

CMSC 423: Data Clustering

Part 3

Hierarchical clustering

- Distance-based clustering: need definition of distance between data-points (e.g. individual genes)

	1 hr	2 hr	3 hr
g_1	10.0	8.0	10.0
g_2	10.0	0.0	9.0
g_3	4.0	8.5	3.0
g_4	9.5	0.5	8.5
g_5	4.5	8.5	2.5
g_6	10.5	9.0	12.0
g_7	5.0	8.5	11.0
g_8	3.7	8.7	2.0
g_9	9.7	2.0	9.0
g_{10}	10.2	1.0	9.2

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

Hierarchical clustering

- Distance-based clustering: need definition of distance between data-points (e.g. individual genes)

- Some measures:

- Euclidean distance

$$D(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

- Manhattan distance

$$D(x, y) = \sum_i |x_i - y_i|$$

- Pearson correlation

$$D(x, y) = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y}$$

- Clustering: group together data points that are most similar and repeat

Hierarchical clustering

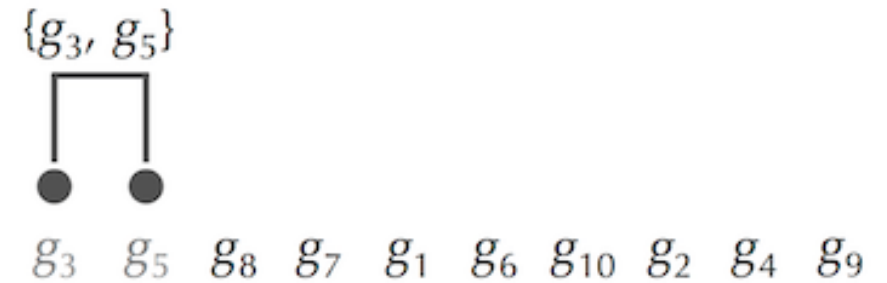
- Key element: how do you compute distance between two clusters, or a point and a cluster?
- UPGMA/average neighbor (average linkage)
 - Average distance between all genes in the two clusters
- Furthest neighbor (complete linkage)
 - Largest distance between all genes in clusters
- Nearest neighbor (single linkage)
 - Smallest distance between all genes in clusters
- Ward's distance
 - Inter-cluster distances is variance of inter-gene distances

Hierarchical clustering

- Irrespective of distance choices, algorithm is the same
 1. Compute inter-gene/cluster distances
 2. Join together pairs of genes/clusters with smallest distance
 3. Recompute distances to include the newly created cluster
 4. Repeat until all points are in a cluster
- Output of program is a tree

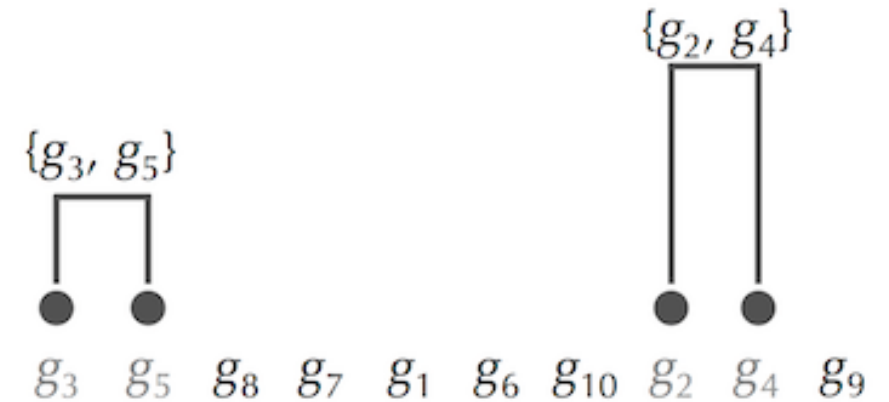
Hierarchical clustering

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0



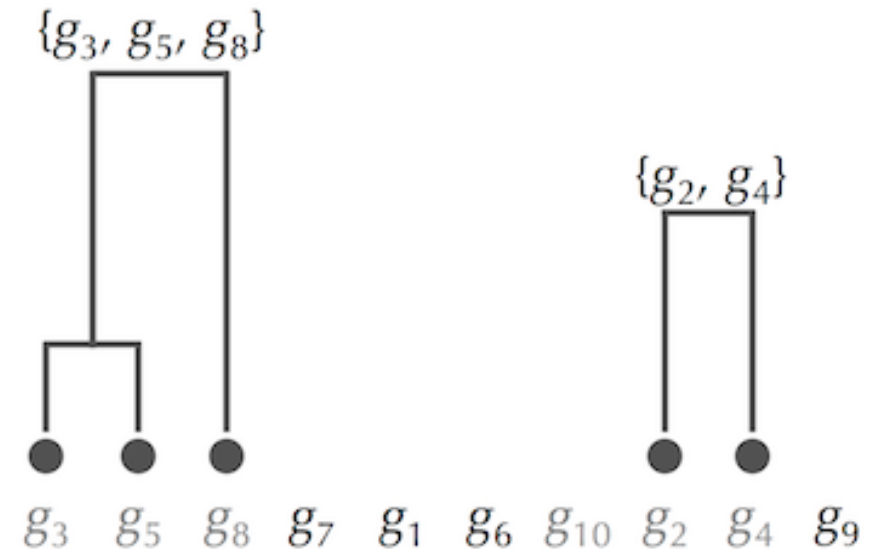
Hierarchical clustering

	g_1	g_2	g_3, g_5	g_4	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	9.5	10.1	12.8	2.0	1.0
g_3, g_5	9.2	12.0	0.0	11.2	11.1	8.1	1.0	10.5	11.5
g_4	7.7	0.9	11.2	0.0	9.2	9.5	12.0	1.6	1.1
g_6	2.3	9.5	11.1	9.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.0	12.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	8.5	9.3	12.4	1.1	0.0

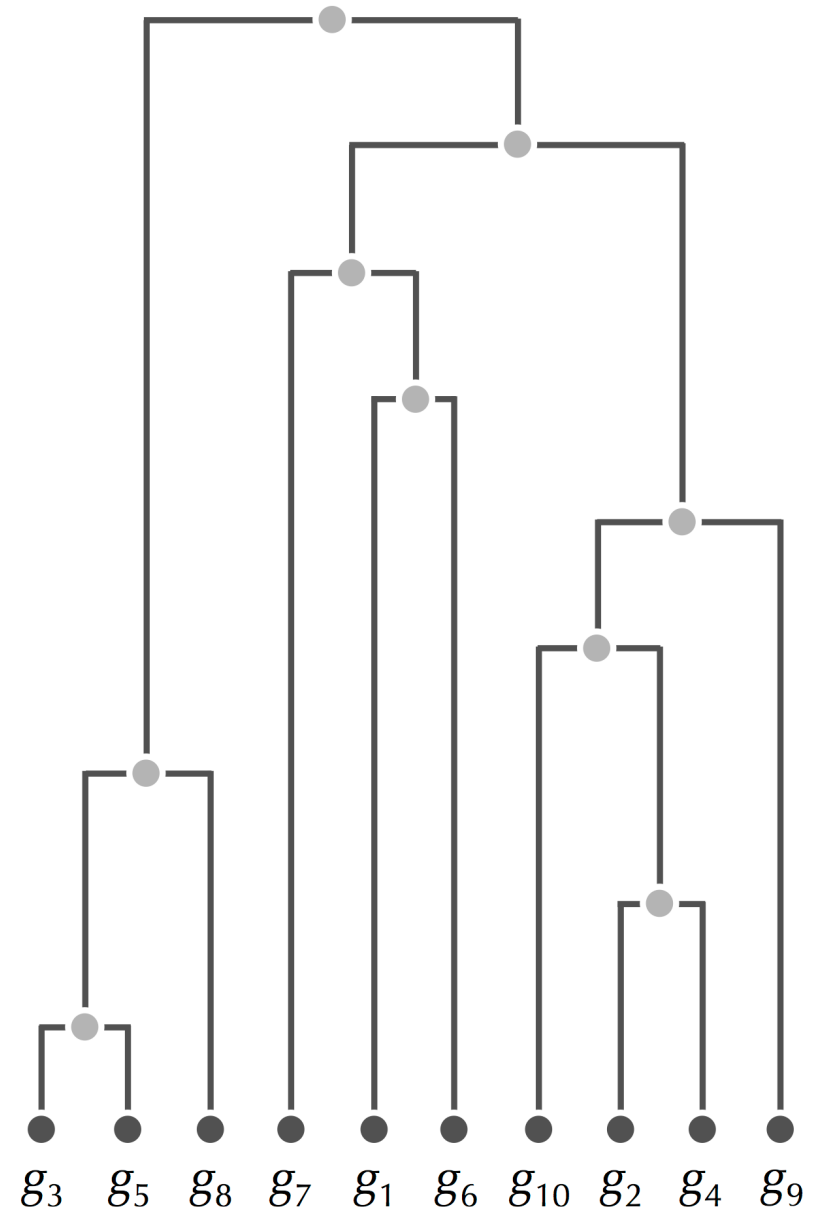


Hierarchical clustering

	g_1	g_2, g_4	g_3, g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	7.7	9.2	2.3	5.1	10.2	6.1	7.0
g_2, g_4	7.7	0.0	11.2	9.2	9.5	12.0	1.6	1.0
g_3, g_5	9.2	11.2	0.0	11.1	8.1	1.0	10.5	11.5
g_6	2.3	9.2	11.1	0.0	5.6	12.1	7.7	8.5
g_7	5.1	9.5	8.1	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	1.6	10.5	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	8.5	9.3	12.4	1.1	0.0



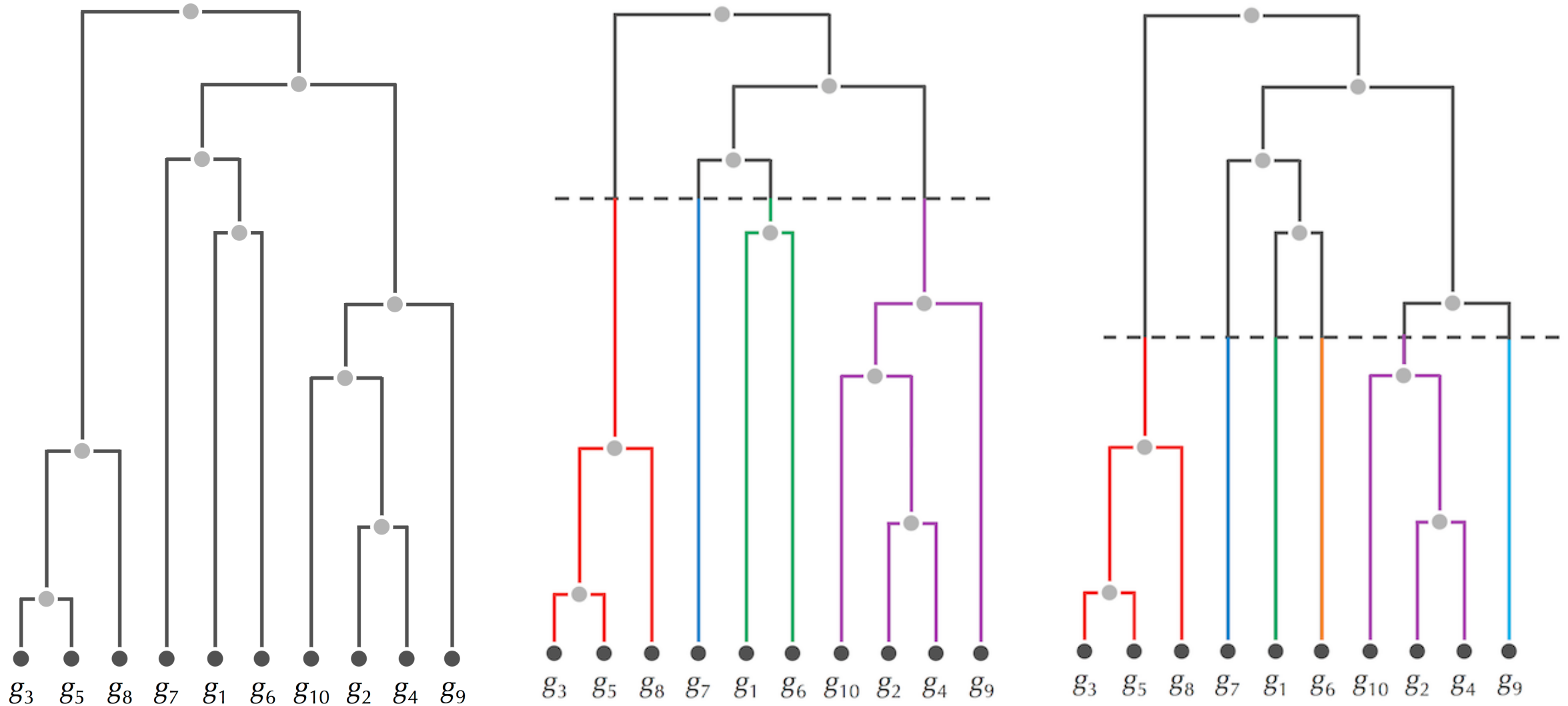
Hierarchical clustering



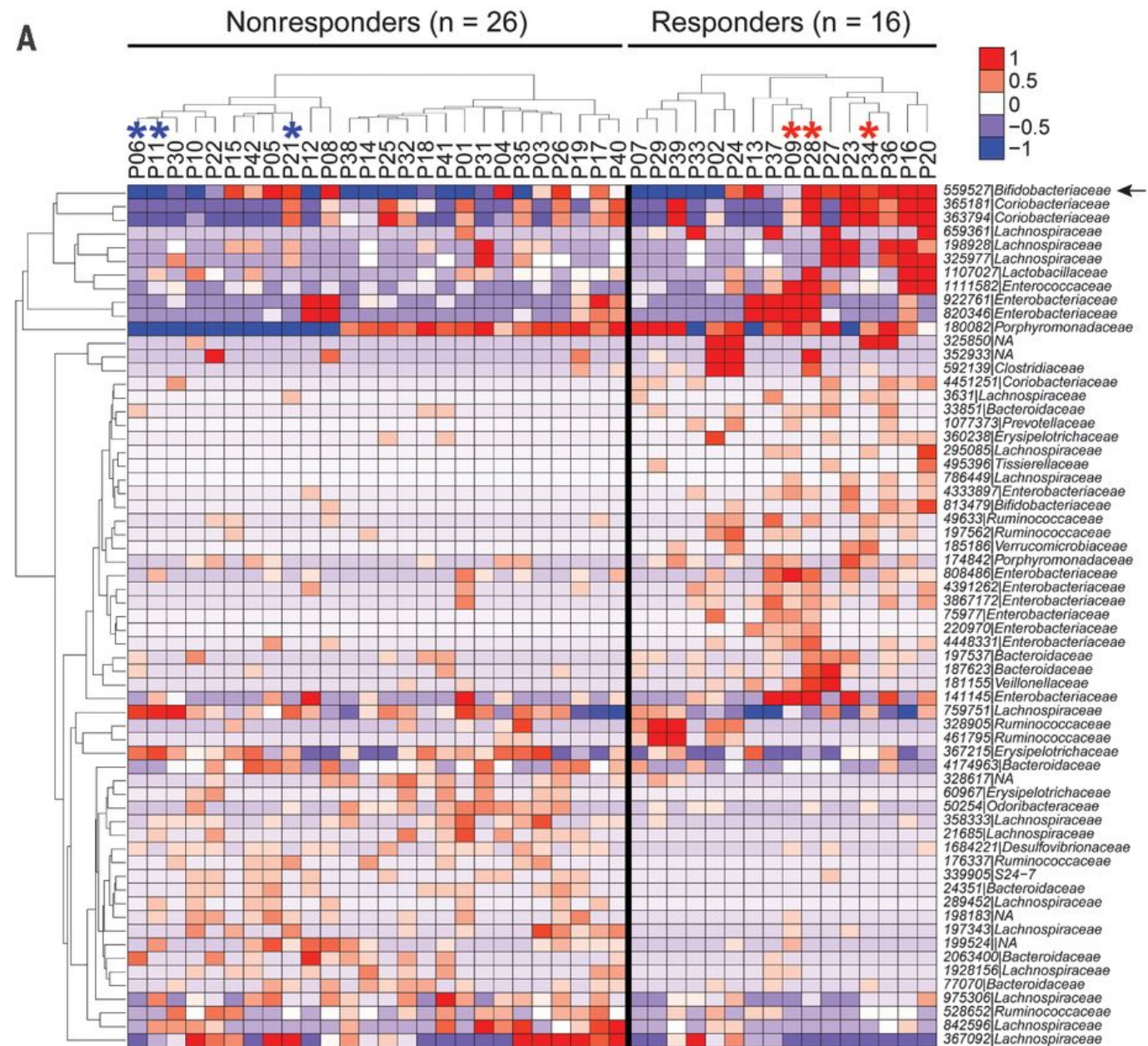
Hierarchical clustering

- Irrespective of distance choices, algorithm is the same
 1. Compute inter-gene/cluster distances
 2. Join together pairs of genes/clusters with smallest distance
 3. Recompute distances to include the newly created cluster
 4. Repeat until all points are in a cluster
- Output of program is a tree
- Cluster sets- defined by “cut” nodes - any subset of internal tree nodes defines a set of clusters – the sets of leaves in the corresponding subtrees
- Choice of cut can be tricky – usually problem-specific

Hierarchical clustering

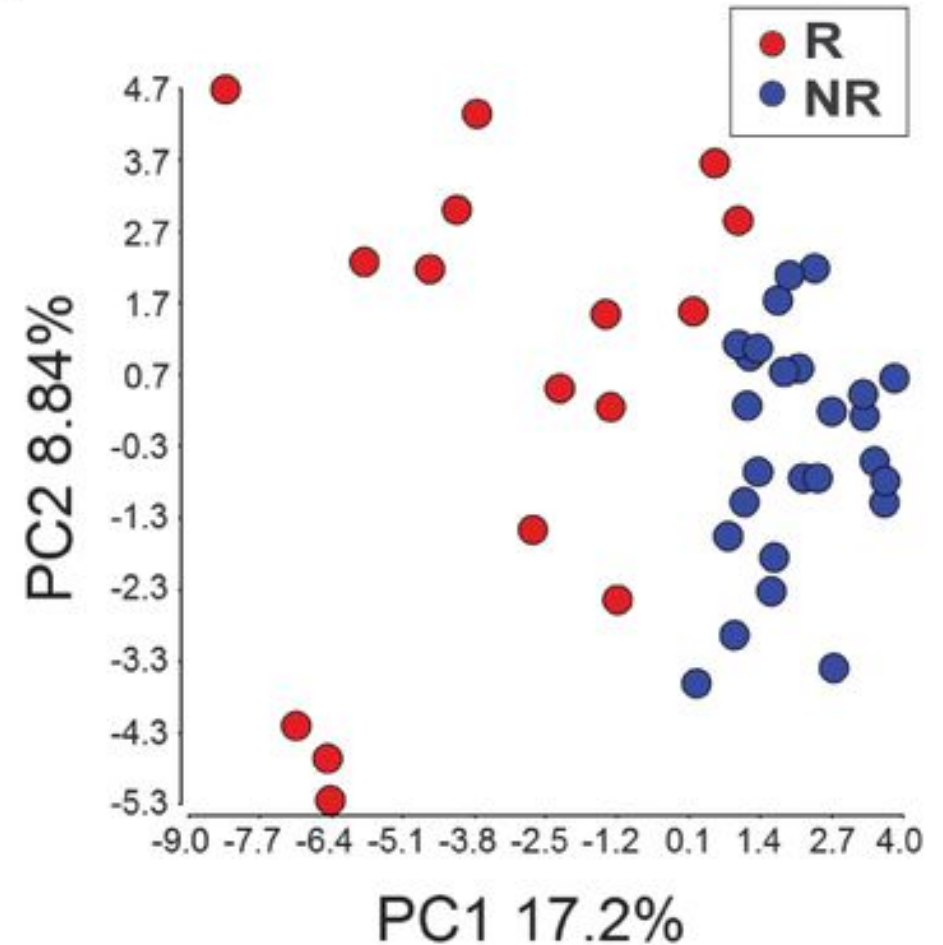


Example: microbiome analysis



Other clustering approaches

- Principal component analysis
 - Identify a direction (vector V) such that the projection of data on V has maximum variance (first principal component)
 - repeat (vector $V' \neq V$ such that project of data on V' has maximum variance)
 - Usually plot the first 2 or 3 principal components



Other clustering approaches

- Self-organizing maps
 - Neural-network based approach
 - Output layer of network are points in a low-dimensional space
- Graph theoretic
 - Points are connected by edges representing strength of "connection" (e.g. similarity or dissimilarity)
 - Pick clusters such that number of "similar" edges spanning boundaries is minimized, or number of "dissimilar" edges within each cluster is minimized
- Markov chain clustering
 - basic idea – a random walk through a graph will stay within a local strongly connected region