

M.

L.

$$V = \{1, 2, 1, \dots, t, \dots, N\}$$

word index.

$$V(t) = \text{"dellu"}$$

$$V(N) =$$

$$\sum_{i=1}^N n(V(i)) = L$$

$$w_i = \{m, n\}$$

K

(latent topic)

$$P(w_i = t) = \sum_{k=1}^K P(\gamma = k | d_m) \cdot P(w_i = t | \gamma = k)$$

$\theta_{m,k}$

$\varphi_{k,t}$

mixing

$$\vec{\theta}_{m,\cdot} \propto \text{Dir}(\vec{\alpha}) = \frac{P(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K P(\alpha_i)} \cdot \prod_{i=1}^K \alpha_i^{-1}$$

$$\vec{\varphi}_{k,\cdot} \propto \text{Dir}(\vec{\beta}) = \frac{P(\sum_{i=1}^N \beta_i)}{\prod_{i=1}^N P(\beta_i)} \cdot \prod_{i=1}^N \varphi_{k,i}^{\beta_i-1} = \frac{1}{\Delta(\vec{\beta})} \cdot \prod_{i=1}^N \varphi_{k,i}^{\beta_i-1}$$

$$\Delta = \frac{\prod_{i=1}^N P(\beta_i)}{P(\sum_{i=1}^N \beta_i)} = \Delta(\vec{\beta})$$

$$P(\vec{w} | \vec{\gamma}, \vec{\varphi}) = \prod_{\gamma=1}^K \prod_{t=1}^N \varphi_{\gamma,t}^{n_{\gamma}(t)}$$

$$P(\vec{w} | \vec{\gamma}, \vec{\beta}) = \int P(\vec{w} | \vec{\gamma}, \vec{\varphi}) \cdot P(\vec{\varphi} | \vec{\beta}) d\vec{\varphi}$$

$$= \int \prod_{\gamma=1}^K \prod_{t=1}^N \varphi_{\gamma,t}^{n_{\gamma}(t)} \cdot \left(\prod_{\gamma=1}^K \frac{1}{\Delta(\vec{\beta})} \cdot \prod_{t=1}^N \varphi_{\gamma,t}^{\beta_{\gamma,t}-1} \right) d\vec{\varphi}$$

$$\vec{n}_{\gamma} = \{n_{\gamma}(t)\}_{t=1}^N$$

$$= \prod_{\gamma=1}^K \frac{1}{\Delta(\vec{\beta})} \cdot \int \prod_{t=1}^N \varphi_{\gamma,t}^{n_{\gamma}(t) + \beta_{\gamma,t}-1} d\vec{\varphi} = \prod_{\gamma=1}^K \frac{\Delta(\vec{n}_{\gamma} + \vec{\beta})}{\Delta(\vec{\beta})}$$

$$P(\vec{y}|\vec{\theta}) = \prod_{m=1}^M \prod_{k=1}^K \theta_{m,k}^{n_m(k)}$$

$$P(\vec{y}|\vec{\alpha}) = \int P(\vec{y}|\vec{\theta}) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta}$$

$$= \int \prod_{m=1}^M \prod_{k=1}^K \theta_{m,k}^{n_m(k)} \cdot \left(\prod_{m=1}^M \frac{1}{\Delta(\vec{\alpha})} \cdot \prod_{k=1}^K \theta_{m,k}^{\alpha_k - 1} \right) d\vec{\theta}$$

$$= \prod_{m=1}^M \frac{1}{\Delta(\vec{\alpha})} \int \prod_{k=1}^K \theta_{m,k}^{n_m(k) + \alpha_k - 1} d\vec{\theta}$$

$$= \prod_{m=1}^M \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}$$

$$\vec{n}_m = \{ n_m(k) \}_{k=1}^K$$

$$P(\vec{w}, \vec{y}|\vec{\alpha}, \vec{\beta}) = P(\vec{w}|\vec{y}, \vec{\beta}) \cdot P(\vec{y}|\vec{\alpha})$$

$$= \prod_{j=1}^K \frac{\Delta(\vec{n}_j + \vec{\beta})}{\Delta(\vec{\beta})} \cdot \prod_{m=1}^M \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}$$

$$\vec{w} = \{ w_i = t, \vec{w}_{-i} \}$$

$$\vec{y} = \{ y_i = k, \vec{y}_{-i} \}$$

$$P(y_i = k | \vec{y}_{-i}, \vec{w}, \vec{\alpha}, \vec{\beta}) = \frac{P(\vec{w}|\vec{y}, \vec{\beta}) \cdot P(\vec{y}|\vec{\alpha})}{P(w_i) \cdot P(\vec{w}_{-i}|\vec{y}_{-i}, \vec{\beta}) \cdot P(\vec{y}_{-i}|\vec{\alpha})}$$

$$\propto \frac{\prod_{j=1}^K \frac{\Delta(\vec{n}_j + \vec{\beta})}{\Delta(\vec{\beta})}}{\prod_{j=1}^K \frac{\Delta(\vec{n}_{j,-i} + \vec{\beta})}{\Delta(\vec{\beta})}}$$

$$\cdot \frac{\prod_{m=1}^M \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}}{\prod_{m=1}^M \frac{\Delta(\vec{n}_{m,-i} + \vec{\alpha})}{\Delta(\vec{\alpha})}}$$

$$= \frac{\Delta(\vec{n}_k + \vec{\beta})}{\Delta(\vec{n}_{k,-i} + \vec{\beta})} \cdot \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{n}_{m,-i} + \vec{\alpha})} = \frac{n_{k,-i}(t) + \beta_t}{\sum_{t=1}^T (n_{k,-i}(t) + \beta_t)} \cdot \frac{n_{m,-i}(k) + \alpha_k}{\sum_{k=1}^K (n_{m,-i}(k) + \alpha_k)}$$

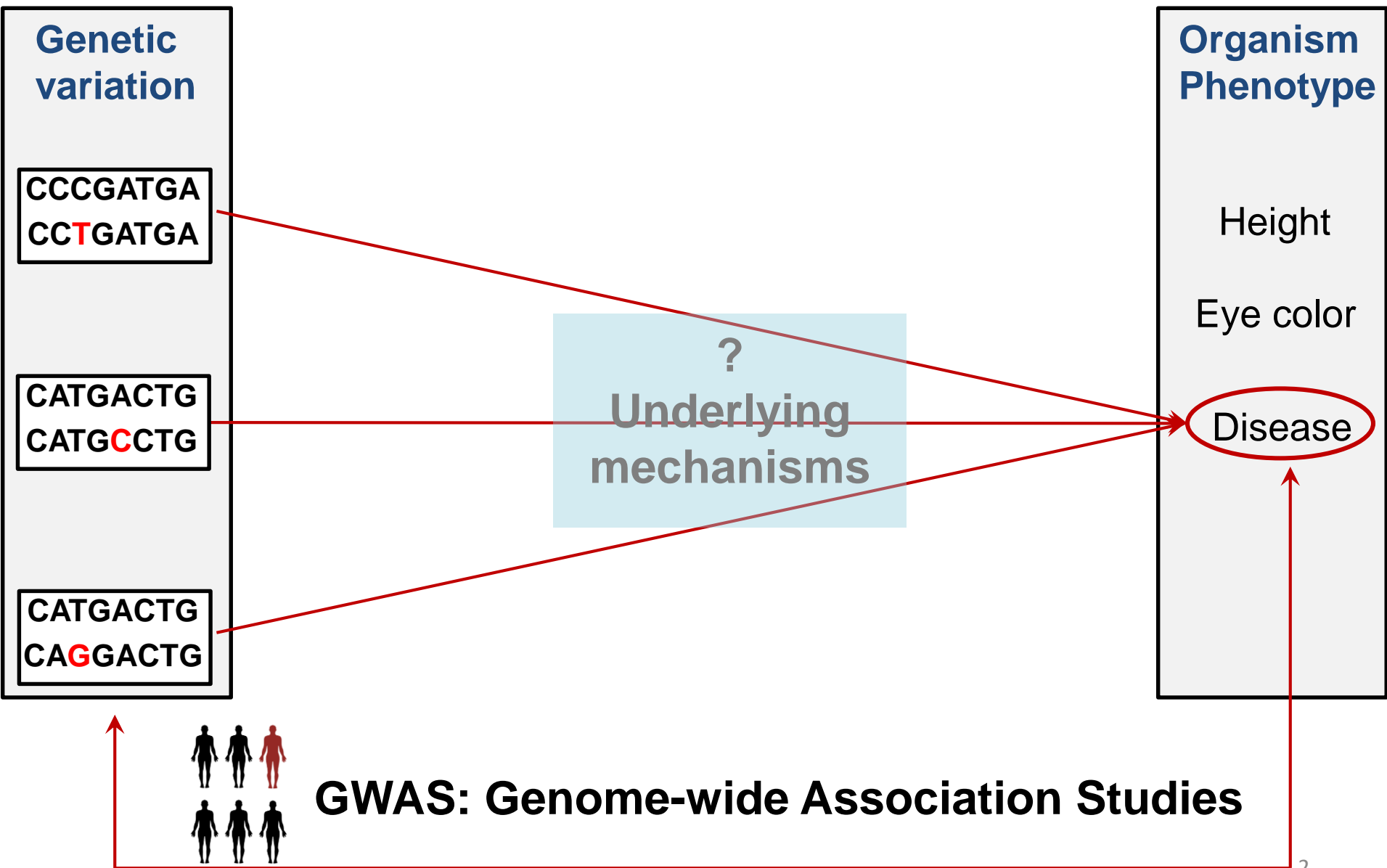


Introduction of Machine Learning in Computational Biology

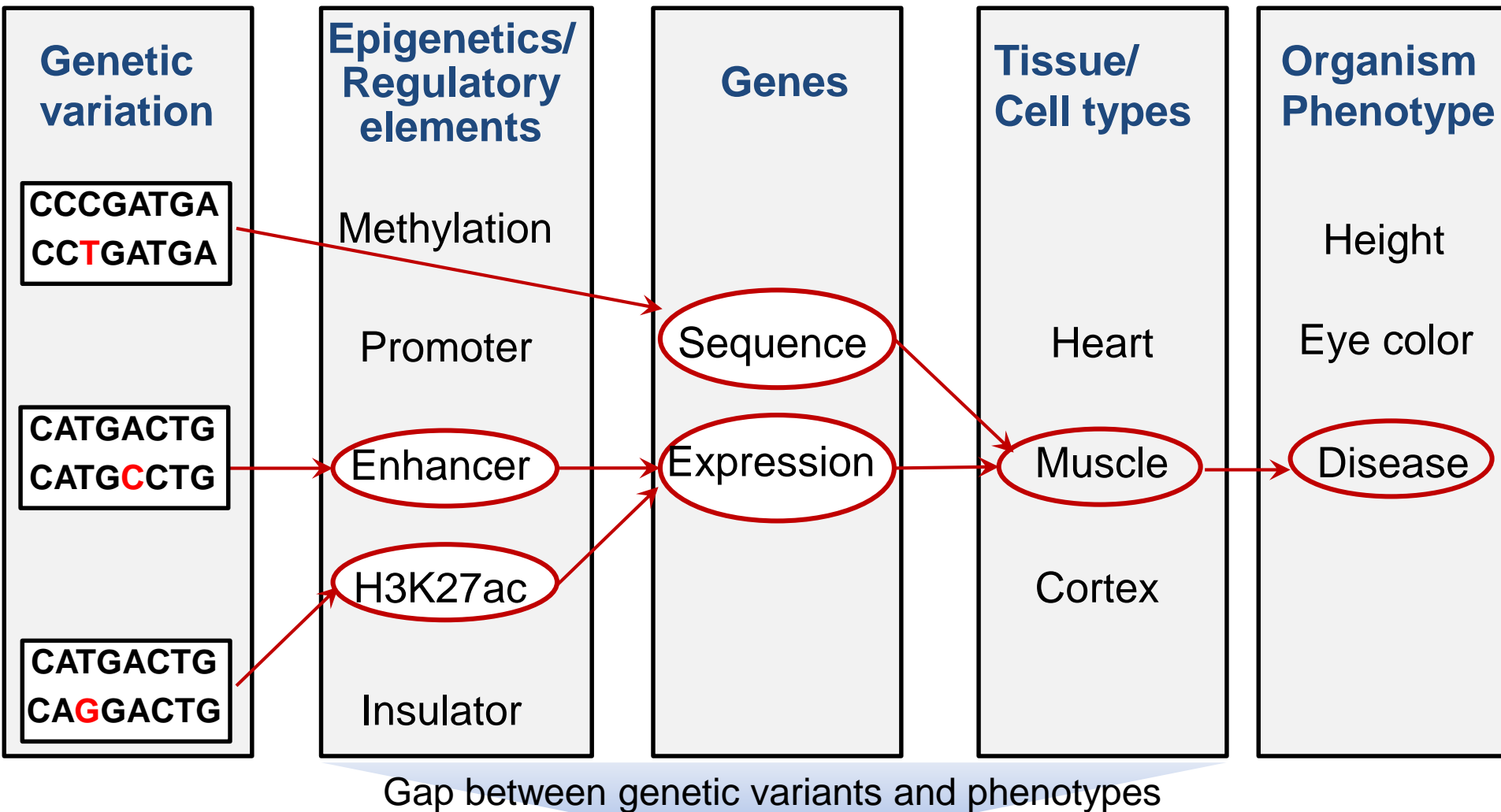
Jianrong Wang

CMSE @MSU

Genetic variants associated with human disease



Understanding disease pathogenesis

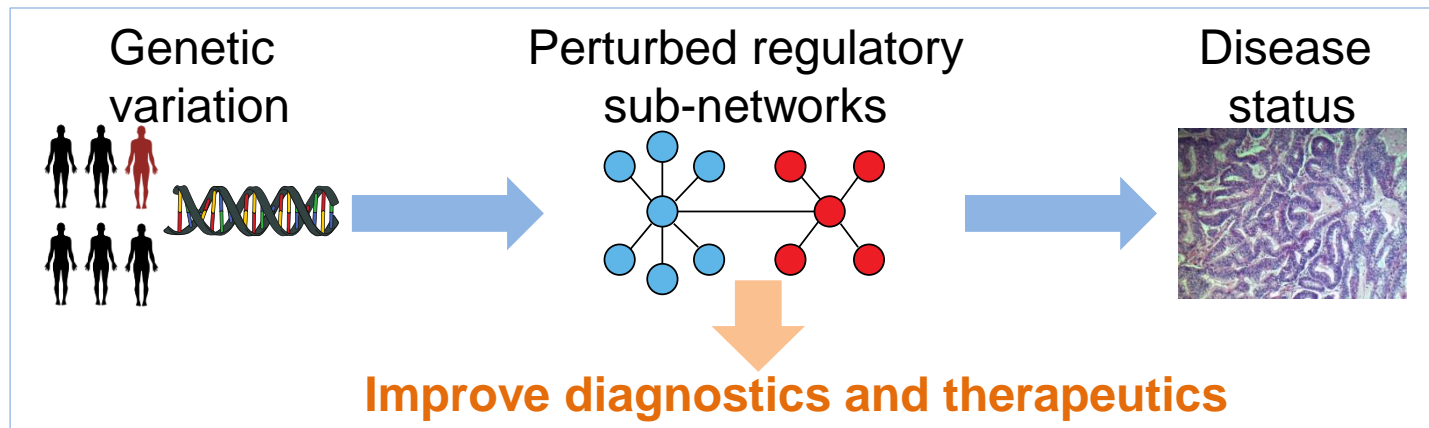


Disrupted molecular phenotypes (regulatory activity, gene expression, and pathways) in relevant cellular contexts

Functional genomics approach to understand disease

Functional genomics:

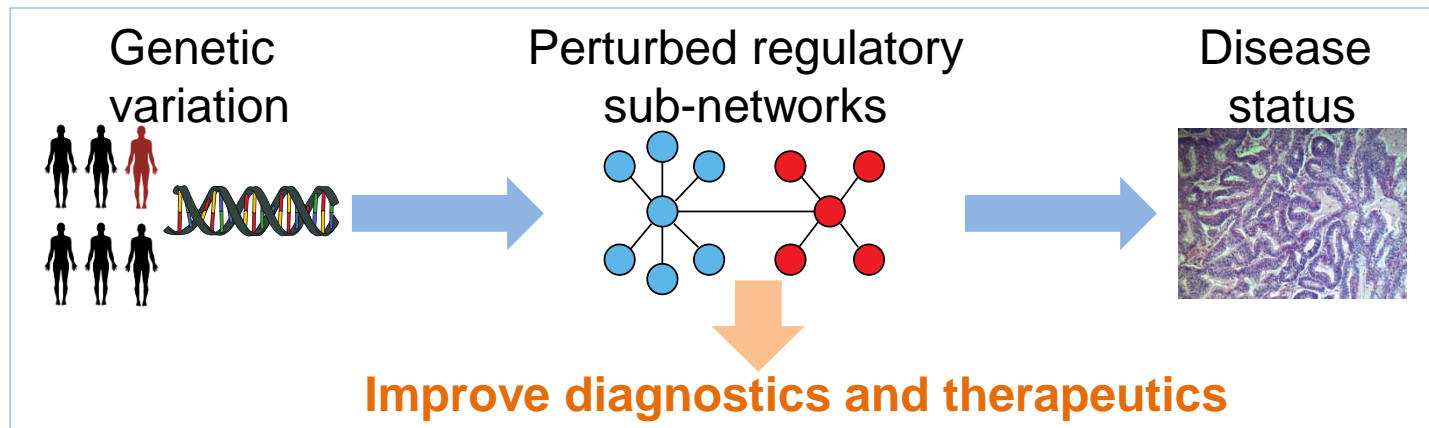
- Elucidate regulatory elements and networks of gene expression in specific cellular contexts.
- Generate reference maps to track how the effects of genetic variants propagate to organism phenotypes through dysregulation of gene expression.



Functional genomics approach to understand disease

Functional genomics:

- Elucidate regulatory elements and networks of gene expression in specific cellular contexts.
- Generate reference maps to track how the effects of genetic variants propagate to organism phenotypes through dysregulation of gene expression.



High-throughput sequencing datasets: unique opportunity for functional genomics and systems biology to address this question.

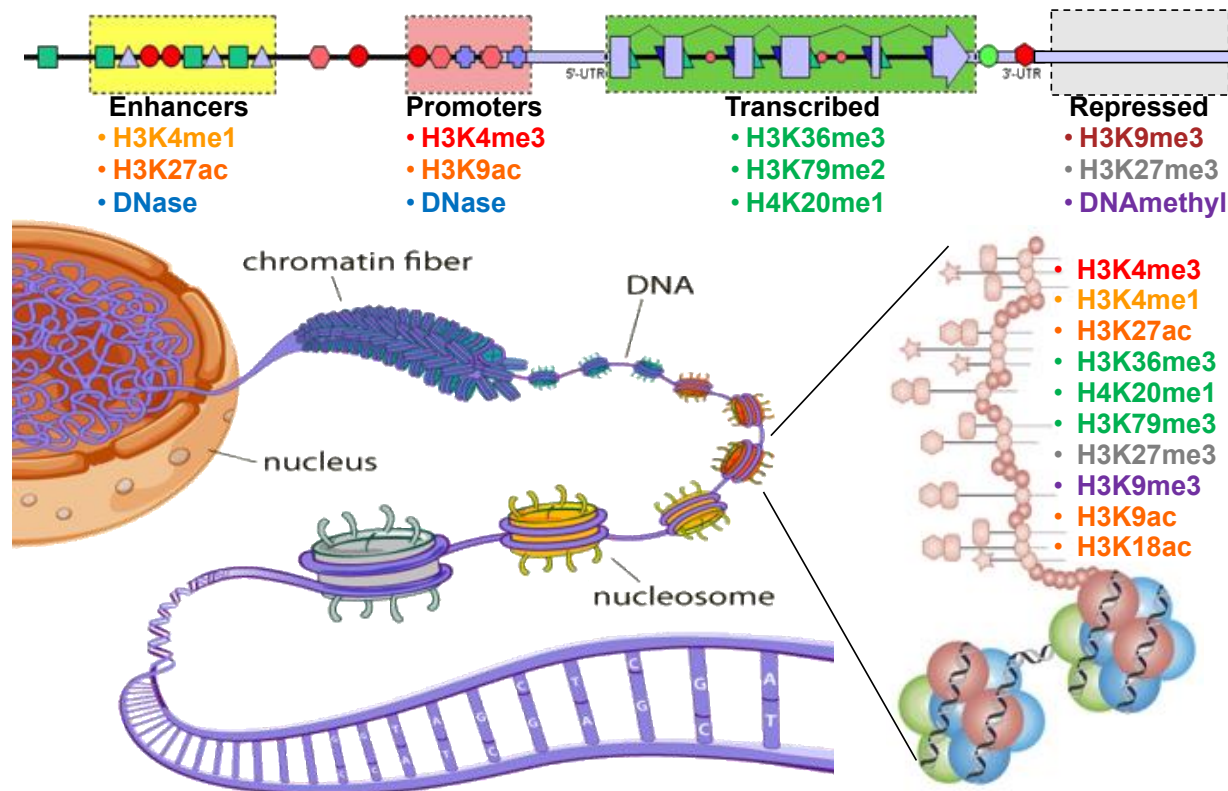
Multi-view of specific cellular contexts ('omic' datasets):

- *Transcriptome for gene expression:* RNA-seq, microarrays;
- *Regulome for transcription factor binding:* ChIP-seq;
- *Epigenome for epigenetic features:* ChIP-seq, bisulfite sequencing;

Epigenetics indicates functional regulatory elements

Epigenetics: histone modifications, chromatin architecture, DNA methylation.

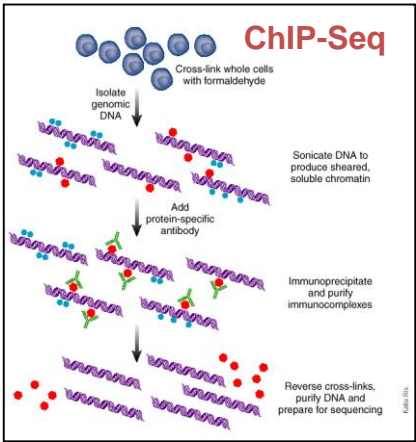
Combinatorial epigenetic signatures indicate different classes of regulatory elements.



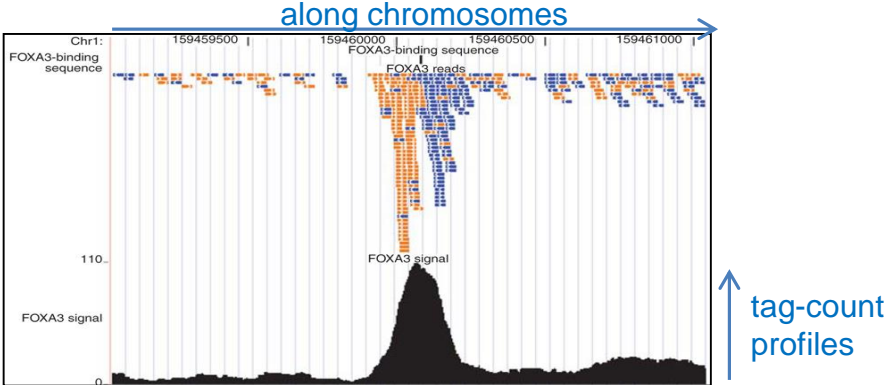
Epigenetic signatures are highly dynamic and represent **cell-type specific regulatory activities**.

Big data: large panels of epigenomic datasets

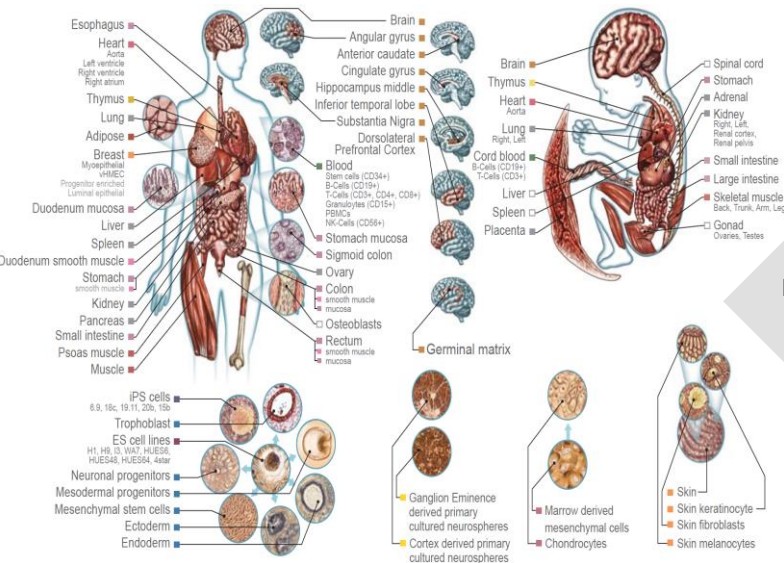
Epigenomics: Genome-wide ChIP-seq histone modification maps enable systematic characterization of cell-type specific combinatorial signatures.



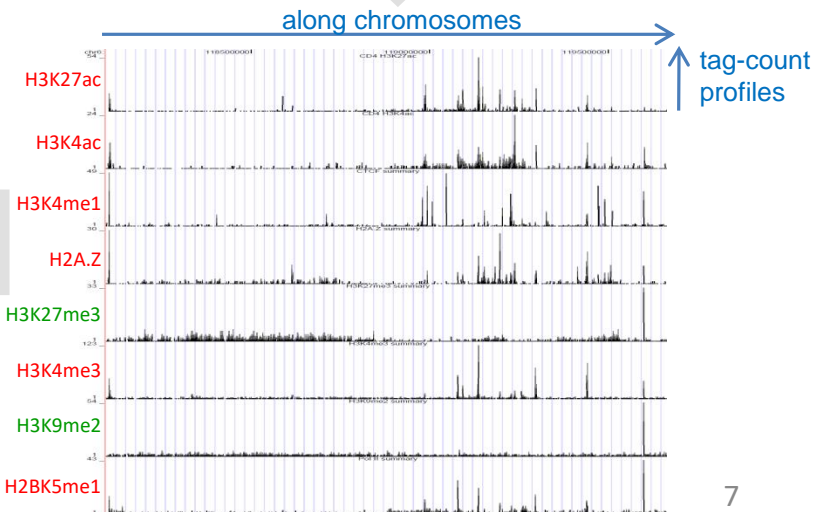
1 feature
1 cell-type



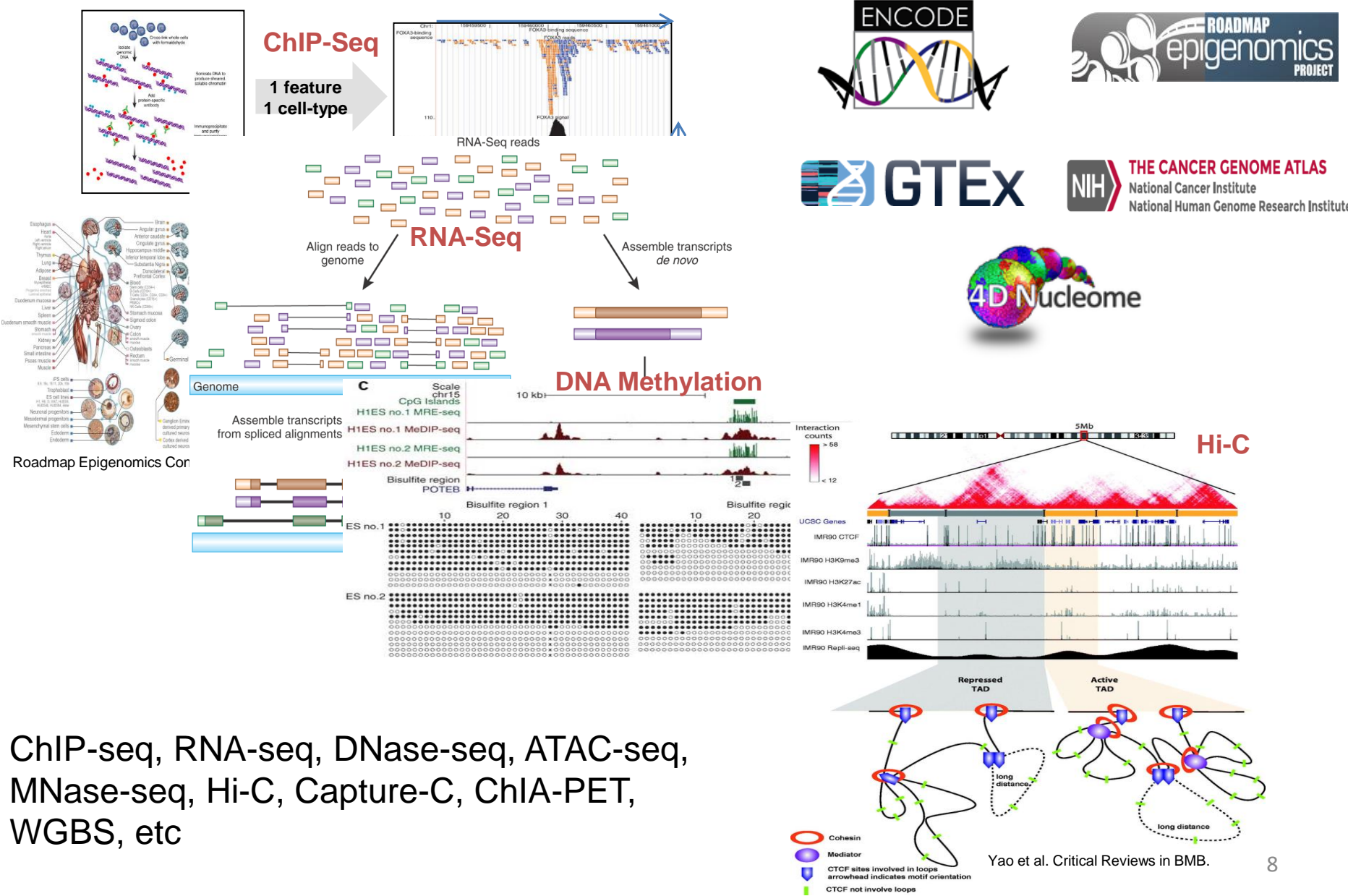
multiple features
1 cell-type



multiple features
100+ cell-types



Big data: large panels of epigenomic datasets



Challenges and Strategy



Challenges:

Biological barriers:

- multiple-layers of regulatory elements/interactions
- dynamic regulation in specific cellular contexts;
- heterogeneity across individuals.

Computational barriers:

- large-scale highly noisy datasets;
- diverse data types with pervasive correlations;
- combinatorial complexity of data structures;



Efficient and robust machine learning algorithms to integrate big datasets and yield real biological discoveries.

Challenges and Strategy


Challenges:

Biological barriers:

- multiple-layers of regulatory elements/interactions
- dynamic regulation in specific cellular contexts;
- heterogeneity across individuals.

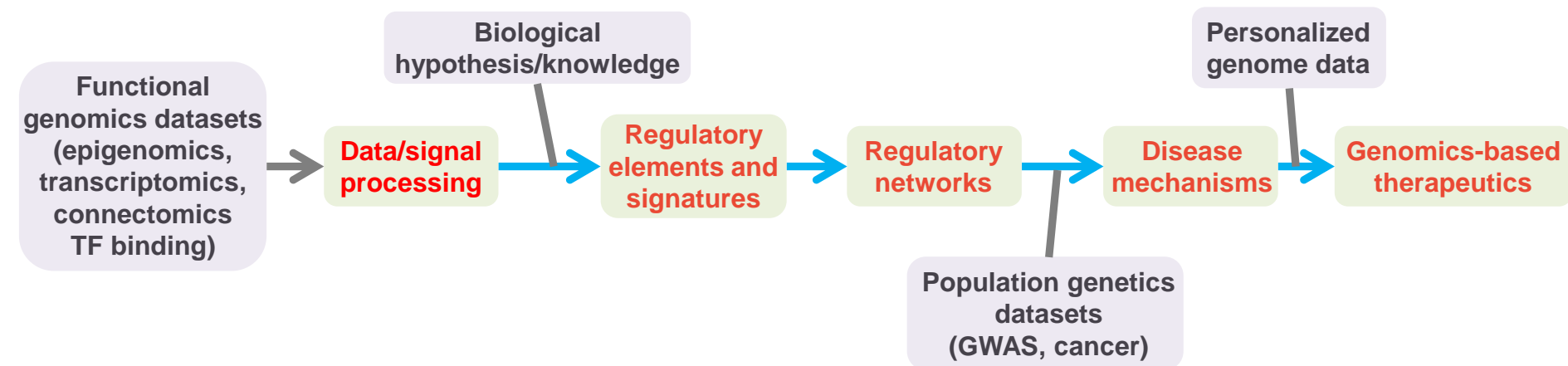
Computational barriers:

- large-scale highly noisy datasets;
- diverse data types with pervasive correlations;
- combinatorial complexity of data structures;



Efficient and robust machine learning algorithms to integrate big datasets and yield real biological discoveries.

Strategy:



Major Methodology



Probabilistic graphical models (HMM, Bayesian Net, Latent Dirichlet Allocation etc):
Gene finding, Combinatorial histone code, Regulatory 'grammar';

Graph theory and algorithms: Alternative-junction prediction, Network analysis, Network
Community detection, Evolving networks, Genome assembly;

Matrix factorization: Cell type deconvolution, Population structure;

Variational methods and MCMC: Motif finding, Network inference;

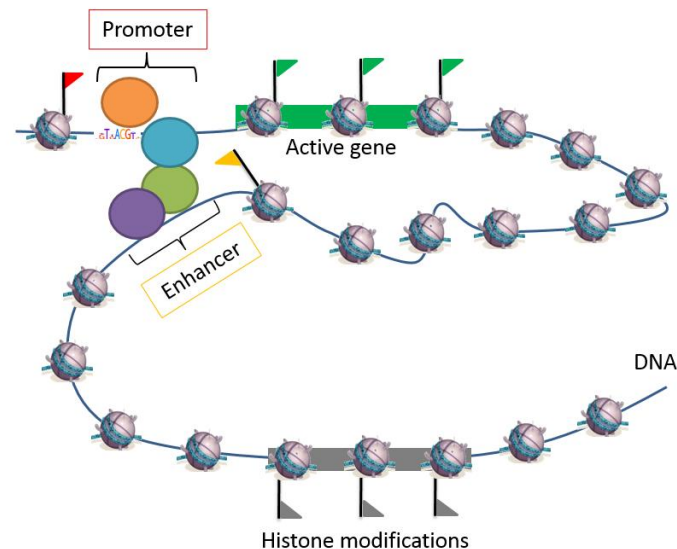
Regularization: GWAS;

Boosting, Clustering, SVM, Deep learning etc;

Three-dimensional enhancer regulation

Enhancers: an important family of regulatory elements activating gene expression.

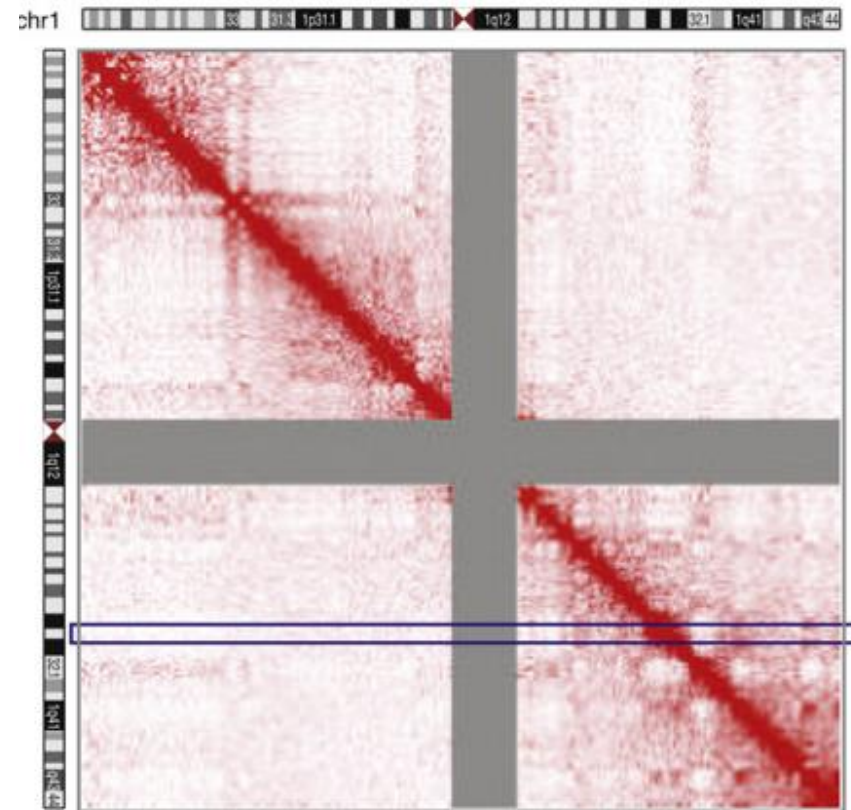
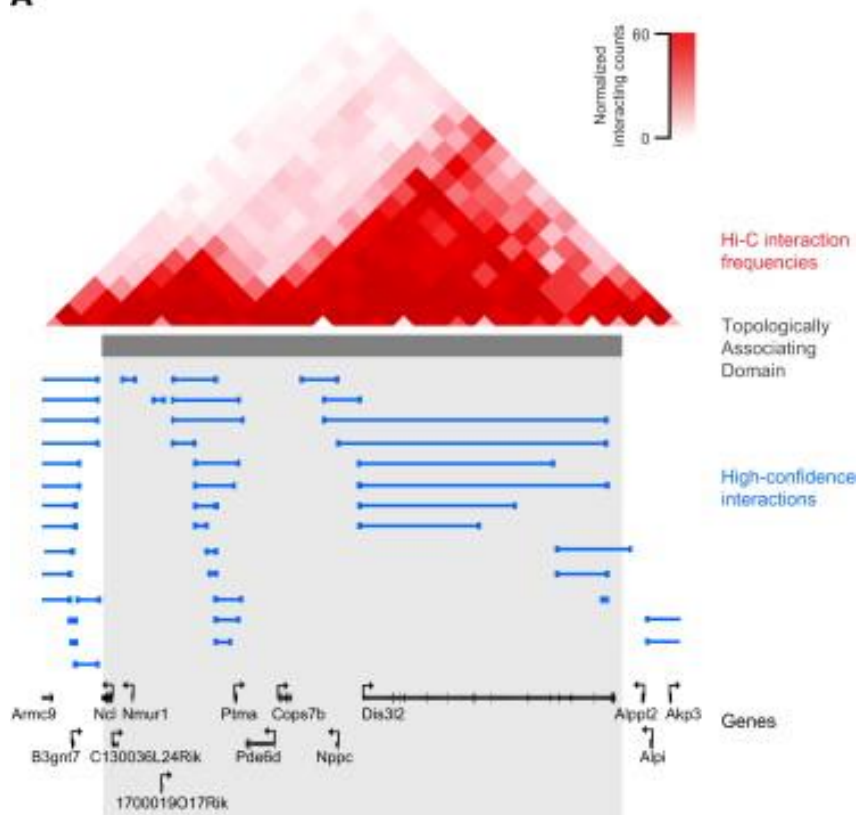
- Abundant across the human genome, especially in non-coding regions;
- Highly dynamic regulatory activity, only active in specific cell-type(s);
- Located distal from the genes that are regulated by them (target genes);
- Nearest genes are not necessarily target genes, and are conditional on cell-types;
- Long-range 3D chromatin interactions mediate enhancer regulation to target genes.



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

Experimental approaches of chromatin interactions

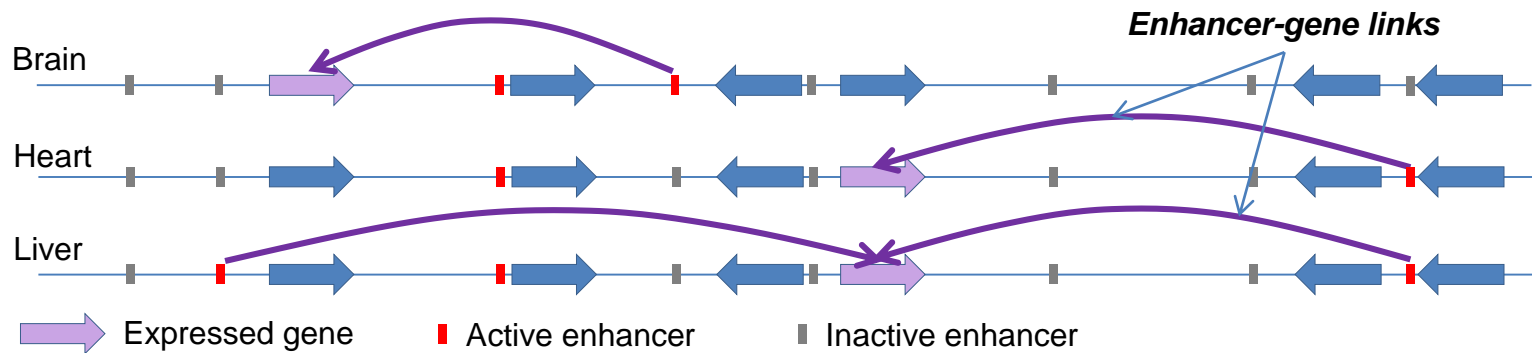
A



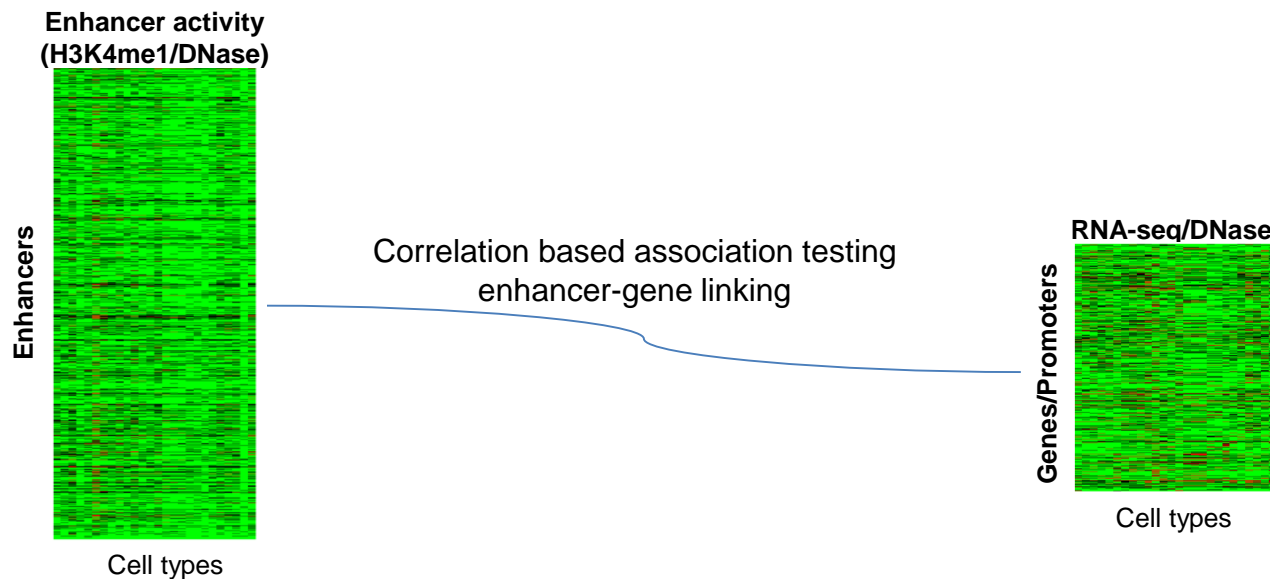
- Hi-C: Coarse-grained low-resolution interaction maps (~5-10Kb fragments)
- ChIA-PET: interactions involving specific proteins, low sensitivity
- Noisy, low signal-to-noise ratios
- High cost(requires billions of reads)
- Only available for a few cell types

Activity-based computational models

Motivation: Target gene expression is expected to be associated with enhancer activity across different cell-types/tissues.



Metric: Marginal association testing using linear or rank correlation between enhancer chromatin activity and gene expression.



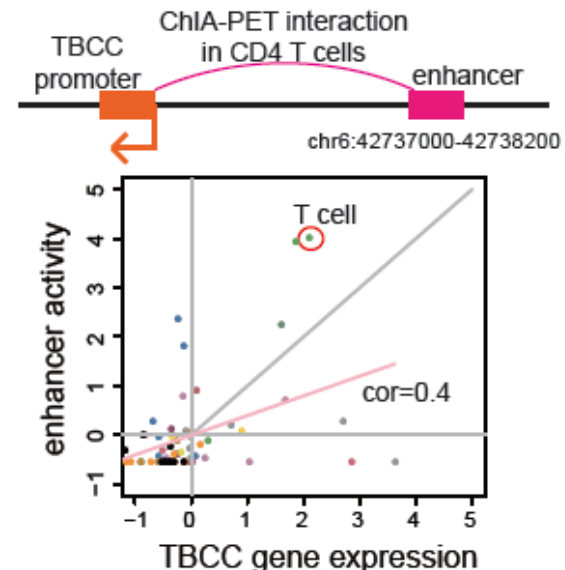
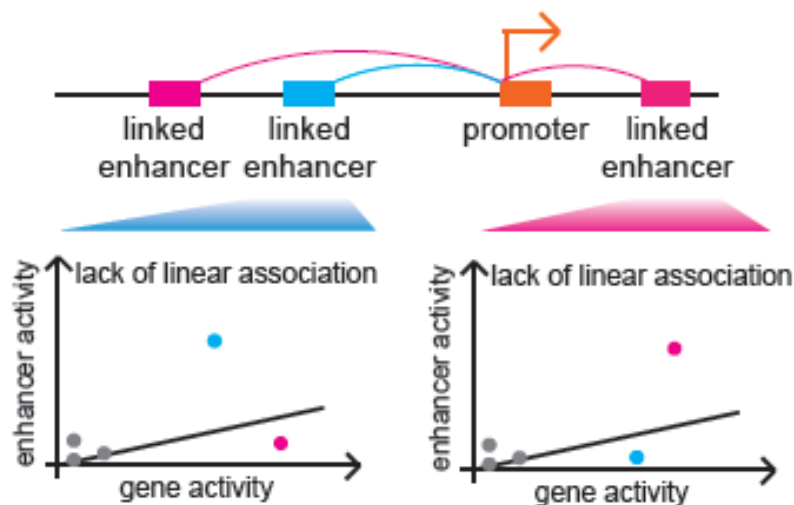
Previous methods and challenges

Supervised learning (classification/regression methods): dependent on experimentally obtained interactions as training samples.

- Training data is highly sparse, overfitting;
- Poor generalization to different cell-types/tissues.

Unsupervised inference (marginal association testing):

- Significantly under-powered due to huge multiple testing burden;
- Sparse enhancer activity (non-Gaussian): not appropriate for correlations;
- Enhancer regulation is highly cell-type specific/restricted: lack of global correlations;
- Need to quantify and assign cell-type specificity to enhancer-gene links;
- Multiple enhancers on multiple genes: Non-linear regulatory effects.

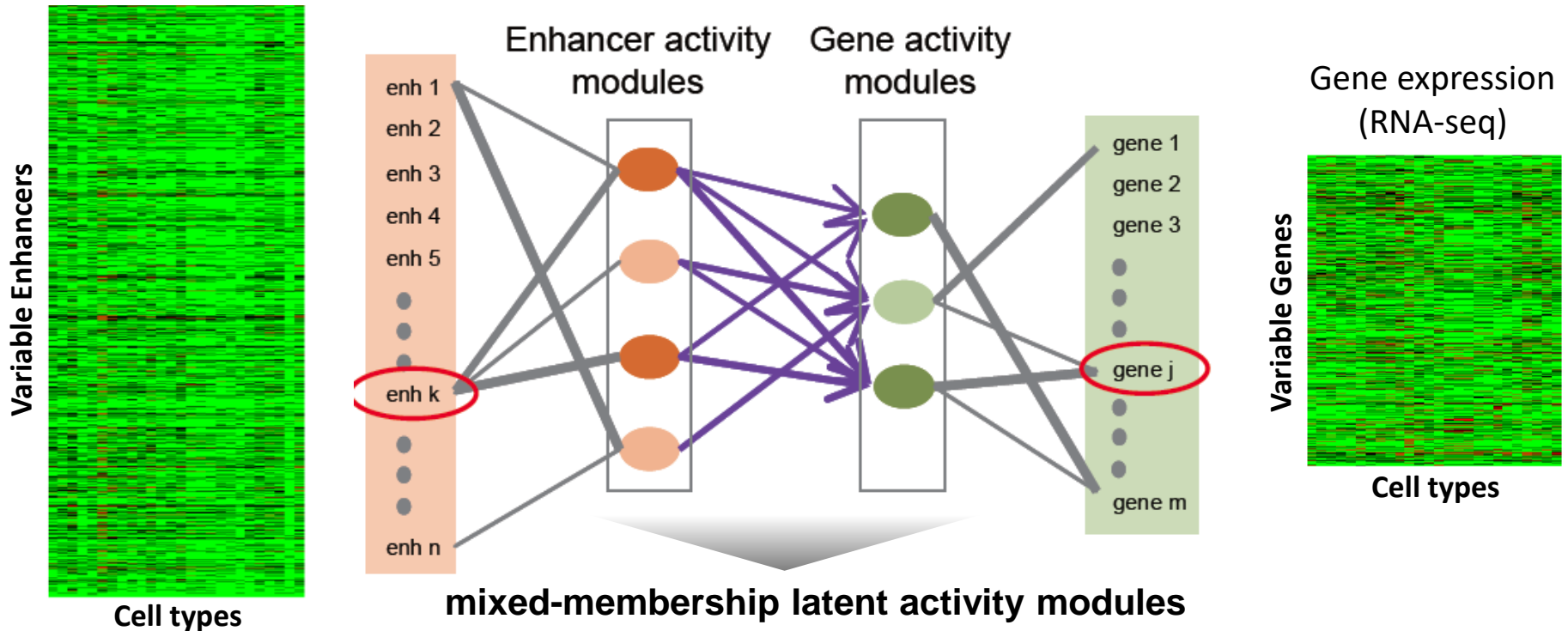


A novel prob. model for enhancer-gene linking



Enhancer activity
(H3K4me1/H3K27me3)

Activity-module based probabilistic model
(a *top-down* approach)



Why modules?

- Increased statistical power: much less number of hypotheses testing;
- Improved robustness: less noisy activity representation than individual enhancers/genes;
- Cell-type specificity: modules are defined by their cell-type specificity parameters.

What kind of modules?

- Mixed-membership prob. modules: capture complex enhancer/gene dynamics across different cell-types.

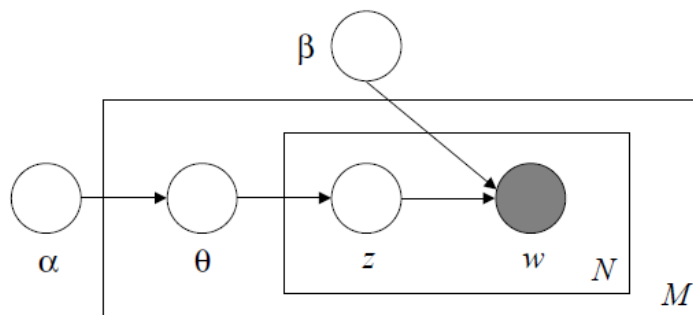
How to link modules?

- Specific non-linear association statistics: reasonable No. of modules make the calculations tractable.

Mixed-membership modules – Latent Dirichlet Allocation

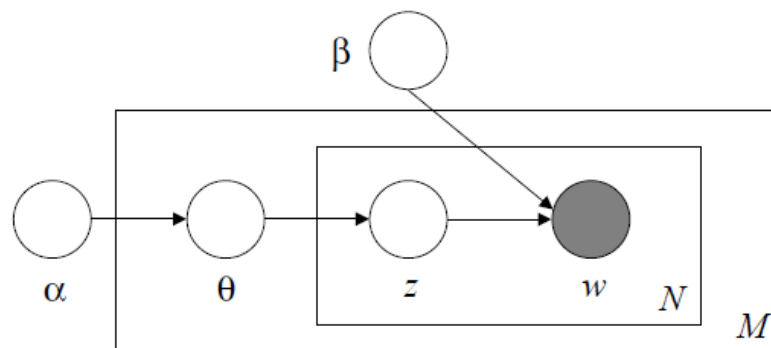
1. ‘Topic’ modeling: Given a collection of books, infer the topics of each book based on observed word counts.
2. Each book is a mixture of topics. And each topic is a mixture of key words.
3. Generative model.

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n :
 - (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
 - (b) Choose a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic z_n .



Mixed-membership modules – Latent Dirichlet Allocation

Hierarchical graphical model:



$$p(\theta | \alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1}$$

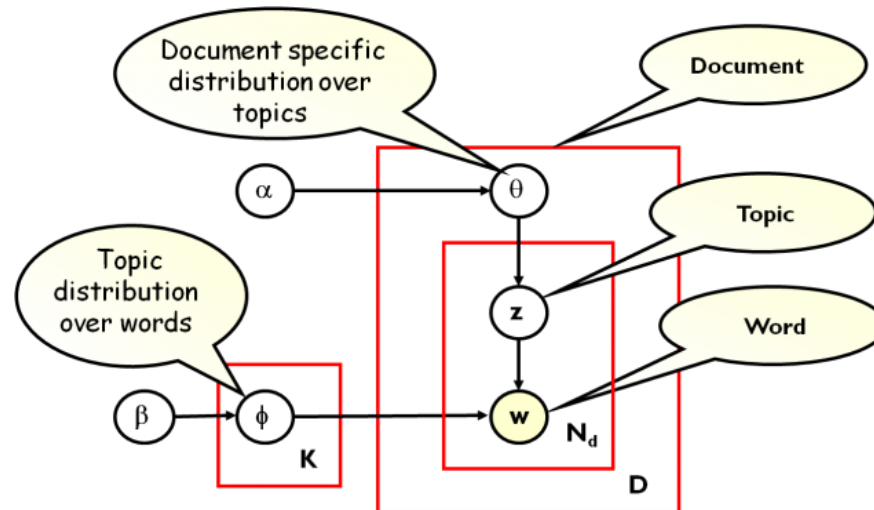
$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta)$$

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left(\prod_{n=1}^N \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta.$$

$$p(D | \alpha, \beta) = \prod_{d=1}^M \int p(\theta_d | \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d\theta_d$$

Mixed-membership modules – Latent Dirichlet Allocation

Non-negative matrix factorization: more ‘part-like’ deconvolution than PCA.



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

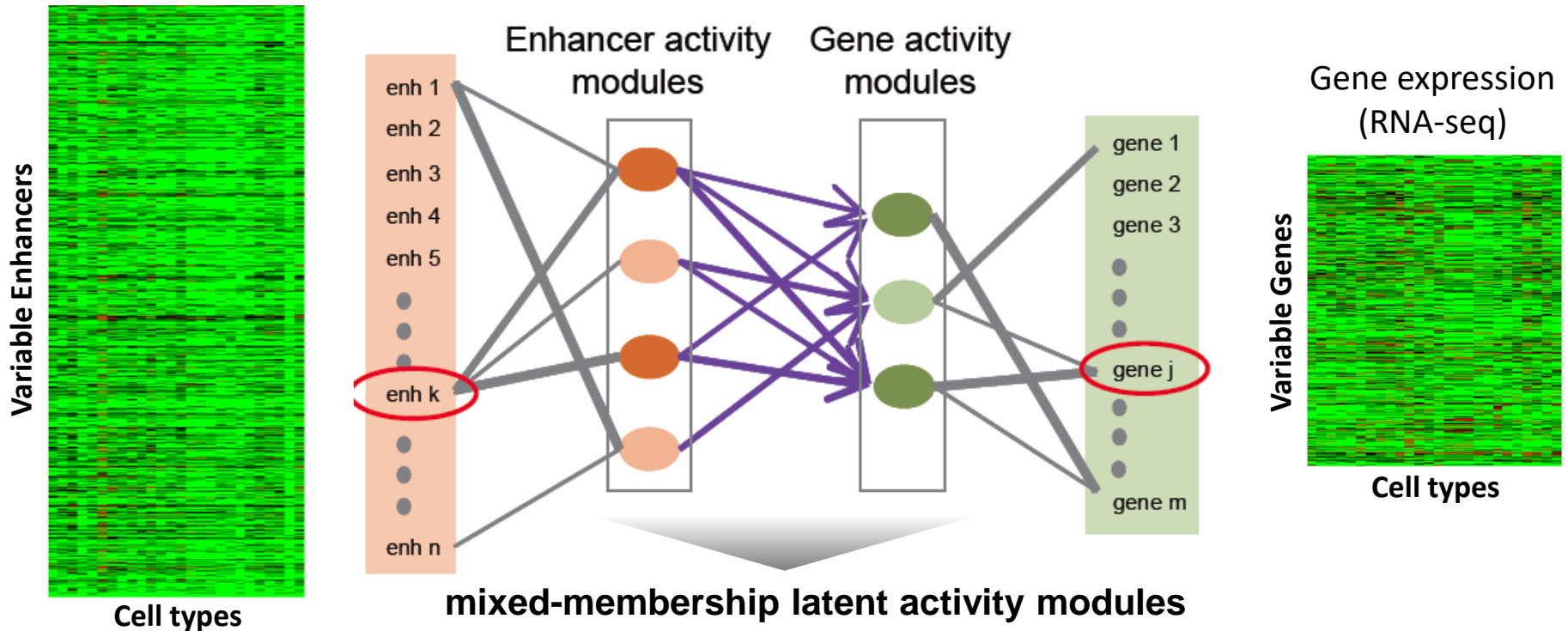
Word occurrence Topic Book

A novel prob. model for enhancer-gene linking



Enhancer activity
(H3K4me1/H3K27me3)

Activity-module based probabilistic model
(a *top-down* approach)



Why modules?

- Increased statistical power: much less number of hypotheses testing;
- Improved robustness: less noisy activity representation than individual enhancers/genes;
- Cell-type specificity: modules are defined by their cell-type specificity parameters.

What kind of modules?

- Mixed-membership prob. modules: capture complex enhancer/gene dynamics across different cell-types.

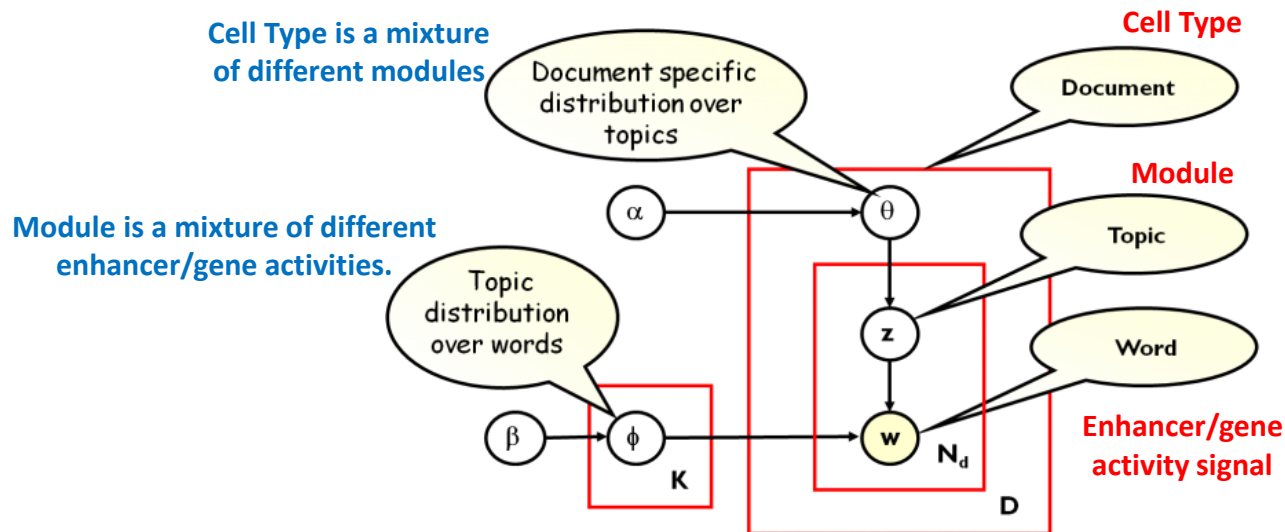
How to link modules?

- Specific non-linear association statistics: reasonable No. of modules make the calculations tractable.

A novel prob. model for enhancer-gene linking

Latent Dirichlet Allocation (LDA) “topic” model allows:

- each gene/enhancer to belong to multiple latent modules;
- learn the cell-type specific module structures of genes/enhancers.



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

Activity signal
Module
Cell Type

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

Membership probabilities of modules to specific cell-types (before normalization).

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

Membership probabilities of enhancers/genes to specific latent modules (before normalization).

Parameter inference – Gibbs sampling



1. Infer model parameters based on observed data: enhancer activity matrix;
2. Gibbs sampling: general idea;

Algorithm 1 Gibbs sampler

Initialize $x^{(0)} \sim q(x)$

for iteration $i = 1, 2, \dots$ do

$$x_1^{(i)} \sim p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, \dots, X_D = x_D^{(i-1)})$$

$$x_2^{(i)} \sim p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, \dots, X_D = x_D^{(i-1)})$$

\vdots

$$x_D^{(i)} \sim p(X_D = x_D | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, \dots, X_{D-1} = x_{D-1}^{(i)})$$

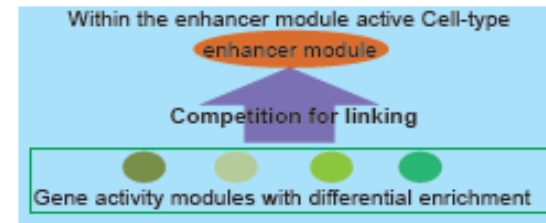
end for

from Ilker Yildirim

3. Gibbs sampling for Latent Dirichlet Allocation models.

Non-linear linking of enhancer modules to gene modules

- Two latent module structures (enhancers and genes) need to be connected.
- 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.



Module-to-tissue probability matrices

	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	$\vartheta_{1,2}$	$\vartheta_{2,2}$	$\vartheta_{3,2}$	$\vartheta_{4,2}$	$\vartheta_{5,2}$
Enhancer module 3	$\vartheta_{1,3}$	$\vartheta_{2,3}$	$\vartheta_{3,3}$	$\vartheta_{4,3}$	$\vartheta_{5,3}$

	t ₁	t ₂	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	$\vartheta_{1,2}$	$\vartheta_{2,2}$	$\vartheta_{3,2}$	$\vartheta_{4,2}$	$\vartheta_{5,2}$

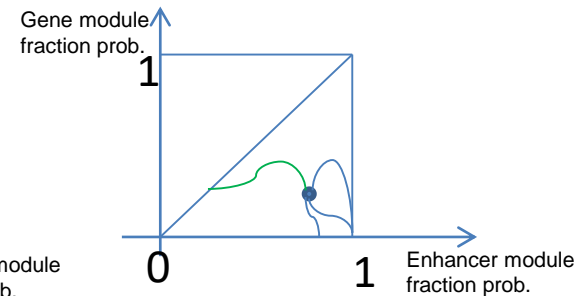
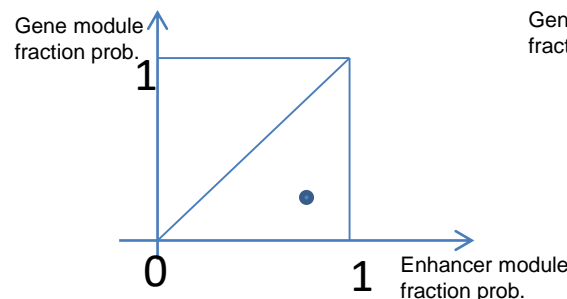
$$A = (a_{ij})_{K_1 \times K_2}$$

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

$$a_{ij} = p(z^1_i \sim z^2_j | \bar{\vartheta}^1, \bar{\theta}^2) = \frac{P(\bar{\vartheta}^1, \bar{\theta}^2 | z^1_i \sim z^2_j)}{\sum_{l=1}^{K_2} P(\bar{\vartheta}^1, \bar{\theta}^2 | z^1_i \sim z^2_l) + P(\bar{\vartheta}^1, \bar{\theta}^2 | z^1_i \sim NA)}$$

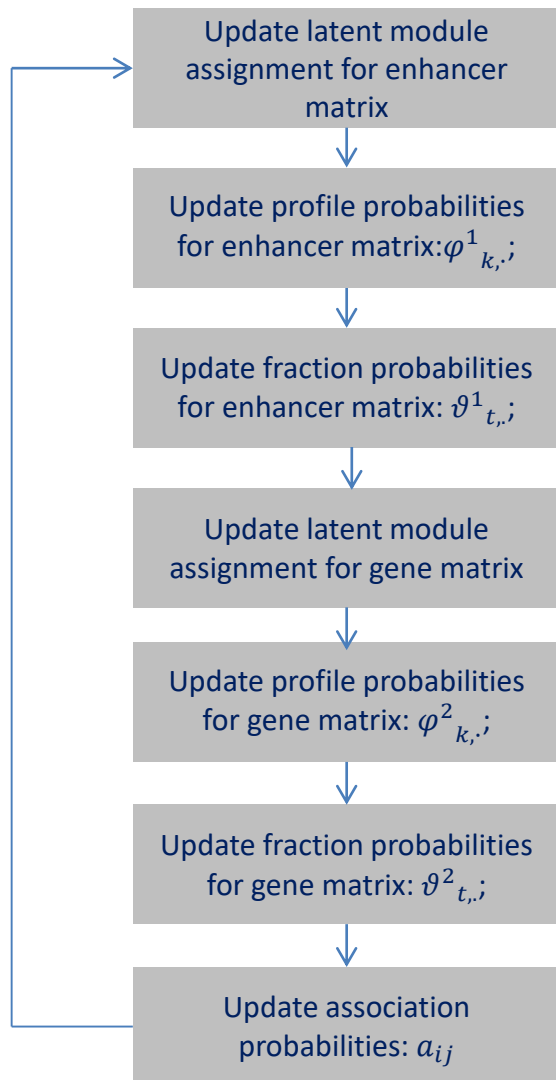
We use a diffusion model to estimate

$$P(\bar{\vartheta}^1, \bar{\theta}^2 | z^1_i \sim z^2_j)$$



Gibbs sampling approach to jointly infer all parameters

Joint Gibbs sampling inference on
2 connected LDA models



Key parameters inferred for the model:

1. Enhancer-to-module and gene-to-module probabilities ($\varphi_{i,\cdot}$);
2. Module-to-tissue probabilities ($\vartheta_{t,\cdot}$);
3. Module-association probabilities ($a_{i,j}$).

Sparsity-inducing regularization is used to deal with the smaller number of cell types compared to the number of modules.

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

A specific pair of individual
enhancer and gene

Statistically significant links

- Focus on links that are within 1Mb of each other
- Null distribution: Linking probabilities on shuffled datasets
- Generate P -values for enhancer-gene links on the real data
- The Benjamini-Hochberg method is used for multiple hypothesis correction.
- Thresholds: FDR 1% (stringent) and 5% (relaxed)

Epigenomes and transcriptomes of 56 cellular contexts

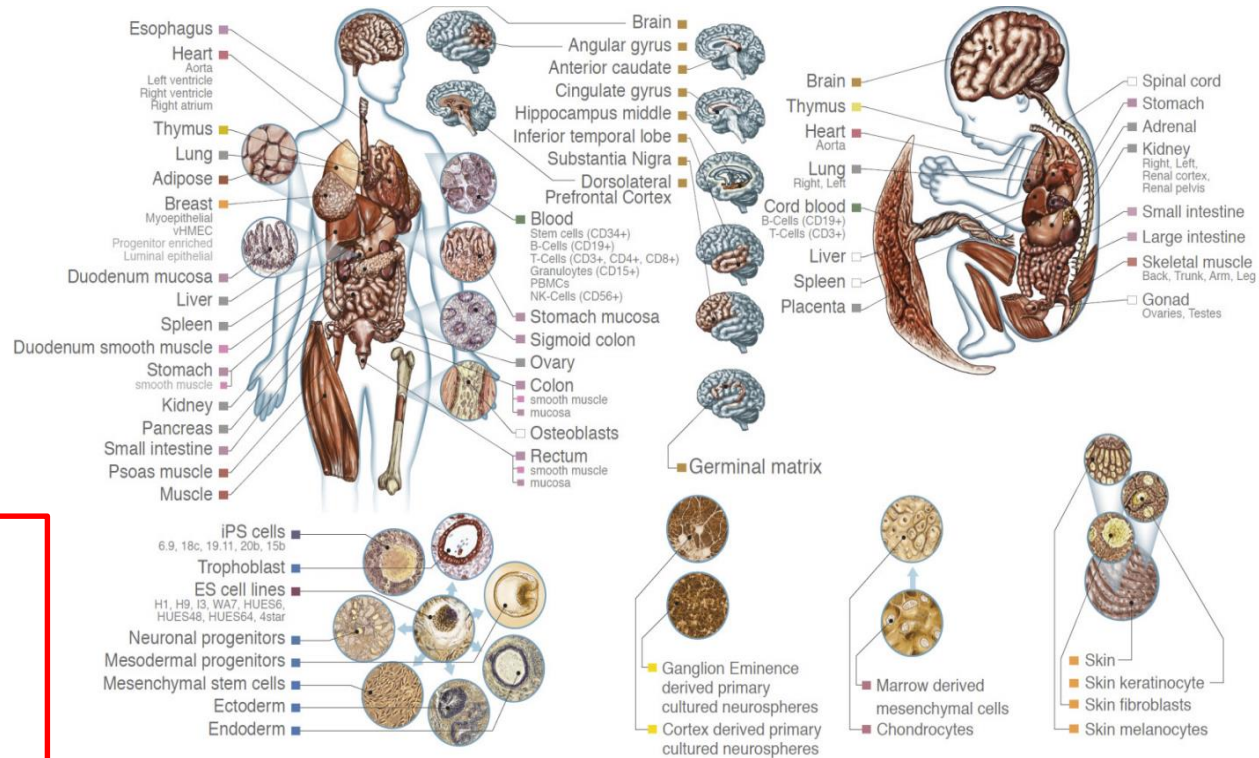
The model is applied on 56 human cellular contexts with both epigenome and transcriptome data available from NIH Roadmap epigenome and ENCODE project.

- 697,876 enhancer elements;
- 19,003 genes

Enhancer activity signals:
H3K4me1/H3K27me3.

Gene expression:
RPKM of RNA-seq data.

New version is coming with
cell-type specific enhancer-
gene networks for 200+
human cell-types/tissues.

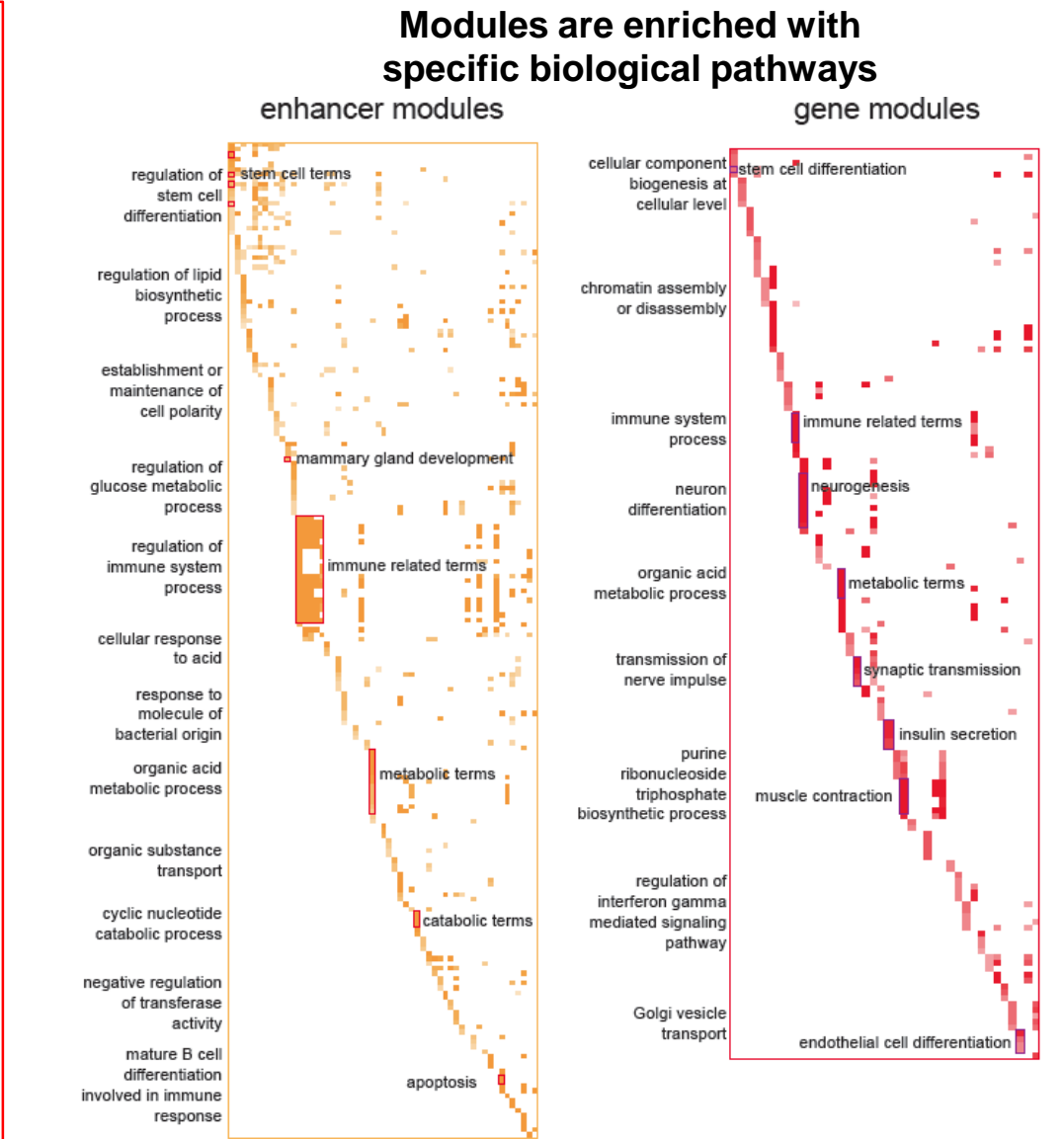
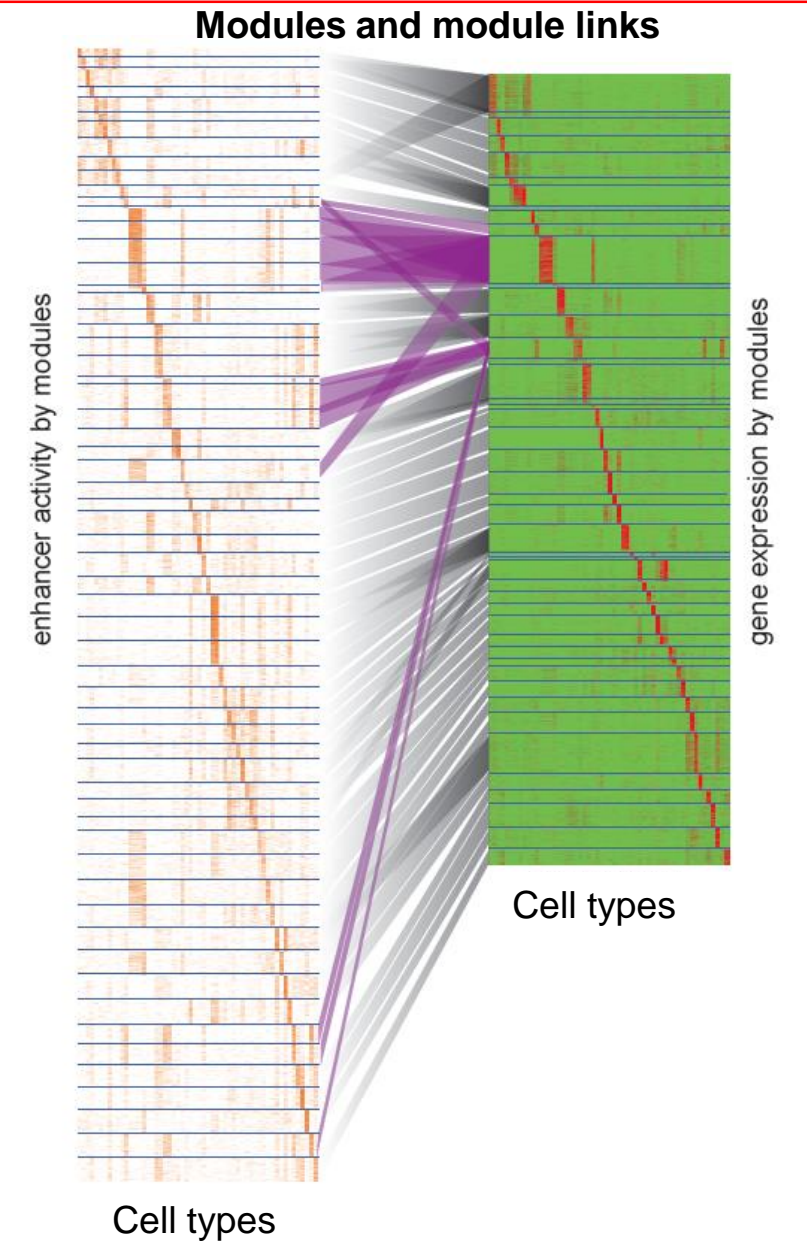


- 6+ key histone marks (Histone ChIP-seq)
- Open chromatin (DNase-seq)
- DNA methylation
- Gene expression (RNA-seq)



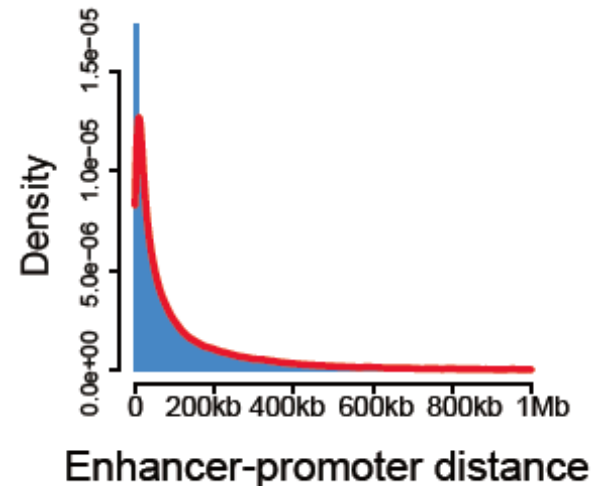
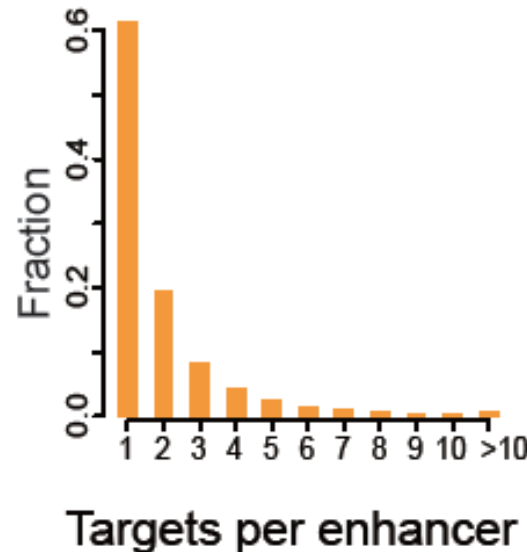
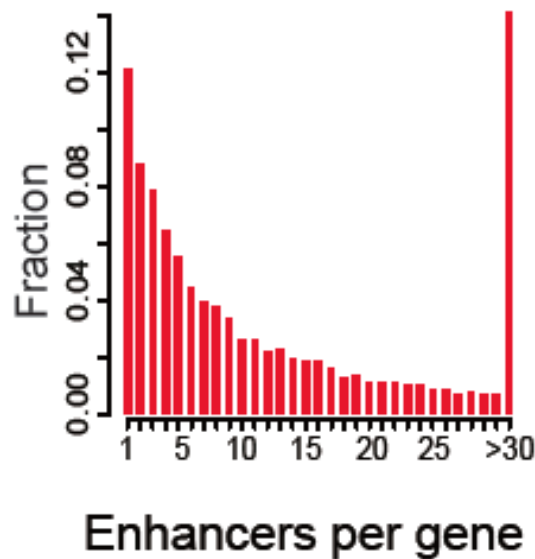
<http://roadmapepigenomics.org>

Learned enhancer/gene modules and associations



Basic properties of long-range enhancer-gene networks

- At the threshold of $\text{FDR}=0.01$, totally ~250k significant enhancer-gene links are predicted across the panel of 56 cellular contexts.
- Each enhancer-gene link is assigned with the cell-type specificity information.



The enhancer-gene network is highly connected

- 88% of genes and 39% of enhancers are multiply linked

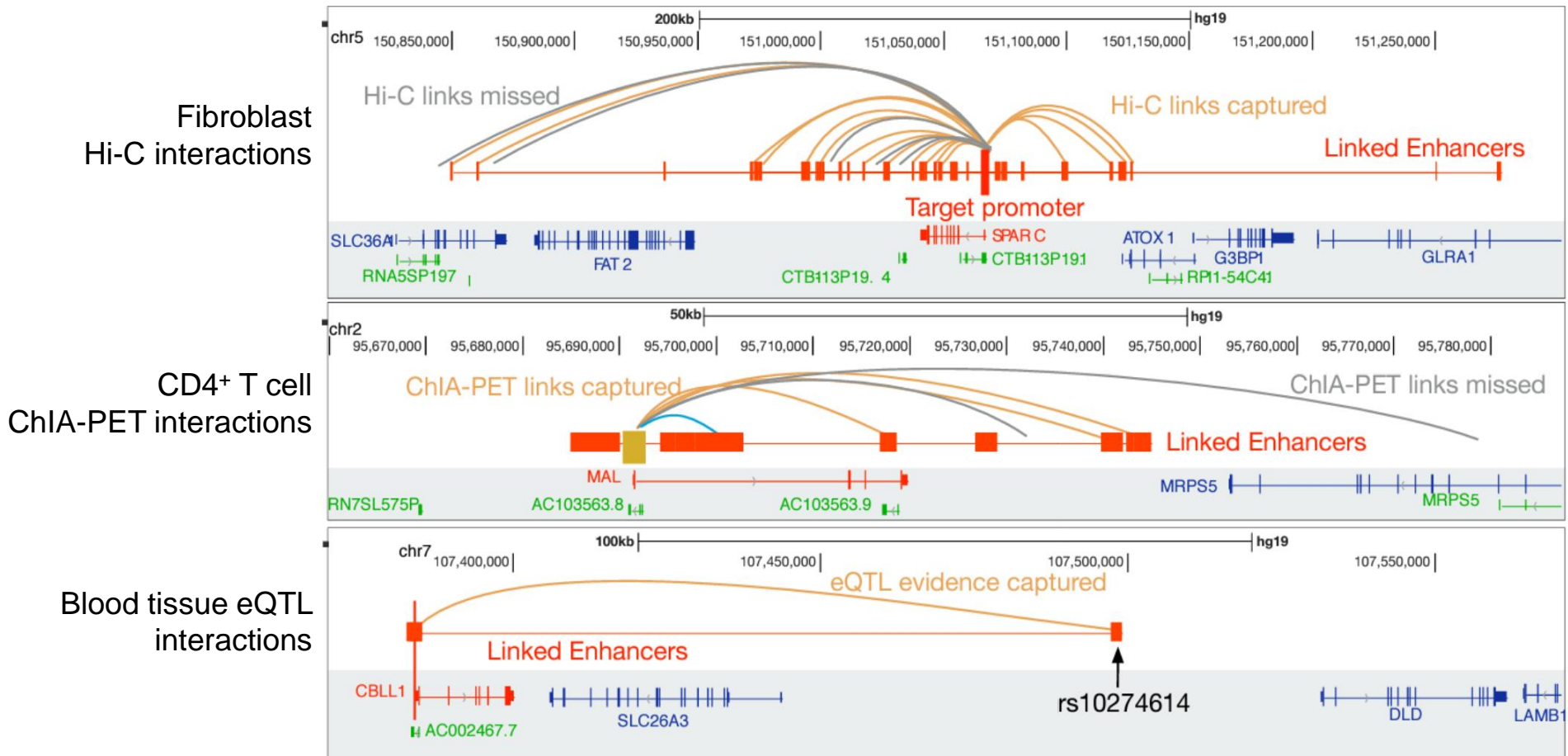
Links are highly tissue-specific

- 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**

Predictions are supported by Hi-C, ChIA-PET and eQTLs

Comparison to experimental interactions in matched cell-types/tissues



Comparison to existing methods

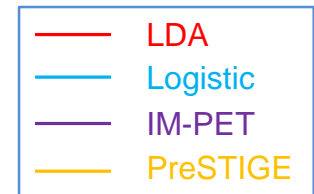
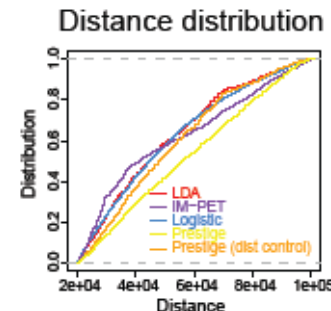
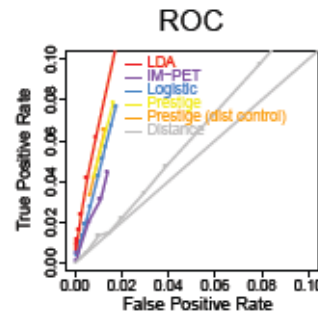
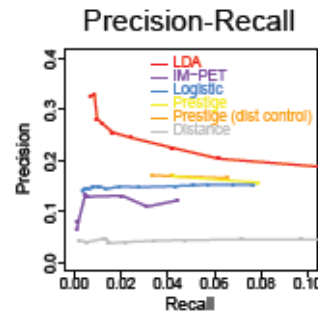
Model performance compared to 3 existing algorithms (supervised and unsupervised).

Gold standards: totally 17 experimental interaction datasets of different cell-types/tissues used for evaluation.

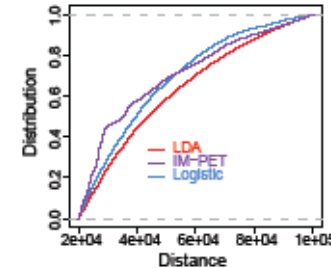
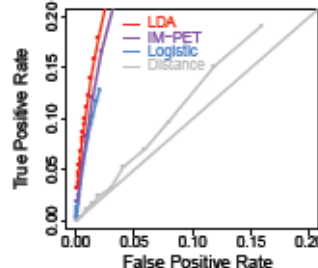
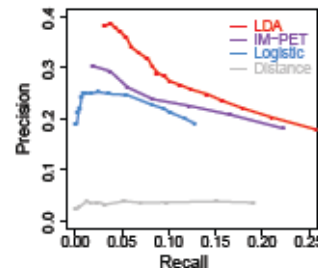
Consistently better performance regardless of the specific experimental datasets or cell-types:

- Area under Precision-recall curves;
- Area under ROC;

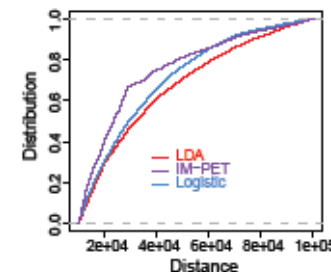
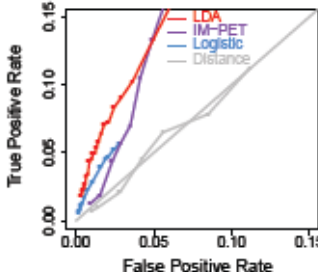
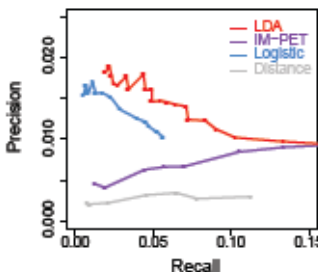
IMR90 Hi-C
Dixon et al. Nature 2015



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



whole blood eQTL
Battle et al. Genome Research 2014



Factors controlled:

- Common enhancer sets;
- Common gene sets;
- Distance distribution;
- Enhancer size.

Enhancer-gene networks in functional GWAS annotation

Population genetics (e.g. GWAS) can tell us *which* genetic variants are associated with different human diseases/phenotypes.

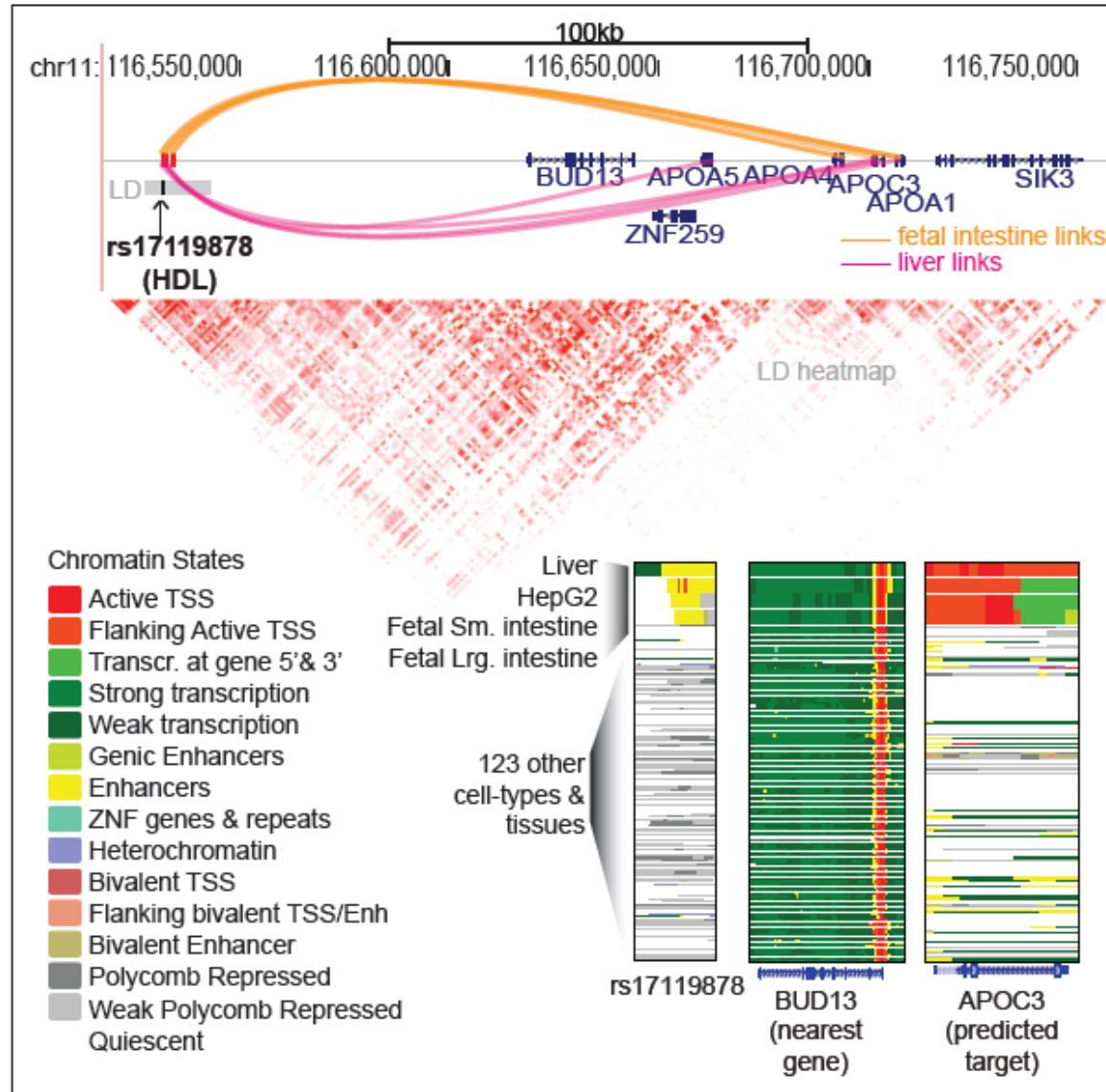
But *how*?

Challenge: vast majority of GWAS variants are located in non-coding regions.

Goal: Identify target genes and pathways affected by non-coding disease variants.

Enhancer-gene networks in functional GWAS annotation

Examples of identified distal target genes of GWAS variants by tissue specific enhancer-gene networks.

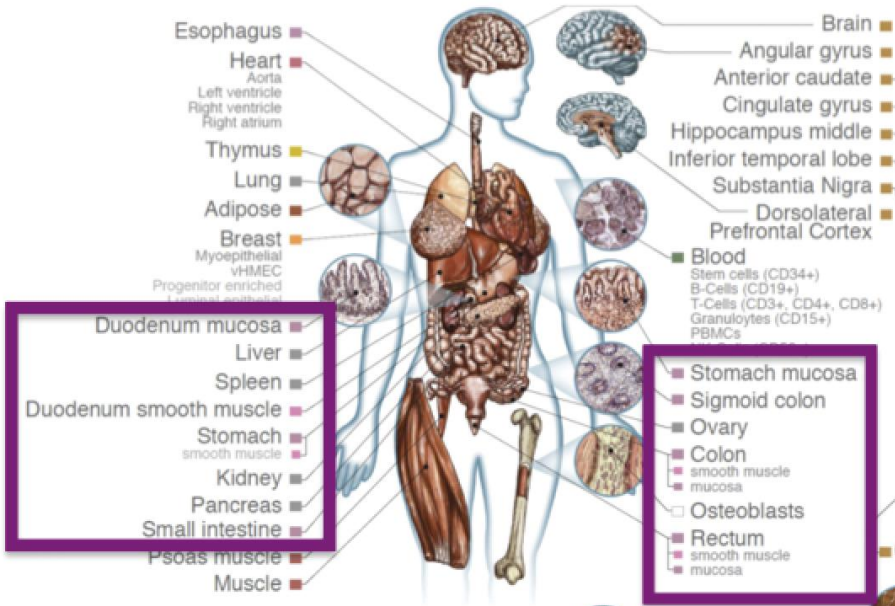


Case study: Colorectal cancer interpretation

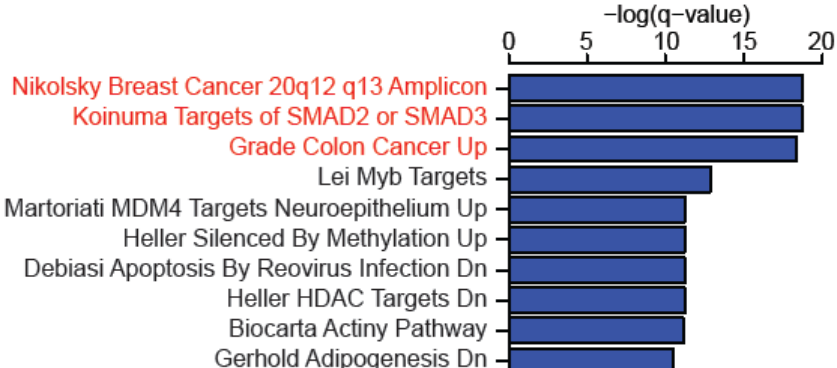
Link enhancers containing colorectal cancer (CRC) variants to genes using predicted interactions in relevant tissues.

60 target genes identified

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



Enriched Gene Sets for CRC GWAS



Top enriched pathways are supported by literature.

As comparison, other methods can only identify 6 genes: not sufficient to look for any pathways.