Dimension reduction and integrative clustering on genomics data

Zhixiang Lin

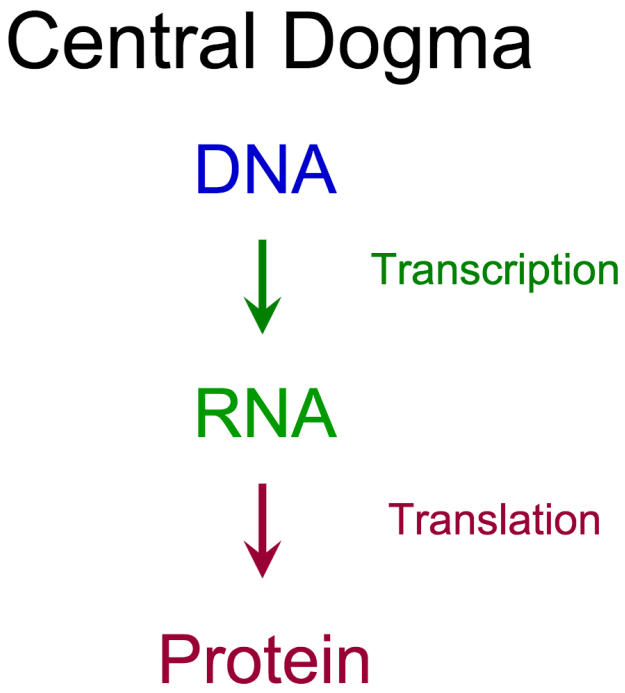
Stanford University

11-9-2017

AC-PCA: simultaneous dimension reduction and Adjustment for Confounding variation

Joint work with Can Yang, John C. Duchi, Hongyu Zhao and Wing Hung Wong

High throughput gene expression data



- Microarray and RNA-Seq measure the genome-wide expression level of RNAs

Dimension Reduction - Motivation

- For single cell genomics, dimension reduction and clustering methods are important
  - A mixture of cells with unknown “identity”

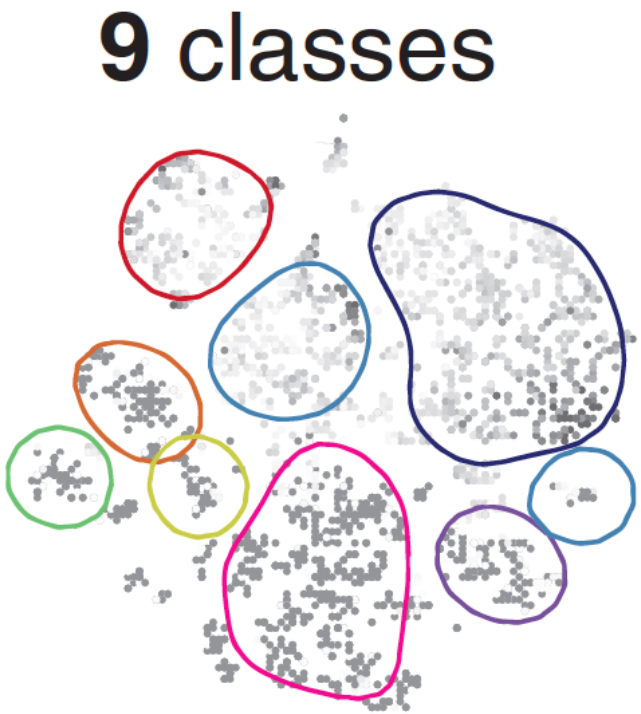


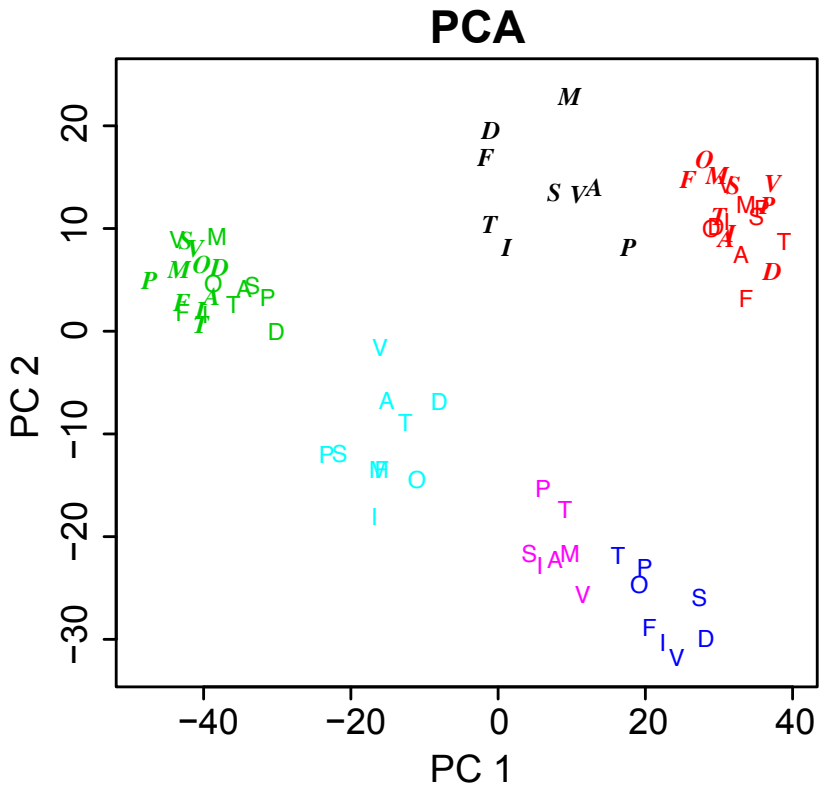
Figure: Mouse brain single cell RNA-Seq (Zeisel A et al. 2015)

Confounding factors

- Commonly observed in high throughput biological data
  - Technical: reagents, temperature...
  - Biological: race, age, donor, species...
- Confounding factors “hurts”
  - Wrong patterns: confounding variation may dominate over the desired variation
  - Wrong genes: select genes associated with the confounding factors

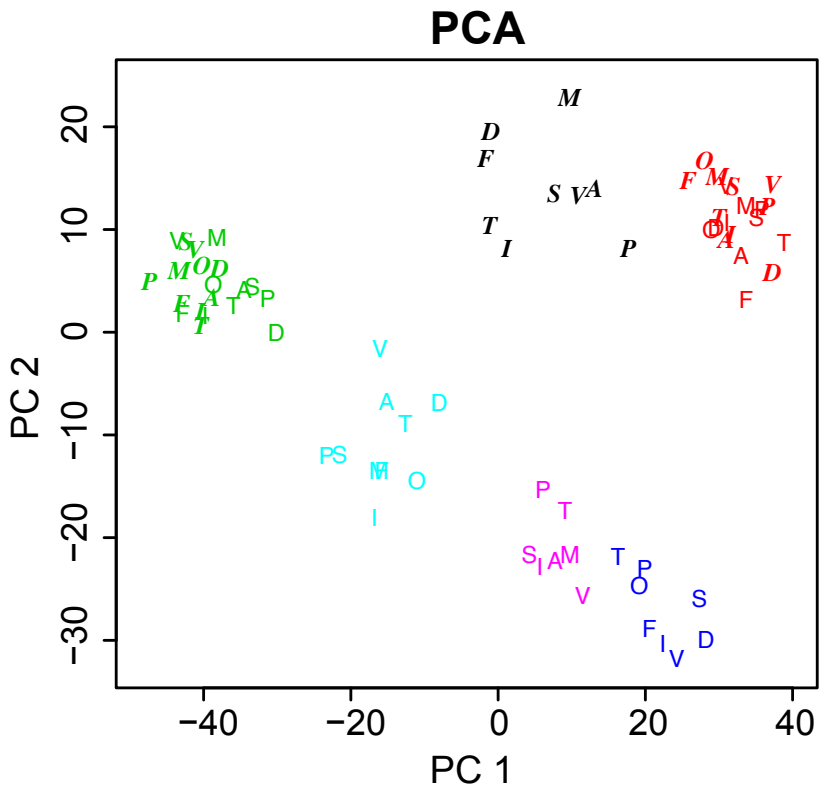
Motivating example: human brain exon array data

- Expression levels for 17,565 genes in 16 brain regions and 9 time windows (Kang HJ et al. 2011)



Motivating example: human brain exon array data

- Expression levels for 17,565 genes in 16 brain regions and 9 time windows (Kang HJ et al. 2011)
- Challenging to identify the interregional variation shared among donors



Methods for confounder adjustment (1)

ComBat (Johnson WE et al. 2007):
The expression level for gene g in sample j from batch i
$Y_{ijg} = \alpha_g + x_j\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg}$

- x_j consist of the covariates of interest
  - Brain region label
- γ_{ig} and δ_{ig} characterize the additive and multiplicative batch effects
  - For gene g, batch effect is the same across j
  - Donor label
- $Y^*_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - x_j\hat{\beta}_g - \hat{\gamma}_{ig}}{\hat{\delta}_{ig}} + \hat{\alpha}_g + x_j\hat{\beta}_g$

Methods for confounder adjustment (2)

Suppose the batch is unobserved. The expression level for gene g in sample j: $Y_{jg} = \alpha_g + x_j\beta_g + w_j h_g + \epsilon_{jg}$

- Y_{jg} and x_j are observed
- rank k for the unwanted variation

SVA (Leek JT et al. 2007)

- Fit $Y_{jg} = \alpha_g + x_j\beta_g + \epsilon_{jg}$
- SVD on the residuals

RUV

- For a set of negative control genes, assume $\beta_g = 0$ (Gagnon-Bartsch JA et al. 2012)
- Other extensions (Gagnon-Bartsch et al. 2013, Risso et al. 2014, and Jacob et al. 2016)

AC-PCA in a general form

- Data matrix $X_{N \times p}$

AC-PCA in a general form

- Data matrix $X_{N \times p}$
- The projection vector v

AC-PCA in a general form

- Data matrix $X_{N \times p}$
- The projection vector v
- Projection for the ith observation $t_i = x_{(i)} \cdot v$

AC-PCA in a general form

- Data matrix $X_{N \times p}$
- The projection vector v
- Projection for the ith observation $t_i = x_{(i)} \cdot v$
- Total variation after the projection $\sum_{i=1}^{N} t_i^2 = v^T X^T X v$

AC-PCA in a general form

- Data matrix $X_{N\times p}$
- The projection vector v
- Projection for the ith observation $t_i = x_{(i)} \cdot v$
- Total variation after the projection $\sum_{i=1}^{N} t_i^2 = v^T X^T X v$
- Confounder matrix $Y_{N\times l}$ is chosen so that $v^T X^T Y Y^T X v$ represent the confounding variation in t

AC-PCA in a general form

- Data matrix $X_{N\times p}$
- The projection vector v
- Projection for the ith observation $t_i = x_{(i)} \cdot v$
- Total variation after the projection $\sum_{i=1}^{N} t_i^2 = v^T X^T X v$
- Confounder matrix $Y_{N\times l}$ is chosen so that $v^T X^T Y Y^T X v$ represent the confounding variation in t
- The optimization problem:

$$\begin{aligned} \underset{v\in\mathbb{R}^p}{\text{maximize}} \quad & v^T X^T X v - \lambda v^T X^T Y Y^T X v \\ \text{subject to} \quad & ||v||_2^2 \leq 1, \end{aligned} \tag{1}$$

AC-PCA in a general form

- Data matrix $X_{N\times p}$
- The projection vector v
- Projection for the ith observation $t_i = x_{(i)} \cdot v$
- Total variation after the projection $\sum_{i=1}^{N} t_i^2 = v^T X^T X v$
- Confounder matrix $Y_{N\times l}$ is chosen so that $v^T X^T Y Y^T X v$ represent the confounding variation in t
- The optimization problem:

$$\begin{aligned} \underset{v\in\mathbb{R}^p}{\text{maximize}} \quad & v^T X^T X v - \lambda v^T X^T Y Y^T X v \\ \text{subject to} \quad & ||v||_2^2 \leq 1, \end{aligned} \tag{1}$$

- Capture variation that is "invariant" to the confounding factors

AC-PCA in a general form

- The optimization problem:

$$\begin{aligned} \underset{v \in \mathbb{R}^p}{\text{maximize}} \quad & v^T X^T X v - \lambda v^T X^T Y Y^T X v \\ \text{subject to} \quad & ||v||_2^2 \leq 1, \end{aligned} \tag{2}$$

- When $\lambda \to \infty$, Xv tends to be orthogonal to the columns in Y

AC-PCA adjusting for variations of individual donors

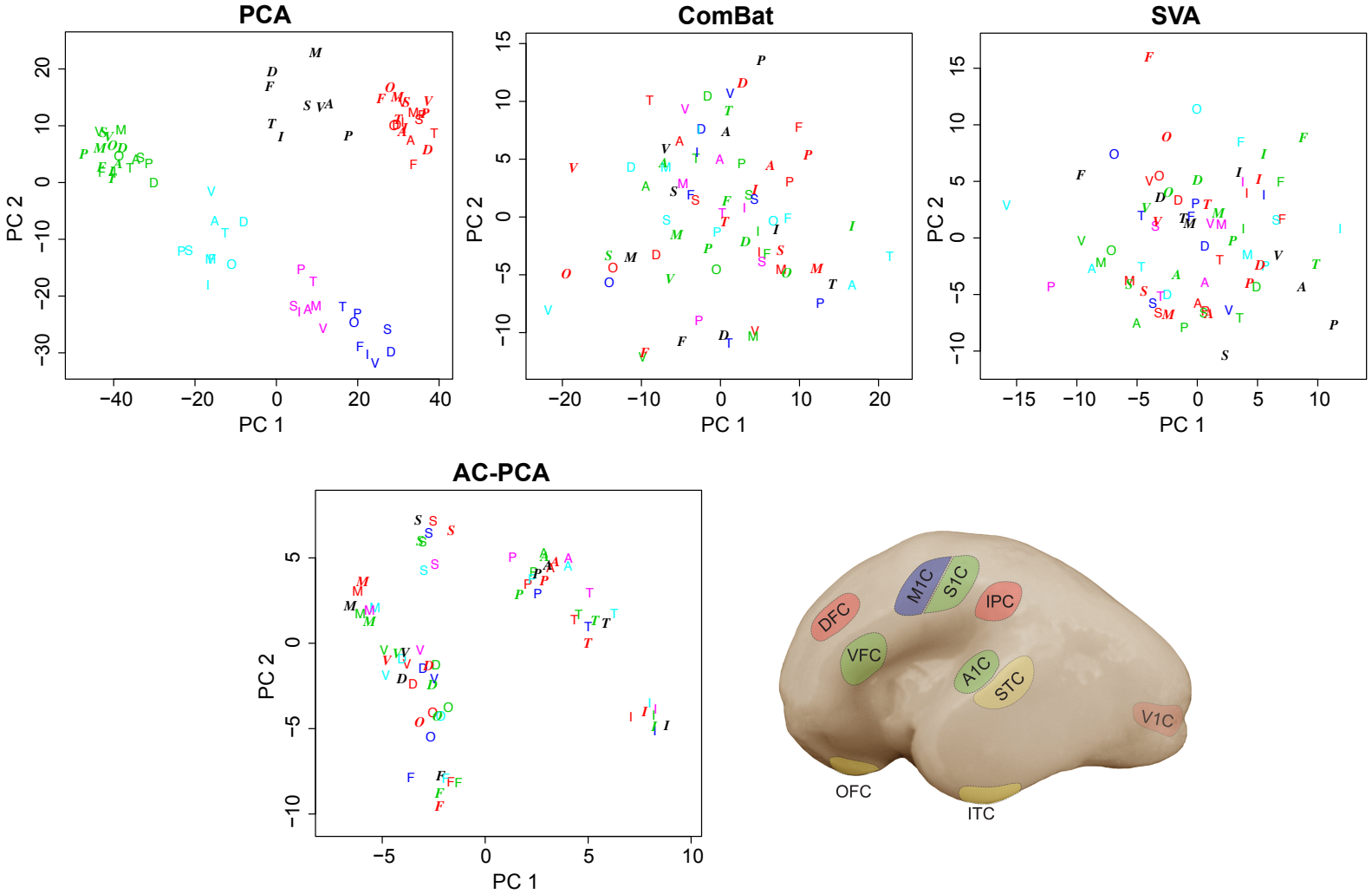
- Capture the interregional variation shared among donors
- $X^{(i)}$, $b \times p$ data matrix for donor i, $i = 1, \cdots, n$
- The optimization problem:

$$\underset{v \in \mathbb{R}^p}{\text{maximize}}\ v^T X^T X v - \frac{\lambda}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} v^T (X^{(j)} - X^{(i)})^T (X^{(j)} - X^{(i)}) v$$

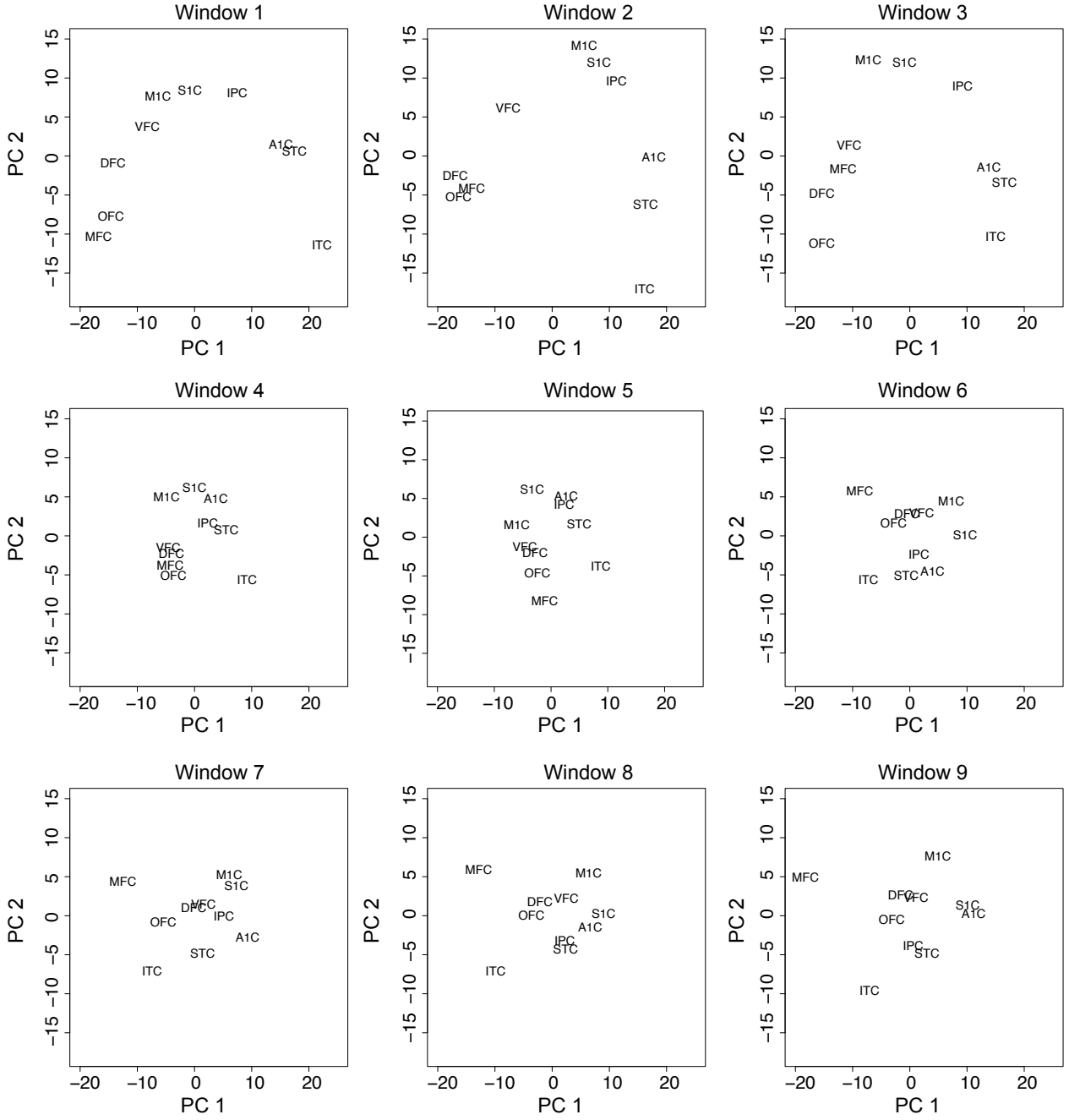
$$\text{subject to } ||v||_2^2 \leq 1. \tag{3}$$

- The regularization term encourages the clustering of regions, after the projection:
$(X^{(j)} - X^{(i)})v = \sum_{k=1}^{b} (X_{k\cdot}^{(j)} v - X_{k\cdot}^{(i)} v)$

Application to the human brain exon array data

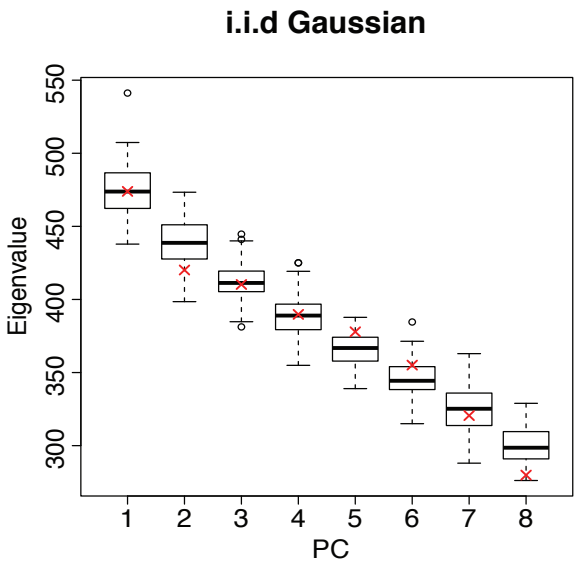


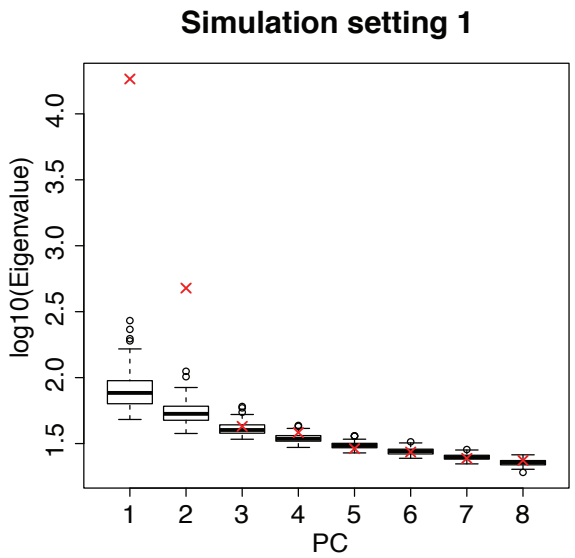
- Time window 5

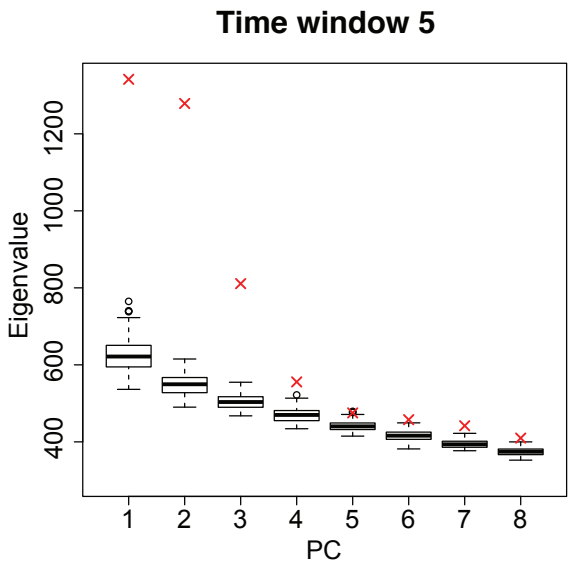

Window 1
Window 2
Window 3
Window 4
Window 5
Window 6
Window 7
Window 8
Window 9
PC 1
PC 2
M1C
S1C
IPC
VFC
A1C
STC
DFC
OFC
MFC
ITC

The significant PCs

- The penalty term can induce artificial clusters
- To evaluate the PCs, the rows in X are permutated and keep Y the same
  - Compare the eigenvalues and variance, data vs. permutation







The analysis of variance interpretation

For a number of observations that can be divided in K groups

- t_{jk}, jth observation in group k, $j = 1, \cdots, n_k$
- Group mean $\bar{t}_{.k}$ and grand mean $\bar{t}_{..}$
- The total sum of squares $SS_T = \sum_{k=1}^{K} \sum_{j=1}^{n_k} (t_{jk} - \bar{t}_{..})^2$
- The between groups sum of squares $SS_B = \sum_{k=1}^{K} n_k (\bar{t}_{.k} - \bar{t}_{..})^2$
- We have $SS_T = SS_B + SS_R$

The analysis of variance interpretation

Consider the samples for region r in the brain data. The donor labels represent the groups.

- We have $K = n$ and $n_k = 1, \forall k$
- $SS_B^{(r)}(v) = \sum_{k=1}^{n}(t_{rk} - \bar{t}_{r\cdot})^2 = \frac{1}{n}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}(t_{rk} - t_{rl})^2$
  Let $SS_B^*(v) \equiv \sum_{r=1}^{b} SS_B^{(r)}(v)$
- $SS_T(v) = \sum_{r=1}^{b}\sum_{k=1}^{n}(t_{rk} - \bar{t}_{\cdot\cdot})^2 = \sum_{r=1}^{b}\sum_{k=1}^{n}(t_{rk})^2$

$$v^T X^T X v - \frac{\lambda}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} v^T (X^{(j)} - X^{(i)})^T (X^{(j)} - X^{(i)}) v$$

$$= SS_T(v) - \lambda SS_B^*(v)$$

The penalty term can be formulated based on the ANOVA interpretation

AC-PCA with sparse loading

$$\begin{aligned} \underset{v \in \mathbb{R}^p}{\text{maximize}} \quad & v^T X^T X v \\ \text{subject to} \quad & v^T X^T Y Y^T X v \leq c_1, \\ & ||v||_1 \leq c_2, \\ & ||v||_2^2 \leq 1. \end{aligned} \tag{4}$$

- c_1 and c_2 are the tuning parameters

Algorithm for sparse AC-PCA

The optimization problem is equivalent to (Witten et al. 2009)

$$\begin{aligned} \underset{u,v}{\text{minimize}} \quad & -u^T X v \\ \text{subject to} \quad & v^T X^T Y Y^T X v \leq c_1 \\ & ||v||_1 \leq c_2 \\ & ||v||_2^2 \leq 1 \\ & ||u||_2^2 \leq 1 \end{aligned}$$

- Biconvex in u and v
- Find local optimal by iteratively update between u and v

Updating u and v

- The update $u = \frac{Xv}{||Xv||_2}$
- To update v

$$\begin{aligned} \underset{v}{\text{minimize}} \quad & -u^T Xv \\ \text{subject to} \quad & v^T X^T Y Y^T Xv \leq c_1 \\ & ||v||_1 \leq c_2 \\ & ||v||_2^2 \leq 1 \end{aligned}$$

- Convex. Can be slow when v is large

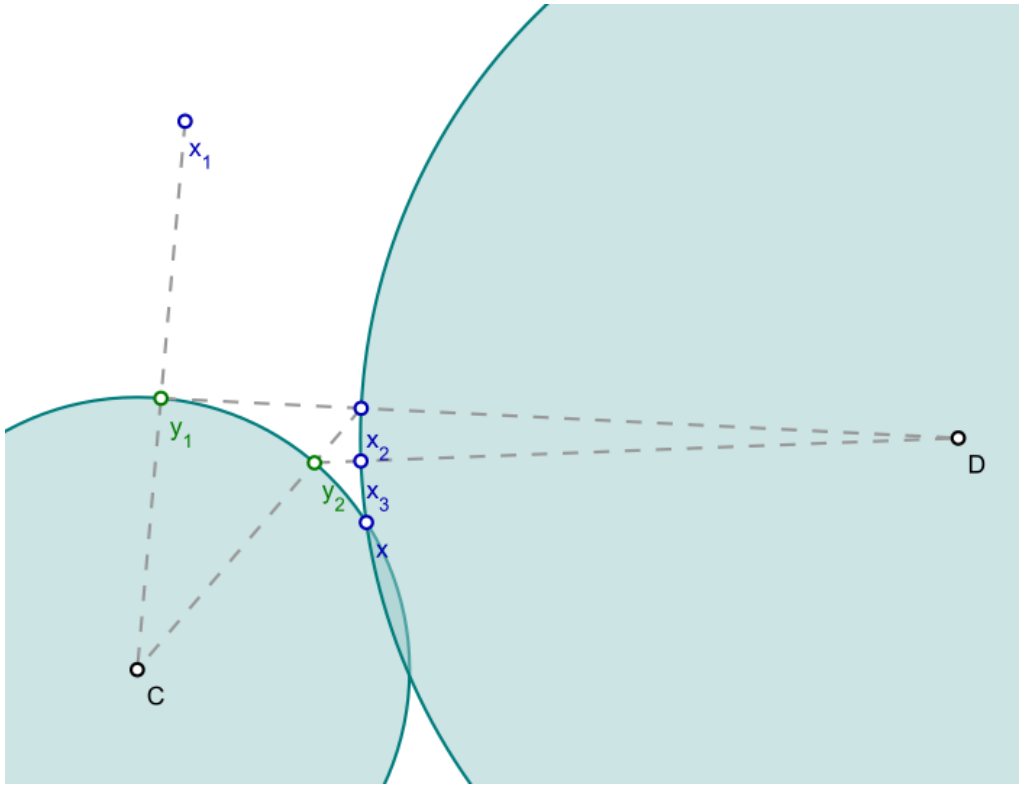
Updating v

Formulate it as a feasibility problem (joint work with John C. Duchi):

$$\begin{aligned} \text{find} \quad & v \\ \text{subject to} \quad & -u^T X v \le t \\ & v^T X^T Y Y^T X v \le c_1 \\ & ||v||_1 \le c_2 \\ & ||v||_2^2 \le 1 \end{aligned}$$

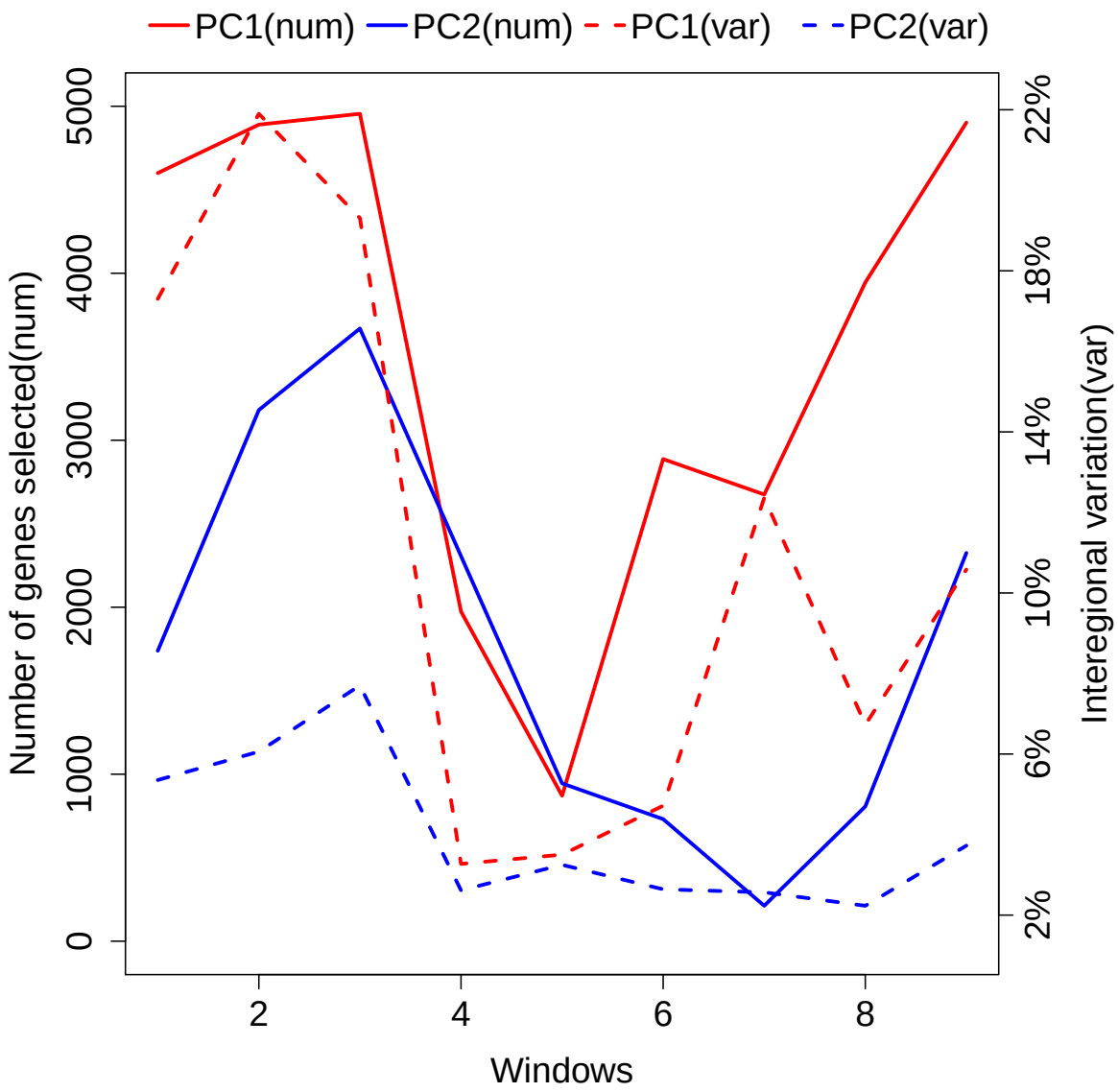
- Binary search on t
- Alternating projection to check feasibility
- $\sim$ 10-fold faster than cvx

Alternating projection



Adapted from Wikipedia

Application to the human brain exon array data



Summary

- We have proposed a general class of penalty functions in PCA for simultaneous dimension reduction and adjustment for confounding variation
- We have demonstrated its performance through real data analysis and simulations

Lin Z, Yang C, Zhu Y, Duchi JC, Fu Y, Wang Y, Jiang B, Zamanighomi M, Xu X, Li M, Sestan N, Zhao H, and Wong WH. *Simultaneous dimension reduction and adjustment for confounding variation*. Proceedings of the National Academy of Sciences of the United States of America, 2016.

Integrative clustering of chromatin accessibility and gene expression with application to single cell genomics

Joint work with Mahdi Zamanighomi, Timothy Daley, Shining Ma and Wing Hung Wong

Motivation

- Multi-omic data on the same sample provide rich insight in biomedical research
  - Better characterization of immune cells, tumor subtypes
  - More information on transcriptional regulation
  - ...
- Technical challenges for single cell multi-omic experiment
  - Lower throughput (fewer cells, fewer reads per cell)
  - More prone to noise: loss of nucleic acids
- Most clustering methods for single cell data do not consider the uncertainty in the cluster assignment
  - Intermediate stage
  - “Problematic” cells

Chromatin Accessibility

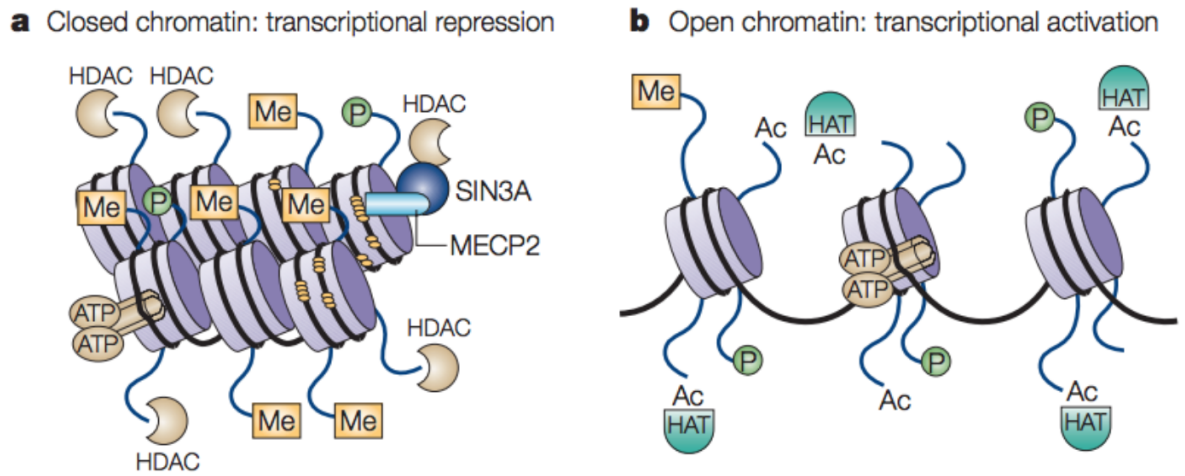
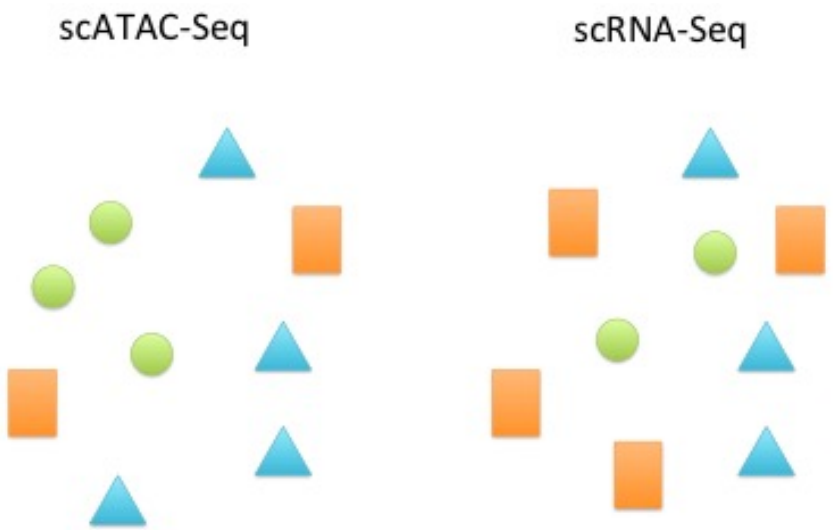


Figure: Chromatin structure regulates transcriptional activity (Johnstone RW 2002)

- Open chromatin associates with active transcription
- DNase-Seq, FAIRE-Seq and ATAC-Seq
- Technically challenging to measure chromatin accessibility and gene expression for the same cell

Our goal



Assume that scATAC-Seq and scRNA-Seq experiments are performed for similar cell populations but on different cells

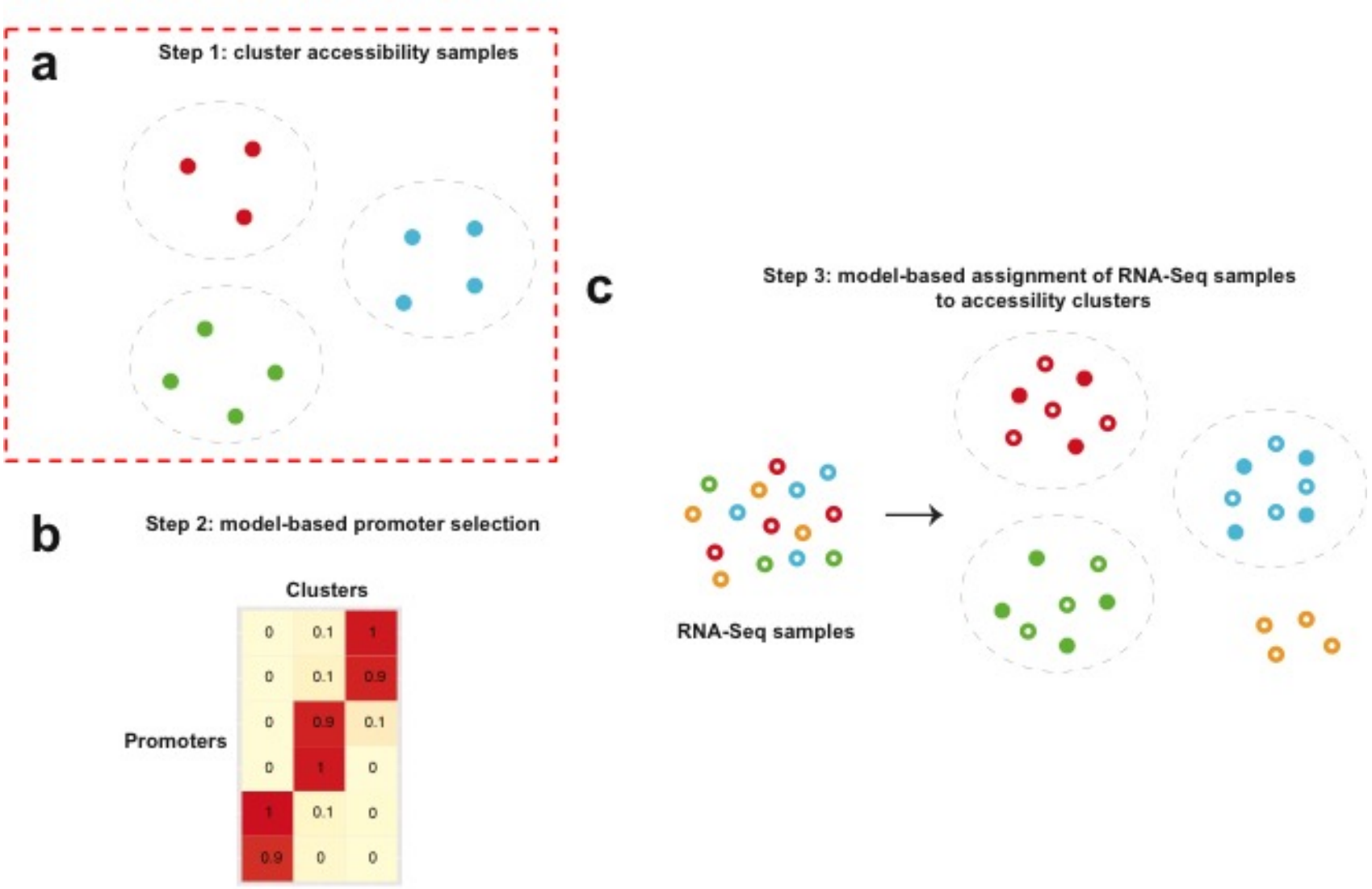
- Match the cell sub-populations in the two data types
- One data type (accessibility) guides the clustering of the other (gene expression)
- Model-based method enables inference on the cluster assignment

Data example

in silico heterogeneous population from mixing public cell line data

- scATAC-Seq data (Buenrostro et al. 2015, Nature)
  - K562 chronic myelogenous leukaemia: 223 samples
  - HL-60 promyeloblasts: 91 samples
- scRNA-Seq data (Pollen et al. 2014, Nat Biotech)
  - K562: 42 samples
  - HL-60: 54 samples
- Assume cell type information are not given

The workflow for our approach

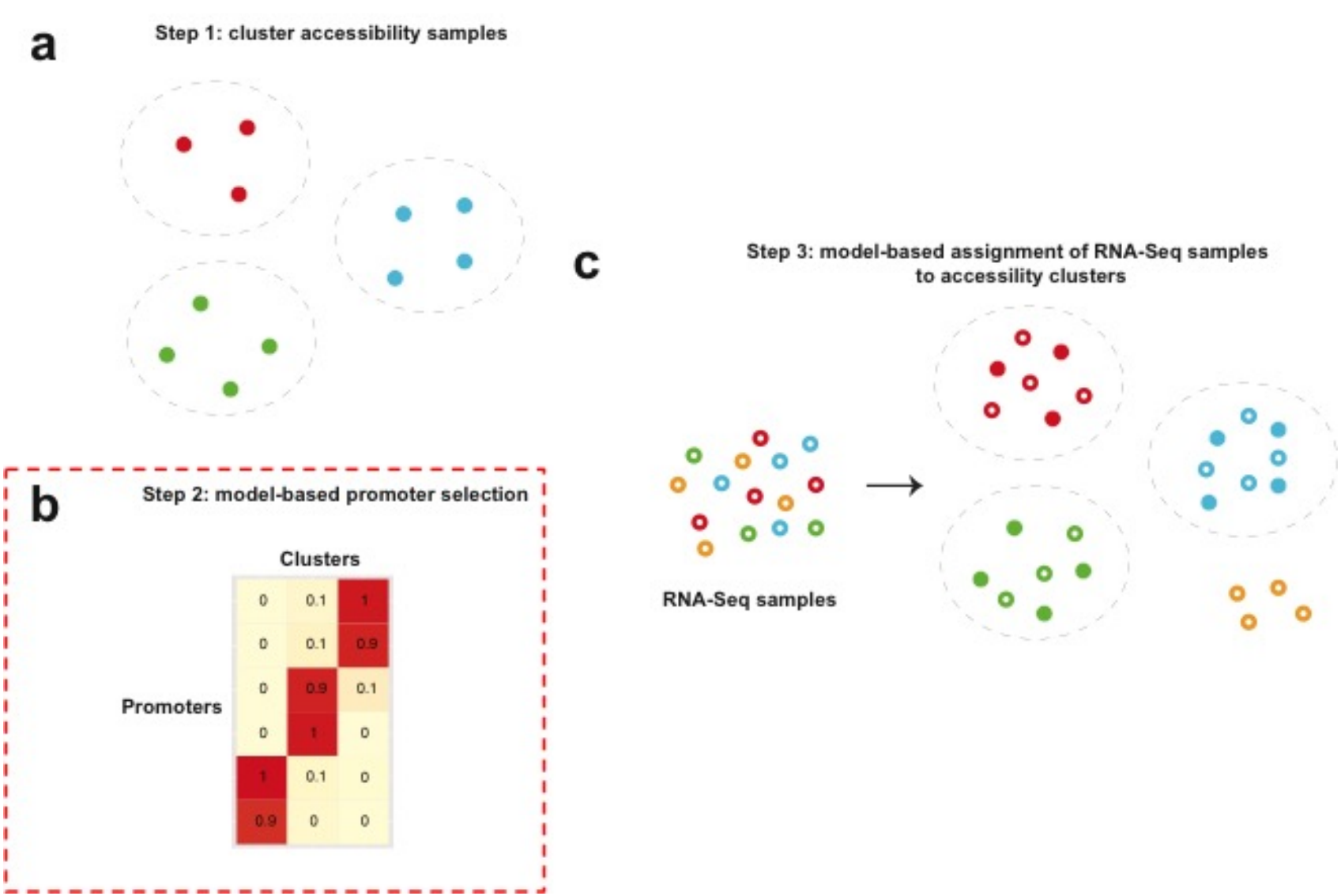


Step 1: cluster accessibility samples

- Implement *scABC* to cluster scATAC-Seq samples
  - All open regions
- Good separation of K562 and HL-60

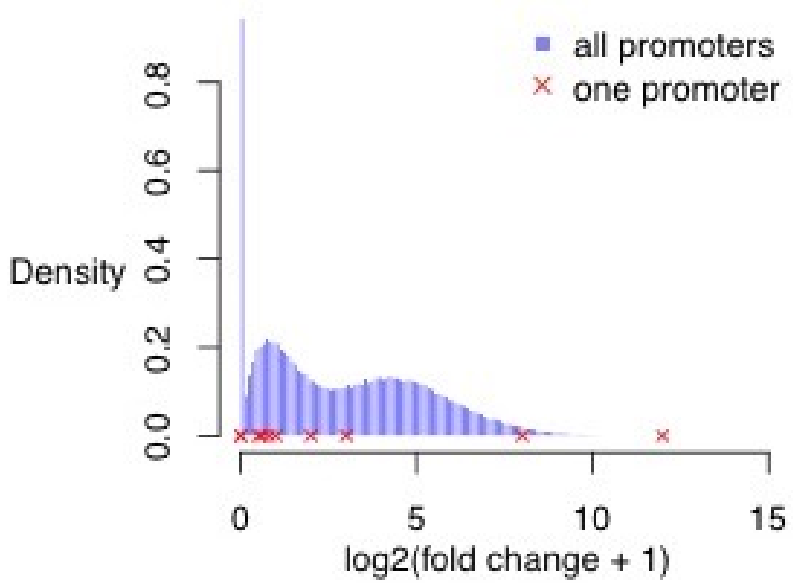
		scATAC-Seq samples	
		K562	HL60
scATAC-Seq clusters	Cluster 1	223	0
	Cluster 2	0	91

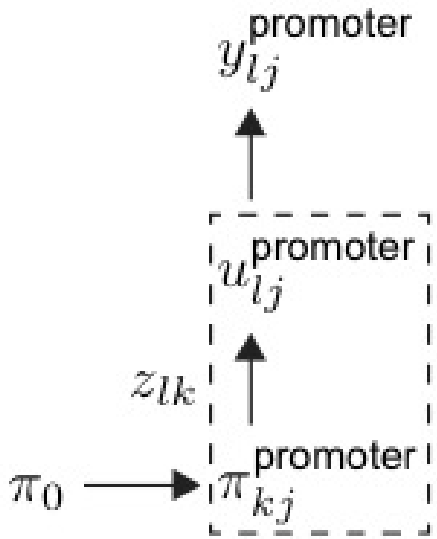
scABC: Unsupervised Clustering And Epigenetic Classification Of Single Cells. Zamanighomi M*, Lin Z*, Daley T*, Schep A, Greenleaf WJ, Wong WH. doi: https://doi.org/10.1101/143701

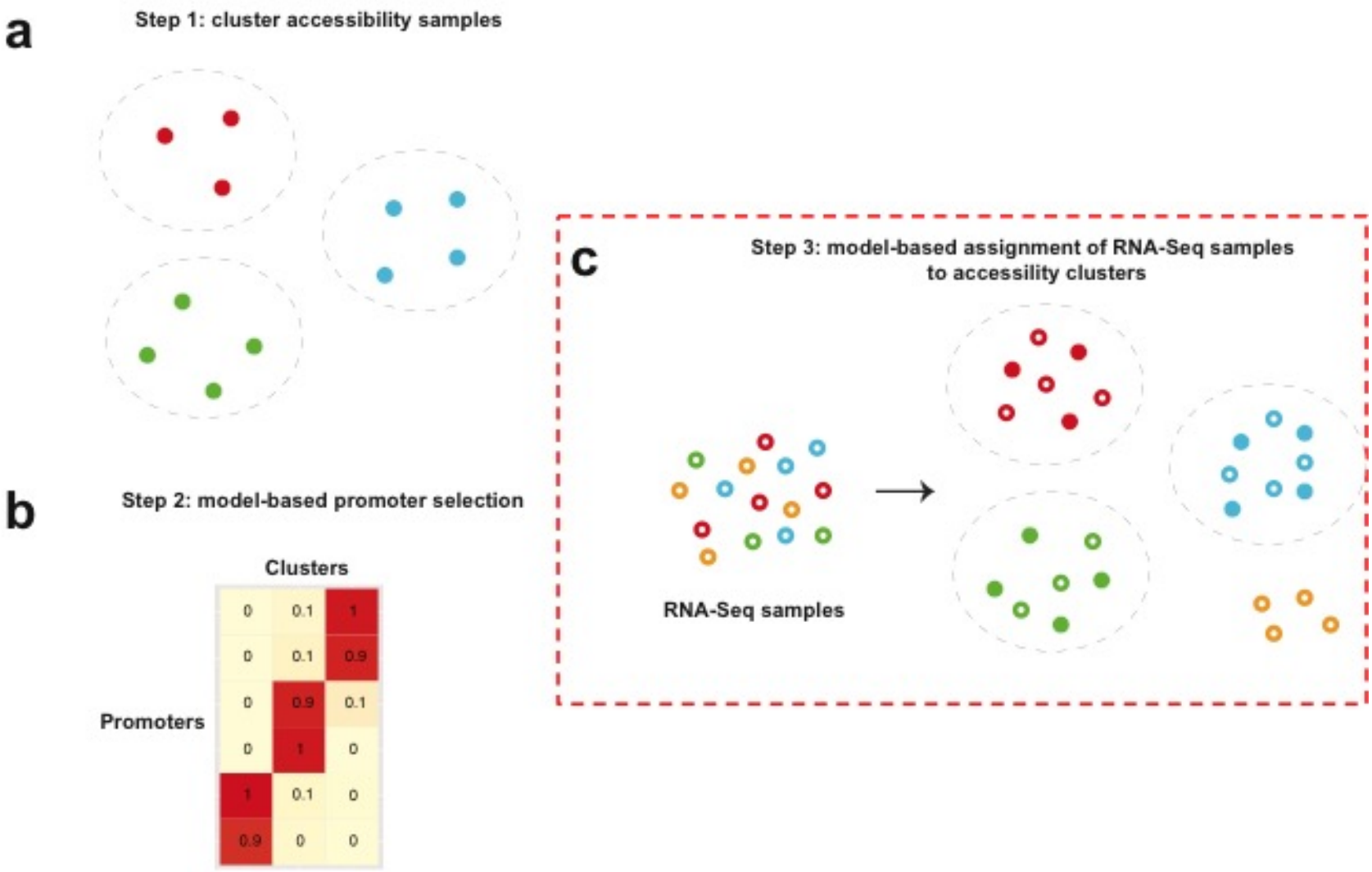

a
Step 1: cluster accessibility samples
b
Step 2: model-based promoter selection
Clusters
Promoters
0 0.1 1
0 0.1 0.9
0 0.9 0.1
0 1 0
1 0.1 0
0.9 0 0
c
Step 3: model-based assignment of RNA-Seq samples to accessility clusters
RNA-Seq samples

Step 2: model-based promoter selection

- Summarize promoter activity in each cluster
  - Combine information over all samples within a cluster
  - Infer promoter open/closed probability by Bayesian hierarchical model
- Select a subset of informative promoters

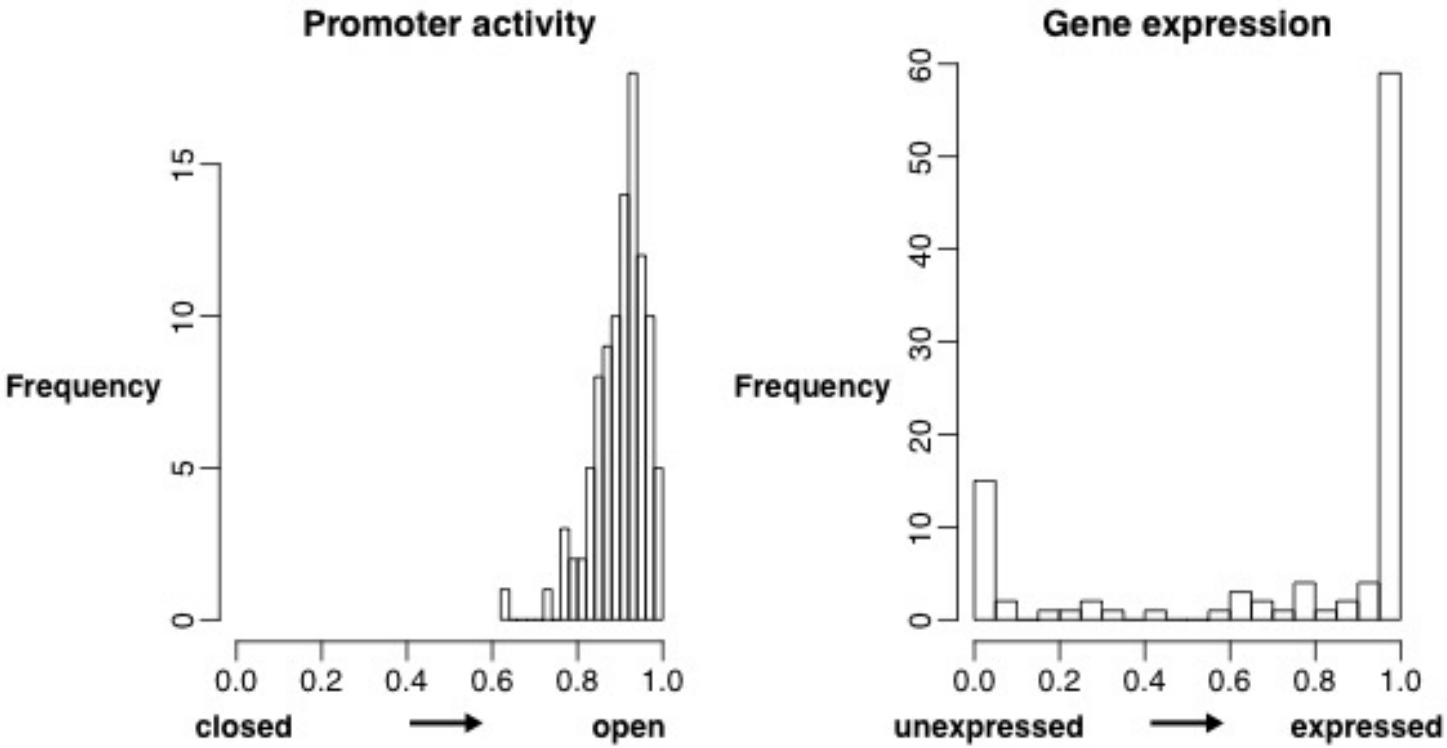




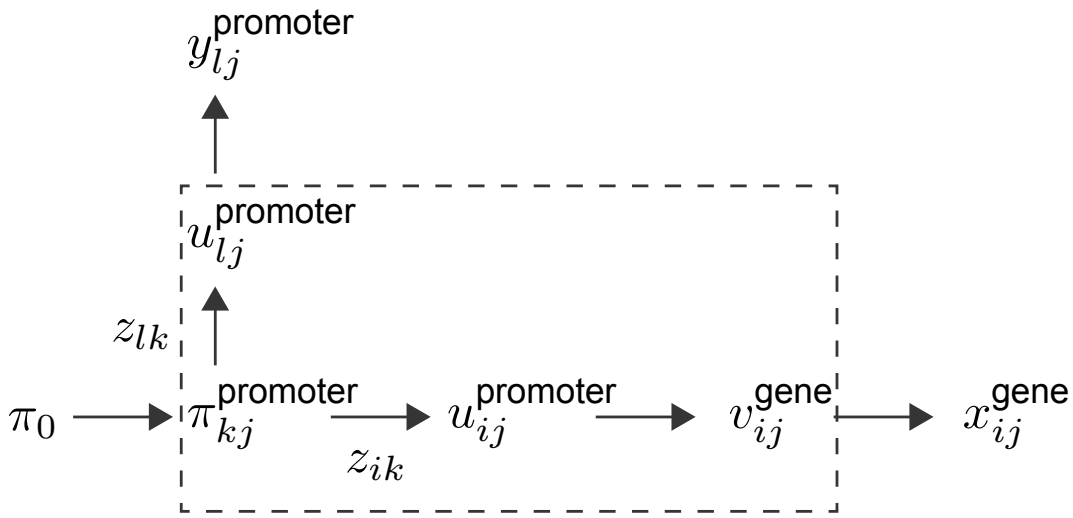

a
Step 1: cluster accessibility samples
b
Step 2: model-based promoter selection
Clusters
Promoters
0 0.1 1
0 0.1 0.9
0 0.9 0.1
0 1 0
1 0.1 0
0.9 0 0
c
Step 3: model-based assignment of RNA-Seq samples to accessility clusters
RNA-Seq samples

Step 3: cluster RNA-Seq samples

- Assign scRNA-Seq samples to scATAC-Seq clusters
  - Use the promoter-gene pair to link these two data types

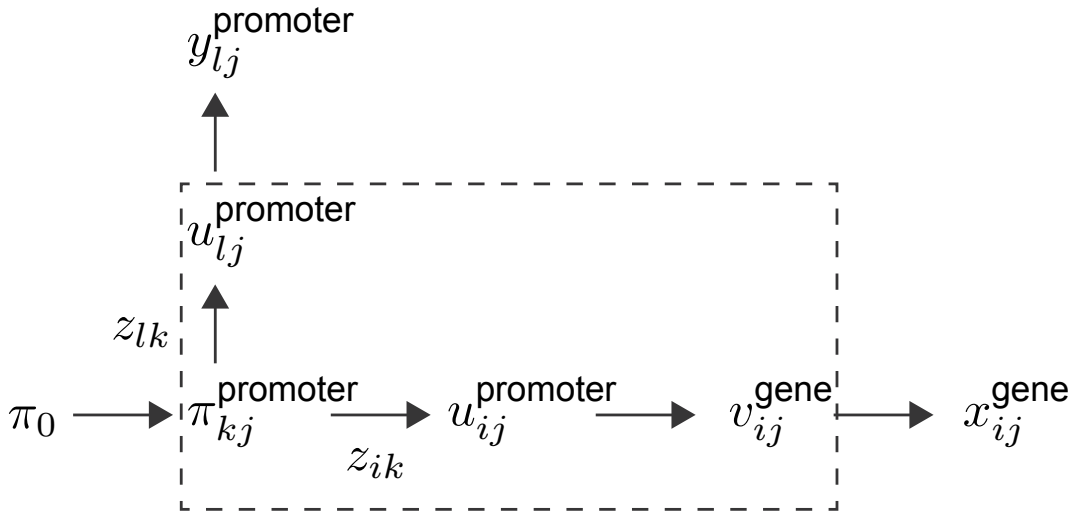


Probabilistic model for stochasticity



- Samples: $I = 1, \cdots, N^{Acc}$, $i = 1, \cdots, N^{Rna}$
- Features: $j = 1, \cdots, R$
- Clusters: $k = 1, \cdots, K$
- Z is the cluster assignment, X and Y are the observed data
- U and V represent the mixture component

Key of the model (1)



- Sample specific model parameter $\{\pi_{i1}, \pi_{i0}\}_{i=1,\dots,N^{Rna}}$

$$p(x_{ij}^{\text{gene}} \mid v_{ij}^{\text{gene}}) = v_{ij}^{\text{gene}} f_1(x_{ij}^{\text{gene}}) + (1 - v_{ij}^{\text{gene}}) f_0(x_{ij}^{\text{gene}}),$$

$$v_{ij}^{\text{gene}} \mid u_{ij}^{\text{promoter}} = 1 \sim \text{Bernoulli}(\pi_{i1}),$$

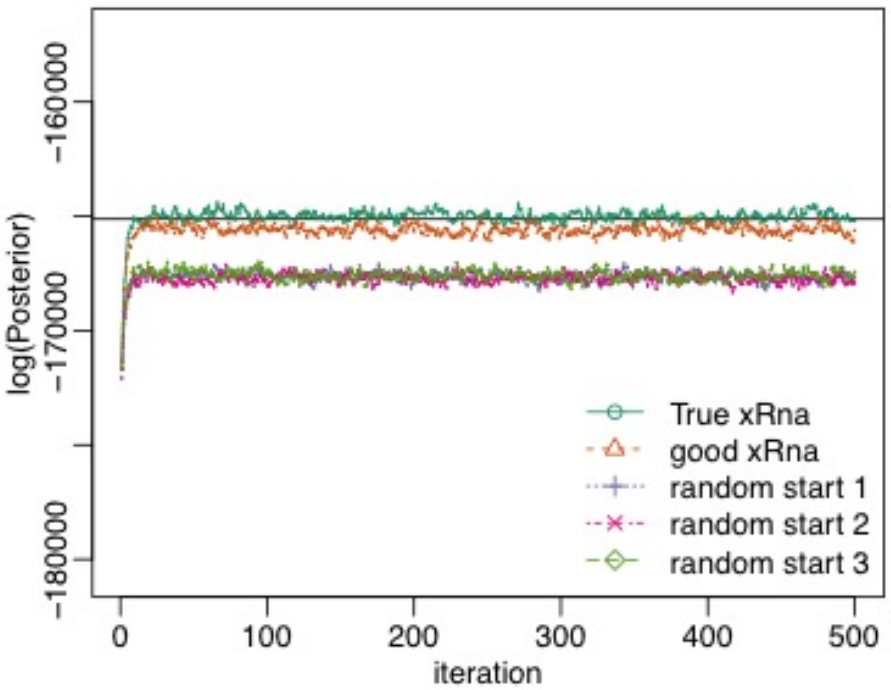
$$v_{ij}^{\text{gene}} \mid u_{ij}^{\text{promoter}} = 0 \sim \text{Bernoulli}(\pi_{i0}),$$

$$\pi_{i1} \mid \pi_{i0} \sim \text{Uniform}(\pi_{i0}, 1),$$

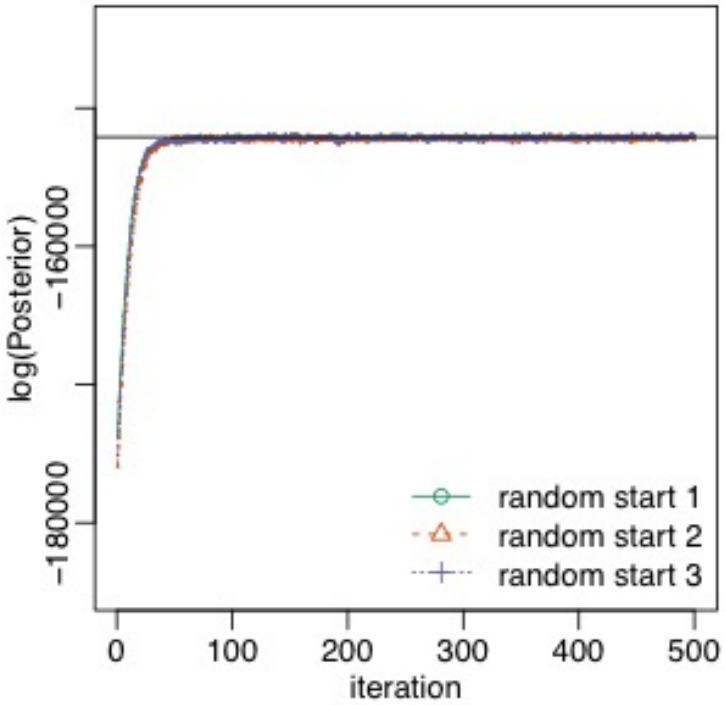
$$\pi_{i0} \sim \text{Uniform}(0, 1)$$

Key of the model (2)

- Integrate out U in the MCMC
  - The Gibbs sampler gets “sticky”



Before



After

Step 3: cluster RNA-Seq samples

- The scRNA-Seq samples are correctly matched to the scATAC-Seq clusters

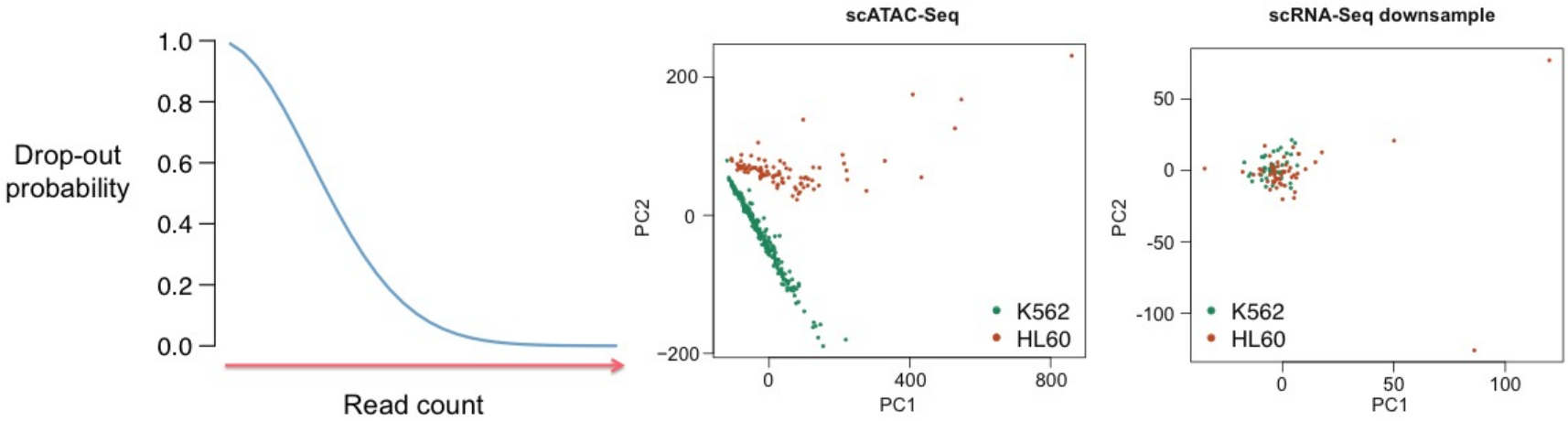
Step 1

		scATAC-Seq samples	
		K562	HL60
scATAC-Seq clusters	Cluster 1	223	0
	Cluster 2	0	91

Step 3

		scRNA-Seq samples	
		K562	HL60
scATAC-Seq clusters	Cluster 1	42	0
	Cluster 2	0	54

Downsample the reads in scRNA-Seq data



- $\sim 30,000$ reads, and ~ 150 non-zero genes per cell

scATAC-Seq helps separate scRNA-Seq samples

Step 1

		scATAC-Seq samples	
		K562	HL60
scATAC-Seq clusters	Cluster 1	223	0
	Cluster 2	0	91

Step 3

		scRNA-Seq downsample	
		K562	HL60
scATAC-Seq clusters	Cluster 1	30	8
	Cluster 2	12	46

- Our method enables statistical inference on the clustering result
  - Accuracy increases from 79%(76/96) to 90%(54/60) after removing the noisy cells

Step 3

		scRNA-Seq downsample(0.65)	
		K562	HL60
scATAC-Seq clusters	Cluster 1	21	3
	Cluster 2	3	33

Summary

- Our method matches the cell sub-populations in scRNA-Seq and scATAC-Seq data
- In single cell and bulk data, chromatin accessibility can guide the clustering of RNA-Seq samples, when RNA-Seq data alone cannot distinguish the cell types
- Model-based method can improve the clustering result by not clustering the “problematic” cells

References I


Wang Z et al.
RNA-Seq: a revolutionary tool for transcriptomics.
Nature Review Genetics, 2009.


Zeisel A et al.
Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.
Science, 2015.


Kang HJ et al.
Spatio-temporal transcriptome of the human brain.
Nature, 2011.


Johnson WE et al.
Adjusting batch effects in microarray expression data using Empirical Bayes methods.
Biostatistics, 2007.


Leek JT et al.
Identifying and estimating surrogate variables for unknown sources of variation in high-throughput experiments.
PLoS Genetics, 2007.

References II


Gagnon-Bartsch JA et al.
Using control genes to adjust for unwanted variation in microarray data.
Biostatistics, 2012.


Gagnon-Bartsch JA et al.
Removing Unwanted Variation from High Dimensional Data with Negative Controls.
IMS Monographs series (Accepted). Cambridge University Press.


Risso D et al.
Normalization of RNA-seq data using factor analysis of control genes or samples.
Nature Biotechnology, 2014.


Jacob L et al.
Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed.
Biostatistics, 2016.


Zang C, Wang T et al.
High-dimensional genomic data bias correction and data integration using MANCIE.
Nature Communications, 2016.

References III


Celniker SE et al.
Unlocking the secrets of the genome.
Nature, 2009.


Gerstein MB et al.
Comparative analysis of the transcriptome across distant species.
Nature, 2014.


Witten DM et al.
A penalized matrix decomposition, with applications to sparse principal components and canonical correlation.
Biostatistics, 2009.


Besag JE et al.
On the Statistical Analysis of Dirty Pictures.
Journal of the Royal Statistical Society, Series B, 1986.

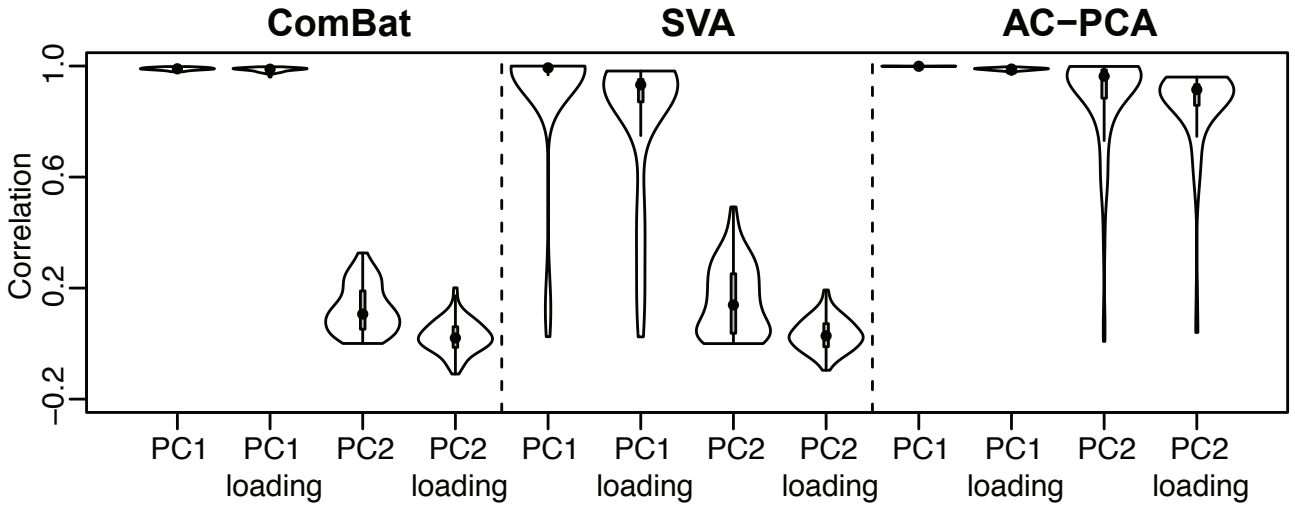
Thank you!

Conclusions and future directions

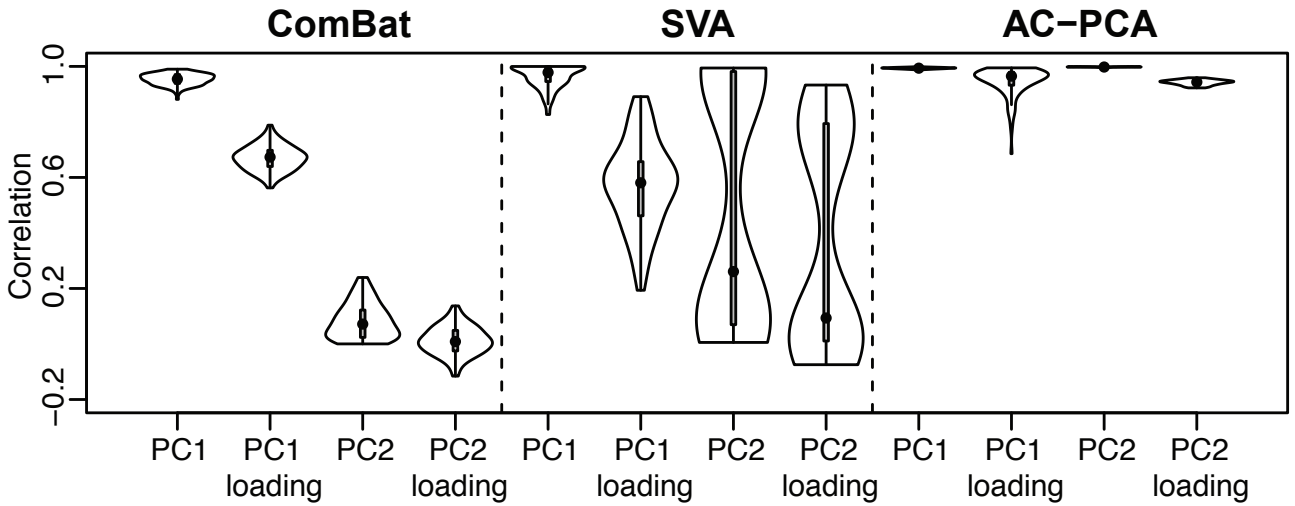
- We have proposed a general class of penalty functions in PCA for simultaneous dimension reduction and adjustment for confounding variation
- We have demonstrated its performance through real data analysis and simulations
- An extension: $K(Y, Y)$ in place of YY^T
- Other applications
  - Discrete data (binary and count data)
  - Clustering
  - Meta-analysis, integrate multiple types of heterogeneous data
- Theoretical properties

Simulations - result (2)

Setting 1



Setting 2

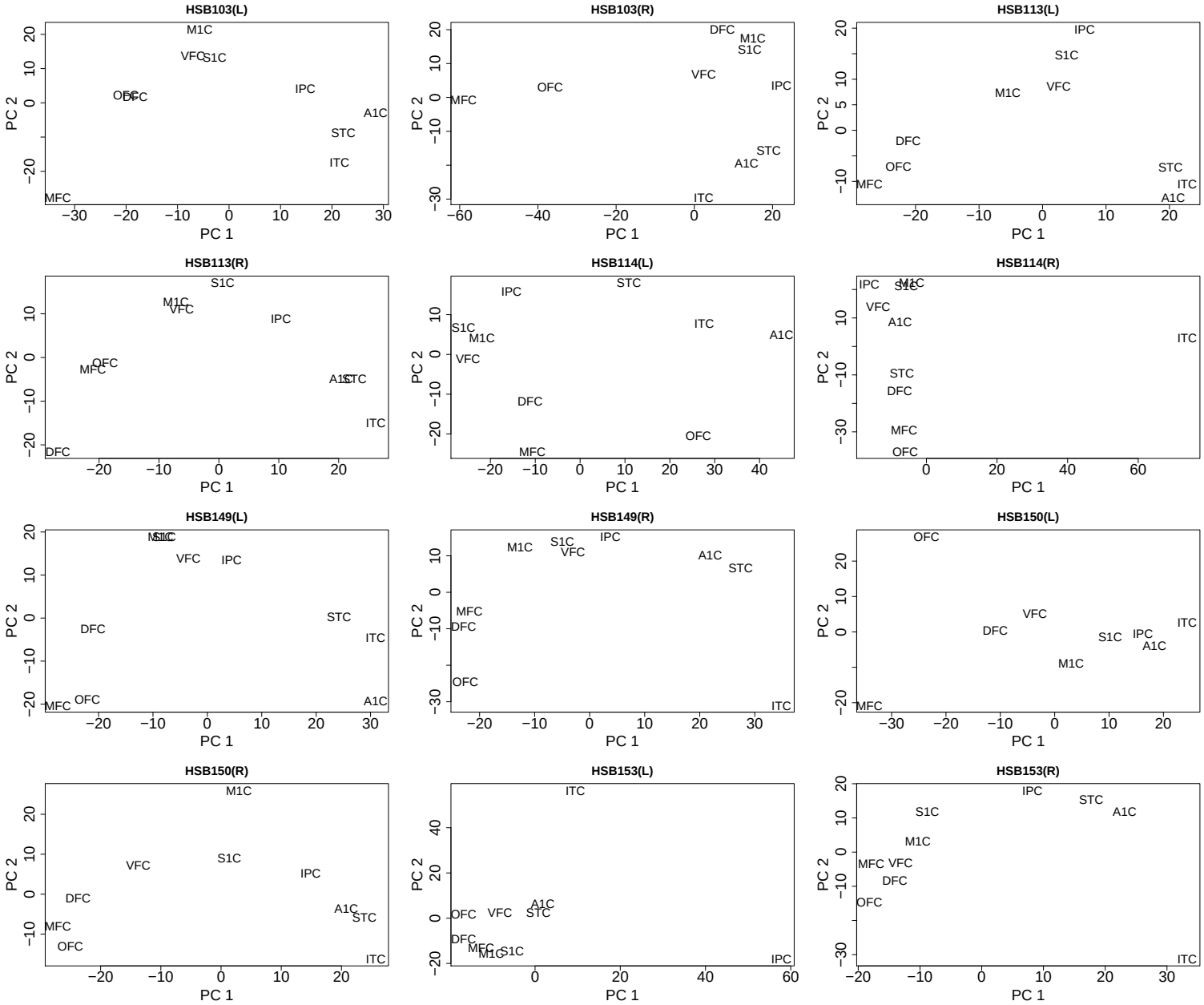


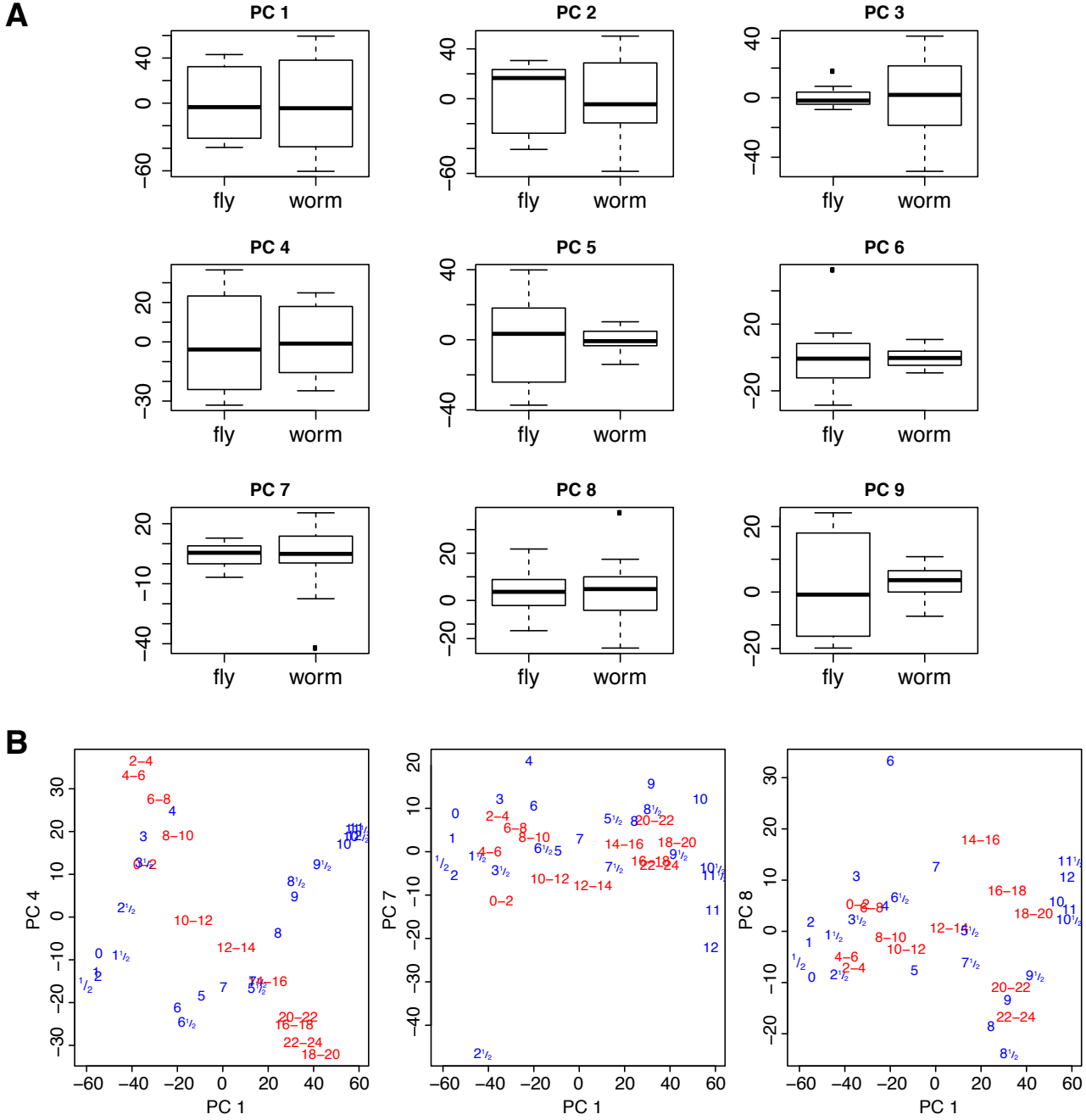
Simulation

Table S3. Sparsity estimation and sensitivity

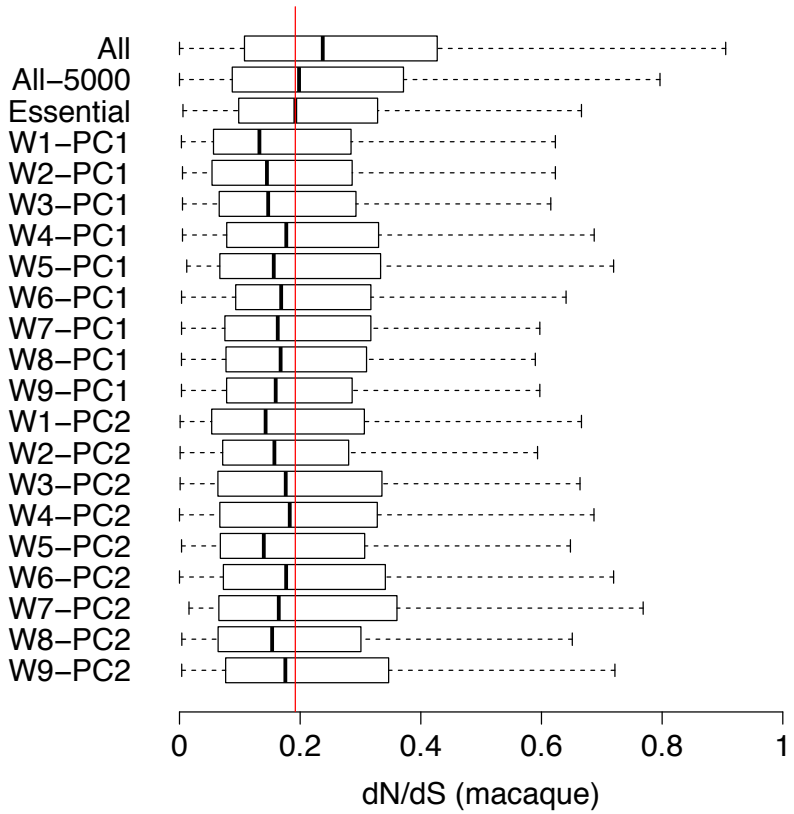
		Estimated non-zeros		Sensitivity‡	
		$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 0.2$	$\sigma = 0.5$
100*	$\alpha = 1.5$	113.7(34.2)	120.6(42.5)	0.85(0.06)	0.80(0.07)
	$\alpha = 2.0$	88.7(20.9))	91.6(31.1)	0.83(0.08)	0.79(0.08)
	$\alpha = 2.5$	81.4(20.5)	79.4(24.4)	0.83(0.08)	0.80(0.05)
40*	$\alpha = 2.0$	37.4(14.7)	38.9(15.6)	0.74(0.11)	0.62(0.10)
	$\alpha = 2.0^{\dagger}$	46.7(16.5)	54.6(19.0)	0.77(0.10)	0.69(0.11)

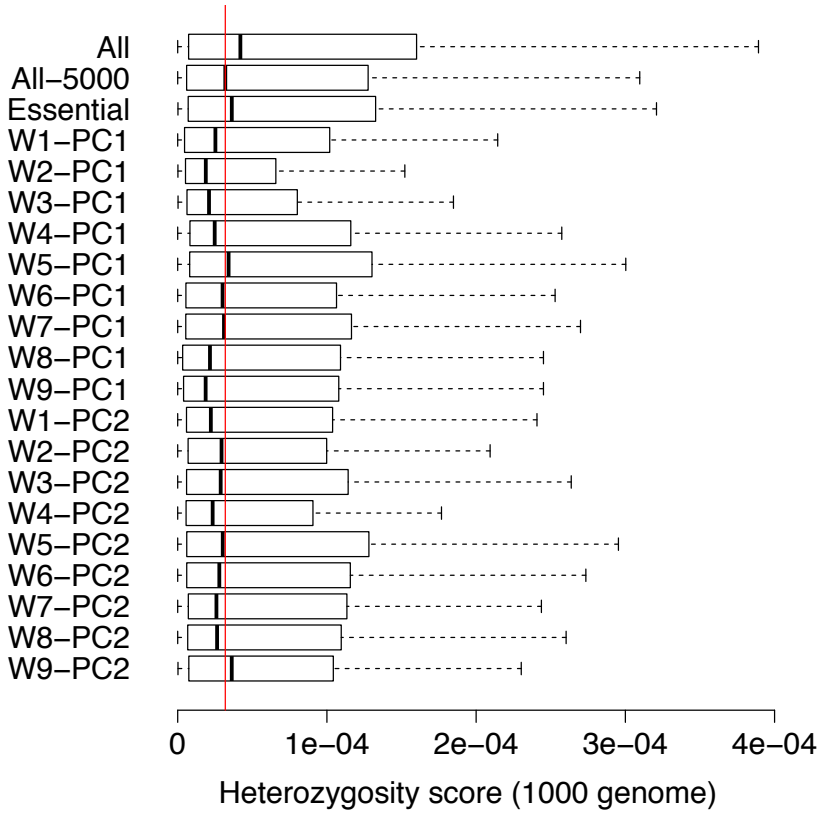
- $n = 5$, $b = 10$ and $p = 400$
- α, strength of the confounding variation
- σ, the noise level


HSB103(L)
HSB103(R)
HSB113(L)
HSB113(R)
HSB114(L)
HSB114(R)
HSB149(L)
HSB149(R)
HSB150(L)
HSB150(R)
HSB153(L)
HSB153(R)
PC 1
PC 2


A
PC 1
PC 2
PC 3
PC 4
PC 5
PC 6
PC 7
PC 8
PC 9
fly
worm
B
PC 4
PC 7
PC 8
PC 1

Application to the human brain exon array data





- The top 200 selected genes tend to be functionally conserved

Two-step approach to tune c_1 and c_2

To tune c_1:

1. Use the non-sparse version to tune λ and calculate v
2. Set $c_1 = v^T X^T Y Y^T X v$

c_2 is tuned based on matrix completion (Witten 2009):

1. From X, we construct 10 data matrices $X_1, \cdots, X_{10}$, each of which is missing a non-overlapping one-tenth of the elements of X
2. For $X_1, \cdots, X_{10}$, fit formula (4) and obtain $\hat{X}_i = duv^T$, the resulting estimate of X_i and $d = u^T X_i v$
3. Calculate the mean squared errors of $\hat{X}_i$, for $i = 1, \cdots, 10$, using only the missing entries
4. Choose c_2 that minimizes the sum of mean squared errors

Differential Expression

Let $\mathbf{y}_{btg}$ denote the vectors of expression values for gene g in region b and time t. The two-sample t-statistic :

$$t_{btg} = \frac{\bar{\mathbf{y}}_{b(t+1)g} - \bar{\mathbf{y}}_{btg}}{s},$$

The test statistic t_{btg} is then transformed into z_{btg}:

$$z_{btg} = \Phi^{-1}(F_{n_{b(t+1)}+n_{bt}-2}(t_{btg})),$$

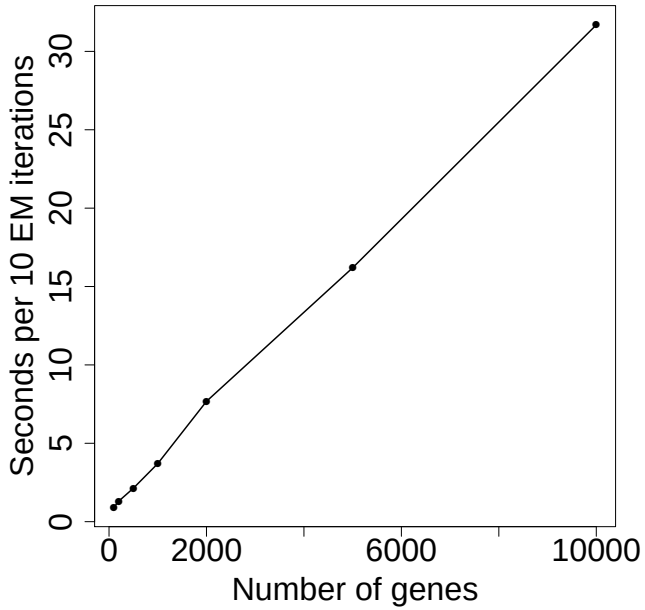
Introduce binary latent variable γ_{btg}:

$$f(z_{btg} \mid \gamma_{btg}) = (1 - \gamma_{btg}) f_0(z_{btg}) + \gamma_{btg} f_1(z_{btg}),$$

Use the nonparametric empirical Bayesian framework (*locfdr*) to estimate the densities (Efron B., 2004)

EM algorithm with mean field-like approximation

- The key idea is that the fluctuation of the neighbors of a node is ignored by fixing the neighbors to some configuration (Zhang, 1992; Celeux et al., 2003)
- The configuration is also updated iteratively
- Much faster



Gaussian Graphical Model (GGM)

$$X = (X_v, v \in V) \sim \mathcal{N}(\mu, \Sigma) \tag{5}$$

- The inverse covariance matrix $\Theta = \Sigma^{-1}$
- $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}} \iff \Theta_{ij} = 0$
- Estimate the conditional independent graph by estimating Θ

Some properties of Gaussian Distribution

Consider the p-dimensional random variable $X \sim \mathcal{N}(0, \Sigma)$.
Let the $n \times p$ matrix $\mathbf{X}$ contain n independent observations of X.
Let $\mathbf{X}_1$ denote the $1st$ column and $\mathbf{X}_2$ denote the sub-matrix of $\mathbf{X}$ excluding the $1st$ column. The following holds:

$$\mathbf{X}_1 \mid \mathbf{X}_2 \sim \mathcal{N}(-\mathbf{X}_2 \Theta_{12}^T \Theta_{11}^{-1}, \Theta_{11}^{-1} \boldsymbol{I}),$$
$$\text{rewrite as } \mathbf{X}_1 \mid \mathbf{X}_2 \sim \mathcal{N}(\mathbf{X}_2 \beta, \sigma^2 \boldsymbol{I}),$$

Lasso for neighborhood selection (Bühlmann et al. 2006)

Bayesian neighborhood selection procedure

Consider estimating the neighborhood of the $1st$ node. Spike and Slab prior on $\boldsymbol{\beta}$:

$$\beta_j \mid \gamma_j, \tau_j \sim (1-\gamma_j)\mathcal{N}(0, c^2\tau_j^2) + \gamma_j\mathcal{N}(0, \tau_j^2),$$

where $0 < c < 1$, $\boldsymbol{\tau}$ are prefixed. The binary latent vector $\boldsymbol{\gamma}$ is the goal of inference.

- For single graph, bernoulli priors on $\boldsymbol{\gamma}$
- For multiple graphs, MRF priors on $\boldsymbol{\gamma}$
- Use MCMC for statistical inference

Posterior sampling procedure

Updating β:

$$\beta \mid . \sim \mathcal{N}((\boldsymbol{X}_2'\boldsymbol{X}_2 + D)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1, \sigma^2(\boldsymbol{X}_2'\boldsymbol{X}_2 + D)^{-1}). \tag{6}$$

where

$$D_{jj} = \begin{cases} \sigma_i^2/\tau_j^2, & \text{if } \gamma_j = 1, \\ \sigma_i^2/(c^2\tau_j^2), & \text{if } \gamma_j = 0. \end{cases}$$

The most computationally intensive step, involves the inversion of $p \times p$ matrix.

Some tricks for faster computation

$$\beta \mid . \sim \mathcal{N}((\boldsymbol{X}_2'\boldsymbol{X}_2 + D)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1, \sigma^2(\boldsymbol{X}_2'\boldsymbol{X}_2 + D)^{-1}),$$

1. Cholesky decomposition: $\boldsymbol{X}_2'\boldsymbol{X}_2 + D = R'R$
2. sample Z from $\mathcal{N}(0, I)$
3. $(\boldsymbol{X}_2'\boldsymbol{X}_2 + D_i)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1 + \sigma R^{-1}Z = R^{-1}((R^{-1})'\boldsymbol{X}_2'\boldsymbol{X}_1 + \sigma Z)$ can be solved by forward and backward substitution.

Can do parallel computing when updating β.

Theoretical properties for single graph estimation

Joint work with Tao Wang.

Sparse Riesz condition

For $A \subseteq \{1, \ldots, p\}$, define $\boldsymbol{X}_A = (\boldsymbol{X}_j, j \in A)$. Let $1 \leq p^* \leq p$. Throughout, we assume that $\boldsymbol{X}$ satisfies the sparse Riesz condition (Zhang and Huang 2008) with rank p^*; that is, there exist some constants $0 < c_1 < c_2 < \infty$ such that

$$c_1 \leq \frac{\|\boldsymbol{X}_A u\|^2}{n\|u\|^2} \leq c_2$$

for any $A \subseteq \{1, \ldots, p\}$ with size $|A| = p^*$ and any nonzero vector $u \in \mathbb{R}^{p^*}$.

Notations

(Follows Narisetty et al. 2014) Consider estimating the neighborhood for the ith node. Let the binary vector k^i index an arbitrary model. For $v > 0$, define

$$m(v) \equiv m_n(v) = (p-1) \wedge \frac{n}{(2+v)\log(p-1)}$$

and

$$\lambda_{m,i}(v) = \min_{k^i : |k^i| \leq m(v)} \lambda_{\min} \left(\frac{\boldsymbol{X}'_{k^i} \boldsymbol{X}_{k^i}}{n} \right).$$

For $K > 0$, let

$$\Delta_i(K) = \min_{\{k^i : |k^i| \leq K|t^i|, k^i \not\supset t^i\}} \|(I - P_{k^i}) \boldsymbol{X}_{t^i} \beta_{t^i}\|_2^2,$$

where $|k^i|$ denotes the size of the model k^i and P_{k^i} is the projection matrix onto the column space of $\boldsymbol{X}_{k^i}$.

Conditions

(A) $p \to \infty$ and $p = O(n^\theta)$ for some $\theta > 0$;

(B) $q = p^{\alpha - 1}$ for some $0 \le \alpha < 1 \wedge (1/\theta)$;

(C) $n\tau_0^2 = o(1)$ and $n\tau_1^2 \sim n \vee p^{2+2\delta_1}$ for some $\delta_1 > 1 + \alpha$;

(D) $|t^i| \prec n/\log p$ and $\|\beta_{t^i}\|_2^2 \prec \tau_1^2 \log p$;

(E) there exist $1 + \alpha < \delta_2 < \delta_1$ and $K > 1 + 8/(\delta_2 - 1 - \alpha)$ such that, for some large $C > 0$, $\Delta_i(K) > C\sigma^2 \log(n \vee p^{2+2\delta_1})|t^i|$;

(F) $p^* \ge (K+1)|t^i|$;

(G) $\lambda_{\max}(\boldsymbol{X}'\boldsymbol{X}/n) \prec (n\tau_0^2)^{-1} \wedge (n\tau_1^2)$ and there exist some $0 < v < \delta_2$ and $0 < \kappa < 2(K-1)$ such that

$$\lambda_{m,i}(v) \succeq \frac{(n \vee p^{2+2\delta_2})}{n\tau_1^2} \vee p^{-\kappa}.$$

Theorem 1

Assume conditions (A)-(G). For some $c > 0$ *and* $s > 1$ *we have, with probability at least* $1 - cp^{-s}$, $P(\gamma = t^i | \boldsymbol{X}, \sigma^2) > 1 - r_n$, *where* r_n *goes to 0 as the sample size increases to* ∞.

Conditions (con't)

To establish graph-selection consistency, we need slightly stronger conditions than (D)-(G). Let

$$t^* = \max_{1\le i\le p} |t^i|, \Delta^*(K) = \min_{1\le i\le p} \Delta_i(K) \text{ and } \lambda^*_m(v) = \min_{1\le i\le p} \lambda_{m,i}(v).$$

(D') $t^* \prec n/\log p$ and $\max_{1\le i\le p} \|\beta_{t^i}\|_2^2 \prec \tau_1^2 \log p$;

(E') there exist $1+\alpha < \delta_2 < \delta_1$ and $K > 1 + 8/(\delta_2 - 1 - \alpha)$ such that, for some large $C > 0$, $\Delta^*(K) > C\sigma^2 \log(n \vee p^{2+2\delta_1}) t^*$;

(F') $p^* \ge (K+1)t^*$;

(G') $\lambda_{\max}(\boldsymbol{X}'\boldsymbol{X}/n) \prec (n\tau_0^2)^{-1} \wedge (n\tau_1^2)$ and there exist some $0 < v < \delta_2$ and $0 < \kappa < 2(K-1)$ such that

$$\lambda^*_m(v) \succeq \frac{(n \vee p^{2+2\delta_2})}{n\tau_1^2} \vee p^{-\kappa}.$$

Theorem 2

Assume conditions (A)-(C) and (D')-(G'). We have, as $n \to \infty$, $P\{\hat{\mathcal{G}} = \mathcal{G}\} \to 1$.

Affymetrix GeneChip microarray

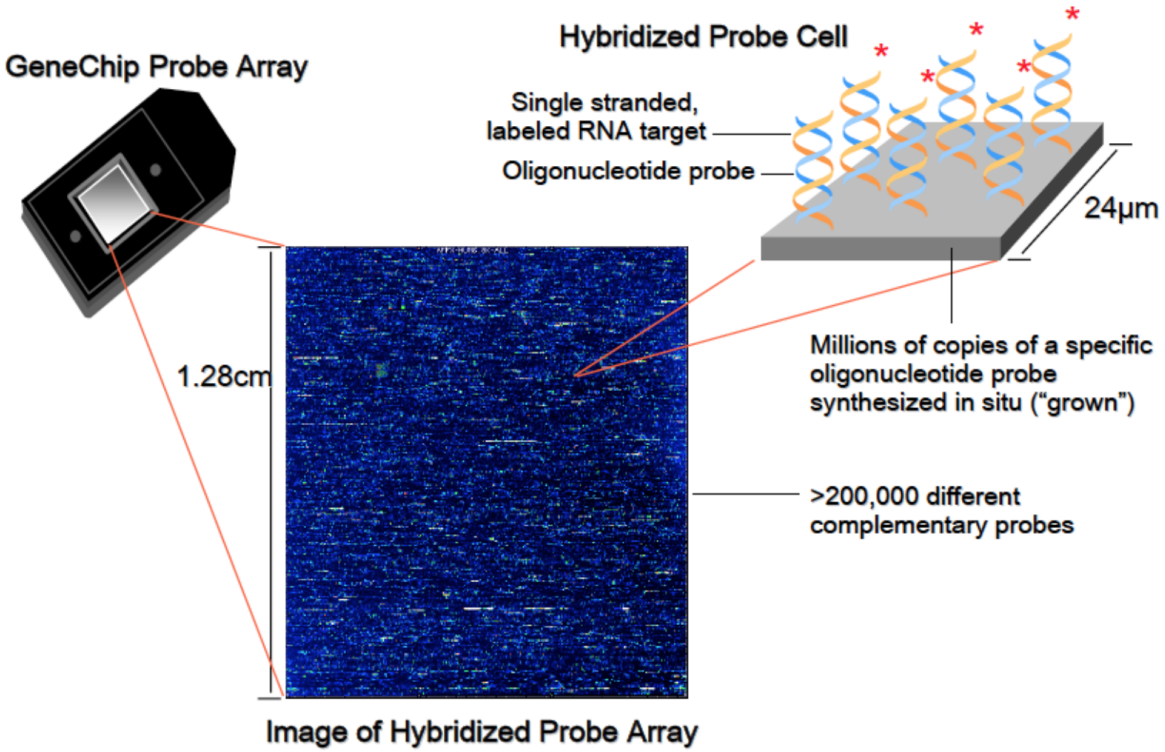


Figure: Adapted from http://www.affymetrix.com/

- Expression level is quantified by the fluorescent intensity

RNA-Seq

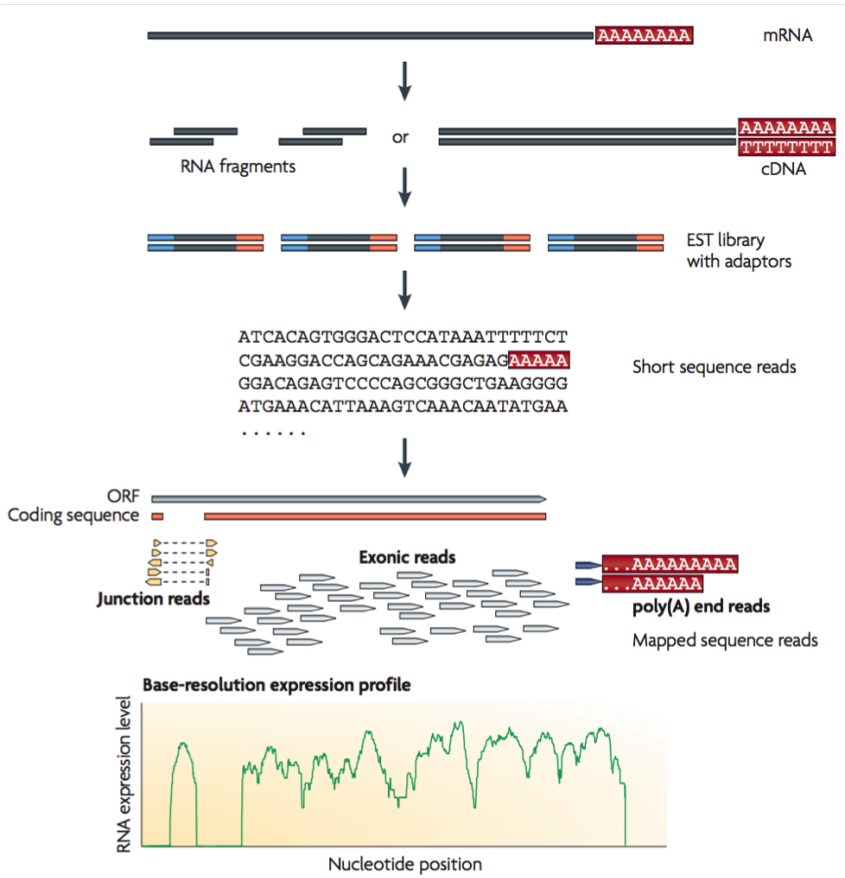
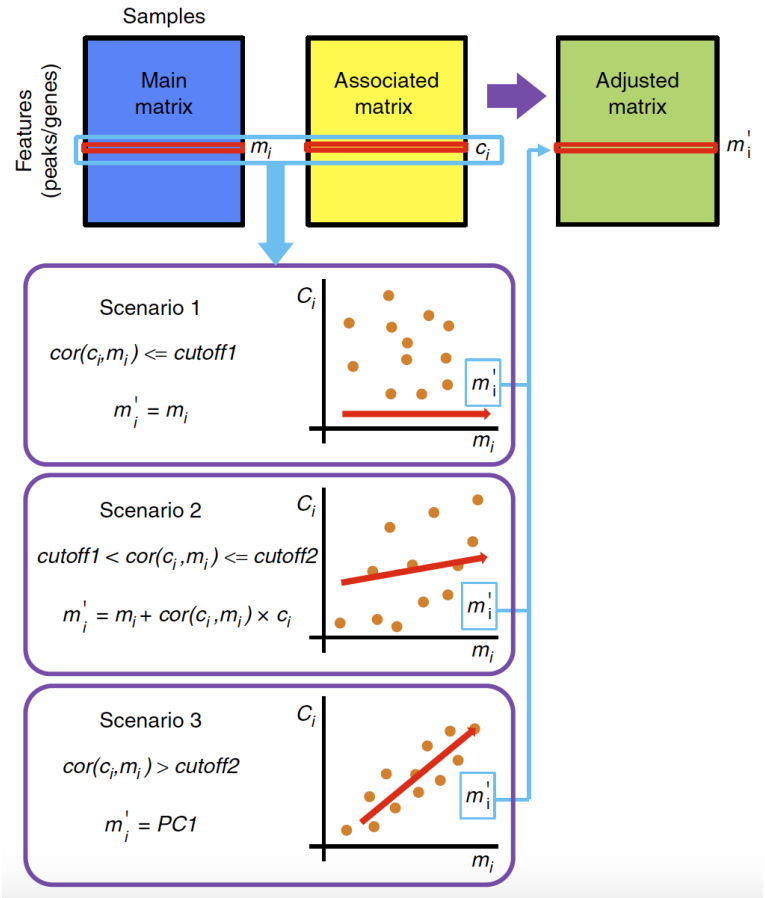


Figure: RNA-Seq technology (Wang Z et al. 2009)

- Expression level is quantified by the number of mapped reads

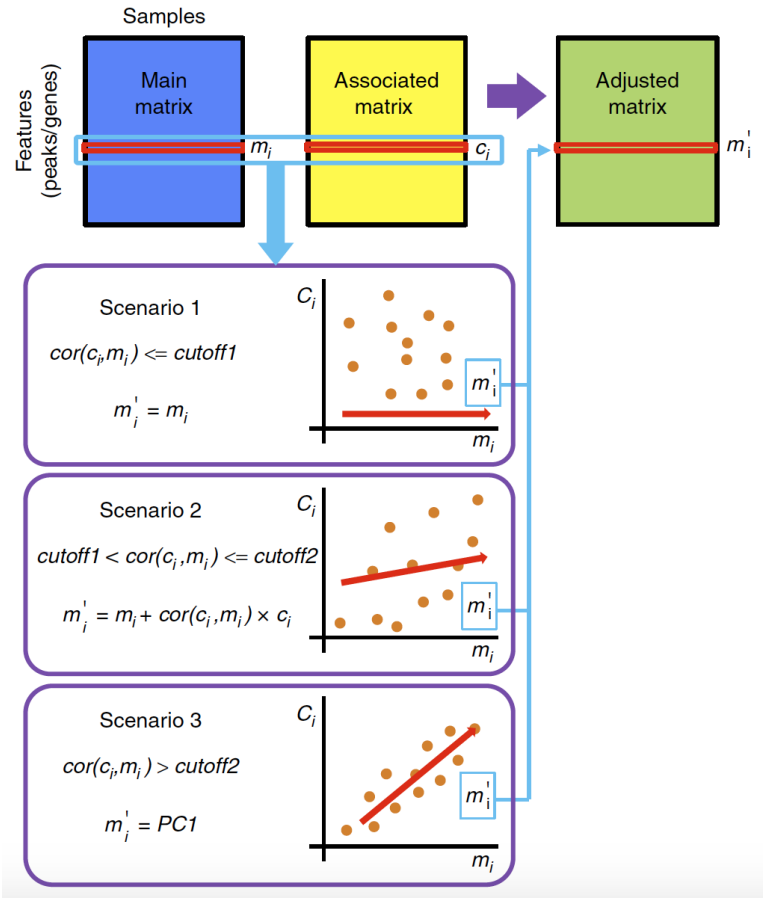
Methods for confounder adjustment (3)

Zang C, Wang T, Deng K et al. 2016



Methods for confounder adjustment (3)

Zang C, Wang T, Deng K et al. 2016



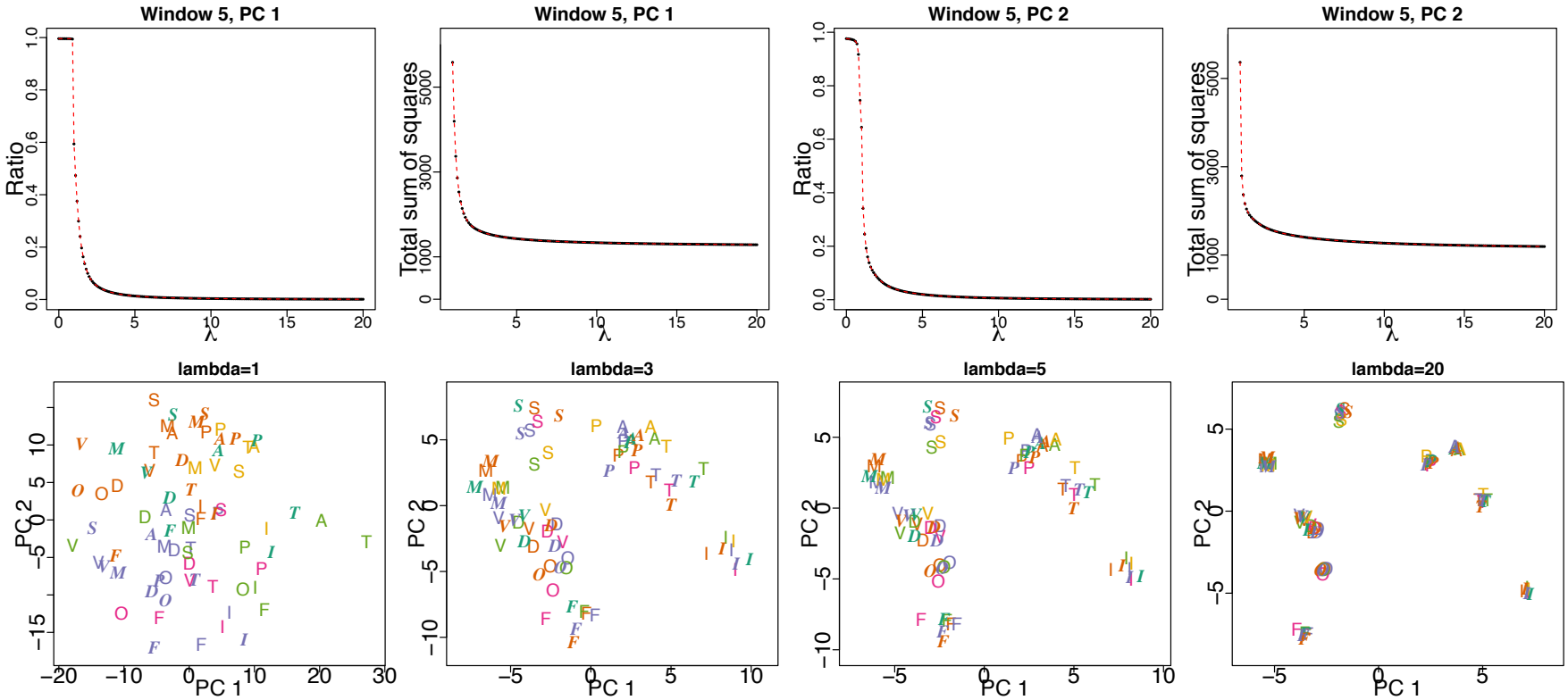
- Limited work has been done in the context of dimension reduction

Tuning λ

- The ratio $R(\lambda) = v_\lambda^T X^T Y Y^T X v_\lambda / v_\lambda^T X^T X v_\lambda$
  - ANOVA: smallest λ such that $R(\lambda) \leq \alpha$
  - Other: smallest λ such that $R(\lambda) \leq \alpha R(\lambda = 0)$

Tuning λ

- The overall pattern is robust to a wide range of λs



Thesis work

Developed statistical methods in a joint estimation framework to model spatial and temporal data structure, and to explore:

- Temporal dynamics of gene expression
  - Differential expression between two adjacent time periods
- Gene-gene interaction networks
  - Gaussian Graphical Model (GGM) setting to capture the conditional independence graphs

Introduce binary latent variable γ

- Differentially expressed/equally expressed
- Presence/absence of edges

MRF prior on γ

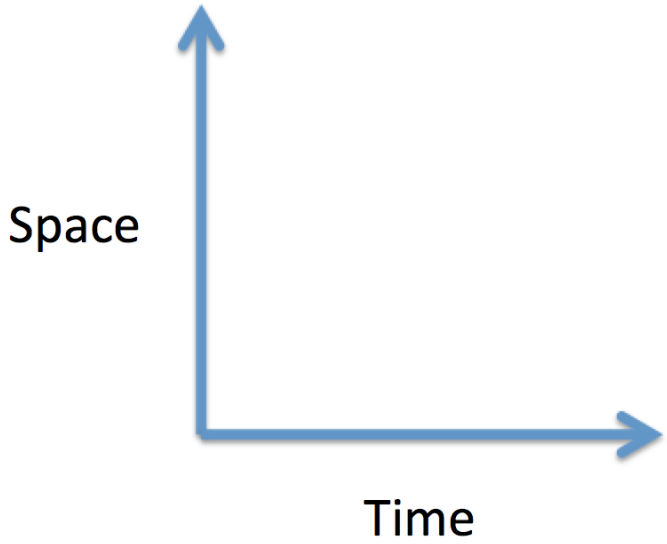
Denote $\gamma = \{\gamma_{btg} : \forall b \in B, \forall t \in (T-1), \forall g \in G\}$. Pairwise Markov Random Field (MRF) prior on γ (Besag J., 1986):

$$p(\gamma \mid \Phi) \propto \prod_{g} \exp\left\{ \alpha_0 \sum_{b,t}(1-\gamma_{btg}) + \alpha_1 \sum_{b,t} \gamma_{btg} + \sum_{e_g \in E_g} \beta(e_g)\left[\gamma_{btg}\gamma_{b't'g} + (1-\gamma_{btg})(1-\gamma_{b't'g})\right] \right\},$$

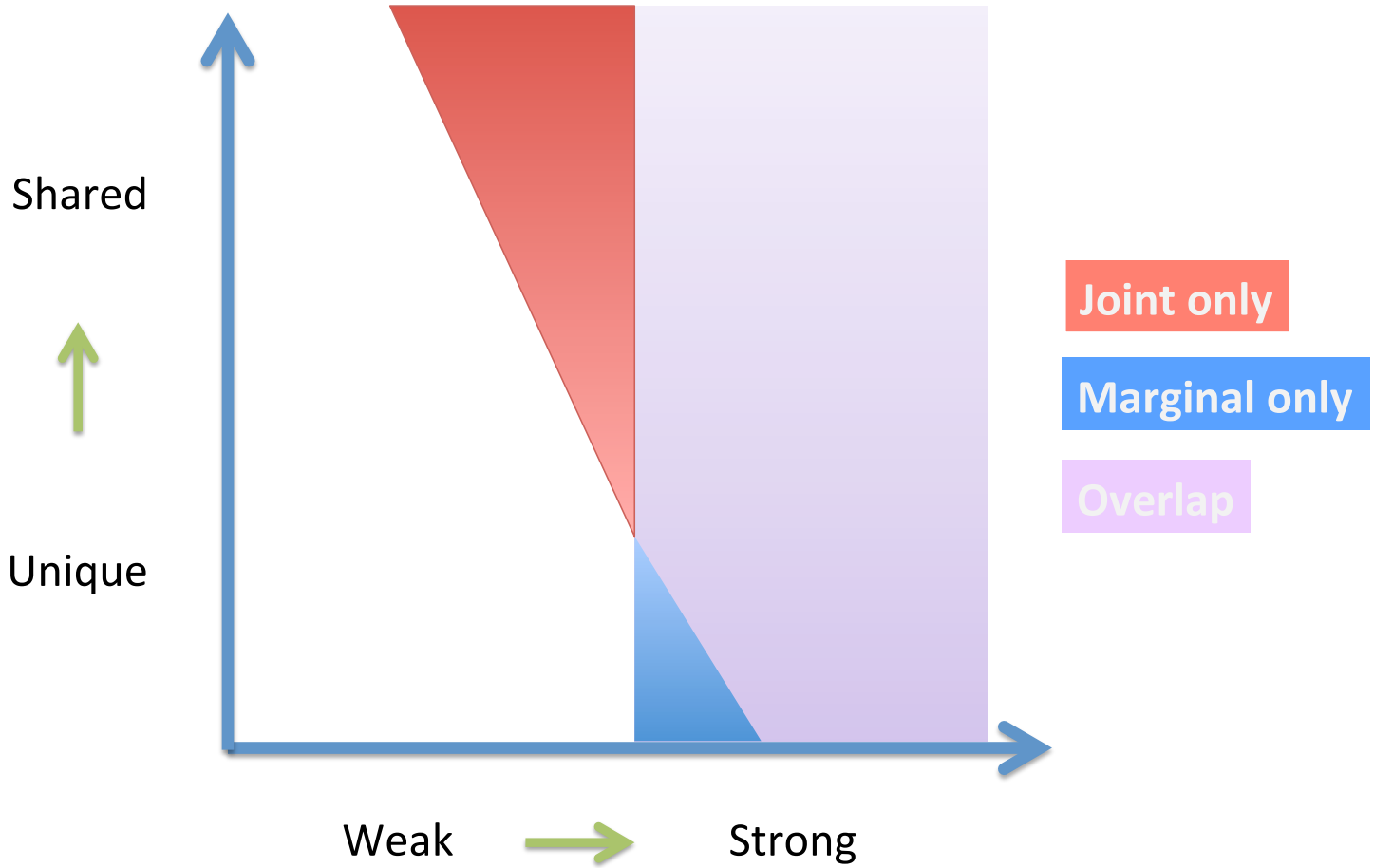
where $E_g = \{(\gamma_{btg}, \gamma_{b't'g}) : b \neq b',\ t = t' \text{ or } b = b',\ |t-t'| = 1\}$, and $\Phi = \{\alpha_1 - \alpha_0, \beta_t, \beta_{cc}, \beta_{cn}, \beta_{nn}\}$.

Intuition

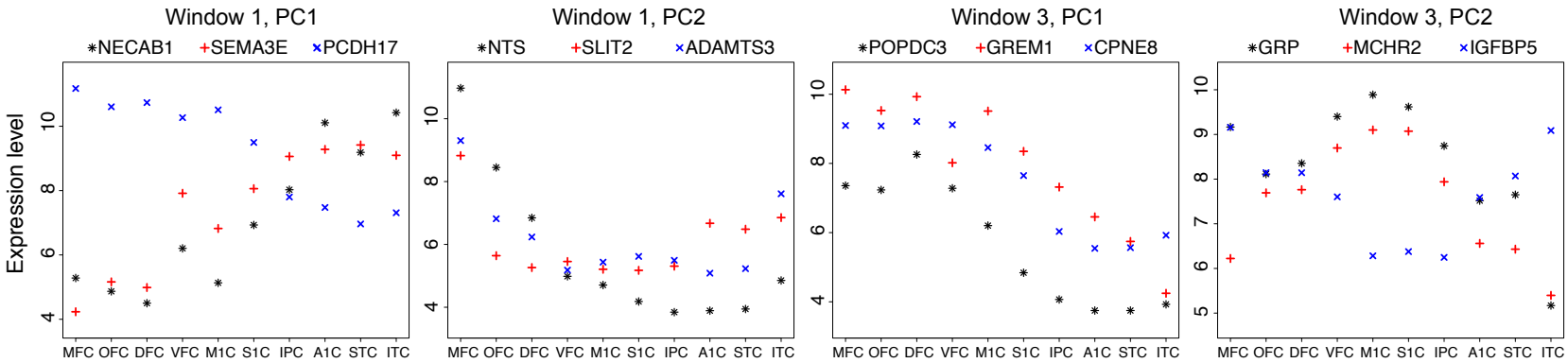
- Smoothing over γ



- The model parameters Φ are estimated from the data


Shared
Unique
Weak
Strong
Joint only
Marginal only
Overlap

Application to the human brain exon array data



- Through literature search, most genes are directly related to the developmental process

Detection of abnormal samples, brain data (con't)

- Predict a sample in $X_{(C)}$ by assigning it to the k-closest clusters of regions in $X_{(-C)}$
- The prediction accuracy:
  - Random guess, 10%($k = 1$), 20%($k = 2$), and 30%($k = 3$)
  - Leave-one-donor-out, 32%($k = 1$), 65%($k = 2$), and 80%($k = 3$)
  - Leave-one-sample-out, 43%($k = 1$), 67%($k = 2$) and 85%($k = 3$)

Detection of abnormal samples, brain data (con't)

- Predict a sample in $X_{(C)}$ by assigning it to the k-closest clusters of regions in $X_{(-C)}$
- The prediction accuracy:
  - Random guess, 10%($k = 1$), 20%($k = 2$), and 30%($k = 3$)
  - Leave-one-donor-out, 32%($k = 1$), 65%($k = 2$), and 80%($k = 3$)
  - Leave-one-sample-out, 43%($k = 1$), 67%($k = 2$) and 85%($k = 3$)
- Since the confounding effect of the left-out donor cannot induce any bias in the clustering of regions, this result shows that the penalty in AC-PCA enable the learning of dimension with significantly reduced confounder influence

Algorithm for sparse AC-PCA

The optimization problem is equivalent to (Witten et al. 2009)

$$\begin{aligned} \underset{u,v}{\text{minimize}} \quad & -u^T X v \\ \text{subject to} \quad & v^T X^T Y Y^T X v \leq c_1 \\ & ||v||_1 \leq c_2 \\ & ||v||_2^2 \leq 1 \\ & ||u||_2^2 \leq 1 \end{aligned}$$

- Biconvex in u and v
- Find local optimal by iteratively update between u and v

Updating u and v

- The update $u = \frac{Xv}{||Xv||_2}$
- To update v

$$\begin{aligned} \underset{v}{\text{minimize}} \quad & -u^T Xv \\ \text{subject to} \quad & v^T X^T Y Y^T Xv \leq c_1 \\ & ||v||_1 \leq c_2 \\ & ||v||_2^2 \leq 1 \end{aligned}$$

- Convex. Can be slow when v is large

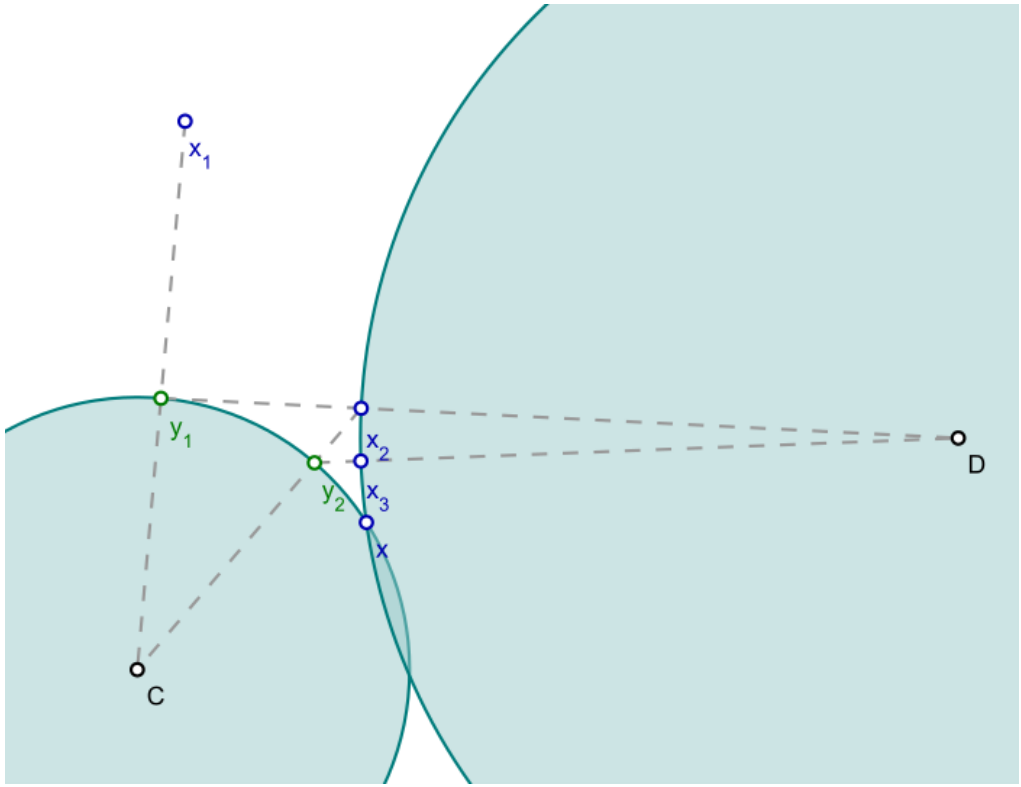
Updating v

Formulate it as a feasibility problem (joint work with John C. Duchi):

$$\begin{aligned} \text{find} \quad & v \\ \text{subject to} \quad & -u^T X v \le t \\ & v^T X^T Y Y^T X v \le c_1 \\ & ||v||_1 \le c_2 \\ & ||v||_2^2 \le 1 \end{aligned}$$

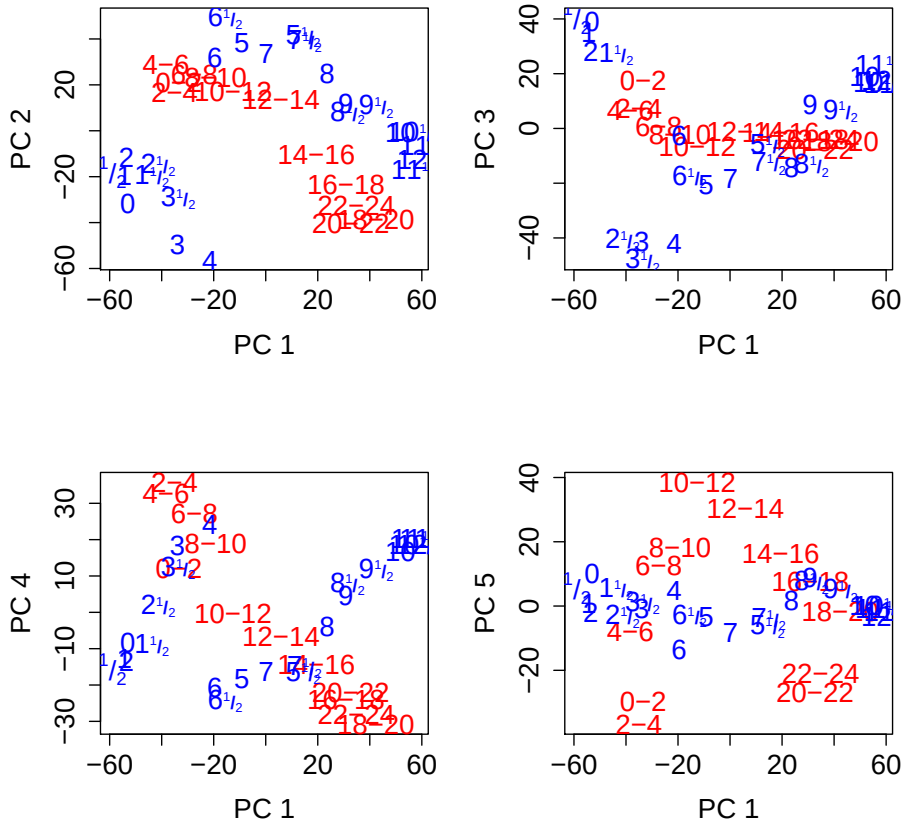
- Binary search on t
- Alternating projection to check feasibility
- $\sim$ 10-fold faster than cvx

Alternating projection



Adapted from Wikipedia

modENCODE RNA-Seq data, ComBat



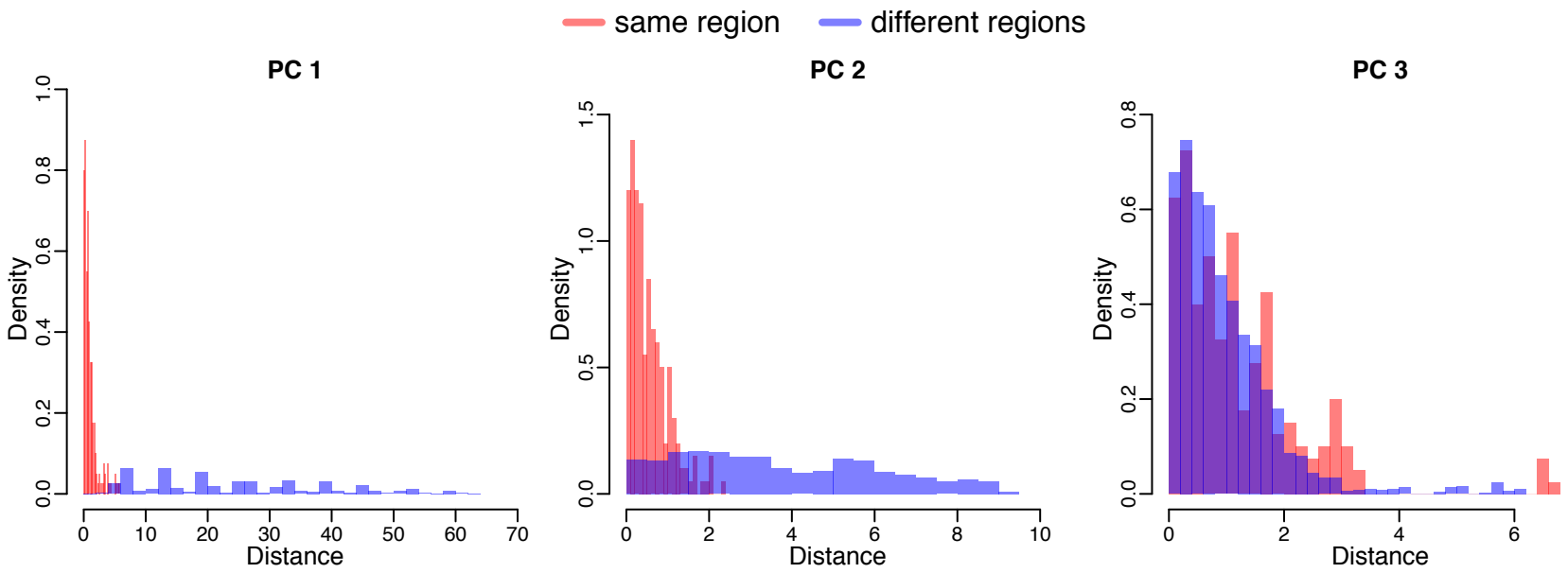
- ComBat using species label for the adjustment

Detection of abnormal samples, simulation

PCA has been used to detect abnormal samples. This can be achieved in AC-PCA with cross-validation:

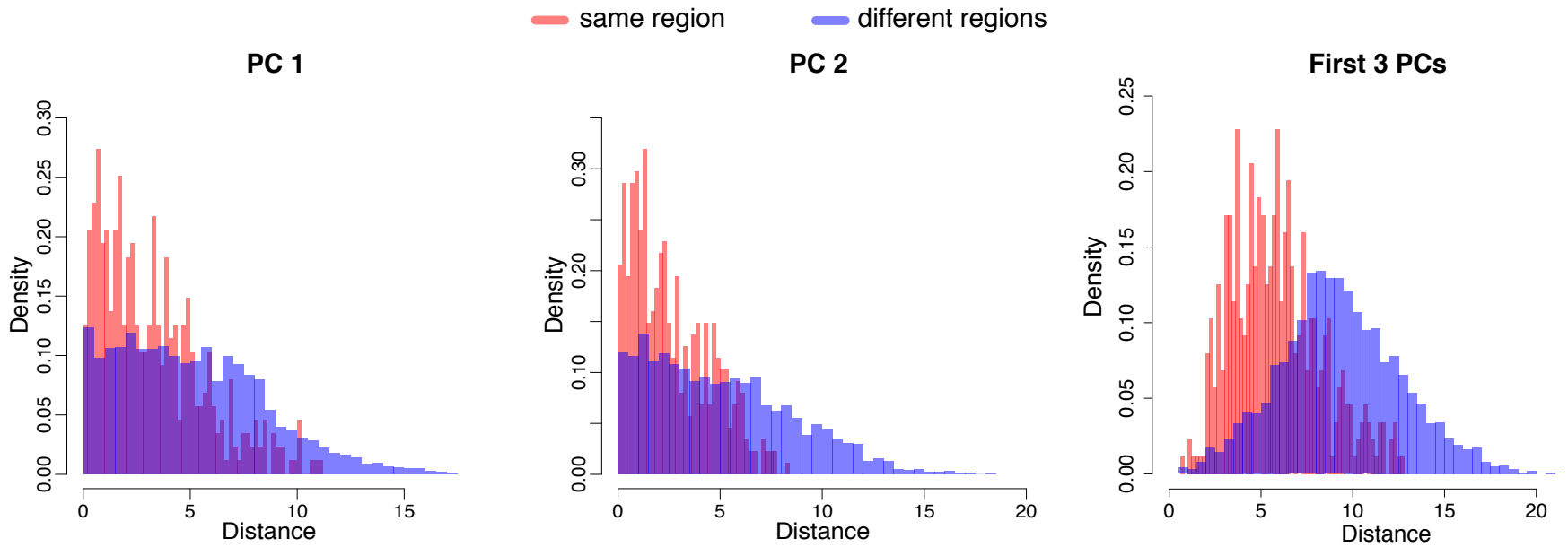
- For a test set C, leave out $X_{(C)}$, use $X_{(-C)}$ to calculate v
- For each PC, calculate the pairwise distance between samples in $X_{(C)}$ and that in $X_{(-C)}$

Leave-one-sample-out:



Detection of abnormal samples, brain data

- Leave-one-donor-out

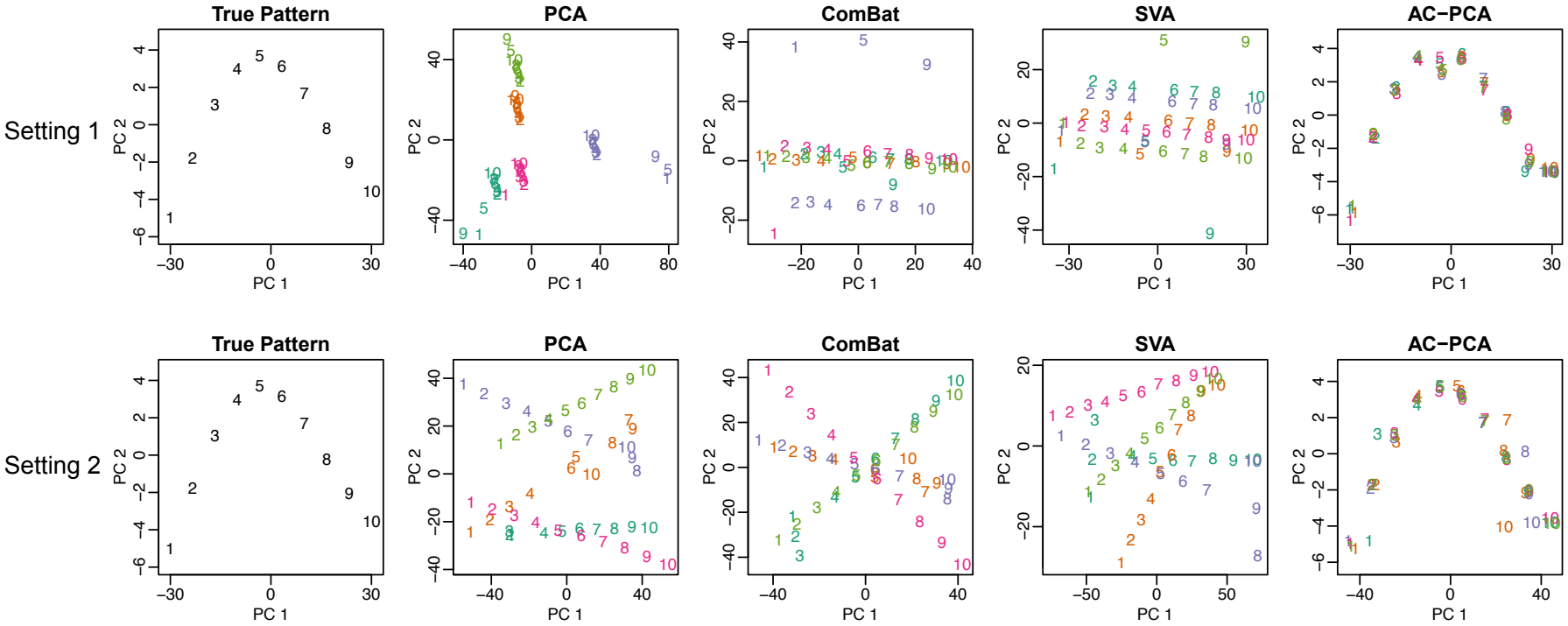


Simulations

- $n = 5$, $b = 10$ and $p = 400$
- Data matrix for the ith donor $X^{(i)} = \Omega + \Gamma^{(i)} + \epsilon^{(i)}$
- Goal: capture the shared component Ω (low rank)
- The donor specific component $\Gamma^{(i)} = \Lambda_1^{(i)} + \Lambda_2^{(i)}$
  - $\Lambda_1^{(i)}$: donor's effect is the same among the regions
  - $\Lambda_2^{(i)}$: donor's effect is not the same, considered two settings

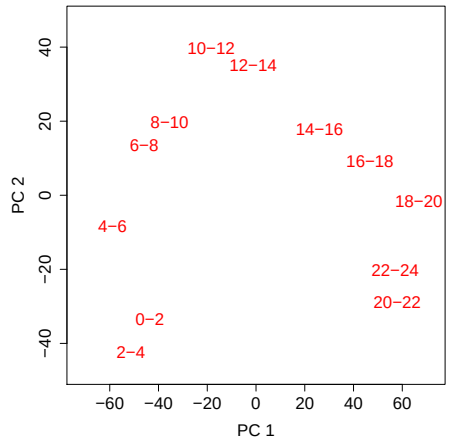
Simulations - result

- Setting 1: three random regions are affected in $\Lambda_2^{(i)}$
- Setting 2: the latent structure in $\Lambda_2^{(i)}$ is correlated with that in Ω

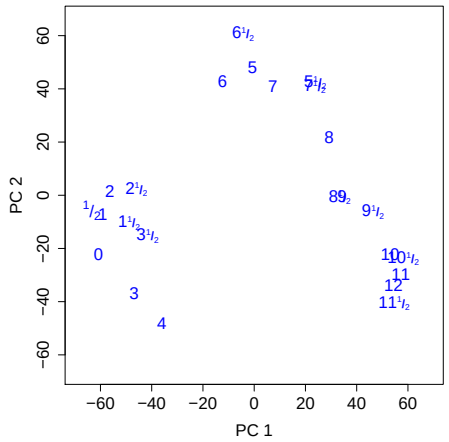


Data example (2): modENCODE RNA-Seq data

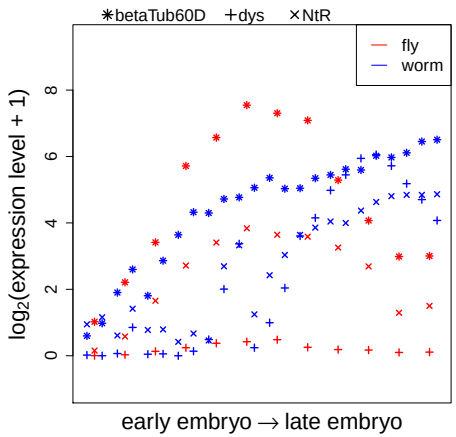
- Fly and worm, 4831 ortholog (gene) pairs, time course data during embryonic development (Celinker SE et al. 2009, Gerstein MB et al. 2014)
- Fly embryo, 12 time windows: 0-2h, 2-4h, ..., 22-24h
- Worm embryo, 24 time points: 0h, 0.5h, ..., 4h, 5h, ..., 12h



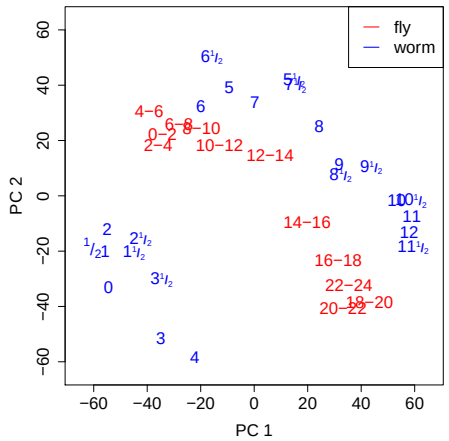
(a) Fly, PCA



(b) Worm, PCA



(c) Fly, PCA, PC 2



(d) Fly+Worm, PCA

- Challenging to identify the shared temporal variation

AC-PCA adjusting for variations of species

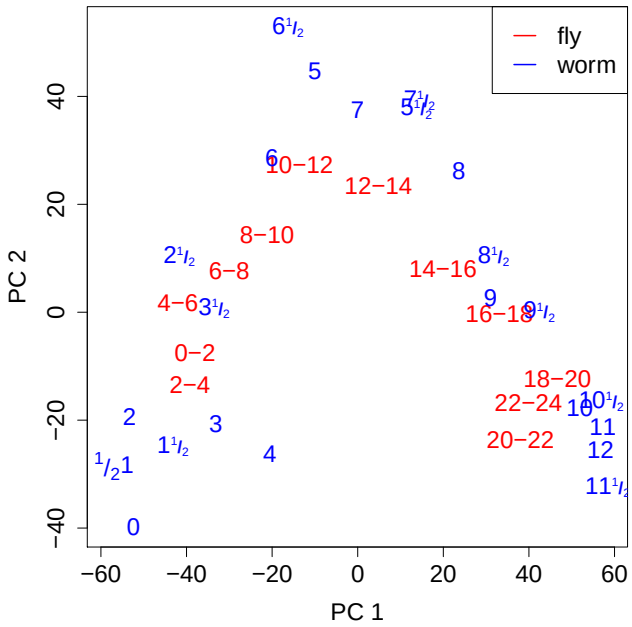
- Fly data matrix $X^{(f)}$; Worm data matrix $X^{(w)}$
- Fly 0-$2h$ $\rightarrow$ worm $\{0h, 0.5h, 1h\}$

$$\begin{aligned} &\underset{v \in \mathbb{R}^p}{\text{maximize}}\ v^T X^T X v - \lambda \sum_{t=1}^{12} v^T (X_t^{(f)} - f(X^{(w)}, t))^T (X_t^{(f)} - f(X^{(w)}, t)) v \\ &\text{subject to } ||v||_2^2 \leq 1, \end{aligned} \tag{7}$$

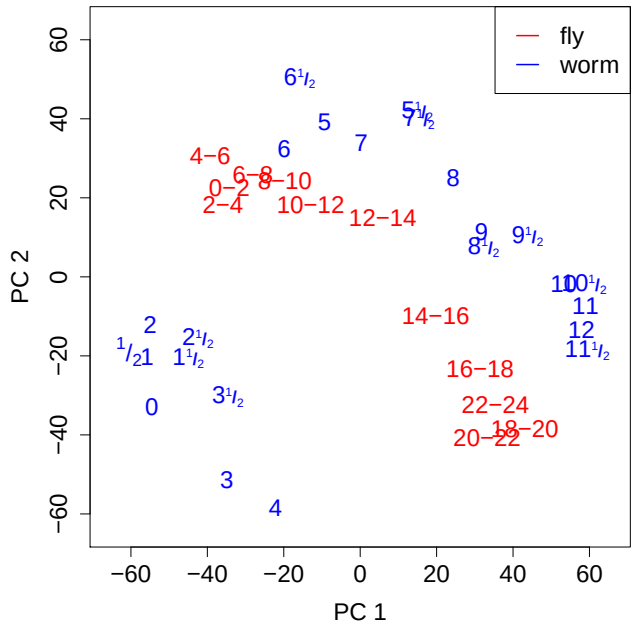
where

$$f(X^{(w)}, t) = \begin{cases} \frac{1}{2}(X_{2t-1}^{(w)} + X_{2t+1}^{(w)}), \text{ if } t = 5 \\ \frac{1}{3}(X_{2t-1}^{(w)} + X_{2t}^{(w)} + X_{2t+1}^{(w)}), \text{ otherwise} \end{cases}$$

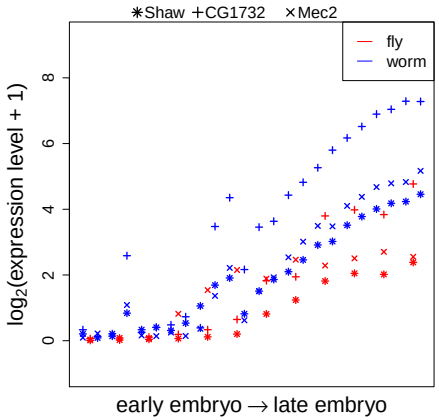
modENCODE RNA-Seq data



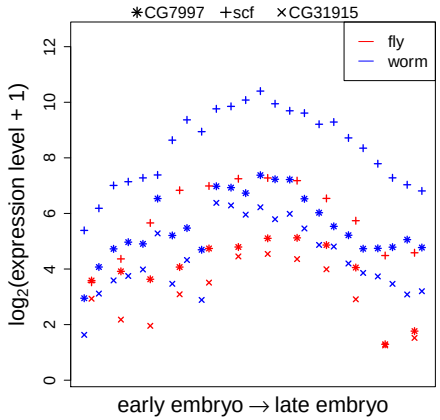
(a) Fly+Worm, AC-PCA



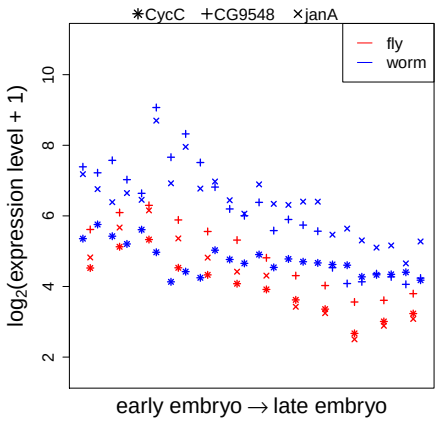
(b) Fly+Worm, PCA



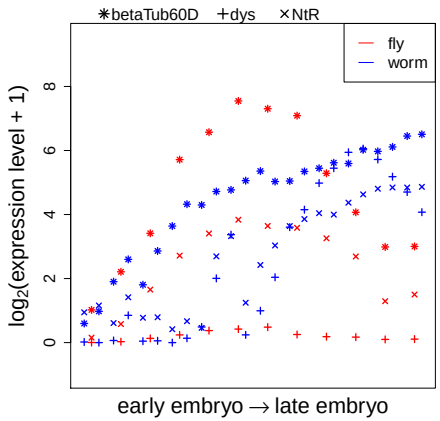
(a) AC-PCA, PC 1



(b) AC-PCA, PC 2



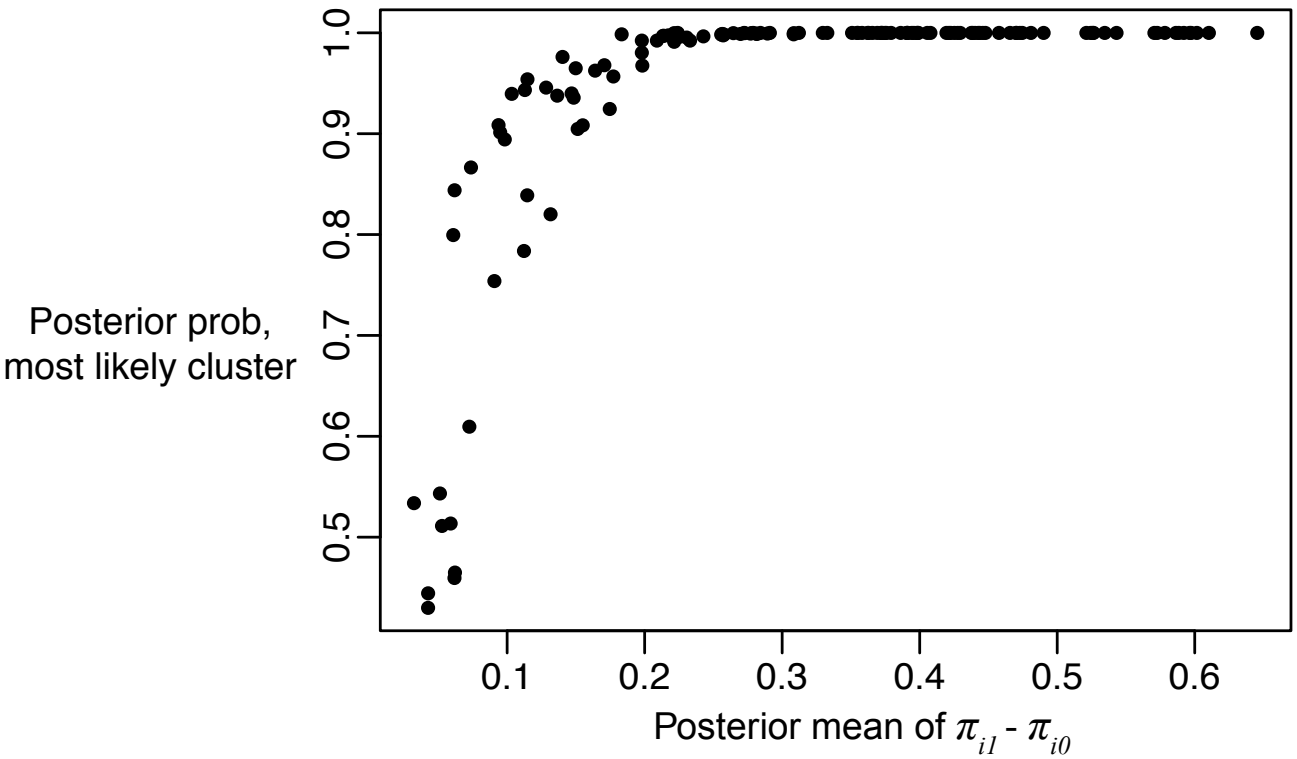
(c) Fly, PCA, PC 1



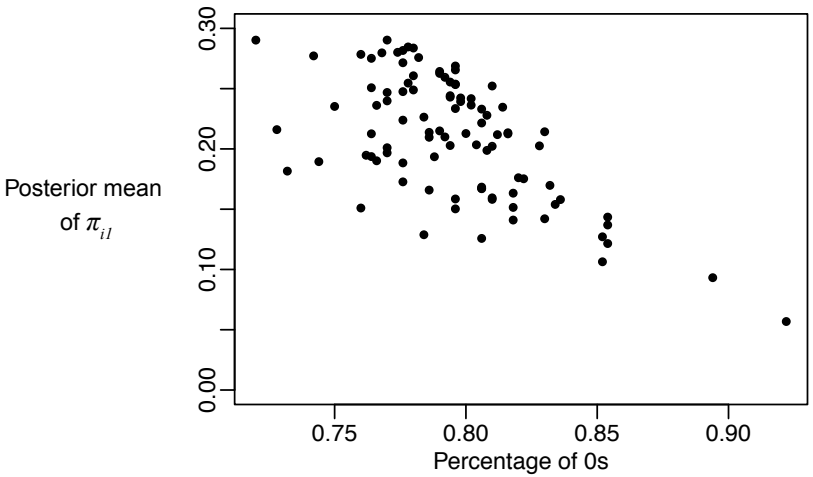
(d) Fly, PCA, PC 2

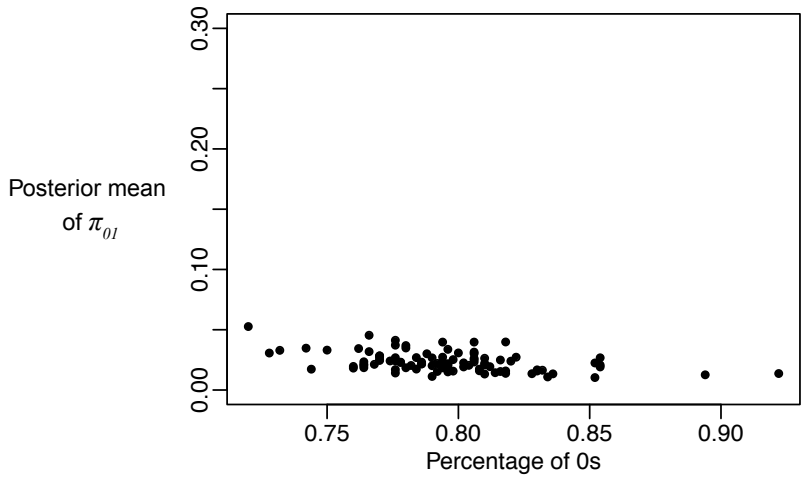
- The joint analysis tends to identify genes with consistent patterns in fly and worm

$\pi_{i1} - \pi_{i0}$ associates with the cluster probability



π_{i1} associates with the drop-out rate



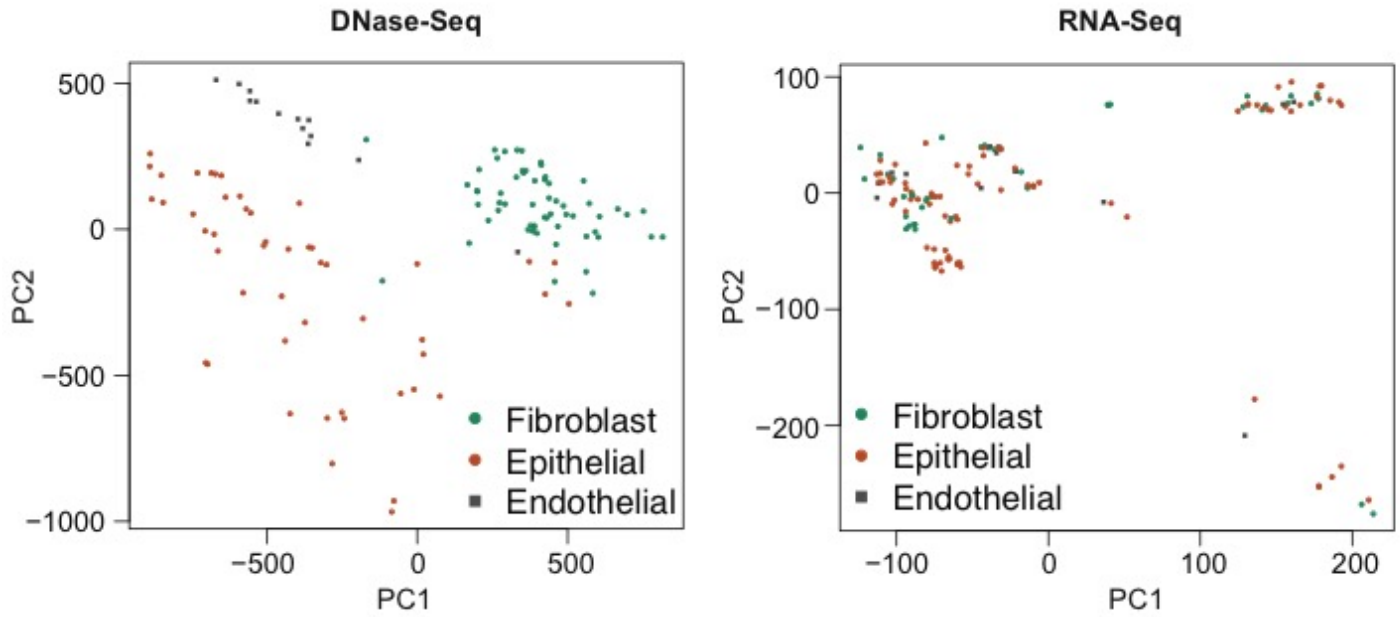


Acknowledgement

- Hongyu Zhao (Yale), Matthew State (UCSF) and Wing Hung Wong (Stanford)
- Can Yang (Hong Kong Baptist University)
- Tao Wang (Shanghai Jiao Tong University)
- Stephan Sanders (UCSF), Yao Fu, Feng Cheng, Ying Zhu, Xuming Xu, Mingfeng Li and Nenad Sestan
- Yong Wang (Chinese Academy of Sciences), Rui Jiang (Tsinghua), John C. Duchi, Bai Jiang and Mahdi Zamanighomi

Data example (2)

- Data collected from human ENCODE and Roadmap
- Fibroblast, epithelial and endothelial cells
- Bulk DNase-Seq, 121 samples
- Bulk RNA-Seq, 135 samples



- Better separation by accessibility data

Accessibility guided separation of RNA-Seq samples

- Step 1: cluster the accessibility samples

		DNase-Seq samples		
		Fibroblast	Epithelial	Endothelial
DNase-Seq clusters	Cluster 1	57	4	1
	Cluster 2	2	45	0
	Cluster 3	0	0	12

- Step 2: model-based promoter selection (166)
- Step 3: model-based assignment of RNA-Seq samples to accessibility clusters

		RNA-Seq samples		
		Fibroblast	Epithelial	Endothelial
DNase-Seq clusters	Cluster 1	38	1	0
	Cluster 2	2	79	1
	Cluster 3	0	2	12