Bioinformatics Lab

Ivan Gesteira Costa, Mingbo Cheng, James Nagai
Institute for Computational Genomics

Institute for Computational Genomics
01011011010
10100100101

RWTH AACHEN UNIVERSITY

Objectives

- Hands on introduction to bioinformatics programming
- Review basic biological/computational aspects
    1. basics of molecular biology
    2. basics of sequencing
    3. basics bioinformatics problems
        - short sequences read alignment
        - gene expression matrix
        - clustering and interpretation

Objectives

- **Introduction to Bioinformatics Frameworks/Tools**

  1. **biological sequence data formats/handling**
     - **Biopython, Pysam, R/bioconductor**
  2. **bioinformatics tools**
     - **BWA (aligner), Seurat, Cell Ranger, …**

Grading/Online material

Evaluation:

- **20% prototypes**
- **60% final project**
- **20% presentation**

Extra-work for media informatics:

- **research report**

References/Courses Online

http://costalab.org/teaching/bioinformatics-software-lab-2020/

Introduction to Molecular Biology

Institute for
Computational Genomics
01011011010
10100100101
RWTH AACHEN
UNIVERSITY

Understanding Live in a Molecular Level

How is genetic information inherited?

How the genetic information influence cellular processes?

How genes work together to promote particular molecular functions?

Genetic Information - DNA

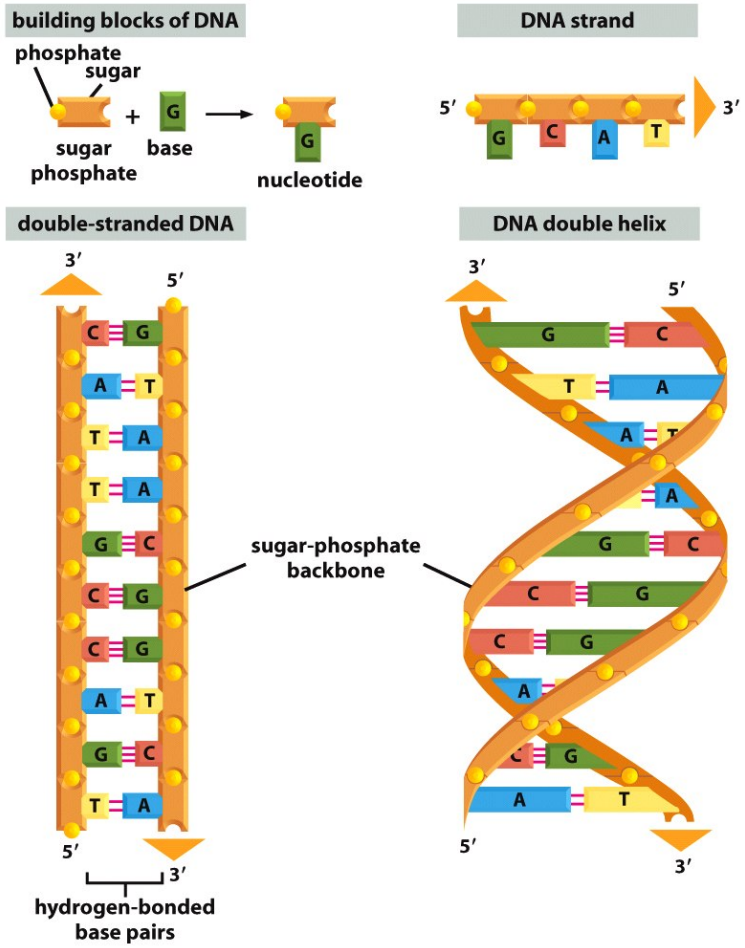
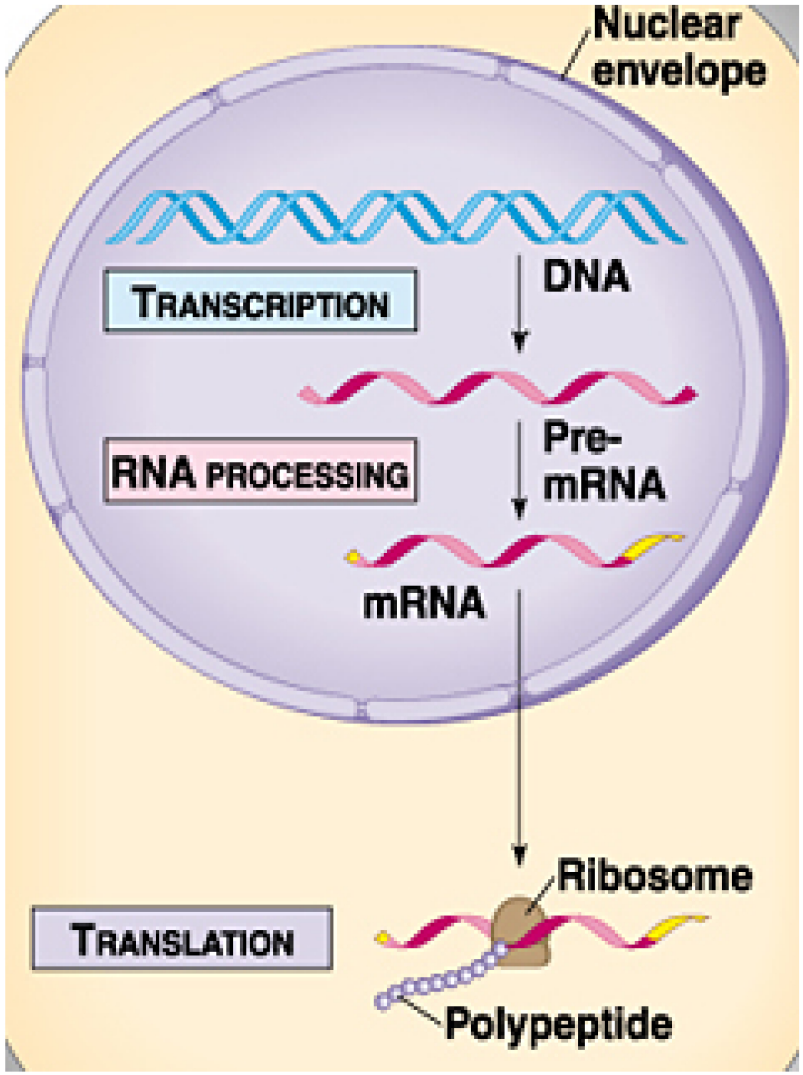


Figure 4-3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

DNA *(Deoxyribonucleic)*

- chain of nucleic acids
- 4 bases: A;C;G;T
- forms DNA duplexes with paring A = T e C = G

Central Dogma - Transcription



Transcription

- ***DNA to RNA***

RNA (ribonucleic acid)

- **single stranded**
- **4 bases: A;C;G;U**
- **unstable**
- **transport of information from nucleus to cytoplasm**

Central Dogma - Transcription

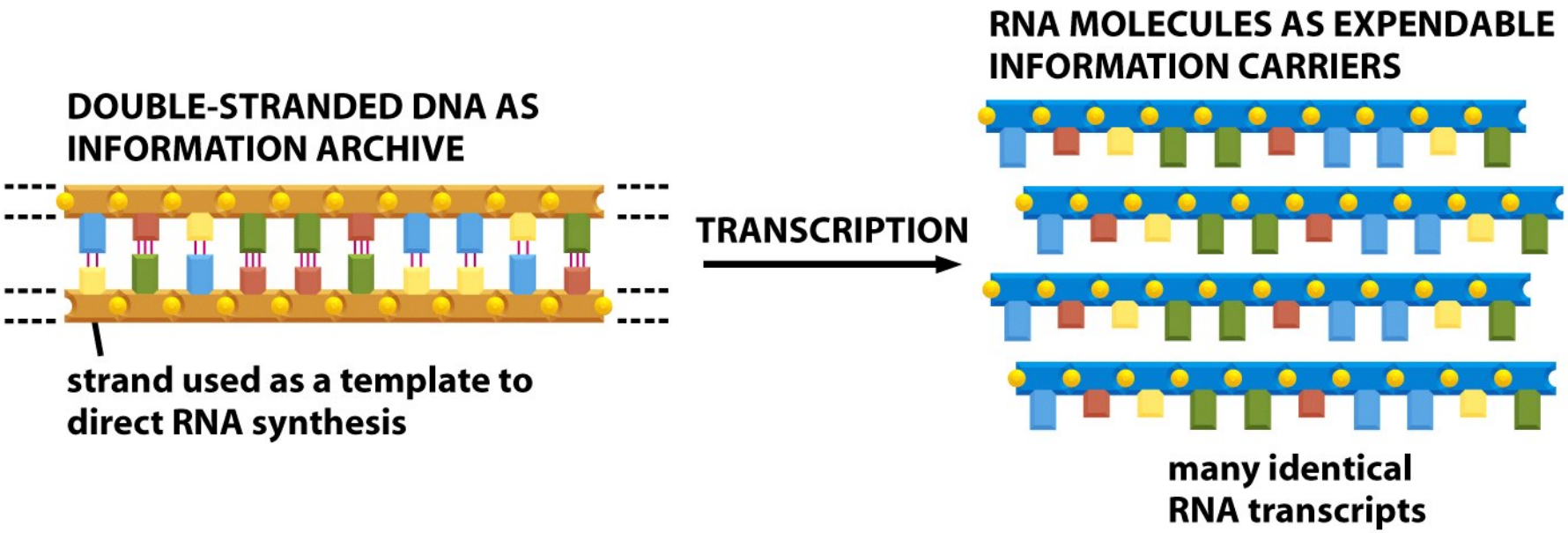
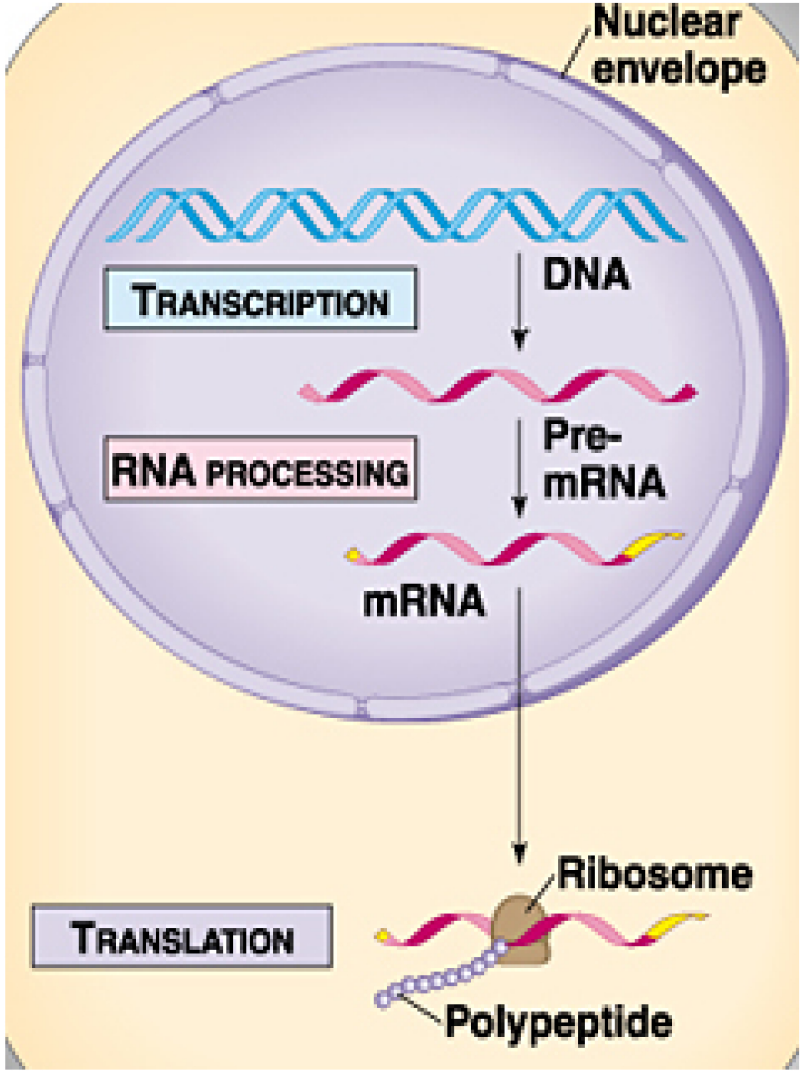


Figure 1-5 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Transcription - copy of DNA information to RNA (T to U)

Central Dogma - Translation



Translation

- ***RNA to Protein***
- **performed by the ribosome**
- **follows the genetic code**

Proteins

- **single stranded chain**
- **20 amino acids**
- **assumes 3D structure**
- **main functional entities in the cell**

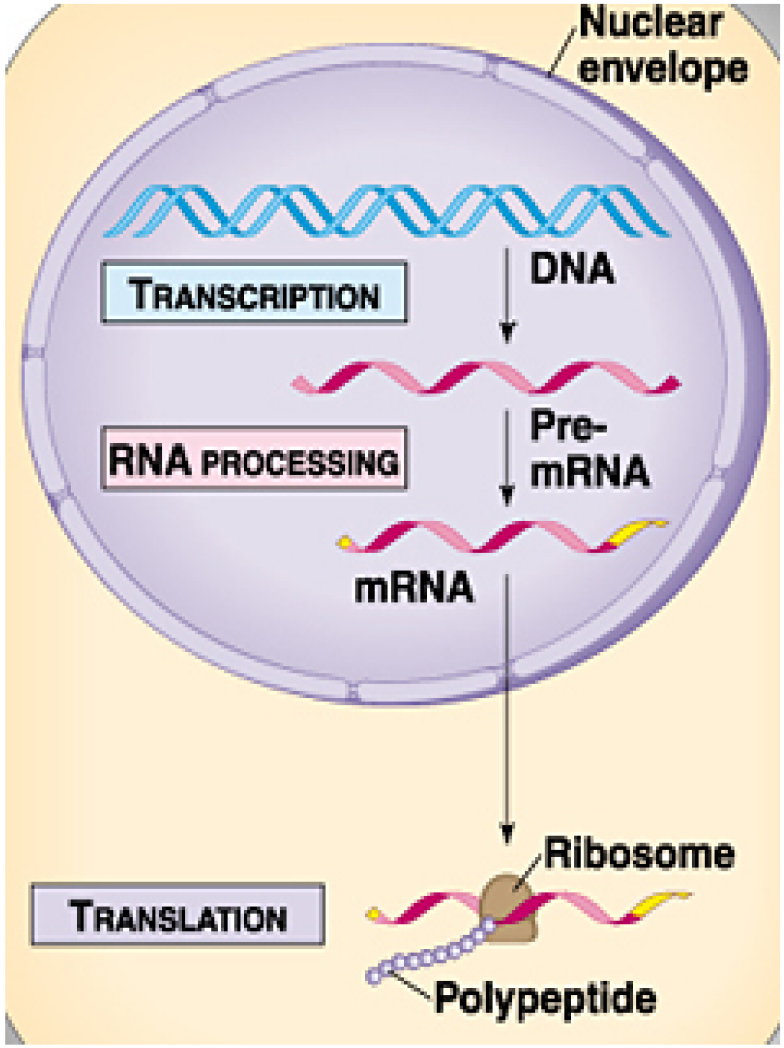
Genetic Code - Translation

GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Figure 6-50 Molecular Biology of the Cell 5/e (© Garland Science 2008)

triples of RNA bases encodes a amino acid

Central Dogma



- Dogma: information flux DNA -> mRNA -> Proteins
- Gene: DNA segment coding a protein.
- Transcript: RNA segment associated to a gene.
- Genes is associated to one proteins and one function*

* Genes might be associated to many proteins

Control of Gene Expression

How is the expression of genes controlled?

Certain proteins (transcription factors) bind to DNA and initiate transcription

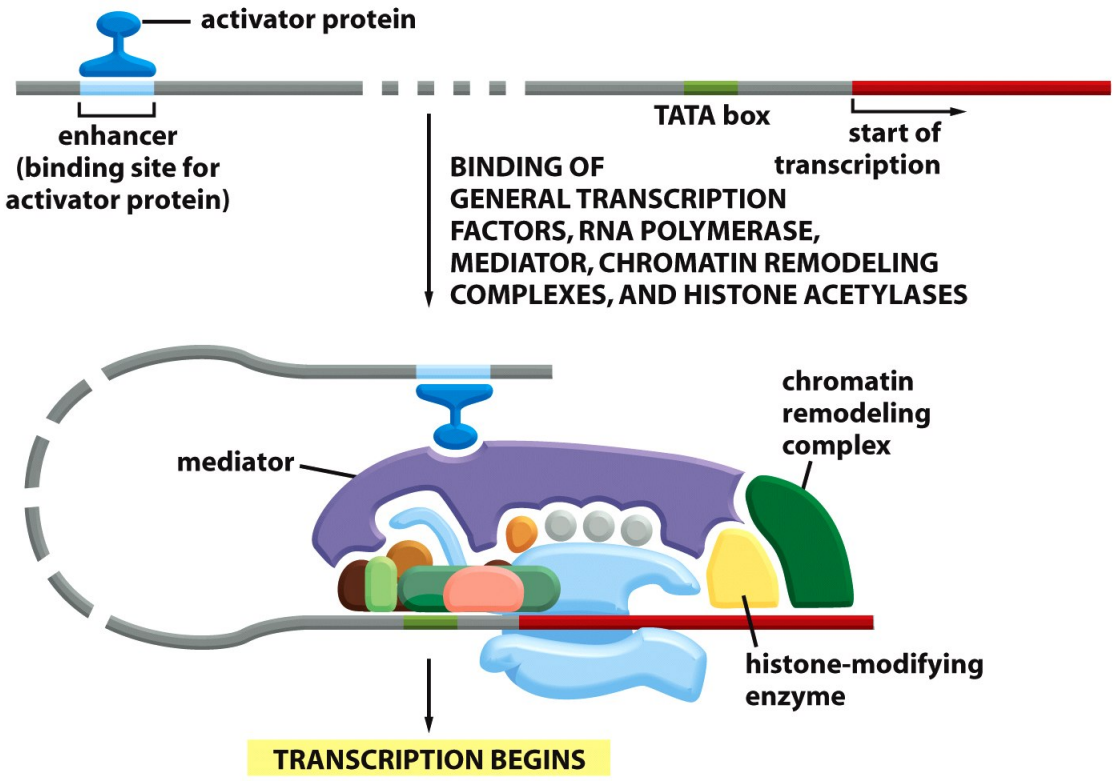


Figure 6-19 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Gene Expression

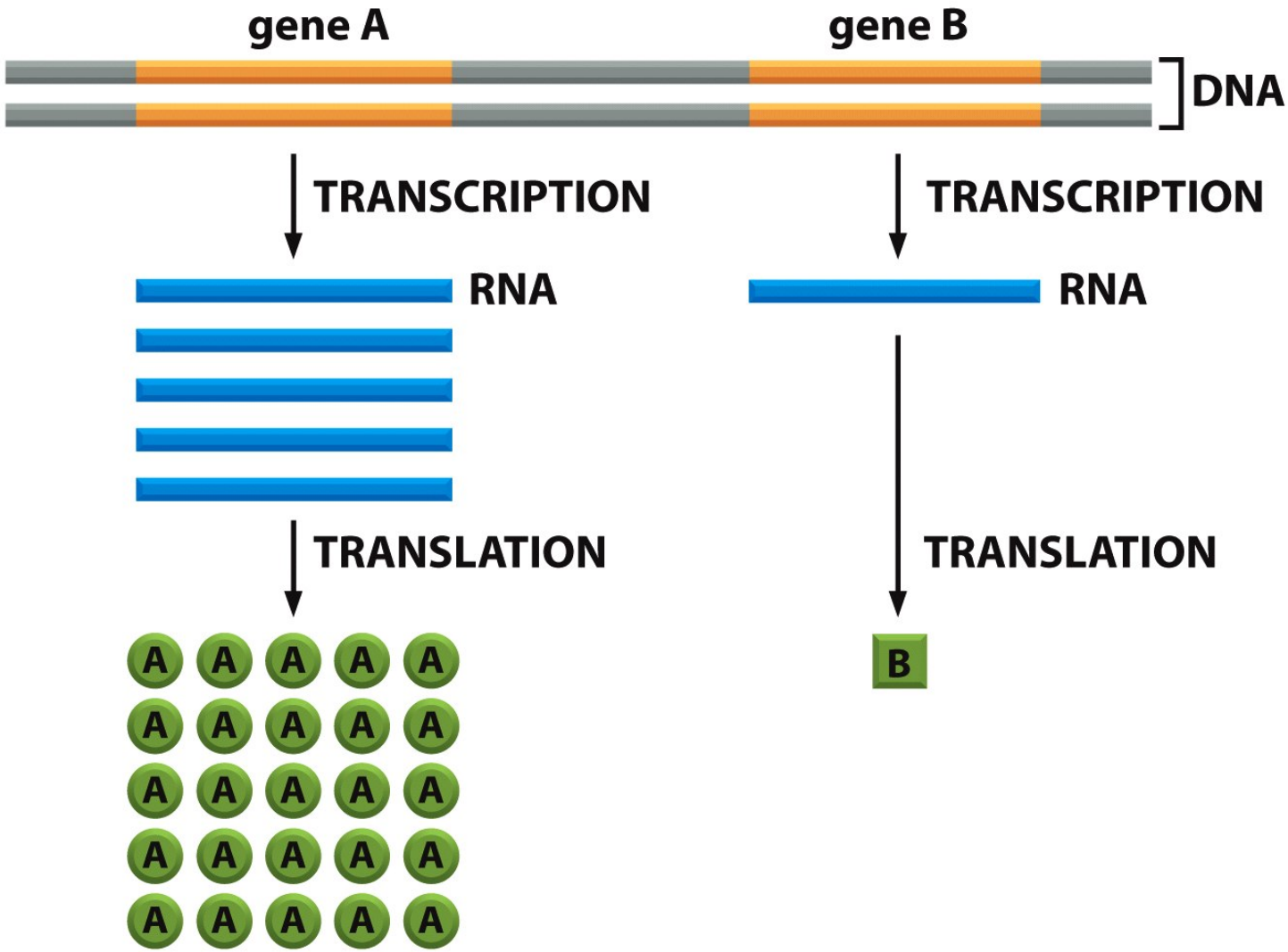
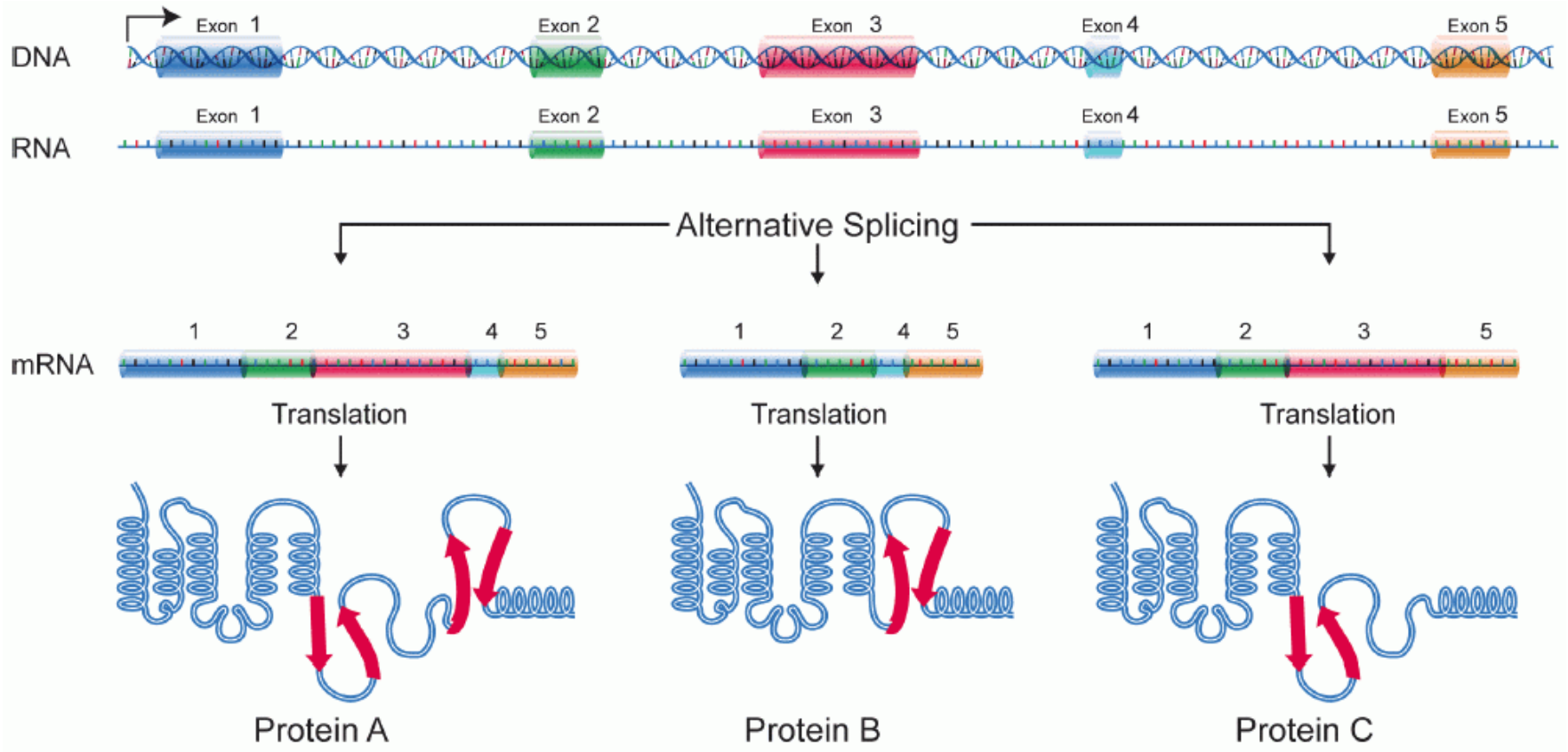


Figure 6-3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Gene / Alternative Splicing



Cellular Complexity

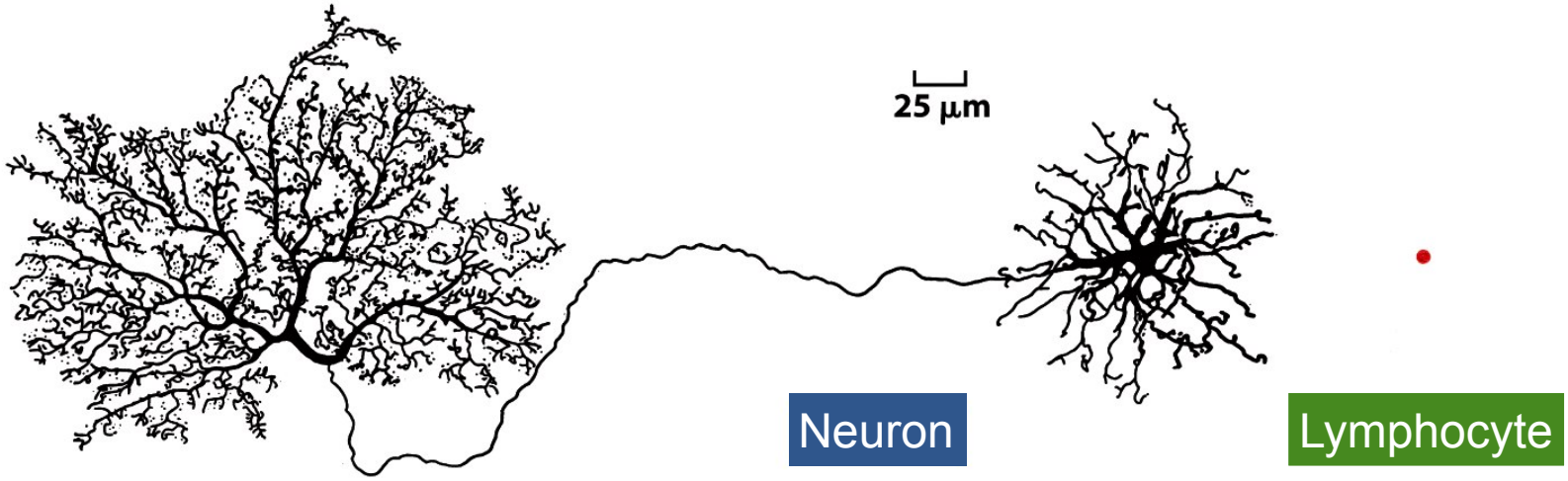


Figure 7-1 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Two cells of a organism have exactly* the same DNA

How does this differences arise?

How is cell fate remembered?

*** with exception of somatic mutations and rearrangements of immunological loci**

Cellular Complexity & Gene Expression

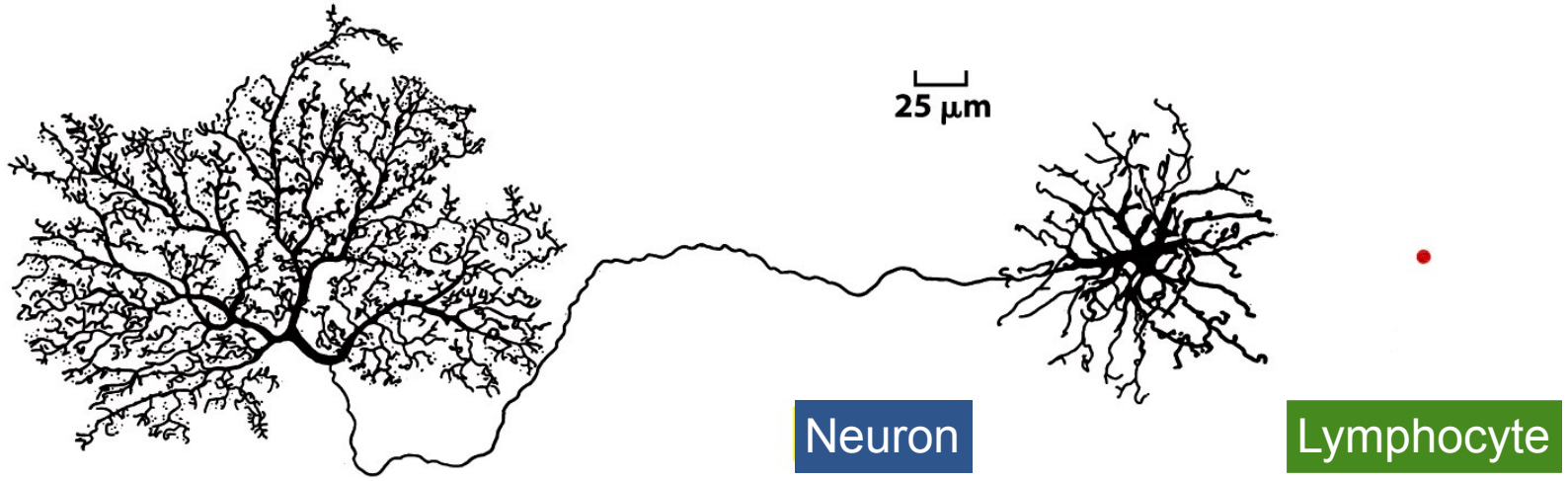
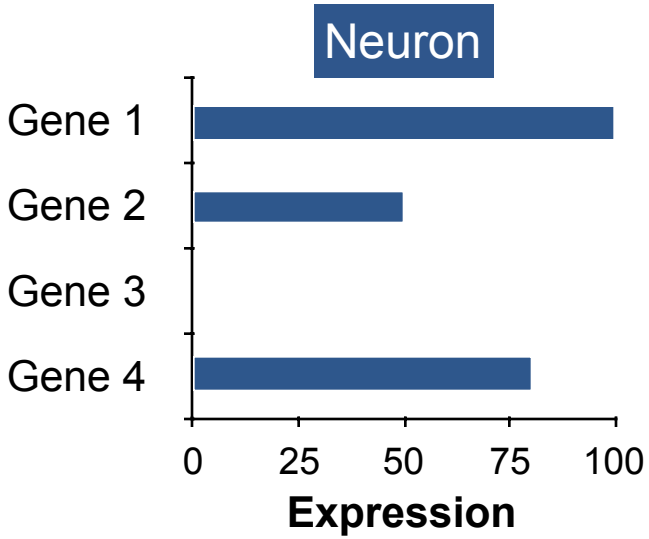
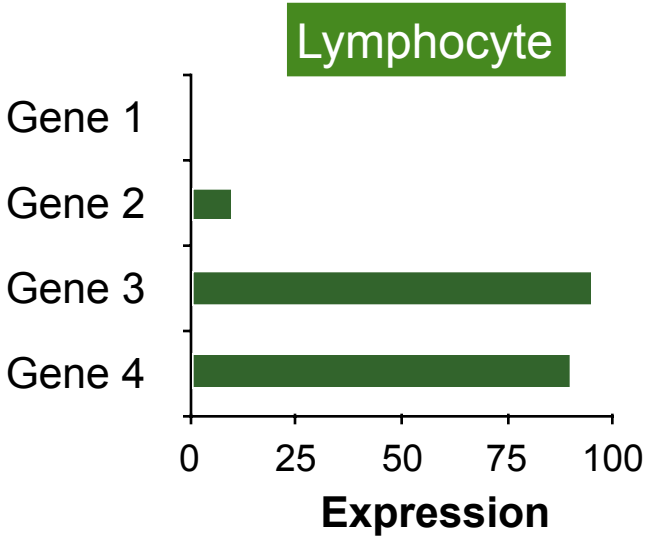


Figure 7-1 Molecular Biology of the Cell 5/e (© Garland Science 2008)





Sequencing

Sequencing

Read the bases of a particular DNA/RNA sequence

Applications:

- sequence DNA of known and unknown organism
- detect variants on patients
- sequence the RNA of a cell
- detect location of proteins interacting with DNA

Problem:

- only short DNA sequences (<1.000 bs) can be read

Solution:

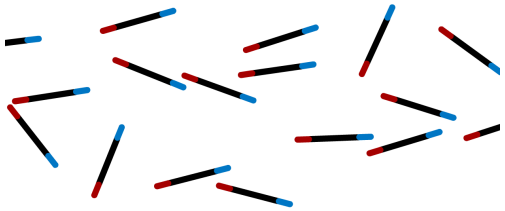
break DNA in several small pieces and use **bioinformatics**

Next Generation Sequencing

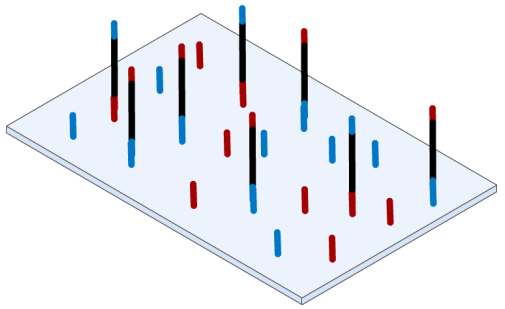
- NGS take advantage of parallelization
  - reads millions/billions of reads for a time
  - short reads (50-100 bps)
  - moderate error rates (0.1%)
- commercial products:
  - 454
  - SOLiD
  - Solexa (Illumina)

Illumina Flow Cell - NGS Sequencing

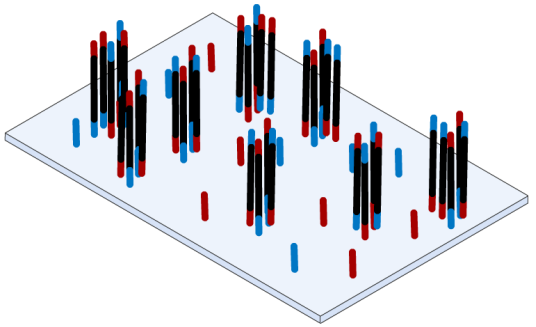
1- fragment sample DNA, insert adapters, attach to flow cell



2- use (bridge) PCR to copy fragments (close to origin)



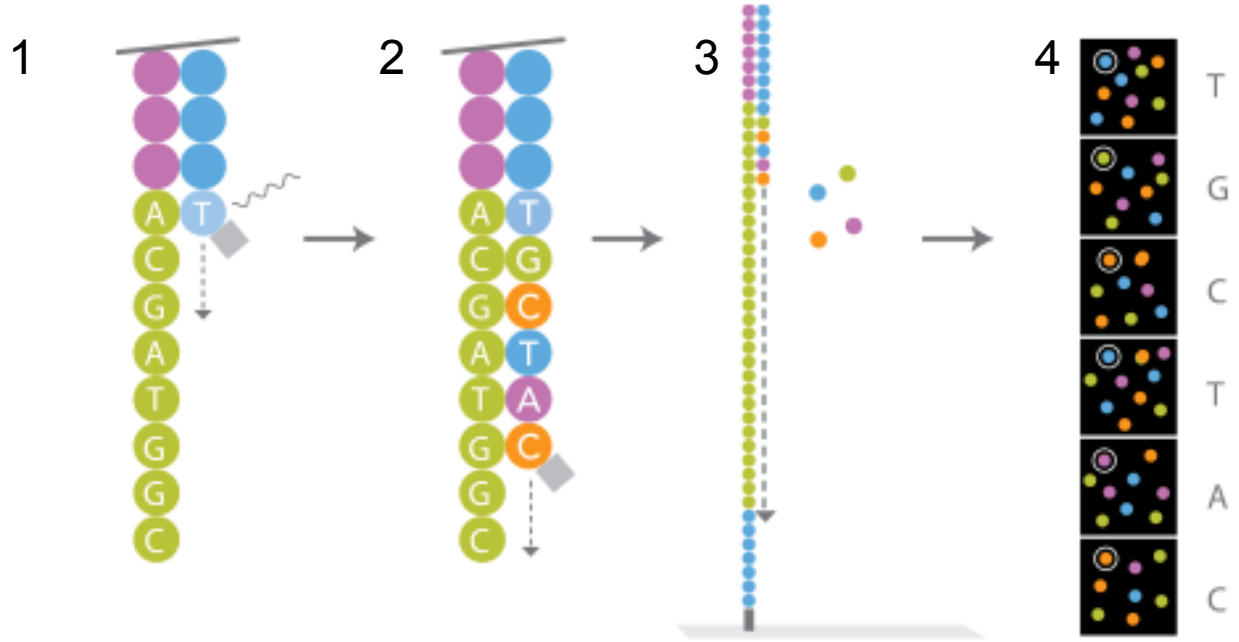
3- clusters of single stranded DNA (200m clusters with 2k DNA strands



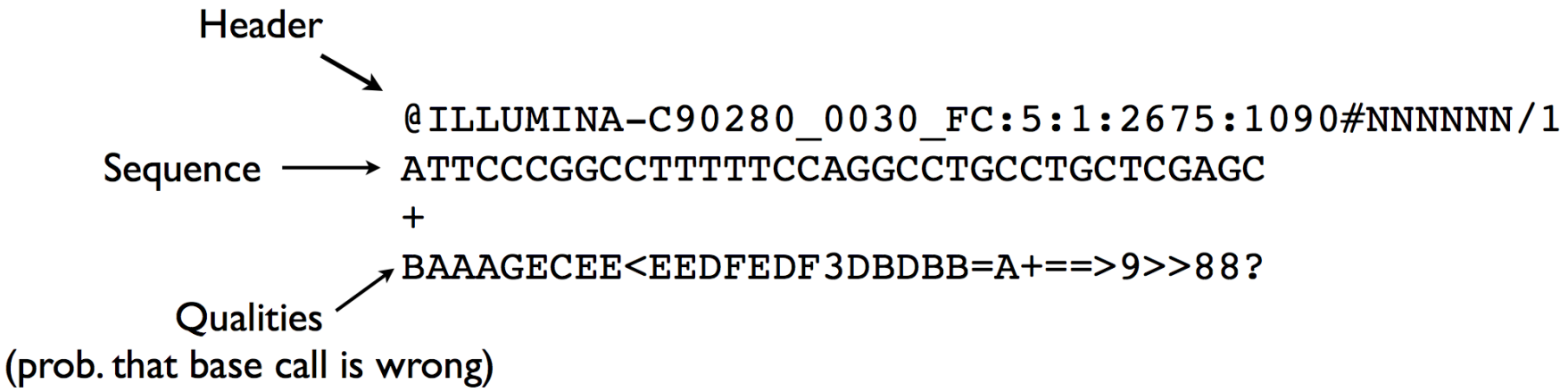
See video http://www.wellcome.ac.uk/Education-resources/Education-and-learning/Resources/Animation/WTX056051.htm

Illumina Flow Cell - NGS Sequencing

- **Iterative evaluation process:**
  1. **add RT-bases, polymerases integrate them**
  2. **wash away all not integrated elements**
  3. **take picture of flow cell to determine current base by dye**
  4. **derive reads from pictures**



Sequencing Results



One character encodes a number using ascii table (0-255)

This number (*Q*) can be converted to *P*

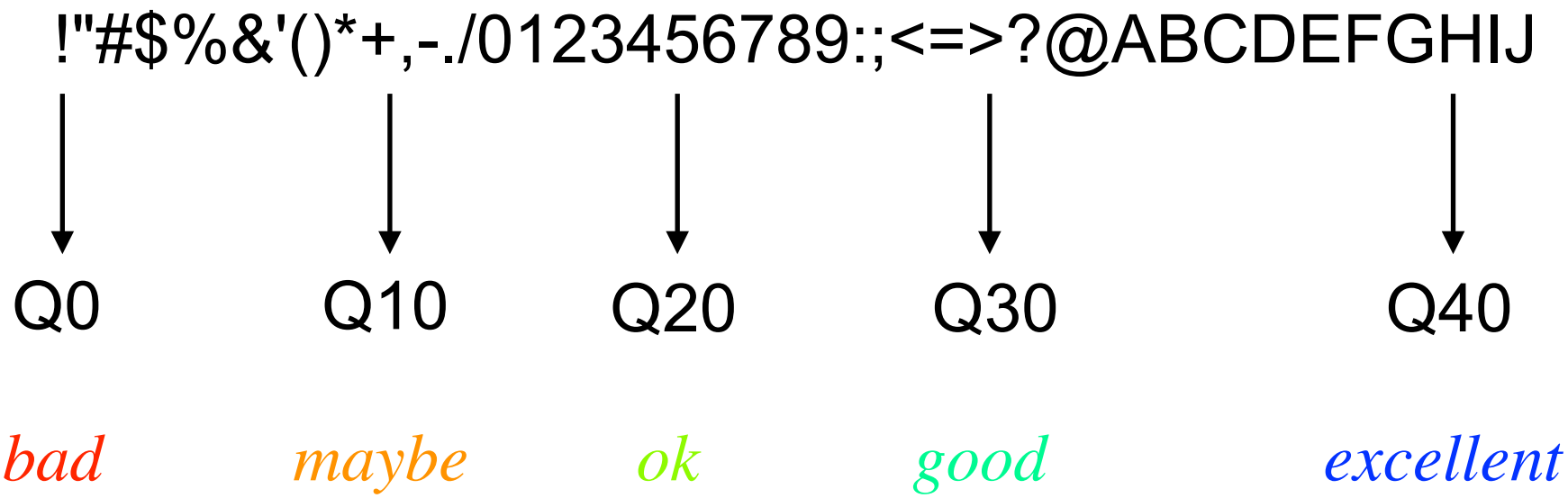
Phred-scale

Q = -10 * log10 P

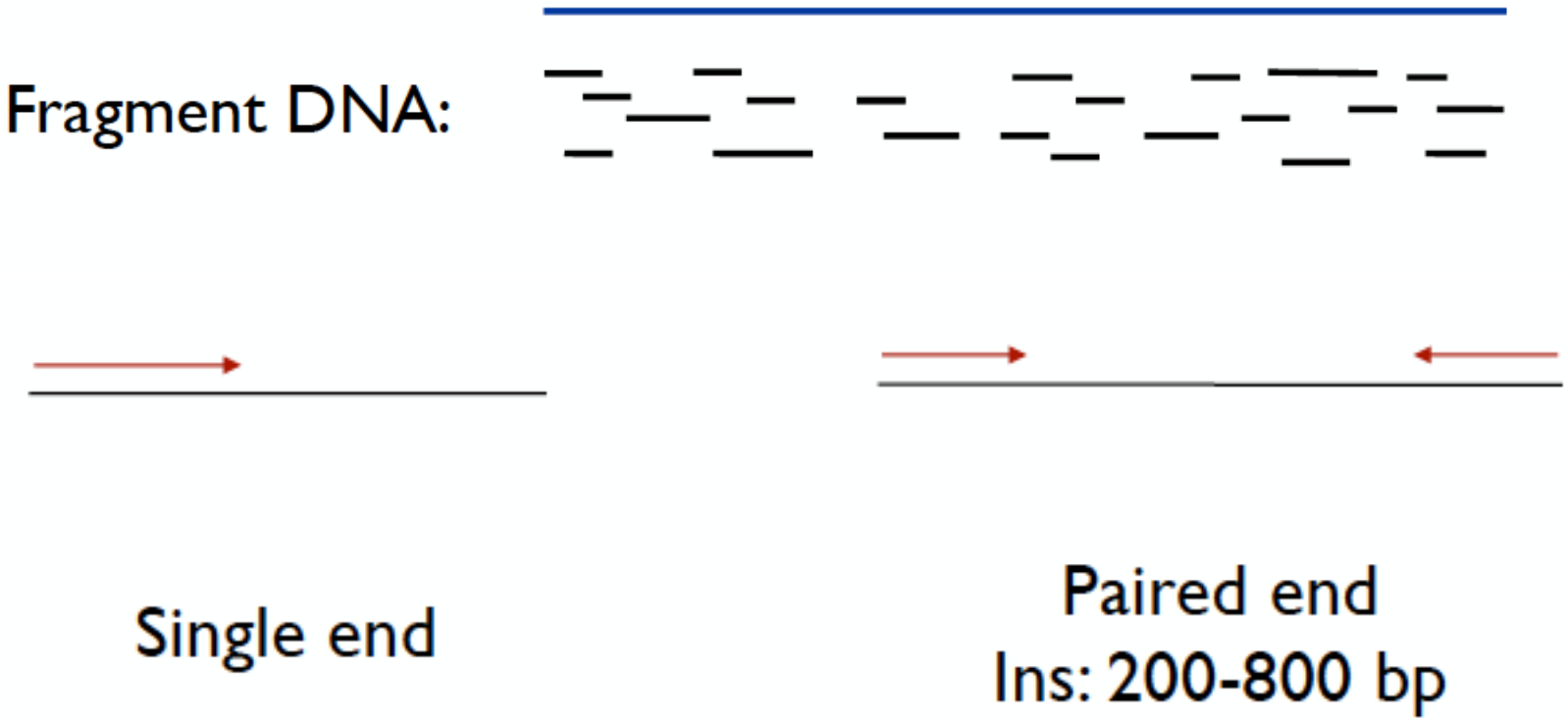
P = 10^(-Q/10)

Sequencing Results / Phred scores

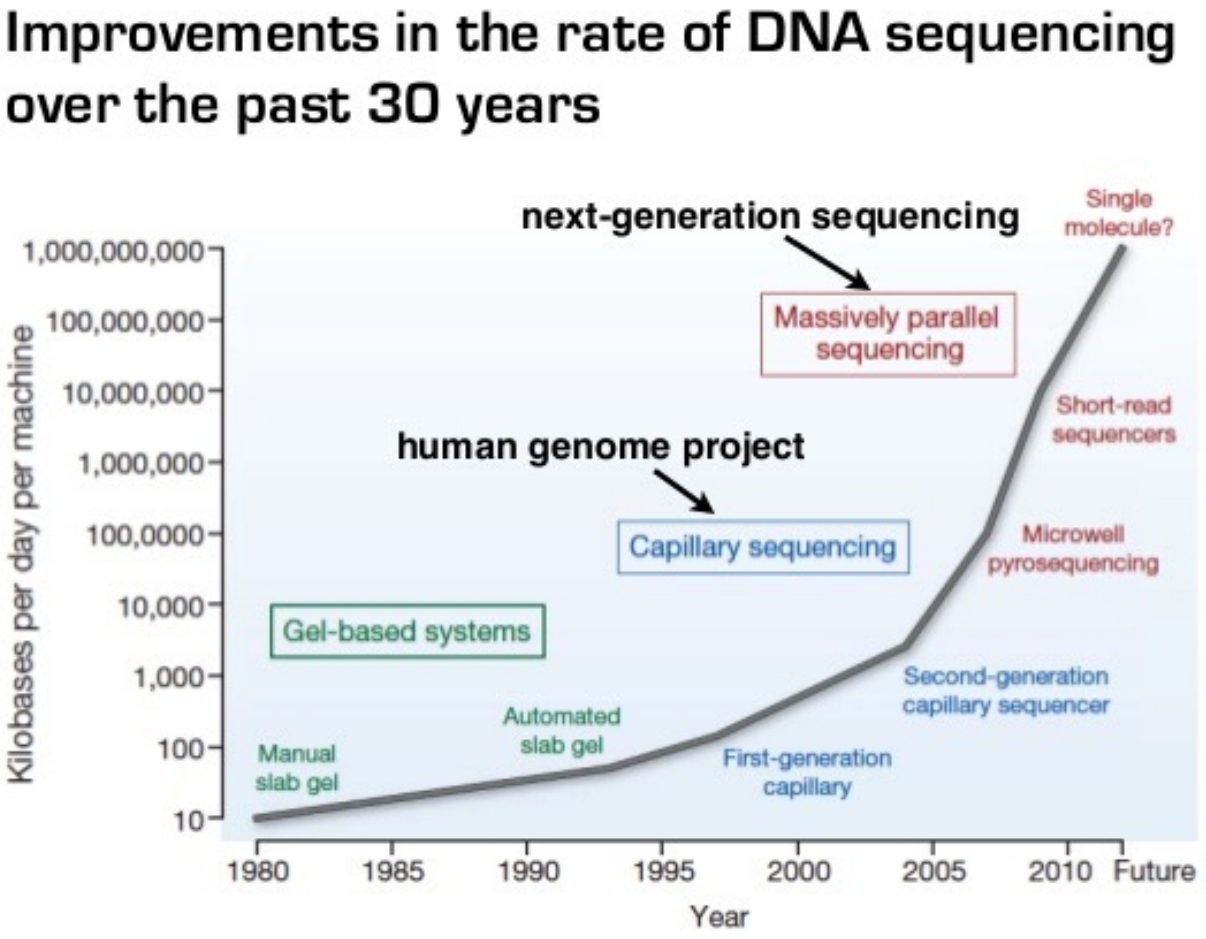
Uses letters/symbols to represent numbers:



Read Types

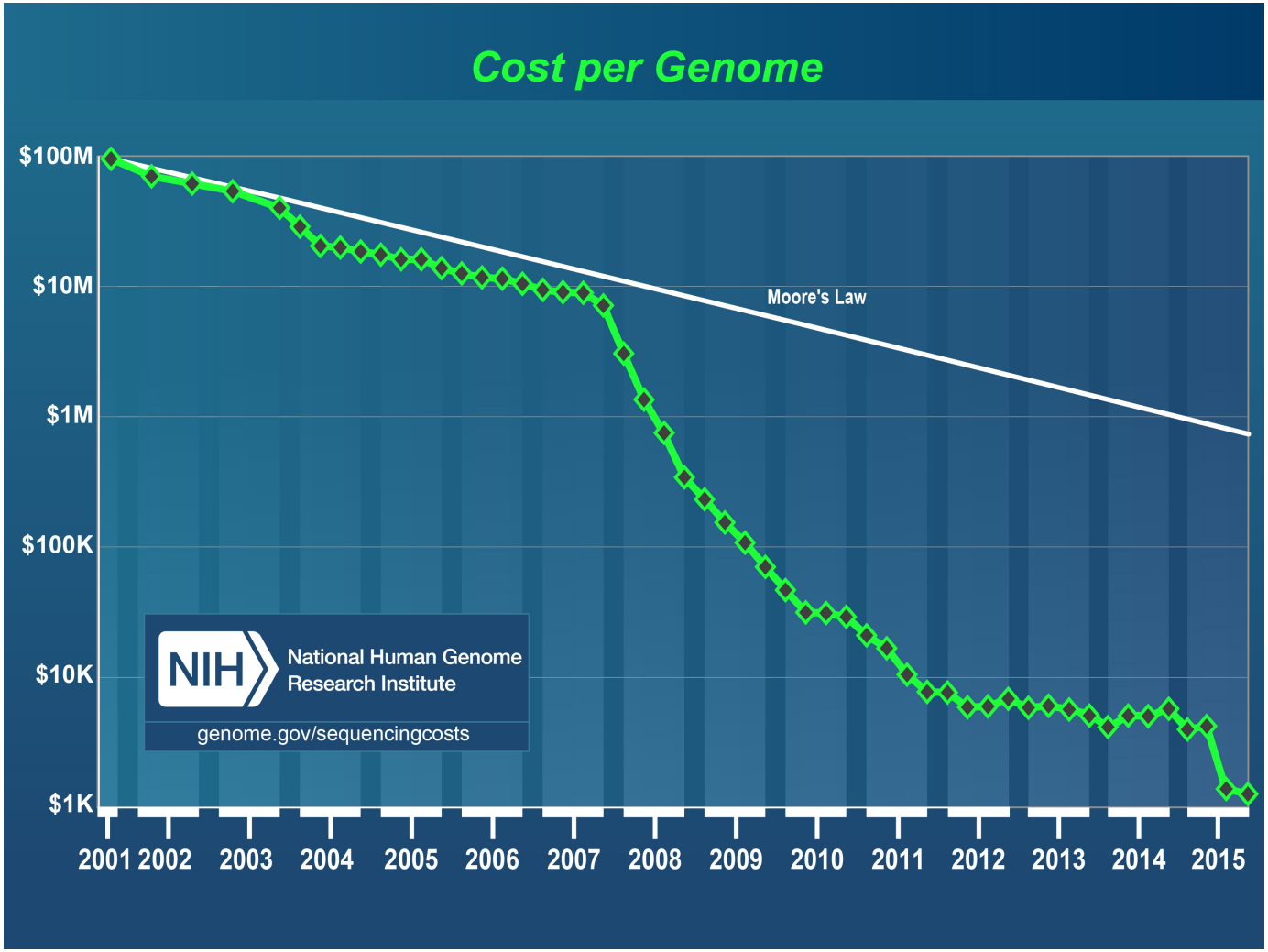


Next Generation Sequencing



Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

Sequencing Costs



Sequence Alignment

Sequence Alignment

NGS

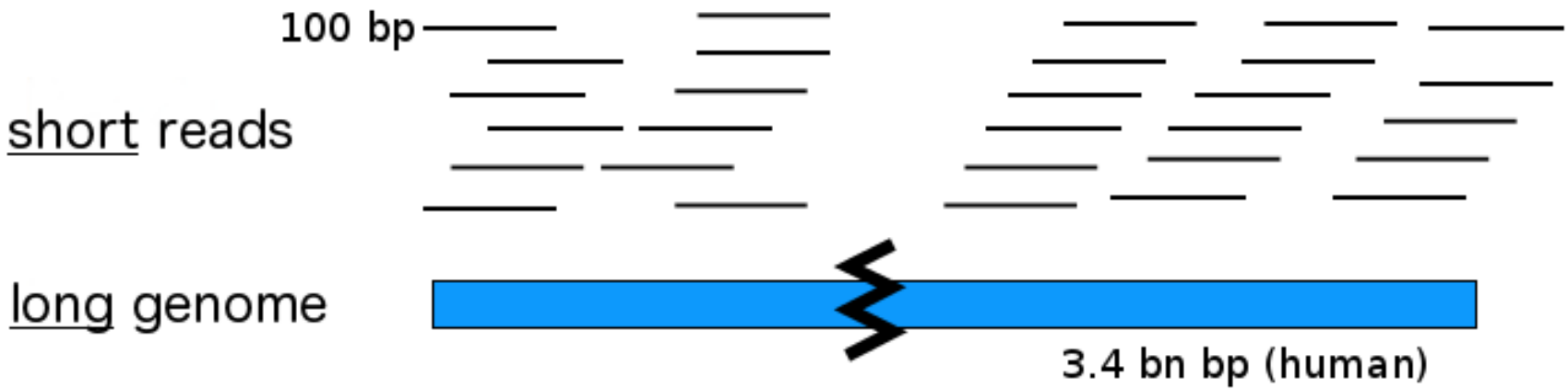
- reads from DNA fragments
- position in genome is unknown
- solution: alignment

DNA Sequencing

- de-novo assembly
  - construct unknown reference sequence from scratch
- resequencing / mapping
  - reference sequence given (applies to human- and mouse-studies)
  - build sequence that is similar but not necessarily identical to reference sequence

Alignment Problem

- a large reference sequence is given (genome)
  - up to billions of base pairs
- millions of short reads (<200bps)
- find most probable position of the read in the genome (by inexact string matching)



Pitfals

- (Unknown) divergent of sample and reference genome
- Repeats in the genome (larger than read size)
- Recombinations
- Poor genome reference quality
- Sequencing/read errors

Algorithms - Alignment

Alignment/Mapping is a typical inexact string match problem

Algorithmic Solutions: ?

Algorithms - Alignment

Alignment/Mapping is a typical inexact string match problem

Algorithmic Solutions:

- **Smith & Waterman - dynamic programming (quadratic time/memory)**

Algorithms - Alignment

Alignment/Mapping is a typical inexact string match problem

Algorithmic Solutions:

- **Smith & Waterman - dynamic programming (quadratic time/memory)**
- **Blast - k-mer search for seeding followed by dynamic programming**
  - **large memory requirement**
  - **local alignment**

11th and 13th most cited papers ever!!!

Algorithms - Alignment

Short read alignment is a special problem

- **reference sequence is large and fixed**
- **query sequence (reads) are short and many**

Solution: ?

Algorithms - Alignment

Short read alignment is a special problem

- **reference sequence is large and fixed**
- **query sequence (reads) are short and many**

Solution: ?

1. Use a data structure to represent reference

- **k-mer hash table (>40GB for k=8)**
- **suffix trees (> 4GB)**

Algorithms - Alignment

Short read alignment is a special problem

- **reference sequence is large and fixed**
- **query sequence (reads) are short and many**

Solution: ?

1. Use a data structure to represent reference

- **k-mer hash table (>40GB for k=8)**
- **suffix trees (> 4GB)**

2. Find candidate (k-mer) hits on genome (>100)

Algorithms - Alignment

Short read alignment is a special problem

- **reference sequence is large and fixed**
- **query sequence (reads) are short and many**

Solution: ?

1. Use a data structure to represent reference

- **k-mer hash table (>40GB for k=8)**
- **suffix trees (> 4GB)**

2. Find candidate (k-mer) hits on genome (>100)

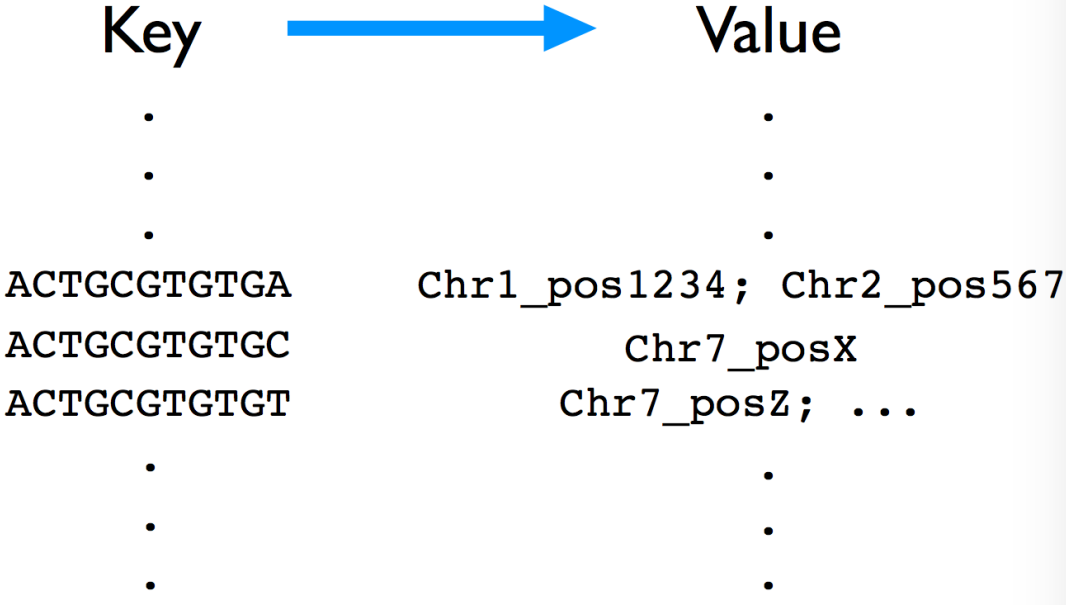
3. Improve alignment with Smith-Waterman

Methods work on linear time (query sequence)

Hash based algorithm

Lookups in hashes are *fast!*

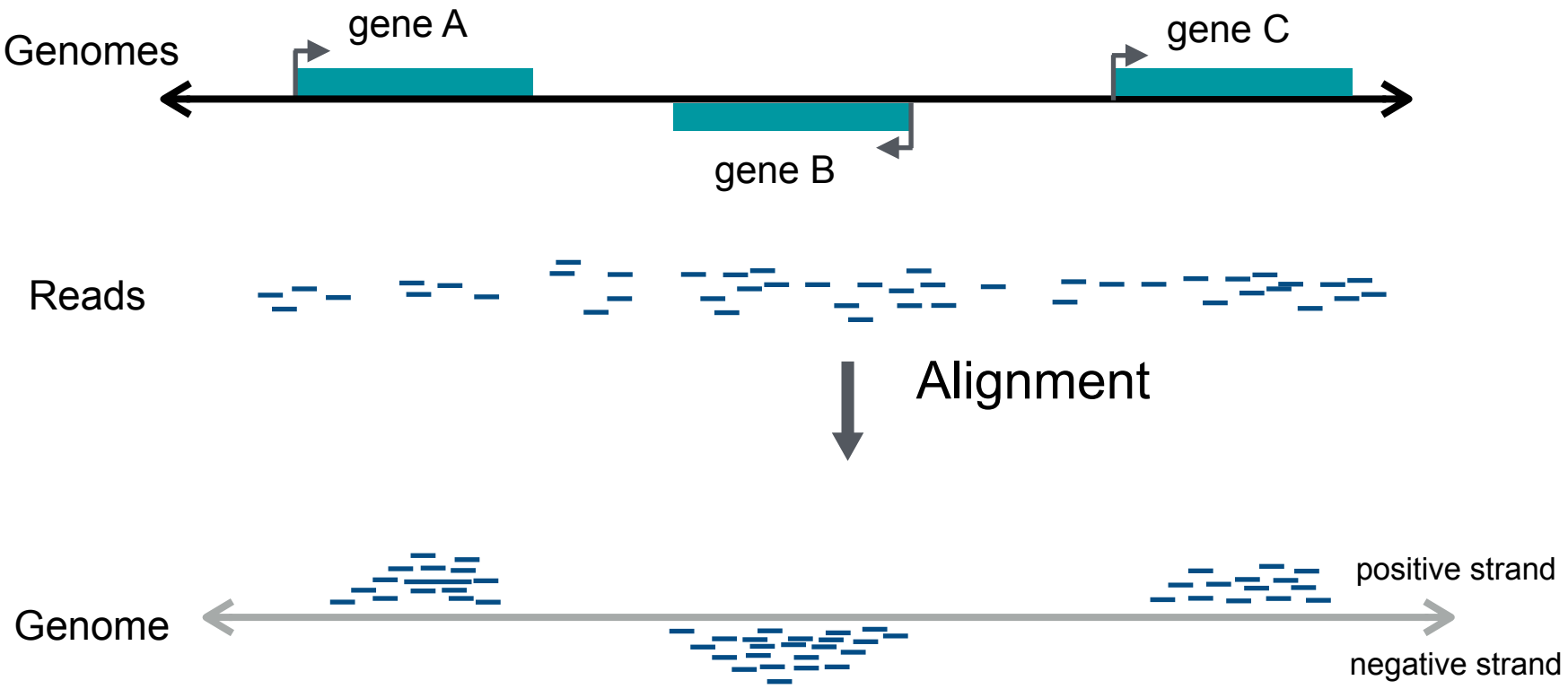
1. Index the reference using *k*-mers.
2. Search reads vs. hash *k*-mers
3. Perform alignment of entire read around seed
4. Report best alignment



Also known as *Seed and extend*

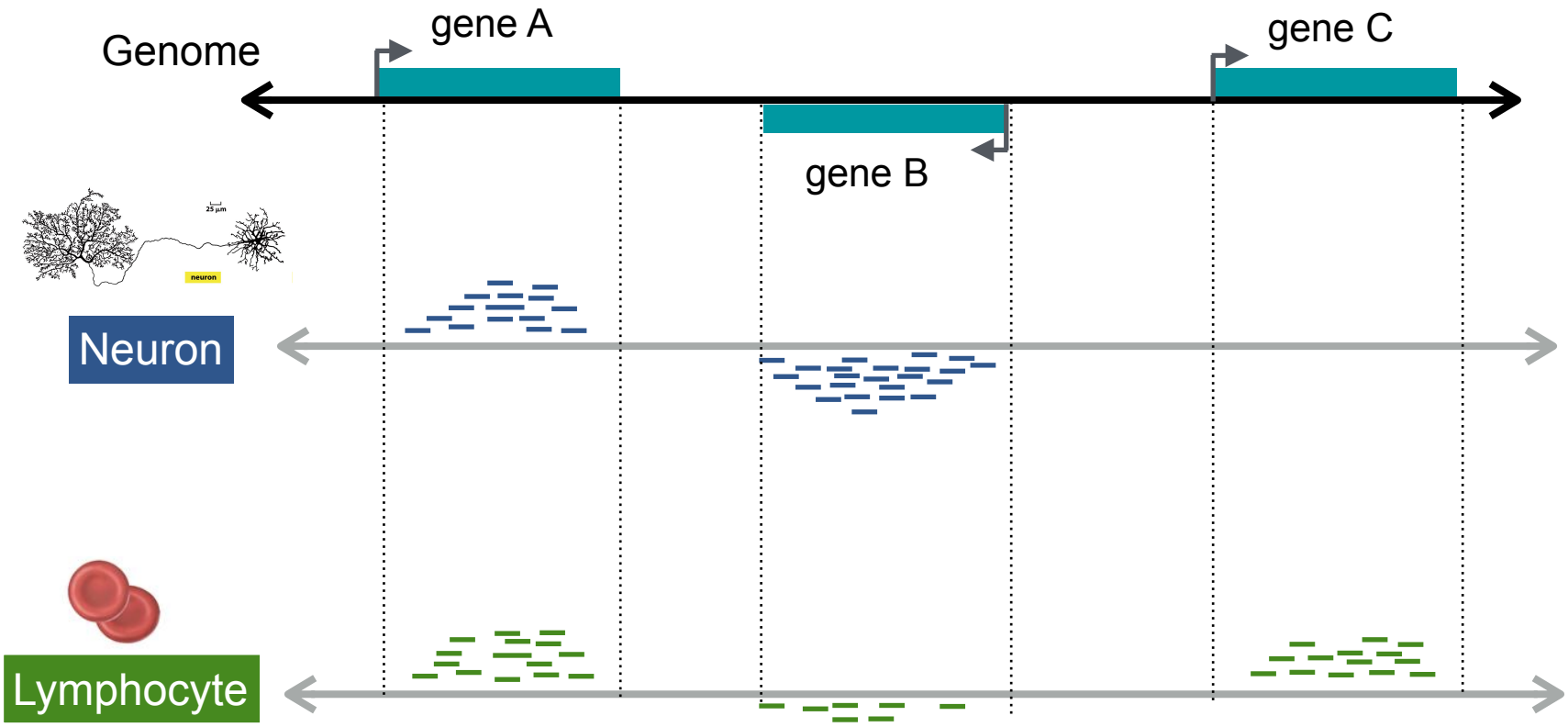
Alignment Results / RNA sequencing

- Position and strand of reads aligned to the genome



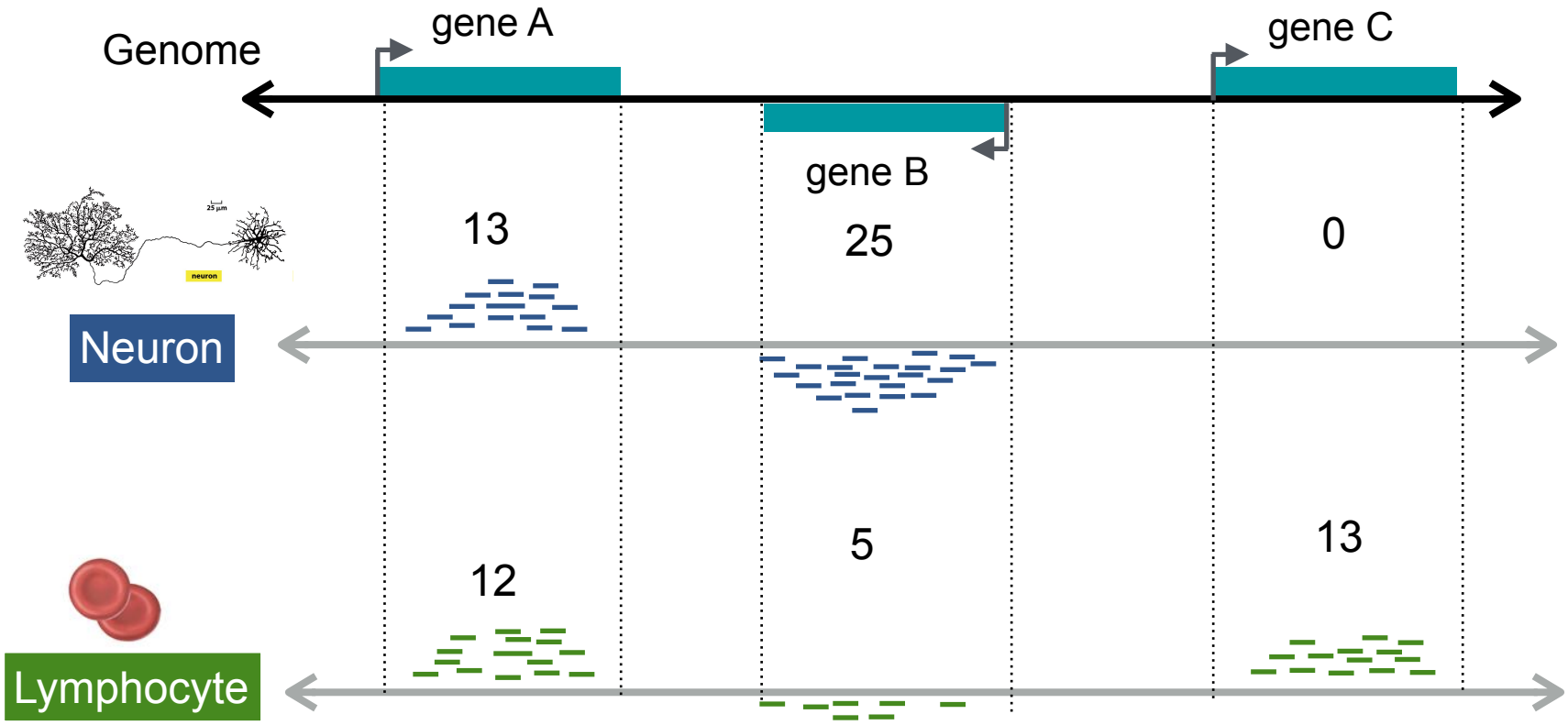
Gene Quantification

- Perform sequencing for each cell (neuron, lymphocyte)
- Align reads to genome

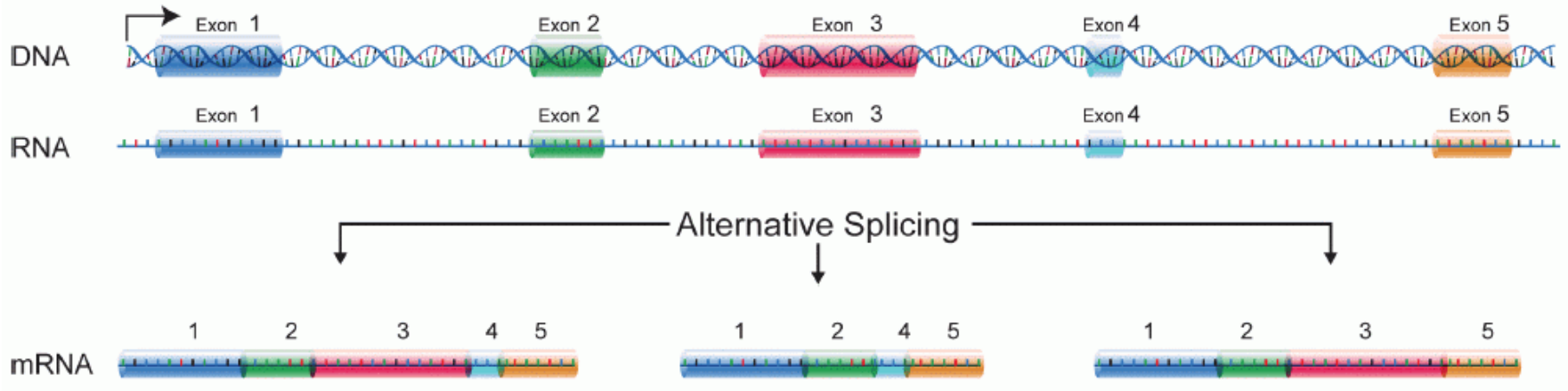


Gene Quantification

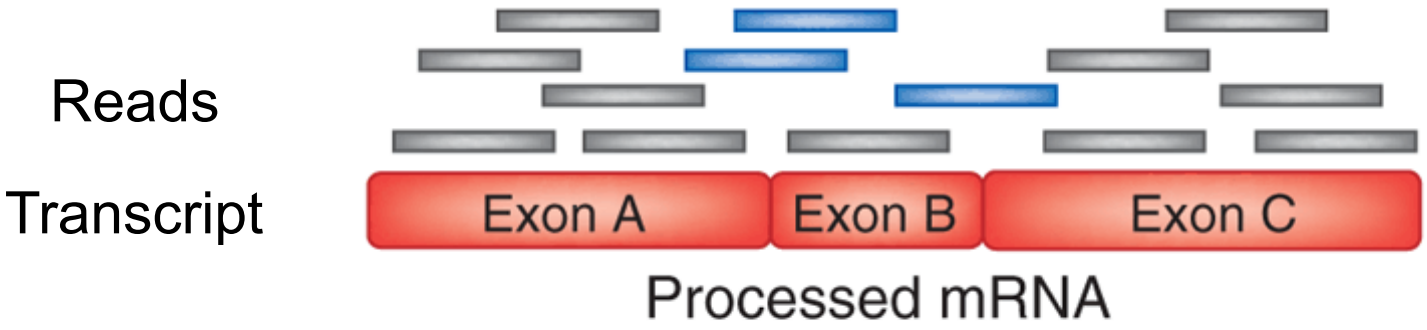
- Perform sequencing for each cell (neuron, lymphocyte)
- Align reads to genome
- Count number of reads inside genes (using known genes annotation)



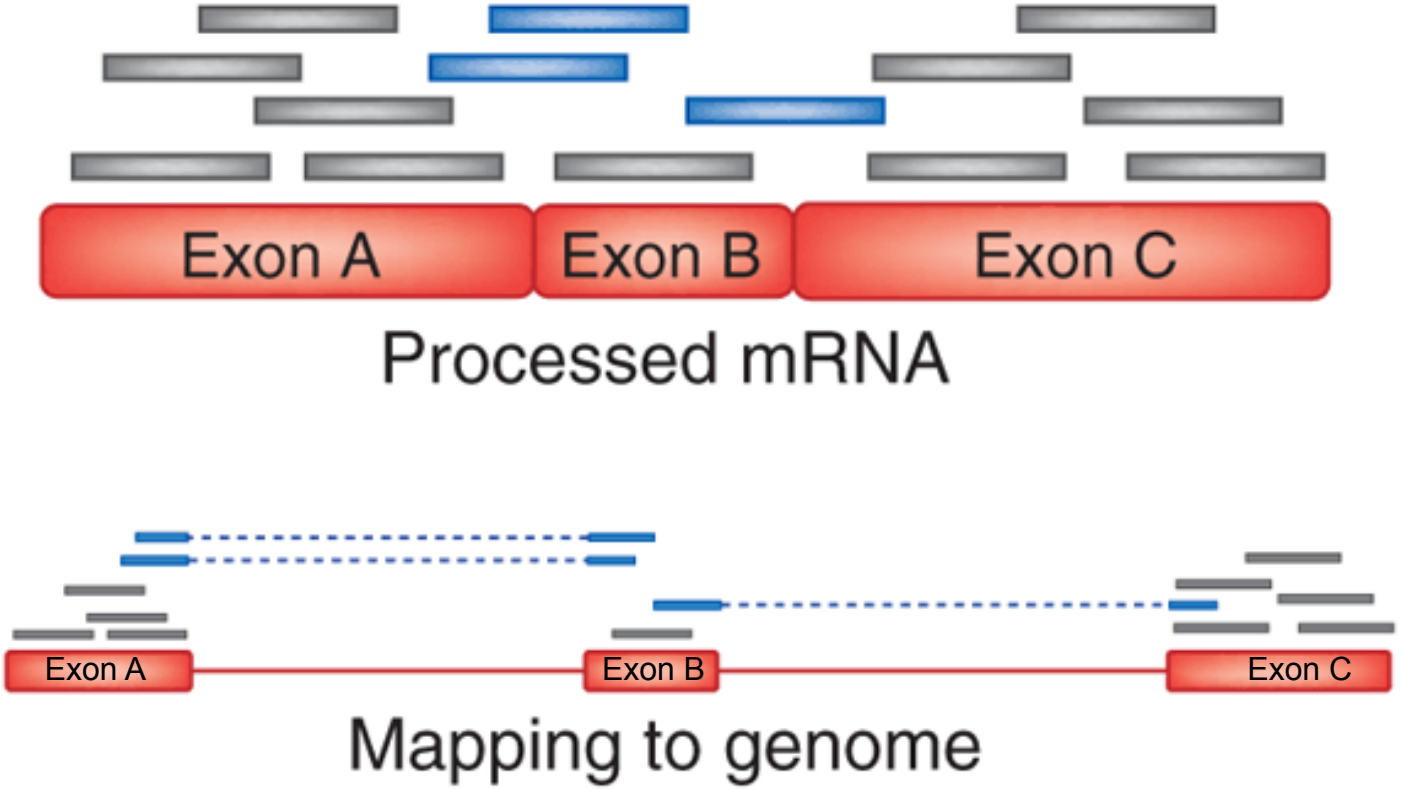
Gene Quantification - Transcripts



Transcripts

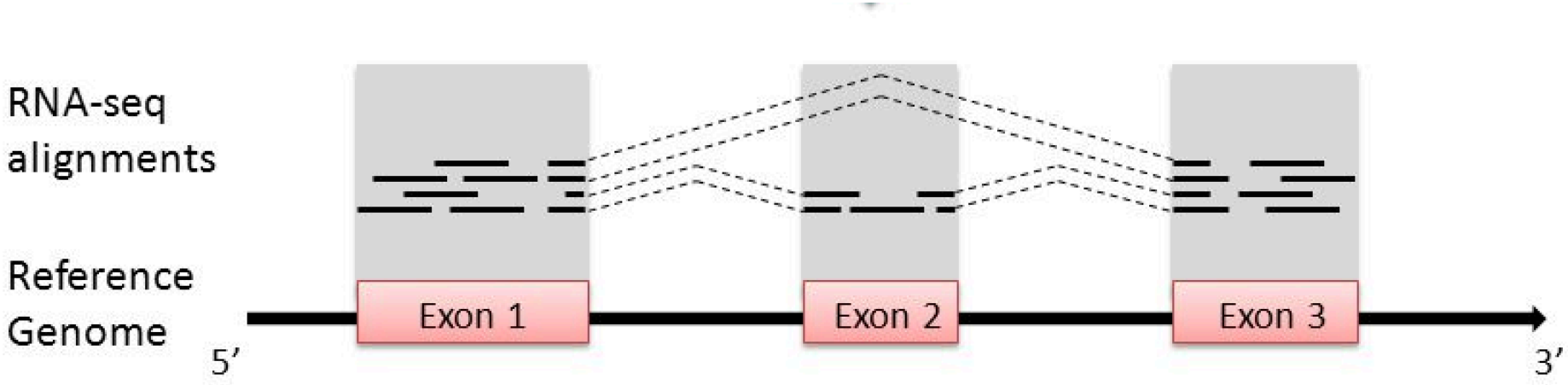


Alignment - Split Read Mapping (RNA-Seq)



- **reads needs to be split within intros when mapped to genome (special aligners / STAR)**

Quantification - Gene vs. Transcript vs. Exon



Counting Strategies

Gene Level - 17 reads
Exon level - exon 1 (8 reads), exon 2 (3 reads), exon 3 (6 reads)
Transcript Level - Exons 1,2 & 3 (10 reads) and exon 1 & 3 (7 reads) *
* complex computational methods required (TopHAT)

Quantificaiton - Normalization

- **Correct for:**
  - Genes having distinct size
  - Sequencing efficiency differs between cell (usually same RNA quantity provided for sequencing)

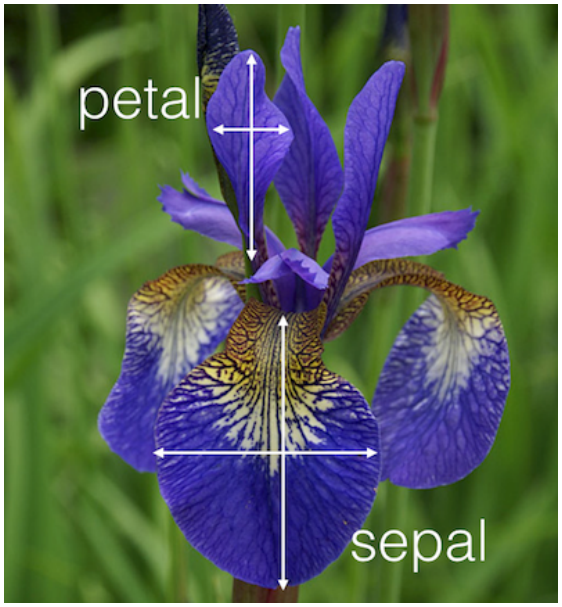
	Cell A	*Cell B*	*…*
GeneA (1kb)	*20*	*15*	*30*
GeneB (2kb)	*100*	*300*	*10*
GeneC (1.5kb)	*10*	*20*	*100*
Gene D (3kb)	*300*	*200*	*100*
Total Library	*430*	*535*	*240*

$$\text{Reads per kilobase million (RPKM)} = \#reads * \frac{gene\ size}{1.000} * \frac{total\ library}{1.000.000}$$

Clustering

Clustering

- **Given a data description**
  - i.e. measurement of size of iris flowers
- **Find groups of similar observations**
  - i.e. iris flower sub-types

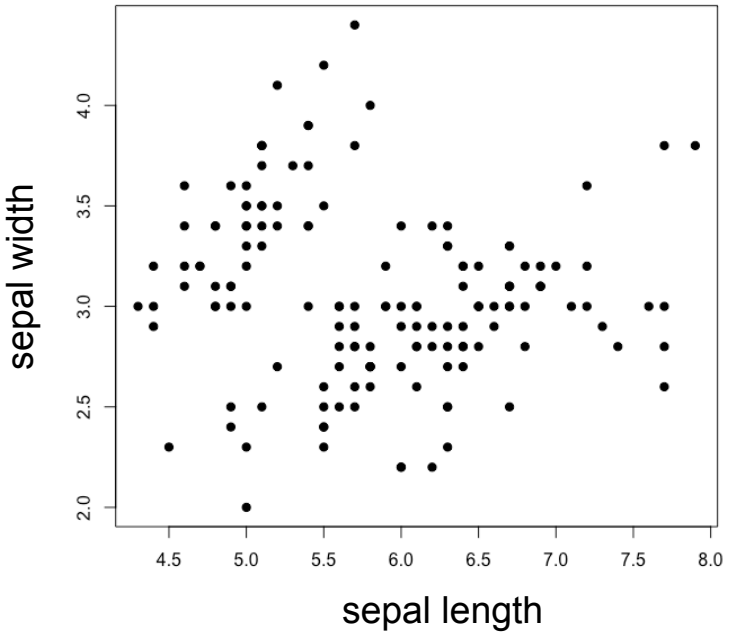


	Sepal Length	Sepal Width	Petal Length	Petal Width
Flower 1	5.1	3.5	1.4	0.2
Flower 2	4.9	3.0	1.4	0.2
Flower 3	4.7	3.2	1.3	0.2
Flower 4	4.6	3.1	1.5	0.2
…	…	…	…	…

Clustering

- **Given a data description**
  - i.e. measurement of size of iris flowers
- **Find groups of similar observations**
  - i.e. iris flower sub-types

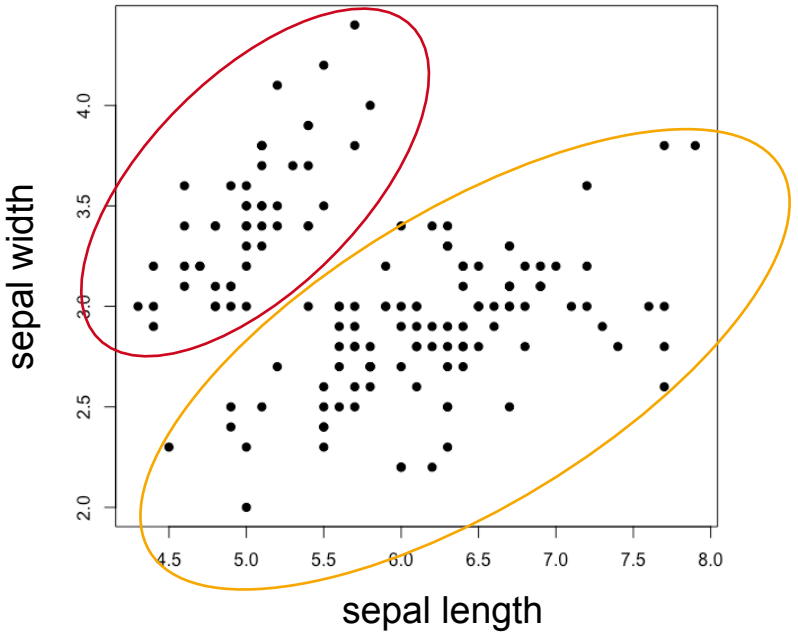
	Sepal Length	Sepal Width	Petal Length	Petal Width
Flower 1	5.1	3.5	1.4	0.2
Flower 2	4.9	3.0	1.4	0.2
Flower 3	4.7	3.2	1.3	0.2
Flower 4	4.6	3.1	1.5	0.2
…	…	…	…	…



Clustering

- **Given a data description**
  - i.e. measurement of size of iris flowers
- **Find groups of similar observations**
  - i.e. iris flower sub-types

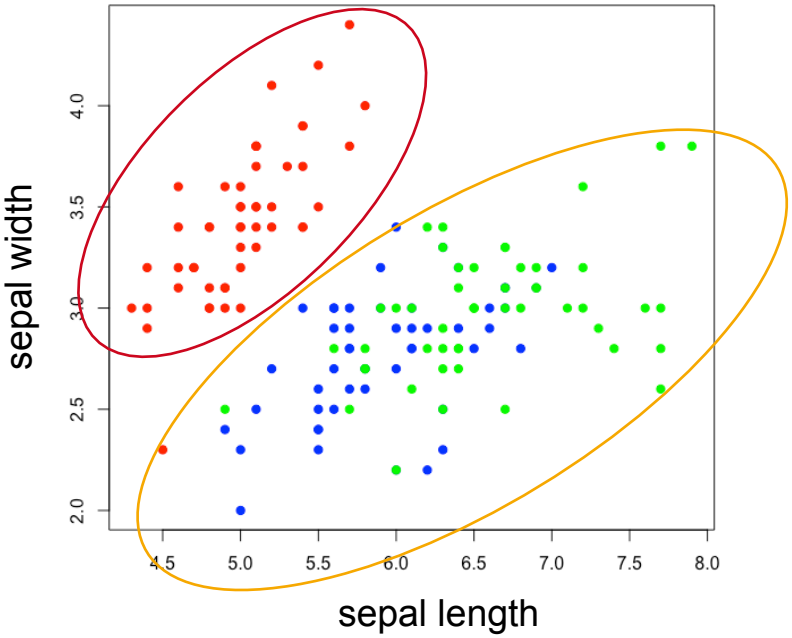
	Sepal Length	Sepal Width	Petal Length	Petal Width
Flower 1	5.1	3.5	1.4	0.2
Flower 2	4.9	3.0	1.4	0.2
Flower 3	4.7	3.2	1.3	0.2
Flower 4	4.6	3.1	1.5	0.2
…	…	…	…	…



Clustering

- **Given a data description**
  - i.e. measurement of size of iris flowers
- **Find groups of similar observations**
  - i.e. iris flower sub-types

Iris Setosa



Iris Virginia

Iris Versicolor

Clustering Formalism

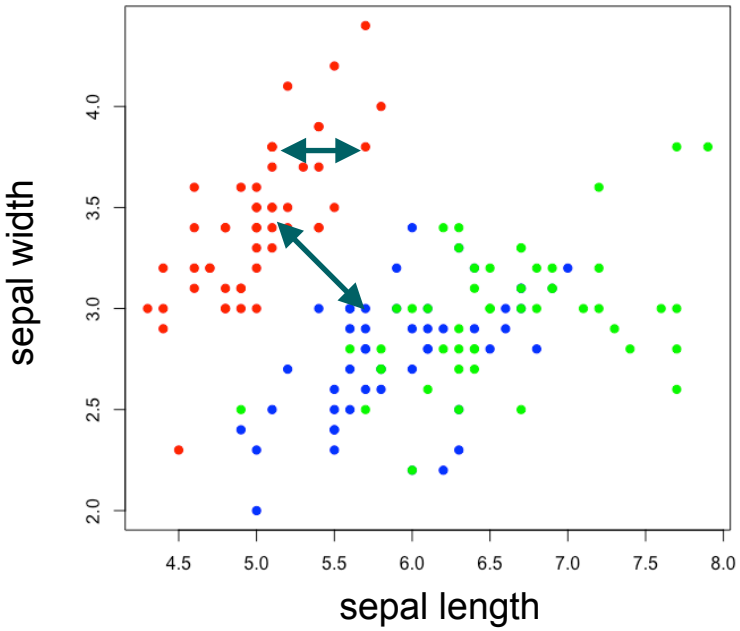
- **For a given data:**
  - Matrix X with N observations and L dimensions
    where x_i is a vector representing observation i

x_{11}	x_{12}	…	x_{1L}
x_{21}	x_{22}	…	x_{2L}
x_{31}	x_{32}	…	x_{3L}
…	…	…	…
x_{N1}	x_{N2}	…	x_{NL}

- **find groups of similar observations**
  - *vector* $Y = (y_1, \ldots, y_N)$
    where $y_i \in \{1, \ldots, K\}$ indicates the cluster of observation i

Distance

- **A important concept in clustering is a distance (similarity) between a pair of objects x_i and x_j**
  - Observations of a same group should be close in space

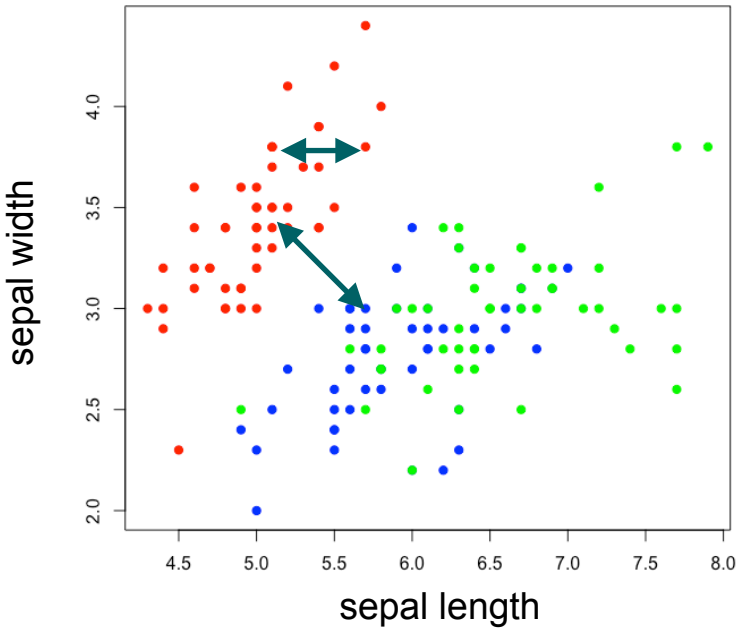


Euclidean distance
(sensitive to scale)

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{L} (x_{il} - x_{jl})^2}$$

Distance

- **A important concept in clustering is a distance (similarity) between a pair of objects x_i and x_j**
  - Observations of a same group should be close in space



Euclidean distance
(sensitive to scale)

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{L} (x_{il} - x_{jl})^2}$$

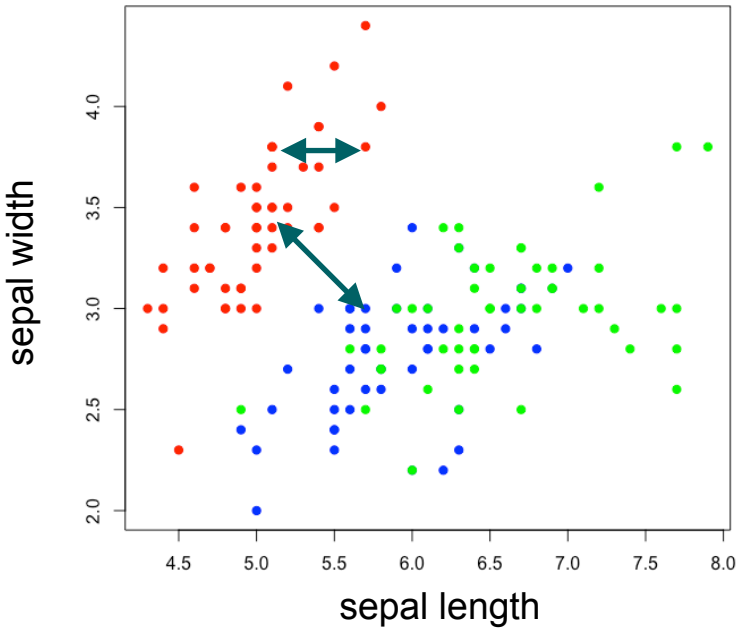
Pearson Correlation

(scale insensitive/ similarity)

$$d(x_i, x_j) = \frac{\sum_{l=1}^{L} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{\sigma_i^2 \sigma_j^2}$$

Distance

- **A important concept in clustering is a distance (similarity) between a pair of objects x_i and x_j**
  - Observations of a same group should be close in space



Euclidean distance
(sensitive to scale)

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{L} (x_{il} - x_{jl})^2}$$

Pearson Correlation

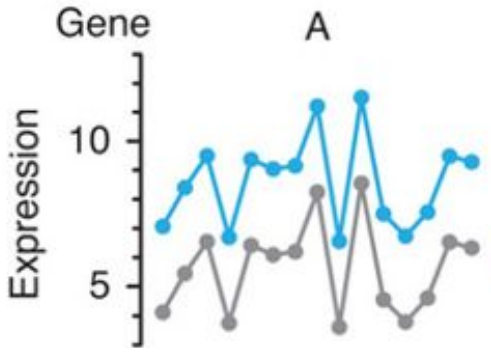
(scale insensitive/ similarity)

$$d(x_i, x_j) = \frac{\sum_{l=1}^{L} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{\sigma_i^2 \sigma_j^2}$$

Distance and Scale

- **In some problems scale can be important!**
  - Similarly in changes are more important / not absolute values.

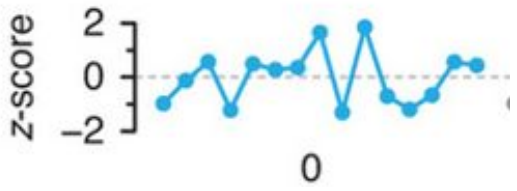
unscaled data



Euclidean - not similar

Correlation - similar

z-score normalised data



$$z = \frac{x_{ij} - \mu_i}{\sigma_i}$$

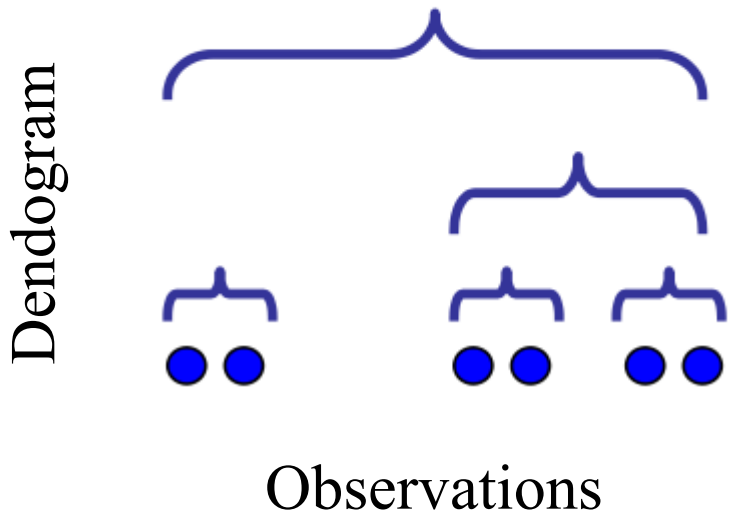
Euclidean - similar

Correlation - similar

Clustering Methods

- **Hierarchical methods**
  - Mostly bottom up
  - based on distance / simple to interpret
- **Partitional methods (k-means or mixture models)**
  - Mostly top down
  - Use models of groups, centroids
- **Graph based methods**
  - Use graph formalisms to represent data:
    - nodes are representations
    - edges weights represent distances
    - Explore graph topologies

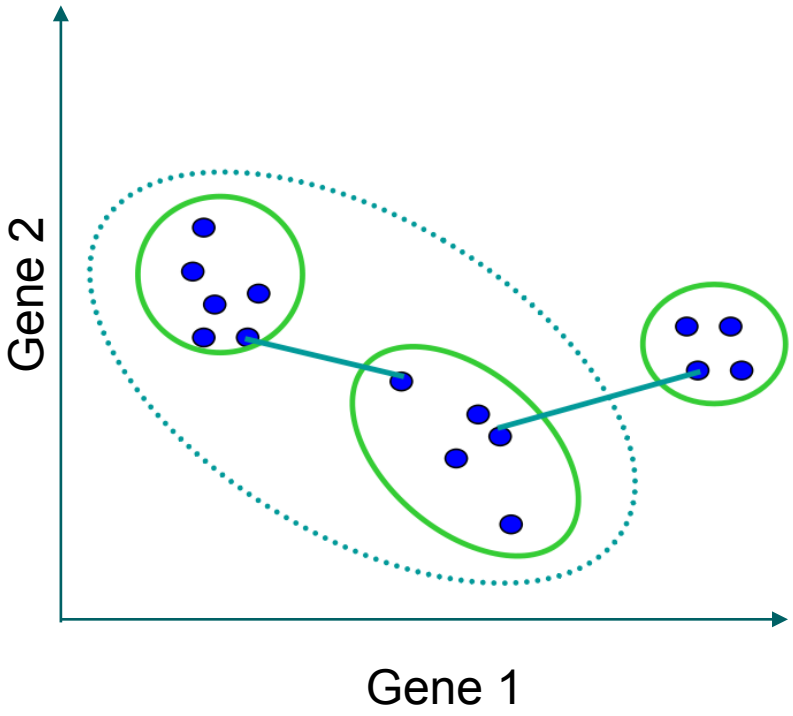
Hierarchical Clustering



- Botton up method
- Starting with a distance (similarity) matrix and each object as a group
- Repeat:
  - Joint two most similar groups
- Until the dendrogram has only one group

Source: Slides adapted from Ziv-Bar Joseph

Hierarchical Clustering



Single-Linkage

- Join two groups where two examples are close
- Find groups with linear shapes

Source: Slides adapted from Ziv-Bar Joseph

Hierarchical Clustering

Distance Matrix

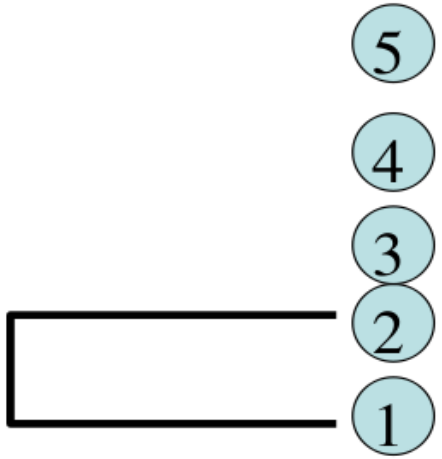
	1	2	3	4	5
1	0				
2	2	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0

Source: Slides adapted from Ziv-Bar Joseph

Hierarchical Clustering

Distance Matrix

	1	2	3	4	5
1	0				
2	2	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0



Source: Slides adapted from Ziv-Bar Joseph

Hierarchical Clustering

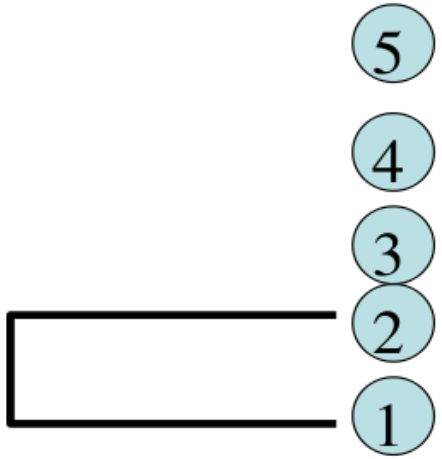
Distance Matrix

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 9 & 7 & 0 & \\ 8 & 5 & 4 & 0 \end{bmatrix} \end{array}$$

$$d_{(1,2),3} = \min\{d_{1,3}, d_{2,3}\} = \min\{6,3\} = 3$$

$$d_{(1,2),4} = \min\{d_{1,4}, d_{2,4}\} = \min\{10,9\} = 9$$

$$d_{(1,2),5} = \min\{d_{1,5}, d_{2,5}\} = \min\{9,8\} = 8$$



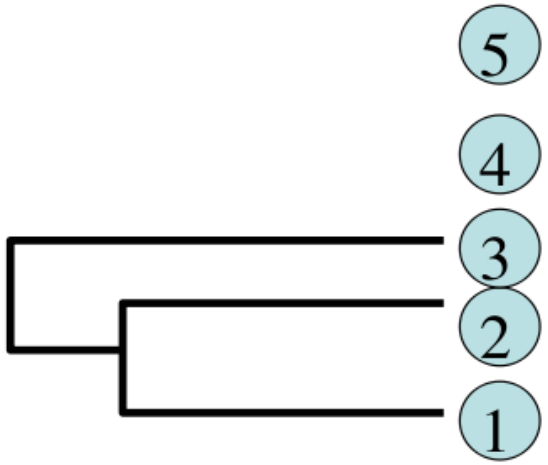
Source: Slides adapted from Ziv-Bar Joseph

Hierarchical Clustering

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2,3) & 4 & 5 \\ (1,2,3) & 0 & & \\ 4 & 7 & 0 & \\ 5 & 5 & 4 & 0 \end{array}$$

$$d_{(1,2,3),4} = \min\{d_{(1,2),4}, d_{3,4}\} = \min\{9,7\} = 7$$

$$d_{(1,2,3),5} = \min\{d_{(1,2),5}, d_{3,5}\} = \min\{8,5\} = 5$$

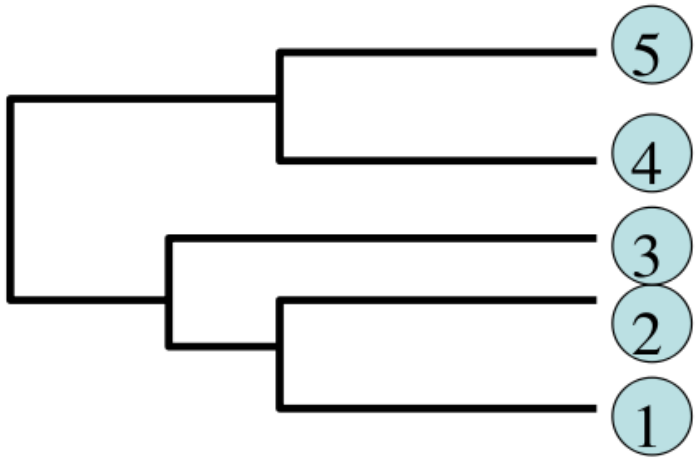


Source: Slides adapted from Ziv-Bar Joseph

Hierarchical Clustering

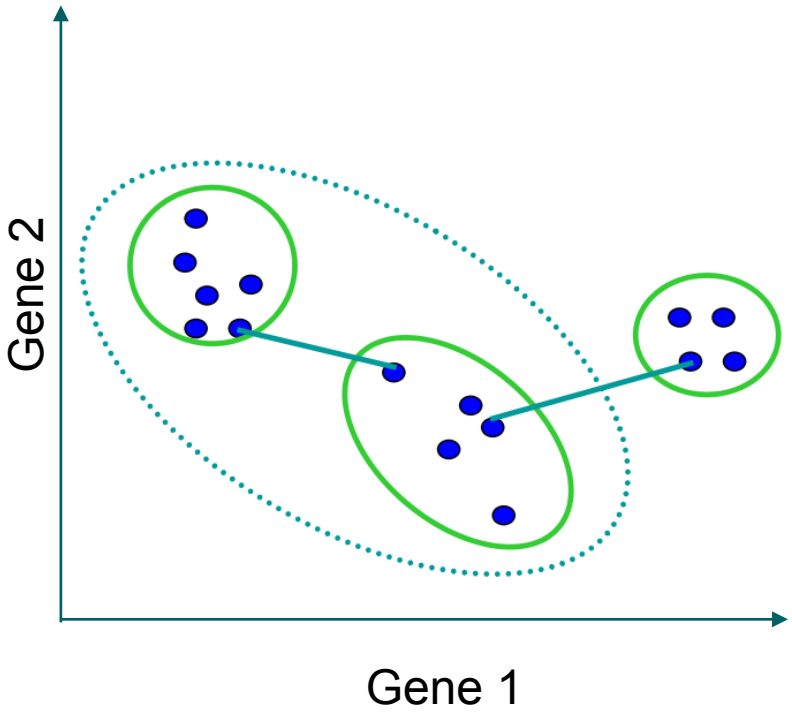
$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2,3) & 4 & 5 \\ (1,2,3) & 0 & & \\ 4 & 7 & 0 & \\ 5 & 5 & 4 & 0 \end{array}$$

$$d_{(1,2,3),(4,5)} = \min\{d_{(1,2,3),4}, d_{(1,2,3),5}\} = 5$$



Source: Slides adapted from Ziv-Bar Joseph

Hierarchical Clustering



Single-Linkage

- Groups with closest genes
- linear shapes

Complete-Linkage

- Closest groups with more far genes
- Compact clusters

Average Linkage

- Groups with closest centroids (middle)
- Outlier robust

Source: Slides adapted from Ziv-Bar Joseph

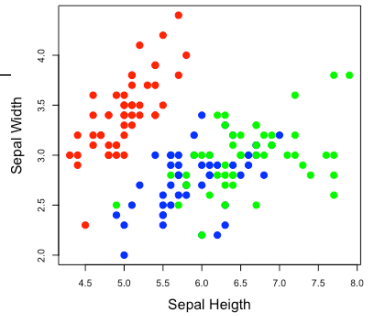
Hierarchical Clustering

Which linkage?

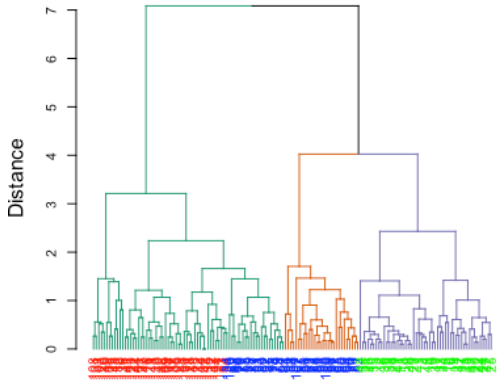
Which distance?

Hierarchical Clustering of Iris

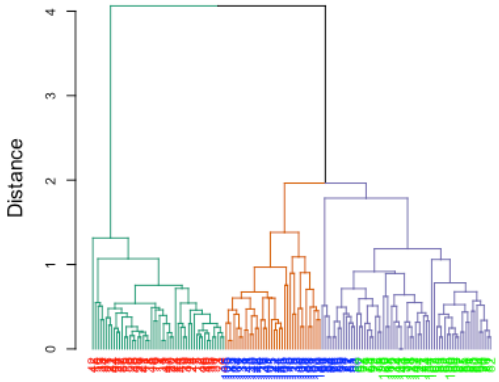
True labels



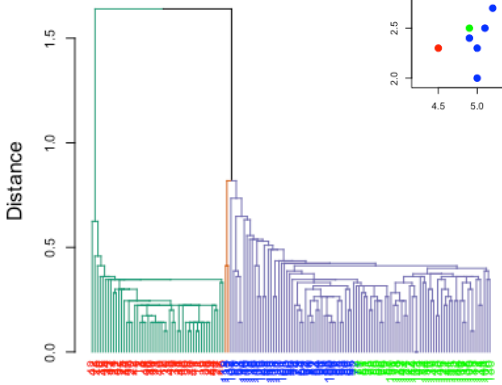
Euclidean distance



Complete



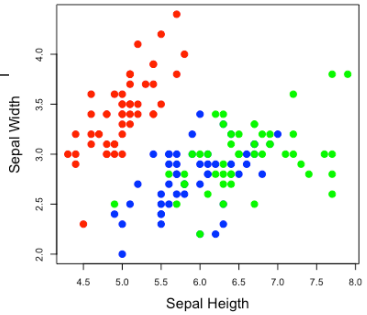
Average



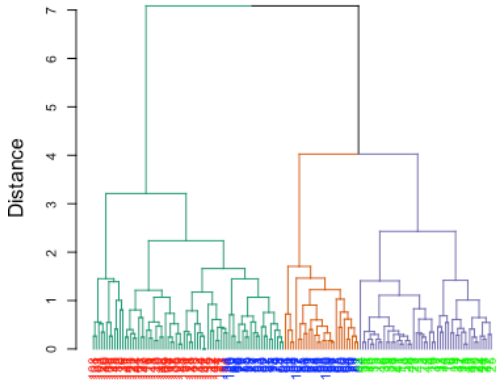
Single

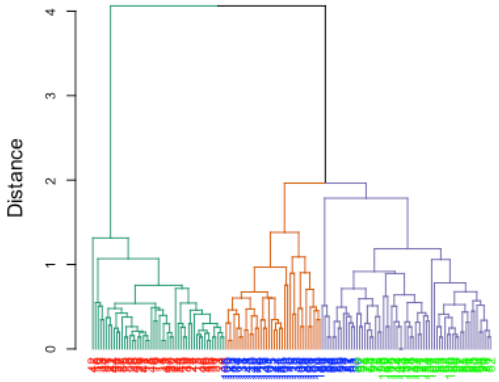
Hierarchical Clustering of Iris

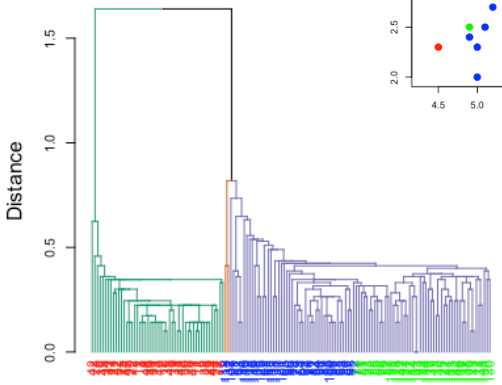
True labels



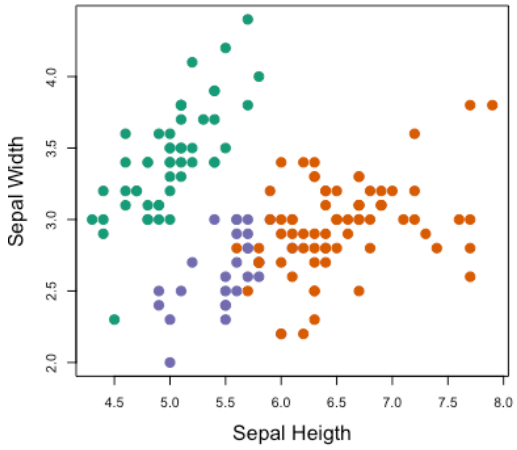
Euclidean distance



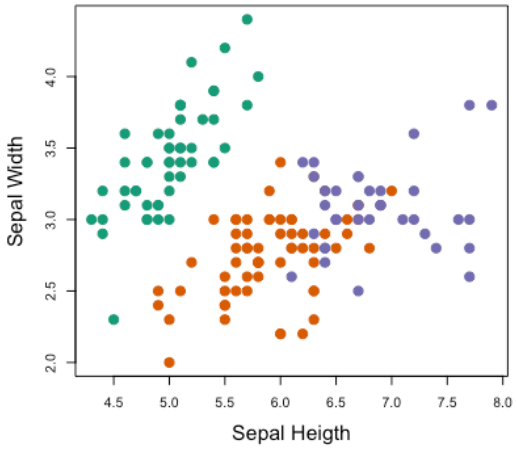




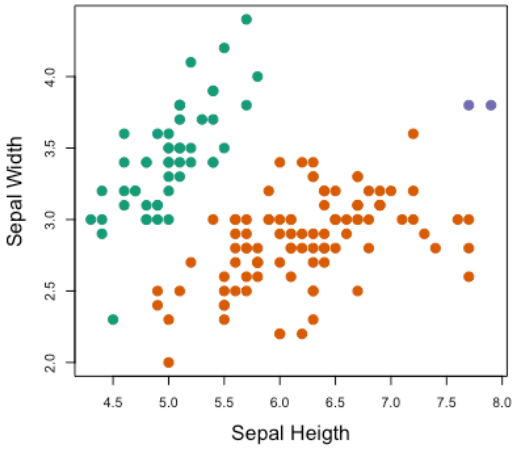
Complete



Average



Single

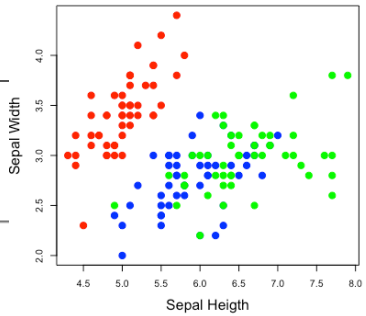


Institute for Computational Genomics

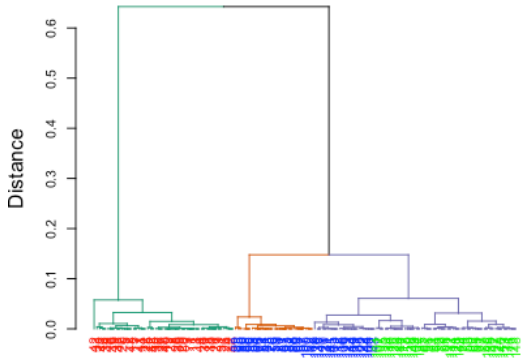
RWTH AACHEN UNIVERSITY

Hierarchical Clustering of Iris

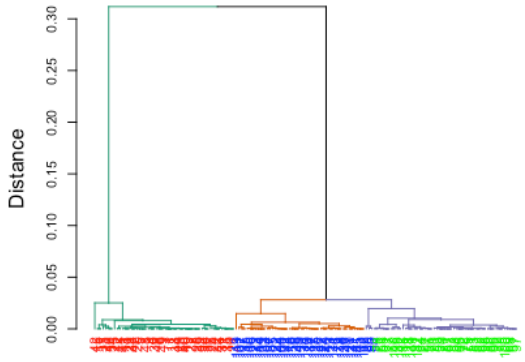
True labels



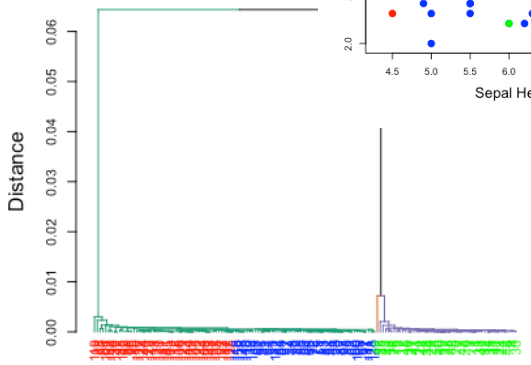
Pearson distance



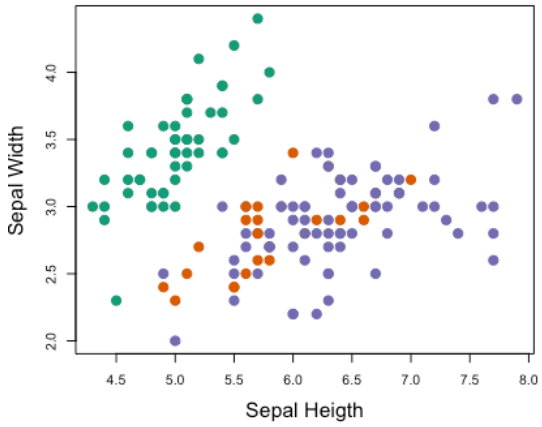
Complete

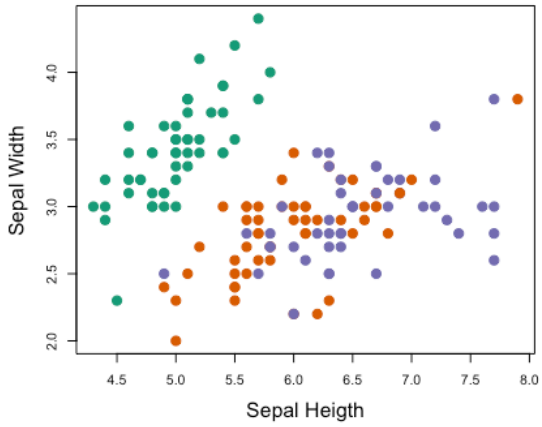


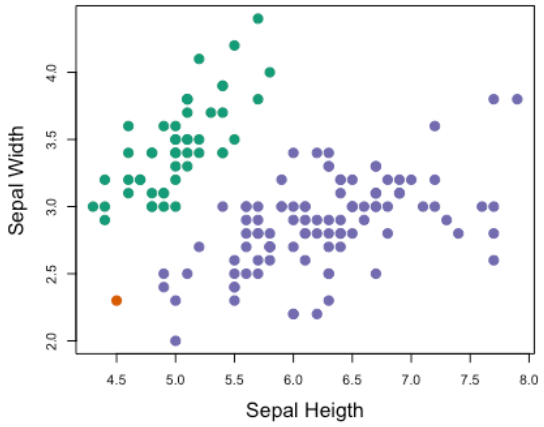
Average



Single







- Hierarchical cluster is sensitive to noise/outliers
- High computational cost O(n^2)

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

K-means

Iterative algorithm using **centroids** as cluster representations

Requires specification of number of clusters (***K***)

Algorithm:

Start cluster (Y) randomly

Repeat for a number of iterations

- estimate centroid (m_k) for each cluster

$$m_k = \frac{\sum_{i=1}^{N} 1(y_i = k) x_i}{\sum_{i=1}^{N} 1(y_i = k)}$$

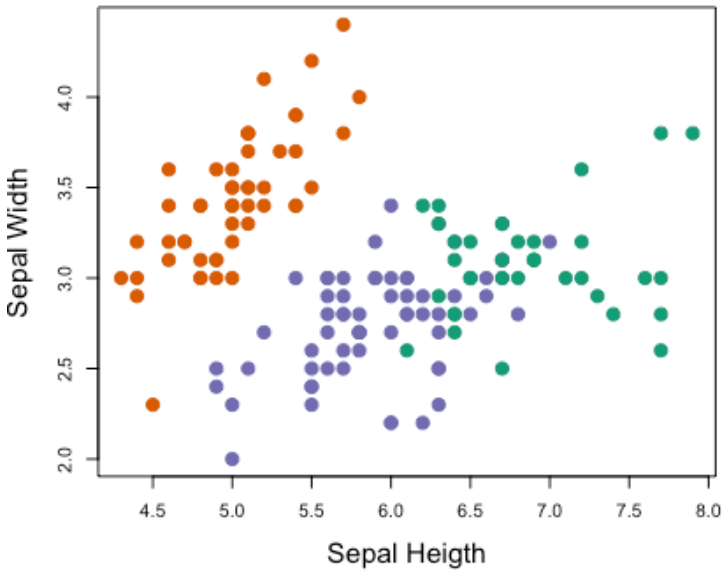
- Assign objects to closest centroid:

$$y_i = \text{argmin}_k \, d(x_i, m_k)$$

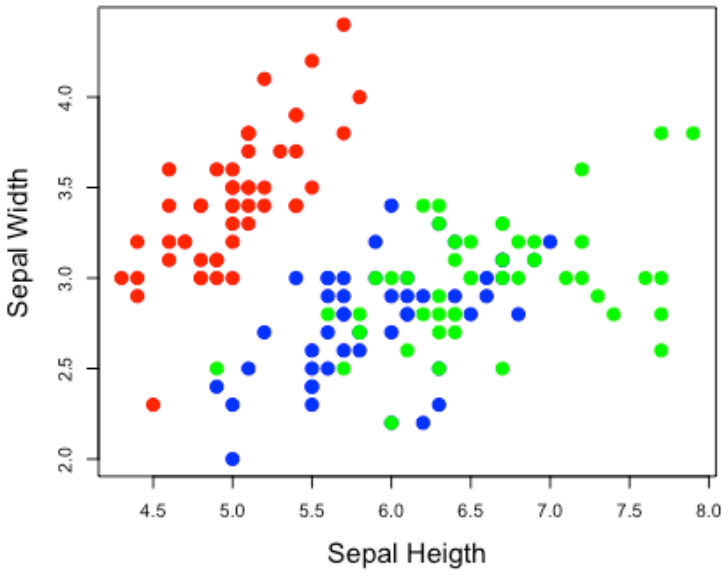
* convergence is only guaranteed for Euclidean distance

K-means on Iris

K-means solutions



True labels



- K-means tends to find spherical clusters
- Sensitive to initialisation

Resume / Clustering Methods

- K-means and hierarchical clustering

  - standard algorithms with standard performance on simple clustering problems

- State-of-art methods explore characteristics of the data (images, genomic data, text) at hand as type of features, dimensionality)

- Further issues:
  - Validation:
    - How many clusters is present in the data?
    - Which is the best method?
  - Data dimensionality:
    - distances do not work well on high dimension
    - visualisation is easier in low level space

More details on clustering

- Hastie, Tibshirani and Friedman, The Elements of Statistical Learning, Chapter 14
- Bishop, Pattern Recognition and Machine Learning, Chapter 9

Cluster Validation

- How to evaluate clustering results? Which is the best method? How many clusters?
- Internal/relative validation:
  - Measure of cluster coherence:
    - Distance within a cluster -> small (compactness)
    - Distance between clusters -> high (separation)
  - Stability measures:
    - Cluster data in part of the data and compare results
- External validation:
  - Compare clusters with class labels (iris data)
    - Not possible in real word problems!

Silhouette - Internal Index

The silhouette for a given object *i* is defined as:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

where

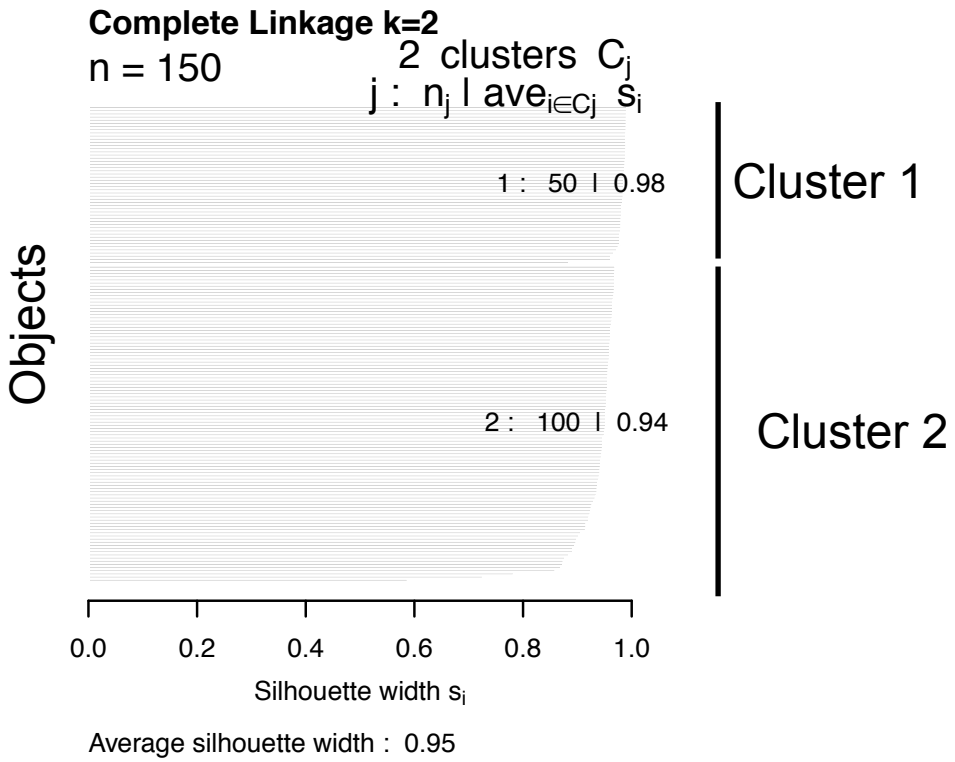
a(i) –mean distance of *i* to objects on same cluster **(compactness)**
d(i,k) – mean distance of *i* to objects of cluster *k* (not own)
b(i) = min_k *(d(i,k))* **(separation)**

Average of *s(i)* -> quality of all results or clusters

Value of 1 indicate perfect solutions!

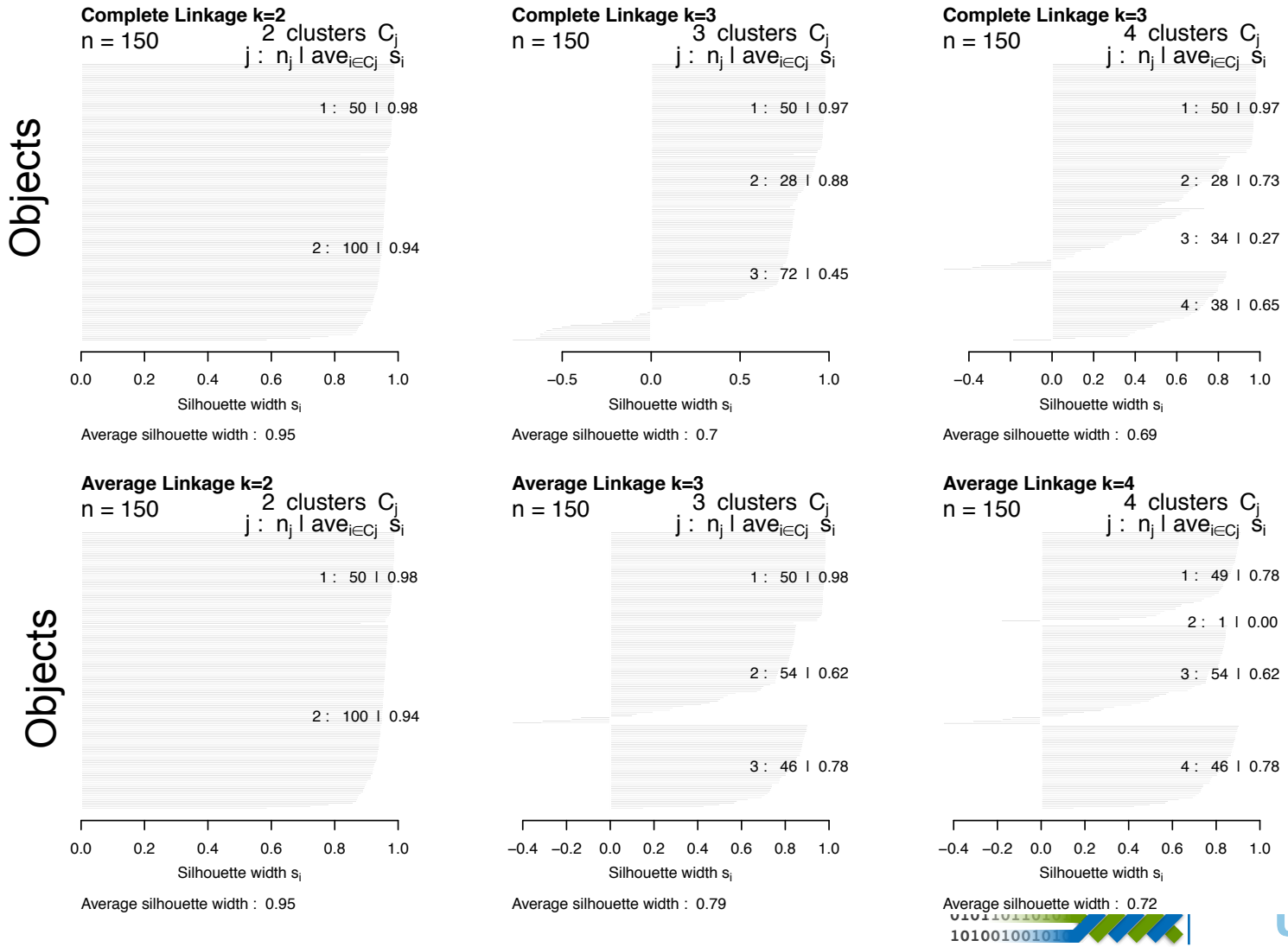
Silhouette - Internal Index / Iris

- silhouette values for hierarchical clustering with Pearson



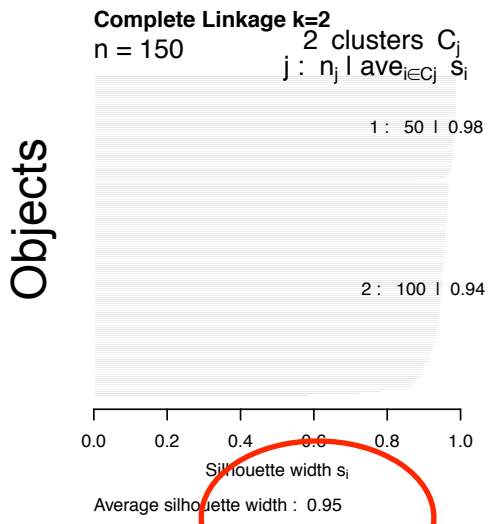
Silhouette - Internal Index / Iris

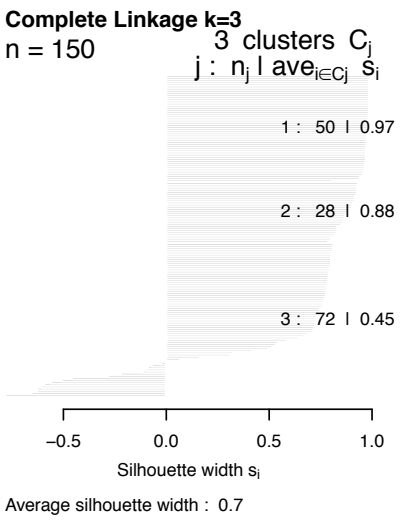
- silhouette values for hierarchical clustering with Pearson

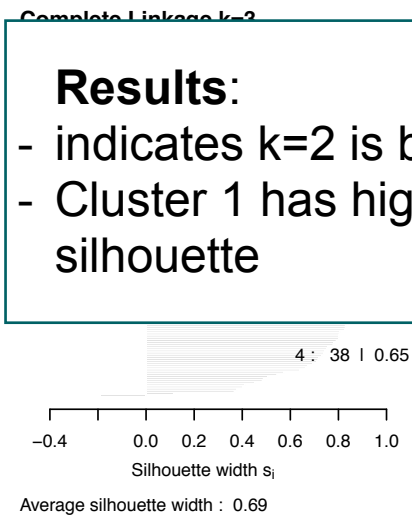


Silhouette - Internal Index / Iris

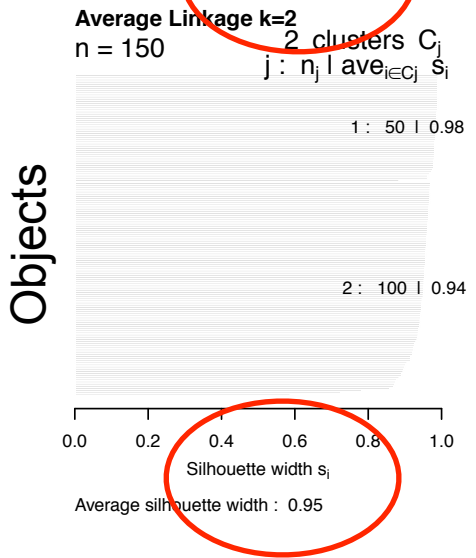
- silhouette values for hierarchical clustering with Pearson

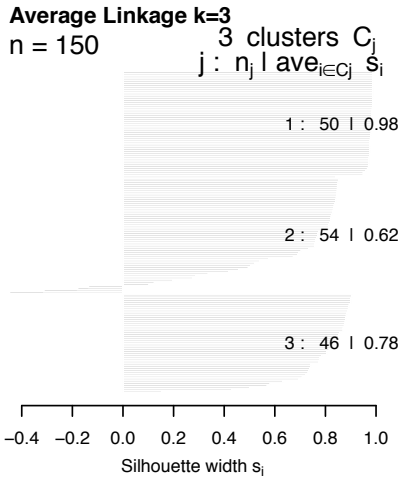


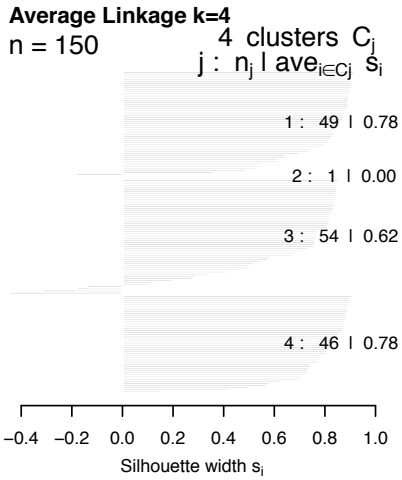




Results:
- indicates k=2 is best
- Cluster 1 has high silhouette



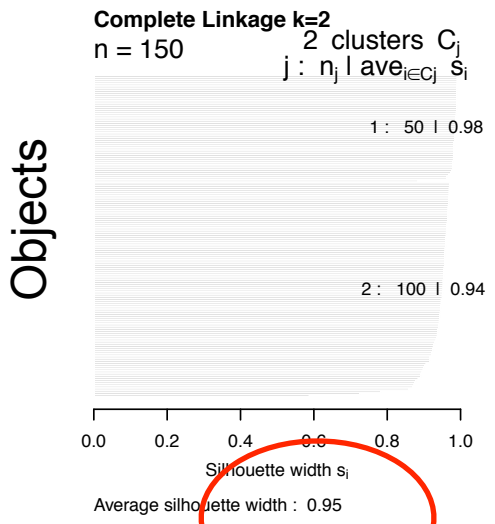


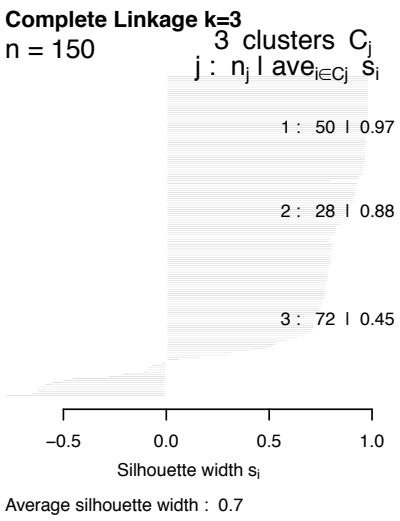


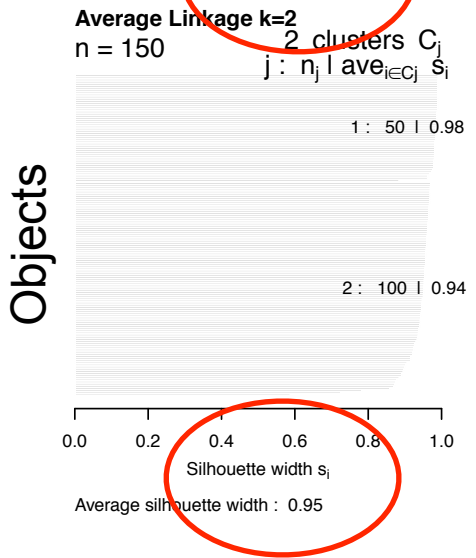
RWTH AACHEN UNIVERSITY

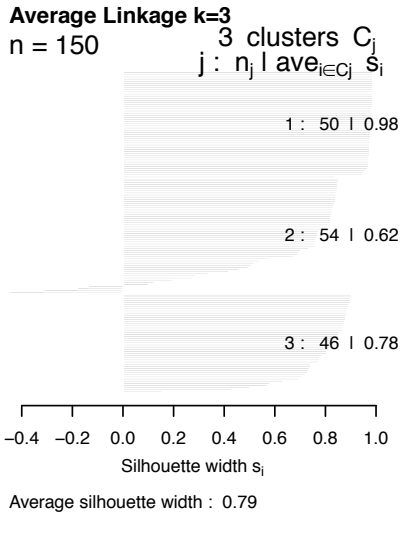
Silhouette - Internal Index / Iris

- silhouette values for hierarchical clustering with Pearson

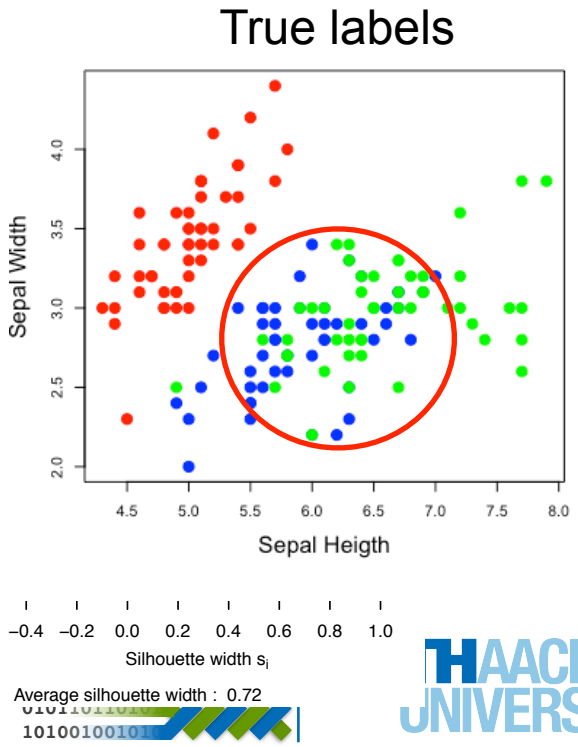








Results:
- indicates k=2 is best
- Cluster 1 has high silhouette



Gap statistic - Internal Index

For a given solution with *K clusters*

$$W_K = \sum_{k=1}^{K} \sum_{y_i=k} \sum_{y_j=k} ||x_i - x_j||^2$$

W_K - measures cluster compactness
W_K - tends to 0 for increasing K

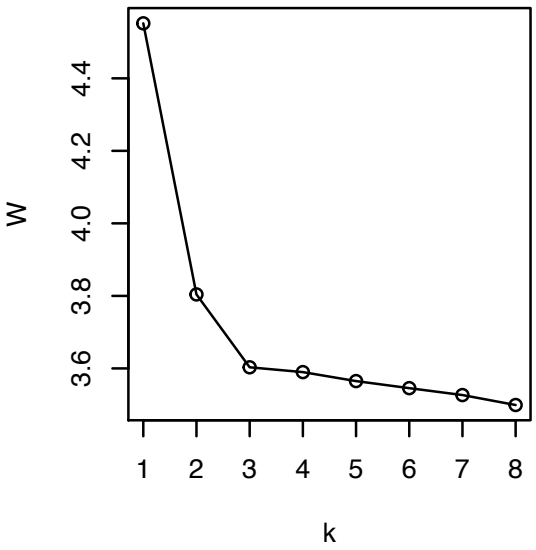
The Gap Statistic consider clustering of random data *W**

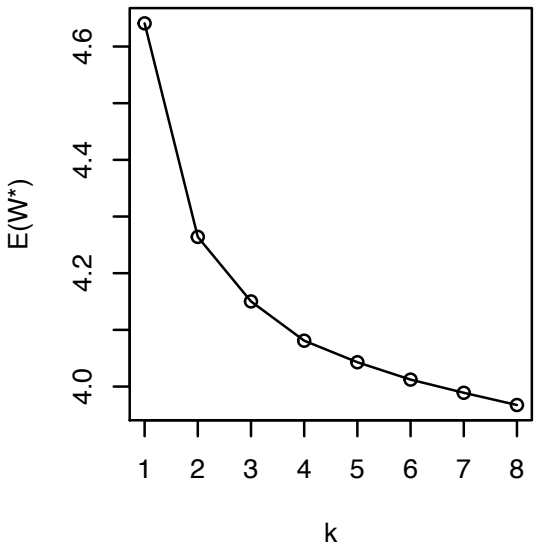
$$GAP(k) = E_r[logW_K^*] - logW_K$$

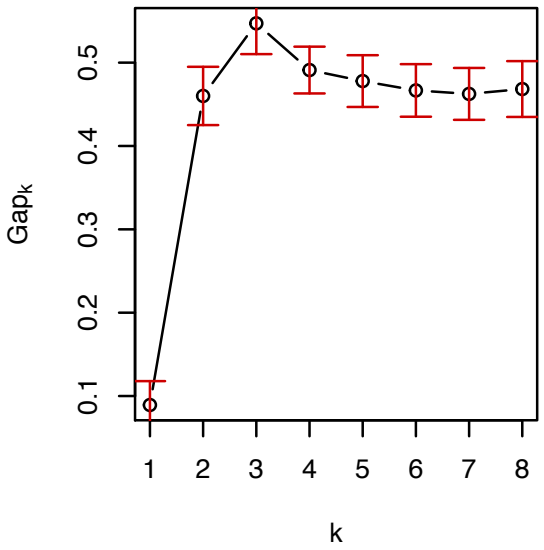
where *W** estimated from clustering random points at the same data space of *X*

Gap statistic - Iris

- GAP statistics for Iris / Average Linkage with Pearson



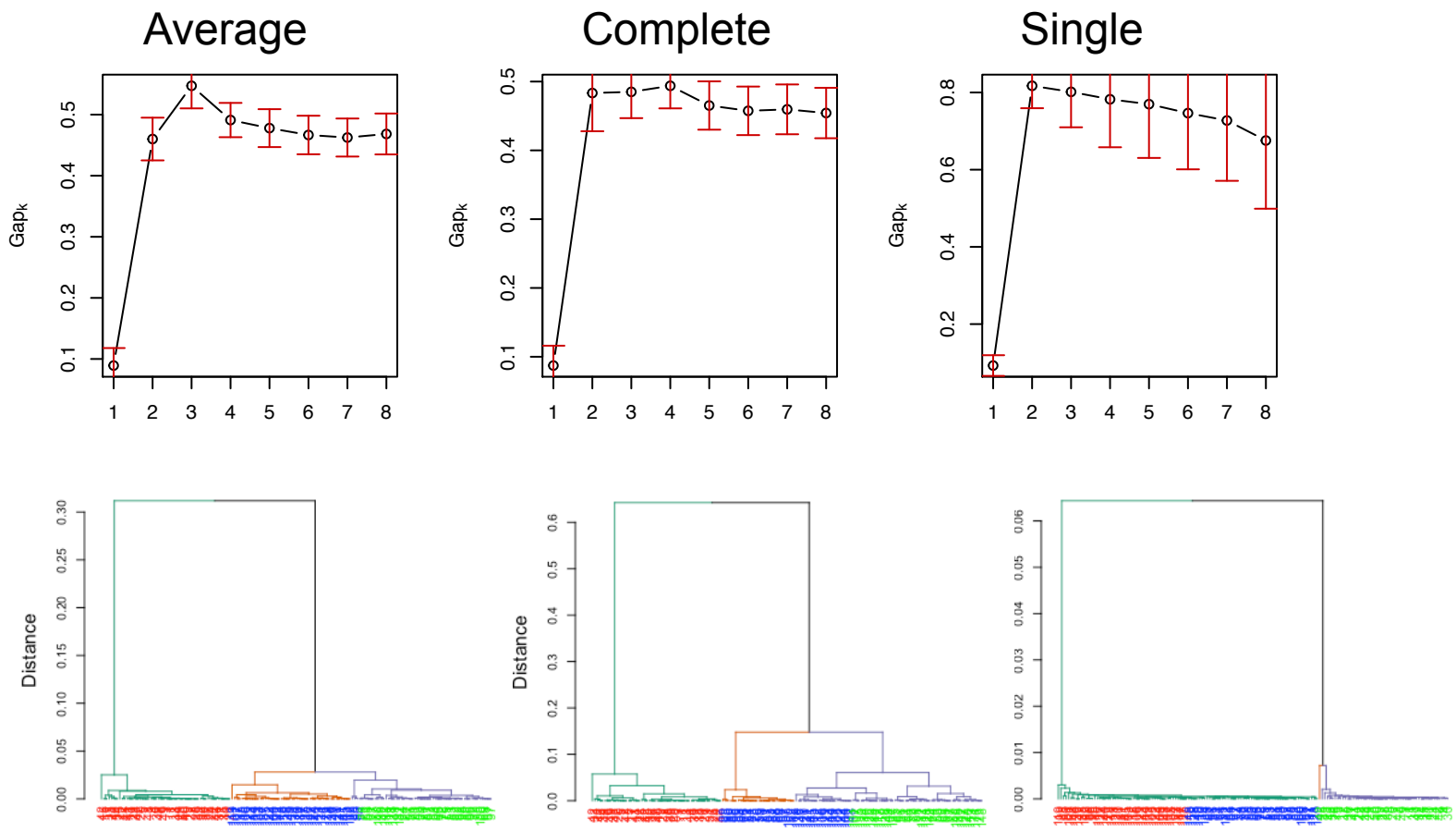




3 clusters has highest Gap !!!

Gap statistic - Iris

- GAP statistics for distinct linkage methods



Resume / Validation

- Help detection of number of clusters / real clusters
  - Do not work perfectly!
- GAP statistics is widely used
  - Requires *r* data randomisations
    - high computational costs
    - random datasets uniformly distributed (unreal assumption)
- Expert interpretation is important!

Dimension Reduction

- Distances lose meaning at high dimensional space (curse of dimensionality)

- Unspecific Filtering (without class labels):
  - Keep variables with highest variance
    - ***rational**: i*mportant features change values across groups

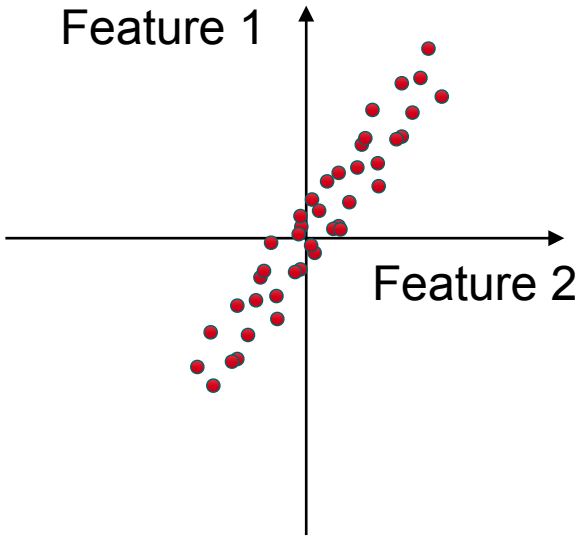
- Dimensionality Reduction by Transformation:
  - linear: principal component analysis (PCA)
  - Non-linear / manifold learning: t-SNE & UMAP

Principal Component Analysis

- For a data *X*, find linear combination of features (*w)* capturing most of data variance

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|^2\}$$

- Can be solved by linear algebra / eigen vector decomposition



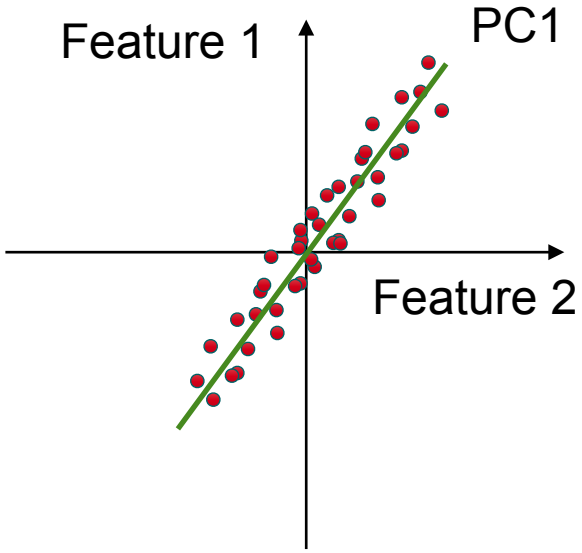
Recommended reading:
Ringner M., *Nature Biotechnology* 26, 303 - 304 (2008)

Principal Component Analysis

- For a data *X*, find linear combination of features (*w)* capturing most of data variance

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|^2\}$$

- Can be solved by linear algebra / eigen vector decomposition



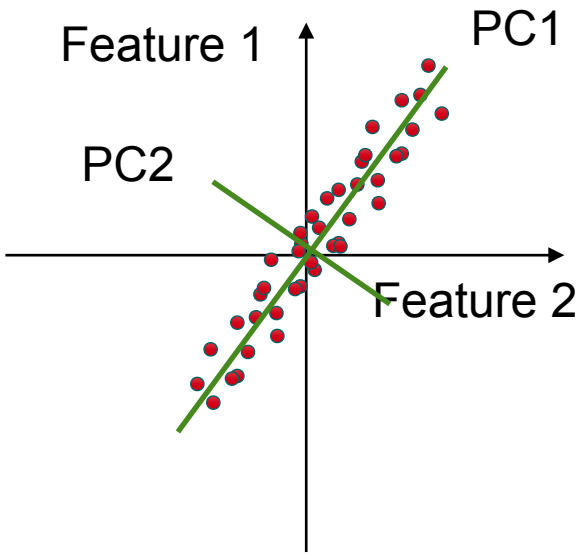
Recommended reading:
Ringner M., *Nature Biotechnology* 26, 303 - 304 (2008)

Principal Component Analysis

- For a data *X*, find linear combination of features (*w)* capturing most of data variance

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|^2\}$$

- Can be solved by linear algebra / eigen vector decomposition



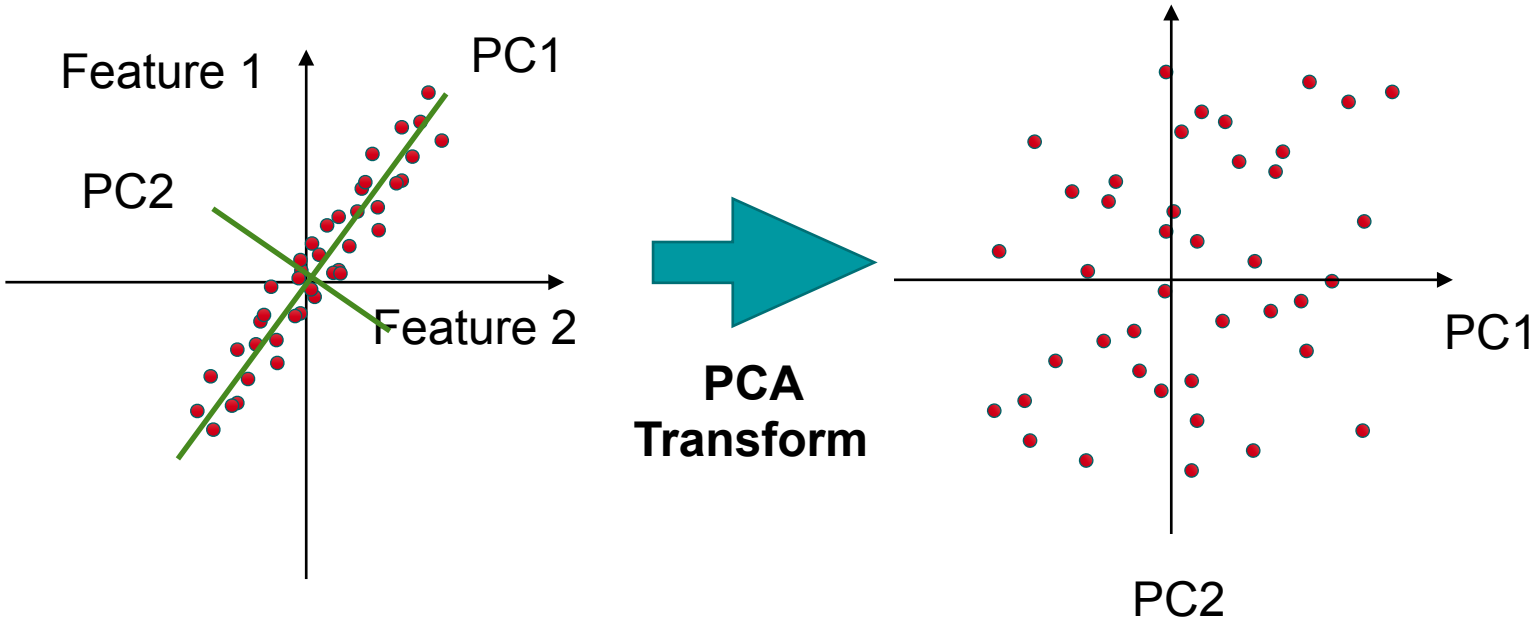
Recommended reading:
Ringner M., *Nature Biotechnology* 26, 303 - 304 (2008)

Principal Component Analysis

- For a data *X*, find linear combination of features (*w)* capturing most of data variance

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|^2\}$$

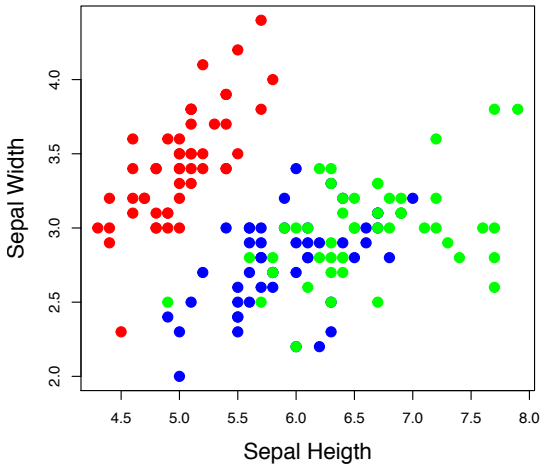
- Can be solved by linear algebra / eigen vector decomposition

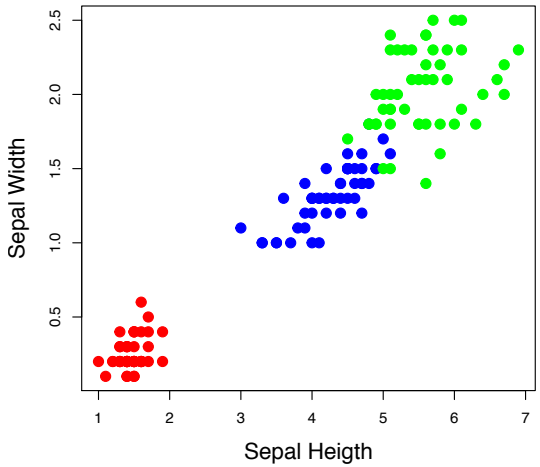


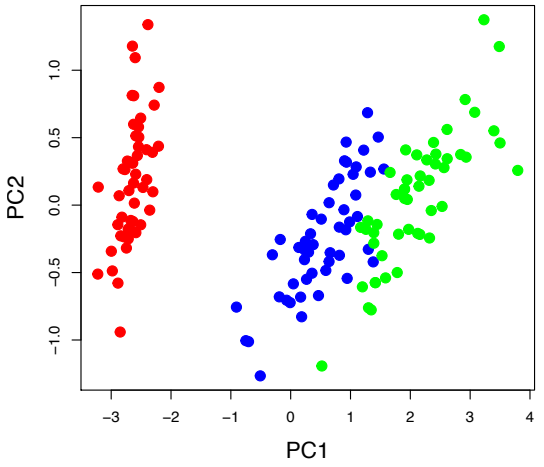
Recommended reading:
Ringner M., *Nature Biotechnology* 26, 303 - 304 (2008)

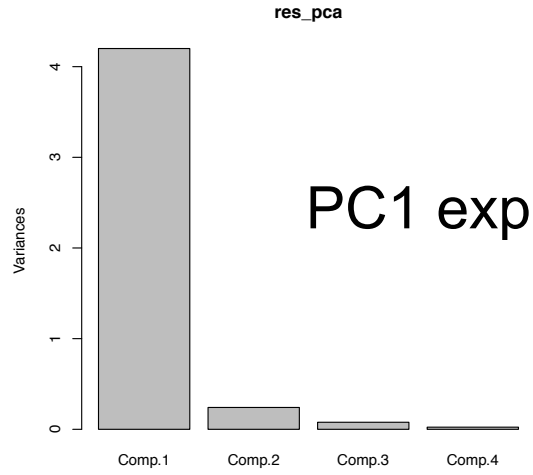
PCA - Iris

- Original iris data had 4 variables





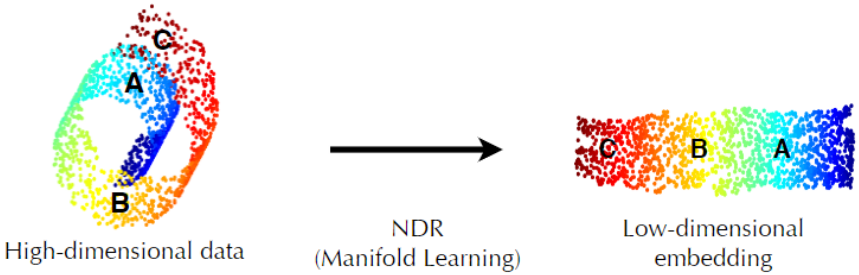




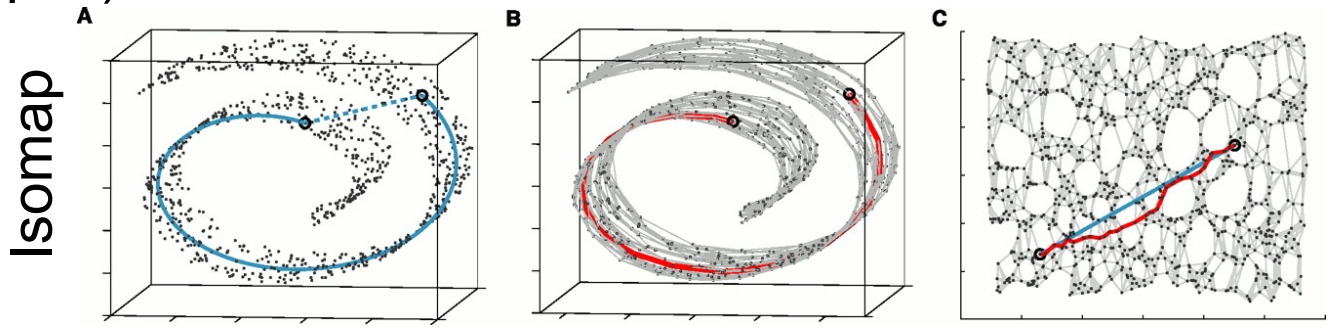
PC1 explains most of variance

Non-linear / Manifold methods

- Data might be distributed at particular regions of a high dimensional space



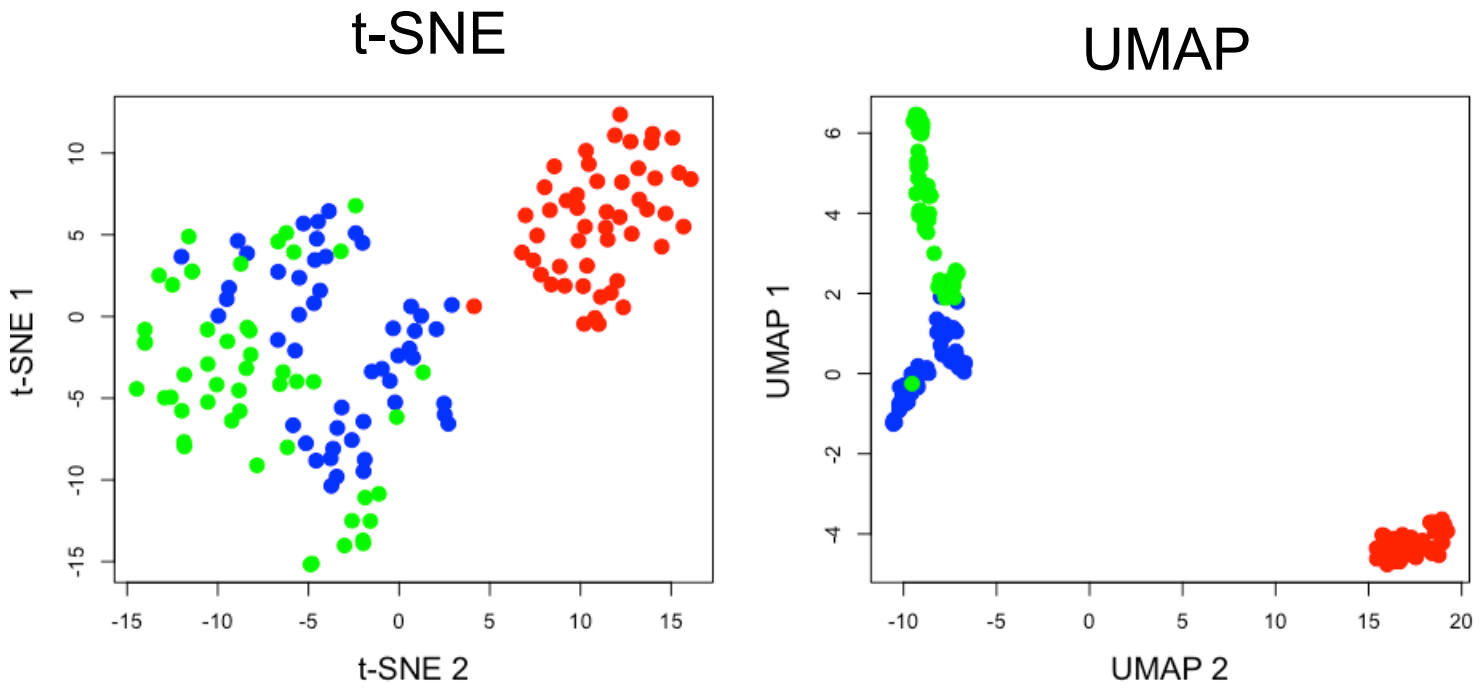
- Manifold methods use topological distance (nearest neighbour graphs)



- t-SNE and UMAP are newer/widely used methods

Adapted from Tenembaum, et al. 2000

Manifold learning and IRIS



- Nice low dimensional visualisation of the data
- Caution: These methods fail capturing global structures (distance between clusters!)

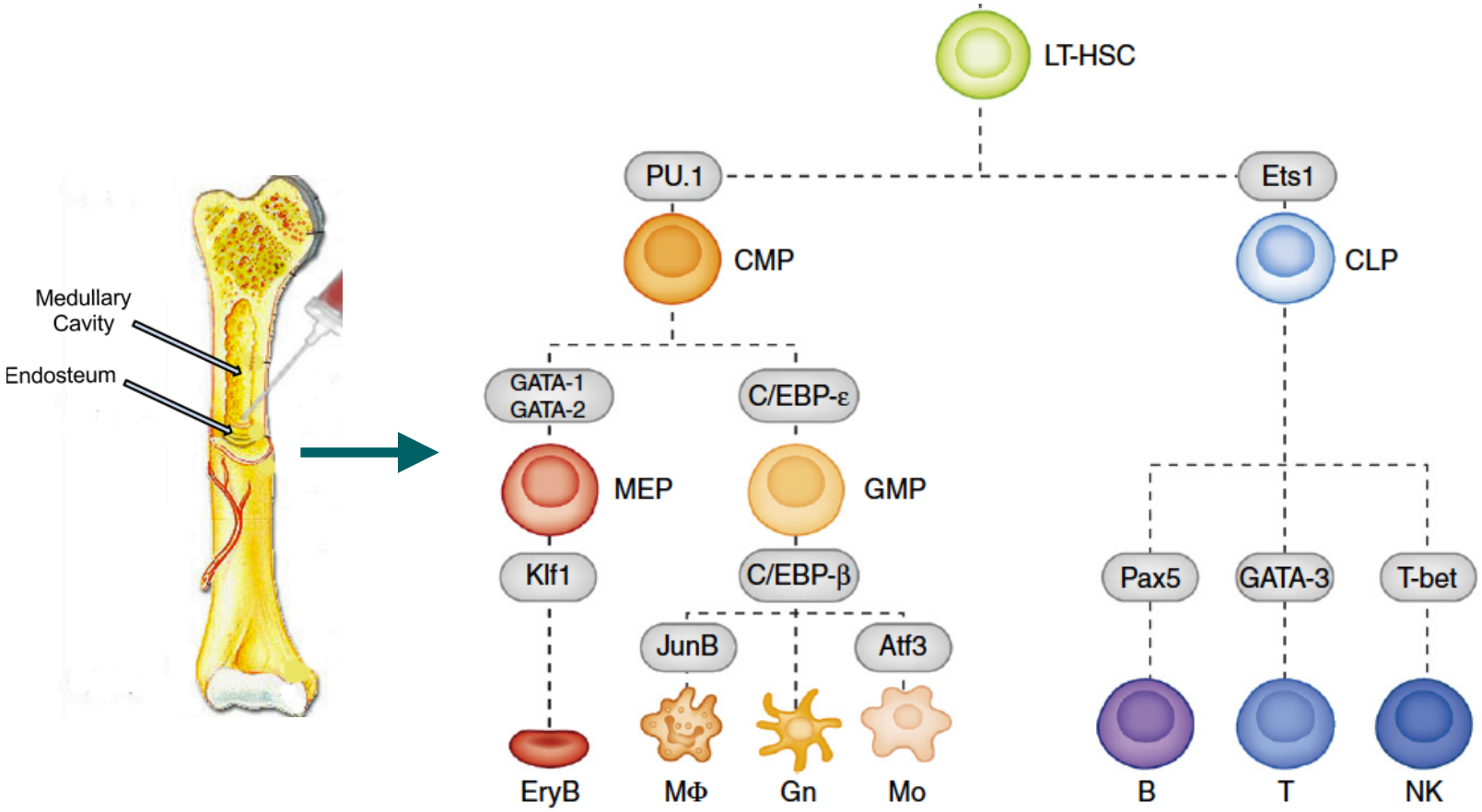
See for more details: https://www.youtube.com/watch?v=9iol3Lk6kyU&t=350s

Resume / Dimension Reduction

- PCA analysis is a wide spread technique to reduce dimension!
  - Loses importance of individual variables
- Manifold methods
  - Nice low dimensional representation of data
  - Require parametrisation and loose global distance information

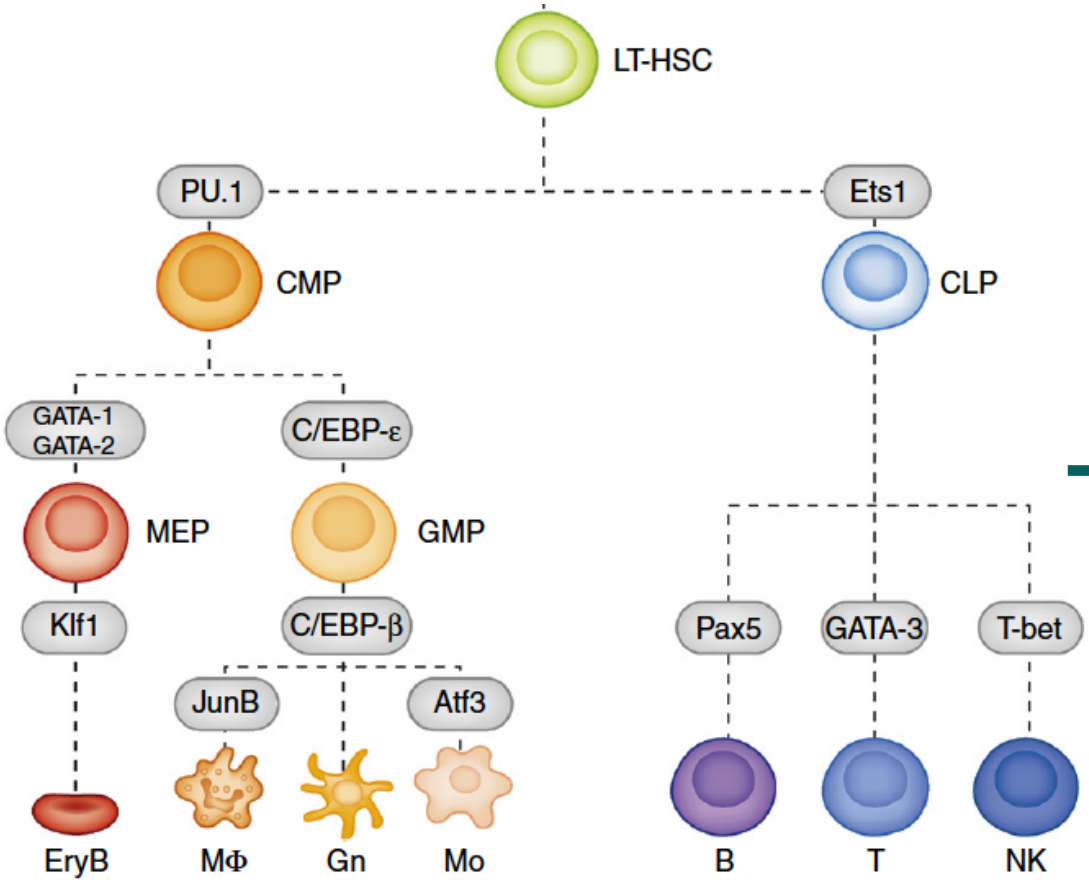
Expression at Single Cell Level

Cell Differentiation



Source: Amit (2016), *Nature Immunoloy.*

Cell Differentiation & Gene Expression

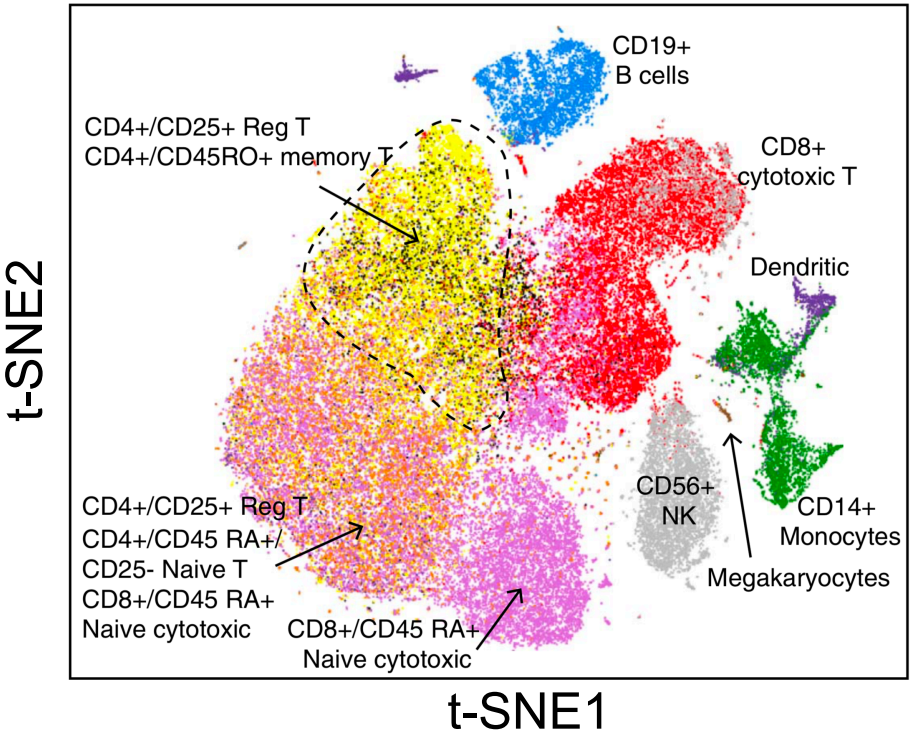


	Cell 1	Cell 2	...
Gene 1	25	918	
Gene 2	0	456	
Gene 3	20	342	
Gene 4	0	214	
...			

Source: Amit (2016), *Nature Immunoloy.*

Gene Expression of Lymphoid Cells

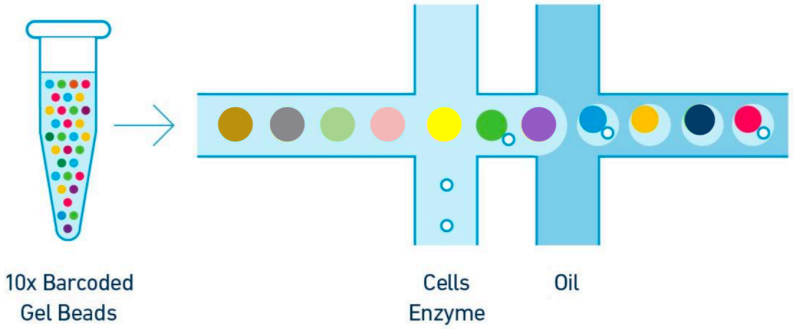
PBMCs from Humans

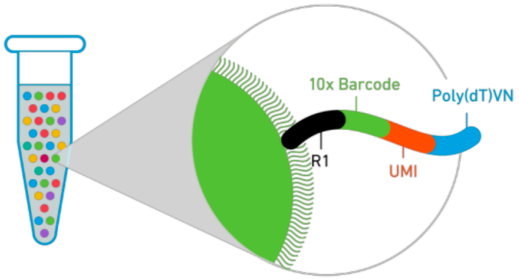


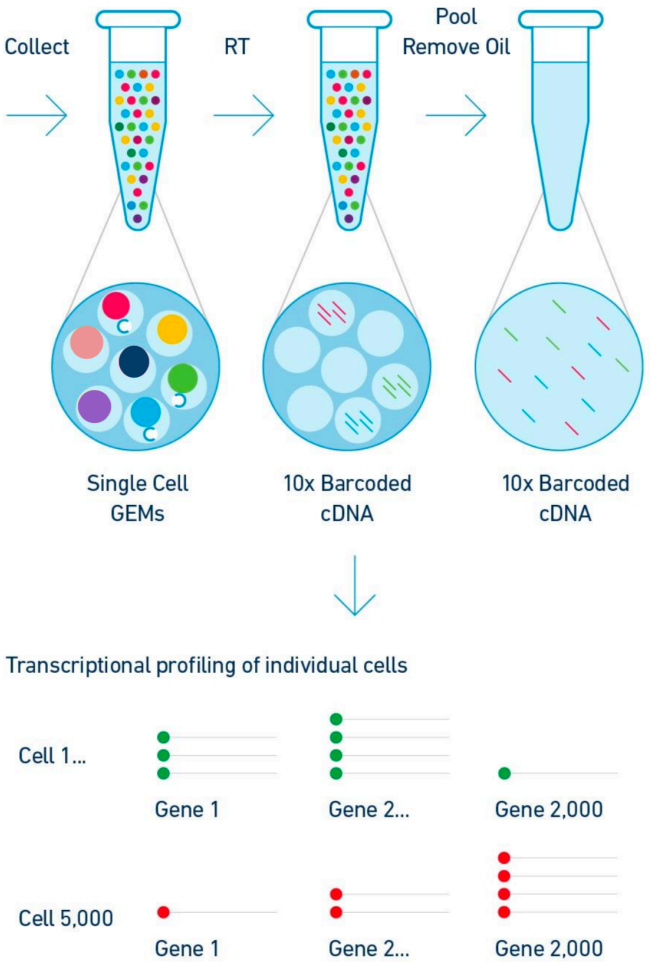
Single cell RNA-seq from 68k cells

Source: Zheng et al. 2017 & Buenrostro et al. 2018

Droplet based RNA single cell sequencing





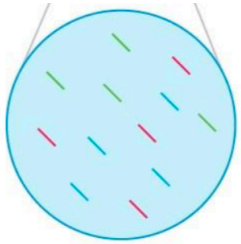


- Input: Single cells in suspension + 10x Gel Beads and Reagents
- Output: Digital gene expression profiles from every partitioned cell

Source: 10x genomics

Basics Bioinformatics - single cell RNA-seq

Sequences



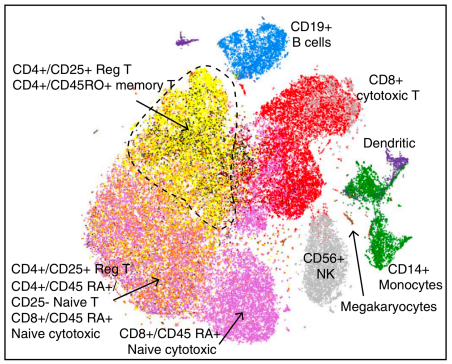
10x Barcoded cDNA

Expression matrix

Read counts

	Cell 1	Cell 2	...
Gene 1	25	918	
Gene 2	0	456	
...			

Clustering

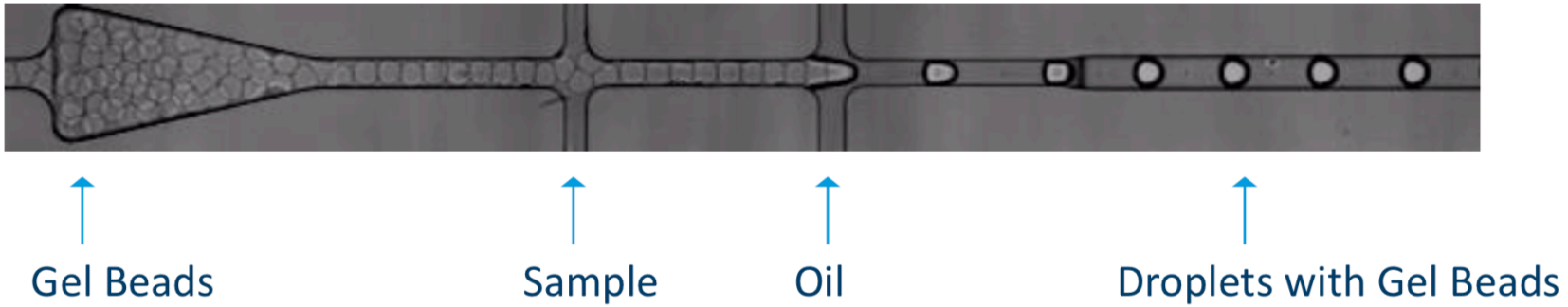


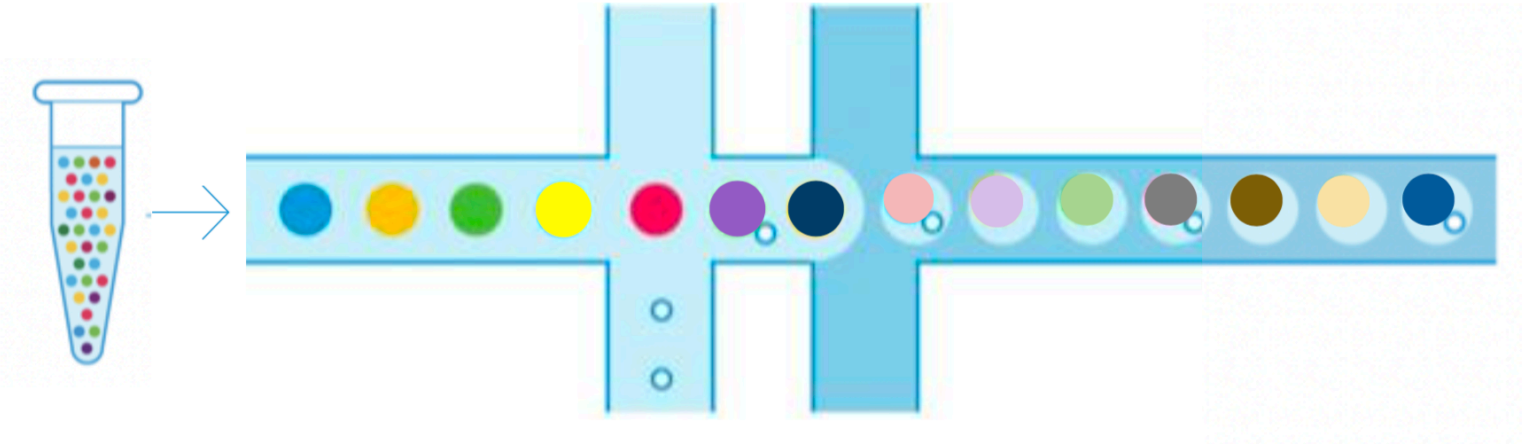
1. Alignment / Transcript count
2. Cell filtering
3. Removal of biological variation
4. Dimension reduction / cell clustering

cell ranger 10x genomics

Seurat -R

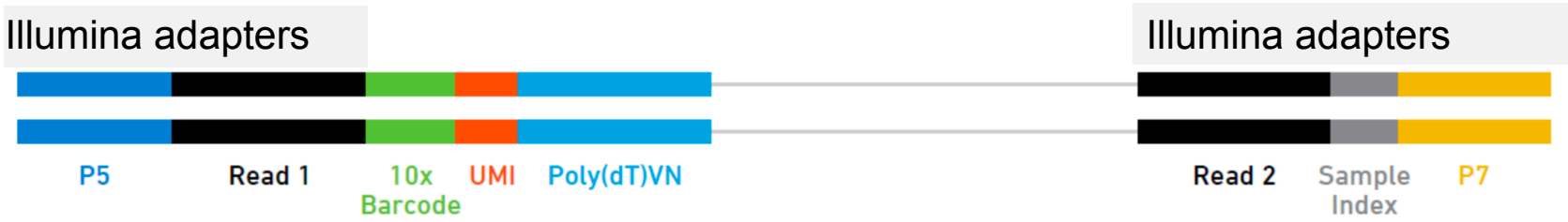
Droplet based RNA single cell sequencing





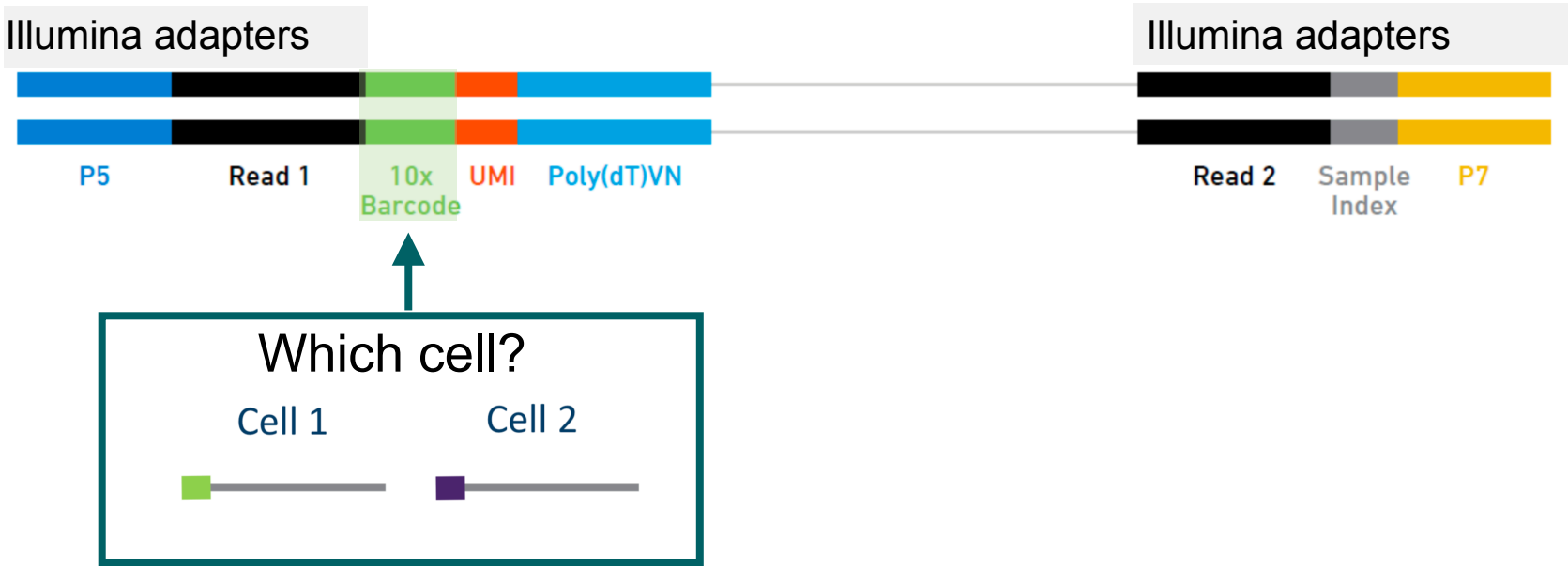
Source: 10x genomics

Basics Bioinformatics - Transcript Counts



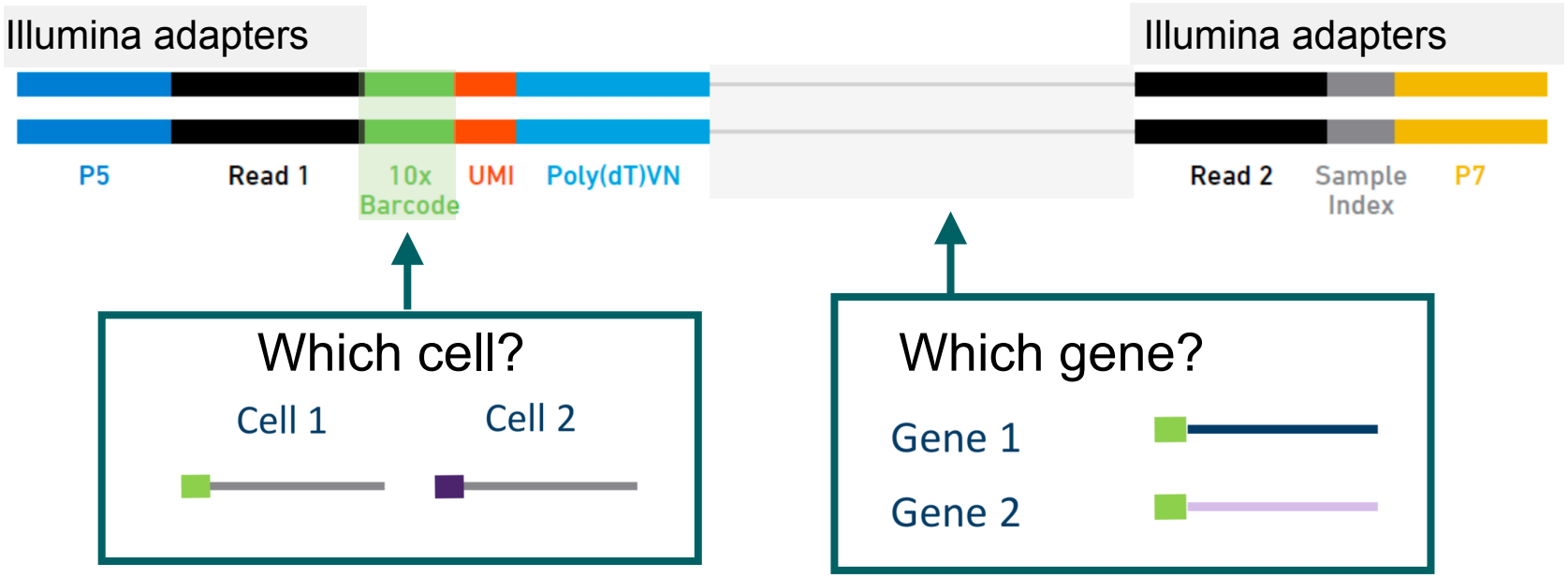
Source: 10x genomics

Basics Bioinformatics - Transcript Counts



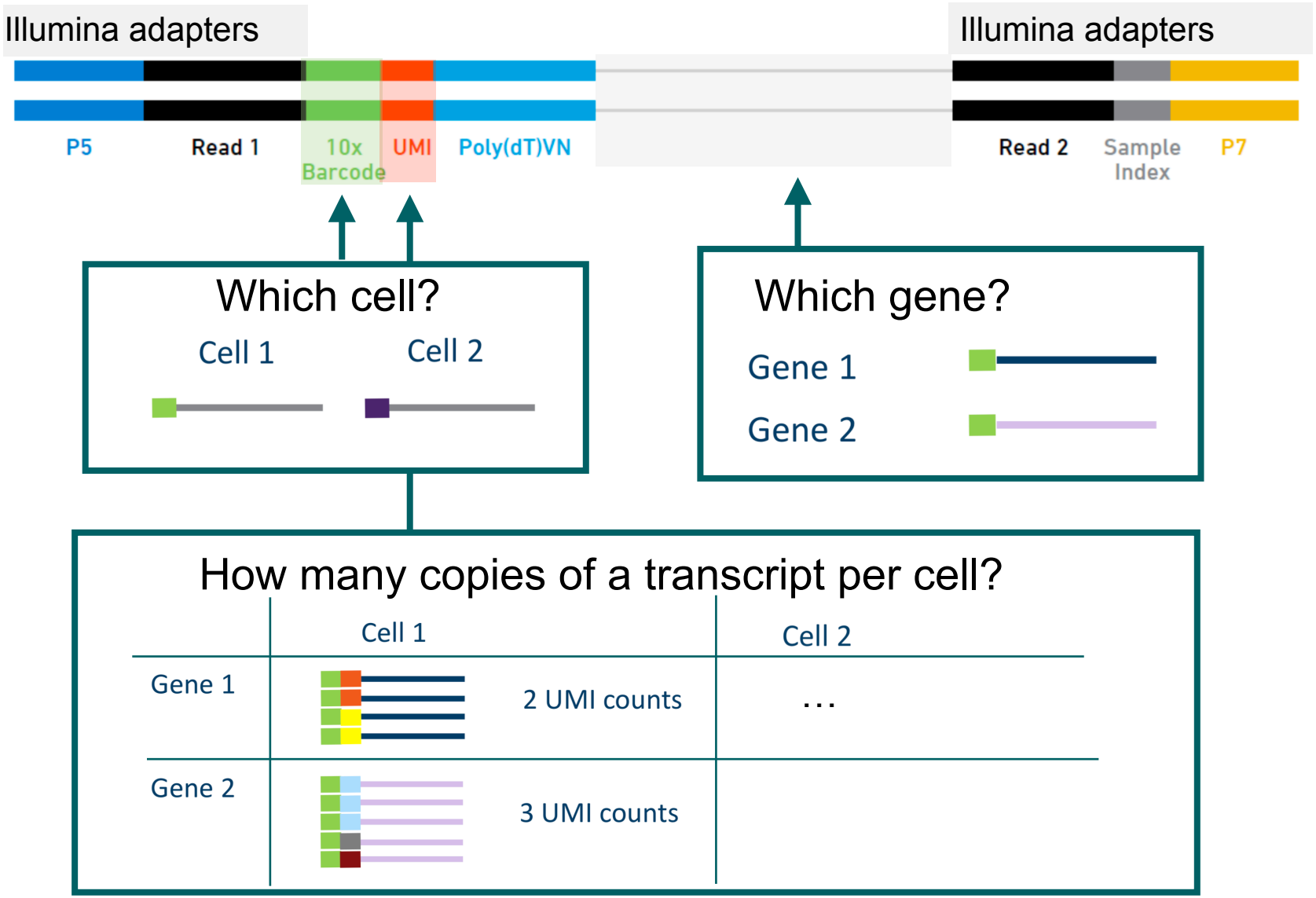
Source: 10x genomics

Basics Bioinformatics - Transcript Counts



Source: 10x genomics

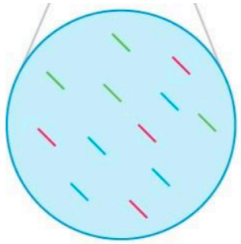
Basics Bioinformatics - Transcript Counts



Source: 10x genomics

Basics Bioinformatics - single cell RNA-seq

Sequences

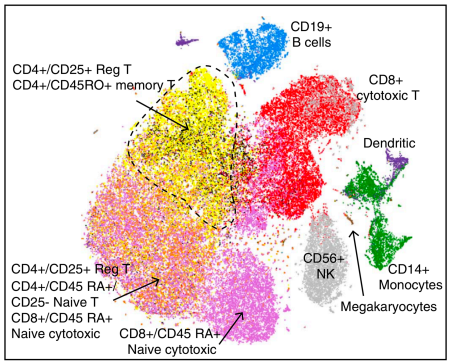


Expression matrix

Read counts

	Cell 1	Cell 2	...
Gene 1	25	918	
Gene 2	0	456	
...			

Clustering



1. Alignment / Transcript count

2. Cell filtering

3. Removal of biological variation

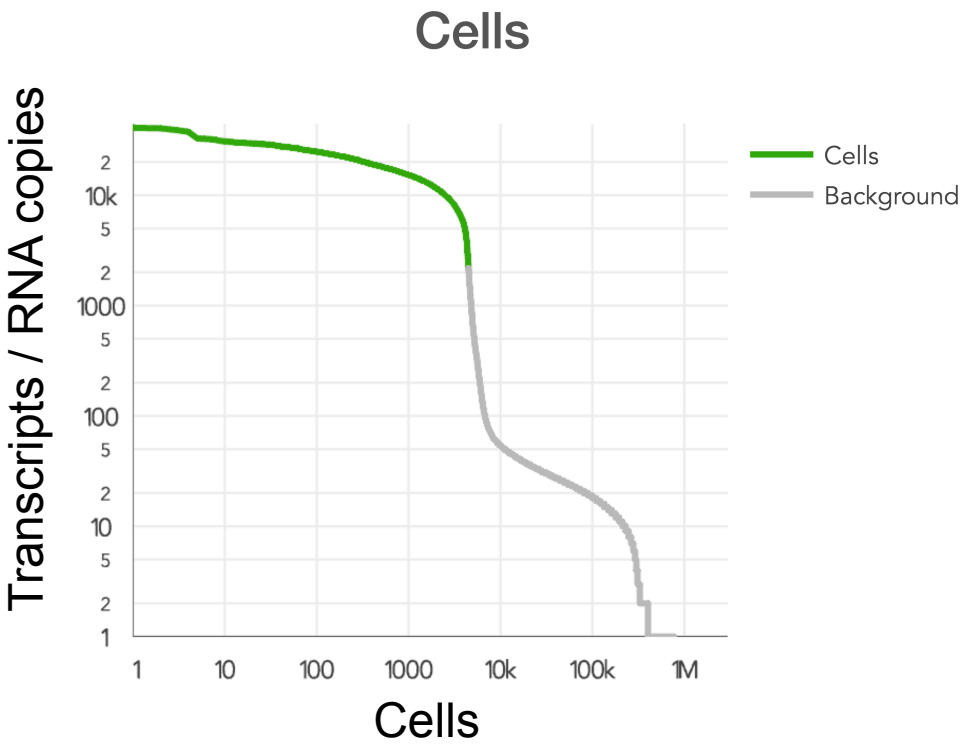
4. Dimension reduction / cell clustering

cell ranger
10x genomics

Seurat -R

Basics Bioinformatics - Cell Filtering

1. sum UMIs (copy of transcripts) per cell
2. consider cells with total UMI count > 99th of expected recovered cells



Estimated Number of Cells
4,495

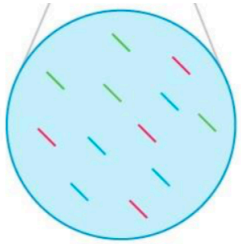
Post-Normalization Mean Reads per Cell
89,289

Median Genes per Cell
2,504

cell ranger - 10x genomics

Basics Bioinformatics - single cell RNA-seq

Sequences



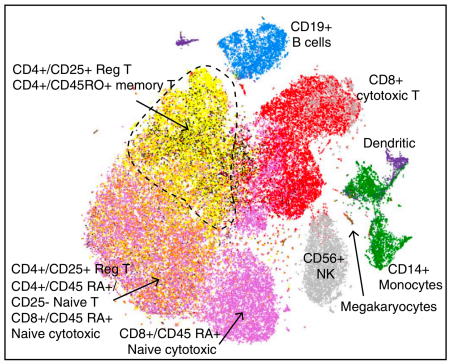
10x Barcoded cDNA

Expression matrix

Read counts

	Cell 1	Cell 2	...
Gene 1	25	918	
Gene 2	0	456	
...			

Clustering



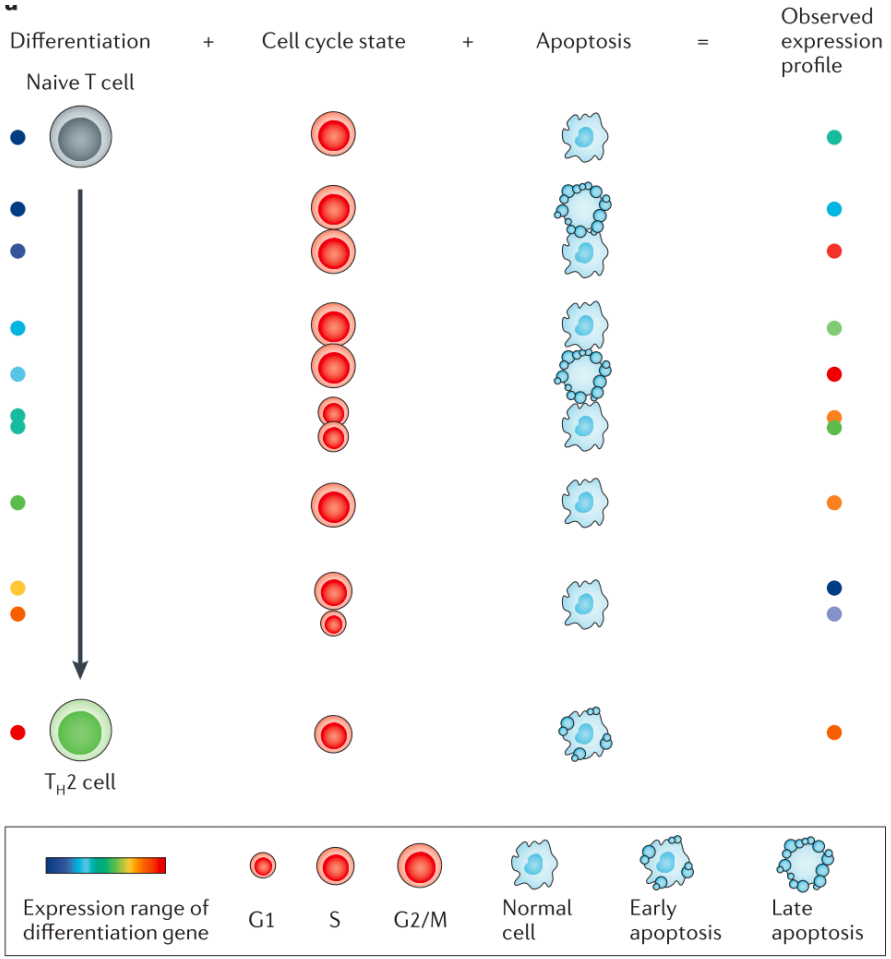
1. Alignment / Transcript count
2. Cell filtering
3. Removal of biological variation
4. Dimension reduction / cell clustering

cell ranger
10x genomics

Seurat -R

Removal of Unwanted Biological Variation

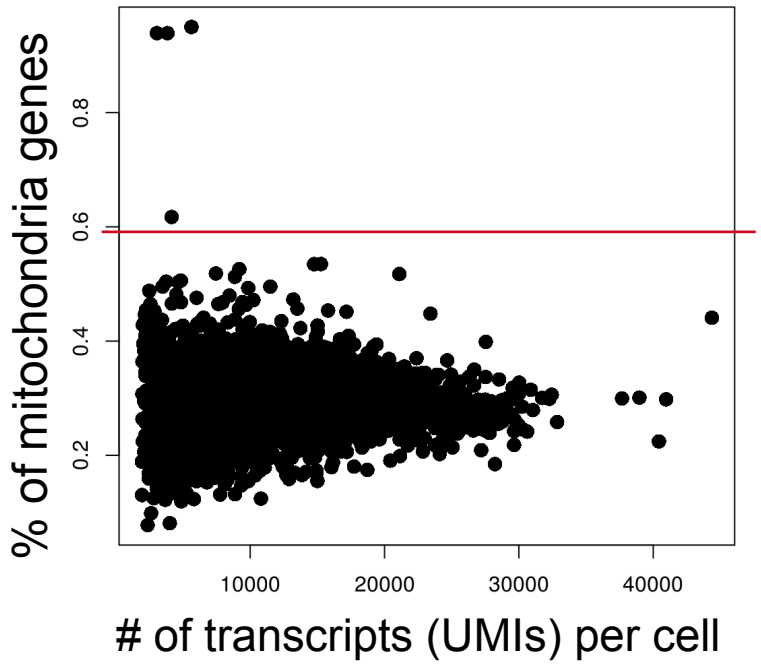
some biological changes might not be of interest for your study



Source: Stegle et al. 2015

Removal of Unwanted Biological Variation

Cells with high % of mitochondria genes

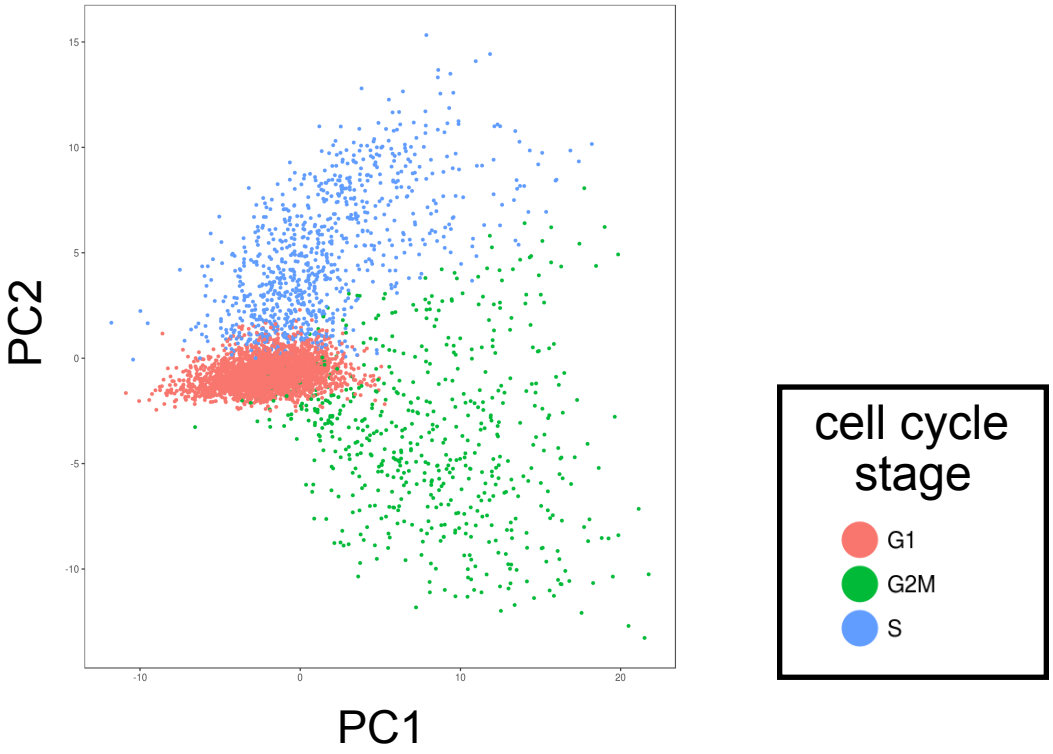


Data: dendritic cells of Peyer patches
with Torrow & Hornef UK Aachen

Removal of Unwanted Biological Variation

Remove genes associated to cell cycle

PCA based on cell cycle genes

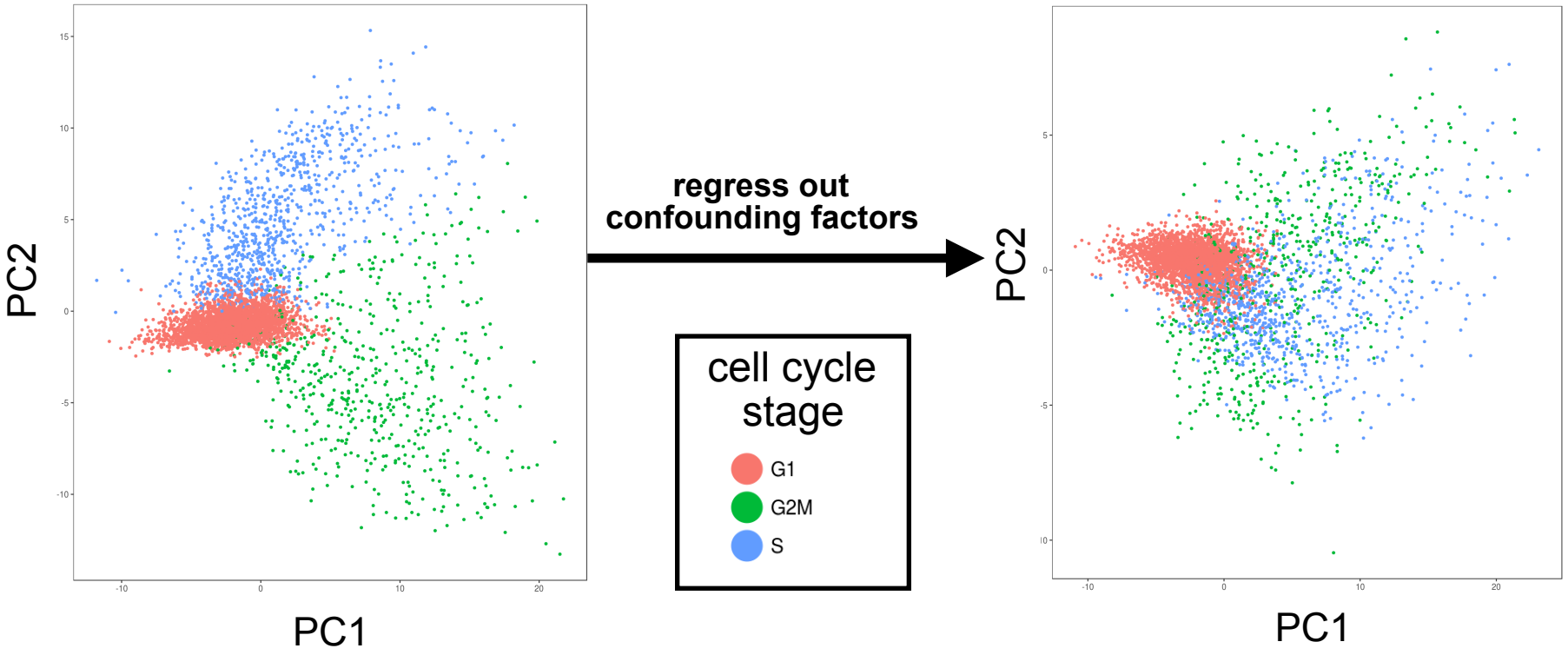


Data: dendritic cells of Peyer patches with Torrow & Hornef UK Aachen

Removal of Unwanted Biological Variation

Remove genes associated to cell cycle

PCA based on cell cycle genes



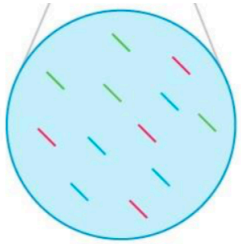
Other confounding factors: experimental replicates, individual variation

Data: dendritic cells of Peyer patches
with Torrow & Hornef UK Aachen

Institute for
Computational Genomics
RWTH AACHEN
UNIVERSITY

Basics Bioinformatics - single cell RNA-seq

Sequences



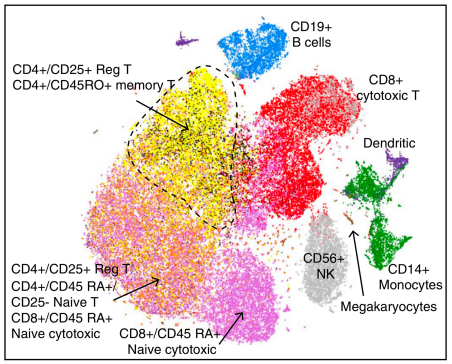
10x Barcoded cDNA

Expression matrix

Read counts

	Cell 1	Cell 2	...
Gene 1	25	918	
Gene 2	0	456	
...			

Clustering



1. Alignment / Transcript count

2. Cell filtering

3. Removal of biological variation

4. Dimension reduction / cell clustering

cell ranger 10x genomics

Seurat -R

Basics Bioinformatics - Dimension Reduction

Expression matrix

Read counts

	Cell 1	Cell 2	...
Gene 1	25	918	
Gene 2	0	456	
Gene 3	20	342	
Gene 4	0	214	
...			

- High dimension matrix:
  - 4945 cells vs. 17328 genes
- Sparse matrix:
  - 50% zeros (90k reads per cell)

Data: dendritic cells of Peyer patches with Torrow & Hornef UK Aachen

Basics Bioinformatics - Dimension Reduction

Expression matrix

Read counts

	Cell 1	Cell 2	...
Gene 1	25	918	
Gene 2	0	456	
Gene 3	20	342	
Gene 4	0	214	
...			

Reduction with t-SNE or PCA

t-SNE matrix

t-SNE Scores

	Cell 1	Cell 2	...
t-SNE1	3.1	0.3	
t-SNE2	-2.1	2.1	

- High dimension matrix:
  - 4945 cells vs. 17328 genes
- Sparse matrix:
  - 50% zeros (90k reads per cell)

PCA - distance preserving
- used for clustering

t-SNE - local neighbourhood preserving
- used for visualisation

Basics Bioinformatics - Dimension Reduction

Expression matrix

Read counts

	Cell 1	Cell 2	...
Gene 1	25	918	
Gene 2	0	456	
Gene 3	20	342	
Gene 4	0	214	
...			

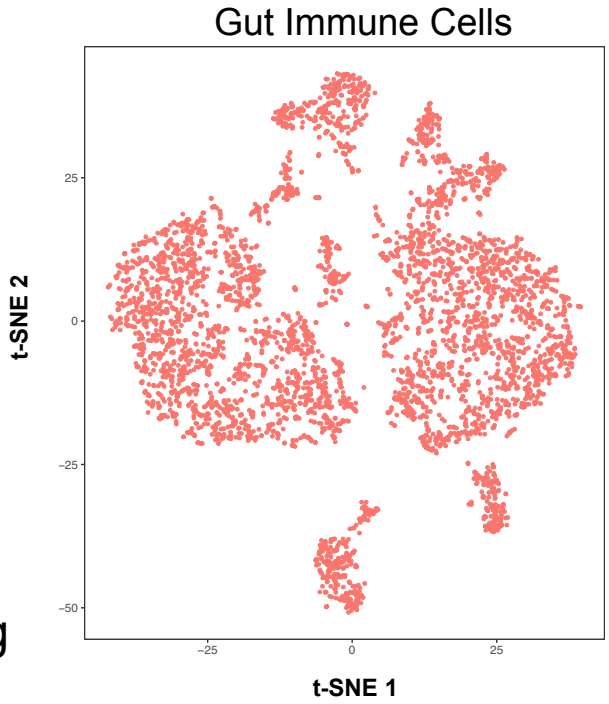
Reduction with t-SNE or PCA

t-SNE matrix

t-SNE Scores

	Cell 1	Cell 2	...
t-SNE1	3.1	0.3	
t-SNE2	-2.1	2.1	

- High dimension matrix:
  - 4945 cells vs. 17328 genes
- Sparse matrix:
  - 50% zeros (90k reads per cell)

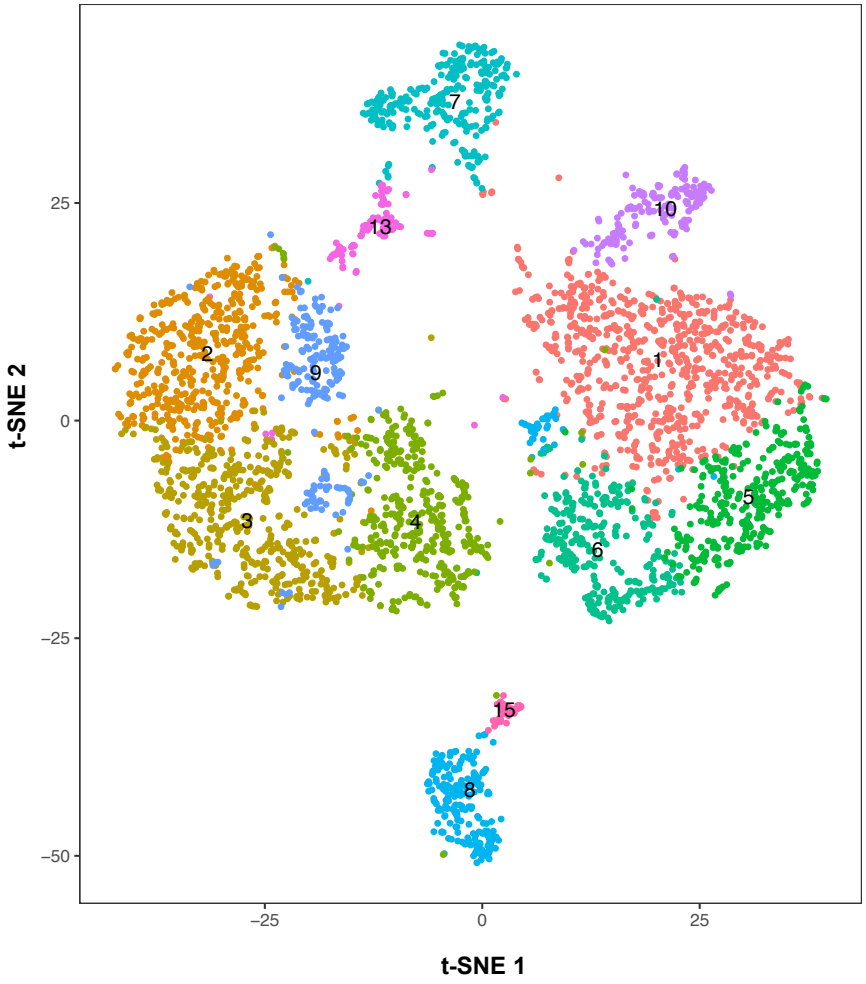


PCA - distance preserving
- used for clustering

t-SNE - local neighbourhood preserving
- used for visualisation

Basics Bioinformatics - Clustering

Gut Immune Cells - 12 groups

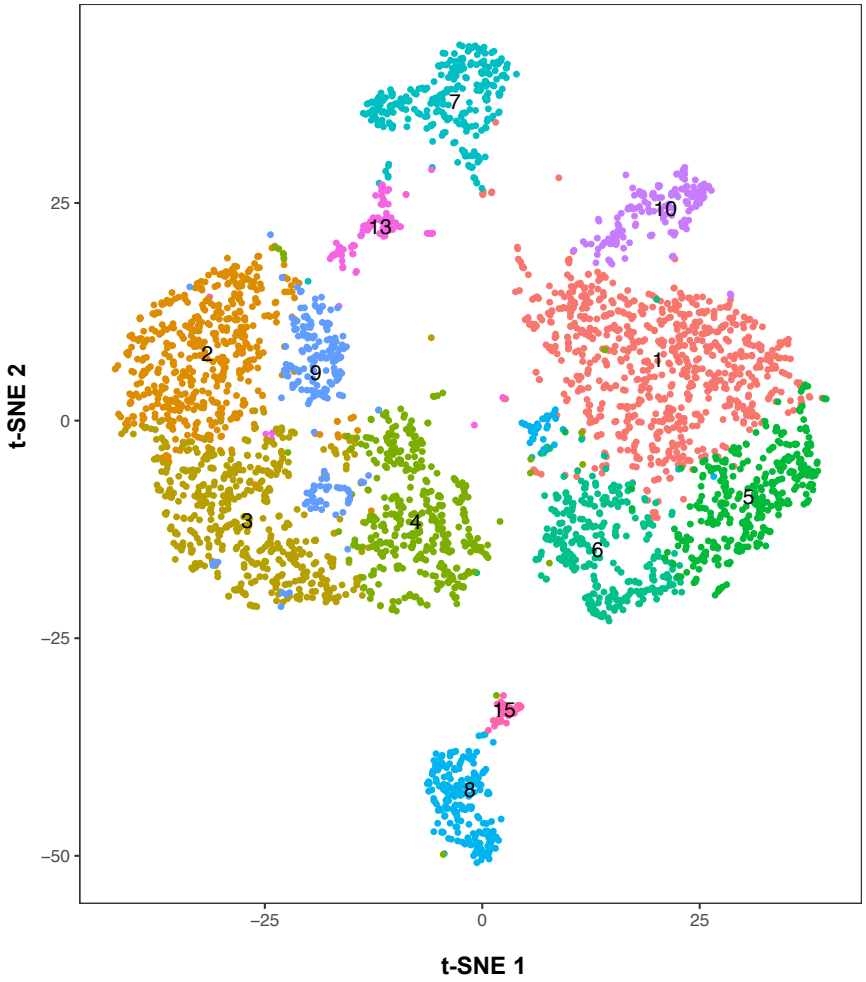


Clustering - identify cells with similar expression patterns
- based on PCA (20 dimension)

How to identify cell types?

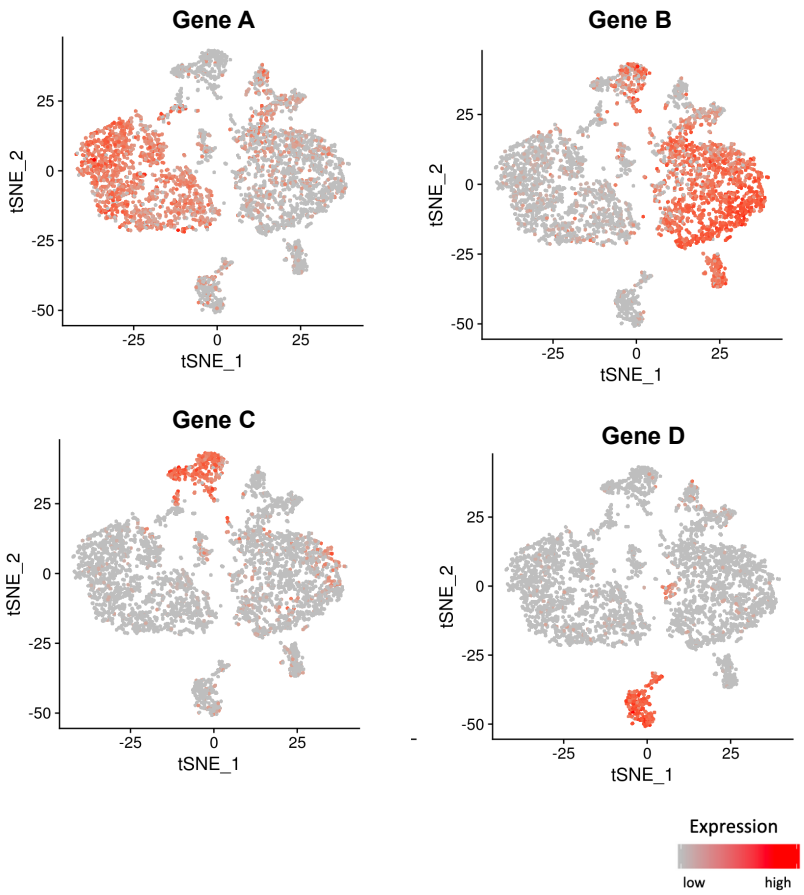
Cell Identity with an Expert

Gut Immune Cells - 12 groups



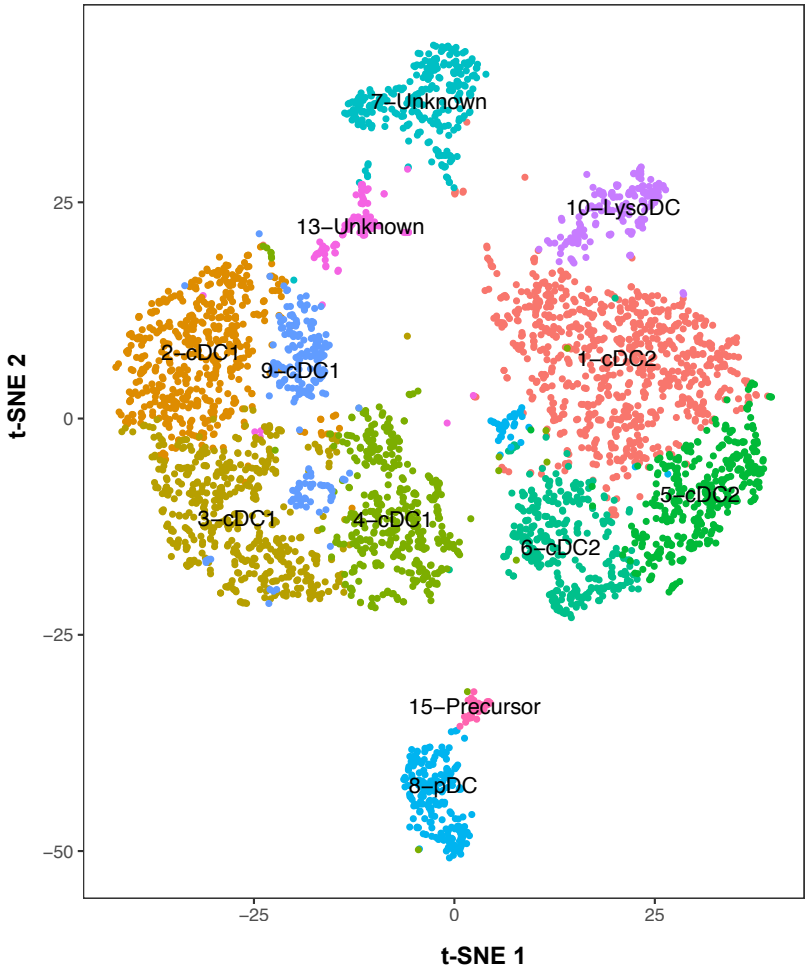
Check expression of:

1. known genes



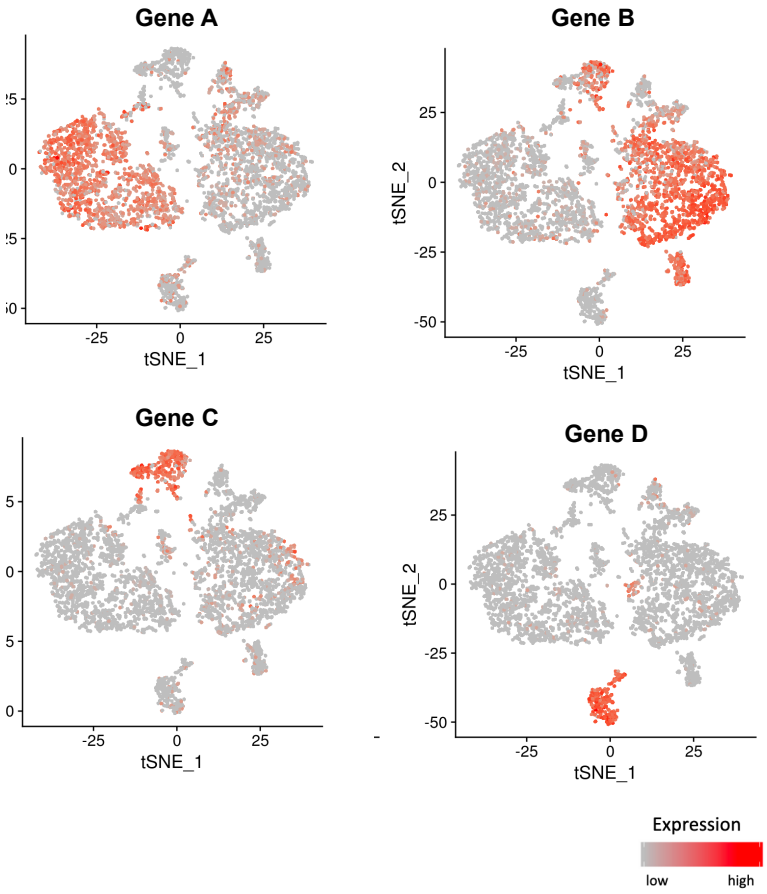
Cell Identity with an Expert

Gut Immune Cells - 12 groups



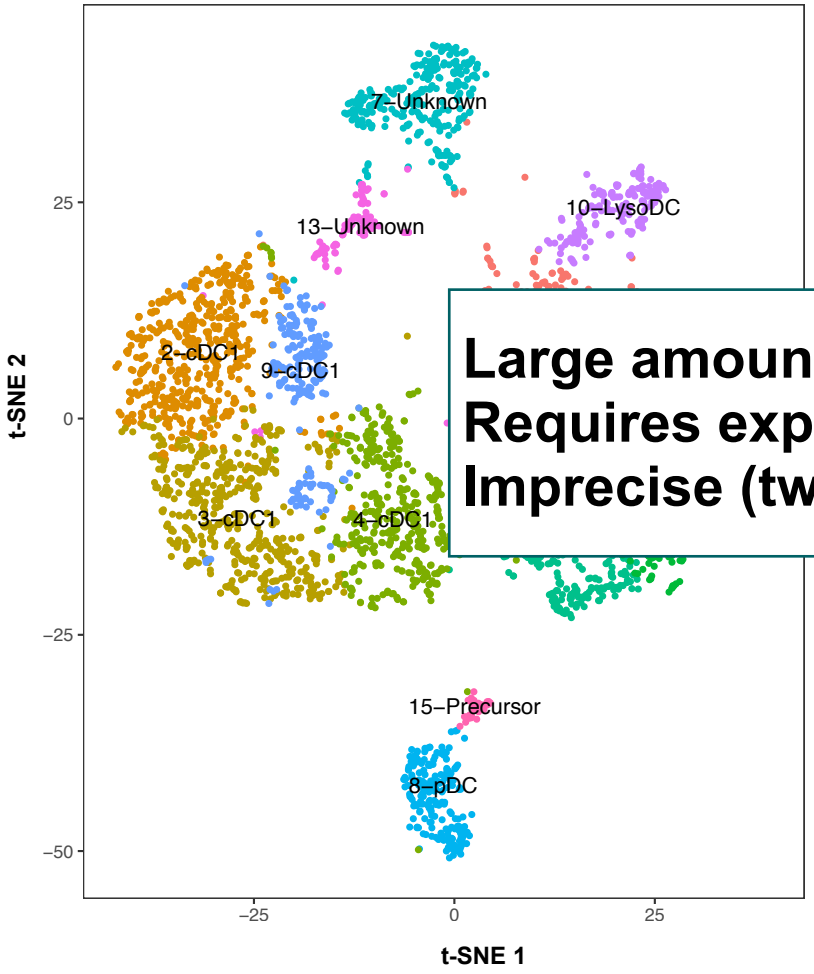
Check expression of:

1. known genes



Cell Identity with an Expert

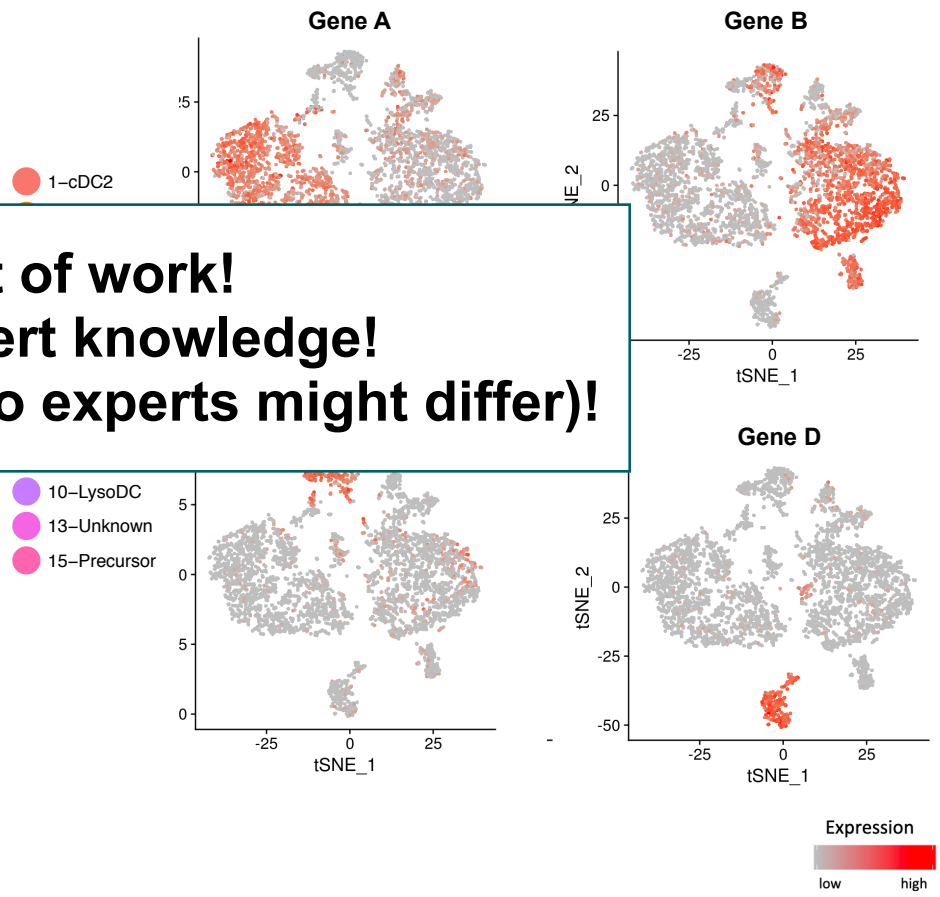
Gut Immune Cells - 12 groups



Check expression of:

1. known genes

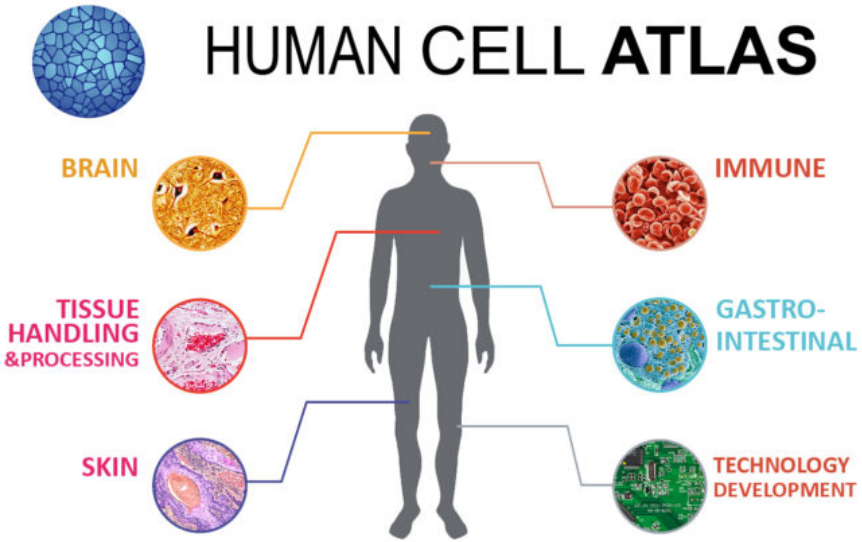
Large amount of work!
Requires expert knowledge!
Imprecise (two experts might differ)!



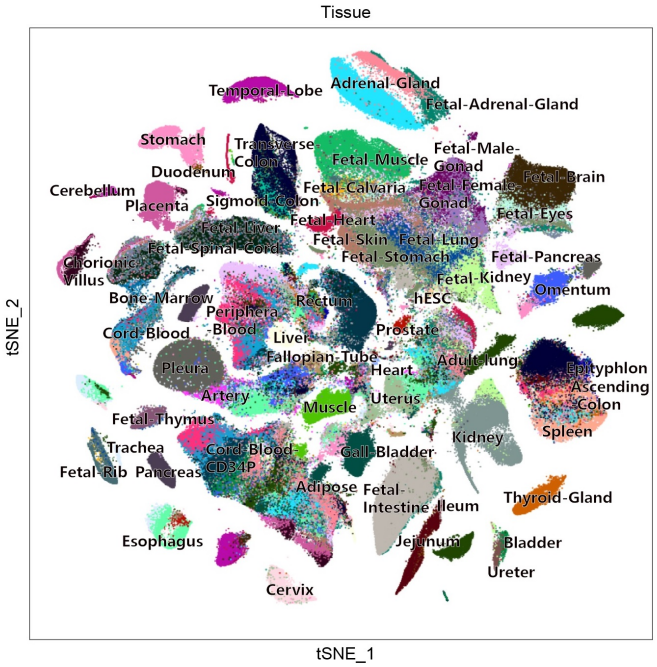
Resume / Single cell clustering

- Finding groups of single cells require complex pipeline:
  - Cell filtering
  - Normalisation
  - Artefact removal
  - Dimension reduction
  - Clustering
  - Cell annotation / visualisation
- Open points:
  - How to do dimension reduction?
  - How to detect cells of rare populations?
  - How to deal with large data sets (millions of cells)?

Clustering of cells



HCL - Human Cell Landscape



>700,000 single cells from >50 human tissue

- Open points:
  - How to do dimension reduction?
  - How to detect cells of rare populations?
  - How to deal with large data sets (millions of cells)?

Calendar

20.04.2020 –Introduction to Bioinformatics, Next Generation Sequencing and Single Cell Sequencing

28.04.2020 – Practical Course in NGS data analysis

4.05.2020 – Project Description / Introduction to HPC clusters and GPUs

11.05.2020 to 29.06.2020 – Project Development

6.07.2020 – Project Presentation

Thank you!