Exploratory questions

PCA

Computing PCA

- Which groups are clearly separated by PC1 and PC2?
- Which groups are clearly separated by PC3 and PC4?
- What happens if we include 2000 genes (rather than 500) in the PCA?
- What happens if we include all genes in the PCA?
- What happens if we use the raw CPM, rather than log2CPM for the PCA?
- What happens if we don't scale or centralize the data?
- How does PC3 and PC4 look? Go back to the code above and change which PCs to plot.
- Do PC10 and PC11 still separate your samples as well as PC1 and PC2?

Computing PC variance

- Which PCs explain at least 2% of variance?
- Instead of using our whole dataset, could we use only the top PCs for downstream analysis ? Explain.

Leading genes (optional)

- Which genes impact the most in PC1?
- If two genes highly impact on PC1 (i.e. the top 2), are their expression correlated?
- Which genes impact the most in PC2?

Hierachical clustering

- When printing the distance object, why is only half of the matrix shown?

Defining distance between samples

- What is the adjacency distance between two samples with correlation r of 0.8 ?
- What is the adjacency distance between two samples with correlation r of -0.6 ?
- What happens if instead of using the formula above, we used the absolute value of r ($|r|$) as adjacency (this way all values will be also between 0-1). Does this change affects the interpretation of adjacency distances ?
- What happens if instead of using the formula above, we used r^2 as adjacency (this way all values will be also between 0-1). Does this change affects the interpretation of adjacency distances ?

Clustering samples

- Does the ordering of the samples in dendrogram has a meaning ?
- Why are the scales between those dendrograms so different ? Look at the distance matrices to get the intuition.

- Do the two methods represent the same results ? Which one would you trust the most ?
- Does the ordering of the dendrogram have a meaning ?
- Change the clustering method to `ward.D2`. How does it affect your results ?
- We used the whole dataset for clustering. Re-calculate the distances now using only the top 500 genes with highest CV. Does it change the results ?
- Instead of clustering on the raw data, could we cluster on the top N principal components? Justify.

Defining clusters

- Check the dendrogram above, what is a sensible height to cut the tree?
- What is the maximum sensible amount of clusters I could have in my data?

Clustering on Heatmap (optional)

- Check the dendrograms on the side of the heatmap. Do they look familiar with the previous ones?
- What does the 'scale' parameter mean? What happens with the clustering if we change it to 'none' or to 'columns'?
- Can you change the clustering to 'ward.D2'? Explore the `pheatmap` function pressing tab.
- Can you cut your gene tree into 4 clusters? Explore the `pheatmap` function pressing tab.

Clustering bootstrapping (optional)

- With what percentages the samples 10, 11 and 12 fall together in the same cluster?
- With what percentages the samples other than 10, 11 and 12 fall together in the same cluster?
- With what percentages the samples 8, 7 and 9 fall together in the same cluster?
- With what percentages the samples 4 and 5 fall together in the same cluster?
- The orange rectagles represent 2 clusters. What do you think is the criteria used to define this clusters ? PS: it is not the height as in the previous examples.