ChIP-X: Data Integration

Statistics in Genomics
Hongkai Ji (hji@jhsph.edu)
dPCA: differential principal component analysis of ChIP-seq

Motivation: How to compare multiple ChIP-seq profiles between two biological conditions?
Basic Questions

1) What are the major differential patterns?

2) How to detect and prioritize differential loci for follow-up studies?

3) How to assess significance given the background biological or technical variation among replicate samples?
Data Structure

Cell Type 1

Dataset 1 (H3K4me3)
Rep 1
... Rep K_{11}

Dataset 2 (H3K27me3)
Rep 1
... Rep K_{12}

... Dataset M (Myc)
Rep 1
... Rep K_{1M}

Cell Type 2

Dataset 1 (H3K4me3)
Rep 1
... Rep K_{21}

Dataset 2 (H3K27me3)
Rep 1
... Rep K_{22}

... Dataset M (Myc)
Rep 1
... Rep K_{2M}

Locus 1
Locus 2
... Locus G

\( x_{g1mk} \)
Intensity for locus \( g \), condition 1, dataset \( m \), replicate \( k \)

\( x_{g2mk} \)
Intensity for locus \( g \), condition 2, dataset \( m \), replicate \( k \)
Limitations of Existing Methods

Dataset 1 \[\rightarrow\] condition 1 \[\rightarrow\] Dataset 1

Dataset 1 \[\rightarrow\] condition 2 \[\rightarrow\] Dataset 2

Dataset m \[\rightarrow\] condition m \[\rightarrow\] Dataset m

Dataset M

**Differential Expression Approach:**
Smyth (2004); Anders & Huber (2010); Laajala et al. (2009); Xu et al. (2008); Choi et al. (2009); Johannes et al. (2010); Shao et al. (2012)

(1) Ignores correlation among proteins
(2) $3^M$ possible differential patterns

**Venn Diagram Approach:**
(1) Comparison is based on binary peak calls
(2) Loses important quantitative information
(3) Cutoff dependent
dPCA Model

Intensity for locus $g$, condition 1, dataset $m$, replicate $k$ :

$x_{g1mk} \sim N(\mu_{g1m}, \sigma^2)$

Intensity for locus $g$, condition 2, dataset $m$, replicate $k$ :

$x_{g2mk} \sim N(\mu_{g2m}, \sigma^2)$

$\bar{X}_1 = (x_{g1m})_{G \times M}$

$\bar{X}_2 = (x_{g2m})_{G \times M}$

$D_{G \times M} = \text{True Difference} \Delta + \text{Noise } E$
Modeling True Differences

\[ \Delta_{G \times M} = \delta_g^T = \left[ \beta_{g1} \ldots \beta_{gM} \right] \left[ \begin{array}{c} v_1^T \\ \vdots \\ v_M^T \end{array} \right] \]

s.t.

(1) \[ \|v_j\|_2 = 1 \]
\[ v_i \perp v_j \ (i \neq j) \]

(2) \[ b_{gj} \sim \text{Bernoulli}(\pi_j) \]
\[ w_{gj} \sim H_j(w; 0, \tau_j^2) \]
\[ \beta_{gi} = b_{gl}w_{gj} \]
dPCA Model in Matrix Form

\[ D = \overline{X}_1 - \overline{X}_2 = \Delta + E \]

\[ \equiv \Delta = BV^T = \]

\[ V^TV = I \]

\[ \text{Var}(\beta_{gi}) \equiv \lambda_j = \pi_j \tau_j^2 \]

\[ \lambda_1 > \lambda_2 > \cdots > \lambda_M \]
Dimension Reduction

\[ D = \overline{X}_1 - \overline{X}_2 = \Delta + E \]

\[ \equiv \Delta = BV^T = \]

\[ V^T V = I \]
Analysis Goals

1. Estimate

\[ V = \begin{bmatrix} \vdots \end{bmatrix} \]

2. Infer and for each dPC \( j \), rank loci based on \( \beta_{gj} \).

\[ B = \begin{bmatrix} \vdots \end{bmatrix} \]

3. For each locus \( g \) and dPC \( j \), test if \( \beta_{gj} = 0 \).
dPCA Algorithm

1. Estimate \( V \) by \( \hat{V} \), which are eigenvectors of \( \text{Var}(D) - \text{Var}(E) \).

2. Project data to \( \hat{V} \) to estimate \( B \):

\[
D = BV^T + E
\]

\[
d_g = \delta_g + e_g \implies v_j^T d_g = v_j^T \delta_g + v_j^T e_g = \beta_{gj} + \varepsilon_{gj}
\]
Statistical Tests

Under $H_0: \beta_{gj} = 0$,

$$\mathbf{v}_j^T \mathbf{d}_g \sim \mathcal{N}(0, \sigma^2 \mathbf{v}_j^T \mathbf{\Omega} \mathbf{v}_j)$$

$$T_{gj} = \frac{\mathbf{v}_j^T \mathbf{d}_g}{\sqrt{\hat{\sigma}^2 \mathbf{v}_j^T \mathbf{\Omega} \mathbf{v}_j}} \xrightarrow{t_v} p_{gj} \quad \text{(Two-sided p-value)}$$

Since we don’t know $\mathbf{V}$, we have

$$\hat{T}_{gj} = \frac{\hat{\mathbf{v}}_j^T \mathbf{d}_g}{\sqrt{\hat{\sigma}^2 \hat{\mathbf{v}}_j^T \mathbf{\Omega} \hat{\mathbf{v}}_j}} \xrightarrow{t_v} \hat{p}_{gj}$$

$$\mathbf{D} = \mathbf{B} \mathbf{V}^T + \mathbf{E}$$
Which dPCs to Report?

Signal-to-Noise Ratio (SNR) for dPC \( j \):

\[
SNR_j = \frac{\text{Var}(v_j^T d_g)}{\text{Var}(v_j^T e_g)} \approx \frac{\hat{v}_j^T (D^T D / G) \hat{v}_j}{\hat{v}_j^T (\hat{\sigma}^2 \Omega) \hat{v}_j}
\]

\[
= \frac{\text{Projected Total Variation}}{\text{Projected Error Variation}}
\]

Report dPC \( j \) if \( SNR_j \geq 5 \)
Which dPCs to Report: Intuition

High SNR

Low SNR
Example: differential chromatin patterns at DNA motif sites

Conditions: K562 vs. Huvec

$M = 18$ ENCODE data sets

$G = 58,997$ MYC motif sites in human genome
dPCA can provide meaningful ranking

![Scatter plots showing correlation]

Correlation:
- Cor = 0.647
- Cor = -0.061
dPCA improves differential loci ranking over single dataset analysis
False Discovery Rate Control

A
Signal-to-Noise Ratio

B
Eigenvector Estimation Error

C
Eigenvalue Estimation Error

D
Variance Structure

E
FDR: dPC1

F
FDR: dPC2

G
FDR: dPC3
dPCA vs. PCA

**dPCA:**
1. More effective dimension reduction (75% vs. 57% variance by top two PCs)
2. Incorporates replicate variability
3. Allows assessment of statistical significance and reproducibility
Application: predict differential TF binding

ChIP-seq is

**High-throughput** with respect to surveying the whole genome

BUT

**Low-throughput** with respect to surveying many TFs

However,

Among ~1400 TFs, ~500 have DNA motifs
Many HMs have good antibodies
Therefore, coupling dPCA with chromatin data may allow one to predict differential TF binding for many TFs!
Additional tests produce similar results
Summary

- Data Integration Can Help You to Better Utilize ChIP-X Data
- dPCA: a better way to analyze differential binding