Gene Expression Data Visualization with R

Theory Session

Dr. Cihan Erkut
Applied Functional Genomics (B290)
Translational Medical Oncology (B340)

Main topics

- Principal component analysis (PCA)
  - Dimensionality reduction
  - Grouping by similarity

- Heatmaps
  - Hierarchical clustering
  - Sample correlation / difference

- Mean – variance plots
  - Variance stabilization

- MA plots
  - Visualization of fold changes
  - Fold change shrinkage

Principal component analysis (PCA)

https://blog.freepeople.com/2013/02/book-club-meeting-deepest/

Principal component analysis (PCA)

- Orthogonal projection of
  - an N-dimensional object viewed from a **random perspective**
  - into an M-dimensional object viewed from from **another perspective**

- The projection has the following features
  - First axis (principal component) explains **as much of total variation between data points as possible**
  - Second PC explains **as much of remaining variation as possible**
  - **…**
  - Mth (last) PC explains **the last remaining variation**
  - No axis depends on another

- Total variation is preserved
- No data is created or removed

Principal component analysis (PCA)

- A gene expression dataset is a multidimensional object
  - *n* genes in *m* samples = *n* x *m* matrix
  - Every sample can be represented as a point in an n-dimensional space
  - Coordinates of a point are expression values of all genes

	S_1	S_2	S_3	S_4
G_1	E_{11}	E_{12}	E_{13}	E_{14}
G_2	E_{21}	E_{22}	E_{23}	E_{24}
…	…	…	…	…
G_n	E_{n1}	E_{n2}	E_{n3}	E_{n4}

$$S_1 = < E_{11}, E_{21}, \ldots, E_{n1} >$$

Principal component analysis (PCA)

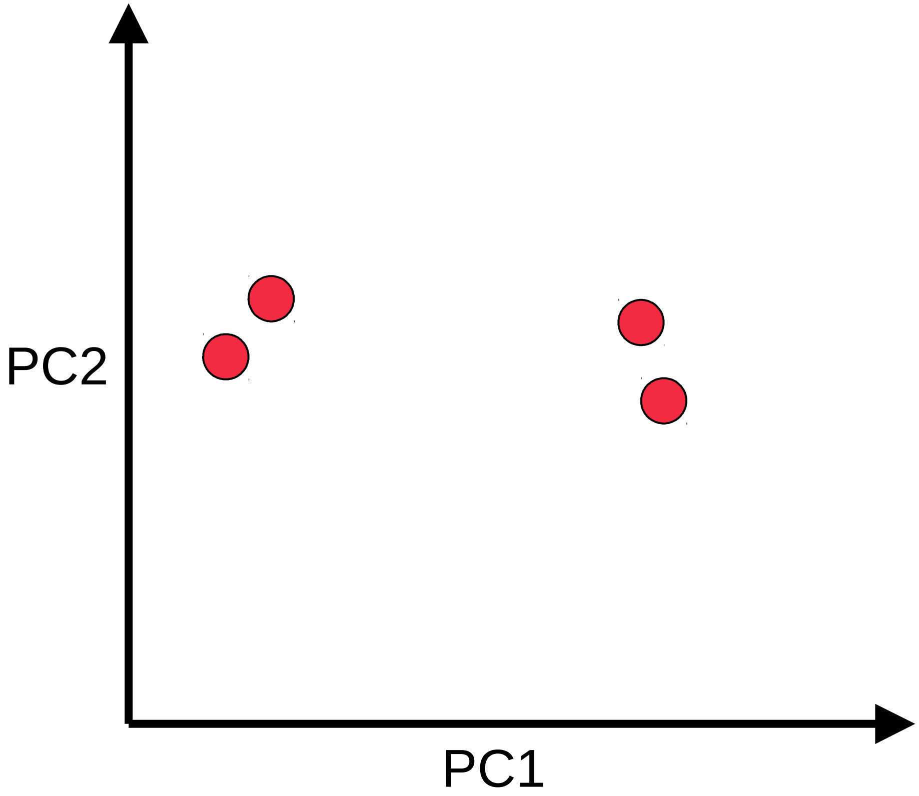
- Can you imagine a point in a 5-dimensional space?
  - If yes, I want to talk to you after the seminar □

- Solution: Reduce dimensionality

- Based on which criteria?
  - Variation among genes!
  - A unique signature of the sample

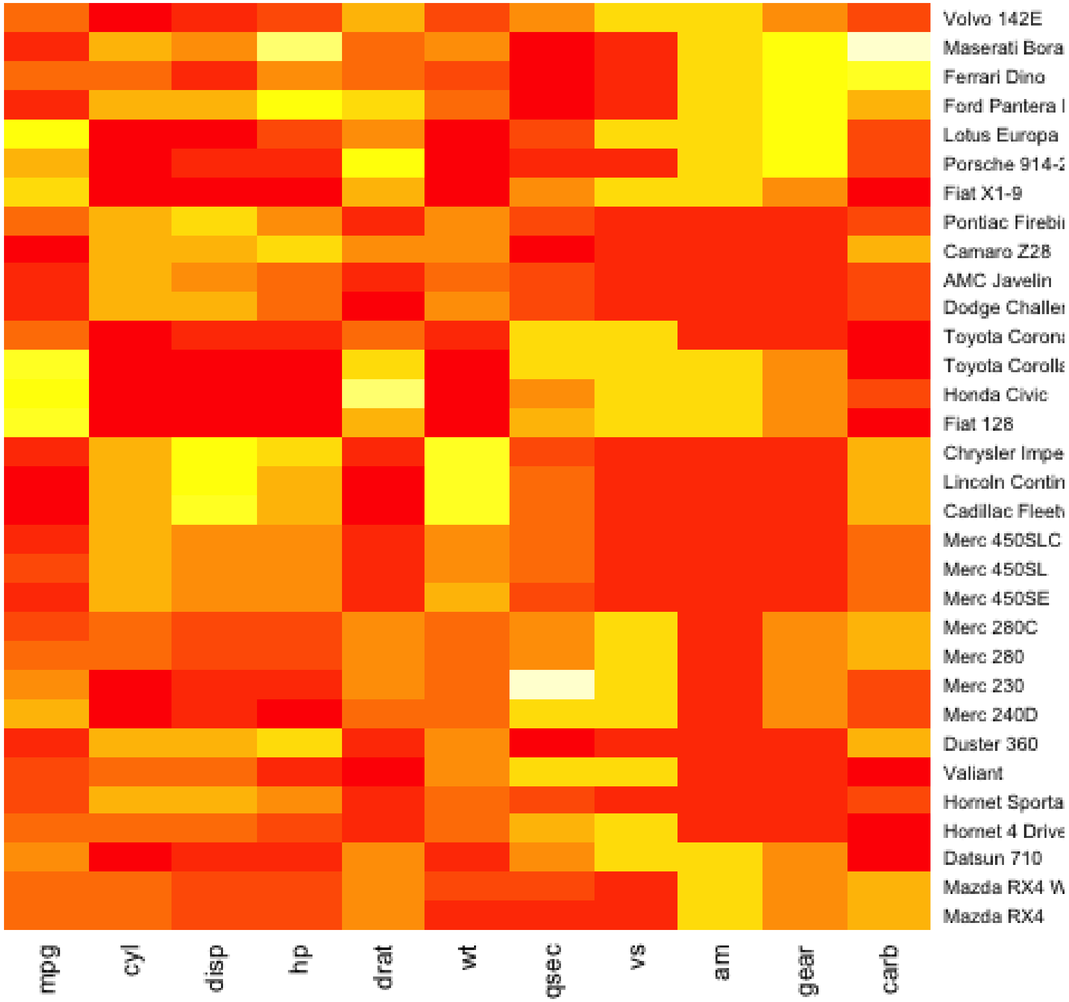
- How many dimensions are enough?
  - Usually two, at most 3 (not recommended)

Principal component analysis (PCA)



Questions?

Heatmaps



www.r-graph-gallery.com

Heatmaps

- A method to visualize 3-dimensional data on a 2-dimensional space
- Takes advantage of human color perception
  - Dimension 1: x-axis
  - Dimension 2: y-axis
  - Dimension 3: color

- Extra information can be encoded via grouping

- Grouping on which criteria?
  - Similarity / dissimilarity (hierarchical clustering)

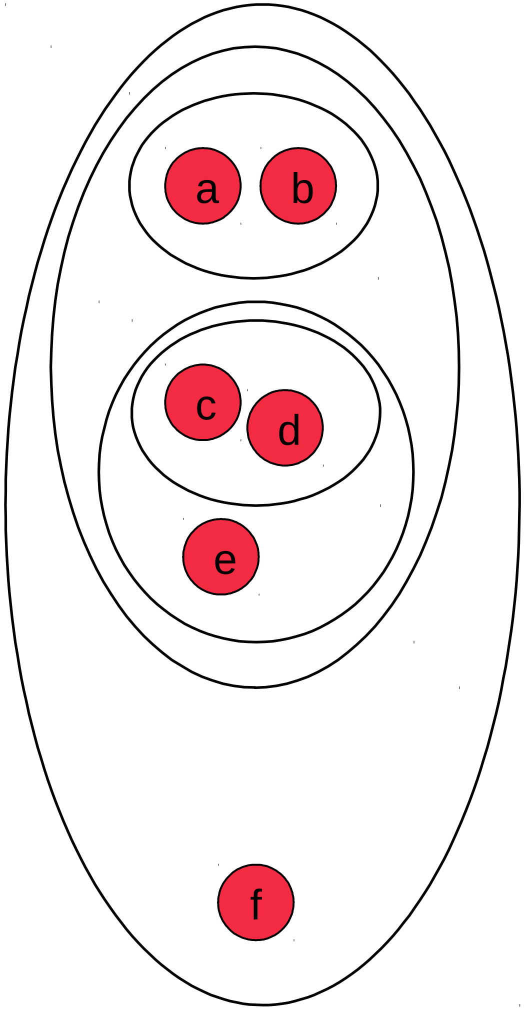
Heatmaps

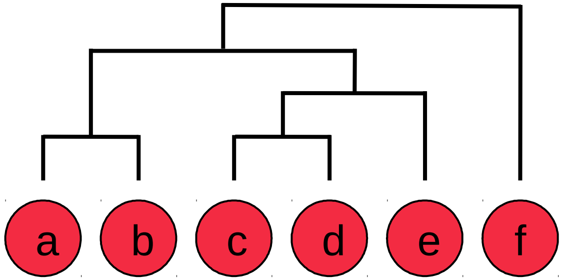
- What does hierarchical clustering do?

- First calculate a distance matrix
  - Remember, every sample / gene is a point in a multidimensional space
  - There always exists a line segment that connects two points!
  - The length of that segment is (an Euclidean) distance!
  - A distance matrix is half of a square matrix. Think why!

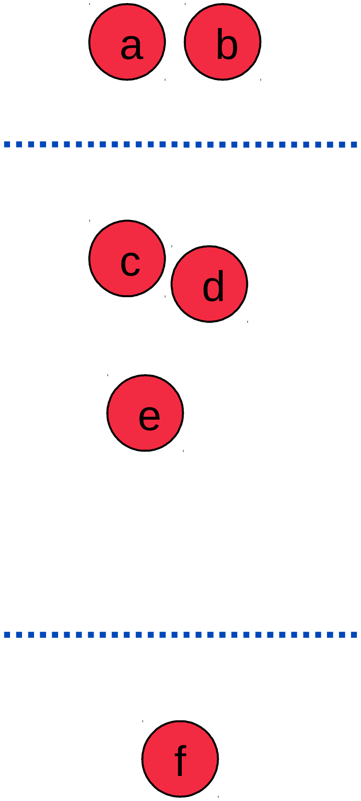
- Iterate over samples / genes and group them based on distances
  - The result is a tree-like structure called a **dendrogram**

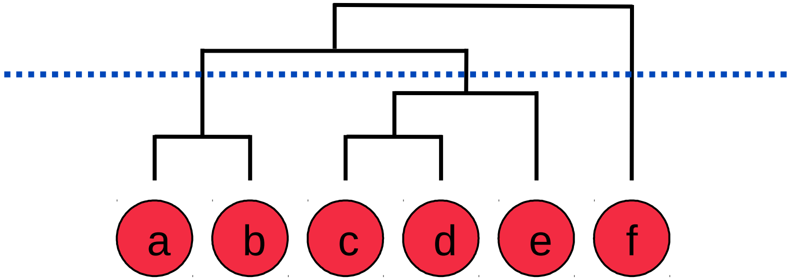
Hierarchical clustering





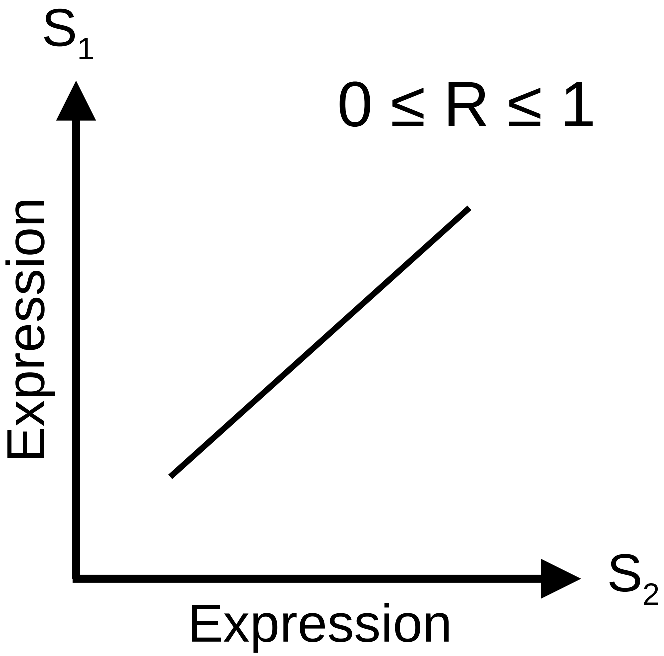
Hierarchical clustering





Correlation

- Correlation matrix instead of distance matrix
  - Based on correlation coefficient
  - Easier to interpret



Questions?

Logarithmic transformation

- Hierarchical clustering and PCA are sensitive to data distribution

- It’s best when the data is:
  - Normally distributed
  - Homoskedastic (variance is stable)

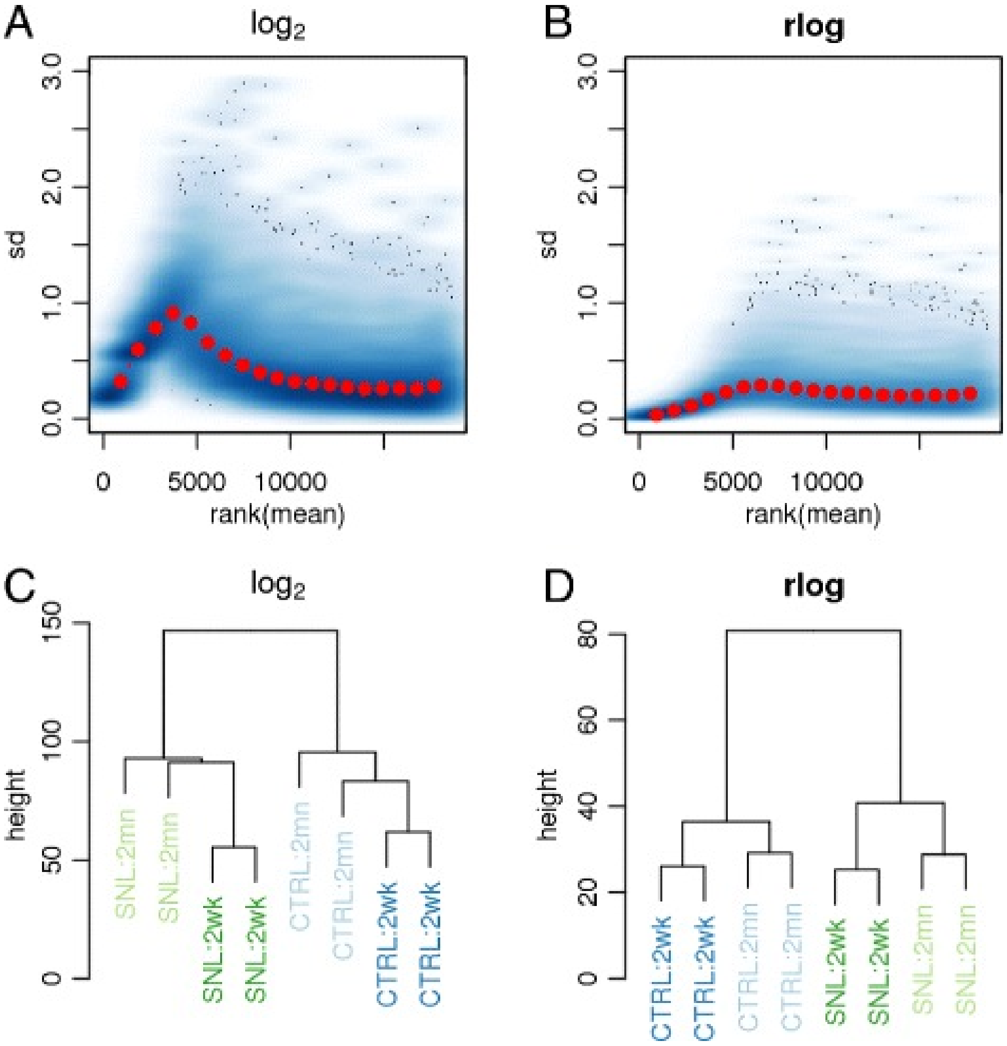
- Gene expression data is naturally skewed
  - A lot of low-expression genes, few high-expression genes
  - A logarithmic transformation helps to normalize the data

Variance stabilization

- Log transformation affects the variance of small numbers more than big numbers
  - Set 1: 5, 6, 7
  - Set 2: 5005, 5006, 5007

		Set 1	**Set 2**
Original	**Mean**	6	5006
	SD	1	1
log2	**Mean**	2.571	12.289
	SD	0.2430	0.0003

Mean-variance plots



Love et al. 2014

Questions?

MA Plots

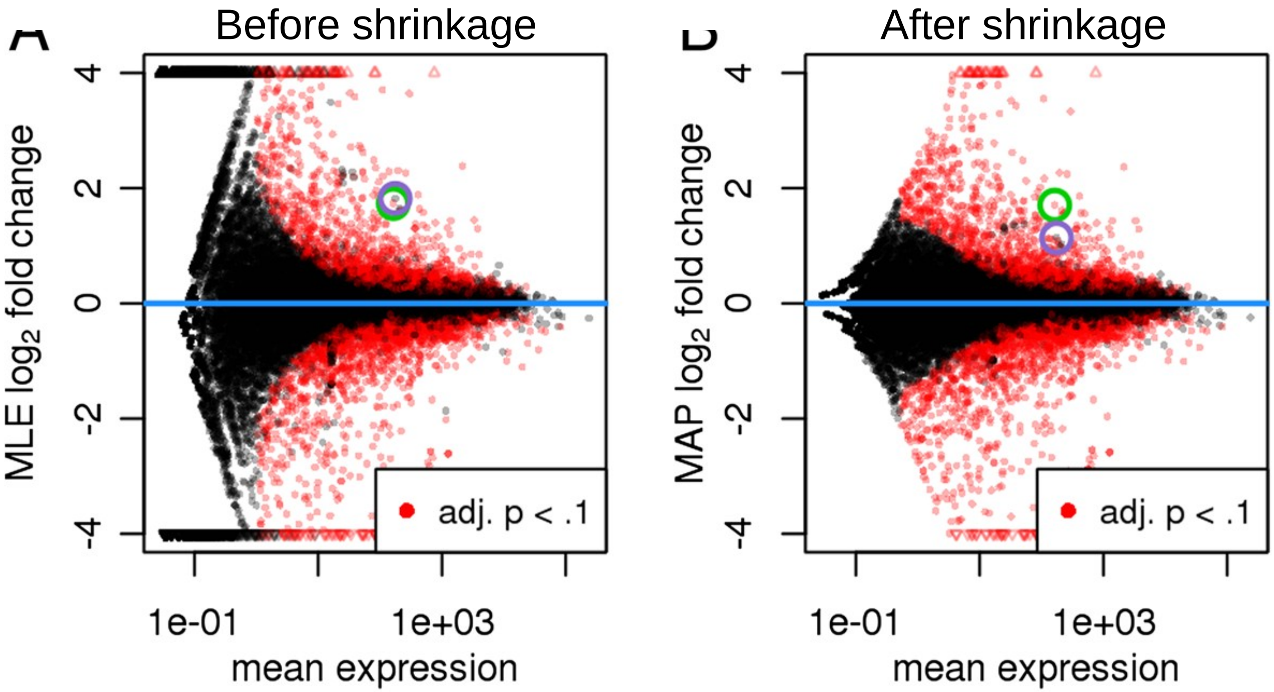
- **Why do we do differential gene expression analysis?**
  - To find out up- or downregulated genes in a **test sample <u>compared to</u> a control sample**

- **M represents log fold change (LFC)**
  - $\text{LFC}_i = \log_2 \frac{E_{i,test}}{E_{i,control}}$
    - $\text{LFC} > 0 \Longrightarrow$ Upregulation
    - $\text{LFC} < 0 \Longrightarrow$ Downregulations
    - $\text{LFC} = 0 \Longrightarrow$ No difference
  - Associated with an adjusted p-value

- A represents average gene expression across samples
  - An approximation of abundance level

LFC shrinkage

- Similar to variance stabilization, more complicated
- Aims to remove very high/low LFCs observed in low counts



Love et al. 2014

Questions?