Clustering and heatmaps

George W. Bell

Bioinformatics and Research Computing (BaRC)

http://barc.wi.mit.edu/hot_topics/

15 April 2021

What is clustering?

- The process of organizing objects (data) into groups based on similar features within the members (data points) of the group.
- An example of unsupervised learning, a tool of machine learning.
- Heatmaps are one way to visualize the results of clustering.

Why perform clustering?

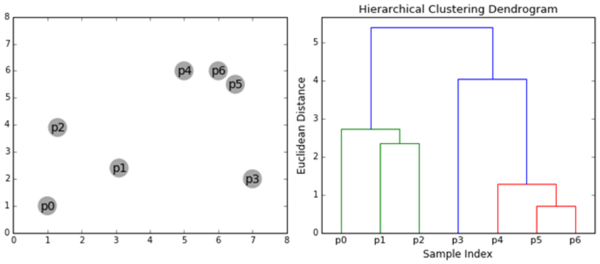
- Explore a large data matrix (such as of expression measurements)
- Organize similar items close to each other
- Get experiment-wide look at an interesting subset of data
- Visually identify patterns for further analysis
- Order features and/or samples in a sensible way
- Split features and/or samples into a predefined number of groups
- As one method of quality control

Why not perform clustering

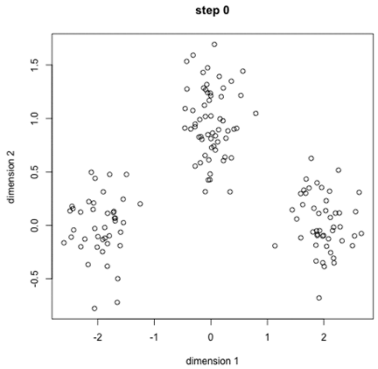
- Clustering is not a substitute for rigorous statistics.
- Sorting or categorizing objects in some other way may be more effective.
- A narrow matrix (few features or samples, for example) may not be informative
- “Getting clusters” is not news -- even noise can be clustered.

Selected clustering algorithms

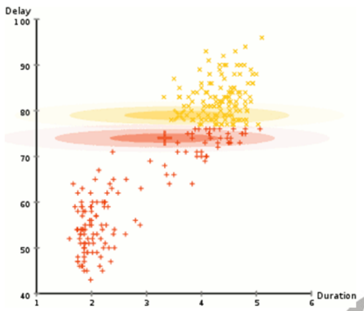
- Hierarchical Clustering algorithm
  - Divisive: start with one big cluster and split it
  - Agglomerative: start with each item in its own cluster and then merge most similar ones



- K-means algorithm
  - Choose k centroids (cluster centers)
  - Assign each point to a centroid
  - Get the centroid of each cluster
  - Repeat steps 2-3 until the centroids stop moving



- Expectation-Maximization (EM) algorithm
  - Optimize some function by iterating over items



- Graph-based methods (like the Louvain method in Seurat's scRNA-seq analysis package)

https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Considerations: Preparing data for clustering

- What data do you actually want to cluster?
  - If you answer ”everything” and it’s a lot, you may encounter some problems:
    - Hierarchical clustering scales with the square of the number of objects (so a 2-fold increase in number of objects take 4 times longer to cluster)
    - When you try to visualize the results, the “signal” may end up obstructed by the “noise”.
  - Merge biological replicates or keep separate?
- How do you want to weight each object your clustering?
  - Some programs let you assign different objects different weights
  - Should larger measurements be given more weight that lower measurements?
- Does your algorithm tolerate missing data?
- Clustering ratios is often more informative than clustering levels

Practical ideas:
Preparing data for clustering

- Start with the most “interesting” subset of your data (ex: genes (rows) x samples (columns))
- Reduce the range of the measurements by log-transformation
- Further transform for better comparisons
  - Mean/median center => subtract the mean or median of each log2-transformed row from the log2-transformed values
  - This creates log2 ratios of each measurement relative to the summary of this item across all samples (so positive and negative values)

Preparing data (example)

- Step 1: Select an interesting submatrix:

Gene name	Sample1	Sample2	Sample3	Sample4	Sample5
Gene1	4132	2042	4003	3076	3168
Gene2	41	49	79	77	377
Gene3	763	719	1203	623	879
Gene4	650	539	1086	392	927
Gene5	160	294	333	231	331

Preparing data (example)

- Step 2: log2-transform

Gene name	**Sample1**	**Sample2**	**Sample3**	**Sample4**	**Sample5**	mean
Gene1	12.01	11.00	11.97	11.59	11.63	11.64
Gene2	5.36	5.61	6.30	6.27	8.56	6.42
Gene3	9.58	9.49	10.23	9.28	9.78	9.67
Gene4	9.34	9.07	10.08	8.61	9.86	9.39
Gene5	7.32	8.20	8.38	7.85	8.37	8.02

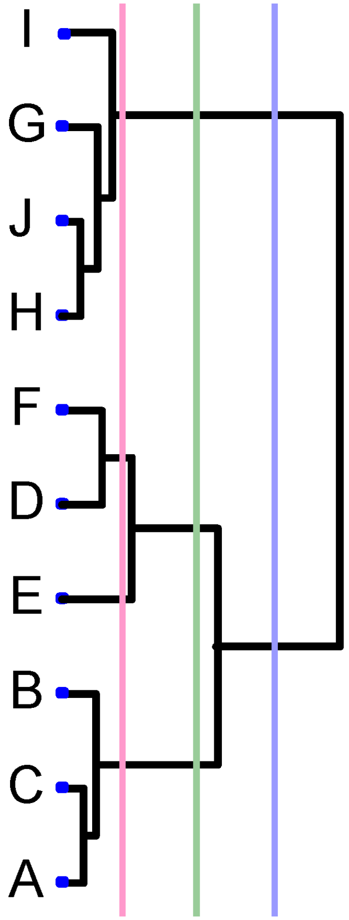
Preparing data (example)

- Step 3: mean-center

Gene name	Sample1	Sample2	Sample3	Sample4	Sample5
Gene1	0.37	-0.64	0.33	-0.05	-0.01
Gene2	-1.06	-0.81	-0.12	-0.15	2.14
Gene3	-0.10	-0.18	0.56	-0.39	0.11
Gene4	-0.05	-0.32	0.69	-0.78	0.46
Gene5	-0.70	0.17	0.35	-0.17	0.35

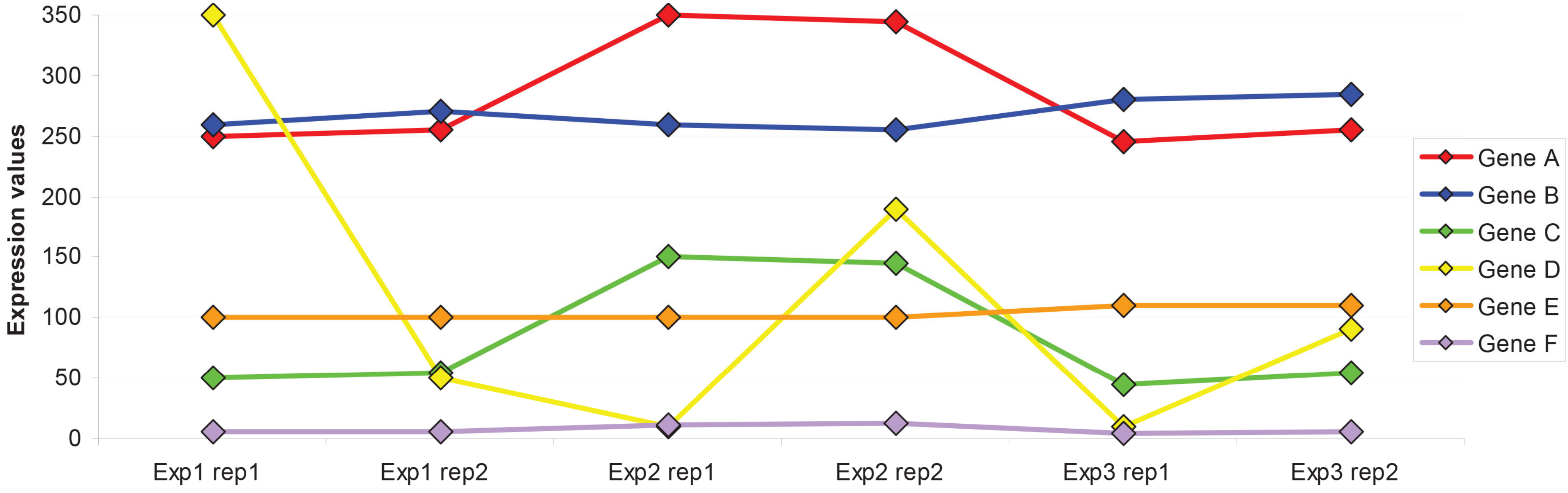
Hierarchical clustering

- Most commonly, this doesn’t produce discrete clusters but rather dendrograms (trees) reflecting relationships between rows and columns
- Choices one needs to make:
  - Matrix pre-processing details
  - Similarity metric: How do we quantify similarity?
  - Clustering method: After merging items into a temporary cluster, how is the similarity metric applied to this “pseudo-item”?
- Output dendrograms can be sliced to partition items into clusters (if desired).



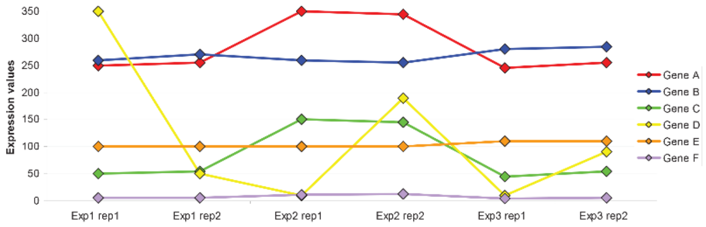
Hierarchical clustering: similarity metrics

- How do we quantify similarity?



Hierarchical clustering: similarity metrics

How do we quantify similarity?

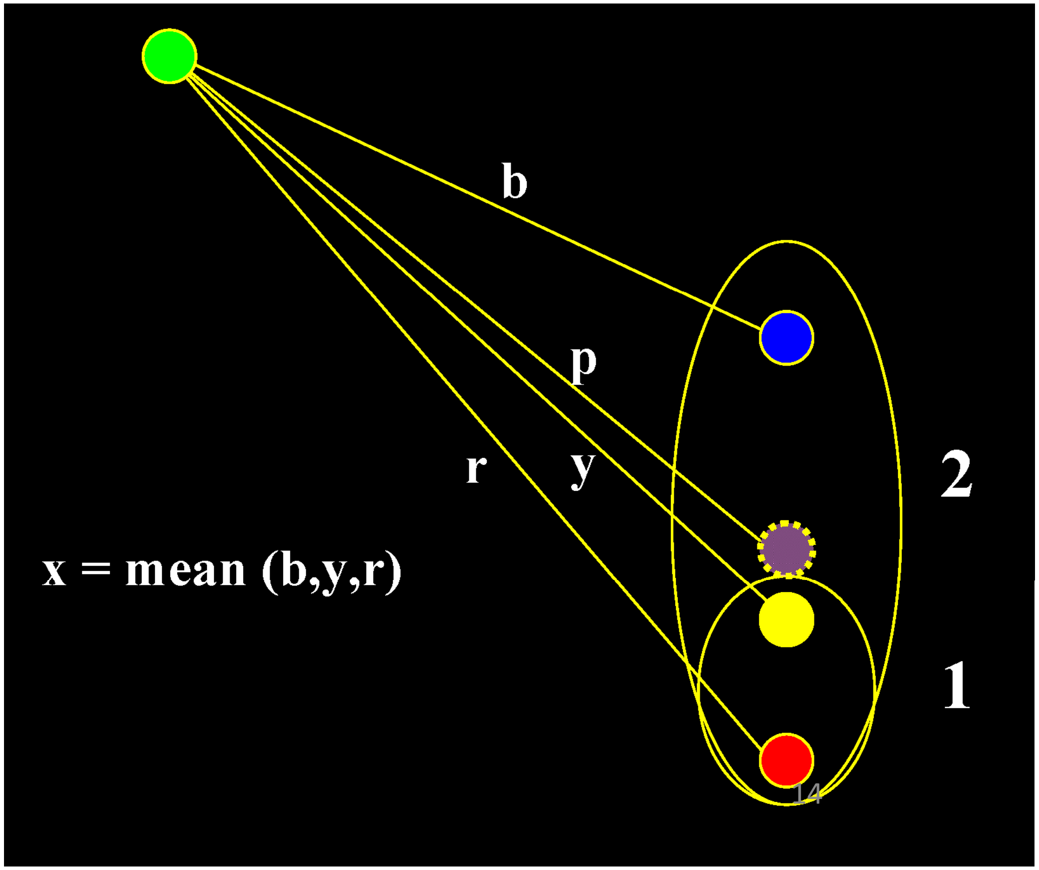


- Pearson correlation
  - Measures the difference in the shape of two curves
  - modifications:
    - uncentered correlation: for offset profiles, coefficient < 1
    - absolute correlation: opposite profiles cluster together
- Euclidean distance: multidimensional Pythagorean Theorem
  - Measures the distance between two curves
- Nonparametric or Rank Correlation
  - Similar to the Pearson correlation but data values are replaced with their ranks
  - Ex: Spearman Rank, Kendall’s Tau
  - Good idea if distribution of data is not normal
  - More robust (against outliers) than other methods

Hierarchical clustering: clustering methods

- How can groups of objects be represented?
- How is distance measured to a cluster of objects?

- Single linkage (b)
  - minimum distance
- Complete linkage (r)
  - maximum distance
- Centroid linkage (p)
  - distance to “centroid” of group
- Average linkage (x)
  - average distance



K-means clustering

- Places objects into k clusters (but with no further organization within each cluster)
- Choices one needs to make:
  - Matrix pre-processing details
  - Number of clusters (k)
    - Visual method: What k produces visually distinct clusters?
    - More quantitative method: Use “elbow method” to identify k when diminishing returns sets in
  - Similarity metric: How do we quantify similarity?

K-means clustering of genes

Input matrix (genes x samples):

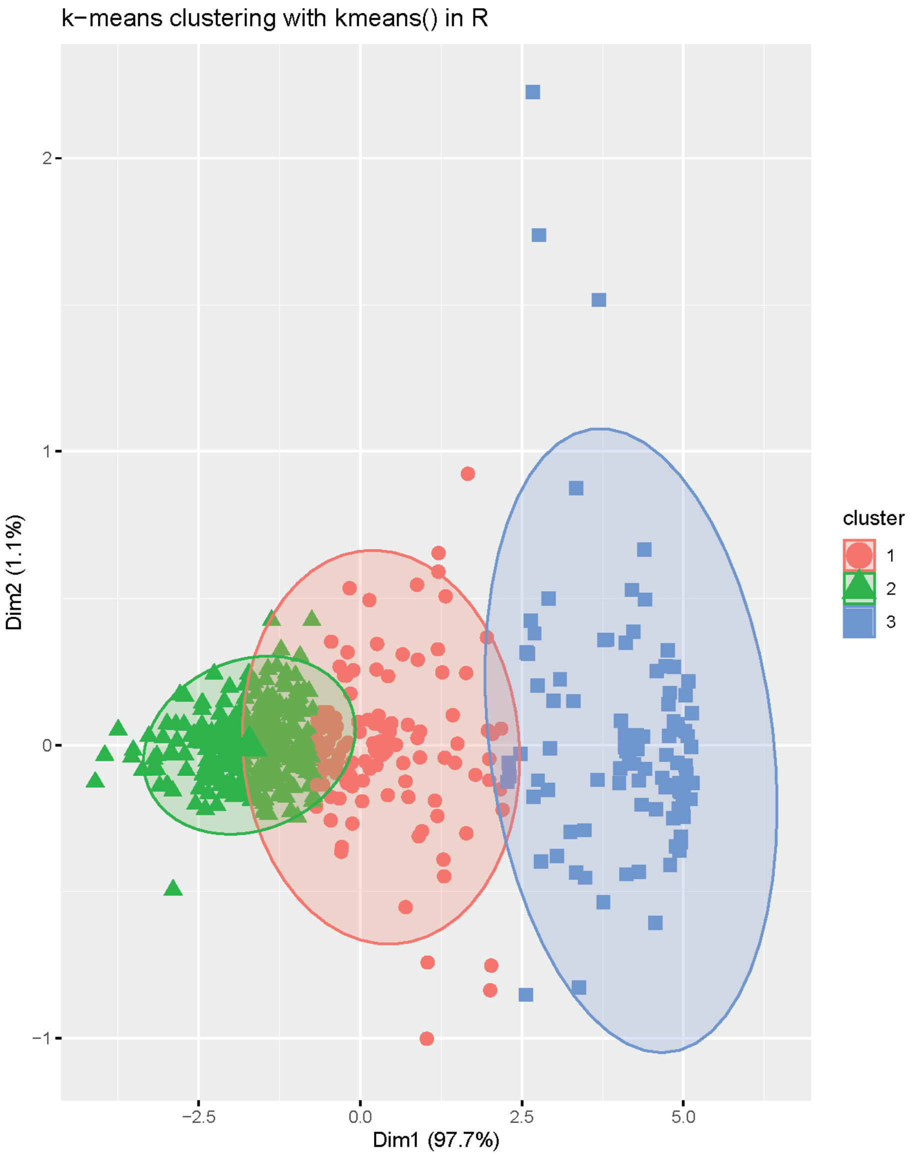
Gene ID	A	B	C	D	E	F
ENSG00000000003	2304.5	1528.0	1895.5	1819.0	1472.5	1856.7
ENSG00000000419	5419.9	2571.8	4654.4	3934.7	3361.6	4431.3
ENSG00000000457	417.1	603.5	493.7	504.3	535.3	470.9
ENSG00000000460	808.4	509.3	1699.7	1402.8	890.3	1402.8
ENSG00000000938	0.0	0.0	0.0	5.8	0.0	0.0
ENSG00000000971	805.8	1153.0	1123.8	1172.1	1281.9	921.4
ENSG00000001036	4518.5	3947.7	7201.9	6961.7	4812.4	7449.3
ENSG00000001084	27439.3	22138.7	35049.3	44486.7	40944.1	41730.4
ENSG00000001167	2444.1	1845.1	2016.3	1786.5	1679.1	2092.5

Preprocessing: matrix is

- log2-transformed
- scaled

Clustering and visualization created with cluster_by_kmeans_and_plot.R

Cluster 3.0 also includes this method.



findClusters() in Seurat

One function of the Seurat package for single-cell RNA-seq analysis

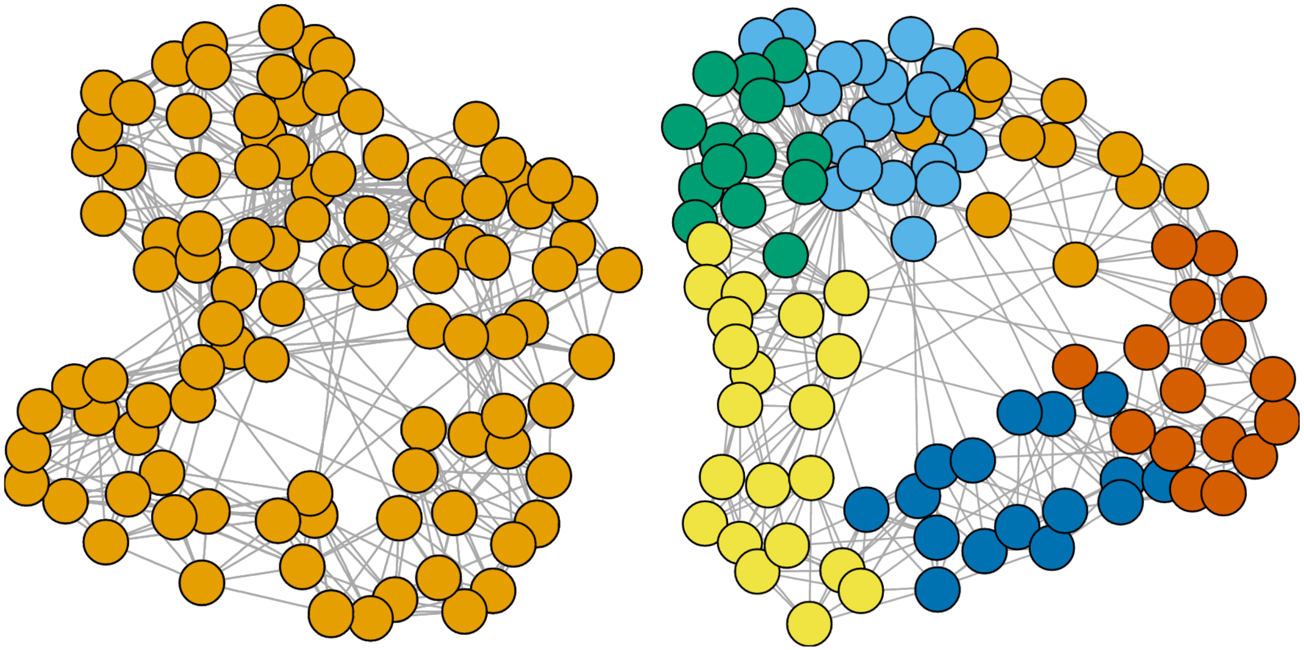
- Start out by representing each cell with multiple principal components (PCs).
- Calculate all pairwise Euclidean distances in PC space.
- Find the n nearest neighbors for every cell.
- Create a shared nearest neighbor (SNN) graph and connect cells that share a certain number of SNNs.
- Using the Louvain clustering method, partition this graph in order to make clusters.
  - This optimizes the density of links inside clusters compared to links between clusters

findClusters(): SNN graph and clusters

The shared nearest neighbor (SNN) graph is in the Seurat object:

object$RNA_snn

The Louvain (or other) method can partition these cells into clusters.



https://www.sixhat.net/finding-communities-in-networks-with-r-and-igraph.html

Clustering vs. heatmaps

- Most clustering methods take a matrix as input and create a matrix as output.
- Visualizing the results often requires another piece of software
- What output is important? Why are you clustering?
  - Ordered matrix
  - Dendrogram(s) of row and column items (ex: genome features, samples)

Heatmap software

- Microsoft Excel
  - Create matrix of log2 ratios or open CDT file from Cluster 3.0
  - Home => Conditional Formatting => Color Scales

- Cluster 3.0 + Java TreeView (Windows, Mac, and Linux)
  - http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm
  - Java is required
  - On tak, command is **`cluster`**

- Morpheus
  - https://software.broadinstitute.org/morpheus/

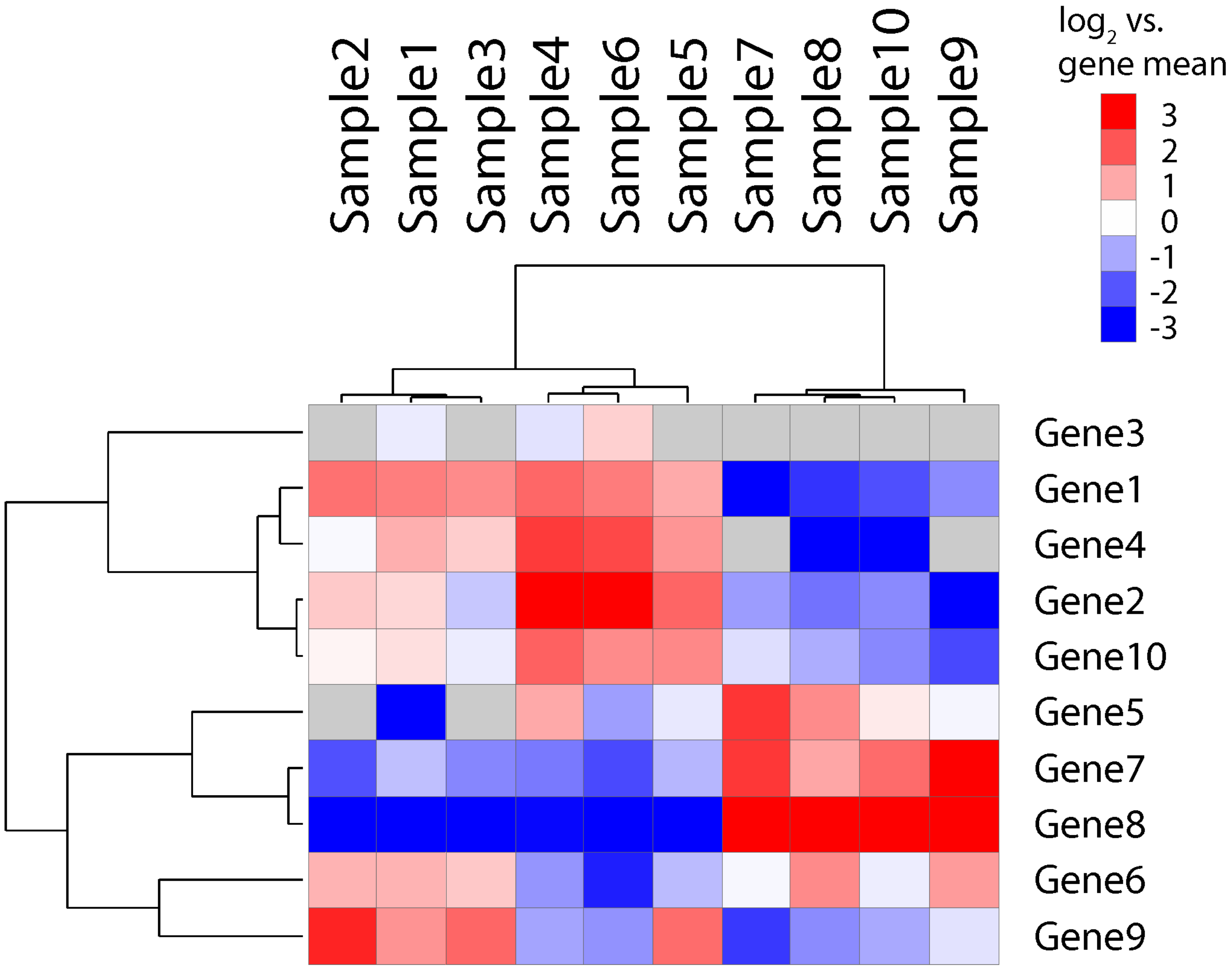
- R/Bioconductor
  - See BaRC R scripts such as

    /nfs/BaRC_Public/BaRC_code/R/drawHeatmap/drawHeatmap.R
    /nfs/BaRC_Public/BaRC_code/R/cluster_draw_pheatmap/cluster_draw_pheatmap.R

- For all software: thoughtfully choose colors and color range

Sample heatmap:

Cluster 3.0 + Java TreeView



Cluster 3.0:

- Log-transform
- Mean center
- Cluster
  - Uncentered corr
  - Average linkage

Java TreeView:

- Visualize
  - Change colors
- Export
  - To postscript
  - Colorbar too

Illustrator

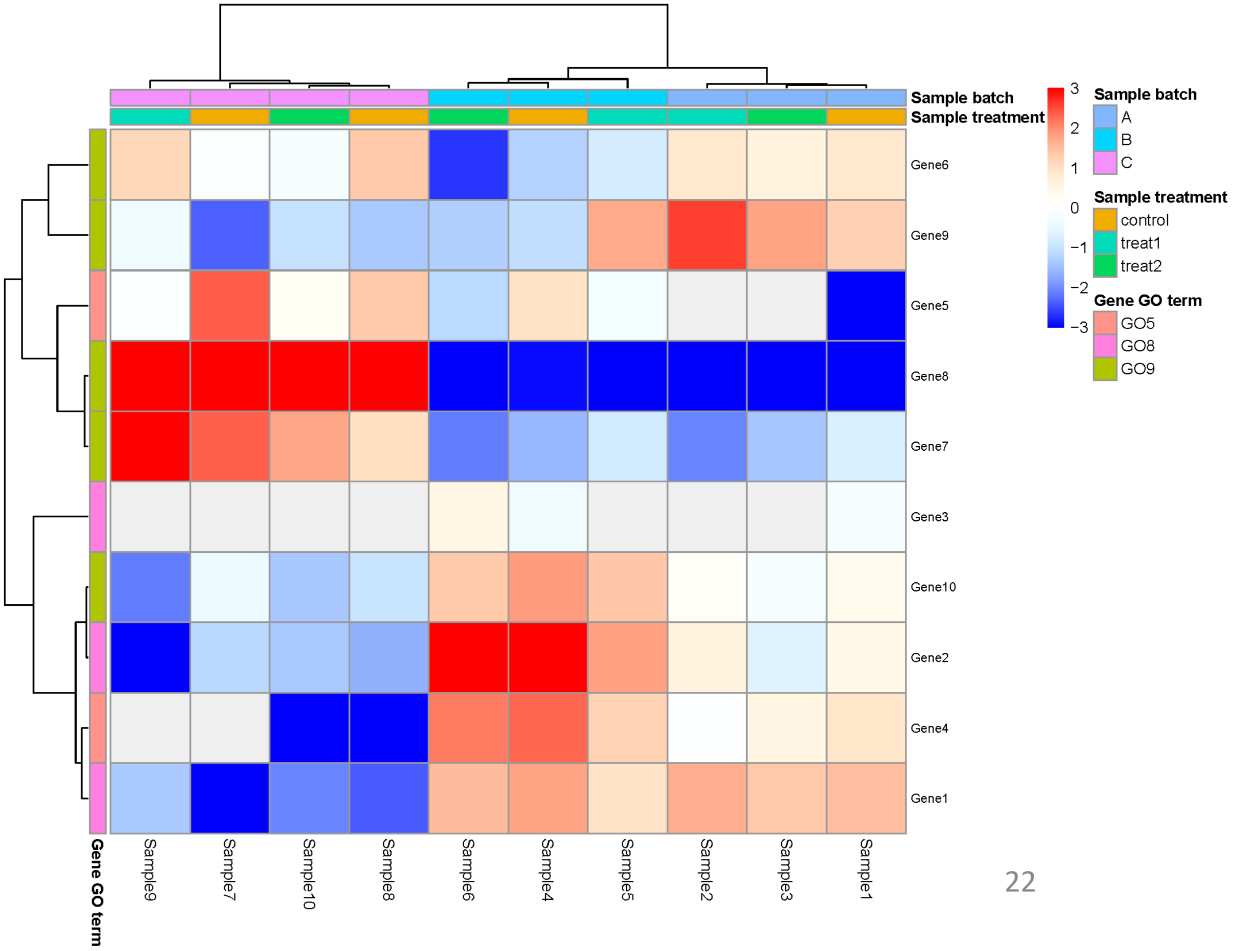
- Assemble

Enhancing a basic heatmap

Even though clustering is unsupervised, it can be useful to add additional information to a heatmap

Samples and/or features can be color-coded.

At right is sample output of cluster_draw_pheatmap.R



Upcoming Hot Topics

- Bioinformatics Resources
- Dimensionality Reduction (part 2)
- ChIP-seq and ATAC-seq
- Analytical project management