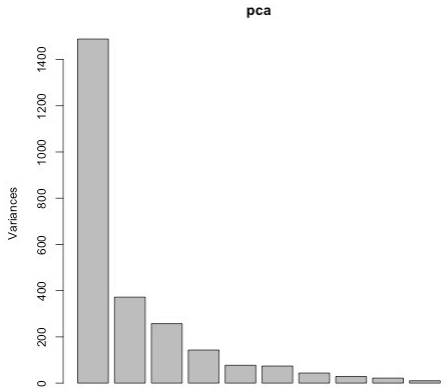
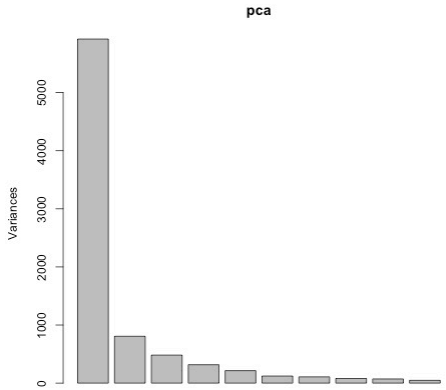
Demonstrating that Histone Modifications are Cell-Type Specific Through Clustering and Identifying Biologically Relevant Domains

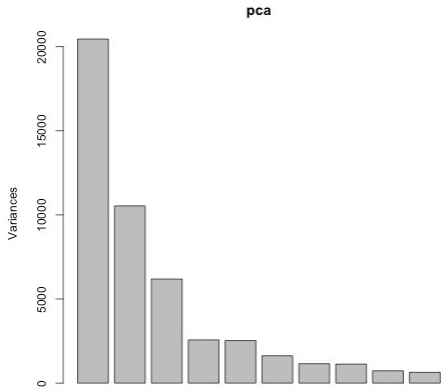
Adriana Sperlea & Douglas Arneson

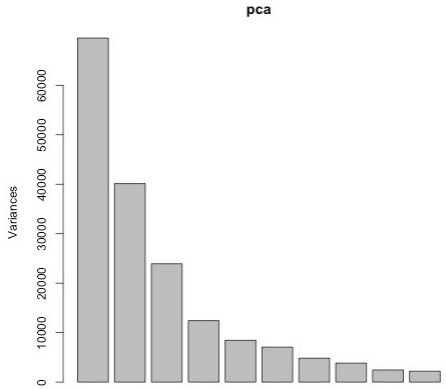
Methods

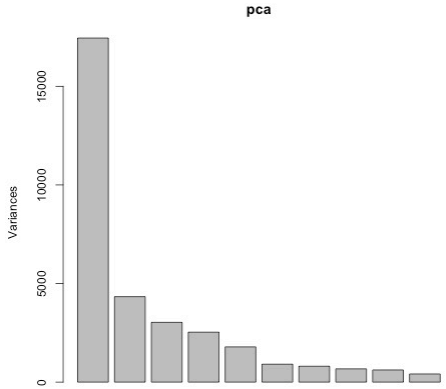
- PCA analysis across cell-types
- K-means clustering to identify cell type specificity of histone modifications
- Analysis of variance to identify important genomic locations

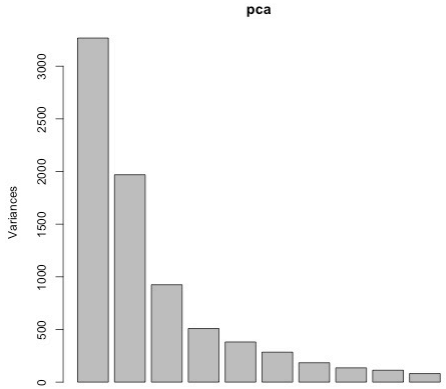

pca
Variances
0
200
400
600
800
1000
1200
1400

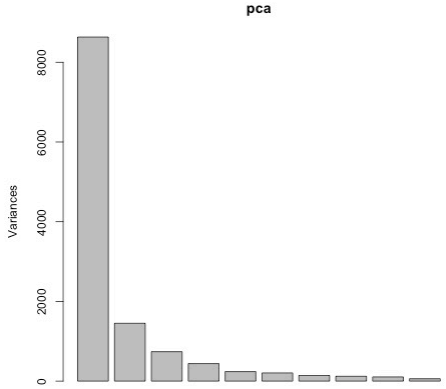

pca
Variances
0
1000
2000
3000
4000
5000

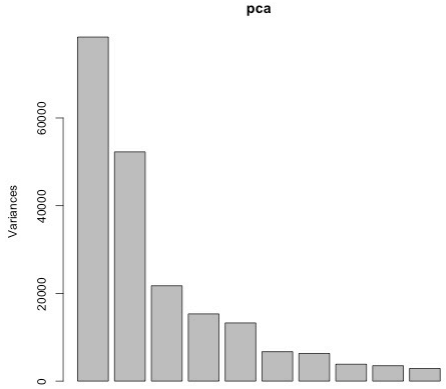

pca
Variances
0
5000
10000
15000
20000

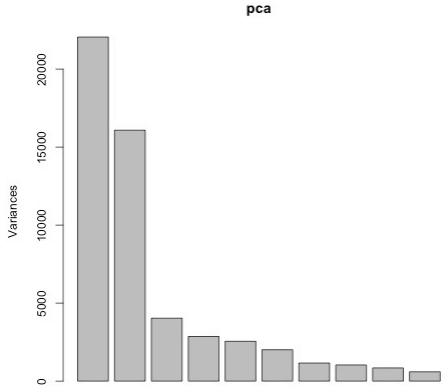

pca
Variances
0
10000
20000
30000
40000
50000
60000


pca
Variances
0
5000
10000
15000

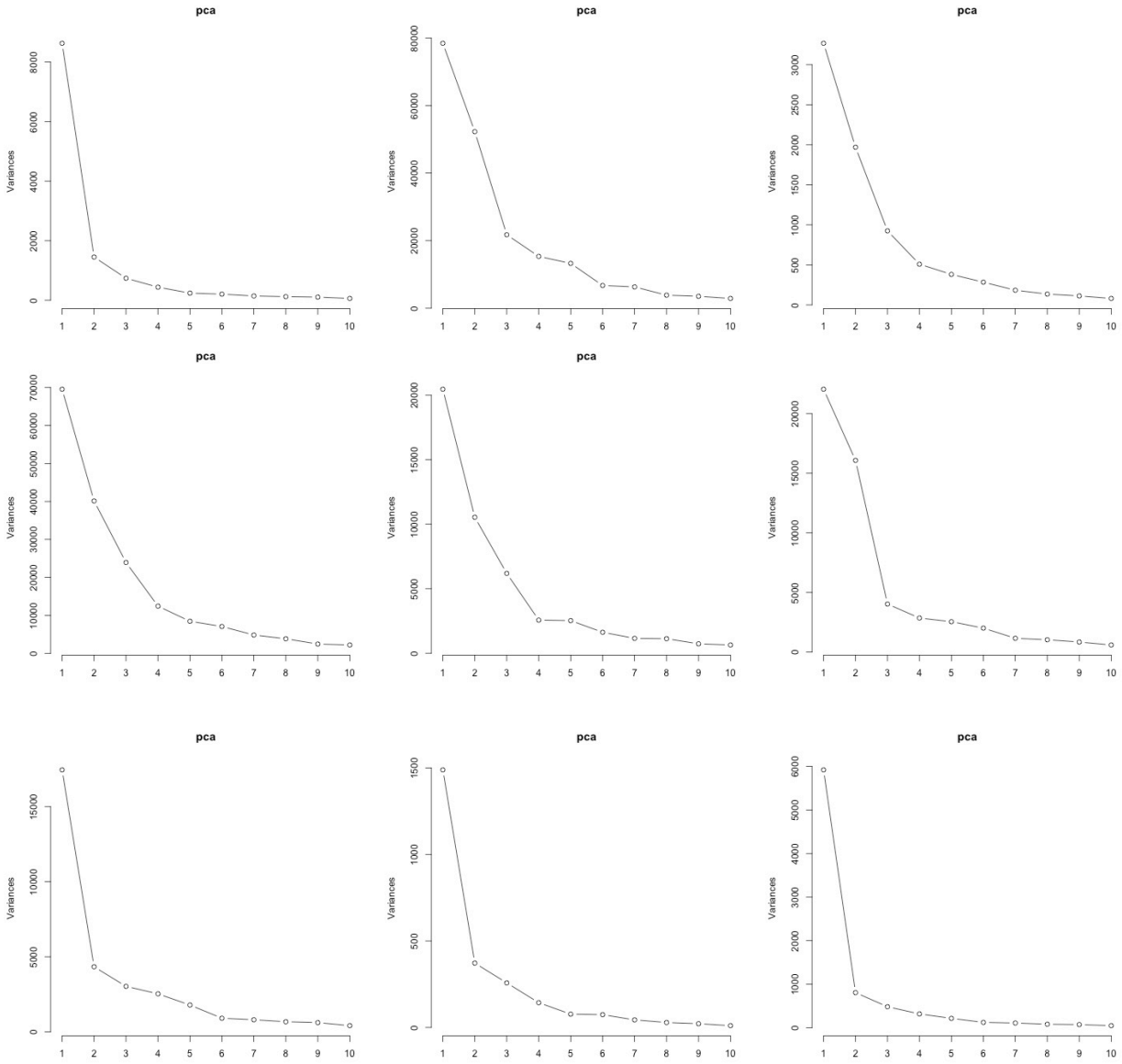

pca
Variances
0
500
1000
1500
2000
2500
3000


pca
Variances
0
2000
4000
6000
8000

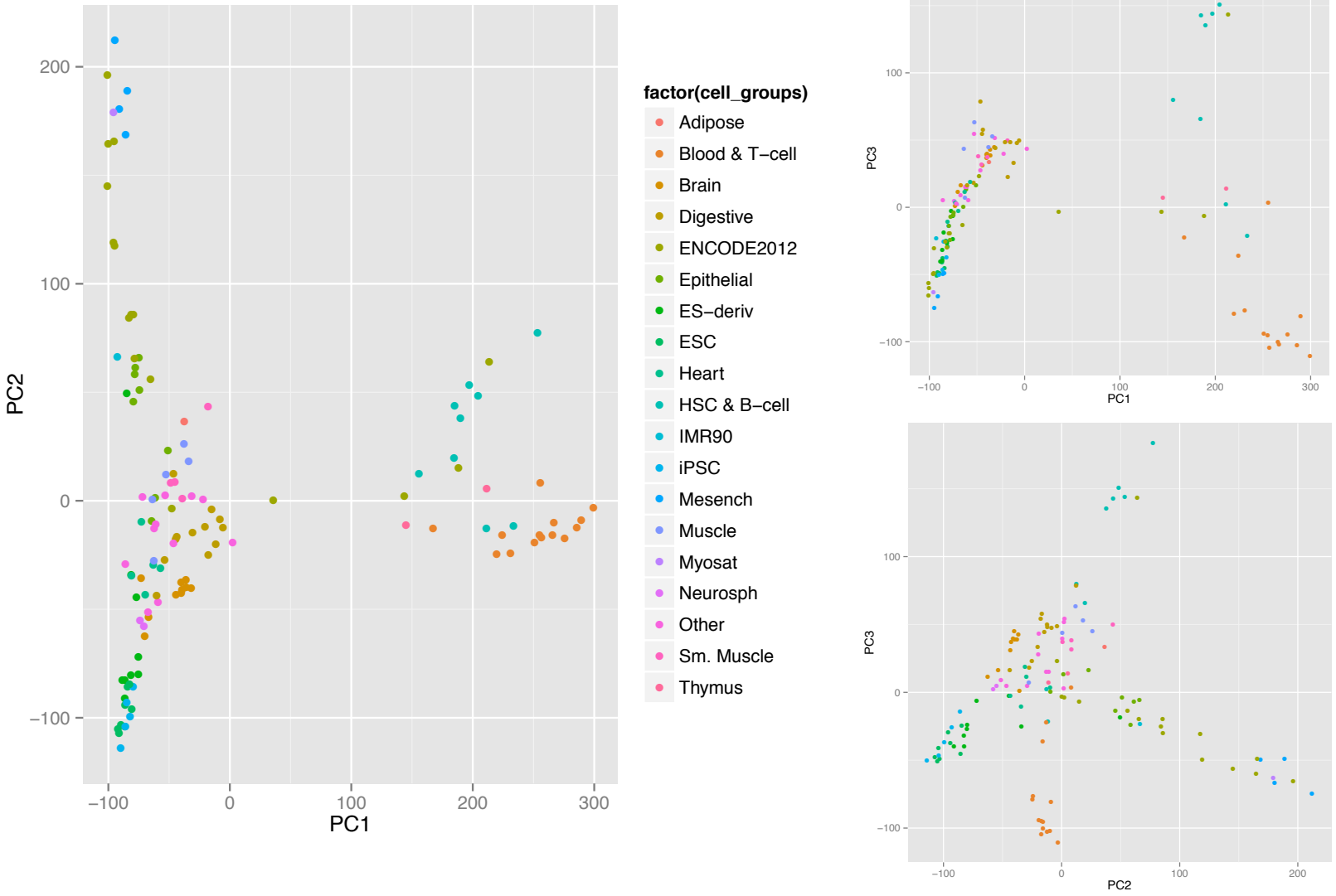

pca
Variances
0
20000
40000
60000


pca
Variances
0
5000
10000
15000
20000

Scree Plots


pca
Variances
pca
Variances
pca
Variances
pca
Variances
pca
Variances
pca
Variances
pca
Variances
pca
Variances
pca
Variances

K-Means Clustering



Subset Data: Remove “Other” and “ENCODE” Cell Types

GROUP					
IMR90	Blood & T-cell	Mesench	Brain	Digestive	ENCODE2012
ESC	Blood & T-cell	Mesench	Brain	Digestive	ENCODE2012
ESC	Blood & T-cell	Mesench	Brain	Digestive	ENCODE2012
ESC	Blood & T-cell	Mesench	Brain	Digestive	ENCODE2012
ESC	Blood & T-cell	Myosat	Adipose	Digestive	ENCODE2012
ESC	Blood & T-cell	Epithelial	Muscle	Digestive	ENCODE2012
ESC	Blood & T-cell	Epithelial	Muscle	Digestive	ENCODE2012
ESC	Blood & T-cell	Epithelial	Muscle	Digestive	ENCODE2012
ESC	Blood & T-cell	Epithelial	Muscle	Other	ENCODE2012
iPSC	Blood & T-cell	Epithelial	Muscle	Other	ENCODE2012
iPSC	Blood & T-cell	Epithelial	Heart	Other	ENCODE2012
iPSC	Blood & T-cell	Epithelial	Heart	Other	ENCODE2012
iPSC	Blood & T-cell	Epithelial	Heart	Other	
iPSC	Blood & T-cell	Neurosph	Heart	Other	
ES-deriv	HSC & B-cell	Neurosph	Heart	Other	
ES-deriv	HSC & B-cell	Thymus	Sm. Muscle	Other	
ES-deriv	HSC & B-cell	Thymus	Sm. Muscle	Other	
ES-deriv	HSC & B-cell	Brain	Sm. Muscle	Other	
ES-deriv	HSC & B-cell	Brain	Sm. Muscle	Other	
ES-deriv	HSC & B-cell	Brain	Digestive	ENCODE2012	
ES-deriv	HSC & B-cell	Brain	Digestive	ENCODE2012	
ES-deriv	HSC & B-cell	Brain	Digestive	ENCODE2012	
ES-deriv	HSC & B-cell	Brain	Digestive	ENCODE2012	

Subset Data: Remove “Other” and “ENCODE” Cell Types

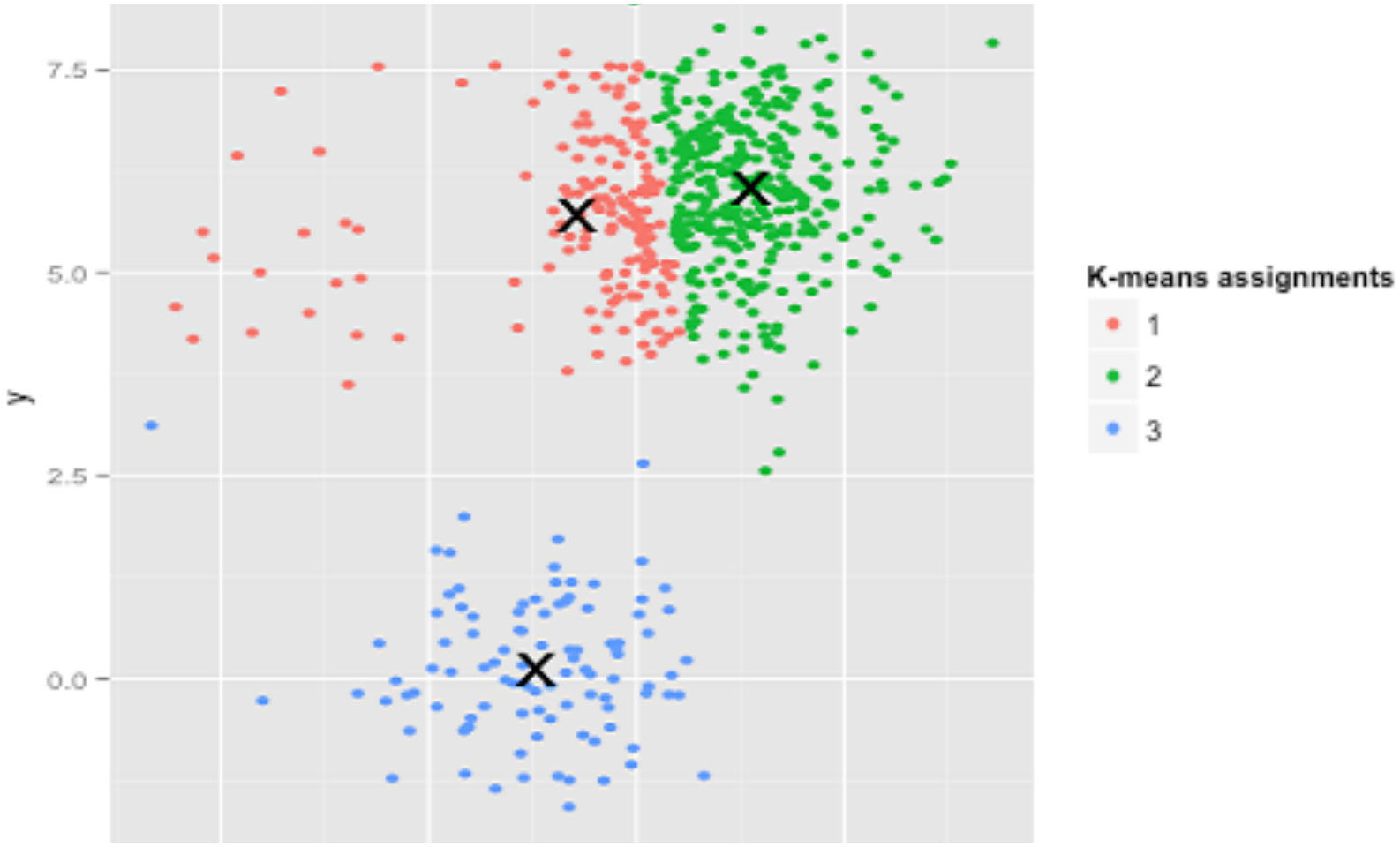
		GROUP		
IMR90	Blood & T-cell	Mesench	Brain	Digestive
ESC	Blood & T-cell	Mesench	Brain	Digestive
ESC	Blood & T-cell	Mesench	Brain	Digestive
ESC	Blood & T-cell	Mesench	Brain	Digestive
ESC	Blood & T-cell	Myosat	Adipose	Digestive
ESC	Blood & T-cell	Epithelial	Muscle	Digestive
ESC	Blood & T-cell	Epithelial	Muscle	Digestive
ESC	Blood & T-cell	Epithelial	Muscle	Digestive
ESC	Blood & T-cell	Epithelial	Muscle	
iPSC	Blood & T-cell	Epithelial	Muscle	
iPSC	Blood & T-cell	Epithelial	Heart	
iPSC	Blood & T-cell	Epithelial	Heart	
iPSC	Blood & T-cell	Epithelial	Heart	
iPSC	Blood & T-cell	Neurosph	Heart	
ES-deriv	HSC & B-cell	Neurosph	Heart	
ES-deriv	HSC & B-cell	Thymus	Sm. Muscle	
ES-deriv	HSC & B-cell	Thymus	Sm. Muscle	
ES-deriv	HSC & B-cell	Brain	Sm. Muscle	
ES-deriv	HSC & B-cell	Brain	Sm. Muscle	
ES-deriv	HSC & B-cell	Brain	Digestive	
ES-deriv	HSC & B-cell	Brain	Digestive	
ES-deriv	HSC & B-cell	Brain	Digestive	
ES-deriv	HSC & B-cell	Brain	Digestive	

Problem with K-means Algorithm

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

- K-means attempts to minimize the objective function J
- The term inside the summations gives the distance measure between the chosen point and the cluster center
- An issue with this algorithm is by minimizing the sum of squares within clusters, more weight is given to larger clusters

Problem with K means Algorithm



http://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

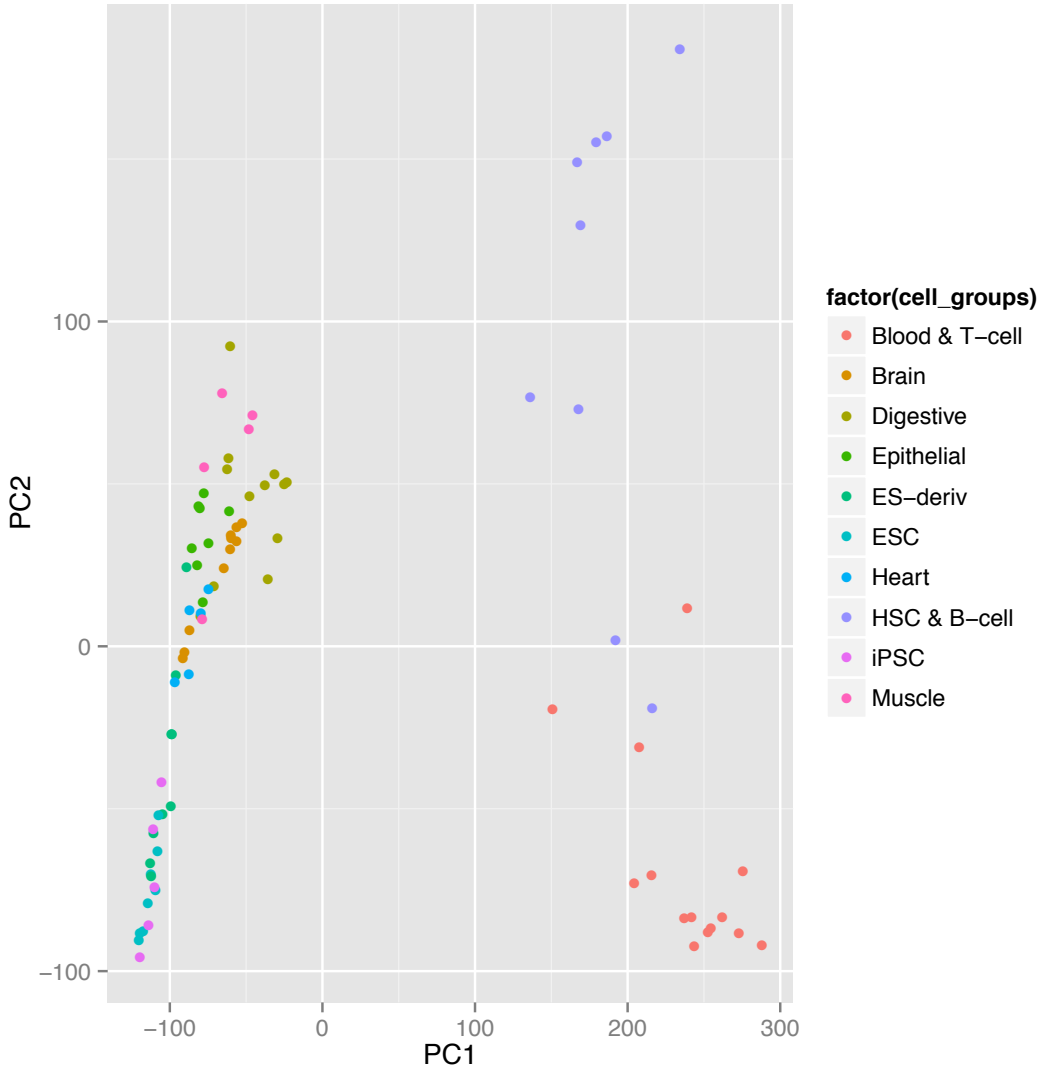
Further Subset Data: Remove Cell Types with Less than Five Entries

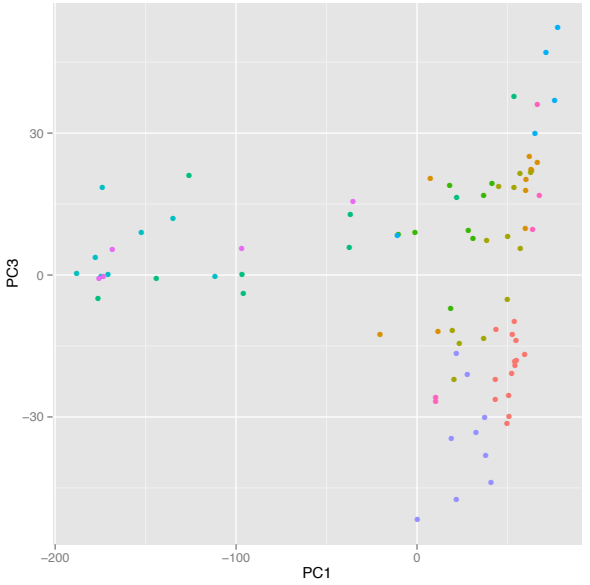
		GROUP		
IMR90	Blood & T-cell	Mesench	Brain	Digestive
ESC	Blood & T-cell	Mesench	Brain	Digestive
ESC	Blood & T-cell	Mesench	Brain	Digestive
ESC	Blood & T-cell	Mesench	Brain	Digestive
ESC	Blood & T-cell	Myosat	Adipose	Digestive
ESC	Blood & T-cell	Epithelial	Muscle	Digestive
ESC	Blood & T-cell	Epithelial	Muscle	Digestive
ESC	Blood & T-cell	Epithelial	Muscle	Digestive
ESC	Blood & T-cell	Epithelial	Muscle	
iPSC	Blood & T-cell	Epithelial	Muscle	
iPSC	Blood & T-cell	Epithelial	Heart	
iPSC	Blood & T-cell	Epithelial	Heart	
iPSC	Blood & T-cell	Epithelial	Heart	
iPSC	Blood & T-cell	Neurosph	Heart	
ES-deriv	HSC & B-cell	Neurosph	Heart	
ES-deriv	HSC & B-cell	Thymus	Sm. Muscle	
ES-deriv	HSC & B-cell	Thymus	Sm. Muscle	
ES-deriv	HSC & B-cell	Brain	Sm. Muscle	
ES-deriv	HSC & B-cell	Brain	Sm. Muscle	
ES-deriv	HSC & B-cell	Brain	Digestive	
ES-deriv	HSC & B-cell	Brain	Digestive	
ES-deriv	HSC & B-cell	Brain	Digestive	
ES-deriv	HSC & B-cell	Brain	Digestive	

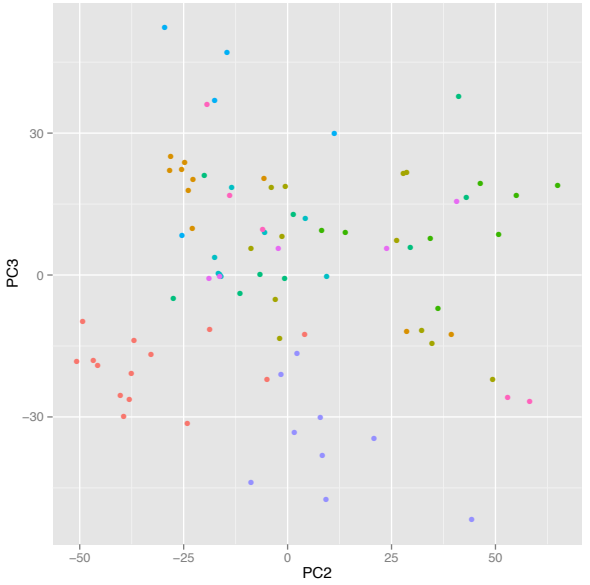
Further Subset Data: Remove Cell Types with Less than Five Entries

	GROUP			
	Blood & T-cell		Brain	Digestive
ESC	Blood & T-cell		Brain	Digestive
ESC	Blood & T-cell		Brain	Digestive
ESC	Blood & T-cell		Brain	Digestive
ESC	Blood & T-cell			Digestive
ESC	Blood & T-cell	Epithelial	Muscle	Digestive
ESC	Blood & T-cell	Epithelial	Muscle	Digestive
ESC	Blood & T-cell	Epithelial	Muscle	Digestive
ESC	Blood & T-cell	Epithelial	Muscle	
iPSC	Blood & T-cell	Epithelial	Muscle	
iPSC	Blood & T-cell	Epithelial	Heart	
iPSC	Blood & T-cell	Epithelial	Heart	
iPSC	Blood & T-cell	Epithelial	Heart	
iPSC	Blood & T-cell		Heart	
ES-deriv	HSC & B-cell		Heart	
ES-deriv	HSC & B-cell			
ES-deriv	HSC & B-cell			
ES-deriv	HSC & B-cell	Brain		
ES-deriv	HSC & B-cell	Brain		
ES-deriv	HSC & B-cell	Brain	Digestive	
ES-deriv	HSC & B-cell	Brain	Digestive	
ES-deriv	HSC & B-cell	Brain	Digestive	
ES-deriv	HSC & B-cell	Brain	Digestive	

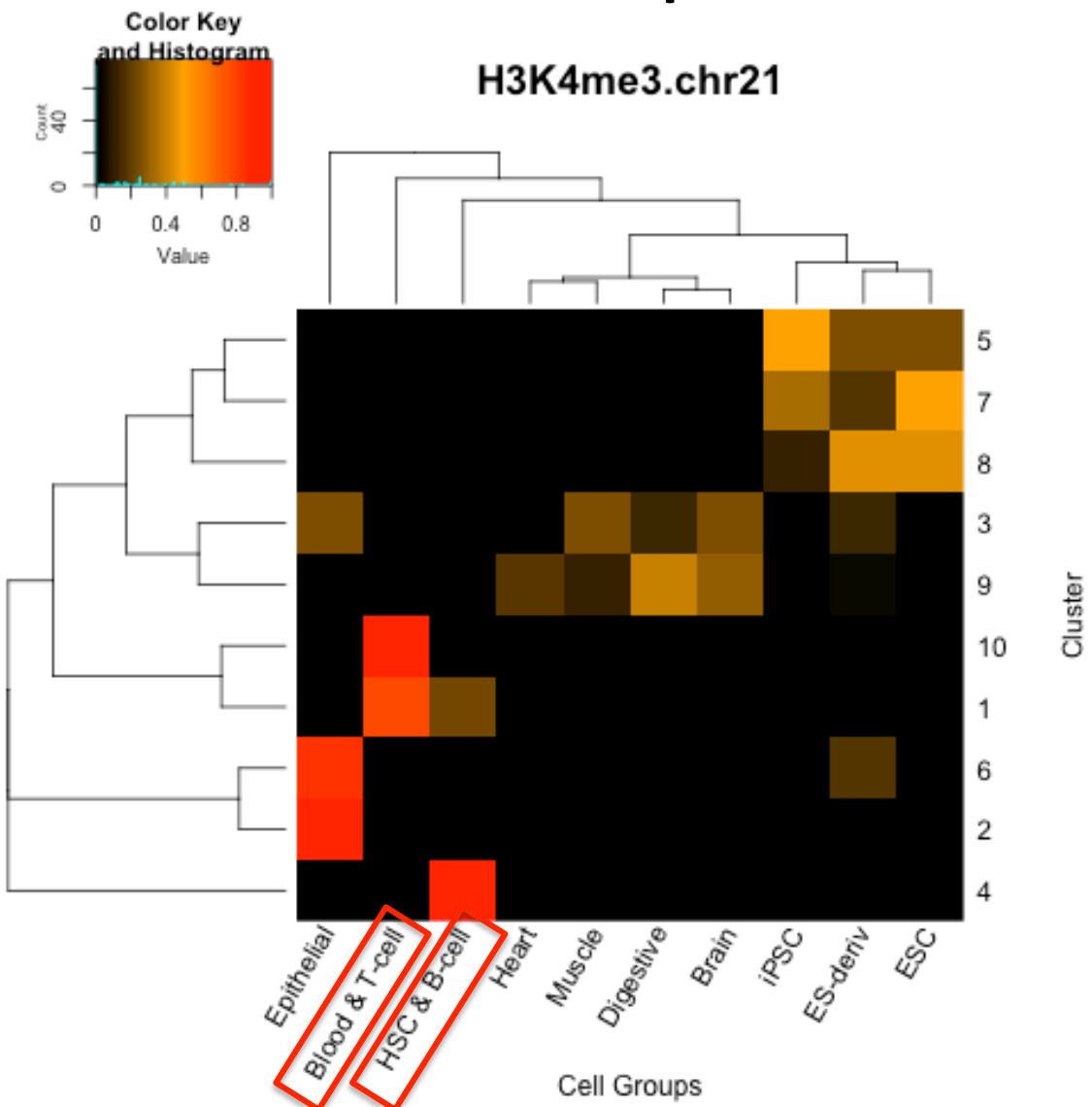
K-Means Clustering: Refined



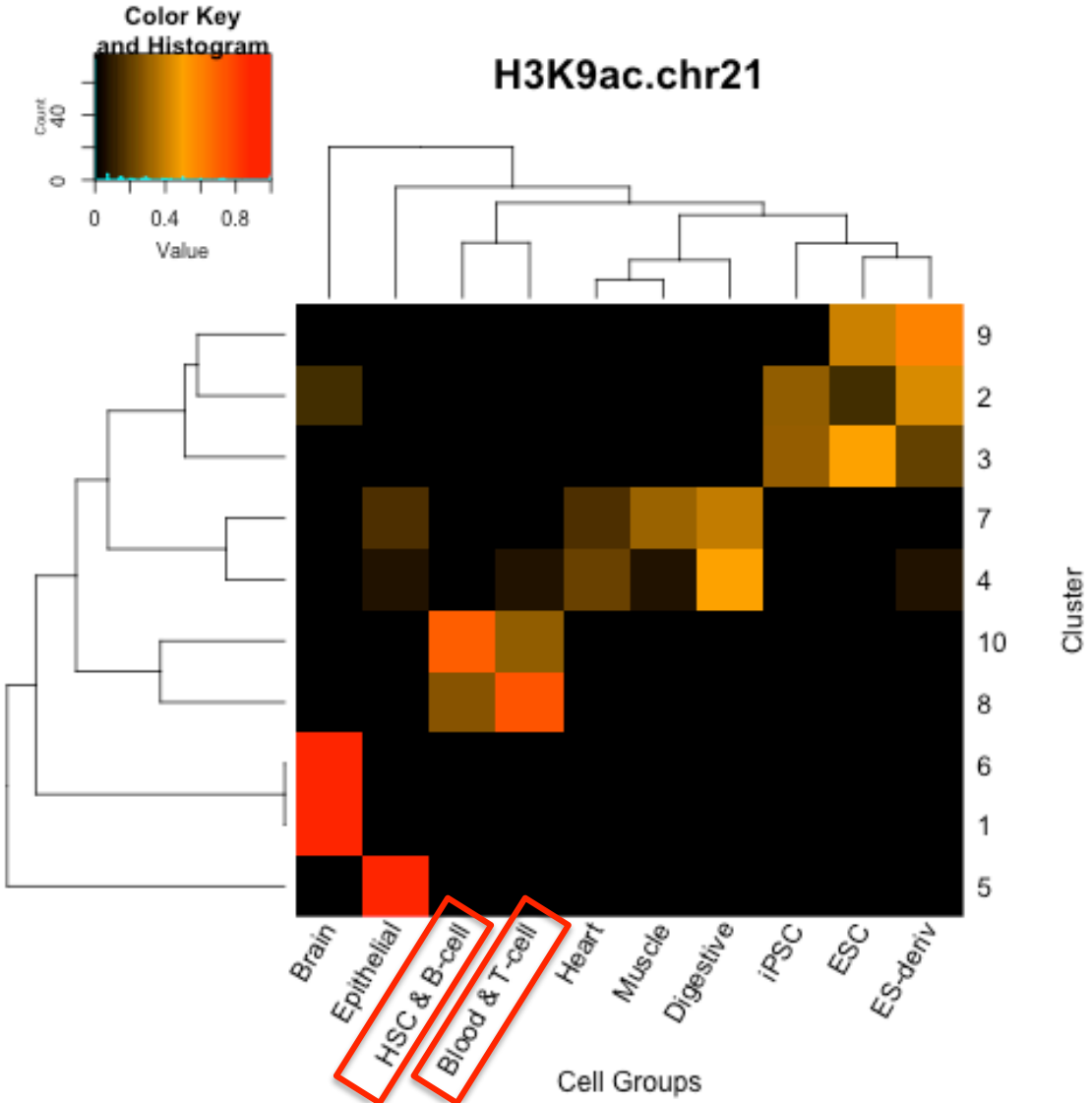




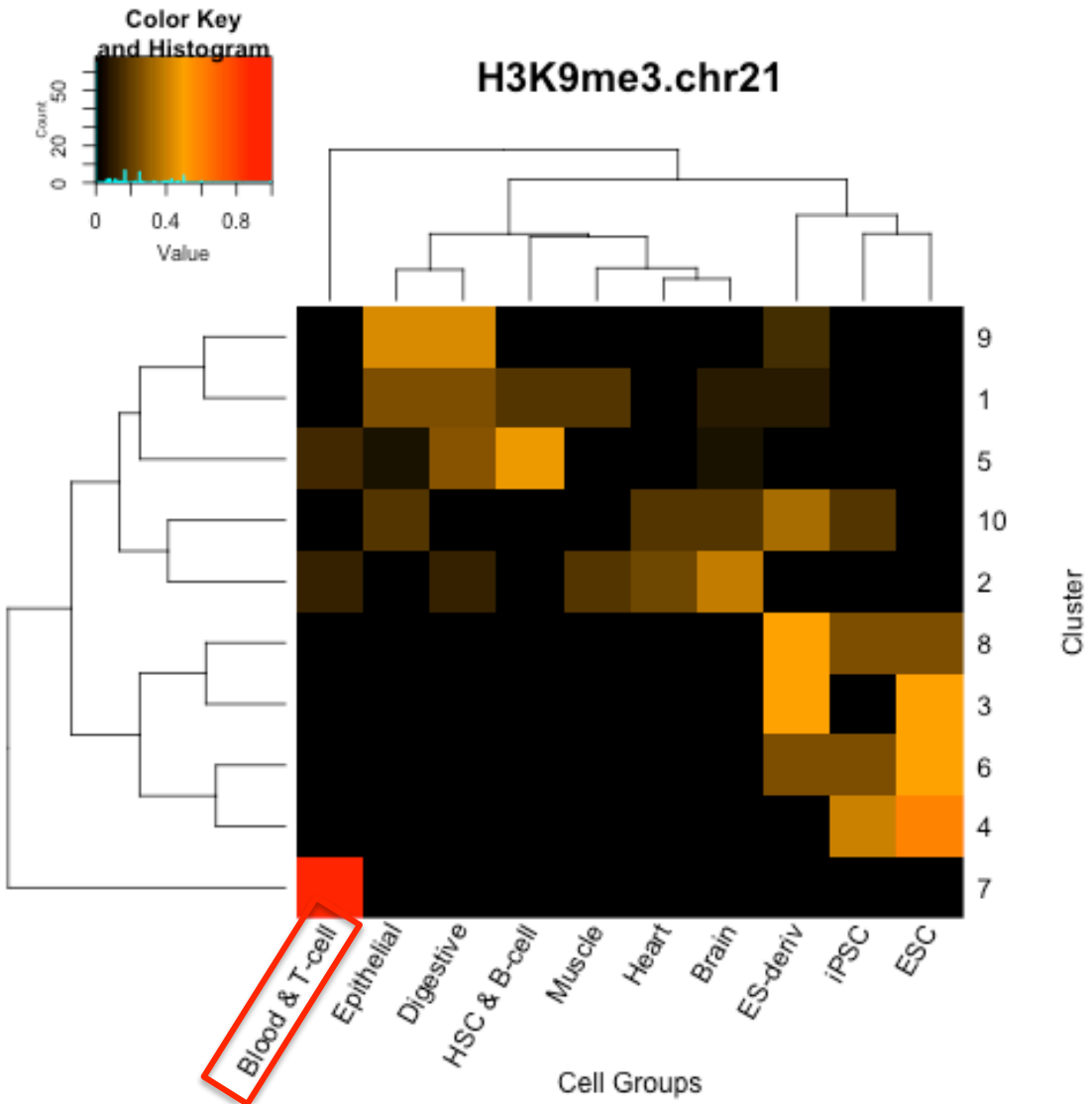
Visualizing Clustering Efficacy: Heatmaps



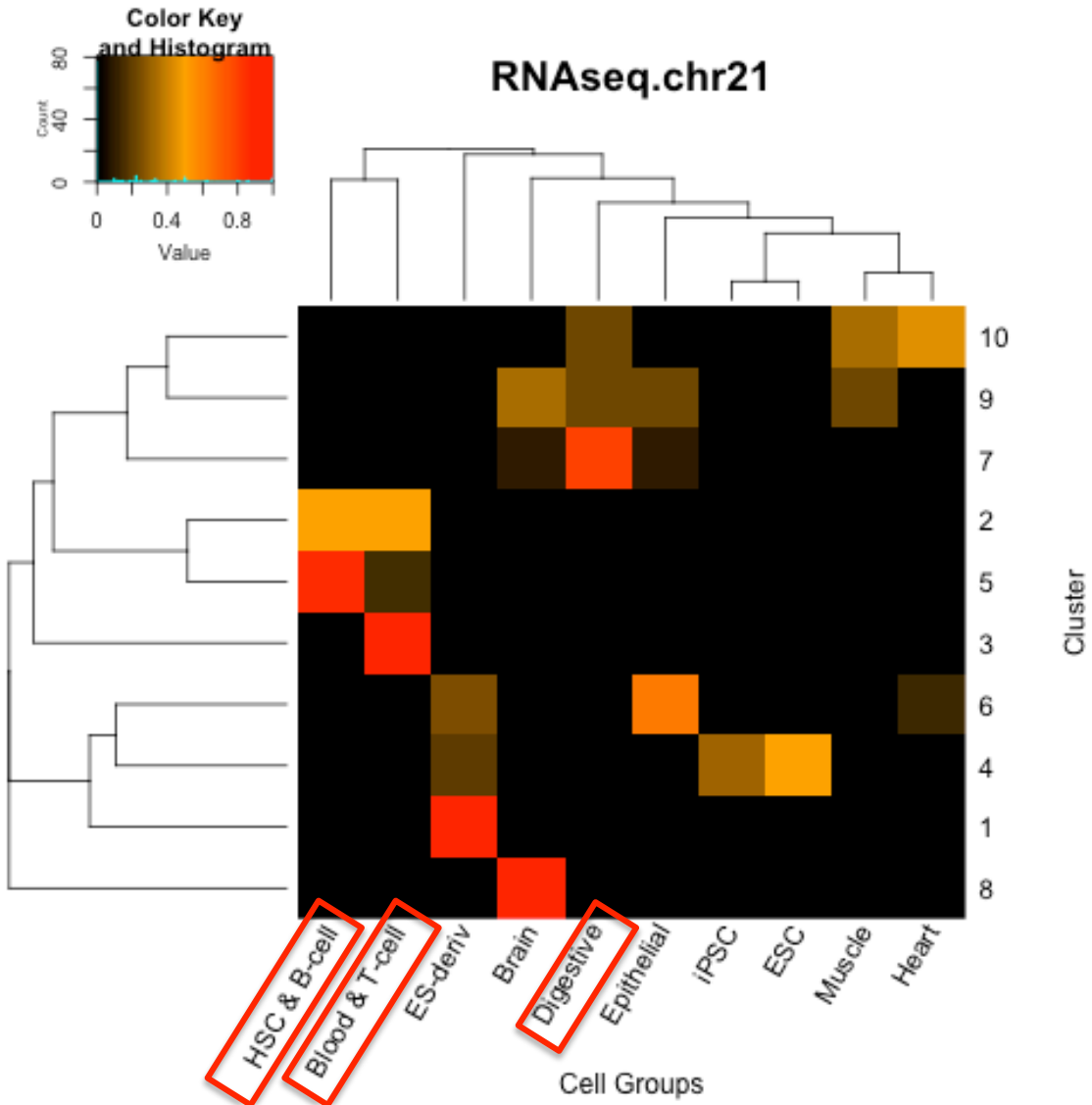
Visualizing Clustering Efficacy: Heatmaps



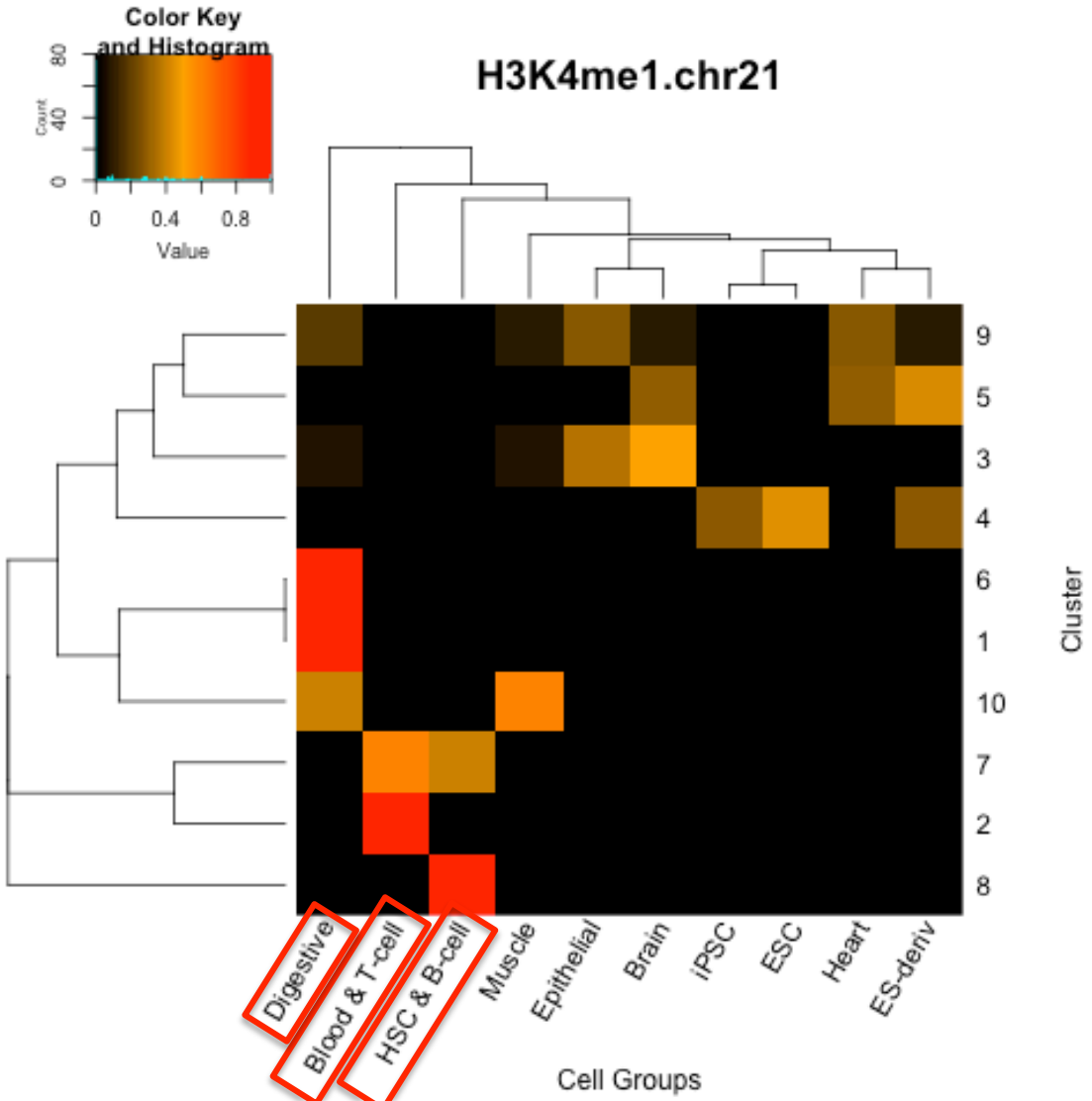
Visualizing Clustering Efficacy: Heatmaps



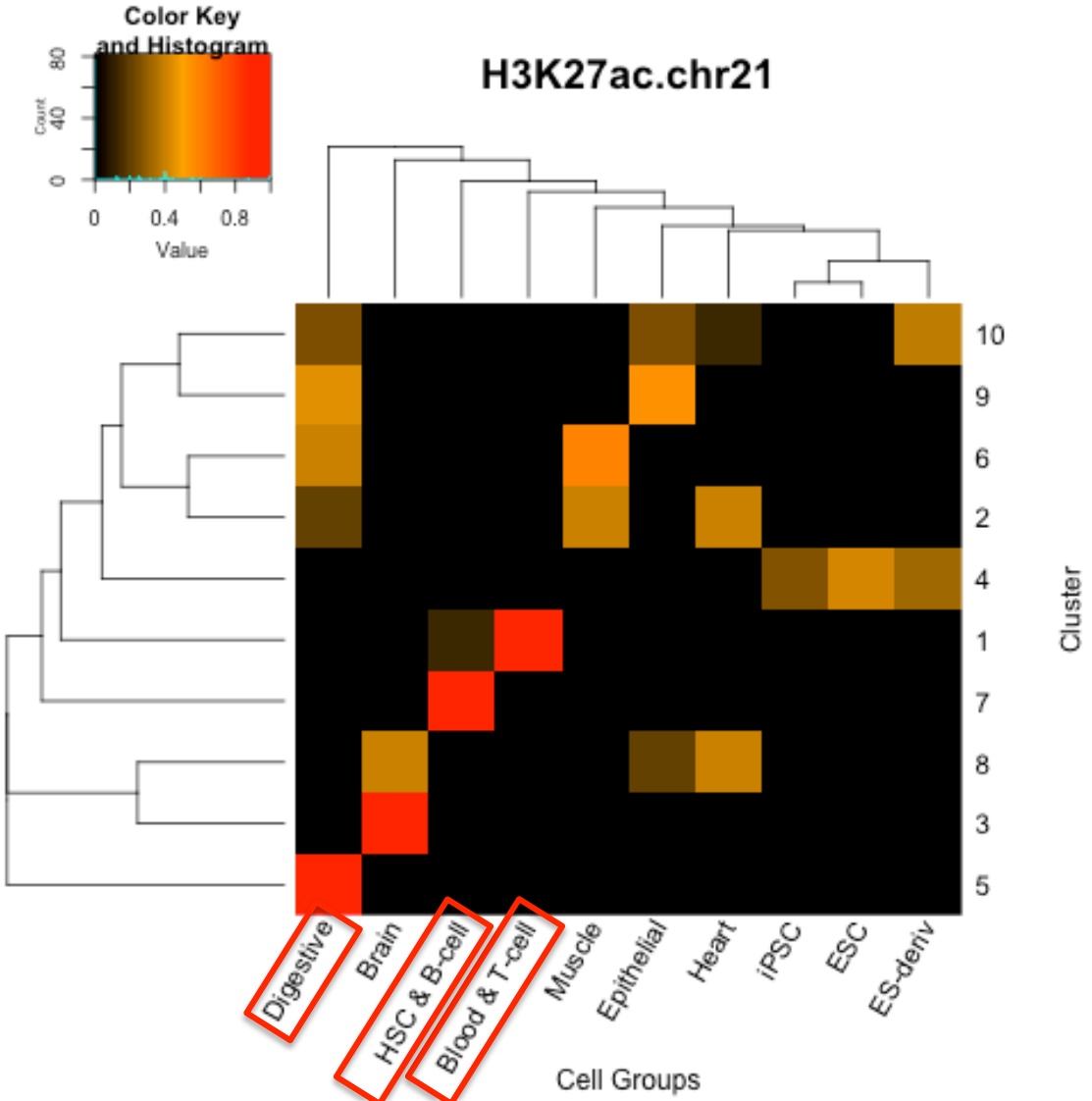
Visualizing Clustering Efficacy: Heatmaps



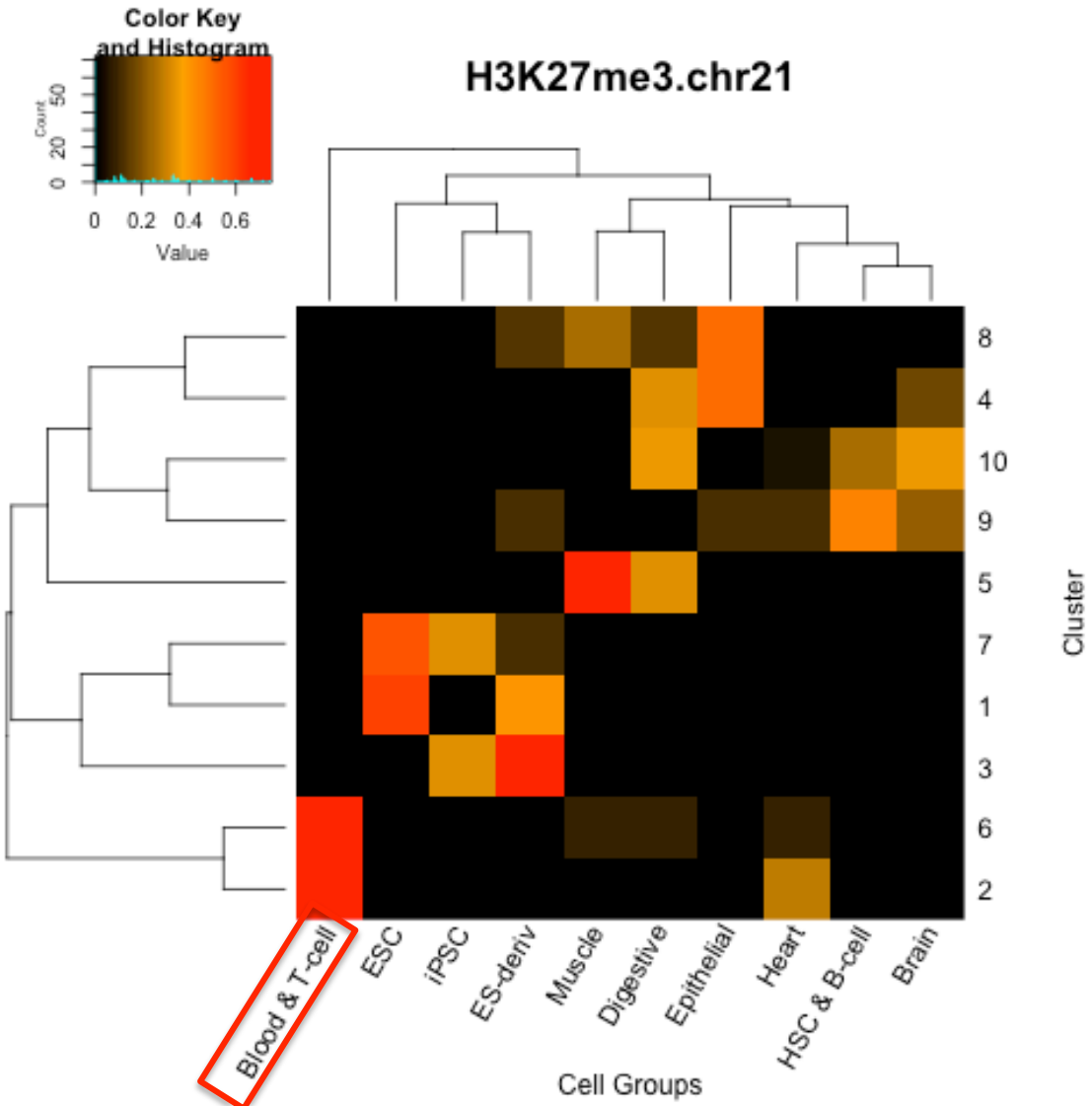
Visualizing Clustering Efficacy: Heatmaps



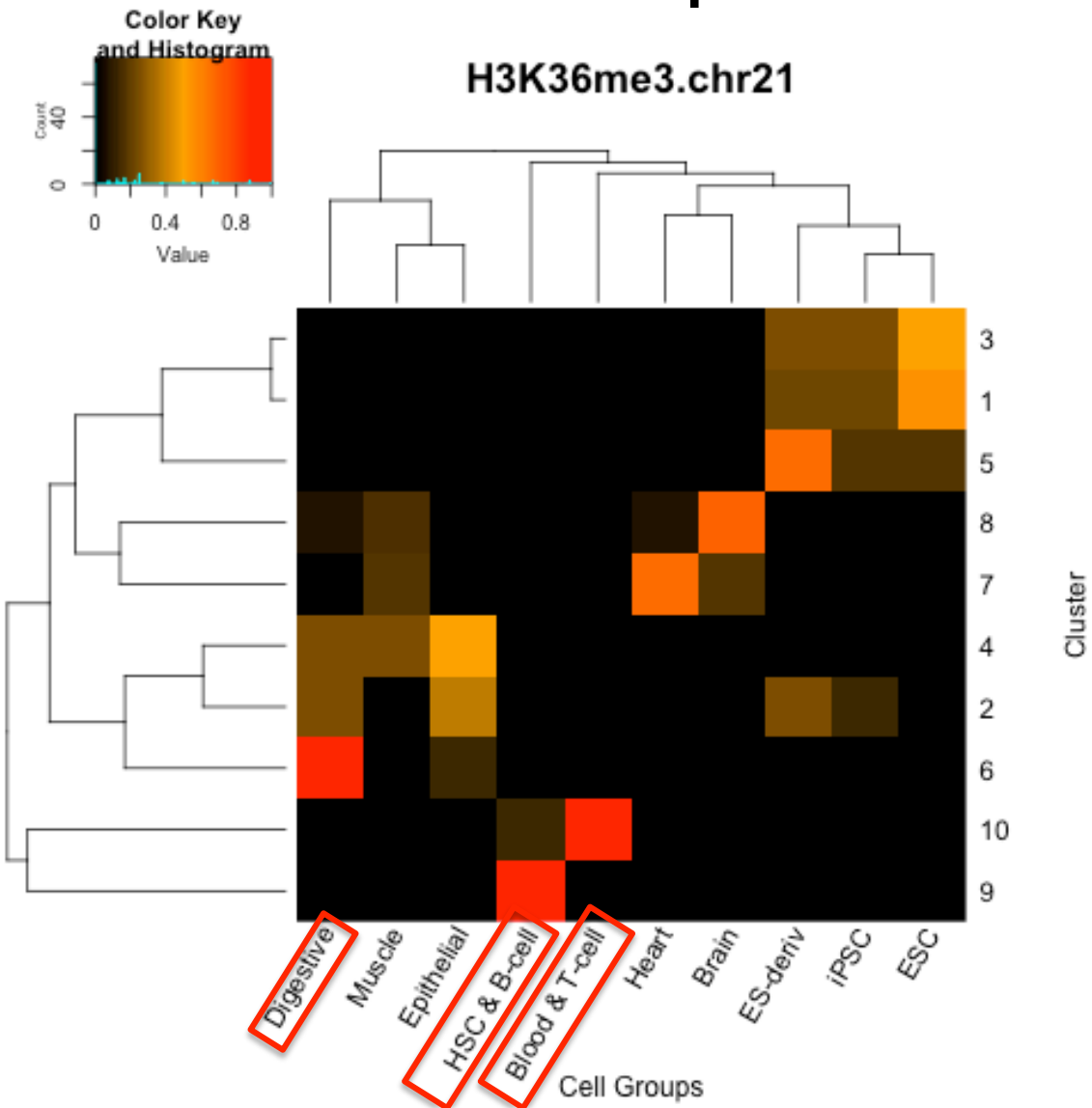
Visualizing Clustering Efficacy: Heatmaps



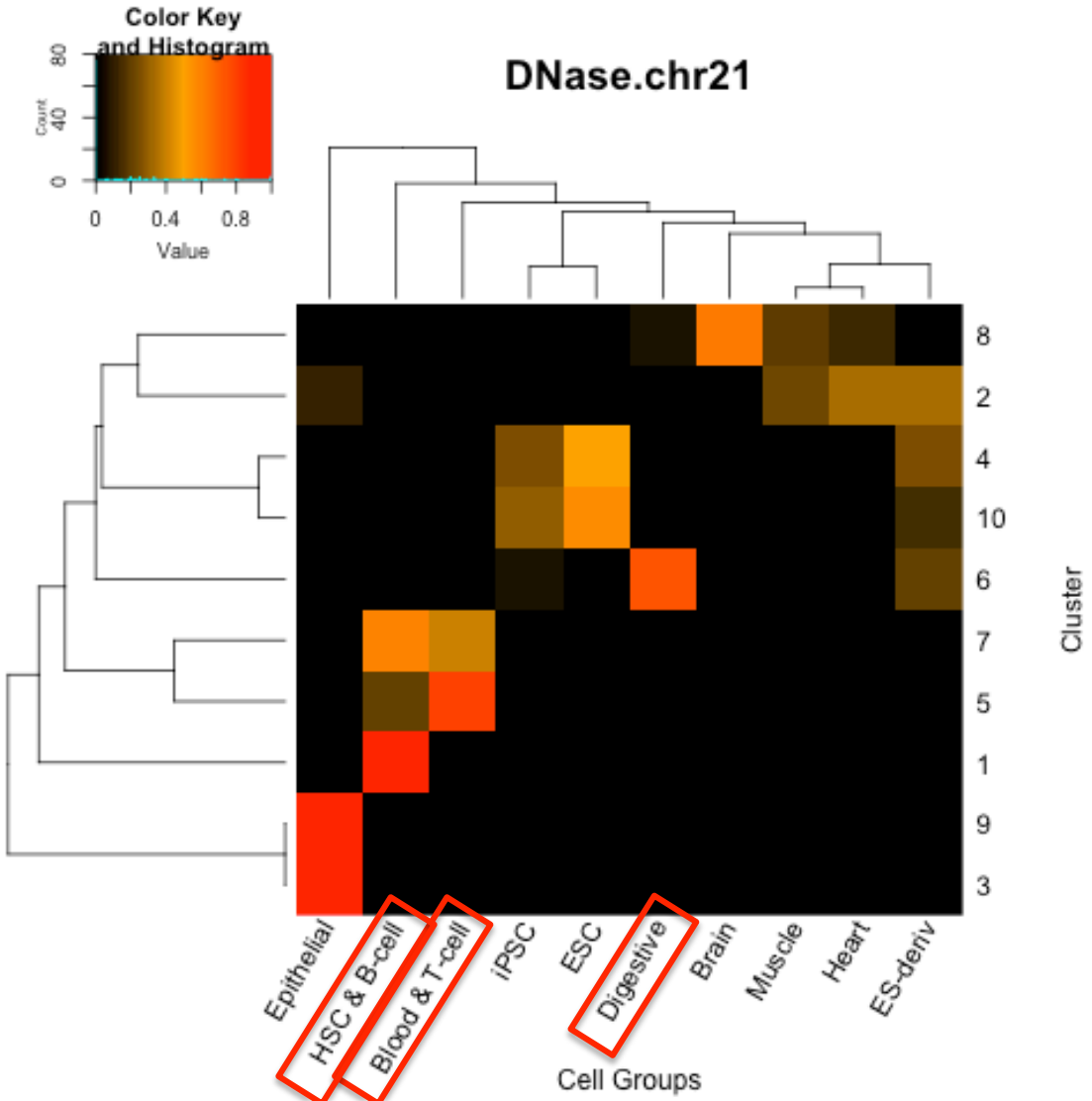
Visualizing Clustering Efficacy: Heatmaps



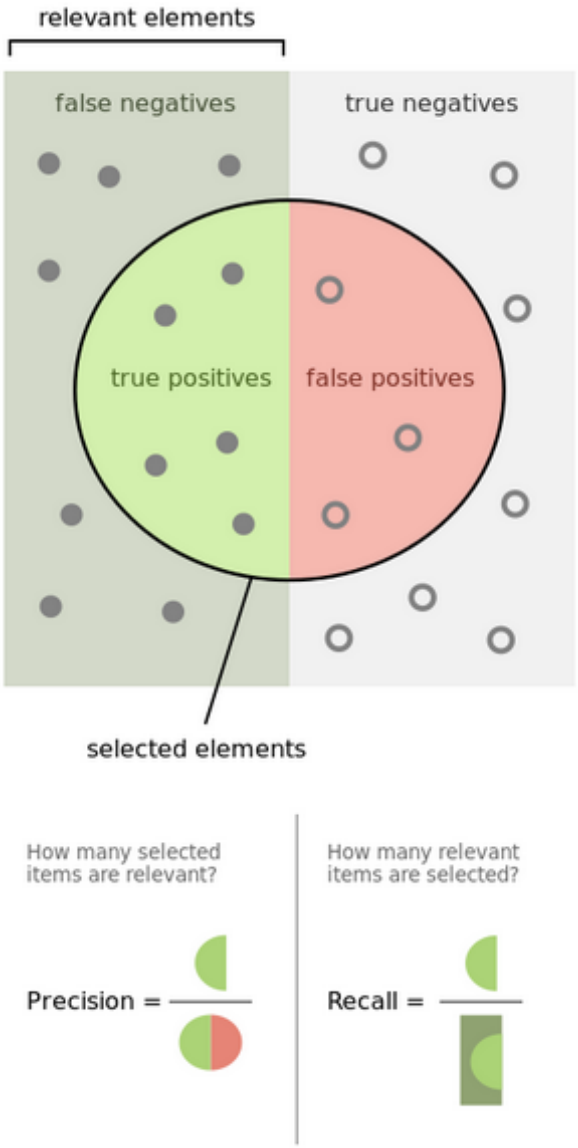
Visualizing Clustering Efficacy: Heatmaps



Visualizing Clustering Efficacy: Heatmaps



Precision and Recall



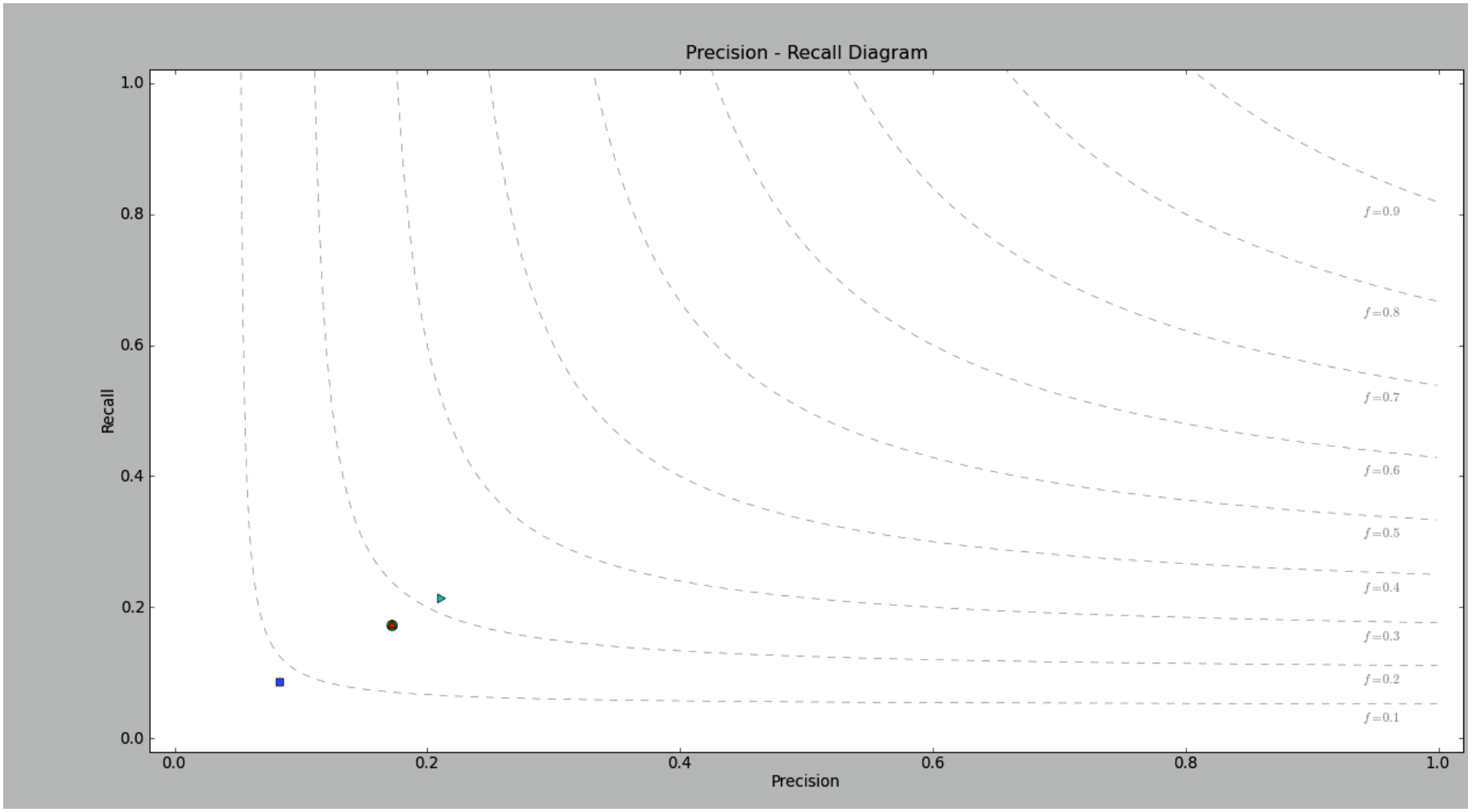
http://upload.wikimedia.org/wikipedia/commons/thumb/2/26/Precisionrecall.svg/2000px-Precisionrecall.svg.png

Evaluating Clustering Efficacy: F-Measure (F-Beta Score)

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- P – precision
- R – recall
- TP – True Positives
- FP – False Positives
- FN – False Negatives
- β – weights recall (for values >1, penalizes false negatives more strongly than false positives)

Precision & Recall vs F Score



http://aimotion.blogspot.com/2011/05/evaluating-recommender-systems.html

Evaluating Clustering Efficacy: F-Measure (F-Beta Score)

- **DNase.chr21 - 0.5944444**
- **H3K27ac.chr21 - 0.6277778**
- H3K27me3.chr21 - 0.3944444
- **H3K36me3.chr21 - 0.5805556**
- H3K4me1.chr21 - 0.5166667
- H3K4me3.chr21 - 0.5416667
- H3K9ac.chr21 - 0.4555556
- H3K9me3.chr21 - 0.3555556
- RNAseq.chr21 - 0.5083333

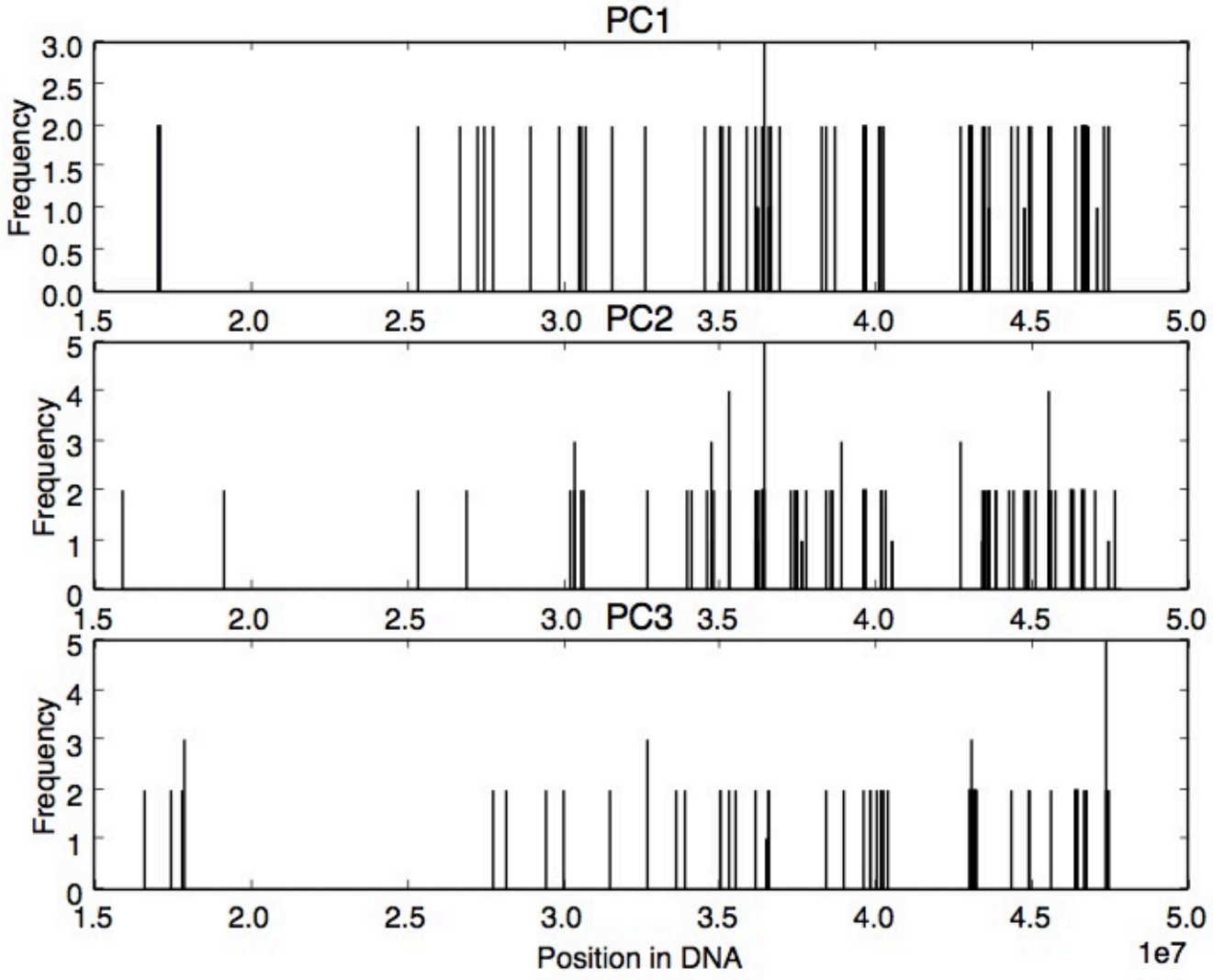
Analysis of Variance

- Question: Are there certain locations in the genome that explain the PCs and to these locations have an important biological function?

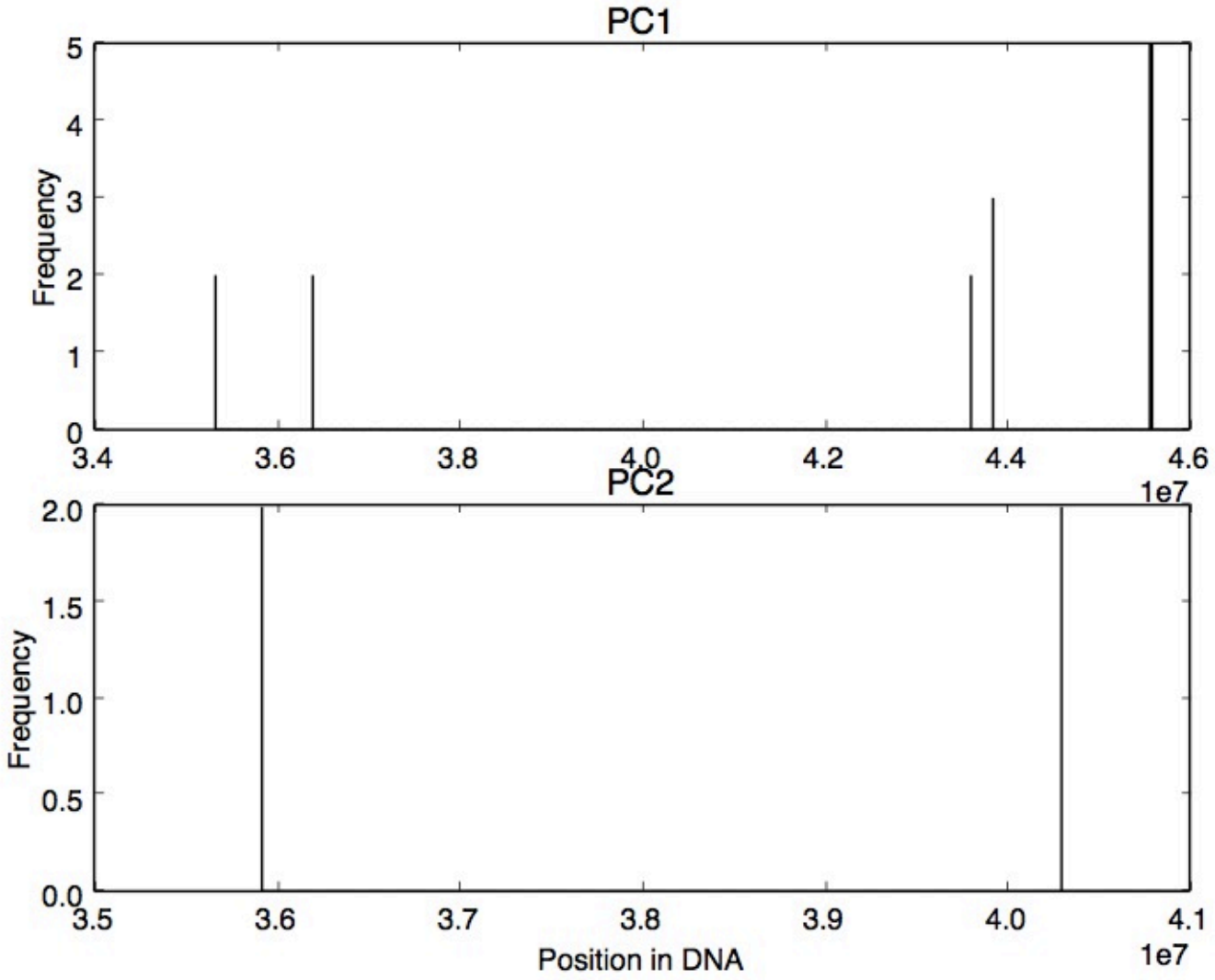
Analysis of Variance

- Sorted components of each PC in descending order of absolute value
- Looked at patterns in the components that contributed more than 5% to the variance
- Binned regions into 5kb segments

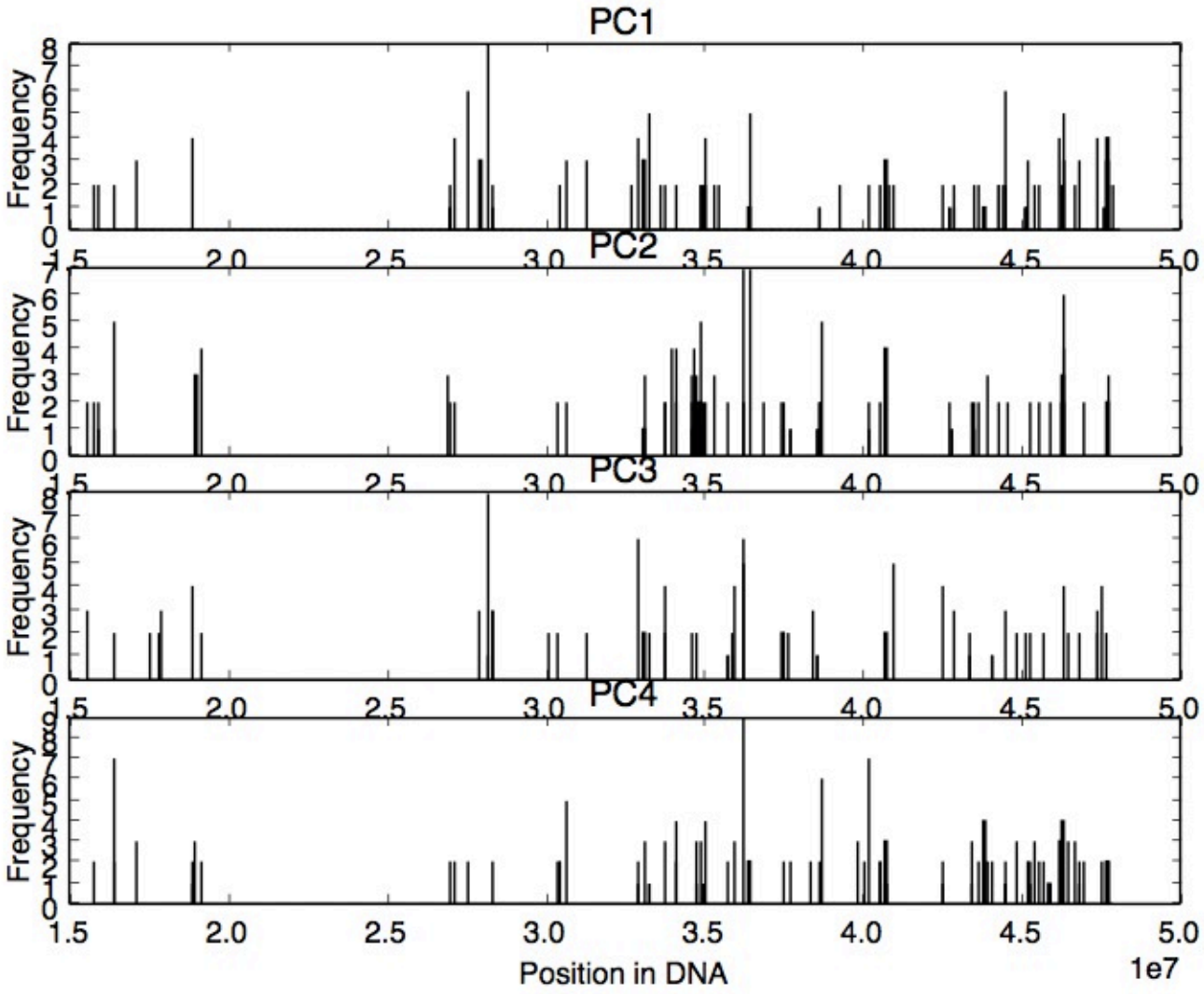
DNase


PC1
Frequency
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1.5
2.0
2.5
3.0
PC2
3.5
4.0
4.5
5.0
Frequency
5
4
3
2
1
0
1.5
2.0
2.5
3.0
PC3
3.5
4.0
4.5
5.0
Frequency
5
4
3
2
1
0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Position in DNA
1e7

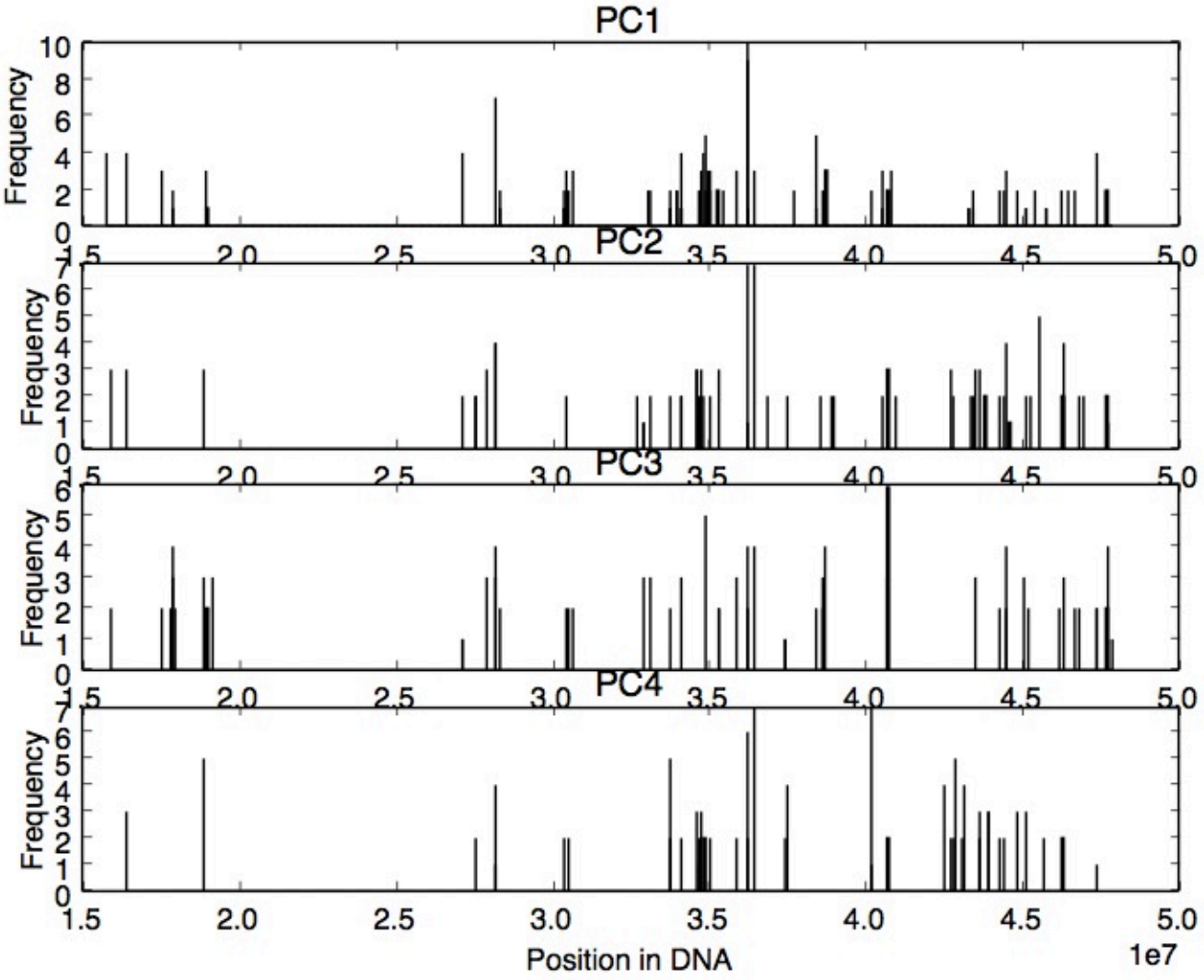
H3K4me1


PC1
Frequency
5
4
3
2
1
0
3.4
3.6
3.8
4.0
4.2
4.4
4.6
1e7
PC2
Frequency
2.0
1.5
1.0
0.5
0.0
3.5
3.6
3.7
3.8
3.9
4.0
4.1
1e7
Position in DNA

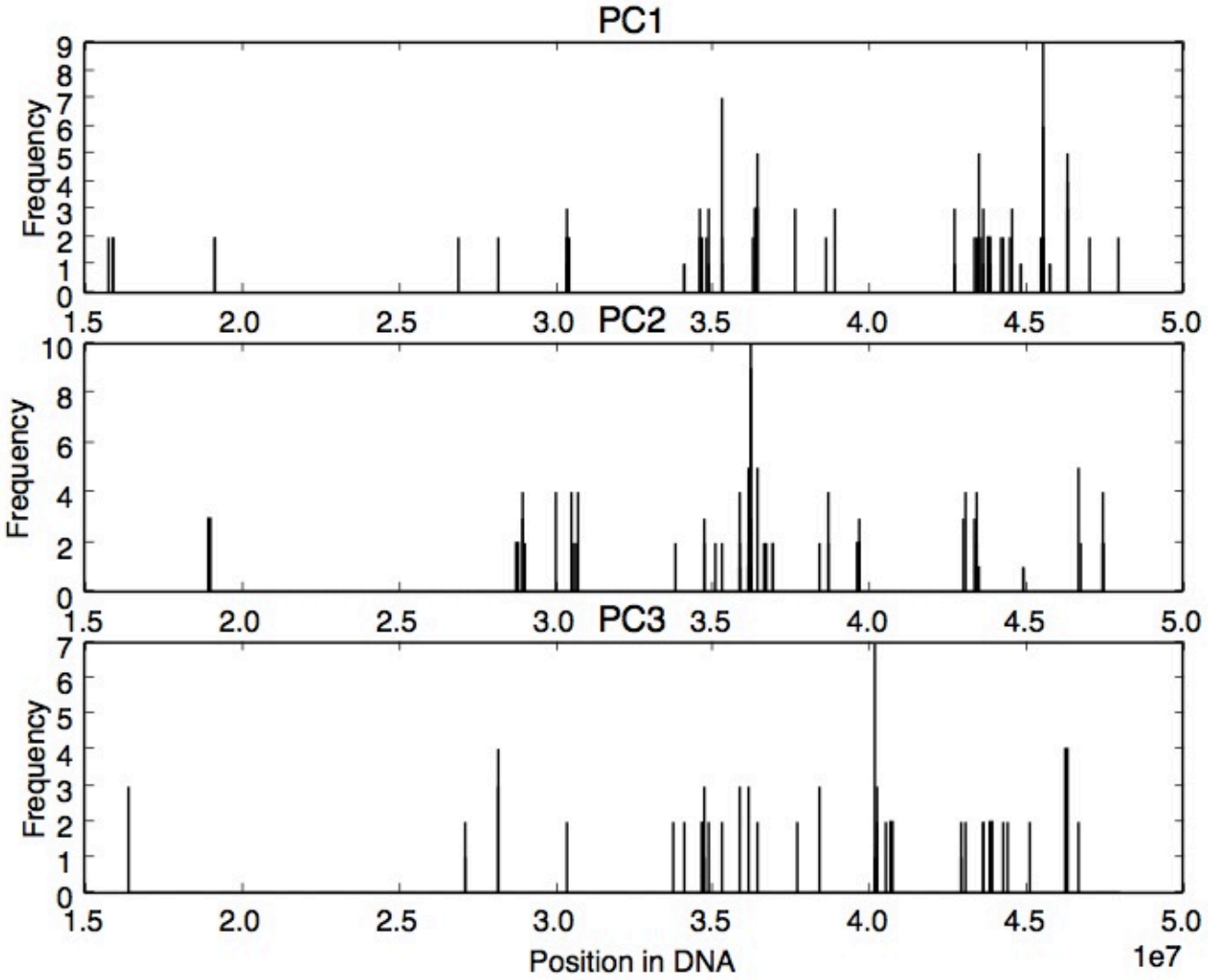
H3K4me3


PC1
PC2
PC3
PC4
Frequency
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Position in DNA
1e7

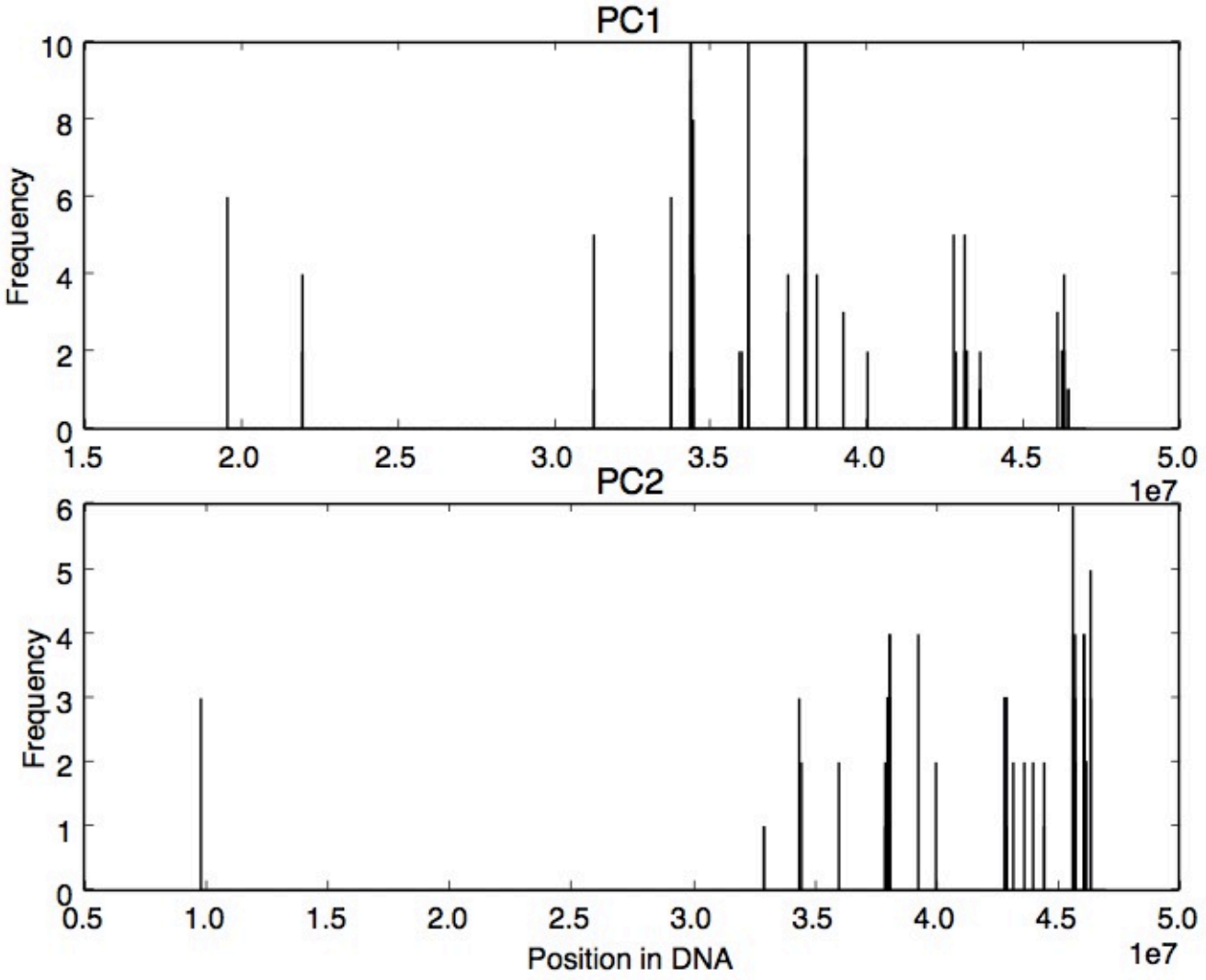
H3K9ac


PC1
PC2
PC3
PC4
Frequency
Position in DNA
1e7
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

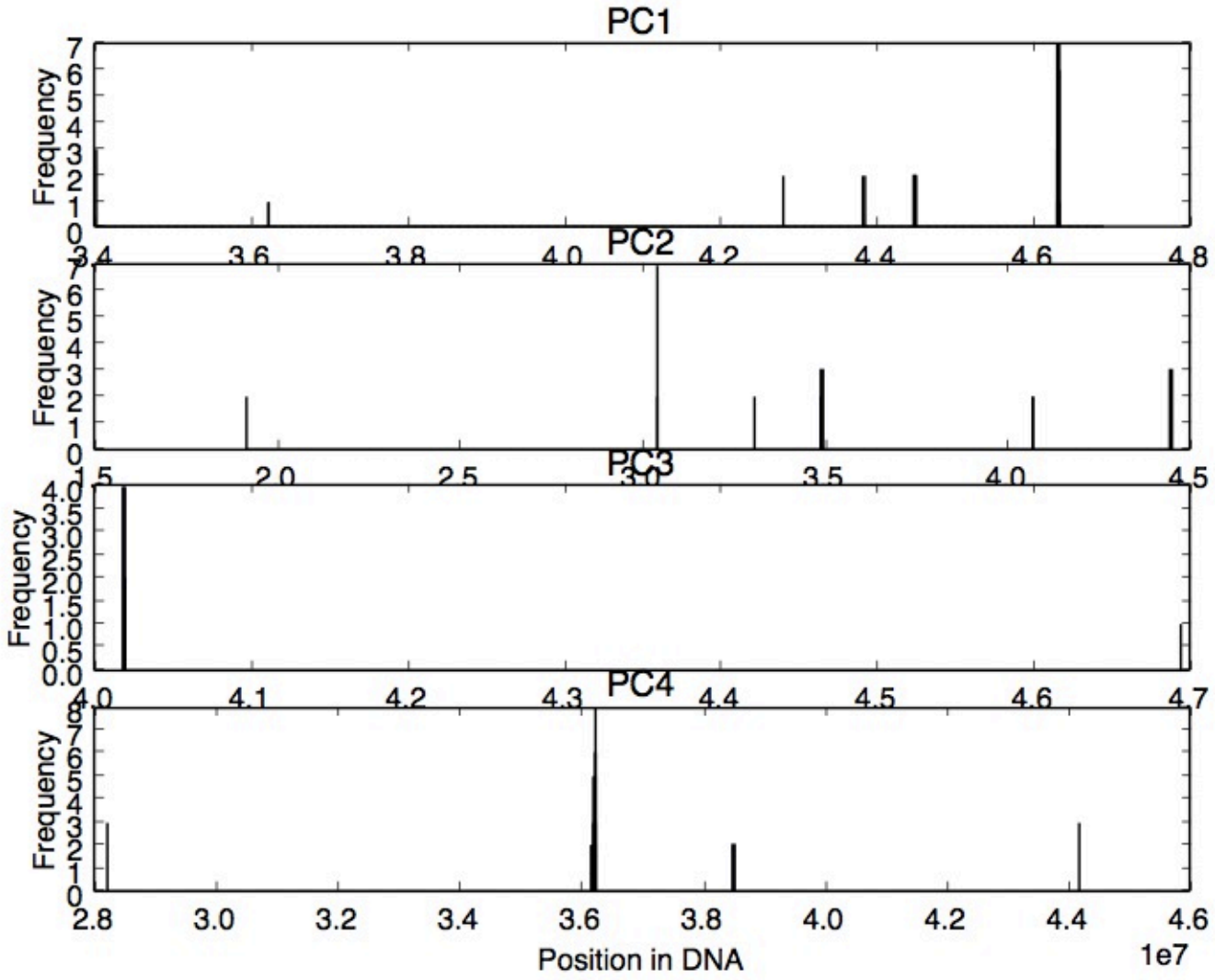
H3K27ac


PC1
Frequency
0 1 2 3 4 5 6 7 8 9
1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
PC2
Frequency
0 2 4 6 8 10
1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
PC3
Frequency
0 1 2 3 4 5 6 7
1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
Position in DNA
1e7

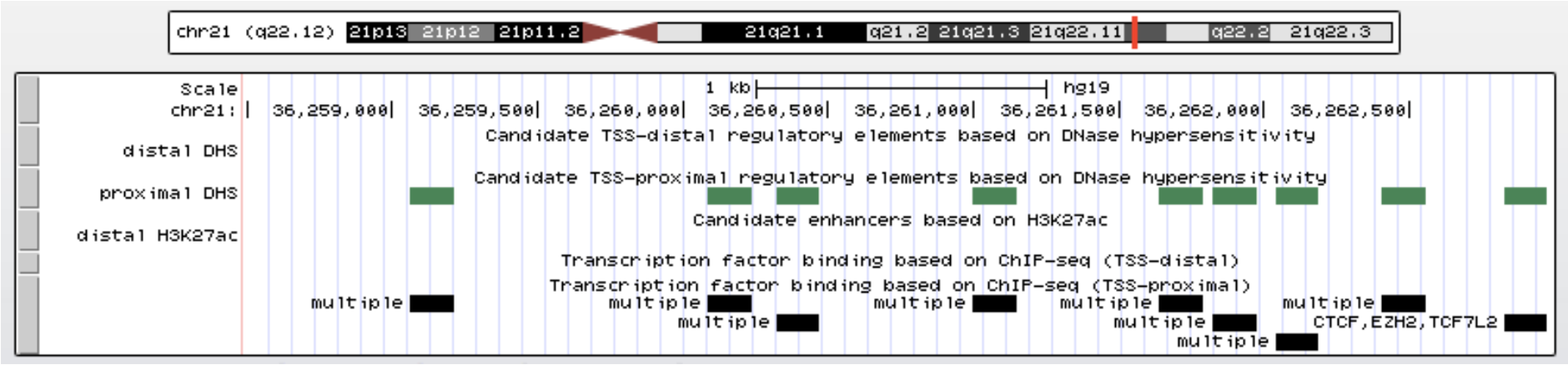
H3K27me3


PC1
Frequency
10
8
6
4
2
0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
1e7
PC2
Frequency
6
5
4
3
2
1
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
1e7
Position in DNA

H3K36me3



Comparison to annotation



- Conclusions:
  - Histone modifications showed the most cell-type specificity in Blood & T-cell, HSC & B-cell and Digestive cells
  - Certain genomic locations contribute more to the variance in cell types
- Future directions:
  - Use a Gaussian mixture model for clustering
  - Use higher resolution data to predict the biological function of important locations in the genome