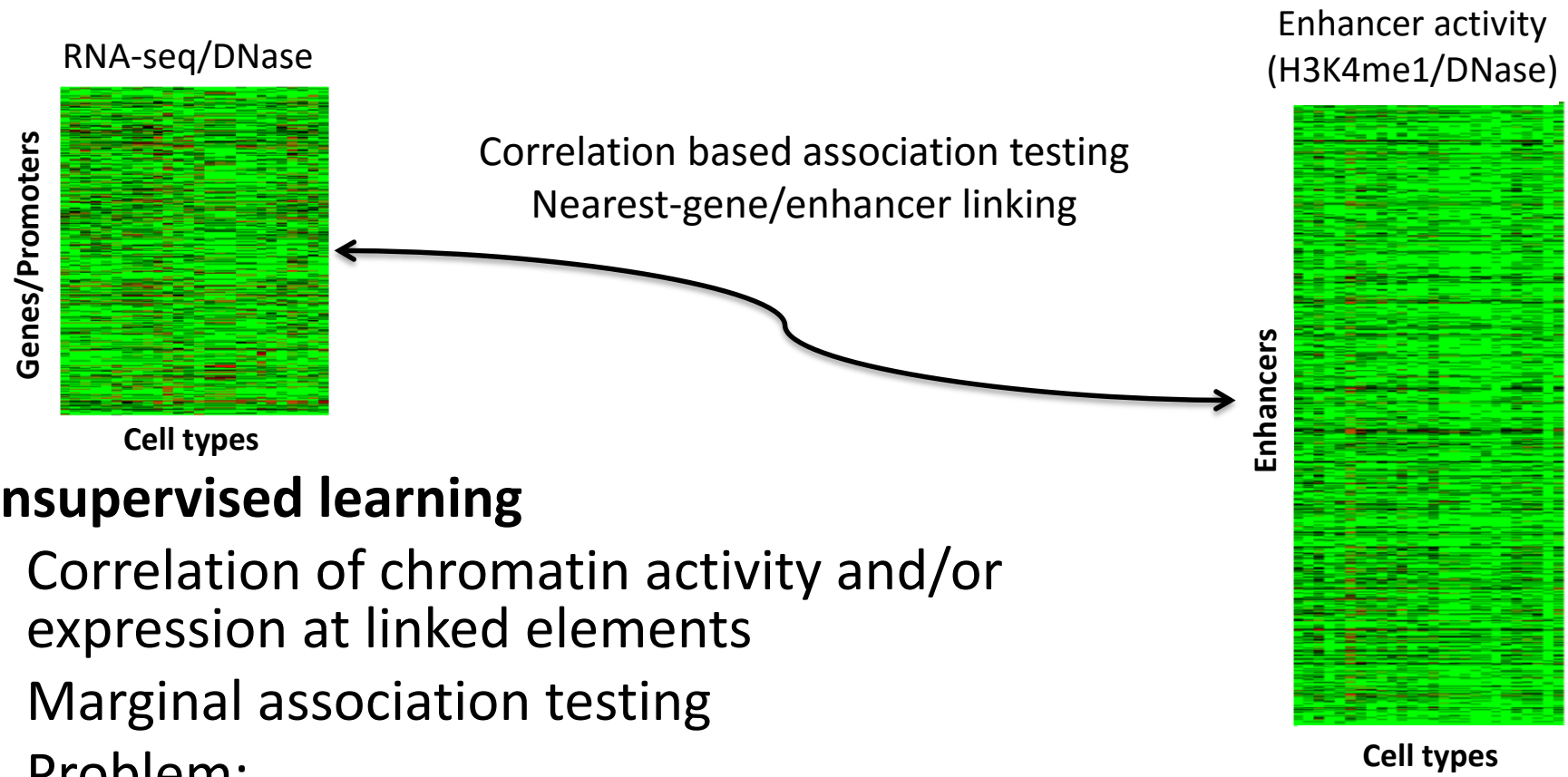

*Rao et al. 2015*




*Rao et al. 2015*

- HiC: Course-grained interaction maps (~5-10Kb fragments with billions of reads)
- ChIA-PET: interactions involving specific proteins
- High cost and requires millions of cells
- Primarily highlight cell-type invariant, stable loops generally involving CTCF/cohesin
- Low signal-to-noise ratio for reliable detection of dynamic enhancer-promoter interactions
- Only available for a few cell types (mostly cell-lines)

# Computational methods for linking distal elements to genes




## Unsupervised learning

- Correlation of chromatin activity and/or expression at linked elements
- Marginal association testing
- Problem:
  - Significantly under-powered due to huge multiple testing burden
  - Expects global correlation between linked elements
  - Difficult to assign cell-type specificity of links

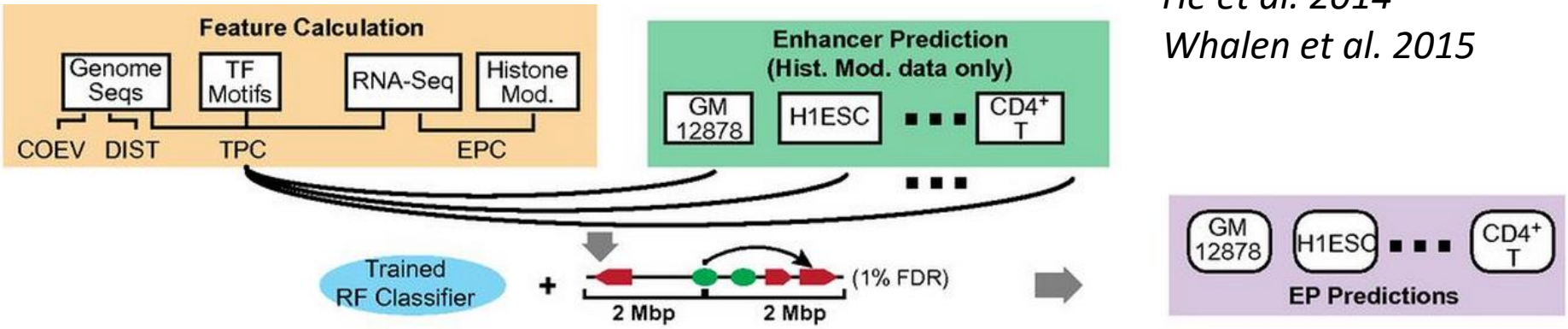
*Ernst et al. 2011*
*Thurman et al 2012*

# Computational methods for linking distal elements to genes

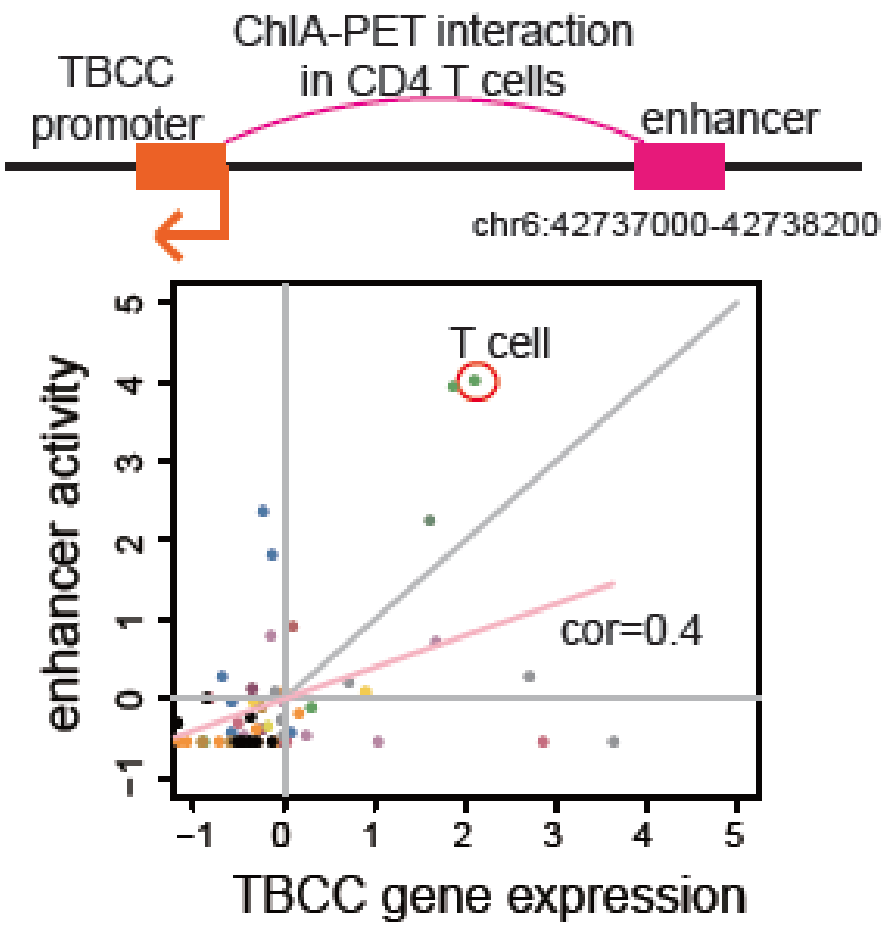
**Supervised learning**

- Training data: Use experimentally obtained links as training examples
- Features: chromatin activity, expression, sequence, TF binding sites
- Use a supervised classification/regression method
- Problems: Training data is usually very sparse (few 1000 positives) and features are often cell type specific

*He et al. 2014*
*Whalen et al. 2015*



# What is the relationship between chromatin and expression co-dynamics at linked enhancer-promoter pairs across cell types

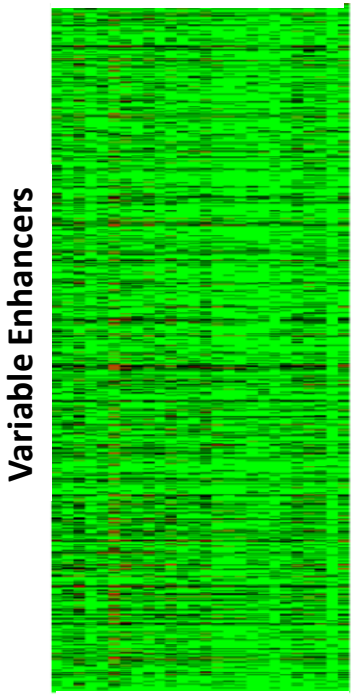


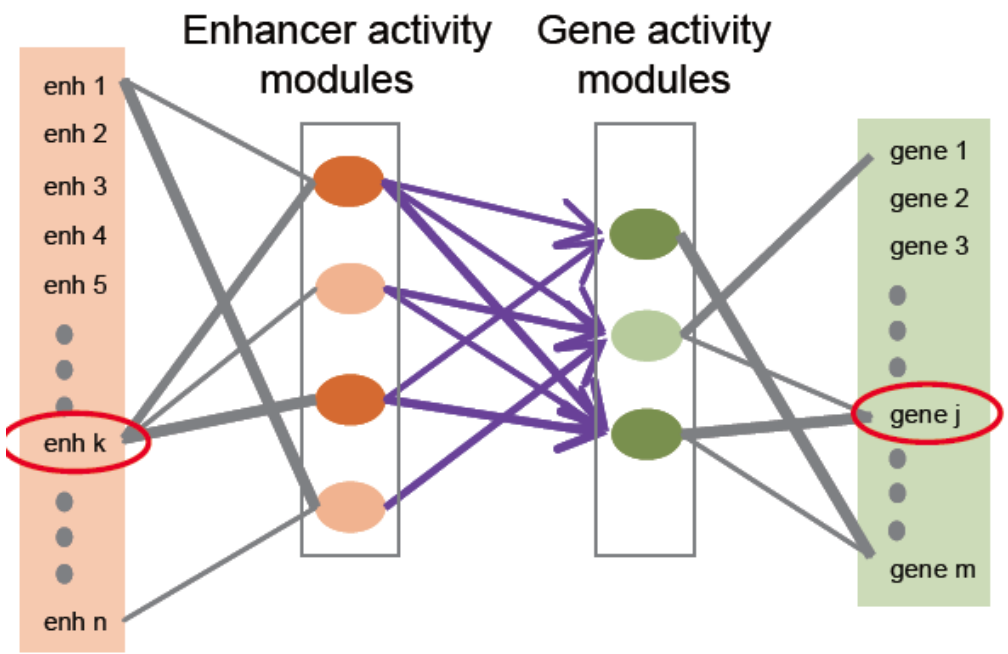

- Enhancer activity is sparse across cell types: **non-linear** effects on expression dynamics
- **many-to-many map**: multiple enhancers with multiple genes
- Links are not invariant across cell-types: **cell-type specific links**

*Marstand et al. 2014*

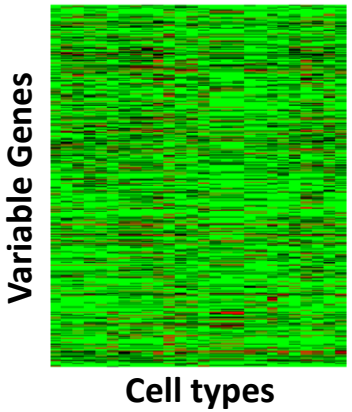
# A novel prob. model for enhancers-gene linking using chromatin-expression dynamics

Enhancer activity
(H3K4me1/H3K27me3)




Gene expression
(RNA-seq)




**Joint learning of mixed-membership probabilistic model**

Mixed membership gene modules (which genes active in which cell types)

Mixed membership enhancer modules (which enhancers active in which cell types)

Prob. non-linear linking of Gene module to enhancer module

Cell-type specific enhancer to gene linking

*With Jianrong Wang, Manolis Kellis MIT*

# Mixed membership model for enhancer/gene modules

Latent Dirichlet Allocation (LDA) like mixed-membership "topic" model allows each gene/enhancer to belong to multiple modules, to learn the module structures of genes/enhancers.

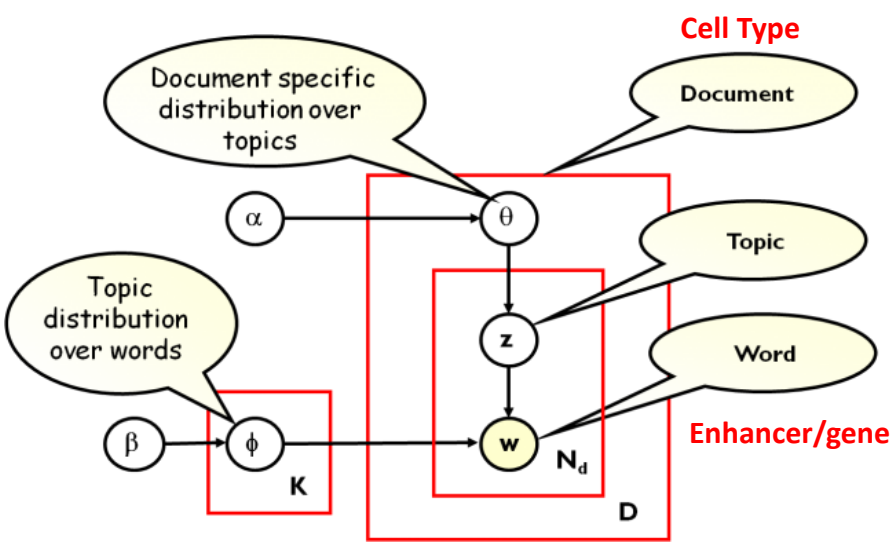
$$\vartheta_{t,k} = p(z = k|t)$$

Fraction probability: the probability of observing the $k_{th}$ module in the $t_{th}$ cell type.

$$\varphi_{k,n} = p(e_n|z = k)$$

Profile probability: the probability of observing a signal unit from the $n_{th}$ loci if the $n_{th}$ loci belong to the $k_{th}$ module.

$$p(e_n|t) = \sum_k p(e_n|z = k)p(z = k|t)$$



# Non-linear linking of enhancer-gene modules

Associated enhancer/gene modules should show similar prob. for certain **critical tissues (not all tissues)**: the tissue which has the maximal prob. for each enhancer module.

<u>Module fraction probability matrices</u>

(Which cell types is a enhancer/gene module most active in?)

| | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ |
|---|---|---|---|---|---|
| Enhancer module 1 | $\vartheta_{1,1}$ | $\vartheta_{2,1}$ | $\vartheta_{3,1}$ | $\vartheta_{4,1}$ | $\vartheta_{5,1}$ |
| Enhancer module 2 | $\vartheta_{1,2}$ | $\vartheta_{2,2}$ | $\vartheta_{3,2}$ | $\vartheta_{4,1}$ | $\vartheta_{5,1}$ |
| Enhancer module 3 | $\vartheta_{1,3}$ | $\vartheta_{2,3}$ | $\vartheta_{3,3}$ | $\vartheta_{4,1}$ | $\vartheta_{5,1}$ |

| | $t_1$ | $t_2$ | $\mathbf{t_3}$ | $t_4$ | $t_5$ |
|---|---|---|---|---|---|
| Gene module 1 | $\vartheta_{1,1}$ | $\vartheta_{2,1}$ | $\vartheta_{3,1}$ | $\vartheta_{4,1}$ | $\vartheta_{5,1}$ |
| Gene module 2 | $\vartheta_{1,2}$ | $\vartheta_{2,2}$ | $\vartheta_{3,2}$ | $\vartheta_{4,2}$ | $\vartheta_{5,2}$ |

$$A = (a_{ij})_{K_1 \times K_2}$$

$a_{ij}$ the posterior probability that the $i_{th}$ enhancer module is associated with the $j_{th}$ gene module

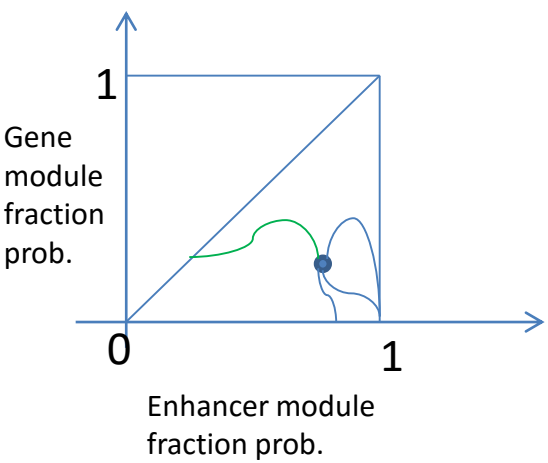
# Non-linear linking of enhancer-gene modules

Associated enhancer/gene modules should show similar prob. for certain **critical tissues (not all tissues)**: the tissue which has the maximal prob. for each enhancer module.

$a_{ij}$ the posterior probability that the $i_{th}$ enhancer module is associated with the $j_{th}$ gene module

$$a_{ij} = P\left(E_i \sim G_j \middle| \vartheta_{t,i}^1;\ \vartheta_{t,1}^2, \vartheta_{t,2}^2 \dots \vartheta_{t,j}^2 \dots \vartheta_{t,K_2}^2\right) = P(E_i \sim G_j | \vartheta_{t,i}^1;\ \overrightarrow{\vartheta_{t,\cdot}^2})$$

$$= \frac{P\left(\vartheta_{t,i}^1;\ \overrightarrow{\vartheta_{t,\cdot}^2} \middle| E_i \sim G_j\right)}{\sum_{h=1}^{K_2} P\left(\vartheta_{t,i}^1;\ \overrightarrow{\vartheta_{t,\cdot}^2} \middle| E_i \sim G_h\right) + P\left(\vartheta_{t,i}^1;\ \overrightarrow{\vartheta_{t,\cdot}^2} \middle| E_i \sim \emptyset\right)}$$




We use a diffusion model to estimate the probabilities

$$P\left(\vartheta_{t,i}^1;\ \overrightarrow{\vartheta_{t,\cdot}^2} \middle| E_i \sim G_j\right) = (1 - |\vartheta_{t,i}^1 - \vartheta_{t,j}^2|) \prod_{\substack{h=1 \\ h \neq j}}^{K_2} |\vartheta_{t,i}^1 - \vartheta_{t,h}^2|$$

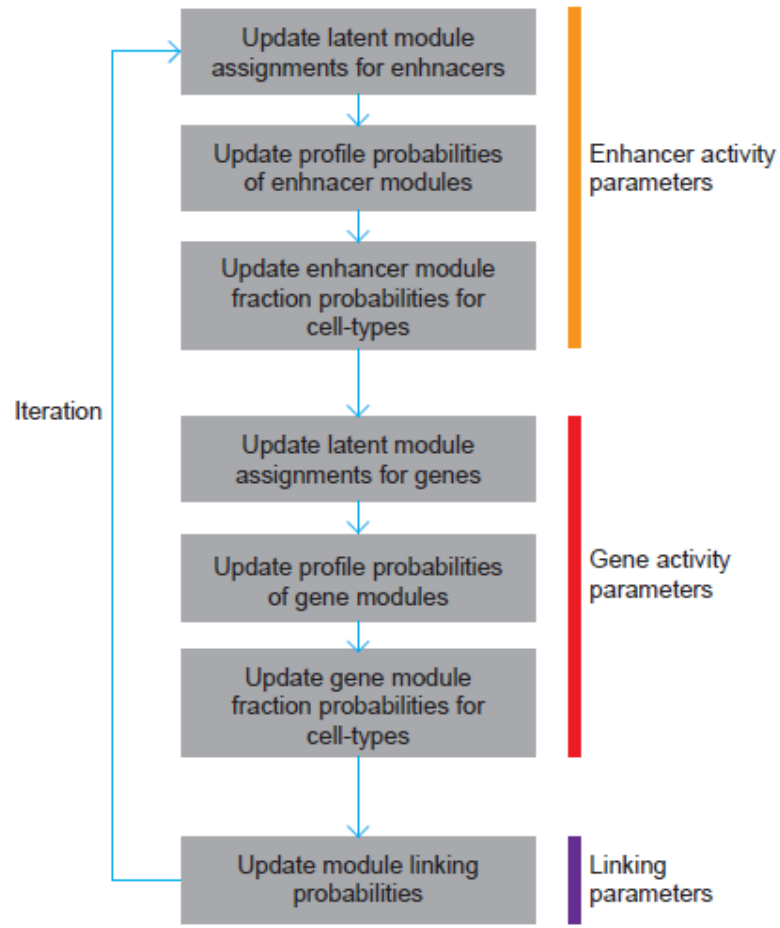
# Joint learning of modules and associations

Key parameters inferred for the model:

1. Module profile probabilities ($\varphi_{i,\cdot}$): indicate most informative enhancers/genes for each module;
2. Module fraction probabilities across cell-types ($\vartheta_{t,\cdot}$): how relevant a module is for each cell-type;
3. Association probabilities ($a_{i,j}$): relations between enhancer modules and gene modules

We use sparsity inducing regularization to deal with the small number of cell types to learn from.

Joint Gibbs Sampling Approach



# Cell-type specific prob. enhancer-gene links

$$P\left(e_i \sim g_j | t\right) = \sum_{k=1}^{K_1} \varphi_{k,i}^1 \, \vartheta_{t,k}^1 \left( \sum_{h=1}^{K_2} a_{kh} \, \vartheta_{t,h}^2 \varphi_{h,j}^2 \right)$$

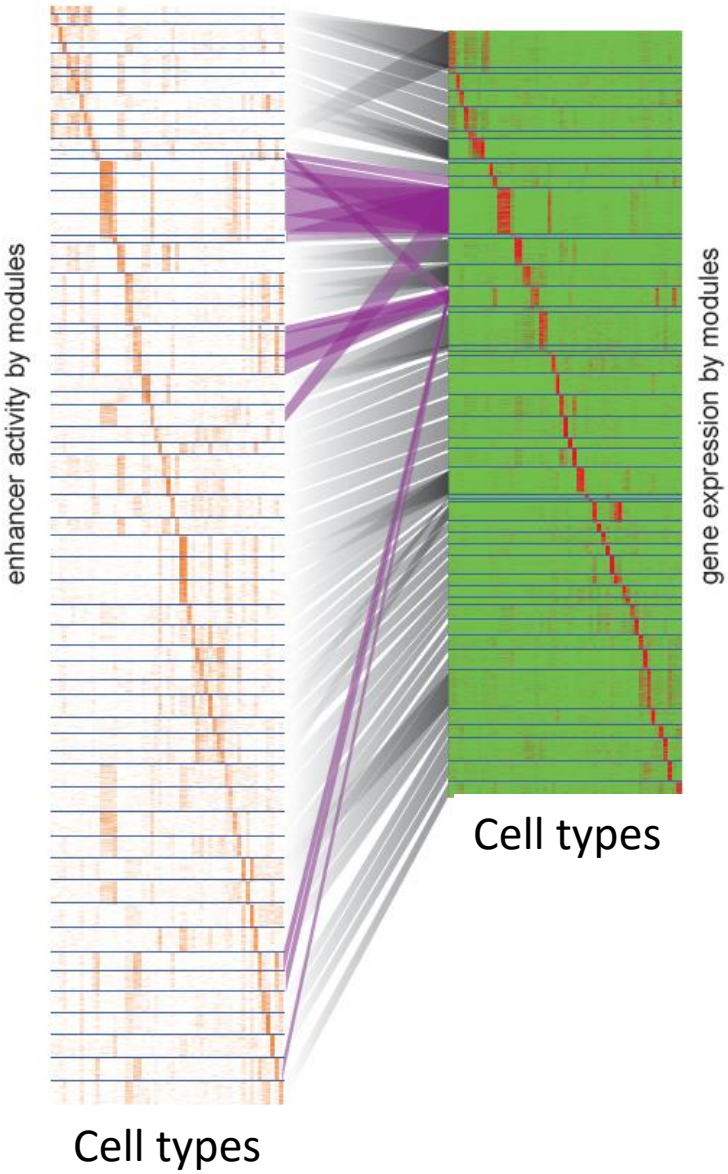
Key parameters inferred for the model:

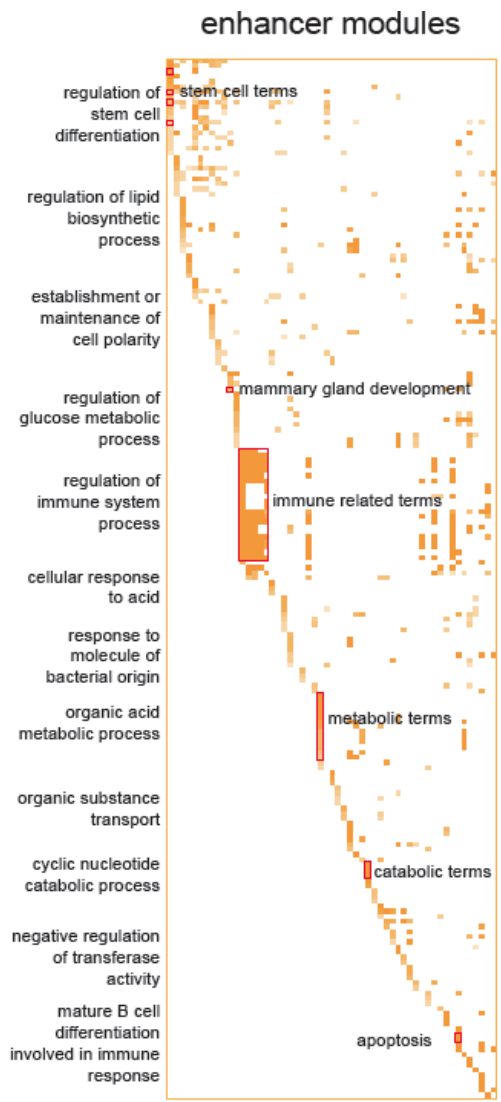
1. Module profile probabilities ($\varphi_{i,\cdot}$): indicate most informative enhancers/genes for each module;
2. Module fraction probabilities across cell-types ($\vartheta_{t,\cdot}$): how relevant a module is for each cell-type;
3. Association probabilities ($a_{i,j}$): relations between enhancer modules and gene modules;

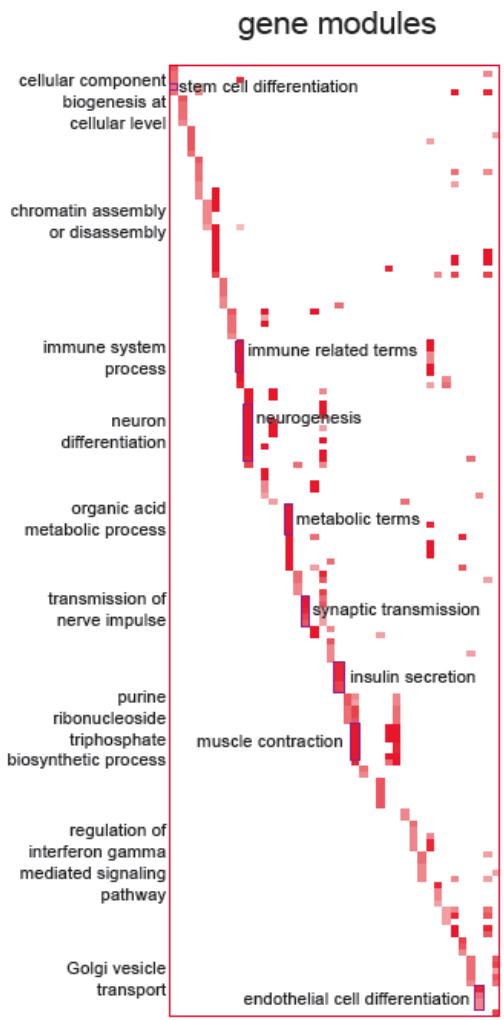
# Calling statistically significant links

- Focus on links that are within 1Mb of each other (or use TADs)
- Null distribution: Linking probabilities on shuffled data with distance-related prior distributions
- Generate *P*-values for enhancer-gene links on the real data
- The Benjamini-Hochberg method is used for multiple hypothesis correction.
- We use FDR 1% (stringent) and 5% (relaxed)

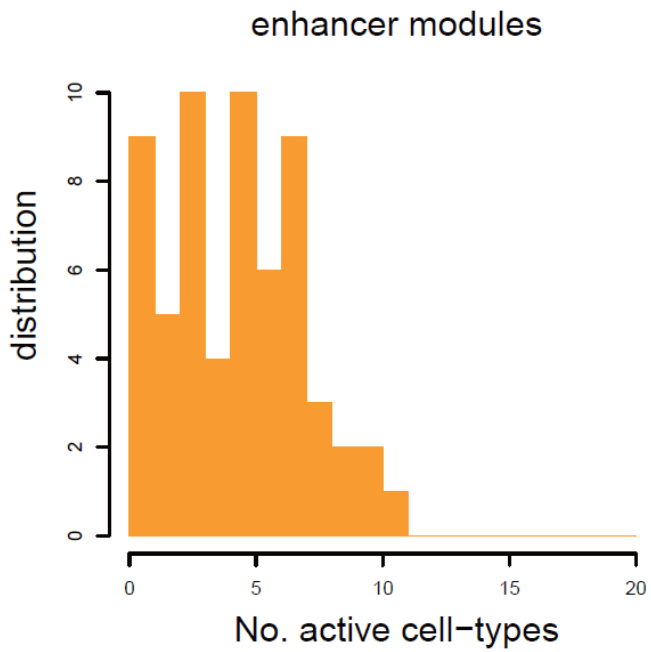
# Learned enhancer-gene modules and their links

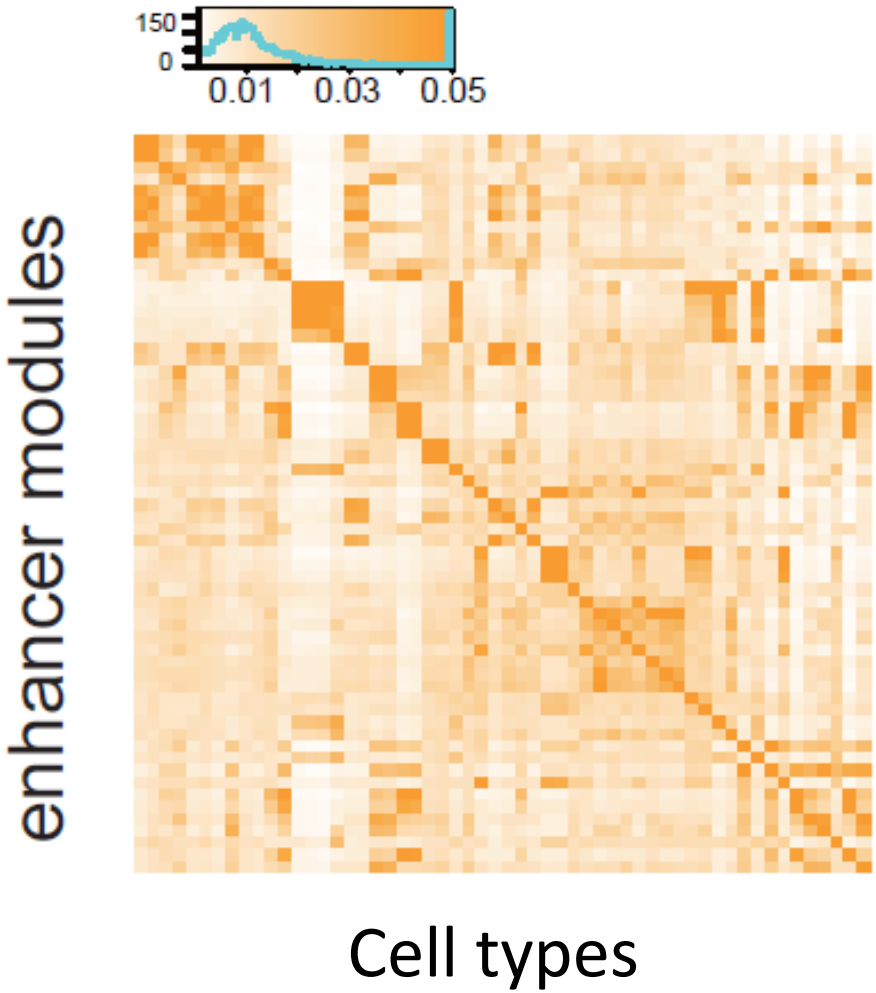


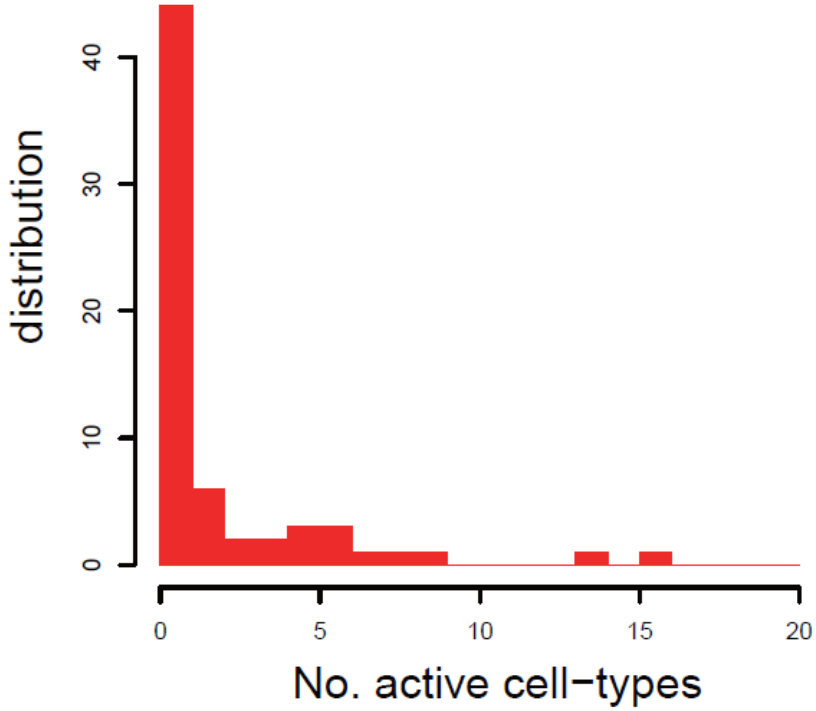



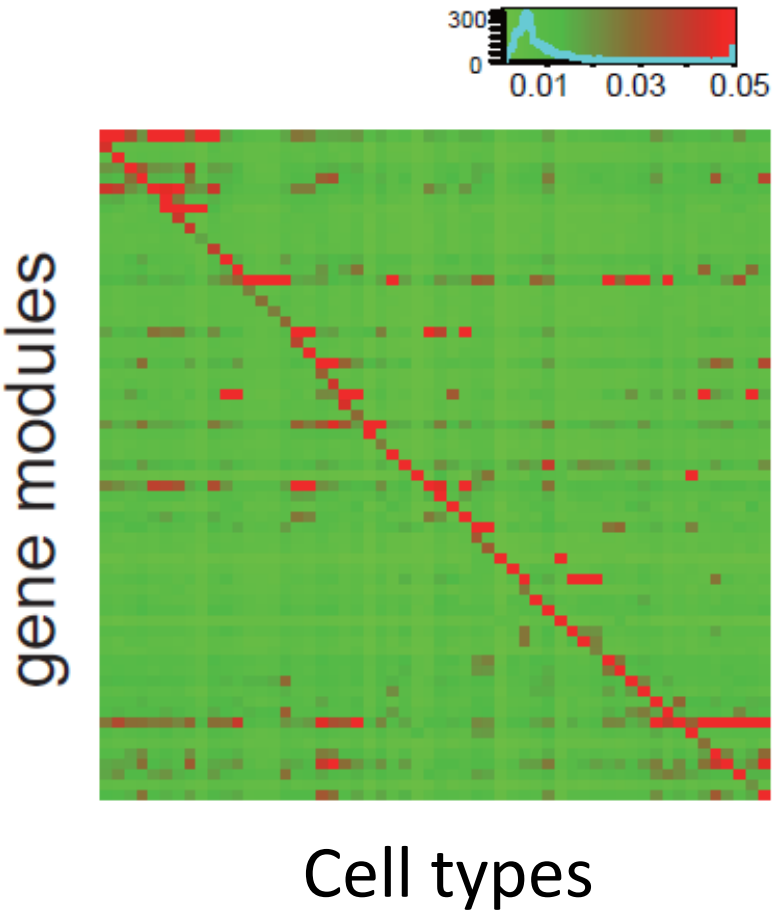
# 61 enhancer modules and their tissue-specificity



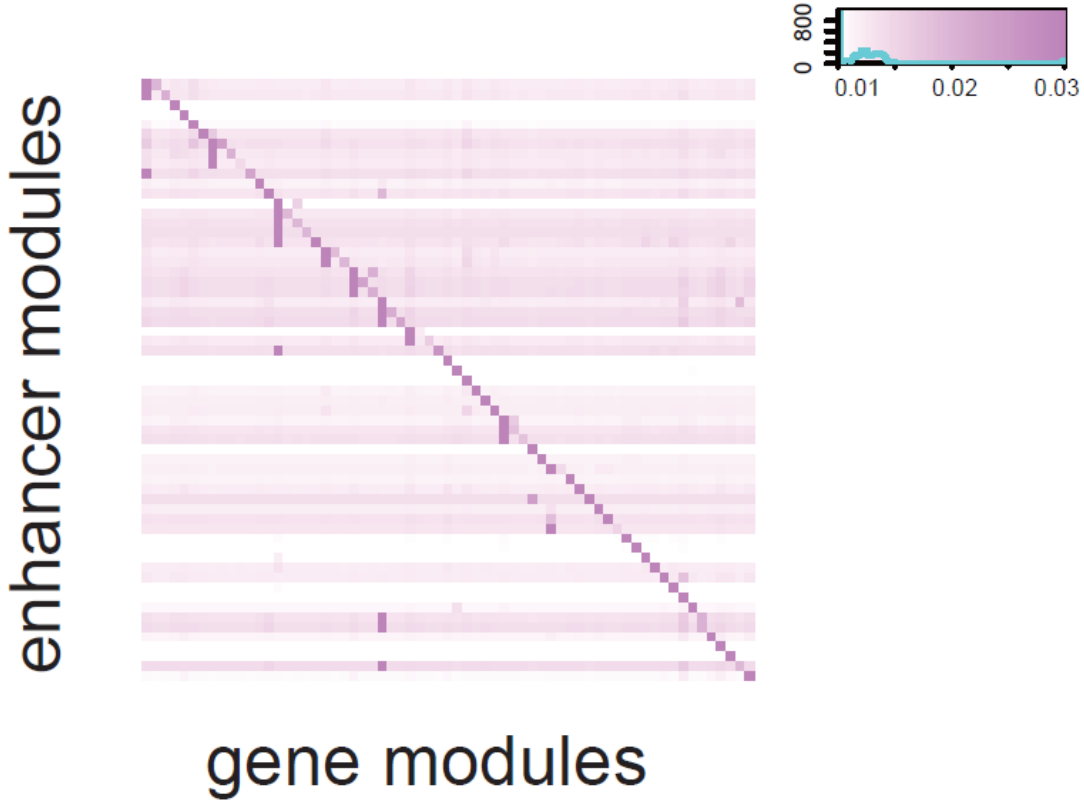



# 65 gene modules and their tissue-specificity



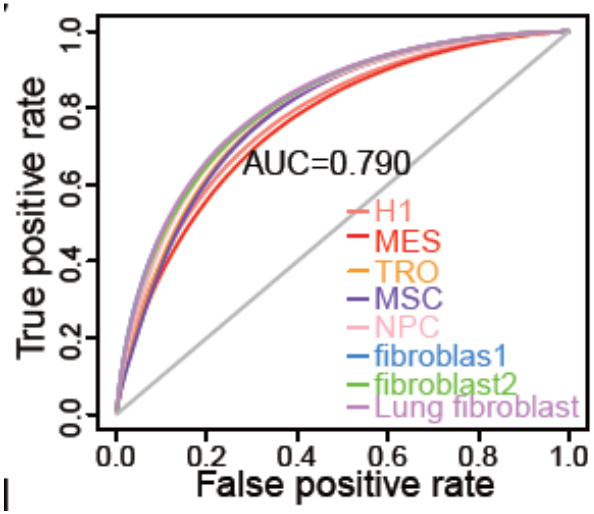
# Module linking probabilities



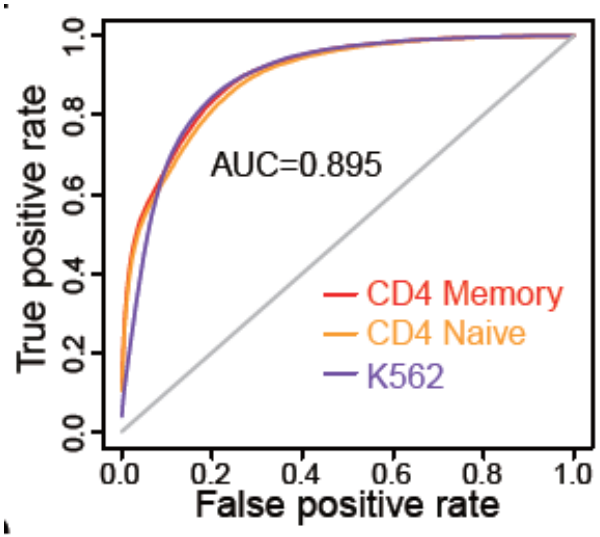

At a false discovery rate (FDR) of 1%, we found 249k statistically significant enhancer-gene links across all cell types, linking 132,419 enhancers to 15,465 target genes

# Predicted cell-type specific enhancer-gene are globally predictive of HiC, ChIA-PET data + eQTLs

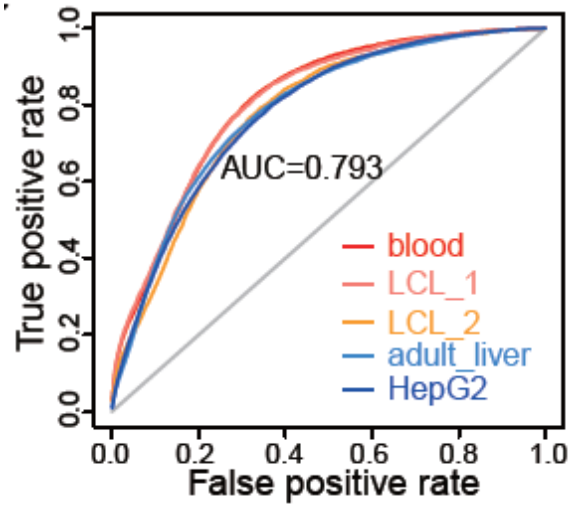
HiC



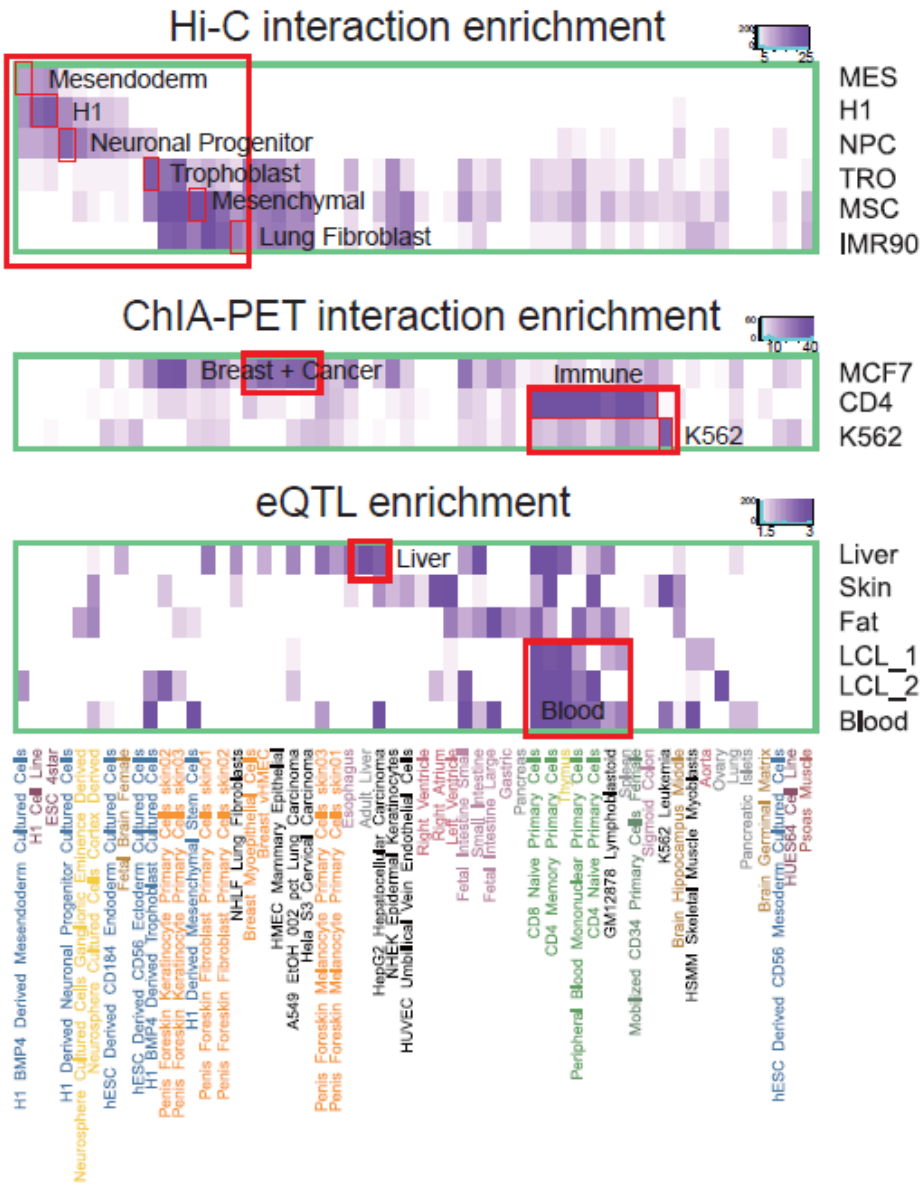

ChIA-PET




eQTLs



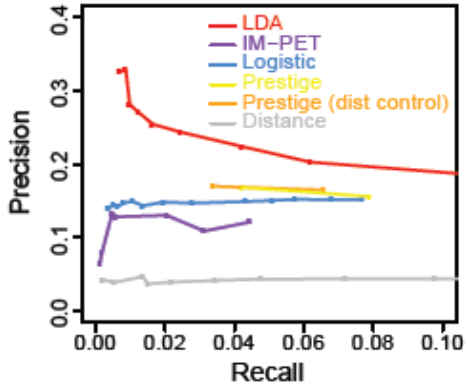
# Predicted cell-type specific enhancer-gene are globally predictive of HiC, ChIA-PET data + eQTLs

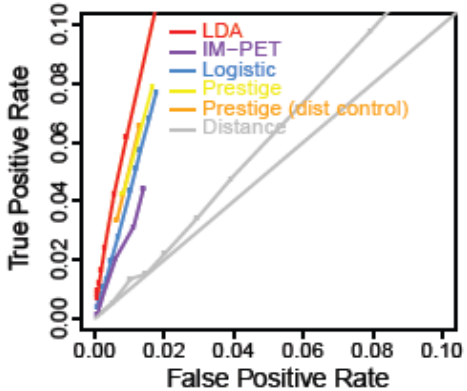


# Comparison to other methods

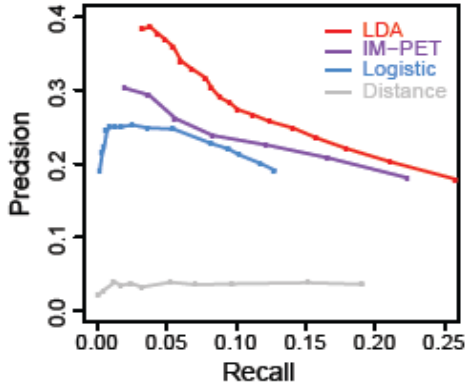
Precision-Recall | ROC

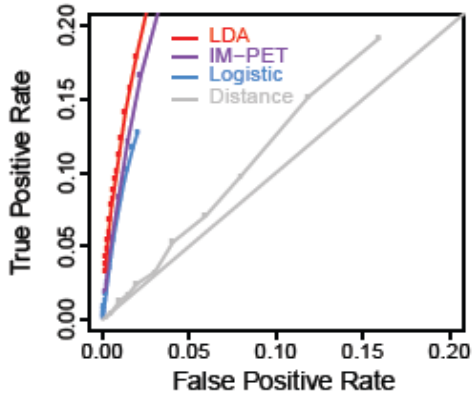
IMR90 Hi-C
Dixon et al. Nature 2015



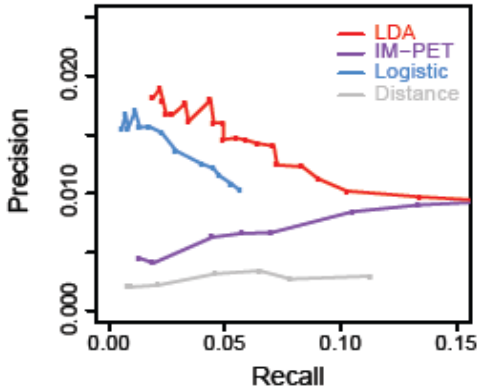



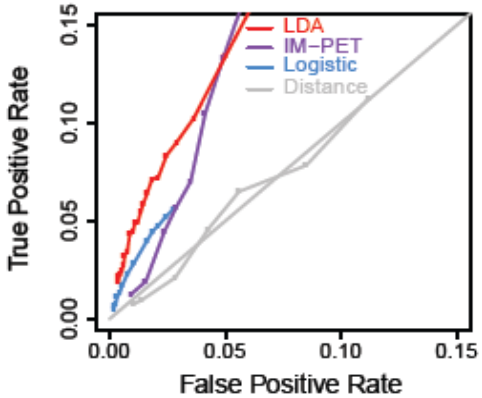

CD4 T cell ChIA-PET
Chepelev et al. Cell Research 2012



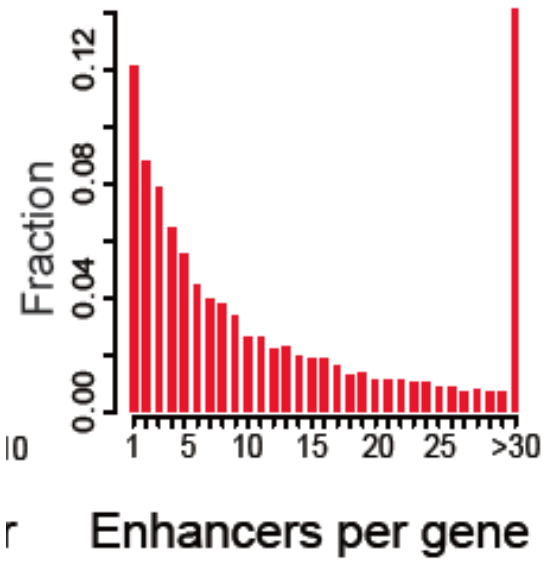



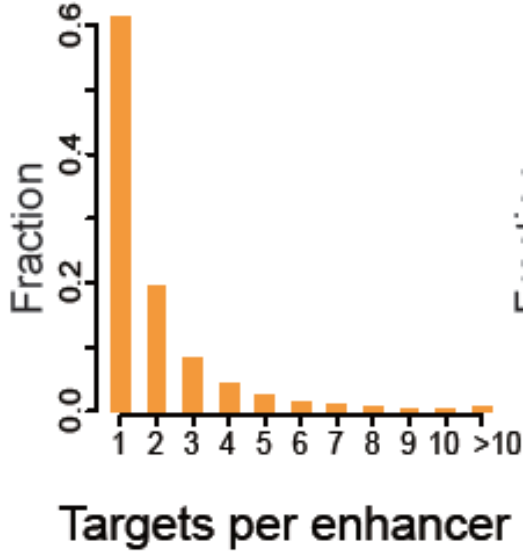

whole blood eQTL
Battle et al. Genome Research 2014

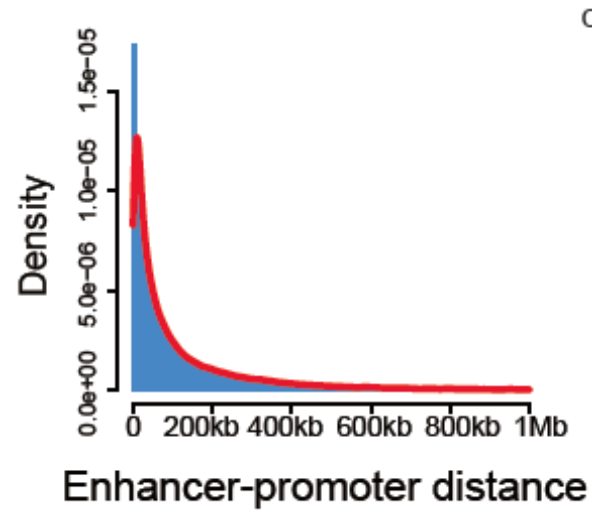


# Properties of long-rang enhancer-gene network

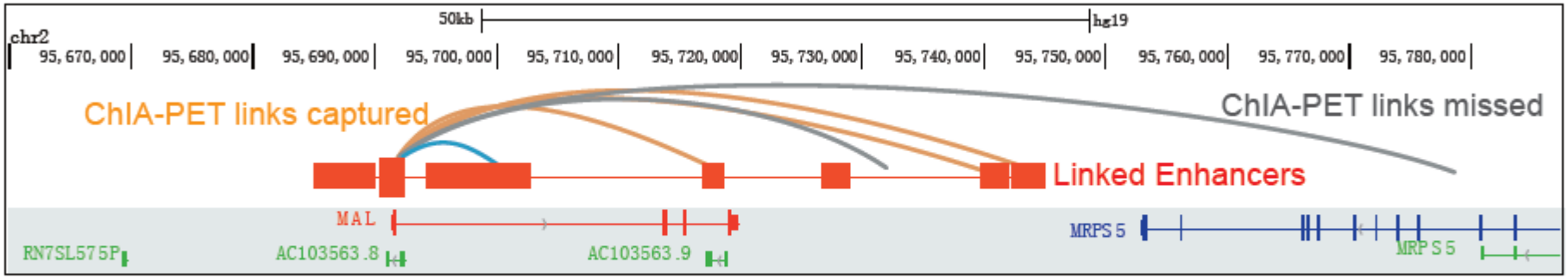


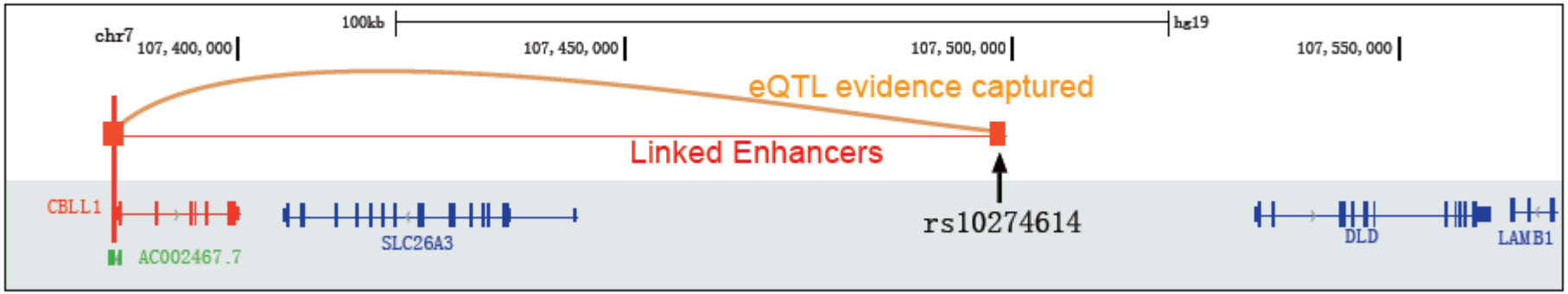




- **The enhancer-gene network is highly connected**
  - 88% of genes and 39% of enhancers are multiply linked
- **Links are highly tissue-specific**
  - 56% of links specific to one lineage
  - Only 26% found in three or more
- Half of predicted links < 50kb apart
- **Only a third of enhancers are linked to a nearest gene**

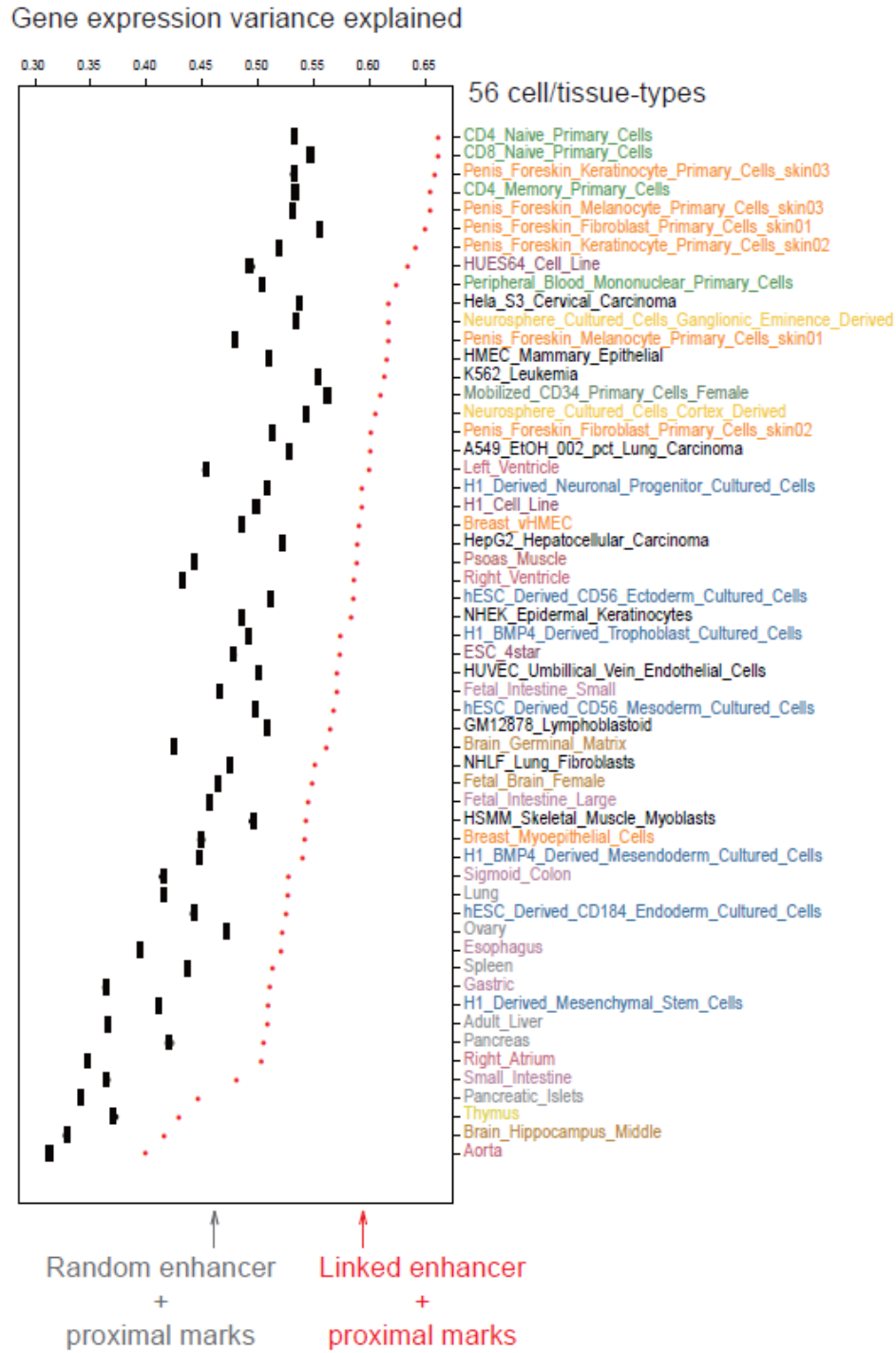
# Enhancers often do NOT associate with their nearest promoters



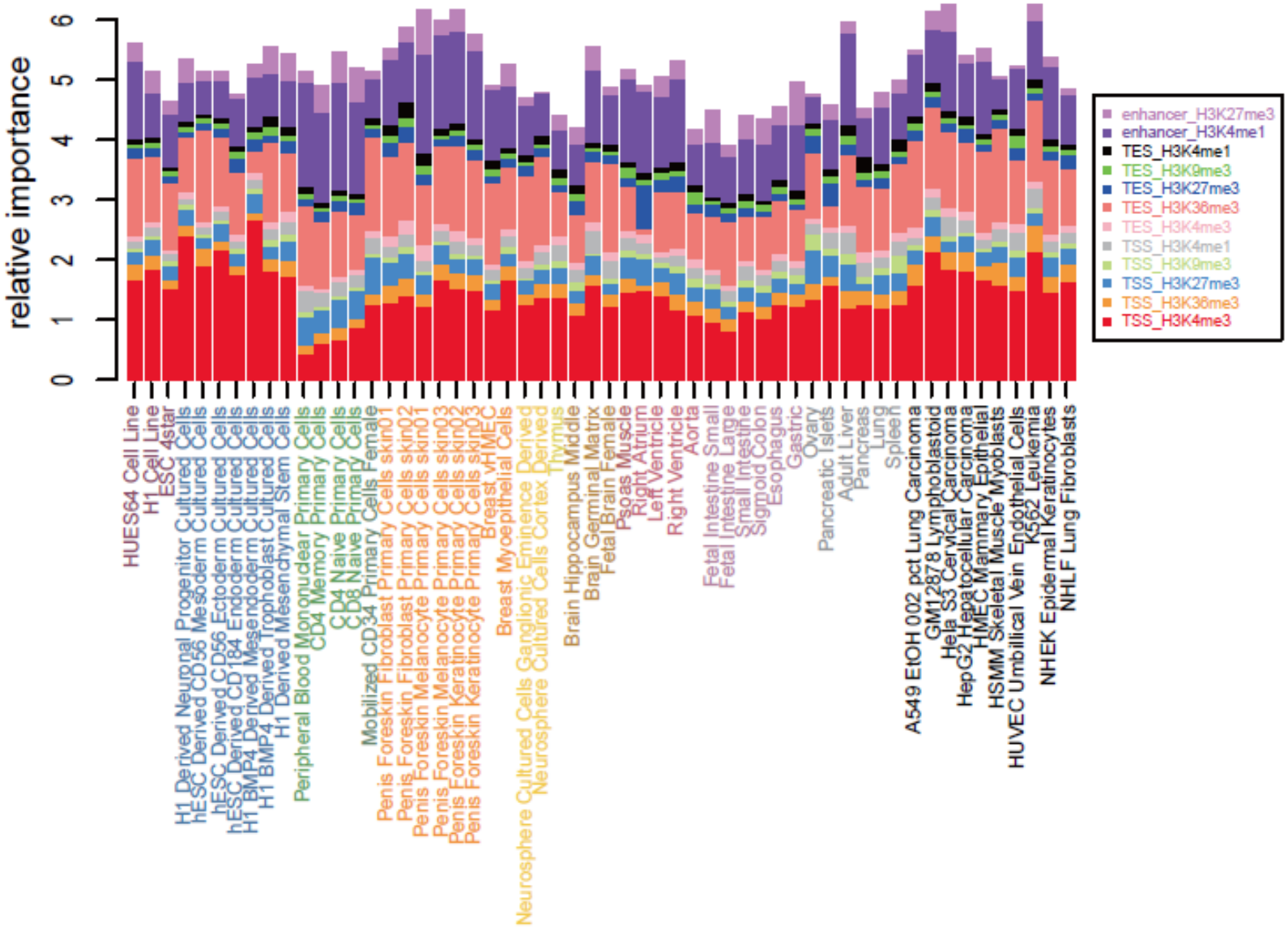



# Enhancers explain significant proportion of gene expression variance in each cell type

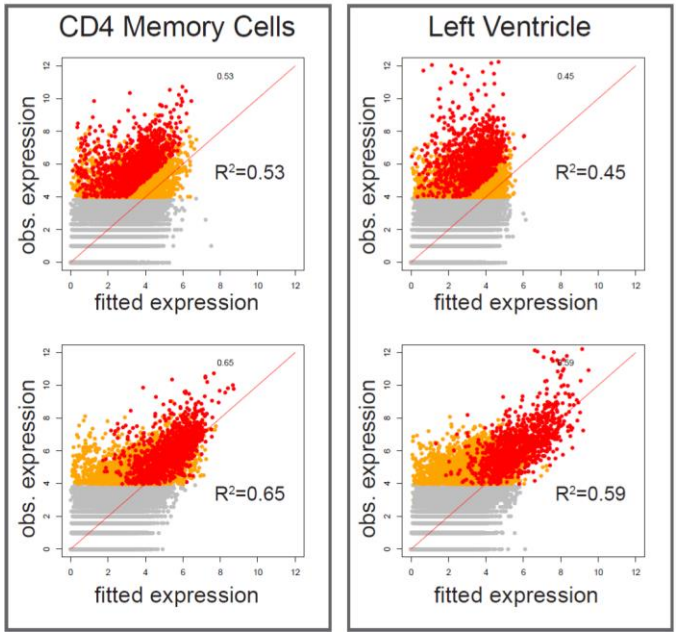
Random Forest Regression models to fit gene expression using promoter associated histone marks AND with linked enhancers vs. random distance matched enhancers



# Enhancers explain significant proportion of gene expression variance in each cell type

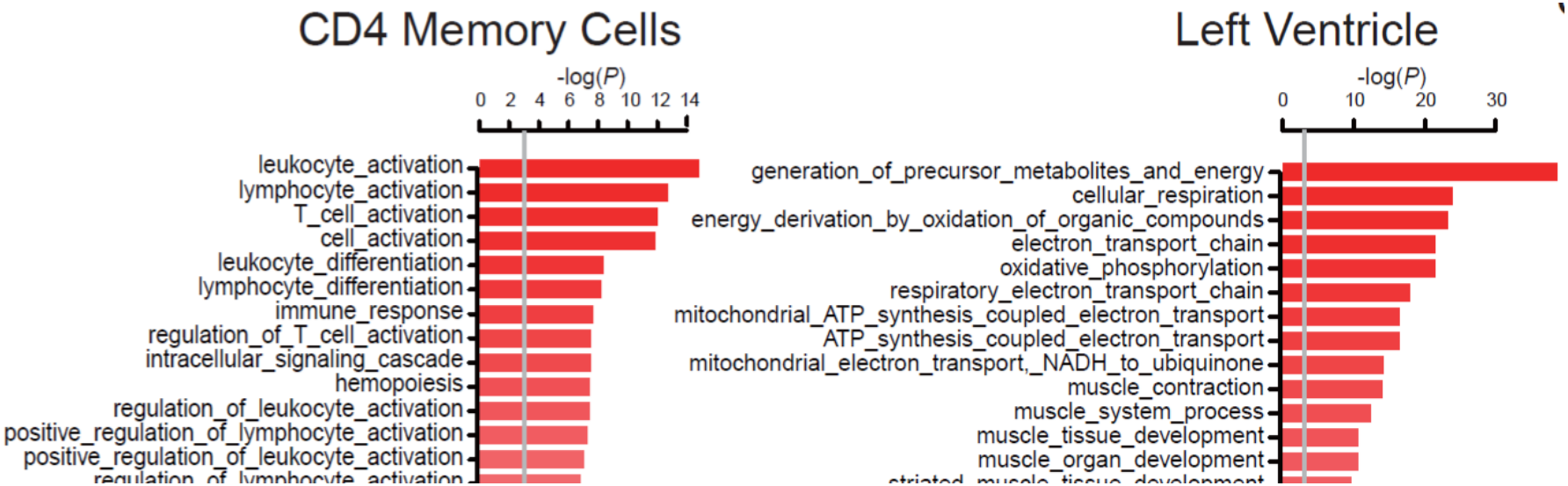


# Enhancers help explain cell-type specific gene expression

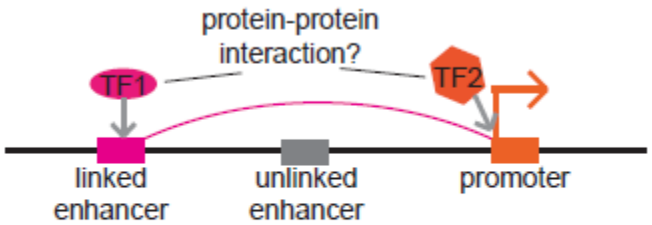


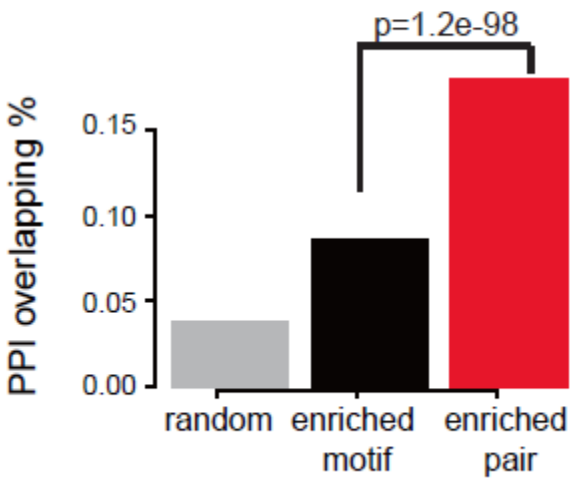

Without enhancers

With enhancers



# Linked enhancer-promoter pairs are enriched for cell-type specific TFs that are involved in protein-protein interactions

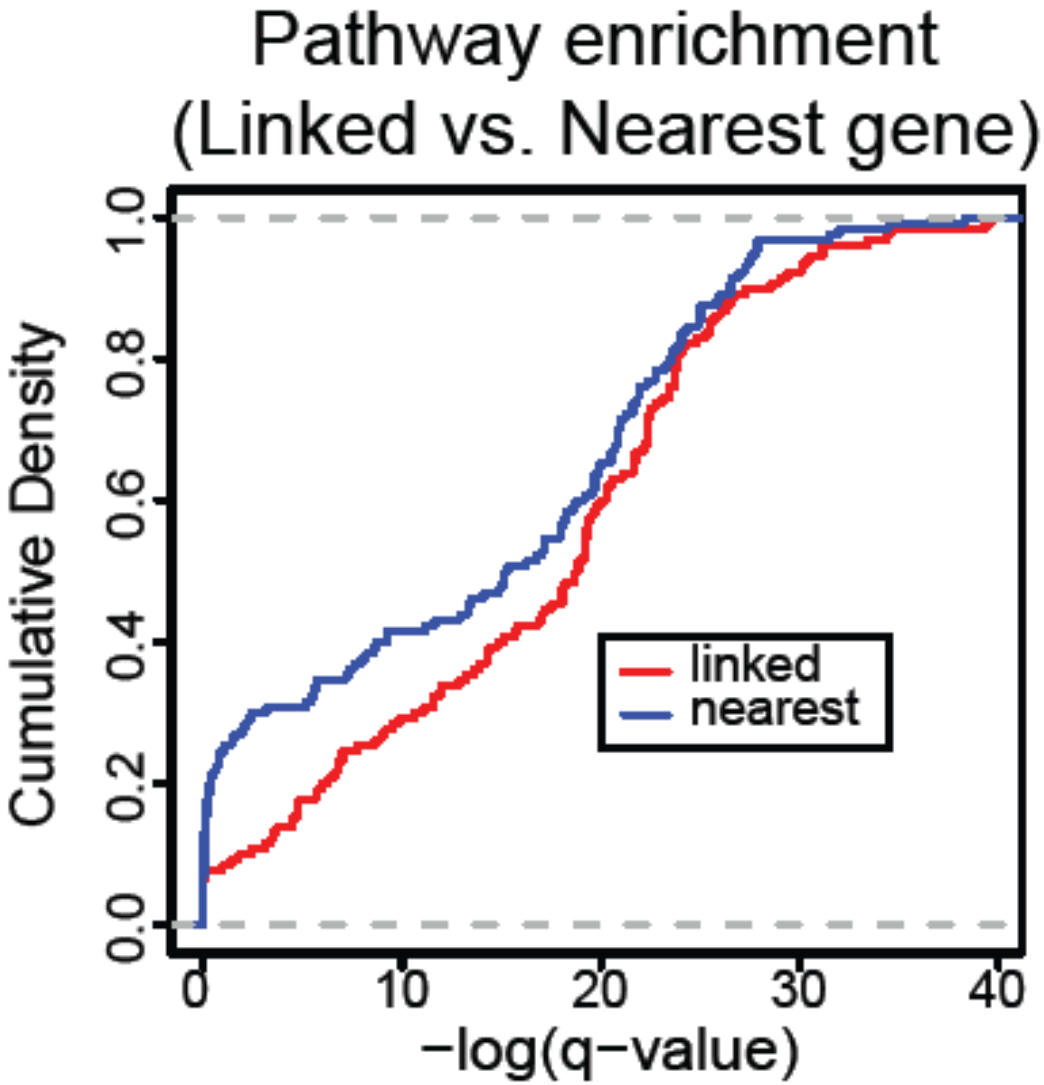



3263 motif pairs that are significantly enriched (*P*<0.05, Binomial test)

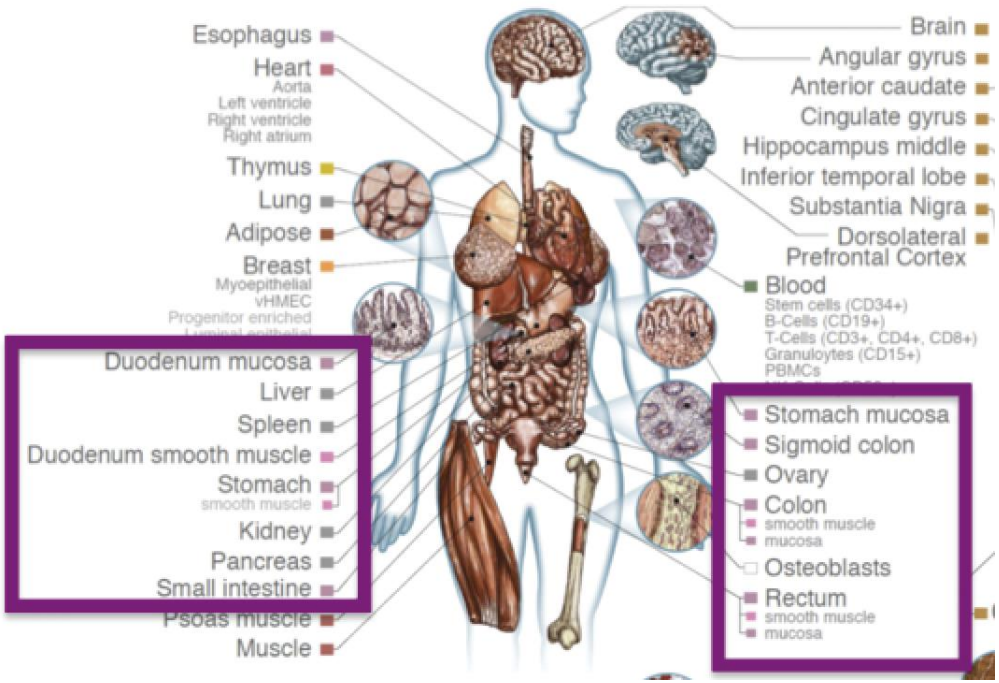
| cell-type | promoter TF | enhancer TF | Z-score |
|---|---|---|---|
| CD4 T cell | IRF1  | NFKB  | 5.90 |
| Liver | GATA4  | HNF4  | 2.75 |

# Pathway enrichments based on target genes associated with enhancers overlapping non-coding GWAS variants



# Case Study: Predict target genes of CRC variants using enhancer gene links

1. Take top CRC GWAS SNPs p-value < 10^-5
2. Expand list to all SNPs in LD with r^2 > 0.8
3. Find all enhancers active in CRC tissues that have an overlapping CRC SNP
4. Use enhancer-gene links to associate CRC SNPs to genes (76 genes)




| | | | |
|---|---|---|---|
| ADNP | DEF6 | NEU1 | TBX2 |
| ADRM1 | DHX9 | PGA3 | TEAD3 |
| ALDH2 | DIP2B | PGA4 | TMBIM1 |
| ARPC2 | DSP | PGA5 | TMBIM6 |
| ARPC5 | F3 | POU5F1 | TMEM138 |
| ATF1 | FGR | PPARD | TMEM189 |
| C11orf92 | FKBP5 | PSMA7 | -UBE2V1 |
| C11orf93 | IER3 | RAD21 | WASF2 |
| C20orf166 | IFI6 | RCSD1 | |
| CABLES2 | KANK1 | RNF114 | |
| CCND2 | LAMA5 | RNF169 | |
| CD247 | LAMC1 | SATB1 | |
| CLPS | LAMC2 | SFN | |
| CNN3 | METTL7A | SH2B3 | |
| CREG1 | MPZL1 | SMAD7 | |
| CTNNB1 | MYC | SRPK1 | |
| DDB1 | MYL2 | TBC1D5 | |

# Enrichment points to SMAD pathway, laminin complex

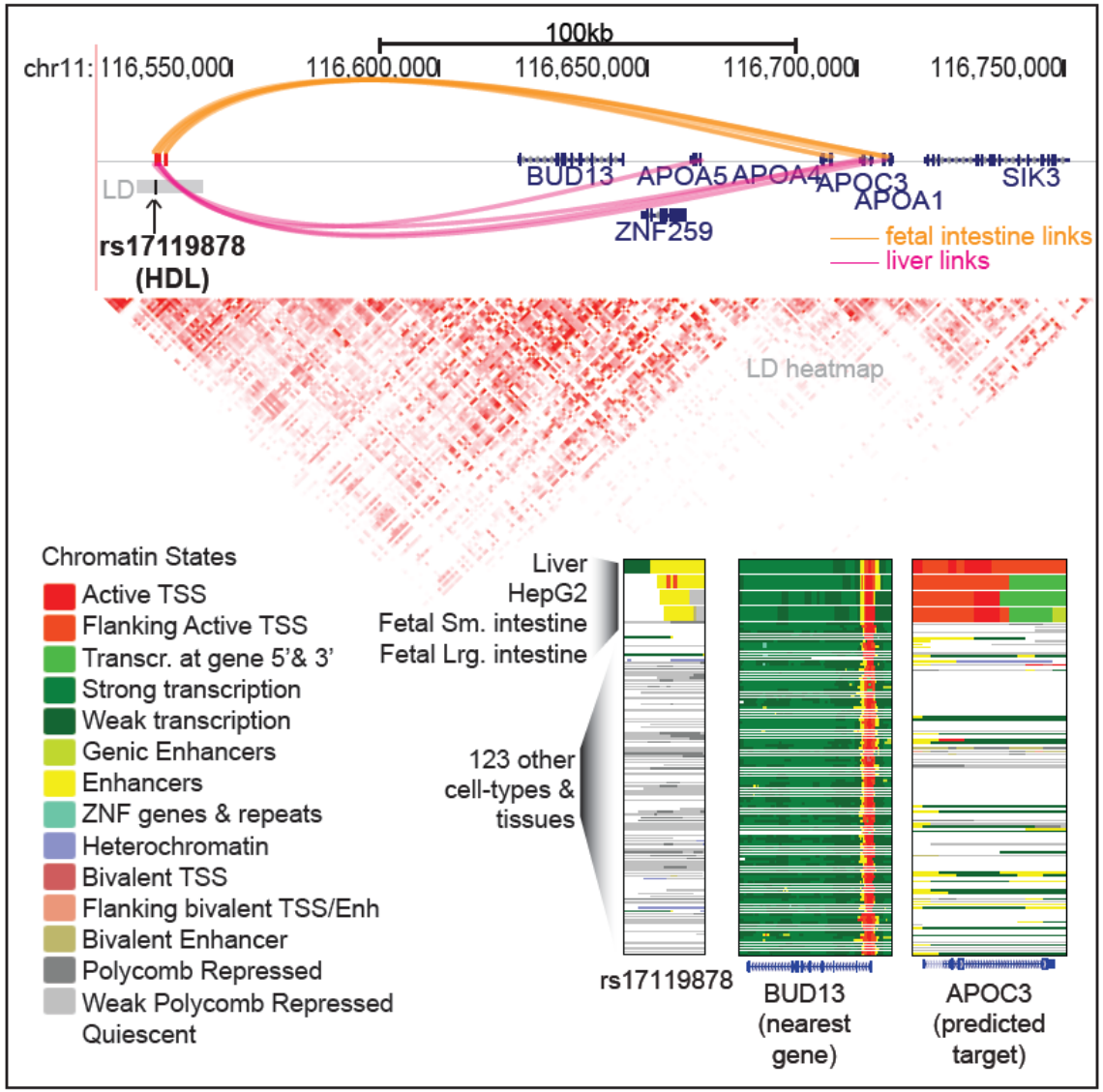
- Pathways also point to SMAD, TGFβ, and integrin

**Pathway Commons (20+ terms)** Global controls

*Table controls:* Export  Shown top rows in this table: 20 Set  Term annotation count: Min: 1 Max: Inf Set  Visualize this table: [select one]

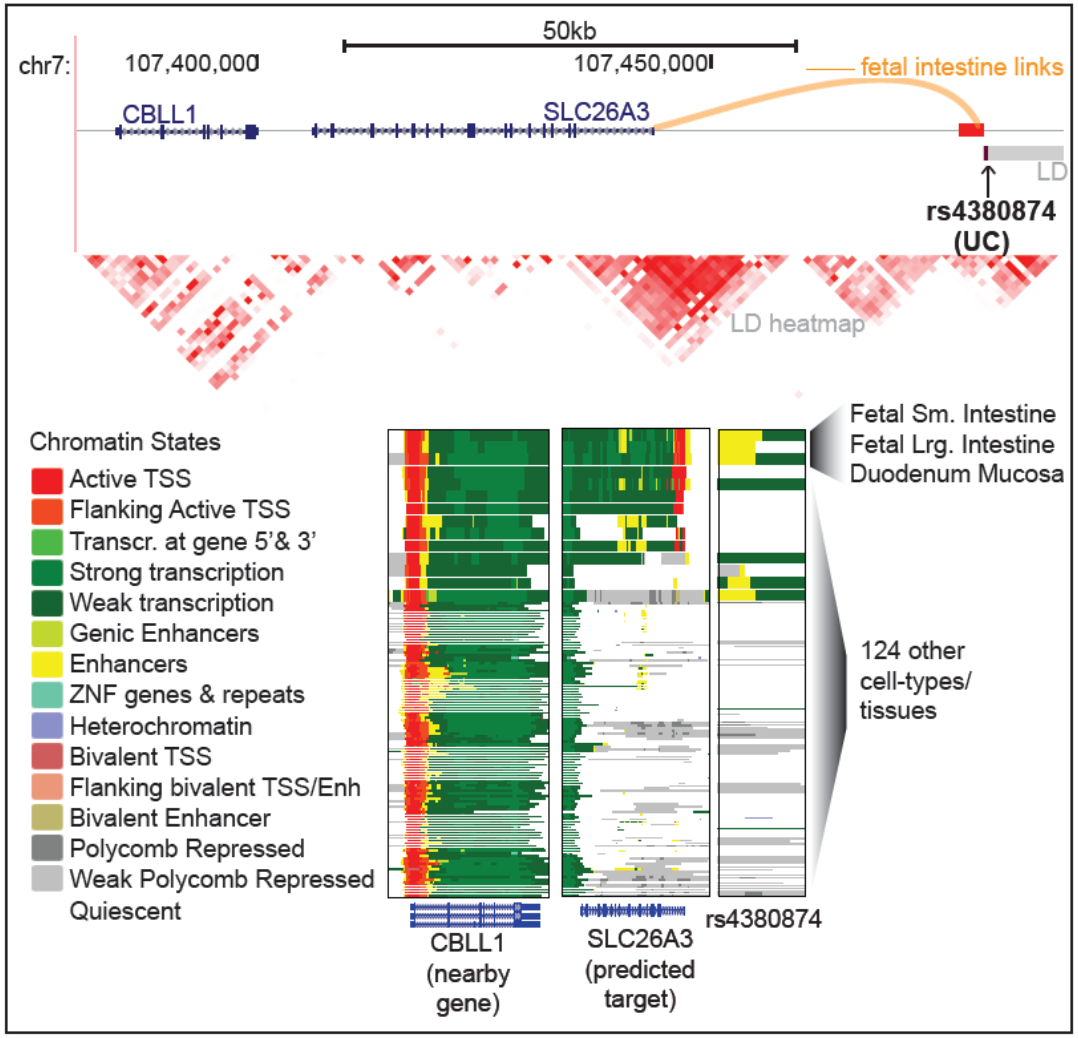
| Term Name | Hyper Rank | Hyper Raw P-Value | *Hyper FDR Q-Val* | Hyper Fold Enrichment | Hyper Foreground Region Hits | Hyper Total Regions | Hyper Region Set Coverage | Hyper Foreground Gene Hits | Total Genes Annotated |
|---|---|---|---|---|---|---|---|---|---|
| CREB phosphorylation | 1 | 2.2665e-39 | 3.6423e-36 | 88.0602 | 24 | 283 | 14.81% | 1 | 7 |
| Alpha6 beta4 integrin-ligand interactions | 2 | 1.0907e-38 | 8.7642e-36 | 71.7111 | 25 | 362 | 15.43% | 3 | 11 |
| TGF-beta receptor signaling | 3 | 5.3801e-36 | 2.8819e-33 | 6.8673 | 63 | 9,526 | 38.89% | 7 | 306 |
| Regulation of nuclear SMAD2/3 signaling | 3 | 5.3801e-36 | 2.8819e-33 | 6.8673 | 63 | 9,526 | 38.89% | 7 | 306 |
| Regulation of cytoplasmic and nuclear SMAD2/3 signaling | 3 | 5.3801e-36 | 2.8819e-33 | 6.8673 | 63 | 9,526 | 38.89% | 7 | 306 |
| AP-1 transcription factor network | 6 | 4.6025e-35 | 1.2327e-32 | 4.6483 | 80 | 17,871 | 49.38% | 10 | 623 |
| ALK1 signaling events | 7 | 4.7661e-35 | 1.0942e-32 | 6.6105 | 63 | 9,896 | 38.89% | 7 | 322 |
| ALK1 pathway | 8 | 6.9264e-35 | 1.3913e-32 | 6.5674 | 63 | 9,961 | 38.89% | 7 | 325 |
| Nuclear Events (kinase and transcription factor activation) | 9 | 2.1999e-33 | 3.9281e-31 | 50.1429 | 24 | 497 | 14.81% | 1 | 15 |
| Integrin-linked kinase signaling | 10 | 2.7267e-33 | 4.3818e-31 | 4.3860 | 80 | 18,940 | 49.38% | 10 | 656 |
| CDC42 signaling events | 11 | 1.2242e-29 | 1.7884e-27 | 3.8250 | 81 | 21,989 | 50.00% | 11 | 759 |
| Regulation of CDC42 activity | 12 | 4.7616e-29 | 6.3765e-27 | 3.7505 | 81 | 22,426 | 50.00% | 11 | 772 |
| MAPK targets/ Nuclear events mediated by MAP kinases | 13 | 7.9386e-29 | 9.8134e-27 | 32.3650 | 24 | 770 | 14.81% | 1 | 21 |
| Type I hemidesmosome assembly | 14 | 2.3894e-28 | 2.7427e-26 | 73.8766 | 18 | 253 | 11.11% | 1 | 8 |
| Signaling mediated by p38-alpha and p38-beta | 15 | 8.0574e-27 | 8.6321e-25 | 20.0115 | 27 | 1,401 | 16.67% | 2 | 50 |
| MAP kinase activation in TLR cascade | 16 | 1.7758e-26 | 1.7836e-24 | 25.7449 | 24 | 968 | 14.81% | 1 | 30 |
| BMP receptor signaling | 17 | 8.2935e-25 | 7.8398e-23 | 6.5405 | 46 | 7,303 | 28.40% | 7 | 227 |
| Beta1 integrin cell surface interactions | 18 | 1.3819e-24 | 1.2337e-22 | 2.6962 | 96 | 36,972 | 59.26% | 18 | 1,352 |
| Syndecan-1-mediated signaling events | 19 | 3.0318e-24 | 2.5643e-22 | 2.7272 | 94 | 35,790 | 58.02% | 17 | 1,300 |
| Integrin family cell surface interactions | 20 | 3.1753e-24 | 2.5513e-22 | 2.6675 | 96 | 37,370 | 59.26% | 18 | 1,379 |

# Case Study: Target genes of key HDL cholesterol GWAS SNP




- rs17119878 is significantly associated with HDL cholesterol ($P$=2.5×10$^{-15}$)
- Two nearest genes include BUD13 and ZNF259 are not predicted to be linked to the SNP.
- SNP to predicted to link to four distal genes APOC3, APOA1, APOA4 and APOA5 specifically in fetal intestine and liver which has primary roles in lipid metabolism.
- All of the linked genes are functionally associated with HDL levels

# Case Study: Novel target gene of SNP associated with Ulcerative Colitis




- rs4380874: candidate causal SNP for Ulcerative Colitis (UC) ($P$=1.5×10$^{-15}$, PICS probability=0.76)
- Located in fetal intestine specific active enhancers.
- The enhancers are significantly linked ($p$-value<7.2×10$^{-10}$) to the gene SLC26A3 (35kb away) in fetal intestine tissues by our model
- SLC26A3 is an important membrane protein for intestinal cells and has been suggested to be associated with UC.
- Both SLC26A3 and the linked enhancers show highly specific active chromatin states in fetal intestines and duodenum mucosa

# Summary

- Novel probabilistic model to learn multi-cell type enhancer-gene linking from chromatin and expression dynamics
- High resolution and cell-type specific predictions are validated by experimental data
- Enhancers are very important for explaining cell-type specific regulation
- Majority of genes are regulated by more than 1 enhancer and often different ones in different cell types
- Majority of enhancers do not loop to their nearest genes!
- Linked enhancers and promoters contain motifs of TFs that form PPIs
- Links are highly predictive of target genes and pathways affected by non-coding disease-associated variants

# Acknowledgements

Jianrong Wang

Peyton Greenside

Manolis Kellis

ROADMAP epigenomics PROJECT

ENCODE

ALFRED P. SLOAN FOUNDATION 1934