Exploring patterns in Hi-C datasets from various mice strains using standard and contrastive principal component analysis

Asiya Falak¹, Zeyuan Chen², Jeffrey N. Chiang³

1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
2 Department of Computer Science, UCLA
3 Department of Computational Medicine, UCLA

Background

- Chromosomes are composed of strands of chromatin that can be analyzed via compartments.
- There are two main compartments: 'A' = open and active chromatin and 'B' = inactive and closed chromatin.
- Differential compartment analysis of Hi-C data (dcHiC) identifies statistically significant differences in compartmentalization among two or more contact maps.

Methodology – dcHiC and PCA

- SVD on correlation matrix is performed; PCs are generated.
- Matrix is checked for artifacts; sign flips and CG enrichment are performed.
- Mahalanobis distances are calculated.
- Chi-square tests performed.
- Outliers are detected.
- Second pass of Mahalanobis distances and chi-square tests.

Preprocessing Pipeline

- Raw contact matrix with Hi-C data undergoes fast KR normalization to control for noise.
- Observed vs. expected (OE) is used to correct for high-value densities on the diagonal.
- A matrix of z-scores is created for standardization (includes column center and unit variance).
- Variance covariance matrix is generated and symmetric.

Results - PCA

PCA calculates PCs to help visualize variance in data.

Contrastive PCA (cPCA)

- Maximum variance is preserved on foreground data in low dimensional space on cPC.
- Simultaneously, variance is minimized on background data.
- Simulated expected data:

\[
\lambda_1(v) = \sqrt{\sum C_{ij} v_i v_j}
\]

Results - cPCA

- Modalities: methylation and RNA follows 3-D, Differential gene, D methylation at DC region
- Pseudo-bulk of single cell Hi-C, cell-type specific DC
- Findings in Fig. 2 suggest that background data (BLK6H) is too similar to foreground data (DBAH) to pick up trends.
- Further studies can modify foreground and background data to better capture differential patterns.

Discussion

- Background data (BLK6H) is too similar to foreground data (DBAH) to pick up trends.
- Further studies can modify foreground and background data to better capture differential patterns.