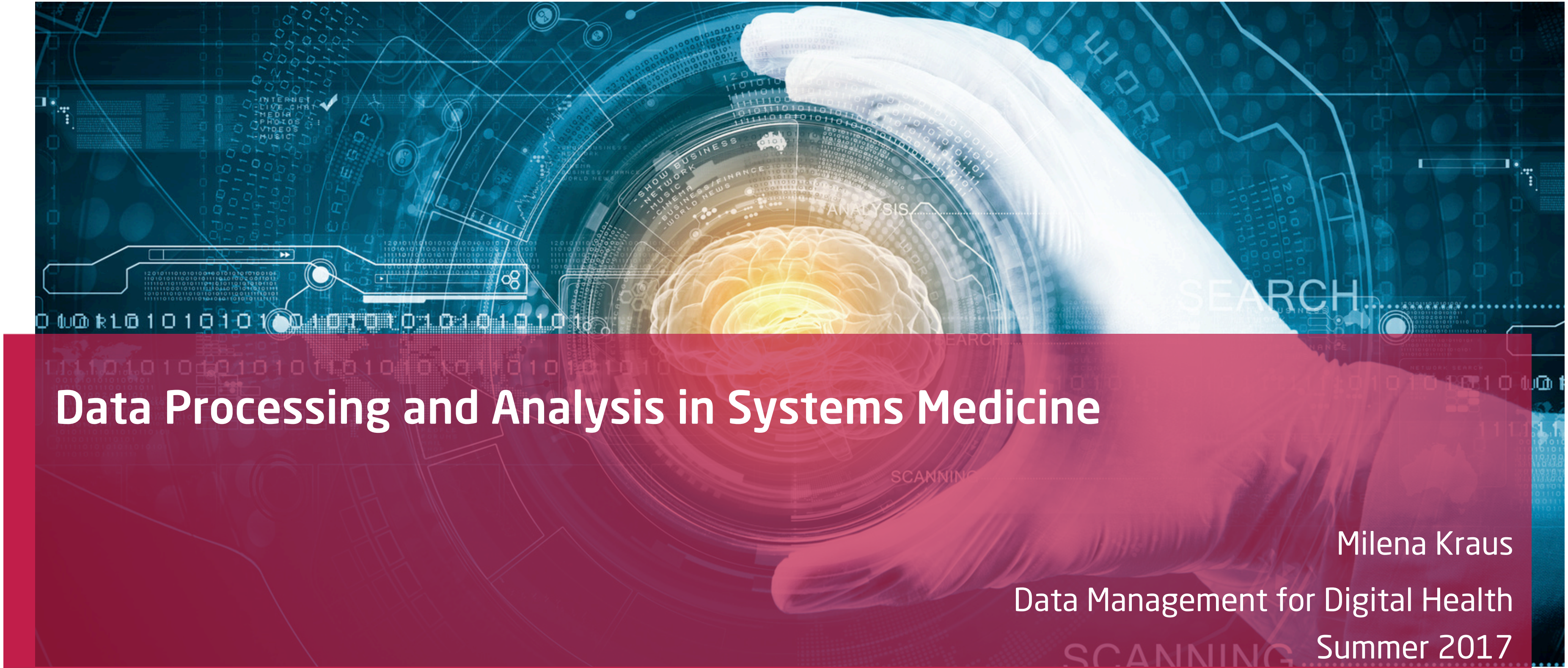
Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

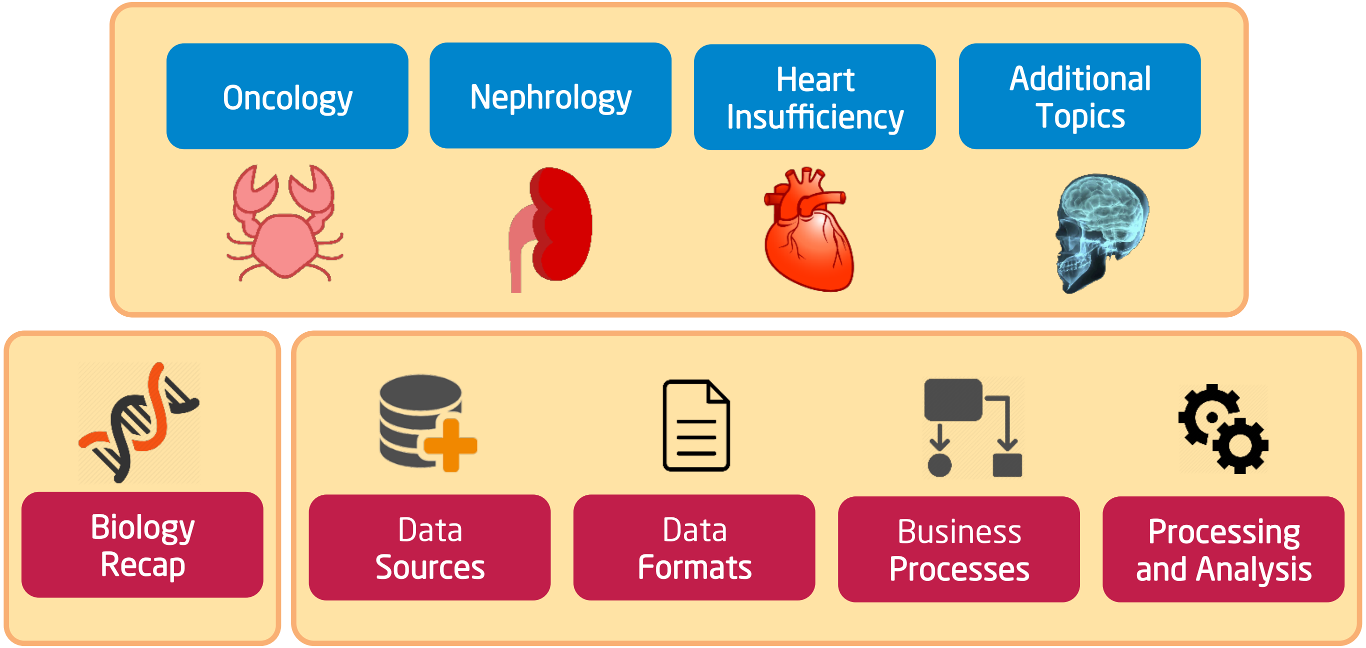


Data Processing and Analysis in Systems Medicine

Milena Kraus

Data Management for Digital Health
Summer 2017

Agenda



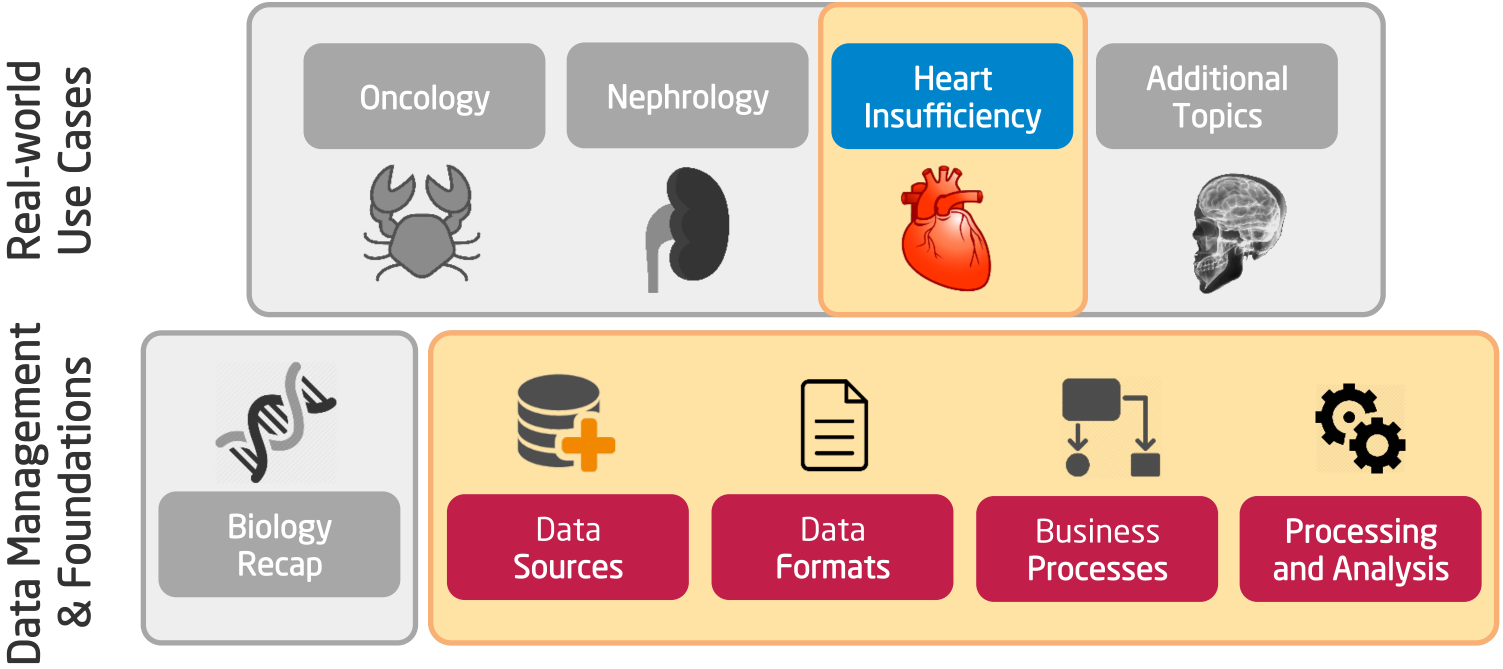
Real-world Use Cases

- Oncology
- Nephrology
- Heart Insufficiency
- Additional Topics

Data Management & Foundations

- Biology Recap
- Data Sources
- Data Formats
- Business Processes
- Processing and Analysis

Where are we?


Real-world Use Cases
Oncology
Nephrology
Heart Insufficiency
Additional Topics
Data Management & Foundations
Biology Recap
Data Sources
Data Formats
Business Processes
Processing and Analysis

In this lecture you will learn…

In addition to e.g., genomic data, we will take a deep dive into the

- Processing and analysis of RNA sequencing data and
- Differential gene expression analysis.

For a better understanding, we will repeat some methods of unsupervised learning:

- Clustering strategies (Hierarchical, K-Means) and
- Dimensionality reduction (PCA, MFA).

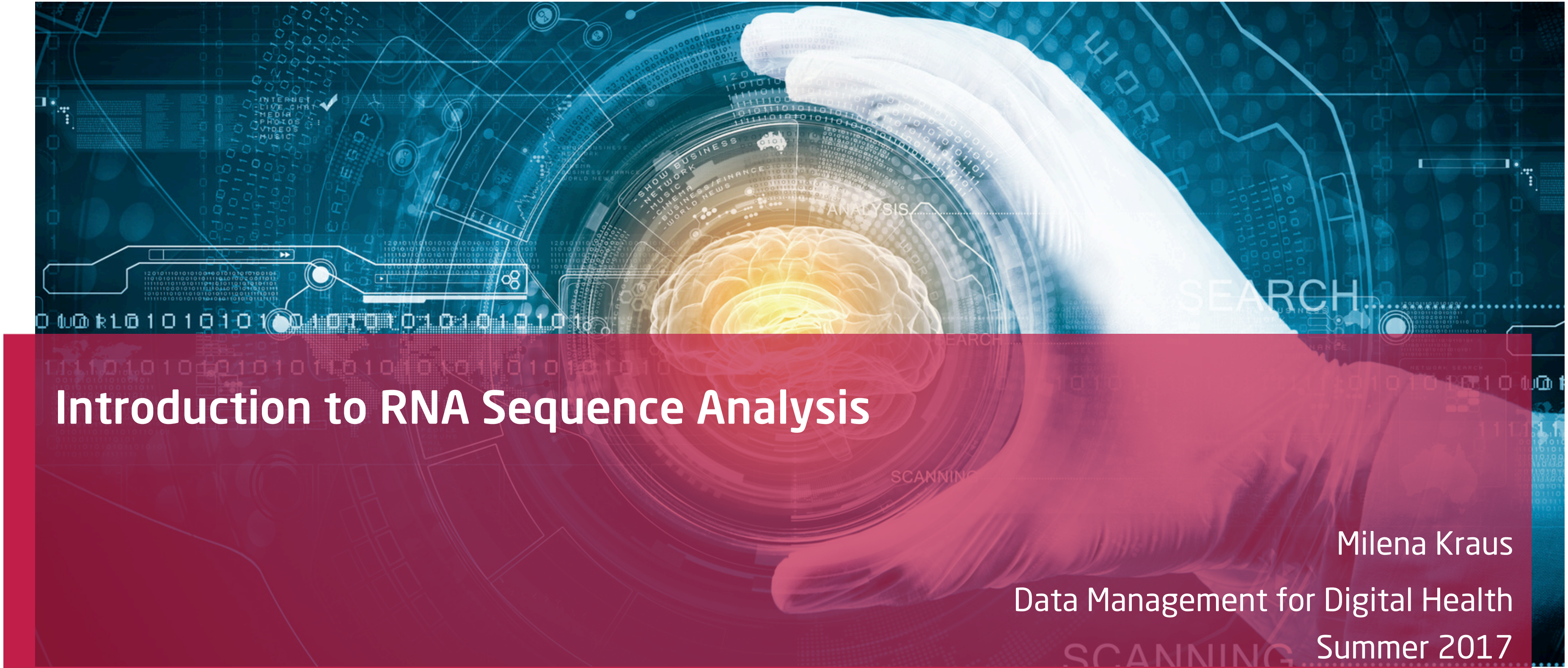
Systems Medicine
Information Levels

- Clinical information, e.g., imaging data, hemodynamics, lab reports, diagnosis etc.
- Patient information, e.g., age, gender, environment, history etc.
- Omics
  - □ Genome → modeled as categorical values
  - □ Transcriptome (Gene Expression)
  - □ Proteome
  - □ Metabolome

  } → modeled as numerical values

Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam



Introduction to RNA Sequence Analysis

Milena Kraus

Data Management for Digital Health

Summer 2017

Agenda

What is RNA sequencing used for?

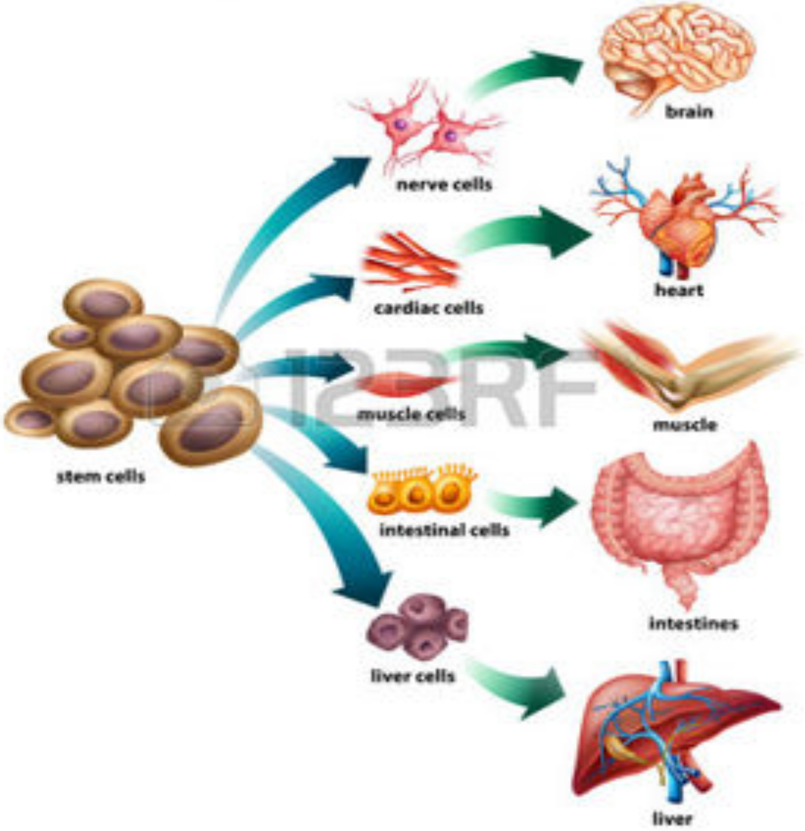
1. Biological background
2. From wet lab sample to transcriptome
   a. Experimental procedure
   b. Raw data
   c. Processing pipeline(s)
   d. Downstream analysis
3. Differential gene expression analysis

How is a muscle cell different from a liver cell?

You've learned so far:

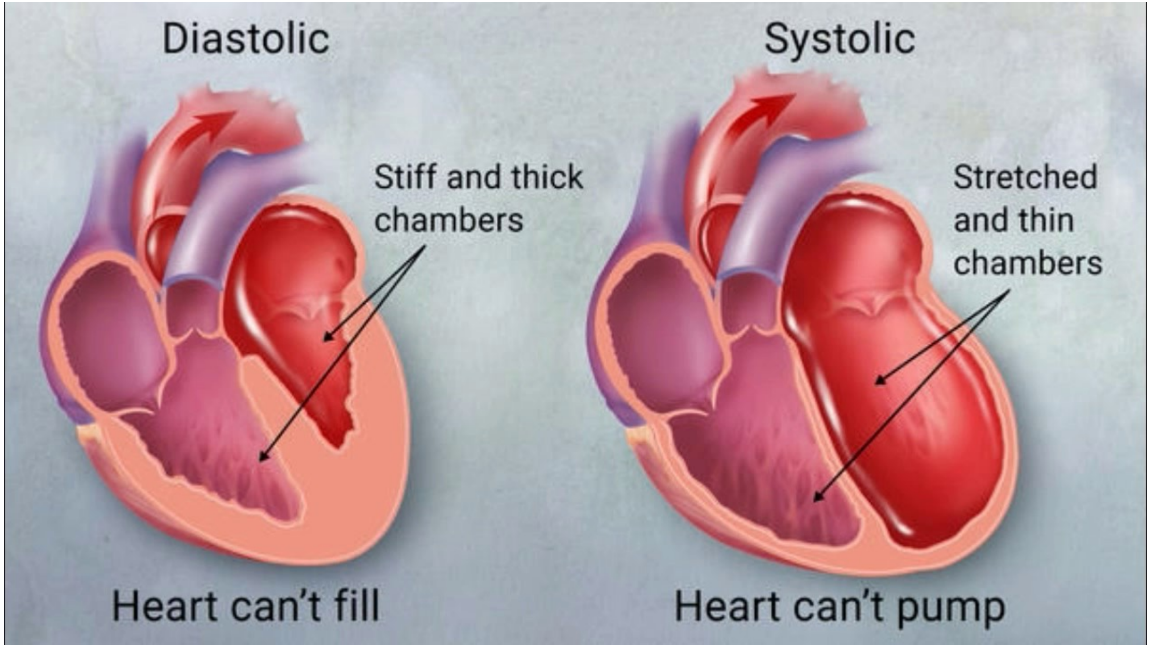
- Every cell in your body contains the same DNA as every other cell
- The DNA codes for every process in the cell

How can one cell be different from another if they use the same genetic information?



What is the difference between a healthy heart and a sick heart?

- How do cells taken from a healthy and a diseased heart differ? And are those differences treatable, e.g, through drugs?



https://www.youtube.com/watch?v=B93TsbJXnMc

Information Content in RNA

- Information retrieved from RNA:
  - □ Quantity (primary, How many RNAs are transcribed from a specific gene?)
  - □ Sequence (secondary, as sequence information is also in DNA)

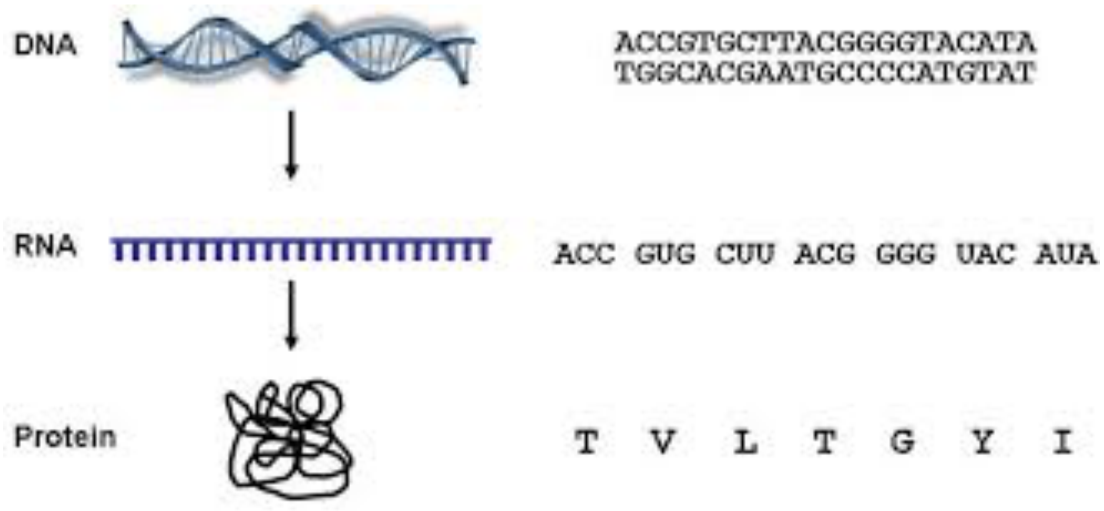


Image source: http://cureangelman.org/understanding-angelman/testing-101/

Challenge in RNAseq
Alternative Splicing

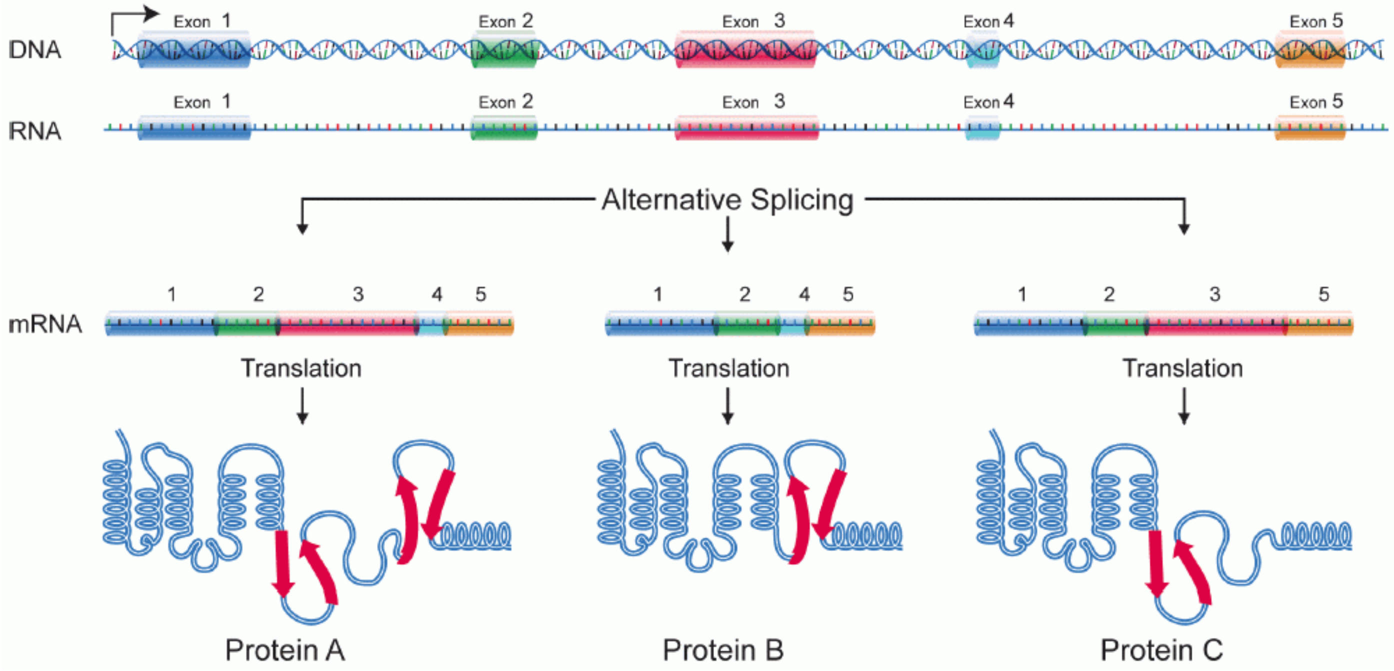


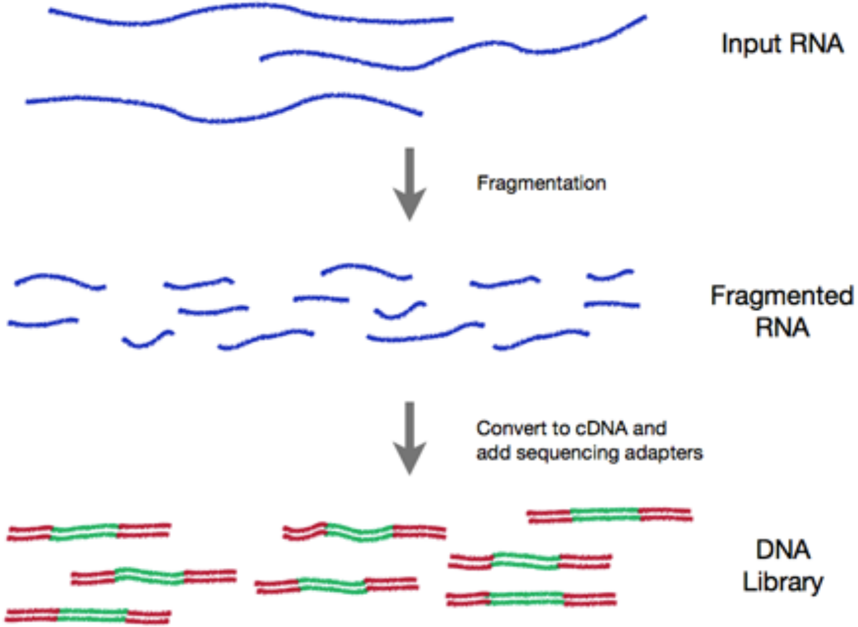
Image Source: National Human Genome Research Institute - http://www.genome.gov/Images/EdKit/bio2j_large.gif, Gemeinfrei, https://commons.wikimedia.org/w/index.php?curid=2132737

Experimental Procedure

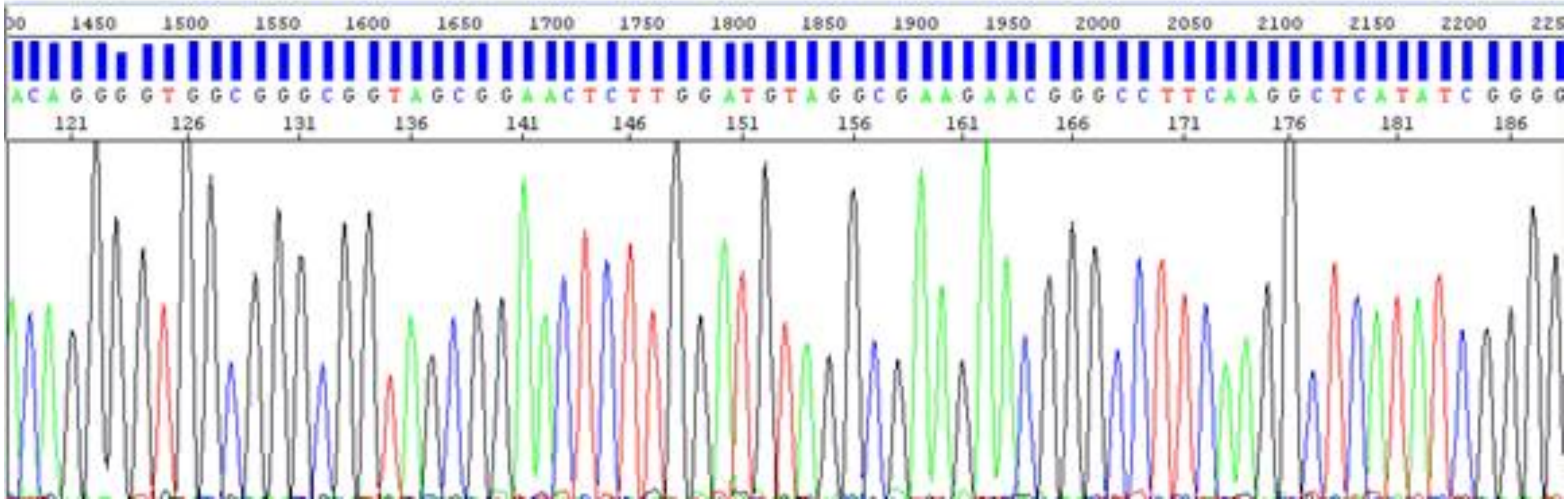
- Generally similar to DNA sequencing
- Over 20.000 single stranded RNAs in variable abundance (1-k times) of 1.500-2.000 nt
- Fragmented into 30-200 nt

Differences:

- RNA is single stranded and needs to be revers-transcribed to DNA for sequencing
- Coverage is dependent on expression value of gene



Sequencing Signal



Differences and Similarities of RNA- and DNAseq Wetlab and Sequencing Procedure

	DNAseq	RNAseq
Input molecule	All DNA molecules extracted from any collection of cells	All RNA molecules of one cell type
Preparation	Extraction – Fragmentation – Adapter Ligation	Extraction – Fragmentation – Reverse Transcription- Adapter Ligation
Sequencing Output File	FASTQ	FASTQ

Raw Data
FASTQ files

```
@SRR831012.1 HWI-ST155_0742:7:1101:1284:1981/1
NGAGATGAAGCACTGTAGCTTGGAATTCTCGGGTGCCAAGGAACTCCAGT
+
%1=DDDFFHHHGFIHHIIIIIIIIIIIIIIIIIEHIIIIIIIFIIIIIII
@SRR831012.2 HWI-ST155_0742:7:1101:2777:1998/1
NGAGATGAAGCACTGTAGCTCTTTGGAATTCTCGGGTGCCAAGGAACTCC
+
%1=DFFFFHHHHHIIIIIIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIIG
```

@SampleID.ReadNr

Experimental Setup

Quality score (increasing from worst to best):

```
!”#$%&’()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

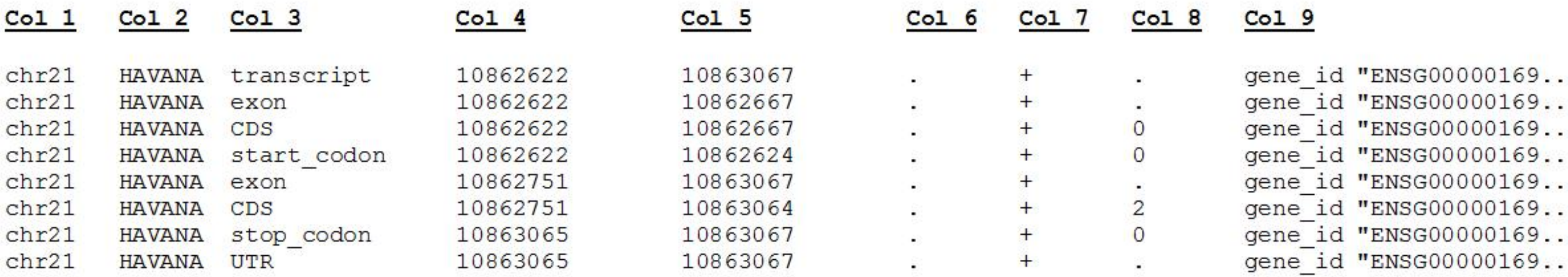
Raw Data
Reference Genome or Reference Transcriptome

- FASTA-file

```
>Sequence 1
;comment A
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCC
TGCCCCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAG
GAAAAGCAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGC
CCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCC
GCGCGCCGGGACAGAATGCCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAA
CCTCACCCATGAATGCTCACGCAAGTTTAATTACAGACCTGAA...
```

Raw Data
Gene library

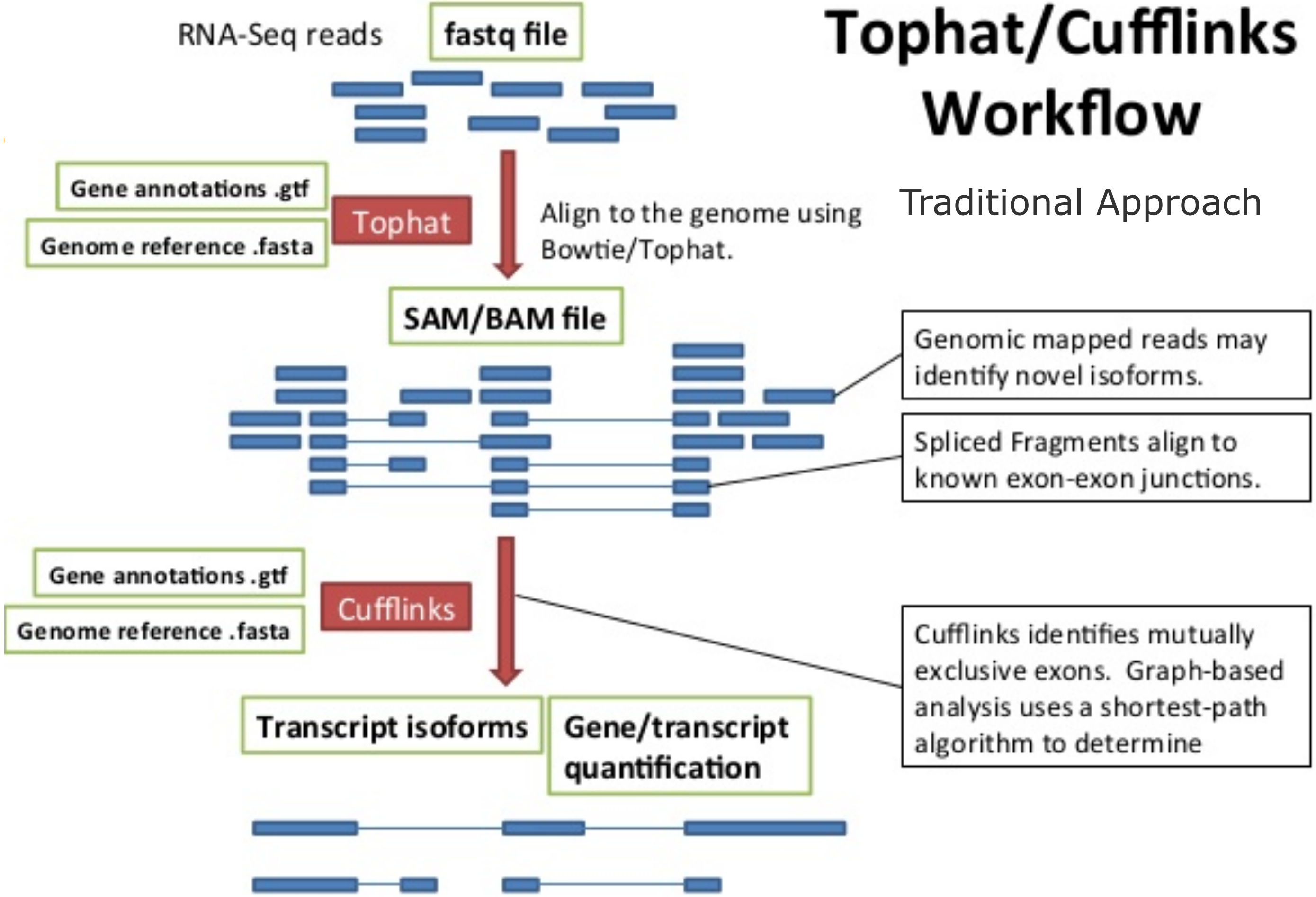
- 20k-25k protein coding genes representing small part of the genome
- Using the annotation to speed up processing
- If the discovery of new genes in a sample is expected, a custom annotation can be calculated from the reads



Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9
chr21	HAVANA	transcript	10862622	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862622	10862667	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862622	10862667	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	start_codon	10862622	10862624	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862751	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862751	10863064	.	+	2	gene_id "ENSG00000169..
chr21	HAVANA	stop_codon	10863065	10863067	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	UTR	10863065	10863067	.	+	.	gene_id "ENSG00000169..

Tophat/Cufflinks Workflow

Traditional Approach



RNA-Seq reads

fastq file

Gene annotations .gtf

Genome reference .fasta

Tophat

Align to the genome using Bowtie/Tophat.

SAM/BAM file

Genomic mapped reads may identify novel isoforms.

Spliced Fragments align to known exon-exon junctions.

Gene annotations .gtf

Genome reference .fasta

Cufflinks

Cufflinks identifies mutually exclusive exons. Graph-based analysis uses a shortest-path algorithm to determine

Transcript isoforms

Gene/transcript quantification

Selected (Splice-aware) Alignment Tools

- **TopHat:** Aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner BOWTIE, and then analyzes the mapping results to identify splice junctions between exons.
- **HISAT2:** Bowtie + multiple small Graph FM-Indices, parallel threads possible
- **STAR**: works by indexing the reference genome first, followed by producing a suffix array index to accelerate the alignment step in further processing. Being capable of running parallel threads on multi-core systems, STAR is faster in comparison with other tools

Differences and Similarities of RNA- and DNAseq Processing and Output

	DNAseq	RNAseq (traditional)
Input	Raw Reads, Reference Genome	Raw Reads, Reference Genome/ Transcriptome, Gene Library
Processing	Alignment and variant calling	Splice aware alignment and abundance estimation/counting
Output	VCF containing all variants	.csv table containing a count/ abundance estimation for all genes/ transcripts/exons
Size	kB to MB size, approx. 30 m variants/human	kB to MB size, approx. 20-25 k genes/human

New Fast Algorithms for RNAseq Quantification

- Sailfish: Facilitates the quantification of RNA-isoform abundance by totally avoiding the time-consuming mapping step. Instead of mapping, it inspects k-mers in reads to observe transcript coverage that results in a fast processing of reads.
- Kallisto: Same lightweight algorithm approach as Sailfish to quantify transcript abundance but improves it with a "pseudoalignment" process

Differential Gene Expression (DGE) Analysis

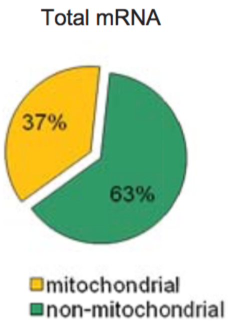
- Goal: Identify genes that change in abundance between conditions, i.e., they differ in counts in different conditions.
- Input: Count table, e.g., GenesxSamples, and design formula
- Processing:
  - □ Cleansing, normalization, log-transformation
  - □ Clustering and PCA
  - □ Estimates variance-mean dependence in count data
  - □ Calculates differences in expression values for groups of samples
- Output: Table containing fold changes

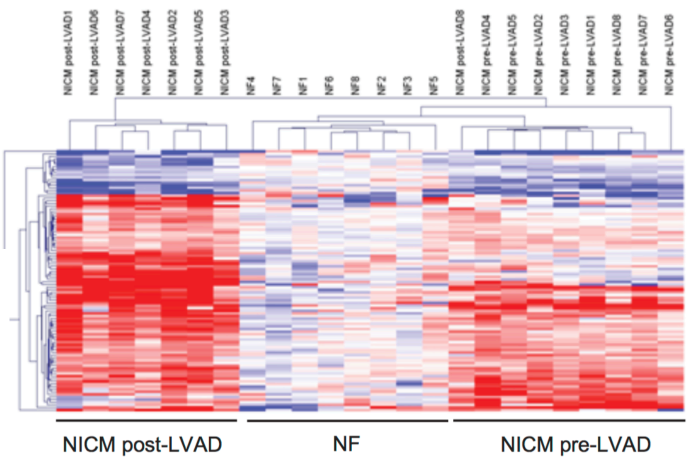
Differential Gene Expression Analysis
Selected Algorithms/Packages

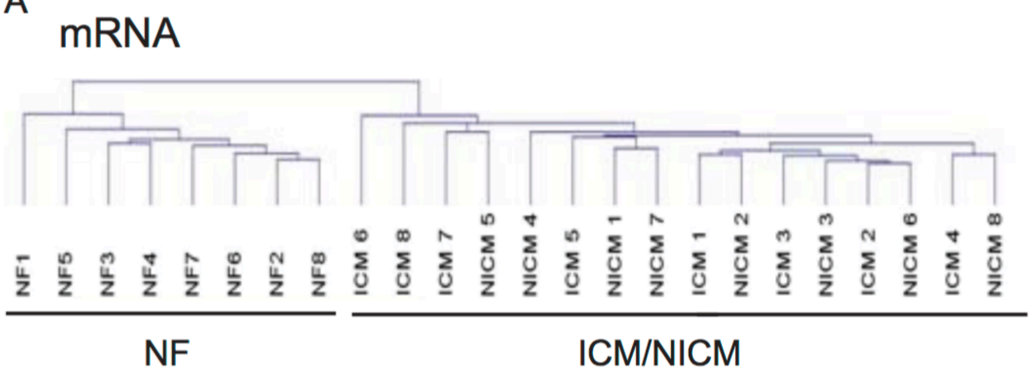
- In order to perform a statistical test on the differences, the distribution of data needs to be known
- Statistical methods needed to calculate significant differences
  - □ Poisson distribution – variance and mean of expression values are equal over samples – PoissonSeq → only for large data sets
  - □ Negative binomial distribution - current gold standard as in DESeq and edgeR as they are applicable for small data sets
  - □ Abundance estimation via bootstrapping – trend as in kallisto/sleuth

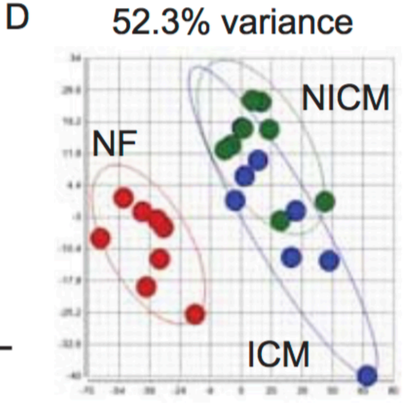
Results of DGE Analysis

- Common analysis:
  - □ Principal Component Analysis
  - □ Volcano and MA plots
  - □ Clustered Heatmaps





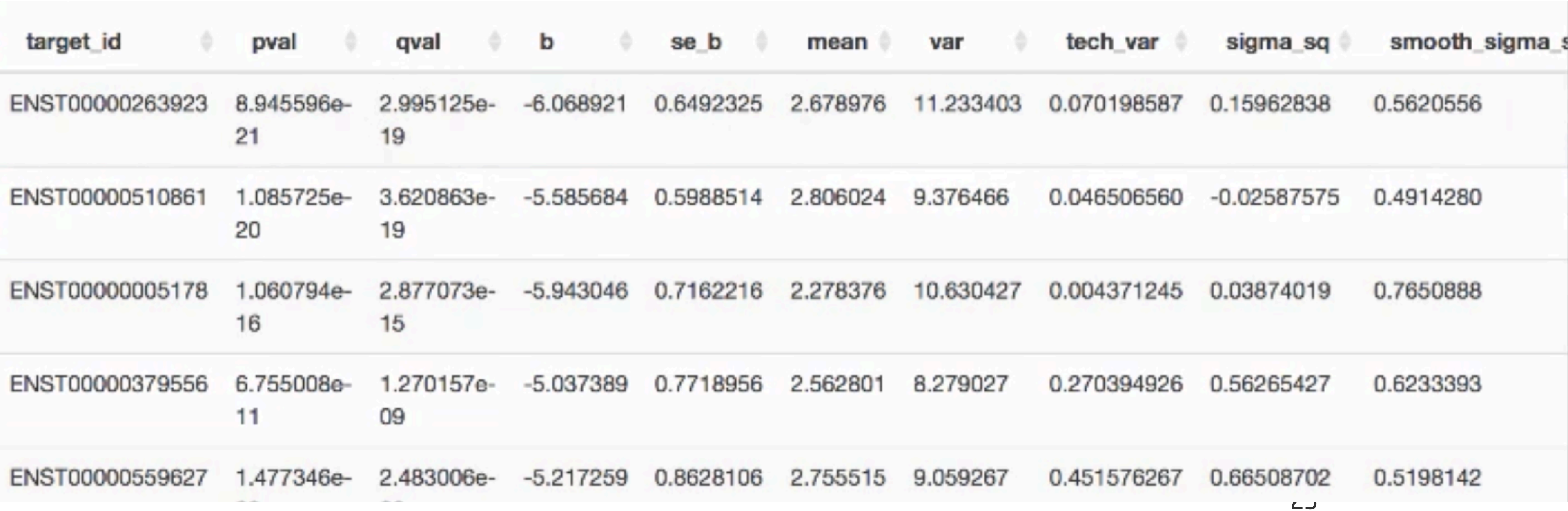




Legend: NF – Non Failing Heart, (N)ICM – (Non-)Ischemic Cardiomyopathy, LVAD – Left Ventricular Assisting Device

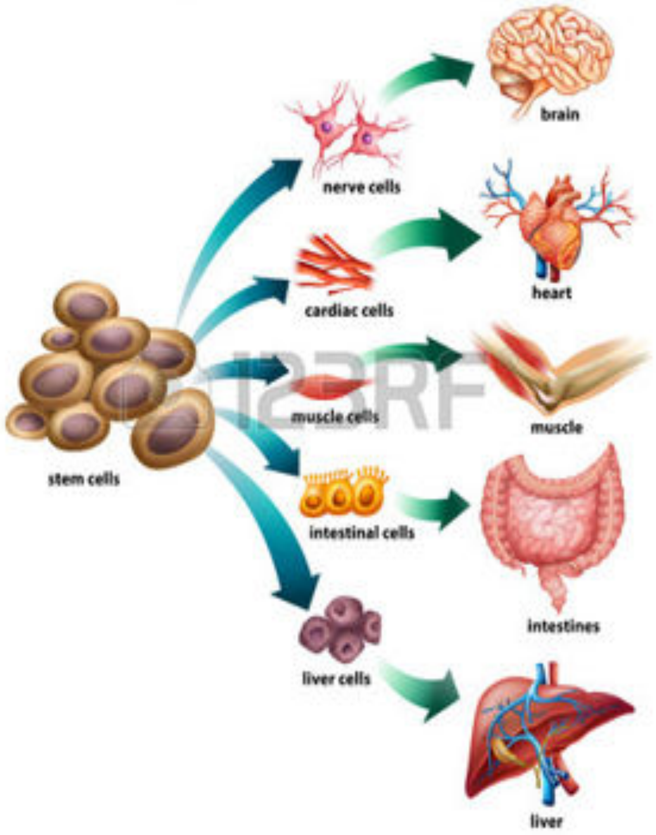
Results of DGE Analysis

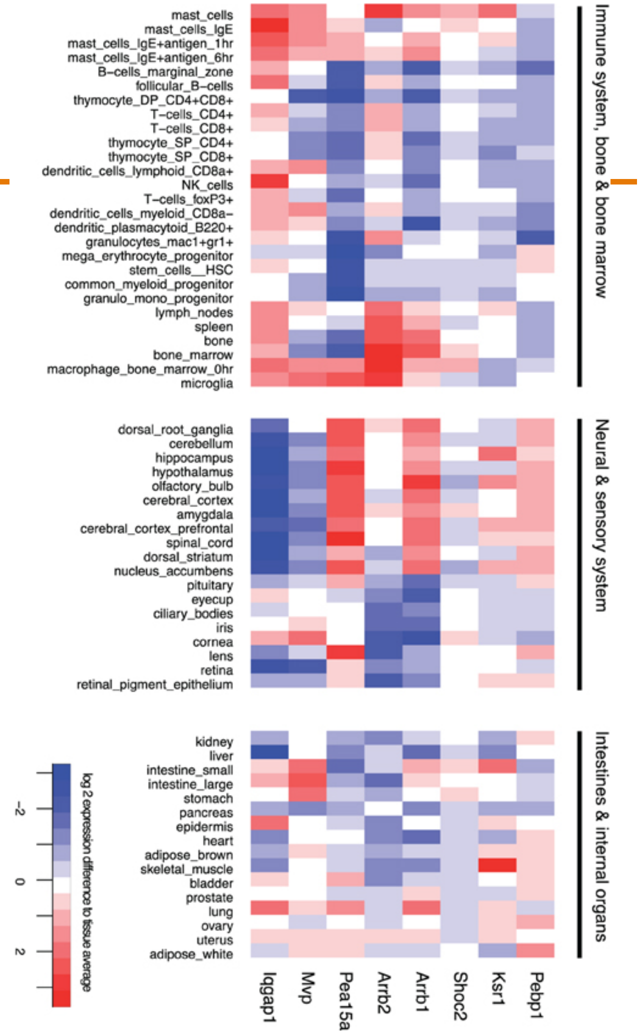
- List of differentially expressed genes and a p-value



target_id	pval	qval	b	se_b	mean	var	tech_var	sigma_sq	smooth_sigma_s
ENST00000263923	8.945596e-21	2.995125e-19	-6.068921	0.6492325	2.678976	11.233403	0.070198587	0.15962838	0.5620556
ENST00000510861	1.085725e-20	3.620863e-19	-5.585684	0.5988514	2.806024	9.376466	0.046506560	-0.02587575	0.4914280
ENST00000005178	1.060794e-16	2.877073e-15	-5.943046	0.7162216	2.278376	10.630427	0.004371245	0.03874019	0.7650888
ENST00000379556	6.755008e-11	1.270157e-09	-5.037389	0.7718956	2.562801	8.279027	0.270394926	0.56265427	0.6233393
ENST00000559627	1.477346e-	2.483006e-	-5.217259	0.8628106	2.755515	9.059267	0.451576267	0.66508702	0.5198142

Results of DGE Analysis
Tissue specificity





Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Recap Clustering

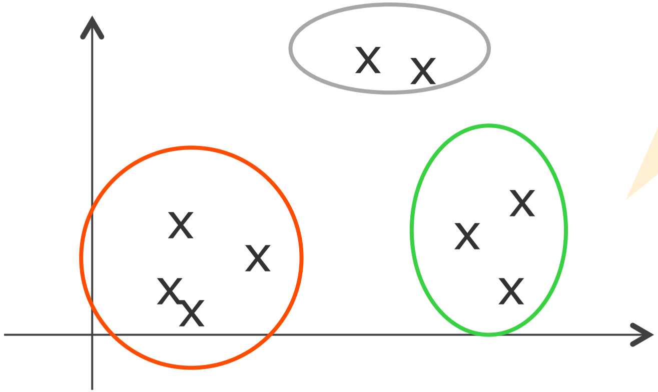
Milena Kraus

Data Management for Digital Health

Summer 2017

Introduction/Recap
Clustering

- The goal is to group data points that are similar to each other and identify such groupings in an unsupervised manner
- For 40% of sequenced genes, functionality cannot be ascertained by comparing to sequences of other known genes
  - □ But, if genes A and B are grouped in the same microarray cluster, then we hypothesize that proteins A and B might interact with each other and conduct experiments to confirm



However, co-expressed genes often don’t always imply similar function!

Example: Gene Expression Data

- Gene expression data are usually transformed into an intensity matrix (below)
- The intensity matrix allows biologists to make correlations between different genes (even if they are dissimilar)
- Make a distance matrix for the distance between every two gene points
- Genes with a small distance share the same expression characteristics and might be functionally related or similar.

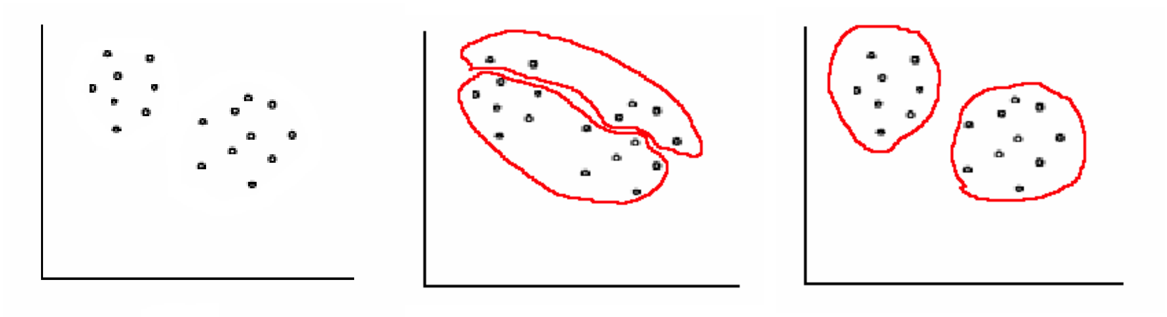
Intensity (expression level) of gene at measured time

Time:	Time X	Time Y	Time Z
Gene 1	10	8	10
Gene 2	10	0	9
Gene 3	4	8.6	3
Gene 4	7	8	3
Gene 5	1	2	3

Separation Principles

There are two key grouping principles to keep in mind

- **Homogeneity**: Elements within a cluster are close to each other
- **Separation**: Elements in different clusters are further apart from each other



Summary of Similarity Measures

- Using different measures for clustering can yield different clusters
- Common choices for gene expression data: Euclidean distance and correlation distance

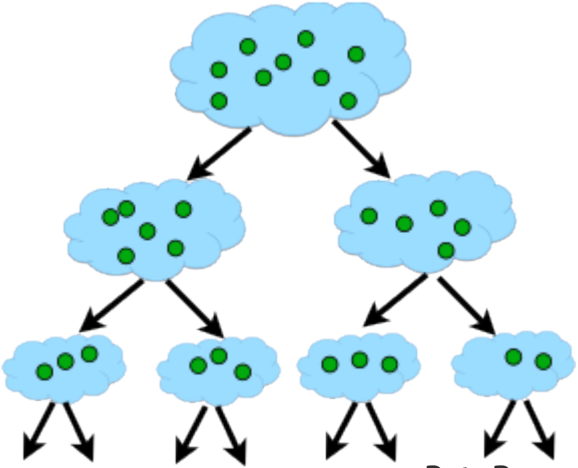
Euclidean vs Correlation Example

- g1 = (1,2,3,4,5)
- g2 = (100,200,300,400,500)
- g3 = (5,4,3,2,1)
- Which genes are similar according to the two different measures?

Hierarchical Clustering

Given input set S, the goal is to produce a hierarchy (dendrogram) in which nodes represent subsets of S.

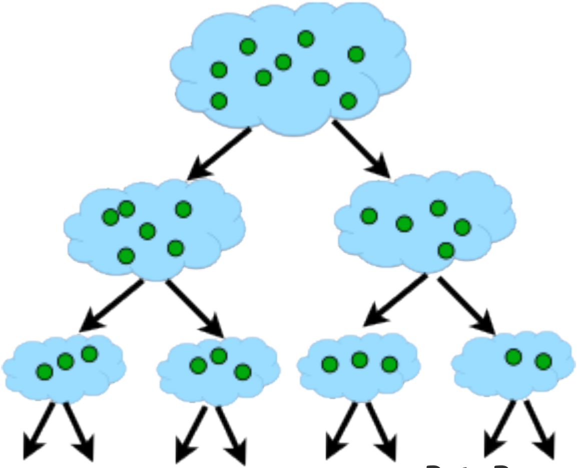
- The **root** is the whole input set S.
- The **leaves** are the individual elements of S.
- The **internal nodes** are defined as the union of their children.



Hierarchical Clustering Approaches

- Agglomerative (bottom-up):
  - Beginning with singletons (sets with 1 element)
  - Merging them until S is achieved as the root.
  - Most common approach.

- Divisive (top-down):
  - Recursively partitioning S until singleton sets are reached.



Hierarchical Clustering Algorithm

- Input: a pairwise matrix involved all instances in S

1. Place each instance of S in its own cluster (singleton), creating the list of clusters L (initially, the leaves of T):

   L= S_1, S_2, S_3, ..., S_{n-1}, S_n.
2. Compute a merging cost function between every pair of elements in L to find the two closest clusters {S_i, S_j} which will be the cheapest couple to merge.
3. Remove S_i and S_j from L.
4. Merge S_i and S_j to create a new internal node S_{ij} in T which will be the parent of S_i and S_j in the resulting tree.
5. Go to **Step 2** until there is only one set remaining.

Similarity measures

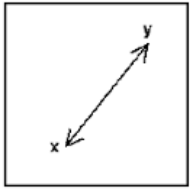
- How to determine similarity between data observations?

- Let **x** = (x_1,...,x_n) and **y** = (y_1,...y_n) be n-dimensional vectors of data points of objects g_1 and g_2
  - □ g_1, g_2 can be two different genes in gene expression data
  - □ n can be the number of samples

Distance measure

- Euclidean distance

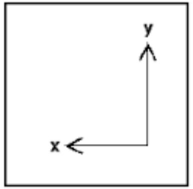
$$d(g_1, g_2) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



Euclidean

- Manhattan distance

$$d(g_1, g_2) = \sum_{i=1}^{n} \left|(x_i - y_i)\right|$$



Manhattan

- Minkowski distance

$$d(g_1, g_2) = \sqrt[m]{\sum_{i=1}^{n} (x_i - y_i)^m}$$

Correlation distance

- Correlation distance

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{(Var(X) \cdot Var(Y)}}$$

- Cov(X,Y) stands for covariance of X and Y
  - □ degree to which two different variables are related

- Var(X) stands for variance of X
  - □ measurement of how a sample differs from their mean

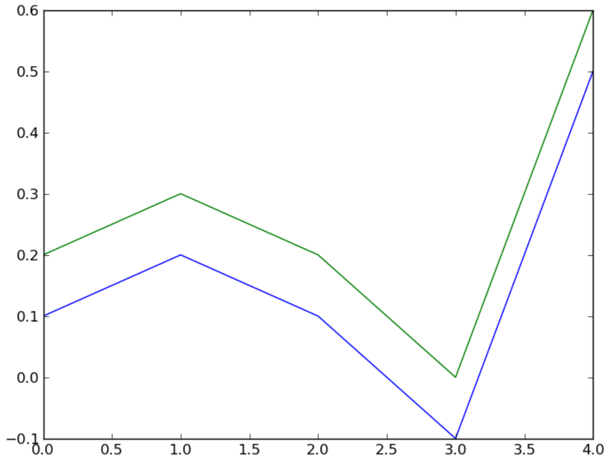
Correlation distance

- Correlation

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{(Var(X)\cdot Var(Y)}}$$

- maximum value of 1 if X and Y are perfectly correlated
- minimum value of -1 if X and Y are exactly opposite
- distance(X,Y) = 1 - r_{xy}

r = 1 (greens and blues variation is equal)



Hierarchical Clustering Algorithm

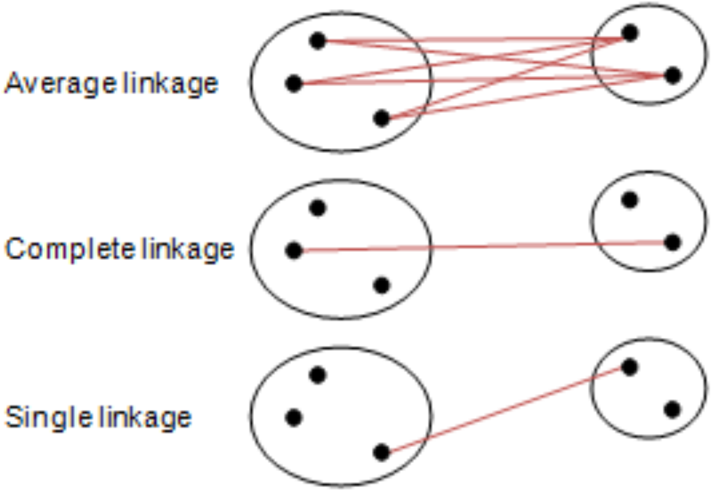
- Input: a pairwise matrix involved all instances in S

1. Place each instance of S in its own cluster (singleton), creating the list of clusters L (initially, the leaves of T):

   L= S_1, S_2, S_3, ..., S_{n-1}, S_n.
2. Compute a merging cost function between every pair of elements in L to find the two closest clusters {S_i, S_j} which will be the cheapest couple to merge.
3. Remove S_i and S_j from L.
4. Merge S_i and S_j to create a new internal node S_{ij} in T which will be the parent of S_i and S_j in the resulting tree.
5. Go to **Step 2** until there is only one set remaining.

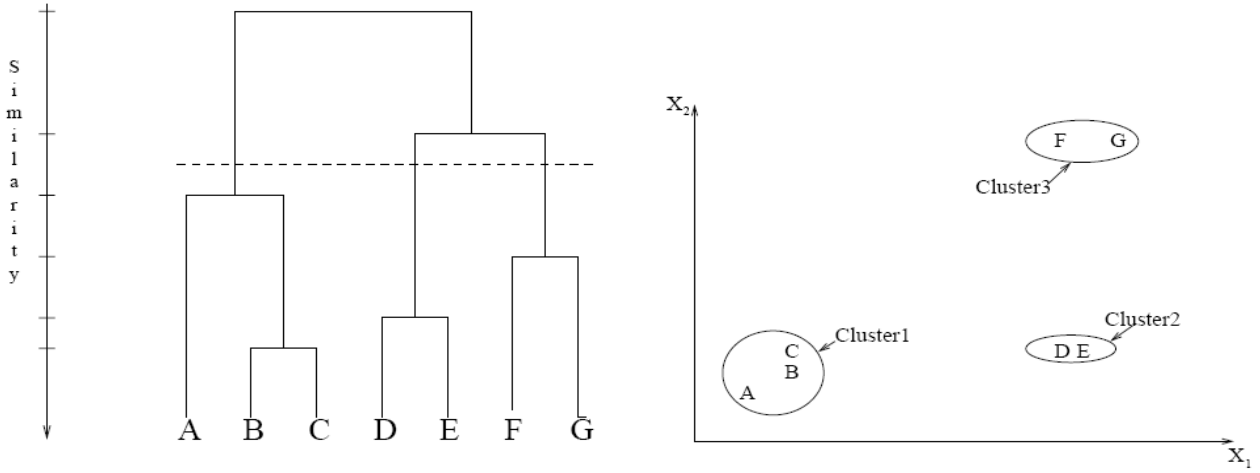
Hierarchical Clustering
Linkage Strategy

- Average linkage: The distance between two clusters is the average of the distances between all the points in those clusters.
- Complete linkage: The distance between two clusters is the distance between the furthest points in those clusters.
- Single linkage: The distance between two clusters is the distance between the nearest neighbors in those clusters.



Hierarchical Clustering
Dendrograms

- The algorithm computes a dendrogram which can then be visualized graphically
- The tree can be pruned to the needed/expected amount of clusters



Hierarchical Clustering

Advantages

- Dendograms are great for visualization and pruning
- Provides hierarchical relations between clusters
- Shown to be able to capture concentric clusters

Disadvantages

- Not easy to define levels for clusters
- Experiments showed that other clustering techniques outperform hierarchical clustering

K-Means Clustering Algorithm

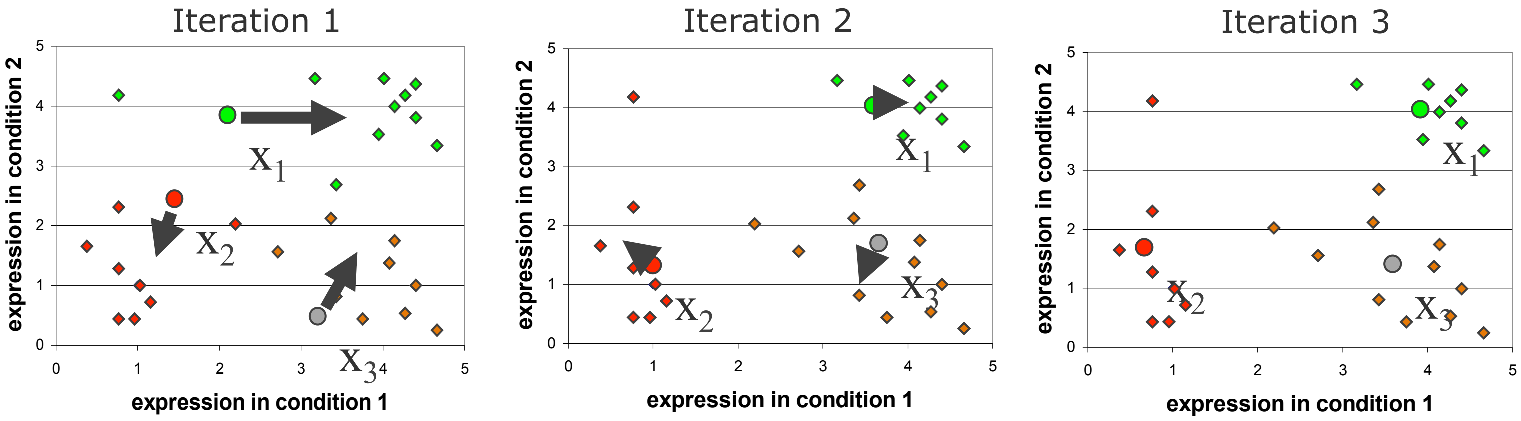
- Input: a pairwise matrix involved all instances in S

1. Randomly place **K points** into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat **Steps 2 and 3** until the stopping criteria is met.

Stopping Criteria

- Convergence (No further changes.)
- Maximum number of iterations.
- Or when the squared error is less than some small threshold value α

K-Means Clustering



K-Means Clustering

Pros:

- Low complexity
- Fast
- Has shown to be efficient in biomedical use cases

Disadvantages:

- Necessity of specifying k
- Sensitive to noise and outlier data points
  - A small number of skewed data can influence the mean value
- Clusters are sensitive to initial assignment of centroids
  - K-means is not a deterministic algorithm
  - Clusters can be inconsistent from one run to another

Evaluation of clusters

Why validity of clusters?

- Given *some* data, any clustering algorithm generates clusters
- So we need to make sure the clustering results are valid and meaningful

Measuring the validity of clustering results usually involve

- Optimality of clusters
- Verification of biological meaning of clusters

Optimality of clusters

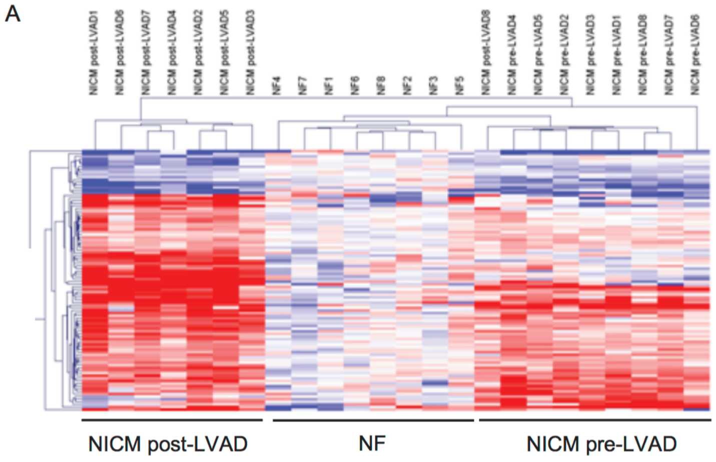
- Optimal clusters should
  - □ minimize distance **within** clusters (intracluster)
  - □ maximize distance **between** clusters (intercluster)
- Example of intracluster measure
  - – Squared error *se*

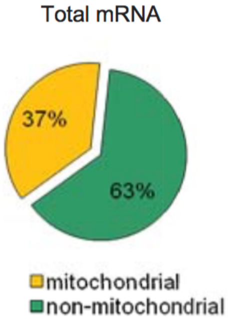
$$se = \sum_{i=1}^{k} \sum_{p \in c_i} \left\| p - m_i \right\|^2$$

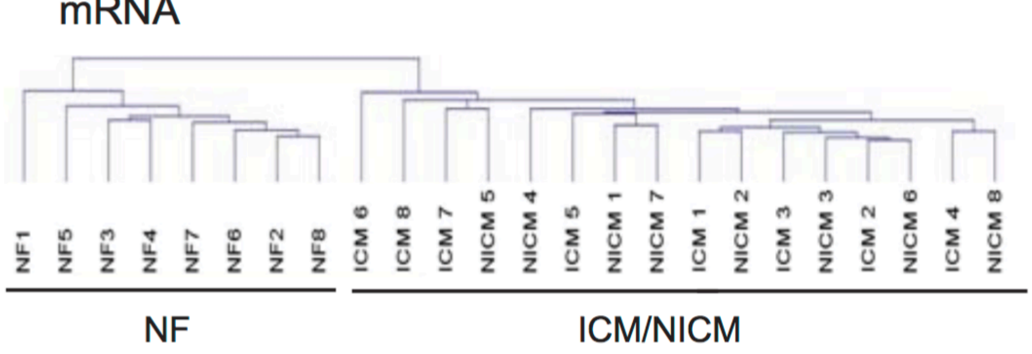
where m_i is the mean of all instances in cluster c_i

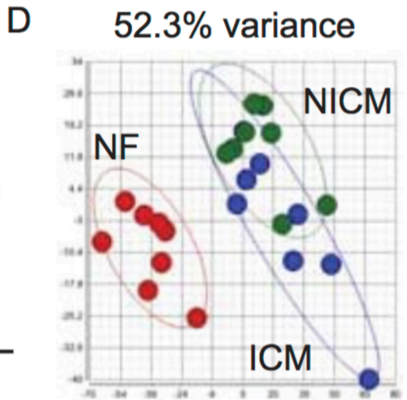
Results of DGE Analysis

- Common analysis:
  - □ Principal Component Analysis
  - □ Volcano and MA plots
  - □ Clustered Heatmaps









Legend: NF – Non Failing Heart, (N)ICM – (Non-)Ischemic Cardiomyopathy, LVAD – Left Ventricular Assisting Device

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Recap Dimensionality Reduction

Milena Kraus

Data Management for Digital Health
Summer 2017

Dimensionality Reduction
PCA/MCA/MFA

The main objective is to sum up and to simplify the data by reducing the dimensionality of the data set. Those methods are used depending on the type of data at hand whether variables are quantitative (numerous) or qualitative (categorical or nominal):

- Principal component analysis (PCA) when observations are described by quantitative variables
- Multiple correspondence analysis (MCA) when observations are described by categorical variables
- Multiple factor analysis (MFA) when observations are described by both numerical and categorical variables

Principle Component Analysis (PCA)
Input from RNAseq analysis

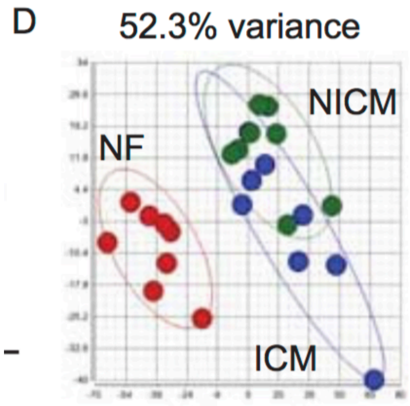
Input:

- Typical data set derived from an RNAseq experiment
- Gene vs. Sample table (e.g. 20 k genes vs. 20 samples), i.e. high dimensional set of numerical values

Gene	Sample 1	Sample 2
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42

Output:

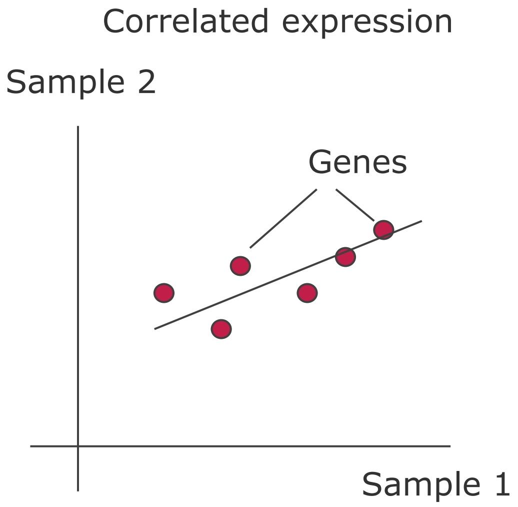
- One (or many) two-dimensional representation of all samples

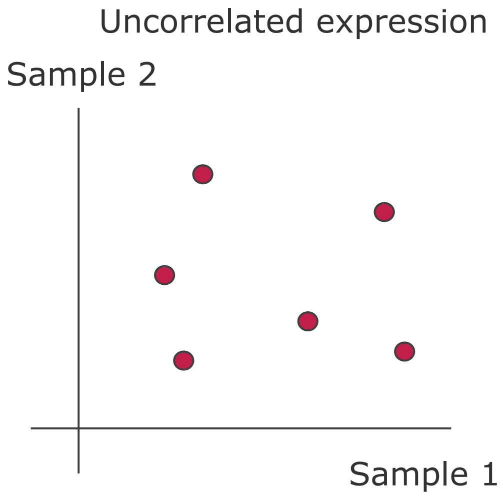


Very good and easy video on PCA on RNAseq data:
https://www.youtube.com/watch?v=_UVHneBUBW0

Principle Components Analysis

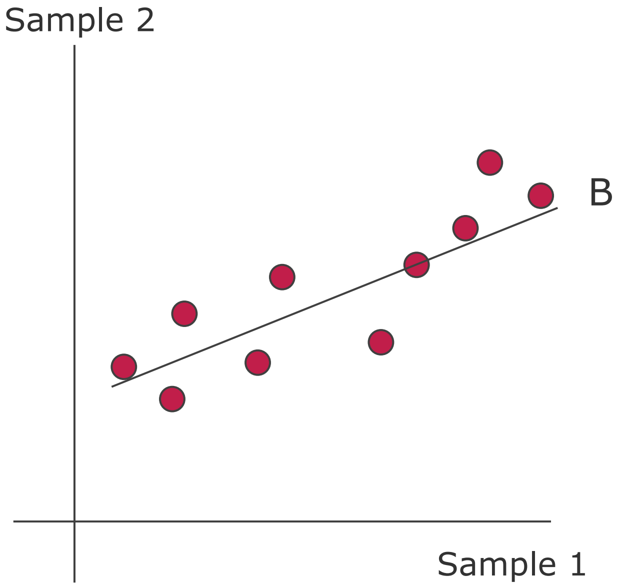
- Consider the expression values of 2 samples and various genes.





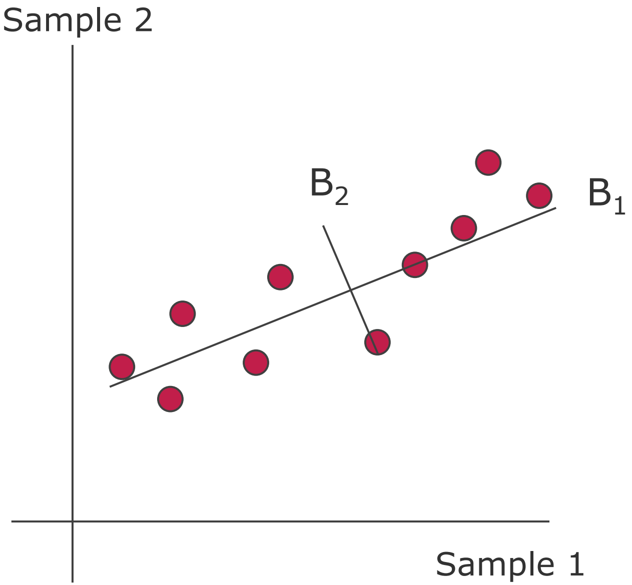
Principle Components Analysis

- Consider the mean of all points m, and a vector B going through the mean
- The vector B (PC1) is stretched along the path of most variation



Principle Components Analysis

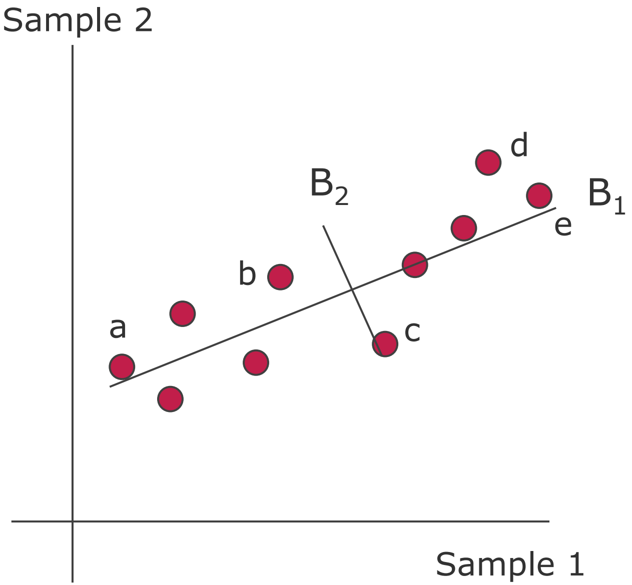
- Consider the mean of all points m, and a vector B going through the mean
- The vector B (PC1) is stretched along the path of most variation
- Vector B_2 (PC2) is stretched along the path of second most variation and orthogonal to B_1
- Length and orientation of the B vectors are most influenced by the outer points



Principle Components Analysis

- Length and orientation of the B vectors are most influenced by the outer points

Gene	Influence on PC1	Influence on PC2
a	High (10)	Low (1)
b	Low (2)	Medium (3)
c	Low (0.5)	High (-5)
d	high (-9)	High (5)
e	High (-10)	Low (1)



Principle Component Analysis

- Principal component scores for each sample

Gene	Sample 1	Sample 2
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42

Gene	Influence on PC1	Influence on PC2
a	10	1
b	2	3
c	0.5	-5
d	-9	5
e	-10	1

PC score(Sample) = (read count (gene a) * PC1_influence (gene a)) + (read count (gene b) * PC1_influence(gene b)) +)

Principle Components Analysis

- Principal component scores for each sample

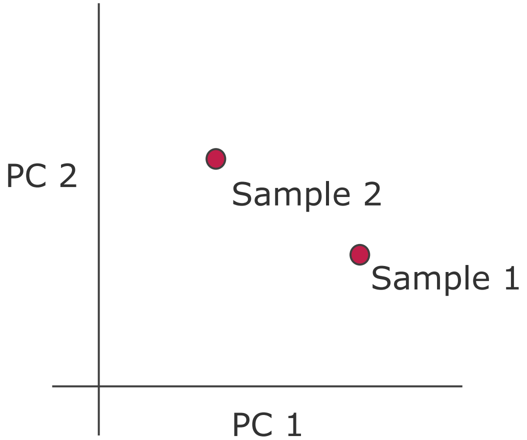
Gene	Sample 1	Sample 2
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42

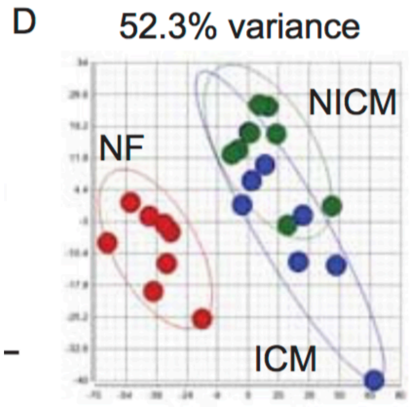
Gene	Influence on PC1	Influence on PC2
a	10	1
b	2	3
c	0.5	-5
d	-9	5
e	-10	1

PC score(Sample1) =(10* 10) + (0 * 2) + (14 * 0.5) + (33*(-9)) +(50*(-10)) = -690

Principle Component Analysis
Individuals Factor Map

- Samples are plotted onto the principal components dimensions
- Clusters may form to discriminate different groups of samples

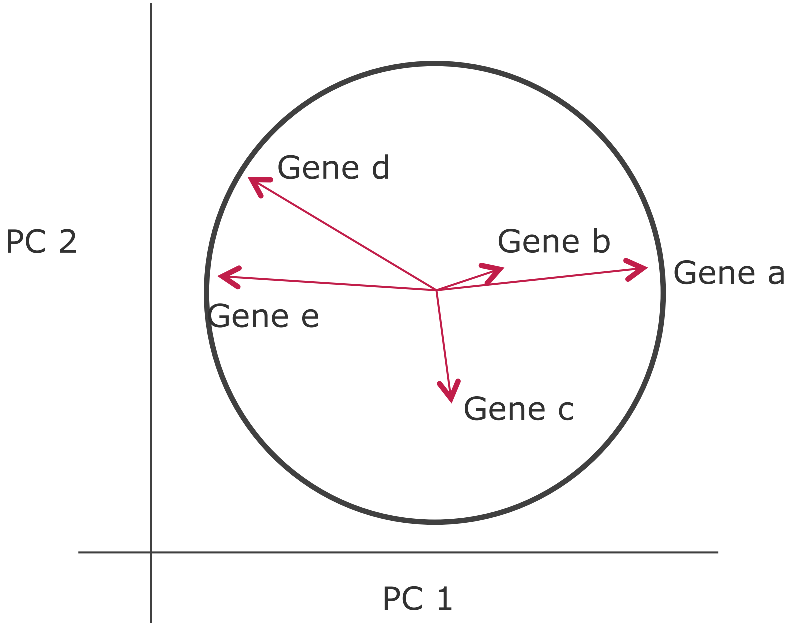




	PC1	PC2
Sample 1	-690	155
Sample 2	-904	232

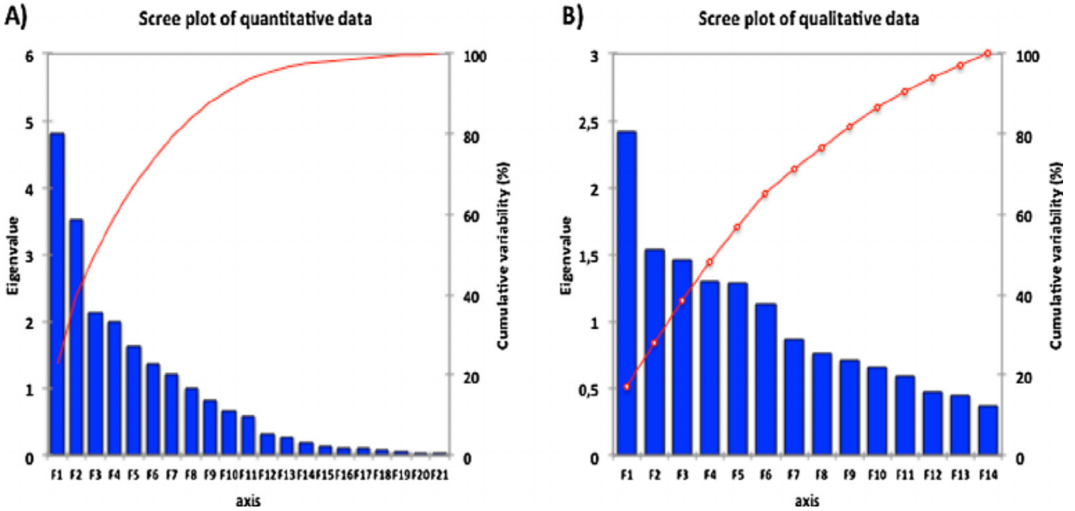
Principle Component Analysis
Variables Factor Map

- The variables factor map visualizes the influences (loadings) of each gene on the principal components
- Genes with high influences are good candidates to conduct further analysis on



Principal Component Analysis Diagnostics

- Scree plot
  - □ Distribution of variance for principal components
  - □ The more variance is explained by the first few components, the better the PCA
  - □ Tail of right-most PC's is mostly noise
- Cumulated variances add up to 100 %
- May be used to reduce/prune for further clustering and cleansing of data



https://www.researchgate.net/publication/282317768_Diversity_of_morpho-physiological_traits_in_worldwide_sweet_cherry_cultivars_of_GeneBank_collection_using_multivariate_analysis/figures?lo=1

Principal Component Analysis
Summary

Advantages:

- Low noise sensitivity
- Decreased requirements for capacity and memory
- Large variance = low covariance = high importance → everything else is supposed to be noise and may be removed

Disadvantages:

- Relies on linear assumptions → If the correlation between, e.g., two genes is not linearly correlated, PCA fails
- Relies on orthogonal transformations → PC's are supposed to be orthogonal to each other, limiting the possibility to find others with higher variance than the orthogonals
- Scale variant

Thank you!