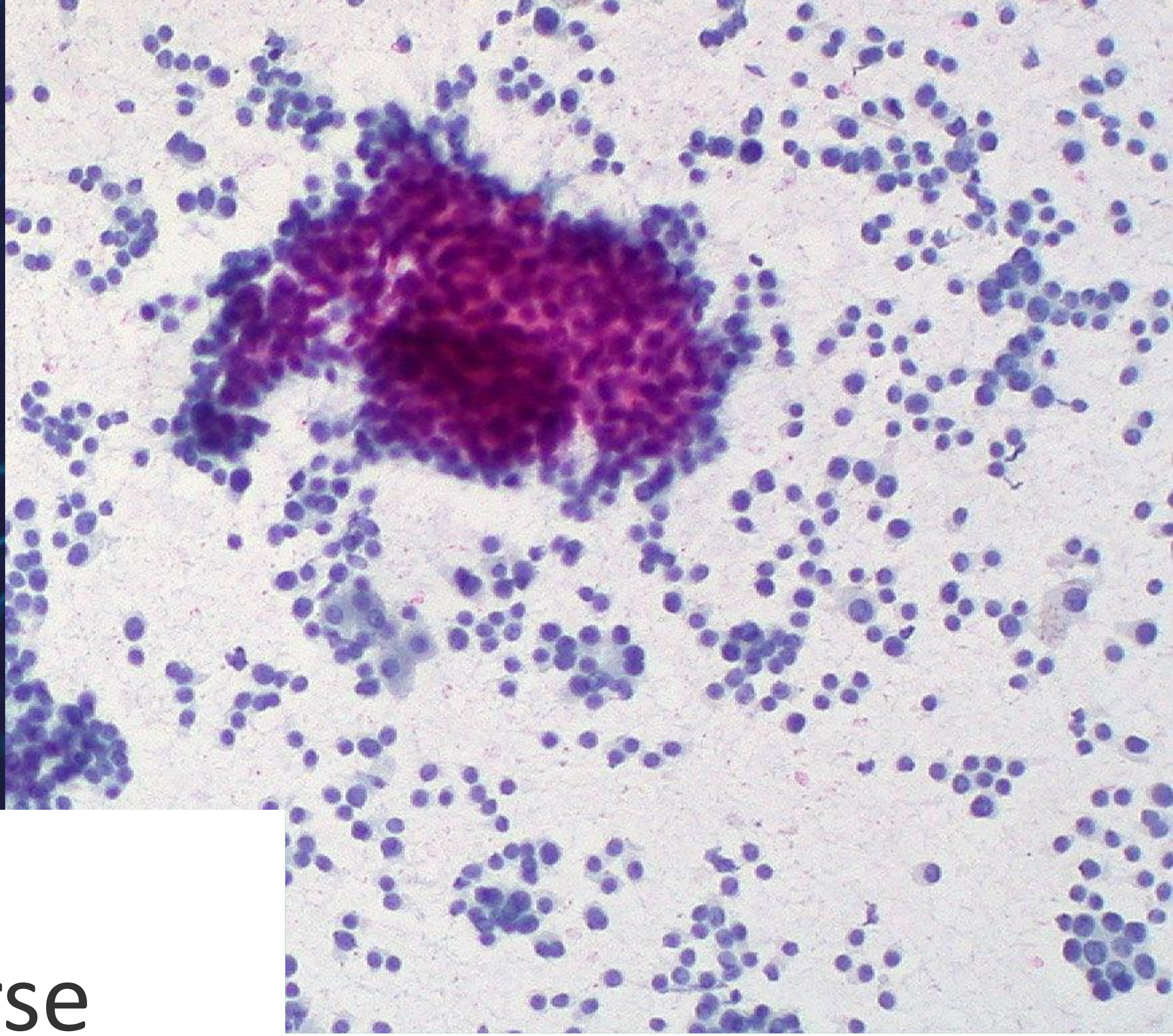


15-16 February 2021

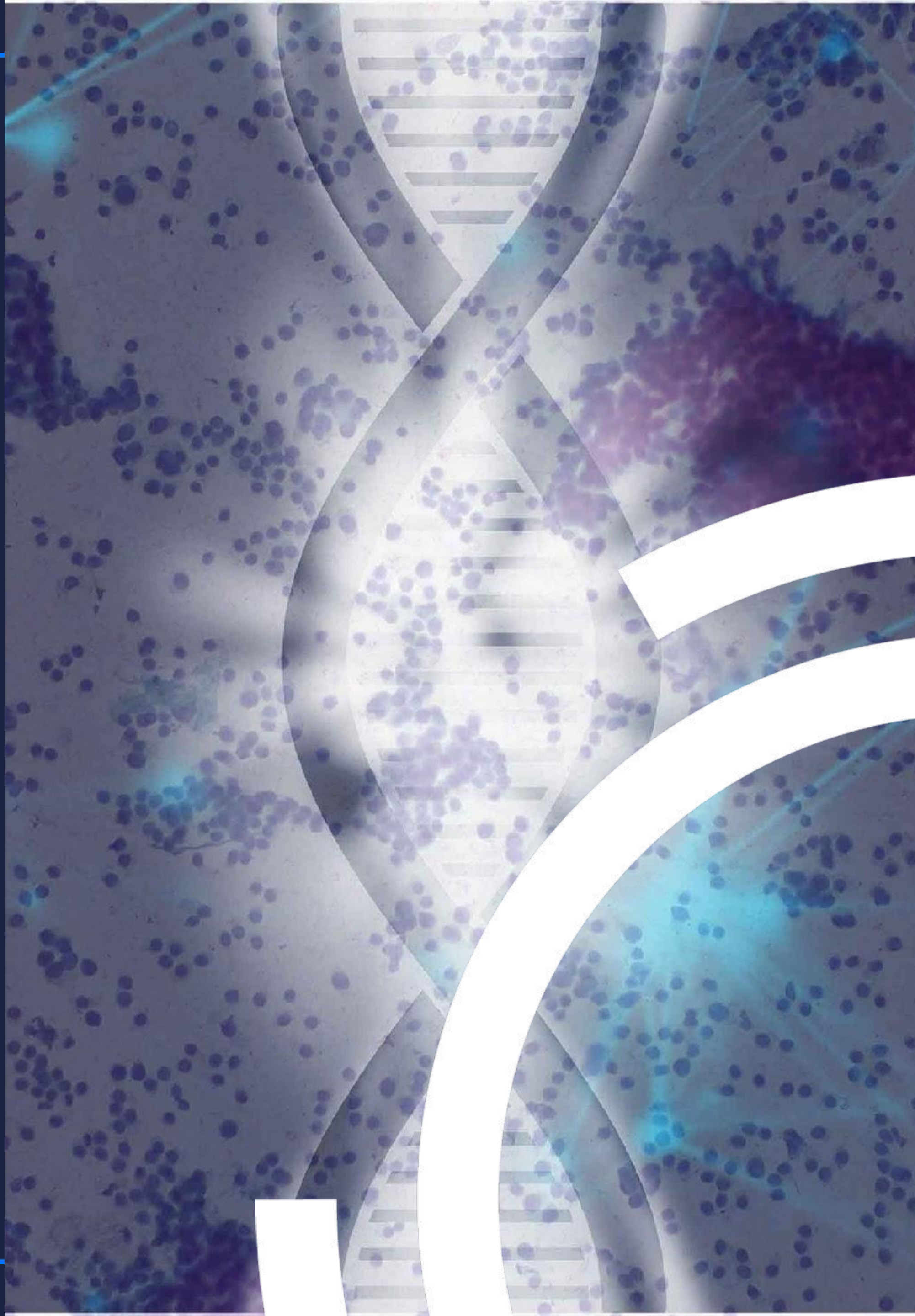
COMETH Training course

From omics data

to tumor heterogeneity quantification

EIT Health is supported by the EIT,
a body of the European Union

How do I start ?



15 January 2021

Clémentine Decamps

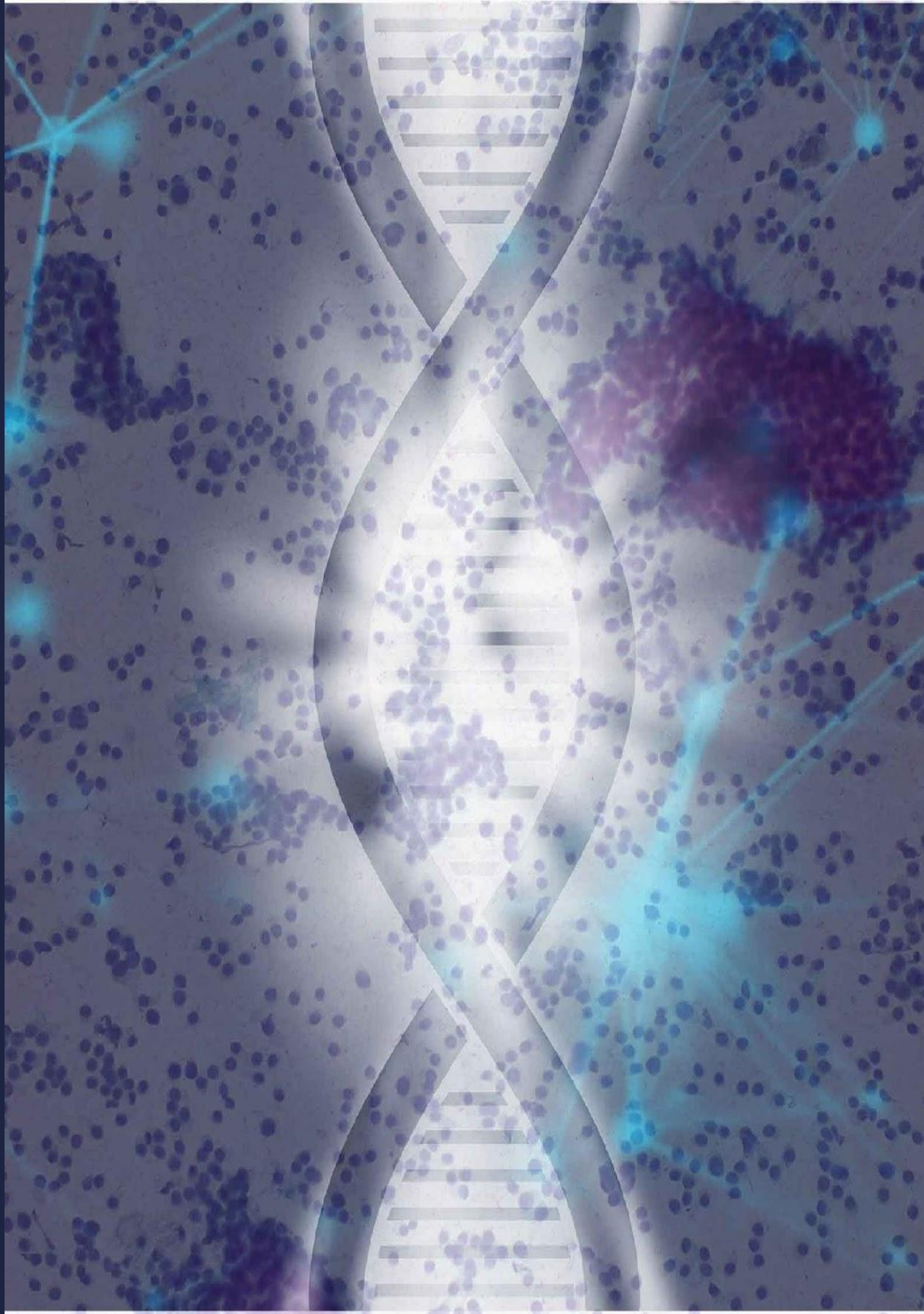
Yasmina Kermezli

EIT Health is supported by the EIT, a body of the European Union

Transcriptomic Data

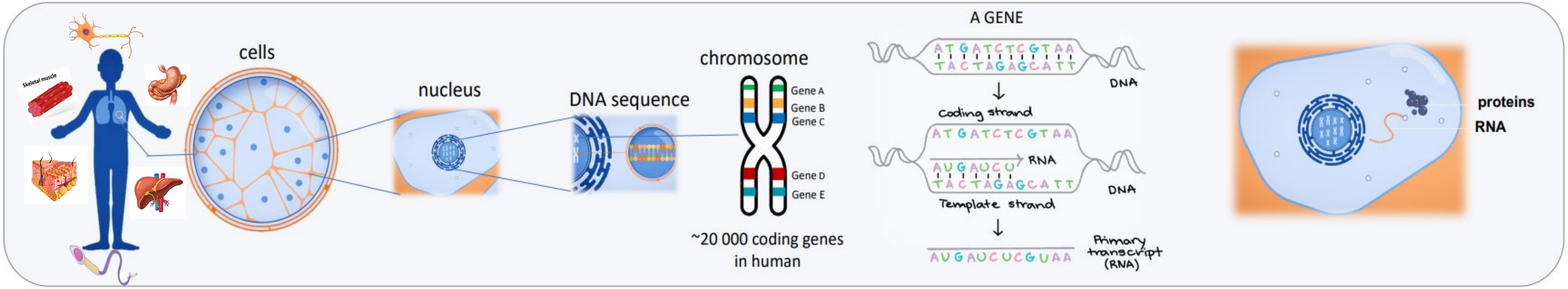
[Manipulation & Normalization]

Yasmina Kermezli



The complexity of human cells

- ❑ 10e^{14} cells Human body
- ❑ All cells with the same genome
- ❑ Differents phenotypes and behaviors

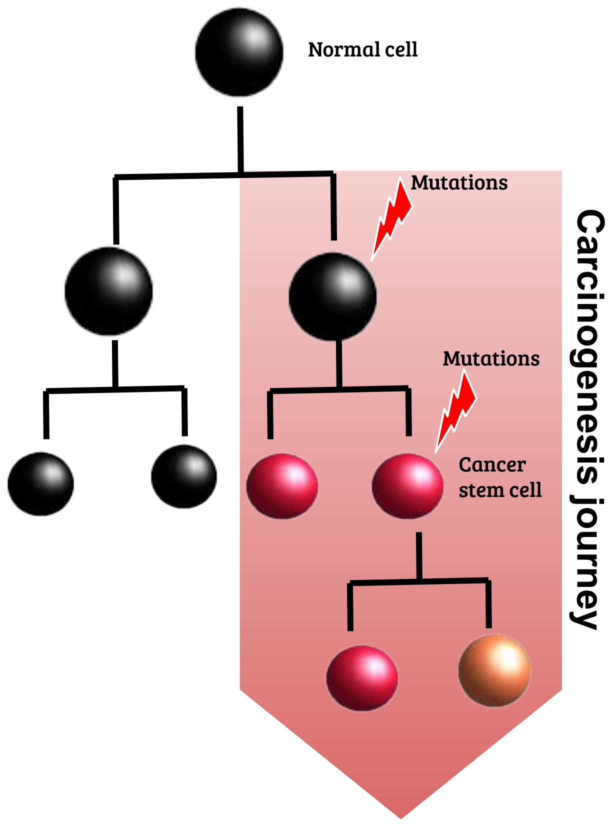


Credits Y. Blum (adapted)

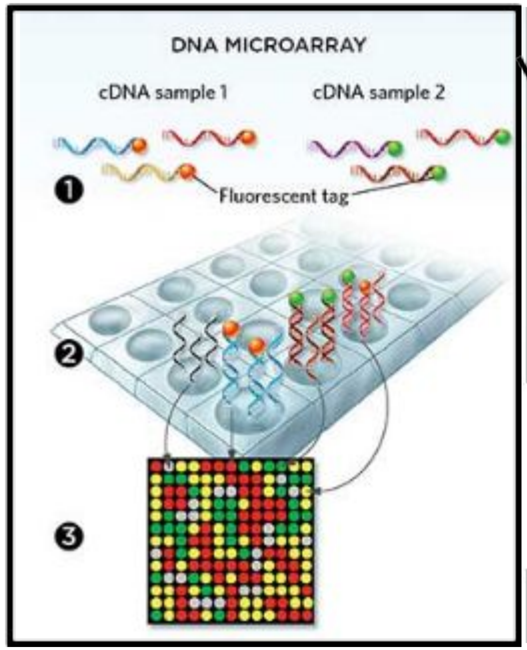
Deregulation of Homeostasis

mutation can:

- ❑ Lead to differences in expressed genes.
- ❑ Affect the type and quantity of RNAs and proteins produced.

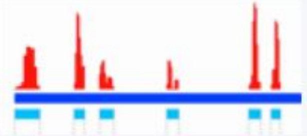


Two types of technologies

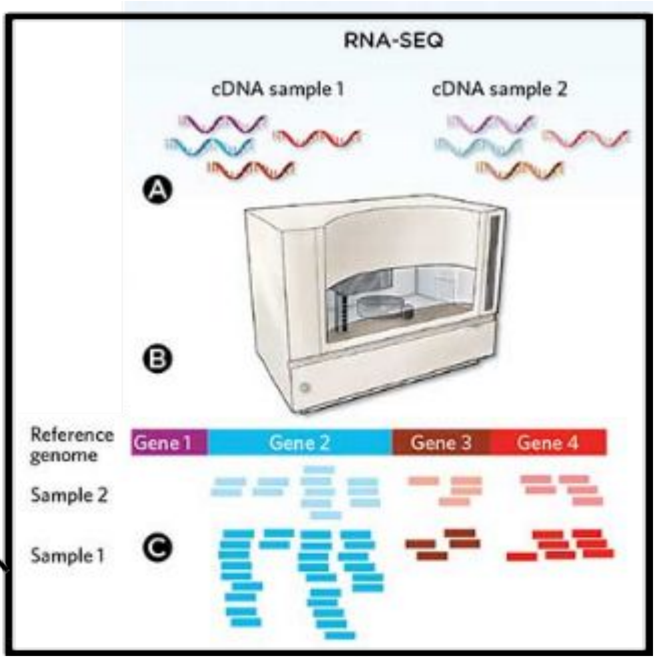


Kate Yandell (2015), TheScientist

Hybridization on a DNA chip (genes)	Massive sequencing of transcripts
Microarray	**RNA-seq**
~150 € per sample	~200-300 € per sample
Technology of the 2000s	Technology of the 2010s
1st publication in 1998	1st publication in 2008



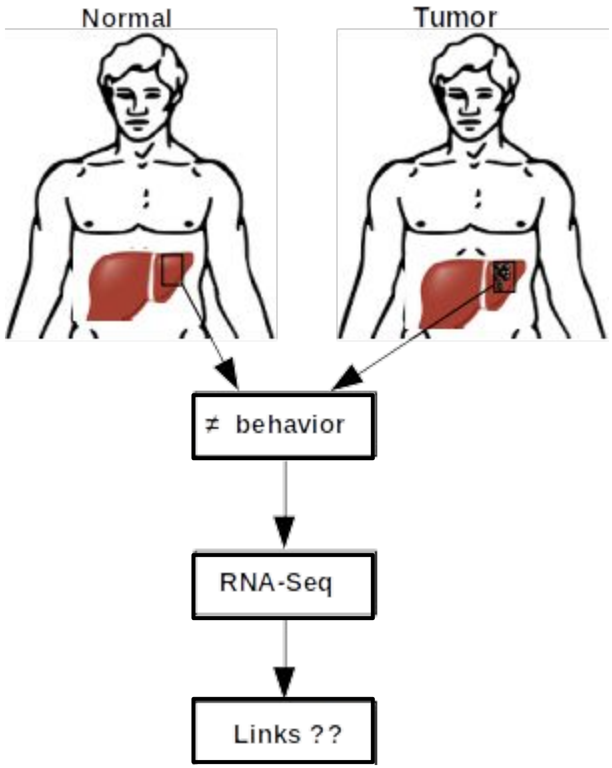
Credits Y. Blum



Kate Yandell (2015), TheScientist

RNA-Seq questions

- Which genes are differentially expressed between sample groups?
- Are there any trends in gene expression over time or across conditions?
- Which groups of genes change similarly over time or across conditions?
- What processes or pathways are enriched for a condition of interest



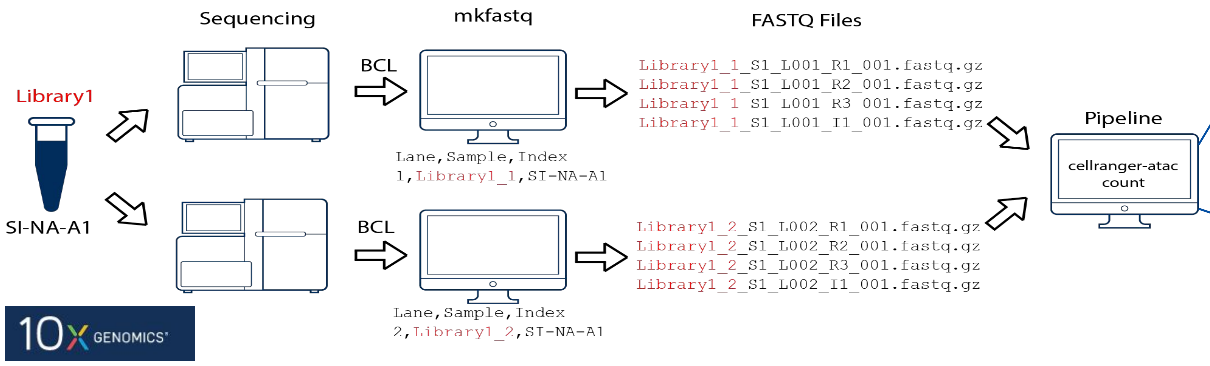
RNA-Seq data analysis

Read : DNA sequence from one fragment (a small section of DNA).

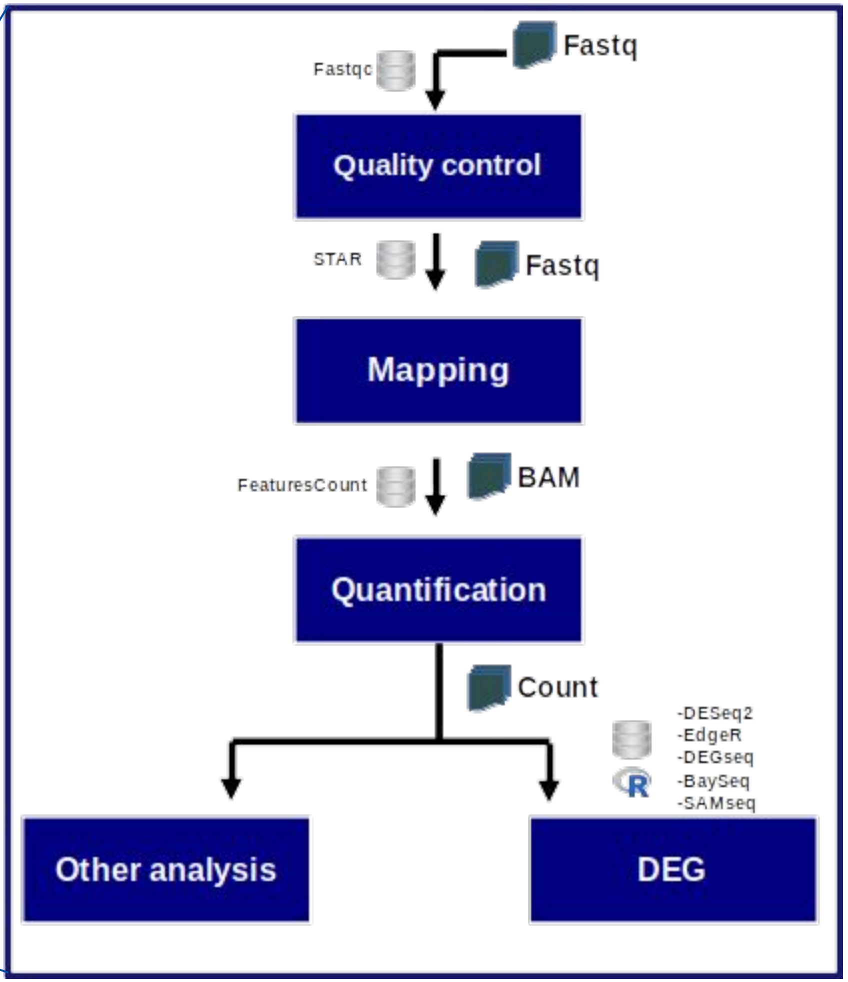
FASTQ : text-based format for storing both a biological sequence and its corresponding quality scores.

BAM(Binary Alignment Map) : contains information about mapped /unmapped reads

Count: the number of reads or fragments aligning to the exons of each gene

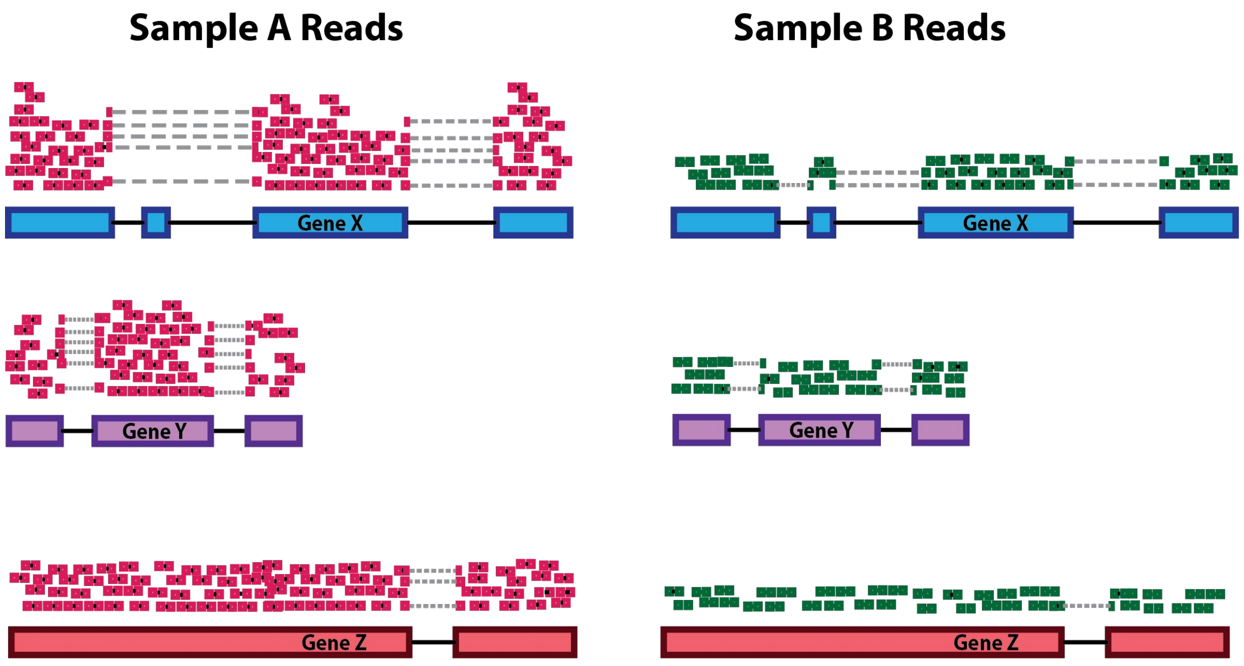


Bioinformatic pipeline

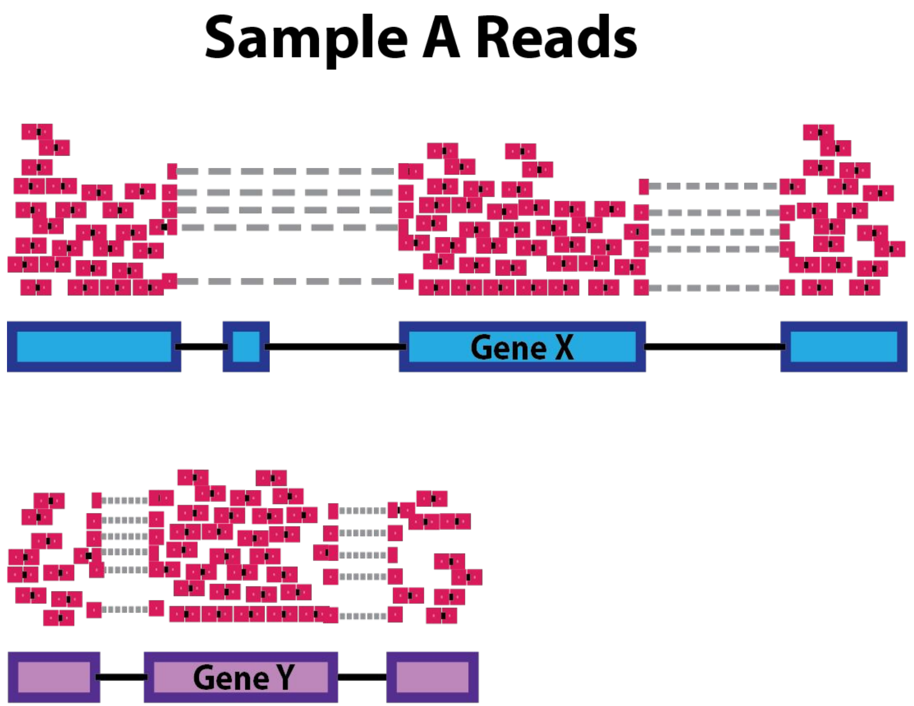


Main factors

Sequencing depth



Gene length



Common normalization methods

$$\text{RPM or CPM} = \frac{\text{Number of reads mapped to gene} \times 10^6}{\text{Total number of mapped reads}}$$

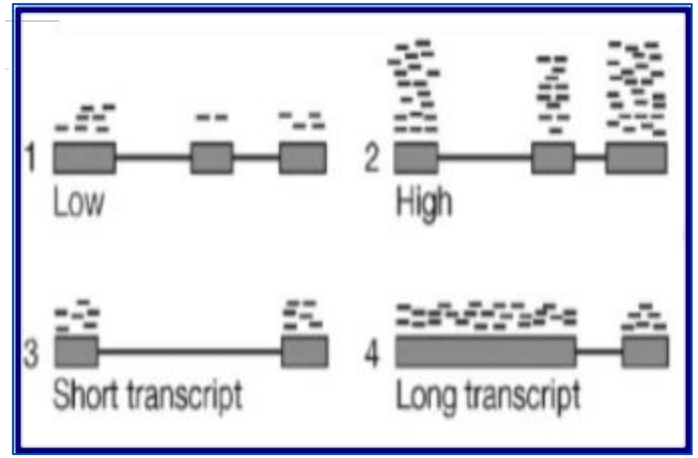
$$\text{RPKM} = \frac{\text{Number of reads mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{gene length in bp}}$$

$$\text{TPM} = A \times \frac{1}{\sum(A)} \times 10^6$$

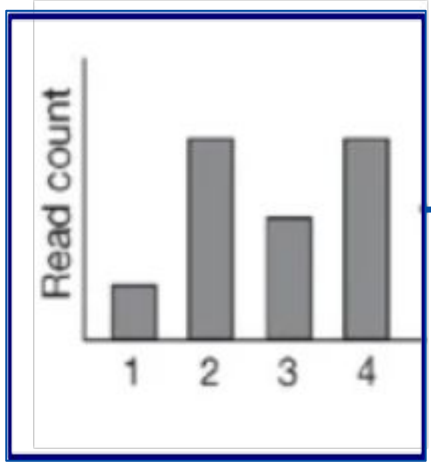
$$\text{Where A} = \frac{\text{total reads mapped to gene} \times 10^3}{\text{gene length in bp}}$$

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

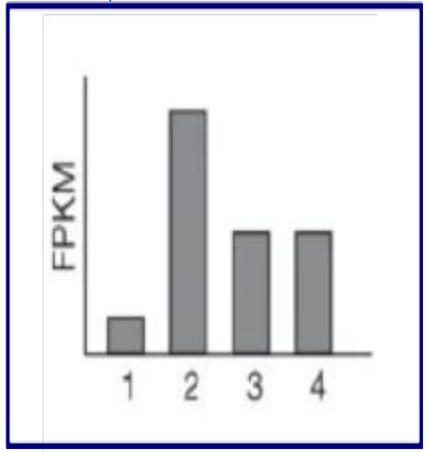
From Counts to FPKM



Mapping



Count



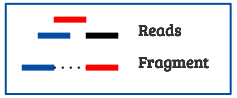
FPKM Norm

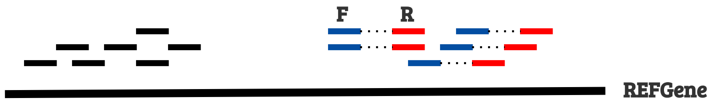
Single-end

RPKM (Reads Per Kilobase Million)

paired-end

FPKM (Fragments Per Kilobase Million)





FPKM takes into account that two reads can map to one fragment (and so it doesn't count this fragment twice)

Computational methods for transcriptome annotation and quantification using RNA-seq

Manuel Garber, Manfred G Grabherr, Mitchell Guttman & Cole Trapnell

Nature Methods **8**, 469–477(2011) | Cite this article

```r

## Install the library if needed then load it

if(!require("DESeq2")){
  install.packages("lazyeval")
  install.packages("ggplot2")

  source("http://bioconductor.org/biocLite.R")
  biocLite("DESeq2")
}

###  call the library

library("DESeq2")


###################


## Create DESeq2Dataset object

dds <- DESeqDataSetFromMatrix(countData = data, colData = meta, design = ~ sampletype)

### Generate normalized counts

dds <- estimateSizeFactors(dds)

### affect them to an object

normalized_counts <- counts(dds, normalized=TRUE)

##save this normalized data matrix to file for later use

write.table(normalized_counts, file="data/normalized_counts.txt", sep="\t", quote=F, col.names=NA)

```

	Raw data	
Gene	S1	S2
A	1234	800
B	23	15
C	1	12

Step 1: creates a pseudo-reference sample (row-wise geometric mean)

Raw data

Gene	S1	S2	PseudoRef
A	1234	800	Sqrt(1234*800) = 993.5794
B	23	15	Sqrt(23*15) = 18.57418
C	1	12	Sqrt(1*12) = 3.464102

Step 1: creates a pseudo-reference sample (row-wise geometric mean)

Step 2: calculates ratio of each sample to the reference

Raw data: S1, S2

Gene	S1	S2	PseudoRef	Ratio[S1]	Ratio[S2]
A	1234	800	Sqrt(1234*800) = 993.5794	1234/993.5794 = 1.241974	800/993.5794 = 0.8051697
B	23	15	Sqrt(23*15) = 18.57418	23/18.57418 = 1.238278	15/18.57418 = 0.8075727
C	1	12	Sqrt(1*12) = 3.464102	1/3.464102 = 0.2886751	12/3.464102 = 3.464101

Step 1: creates a pseudo-reference sample (row-wise geometric mean)

Step 2: calculates ratio of each sample to the reference

Step 3: calculate the normalization factor for each sample (size factor)

Raw data (S1, S2)

Gene	S1	S2	PseudoRef	Ratio[S1]	Ratio[S2]
A	1234	800	Sqrt(1234*800) = 993.5794	1234/993.5794 = 1.241974	800/993.5794 = 0.8051697
B	23	15	Sqrt(23*15) = 18.57418	23/18.57418 = 1.238278	15/18.57418 = 0.8075727
C	1	12	Sqrt(1*12) = 3.464102	1/3.464102 = 0.2886751	12/3.464102 = 3.464101

```
normalization_factor_sampleA <- median(c(1.241974, 1.238278, 0.2886751))

normalization_factor_sampleB <- median(c(0.8051697, 0.8075727, 3.464101))
```

Step 1: creates a pseudo-reference sample (row-wise geometric mean)

Step 2: calculates ratio of each sample to the reference

Step 3: calculate the normalization factor for each sample (size factor)

Raw data

Gene	S1	S2	PseudoRef	Ratio[S1]	Ratio[S2]
A	1234	800	Sqrt(1234*800) = 993.5794	1234/993.5794 = 1.241974	800/993.5794 = 0.8051697
B	23	15	Sqrt(23*15) = 18.57418	23/18.57418 = 1.238278	15/18.57418 = 0.8075727
C	1	12	Sqrt(1*12) = 3.464102	1/3.464102 = 0.2886751	12/3.464102 = 3.464101
				1.238278	0.8075727

```
normalization_factor_sampleA <- median(c(1.241974, 1.238278, 0.2886751))

normalization_factor_sampleB <- median(c(0.8051697, 0.8075727, 3.464101))
```

Step 1: creates a pseudo-reference sample (row-wise geometric mean)

Step 2: calculates ratio of each sample to the reference

Step 3: calculate the normalization factor for each sample (size factor)

Step 4: calculate the normalized count values using the normalization factor

	Raw data					Normalized data	
Gene	S1	S2	PseudoRef	Ratio[S1]	Ratio[S2]	S1	S2
A	1234	800	Sqrt(1234*800) = 993.5794	1234/993.5794 = 1.241974	800/993.5794 = 0.8051697	1234/1.238278 = 996.5452	800/0.8075727 = 990.6229
B	23	15	Sqrt(23*15) = 18.57418	23/18.57418 = 1.238278	15/18.57418 = 0.8075727	23/1.238278 = 18.57418	15/0.8075727 = 18.57418
C	1	12	Sqrt(1*12) = 3.464102	1/3.464102 = 0.2886751	12/3.464102 = 3.464101	1/ 1.238278 = 0.8075731	12/0.8075727 = 14.85934
				1.238278	0.8075727		

Step 1: creates a pseudo-reference sample (row-wise geometric mean)

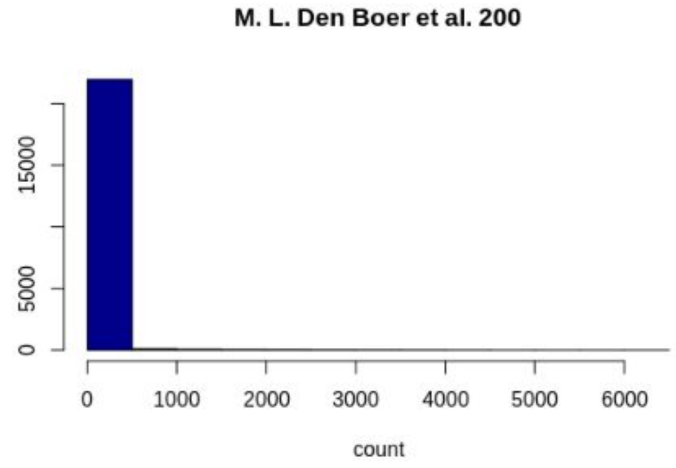
Step 2: calculates ratio of each sample to the reference

Step 3: calculate the normalization factor for each sample (size factor)

Step 4: calculate the normalized count values using the normalization factor

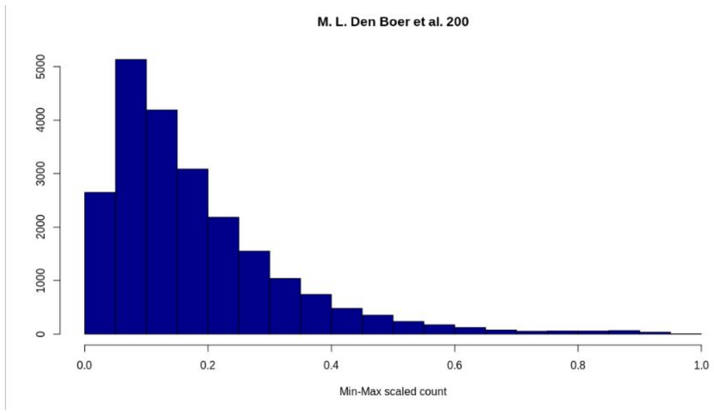
	Raw data		Normalized data	
Gene	S1	S2	S1	S2
A	1234	800	1234/1.238278 = 996.5452	800/0.8075727 = 990.6229
B	23	15	23/1.238278 = 18.57418	15/0.8075727 = 18.57418
C	1	12	1/ 1.238278 = 0.8075731	12/0.8075727 = 14.85934

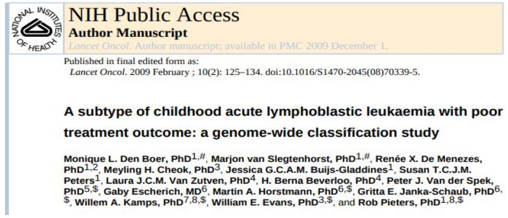
Linear



Min-Max

$$\frac{\textit{Value} - \textit{min}}{\textit{max} - \textit{min}}$$




NIH Public Access
Author Manuscript
Lancet Oncol. Author manuscript; available in PMC 2009 December 1.

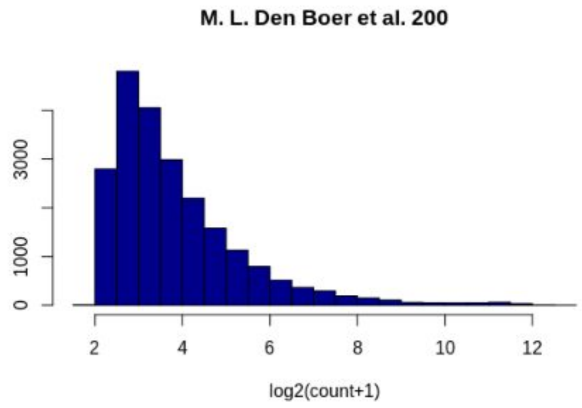
Published in final edited form as:
Lancet Oncol. 2009 February ; 10(2): 125–134. doi:10.1016/S1470-2045(08)70339-5.

A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study

Monique L. Den Boer, PhD[1,#], Marjon van Slegtenhorst, PhD[1,#], Renée X. De Menezes, PhD[1,2], Meyling H. Cheok, PhD[3], Jessica G.C.A.M. Buijs-Gladdines[1], Susan T.C.J.M. Peters[1], Laura J.C.M. Van Zutven, PhD[4], H. Berna Beverloo, PhD[4], Peter J. Van der Spek, PhD[5,$], Gaby Escherich, MD[6], Martin A. Horstmann, PhD[6,$], Gritta E. Janka-Schaub, PhD[6,$], Willem A. Kamps, PhD[7,8,$], William E. Evans, PhD[3,$], and Rob Pieters, PhD[1,8,$]

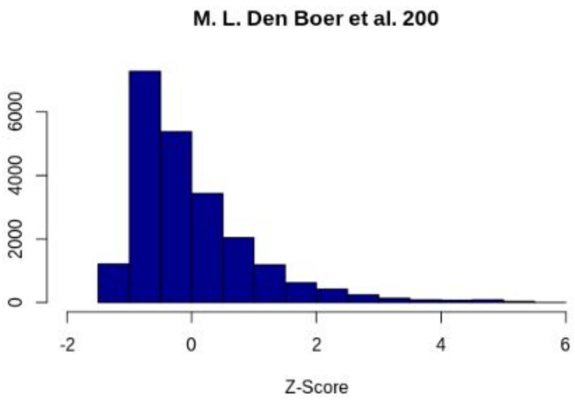
Pseudo-Log

Log2(Value+1)

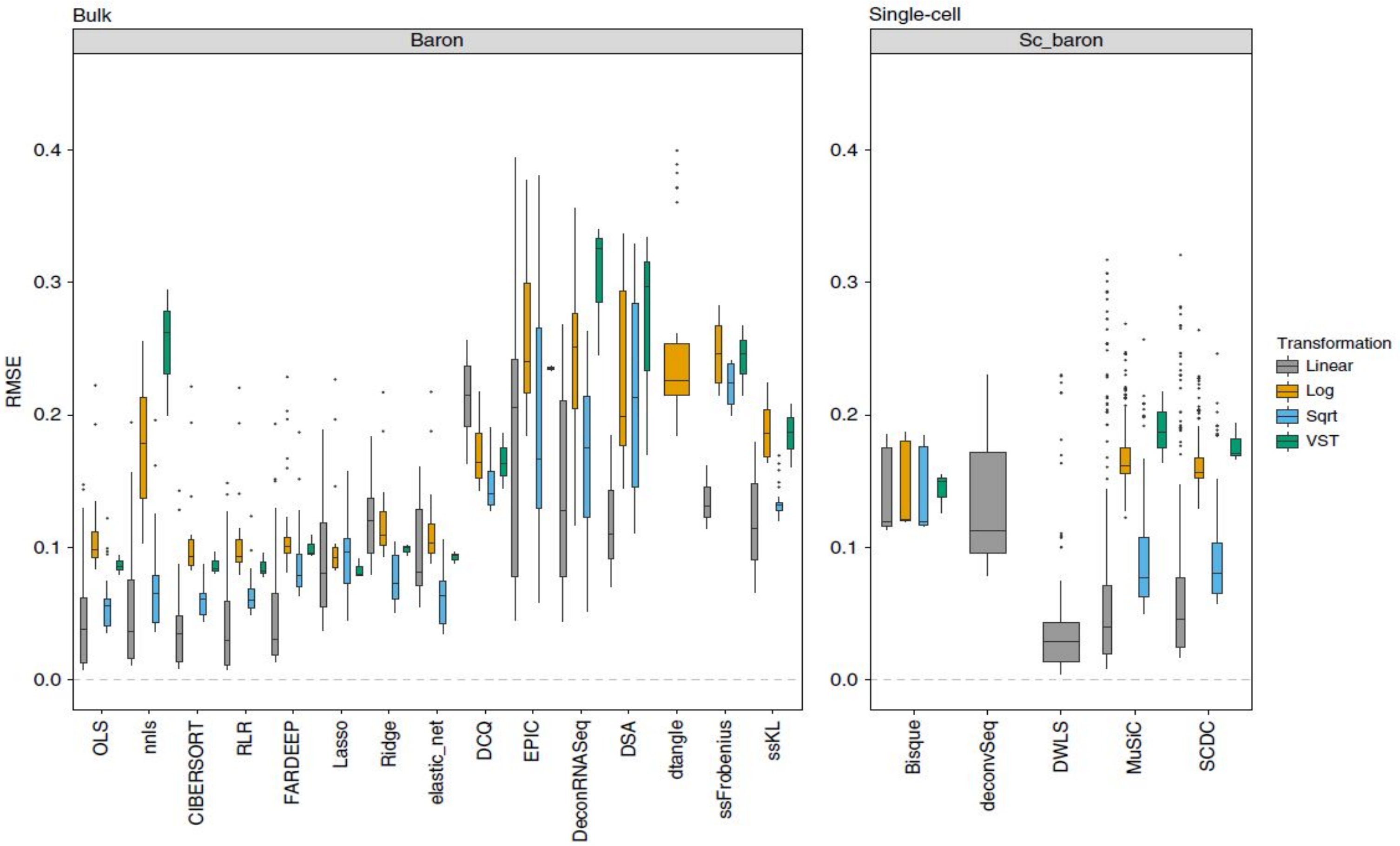


Z-Score

$$\frac{\textit{Value} - \mu}{\sigma}$$



Impact of the data transformation on the deconvolution results



nature COMMUNICATIONS

ARTICLE
Check for updates
https://doi.org/10.1038/s41467-020-19015-1 OPEN

Benchmarking of cell type deconvolution pipelines for transcriptomics data

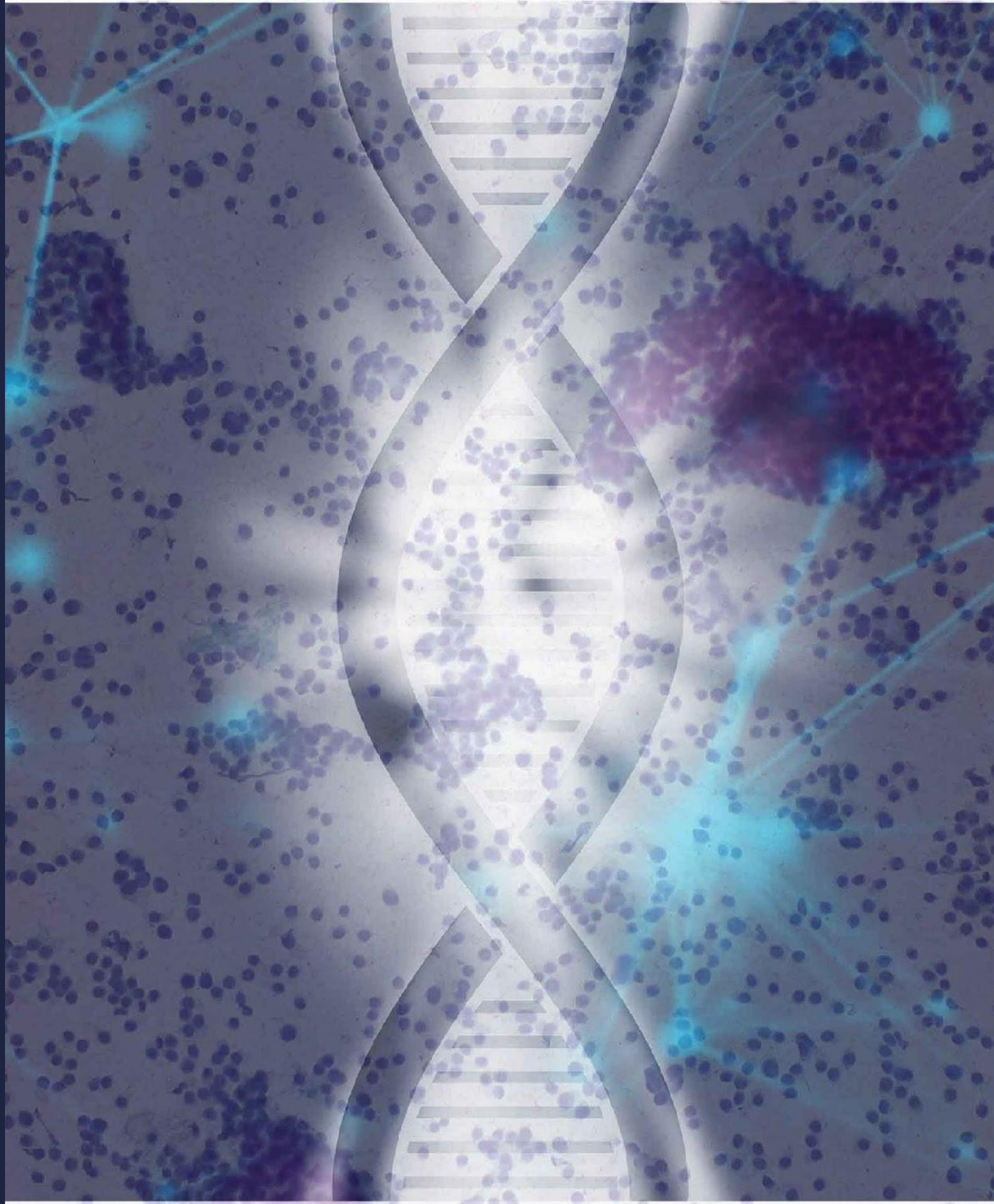
Francisco Avila Cobos [1,2,3], José Alquicira-Hernandez [3,4], Joseph E. Powell [3,4,5], Pieter Mestdagh [1,2,5] & Katleen De Preter [1,2,5]

Thank you for your attention!

DNA methylation Data

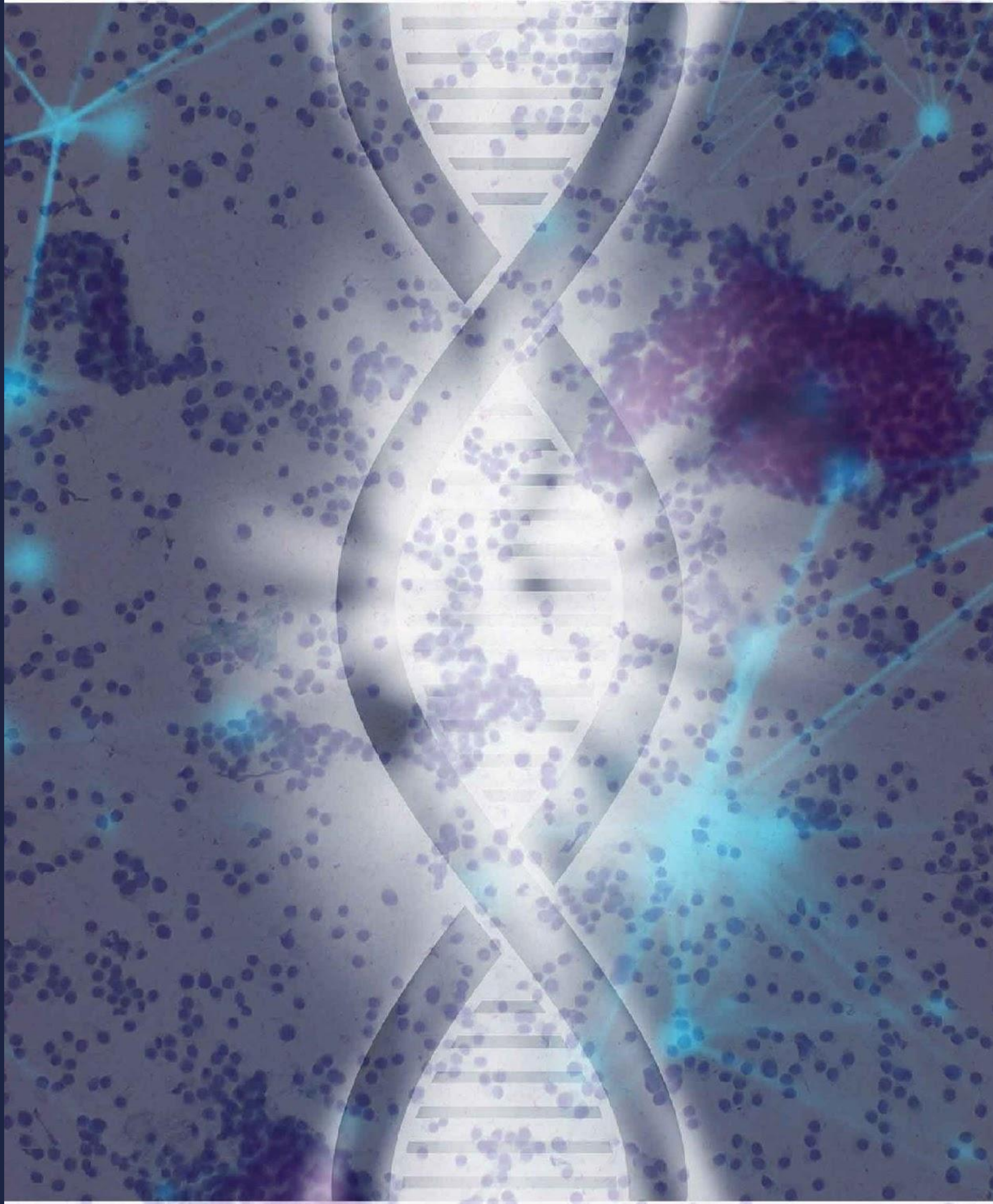
[Manipulation & Normalization]

Clémentine Decamps



DNAm normalization using lumi

Introduction

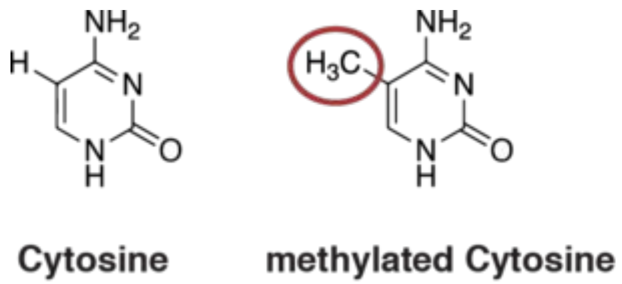


Introduction

DNA methylation

-> Addition of a methyl group on a cytosine of the DNA

DNA methylation and cancer (book)



-> Different technologies: bisulfite sequencing, beadchip,...

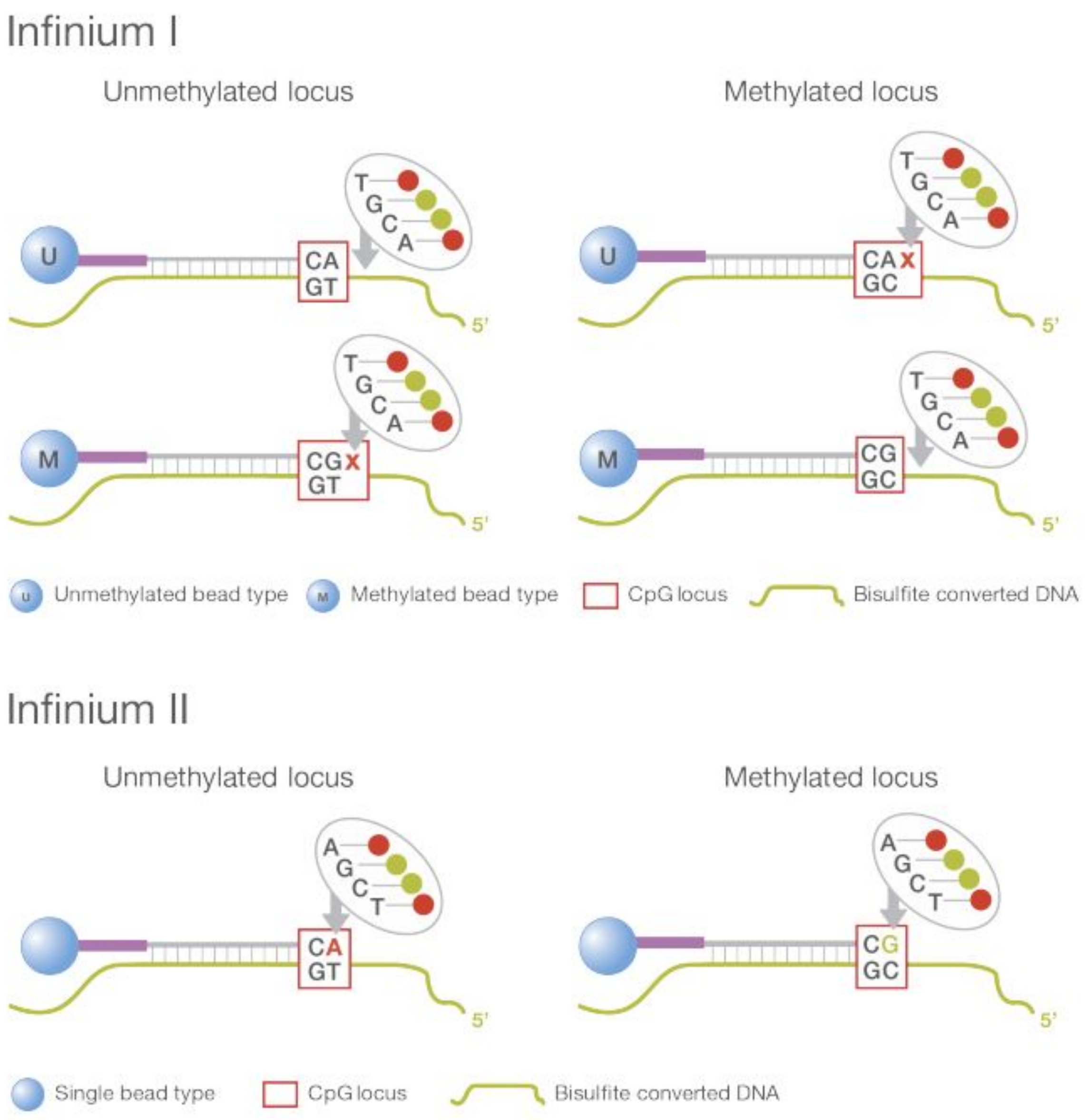
-> BeadChip: 27k, 450k, 850k,...

Infinium HumanMethylation450 BeadChip

-> Here we focus on **850k beadchip**

Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequence | Epigenomics

Introduction



Introduction

850k normalization

-> A lot of different methods, and a big impact on the following analyzes

-> As clinician, you have to ask how the datas was normalized

-> As bioinformatician, stay vigilant about your data !

-> As an example, I will present our pipeline, based on lumi package and Illumina guide.

Lumi package on bioconductor

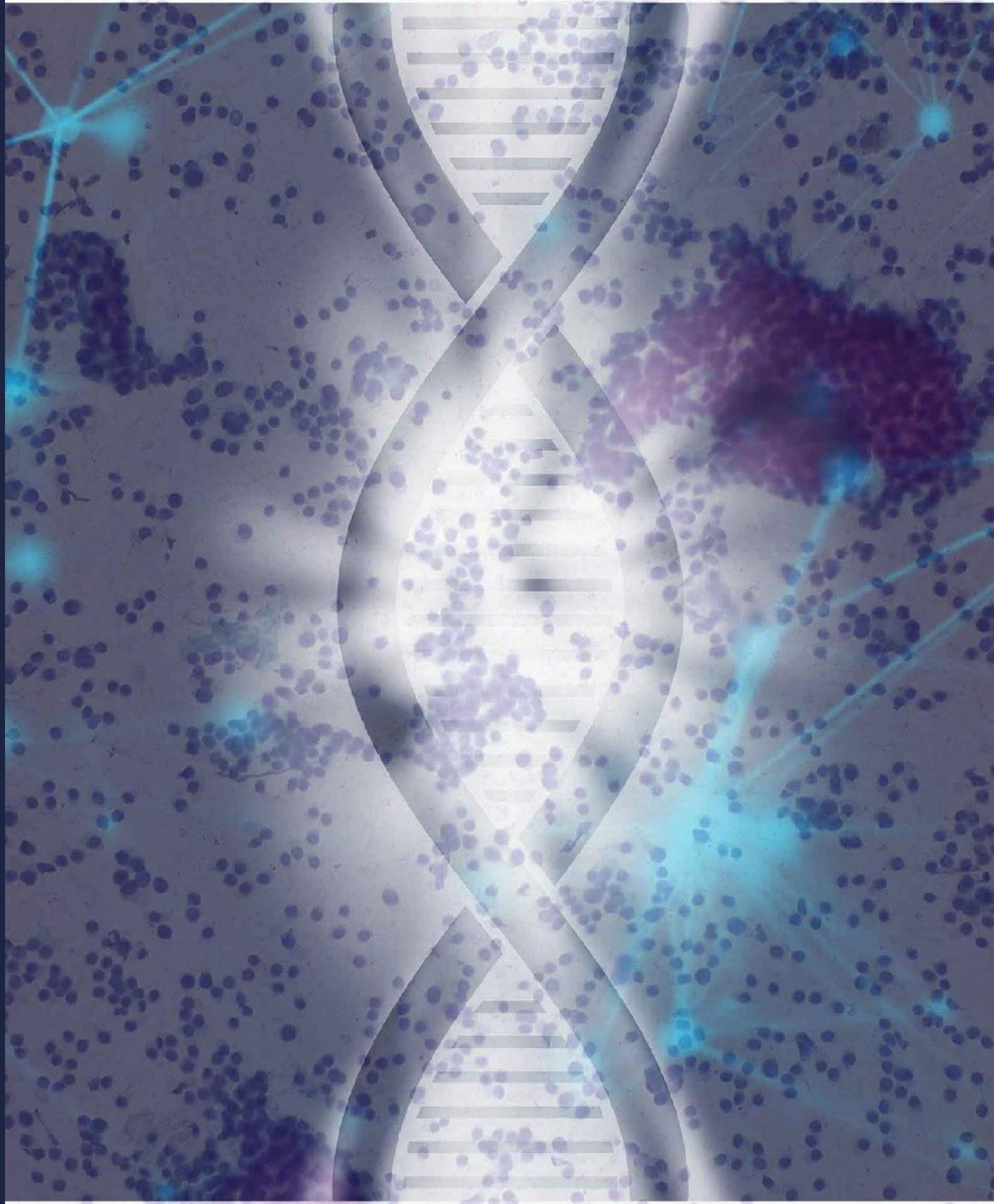
lumi: a pipeline for processing Illumina microarray | Bioinformatics

DNAm normalization using lumi

Introduction

Pre-normalization filtering:

- On probes



Pre-normalization filtering on probes

Probe ID prefix:
- cg: GpG methylation site
- ch: non-CpG methylation site
- rs: non methylated site

-> We only want CpG methylation site

Pre-normalization filtering on probes

Probe ID prefix:
- cg: GpG methylation site
- ch: non-CpG methylation site
- rs: non methylated site

-> We only want CpG methylation site

Probes containing SNP:
-> SNP can distort the signal by removing a citoysine
-> Probes removed

[MethylToSNP: identifying SNPs in Illumina DNA methylation array data]

Pre-normalization filtering on probes

Probe ID prefix:
- cg: GpG methylation site
- ch: non-CpG methylation site
- rs: non methylated site

-> We only want CpG methylation site

Probes containing SNP:
-> SNP can distort the signal by removing a citoysine
-> Probes removed

MethylToSNP: identifying SNPs in Illumina DNA methylation array data

Probes with high intensity:
-> Noise
-> Probes with a mean value > 30,000 between methylated and unmethylated samples are removed

Pre-normalization filtering on probes

Probe ID prefix:
- cg: GpG methylation site
- ch: non-CpG methylation site
- rs: non methylated site
-> We only want CpG methylation site

Probes containing SNP:
-> SNP can distort the signal by removing a citoysine
-> Probes removed

MethylToSNP: identifying SNPs in Illumina DNA methylation array data

Probes with high intensity:
-> Noise
-> Probes with a mean value > 30,000 between methylated and unmethylated samples are removed

Not detected probes:
-> Not informative
-> Probes detected in less than 10% of the samples are removed

Pre-normalization filtering on probes

Probe ID prefix:
- cg: GpG methylation site
- ch: non-CpG methylation site
- rs: non methylated site

-> We only want CpG methylation site

Probes containing SNP:
-> SNP can distort the signal by removing a citoysine
-> Probes removed

[MethylToSNP: identifying SNPs in Illumina DNA methylation array data]

Probes with high intensity:
-> Noise
-> Probes with a mean value > 30,000 between methylated and unmethylated samples are removed

Not detected probes:
-> Not informative
-> Probes detected in less than 10% of the samples are removed

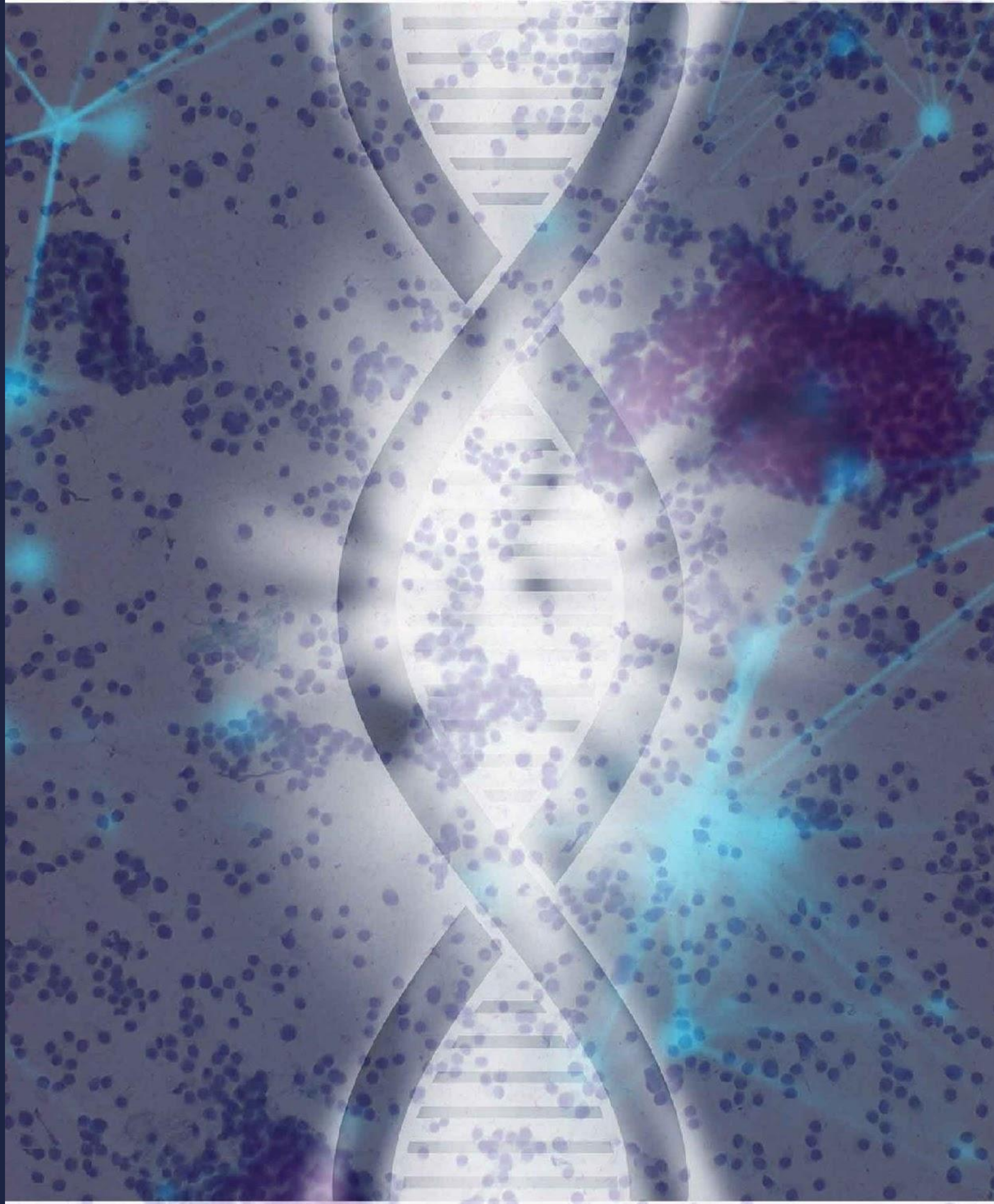
Probes related with sex?
-> Depends a lot of the question
-> We use it as quality check

DNAm normalization using lumi

Introduction

Pre-normalization filtering:

- On probes
- On samples



Pre-normalization filtering on samples

Not detected samples:
-> Not informative
-> Samples with too few probes detected are removed

Pre-normalization filtering on samples

Not detected samples:
-> Not informative
-> Samples with too few probes detected are removed

Aberrant samples:
-> Aberrant samples detected in previous analyzes?
-> Removed

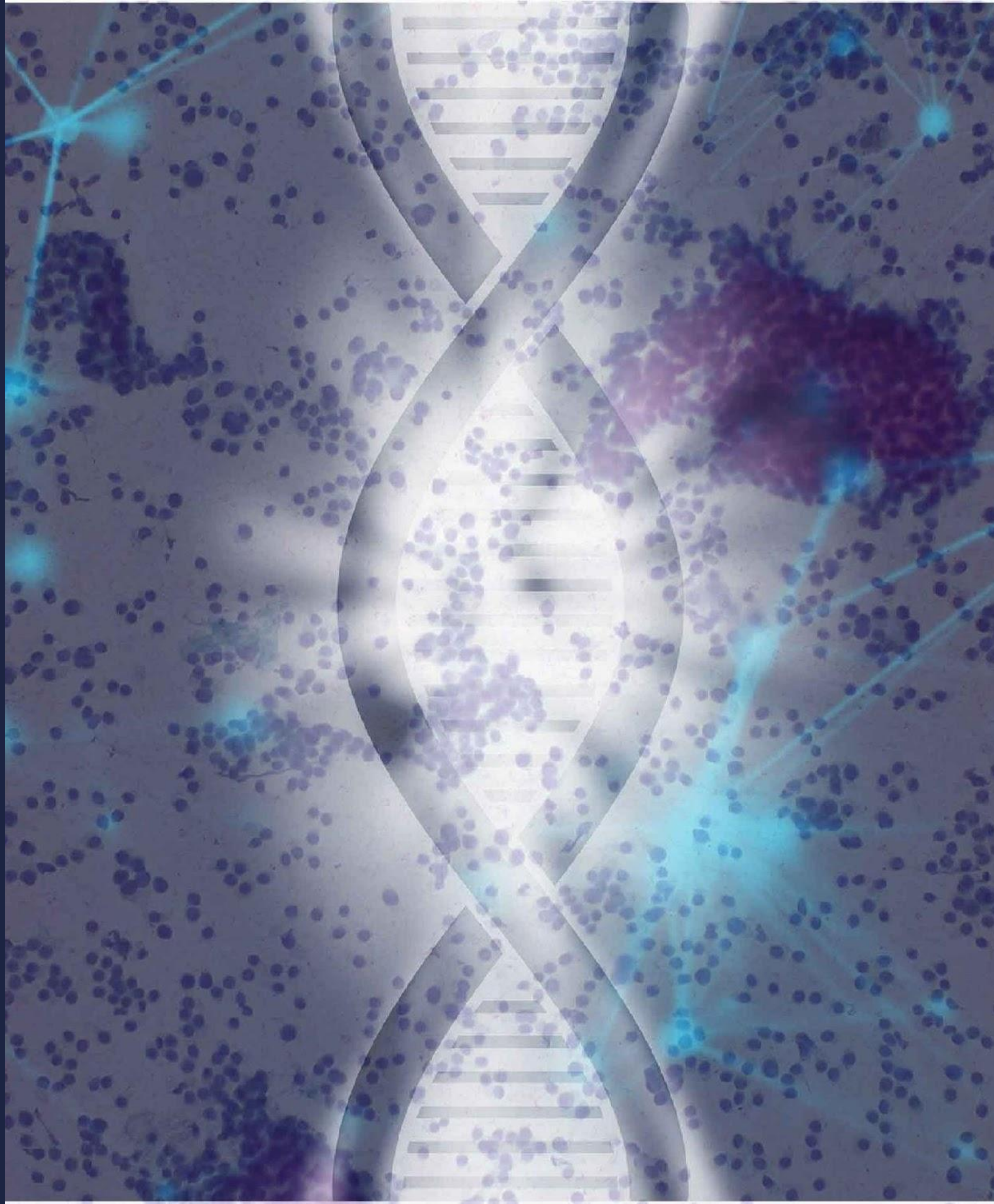
DNAm normalization using lumi

Introduction

Pre-normalization filtering:

- On probes
- On samples

Normalization with lumi

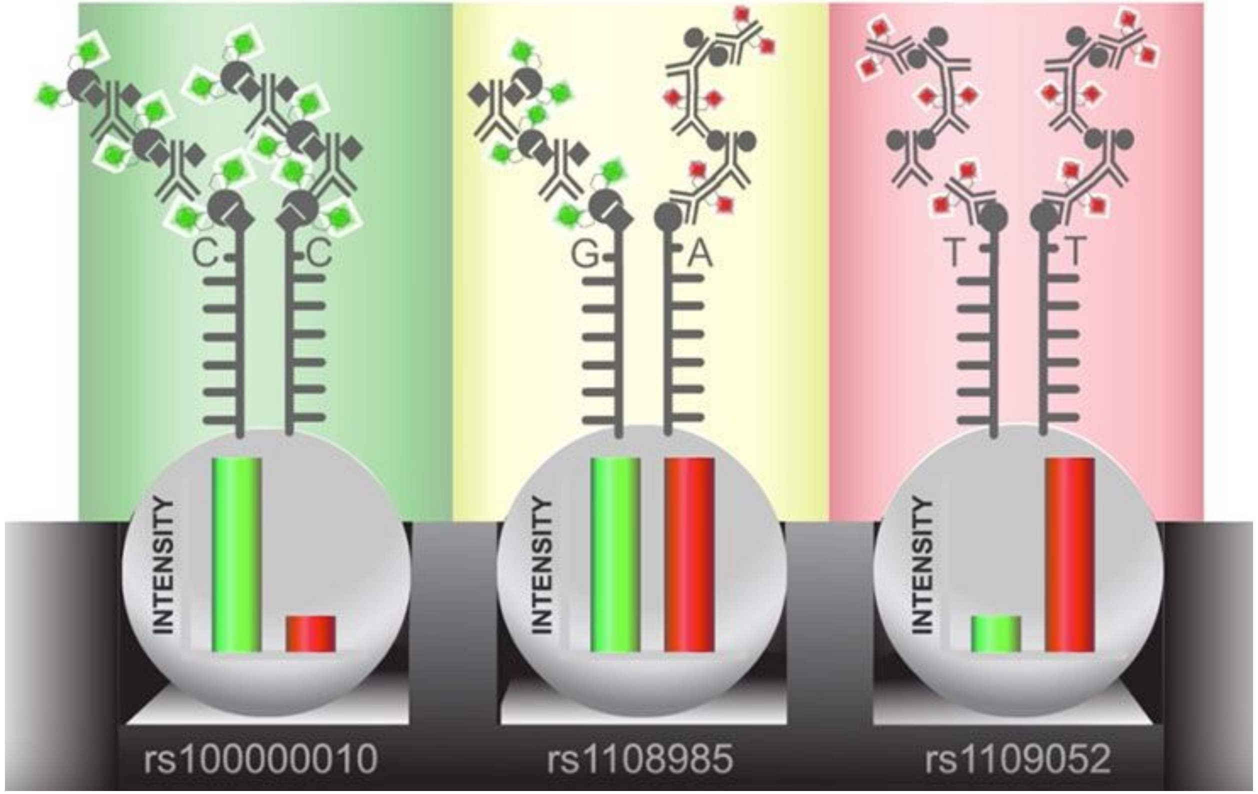


Normalization with lumi

Two steps:

- **Color balance adjustment**
  -> lumiMethyC

- **Normalization between samples**
  -> lumiMethyN



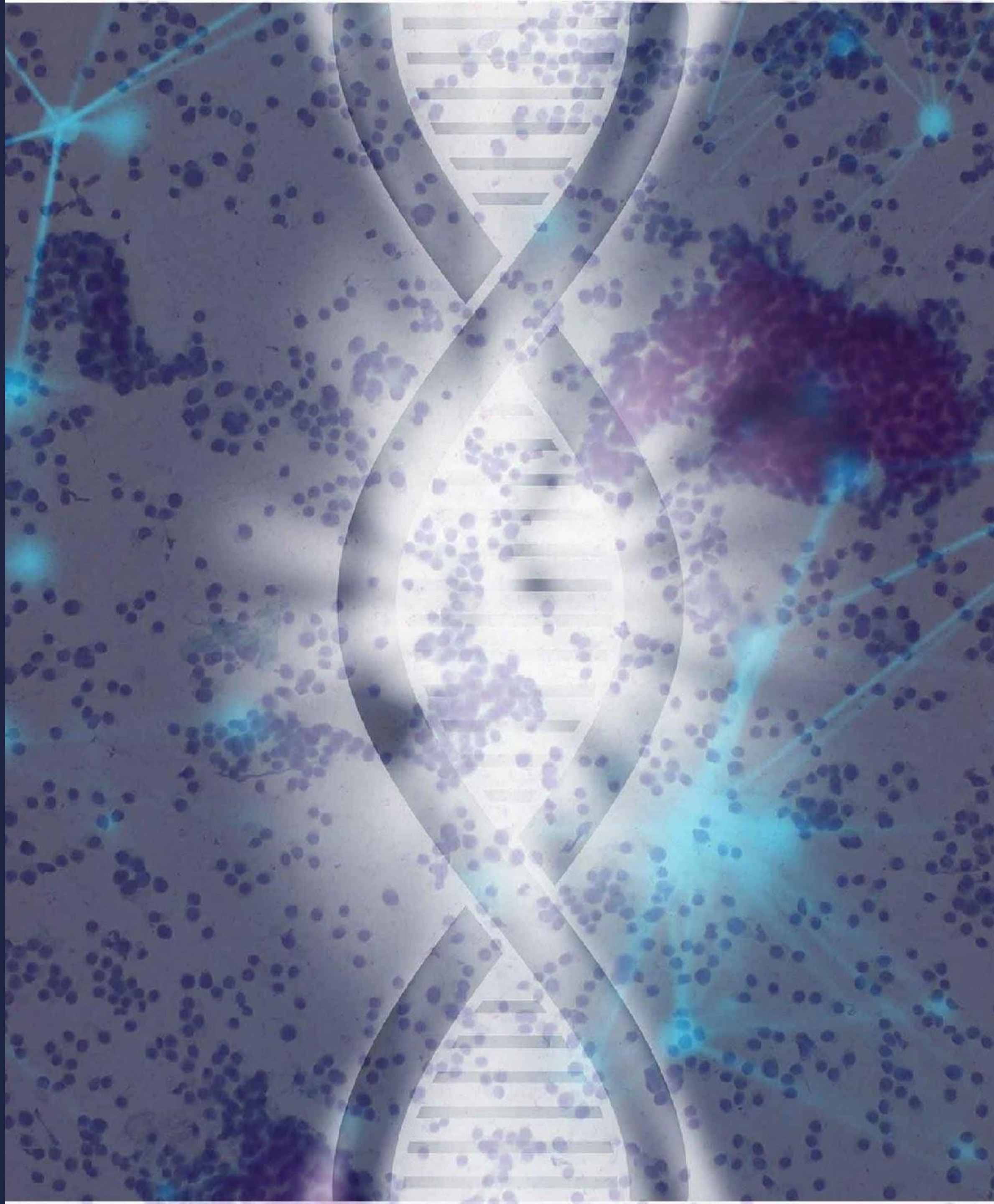
DNAm normalization using lumi

Introduction

Pre-normalization filtering:
- On probes
- On samples

Normalization with lumi

Value transformation



Value transformation

- **Beta-value**
  -> Ratio between unmethylated and methylated probes:
  0 is unmethylated, 1 is fully methylated

$$Beta_i = \frac{\max(y_{i,methy}, 0)}{\max(y_{i,unmethy}, 0) + \max(y_{i,methy}, 0) + \alpha}$$

Value transformation

- **Beta-value**
  -> Ratio between unmethylated and methylated probes:
  0 is unmethylated, 1 is fully methylated

$$Beta_i = \frac{\max(y_{i,methy}, 0)}{\max(y_{i,unmethy}, 0) + \max(y_{i,methy}, 0) + \alpha}$$

- **M-value**
  -> log2 ratio of the intensities of methylated probe versus unmethylated probe
  More statistically valid for the differential analysis of methylation levels

$$M_i = \log_2 \left(\frac{\max(y_{i,methy}, 0) + \alpha}{\max(y_{i,unmethy}, 0) + \alpha} \right)$$

[Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis | BMC Bioinformatics]

Results of the normalization

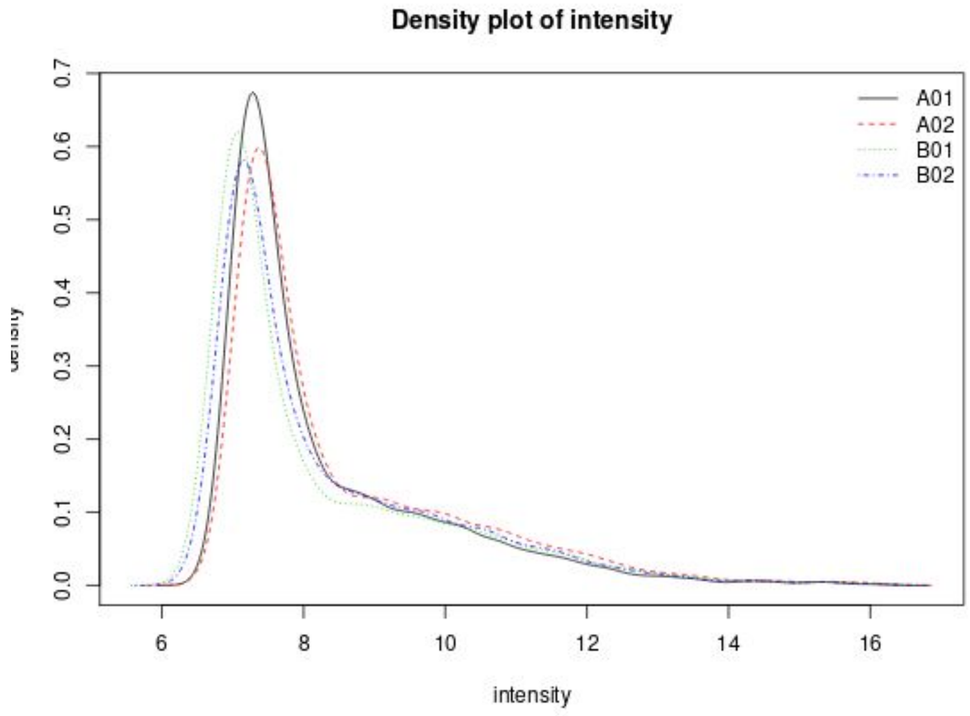


Figure 3: Density plot of Illumina microarrays before normalization

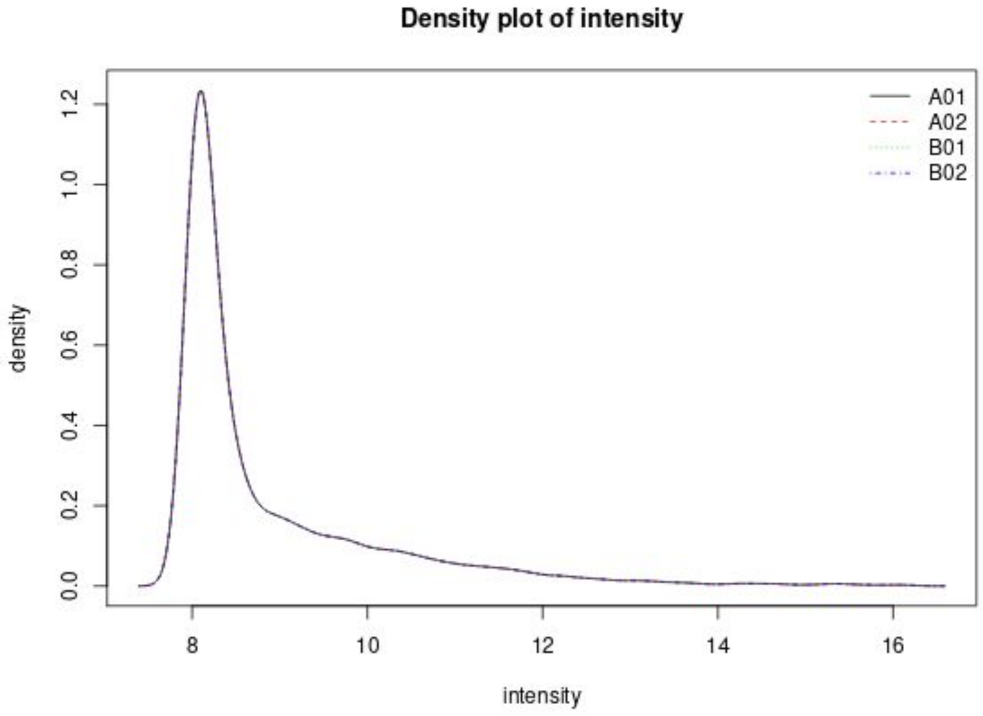


Figure 15: Density plot of Illumina microarrays after normalization

Results of the normalization

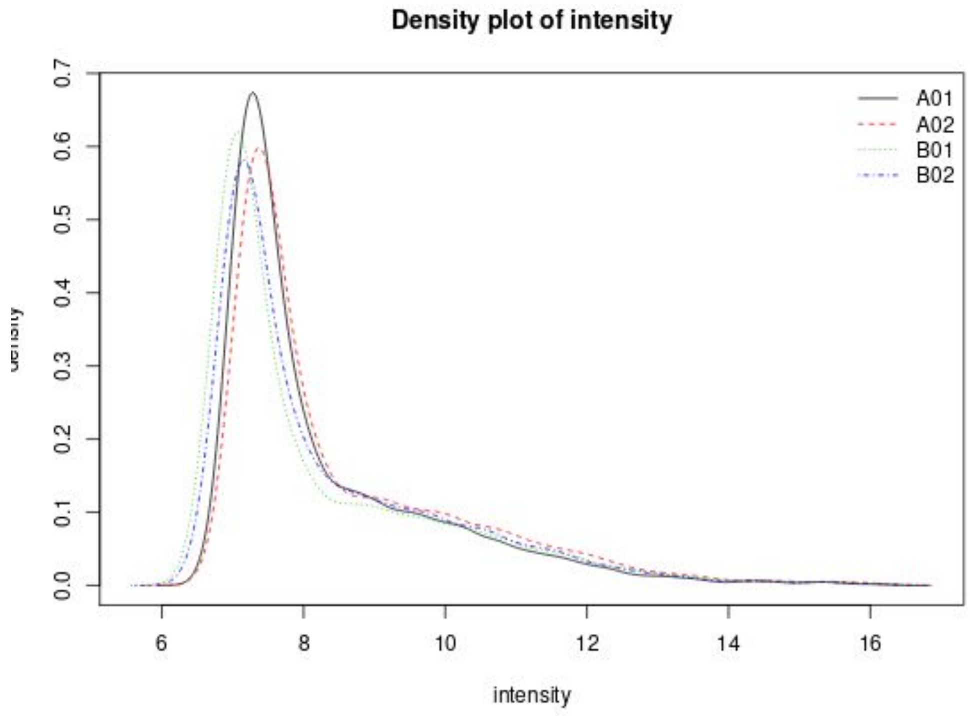


Figure 3: Density plot of Illumina microarrays before normalization

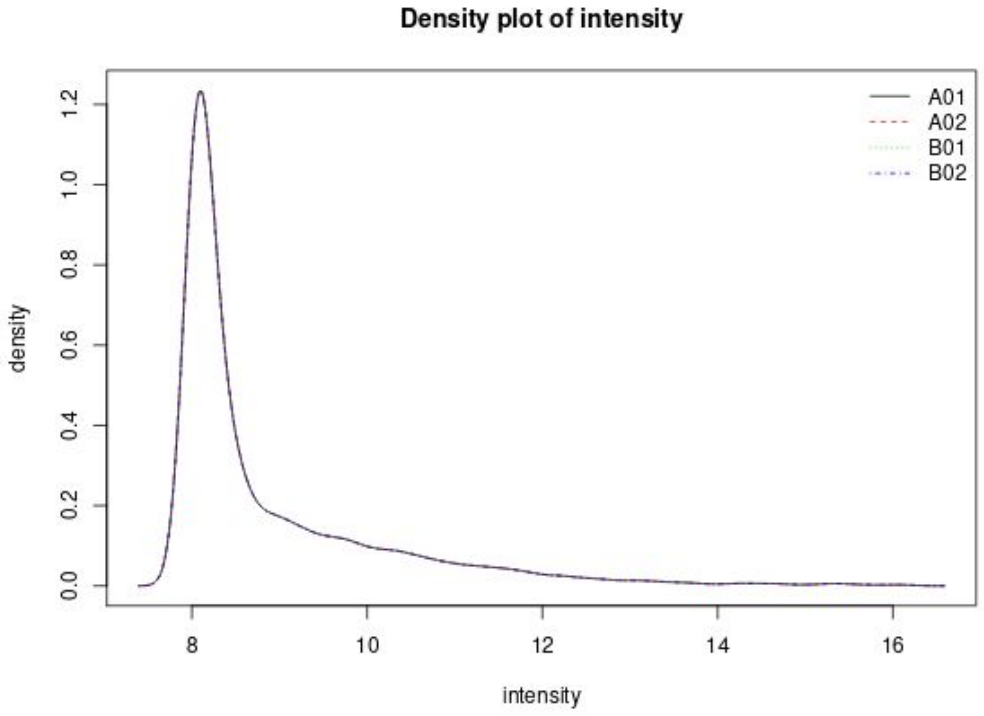


Figure 15: Density plot of Illumina microarrays after normalization

Thank you for your attention!

Yuna Blum, Ligue contre le Cancer

Jérôme Cros, APHP

Clémentine Decamps, Uni Grenoble Alpes

Carl Herrmann, Medical Faculty Heidelberg

Slim Karkar, Uni Grenoble Alpes

Yasmina Kermezli, Uni Grenoble Alpes

Magali Richard, Uni Grenoble Alpes

Ashwini Sharma, Uni Grenoble Alpes

https://cancer-heterogeneity.github.io/cometh_training.html

www.eithealth.eu | info@eithealth.eu

EIT Health is supported by the EIT, a body of the European Union