

# Extracting signatures from combined genomic datasets

**Carl Herrmann**

Health Data Science Unit and BioQuant

[carl.herrmann@uni-heidelberg.de](mailto:carl.herrmann@uni-heidelberg.de)

Allahabad RTG Summer School

April 2019



Medizinische Fakultät Heidelberg

# Combining and understanding multi-omics data

**Carl Herrmann**

Health Data Science Unit and BioQuant

[carl.herrmann@uni-heidelberg.de](mailto:carl.herrmann@uni-heidelberg.de)

Allahabad RTG Summer School

April 2019



Medizinische Fakultät Heidelberg

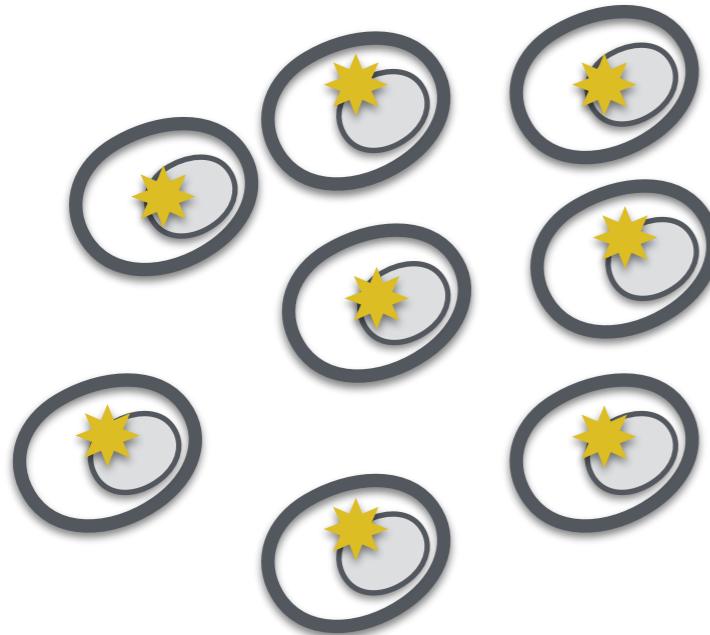


Medizinische Fakultät Heidelberg

# 1. Genomics: where the data comes from

# Identification of genomic variants

## Germline variants



**CENTOGENE**  
THE RARE DISEASE COMPANY



## Neurology

### How can genetic testing help?

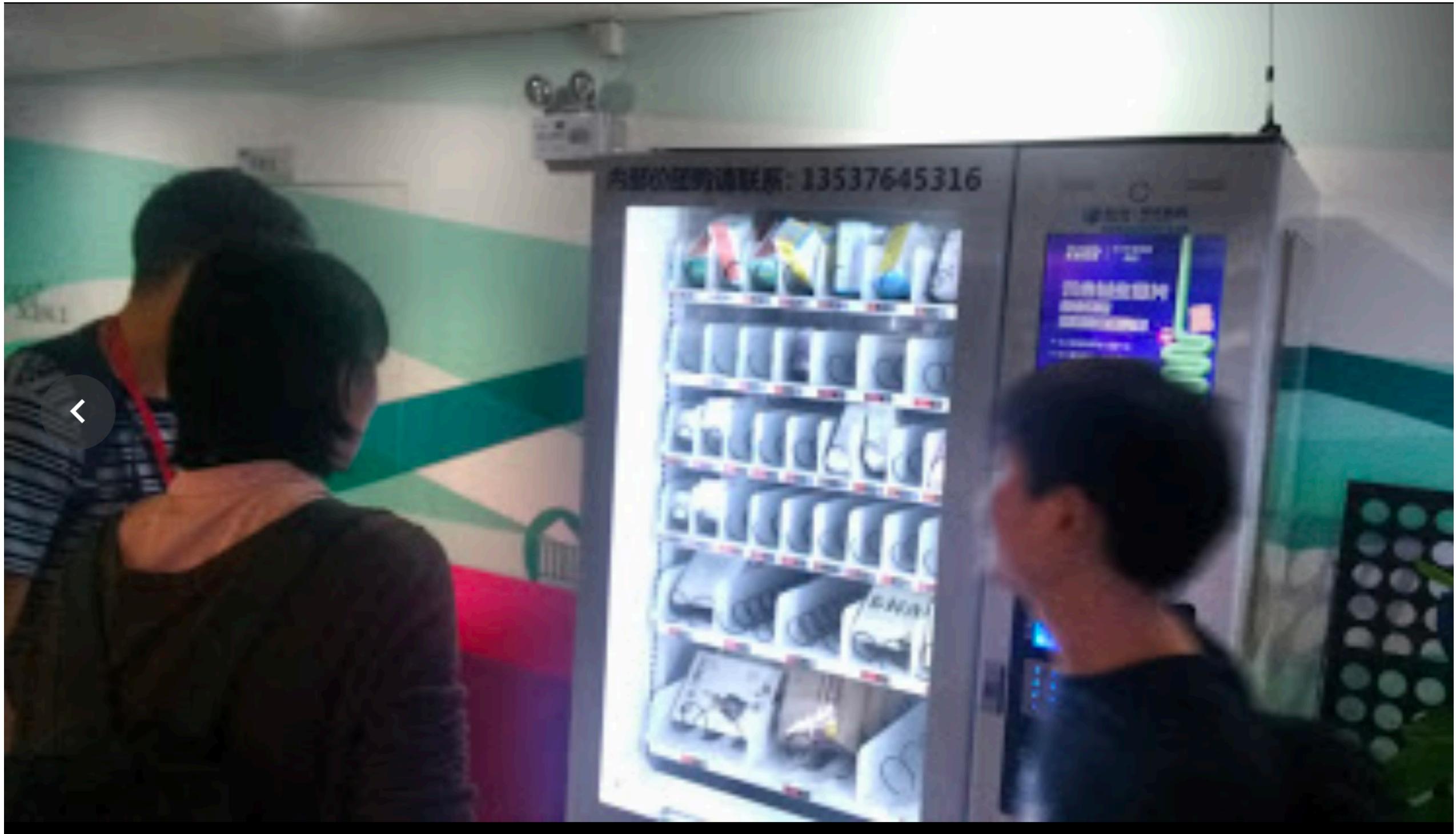
Inherited neurological conditions cover many different types of diseases and are predominantly disorders of muscle control and movement (including convulsions, poor coordination and muscle weakness), delayed mental development, degeneracy and learning disabilities.

Hereditary spastic paraparesis (HSP) is a group of inherited disorders that cause weakness and spasticity of muscles, which gradually gets worse over time. Ataxia is a group of neurological disorders that affect balance, coordination and speech. Many ataxias are inherited conditions. Symptoms typically develop slowly over many years with tremor, gait and coordination problems. The most common inherited progressive ataxia is Friedreich's ataxia.

→ risk alleles  
for hereditary diseases

# Identification of genomic variants

Medizinische Fakultät Heidelberg



# Large scale population genomics projects



- UK: 100k genome project; focus on
  - 8000 rare diseases (< 0.05%)
  - 7 main cancer types
- Iceland: small population, genetically homogeneous, with rich familial history
  - genetic information for 160K individuals (~50% of icelandic population)
- Hopes:
  - identification of rare, disease associated **germline variants**
  - Early monitoring of individuals at risk
  - Personalized treatment adapted to individual genotype

# Identifying risk variants



Medizinische Fakultät Heidelberg

**nature genetics**

Letter | Published: 29 October 2018

Meta-analysis of Icelandic and UK data sets identifies missense variants in SMO, IL11, COL11A1 and 13 more new loci associated with osteoarthritis

Unnur Styrkarsdottir ✉, Sigrun H. Lund, [...] Kari Stefansson ✉

Nature Genetics 50, 1681–1687 (2018) | Download Citation ↴

**nature COMMUNICATIONS**

Article | OPEN | Published: 08 November 2018

Genome-wide associations for benign prostatic hyperplasia reveal a genetic correlation with serum levels of PSA

Julius Gudmundsson ✉, Jon K. Sigurdsson, [...] Kari Stefansson ✉

Nature Communications 9, Article number: 4568 (2018) | Download Citation ↴

**nature COMMUNICATIONS**

Article | OPEN | Published: 20 March 2019

Sequence variation at ANAPC1 accounts for 24% of the variability in corneal endothelial cell density

Erna V. Ivarsdottir, Stefania Benonisdottir, Gudmar Thorleifsson, Patrick Sulem, Asmundur Oddsson, Unnur Styrkarsdottir, Snaedis Kristmundsdottir, Gudny A. Arnadottir, Gudmundur Thorgeirsson, Ingileif Jonsdottir, Gunnar M. Zoega, Unnur Thorsteinsdottir, Daniel F. Gudbjartsson, Fridbert Jonasson, Hilma Holm ✉ & Kari Stefansson ✉

Nature Communications 10, Article number: 1284 (2019) | Download Citation ↴

**Sequence variants associating with urinary biomarkers** ⓘ

Stefania Benonisdottir, Ragnar P Kristjansson, Asmundur Oddsson, Valgerdur Steinhorsdottir, Evgenia Mikaelsdottir, Birte Kehr, Brynjar O Jensson, Gudny A Arnadottir, Gerald Sulem, Gardar Sveinbjornsson, ... Show more

Author Notes

Human Molecular Genetics, Volume 28, Issue 7, 1 April 2019, Pages 1199–1211, <https://doi.org/10.1093/hmg/ddy409>

Published: 24 November 2018 Article history ▾

**nature genetics**

Article | Published: 14 January 2019

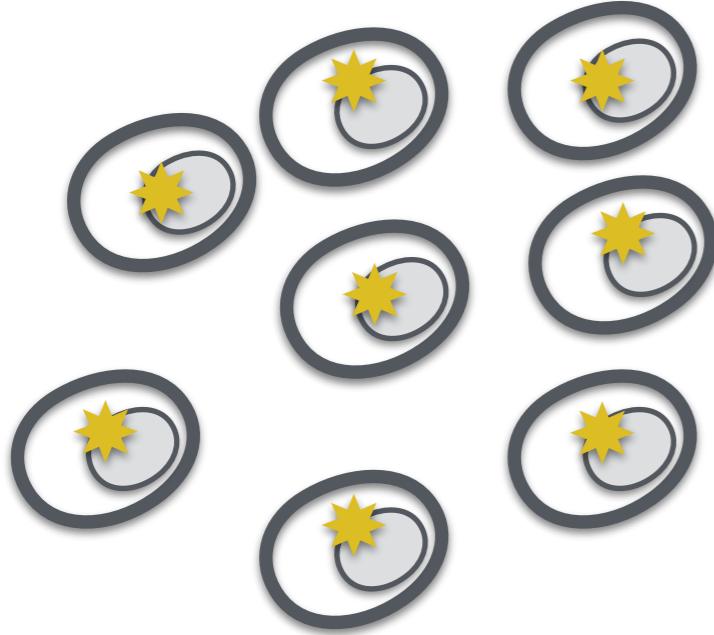
A loss-of-function variant in ALOX15 protects against nasal polyps and chronic rhinosinusitis

Ragnar P. Kristjansson, Stefania Benonisdottir, [...] Kari Stefansson ✉

Nature Genetics 51, 267–276 (2019) | Download Citation ↴

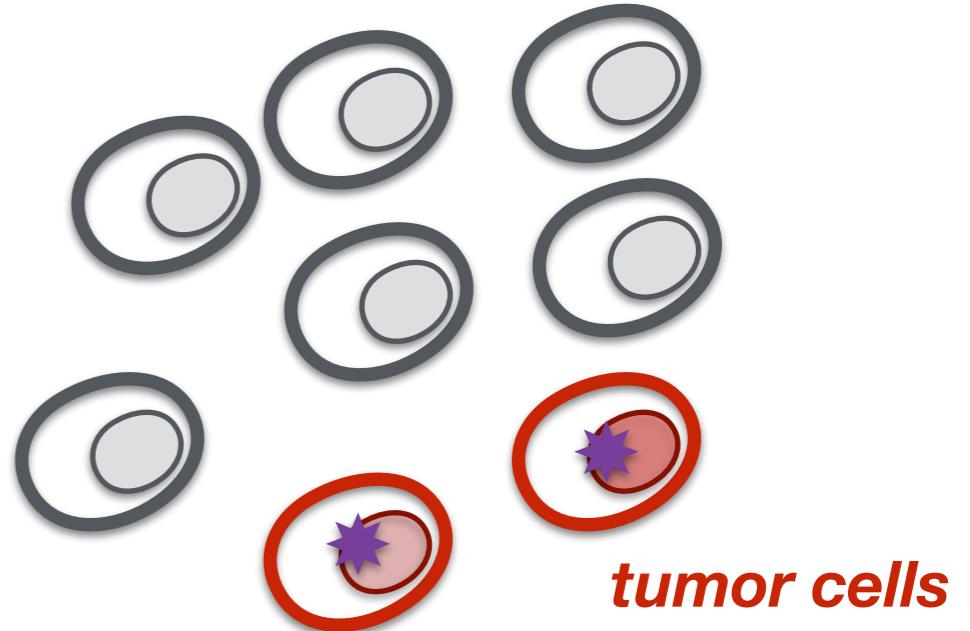
# Identification of genomic variants

## Germline variants



→ risk alleles  
for hereditary diseases

## Somatic variants

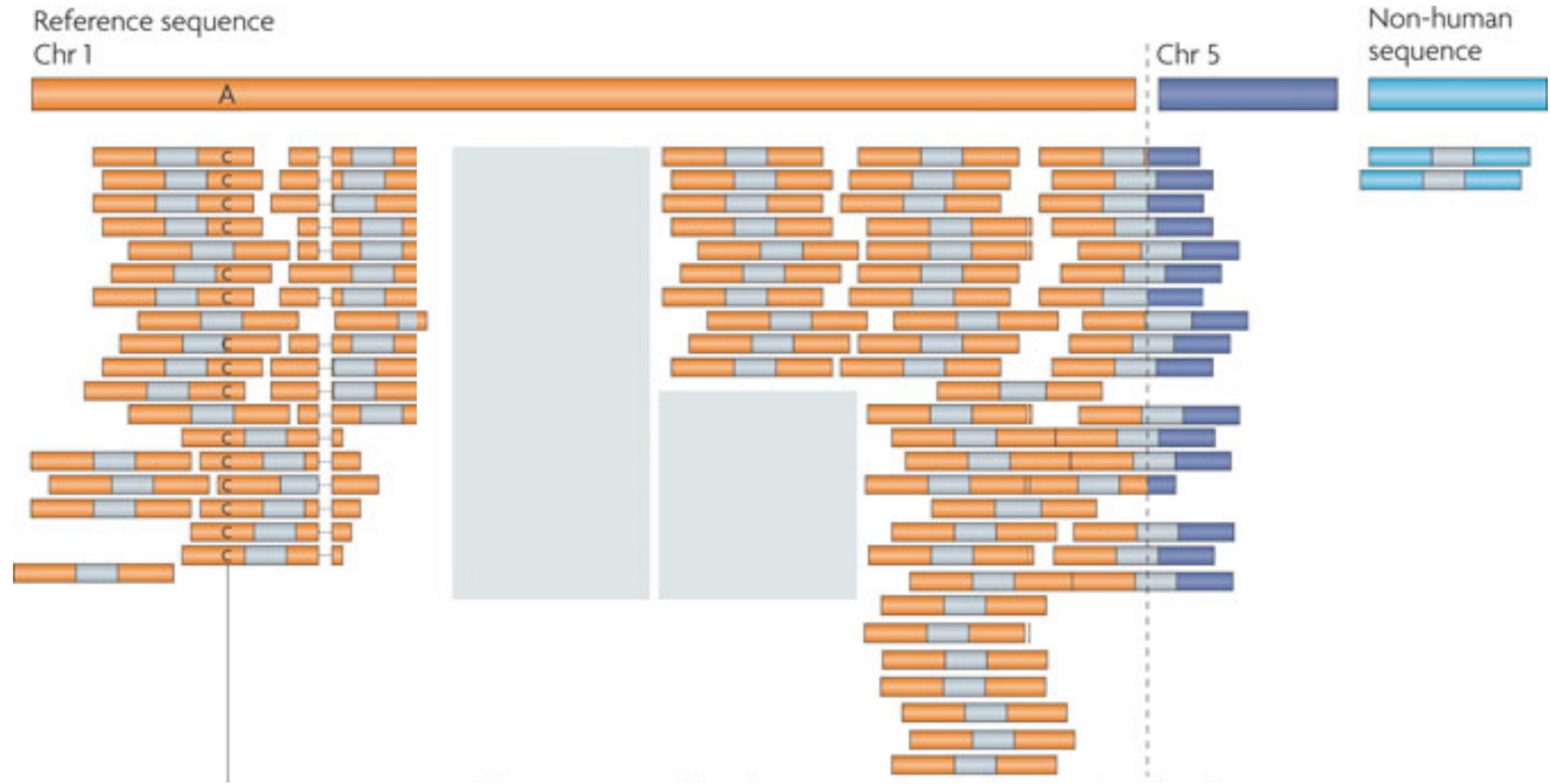


→ oncogenic  
driver mutations

# Cancer genomics



Medizinische Fakultät Heidelberg



Point mutations  
/ indels

copy number  
alterations  
(deletion / LOH / amplification)

structural variant  
viral insertion

# Variant in cancer genomes

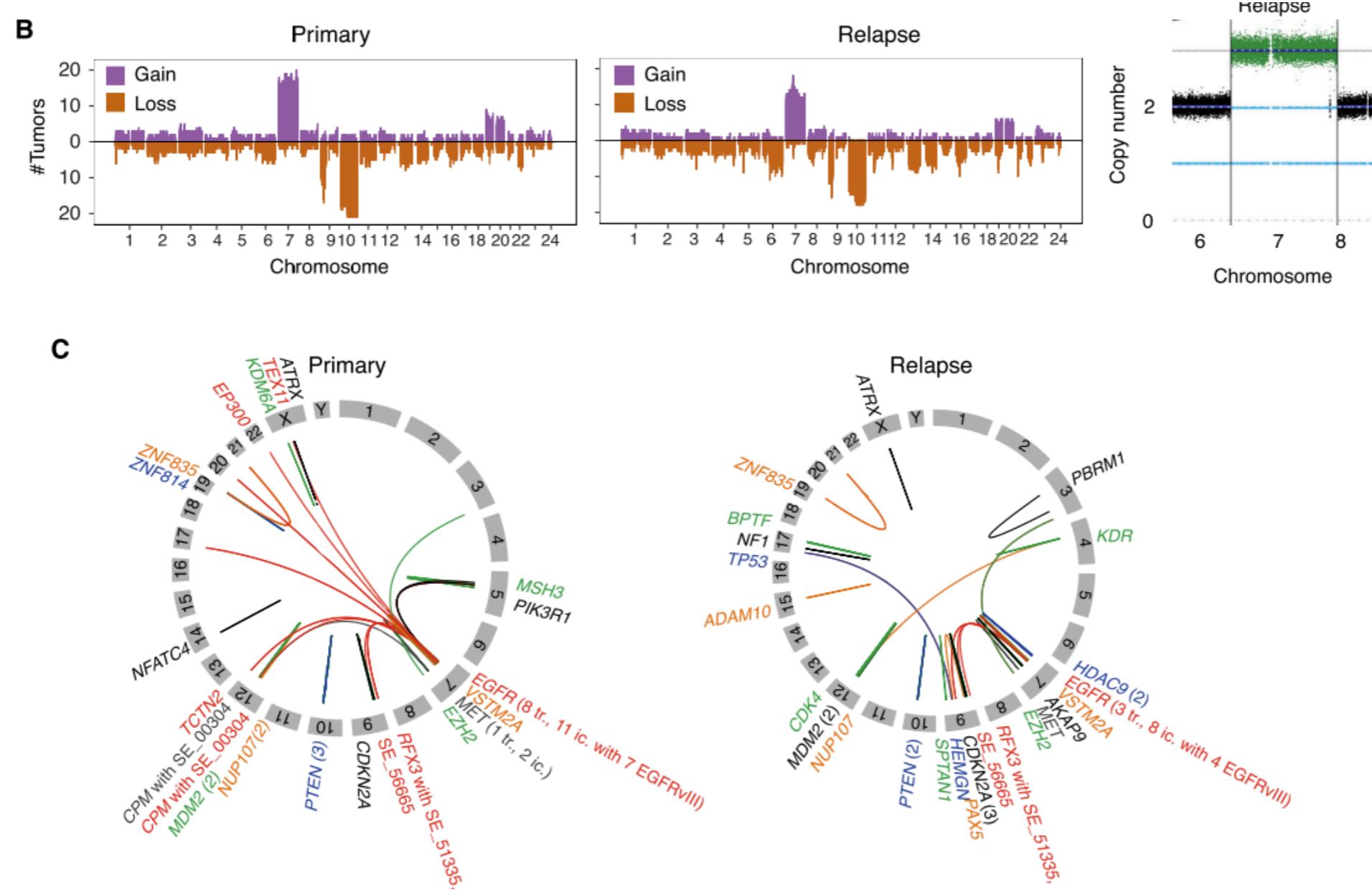


Medizinische Fakultät Heidelberg

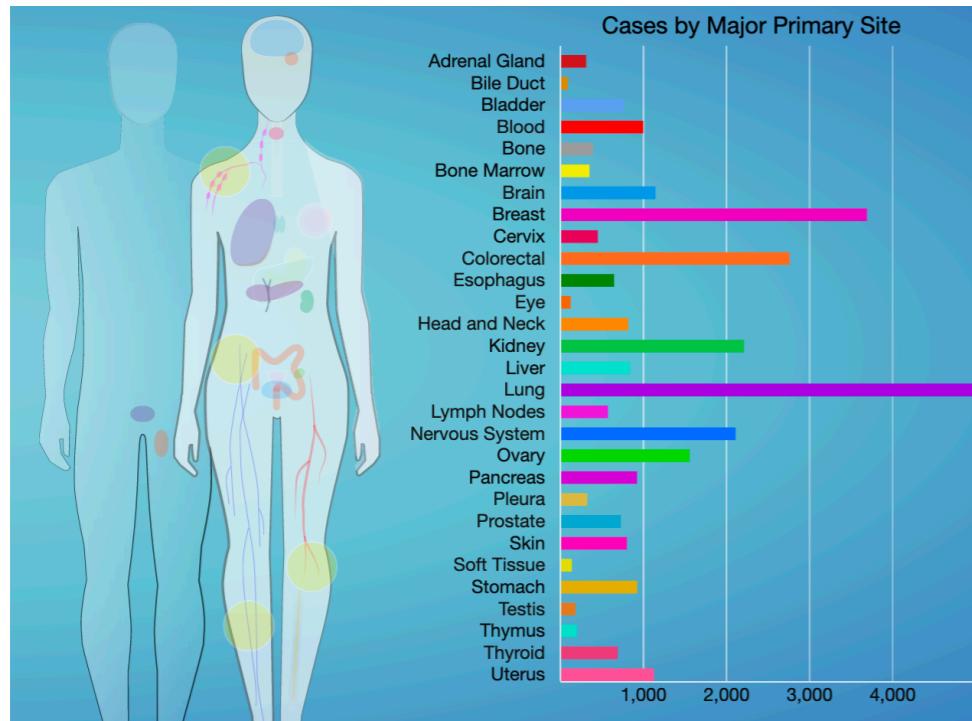


# Structural variants in cancer genomes

## Glioblastoma patients



# The Cancer Genome Analysis (TCGA)



## Data Portal Summary

[Data Release 16.0 - March 26, 2019](#)

### PROJECTS



45

### PRIMARY SITES



68

### FILES



365,463

### GENES



22,872

TCGA produced over

**2.5**  
PETABYTES  
of data

### CASES



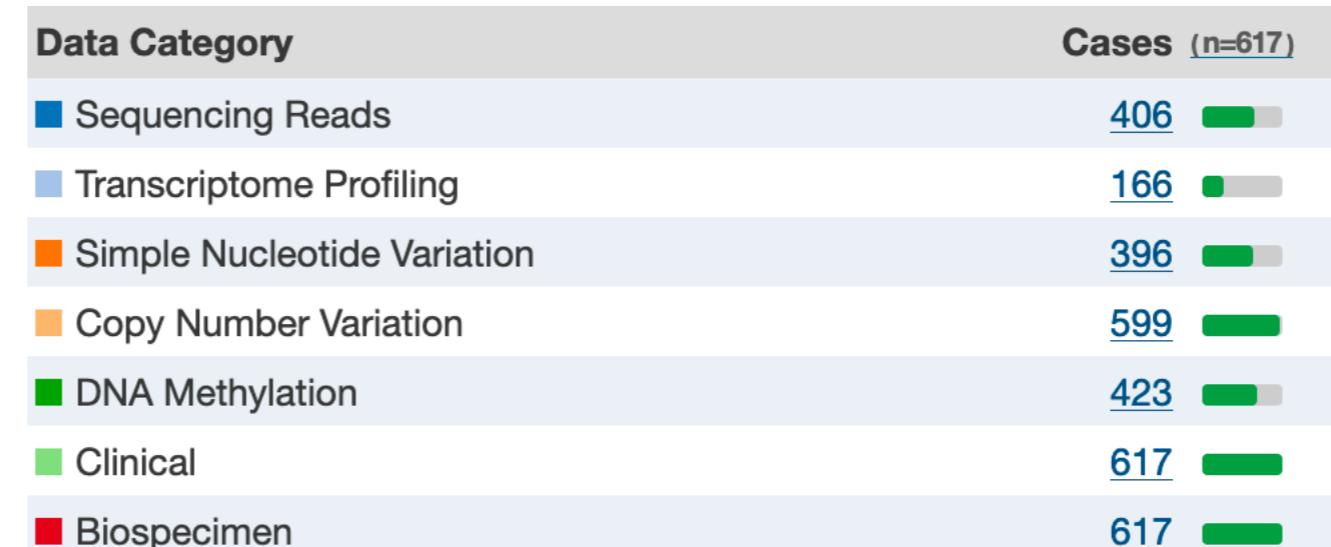
33,549

### MUTATIONS



3,142,246

Example of  
data available  
for glioblastoma

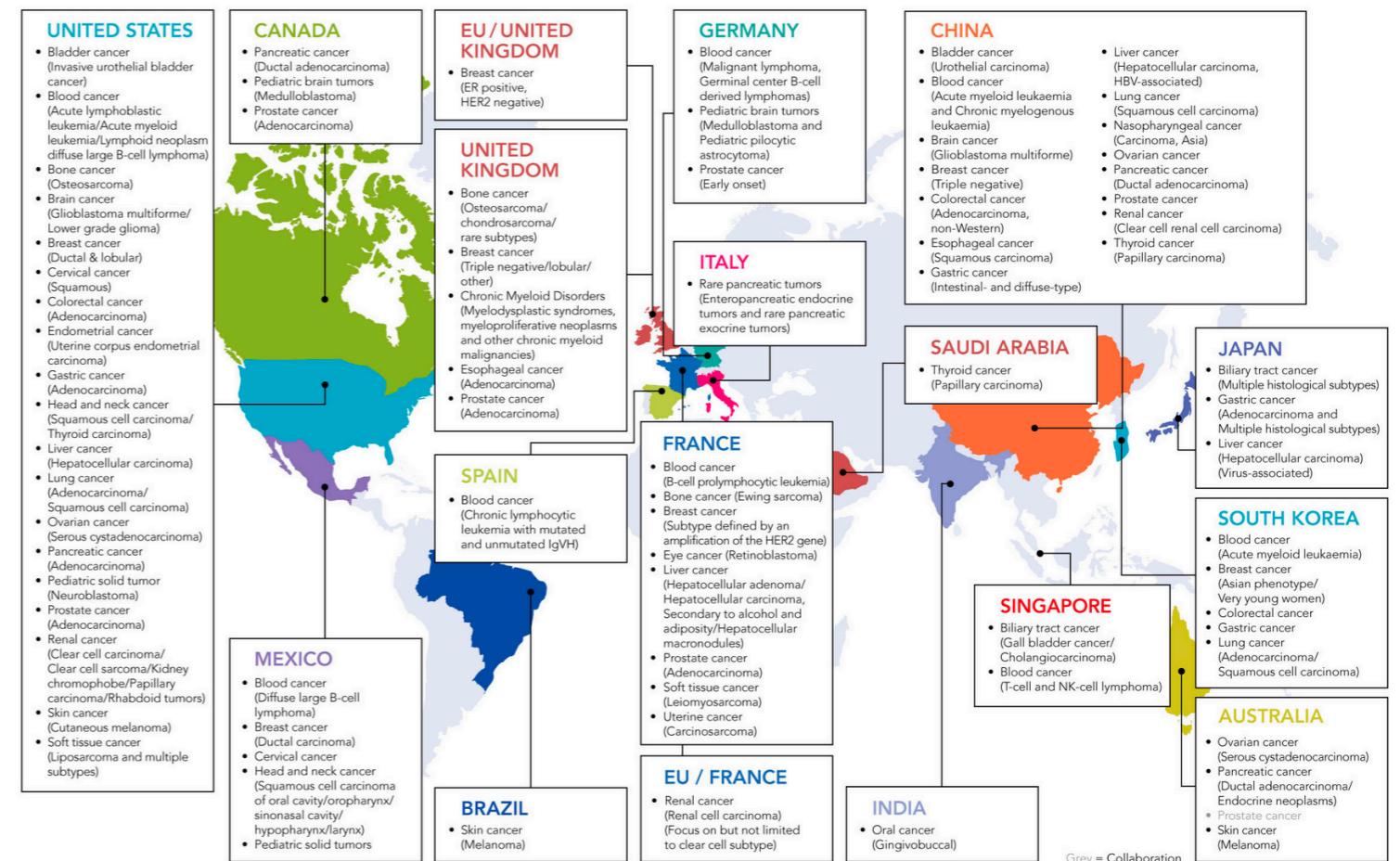


# International Cancer Genome Consortium (ICGC)

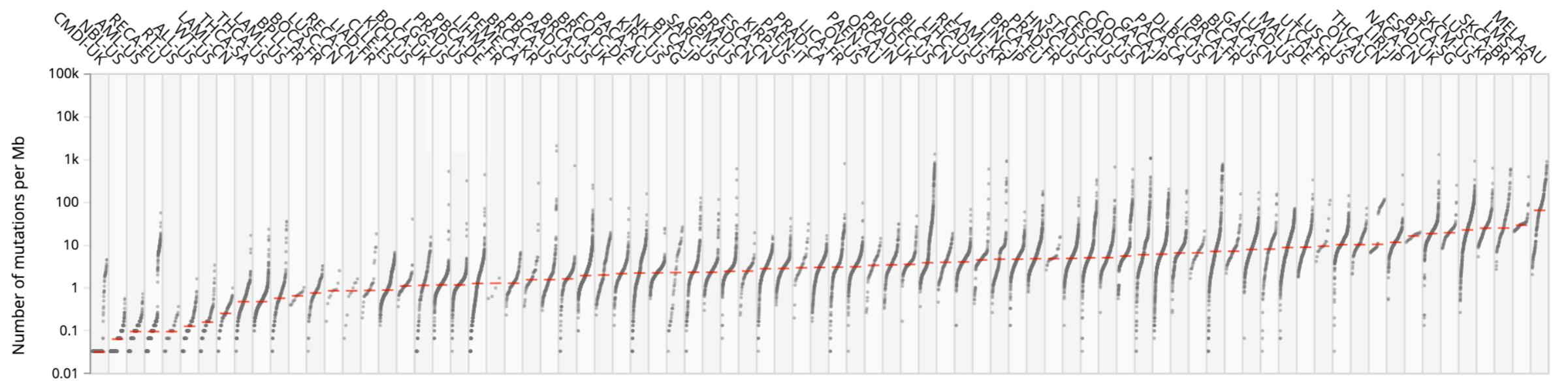


Medizinische Fakultät Heidelberg

- international cancer genome sequencing effort
- 90 genome sequencing projects involving 18 countries



Number of Somatic Mutations in Donor's Exomes Across Cancer Projects



# International Cancer Genome Consortium (ICGC)



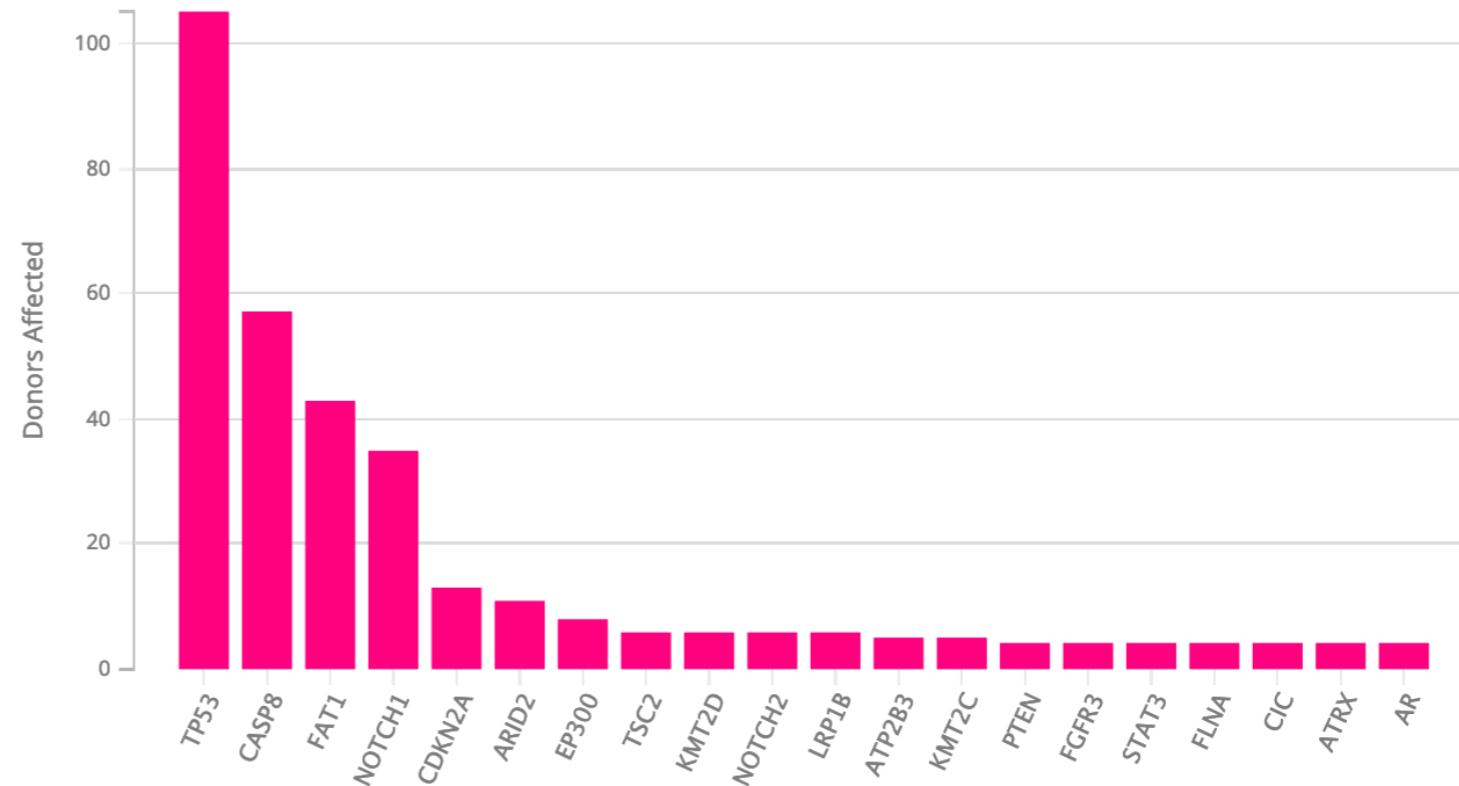
- international cancer genome sequencing effort
- India: oral cancer

## Project Summary

In view of its high prevalence in India and the existence of possible interacting environmental factors, India will focus on oral cancer as a part of the International Cancer Genome Consortium activities. In particular, India will focus on gingivo-buccal cancer. The primary reasons for selection of gingivo-buccal cancer are:

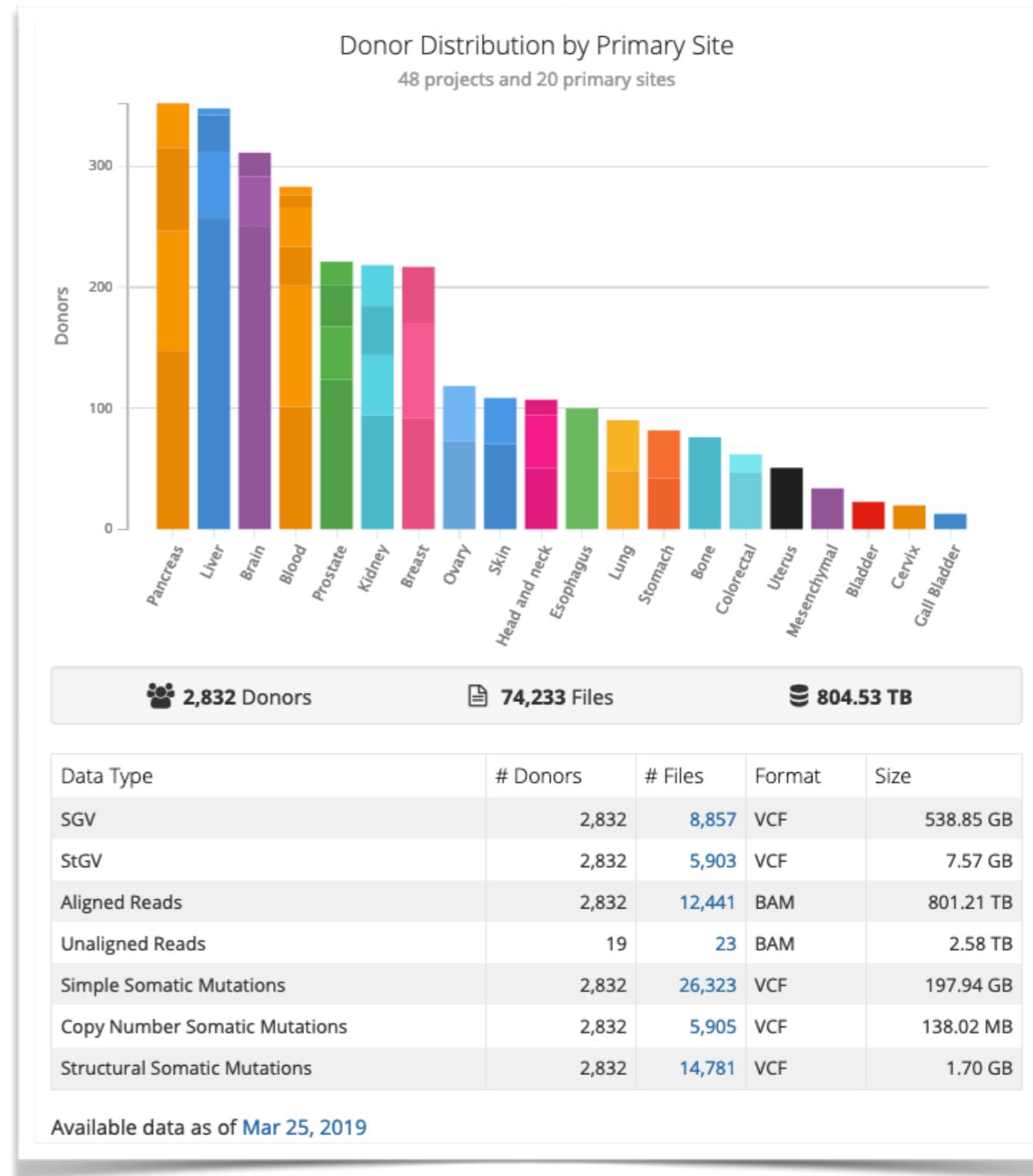
- It is the most common cancer among men in India and has a very high age-adjusted risk
- The site distribution of oral cancer in India is very different from that found in western countries
- While over 70% of tongue cancers have the propensity to metastasize, less than 50% of gingivo-buccal cancers are found to metastasize
- Leukoplakia and sub-mucous fibrosis are two types of pre-cancerous lesion that can help track temporal genomic changes in gingivo-buccal cancer.

Top 20 Mutated Cancer Genes with High Functional Impact SSMs  
178 Unique SSM-Tested Donors



# PanCancer Analysis of Whole Genomes (PCAWG)

- Coordinated analysis of whole-genome sequencing datasets (WGS)
- analysis of mutations
  - in genic regions
  - in regulatory regions
- Interaction between genome and epigenome
- Molecular subtyping
- Check TCGA PanCancer papers  
<https://www.cell.com/pb-assets/consortium/PanCancerAtlas/PanCani3/index.html>



# Genomics is Big Data



Medizinische Fakultät Heidelberg

## Data Phase

Acquisition

Storage

Analysis

Distribution

doi:10.1371/journal.pbio.1002195.t001

[Stephens, PLOS Bio. 2015]

**Petabyte (1 000 000 000 000 000 Bytes)**

**Exabyte (1 000 000 000 000 000 000 Bytes)**

**Zettabyte (1 000 000 000 000 000 000 000 Bytes)**

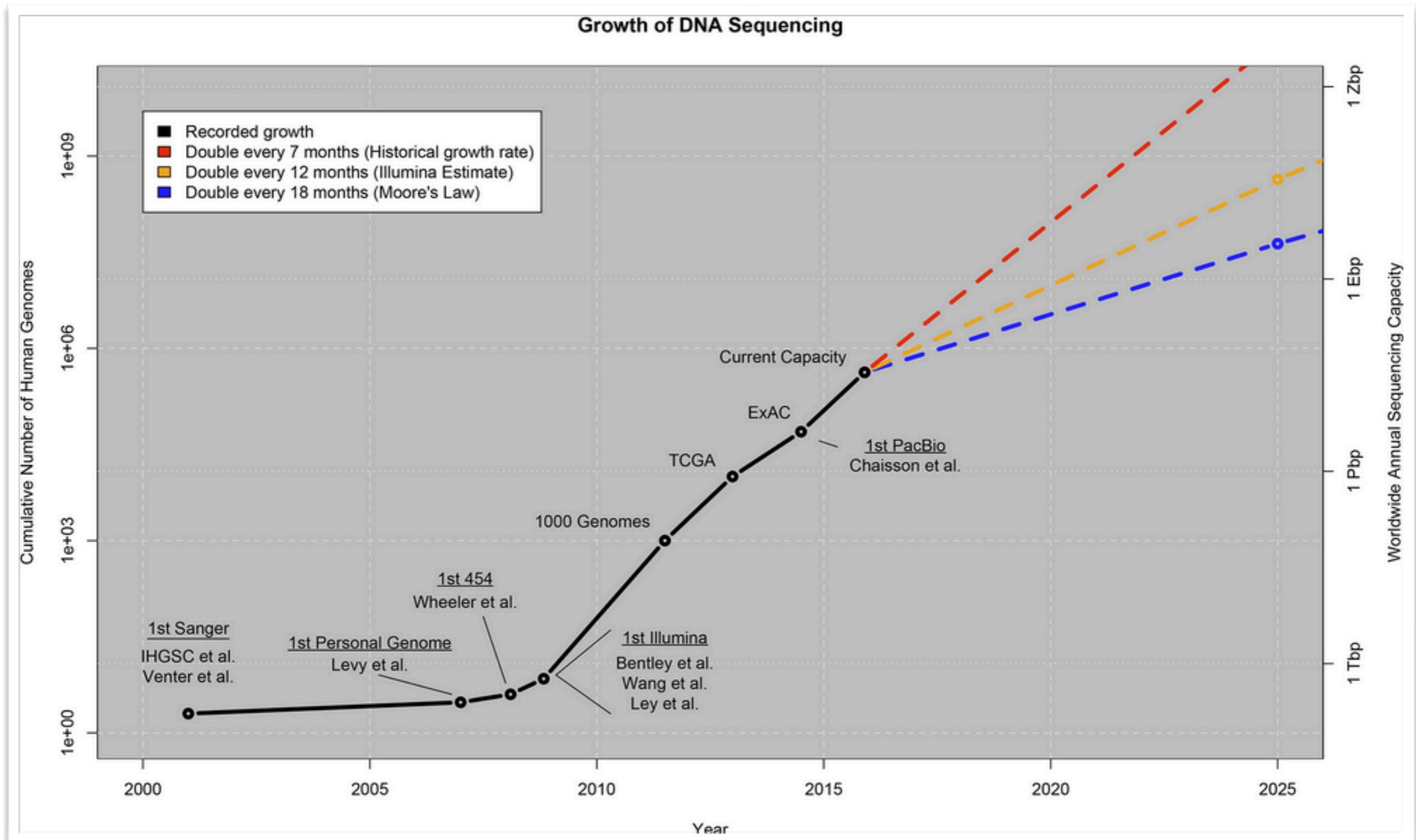
all words ever  
spoken by humans = 5  
Exabyte (text)

all words ever spoken  
by humans = 42 Zettabyte  
(audio)

# Genomics is Big Data



Medizinische Fakultät Heidelberg



[Stephens, PLOS Bio. 2015]

# Genome sequencing



Medizinische Fakultät Heidelberg



Book with  
4.3 million pages  
text written in Arial 12

Illumina X10 sequencer  
1.8 Terabases in 3 Days  
18,000 Genomes / Year

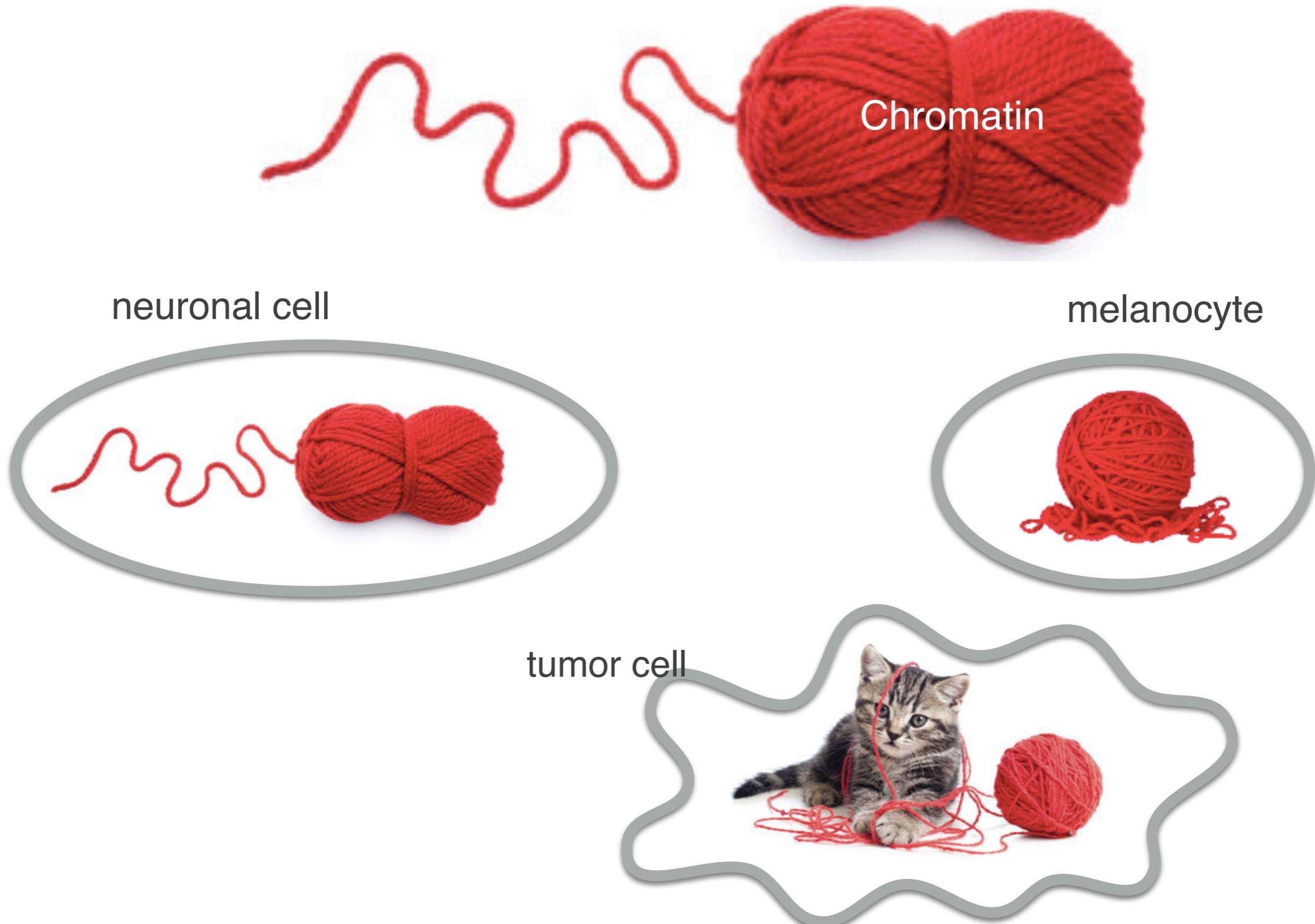
```
@SRR540192.1580 IL34_5480:5:1:6425:1038/1
CATCTTGGCCTCTGTGCAGCATTCTTCATGGT
+
IIIIIIHIIIIIIHIIIIIIIIIIIIIIHIIHID
@SRR540192.1752 IL34_5480:5:1:7005:1052/1
GCTCCCAGAAACCCAGGGCCACTGGCAGCTTCAGGGA
+
GGGGGGGBG@GGGB@>D<GGGF@<?<?9??; (?::2(
@SRR540192.1788 IL34_5480:5:1:10167:1053/1
ATGGGCTTCTCCGGCTTCAGCCACCTGCGCCCTGC
+
GG@G>G@E3<B=B; B<E>EDEAAAB:B.:=>A?;A8D
@SRR540192.2271 IL34_5480:5:1:5889:1093/1
TGATCATCTGGCTGATGCCGGTGAUTGCCACCCCTTGAG
+
IIIGIIIIIIIIIDIIIGIIIHGHHHIIIIHID
```

## 2. Beyond Genome Sequences

# Measuring Chromatin state



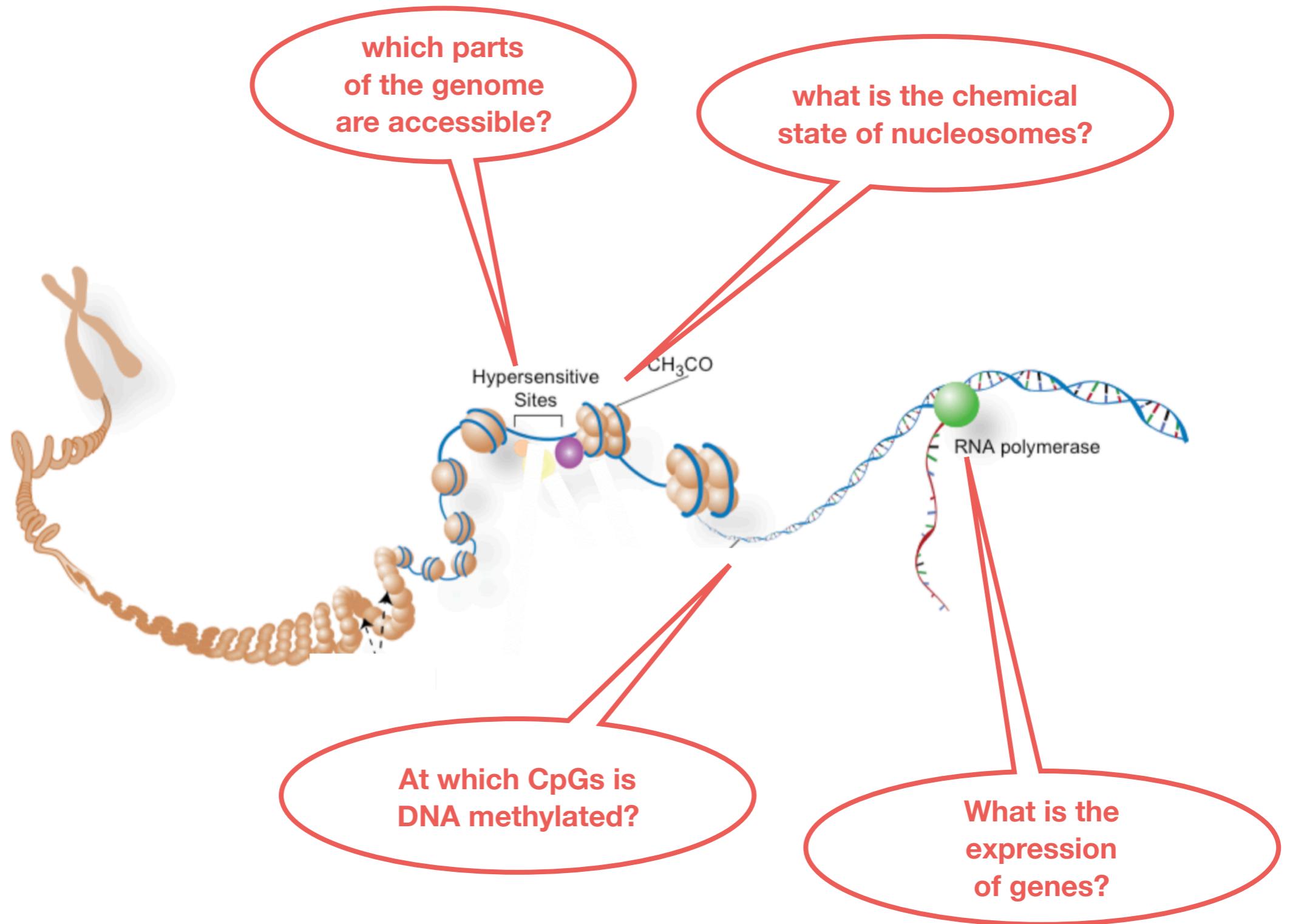
Medizinische Fakultät Heidelberg



# Regulatory genome



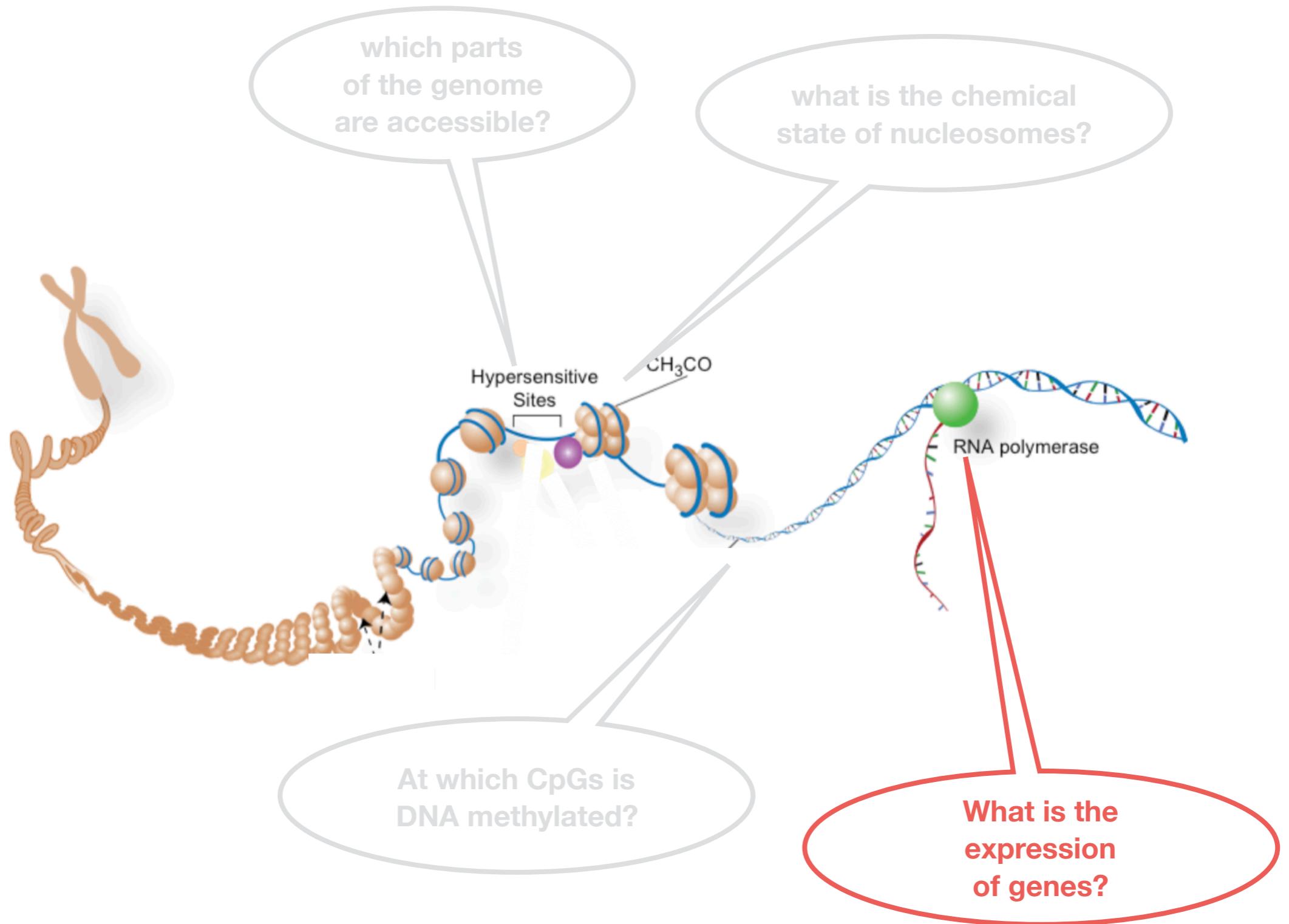
Medizinische Fakultät Heidelberg



# Regulatory genome



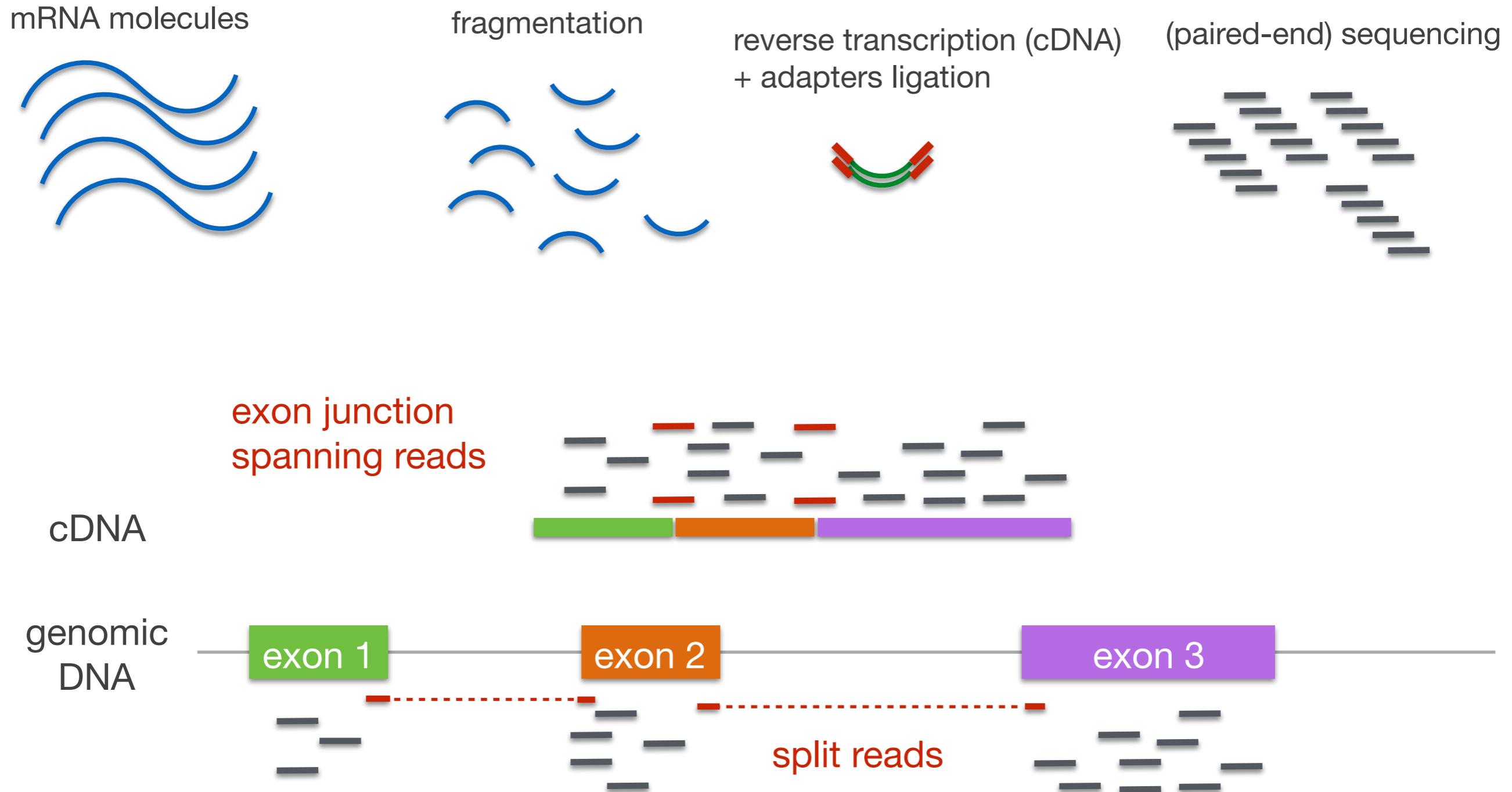
Medizinische Fakultät Heidelberg



# Measuring gene expression



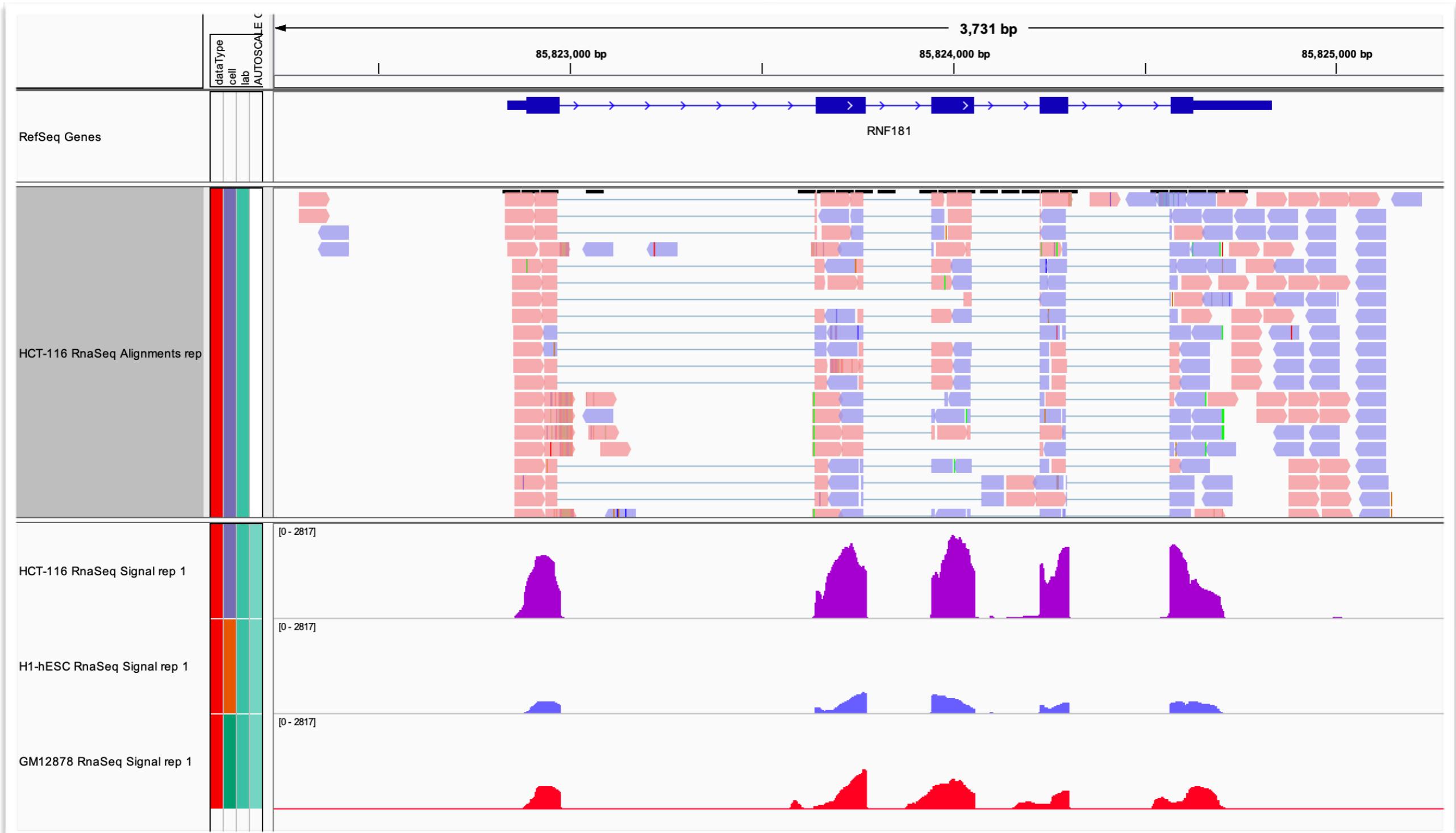
Medizinische Fakultät Heidelberg



# Measuring gene expression



Medizinische Fakultät Heidelberg



# Measuring gene expression



- $n$  = counts of reads mapping to a transcript
- $L$  = length of the transcript in **kilobase** 
- $N$  = total size of the sequenced library  
(i.e. total number of **mapped** reads, **in million**)

**Reads per kilobase per million mapped reads**

$$RPKM_i = \frac{n_i}{L_i} \times \frac{1}{N}$$

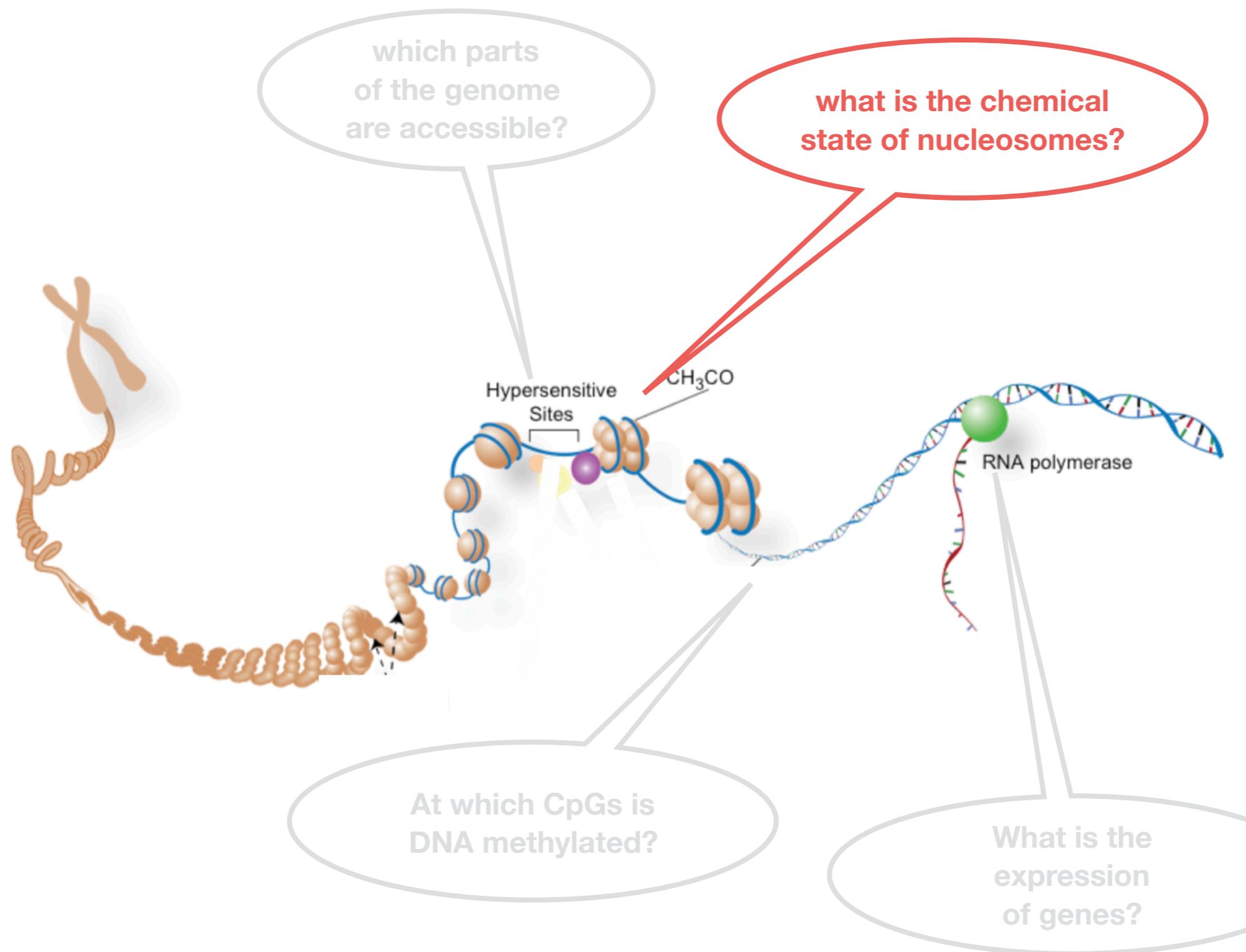
**Transcripts per million**

$$TPM_i = \frac{n_i}{L_i} \times \frac{10^6}{\sum_j \left( \frac{n_j}{L_j} \right)}$$

# Regulatory genome



Medizinische Fakultät Heidelberg

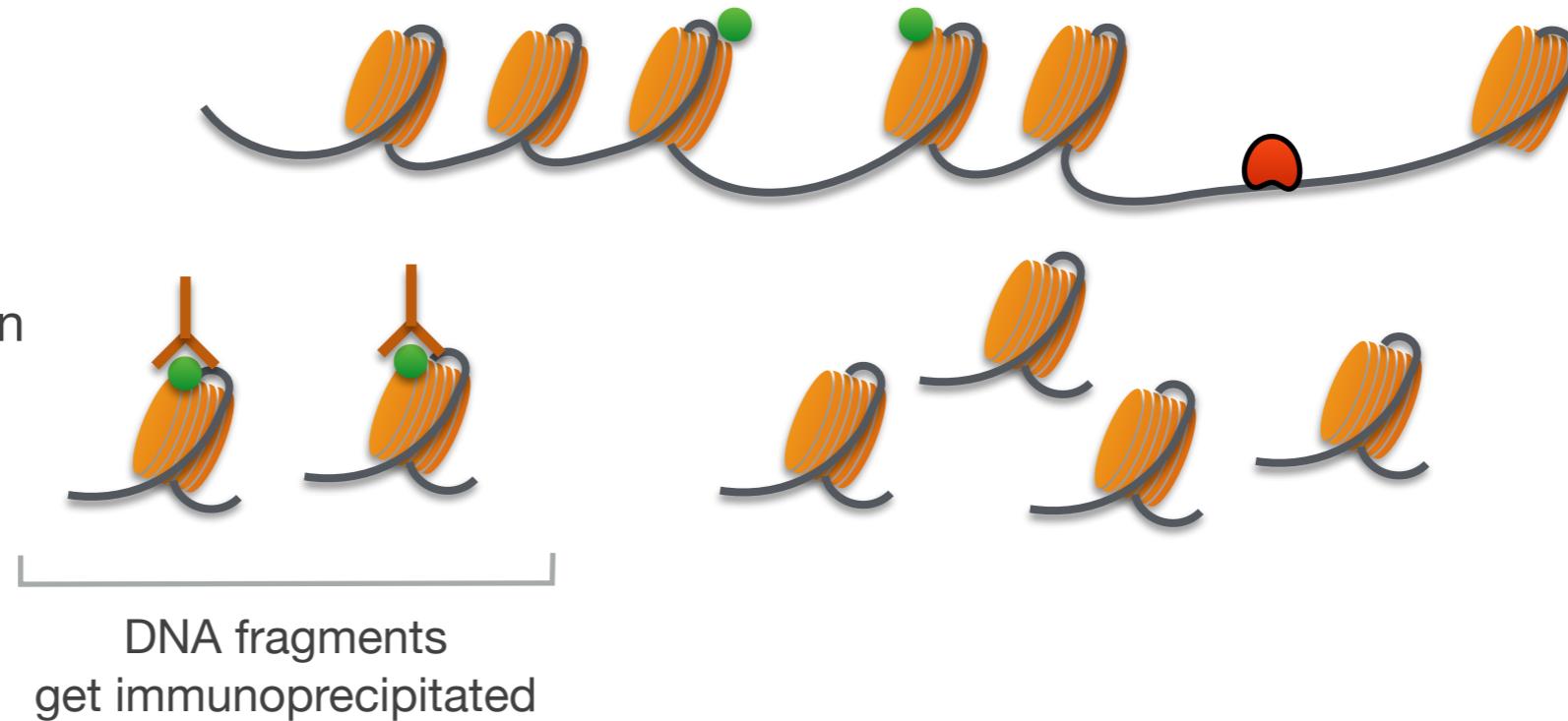


# ChIP-sequencing



Medizinische Fakultät Heidelberg

fragmentation  
(sonication, mnase)  
and immunoprecipitation

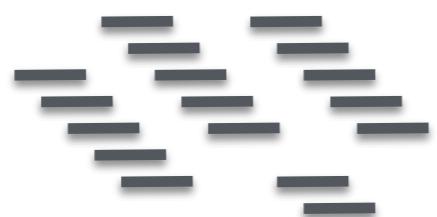


Library preparation  
(PCR,  
sequencing adapters)



Read alignment

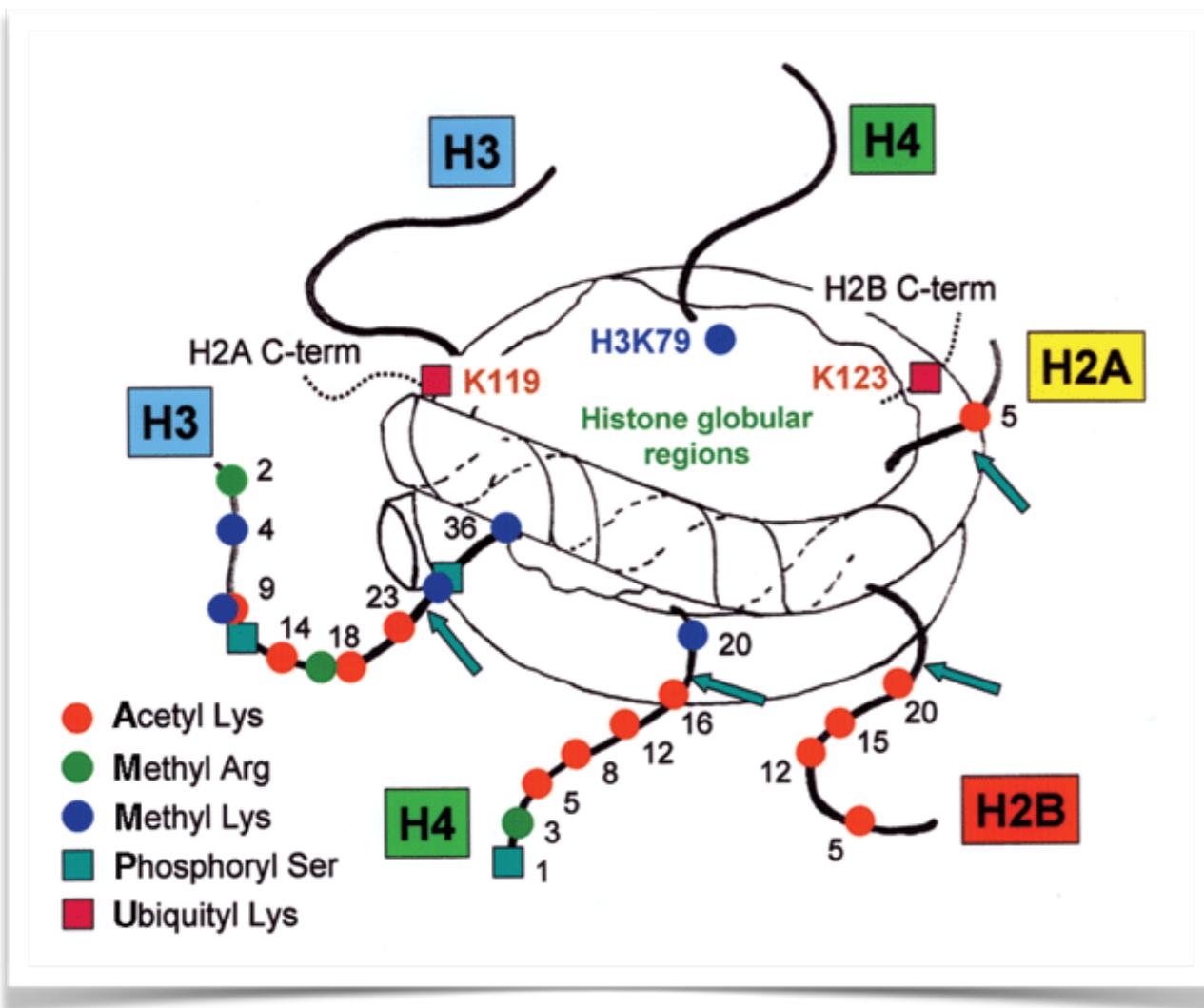
High-throughput  
sequencing



Peak calling



# ChIP-seq for histone modifications



- histones are subject to **post-translational modifications** at their N-terminal tail
  - Lysine methylation
  - Lysine/arginine acetylation
  - Serine phosphorylation
  - ubiquitylation
- they **modify the physical properties of the DNA-nucleosome interactions**

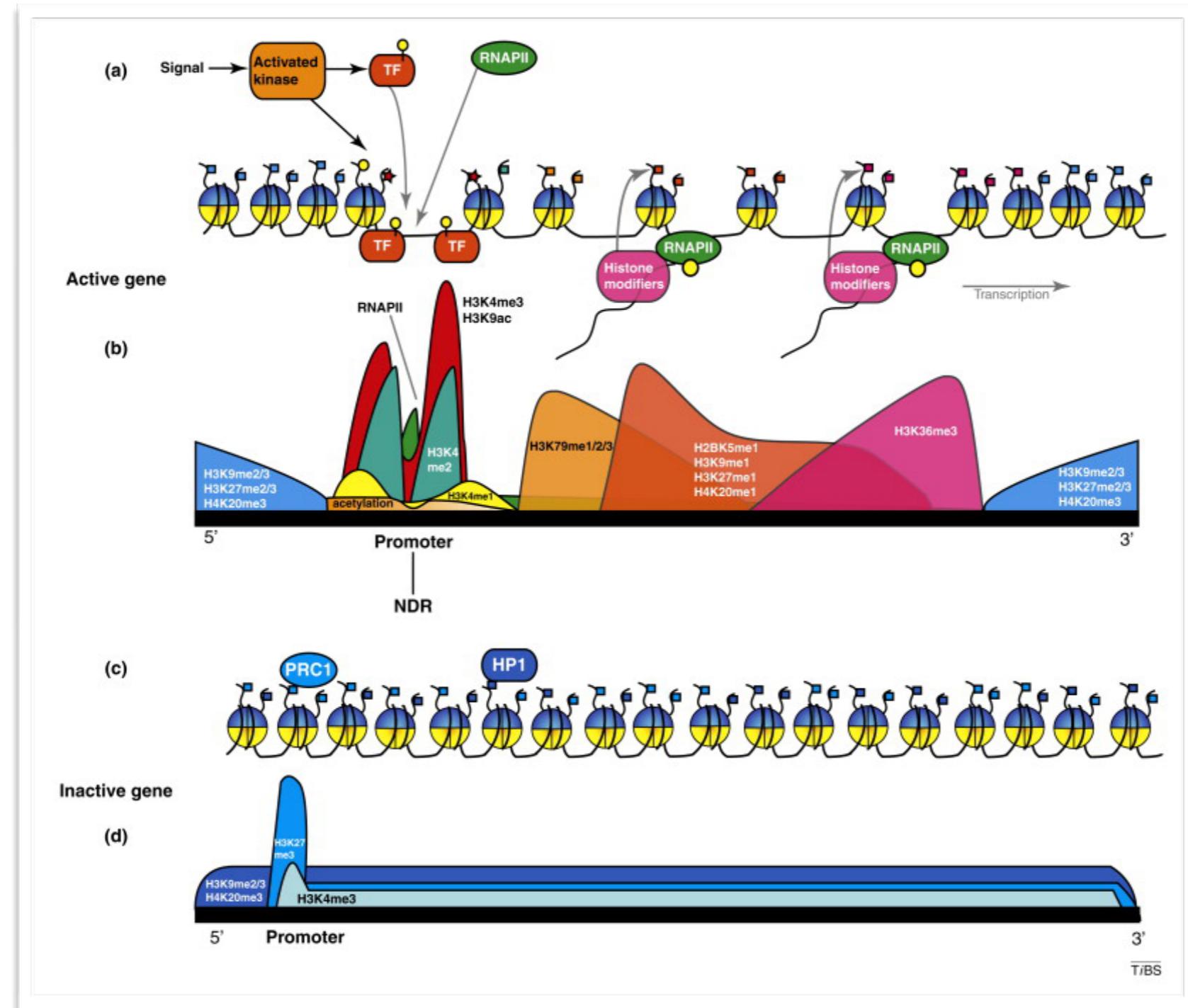
nomenclature: H3K27ac = acetylation of lysine 27 on histone 3

# Histone modifications



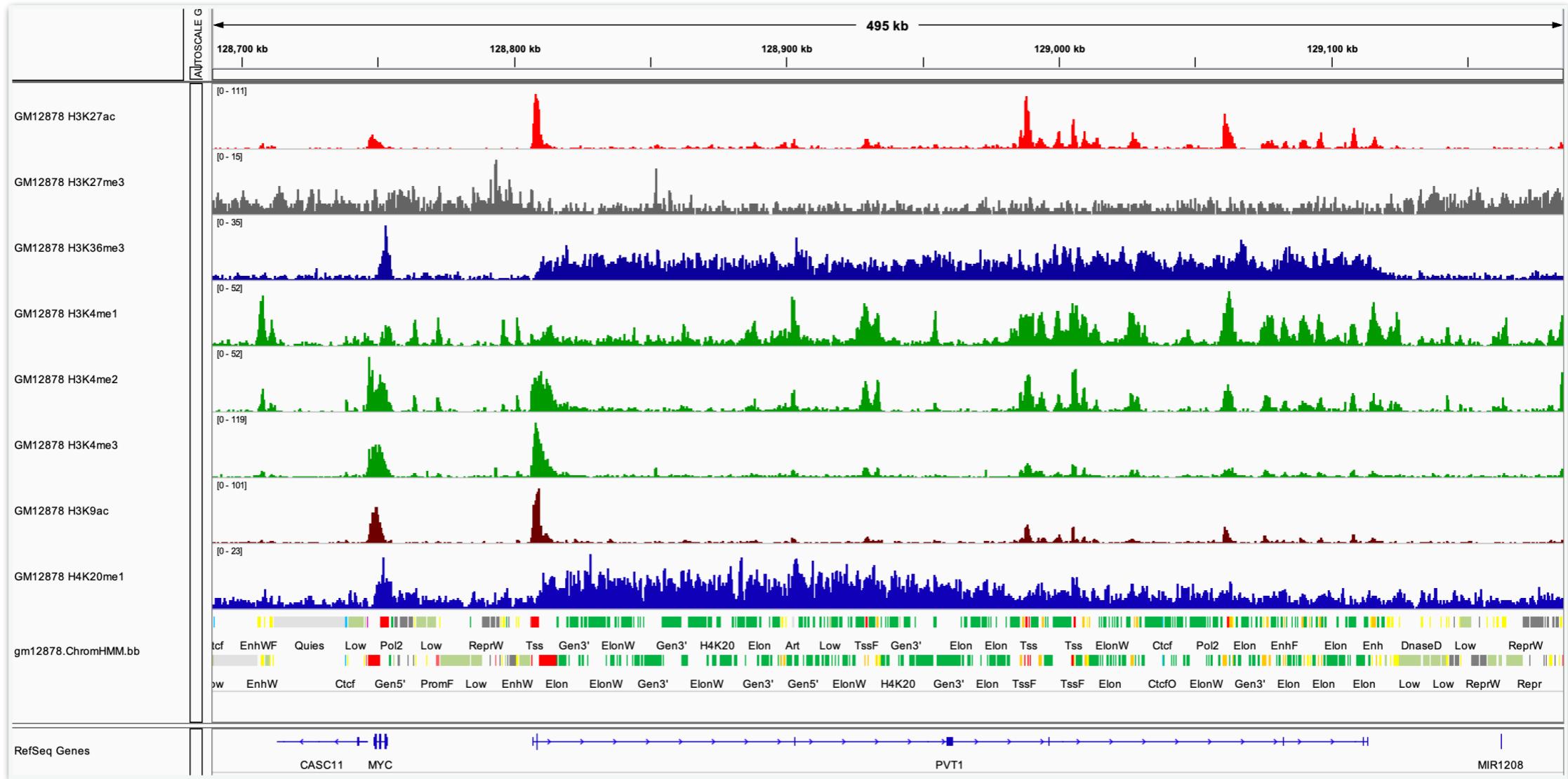
Medizinische Fakultät Heidelberg

histone modifications are a good proxy of gene expression and presence of regulatory elements



**active marks**  
→ open chromatin  
H3K4me1; H3K4me3;  
H3K27ac

**repressive marks**  
→ closed chromatin  
H3K27me3; H3K9me3



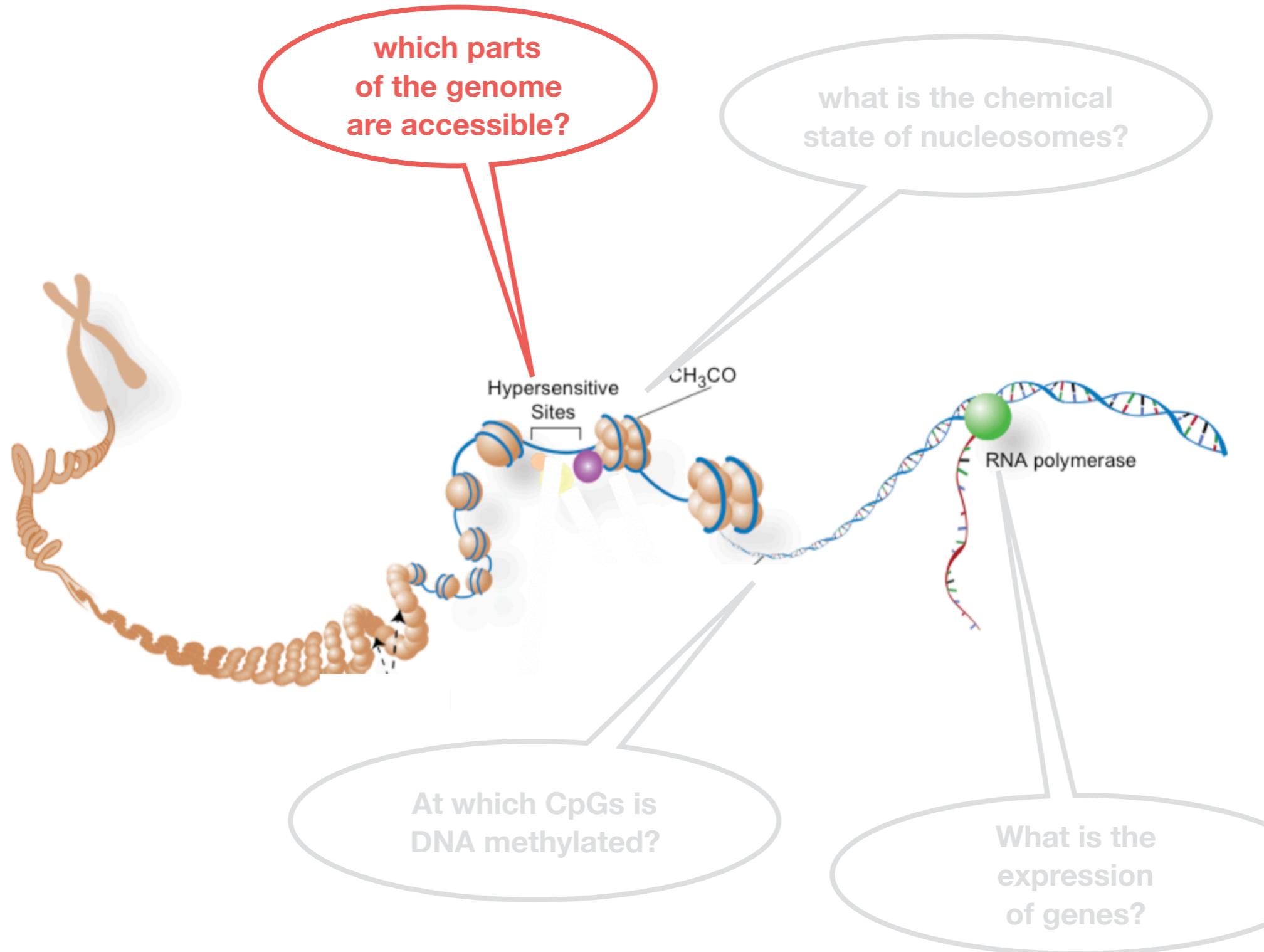
- Histone marks have **distinct signal profiles**

- sharp signal :
  - ▶ H3K4me3 = promoters
  - ▶ H3K27ac = enhancers, ...
- broad signal :
  - ▶ H3K36me3 = transcribed genes
  - ▶ H3K27me3 = repressed regions)

# Regulatory genome



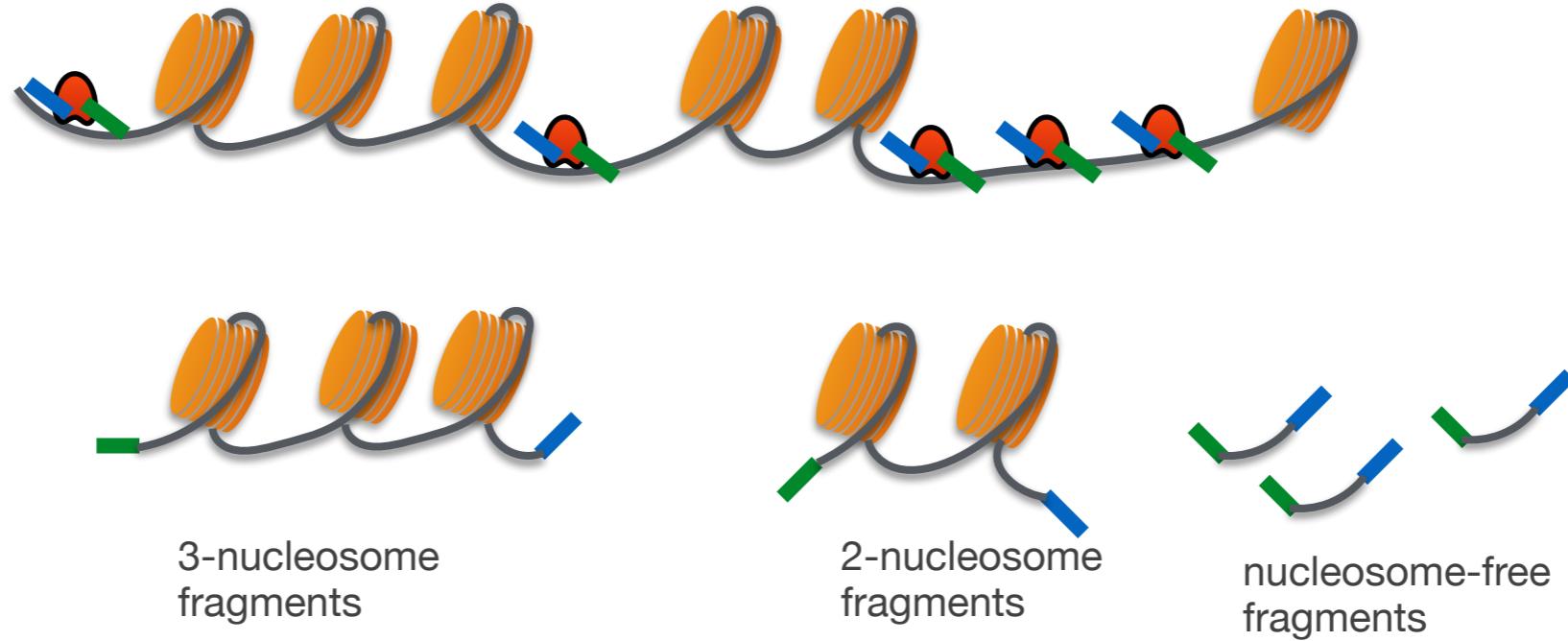
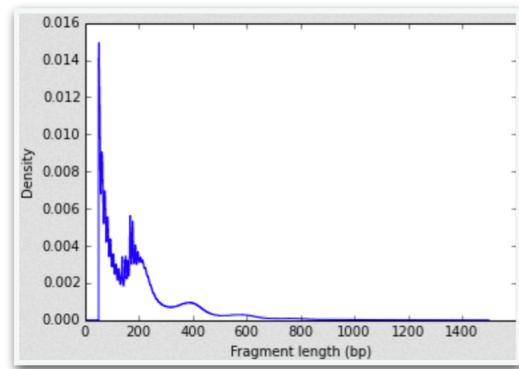
Medizinische Fakultät Heidelberg



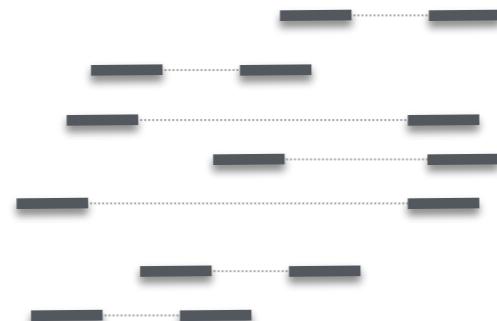
# ATAC-sequencing



Medizinische Fakultät Heidelberg



High-throughput  
paired-end  
sequencing

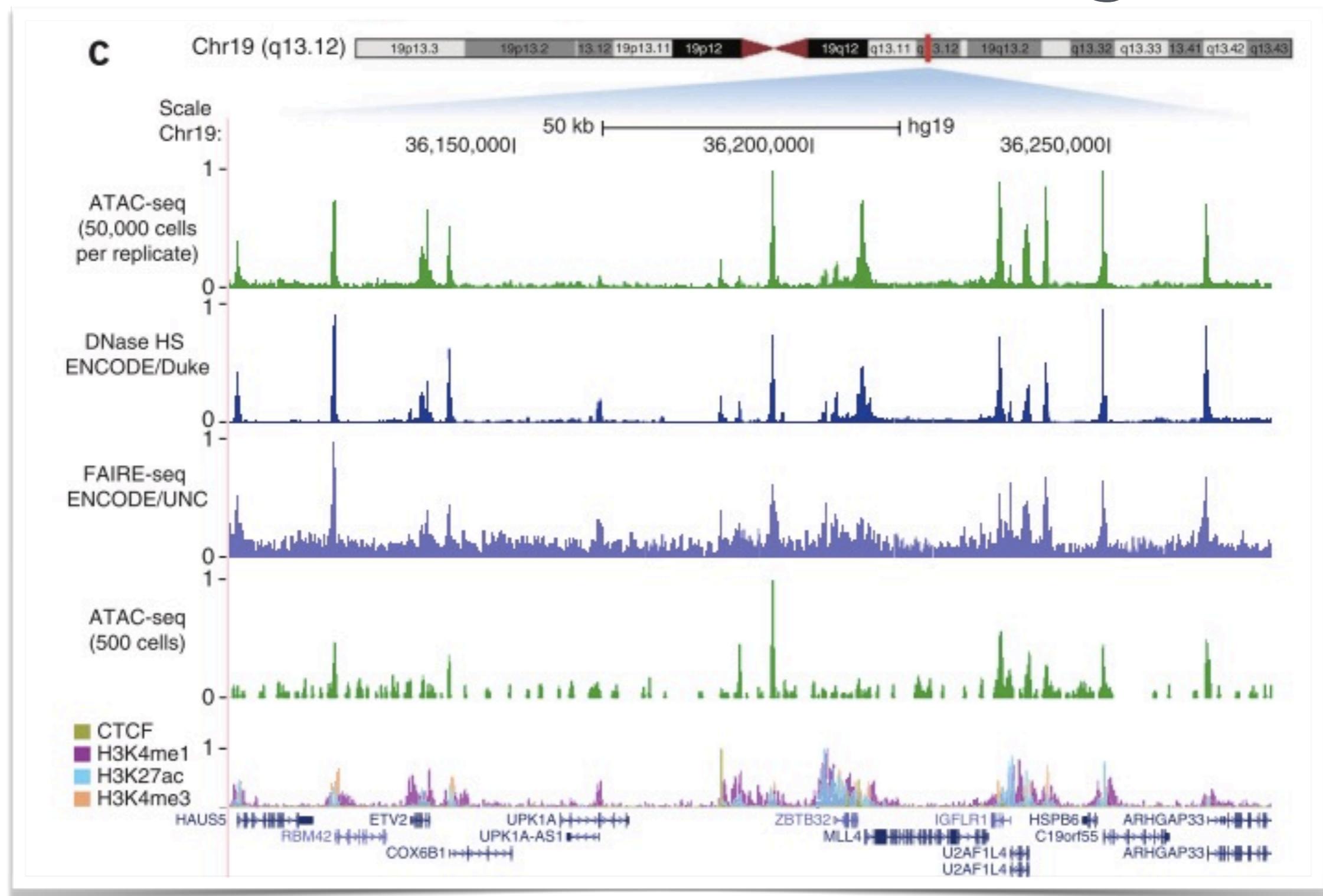


Read alignment



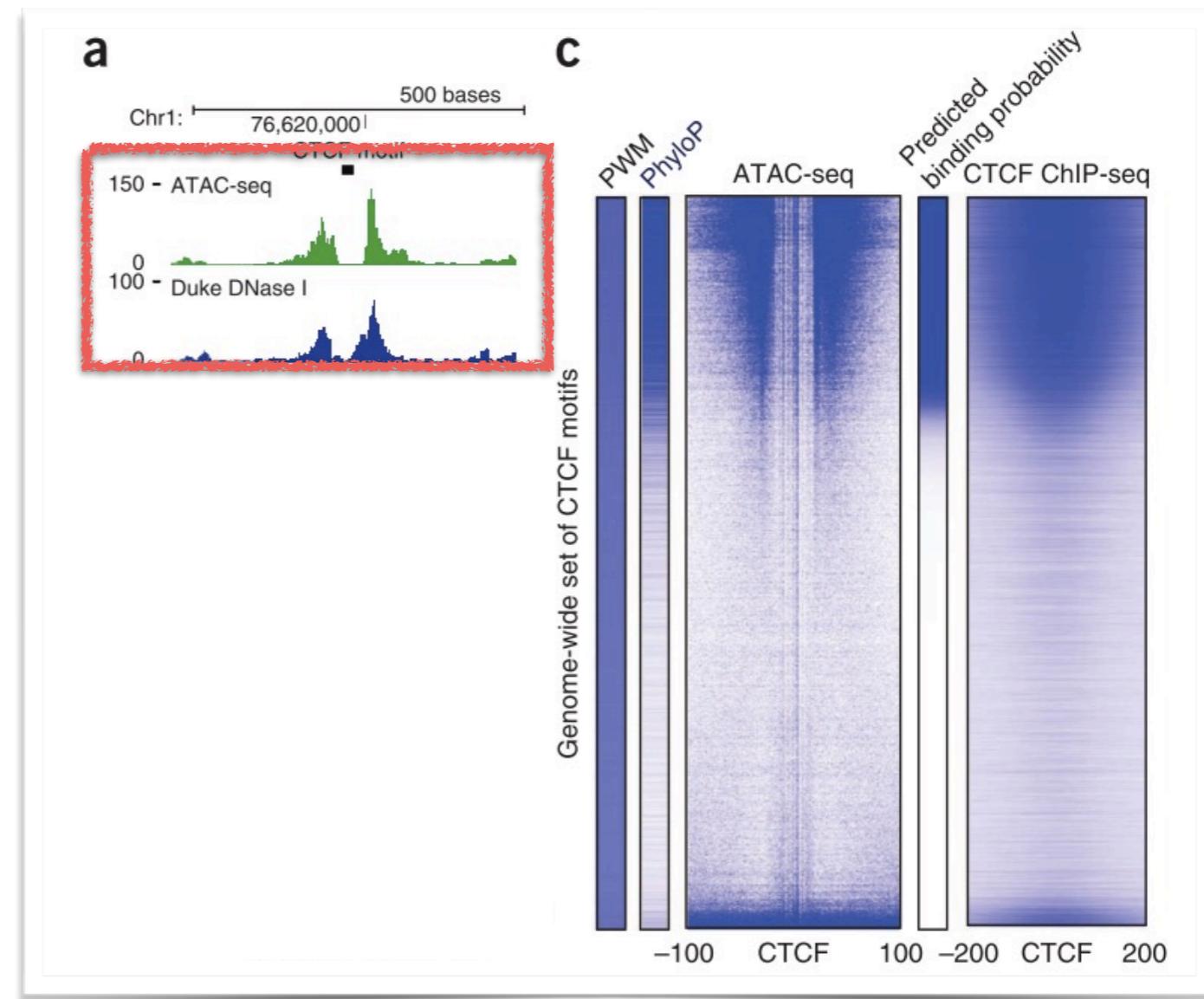
Peak calling

# Identification of accessible regions



# Experimental identification of binding sites

- From open regions to transcription factor binding sites  
→ **footprinting**
- Zooming into the peaks (open regions) : valleys of undigested / un-transposed DNA  
→ **TF binding sites (TFBS)**
- binding sequence can be identified with base-pair resolution

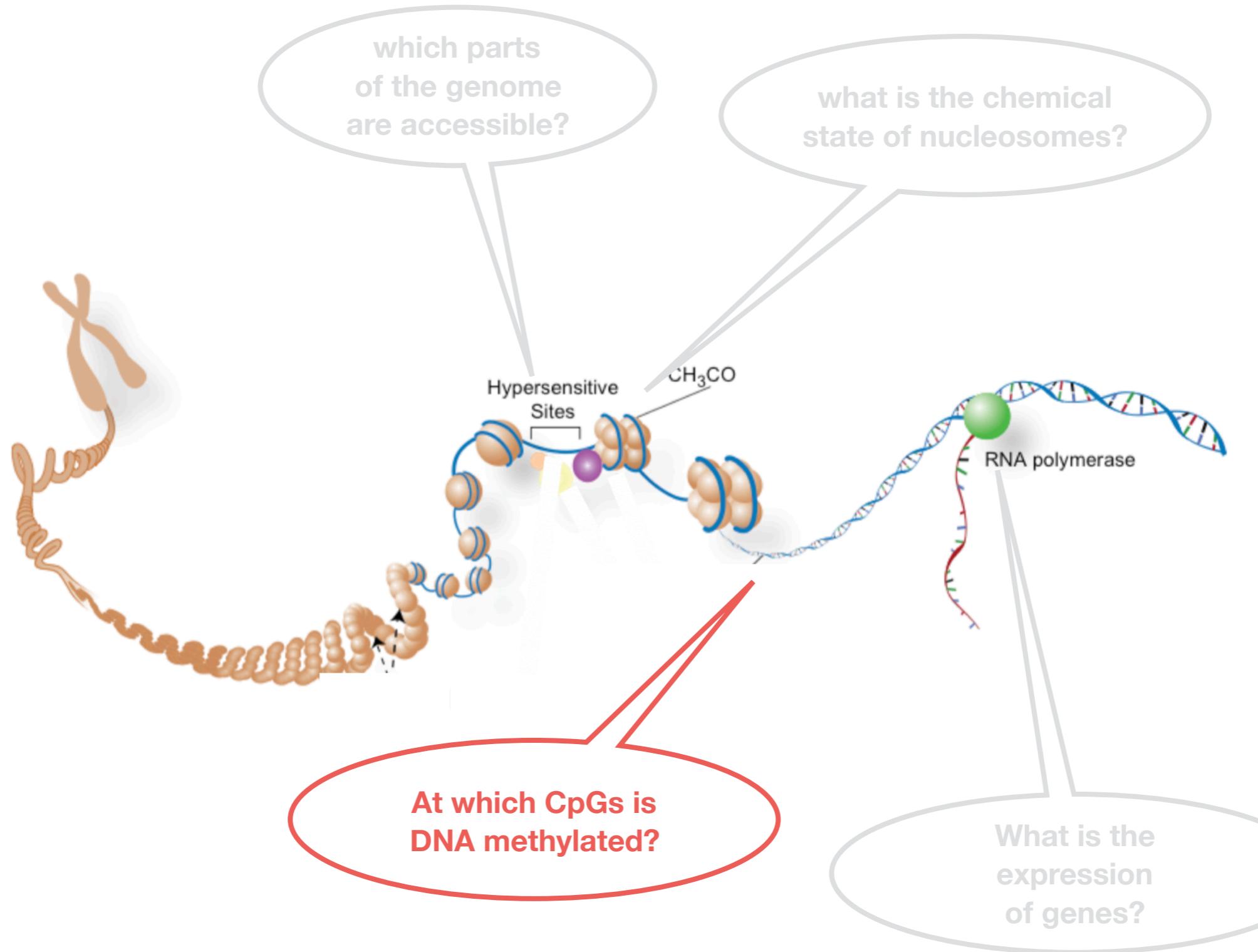


[Greenleaf (2013)]

# Regulatory genome

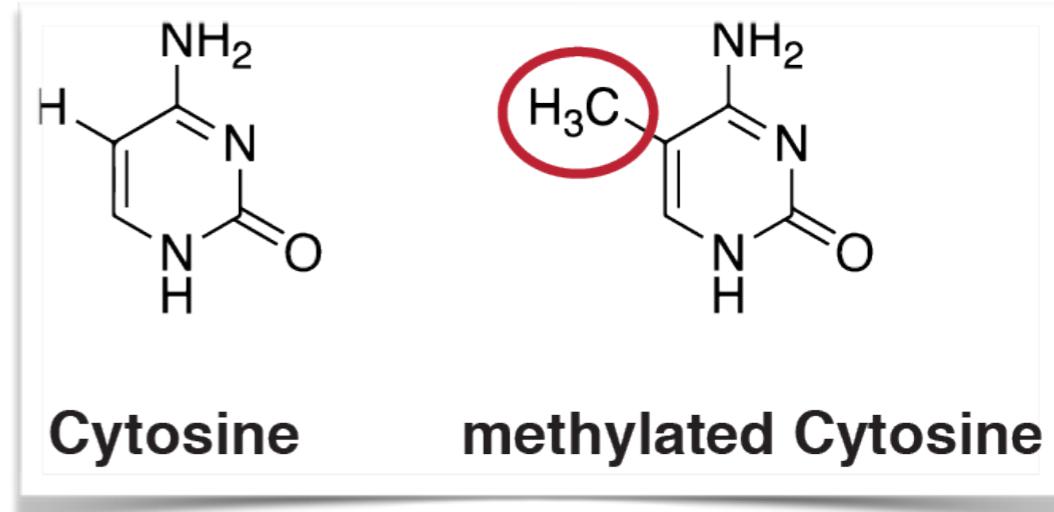


Medizinische Fakultät Heidelberg



# Measuring DNA methylation

- DNA methylation occurs mainly on **cytosines in CpG dinucleotides** in the human genome
- DNA methylation is revealed by using **bisulfite conversion** ( $\text{HSO}_3^-$ ):
  - unmethylated cytosines are converted  
 $\text{C} \rightarrow \text{U} \rightarrow \text{T}$
  - methylated cytosines are protected  
 $\text{mC} \rightarrow \text{mC}$
- unmethylated CpG are identified by the presence of a **mismatch TpG**
- 2 approaches:
  - array based: hybridization to CpG probes on array
  - sequencing: whole genome bisulfite-sequencing



# Measuring DNA methylation

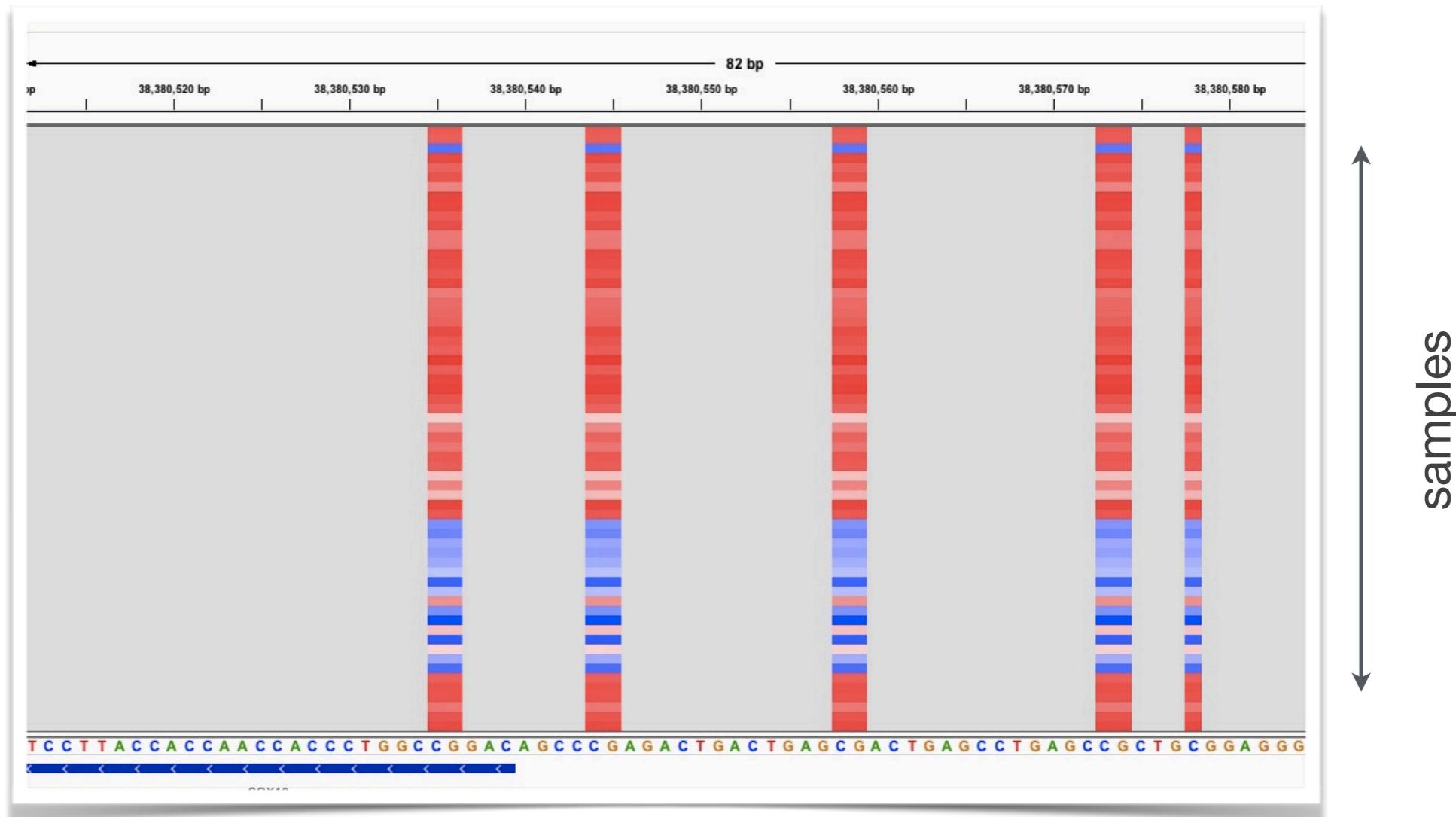
- Array based methods
- CpG containing probes on array
  - 27K probes
  - 450K probes
  - 800K (EPIC)
- all probes contain a methylated (C) and unmethylated (T) version
- Cheap but sparse
- Whole genome bisulfite sequencing
  - unmethylated C → T
  - methylated C → C
- Shearing, conversion and sequencing (Illumina X-10)
- Information about the 28 million CpGs



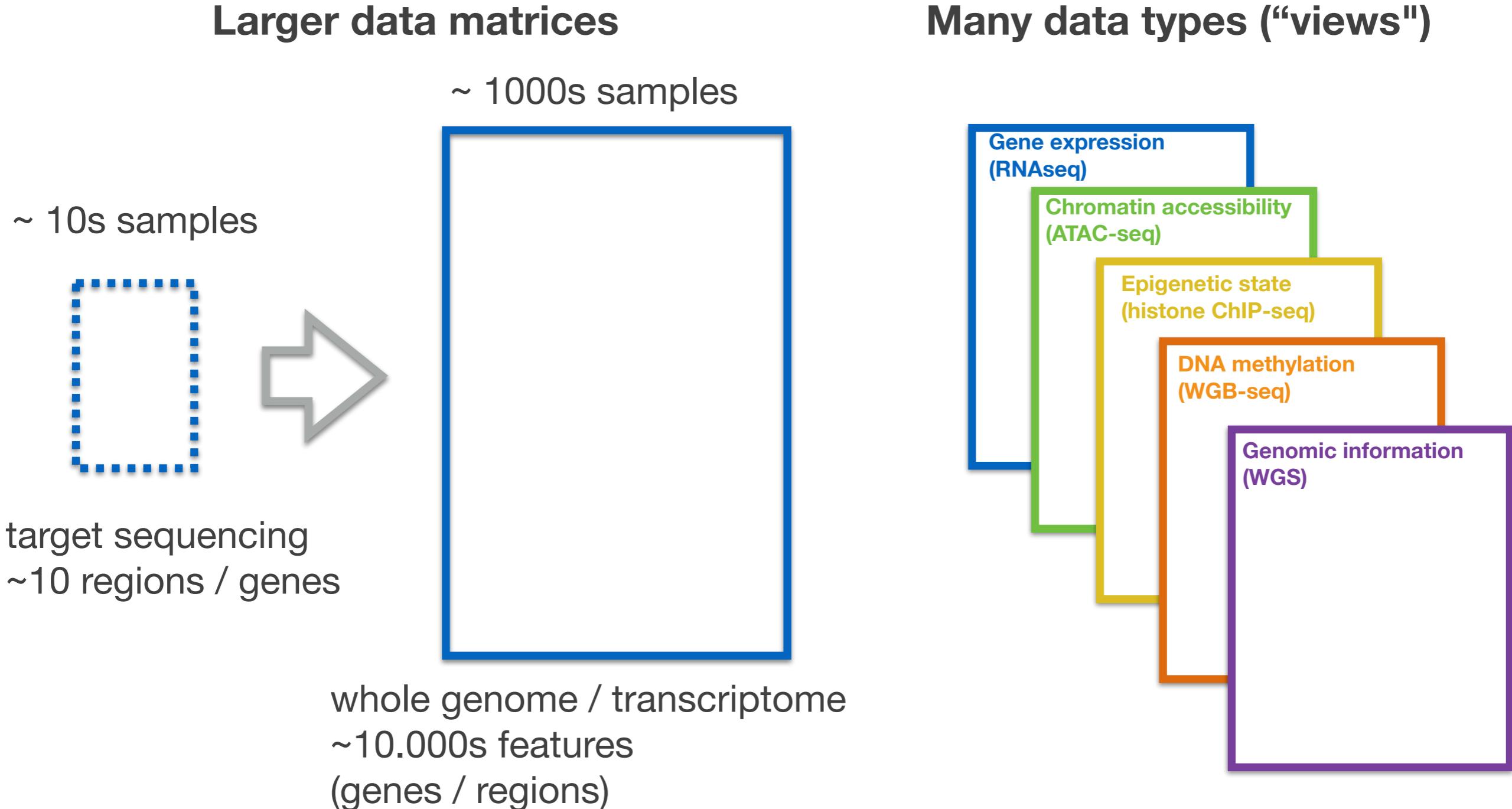
```
--ATGTTCGTAGATTGTACTGTTGAACGTTATGTTAATAGATGCGTTGCGAAT--  
ATGTTCGTAGTTGTA TGTTGAAGTTTATGTTAA  
GTTCGTAGTTGTA TTGAATGTTATGTTAAATA  
TCGTAGTTGTA TGT GAAGTTTATGTTAAATAG  
CGTAGTTGA AATGTTATGTTAAATAGAT  
AGATTGTATGTTGAAG GTTATGTTAAATAGATG  
GATTGTATGTTGAAACG AATAGATGCGTTGGGA  
ATTGTATGTTGAAACGT TAGATGCGTTGGGAAT  
TGTATGTTGAAACGT ATGTTGAACGTTATGTT
```

# Example DNA methylation

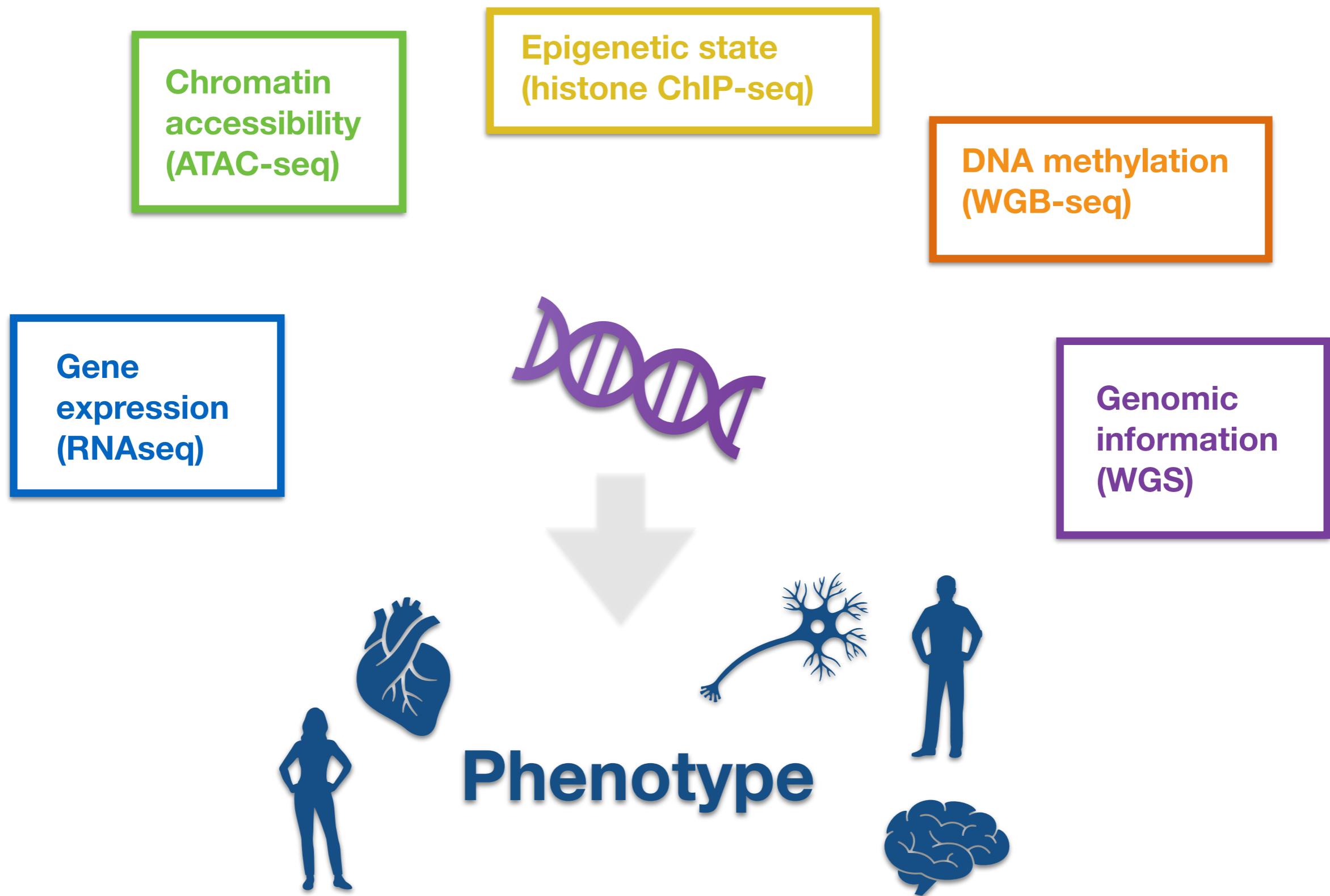
- Whole genome bisulfite sequencing provide information about all CpGs in the genome
- Vertical bars = CpG positions; red = high methylation (100%); blue = no methylation (~10%)



# Multi-omics data



# “Whole more than sum of the parts”

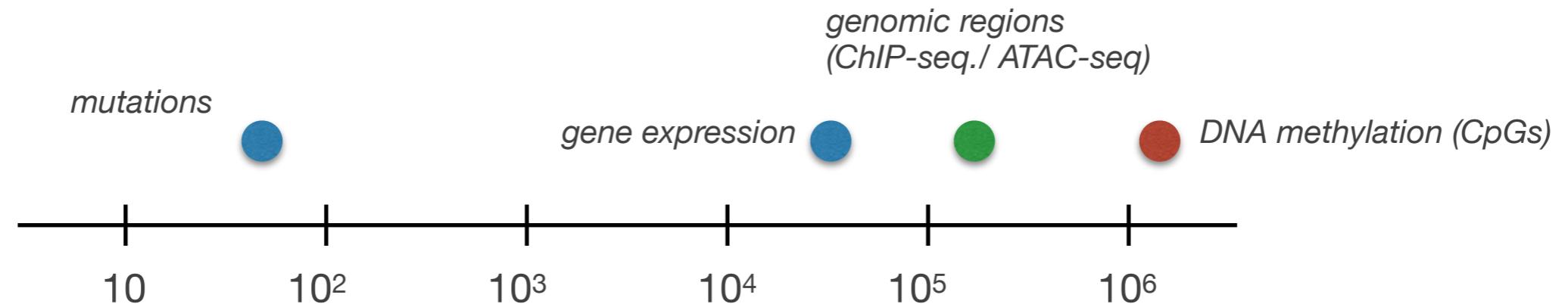


### 3. Dimensional reduction and data integration

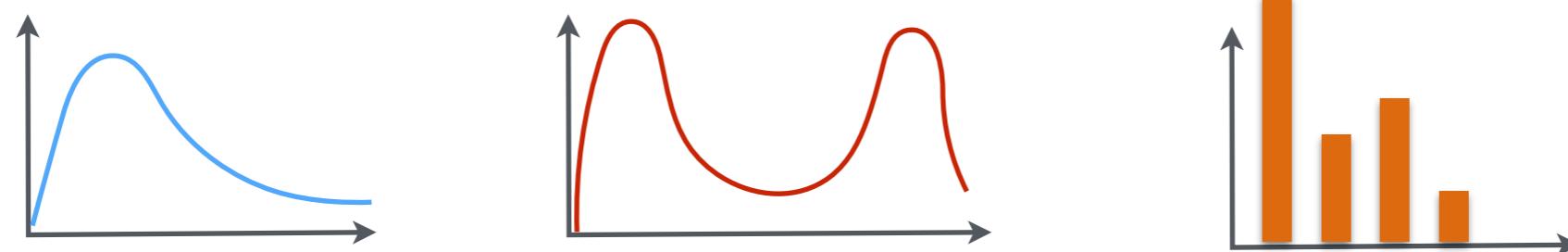
- Clustering; Consensus Clustering
- Clusters of Clusters; iCluster
- Principal Component analysis
- Non-negative matrix factorization and its flavors...

# Challenges

- Different **dimensionalities** and **features**



- Different types / **distributions** of data



- **Missing data:** not all samples have measurements in all features and all views

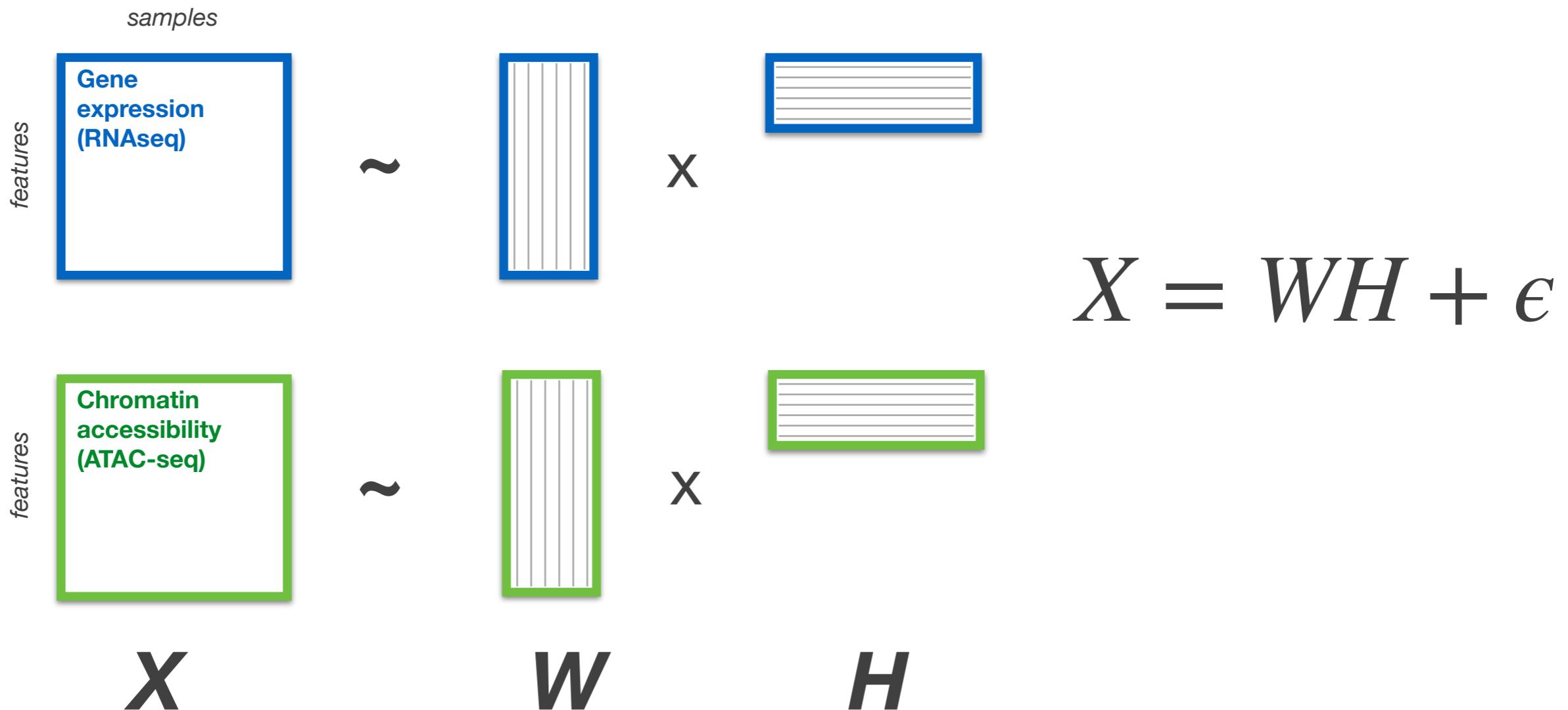
# Various approaches for data integration

- **Consensus clustering** approaches
  - Clusters of Clusters (CoCA)
- **Matrix factorization** approaches
  - Principal component analysis (PCA)
  - Non-negative matrix factorization (NMF)
- **Network based approaches**
  - integrated pathway activity scores (IPA) (PARADIGM, Vaske et al., 2010)
  - patient-patient similarity networks (SNF, Wang et al., 2014)

# Matrix factorization



Medizinische Fakultät Heidelberg

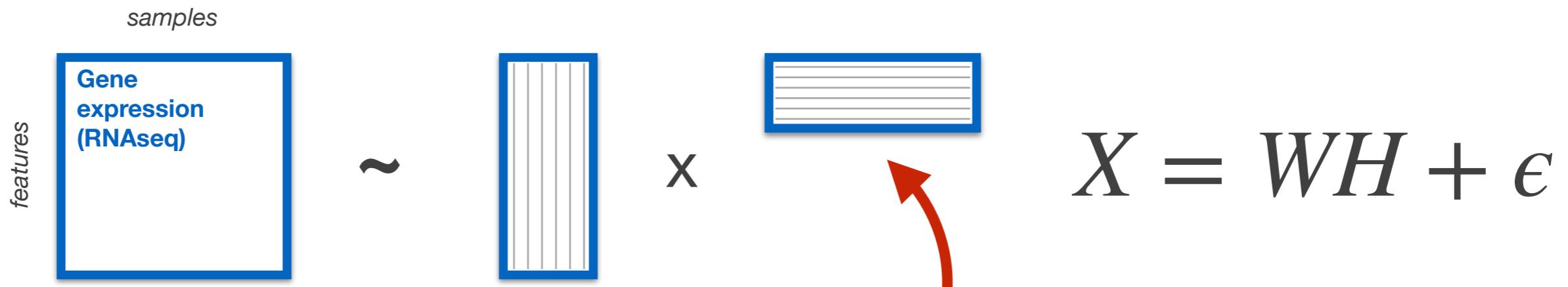


- approximate large data matrix using the product of 2 smaller matrices
- **columns of  $W$  = molecular signatures**

# Matrix factorization



Medizinische Fakultät Heidelberg



$$X = WH + \epsilon$$

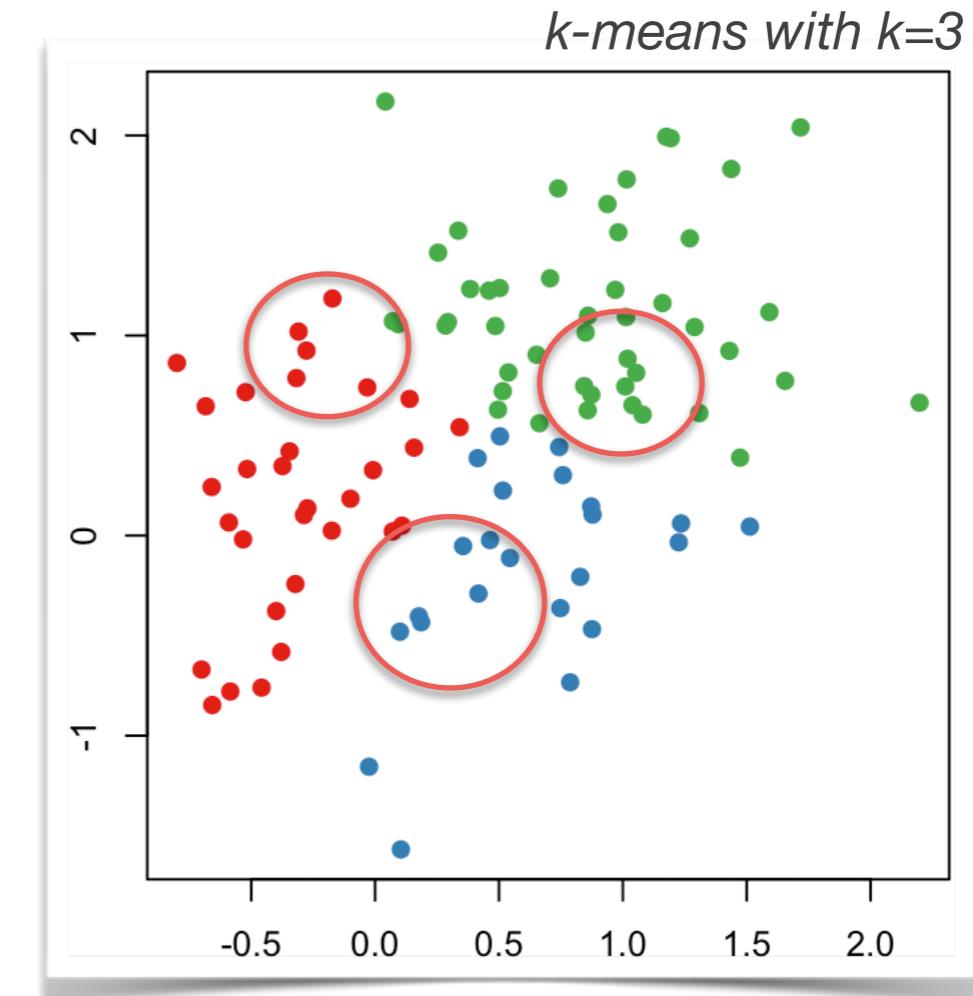
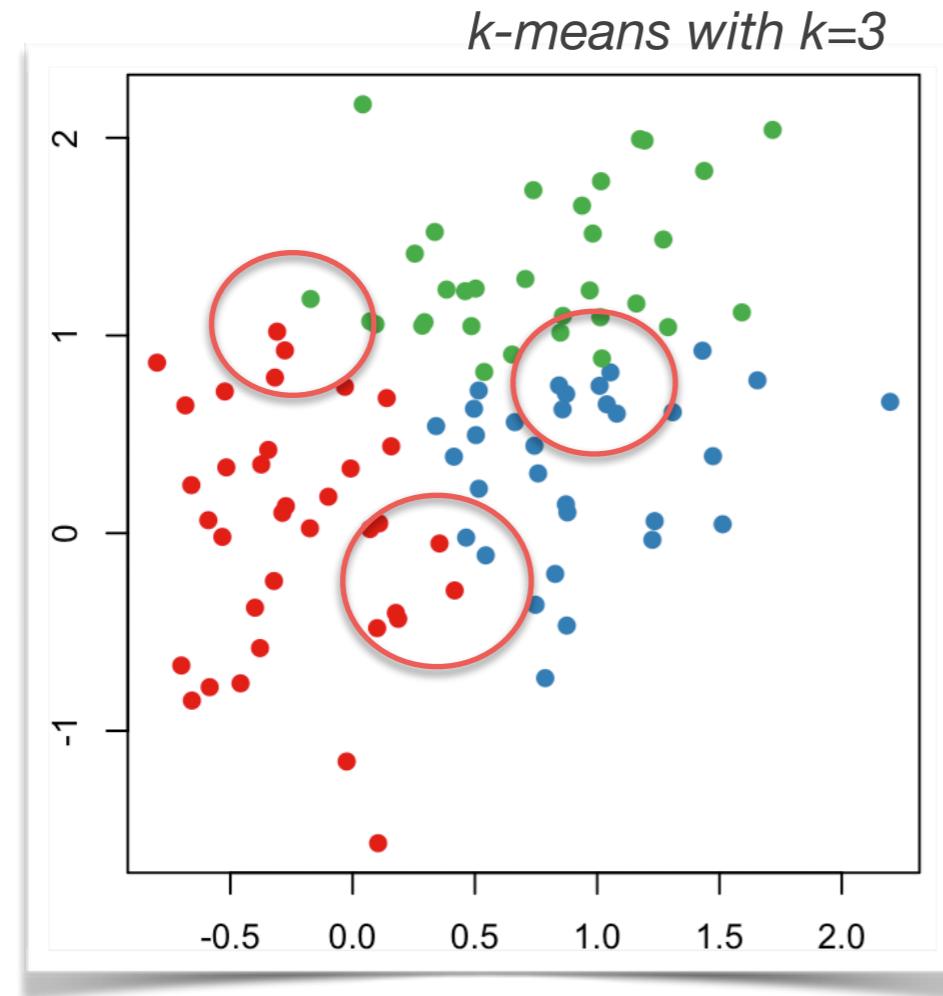
- Many flavors!
  - cluster analysis
  - principal components analysis
  - (group) factor analysis
  - non-negative matrix factorization
  - etc ...

# Clustering



Medizinische Fakultät Heidelberg

- Clustering is the simplest unsupervised dimensional reduction method
- Many clustering methods: k-means, PAM, hierarchical clustering,...
- Sensitive to initialization of procedure, especially if the clusters not well separated!

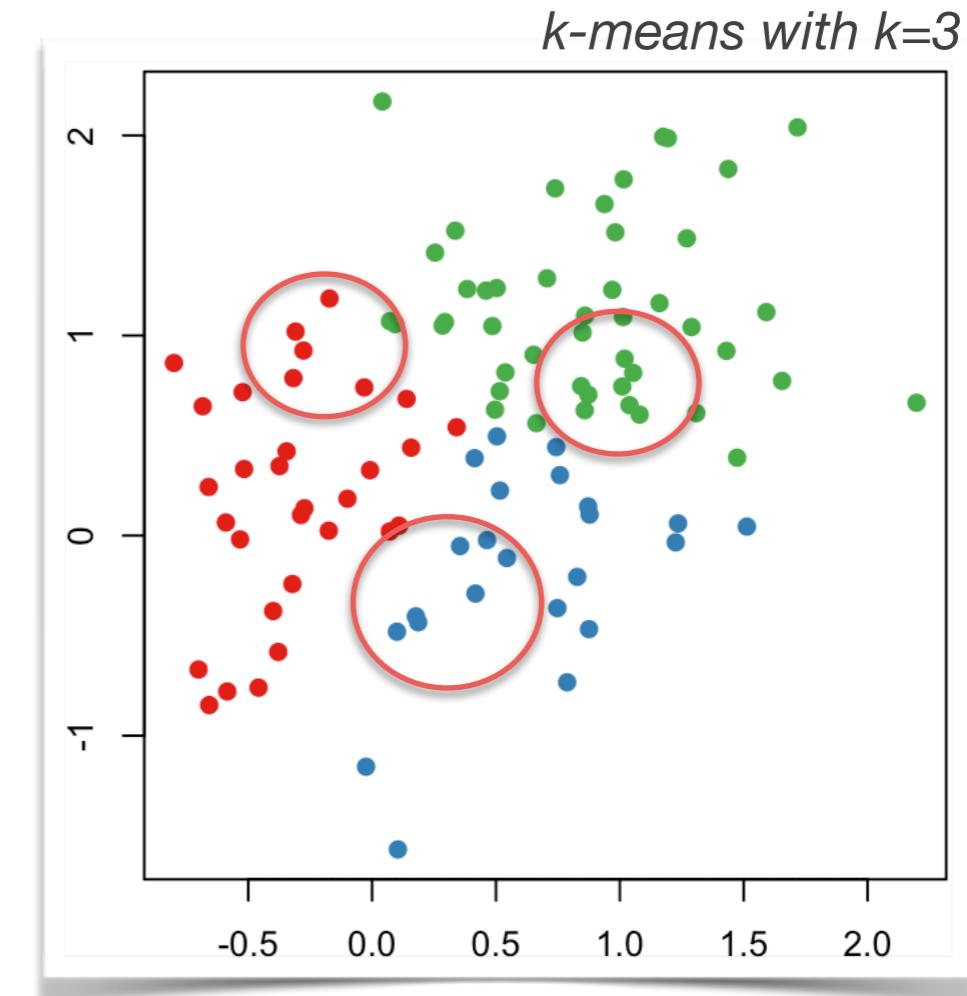
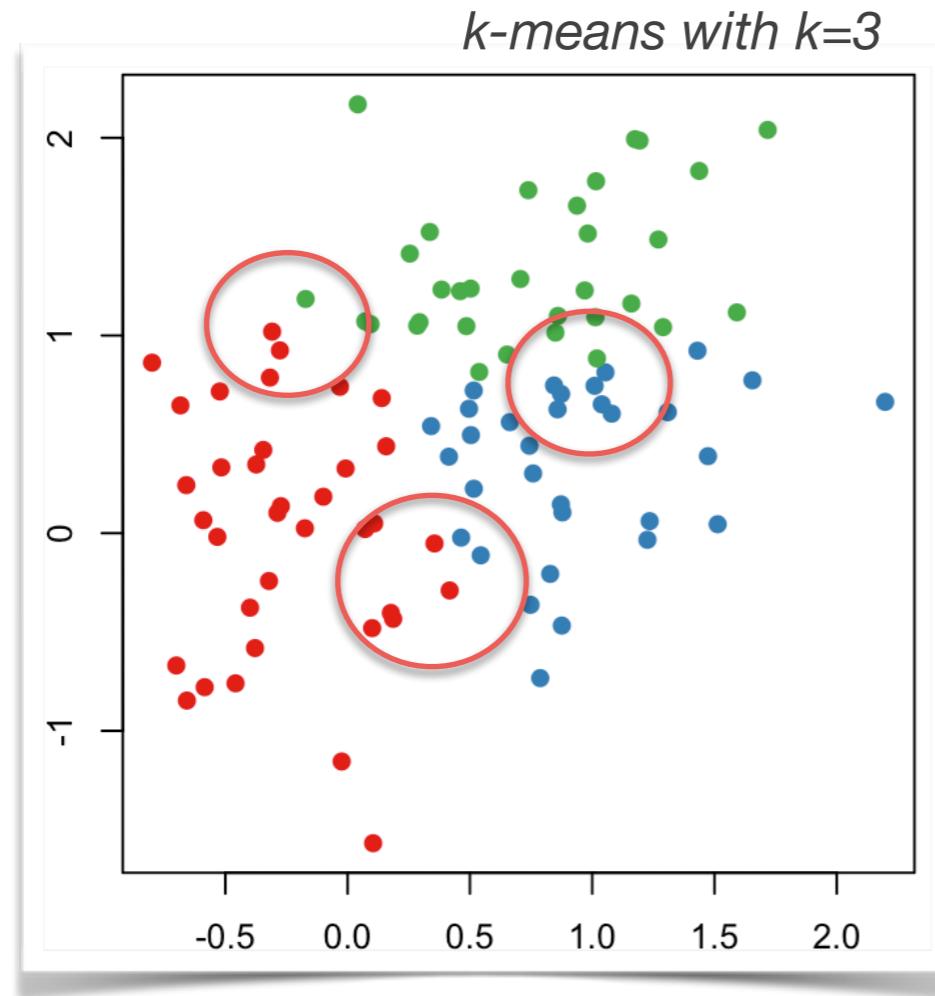


# Clustering



Medizinische Fakultät Heidelberg

- Clustering is the simplest unsupervised dimensional reduction method
- Many clustering methods: k-means, PAM, hierarchical clustering,...
- Sensitive to initialization of procedure, especially if the clusters not well separated!



# Consensus clustering



Medizinische Fakultät Heidelberg

- Idea of **consensus clustering**:

$D = \{e_1, \dots, e_N\}$  expression profiles for N patients

$D^{(h)}$  subset of the patients (e.g. 80%)

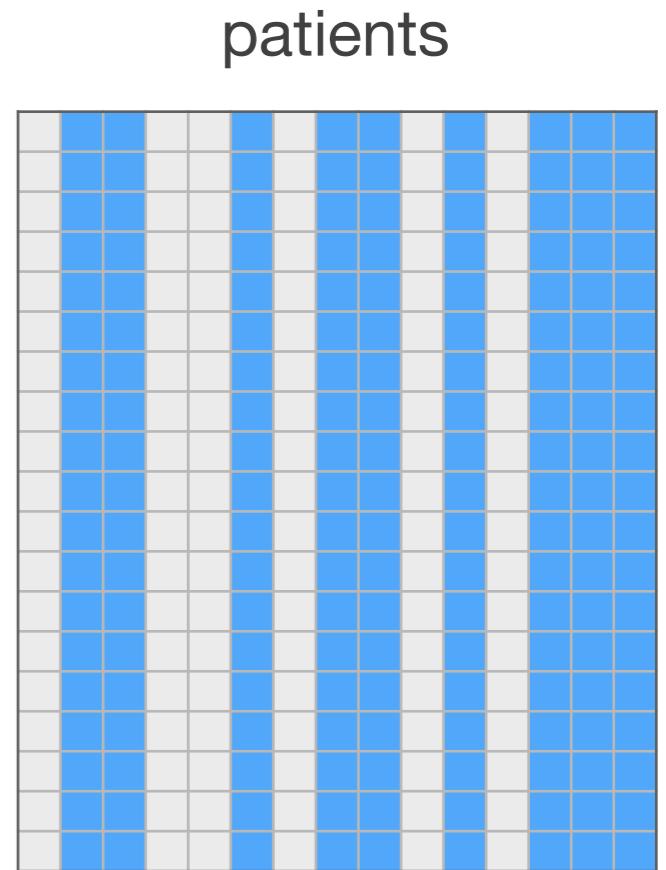
$M^{(h)}$  result of clustering  $D^{(h)}$

$M^{(h)}(i, j) = 1$  if (i,j) belong to the same cluster

$I^{(h)}(i, j) = 1$  if (i,j) both included in  $D^{(h)}$

$$m(i, j) = \frac{\sum_h M^{(h)}(i, j)}{\sum_h I^{(h)}(i, j)}$$

$$d(i, j) = 1 - m(i, j)$$



blue columns = sampled patients

Use the matrix d to perform (hierarchical) clustering

# Consensus clustering



Medizinische Fakultät Heidelberg

```
> results[[2]][["consensusMatrix"]][1:5,1:5]
      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 1.0000000 1.0000000 0.9655172 1.0000000 1.0000000
[2,] 1.0000000 1.0000000 0.8857143 1.0000000 1.0000000
[3,] 0.9655172 0.8857143 1.0000000 0.9166667 0.8823529
[4,] 1.0000000 1.0000000 0.9166667 1.0000000 1.0000000
[5,] 1.0000000 1.0000000 0.8823529 1.0000000 1.0000000
> results[[3]][["consensusMatrix"]][1:5,1:5]
      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 1.0000000 0.3548387 0.8620690 0.2413793 1.0000000
[2,] 0.3548387 1.0000000 0.1142857 1.0000000 0.4000000
[3,] 0.8620690 0.1142857 1.0000000 0.1388889 0.7941176
[4,] 0.2413793 1.0000000 0.1388889 1.0000000 0.3513514
[5,] 1.0000000 0.4000000 0.7941176 0.3513514 1.0000000
```

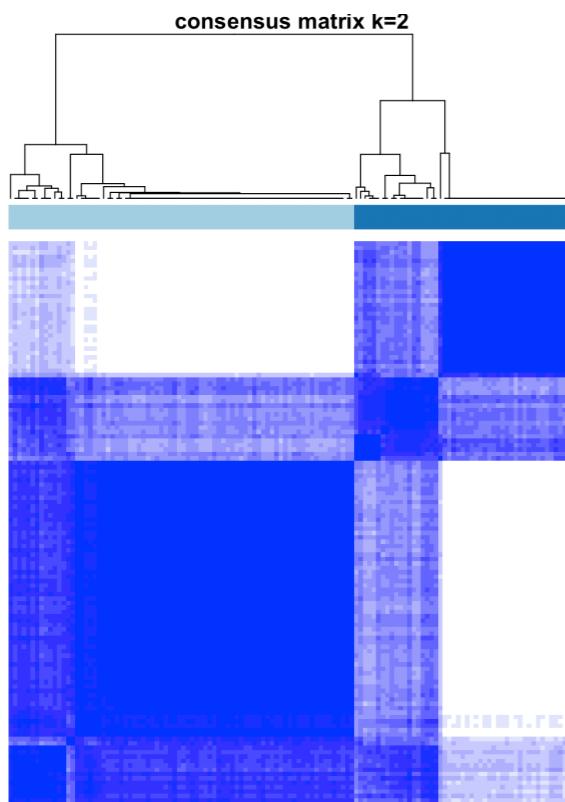
similarity matrix  
for  $k = 2$

similarity matrix  
for  $k = 3$

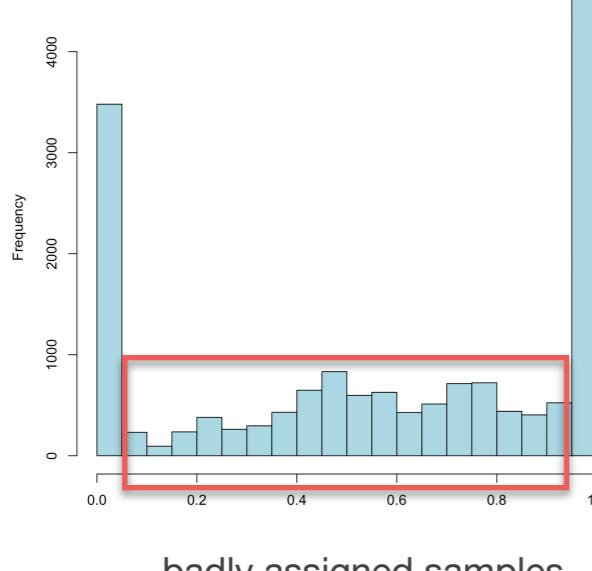
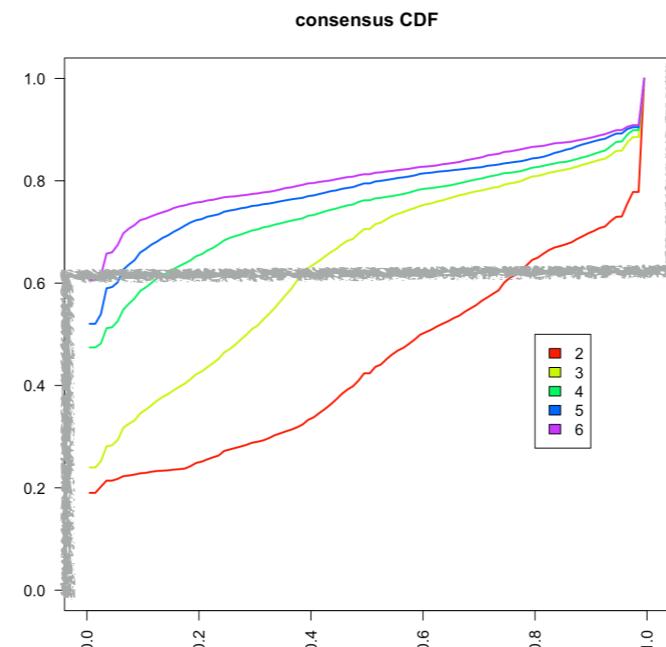
# Consensus Clustering



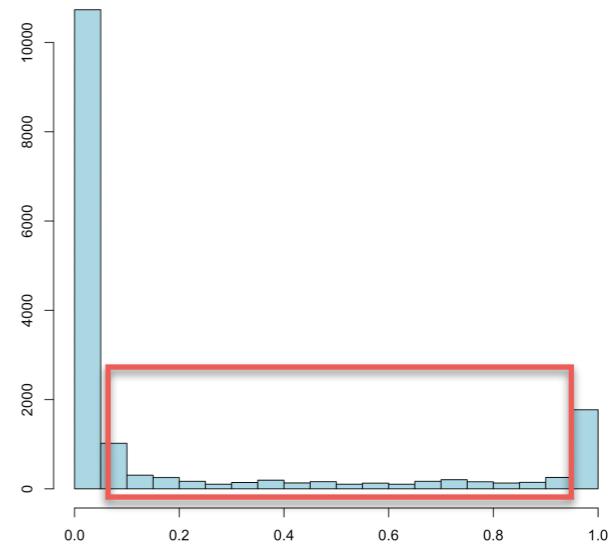
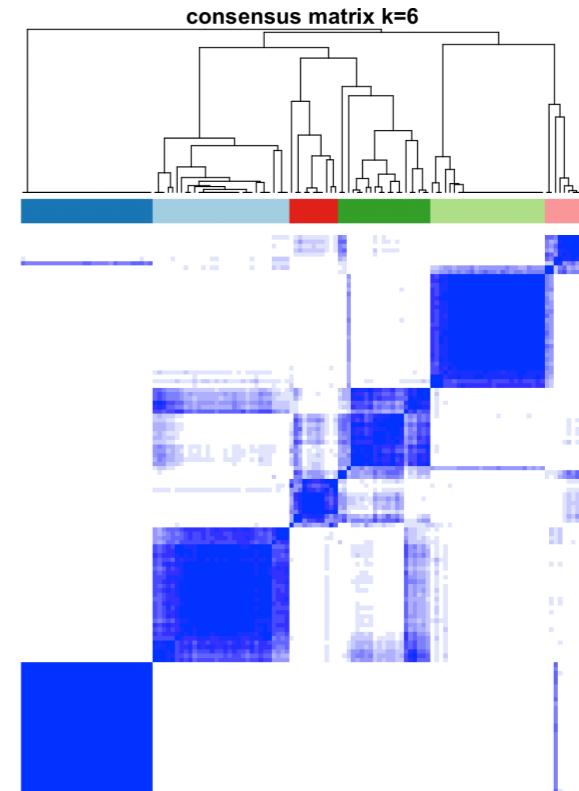
Medizinische Fakultät Heidelberg



Ideal shape would  
be a step function



Optimal K when AUC no longer increases

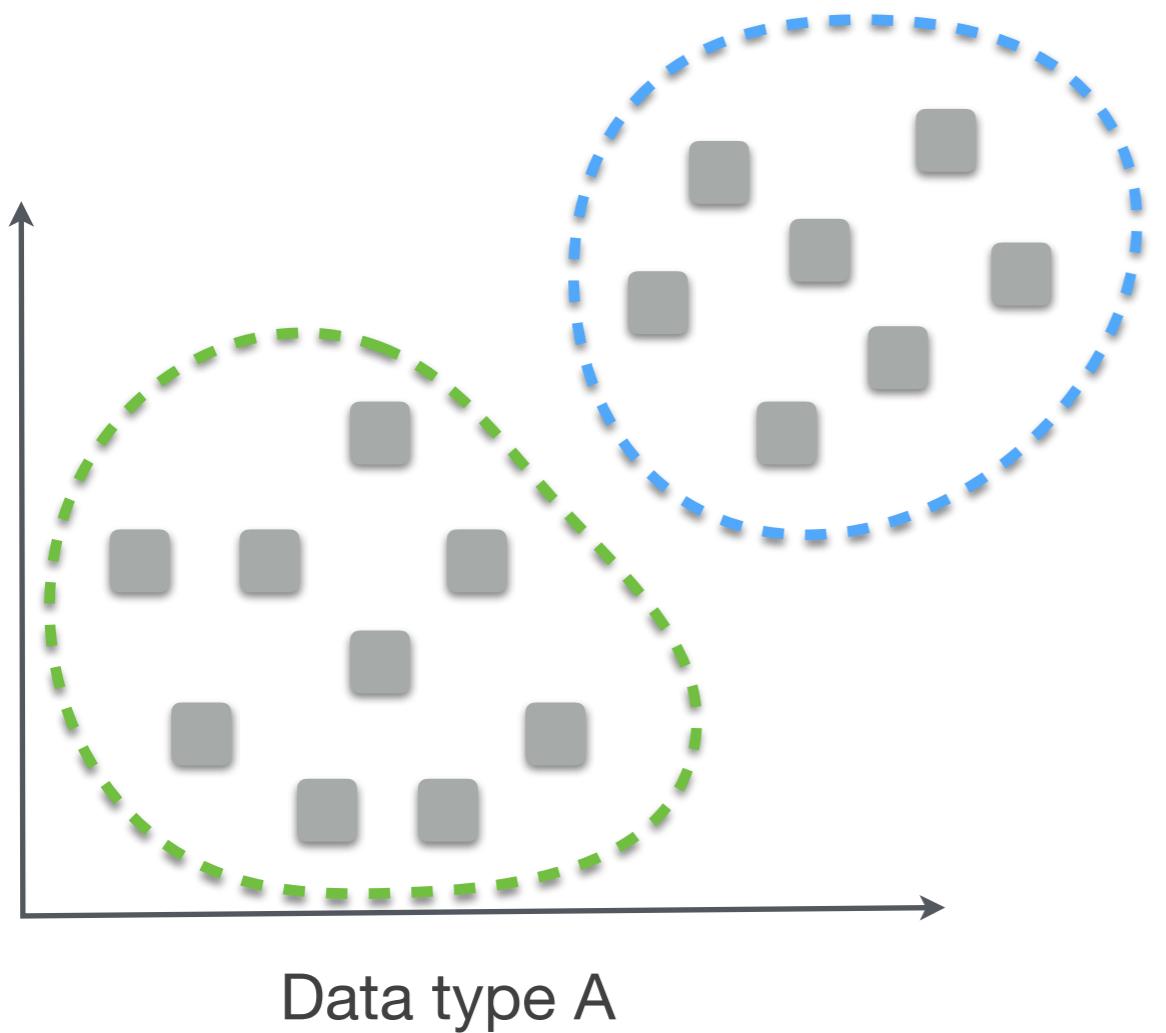


[Monti et al., 2003]

# Clustering over multiple data?



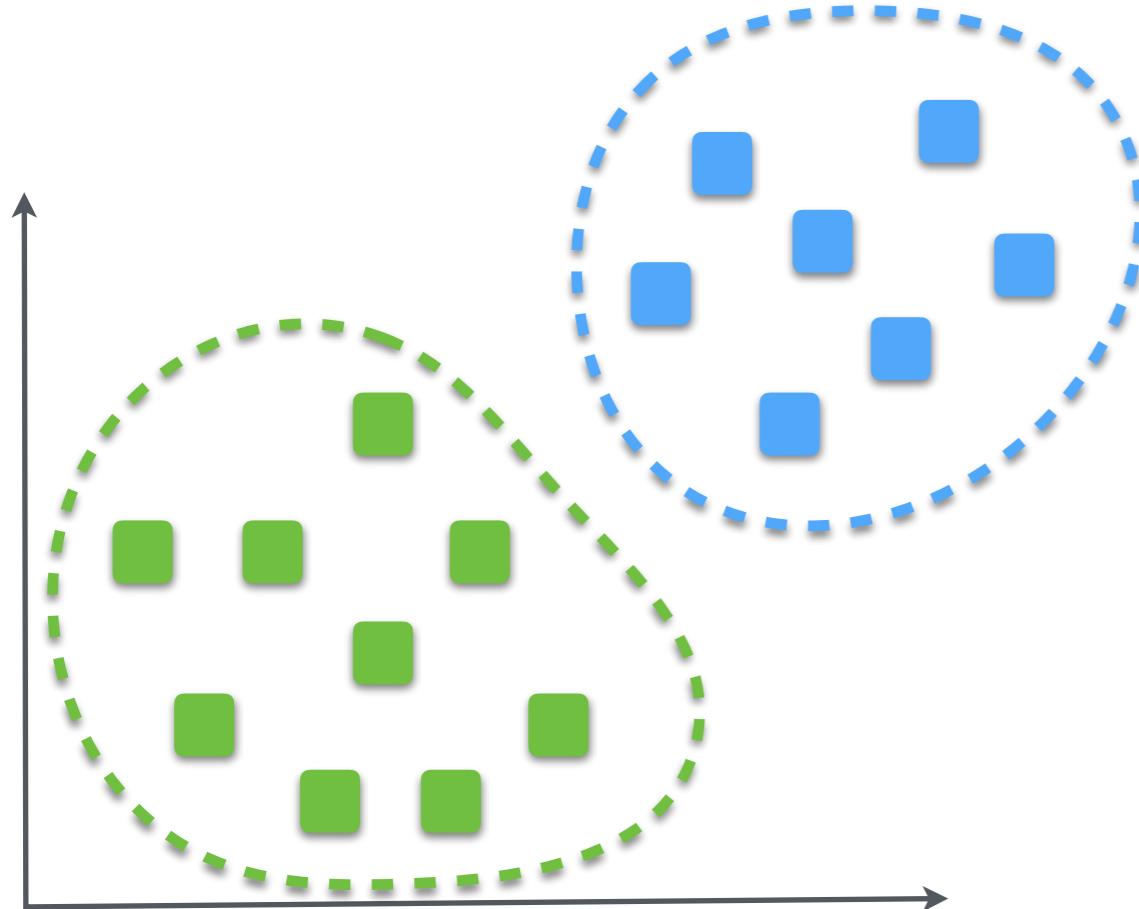
Medizinische Fakultät Heidelberg



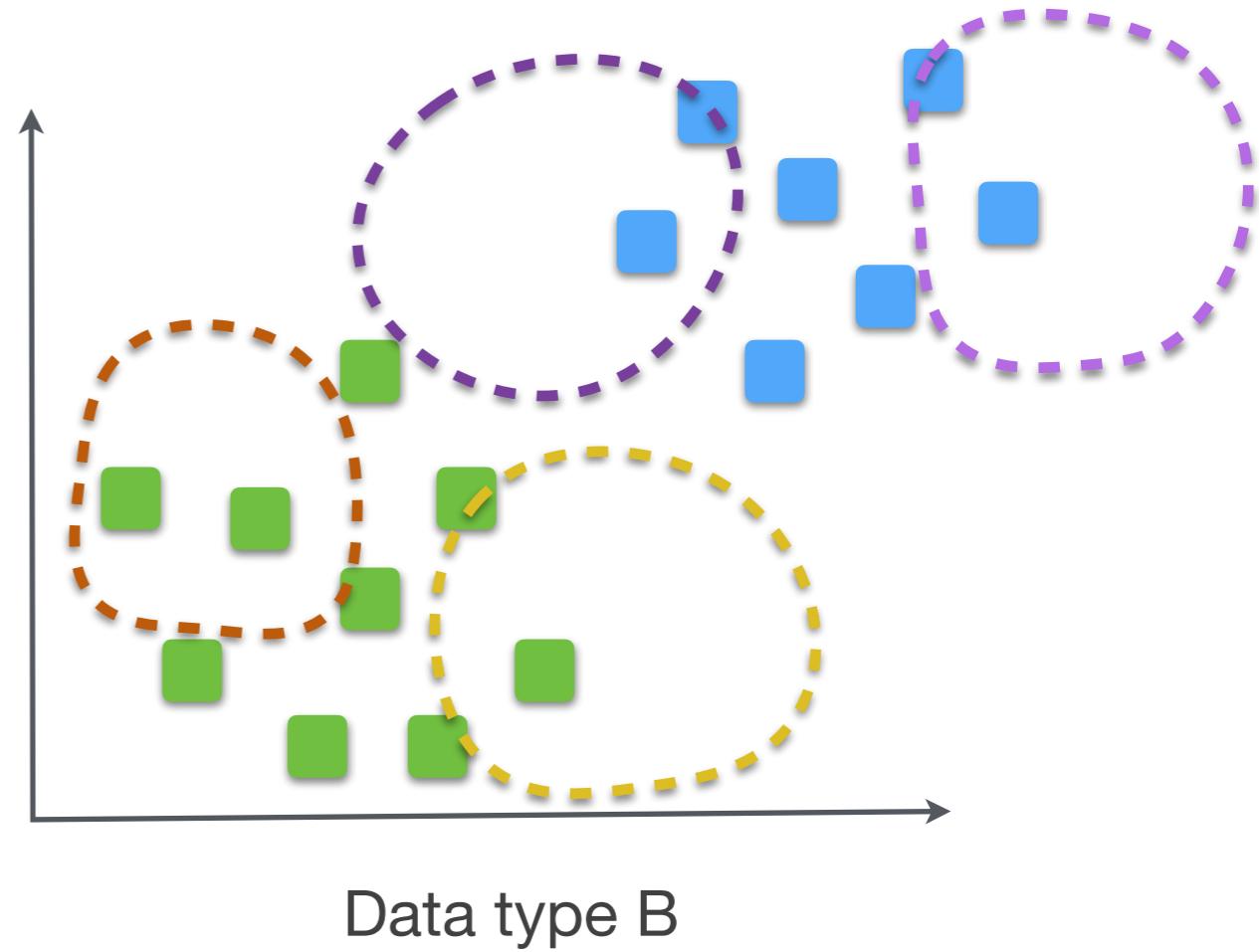
# Clustering over multiple data?



Medizinische Fakultät Heidelberg



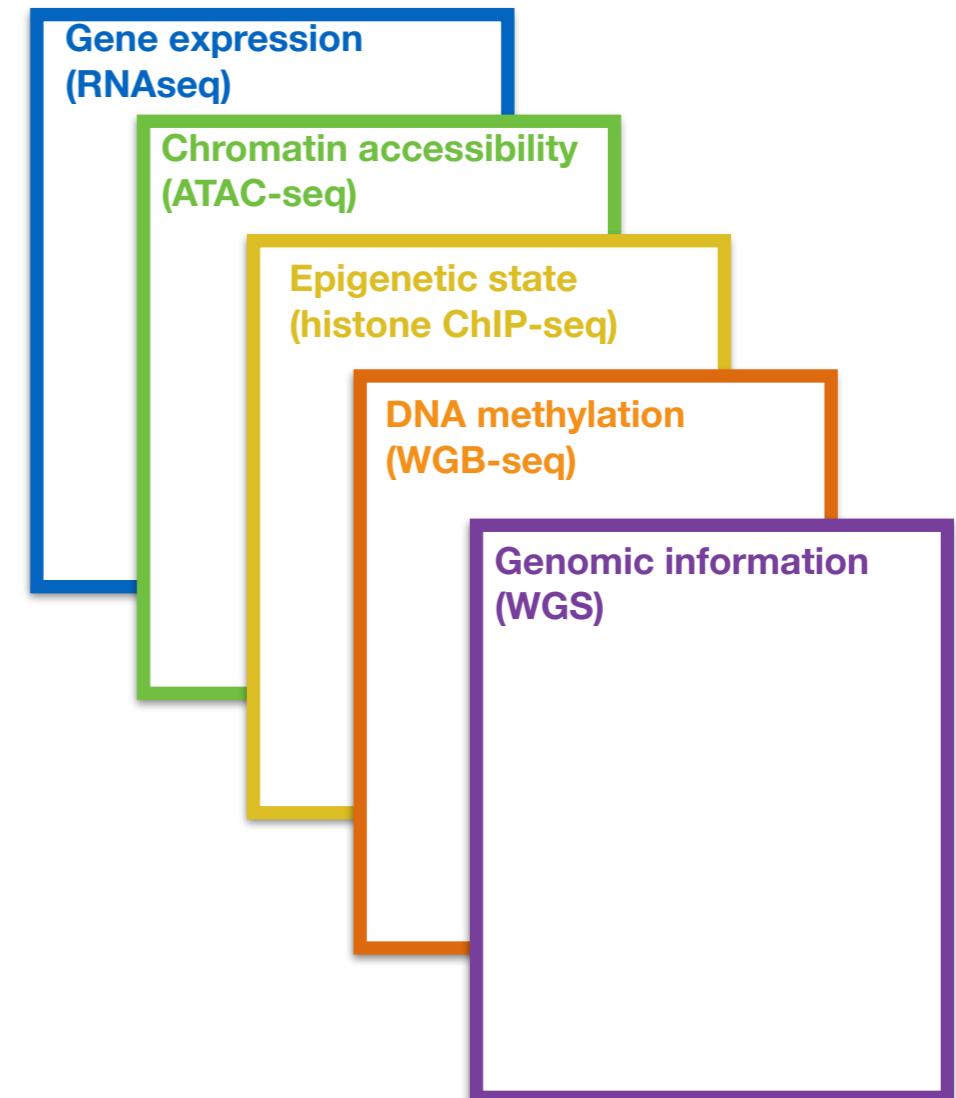
Data type A



Data type B

# Cluster of Cluster Analysis (CoCA)

- Cluster each omics data separately
  - each clustering can use a different clustering algorithm (k-means, PAM,...)
  - each omics datatype can lead to distinct number of clusters
- Represent each sample by an **indicator vector** showing to which cluster it belongs in each omic  
 $s_3 = ( \textcolor{blue}{1}, \textcolor{green}{3}, \textcolor{yellow}{2}, \textcolor{red}{3}, \textcolor{purple}{1} )$
- Cluster the samples based on this indicator vector using **consensus clustering**

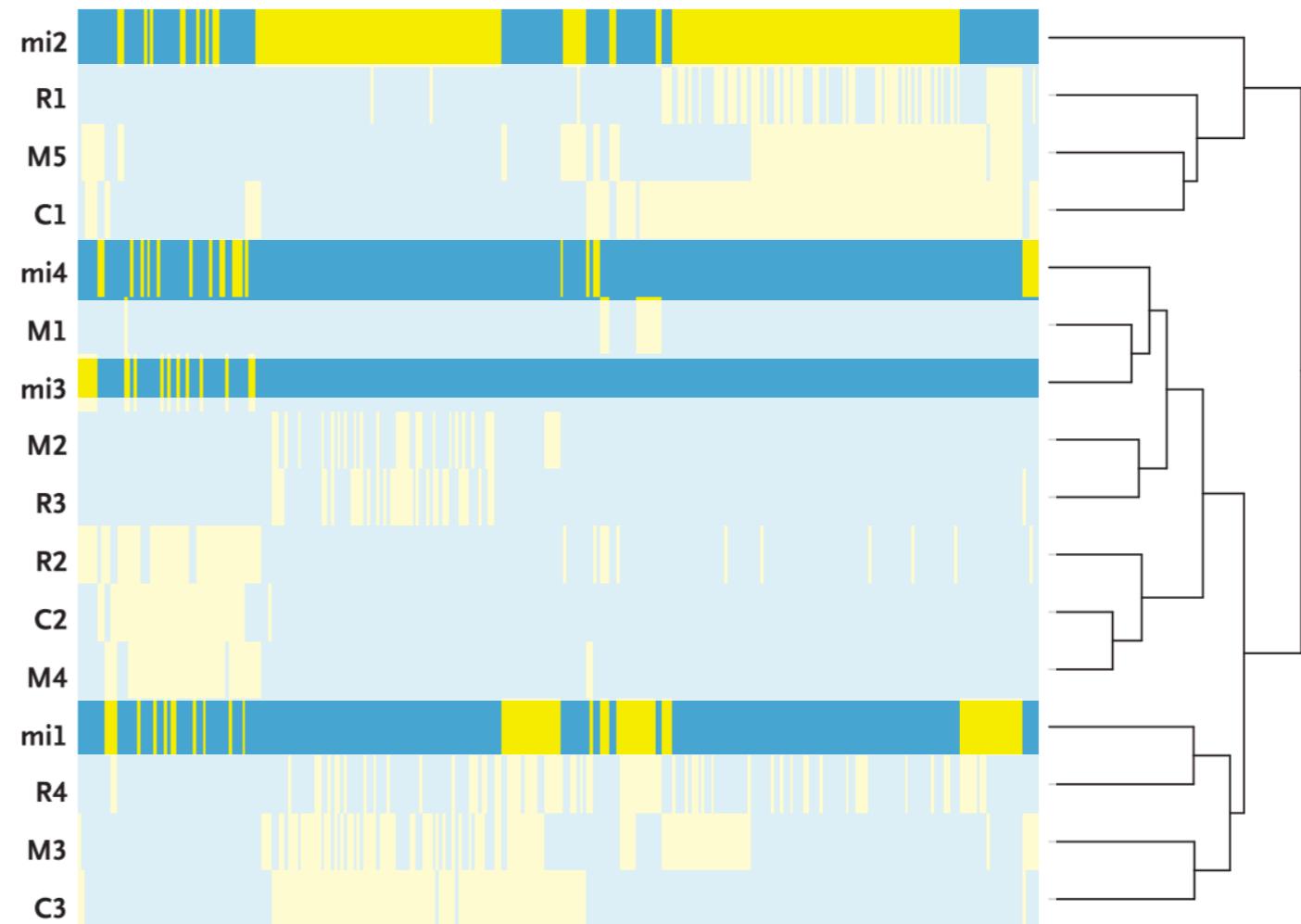


# Application: low-grade glioma



Member  
Nonmember

- TCGA: integrative clustering of low-grade glioma (brain tumor)
- Available data ( $n=293$ ):
  - mRNA expression (R)
  - micro-RNA expression (mi)
  - Copy-number variation (C)
  - DNA-methylation (M)
- Result: 3 robust subtypes which disagree with pathological subtypes!

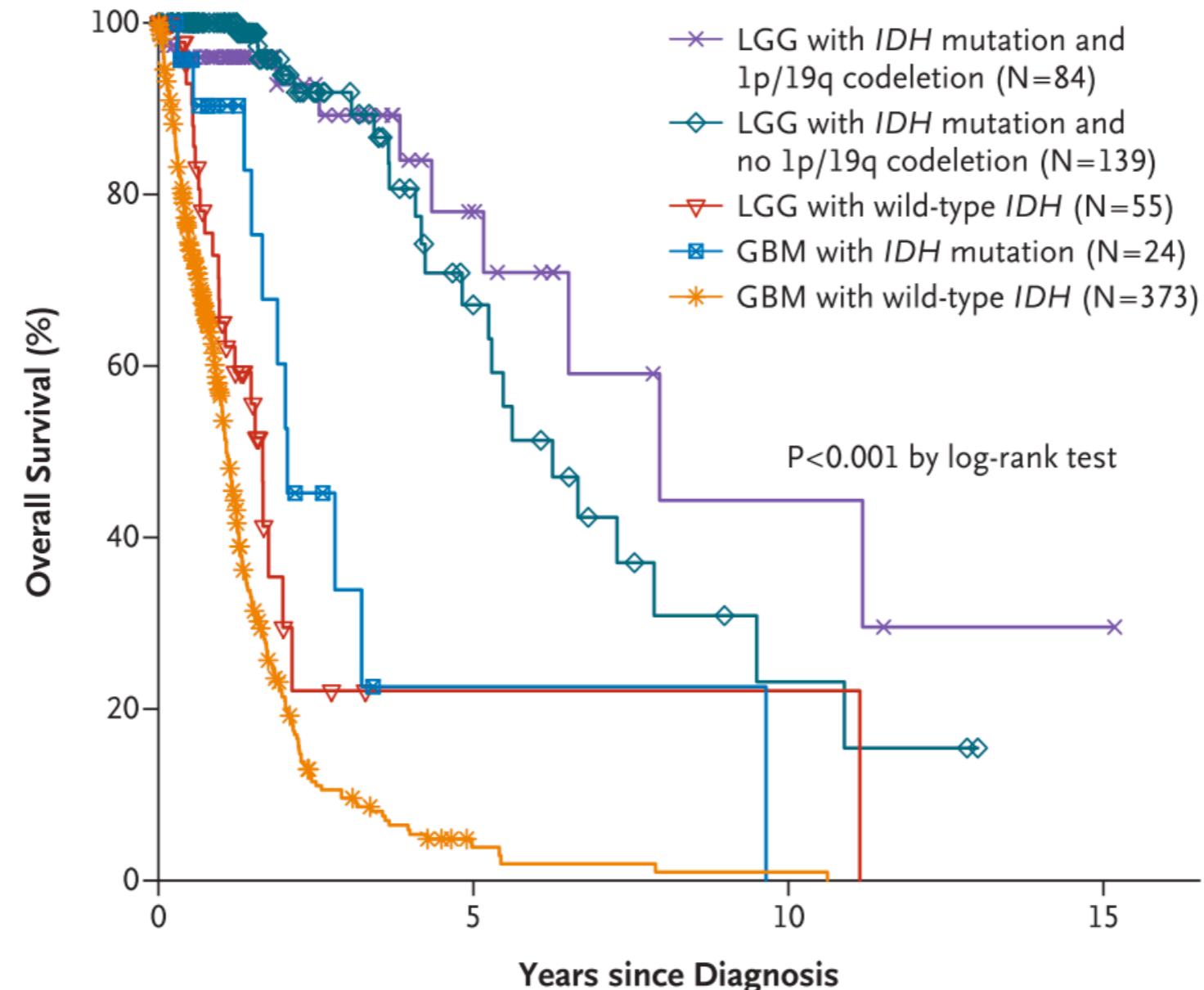


[Brat et al., NJEM, 2015]

# Application: low-grade glioma



B Gliomas Classified According to Molecular Subtype

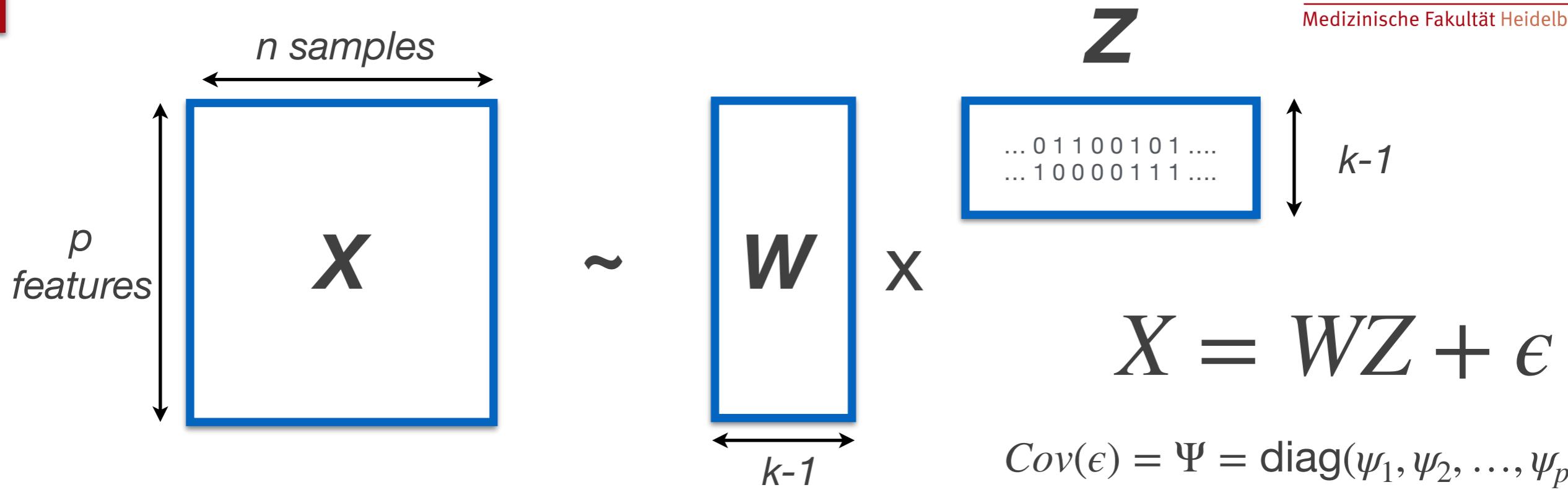


[Brat et al., NJEM, 2015]

# iCluster



Medizinische Fakultät Heidelberg



- Goal: identify  $k$  clusters of samples in the dataset such (i.e.  $Z$ ) that the inter-cluster distance is maximized
- $Z$  is the **indicator function**
  - $Z_{ij} = 1$  : sample  $j$  belongs to cluster  $i$
  - $Z_{ij} = 0$  : sample  $j$  does not belong to cluster  $i$

$$X = WZ + \epsilon$$

$$\text{Cov}(\epsilon) = \Psi = \text{diag}(\psi_1, \psi_2, \dots, \psi_p)$$

- **X is observed**
- **W** and  $\Psi$  are unknown parameters (these are numbers!)
- **Z** is the unknown **latent variable** (this is a random variable!)
- Bayesian formulation: binary  $Z \rightarrow$  continuous  $Z^*$
- Prior distribution :  $Z^* \sim \mathcal{N}(0, I)$
- Goal: maximize posterior probability

$$E[Z^* | X]$$

$$X = WZ + \epsilon$$

$$\text{Cov}(\epsilon) = \Psi = \text{diag}(\psi_1, \psi_2, \dots, \psi_p)$$

- Find optimal solution using **Expectation-Maximization**

Initial random values for  $(W^{(0)}, \Psi^{(0)})$

(Expectation Step)

Estimate  $Z^{(t)}$   
using  $(W^{(t-1)}, \Psi^{(t-1)}, X)$

$$E[Z^* | X] = W'\Sigma^{-1}X$$

with  $\Sigma = WW' + \Psi$

$$X = WZ + \epsilon$$

$$\text{Cov}(\epsilon) = \Psi = \text{diag}(\psi_1, \psi_2, \dots, \psi_p)$$

- Find optimal solution using **Expectation-Maximization**

Initial random values for  $(W^{(0)}, \Psi^{(0)})$

(Expectation Step)

Estimate  $Z^{(t)}$   
using  $(W^{(t-1)}, \Psi^{(t-1)}, X)$

$$E[Z^* | X] = W'\Sigma^{-1}X$$

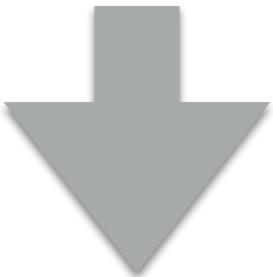
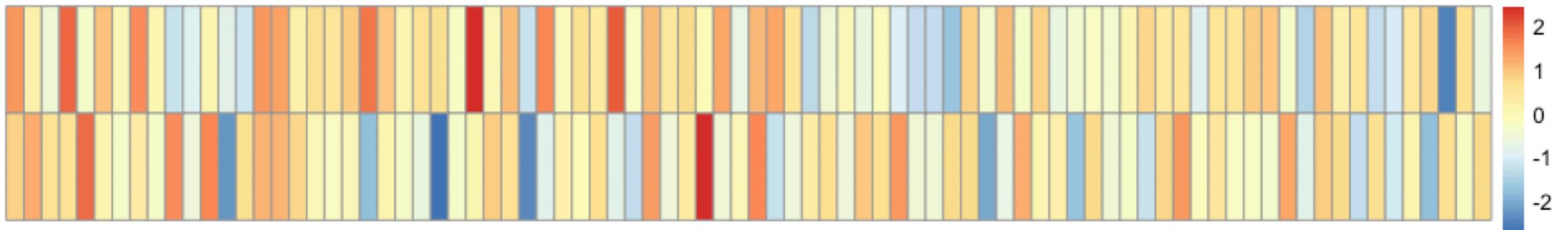
with  $\Sigma = WW' + \Psi$

(Maximization Step)

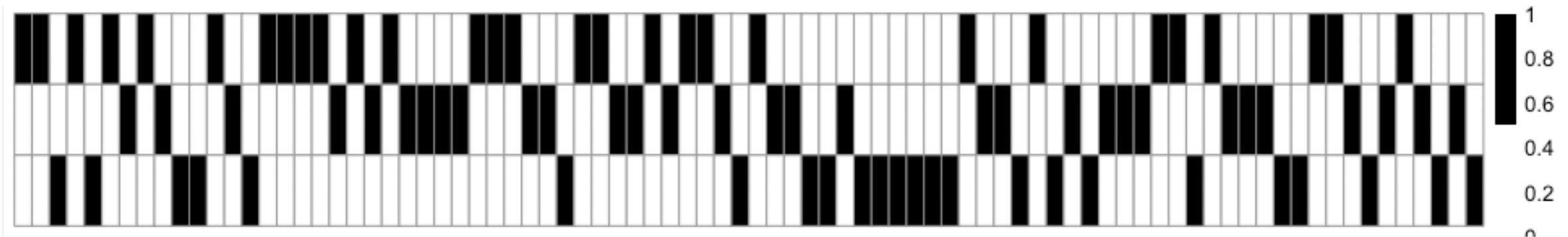
Estimate  $(W^{(t+1)}, \Psi^{(t+1)})$   
using  $(Z^{(t)}, X)$

$$\Psi^{(t+1)} = \frac{1}{n} \text{diag} (XX' - W^{(t)}E[Z^* | X]X)$$
$$W^{(t+1)} = (XE[Z^* | X]')(E[Z^*Z^{*'} | X])^{-1}$$

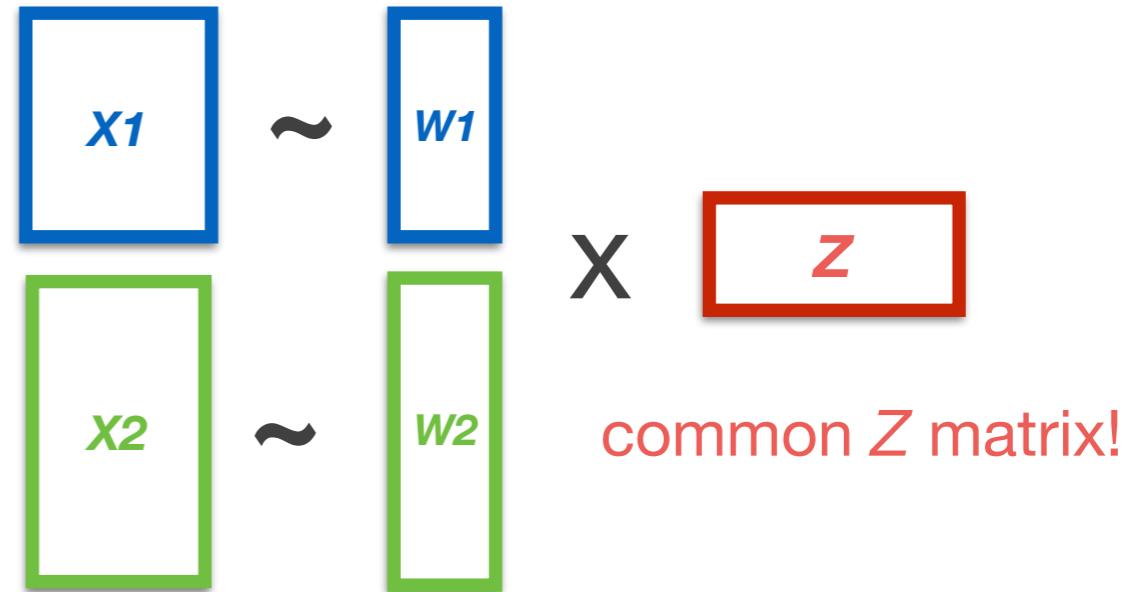
Inferred posterior probability  $E[Z^*|X]$  (for  $k=2$ )



Cluster indicator for  $k+1=3$  clusters



$$\begin{aligned} X_1 &= W_1 Z + \epsilon_1 \\ X_2 &= W_2 Z + \epsilon_2 \\ &\vdots \\ X_p &= W_p Z + \epsilon_p \end{aligned}$$



- Different types of data (binary, count data, continuous data,...) can be taken into account using different conditional probabilities

- $X_i$  is binary: **logistic** regression

$$\log \frac{P(X_i|Z)}{1 - P(X_i|Z)} = \alpha_i + \beta_i Z$$

- $X_i$  is count data: **Poisson** regression

$$\log(\lambda(X_i|Z)) = \alpha_i + \beta_i Z$$

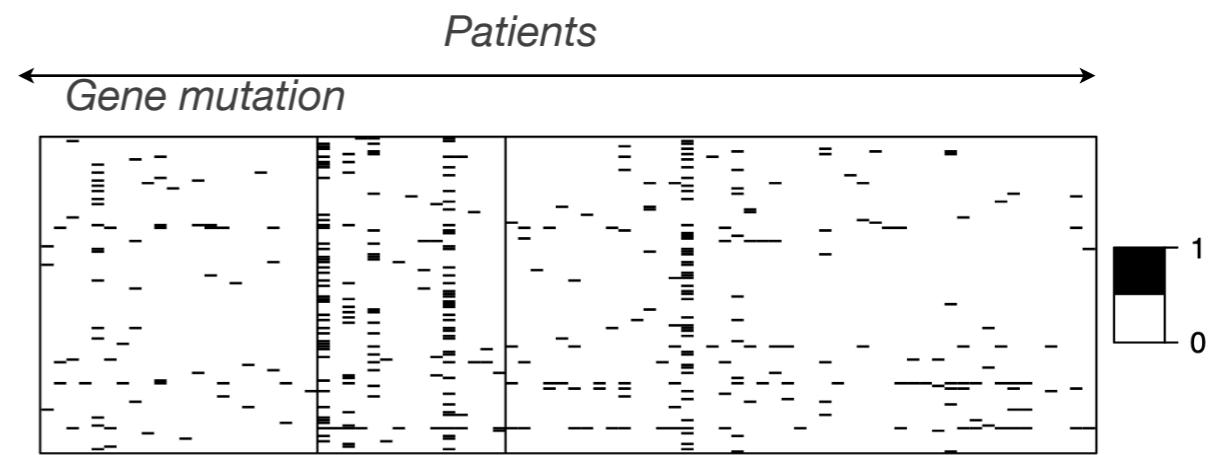
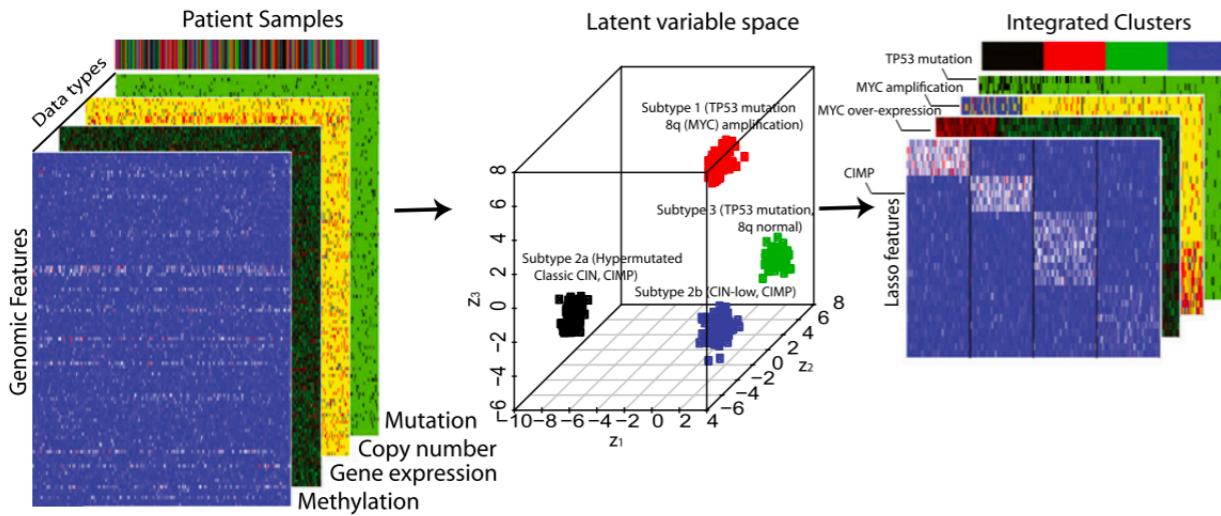
- $X_i$  is continuous: **linear** regression

$$X_i = \alpha_i + \beta_i Z + \epsilon_i$$

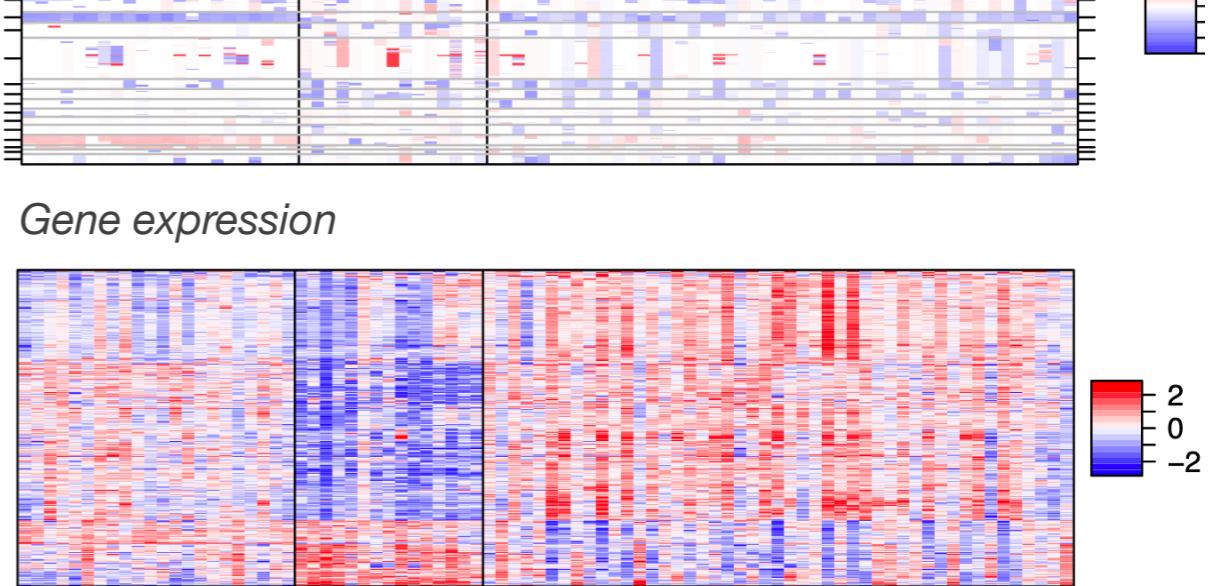
# iCluster



Medizinische Fakultät Heidelberg



- Application: **TCGA glioblastoma datasets**
  - gene mutations**  
(120 genes x 84 patients)
  - copy-number alterations**  
(5512 regions x 84 patients)
  - gene expression**  
(1740 top variable genes x 84 patients)

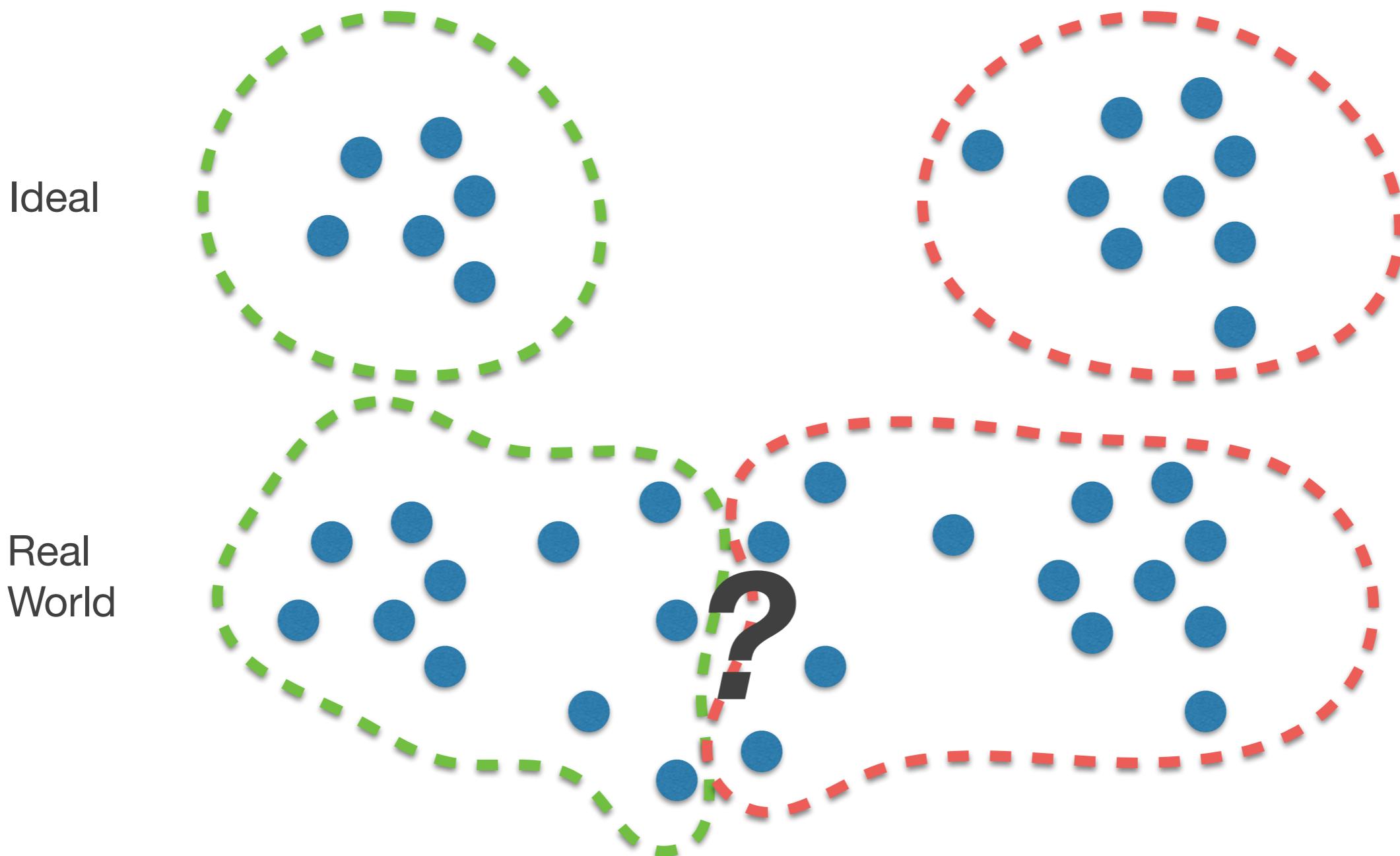


[Olshen et al., PNAS 2013]

# Limitations of Clustering



Medizinische Fakultät Heidelberg

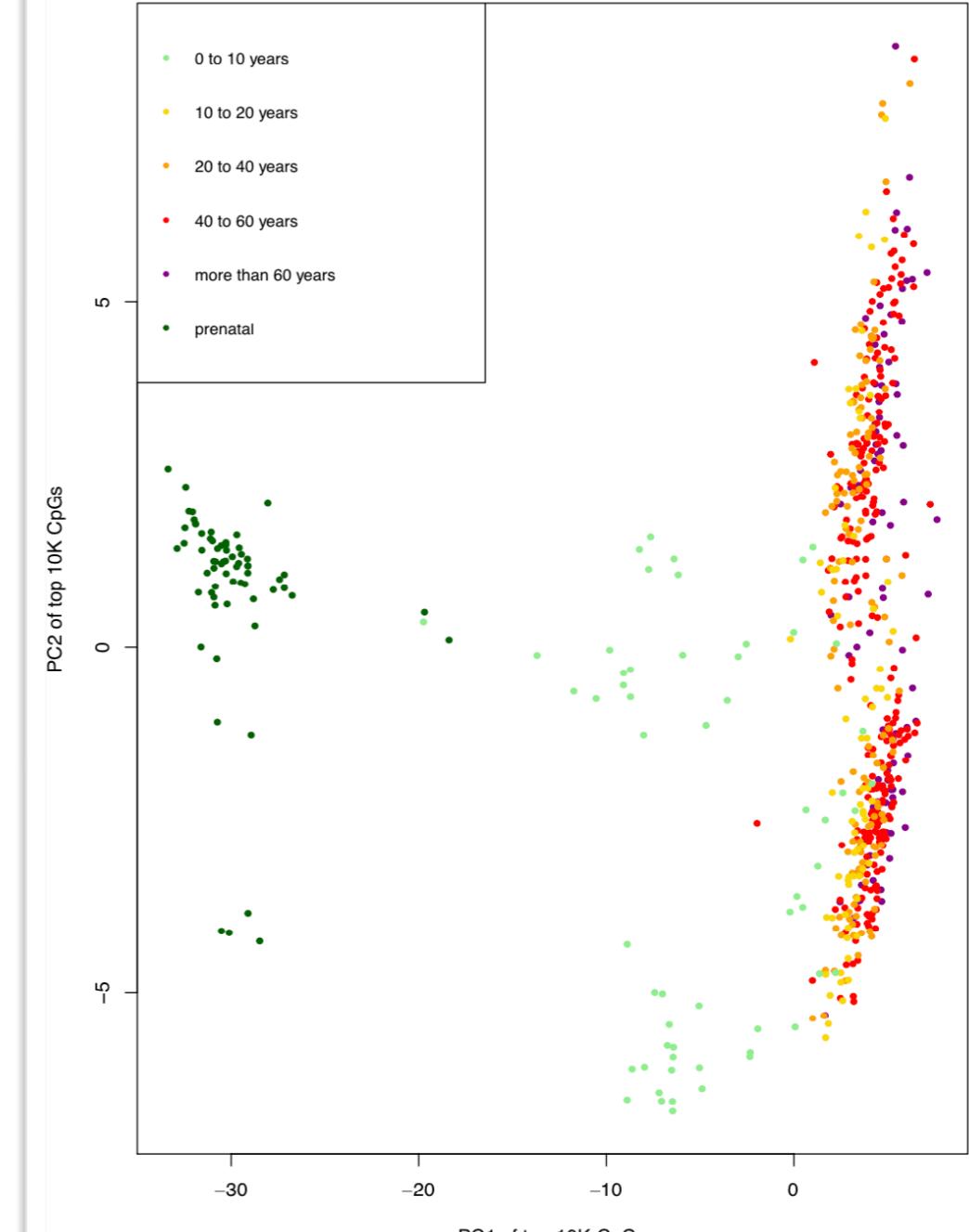


We need methods allowing a “fuzzy” assignment of samples  
clusters → signatures

# Principal component analysis



- Dataset have a very high dimensionality (e.g. number of genes)
- Need to reduce this large number of dimensions to a smaller number of relevant variables
- Relevant variables = variables which carry most of the information ( or variance) of a dataset
- These new variables are **orthogonal**
- Goal:
  - identify **directions** in the data corresponding to **biological effects**



Example of DNA methylation of blood samples in patient cohort (Jana Dalhoff)  
data matrix : 400.000 CpG positions / 250 patients

# Correlation structure



Medizinische Fakultät Heidelberg

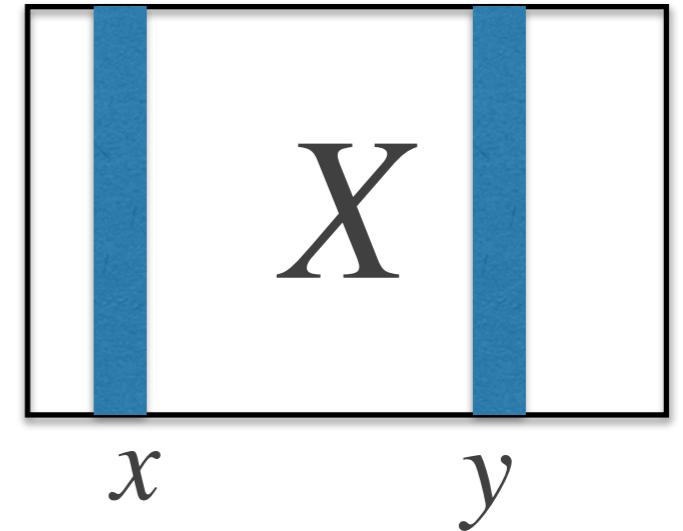
- if two variables are **strongly correlated**, they are partly redundant:  
knowing the variation of one, you have information about how the second variables changes  
→ having 2 variables does not add much information w.r.t. a single variable
- if two variables have **little correlation**, each variable carries information not contained in the other  
→ we need to keep these 2 variables to have information about the dataset
- **The more diagonal a correlation matrix is, the more information is revealed by the variables**

# Correlation and covariance

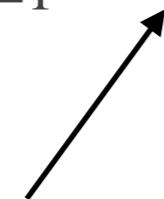


Medizinische Fakultät Heidelberg

$$cov(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N} X_c' \cdot X_c$$



$$cor(x, y) = \frac{1}{N} \sum_{i=1}^N \frac{(x_i - \bar{x})}{\sigma_x} \frac{(y_i - \bar{y})}{\sigma_y} = \frac{1}{N} X_{cs}' \cdot X_{cs}$$

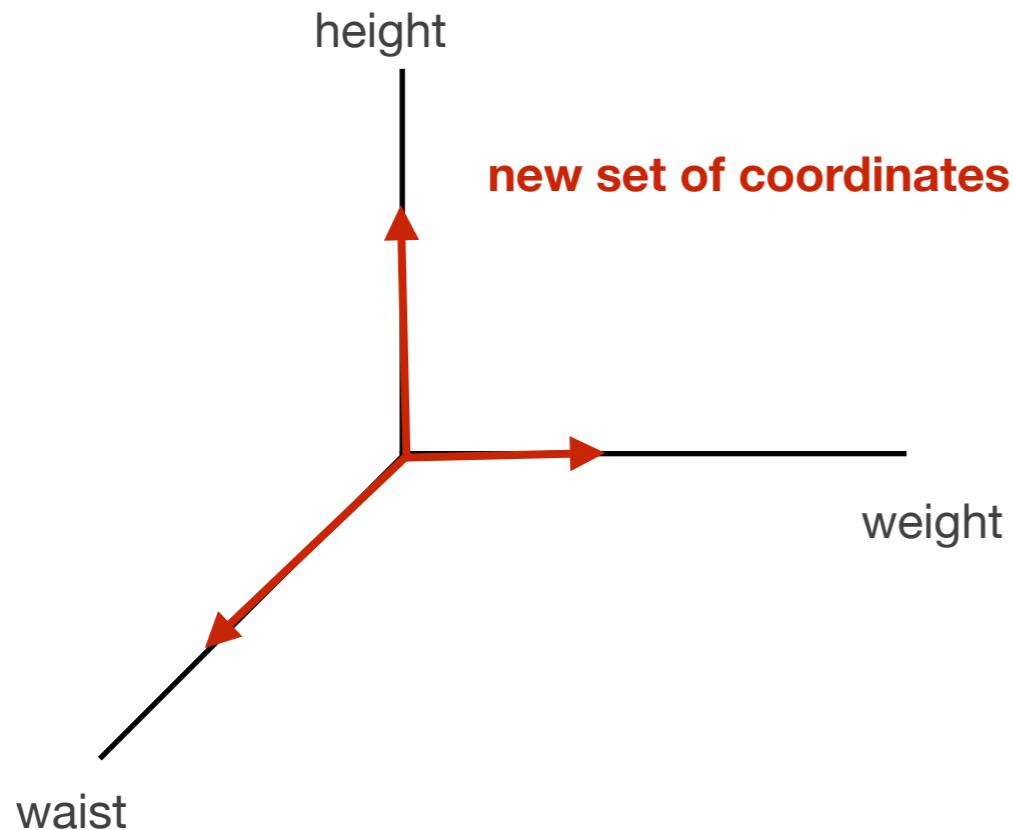


Z-transformation

# Defining new variables



Medizinische Fakultät Heidelberg



- We want to express the dataset in a new set of coordinates
- for each of these coordinate systems, the correlation matrix will change
- **Goal:** find the rotation that makes the correlation matrix diagonal

→ *matrix diagonalization !*

# Matrix diagonalization



Medizinische Fakultät Heidelberg

1. Consider the correlation matrix  $A$

	age	height	chol	waist	weight
age	1.00000000	-0.09479919	0.23990232	0.15255761	-0.06269027
height	-0.09479919	1.00000000	-0.05853973	0.05661532	0.25298143
chol	0.23990232	-0.05853973	1.00000000	0.11245805	0.05932074
waist	0.15255761	0.05661532	0.11245805	1.00000000	0.84955930
weight	-0.06269027	0.25298143	0.05932074	0.84955930	1.00000000

2. Determine its  $n$  eigenvalues and  $n$  eigenvectors and build the  $n \times n$  matrix  $W$  from all the  $n$  eigenvectors

```
$values  
[1] 1.9201374 1.3081302 0.9011191 0.7635241 0.1070892  
  
$vectors  
[,1] [,2] [,3] [,4] [,5]  
[1,] -0.0782536 0.66340112 -0.1957637 0.70124582 -0.15396822  
[2,] -0.2139712 -0.41884235 -0.8557896 0.16410053 0.13957979  
[3,] -0.1338086 0.59835882 -0.3893703 -0.68726253 0.01108988  
[4,] -0.6768556 0.08069999 0.2542954 0.06898591 0.68258976  
[5,] -0.6870622 -0.14115337 0.1141285 -0.06508820 -0.70054229
```

3. Compute  $S = W' \cdot A \cdot W$

$S$  is a diagonal matrix with the eigenvalues as entries!

```
[,1] [,2] [,3] [,4] [,5]  
[1,] 1.92 0.000 0.000 0.000 0.000  
[2,] 0.00 1.308 0.000 0.000 0.000  
[3,] 0.00 0.000 0.901 0.000 0.000  
[4,] 0.00 0.000 0.000 0.764 0.000  
[5,] 0.00 0.000 0.000 0.000 0.107
```

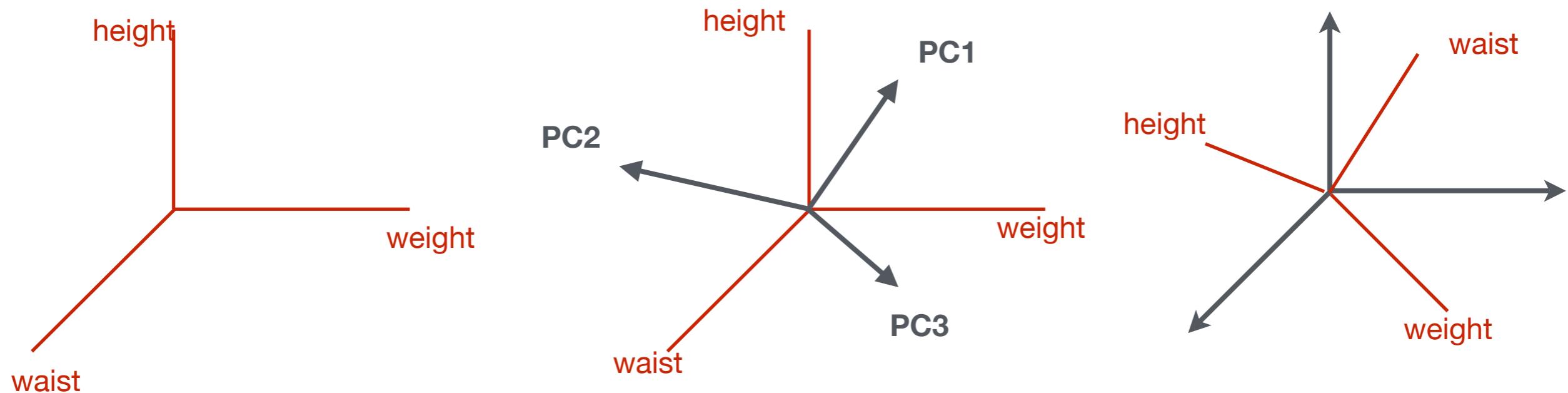


# New coordinate system

$$S = W' \cdot A \cdot W = W' \cdot \left( \frac{1}{N} X'_{cs} \cdot X_{cs} \right) \cdot W$$

$$S = \frac{1}{N} (X_{cs} \cdot W)' \cdot (X_{cs} \cdot W)$$

- $W$  is the **rotation matrix** transforming the initial variables into new variables called principal components



# PCA as a matrix factorization problem

- PCA can be formulated in terms of a **singular value decomposition** problem
- SVD : decomposition of a matrix into the product of 3 matrices

$$X = U\Lambda V' = WH \quad \begin{aligned} W &= U \\ H &= \Lambda V' \end{aligned}$$

$U$  : unitary  $m \times m$  matrix  $UU' = I$

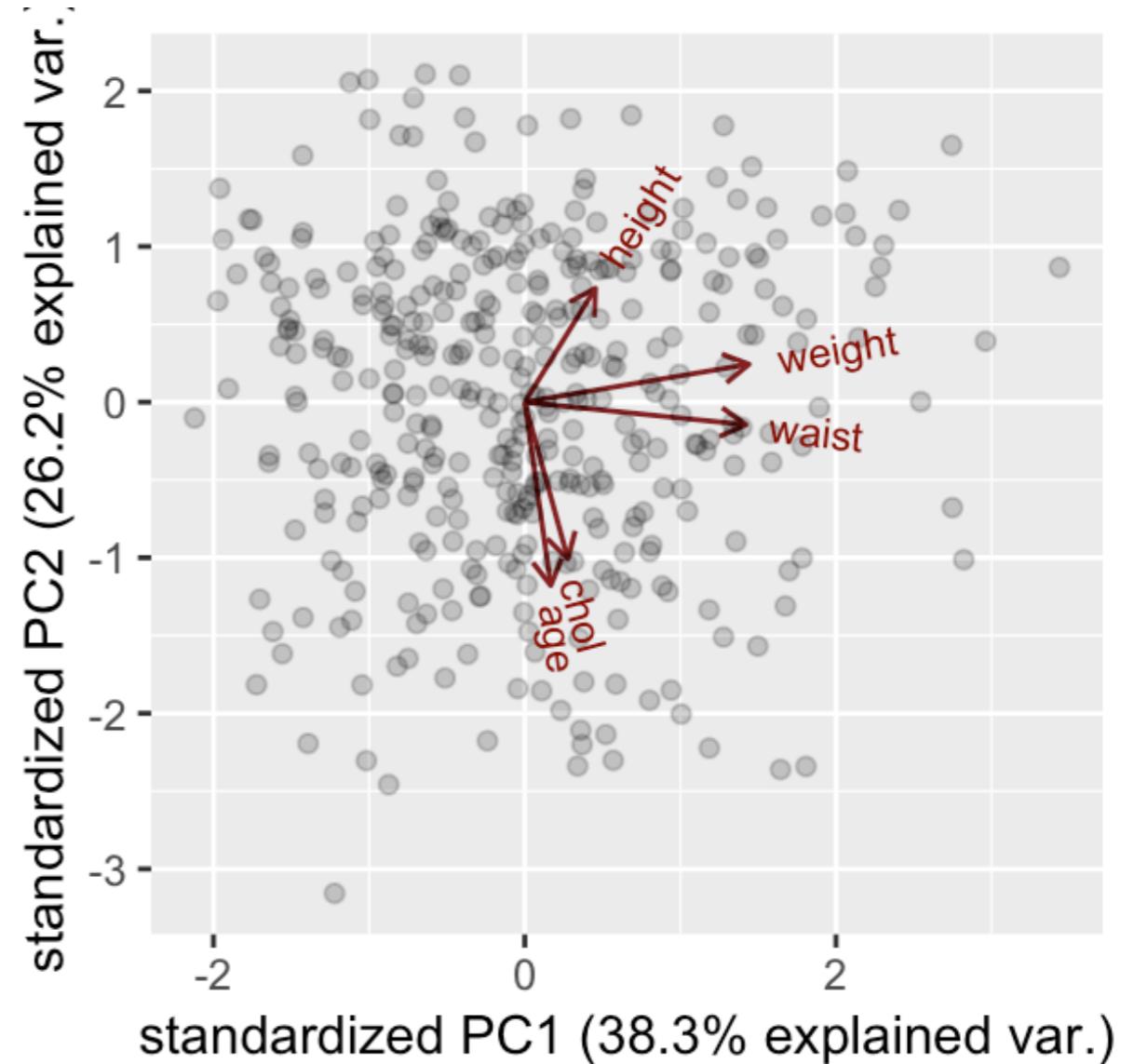
$V$  : unitary  $n \times n$  matrix  $VV' = I$

$\Lambda$  : rectangular diagonal  $m \times n$  matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & -3 \\ 0 & 0 & 0 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & -3 & 0 & 0 \end{bmatrix}$$

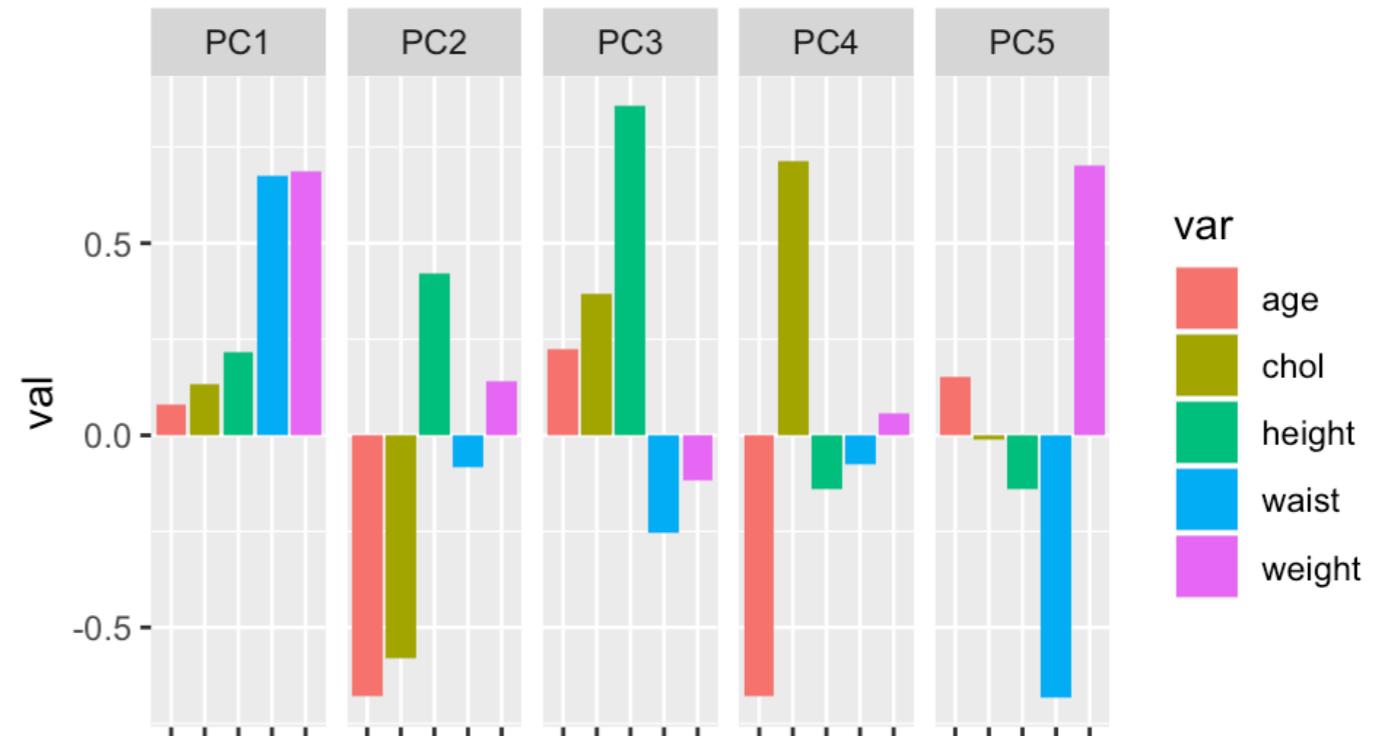
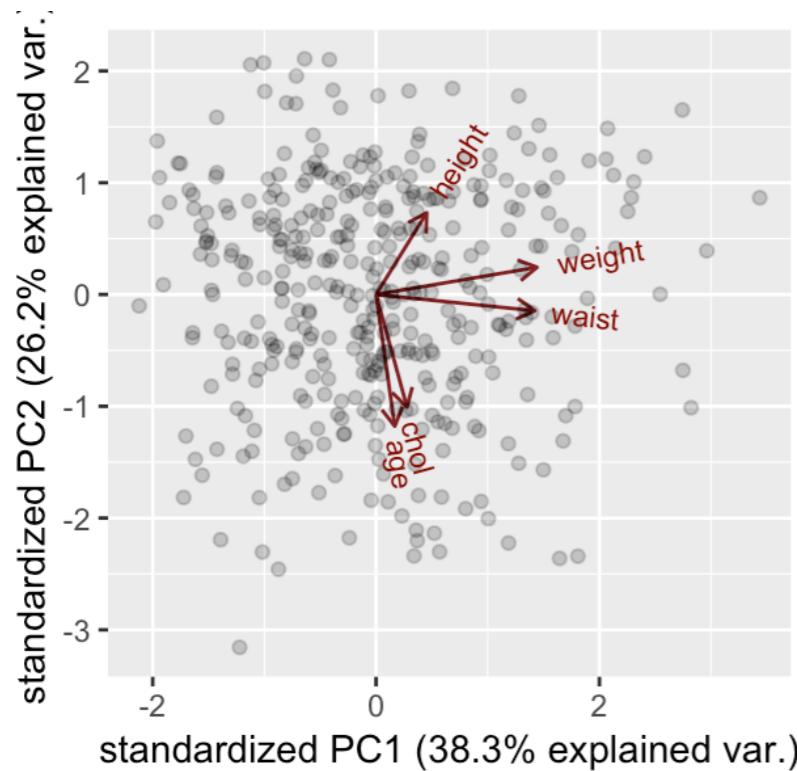
# PCA biplot

- each **dot** is a sample / patient
- new coordinate system is  $(PC_1, PC_2)$
- **Red arrows** indicate the contribution of each “old” coordinate to the PCs



# Principal components

$$PC_i = \alpha_i \cdot \text{age} + \beta_i \cdot \text{chol} + \gamma_i \cdot \text{height} + \delta_i \cdot \text{waist} + \epsilon_i \cdot \text{weight}$$

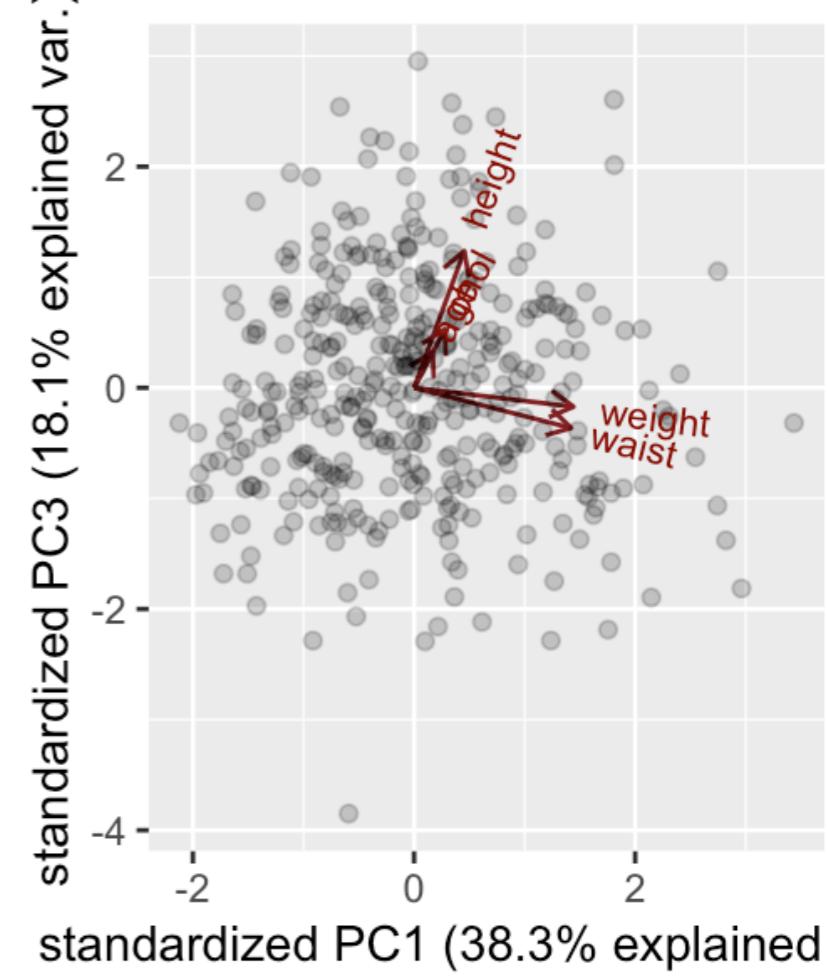
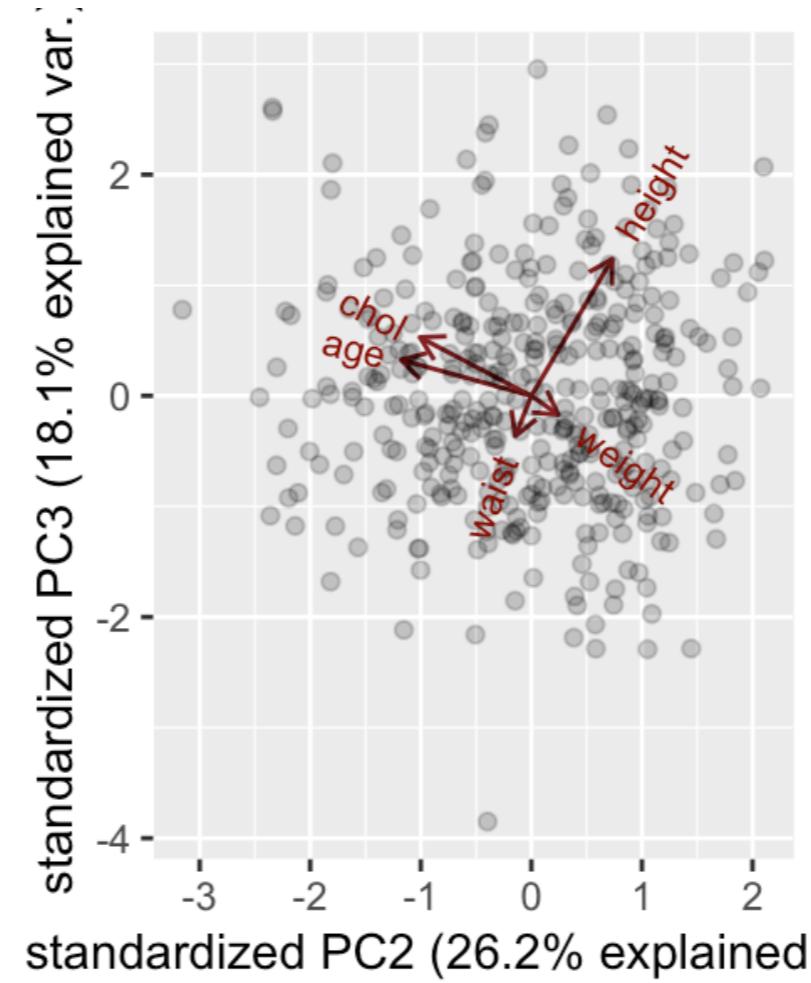
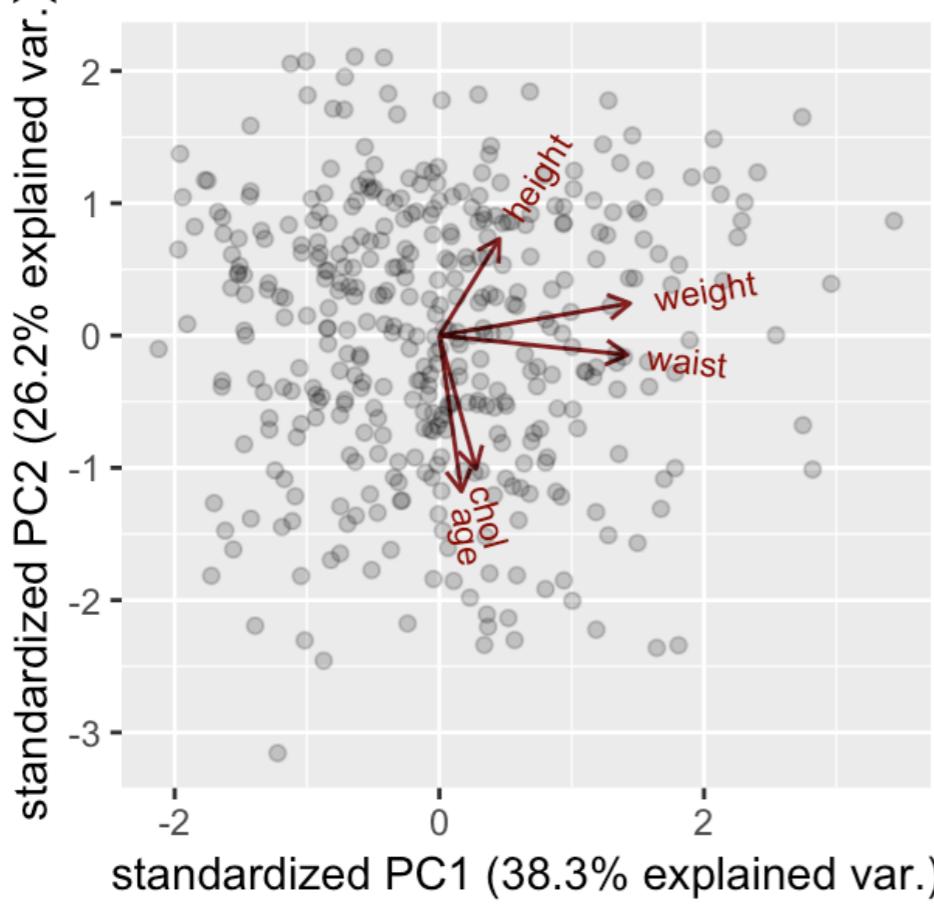


- contribution of each variable to the principal components (coefficients are called "*loadings*")
- some variables contribute in the same direction to some PCs (e.g. waist and height for PC1), but opposite to others (PC5)

# Principal components



Medizinische Fakultät Heidelberg

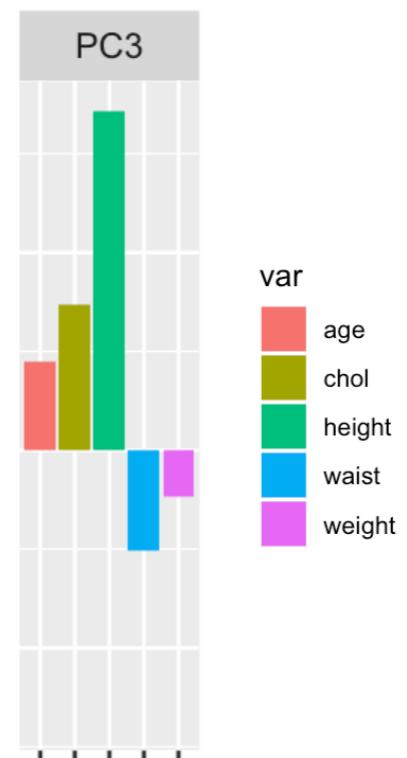
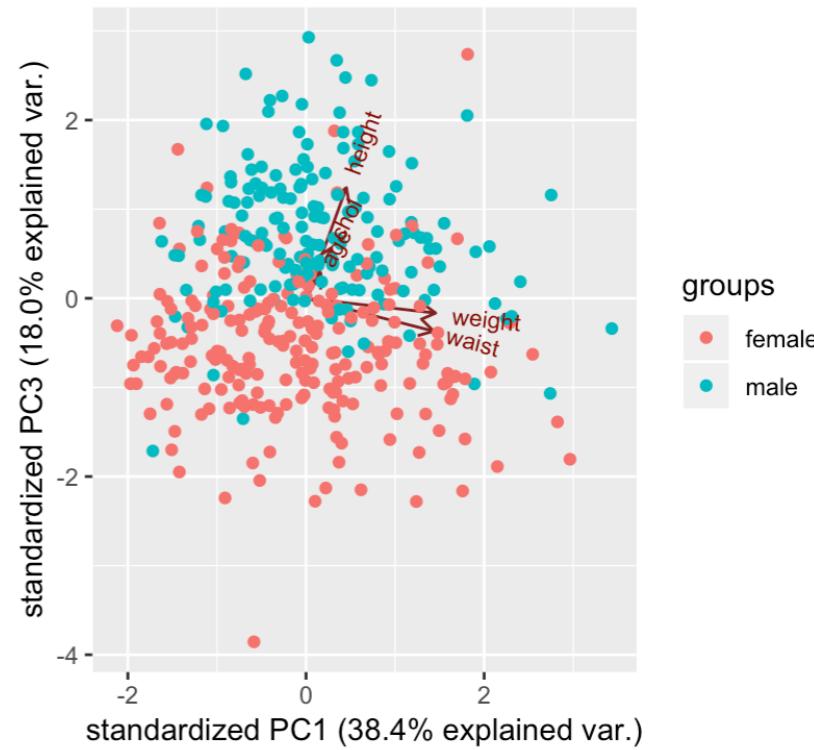


# Identifying interesting PCs



Medizinische Fakultät Heidelberg

- PC plots can highlight a new group structure
- Example: **PC3** seems very associated to gender
- indicates that a combination of height and cholesterol does separate men /women

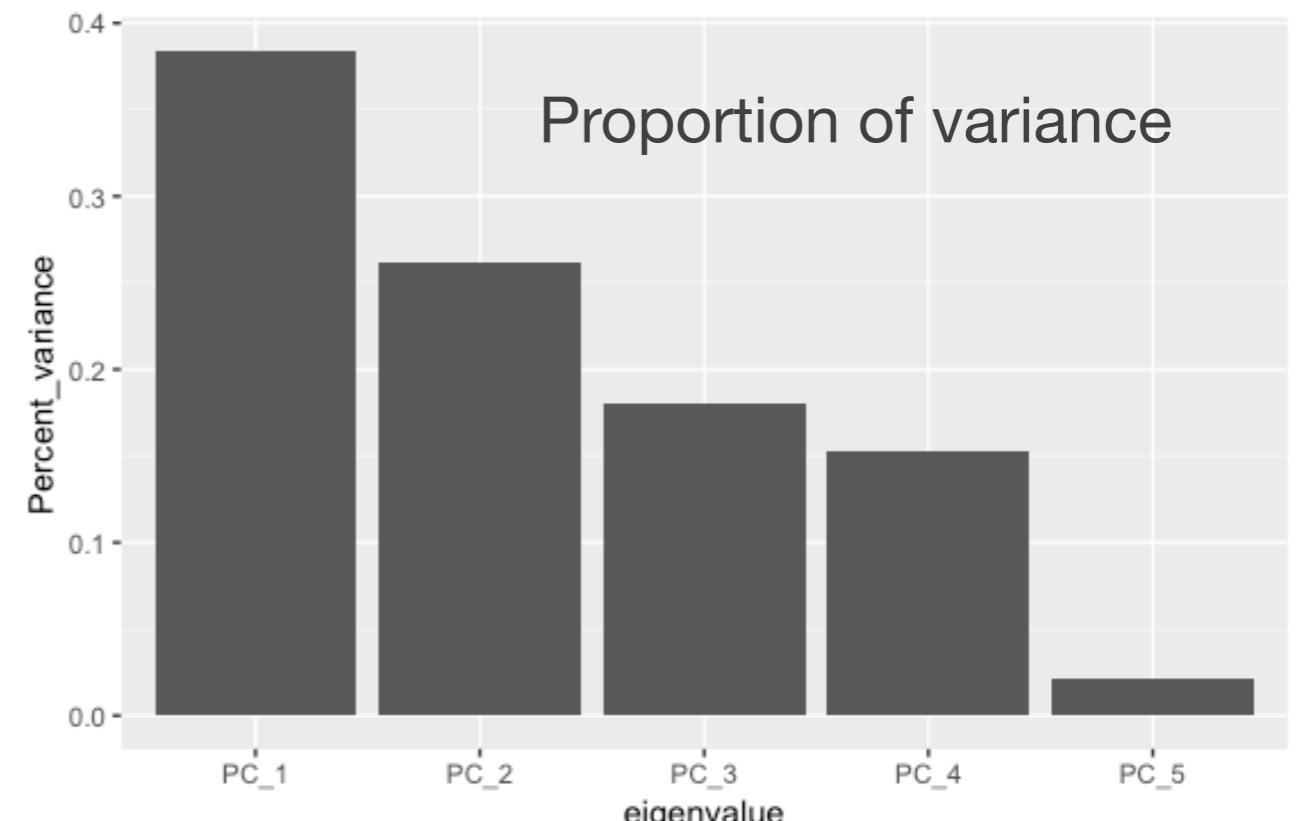
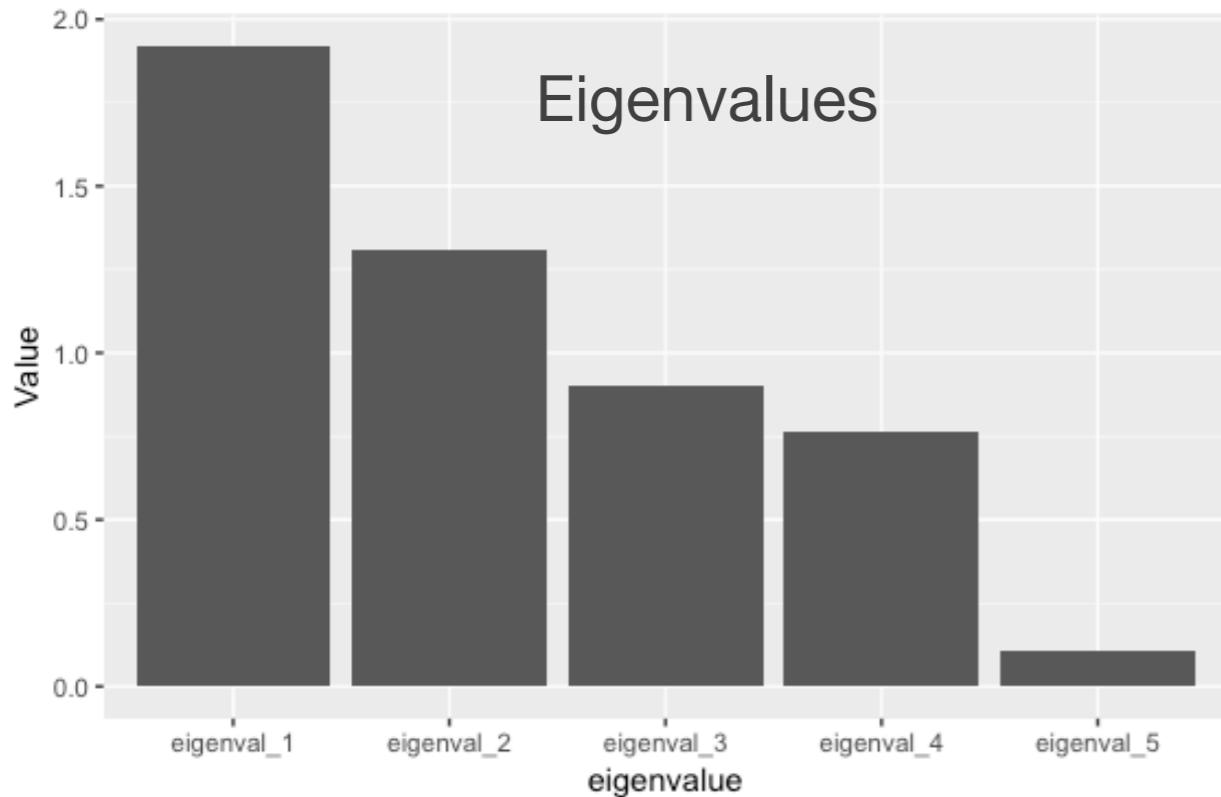


# Number of PCs?



Medizinische Fakultät Heidelberg

- Each PC explains some part of the total variance of the dataset
- This amount is proportional to the corresponding **eigenvalue**
- PCs are ordered by **decreasing eigenvalue** (hence variance)



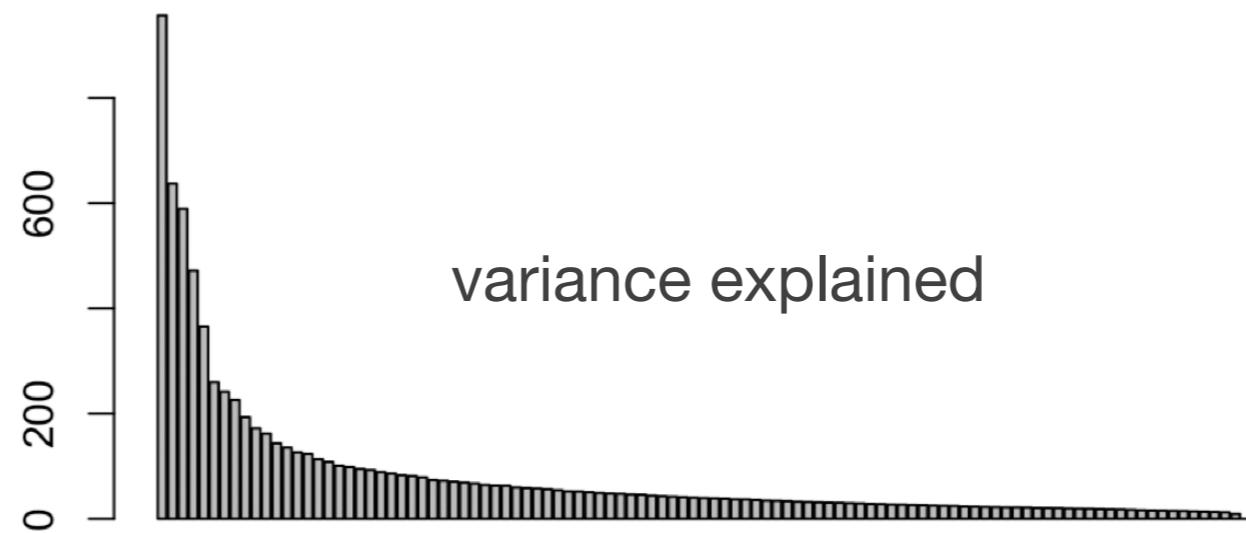
*Considering PC1 & PC2 explains  
63% of the total variance*

# Application to gene expression



Medizinische Fakultät Heidelberg

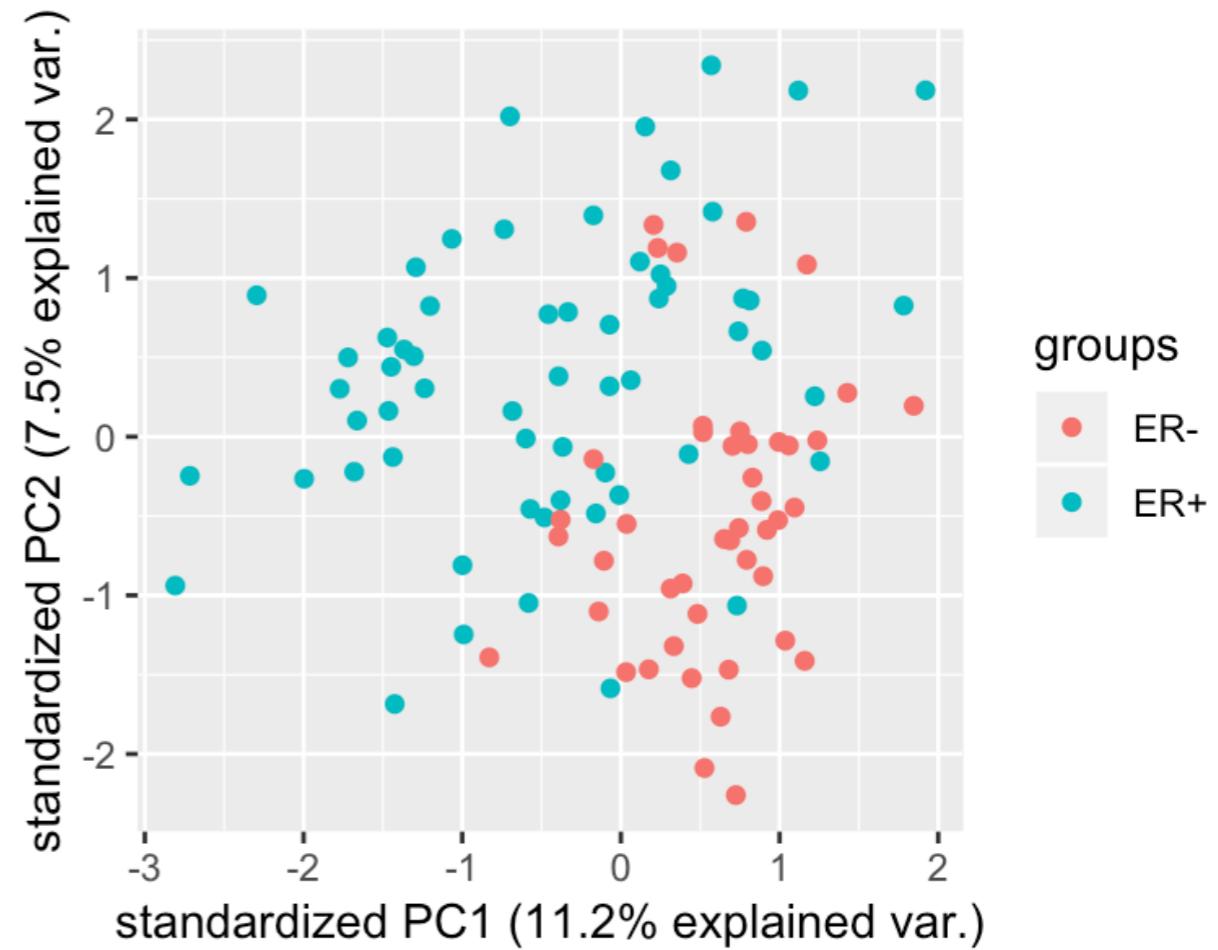
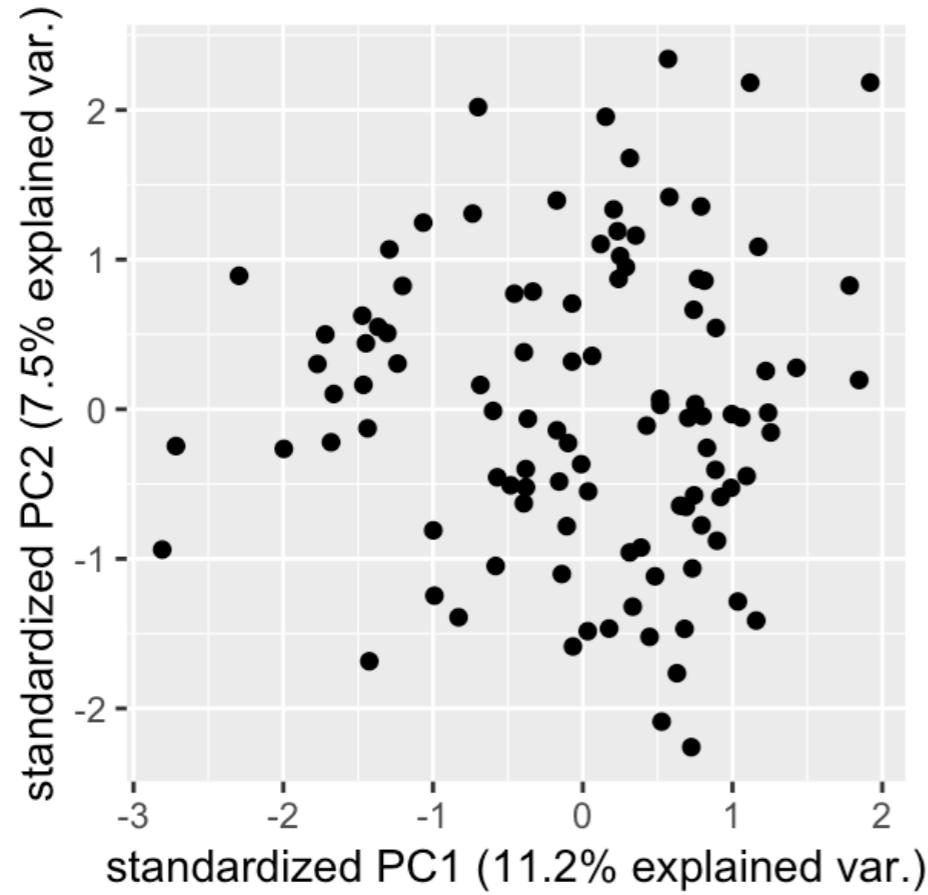
- Gene expression dataset of **breast cancer patients**
- 2 groups: ER+ and ER- patients [M. Ringner, Nature Biotech. (2008)]
- Dimension:  $k = 105$  patients /  $n = 8534$  genes (here:  $n \gg k$ )
- pre-processing:
  - **scale** the gene expression across patients
  - **center** the gene expression across patients
- How many principal components do we get?  
→ **k-1** (this has to do with the rank of the data matrix)



# Application to gene expression



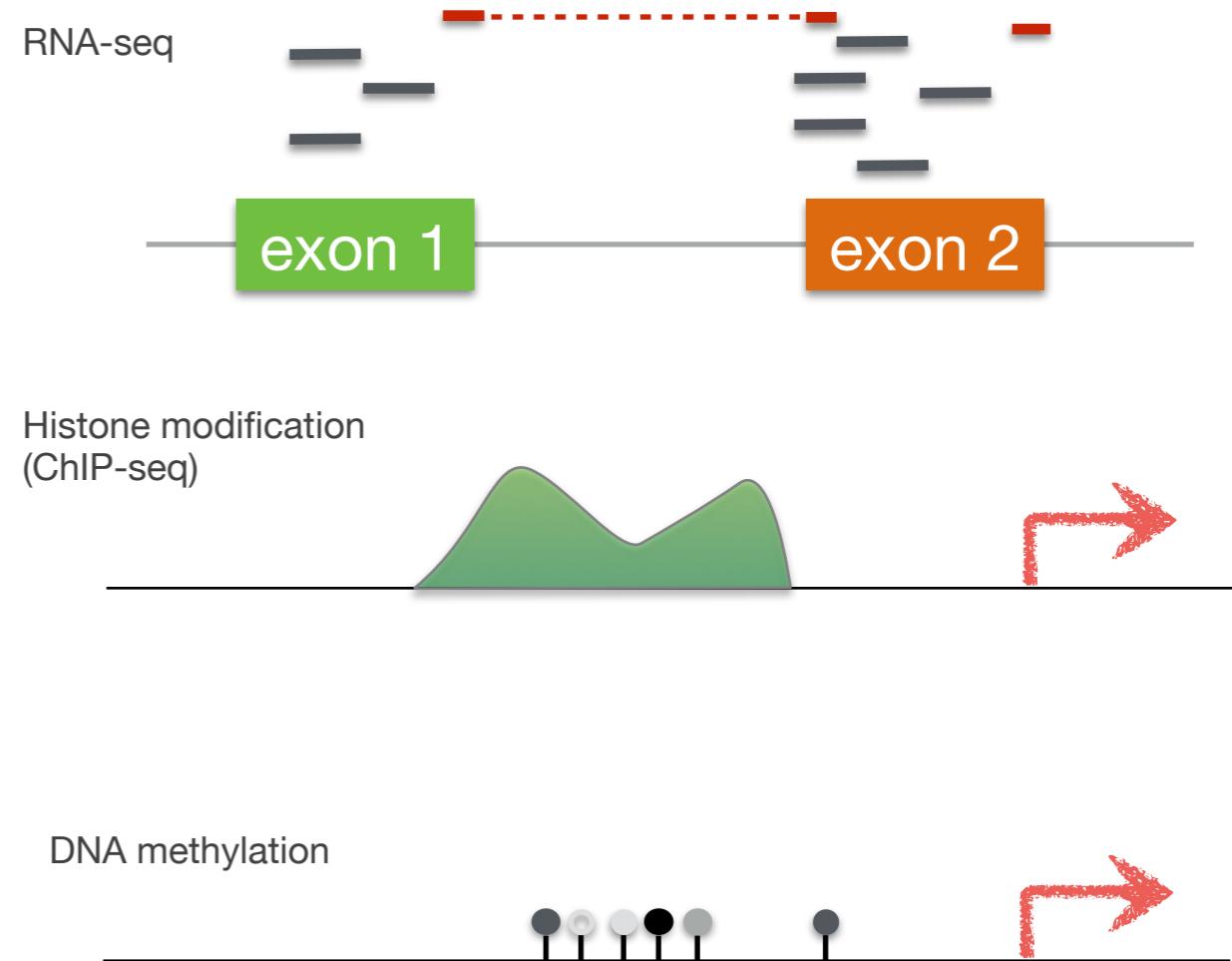
Medizinische Fakultät Heidelberg



- PC1 separates ER+ from ER- patients

# Non-negative matrix factorization

- Most datasets in modern genomics are by essence non-negative
- Read counts in RNA-seq
- Methylation  $b$ -values in DNA methylation arrays
- Integrated signal over genomic regions



***we can apply parts-base decomposition of the data***

# Non-negative matrix factorization

$$X \sim WH \quad \text{with } X \geq 0, W \geq 0, H \geq 0$$

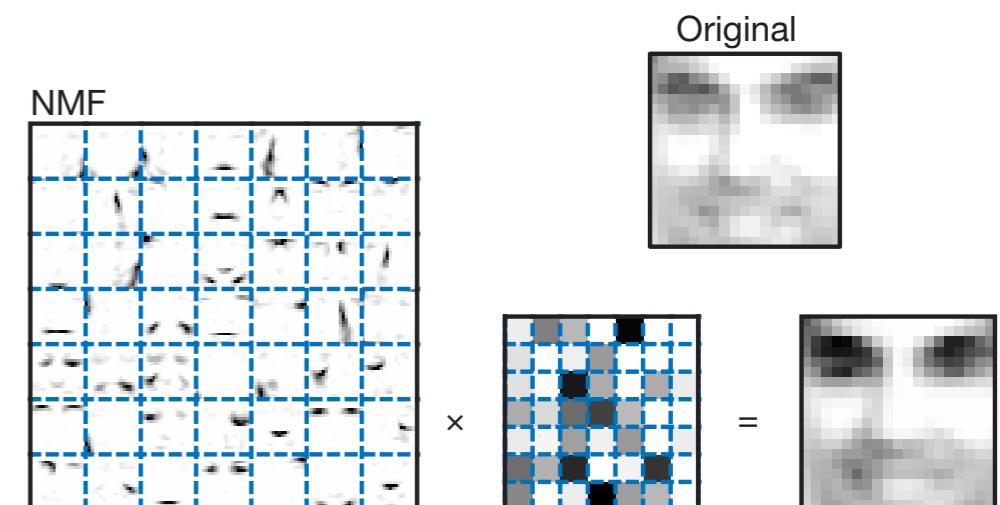
$X : N \times M$  matrix

$N =$  number of features (genes, regions,...)

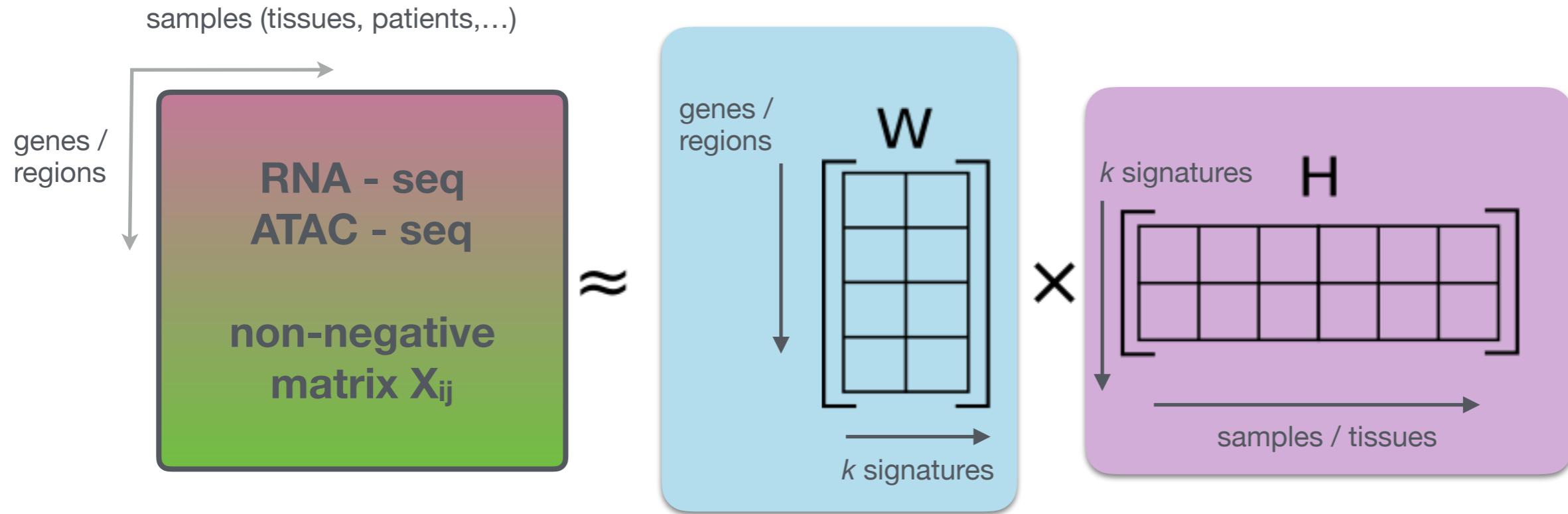
$M =$  number observations (patients,samples,...)

NMF in essence similar to PCA,  
but non-negativity implies

- a better **interpretability** of the signatures
- a natural **sparseness** of the decomposition



[Lee, Seung 1999]



$V$  : original data matrix

columns of  $W$  :  $k$  **signatures** (genes, regions,...)

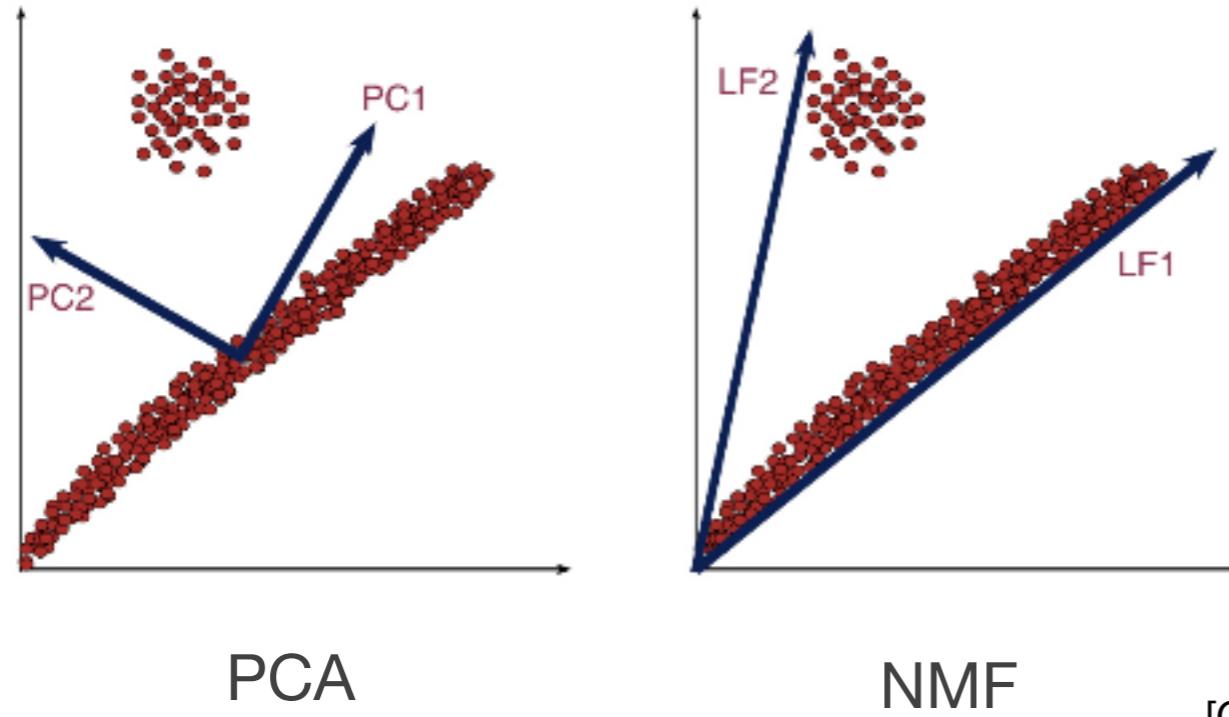
columns of  $H$  : **exposures** to the  $k$  signatures

→ **Genomic signatures + features of the signature**

# NMF vs. PCA



Medizinische Fakultät Heidelberg



*because of the non-negative constraint, only point inside the cone can be reconstructed using the basis vectors*

[Casalino, Buono, Mencar]

- PCA defines orthogonal directions explaining most variance
- NMF signatures (or *latent factors LF*) define the hypercone containing all data points
- There is **no natural ranking of the NMF-signatures** (unlike PCs); choice of the number of signatures is crucial!

# NMF vs. PCA



Medizinische Fakultät Heidelberg



Figure 4.5: Base images of dataset  $\mathbf{D}_{\text{face}}$  after applying the PCA



Part are more easily interpretable in NMF

[Nikolaus]

# Finding the best NMF solution



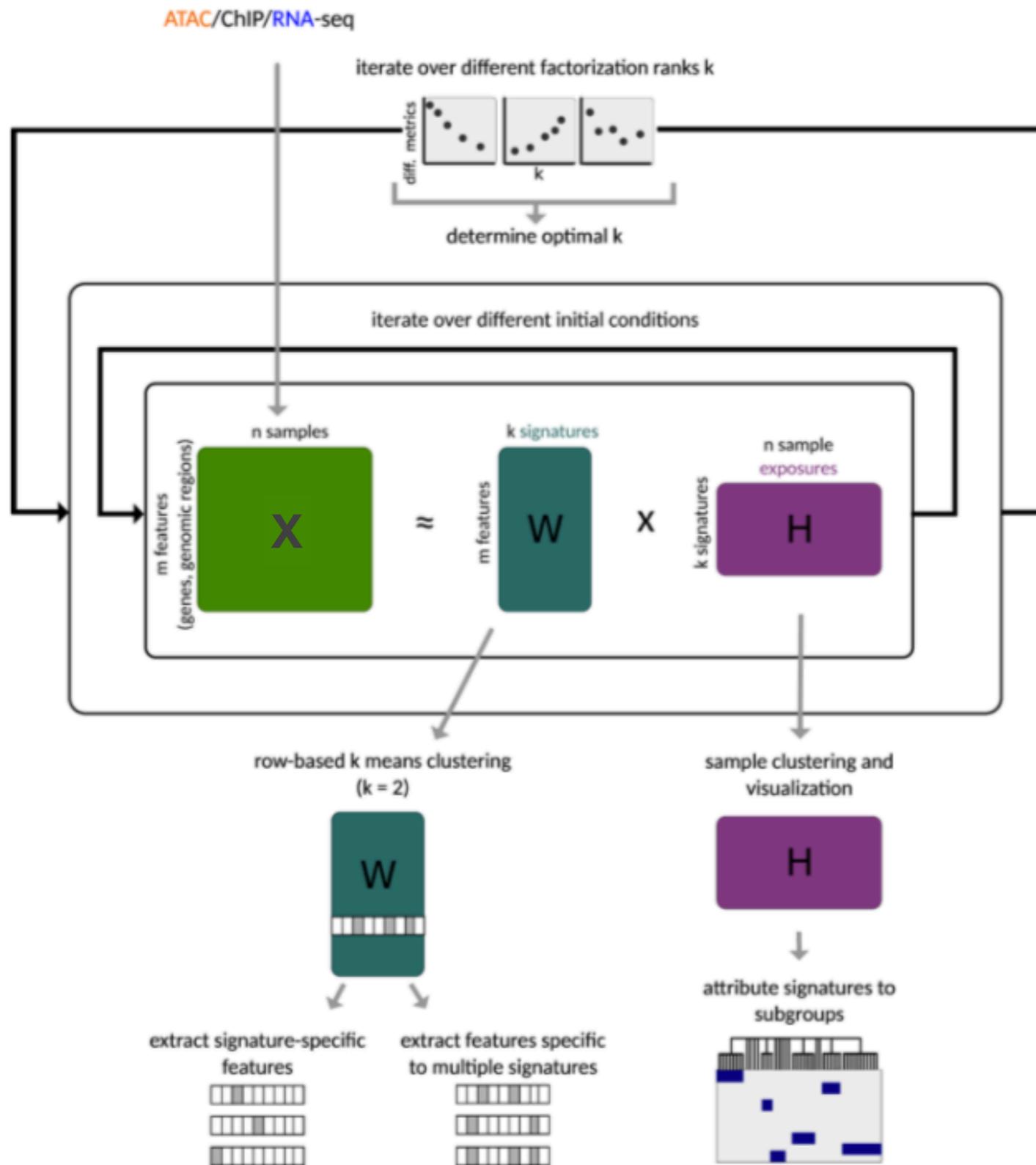
Medizinische Fakultät Heidelberg

**Goal: minimize**  $\|X - WH\|_{Froebenius}^2 = \sum_i \sum_j |(x - wh)_{ij}|^2$

- Iterative algorithm for finding best decomposition:
  - Random initialization of matrices  $H$  and  $W$
  - Multiplicative update rules

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{X_{i\mu}}{(WH)_{i\mu}} H_{a\mu}$$
$$H_{a\mu} \leftarrow H_{a\mu} \sum_i W_{ia} \frac{X_{i\mu}}{(WH)_{i\mu}}$$
$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$$

# Implementation



- Iteration over **update equations** (~ 10.000s, inner iteration)
- Iterate of set of **initial conditions** (~ 10s, outer iteration)
- Iterate over different **number of signatures** to be extracted

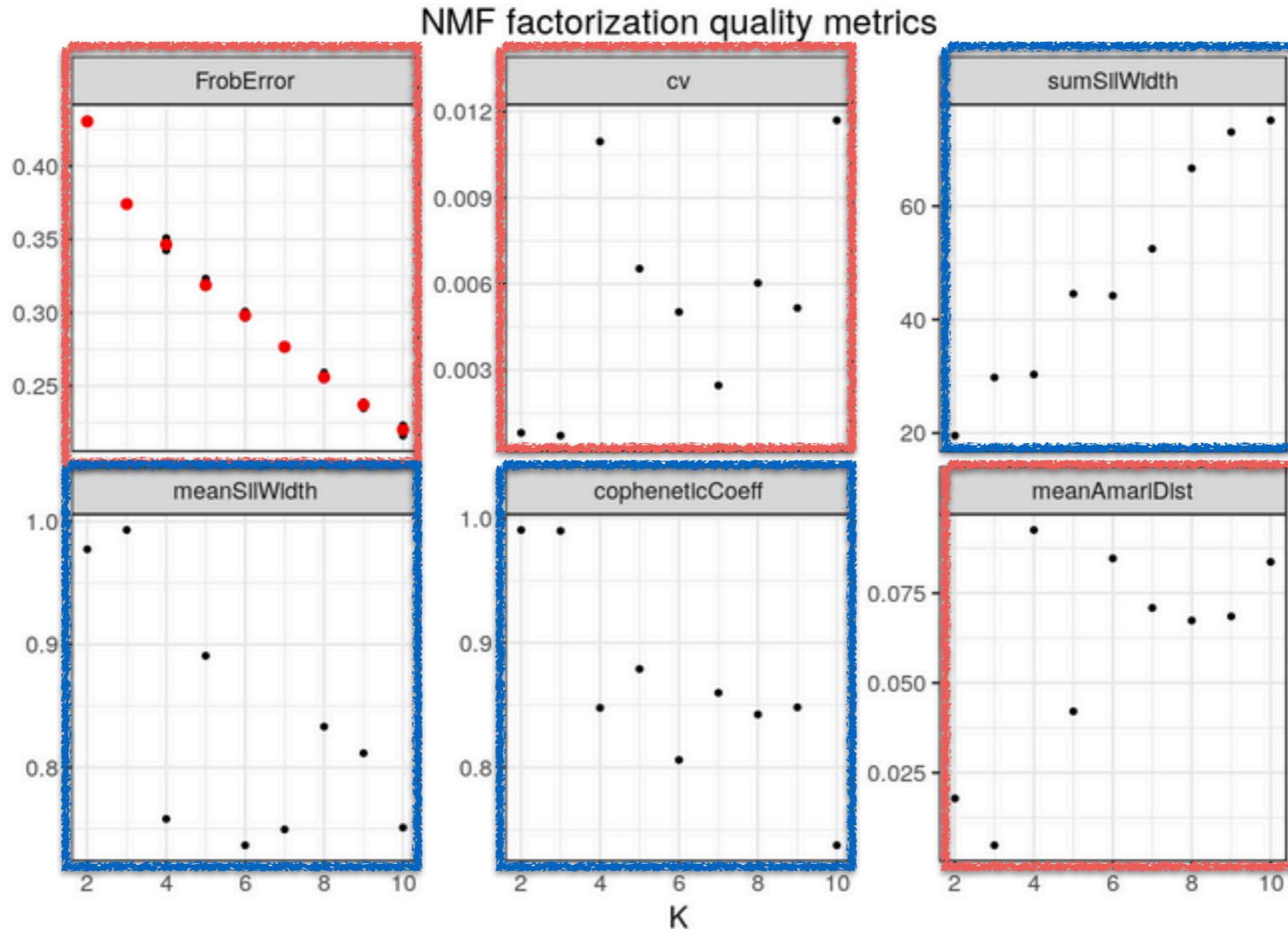
# How to choose k?

- Accuracy of matrix decomposition: **how well does  $WH$  represent  $V$ ?**
  - **Froebenius error** should be small
  - **Amari distance** should be small
- Stability of solutions: **how variable are the solutions using different random initializations?**
  - **Coefficient of variation** should be small
- Groups of samples should be homogeneous: **how well does each sample belong to its group?**
  - **Silhouette coefficient** should be large
- Clustering should well represent the original data
  - **Cophenetic coefficient** should be large

# How to choose k?



Medizinische Fakultät Heidelberg



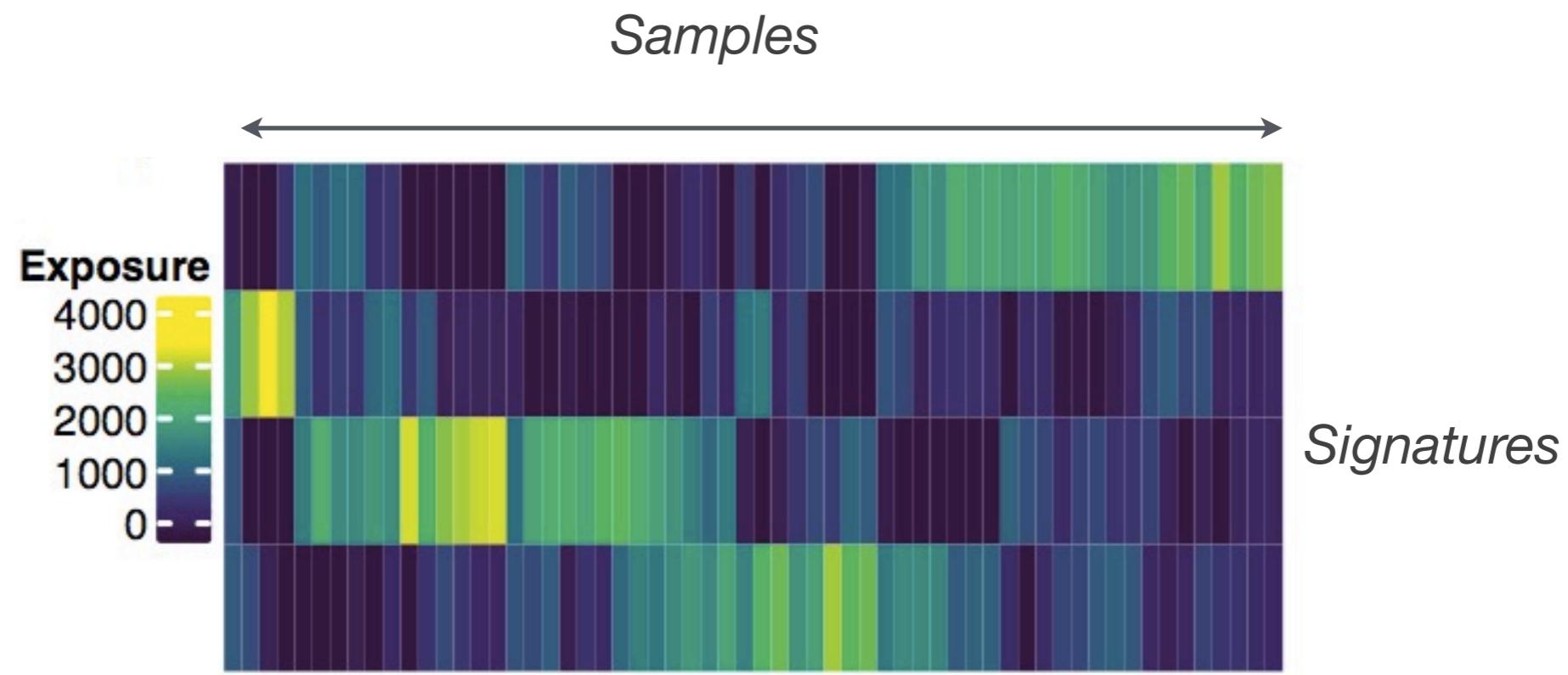
$k = 5$  appears to be a good choice

# Exposure matrix H

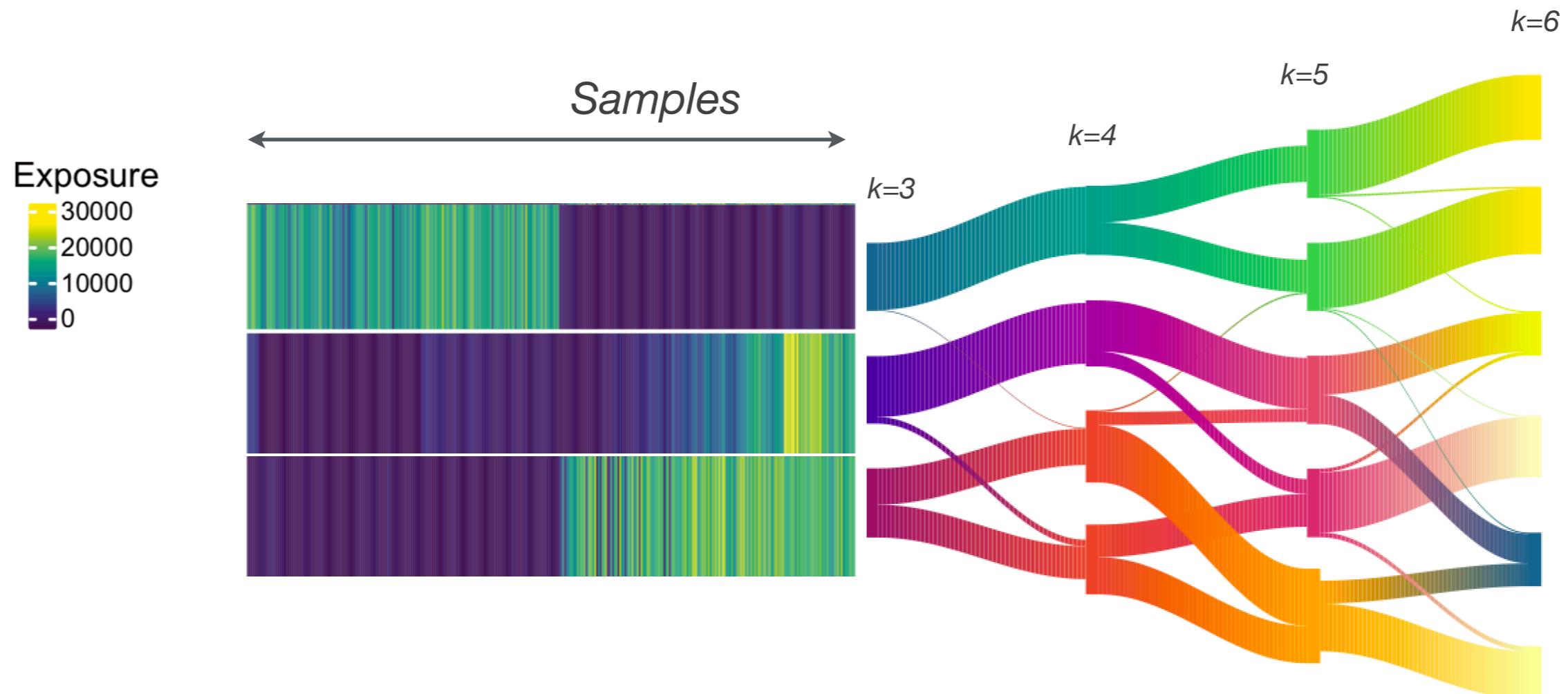


Medizinische Fakultät Heidelberg

- A sample can have “exposure” to multiple signatures
- Gradient of exposures (unlike hard clustering)
- sparseness: many coefficients are (almost) 0 in W and H matrix

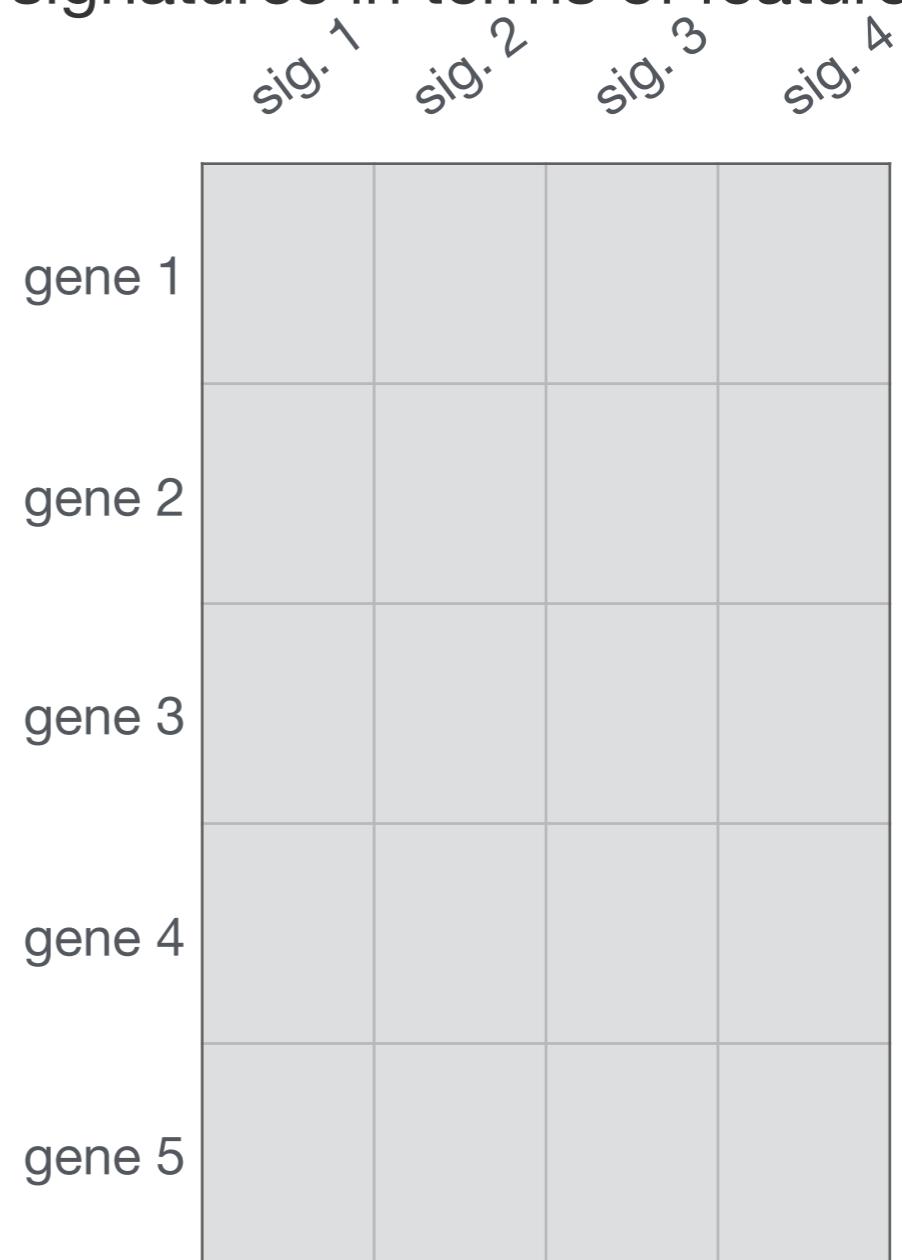


# Stability of signatures



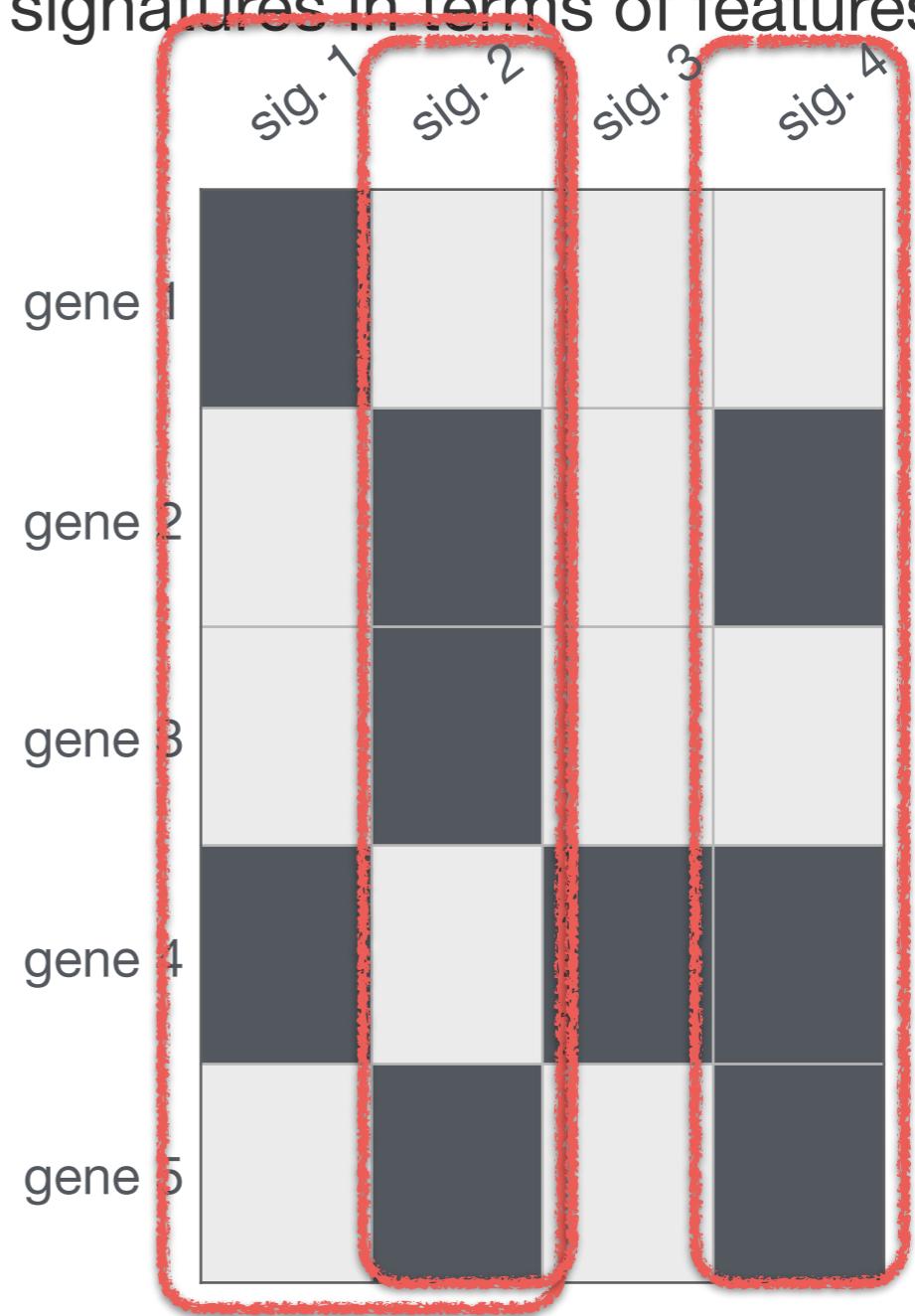
# Signature matrix $\mathbf{W}$

- the  $\mathbf{W}$  matrix gives the “definition” of the signatures in terms of features contributing
- applying k-means ( $k=2$ ) to each row of the  $\mathbf{W}$  matrix



# Signature matrix W

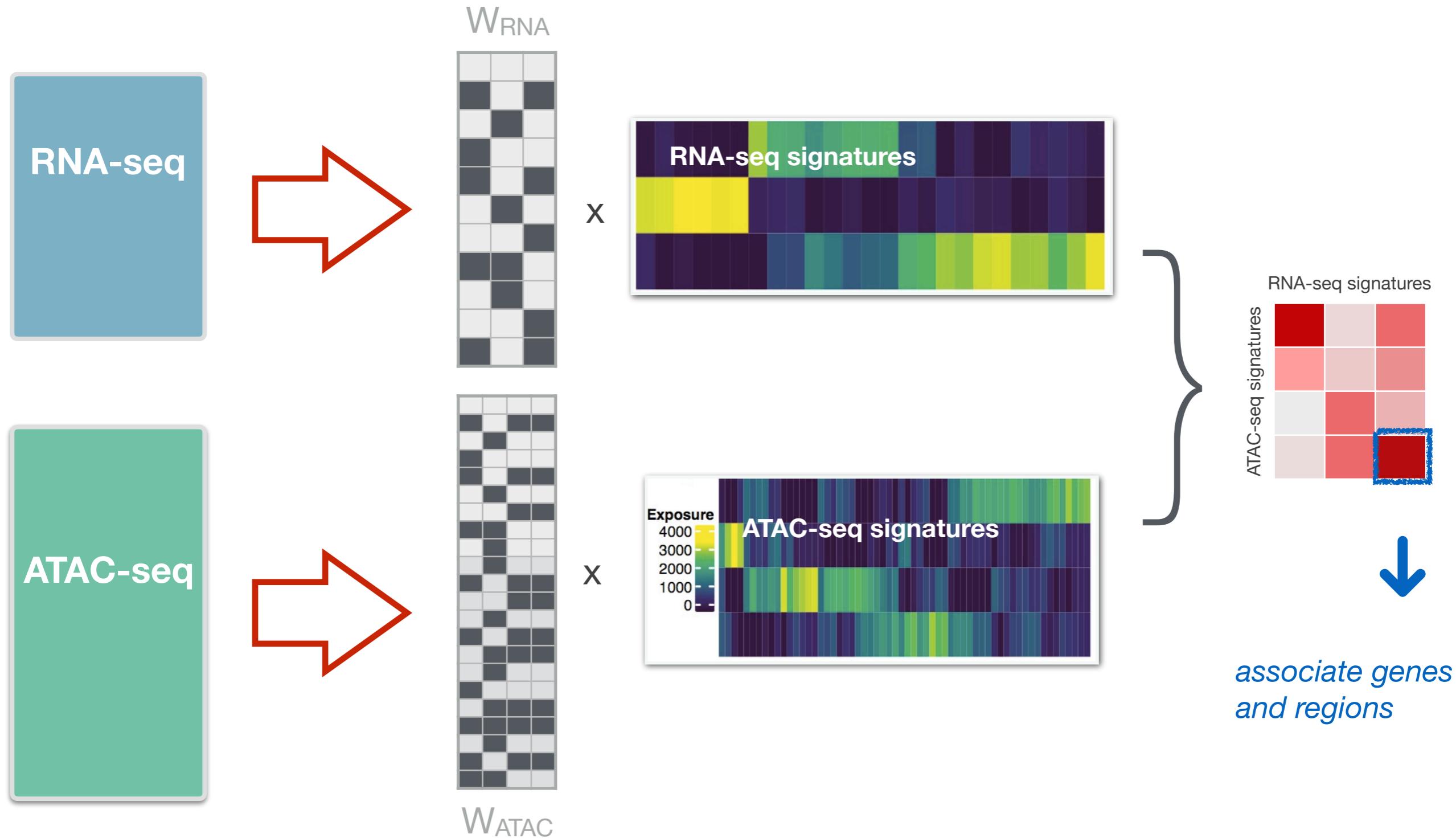
- the W matrix gives the “definition” of the signatures in terms of features contributing
- applying k-means ( $k=2$ ) to each row of the W matrix
  - ▶ **single-signature features:**  
→ gene 1 / 3
  - ▶ **multi-signature features:**  
→ gene 2 / 4 / 5
  - ▶ signatures 1 and 2 share no feature
  - ▶ signatures 2 and 4 share 2 features



# Possible Usages

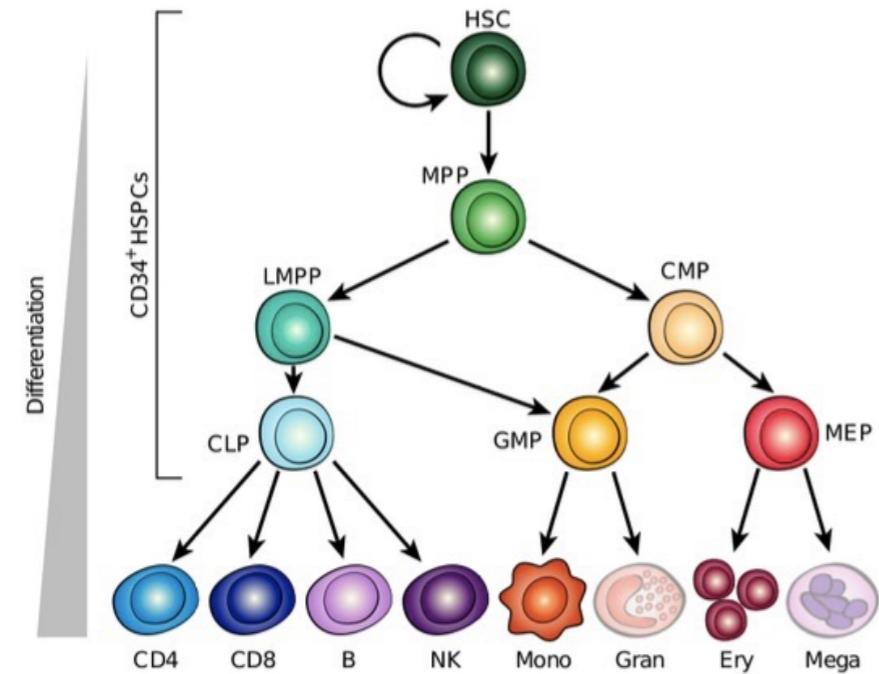


Medizinische Fakultät Heidelberg



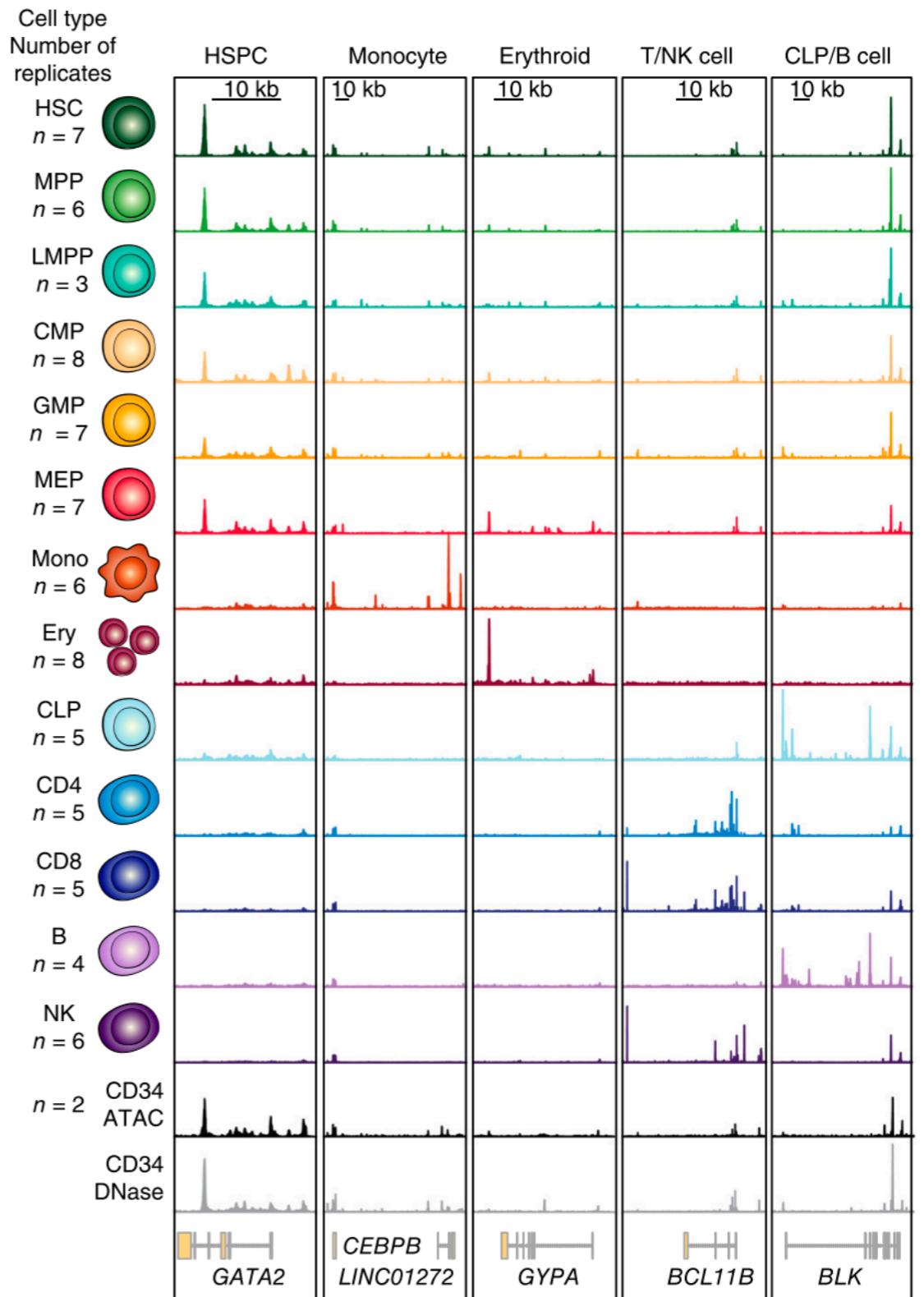


# Example of use case



Combined RNA-seq (gene expression)  
and chromatin accessibility (ATAC-seq) from  
purified blood populations

[ Corces et al. Nat. Gen (2016) ]

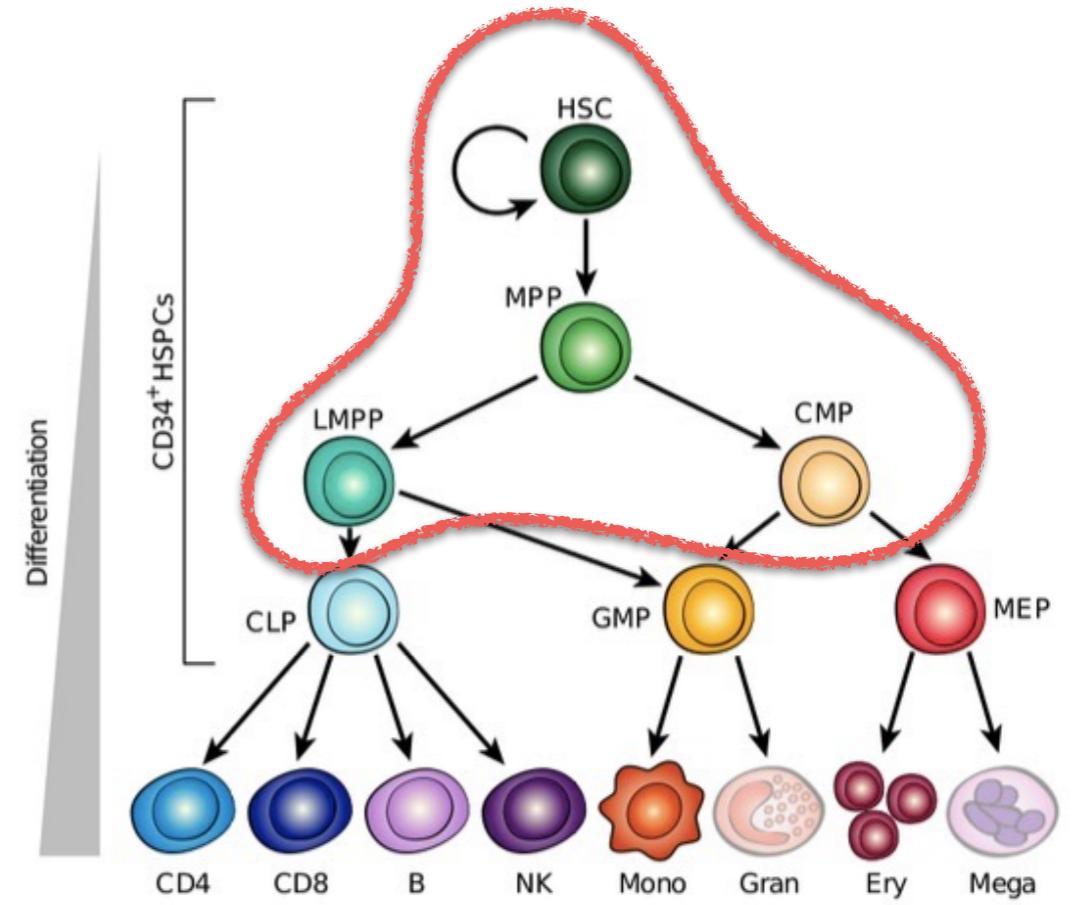
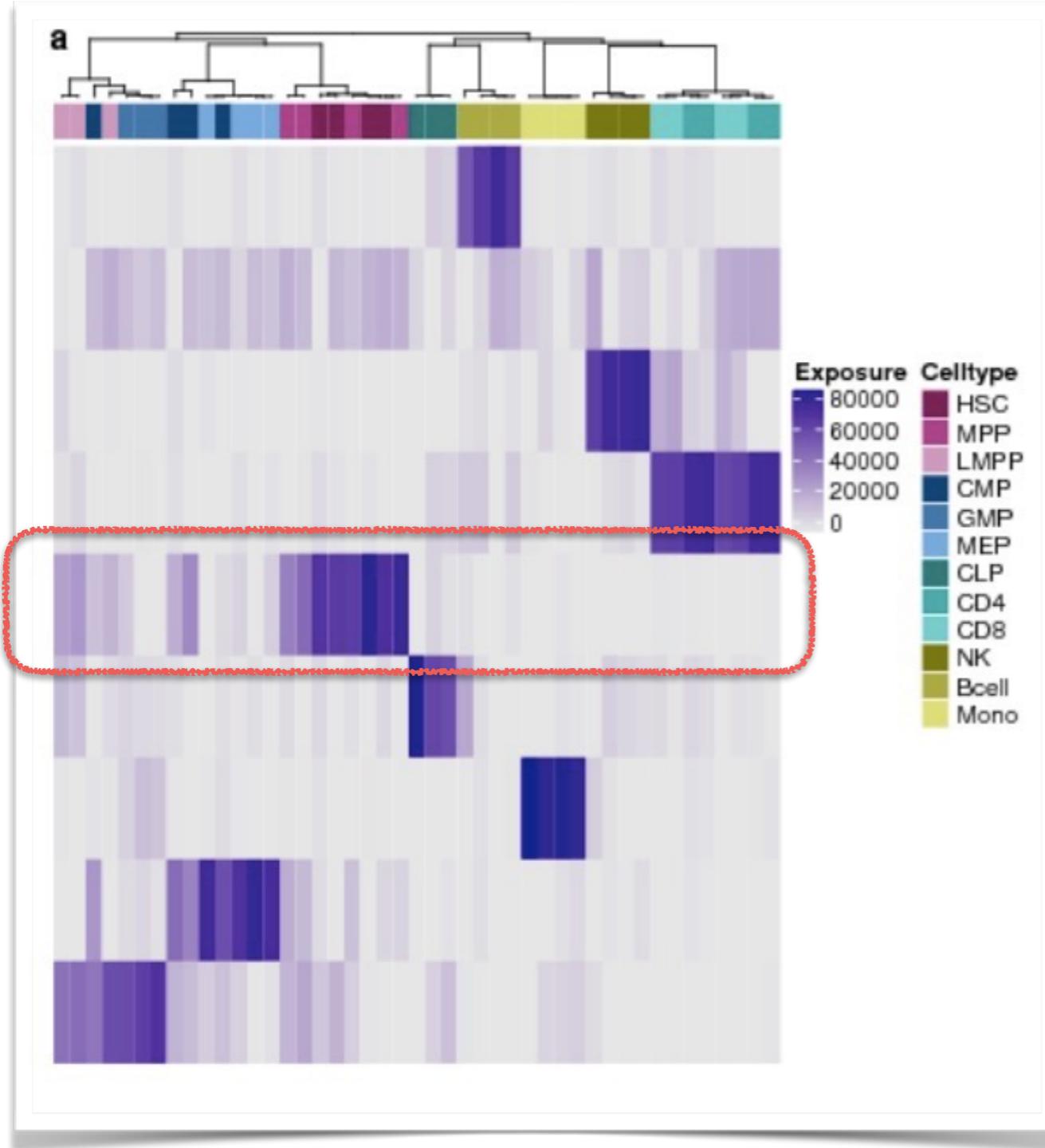


# Interpreting signatures



Medizinische Fakultät Heidelberg

RNA-seq



*Stemness-signature  
fades away, as differentiation  
progresses*

[ Corces et al. Nat. Gen (2016) ]

# Associating signatures

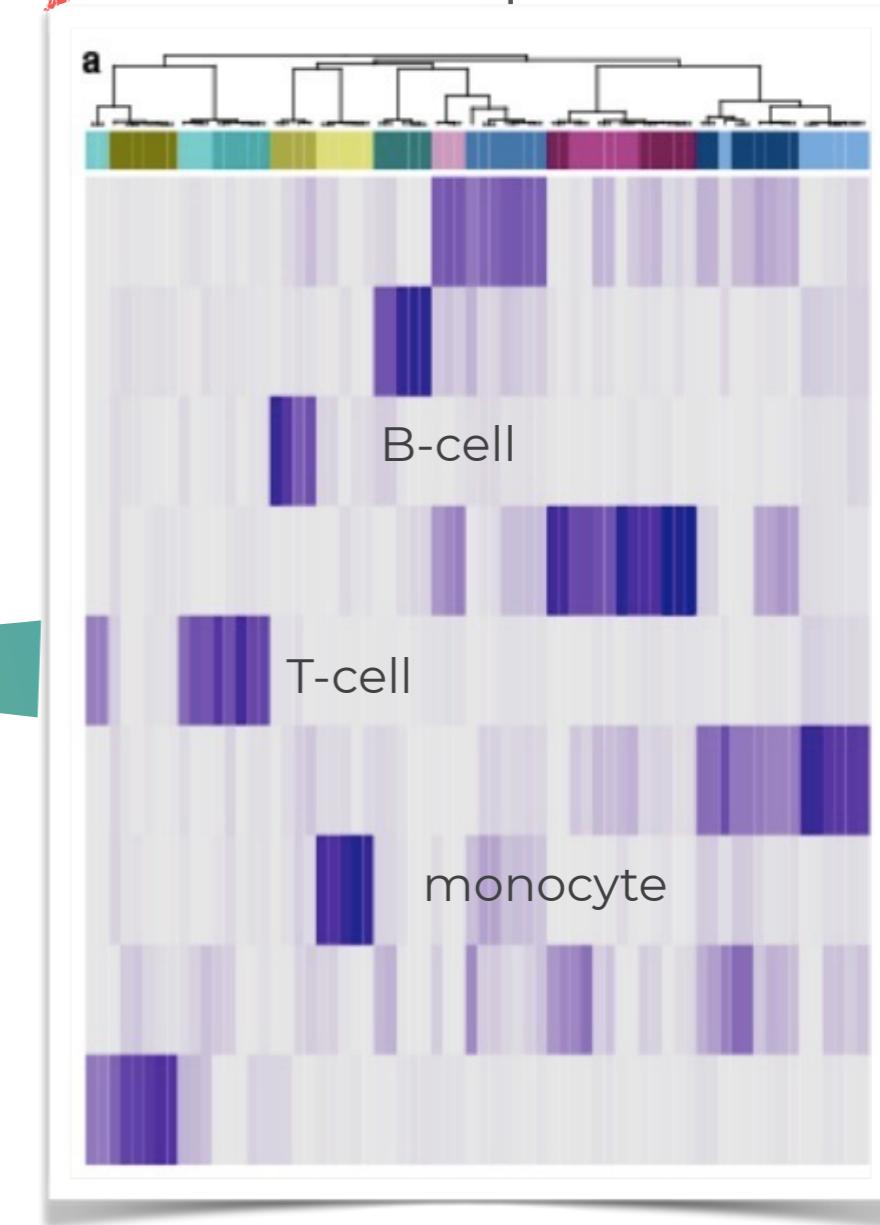
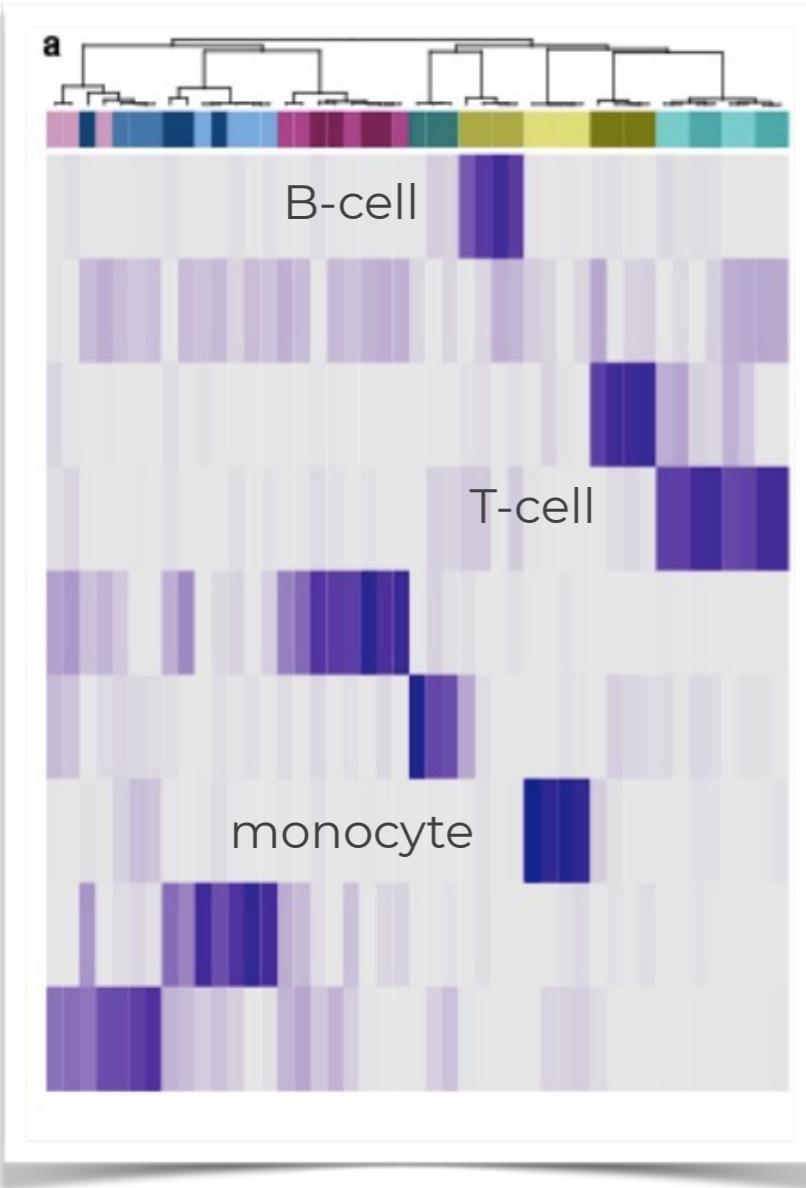


Medizinische Fakultät Heidelberg

RNA-seq

cell-type specific  
capture Hi-C

ATAC-seq

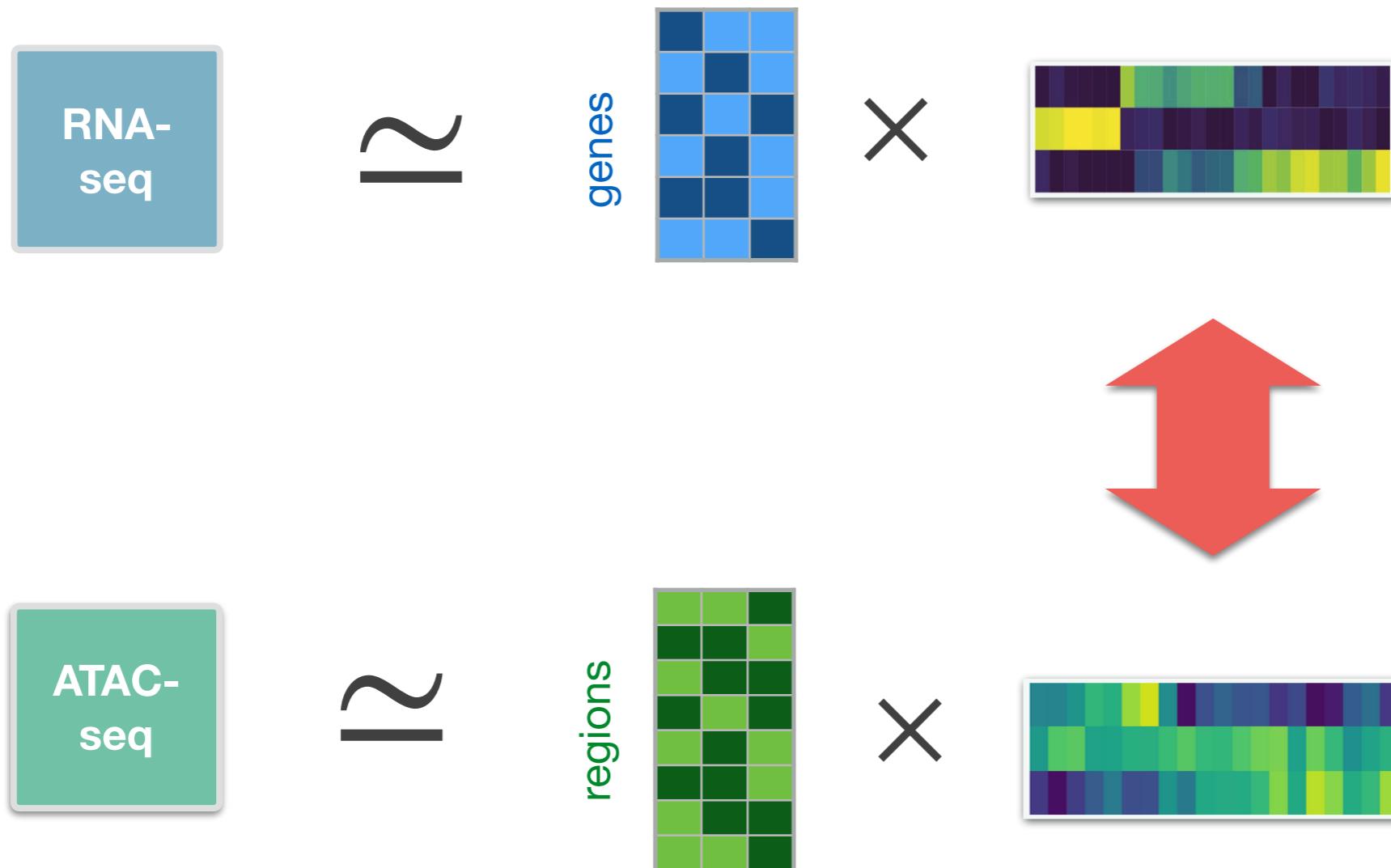


[ Corces et al. Nat. Gen (2016) ]

Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

# Integrating multiple datasets using NMF



# Integrating multiple datasets using NMF

RNA-seq

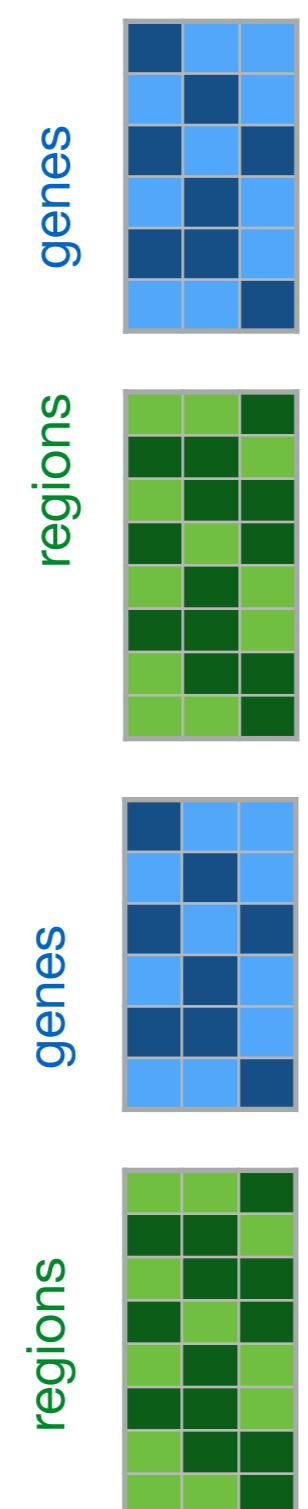
$\approx$

ATAC-seq

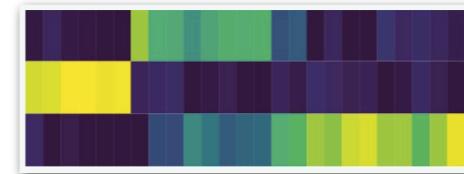
RNA-seq

$\approx$

ATAC-seq



shared  $H$  matrix

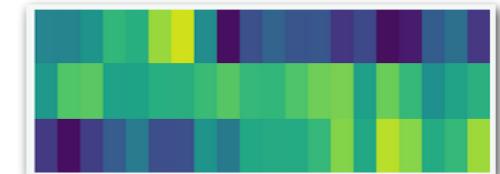


**Joint-NMF**

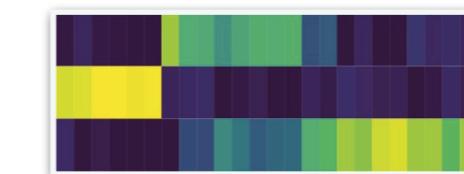
[Chalise, Fridley (2017)]

$$\min_{W, H^i} \sum_i \|X^i - W^i H\|_F^2$$

view specific  $H$  matrix

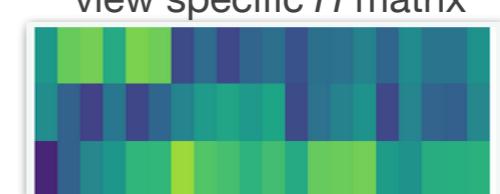


common  $H$  matrix



**integrative-NMF**

[Yang, Michailidis (2015)]



# integrative NMF



Medizinische Fakultät Heidelberg

$$\min_{W^i, H, H^i} \left( \sum_i \|X^i - W^i(H + H^i)\|_F^2 + \lambda \sum_i \|W^i H^i\|_F^2 \right)$$

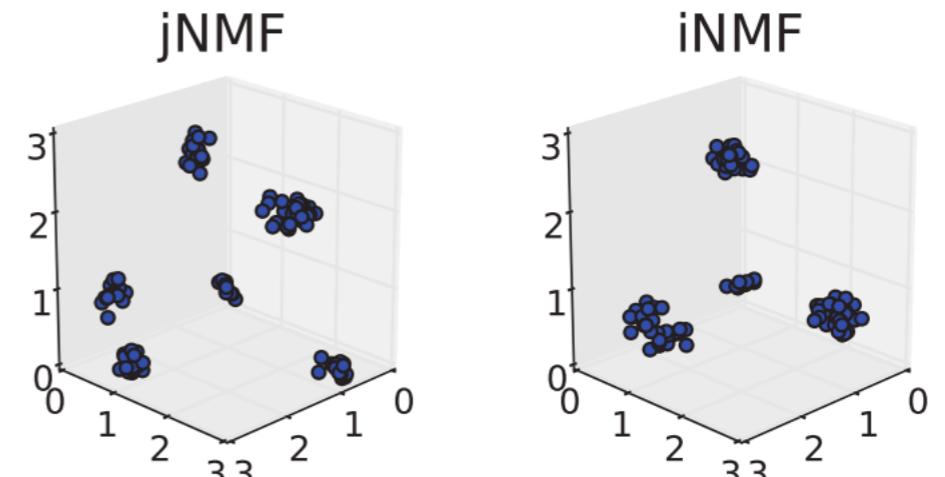
- integrative NMF identifies both **homogeneous** effects between datasets ( $H$ ) as well as **heterogeneous** ( $H^i$ )
- $\lambda$  is a homogeneity parameters
  - large values will penalize the heterogeneous effects
  - small values will promote the heterogeneous effects

simulated data:  
- 3 data types (blue, green, red)  
- 3 sample groups  
with noise and confounding  
effects



[Yang, Michailidis (2015)]

Carl Herrmann



*joint-NMF (jNMF) does not  
recognize the 3 clusters as well  
as integrative NMF (iNMF)*



Medizinische Fakultät Heidelberg

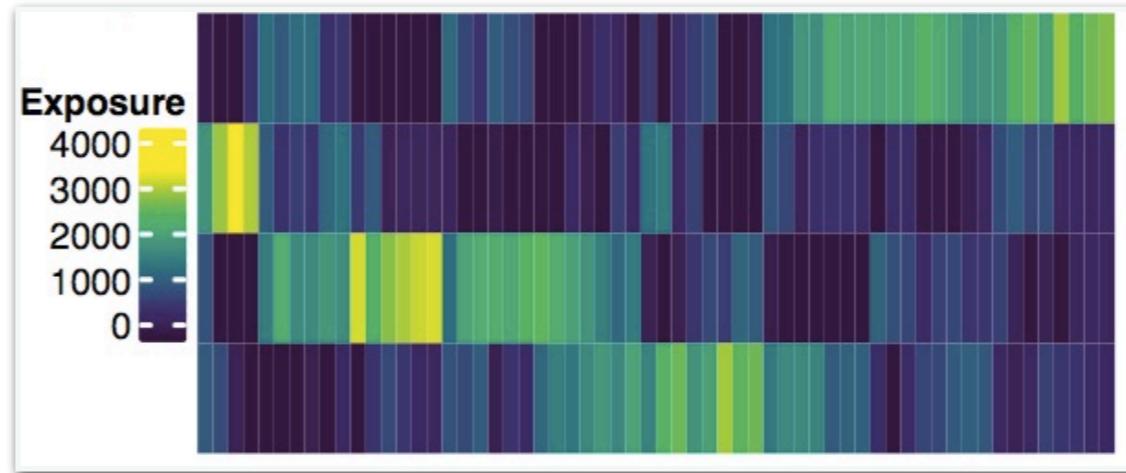
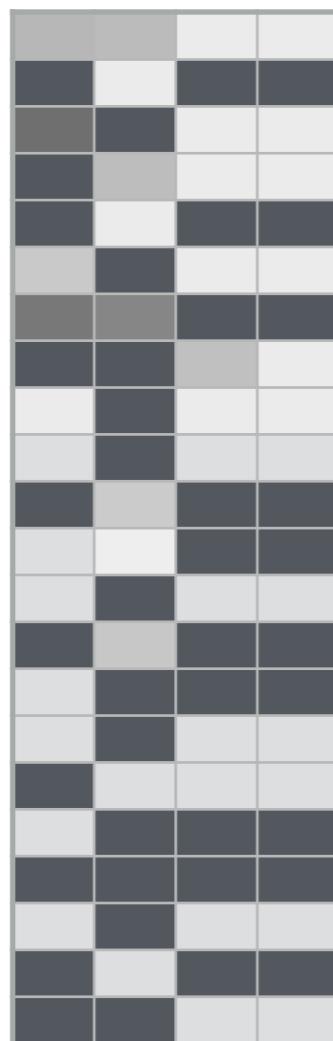
## 4. interpreting signatures

# Interpreting signatures



Medizinische Fakultät Heidelberg

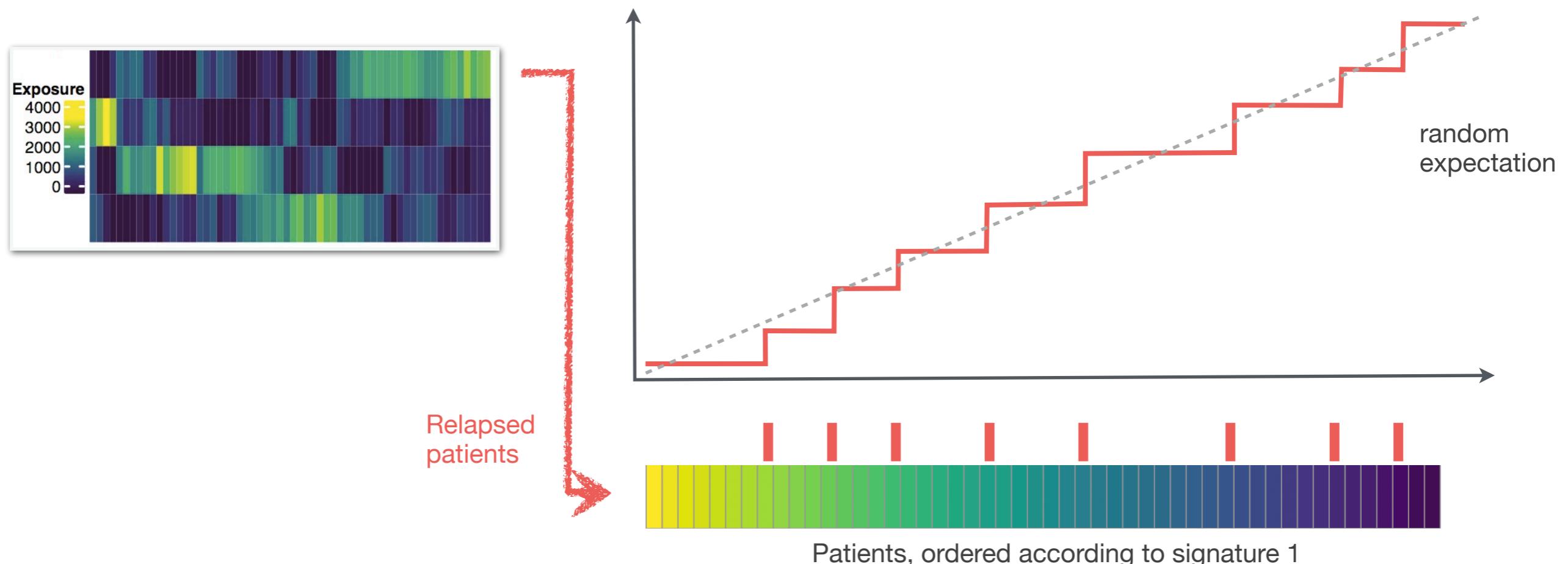
*What is special about  
features highly ranked  
by signature k ?*



*What is special about  
samples highly exposed to  
signature k ?*

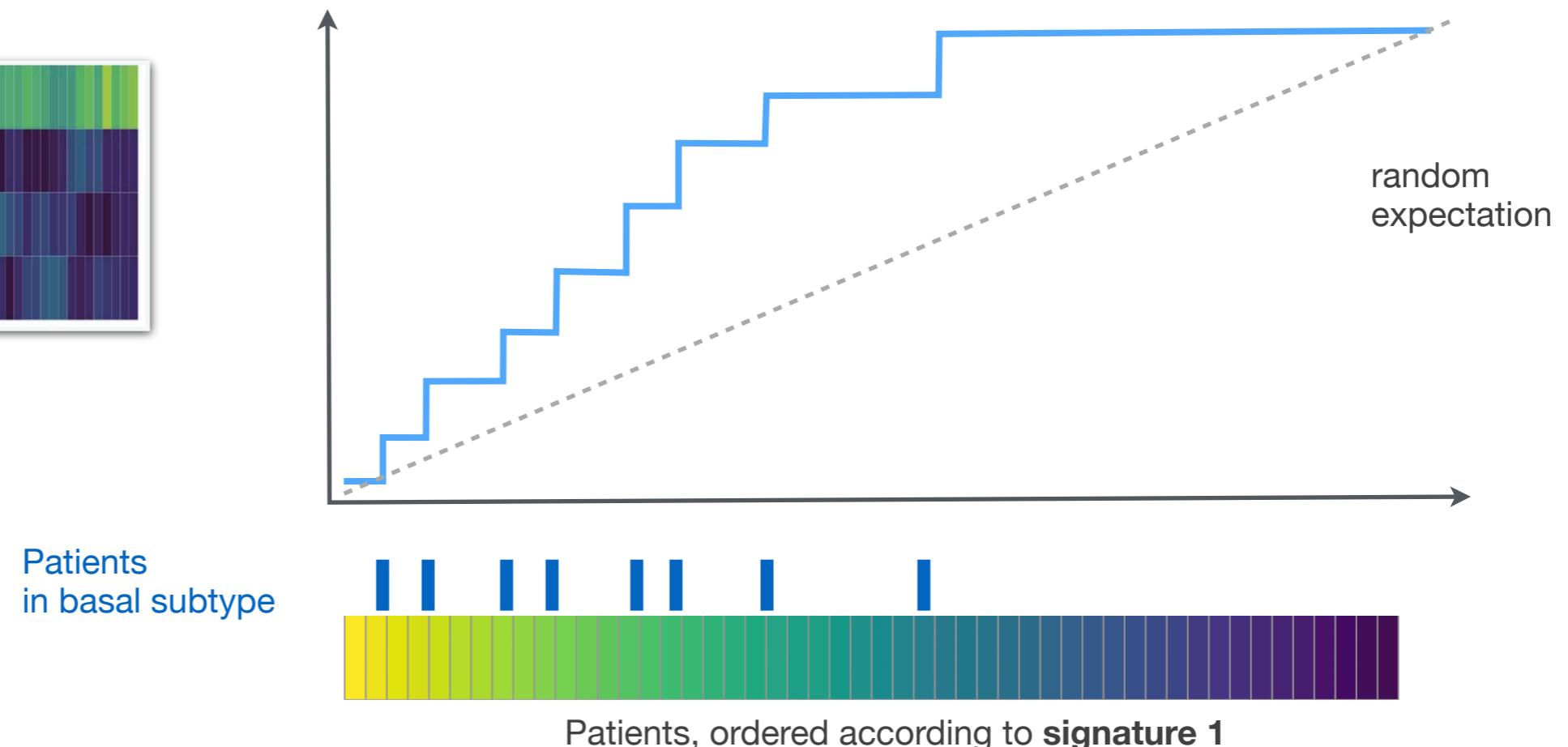
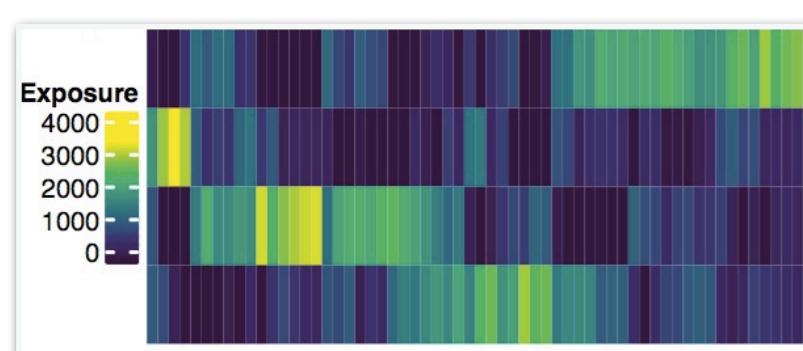
# Recovery analysis

- Rank all samples / features according to the weights / exposures to signature k
- perform recovery analysis using a set of pre-defined regions / genesets / clinical features



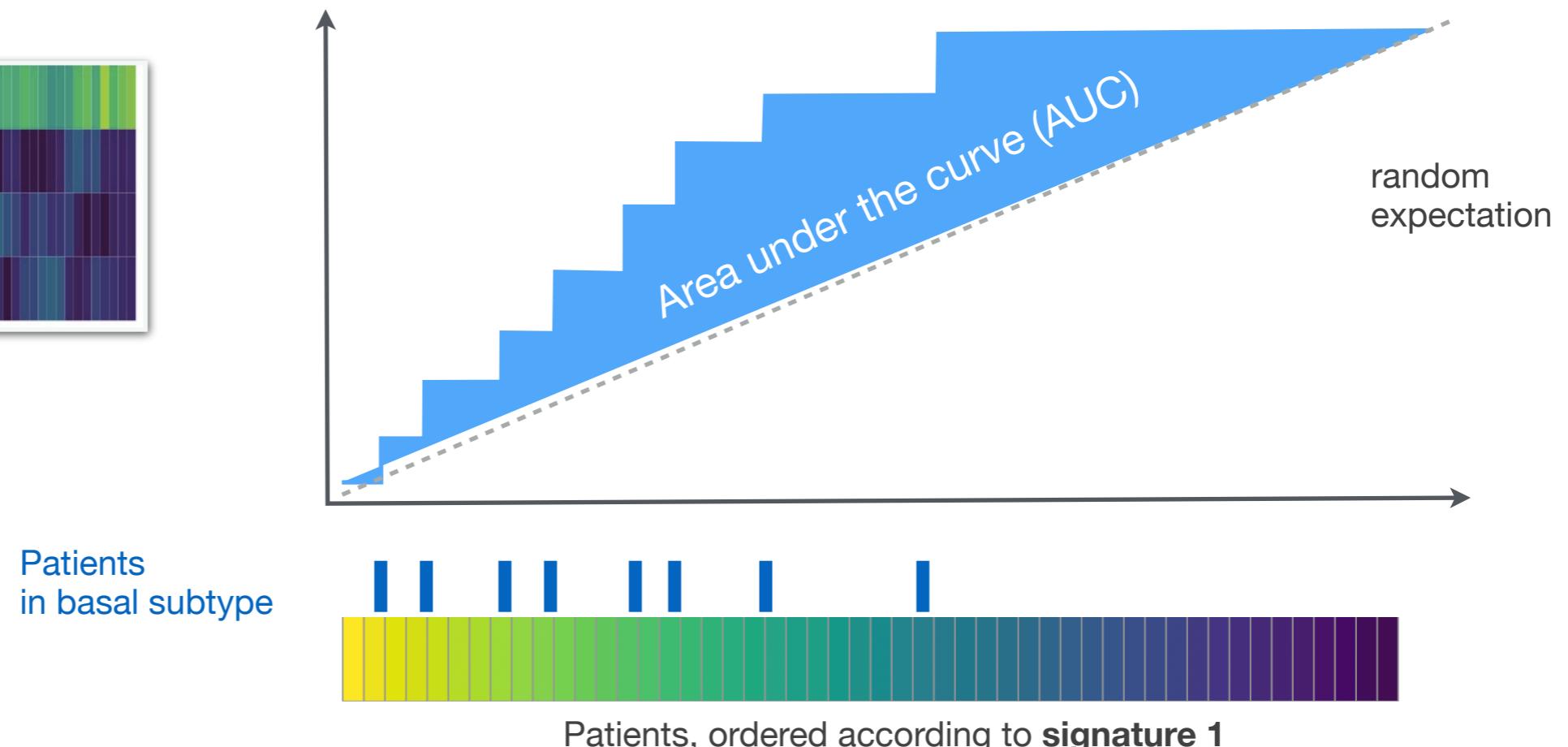
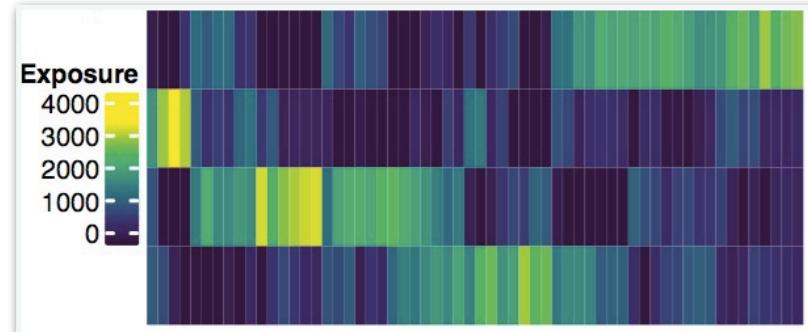
# Recovery analysis

- Rank all samples / features according to the weights / exposures to signature k
- perform recovery analysis using a set of pre-defined regions / genesets / clinical features



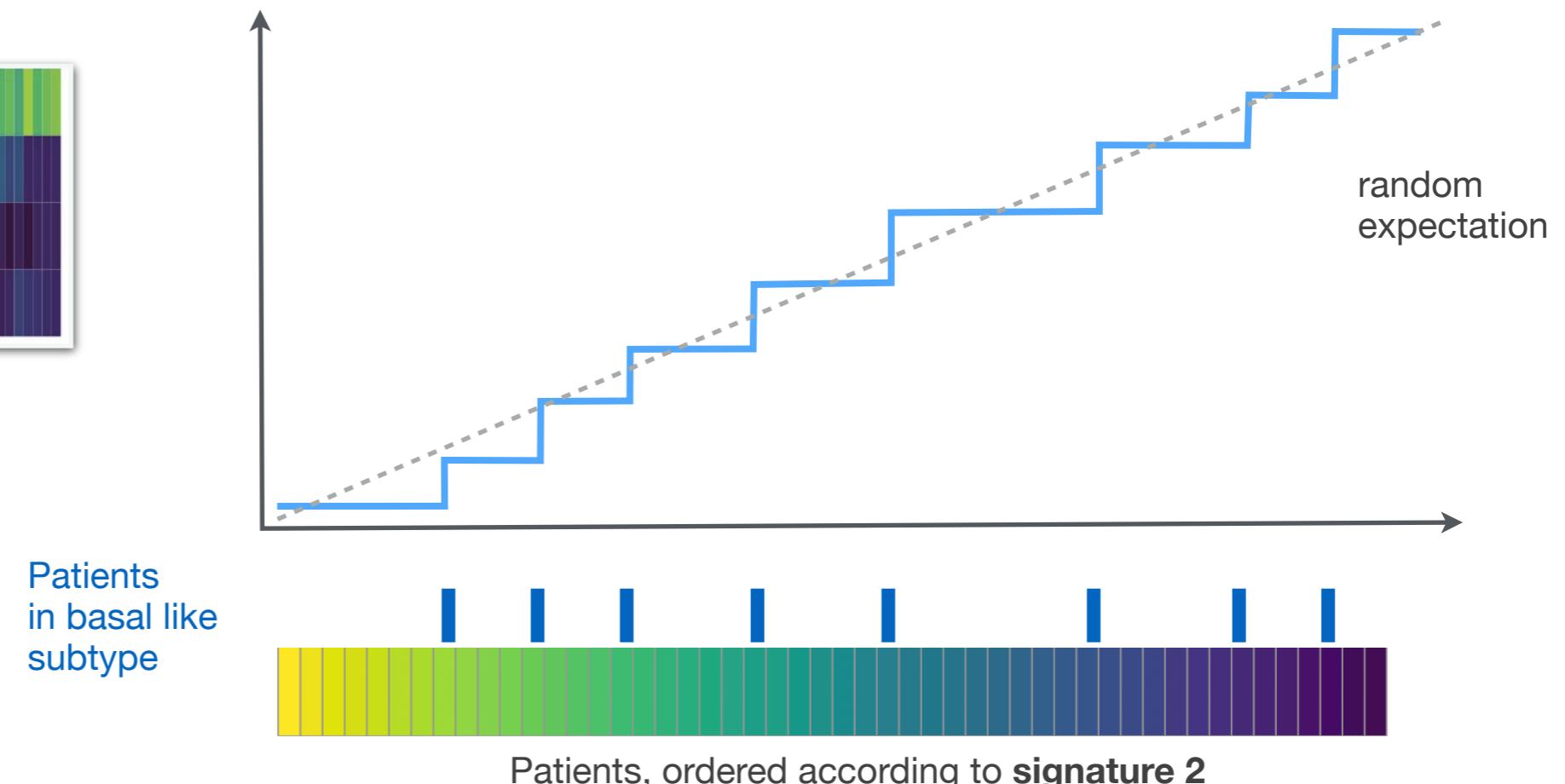
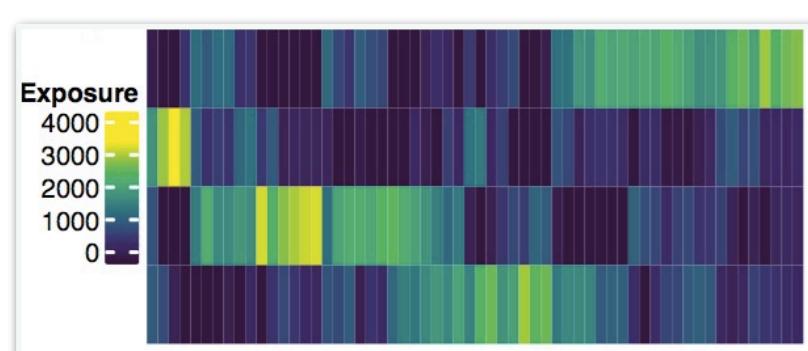
# Recovery analysis

- Rank all samples / features according to the weights / exposures to signature k
- perform recovery analysis using a set of pre-defined regions / genesets / clinical features



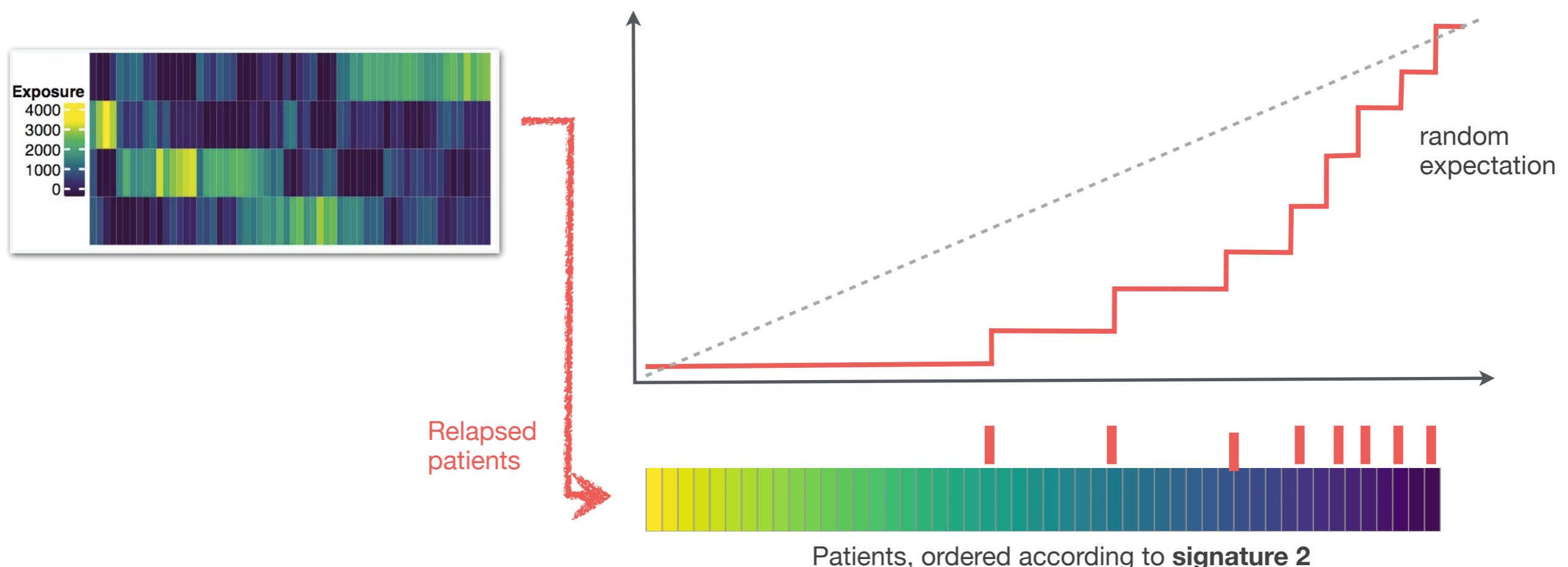
# Recovery analysis

- Rank all samples / features according to the weights / exposures to signature k
- perform recovery analysis using a set of pre-defined regions / genesets / clinical features



# Recovery analysis

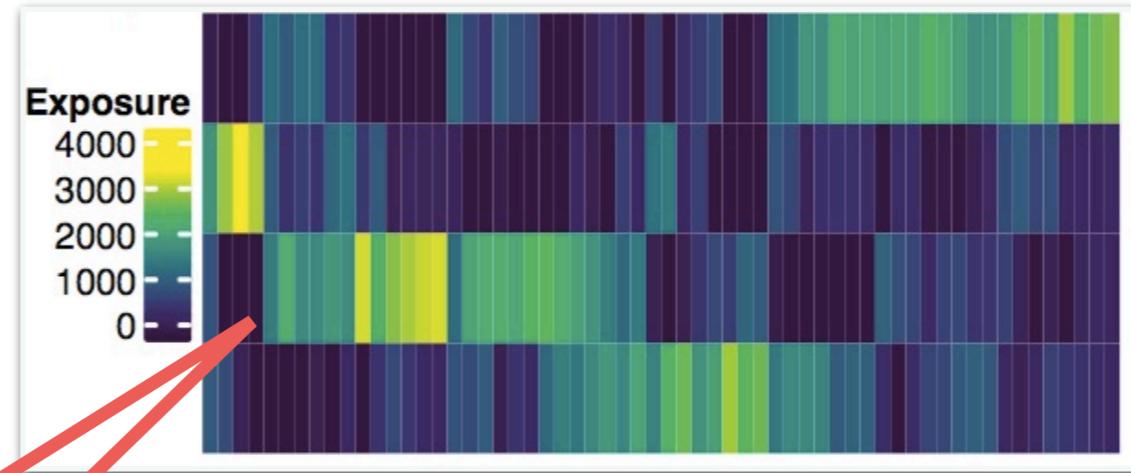
- Rank all samples / features according to the weights / exposures to signature k
- perform recovery analysis using a set of pre-defined regions / genesets / clinical features



# Interpreting signatures



Medizinische Fakultät Heidelberg



***Signature 1 is associated  
with basal like subtype  
Signature 2 is anti-associated  
with relapse patients.***

***What is special about  
samples highly exposed to  
signature k ?***

# Conclusions - important points



- Genomics data is increasingly
  - **large**
  - **heterogeneous** (binary / discrete / continuous / ...)
- Each data type sheds a different light at the data (RNA-seq ≠ DNA-methylation): "*Whole is more than sum of parts*"
- Matrix factorization is a powerfull class of statistical approaches to **reduce dimensions**
- These methods can be adapted to perform **joint learning** accross data types
- NMF and its derivates exploit the natural non-negative nature of the data and **improves interpretation**