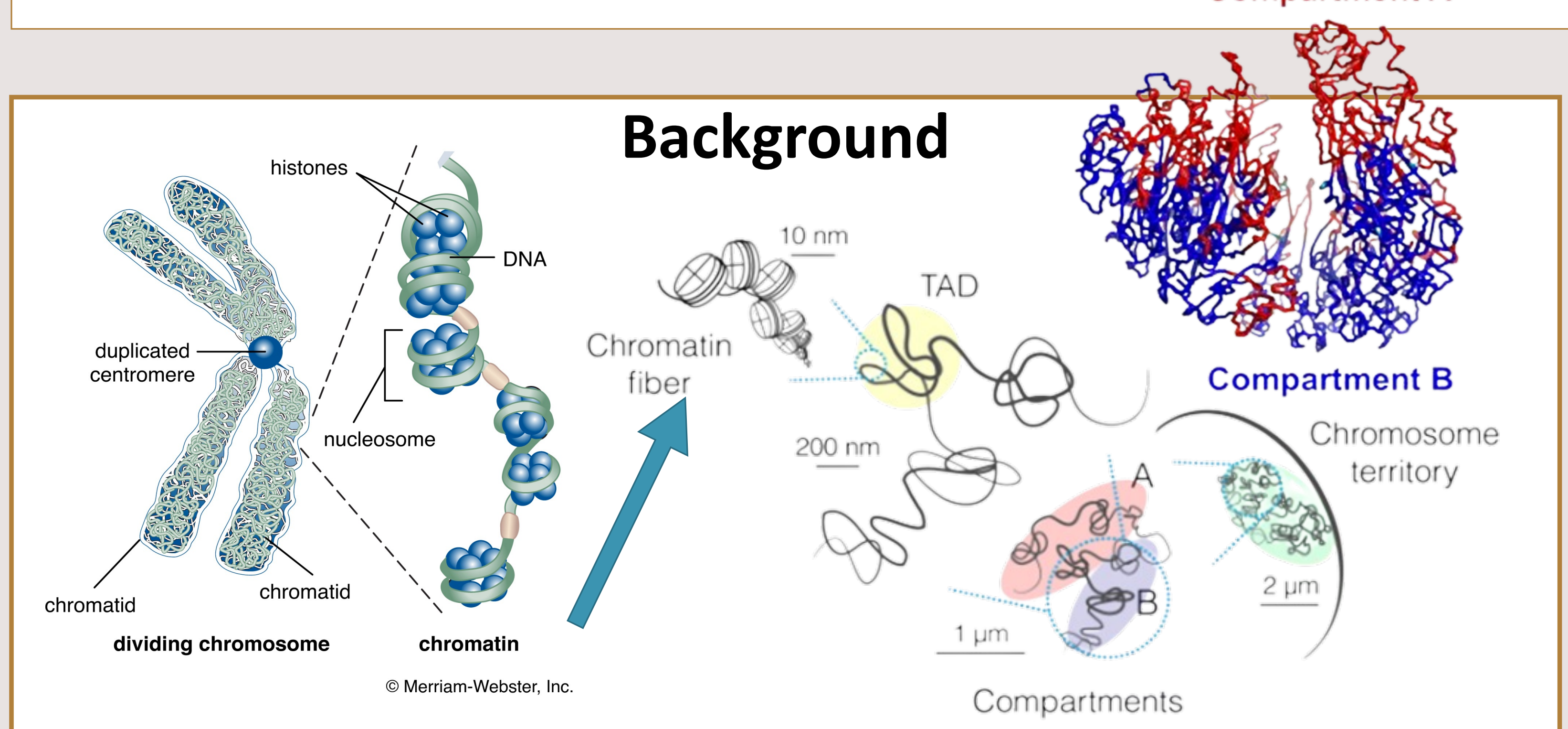
Exploring patterns in Hi-C datasets from various mice strains using standard and contrastive principal component analysis

UCLA QCBio

Asiya Falak[1], Zeyuan Chen[2], Jeffrey N. Chiang[3]
1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
2 Department of Computer Science, UCLA
3 Department of Computational Medicine, UCLA

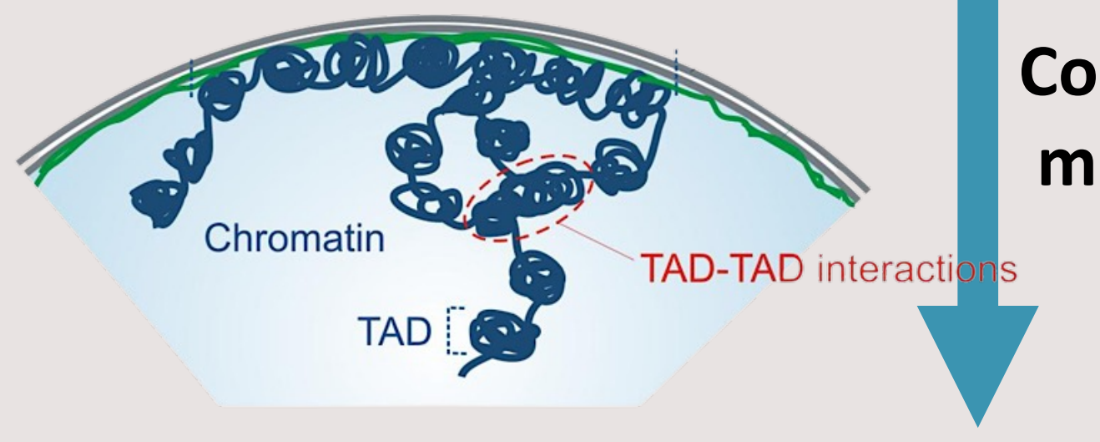
Background



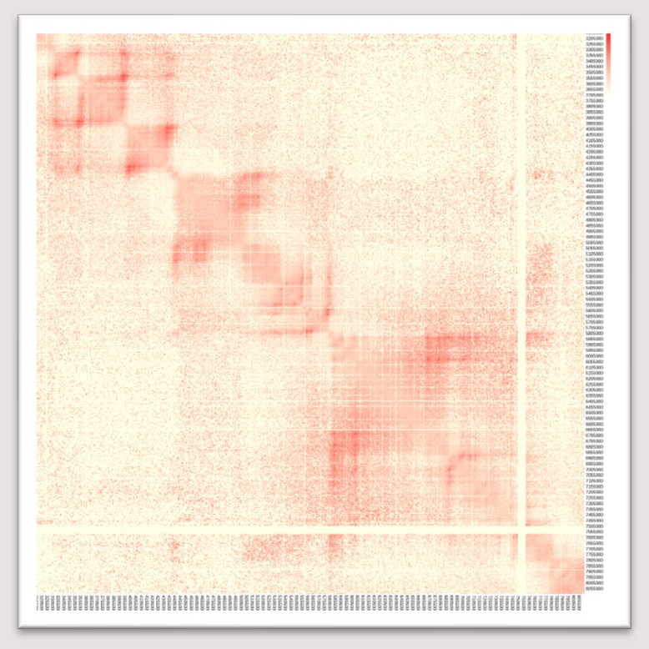
- Chromosomes are composed of strands of chromatin that can be analyzed via compartments.
- There are two main compartments: 'A' = open and active chromatin and 'B' = inactive and closed chromatin.
- Differential compartment analysis of Hi-C data (dcHiC) identifies statistically significant differences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

Preprocessing Pipeline

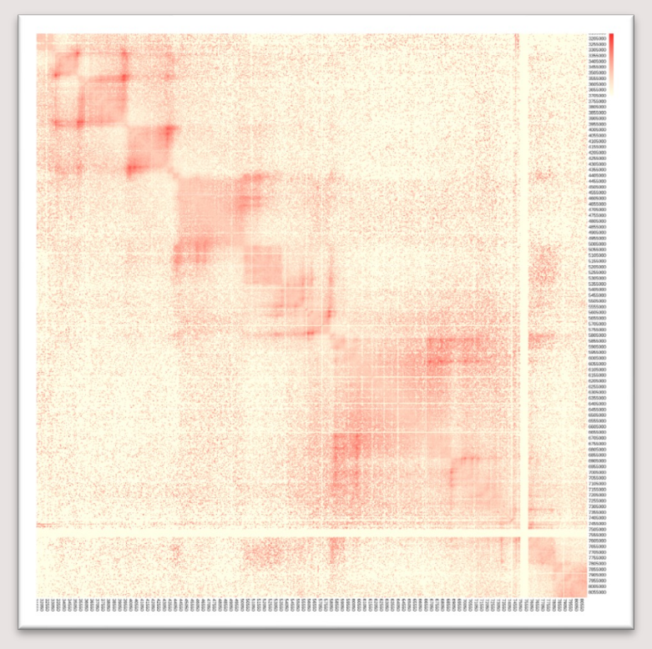
- Raw contact matrix with Hi-C data undergoes fast KR normalization to control for noise.
- Observed vs. expected (OE) is used to correct for high-value densities on the diagonal.
- A matrix of z-scores is created for standardization (includes column center and unit variance).
- Variance covariance matrix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



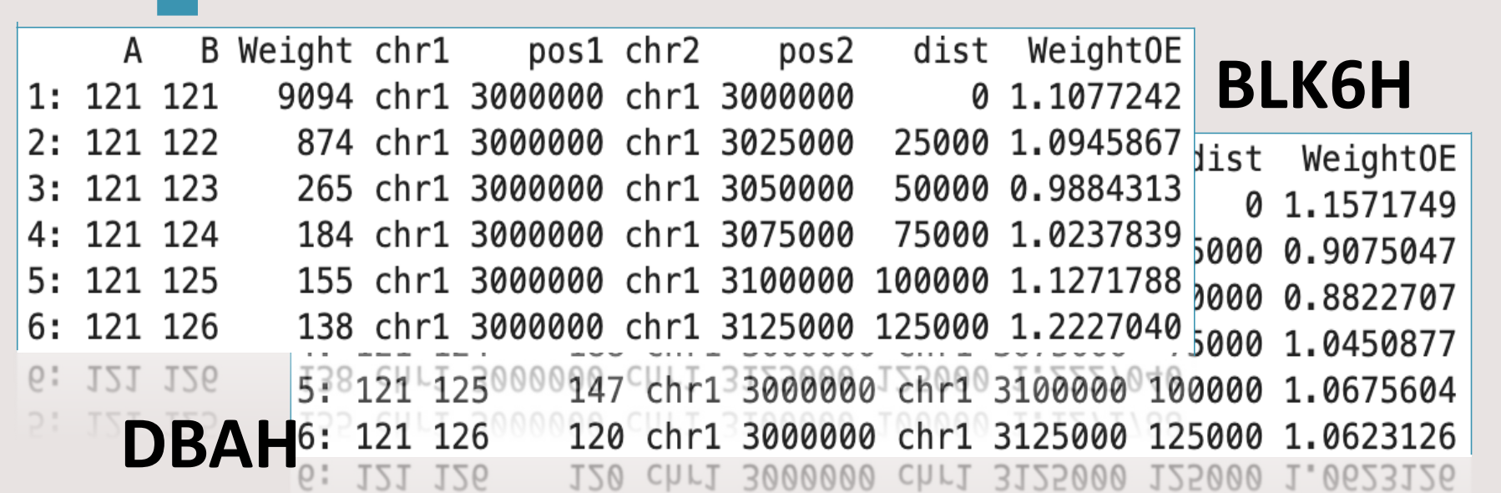
Contact matrix



BLK6H raw Hi-C data



DBAH raw Hi-C data



BLK6H

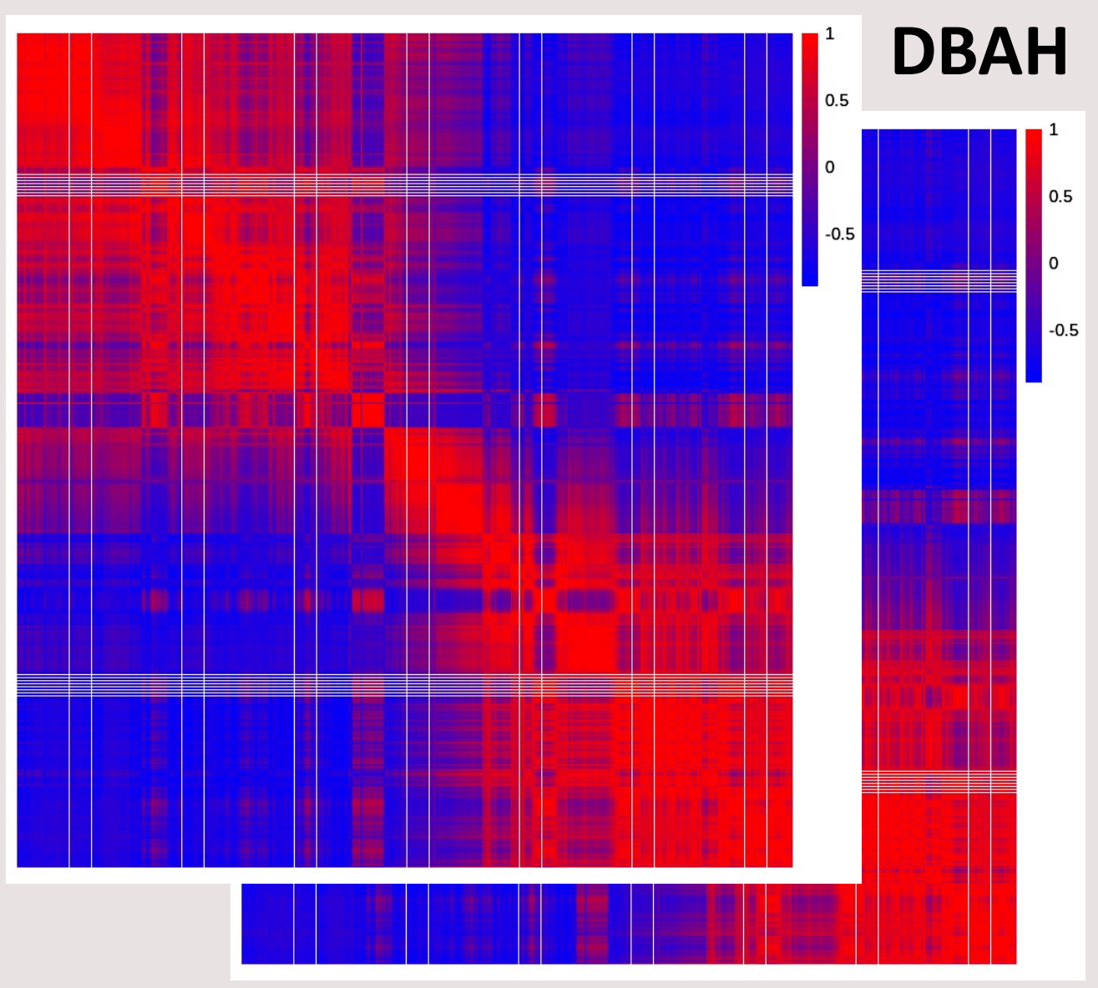
```
      A    B Weight chr1    pos1 chr2    pos2   dist  WeightOE
1: 121 121   9094 chr1 3000000 chr1 3000000      0 1.1077242
2: 121 122    874 chr1 3000000 chr1 3025000  25000 1.0945867
3: 121 123    265 chr1 3000000 chr1 3050000  50000 0.9884313
4: 121 124    184 chr1 3000000 chr1 3075000  75000 1.0237839
5: 121 125    155 chr1 3000000 chr1 3100000 100000 1.1271788
6: 121 126    138 chr1 3000000 chr1 3125000 125000 1.2227040
```

DBAH

```
dist  WeightOE
   0 1.1571749
5000 0.9075047
0000 0.8822707
5000 1.0450877
5: 121 125    147 chr1 3000000 chr1 3100000 100000 1.0675604
6: 121 126    120 chr1 3000000 chr1 3125000 125000 1.0623126
```

Data files include loci, distance, and weight.

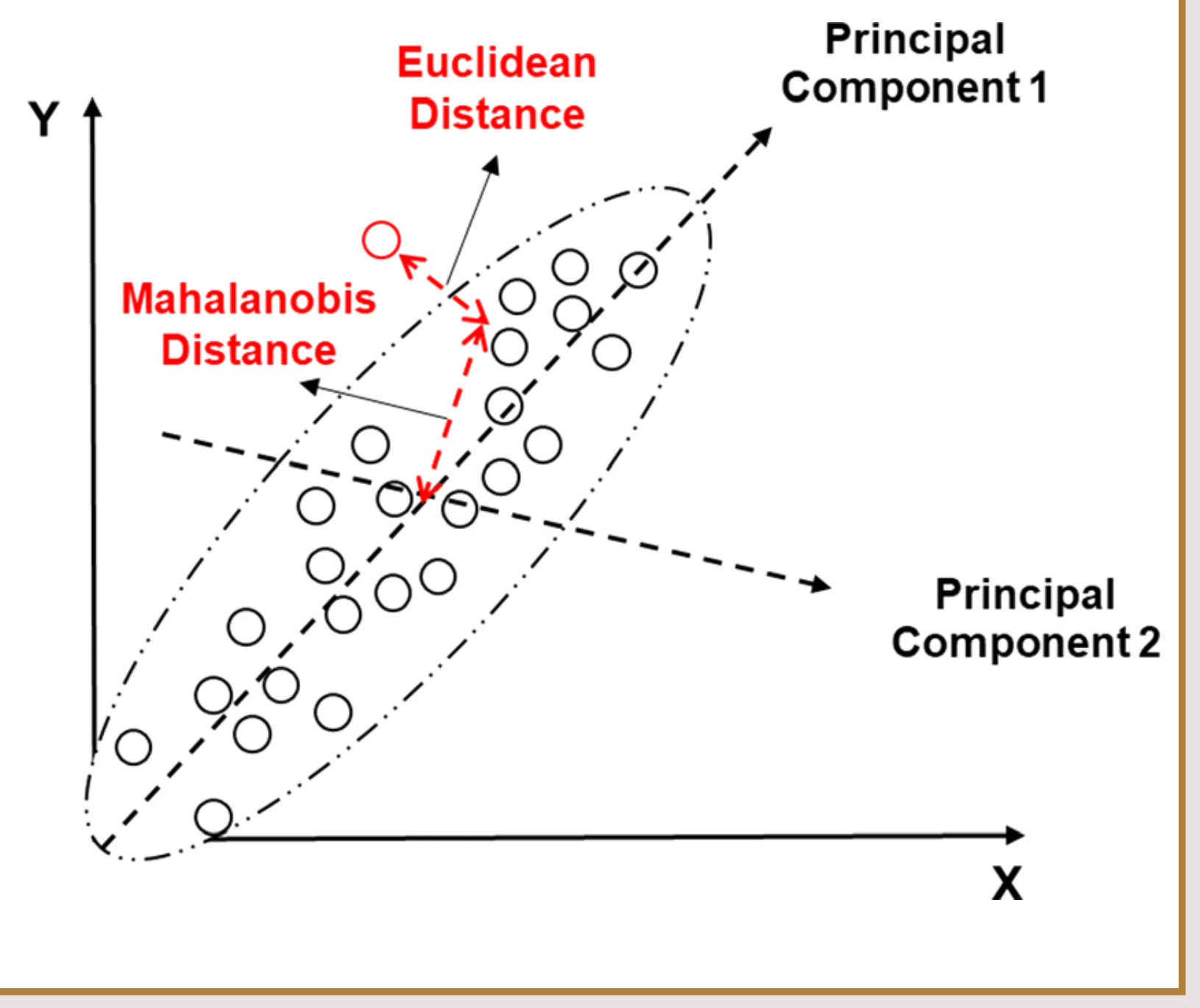
Covariance Matrices



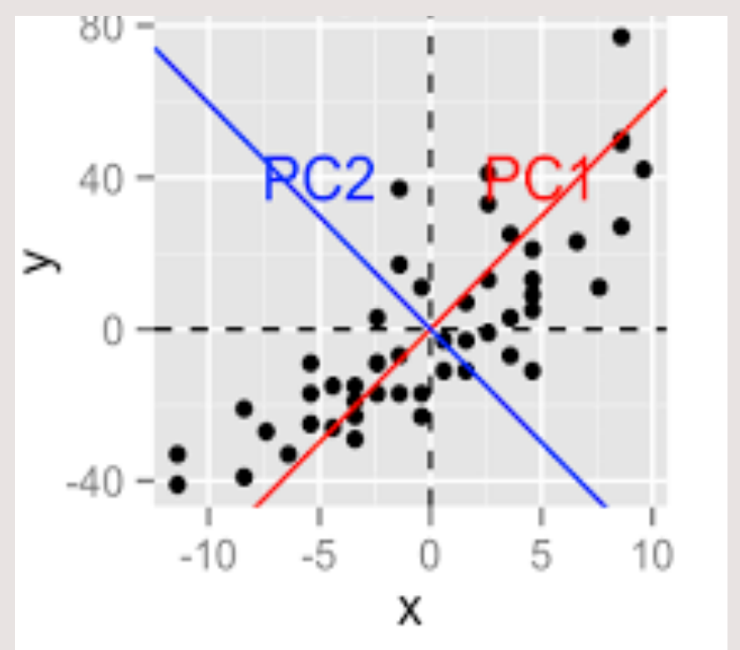
SVD

Methodology – dcHiC and PCA

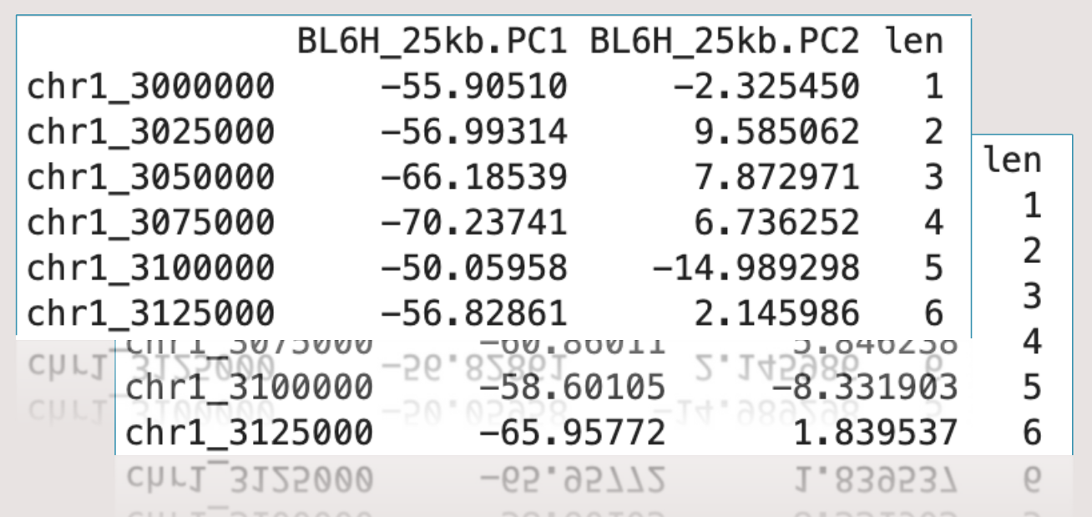
- SVD on correlation matrix is performed; PCs are generated.
- Matrix is checked for artifacts; sign flips and CG enrichment are performed.
- Mahalanobis distances are calculated.
- Chi-square tests performed.
- Outliers are detected.
- Second pass of Mahalanobis distances and chi-square tests.



Results - PCA



PCA calculates PCs to help visualize variance in data.



```
              BL6H_25kb.PC1 BL6H_25kb.PC2 len
chr1_3000000      -55.90510     -2.325450   1
chr1_3025000      -56.99314      9.585062   2
chr1_3050000      -66.18539      7.872971   3
chr1_3075000      -70.23741      6.736252   4
chr1_3100000      -50.05958    -14.989298   5
chr1_3125000      -56.82861      2.145986   6
```

```
                                          len
                                            1
                                            2
                                            3
                                            4
chr1_3100000      -58.60105     -8.331903   5
chr1_3125000      -65.95772      1.839537   6
```

PC1 and PC2 are calculated.

Figure 1.1: PCA Differential Results DBA vs. BL6

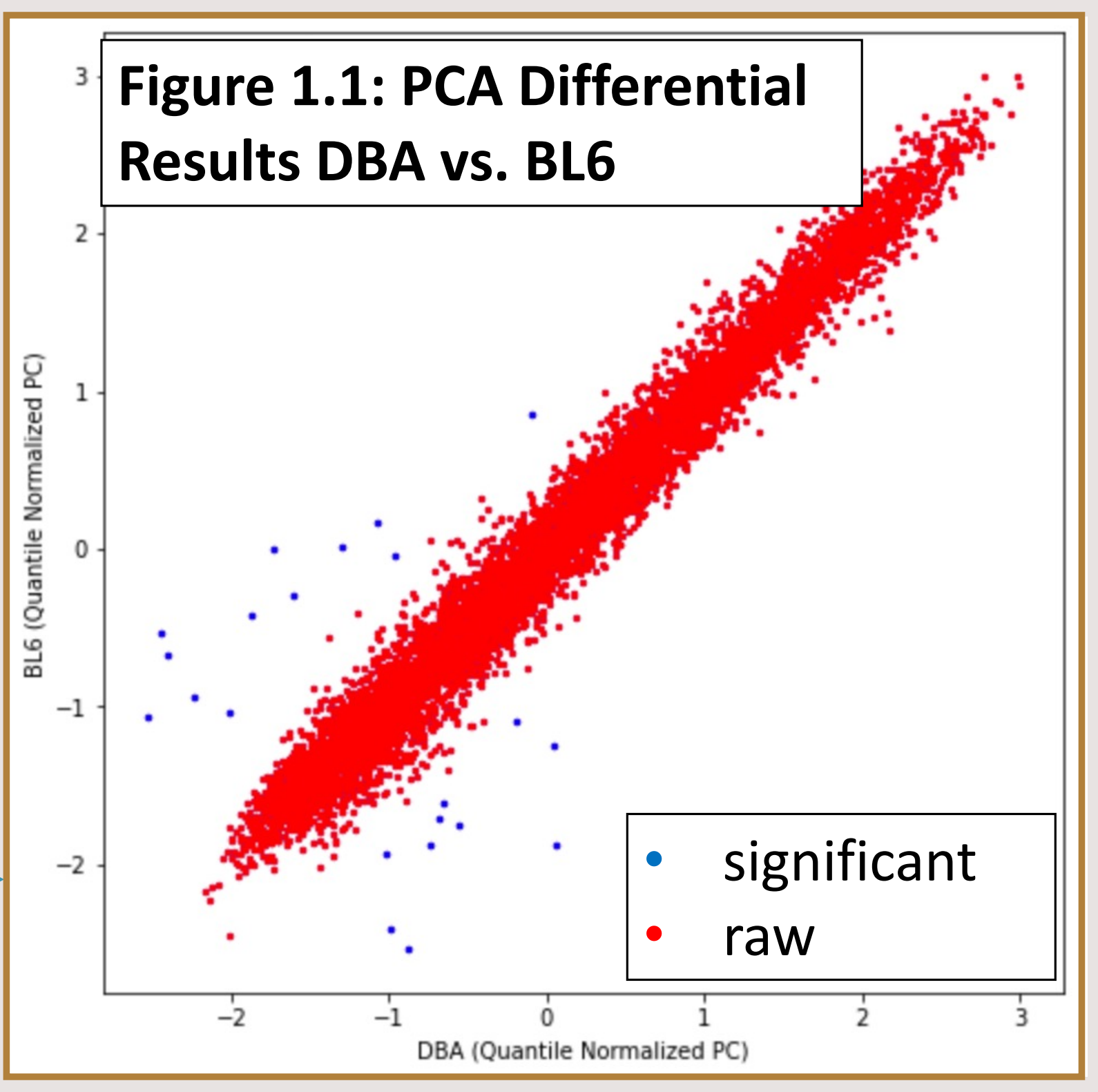


Figure 1.1 shows large resemblance between both strains.

Figure 1.2: AB Compartmentalization of Chromosome 5

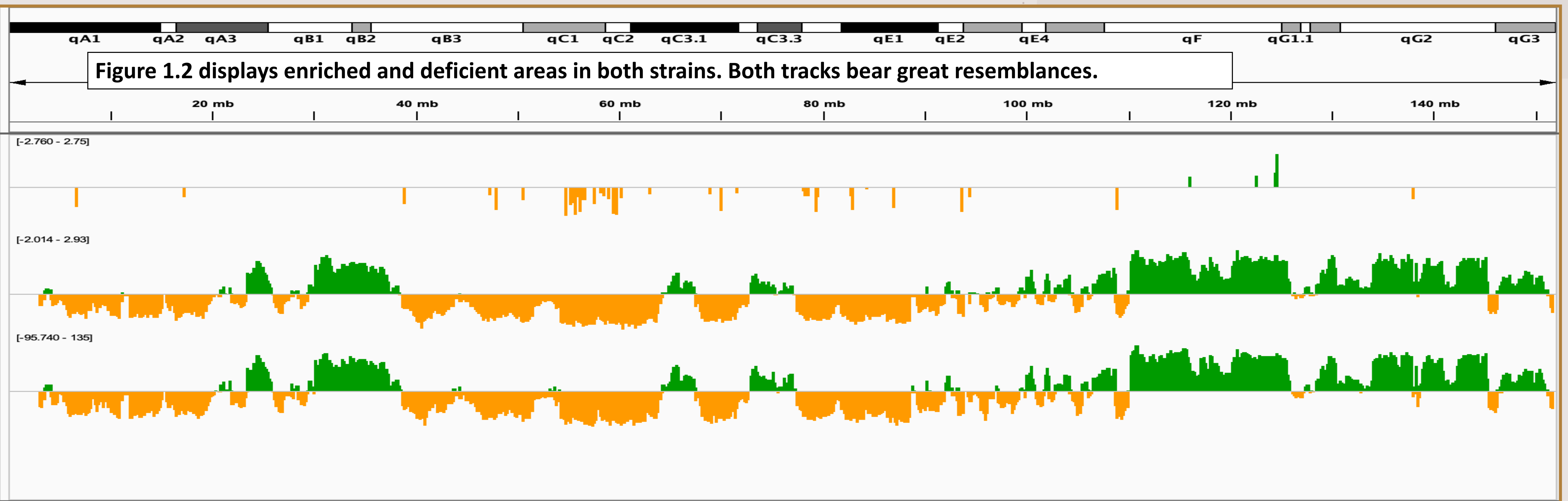


Figure 1.2 displays enriched and deficient areas in both strains. Both tracks bear great resemblances.

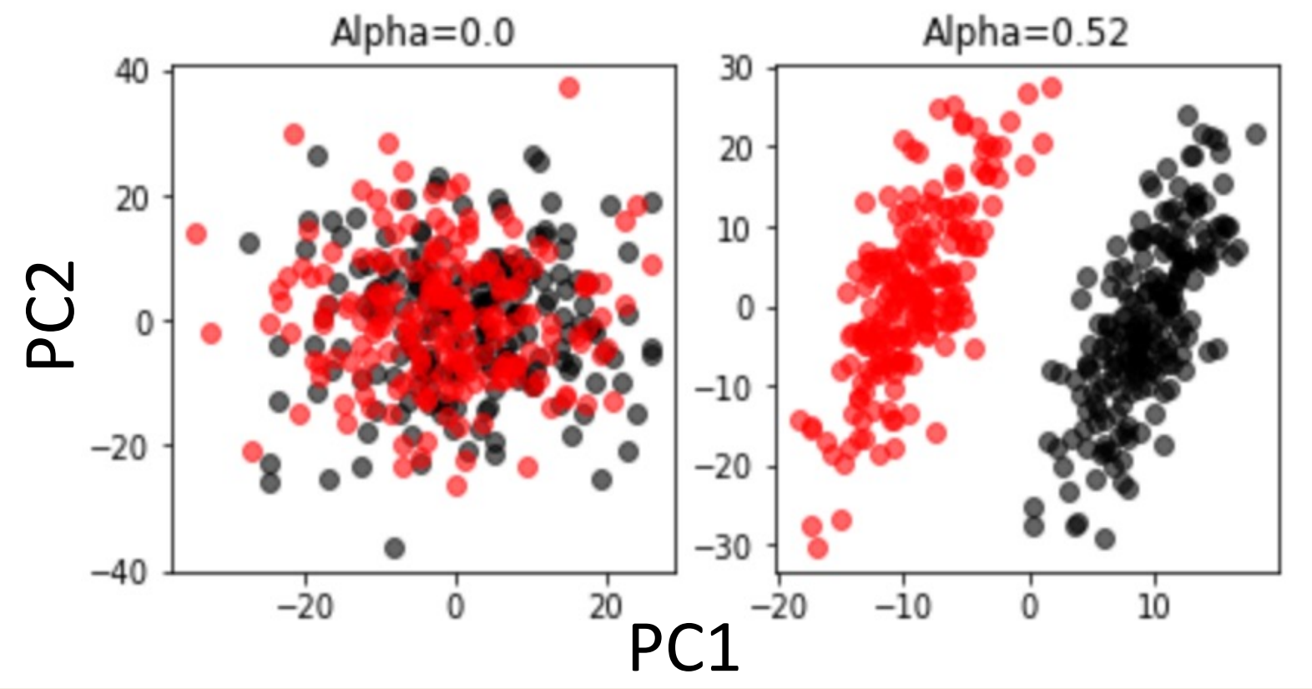
Contrastive PCA (cPCA)

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

Target variance: $\lambda_X(\mathbf{v}) \overset{\text{def}}{=} \mathbf{v}^T C_X \mathbf{v}.$

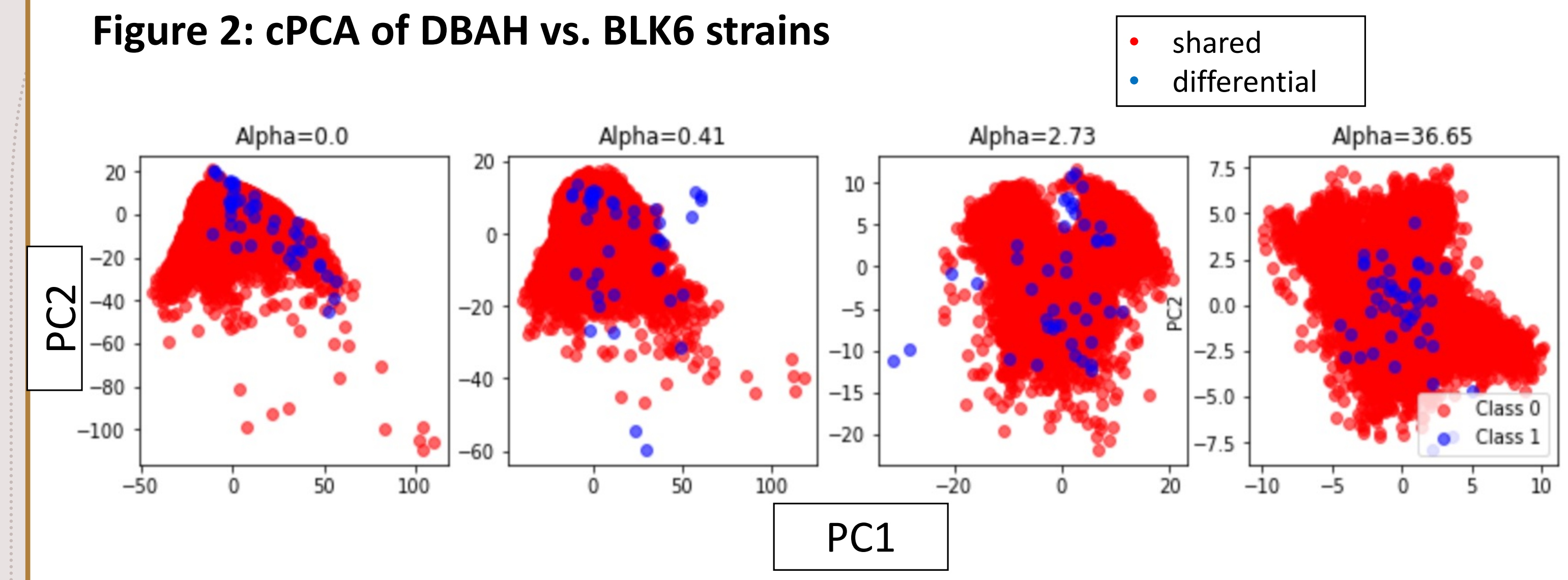
Background variance: $\lambda_Y(\mathbf{v}) \overset{\text{def}}{=} \mathbf{v}^T C_Y \mathbf{v}.$

$$\operatorname{argmax}_{\mathbf{v} \in \mathbb{R}^d_{\text{unit}}} \lambda_X(\mathbf{v}) - \alpha \lambda_Y(\mathbf{v})$$



Results - cPCA

Figure 2: cPCA of DBAH vs. BLK6 strains



Discussion

- 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.