Introduction to Bioinformatics

Introduction to Omics and Application in R

Haries Ramdhani. 2021.

Data Frames

Data frames:

- are tabular data objects
- can contain different types of data inside it
- contain vector of equal length

(Follow up on) Data Frames

Data frames:

- are tabular data objects
- can contain different types of data inside it
- contain vector of equal length

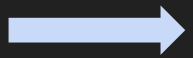
```
> data.frame(
      gender = c("Male", "Male","Female"),
      height = c(152, 171.5, 165),
      weight = c(81,93, 78),
      Age = c(42,38,26)
  )
```

	gender	height	weight	Age
1	Male	152.0	81	42
2	Male	171.5	93	38
3	Female	165.0	78	26

DPLYR

dplyr is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges

Previous subsets [ ]



- select()
- filter()
- group_by()
- summarize()
- mutate()

Installing and using DPLYR (and other packages)

Installing is simple, you just have to type the following
install.packages("name_of_the_packages")

Installing and using DPLYR (and other packages)

Installing is simple, you just have to type the following
install.packages("name_of_the_packages")

Installing packages from BioConductor is a bit different

```
if (!requireNamespace('BiocManager', quietly = TRUE))
    install.packages('BiocManager')

BiocManager::install("minfi")
```

Installing and using DPLYR (and other packages)

Installing is simple, you just have to type the following
install.packages("name_of_the_packages")

Installing packages from BioConductor is a bit different

```
if (!requireNamespace('BiocManager', quietly = TRUE))
    install.packages('BiocManager')

BiocManager::install("minfi")
```

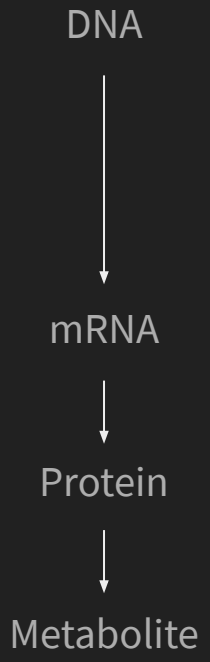
To load the installed library we can use
library(“package_name”)

DPLYR

Common functions in DPLYR

- select() : select our intended column
- filter() : filter the data by some conditions
- group_by() : group the data by column
- summarize() : summarize the results of grouping
- mutate() : mutate a column, create new column with new logic

A Quick Glimpse on Omics



A Quick Glimpse on Omics

Genomics	DNA
Epigenomics	Epigenetic Modification
Transcriptomics	mRNA
Proteomics	Protein
Metabolomics	Metabolite

A Quick Glimpse on Omics

Genomics	DNA	Exploring Variants
Epigenomics	Epigenetic Modification	Quantifying Epigenetic Modification
Transcriptomics	mRNA	Gene Expression
Proteomics	Protein	Post Translational Modifications
Metabolomics	Metabolite	Determining phenotypes

A Quick Glimpse on Omics

Genomics	DNA
Epigenomics	Bisulfite treated DNA, Immunoprecipitation
Transcriptomics	mRNA -> cDNA
Proteomics	Protein
Metabolomics	Metabolite

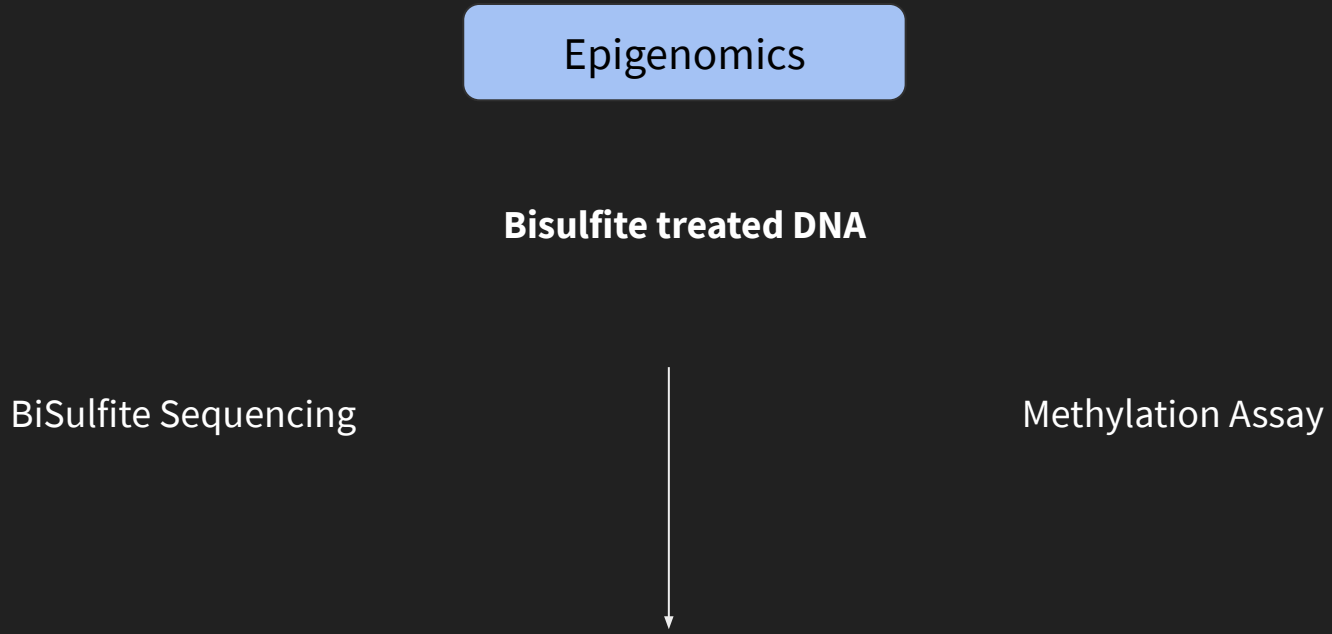
A Quick Glimpse on Omics

Genomics	DNA	DNA-Seq (WGS, WES)
Epigenomics	Bisulfite treated DNA, Immunoprecipitation	BiSulfite Sequencing, EpicArray, ChipSEQ
Transcriptomics	mRNA -> cDNA	RNA-Seq, Expression Microarray
Proteomics	Protein	Protein Microarray, Spectometry
Metabolomics	Metabolite	Liquid Chromatography

A Quick Glimpse on Omics

Genomics	DNA	DNA-Seq (WGS, WES)
Epigenomics	**Bisulfite treated DNA**	**BiSulfite Sequencing, EpicArray**
Transcriptomics	mRNA -> cDNA	RNA-Seq, Expression Microarray
Proteomics	Protein	Protein Microarray, Spectometry
Metabolomics	Metabolite	Liquid Chromatography

Epigenomics



Methylation Array Data

Samples

Array

Data

Methylation Array Data

- For each CpG sites, there are two things that we measure:
  - Unmethylated Intensity (U)
  - Methylated Intensity (M)

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Methylation Array Data

- For each CpG sites, there are two things that we measure:
  - Unmethylated Intensity (U)
  - Methylated Intensity (M)
- These value then can be used to determine the methylation levels
  - M-value ($\log_2(M/U)$)
  - Beta- value (M / (M+ U)) => percentage

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Methylation Array Data

- For each CpG sites, there are two things that we measure:
  - Unmethylated Intensity (U)
  - Methylated Intensity (M)
- These value then can be used to determine the methylation levels
  - M-value ($\log_2(M/U)$)
  - Beta- value (M / (M+ U))
- Beta values are generally preferable for describing the level of methylation at a locus or for graphical presentation because percentage methylation is easily interpretable. However, due to their distributional properties, M-values are more appropriate for statistical testing (Du et al. 2010)

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Methylation Array Data

- Illumina raw data files are usually either in plain text or binary format.
- The binary "IDAT" files (stands for "intensity data file") are generated by the scanner and can be parsed using R/BioConductor packages such as illuminaio)
- You can find the data for training in GEO Omnibus usually written under the ***Methylation profiling by array*** tag

Let’s Remember the Previous Lesson

Pipeline Summary

The pipeline allows you to choose between running either Bismark or bwa-meth / MethylDackel. Choose between workflows by using `--aligner bismark` (default, uses bowtie2 for alignment), `--aligner bismark_hisat` or `--aligner bwameth`.

Step	Bismark workflow	bwa-meth workflow
Generate Reference Genome Index *(optional)*	Bismark	bwa-meth
Raw data QC	FastQC	FastQC
Adapter sequence trimming	Trim Galore!	Trim Galore!
Align Reads	Bismark	bwa-meth
Deduplicate Alignments	Bismark	Picard MarkDuplicates
Extract methylation calls	Bismark	MethylDackel
Sample report	Bismark	-
Summary Report	Bismark	-
Alignment QC	Qualimap	Qualimap
Sample complexity	Preseq	Preseq
Project Report	MultiQC	MultiQC

For array data it’s a bit shorter

Main flow:

- Quality control
- Filtering
- Normalization
- Data exploration

Downstream Analysis:

- Probe wise differential methylation analysis
- Differential variability analysis
- GO analysis
- etc.

Loading the data

Before loading the data you have to import the needed libraries in R

- **Minfi** : provides tools for analyzing Illumina's Methylation arrays, specifically the 450k and EPIC
- **IlluminaHumanMethylation450kanno.ilmn12.hg19** : Annotation for Illumina's 450k methylation arrays
- **IlluminaHumanMethylation450kmanifest** : Annotation for Illumina's 450k methylation arrays
- **missMethyl** : normalization, removal of unwanted variation in differential methylation analysis, differential variability testing and gene set analysis

Loading the data

Before loading the data you have to import the needed libraries in R

- **Minfi** : provides tools for analyzing Illumina's Methylation arrays, specifically the 450k and EPIC
- **IlluminaHumanMethylation450kmanifest** : Annotation for Illumina's 450k methylation arrays
- **missMethyl** : normalization, removal of unwanted variation in differential methylation analysis, differential variability testing and gene set analysis

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

How do I Install those packages?

Installing is simple, you just have to type the following
install.packages("name_of_the_packages")

How do I Install those packages?

Installing is simple, you just have to type the following
install.packages("name_of_the_packages")

Installing packages from BioConductor is a bit different

```
if (!requireNamespace('BiocManager', quietly = TRUE))
    install.packages('BiocManager')

BiocManager::install("minfi")
```

How do I Install those packages?

Installing is simple, you just have to type the following
install.packages("name_of_the_packages")

Installing packages from BioConductor is a bit different

```
if (!requireNamespace('BiocManager', quietly = TRUE))
    install.packages('BiocManager')

BiocManager::install("minfi")
```

To load the installed library we can use
library(“package_name”)

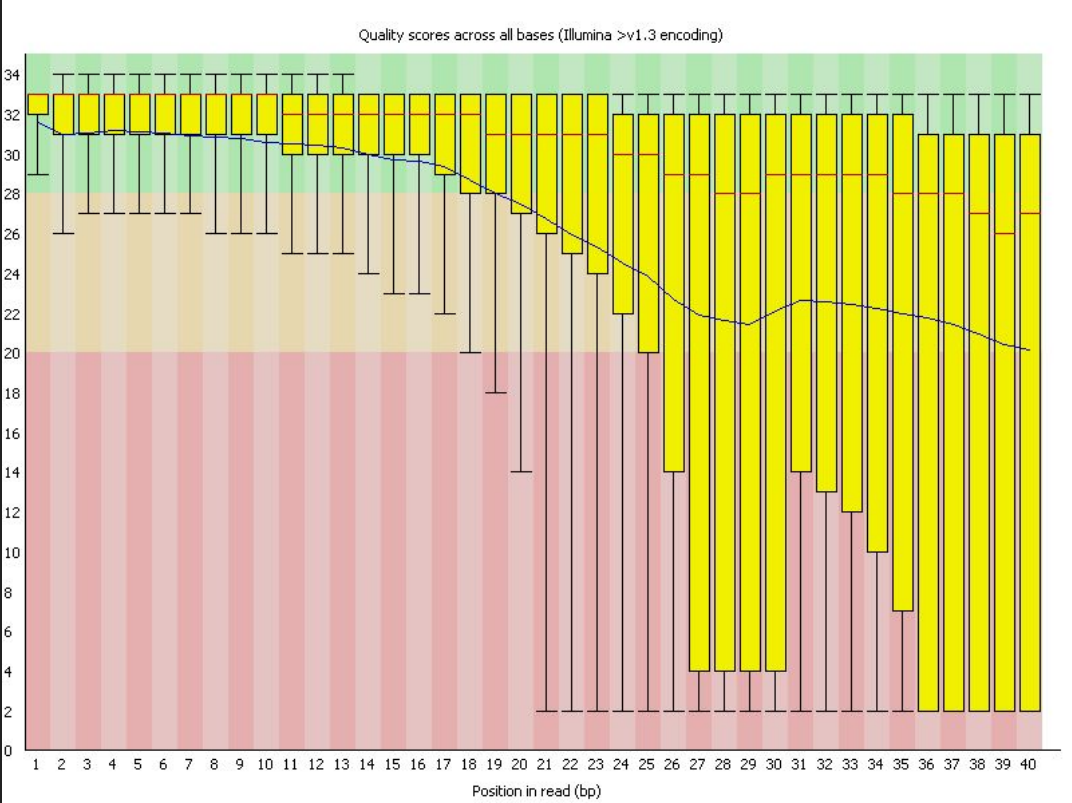
Quality Control

- Some samples may contain outlier

Quality Control

- Some samples may contain outlier
- Similar to the NGS one in quality control we want to measure the signal detection for each sample

Quality Control in NGS



More information on interpreting:

https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html

Quality Control

- Some samples may contain outlier
- Similar to the NGS one in quality control we want to measure the signal detection for each sample
- We calculate the p-value of each probe

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Quality Control

- Some samples may contain outlier
- Similar to the NGS one in quality control we want to measure the signal detection for each sample
- We calculate the p-value of each probe
- Very small p-values are indicative of a reliable signal whilst large p-values, for example >0.01, generally indicate a poor quality signal.

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Quality Control

- Some samples may contain outlier
- Similar to the NGS one in quality control we want to measure the signal detection for each sample
- We calculate the p-value of each probe
- Very small p-values are indicative of a reliable signal whilst large p-values, for example >0.01, generally indicate a poor quality signal.
- The method used by minfi to calculate detection p-values compares the total signal (M+U) for each probe to the background signal level

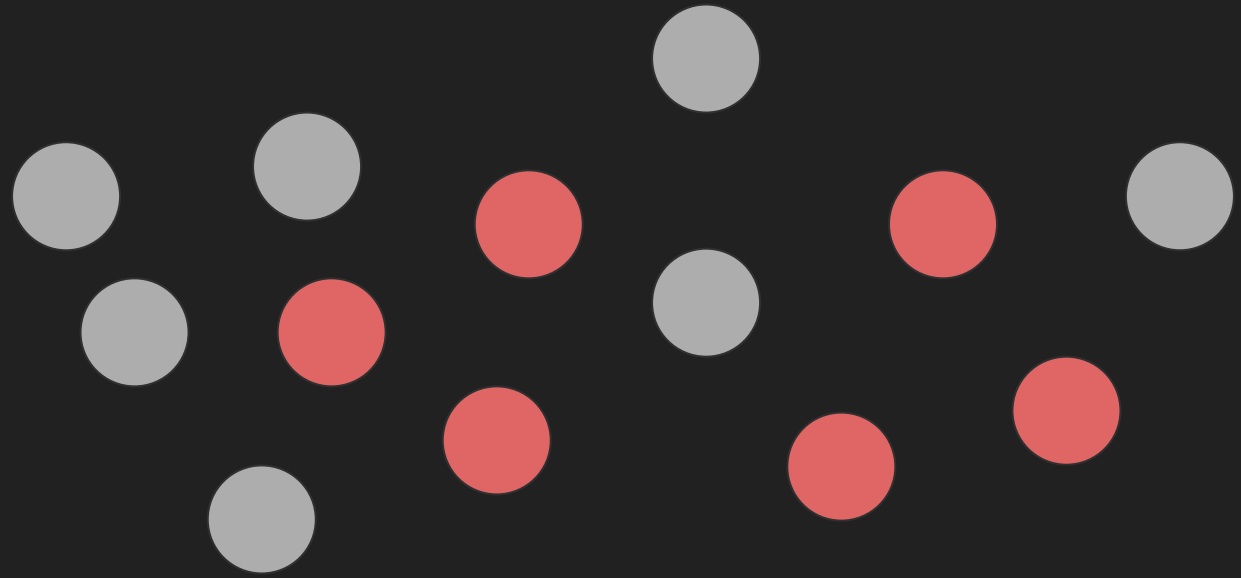
Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Quality Control

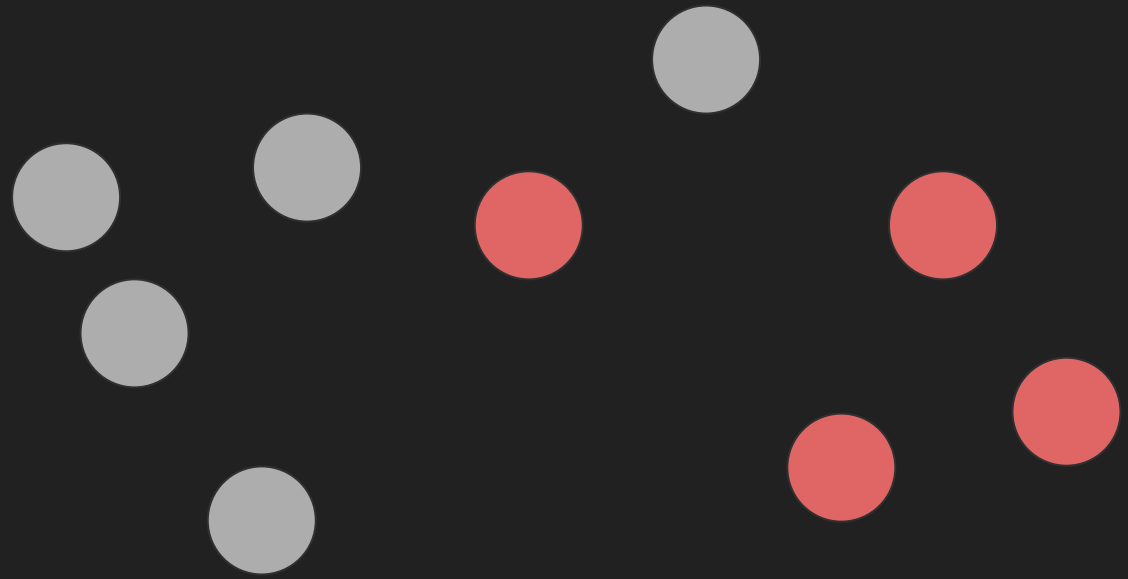
- Some samples may contain outlier
- Similar to the NGS one in quality control we want to measure the signal detection for each sample
- We calculate the p-value of each probe
- Very small p-values are indicative of a reliable signal whilst large p-values, for example >0.01, generally indicate a poor quality signal.
- The method used by minfi to calculate detection p-values compares the total signal (M+U) for each probe to the background signal level
- Poor quality samples can be easily excluded from the analysis using a detection p-value cutoff, for example >0.05

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Quality Control



Quality Control



Normalisation

- Normalisation is being done to minimise the unwanted variation within and between samples. Why?

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Normalisation

- Normalisation is being done to minimise the unwanted variation within and between samples. Why?
- Many different types of normalisation have been developed for methylation arrays

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Normalisation

- Normalisation is being done to minimise the unwanted variation within and between samples. Why?
- Many different types of normalisation have been developed for methylation arrays
- Several methods have been built into minfi and can be directly applied within its framework

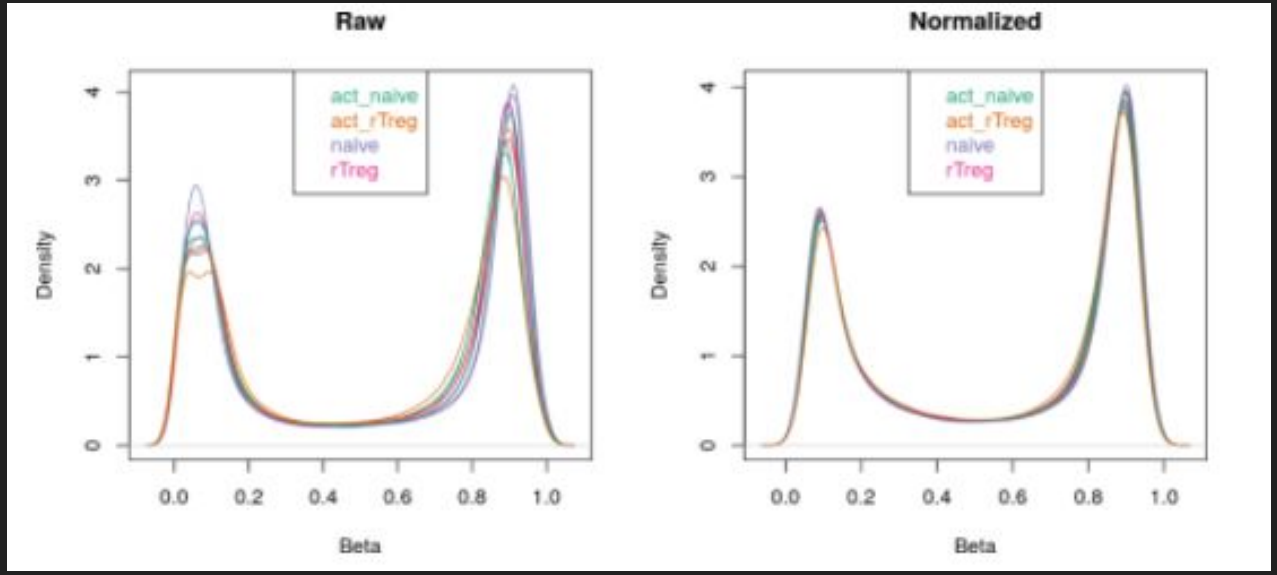
Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Normalisation

- Normalisation is being done to minimise the unwanted variation within and between samples. Why?
- Many different types of normalisation have been developed for methylation arrays
- Several methods have been built into minfi and can be directly applied within its framework
- There is no BEST normalisation method, but a recent study by Fortin et al. (2014) has suggested that a good rule of thumb within the minfi framework to use:
  - preprocessFunnorm (for datasets with global methylation differences such as cancer/normal or vastly different tissue types,
  - whilst the preprocessQuantile function (Touleimat and Tost 2012) is more suited for datasets where you do not expect global differences between your samples, for example a single tissue.

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Normalisation



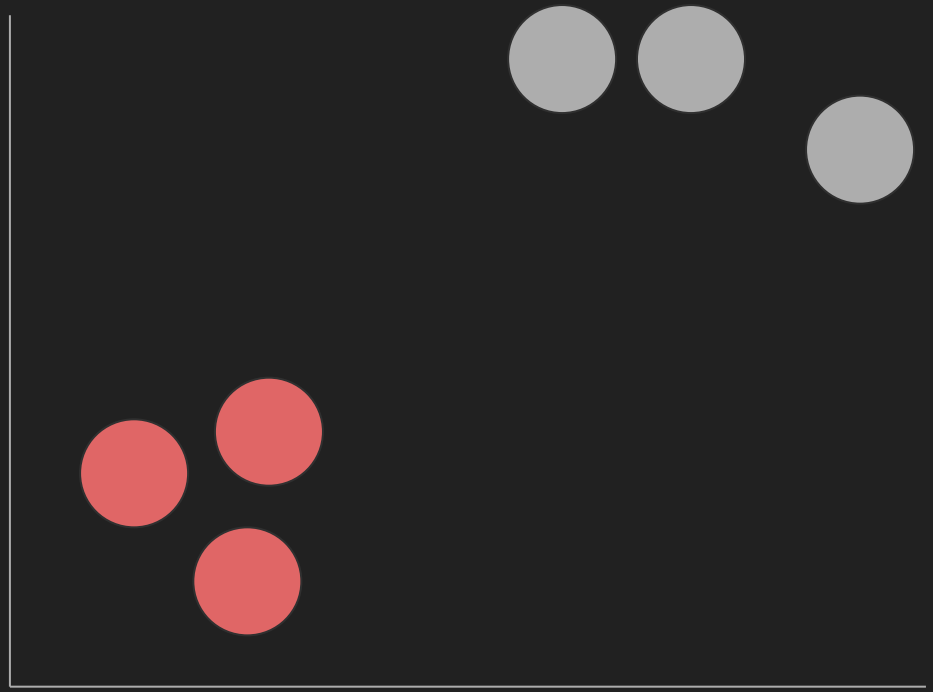
Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Data Exploration

- We perform data exploration in case we want to explore more about our array data
- Maybe this way we can get new insights or new information regarding something that is not seen by our eyes (through data viz)
- Most common approach:
    - MDS (Multidimensional Scaling) : Look similarity between samples
        - Samples that are very similar should be clustered together

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

Data Exploration



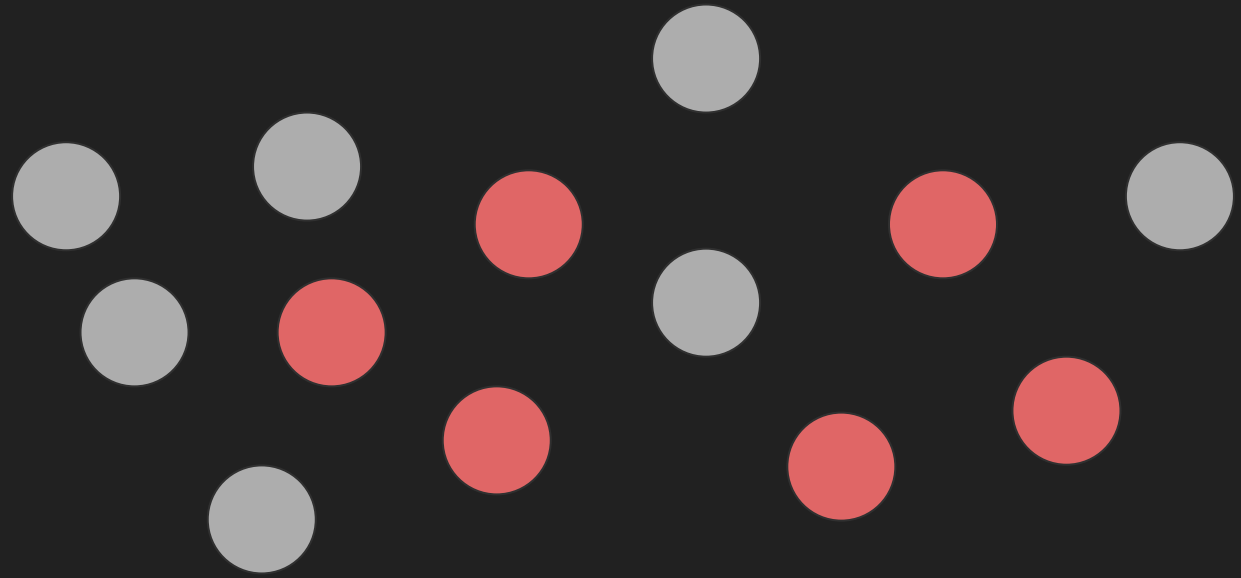
Data Exploration

Filtering

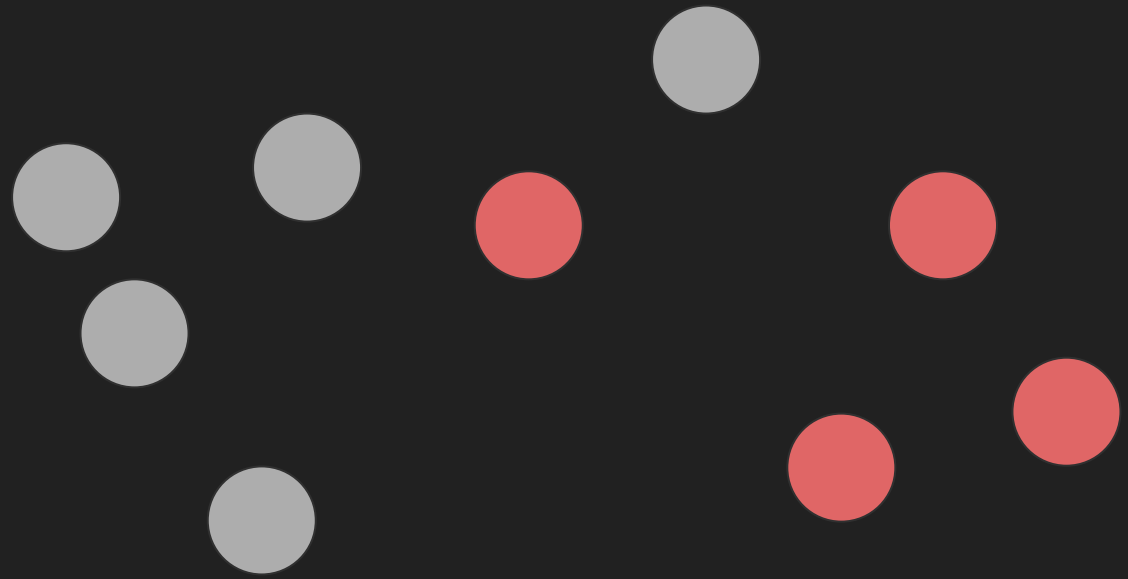
- Poor performing probes are generally filtered out prior to differential methylation analysis.
- As the signal from these probes is unreliable, by removing them we perform fewer statistical tests and thus incur a reduced multiple testing penalty.
- We filter out probes that have failed in one or more samples based on detection p-value.

Maksimovic et al. 2020. A cross-package Bioconductor workflow for analysing methylation array data.

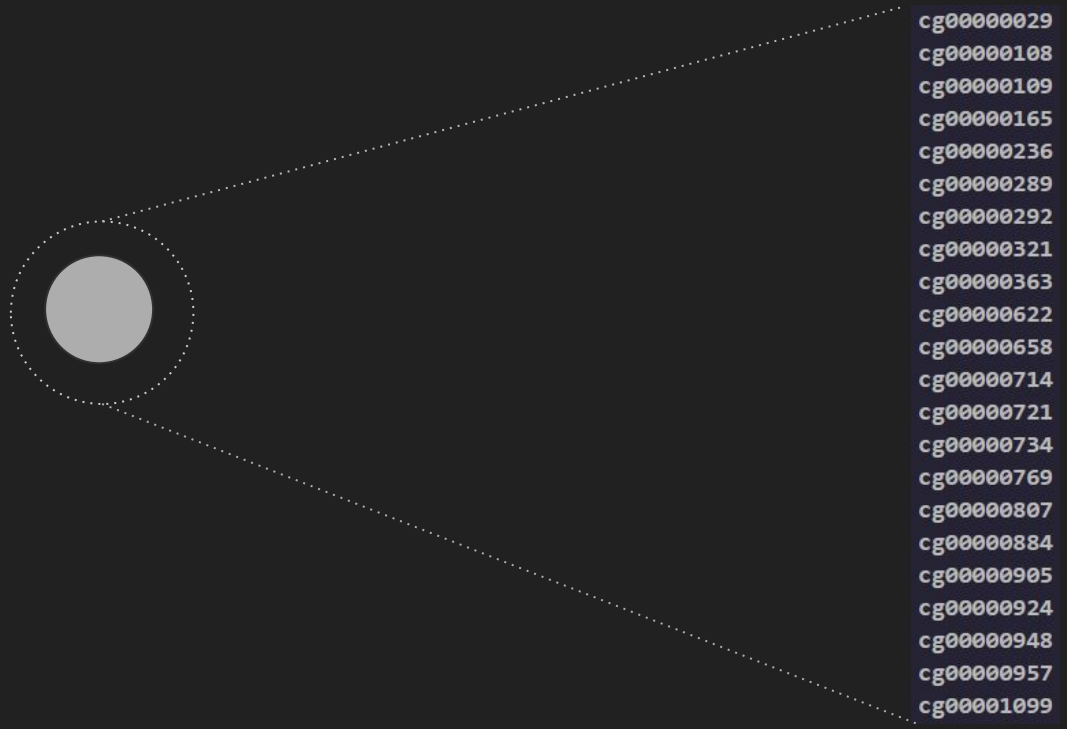
Quality Control



Quality Control



Filtering


cg00000029
cg00000108
cg00000109
cg00000165
cg00000236
cg00000289
cg00000292
cg00000321
cg00000363
cg00000622
cg00000658
cg00000714
cg00000721
cg00000734
cg00000769
cg00000807
cg00000884
cg00000905
cg00000924
cg00000948
cg00000957
cg00001099