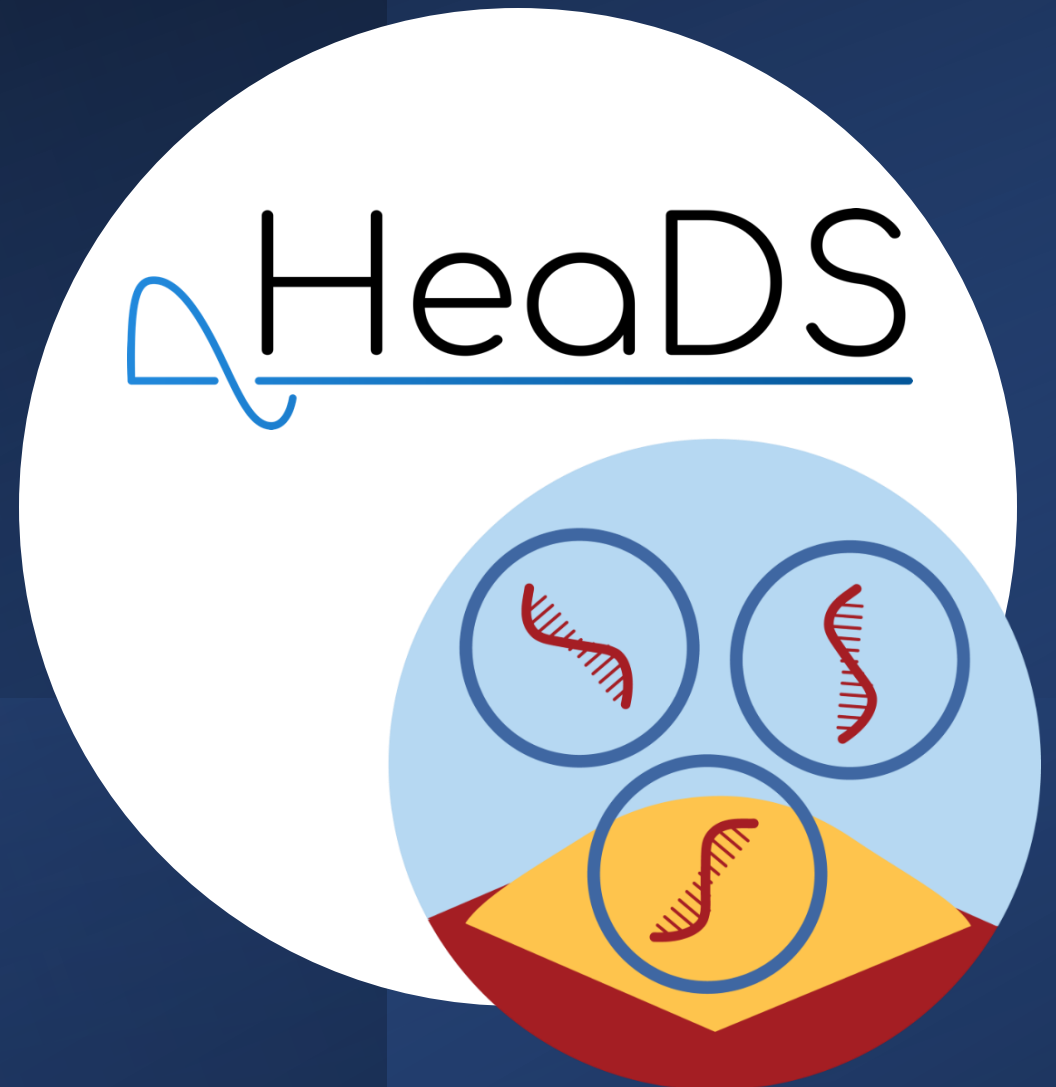


# RNAseq analysis

Center for Health Data Science



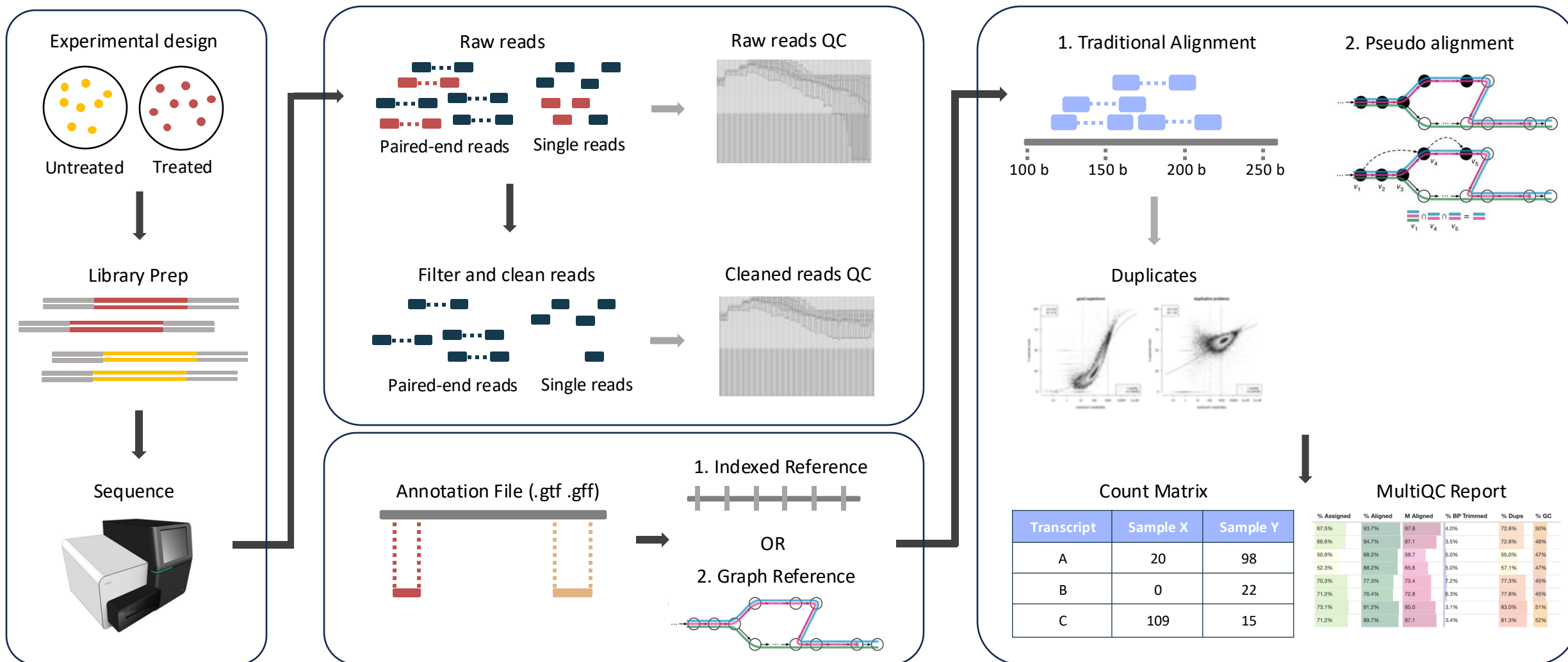
Health Data Science Sandbox

# Upstream Analysis

- **Goal of a bulk RNAseq analysis** is to **compare gene expression** between two or more conditions.
- **Upstream Analysis:**
  - mRNA molecules are **sequenced (reads)** .
  - Reads are **trimmed** to remove experimental artifacts, i.e. primers, barcodes, poor base quality
  - Reads are separated by sample of origin (demultiplexing)
  - **Quality checks (QC)** of reads
  - The reads are **mapped/aligned** to an appropriate **reference genome and annotated**.
  - Quality check of alignment
  - Aligned reads are quantified to counts per gene and sample (count matrix)
- All these **steps are compiled into a pipeline**: nf-core rnased
  - Read more about how to use the nf-core rnaseq pipeline [here](#)

# Upstream Analysis

nf-core/rnaseq<sup>24</sup>



# Downstream Analysis

- We can use R, Python or similar to perform downstream analysis of count matrices
- **Downstream Analysis:**
  - Model distribution of counts
  - Normalization (library size, RNA comp.) and transformation (vst, rlog)
  - Exploratory Data Analysis (PCA, Clustering, ...)
  - Differential Expression Analysis (gene-wise models and post hoc test)
  - Visualization of DE genes (heatmap, volcano, ...)
  - Functional Analysis (GSEA, Pathways, ...)
  - *More... Ranking, simple ML models, co-expression analysis*

# Downstream Analysis



Count Matrix

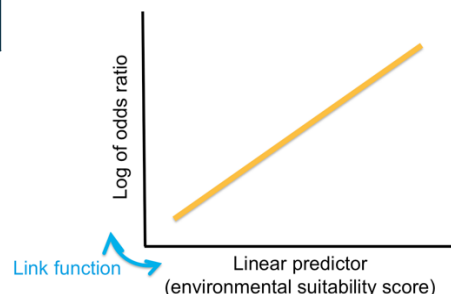
Transcript	Sample X	Sample Y
A	20	98
B	0	22
C	109	15

Meta Data

ID	Treatment	Smoke	Status
ID1	trt	yes	0
ID2	untrt	yes	1
ID3	trt	no	0

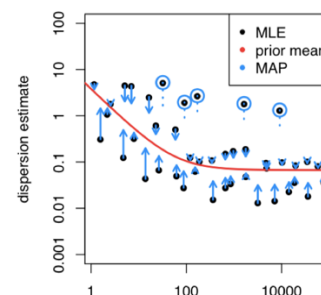
## Modelling

$$Exp_A = \sim Smoke + Treatment$$



## Normalization

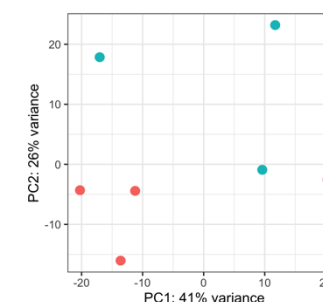
`estimateSizeFactors(dds)`  
`estimateDispersions(dds)`



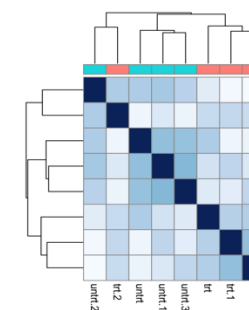
## Exploratory Data Analysis

Transformation  
`vst(dds)`  
`rlog(dds)`

PCA plot



Heatmap



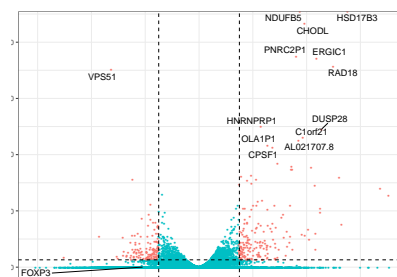
## Differential Expression

Post hoc test:

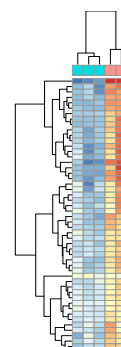
`nbinomWaldTest(dds)`

	LogFC	Padj
Gene A	1.9	0.001
Gene B	-1.1	0.02
Gene C	0.2	0.32

LFCSHrink



Volcano plot

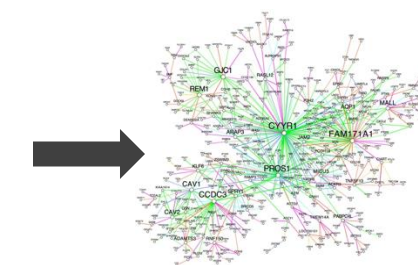


DE Heatmap

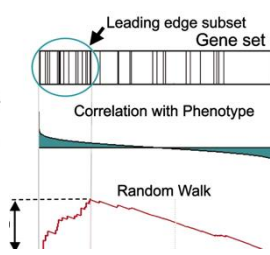
## Functional analysis

Annotated results table

Gene	Entrez ID	LogFC	Padj
Gene A	5506	1.9	0.001
Gene B	63421	-1.1	0.02
Gene C	9932	0.2	0.32



Networks



GSEA



# Ucloud setup

- UCloud is a danish High Performance Computing environment
  - Lots of storage, lots of cpus and RAM (computing power)
- Danish institutions have access to it
  - You personally have 1000dkk in computing resources
- UCloud works in apps, giving you access to different programs
  - All apps have documentation on how to use them!
- This means everyone is using the same versions of software
  - Makes teaching much much easier, results are reproducible