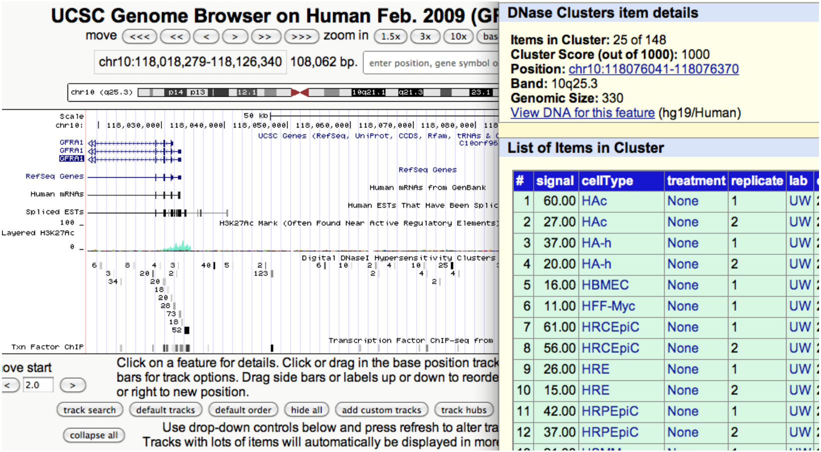
SparseData

Mike Love

Max Planck Institute for Molecular Genetics, Berlin

December 12, 2012

motivation



UCSC Genome Browser on Human Feb. 2009 (GR

chr10:118,018,279-118,126,340 108,062 bp.

Click on a feature for details. Click or drag in the base position track
bars for track options. Drag side bars or labels up or down to reorde
or right to new position.

Use drop-down controls below and press refresh to alter tra
Tracks with lots of items will automatically be displayed in more

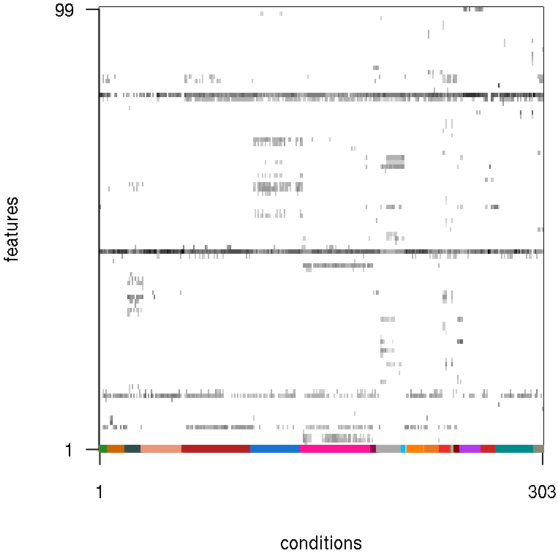
DNase Clusters item details

Items in Cluster: 25 of 148
Cluster Score (out of 1000): 1000
Position: chr10:118076041-118076370
Band: 10q25.3
Genomic Size: 330
View DNA for this feature (hg19/Human)

List of Items in Cluster

#	signal	cellType	treatment	replicate	lab
1	60.00	HAc	None	1	UW
2	27.00	HAc	None	2	UW
3	37.00	HA-h	None	1	UW
4	20.00	HA-h	None	2	UW
5	16.00	HBMEC	None	1	UW
6	11.00	HFF-Myc	None	1	UW
7	61.00	HRCEpiC	None	1	UW
8	56.00	HRCEpiC	None	2	UW
9	26.00	HRE	None	1	UW
10	15.00	HRE	None	2	UW
11	42.00	HRPEpiC	None	1	UW
12	37.00	HRPEpiC	None	2	UW

SparseDataSet



SparseData : basic statistics

- `sds <- calculateMeans(sds, nonzero.ratio=.1)`

SparseData : basic statistics

- `sds <- calculateMeans(sds, nonzero.ratio=.1)`
- `sds <- calculateTStats(sds)`

SparseData : basic statistics

- `sds <- calculateMeans(sds, nonzero.ratio=.1)`
- `sds <- calculateTStats(sds)`
- `fData(sds)[order(-tStats(sds)[["brain"]]),]`

SparseData : combine

- `z <- combine(x,y)`
- `z <- calculateMeans(z)`

SparseData : distances

- Pearson:

$$\frac{\langle \mathbf{x} - \bar{x}, \mathbf{y} - \bar{y} \rangle}{(n-1) s_x s_y}$$

- Euclidean:

$$\|\mathbf{x} - \mathbf{y}\|$$

- Pearson:

$$\frac{\langle \mathbf{x}, \mathbf{y} \rangle - n\bar{x}\bar{y}}{(n-1) s_x s_y}$$

- Euclidean:

$$\sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\langle \mathbf{x}, \mathbf{y} \rangle}$$

- cosine similarity:

$$\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}$$