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

V(\mathbf{v}) = \text{word index}$

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

$W_i = \{m, n\}$

$K = \cdots e \cdot \gamma_i \cdot e$

$P(w_t = t) = \sum_{k=1}^{K} P(\mathbf{v}_t = k | d_m) \cdot P(w_t = t | \mathbf{v}_t = k)$

$\Phi_{m,k}, \Phi_{t,k}, e$

$\tilde{\Theta}_m \propto D_r(\tilde{\Theta}) = \frac{P(\tilde{\Theta}_m)}{\prod_{i=1}^{N} P(\tilde{\Theta}_i)} \cdot \prod_{i=1}^{d-1} \tilde{\Theta}_{m,i}$

$\tilde{\Phi}_k \propto D_r(\tilde{\Phi}) = \frac{P(\tilde{\Phi}_k)}{\prod_{i=1}^{N} P(\tilde{\Phi}_i)} \cdot \prod_{i=1}^{m-1} \tilde{\Phi}_{k,i} \cdot \frac{1}{\Delta(\tilde{\Phi})} \cdot \prod_{i=1}^{m-1} \tilde{\Phi}_{k,i}$

$\Delta = \frac{1}{\prod_{i=1}^{N} P(\tilde{\Phi}_i)} = \Delta(\tilde{\Phi})$

$P(\tilde{\Theta} | \tilde{\Phi}, \tilde{\eta}) = \frac{\prod_{i=1}^{N} \tilde{\Phi}_{k,i}}{\prod_{i=1}^{d-1} \tilde{\Theta}_{m,i}}$

$P(\tilde{\Theta} | \tilde{\Phi}, \tilde{\eta}) = \int P(\tilde{\Theta} | \tilde{\Phi}, \tilde{\eta}) \cdot P(\tilde{\Phi}^t | \tilde{\eta}) d\tilde{\Phi}$

$\tilde{\eta}_y = \left[ \eta_y(t) \right]_{t=1}^{N} = \frac{1}{\gamma(\tilde{\eta})} \int \prod_{i=1}^{N} \phi_{k,i}^{\eta_y(t)} \cdot \left( \frac{K}{\pi} \frac{1}{\Delta(\tilde{\Phi})} \right)^{T/2} d\tilde{\Phi} = \frac{K}{\pi} \frac{1}{\gamma(\tilde{\Phi})} \int \frac{\prod_{i=1}^{N} \phi_{k,i}^{\eta_y(t)+1}}{\Delta(\tilde{\Phi})} d\tilde{\Phi}$
\[ P(\tilde{\beta} | \tilde{\theta}) = \prod_{m=1}^{M} \frac{K}{\pi} \theta_{m,\tilde{\beta}}^{K_{m}(\tilde{\beta})} \]

\[ P(\tilde{\omega} | \tilde{\theta}, \tilde{\beta}) = \int P(\tilde{\omega} | \tilde{\theta}) P(\tilde{\theta} | \tilde{\beta}) d\tilde{\theta} \]

\[ = \int \prod_{m=1}^{M} \frac{K}{\pi} \theta_{m,\tilde{\theta}}^{K_{m}(\tilde{\theta})} \left( \prod_{k=1}^{K} \frac{K_{m}(\tilde{\beta})}{\sigma_{m,k}^{2}} \theta_{m,k}^{\frac{K_{m}(\tilde{\beta})}{2}} \right) d\tilde{\theta} \]

\[ = \frac{M}{\pi^{M}} \left( \int \frac{K}{\sigma_{m,k}^{2}} \theta_{m,k}^{\frac{K_{m}(\tilde{\beta})+2}{2}} d\theta_{m,k} \right) \]

\[ \tilde{\eta}_{m} = \left\{ \frac{K_{m}(\tilde{\beta})}{\sigma_{m,k}^{2}} \right\}^{K_{m}(\tilde{\beta})} \]

\[ P(\tilde{\omega}, \tilde{\theta} | \tilde{\beta}) = P(\tilde{\omega} | \tilde{\theta}, \tilde{\beta}) : P(\tilde{\theta} | \tilde{\beta}) \]

\[ = \frac{\Delta(\tilde{\eta}_{m} + \tilde{\omega})}{\Delta(\tilde{\omega})} \frac{\Delta(\tilde{\eta}_{m} + \tilde{\omega})}{\Delta(\tilde{\omega})} \]

\[ \tilde{\omega} = \{ \omega_{i} = t, \tilde{\omega}_{i} = \tilde{t} \} \]

\[ \tilde{\theta} = \{ \tilde{\gamma}_{i} = k, \tilde{\beta}_{i} = \tilde{\beta}_{i} \} \]

\[ P(\tilde{\gamma}_{i} = k | \tilde{\omega}, \tilde{\beta}, \tilde{\theta}, \tilde{\omega}_{i}, \tilde{\theta}_{i}) = \frac{P(\tilde{\omega}, \tilde{\beta}, \tilde{\theta}) \cdot P(\tilde{\gamma}_{i} | \tilde{\omega}, \tilde{\beta}, \tilde{\theta})}{P(\tilde{\omega}_{i}, \tilde{\beta}, \tilde{\theta}) \cdot P(\tilde{\omega}_{i}, \tilde{\beta}, \tilde{\theta}) \cdot P(\tilde{\gamma}_{i} | \tilde{\omega}, \tilde{\beta}, \tilde{\theta})} \]

\[ = \frac{\Delta(\tilde{\omega}_{i} + \tilde{\beta})}{\Delta(\tilde{\eta}_{m} + \tilde{\omega})} \frac{\Delta(\tilde{\eta}_{m} + \tilde{\omega})}{\Delta(\tilde{\omega})} \]

\[ = \frac{\Delta(\tilde{\omega}_{i} + \tilde{\beta})}{\Delta(\tilde{\eta}_{m} + \tilde{\omega})} \frac{\Delta(\tilde{\omega}_{i} + \tilde{\beta})}{\Delta(\tilde{\eta}_{m} + \tilde{\omega})} \]
Introduction of Machine Learning in Computational Biology

Jianrong Wang

CMSE @ MSU
Genetic variants associated with human disease

Genetic variation

GWAS: Genome-wide Association Studies

Organism Phenotype

Height
Eye color
Disease

? Underlying mechanisms
**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.

![Diagram showing the relationship between genetic variation, perturbed regulatory sub-networks, disease status, and improved diagnostics and therapeutics.](image-url)
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.
Epigenomics: Genome-wide ChIP-seq histone modification maps enable systematic characterization of cell-type specific combinatorial signatures.

Big data: large panels of epigenomic datasets

ChIP-seq, RNA-seq, DNase-seq, ATAC-seq, MNase-seq, Hi-C, Capture-C, ChIA-PET, WGBS, etc
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:**

- Data/signal processing
- Regulatory elements and signatures
- Regulatory networks
- Disease mechanisms
- Genomics-based therapeutics

Functional genomics datasets (epigenomics, transcriptomics, connectomics TF binding)

Biological hypothesis/knowledge

Population genetics datasets (GWAS, cancer)

Personalized genome data
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?
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

**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 \( z_n \).

\[ \alpha \rightarrow \theta \rightarrow z \rightarrow w \rightarrow N \]

\[ \beta \]
Mixed-membership modules – Latent Dirichlet Allocation

Hierarchical graphical model:

\[
p(\theta | \alpha) = \frac{\Gamma \left( \sum_{i=1}^{k} \alpha_i \right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \cdots \theta_k^{\alpha_k-1}
\]

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

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

\[
p(D | \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d | \alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d)p(w_{dn} | z_{dn}, \beta) \right) d\theta_d
\]
Mixed-membership modules – Latent Dirichlet Allocation

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

\[ p(e_{nt}) = \sum_k p(e_{nt} | z = k) p(z = k | t) \]
A novel prob. model for enhancer-gene linking

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.
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.

\[ \vartheta_{t,k} = p(z = k|t) \]  
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;

\[
\begin{align*}
\text{Algorithm 1 Gibbs sampler} \\
\text{Initialize } x^{(0)} &\sim q(x) \\
\text{for iteration } i = 1, 2, \ldots \text{ do} \\
&x_1^{(i)} \sim p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, \ldots, X_D = x_D^{(i-1)}) \\
&x_2^{(i)} \sim p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, \ldots, X_D = x_D^{(i-1)}) \\
&\vdots \\
&x_D^{(i)} \sim p(X_D = x_D | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, \ldots, X_D = x_D^{(i-1)}) \\
\text{end for}
\end{align*}
\]

from Ilker Yildirim

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

<table>
<thead>
<tr>
<th>Enhancer module 1</th>
<th>t₁</th>
<th>t₂</th>
<th>t₃</th>
<th>t₄</th>
<th>t₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>θ₁,1</td>
<td>θ₂,1</td>
<td>θ₃,1</td>
<td>θ₄,1</td>
<td>θ₅,1</td>
<td></td>
</tr>
<tr>
<td>Enhancer module 2</td>
<td>θ₁,2</td>
<td>θ₂,2</td>
<td>θ₃,2</td>
<td>θ₄,2</td>
<td>θ₅,2</td>
</tr>
<tr>
<td>Enhancer module 3</td>
<td>θ₁,3</td>
<td>θ₂,3</td>
<td>θ₃,3</td>
<td>θ₄,3</td>
<td>θ₅,3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene module 1</th>
<th>t₁</th>
<th>t₂</th>
<th>t₃</th>
<th>t₄</th>
<th>t₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>θ₁,1</td>
<td>θ₂,1</td>
<td>θ₃,1</td>
<td>θ₄,1</td>
<td>θ₅,1</td>
<td></td>
</tr>
<tr>
<td>Gene module 2</td>
<td>θ₁,2</td>
<td>θ₂,2</td>
<td>θ₃,2</td>
<td>θ₄,2</td>
<td>θ₅,2</td>
</tr>
</tbody>
</table>

\[
A = (a_{ij})_{K₁ \times K₂}
\]

\[
a_{ij} \text{ the posterior probability that the } i\text{th enhancer module is associated with the } j\text{th gene module}
\]

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

We use a diffusion model to estimate

\[
P(\bar{\theta}^1, \bar{\theta}^2 | z^1_i \sim z^2_j)
\]
Gibbs sampling approach to jointly infer all parameters

Joint Gibbs sampling inference on 2 connected LDA models

- Update latent module assignment for enhancer matrix
- Update profile probabilities for enhancer matrix: \( \varphi^1_{k,i} \);
- Update fraction probabilities for enhancer matrix: \( \vartheta^1_{t,i} \);
- Update latent module assignment for gene matrix
- Update profile probabilities for gene matrix: \( \varphi^2_{k,i} \);
- Update fraction probabilities for gene matrix: \( \vartheta^2_{t,i} \);
- Update association probabilities: \( a_{ij} \)

Key parameters inferred for the model:
1. Enhancer-to-module and gene-to-module probabilities \( (\varphi_{i,.}) \);
2. Module-to-tissue probabilities \( (\vartheta_{t,.}) \);
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)

Wang et al. 2016  In Revision Nature Genetics
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)
Learned enhancer/gene modules and associations

Cell types

Modules and module links

Gene expression by modules

Enhancer activity by modules

Modules are enriched with specific biological pathways

Gene modules

Stem cell differentiation

Cellular component biogenesis at cellular level

Chromatin assembly or disassembly

Immune system process

Neuron differentiation

Organic acid metabolic process

Transmission of nerve impulse

Nucleotide transport

Regulation of interferon-γ-mediated signaling pathway

Apoptosis

Marrow cell differentiation

Immune related terms

Metabolism

Insulin secretion

Muscle contraction

Endothelial cell differentiation

Stem cell differentiation

Immune system process

Neuron differentiation

Organic acid metabolic process

Transmission of nerve impulse

Nucleotide transport

Regulation of interferon-γ-mediated signaling pathway

Apoptosis

Marrow cell differentiation

Immune related terms

Metabolism

Insulin secretion

Muscle contraction

Endothelial cell differentiation
Basic properties of long-range enhancer-gene networks

- At the threshold of 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

Fibroblast Hi-C interactions

CD4\(^+\) T cell ChIA-PET interactions

Blood tissue eQTL interactions
Overall performance: accuracy and cell-type specificity

**Accuracy: high Area Under ROC**
(gold standard: experimental interactions from matched cell-types)

- **Hi-C**
  - AUC = 0.790

- **ChIA-PET**
  - AUC = 0.895

- **eQTL**
  - AUC = 0.793

**Cell-type specificity**
(significant overlaps only observed for experimental datasets from matched cell-types)

**Hi-C interaction enrichment**
- Mesendoderm
- H1
- Neuronal Progenitor
- Trophoblast
- Mesenchymal
- Lung Fibroblast

**ChIA-PET interaction enrichment**
- Breast + Cancer
- Immune
- MCF7
- CD4
- K562

**eQTL enrichment**
- Liver
- Skin
- Fat
- LCL_1
- LCL_2
- Blood

Predictions in 56 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;

Factors controlled:
- Common enhancer sets;
- Common gene sets;
- Distance distribution;
- Enhancer size.
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**

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
<th>Gene 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADNP</td>
<td>DDB1</td>
<td>MYC</td>
<td>SMAD7</td>
</tr>
<tr>
<td>ADRM1</td>
<td>DEF6</td>
<td>MYL2</td>
<td>SRPK1</td>
</tr>
<tr>
<td>ALDH2</td>
<td>DHX9</td>
<td>NEU1</td>
<td>TBC1D5</td>
</tr>
<tr>
<td>ARPC2</td>
<td>DIP2B</td>
<td>PGA3</td>
<td>TBX2</td>
</tr>
<tr>
<td>ARPC5</td>
<td>DSP</td>
<td>PGA4</td>
<td>TEAD3</td>
</tr>
<tr>
<td>ATF1</td>
<td>F3</td>
<td>PGA5</td>
<td>TMBIM1</td>
</tr>
<tr>
<td>C11orf92</td>
<td>FGR</td>
<td>POU5F1</td>
<td>TMBIM6</td>
</tr>
<tr>
<td>C11orf93</td>
<td>FKBP5</td>
<td>PPARD</td>
<td>TMEM138</td>
</tr>
<tr>
<td>C20orf166</td>
<td>IER3</td>
<td>PSMA7</td>
<td>TMEM189</td>
</tr>
<tr>
<td>CABLES2</td>
<td>IFI6</td>
<td>RAD21</td>
<td>-UBE2V1</td>
</tr>
<tr>
<td>CCND2</td>
<td>KANK1</td>
<td>RCSD1</td>
<td>WASF2</td>
</tr>
<tr>
<td>CD247</td>
<td>LAMA5</td>
<td>RNF114</td>
<td></td>
</tr>
<tr>
<td>CLPS</td>
<td>LAMC1</td>
<td>RNF169</td>
<td></td>
</tr>
<tr>
<td>CNN3</td>
<td>LAMC2</td>
<td>SATB1</td>
<td></td>
</tr>
<tr>
<td>CREG1</td>
<td>METTL7A</td>
<td>SFN</td>
<td></td>
</tr>
<tr>
<td>CTNNB1</td>
<td>MPZL1</td>
<td>SH2B3</td>
<td></td>
</tr>
</tbody>
</table>

**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.