Extracting signatures from combined genomic datasets

Carl Herrmann
Health Data Science Unit and BioQuant
carl.herrmann@uni-heidelberg.de

Allahabad RTG Summer School
April 2019

Combining and understanding multi-omics data

Carl Herrmann
Health Data Science Unit and BioQuant
carl.herrmann@uni-heidelberg.de

Allahabad RTG Summer School
April 2019

Medizinische Fakultät Heidelberg

Outline

Medizinische Fakultät Heidelberg

1. **Big data in genomics and medecine**
   - large scale genomics projects
   - cancer genomics

2. **Multi-omics assays**
   - types of assays (WES / WGS / RNA-seq / 450K / …)
   - Roadmap / IHEC / TCGA / …

3. **Dimensional reduction and integration**

Medizinische Fakultät Heidelberg

1. Genomics: where the data comes from

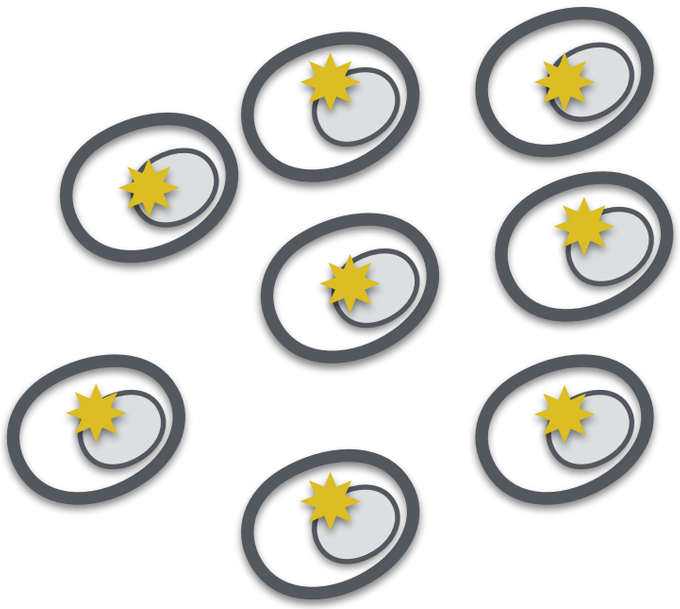
Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

Identification of genomic variants

Medizinische Fakultät Heidelberg

Germline variants



→ **risk alleles for hereditary diseases**

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 paraplegia (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.

Large scale population genomics projects

Medizinische Fakultät Heidelberg

- **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

Precision medecine

Medizinische Fakultät Heidelberg

2019

Characterizing mutagenic effects of recombination through a sequence-level genetic map.

Halldorsson BV, et al.

Science. 2019 Jan 25;363(6425). pii: eaau1043. doi: 10.1126/science.aau1043.

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

Kristjansson er al.

Nat Genet. 2019 Jan 14. doi: 10.1038/s41588-018-0314-6. [Epub ahead of print]

2018

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

Styrkarsdottir U, et al.

Nat Genet. 2018 Dec;50(12):1681-1687. doi: 10.1038/s41588-018-0247-0. Epub 2018 Oct 29. Kristjansson er al.

Genome-wide association meta-analysis yields 20 loci associated with gallstone disease.

Ferkingstad E, et al.

Nat Commun. 2018 Nov 30;9(1):5101. doi: 10.1038/s41467-018-07460-y.

Sequence variants associating with urinary biomarkers.

Benonisdottir S, et al.

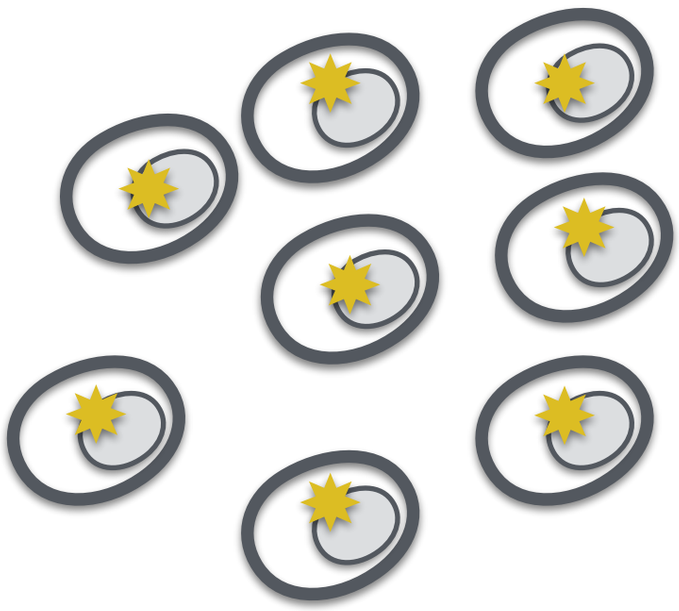
Hum Mol Genet. 2018 Nov 24. doi: 10.1093/hmg/ddy409. [Epub ahead of print]

[publication list deCODE consortium]

Identification of genomic variants

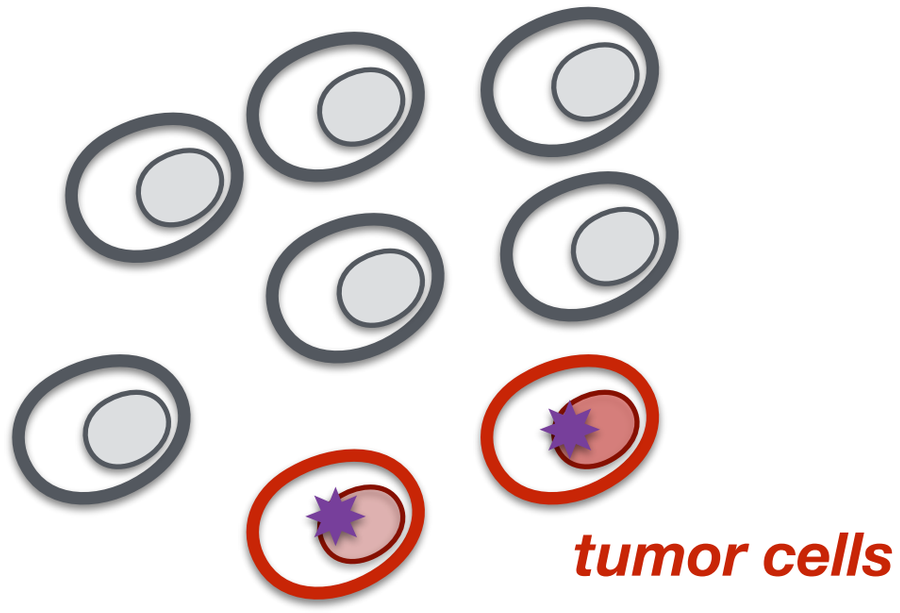
Medizinische Fakultät Heidelberg

Germline variants



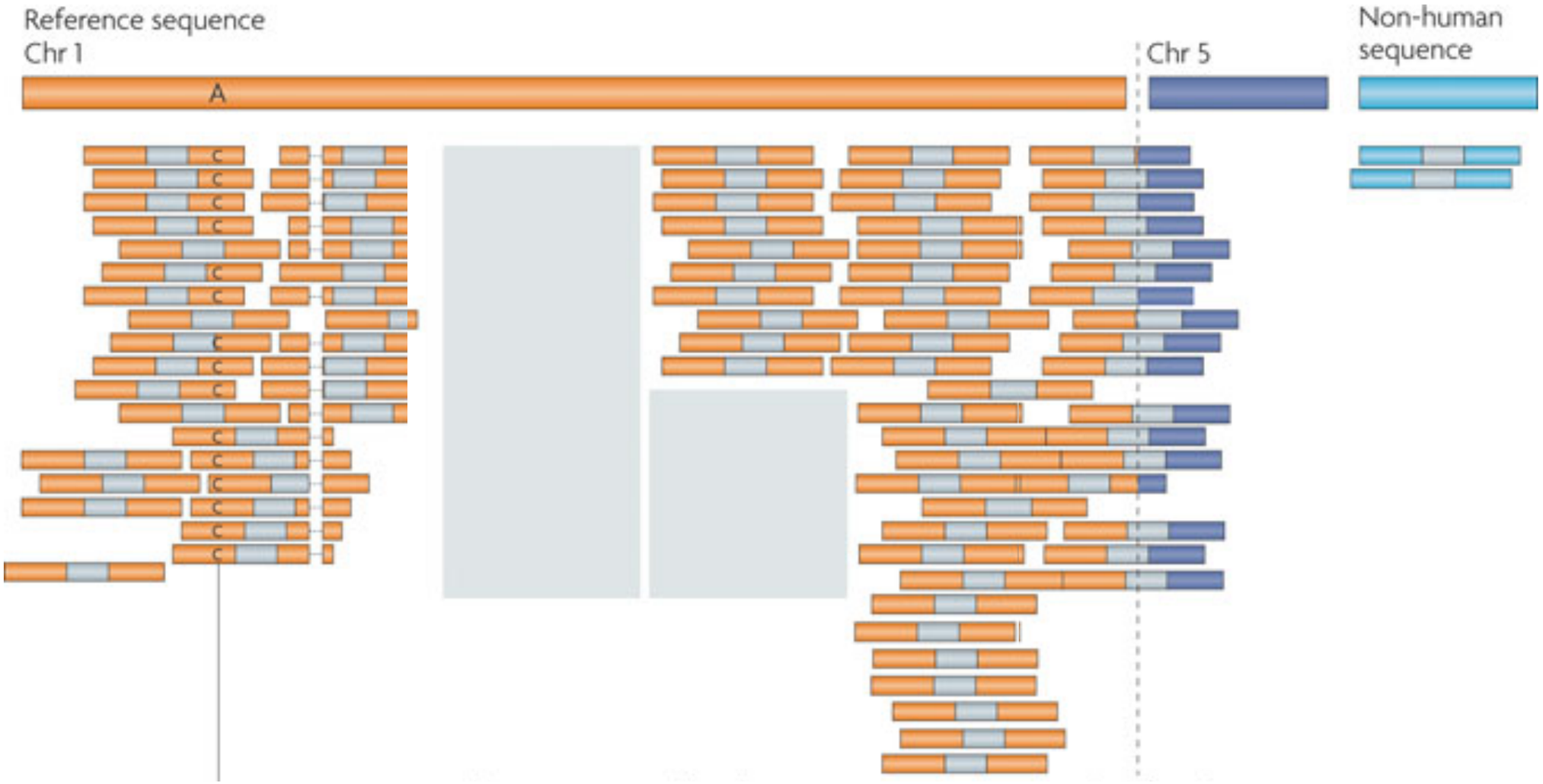
→ **risk alleles for hereditary diseases**

Somatic variants



→ **oncogenic driver mutations**

Cancer genomics



Point mutations
/ indels

copy number
alterations
(deletion / LOH / amplification)

structural variant

viral insertion

Variant in cancer genomes

Medizinische Fakultät Heidelberg

Your selected sets

Type	Set	# Items
Donor	Top 75 PCAWG Liver Donors	75
Gene	Top 75 CGC Genes for Liver	75

Resulting Number of High Impact Mutation Occurrences: 302

Genome Viewer

missense variant · start lost · initiator codon variant · frameshift variant · stop lost · stop gained

Mutation freq. 50 25

TP53 ARID1A LRP1B ACVR2A TSC1 ARID2 KMT2C KMT2D EP300 RPL22 KMT2A APC PTEN NF1 CIC KEAP1 AXIN1 BIRC6 CDKN1A RB1 ATM PBRM1 GATA3 DNMT3A MYH11 COL2A1 CSMD3 GRM3 ROS1 TSC2 CDKN1B CBLB MSH6 PAX8 AKAP9 NOTCH2 BRCA2 POLG BRAF SYK

Mutation freq. 17 35

Gene Sets

Curated Gene Census

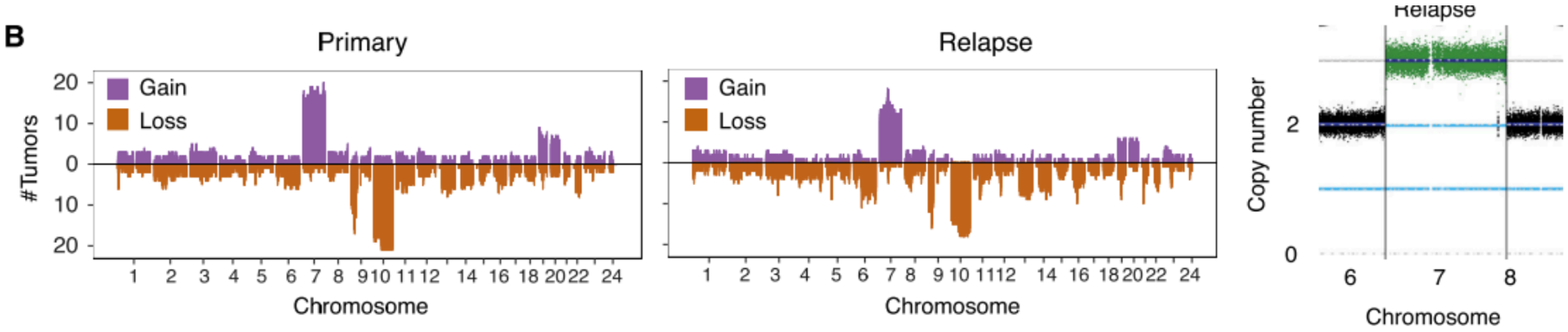
ICGC

Donors affected

Strctural variants in cancer genomes

Medizinische Fakultät Heidelberg

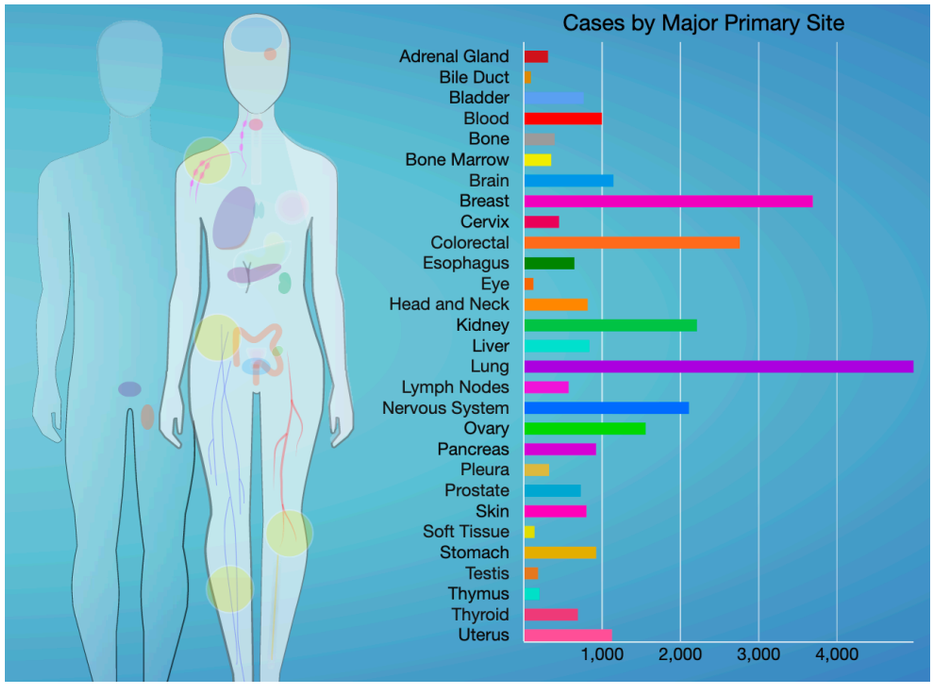
Glioblastoma patients



ell 2019]

The Cancer Genome Analysis (TCGA)

Medizinische Fakultät Heidelberg



Data Portal Summary

Data Release 16.0 - March 26, 2019

TCGA produced over **2.5** PETABYTES of data

PROJECTS	PRIMARY SITES	CASES
45	68	33,549

FILES	GENES	MUTATIONS
365,463	22,872	3,142,246

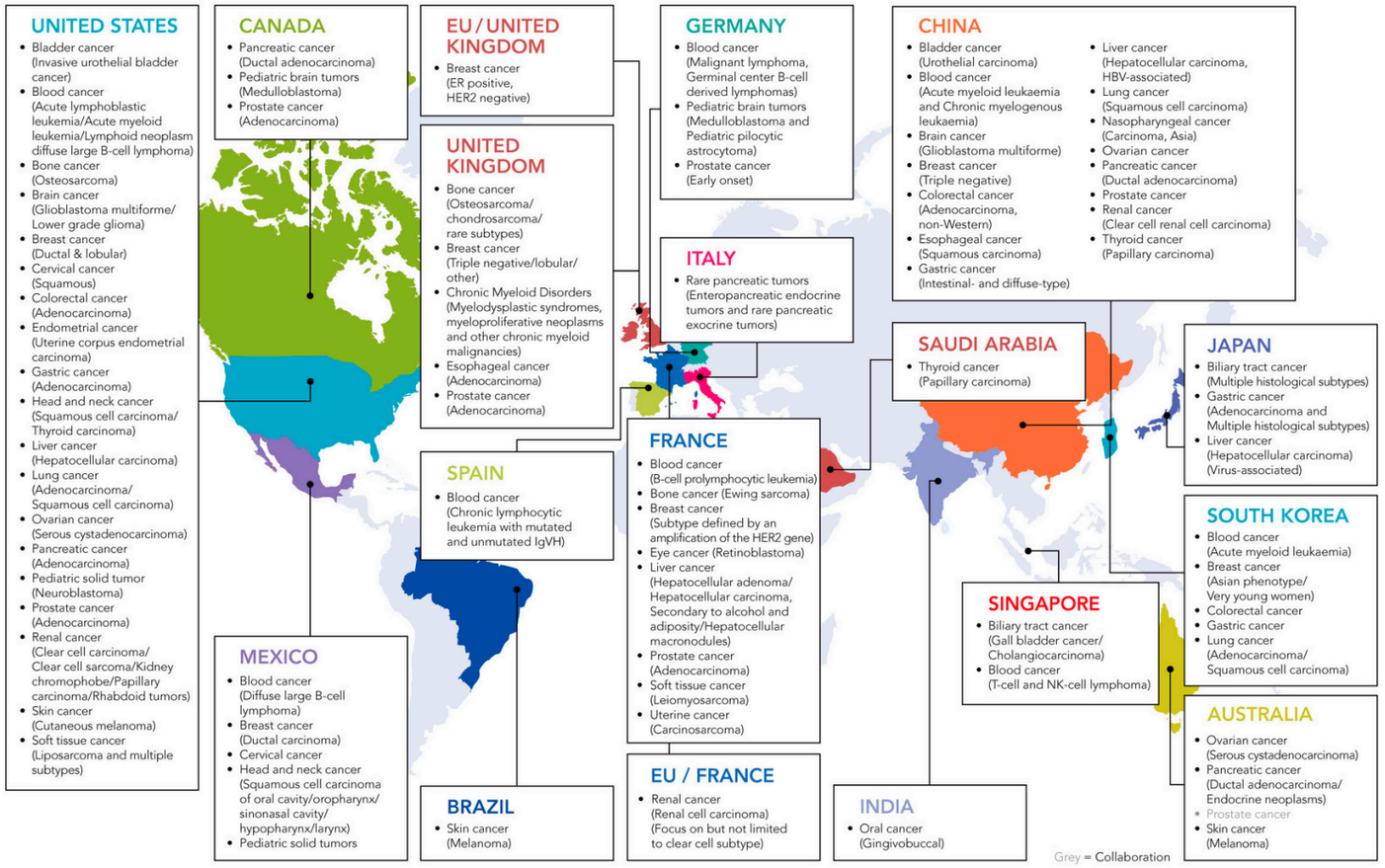
Example of data available for glioblastoma

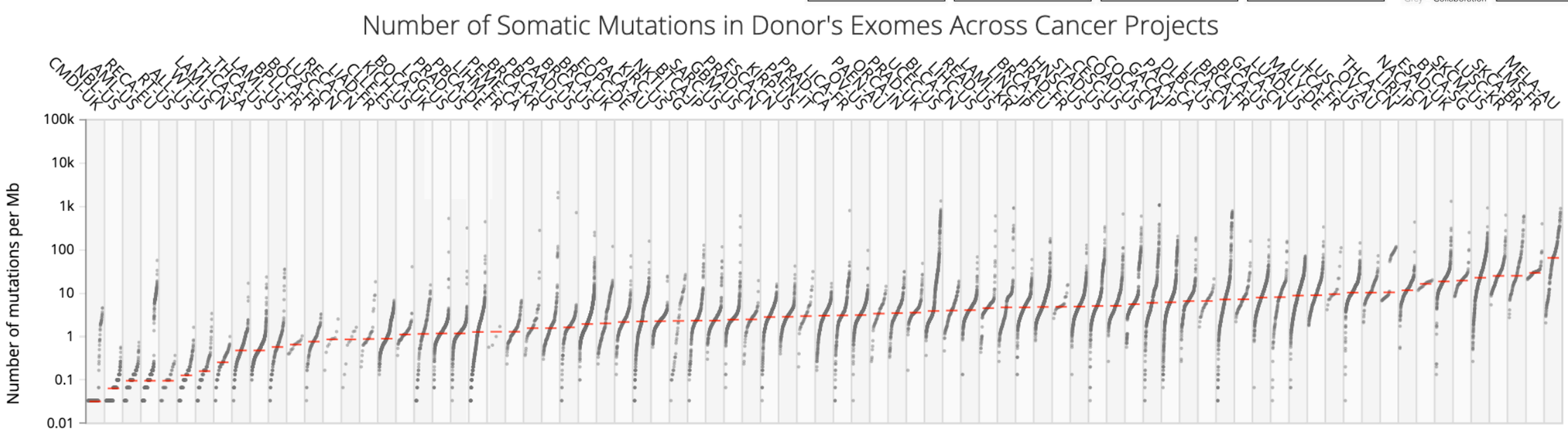
Data Category	Cases (n=617)
Sequencing Reads	406
Transcriptome Profiling	166
Simple Nucleotide Variation	396
Copy Number Variation	599
DNA Methylation	423
Clinical	617
Biospecimen	617

International Cancer Genome Consortium (ICGC)

Medizinische Fakultät Heidelberg

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





International Cancer Genome Consortium (ICGC)

Medizinische Fakultät Heidelberg

- international cancer genome sequencing effort
- India: oral cancer

Project Summary

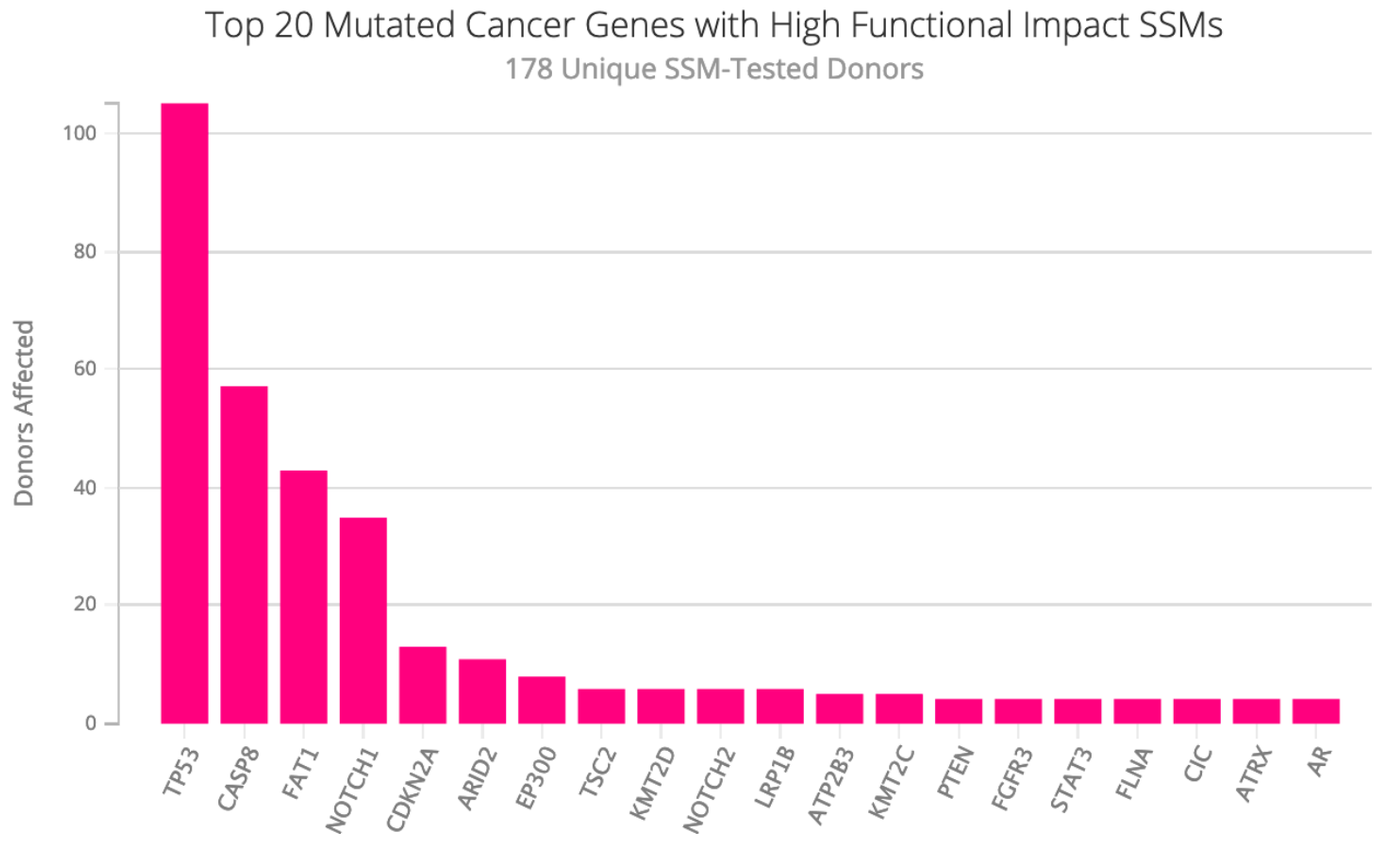
In view of its high prevalence in India and the existence of possible interacting environmental factors, India will focus on oral cancer as a part of the International Cancer Genome Consortium activities. In particular, India will focus on gingivo-buccal cancer. The primary reasons for selection of gingivo-buccal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

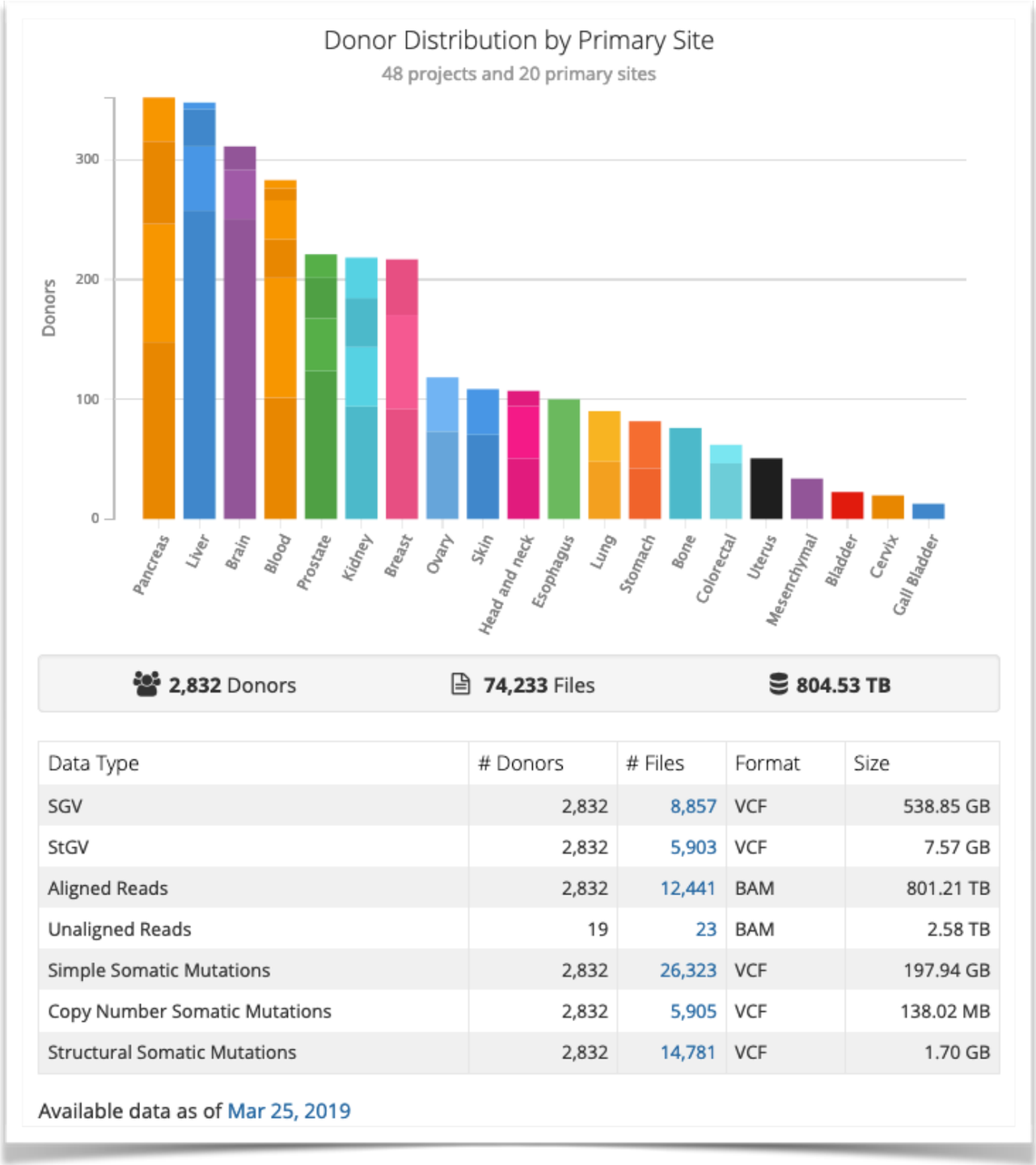
• Leukoplakia and sub-mucous fibrosis are two types of pre-cancerous lesion that can help track temporal genomic changes in gingivo-buccal cancer.



PanCancer Analysis of Whole Genomes (PCAWG)

Medizinische Fakultät Heidelberg

- 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



Data Type	# Donors	# Files	Format	Size
SGV	2,832	8,857	VCF	538.85 GB
StGV	2,832	5,903	VCF	7.57 GB
Aligned Reads	2,832	12,441	BAM	801.21 TB
Unaligned Reads	19	23	BAM	2.58 TB
Simple Somatic Mutations	2,832	26,323	VCF	197.94 GB
Copy Number Somatic Mutations	2,832	5,905	VCF	138.02 MB
Structural Somatic Mutations	2,832	14,781	VCF	1.70 GB

Available data as of Mar 25, 2019

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)

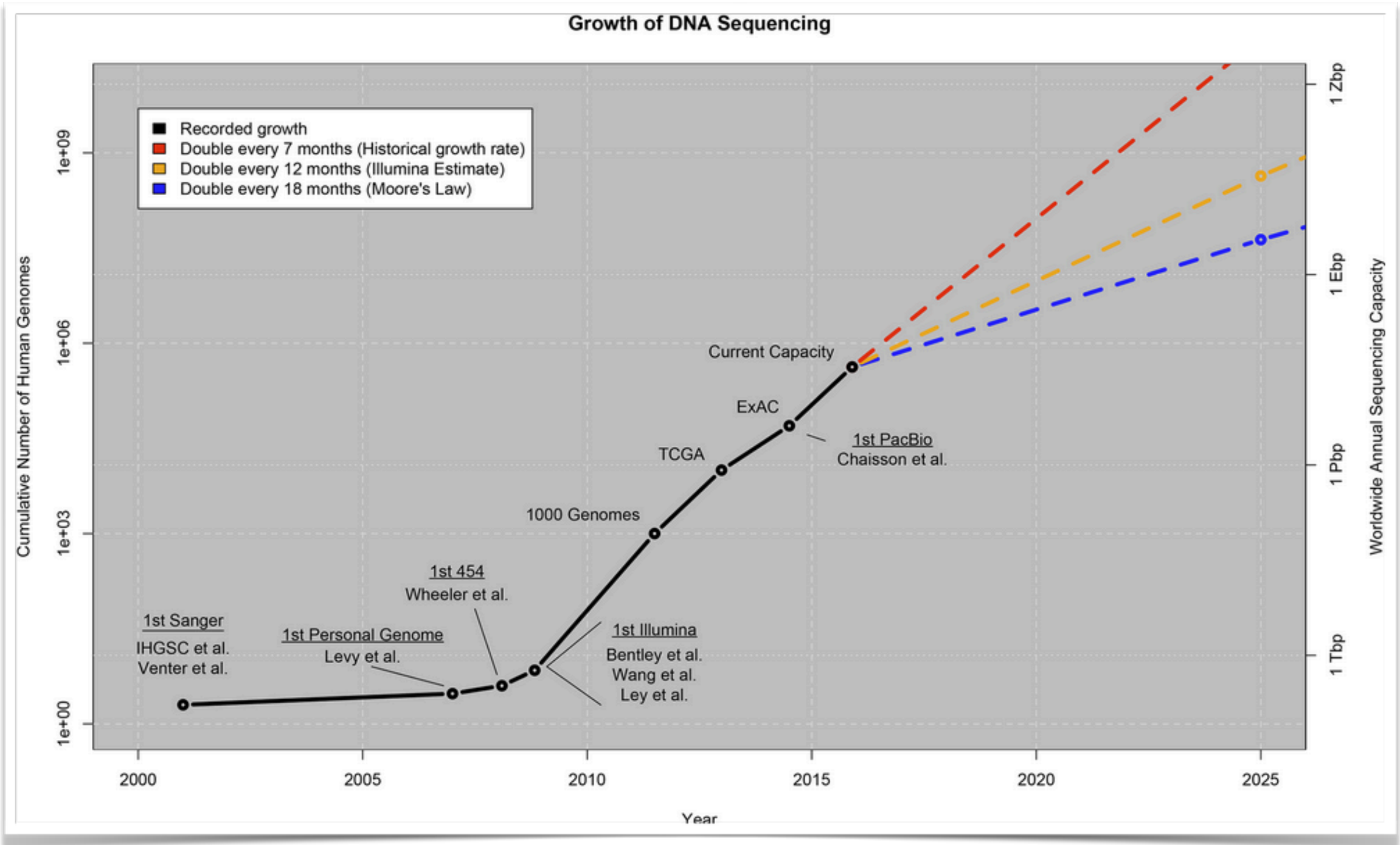
Zettabyte (1 000 000 000 000 000 000 000 Bytes)

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

```
@SRR540192.1580 IL34_5480:5:1:6425:1038/1
CATCTTGGCCTCTGTGCAGCATTCCTTTCTCCATGGT
+
IIIIIIHIIIIIIHIIIIIIIIIIIIIIIHIIHIHID
@SRR540192.1752 IL34_5480:5:1:7005:1052/1
GCTCCCAGAAACCCAGGGCCACTGGCAGCTTCAGGGA
+
GGGGGGGBG@GGGGB@>D<GGGF@<?<?9??;(?:2(
@SRR540192.1788 IL34_5480:5:1:10167:1053/1
ATGGGCTTCCTCCGGCTTTCAGCCACCTGCGCCCTGC
+
GG@G>G@E3<B=B;B<E>EDEAAAB:B.:=>A?;A8D
@SRR540192.2271 IL34_5480:5:1:5889:1093/1
TGATCATCTGGCTGATGCGGTGACTGCCACCCTTGAG
+
IIGIIIIIIIIIIIIDIIIGIIIHGHHHIIIIHIHD
```

Book with
4.3 million pages
text written in Arial 12

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

Medizinische Fakultät Heidelberg

2. Beyond Genome Sequences

Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

Measuring Chromatin state

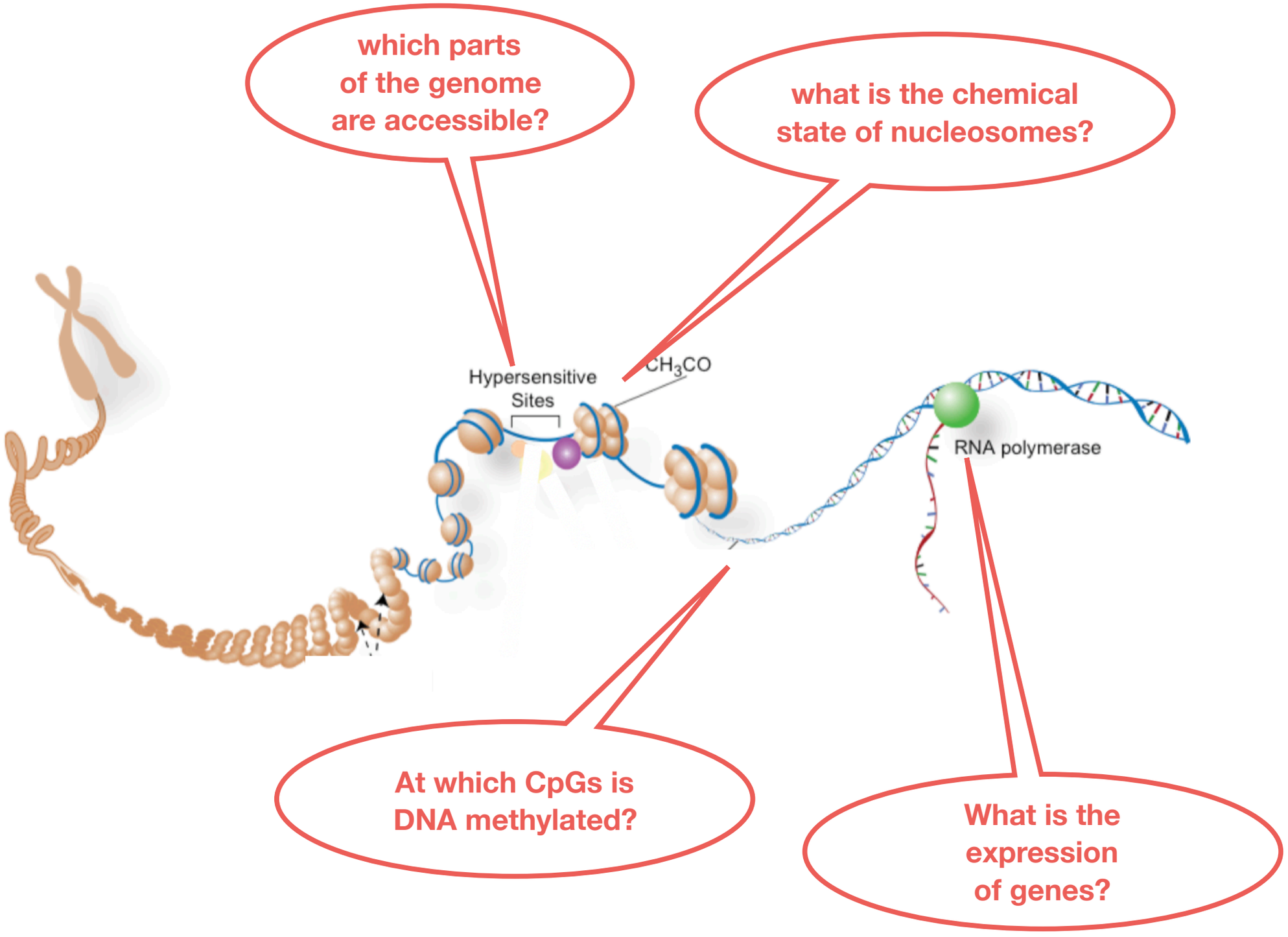
neuronal cell

melanocyte

tumor cell

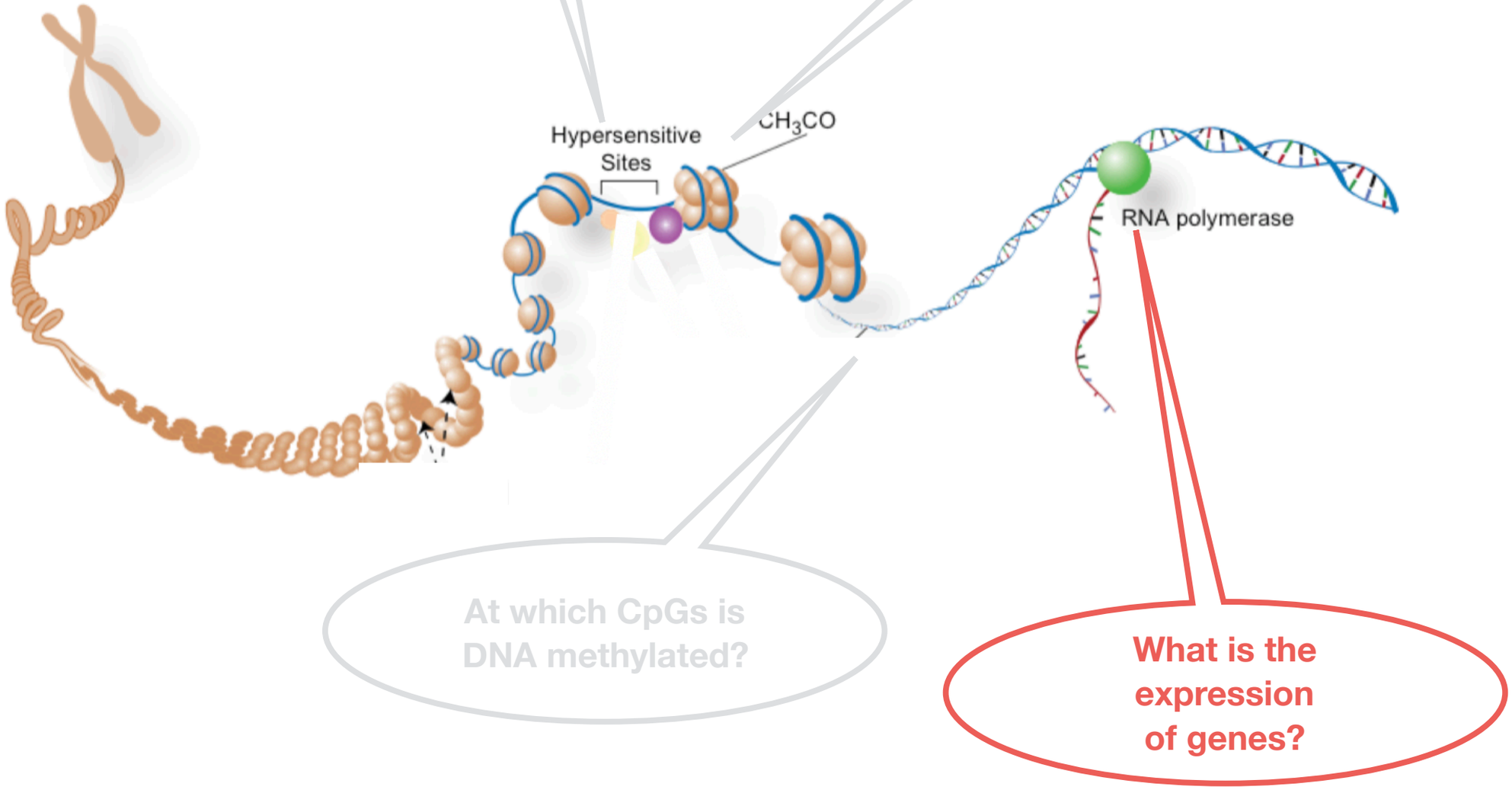
Regulatory genome

Medizinische Fakultät Heidelberg



Regulatory genome

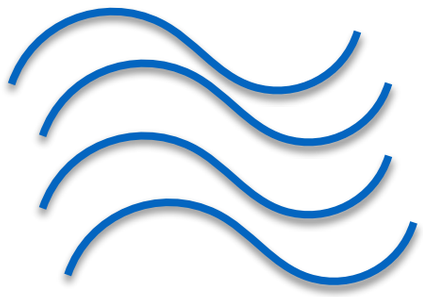
Medizinische Fakultät Heidelberg



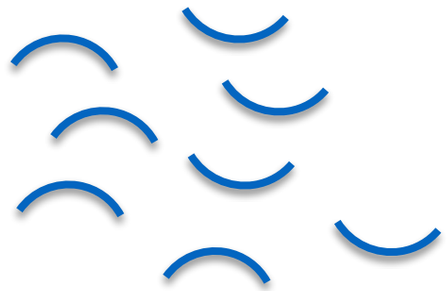
Measuring gene expression

Medizinische Fakultät Heidelberg

mRNA molecules

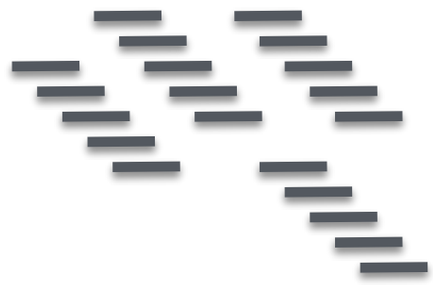


fragmentation

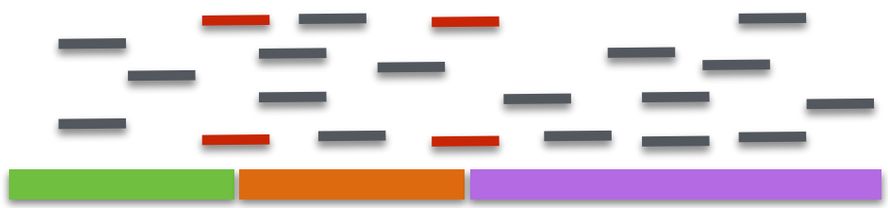


reverse transcription (cDNA)
+ adapters ligation

(paired-end) sequencing



exon junction
spanning reads

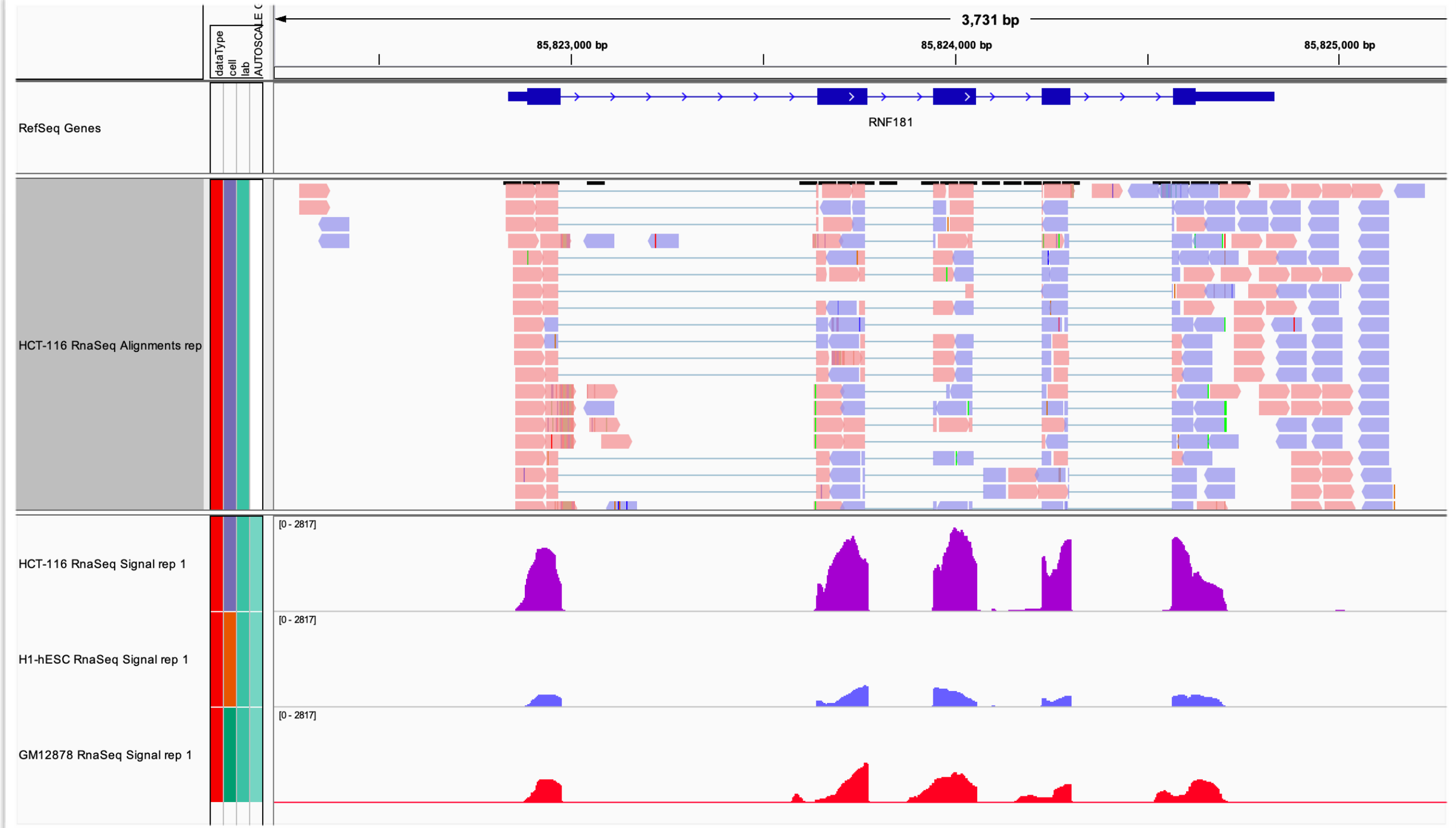


cDNA

genomic
DNA

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

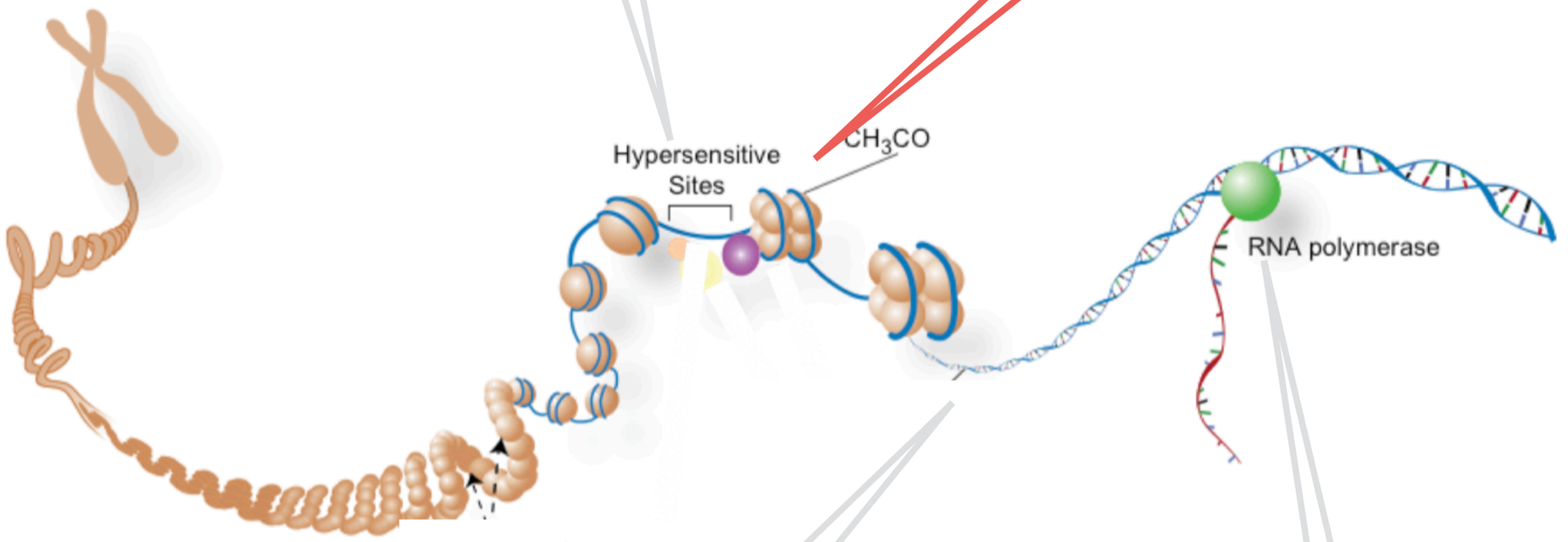
$$TPM_i = \frac{n_i}{L_i} \times \frac{10^6}{\sum_j \left(\frac{n_j}{L_j} \right)}$$

Regulatory genome

Medizinische Fakultät Heidelberg

which parts of the genome are accessible?

what is the chemical state of nucleosomes?



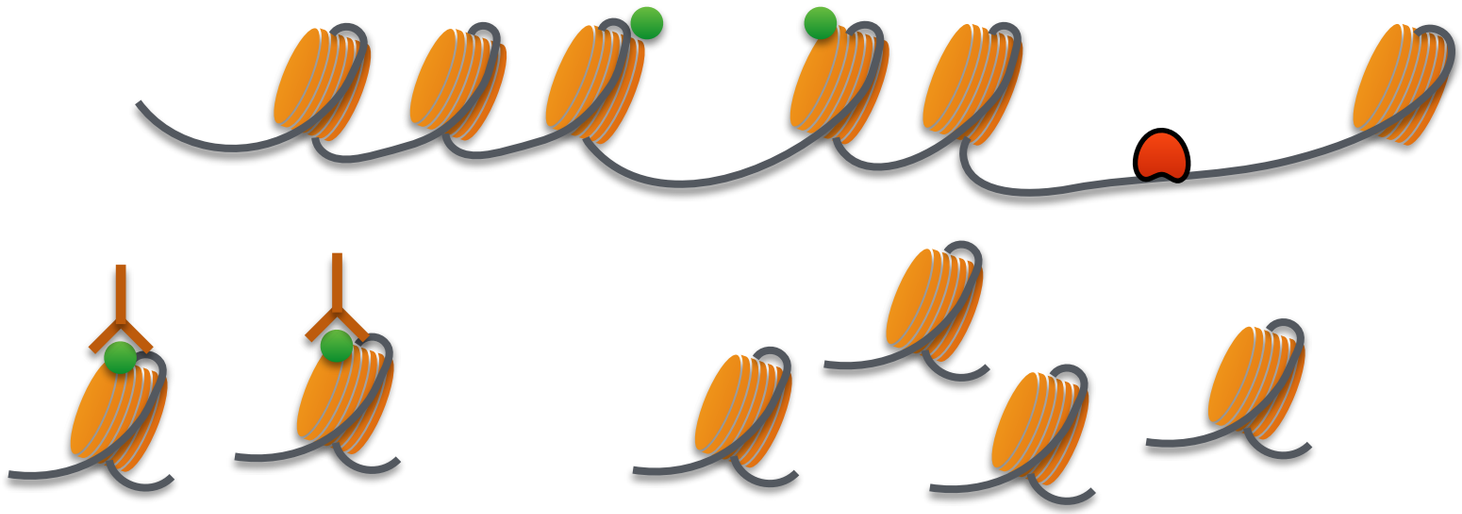
At which CpGs is DNA methylated?

What is the expression of genes?

ChIP-sequencing

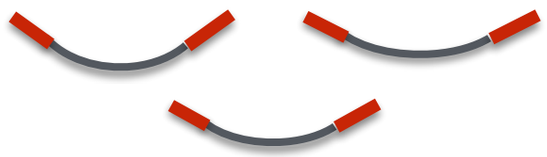
Medizinische Fakultät Heidelberg

fragmentation
(sonication, mnase)
and immunoprecipitation

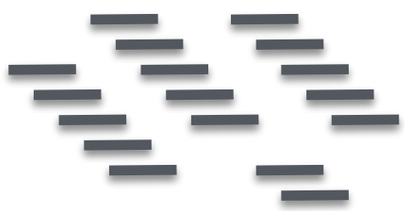


DNA fragments
get immunoprecipitated

Library preparation
(PCR,
sequencing adapters)



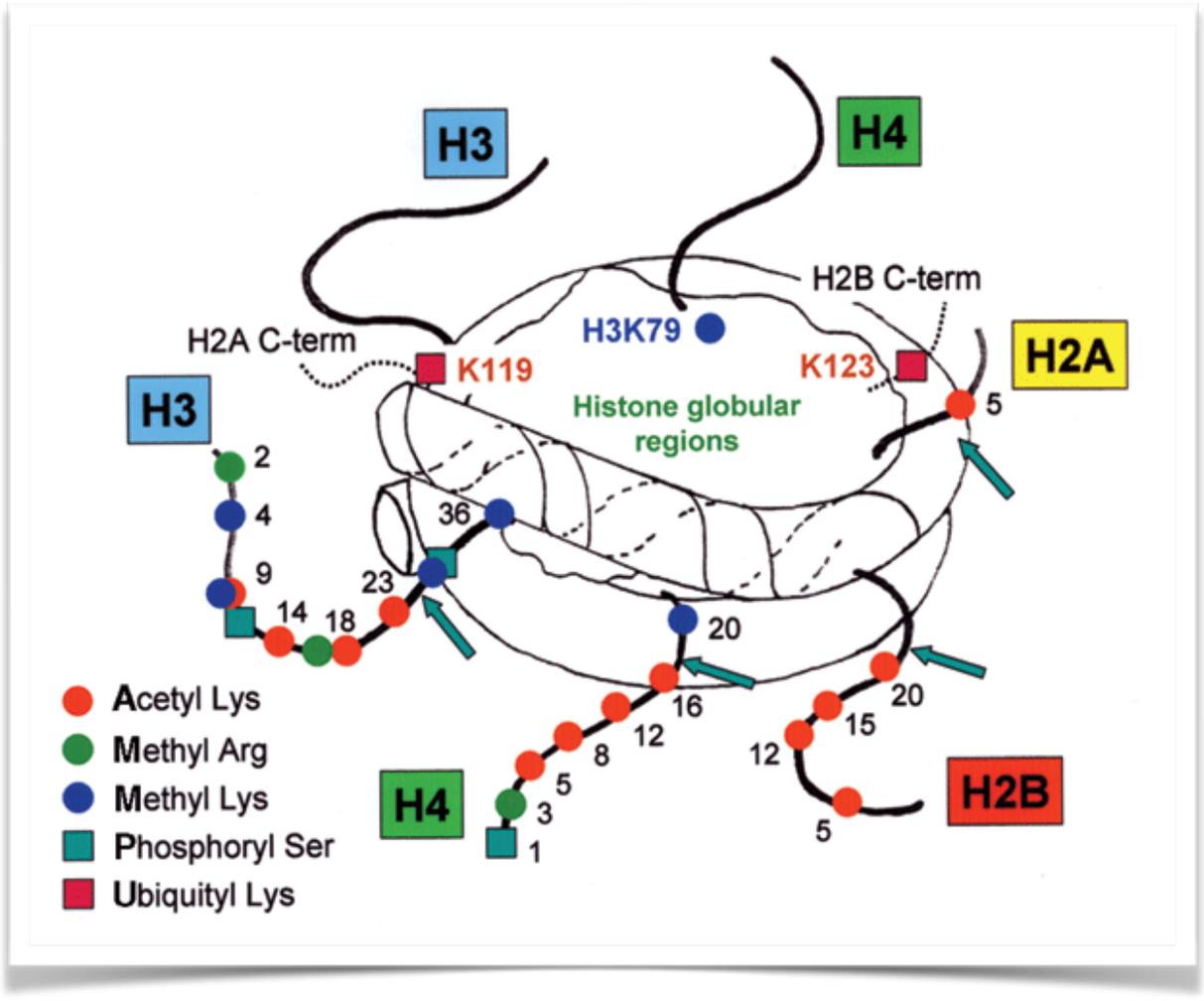
High-throughput
sequencing



Read alignment

Peak calling

ChIP-seq for histone modifications



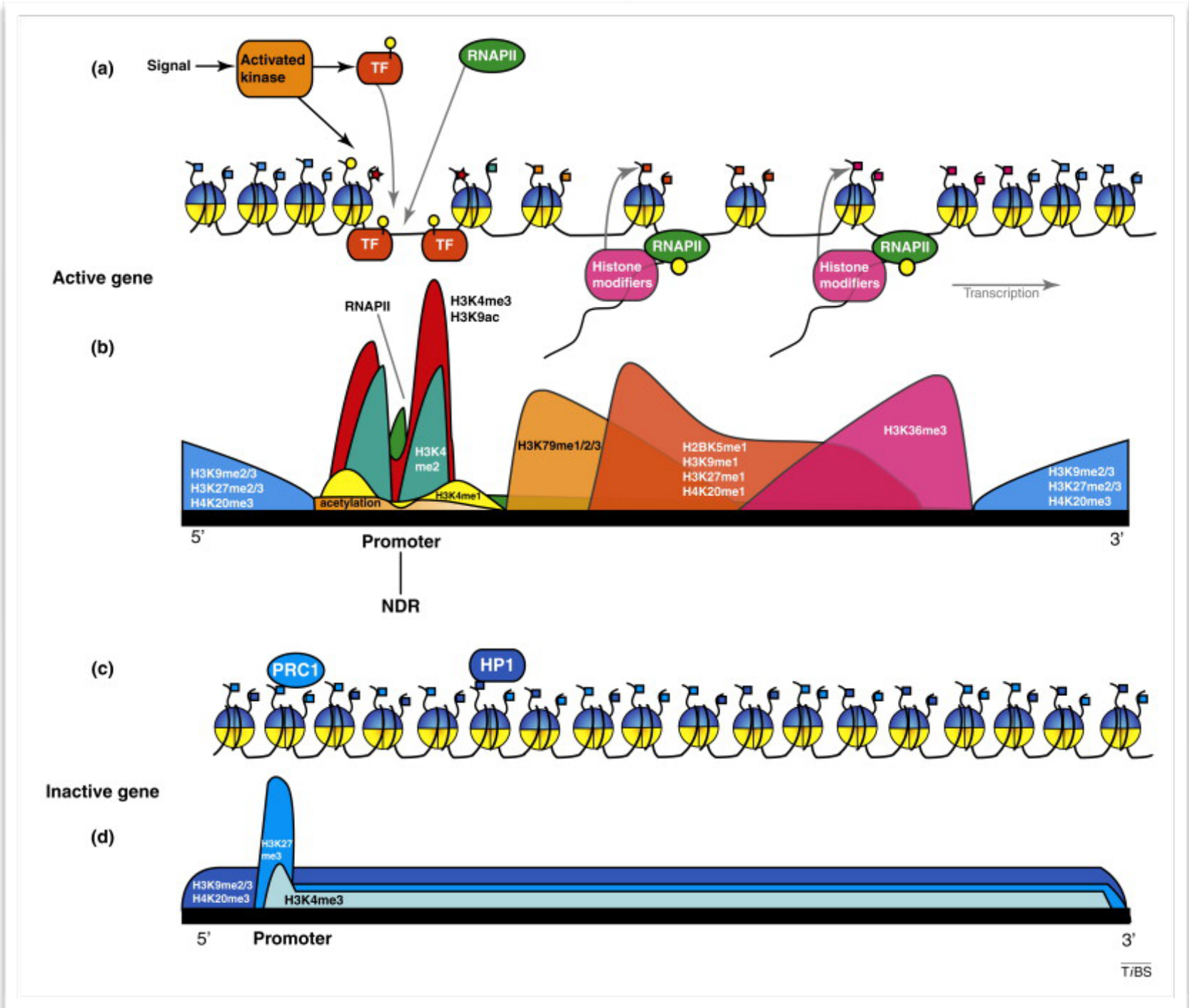
- histones are subject to **post-translational modifications** at their N-terminal tail
  - Lysine methylation
  - Lysine/arginine acetylation
  - Serine phosphorylation
  - ubiquitylation
- they **modify the physical properties of the DNA-nucleosome interactions**

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

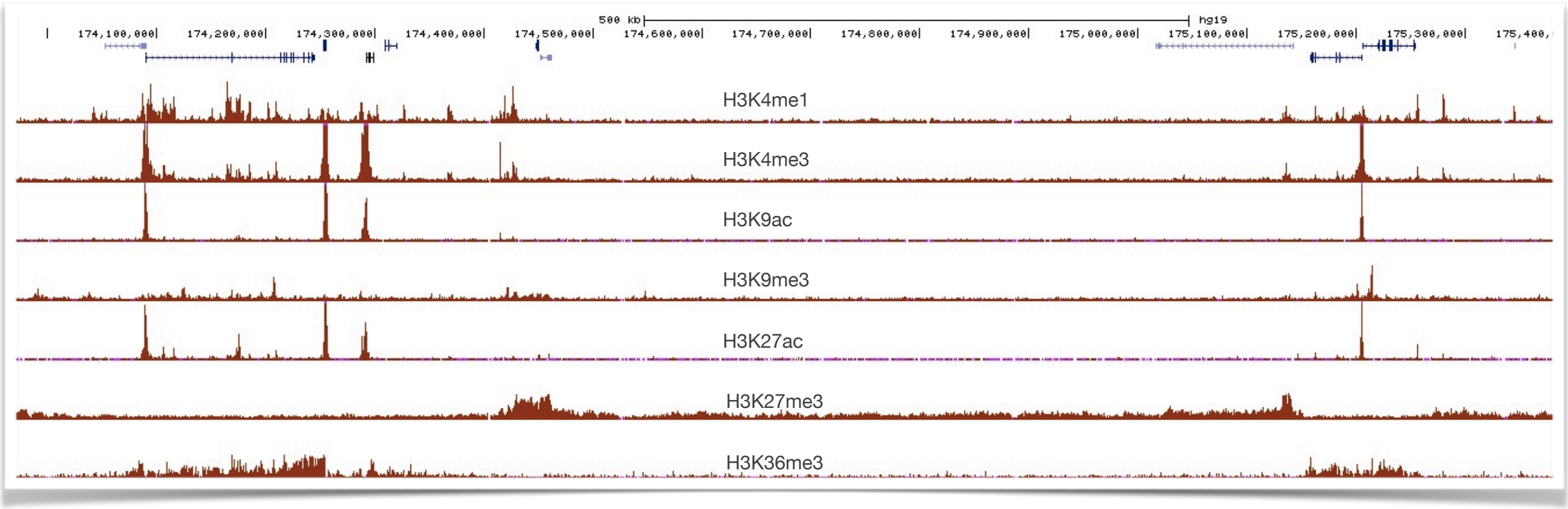


active marks
→ open chromatin
H3K4me1; H3K4me3; H3K27ac

repressive marks
→ closed chromatin
H3K27me3; H3K9me3

Histone modifications

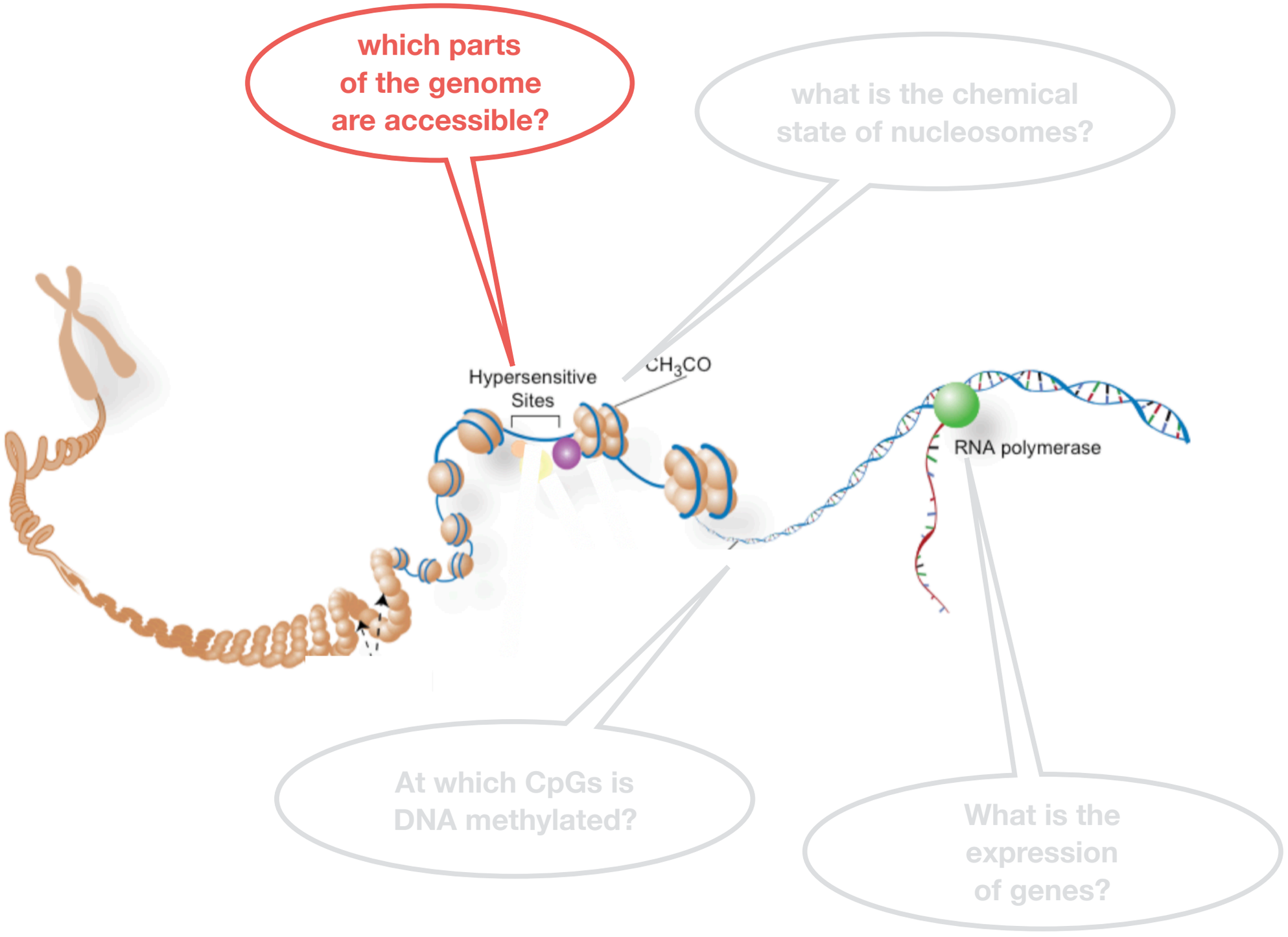
Medizinische Fakultät Heidelberg



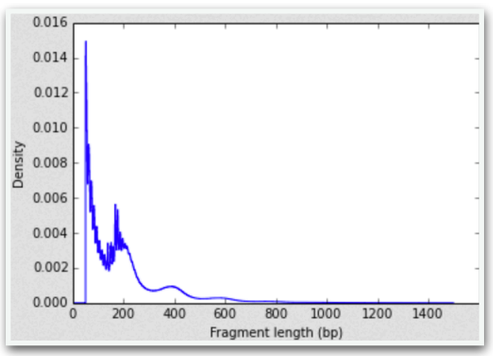
- Histone marks have **distinct signal profiles**
  - sharp signal :
    - H3K4me3 = promoters
    - H3K27ac = enhancers,…
  - broad signal :
    - H3K36me3 = transcribed genes
    - H3K27me3 = repressed regions)

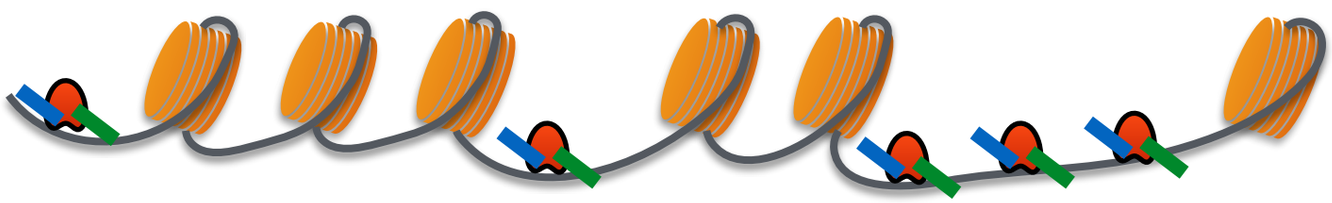
Regulatory genome

Medizinische Fakultät Heidelberg

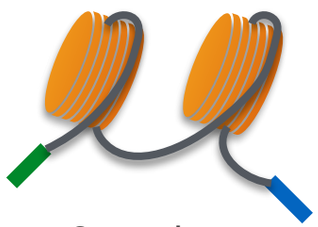


ATAC-sequencing





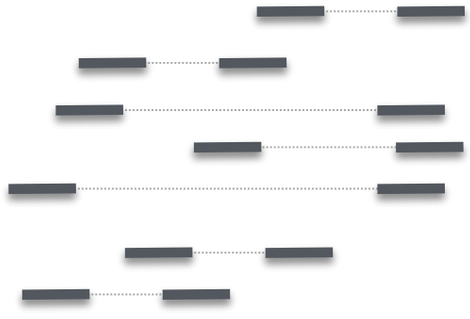
3-nucleosome fragments



2-nucleosome fragments

nucleosome-free fragments

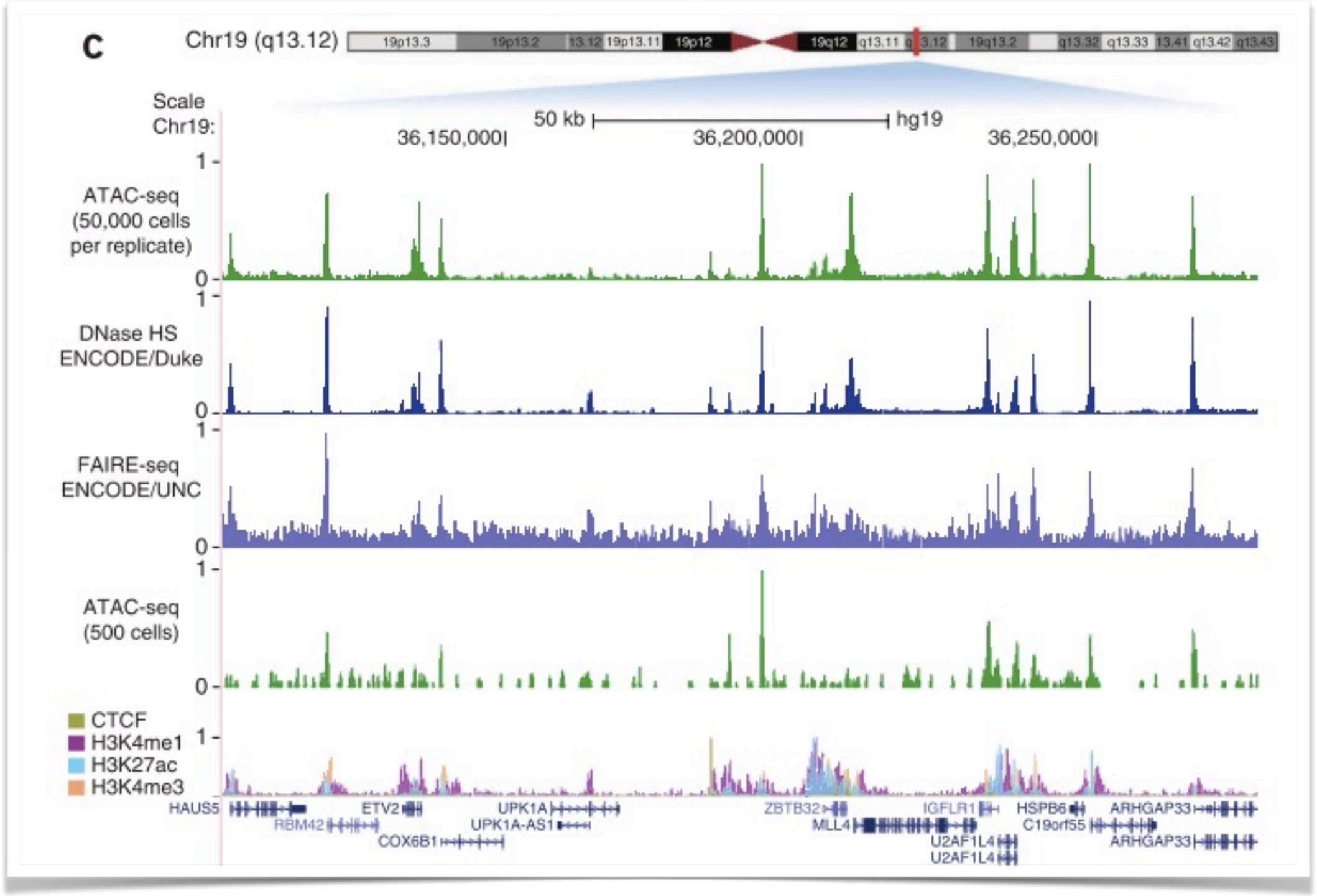
High-throughput paired-end sequencing



Read alignment

Peak calling

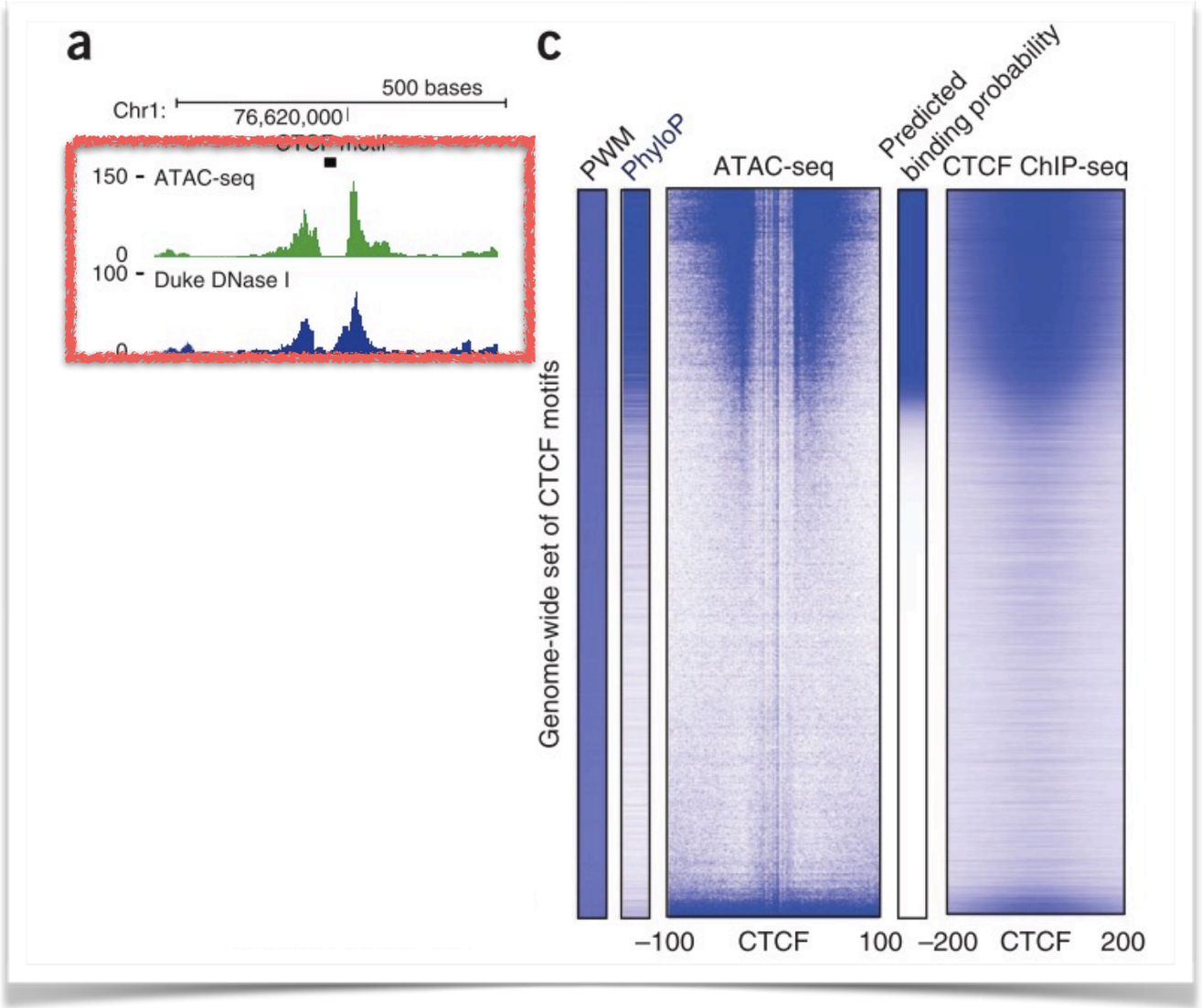
Identification of accessible regions



[Greenleaf (2013)]

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



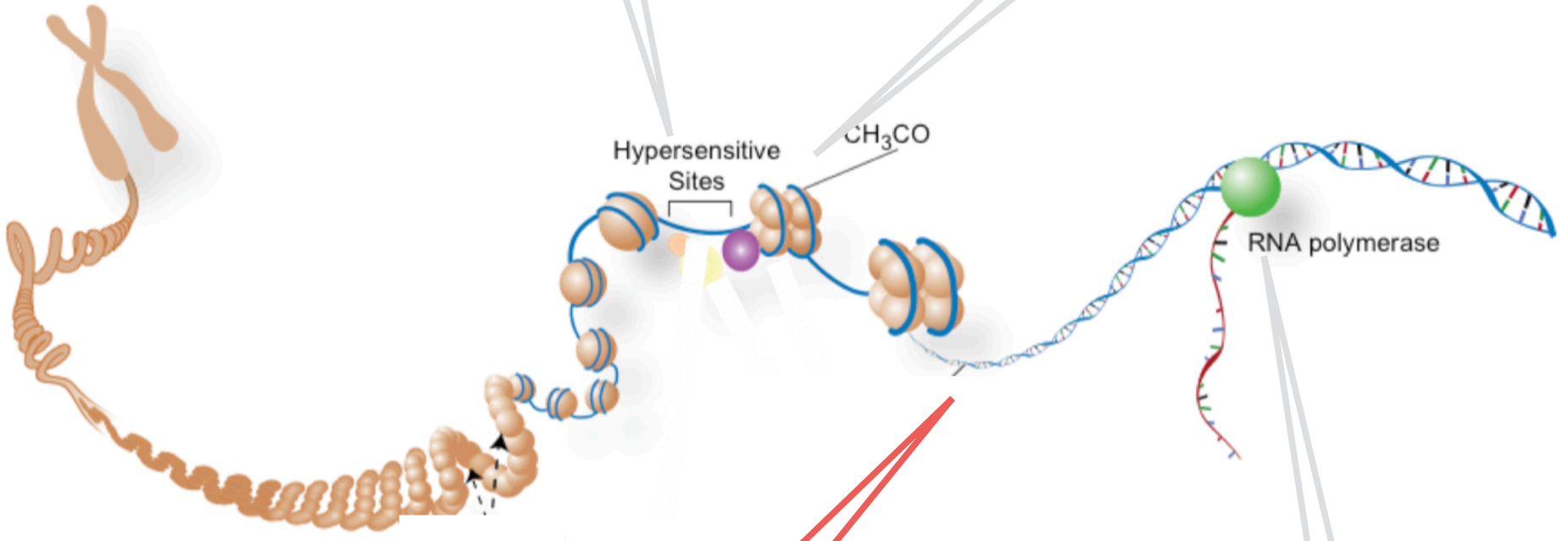
[Greenleaf (2013)]

Regulatory genome

Medizinische Fakultät Heidelberg

which parts of the genome are accessible?

what is the chemical state of nucleosomes?

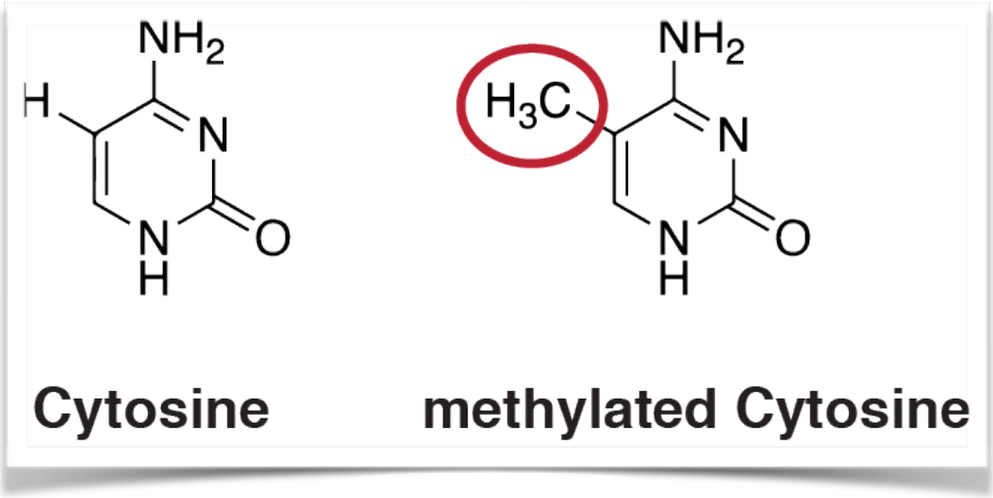


At which CpGs is DNA methylated?

What is the expression of genes?

Measuring DNA methylation

- DNA methylation occurs mainly on **cytosines in CpG** dinucleotides in the human genome
- DNA methylation is revealed by using **bisulfite conversion** (HSO_3^-):
  - unmethylated cytosines are converted C → U → T
  - methylated cytosines are protected mC → mC

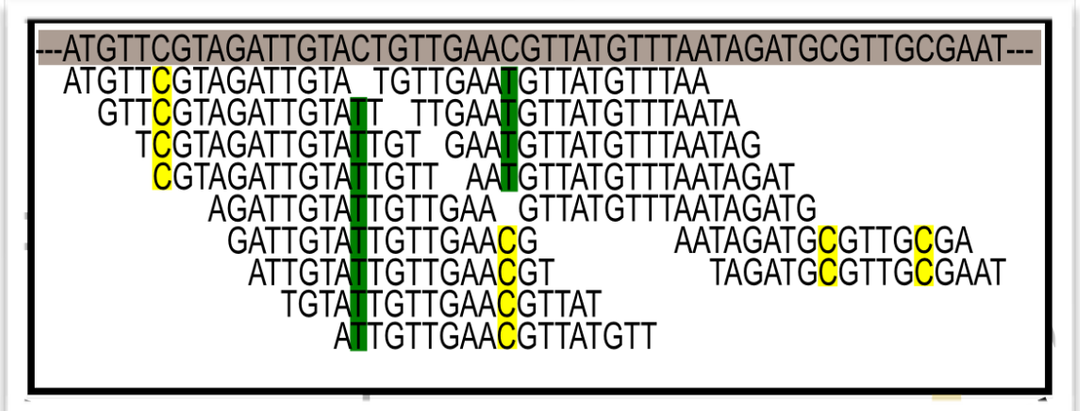


- unmethylated CpG are identified by the presence of a **mismatch TpG**
- 2 approaches:
  - array based: hybridization to CpG probes on array
  - sequencing: whole genome bisulfite-sequencing

Measuring DNA methylation

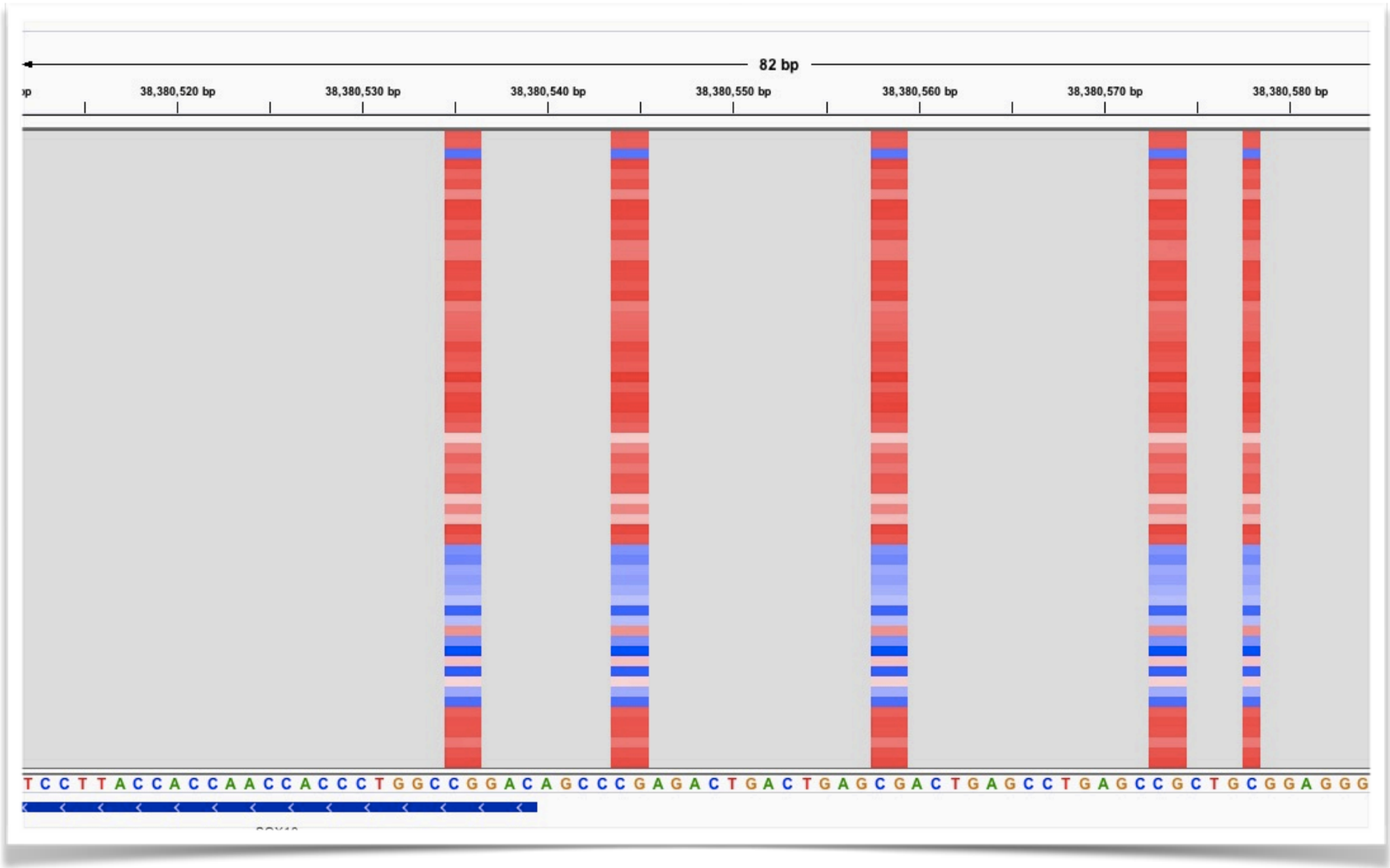
- Array based methods
- CpG containing probes on array
  - 27K probes
  - 450K probes
  - 800K (EPIC)
- all probes contain a methylated (C) and unmethylated (T) version
- Cheap but sparse

- Whole genome bisulfite sequencing
  - unmethylated C → T
  - methylated C → C
- Shearing, conversion and sequencing (Illumina X-10)
- Information about the 28 million CpGs



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

Larger data matrices

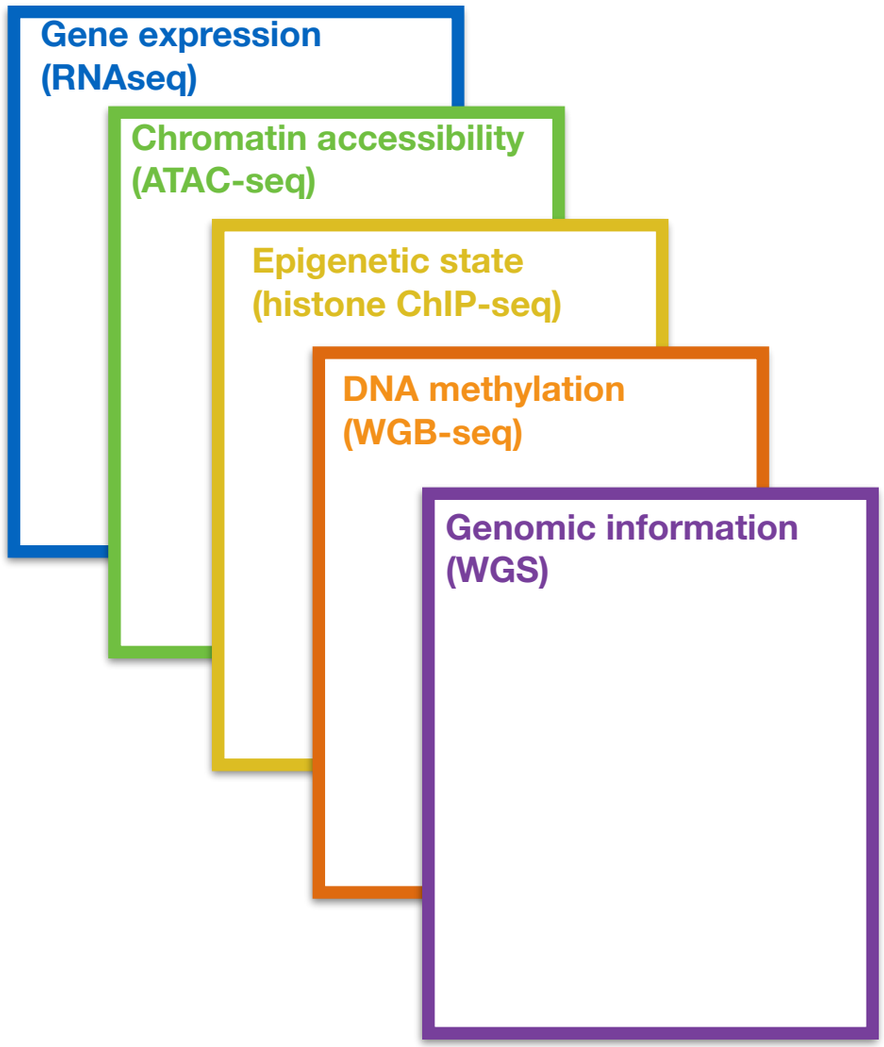
~ 10s samples

target sequencing
~10 regions / genes

~ 1000s samples

whole genome / transcriptome
~10.000s features
(genes / regions)

Many data types (“views")



“Whole more than sum of the parts”

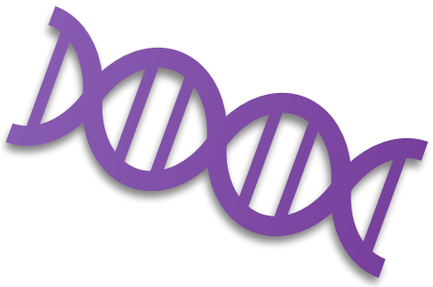
Medizinische Fakultät Heidelberg

Chromatin accessibility (ATAC-seq)

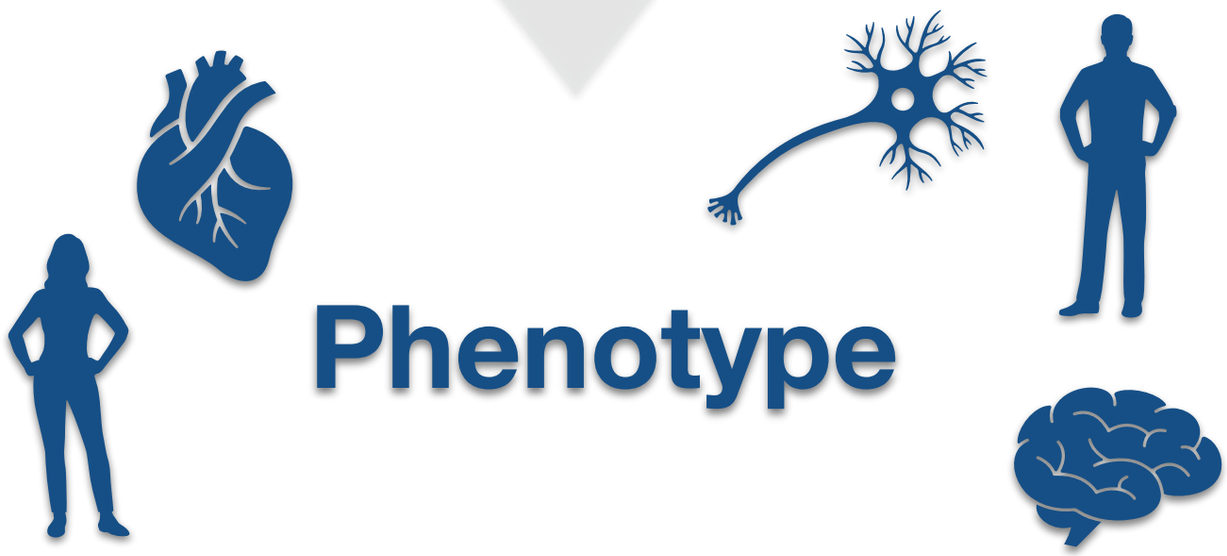
Epigenetic state (histone ChIP-seq)

DNA methylation (WGB-seq)

Gene expression (RNAseq)



Genomic information (WGS)



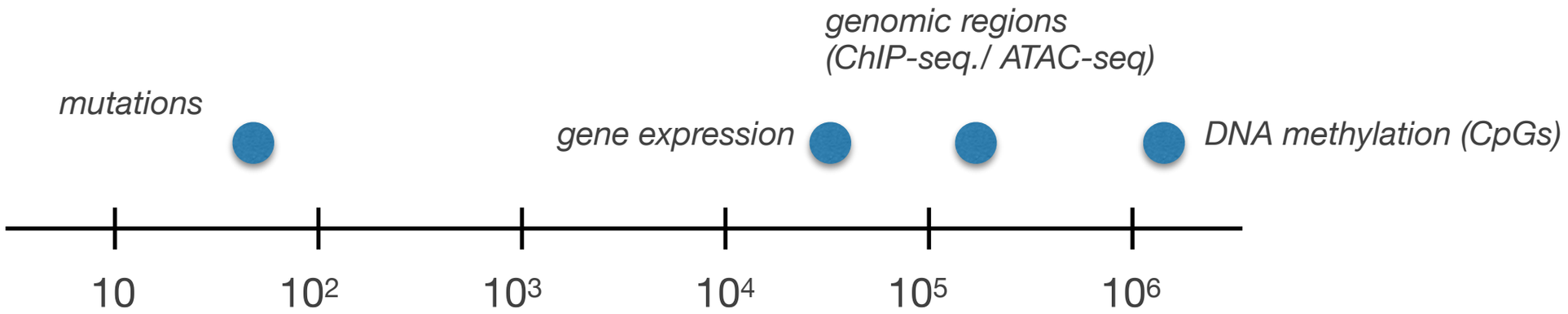
Medizinische Fakultät Heidelberg

3. Dimensional reduction and data integration

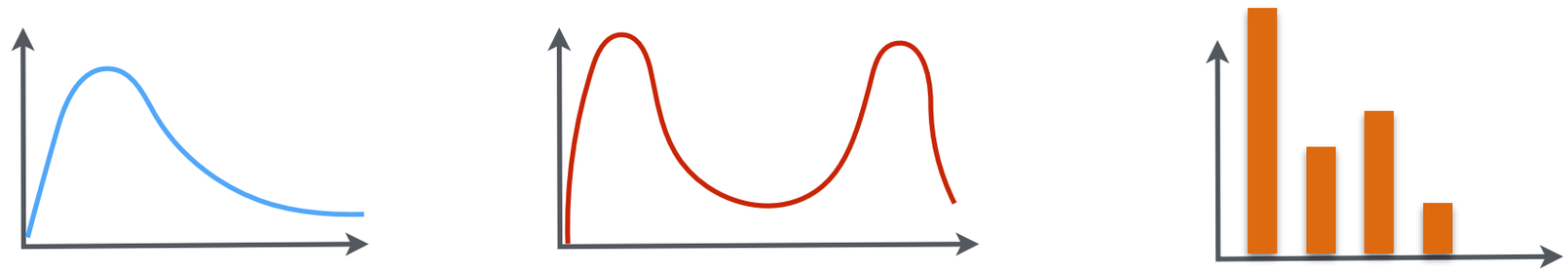
- Clustering; Consensus Clustering
- Clusters of Clusters; iCluster
- Principal Component analysis
- Non-negative matrix factorization and its flavors…

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

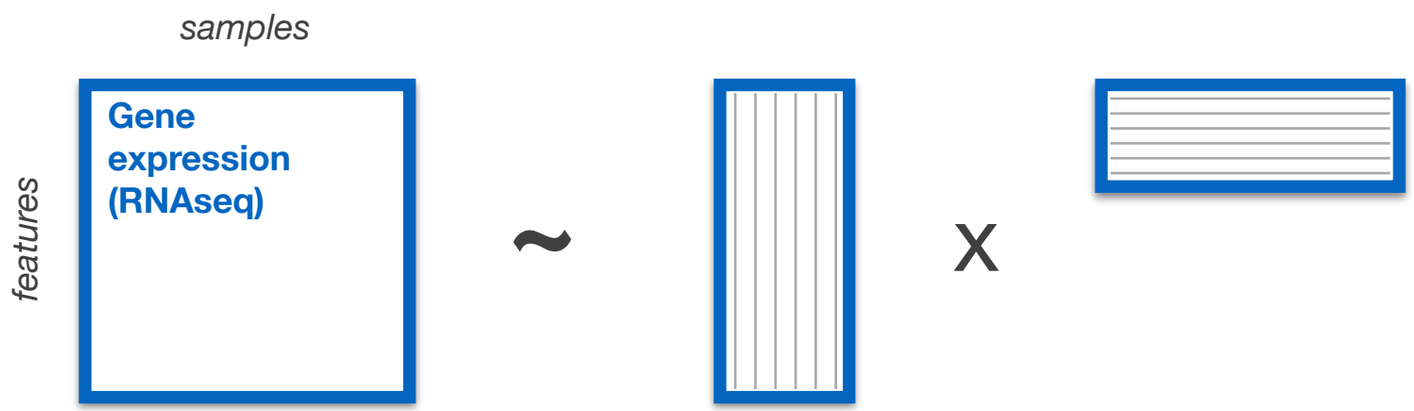
Medizinische Fakultät Heidelberg

- **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)

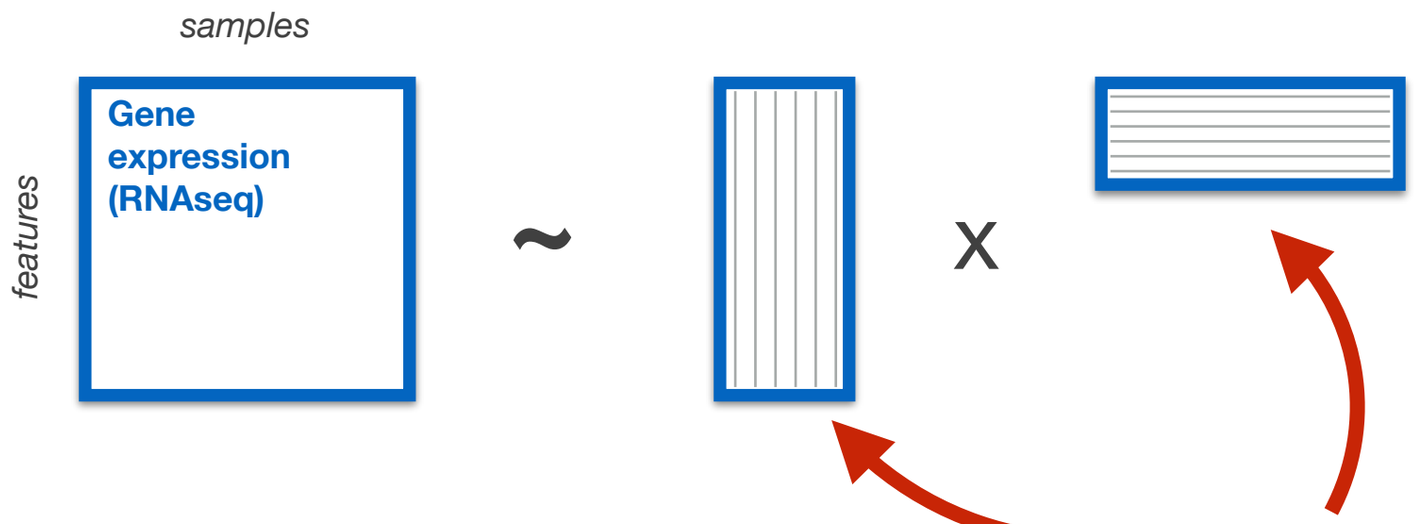
Matrix factorization



$$X = WH + \epsilon$$

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

Matrix factorization



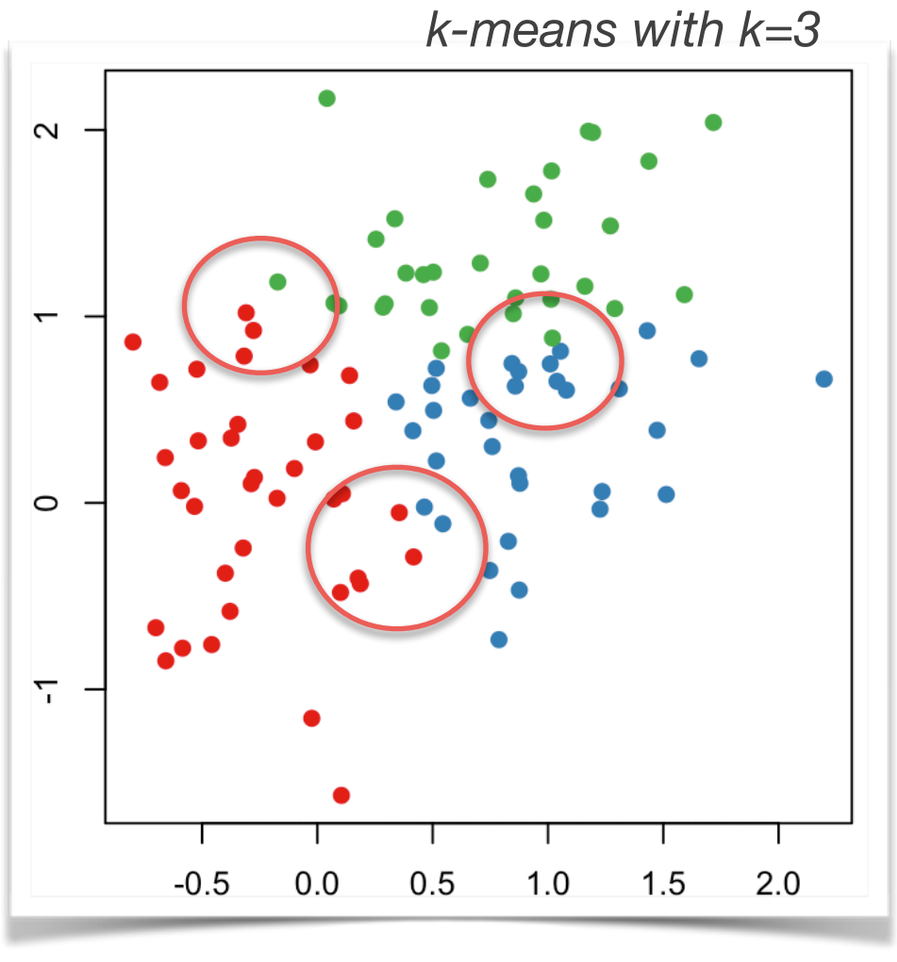
$$X = WH + \epsilon$$

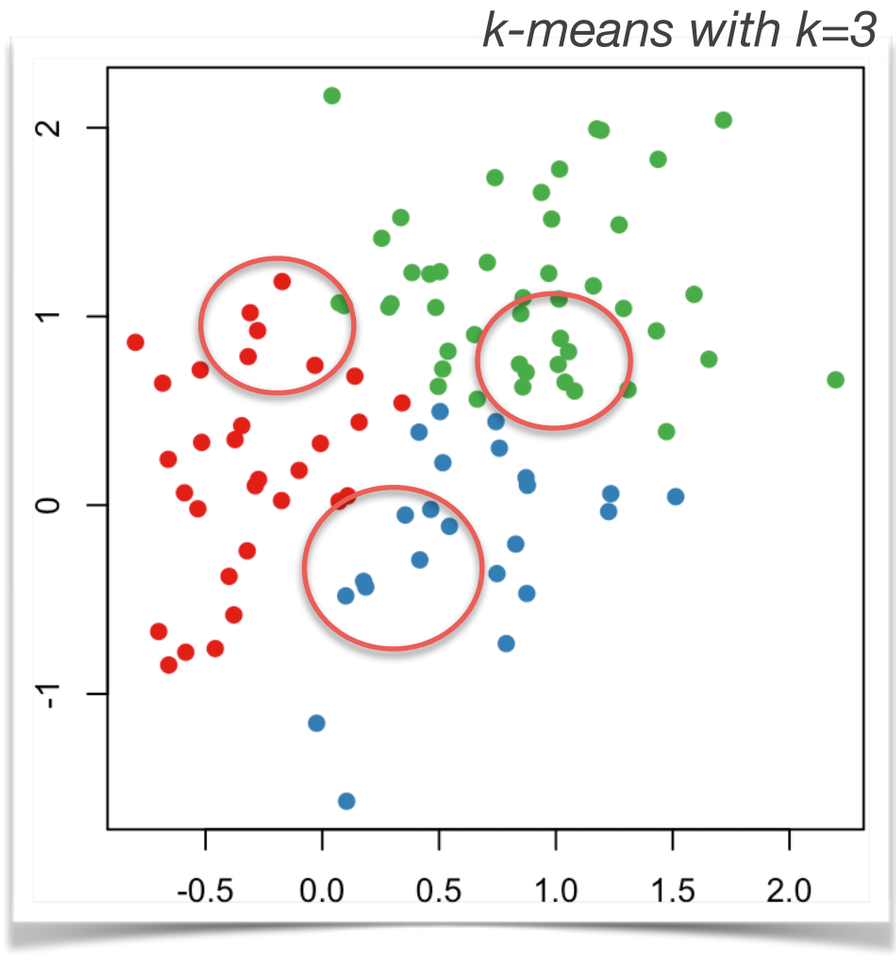
- Many flavors!
  - cluster analysis
  - principal components analysis
  - (group) factor analysis
  - non-negative matrix factorization
  - etc …

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!

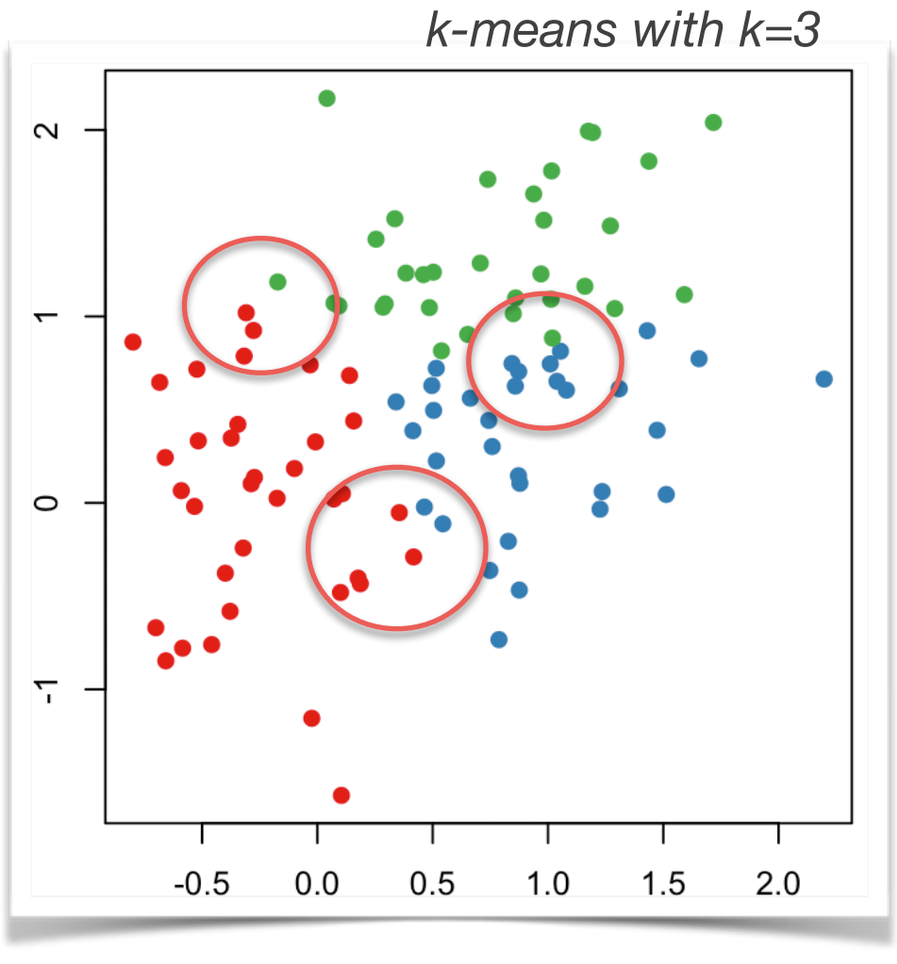


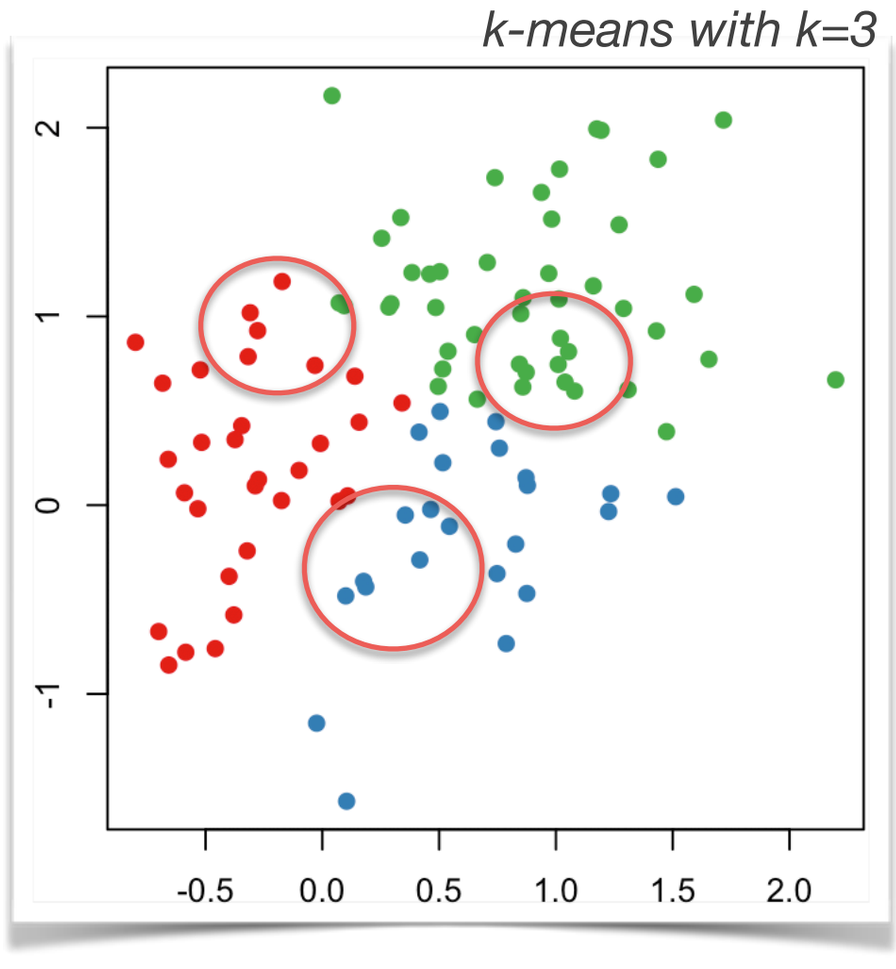


Clustering

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**:

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

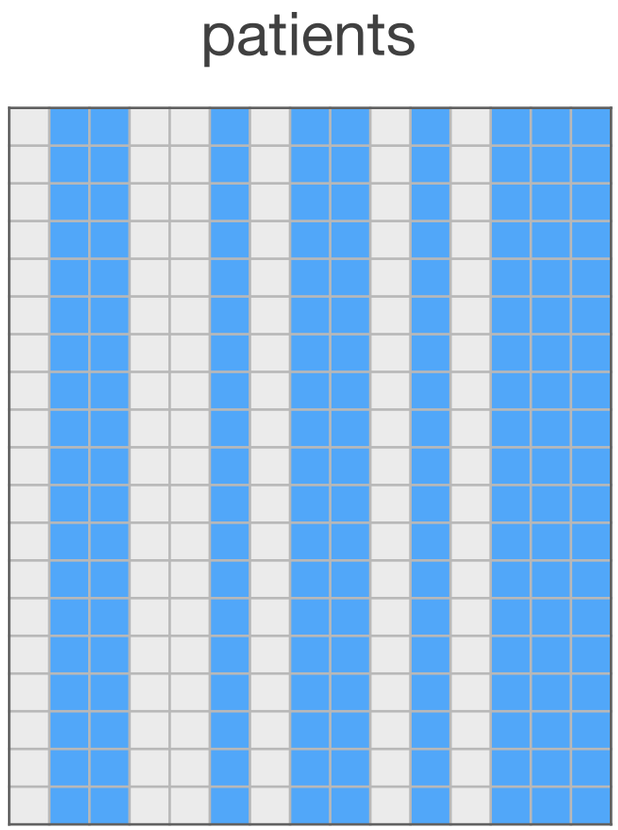
$D^{(h)}$ subset of the patients (e.g. 80%)

$M^{(h)}$ 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 $D^{(h)}$

$$m(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)} \qquad d(i,j) = 1 - m(i,j)$$



blue columns = sampled patients

Use the matrix d to perform (hierarchical) clustering

Consensus clustering

Medizinische Fakultät Heidelberg

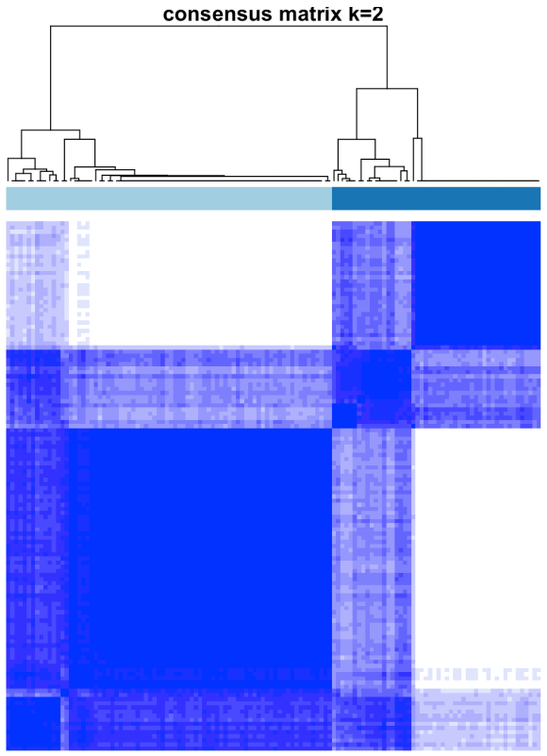
```
> results[[2]][["consensusMatrix"]][1:5,1:5]
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 1.0000000 1.0000000 0.9655172 1.0000000 1.0000000
[2,] 1.0000000 1.0000000 0.8857143 1.0000000 1.0000000
[3,] 0.9655172 0.8857143 1.0000000 0.9166667 0.8823529
[4,] 1.0000000 1.0000000 0.9166667 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 $k = 2$

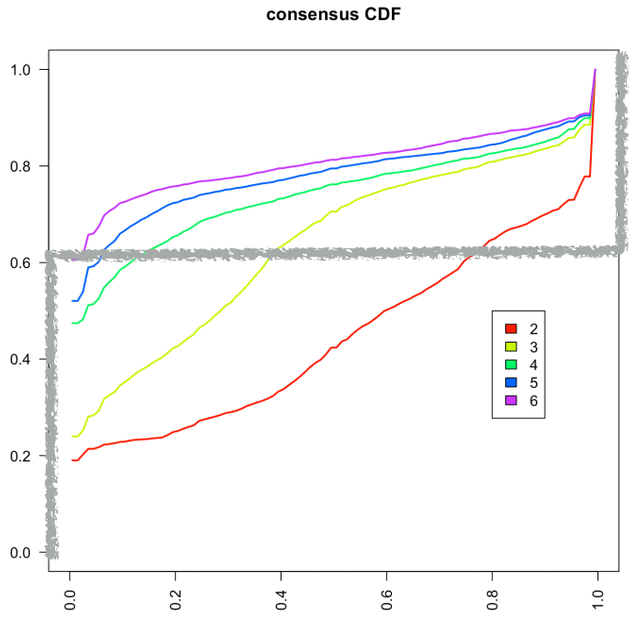
similarity matrix
for $k = 3$

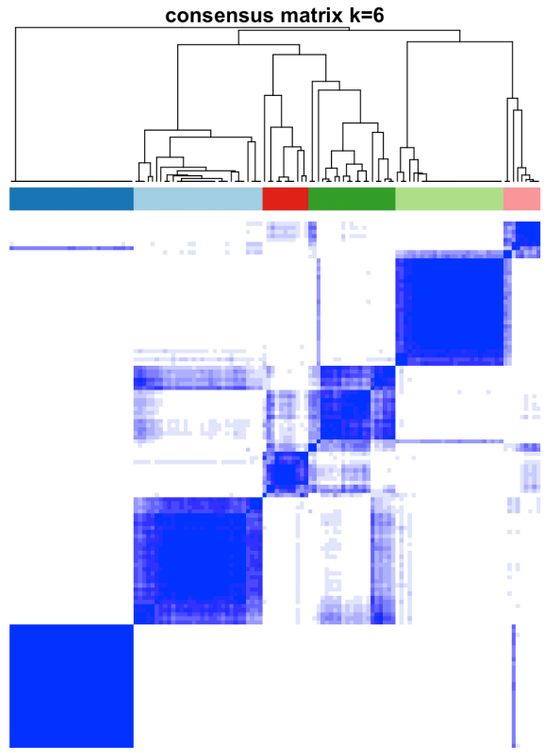
Consensus Clustering

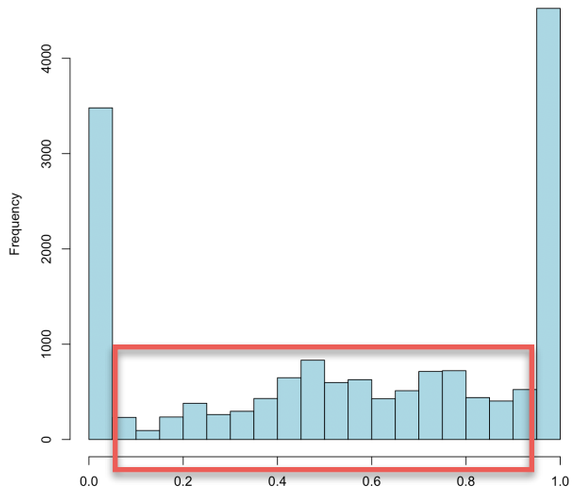
Medizinische Fakultät Heidelberg



Ideal shape would be a step function

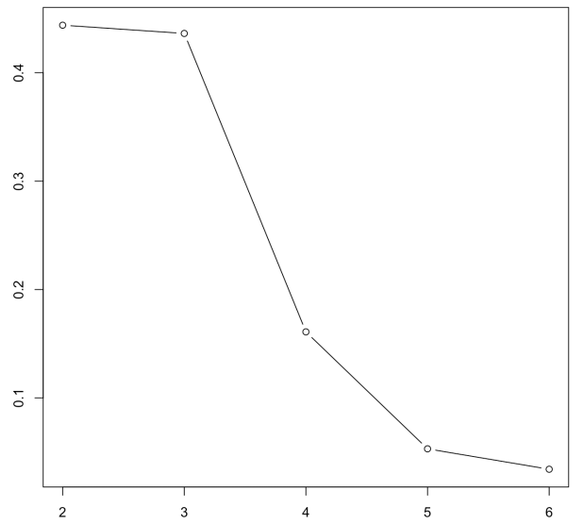




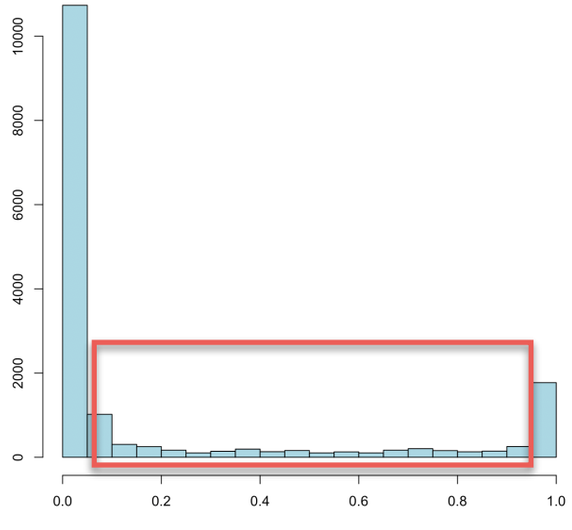


badly assigned samples

Differential AUC



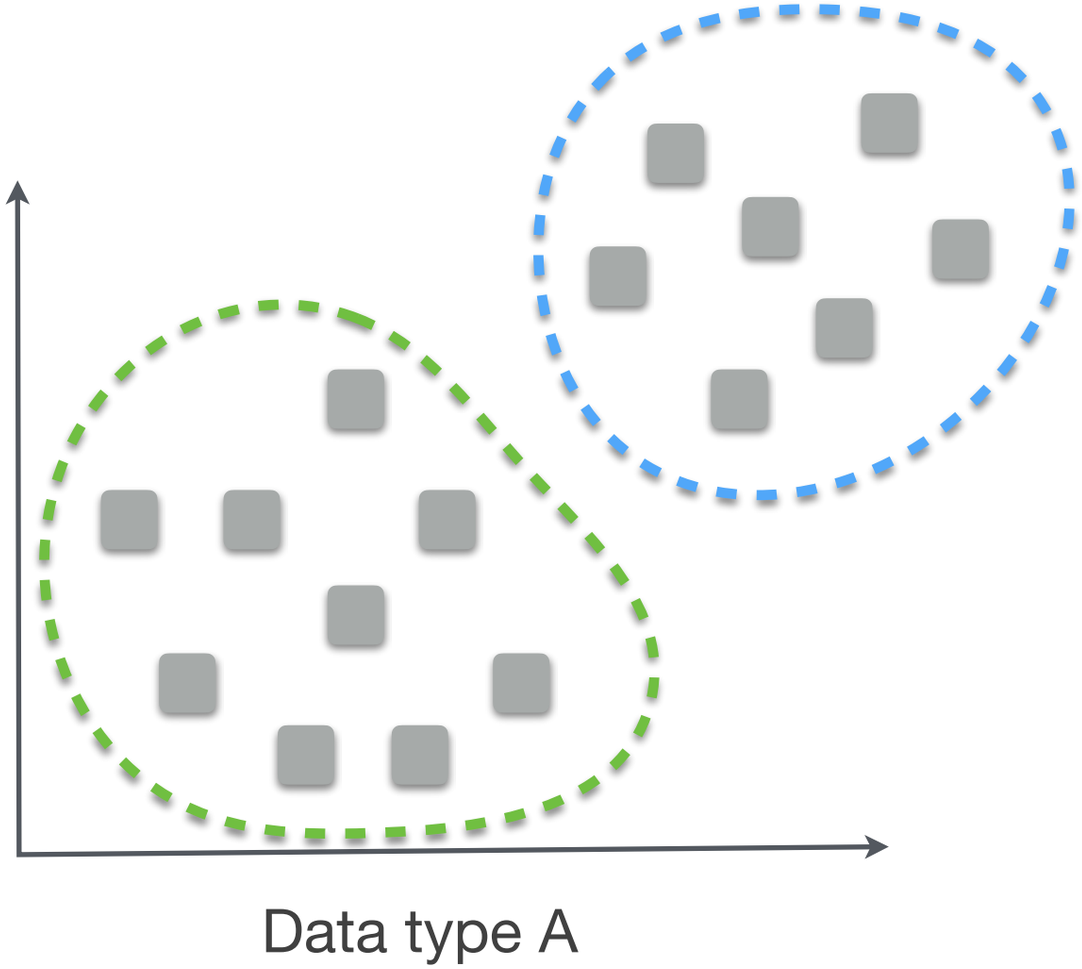
Optimal K when AUC no longer increases



badly assigned samples

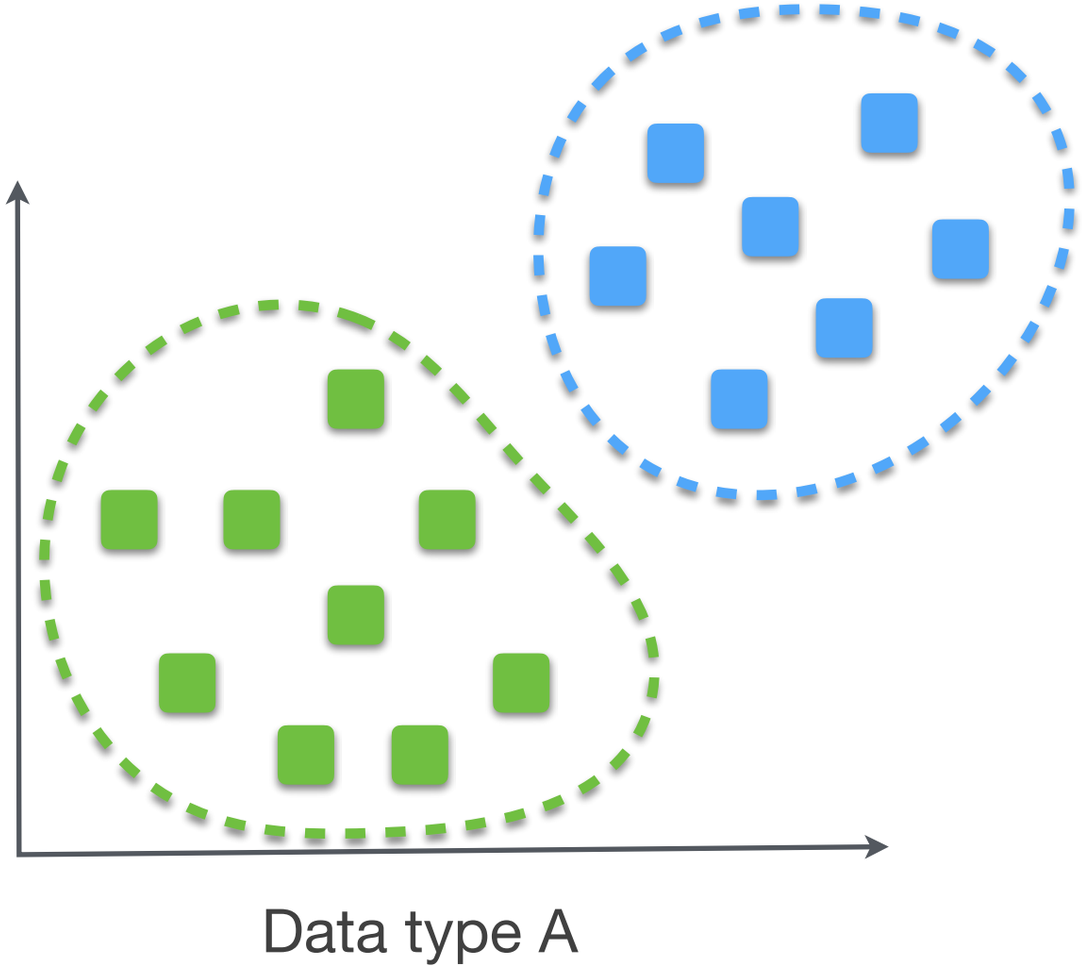
[Monti et al., 2003]

Clustering over multiple data?

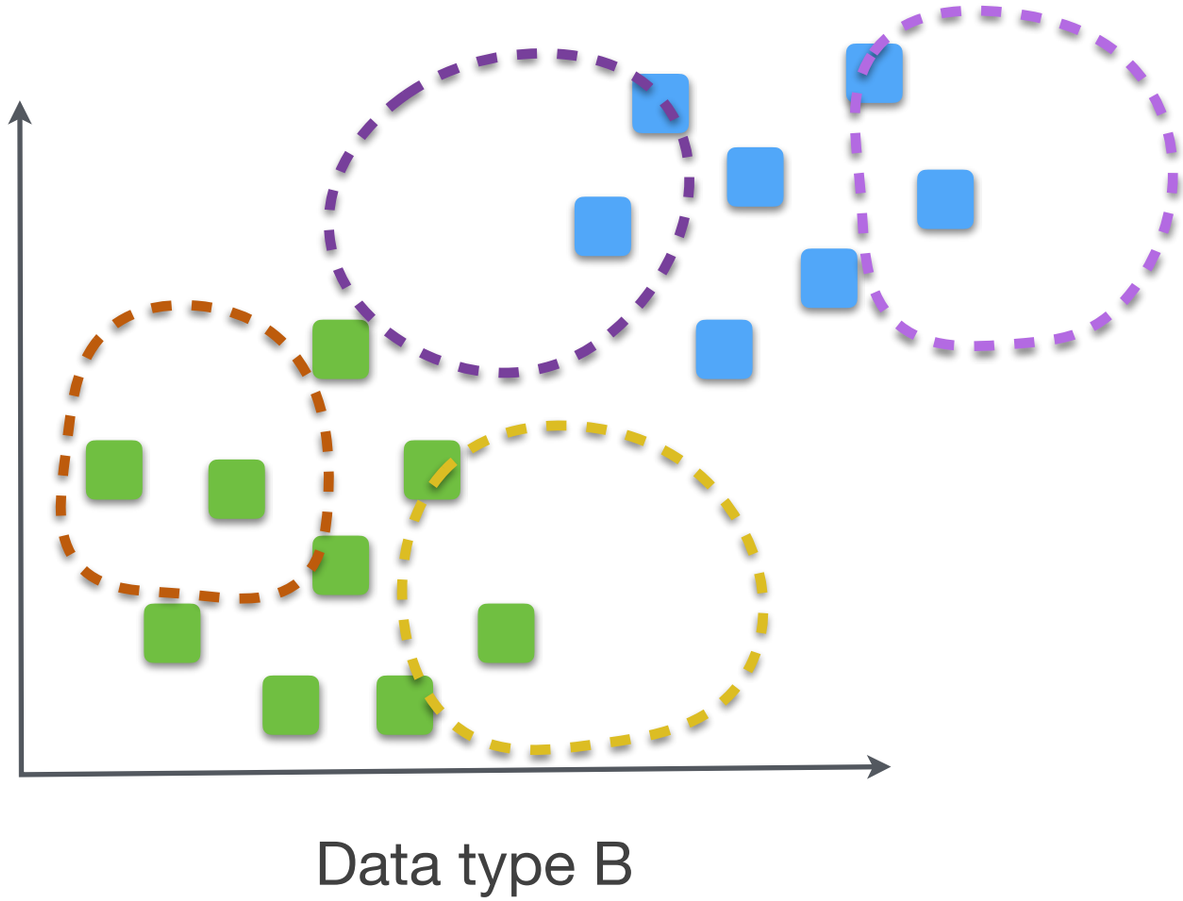


Clustering over multiple data?

Medizinische Fakultät Heidelberg



Data type A



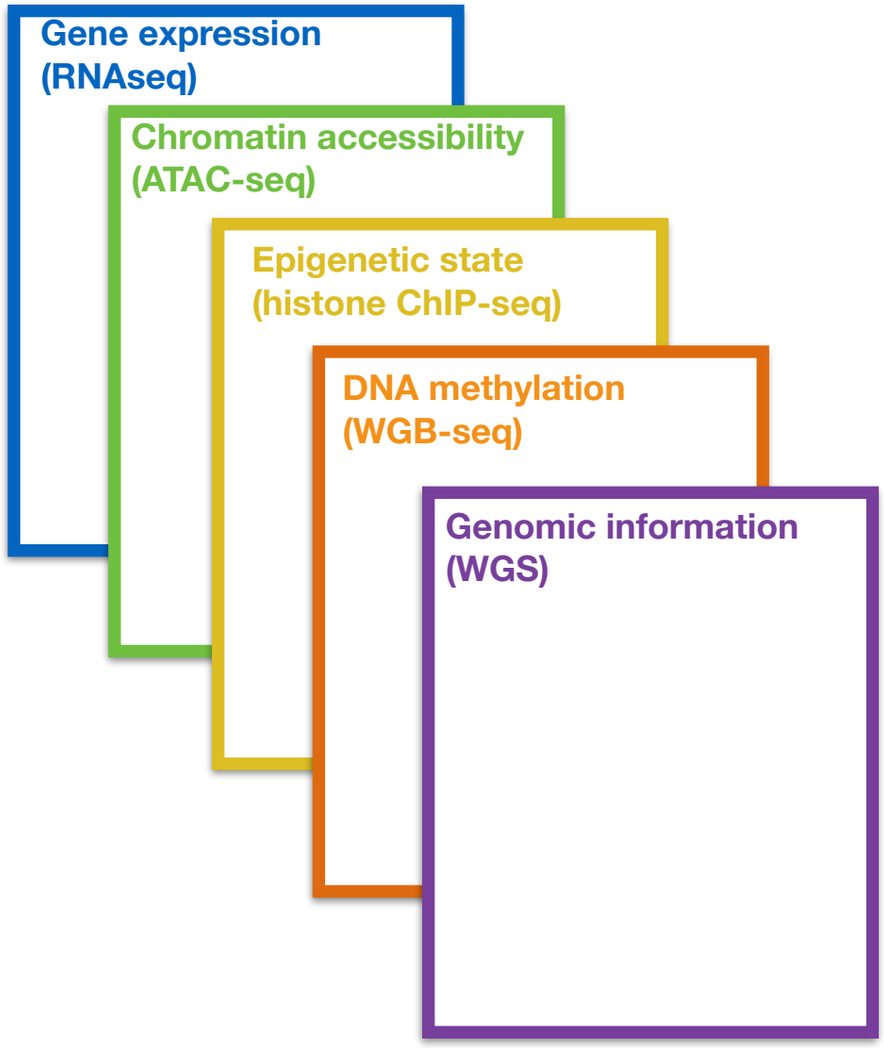
Data type B

Cluster of Cluster Analysis (CoCA)

Medizinische Fakultät Heidelberg

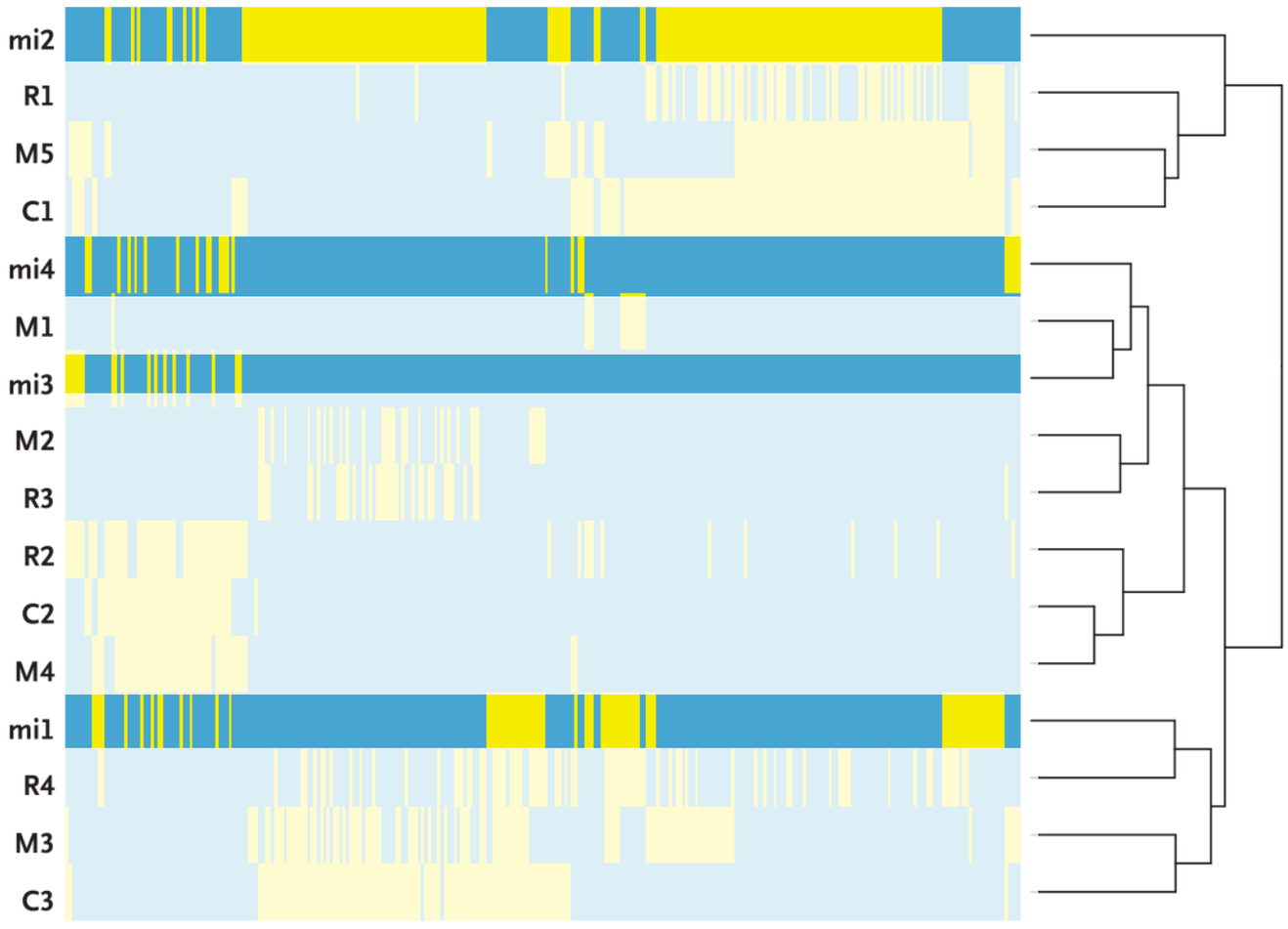
- Cluster each omics data separately
  - each clustering can use a different clustering algorithm (k-means, PAM,…)
  - each omics datatype can lead to distinct number of clusters
- Represent each sample by an **indicator vector** showing to which cluster it belongs in each omic

  s_3 = (**1** , **3**, **2** , **3** , **1**)
- Cluster the samples based on this indicator vector using **consensus clustering**



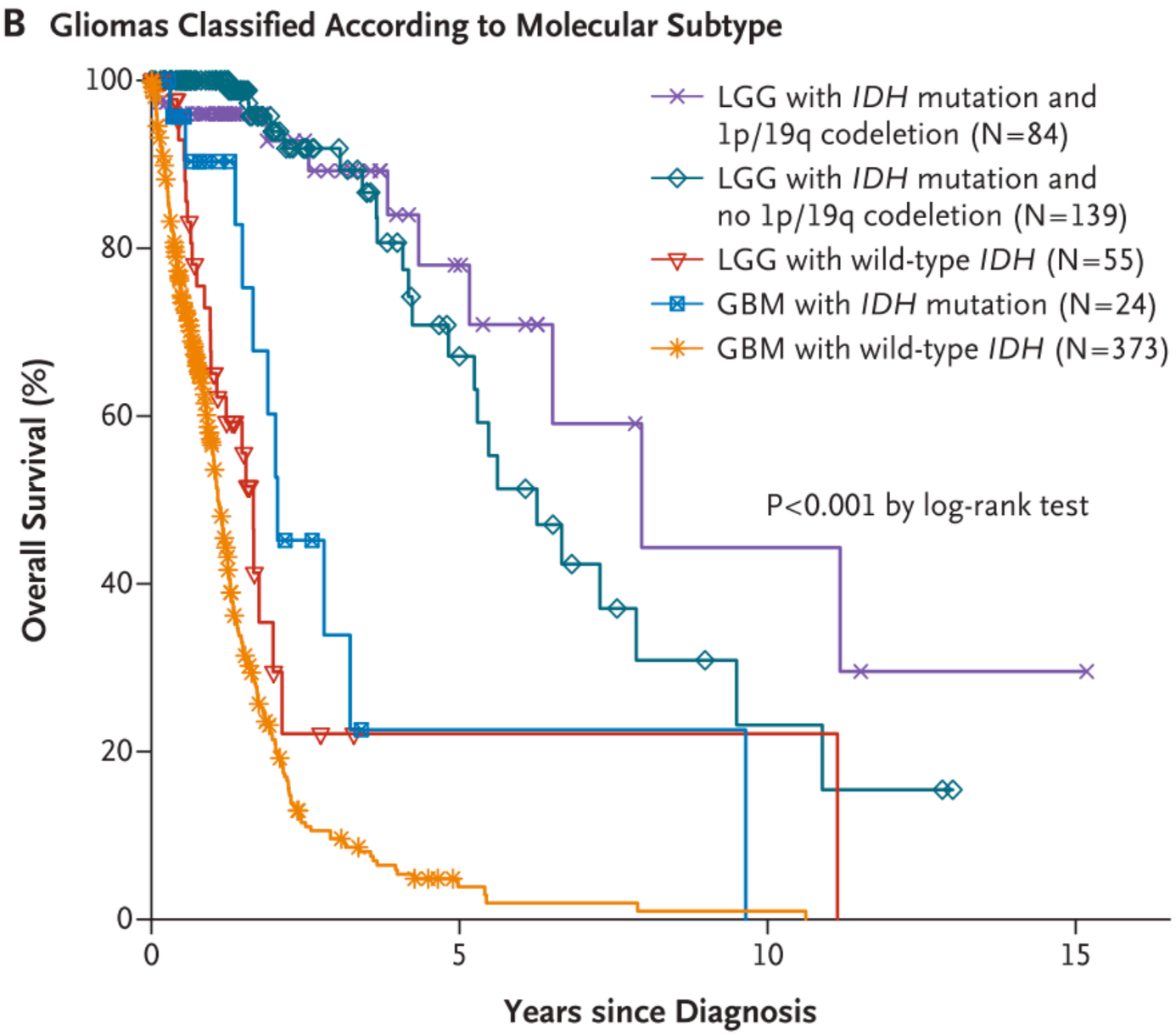
Application: low-grade glioma

- 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!



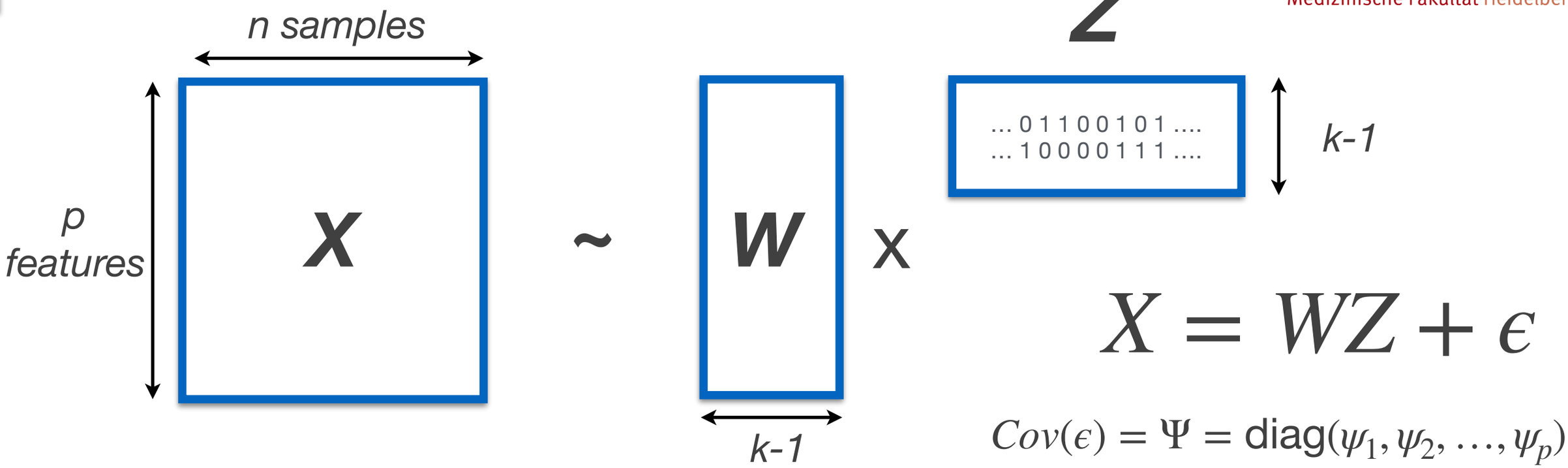
[Brat et al., NJEM, 2015]

Application: low-grade glioma



[Brat et al., NJEM, 2015]

iCluster



$$X = WZ + \epsilon$$

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

- Goal: identify *k* clusters of samples in the dataset such (i.e. Z) such 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

iCluster

Medizinische Fakultät Heidelberg

$$X = WZ + \epsilon$$

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

- ***X*** is **observed**
- ***W*** and Ψ 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

$$E[Z^* \mid X]$$

iCluster

Medizinische Fakultät Heidelberg

$$X = WZ + \epsilon \qquad \text{Cov}(\epsilon) = \Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)$$

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

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

(Expectation Step)

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

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

with $\Sigma = WW' + \Psi$

iCluster

Medizinische Fakultät Heidelberg

$$X = WZ + \epsilon \qquad \text{Cov}(\epsilon) = \Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)$$

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

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

(Expectation Step)

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

$$E[Z^* \mid X] = W'\Sigma^{-1}X$$
$$\text{with} \quad \Sigma = WW' + \Psi$$

(Maximization Step)

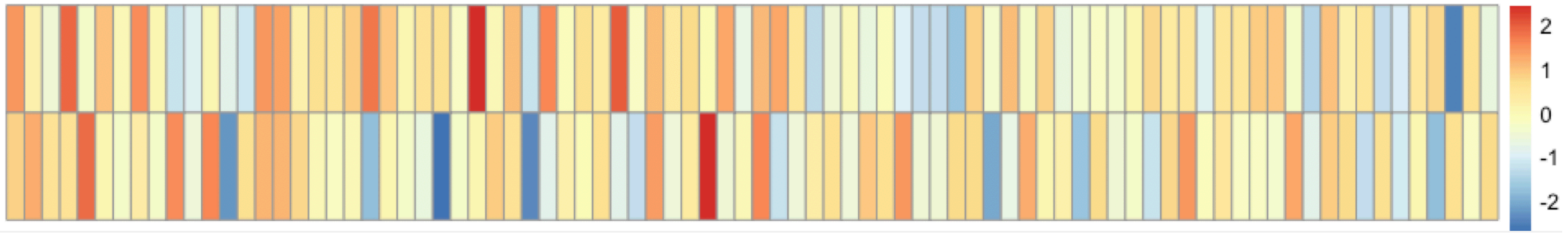
Estimate $(W^{(t+1)}, \Psi^{(t+1)})$
using $(Z^{(t)}, X)$

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

iCluster

Medizinische Fakultät Heidelberg

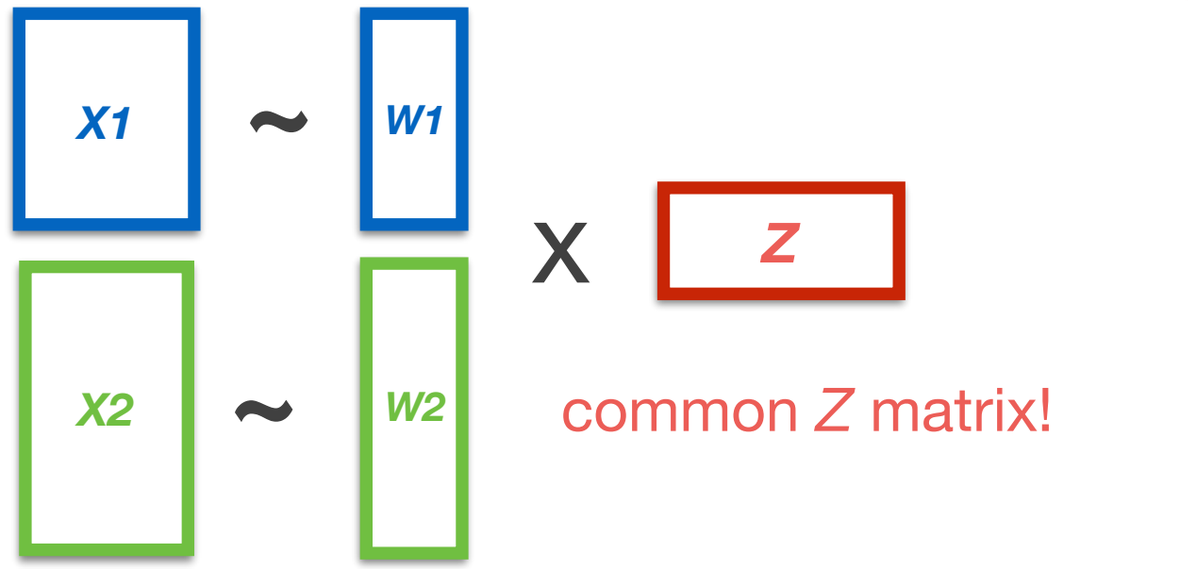
Infered posterior probability $E[Z^*|X]$ *(for k=2)*



Cluster indicator for *k+1=3* clusters

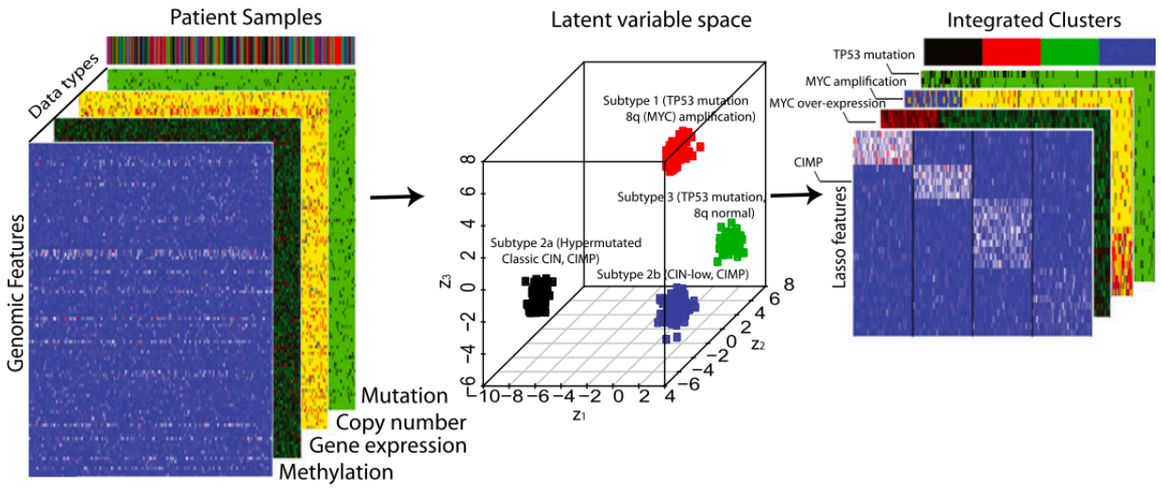
iCluster+

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

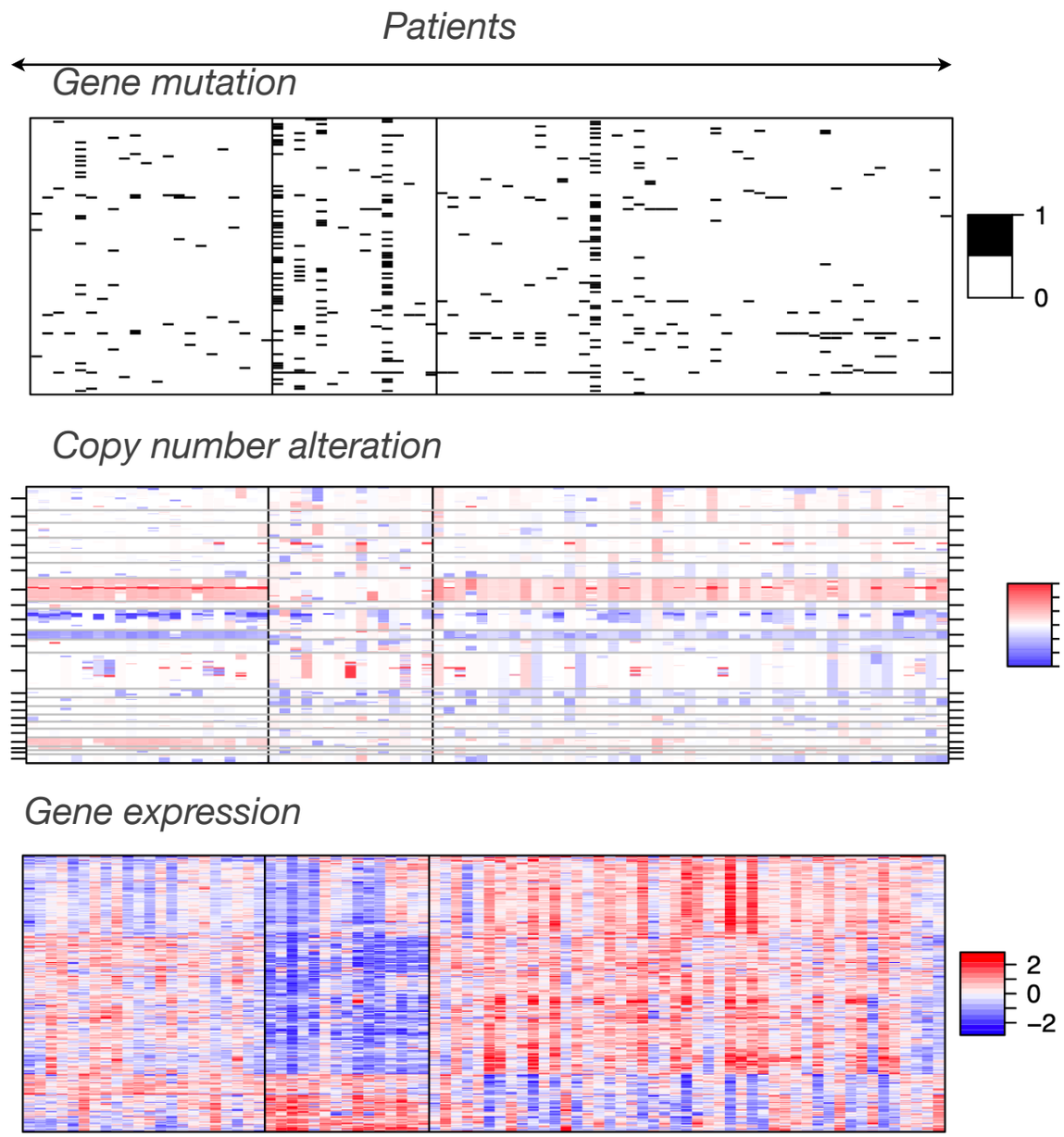


- Different types of data (binary, count data, continuous data,…) can be taken into account using different conditional probabilities
  - X_i is binary: **logistic** regression $\log \frac{P(X_i \mid Z)}{1 - P(X_i \mid Z)} = \alpha_i + \beta_i Z$
  - X_i is count data: **Poisson** regression $\log \left(\lambda(X_i \mid Z)\right) = \alpha_i + \beta_i Z$
  - X_i is continuous: **linear** regression $X_i = \alpha_i + \beta_i Z + \epsilon_i$

iCluster



- 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)

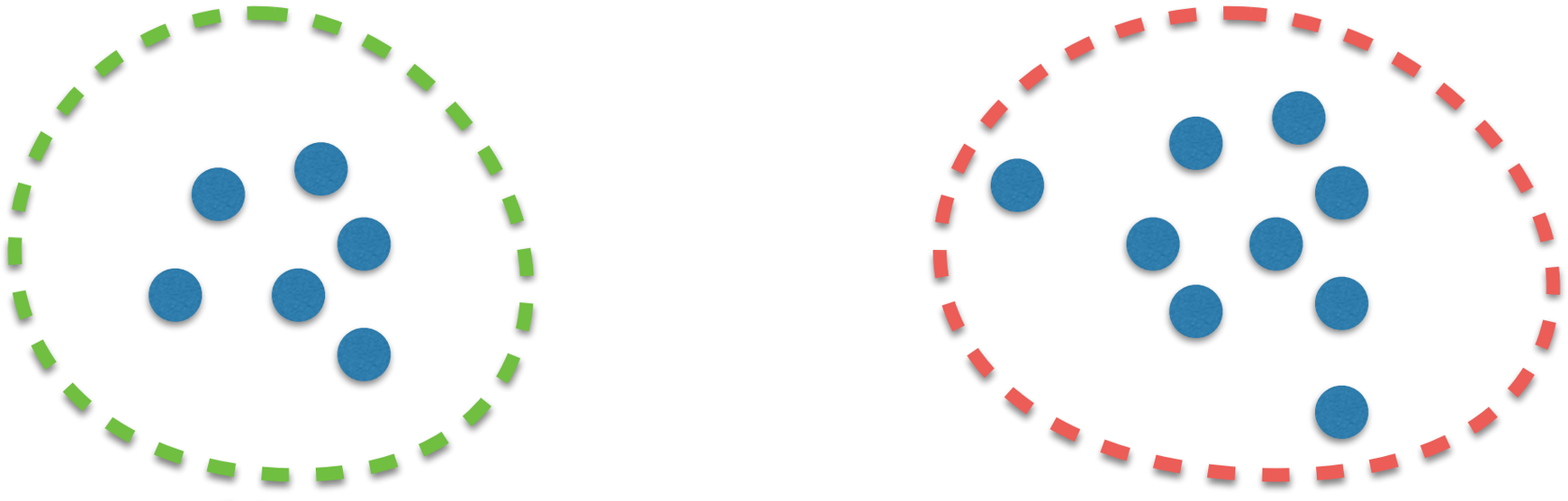


[Olshen et al., PNAS 2013]

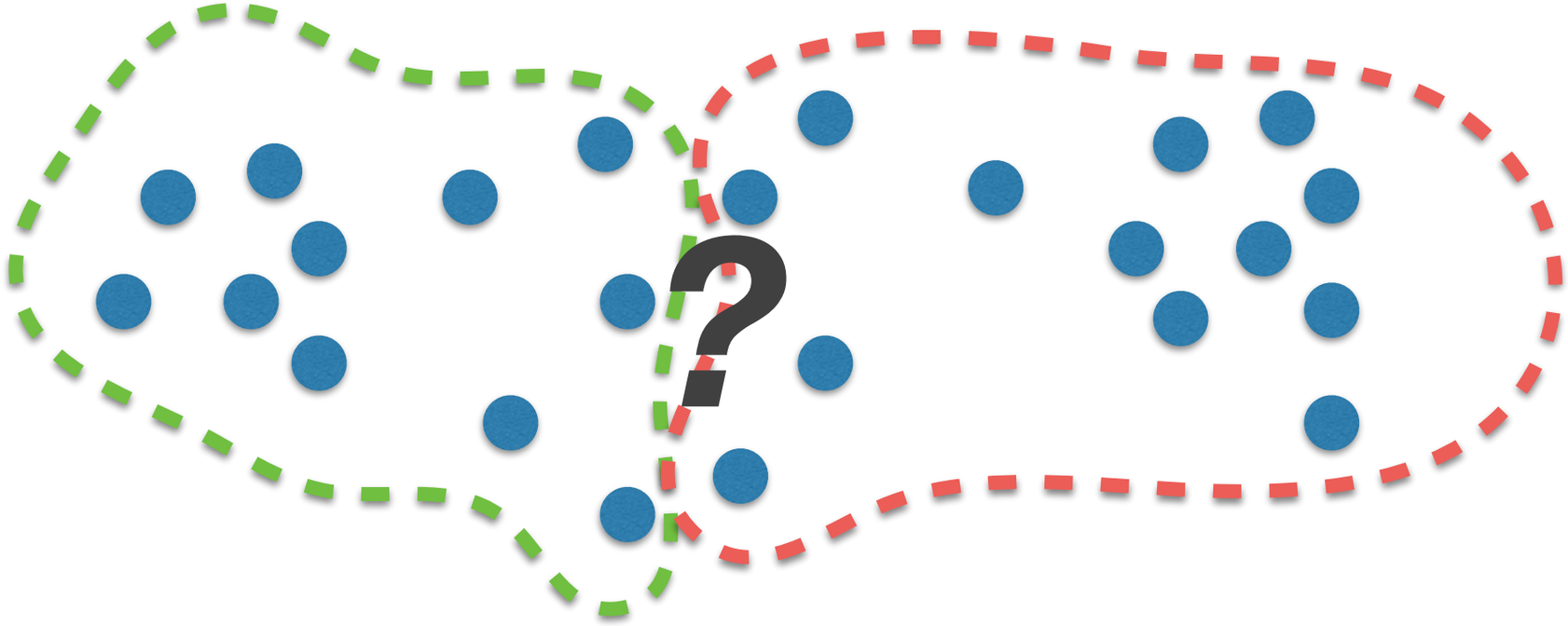
Limitations of Clustering

Medizinische Fakultät Heidelberg

Ideal



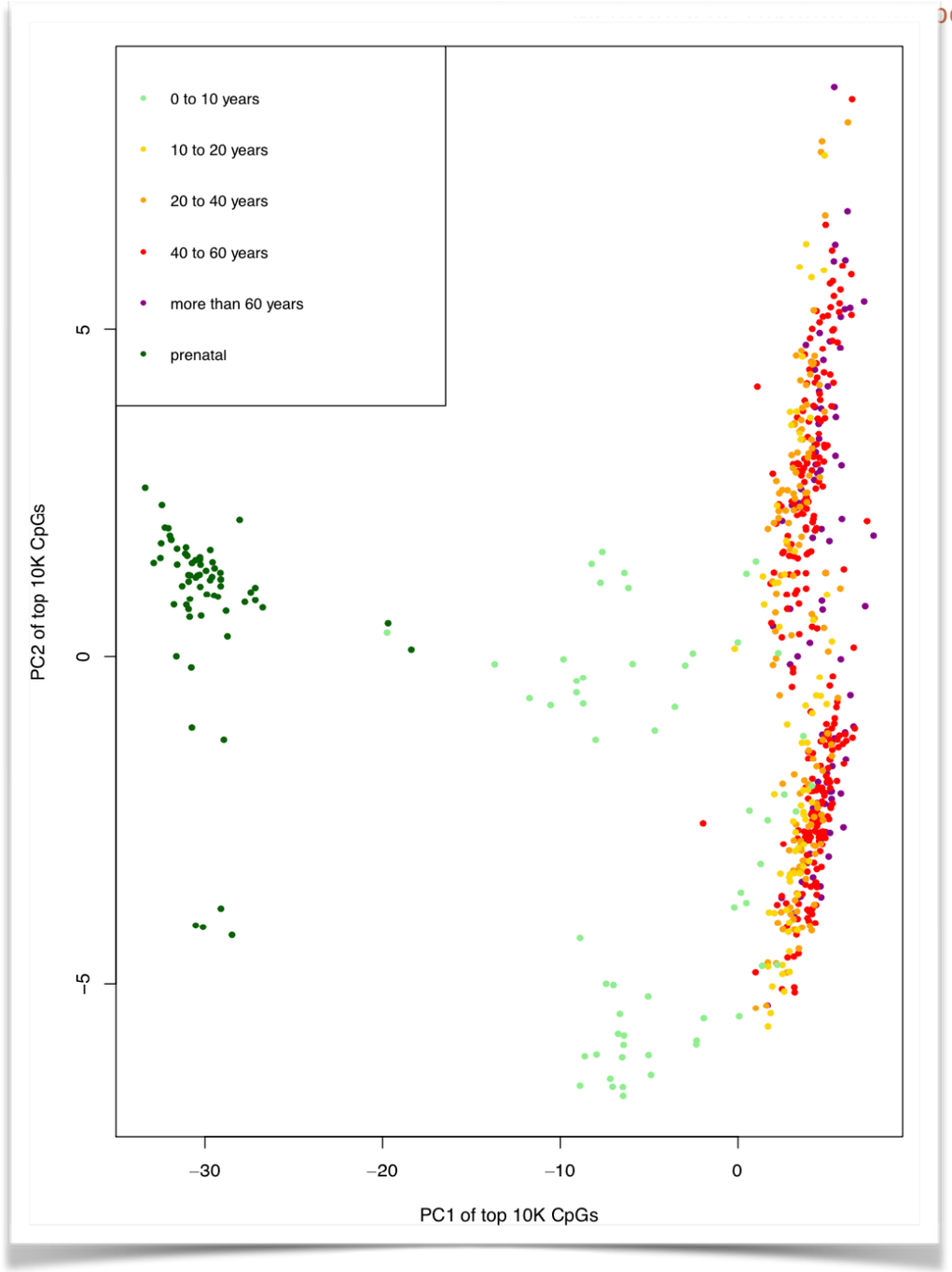
Real World



We needs 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

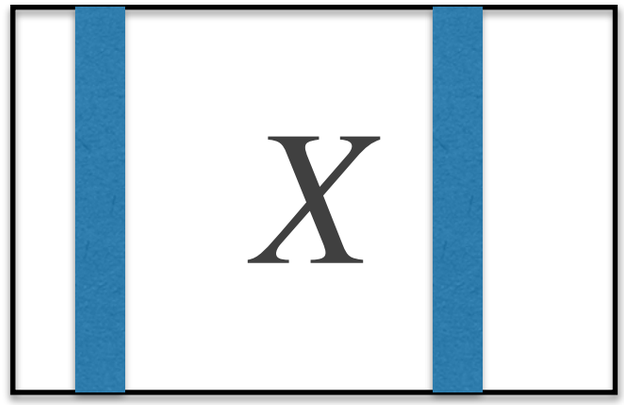
Medizinische Fakultät Heidelberg

- 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

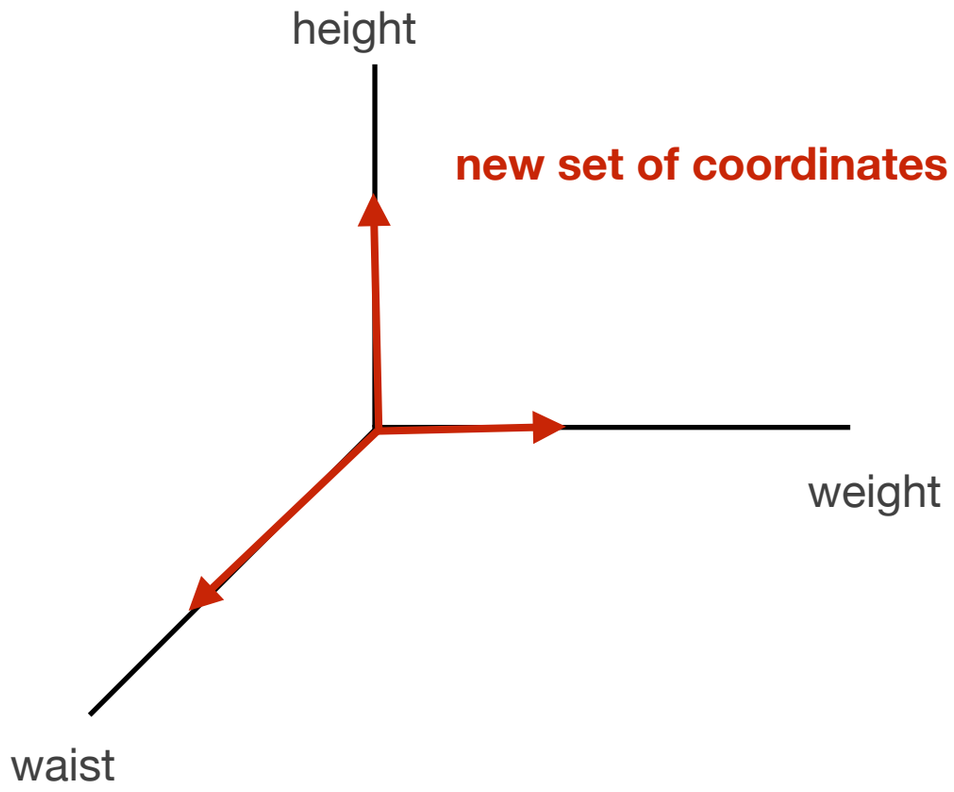
$$cov(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}X'_c \cdot X_c$$



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

Z-transformation

Defining new variables



- 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

- For some square matrices *X*, we can find vectors *v* such that

$$Xv = \lambda v$$

eigenvectors

eigenvalues

- Property: **symmetric matrices** (such as the correlation matrix) are **always diagonalizable**
  → we can find *n* eigenvalues and *n* eigenvectors

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 x 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
```

$$S = W' \cdot A \cdot W$$

3. Compute

S is a diagonal matrix with the eigenvalues as entries!

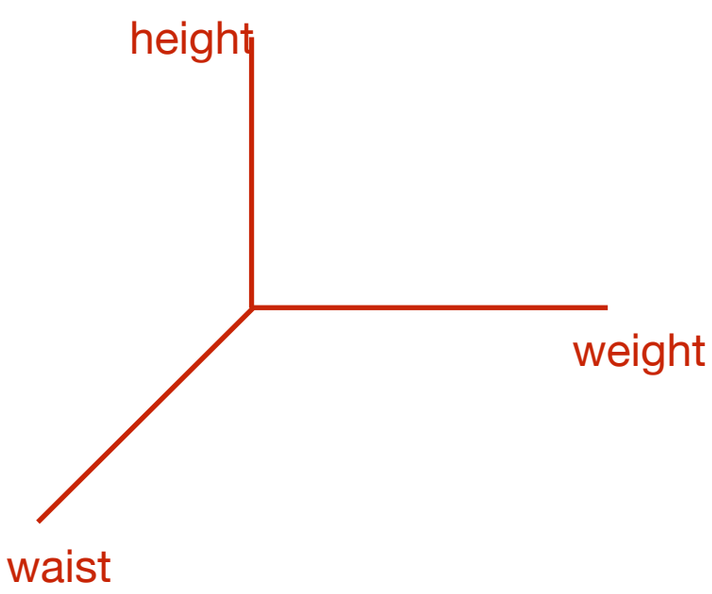
```
     [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 1.92 0.000 0.000 0.000 0.000
[2,] 0.00 1.308 0.000 0.000 0.000
[3,] 0.00 0.000 0.901 0.000 0.000
[4,] 0.00 0.000 0.000 0.764 0.000
[5,] 0.00 0.000 0.000 0.000 0.107
```

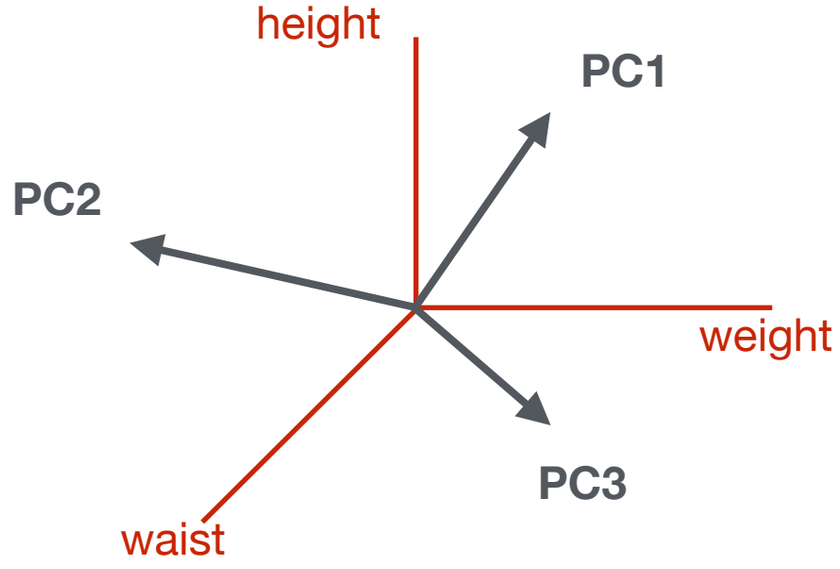
New coordinate system

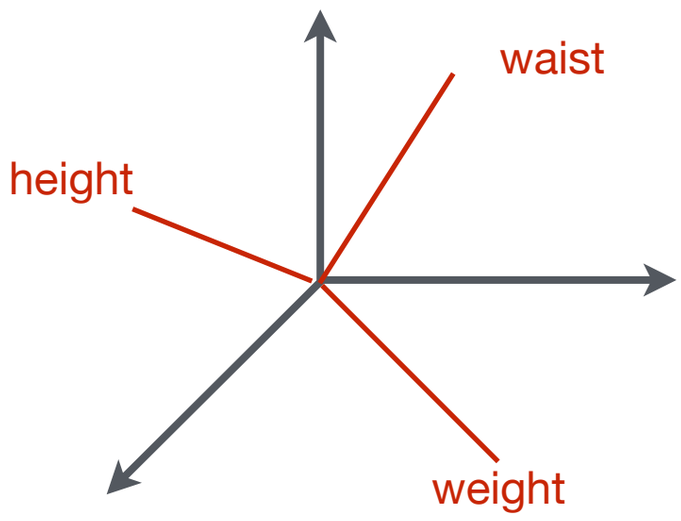
$$S = W' \cdot A \cdot W = W' \cdot \left(\frac{1}{N} X'_{cs} \cdot X_{cs} \right) \cdot W$$

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

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







PCA as a matrix factorization problem

Medizinische Fakultät Heidelberg

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

$$X = U\Lambda V' = WH$$

$$W = U$$
$$H = \Lambda V'$$

U : unitary mxm matrix $UU' = I$

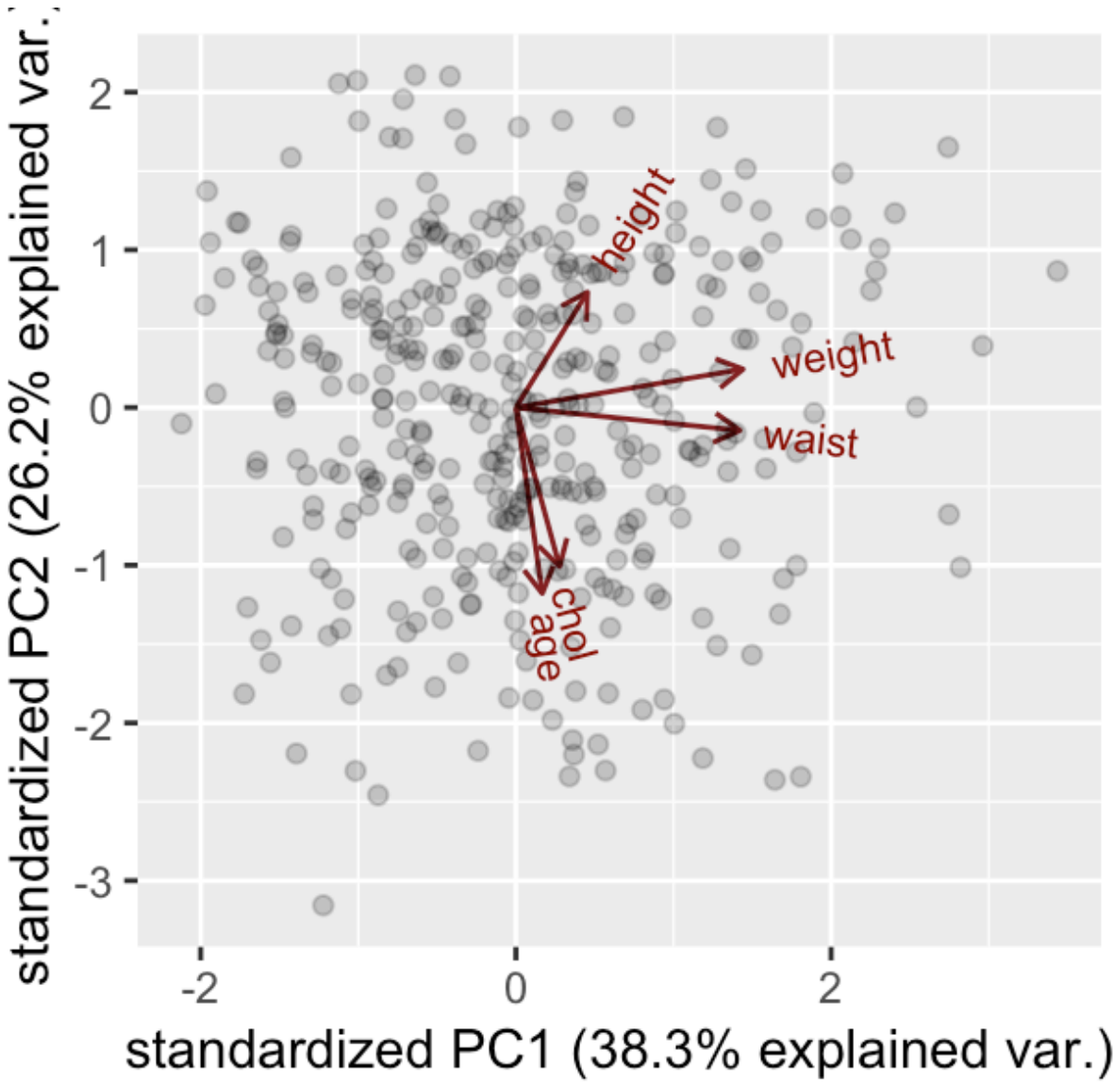
V : unitary nxn matrix $VV' = I$

Λ : rectangular diagonal mxn matrix

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

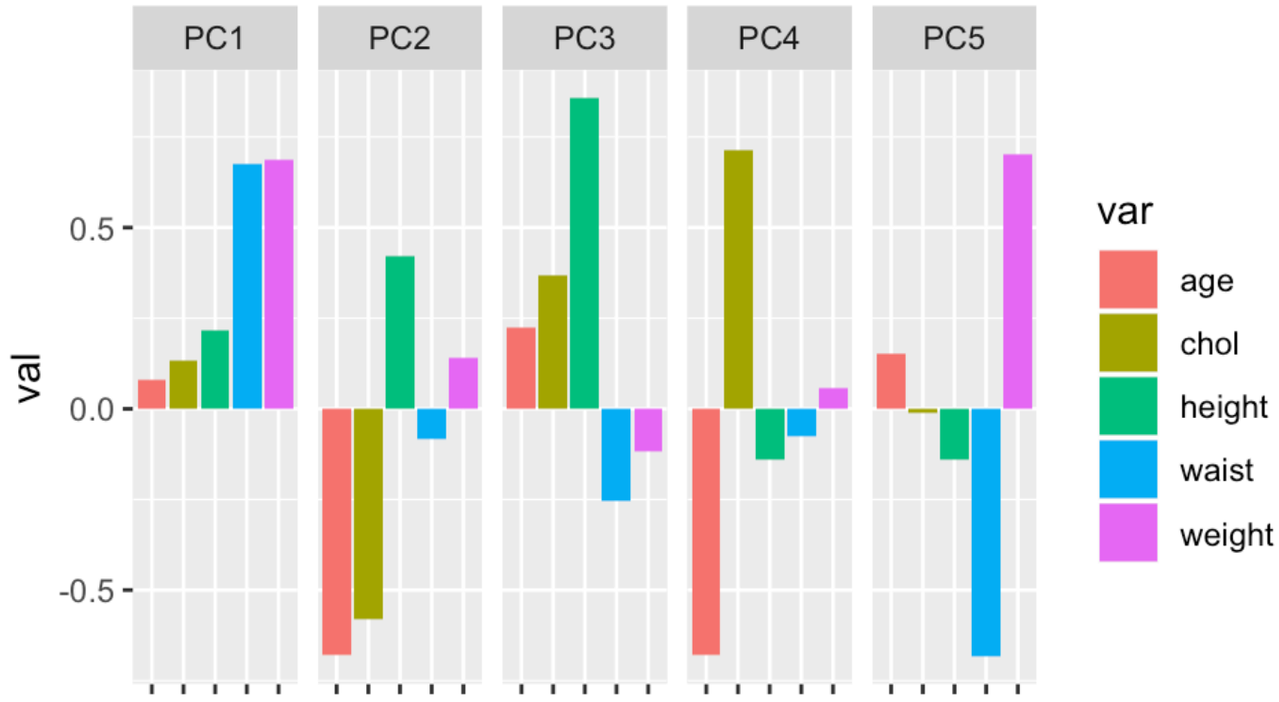
PCA biplot

- each **dot** is a sample / patient
- new coordinate system is (PC1,PC2)
- **Red arrows** indicate the contribution of each “old" coordinate to the PCs



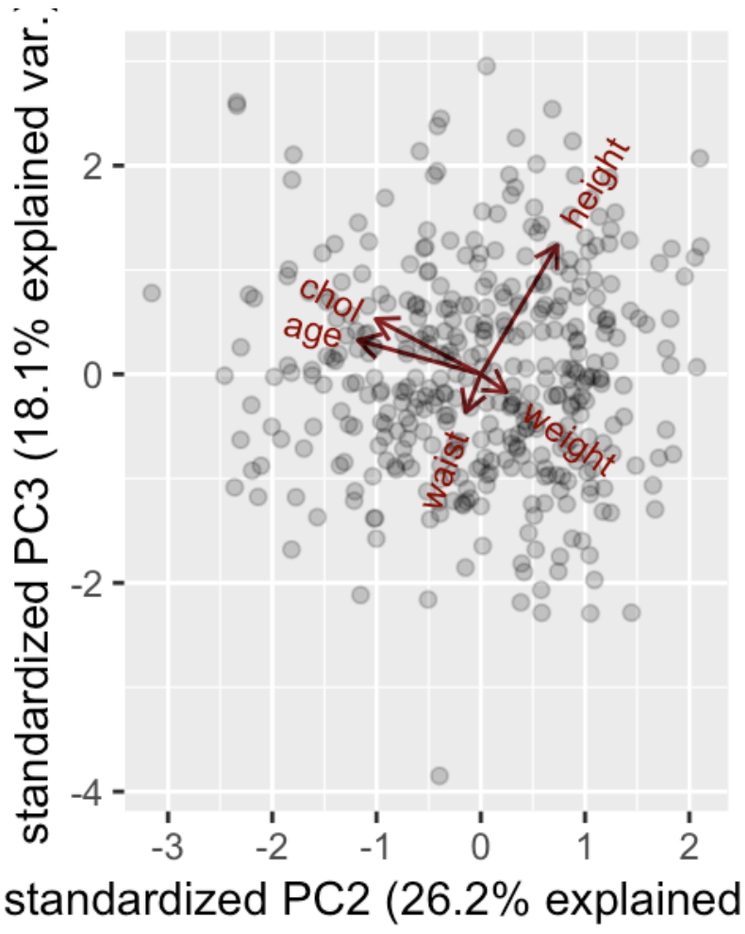
Principal components

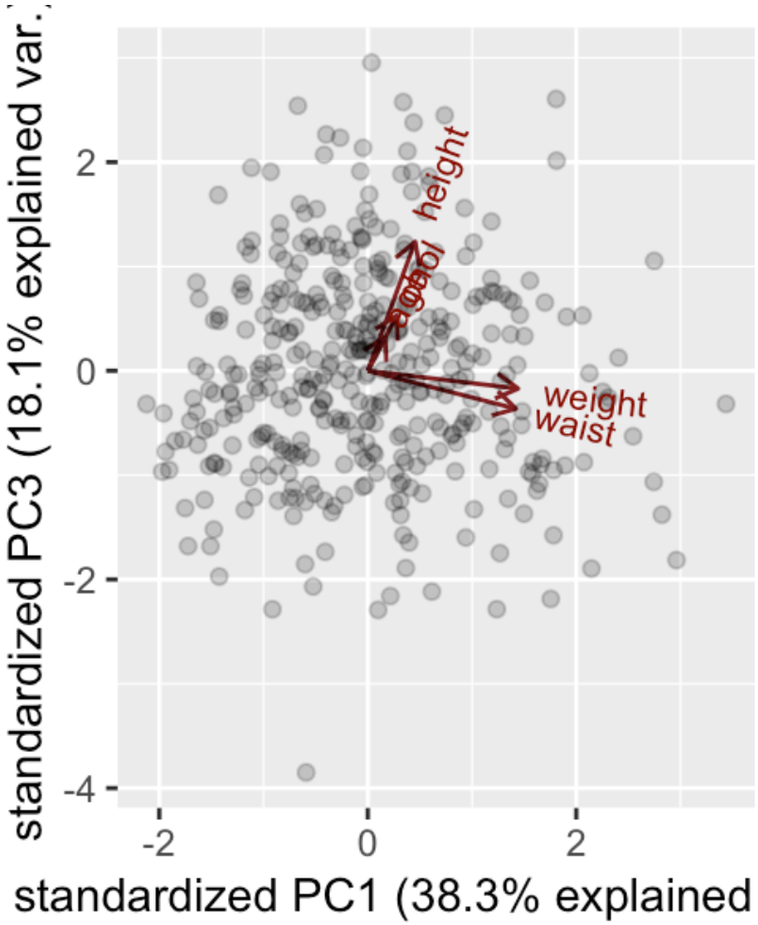
$$PC_i = \alpha_i \cdot \text{age} + \beta_i \cdot \text{chol} + \gamma_i \cdot \text{height} + \delta_i \cdot \text{waist} + \epsilon_i \cdot \text{weight}$$



- 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

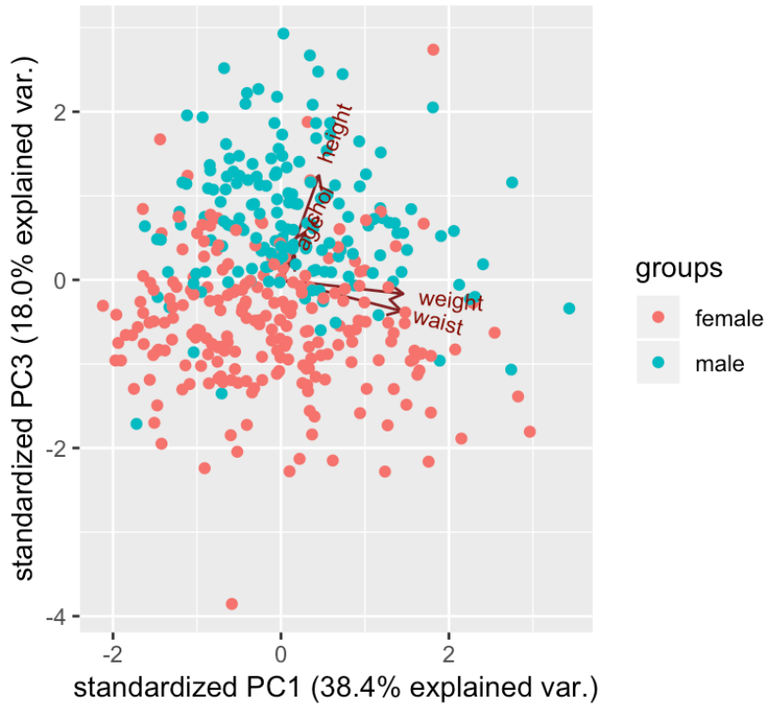




Identifying interesting PCs

Medizinische Fakultät Heidelberg

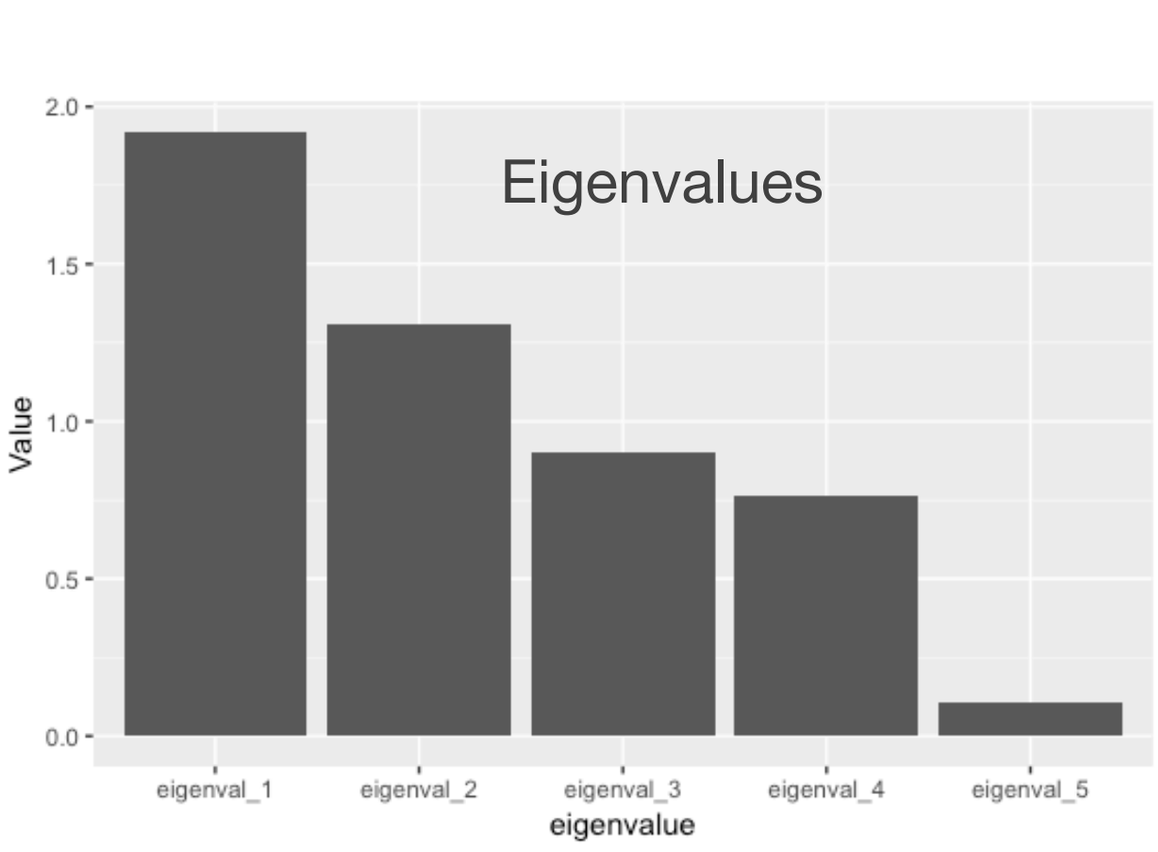
- PC plots can highlight a new group structure
- Example: **PC3** seems very associated to gender
- indicates that a combination of height and cholesterol does separate men /women

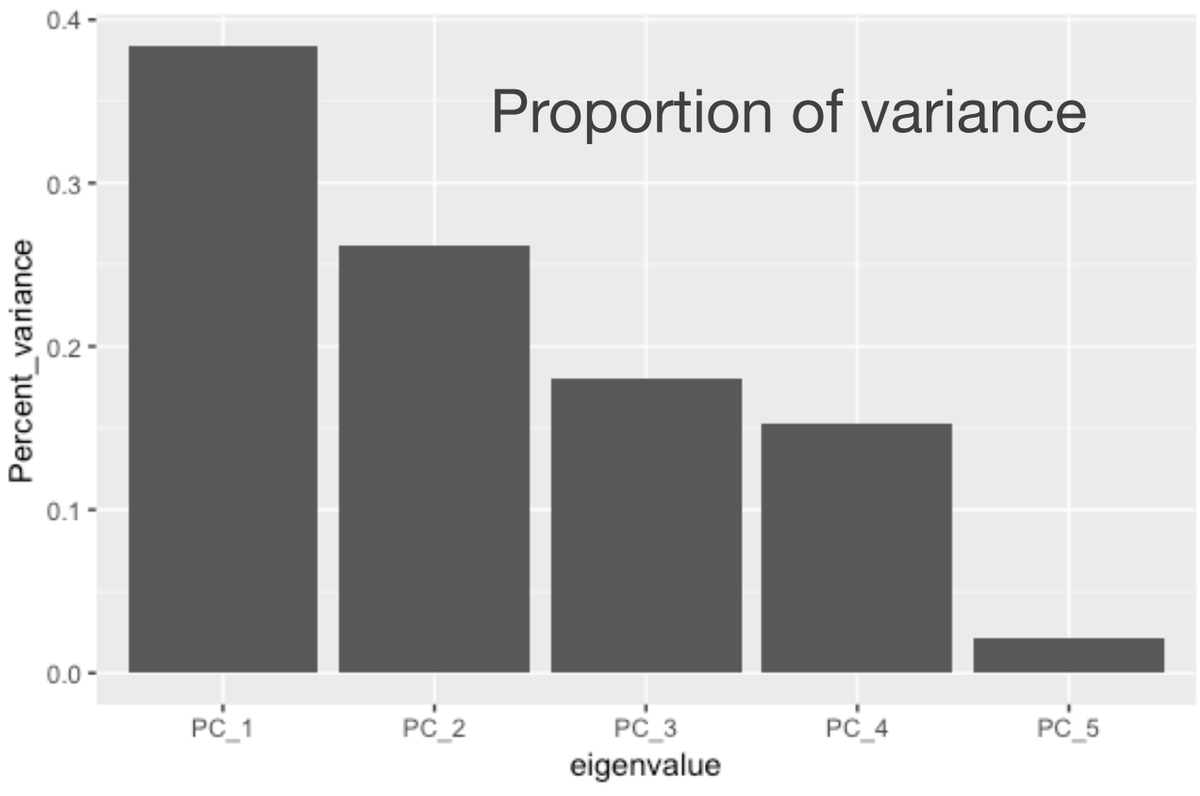


Number of PCs?

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 vairance

Choosing the number of PCs

Medizinische Fakultät Heidelberg

- several criteria to select the optimal subset of PCs, without loosing too much information

- **Proportion of variance:**
  keep PCs such that the cumulative variance is above threshold

$$\sum_{i=1}^{k} \frac{\lambda_i}{\sum \lambda_i} \geq \text{var}_{min}$$

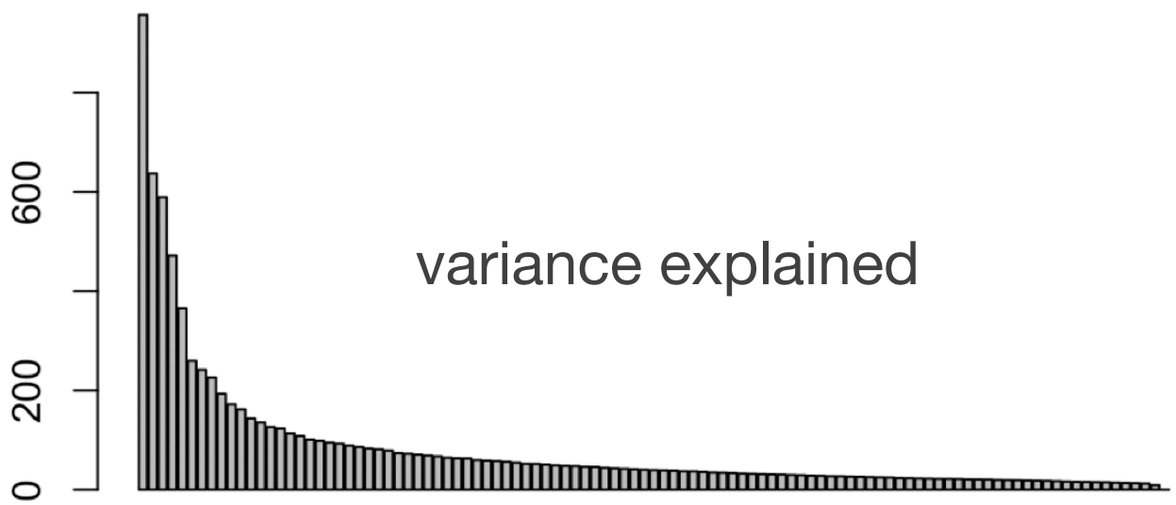
- **Average eigenvalue criteria:**
  keep PCs which have eigenvalue larger than
  - mean eigenvalue (Kaiser rule) or
  - 70% of mean eigenvalue (Jottclife rule)

$$\lambda_i \geq \bar{\lambda}$$

Application to gene expression

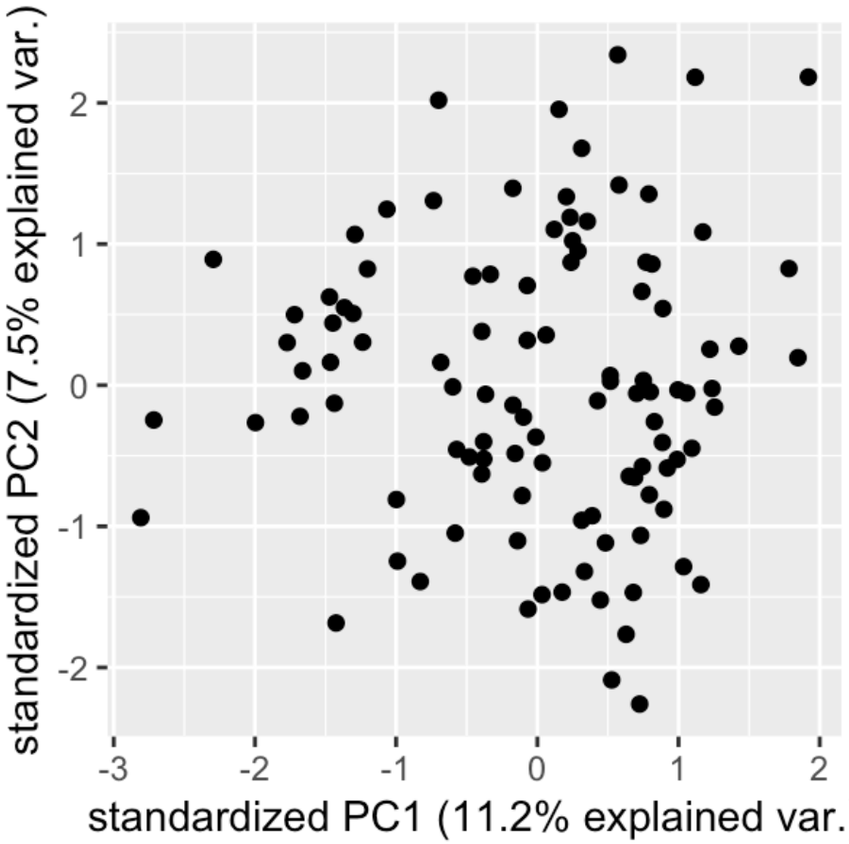
Medizinische Fakultät Heidelberg

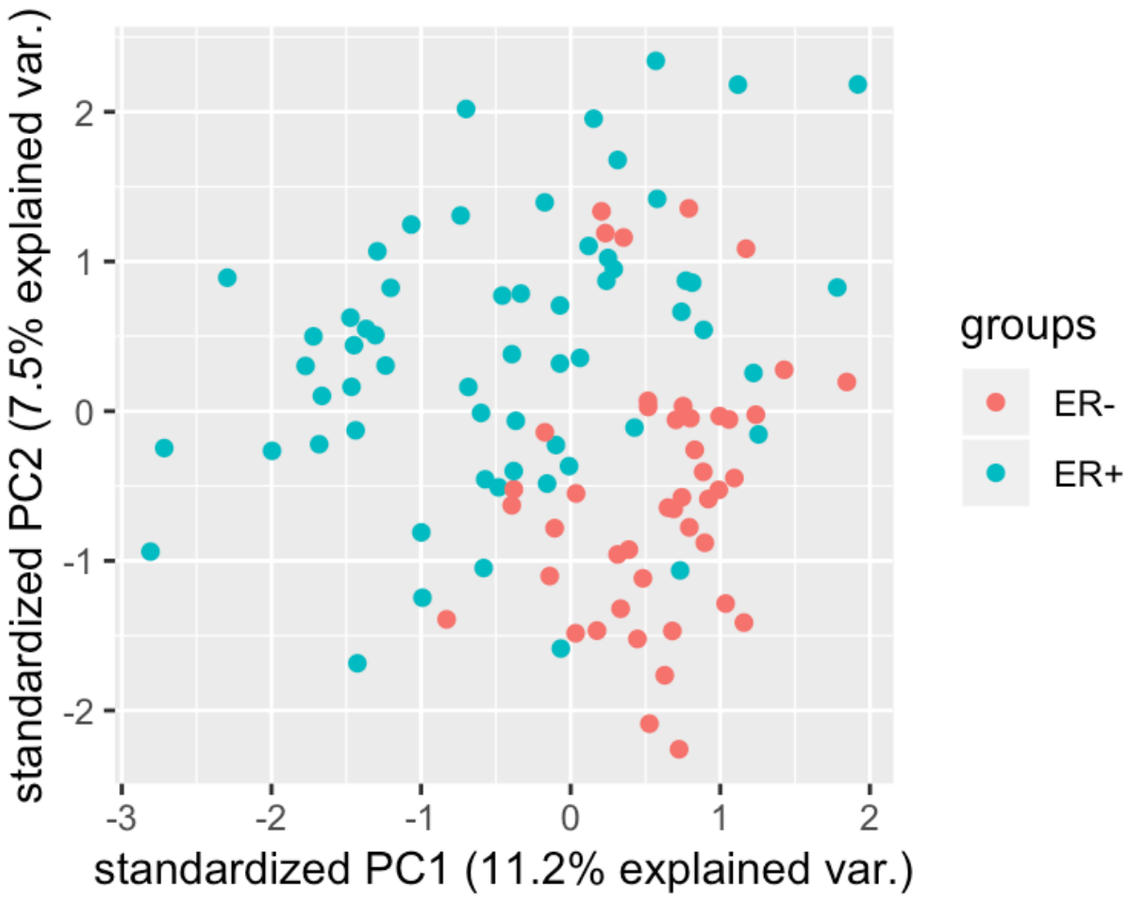
- Gene expression dataset of **breast cancer patients**
- 2 groups: ER+ and ER- patients [M. Ringner, Nature Biotech. (2008)]
- Dimension: k = 105 patients / n = 8534 genes (here: n >> k)
- pre-processing:
  - **scale** the gene expression across patients
  - **center** the gene expression across patients
- 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



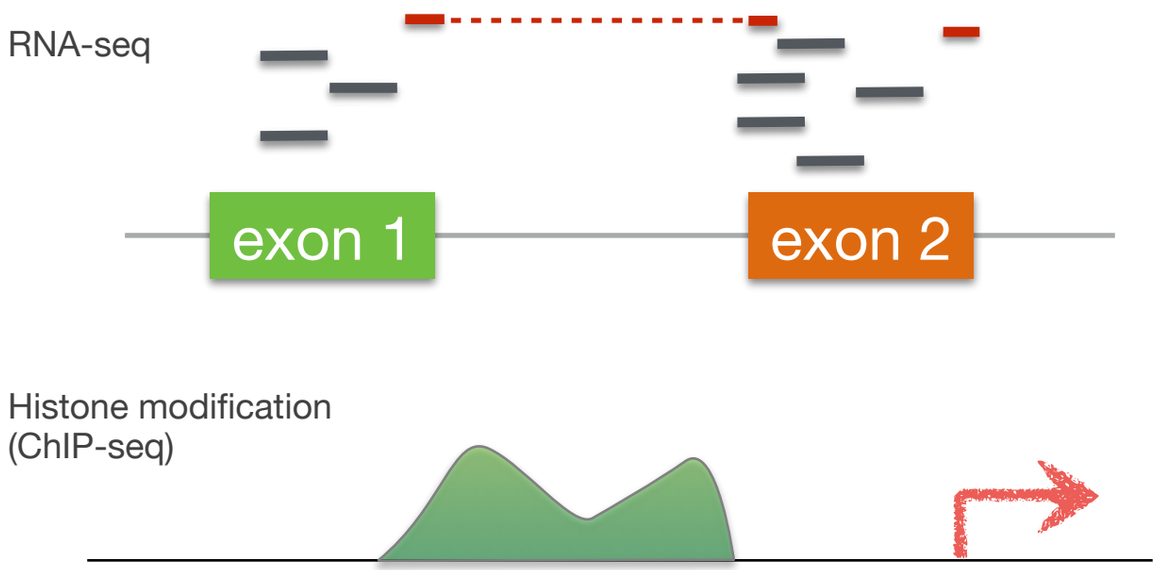


- PC1 separates ER+ from ER- patients

Non-negative matrix factorization

Medizinische Fakultät Heidelberg

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



we can apply parts-base decomposition of the data

Non-negative matrix factorization

Medizinische Fakultät Heidelberg

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

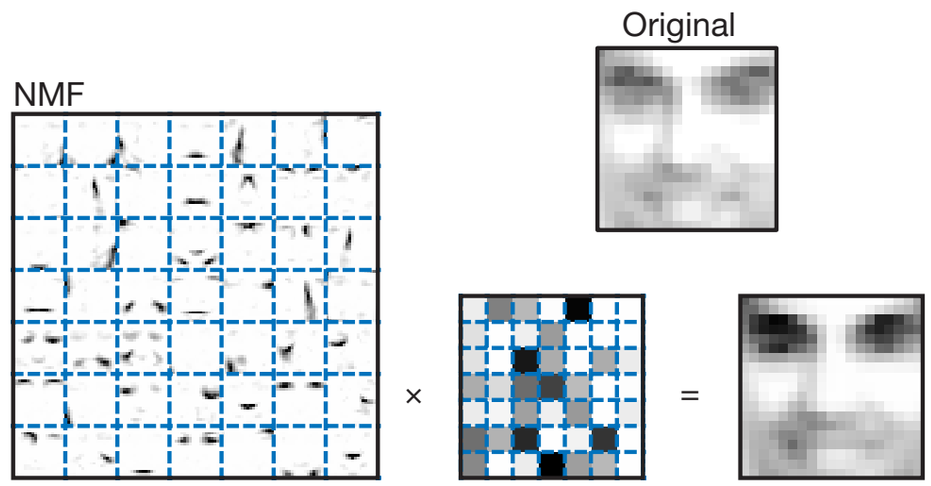
$X : N \times M$ matrix

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

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

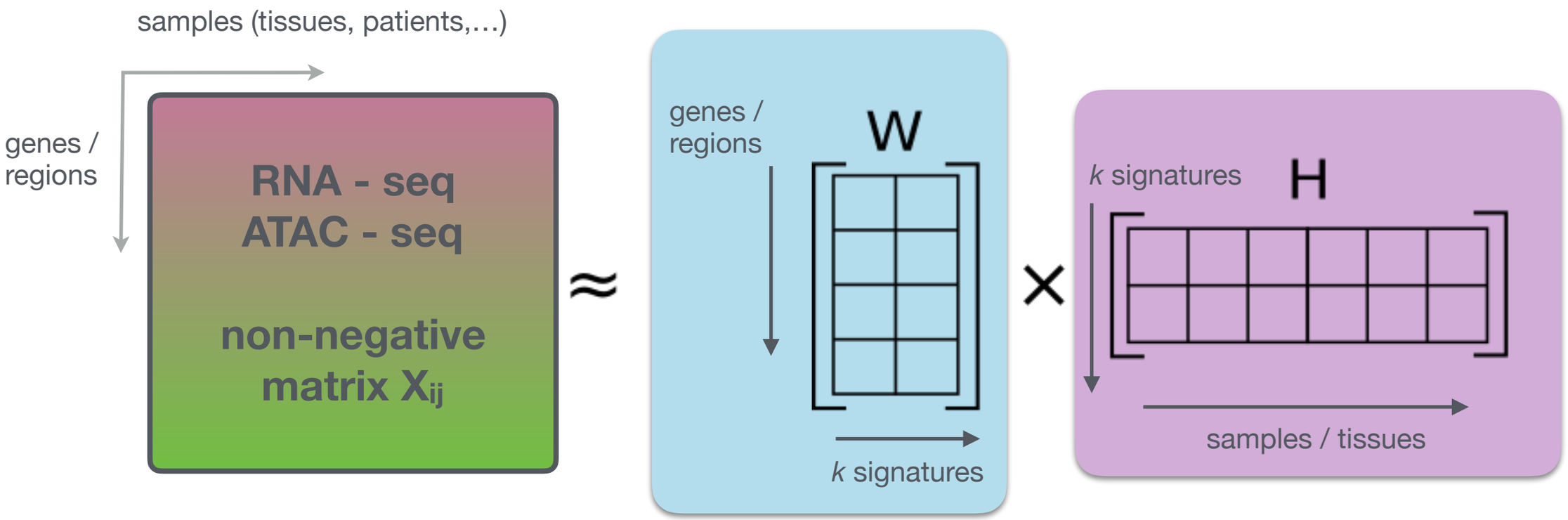
NMF in essence similar to PCA,
but non-negativity implies

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



[Lee, Seung 1999]

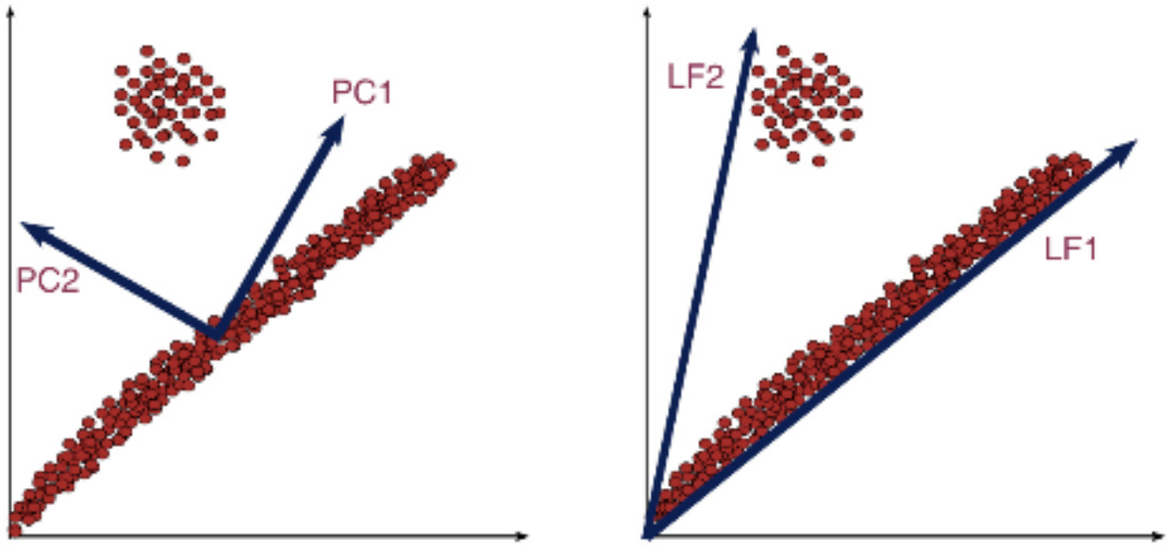
NMF



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



PCA NMF [Casalino, Buono, Mencar]

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

- 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

(a) (b) (c) (d) (e) (f)

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

(a) (b) (c) (d) (e) (f)

Part are more easily interpretable in NMF

[Nikolaus]

Finding the best NMF solution

Medizinische Fakultät Heidelberg

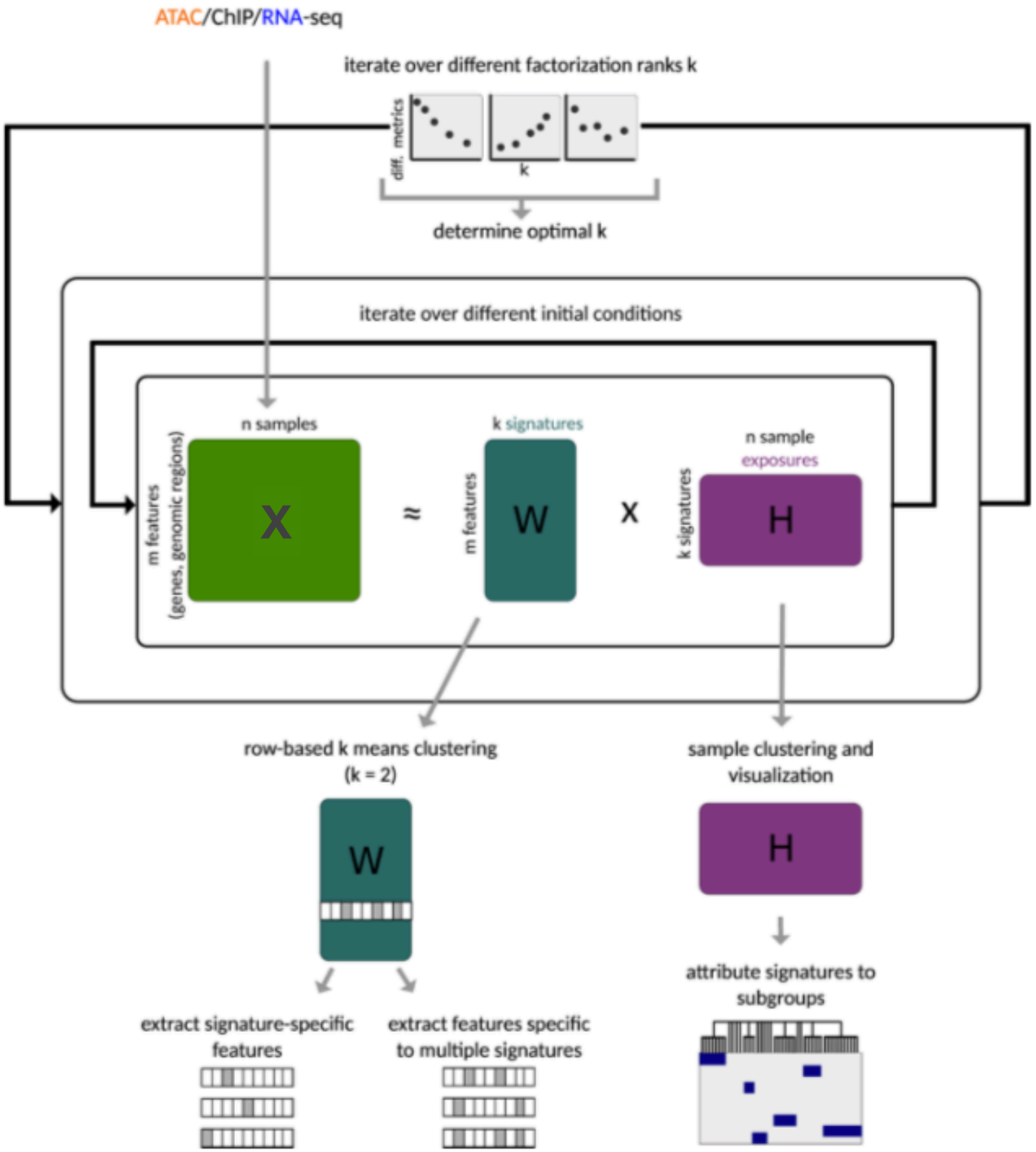
Goal: minimize $\|X - WH\|^2_{Froebenius} = \sum_i \sum_j |(x - wh)_{ij}|^2$

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

$$W_{ia} \longleftarrow W_{ia} \sum_\mu \frac{X_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \qquad H_{a\mu} \longleftarrow H_{a\mu} \sum_i W_{ia} \frac{X_{i\mu}}{(WH)_{i\mu}}$$

$$W_{ia} \longleftarrow \frac{W_{ia}}{\sum_j W_{ja}}$$

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?

Medizinische Fakultät Heidelberg

- Accuracy of matrix decomposition: how well does *WH* represent *V*?
  - **Froebenius error** should be small
  - **Amari distance** should be small

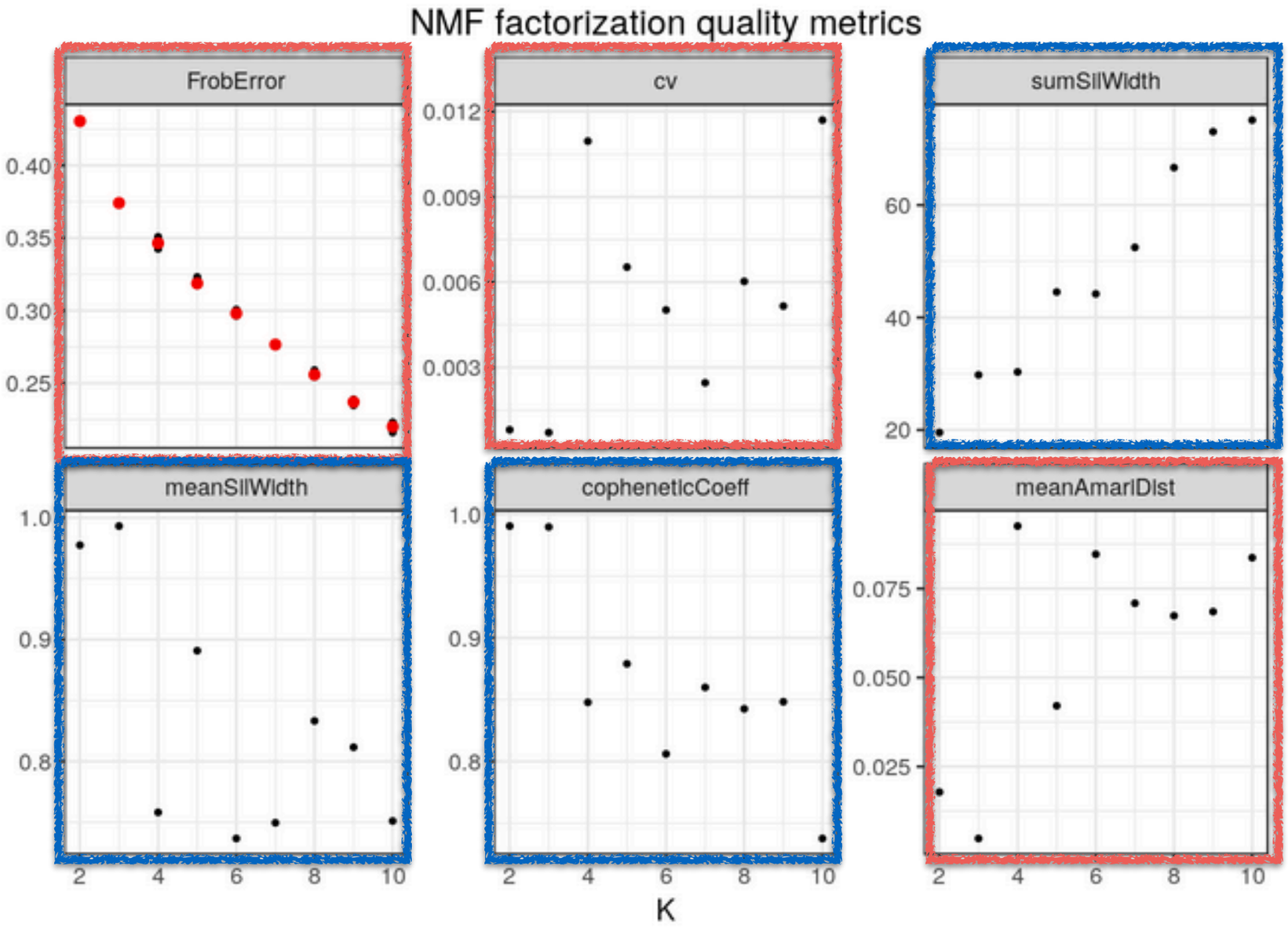
- Stability of solutions: how variable are the solutions using different random initializations?
  - **Coefficient of variation** should be small

- 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?

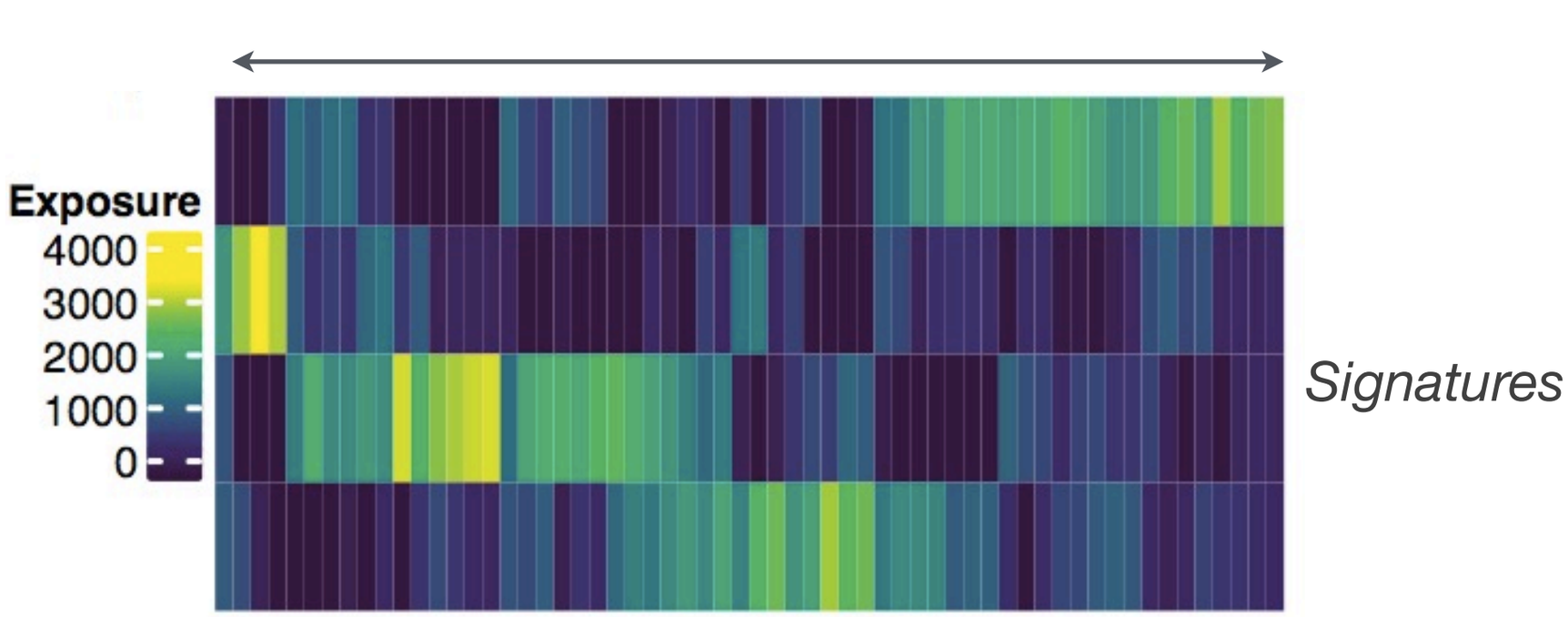
Medizinische Fakultät Heidelberg



k = 5 appears to be a good choice

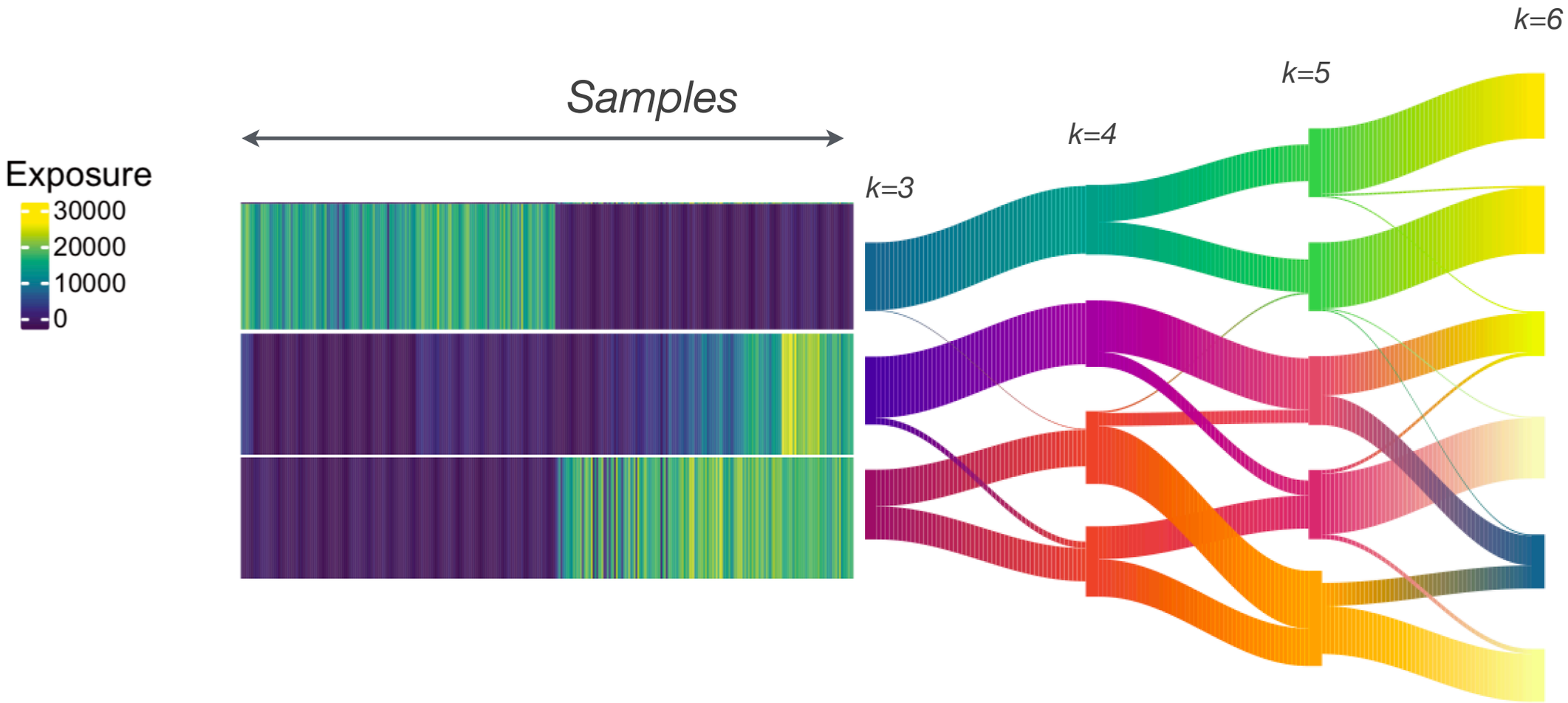
Exposure matrix H

- 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



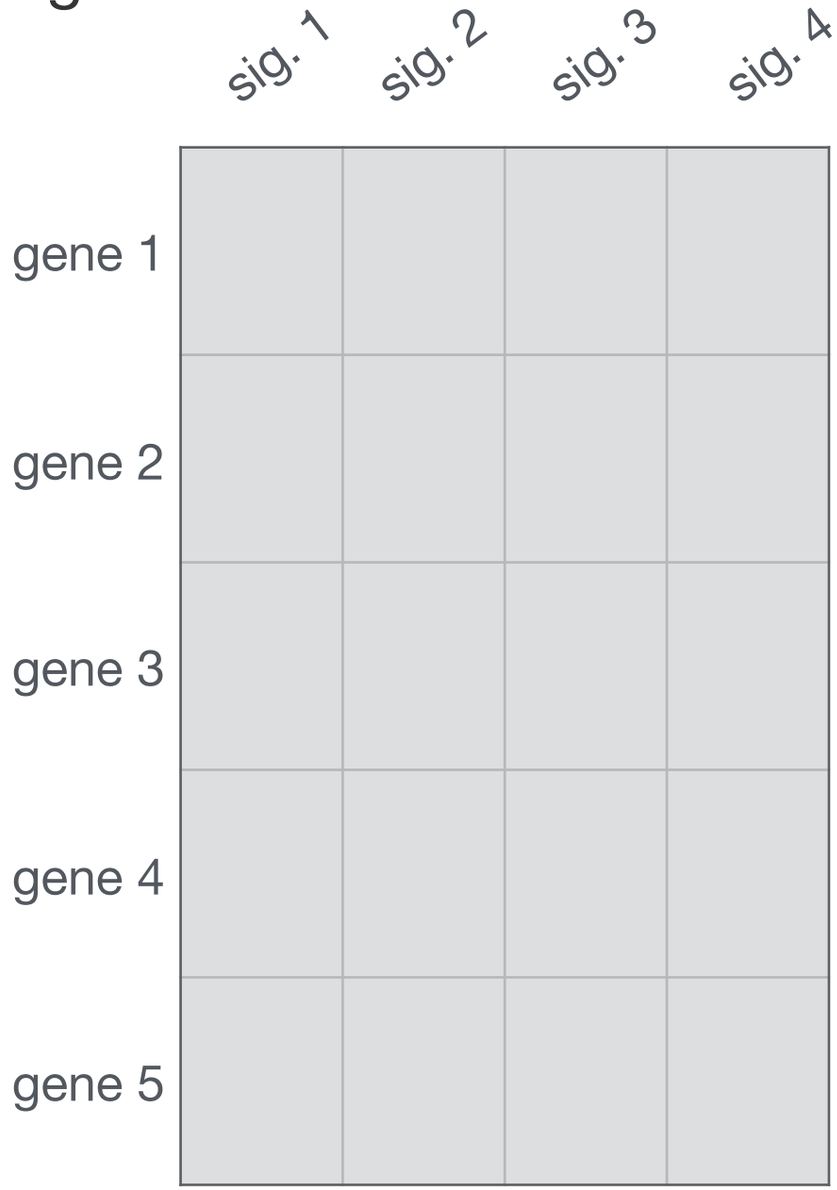
Stability of signatures

Medizinische Fakultät Heidelberg



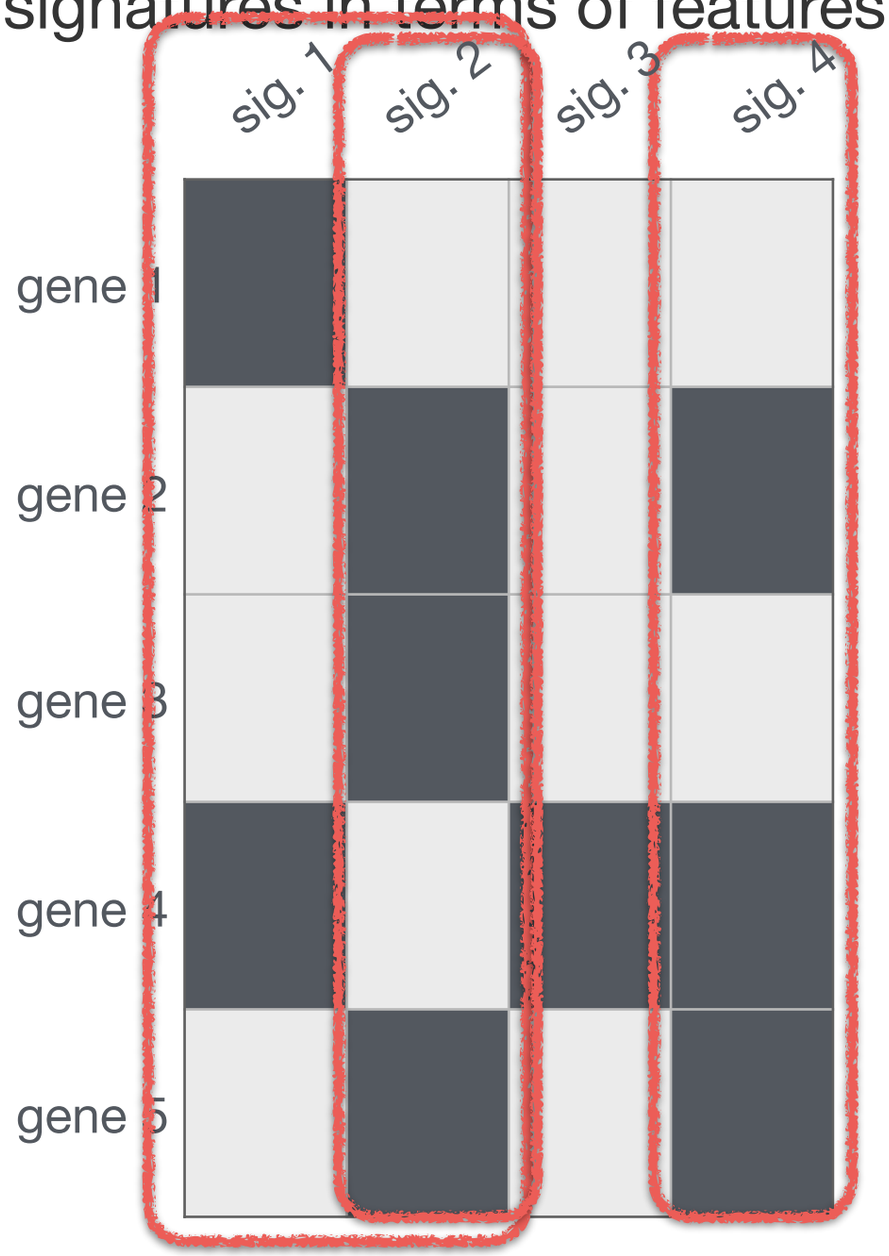
Signature matrix W

- 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



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

RNA-seq

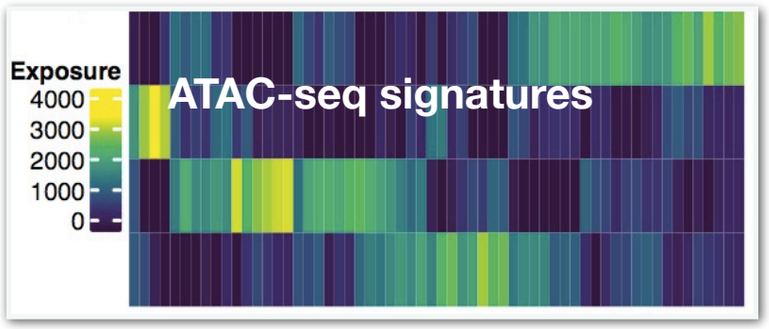
W_{RNA}

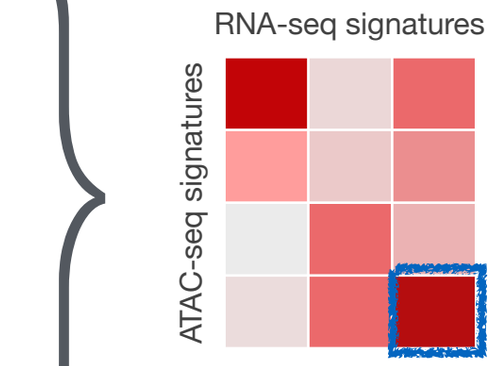
x

ATAC-seq

W_{ATAC}

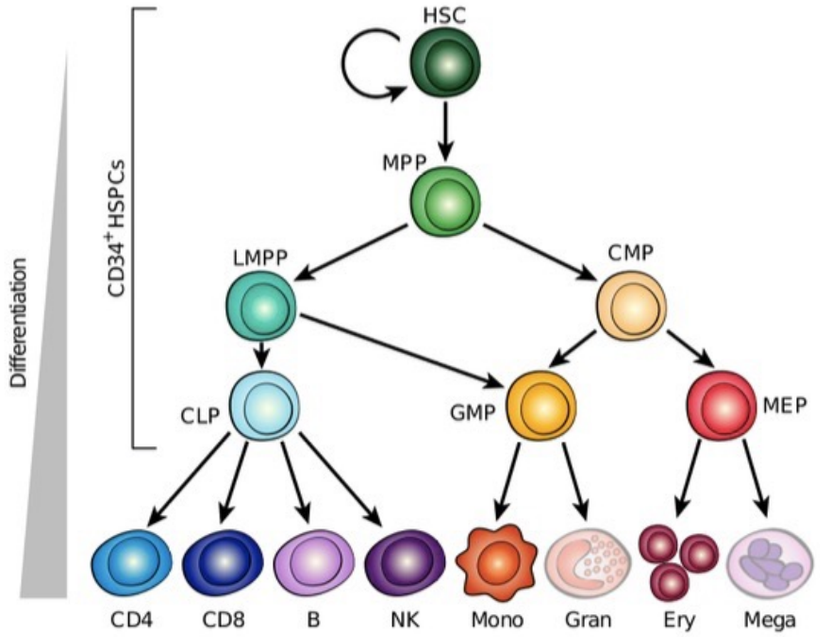
x





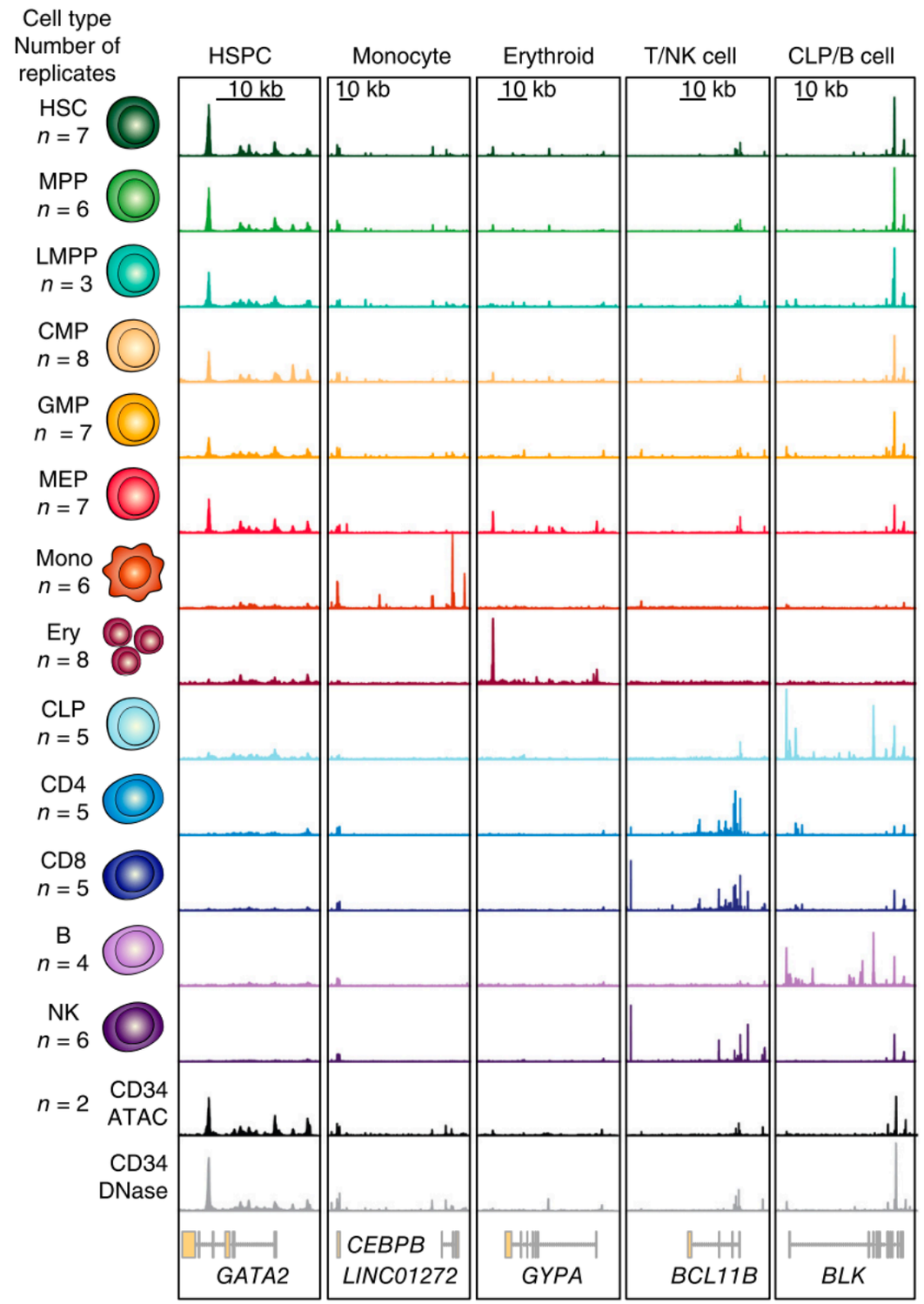
associate genes and regions

Example of use case



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

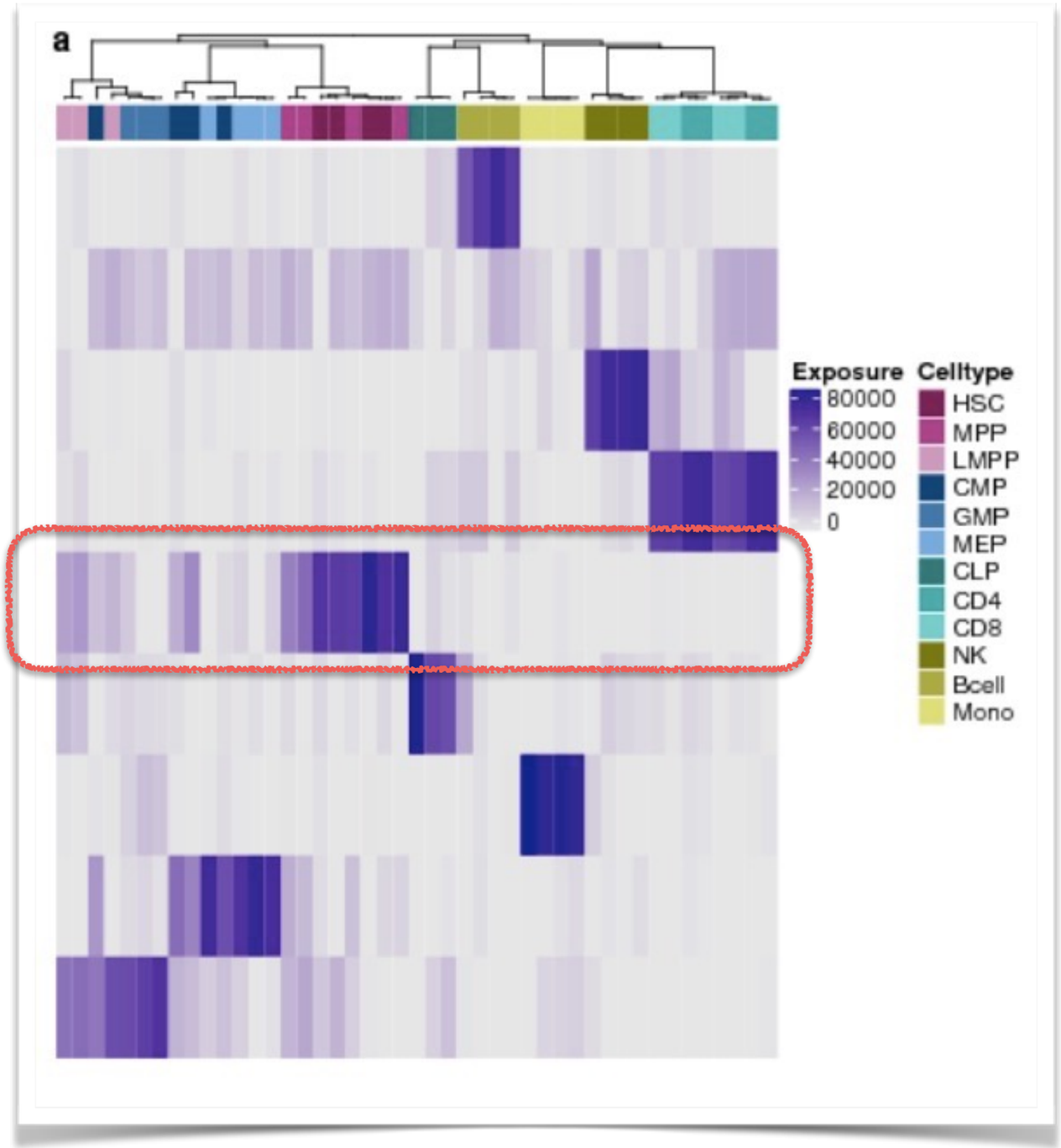
[Corces et al. Nat. Gen (2016)]

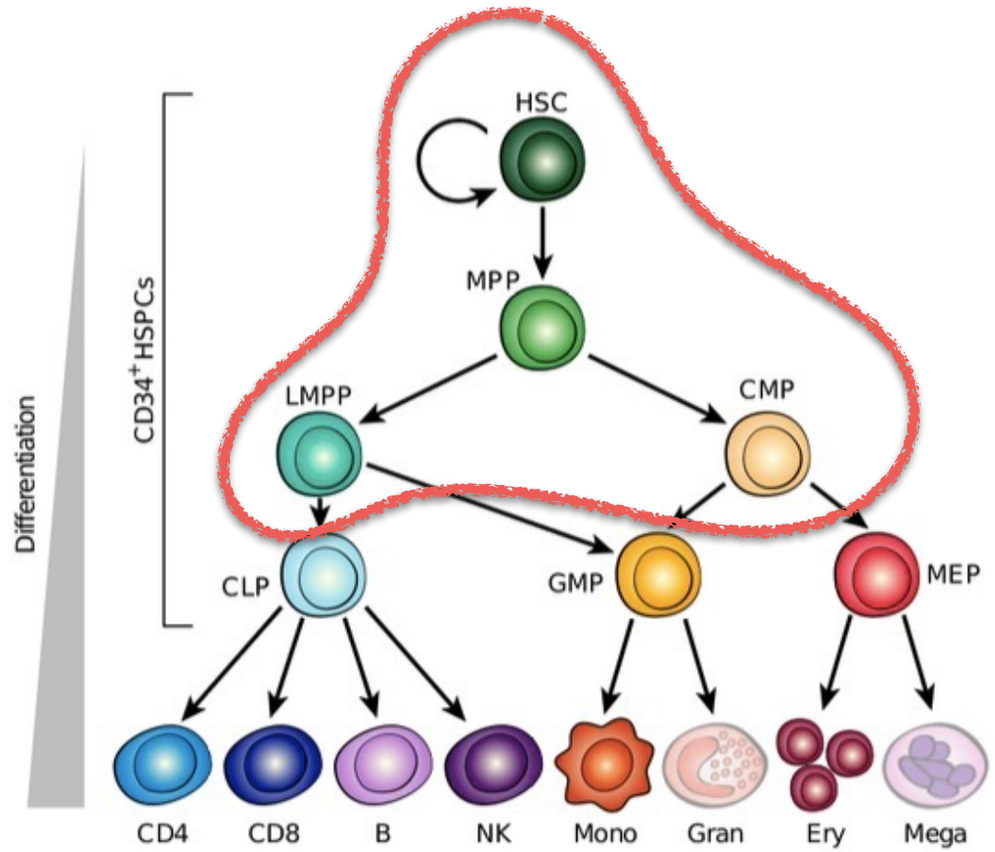


Interpreting signatures

Medizinische Fakultät Heidelberg

RNA-seq





Stemness-signature fades away, as differentiation progresses

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

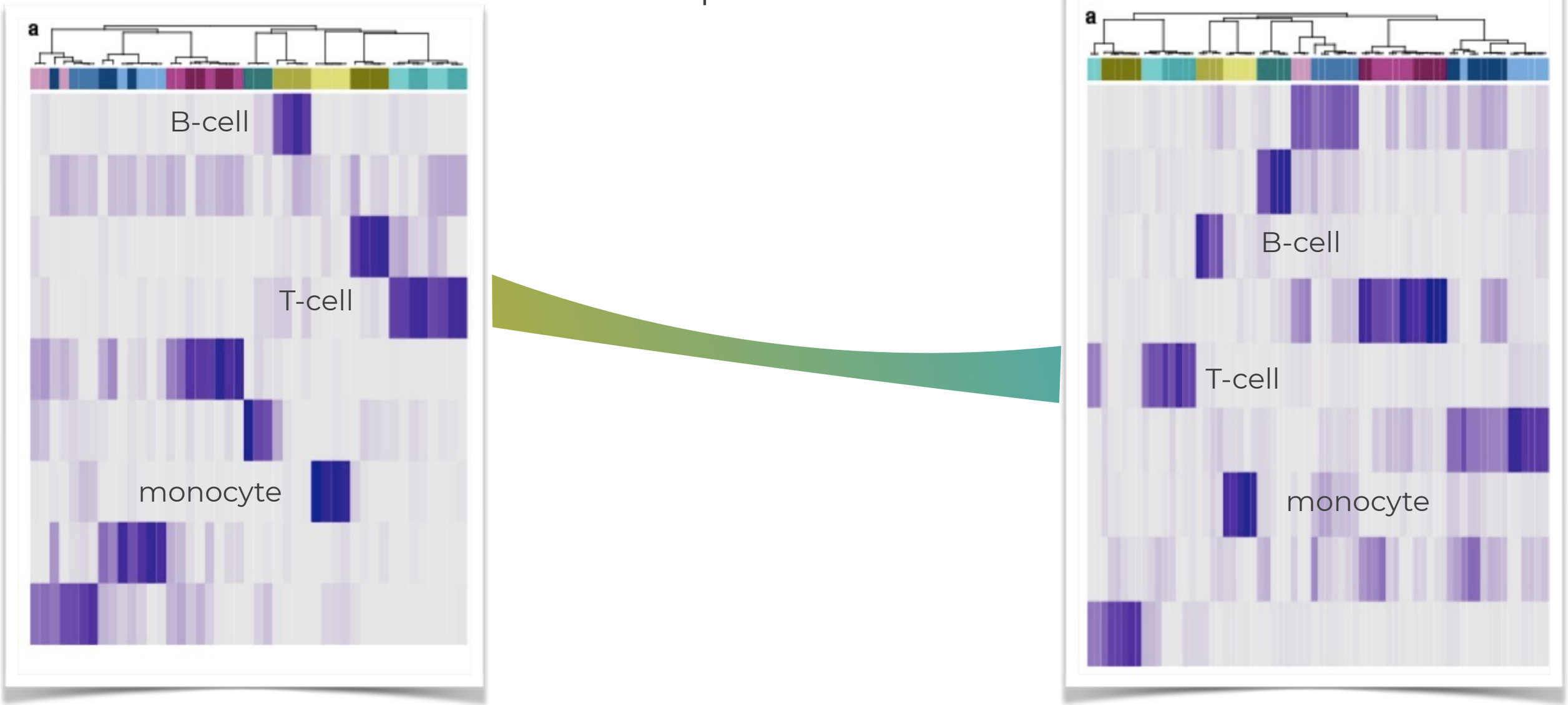
Associating signatures

Medizinische Fakultät Heidelberg

RNA-seq

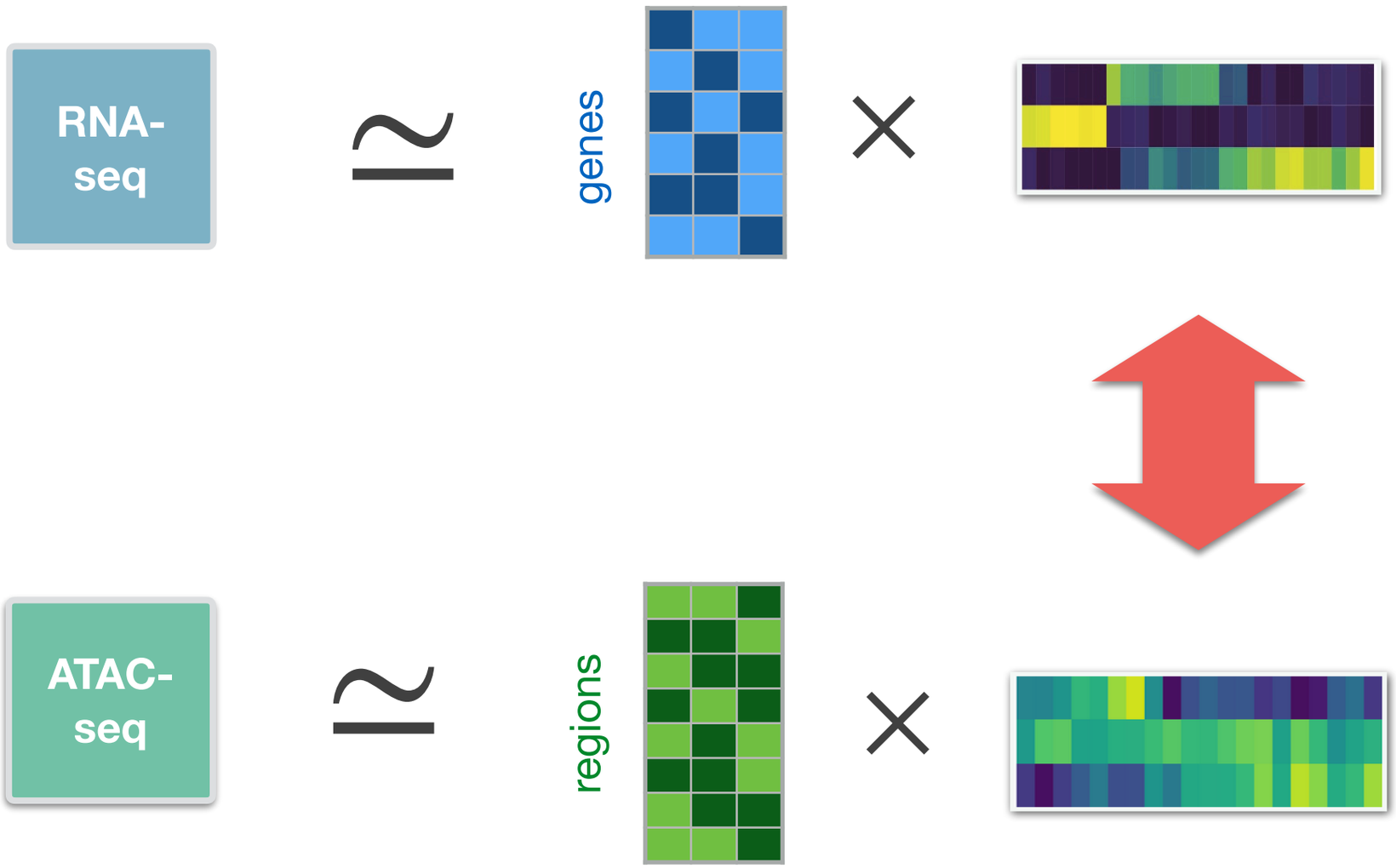
cell-type specific
capture Hi-C

ATAC-seq



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

Integrating multiple datasets using NMF



Integrating multiple datasets using NMF

Medizinische Fakultät Heidelberg

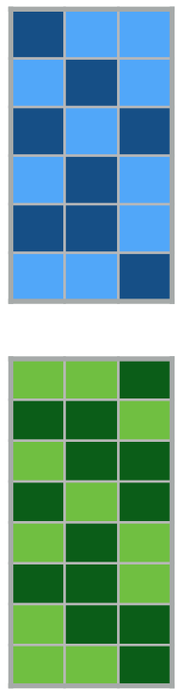
RNA-seq

ATAC-seq

$\simeq$

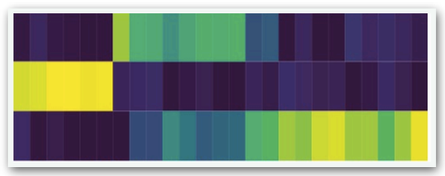
genes

regions



$\times$

shared *H* matrix



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

Joint-NMF [Chalise, Fridley (2017)]

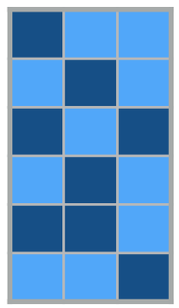
RNA-seq

ATAC-seq

$\simeq$

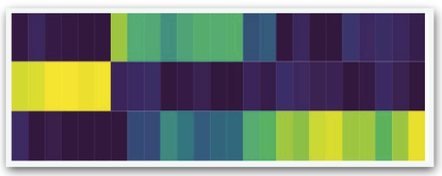
genes

regions



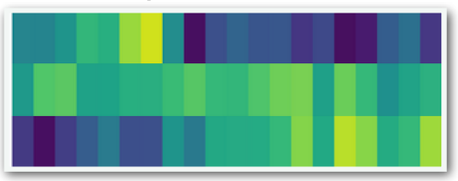
$\times$

common *H* matrix



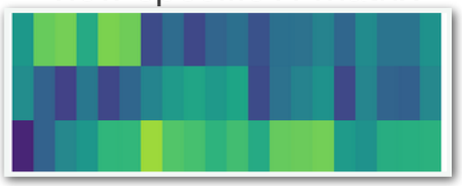
$+$

view specific *H* matrix



$+$

view specific *H* matrix



integrative-NMF [Yang, Michailidis (2015)]

integrative NMF

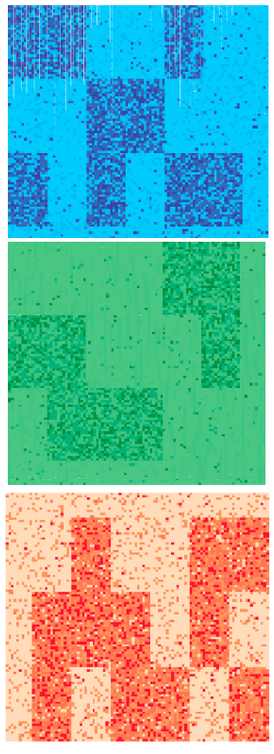
Medizinische Fakultät Heidelberg

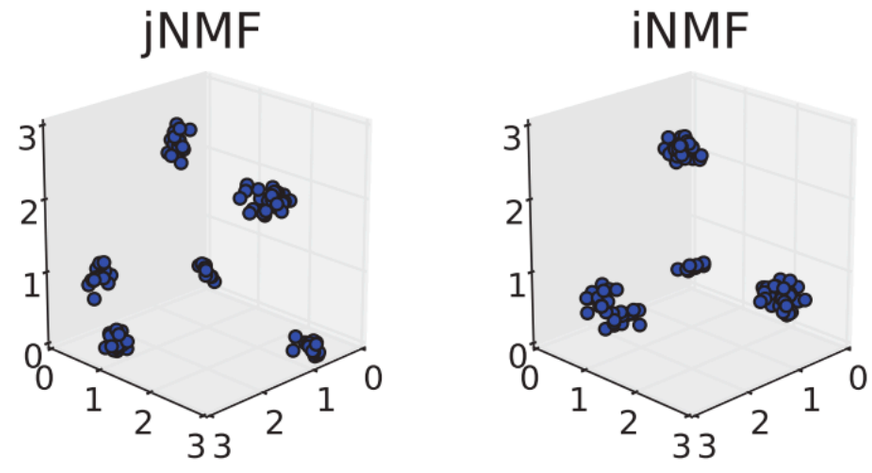
$$\min_{W^i,H,H^i}\left(\sum_i \|X^i - W^i(H+H^i)\|_F^2 + \lambda \sum_i \|W^i H^i\|_F^2\right)$$

- integrative NMF identifies both **homogeneous** effects between datasets (*H*) as well as **heterogeneous** (*H*^{i})
- λ 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





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

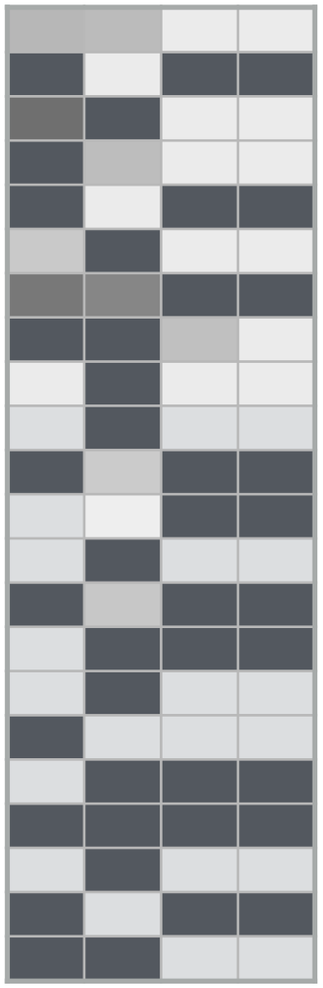
[Yang, Michailidis (2015)]

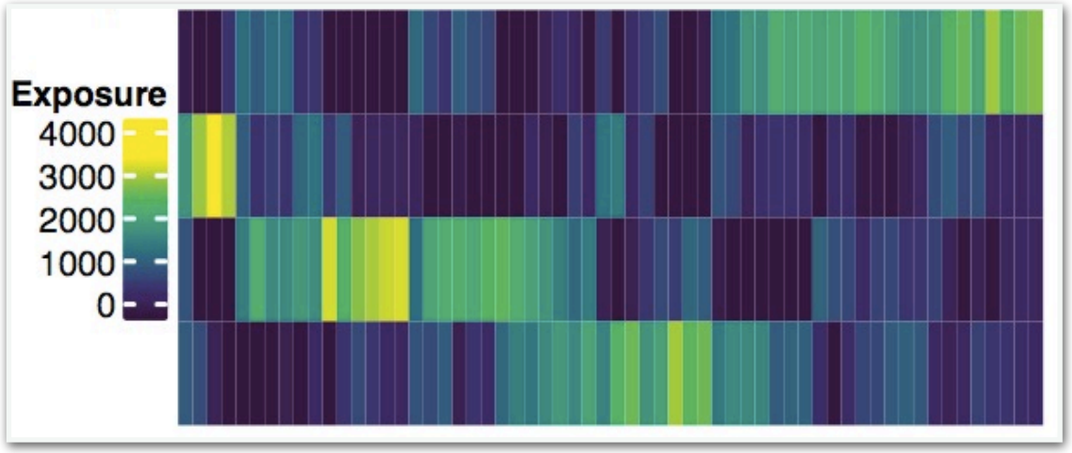
Medizinische Fakultät Heidelberg

4. interpreting signatures

Interpreting signatures

What is special about features highly ranked by signature k ?



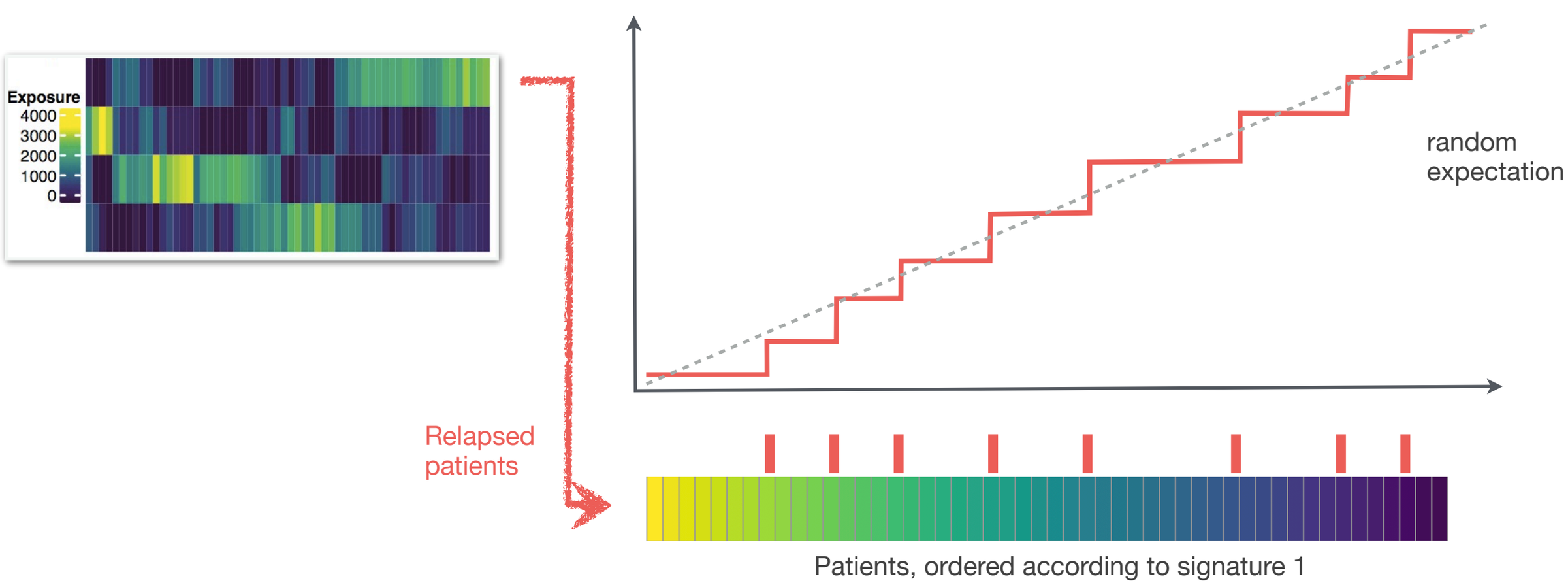


What is special about samples highly exposed to signature k ?

Recovery analysis

Medizinische Fakultät Heidelberg

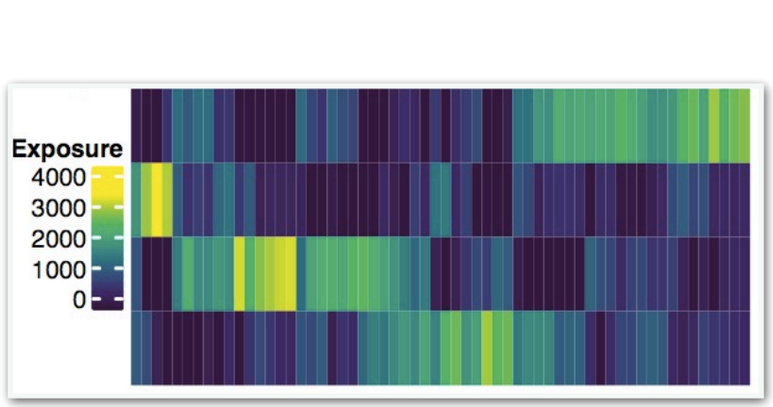
- Rank all samples / features according to the weights / exposures to signature k
- perform recovery analysis using a set of pre-defined regions / genesets / clinical features

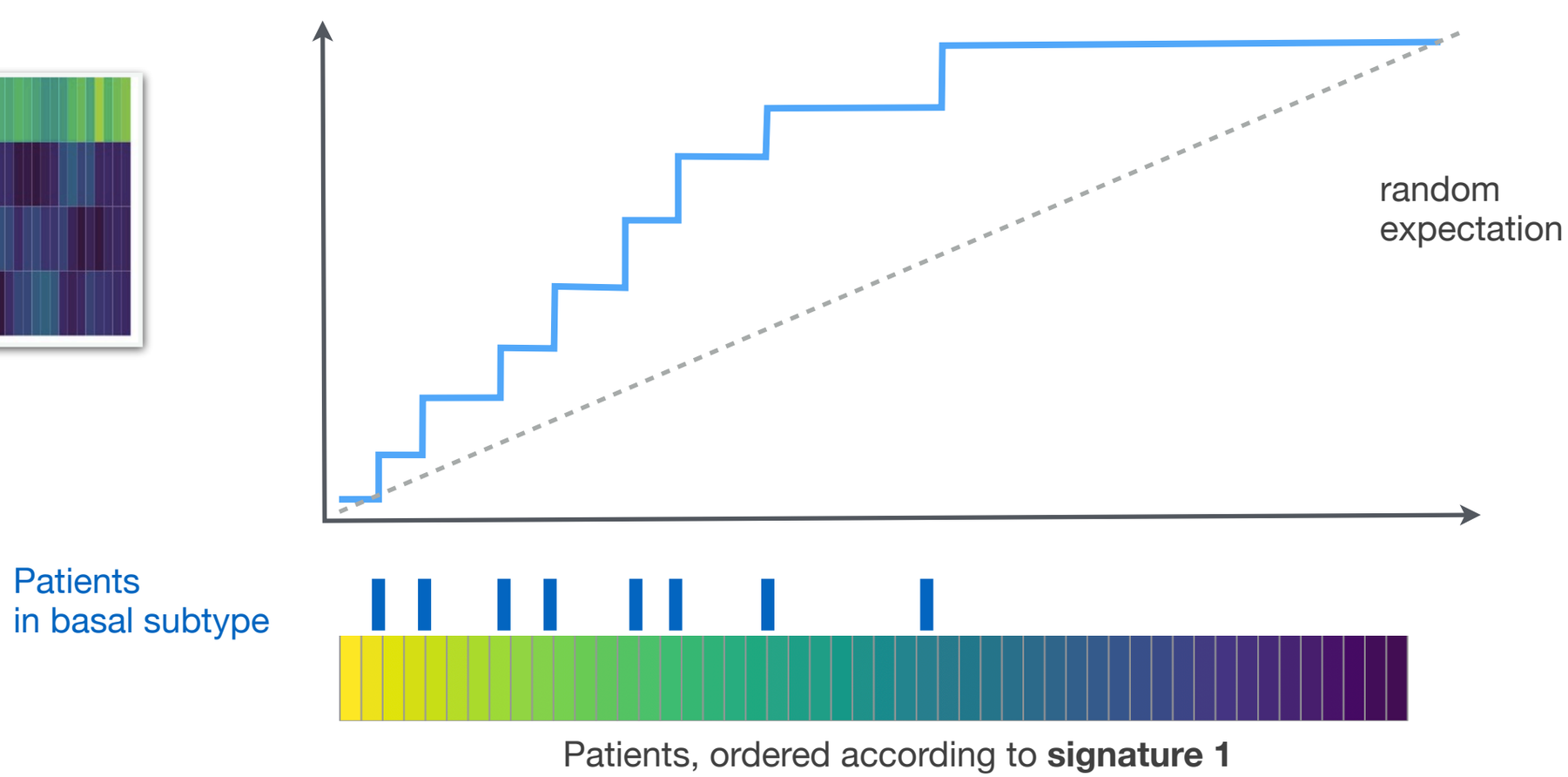


Recovery analysis

Medizinische Fakultät Heidelberg

- 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

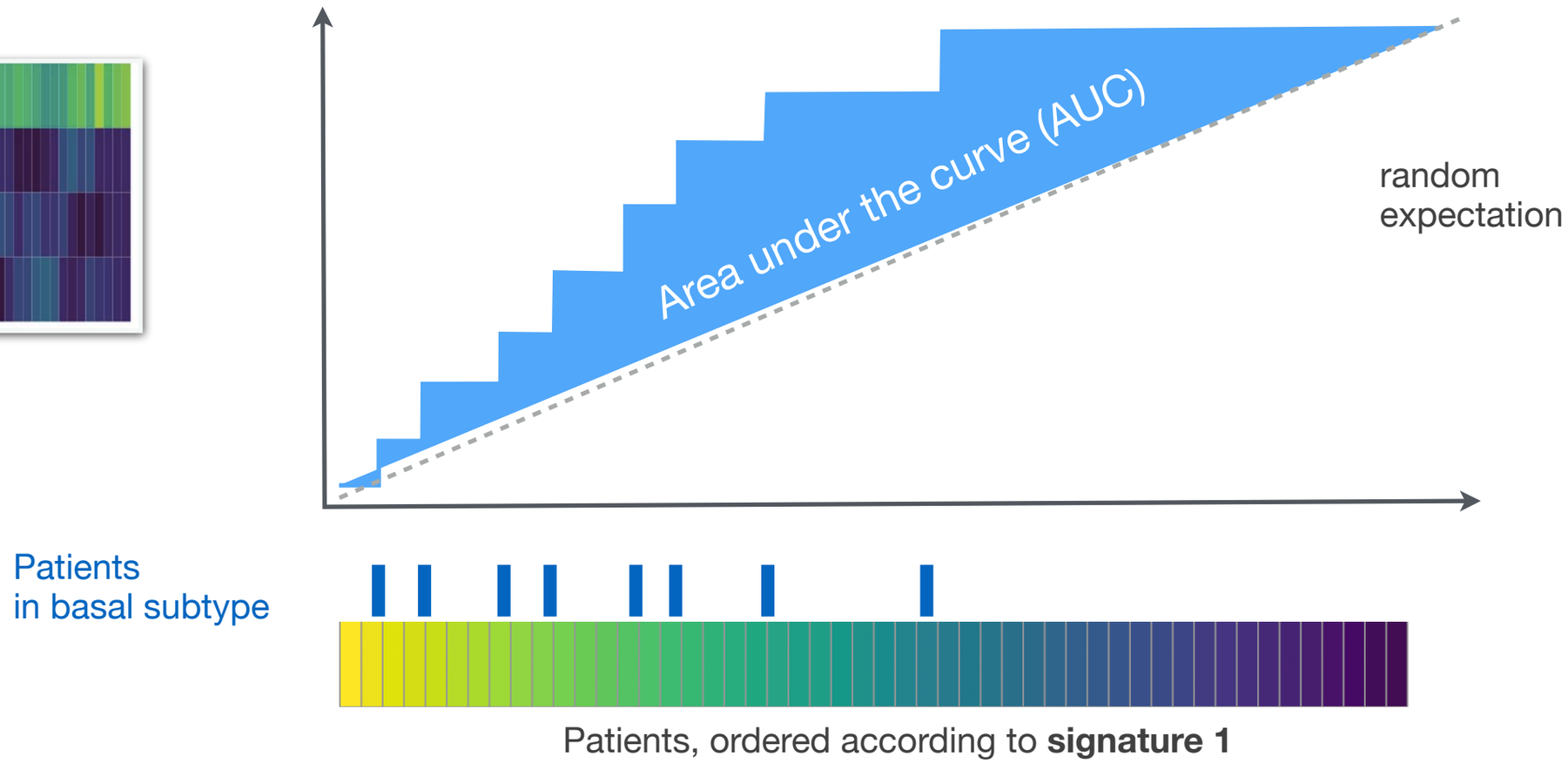




Recovery analysis

Medizinische Fakultät Heidelberg

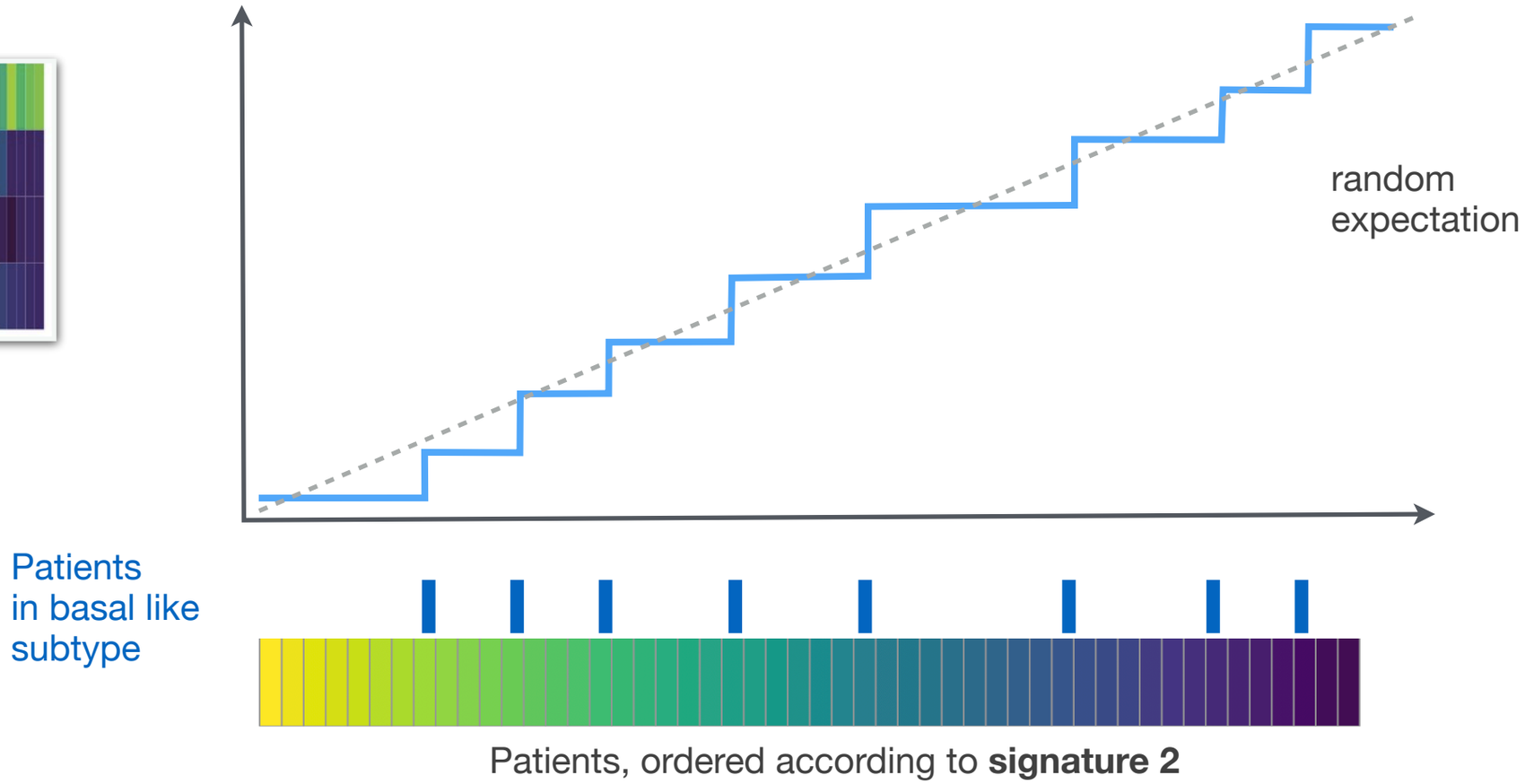
- Rank all samples / features according to the weights / exposures to signature k
- perform recovery analysis using a set of pre-defined regions / genesets / clinical features



Recovery analysis

Medizinische Fakultät Heidelberg

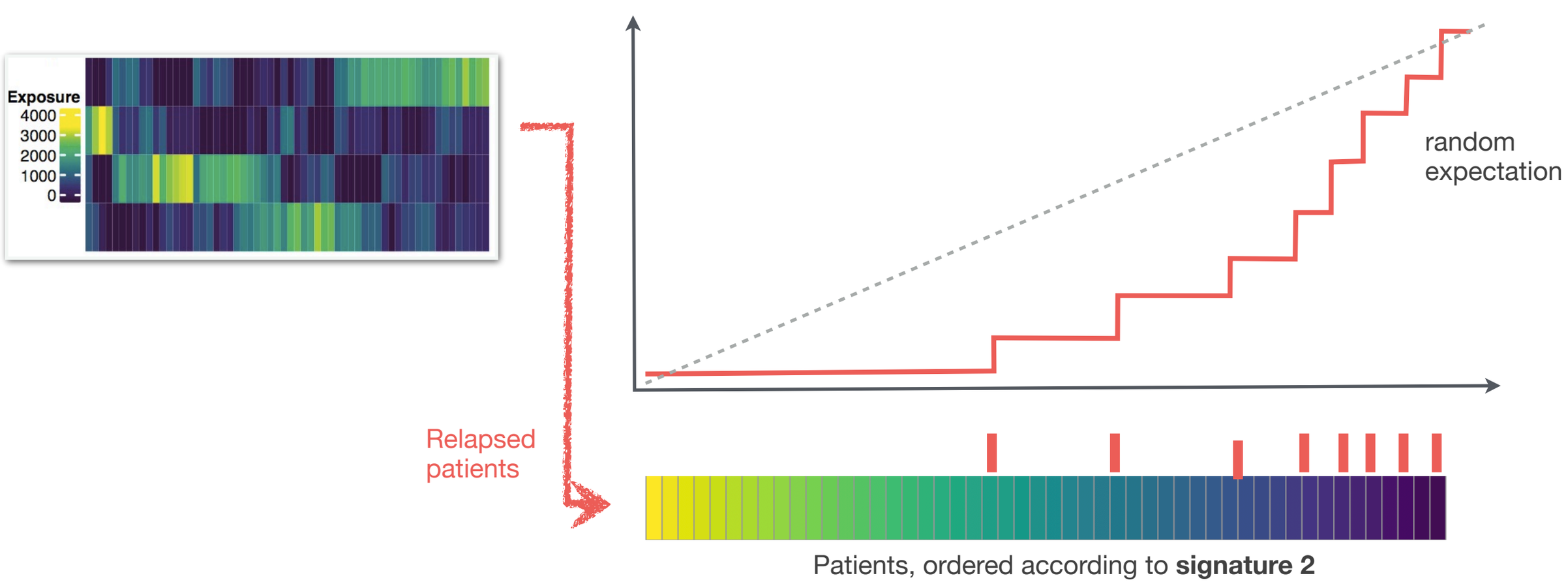
- Rank all samples / features according to the weights / exposures to signature k
- perform recovery analysis using a set of pre-defined regions / genesets / clinical features



Recovery analysis

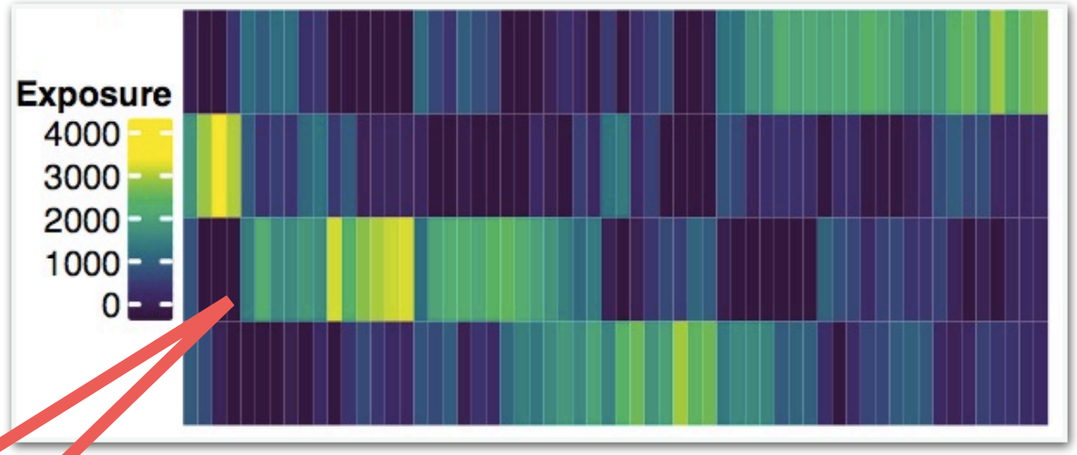
Medizinische Fakultät Heidelberg

- 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

Medizinische Fakultät Heidelberg

- 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**