Bioinformatics Core
Institute of Molecular Biology

RNA Sequencing Data Analysis

Bioinformatics Training Course
11/03/2021

余承欣 (Albert)
albertchyu@gate.sinica.edu.tw
IMB Bioinformatics Core

Bioinformatics Core
Institute of Molecular Biology

Consultation

Experimental Design

Statistical Consultation

Software/Database Support

Routine Analysis

Genome Assembly

Genome Variant Calling

Bulk RNA-Seq

Single-Cell RNA-Seq

Methylation (BS-Seq)

ChIP-Seq and ATAC-Seq

Protein structure prediction

Functional enrichment

Custom Analysis

Custom NGS Analyses

Functional Analysis

Proteomics Analysis

Metabolomics Analysis

Protein Structural Analysis

Image Analysis

Others

余承欣 (Albert), Ph.D.

林信男 (Arith), Ph.D.

葉坤海 (Eason), M.S.

劉謹禎 (Vicky), M.S.

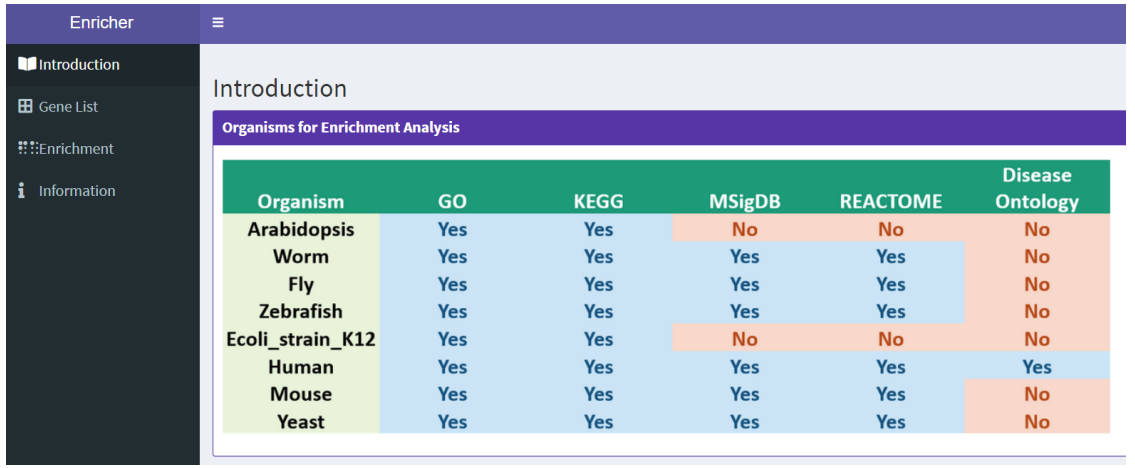
洪鈺荏 (Elly), M.S.

Bioinformatics Core
Institute of Molecular Biology

BBSC Server Portal

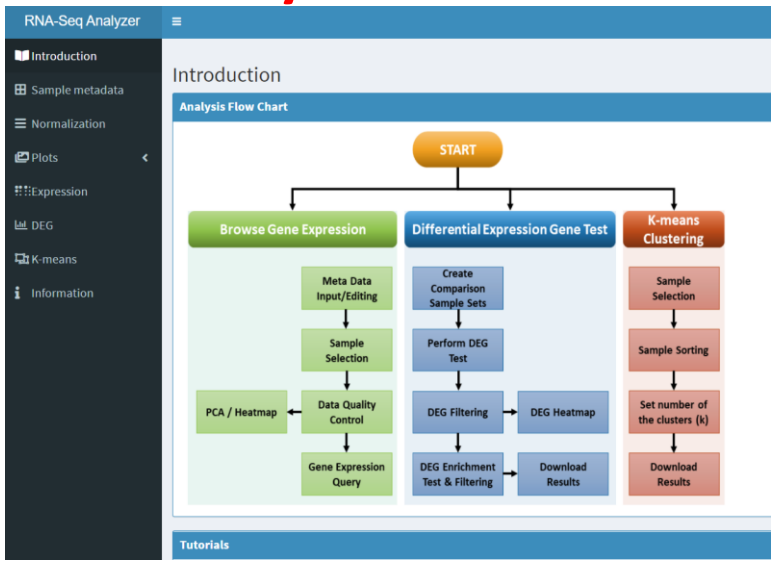
```
****** Welcome to BioInformatics Core Portal ******
    1. RNA-Seq data analysis.
    2. De novo sequence assembly.
    3. Nanopore data manipulation.
    4. Genomic variant calling.
    5. 10x data analysis.
    6. protein data analysis.
    X. Exit.

Your choice:
```

Enricher

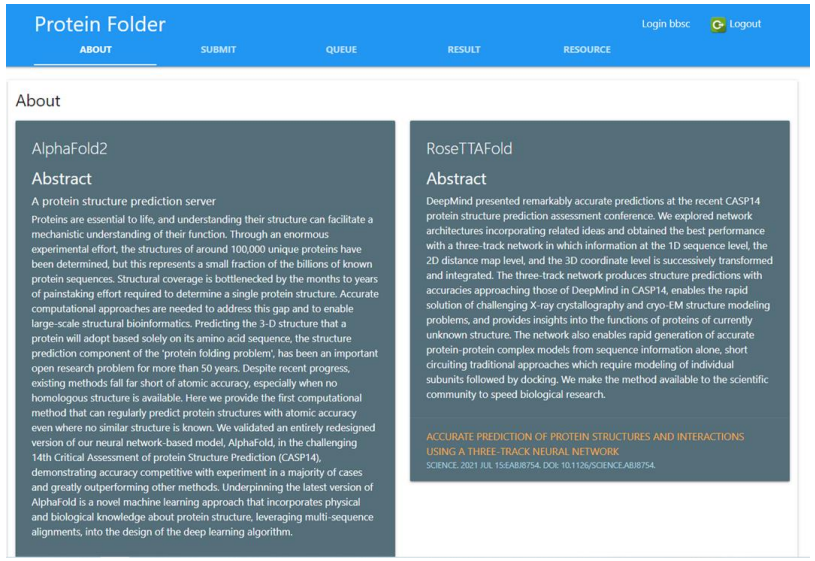


Organism	GO	KEGG	MSigDB	REACTOME	Disease Ontology
Arabidopsis	Yes	Yes	No	No	No
Worm	Yes	Yes	Yes	Yes	No
Fly	Yes	Yes	Yes	Yes	No
Zebrafish	Yes	Yes	Yes	Yes	No
Ecoli_strain_K12	Yes	Yes	No	No	No
Human	Yes	Yes	Yes	Yes	Yes
Mouse	Yes	Yes	Yes	Yes	No
Yeast	Yes	Yes	Yes	Yes	No

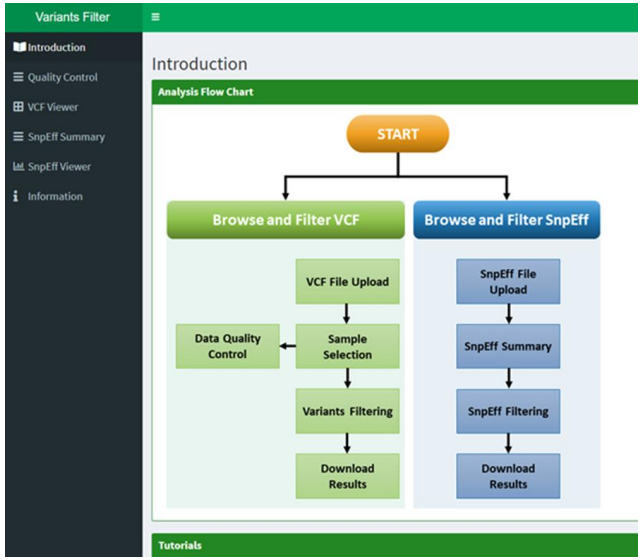
RNA Analyzer



Protein Folder



Variant Filter



Computing & Storage Resources

Server	CPU Type	CPU cores	GPU	Memory
bbsc197	AMD Ryzen 3990X	128	Nvidia GTX 3080 Ti	256 GB
bbsc198	AMD Ryzen 3990X	256	Nvidia RTX-3090	512 GB
bbsc199	Intel Xeon E7	120	NA	512 GB
bbsc200	AMD Ryzen 3990X	128	Nvidia RTX-2070	256 GB

Server	System	Storage	Memory
bbsc204	Qnap	127 TB	32 GB
bbsc205	Qnap	190 TB	32 GB

Core Website

https://bc.imb.sinica.edu.tw/

Online Service System

https://bc.imb.sinica.edu.tw/BioInfoPortal/

Core Project management

https://bc.imb.sinica.edu.tw/ProjectViewer/

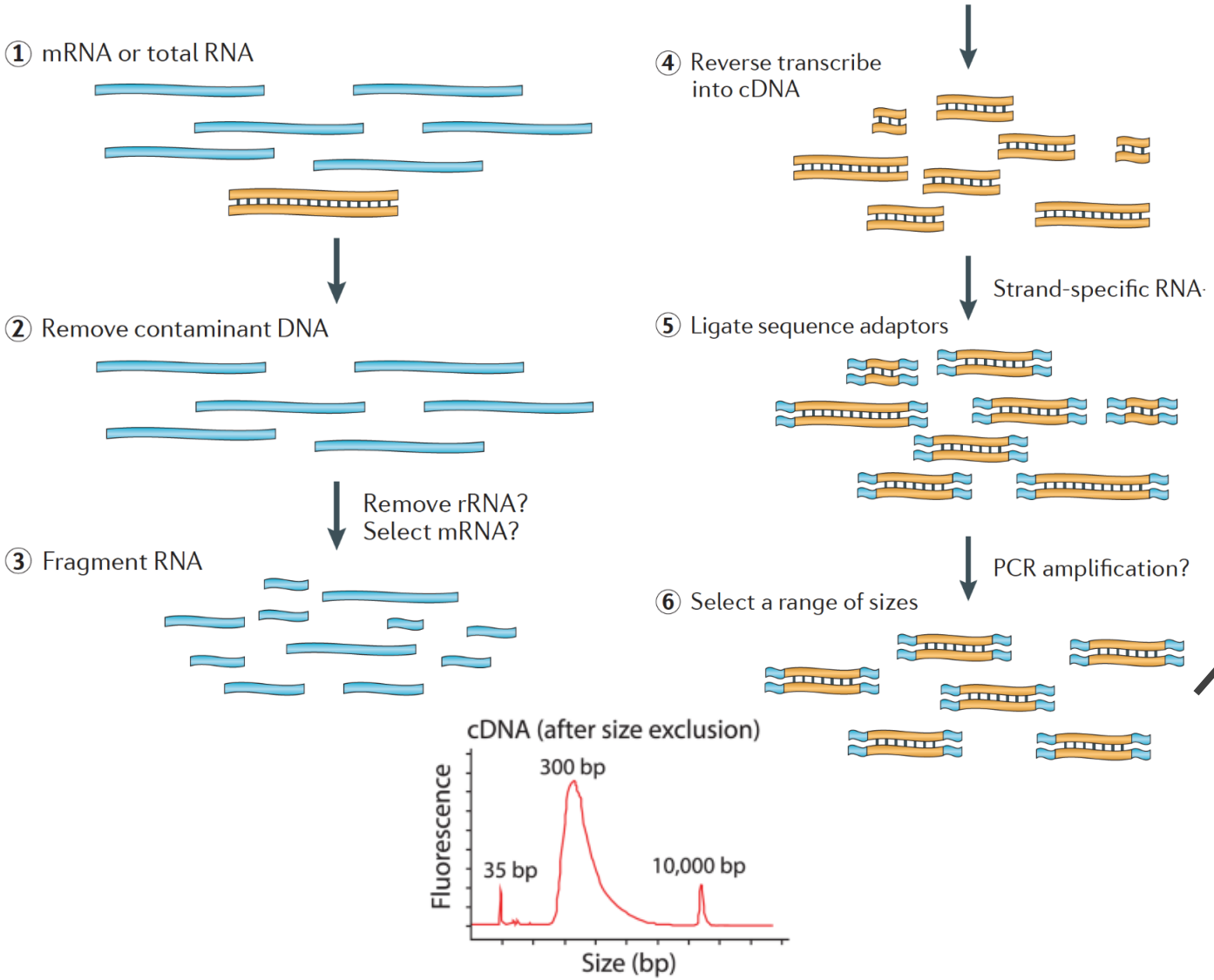
Outline

1. Experimental design and practical considerations

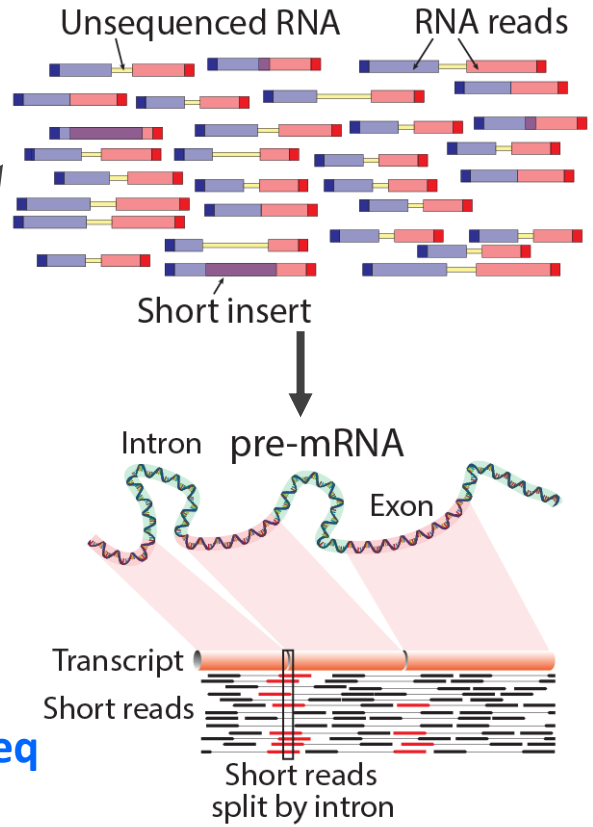
2. Differential gene expression analysis pipeline

3. IMB Bioinformatics Core analysis tools (DEMO)

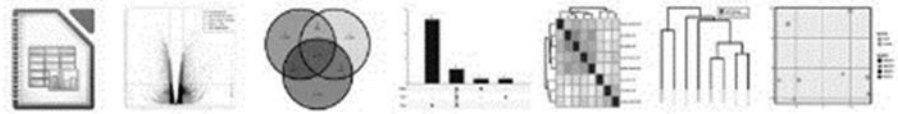
From RNA -> sequence data



Short-read: Illumina HiSeq
Long-read: Nanopore

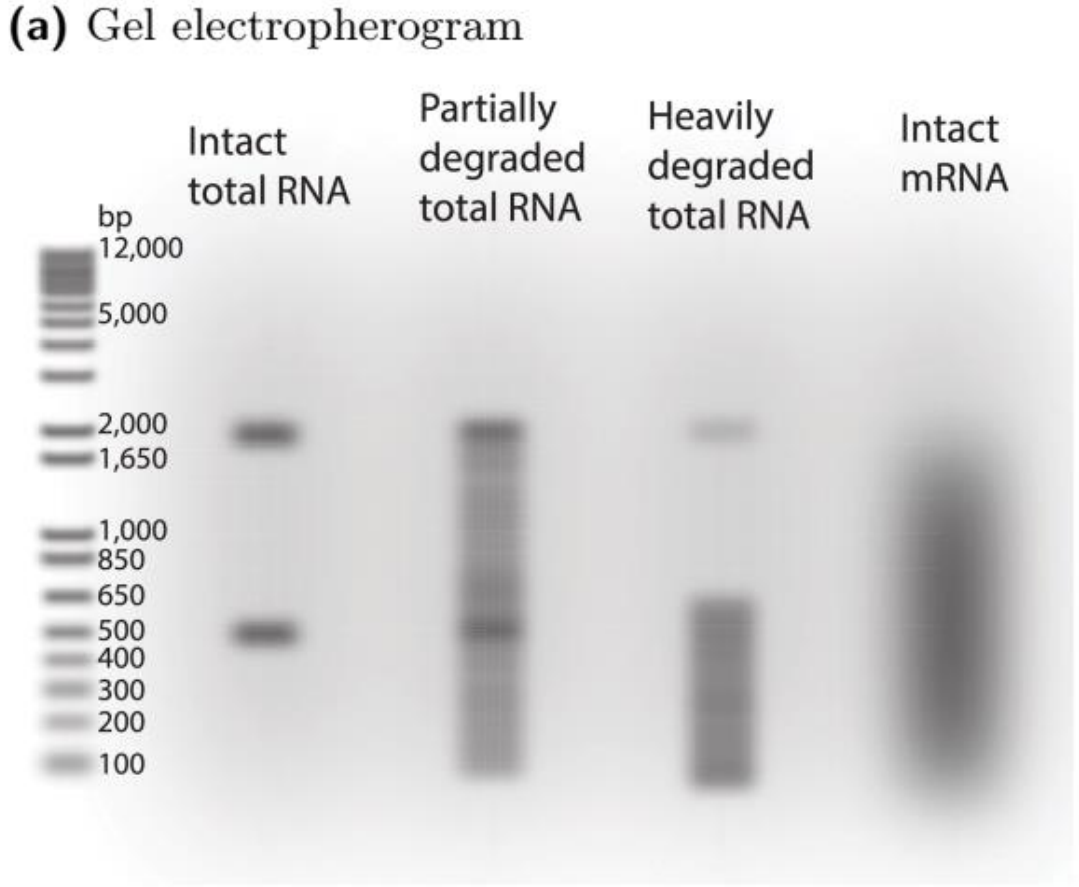


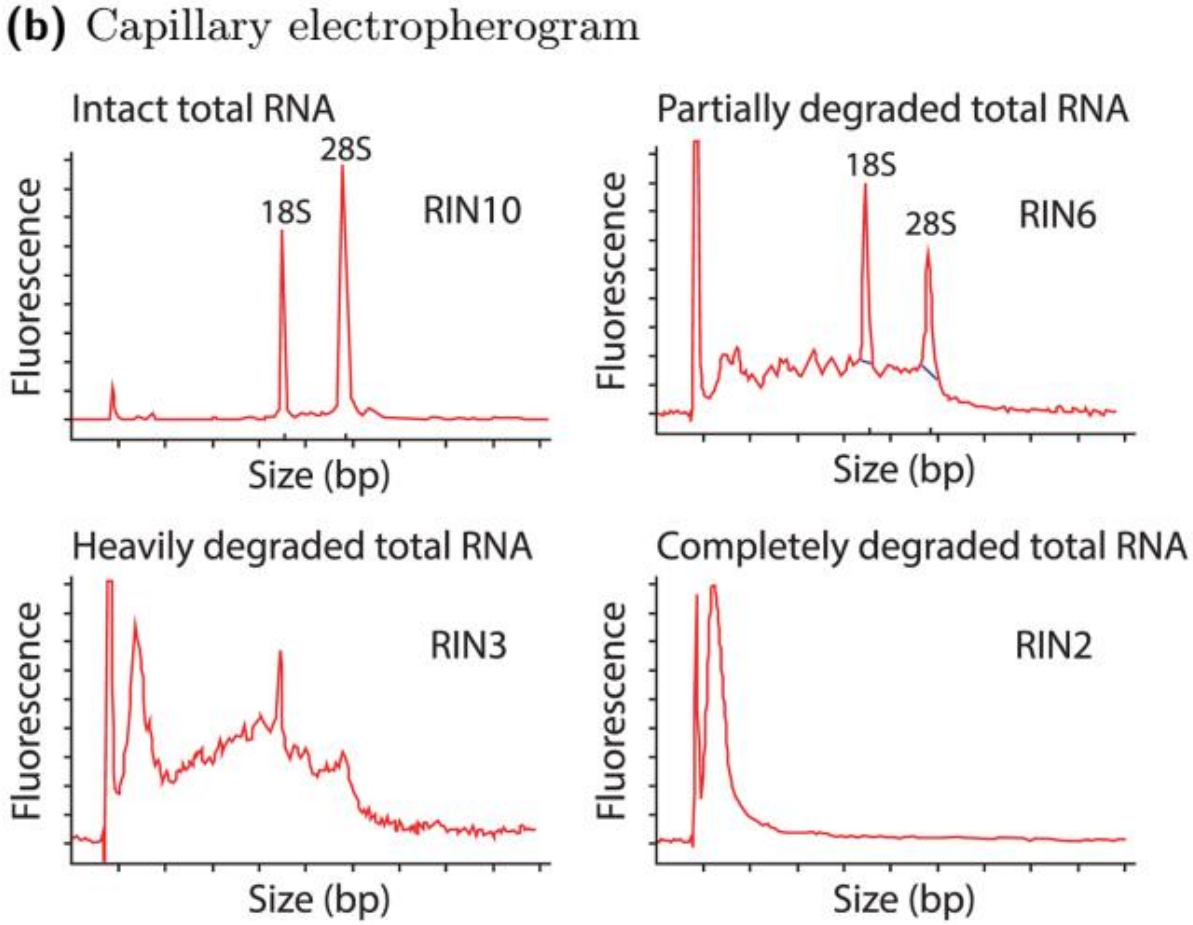
Downstream analysis



Quality control of RNA preparation (RIN)

RNA integrity assessment is based on the 28S:18S rRNA ratio





Removal of rRNA

Type of RNA:

Ribosomal (rRNA)
- Responsible for protein synthesis

Messenger (mRNA)
- Translated into protein in ribosome
- have poly-A tails in eukaryotes

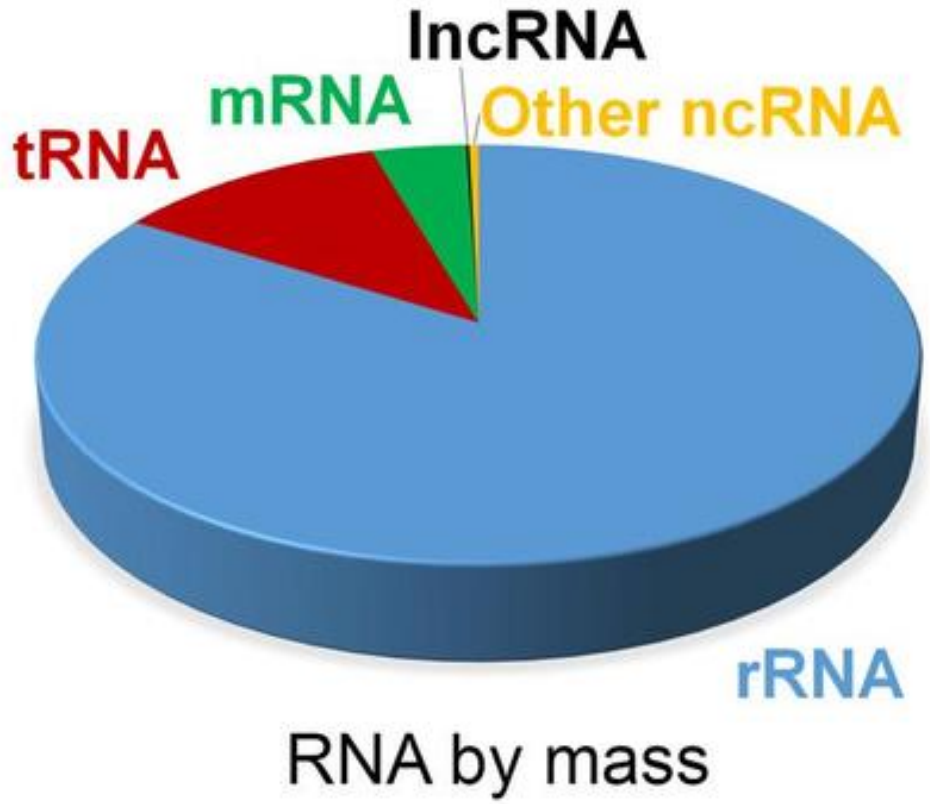
Transfer (tRNA)
- Bring specific amino acids for protein synthesis

Micro (miRNA)
- short non-coding RNA for expression regulation

Others (lncRNA, shRNA, siRNA, snoRNA, etc.)

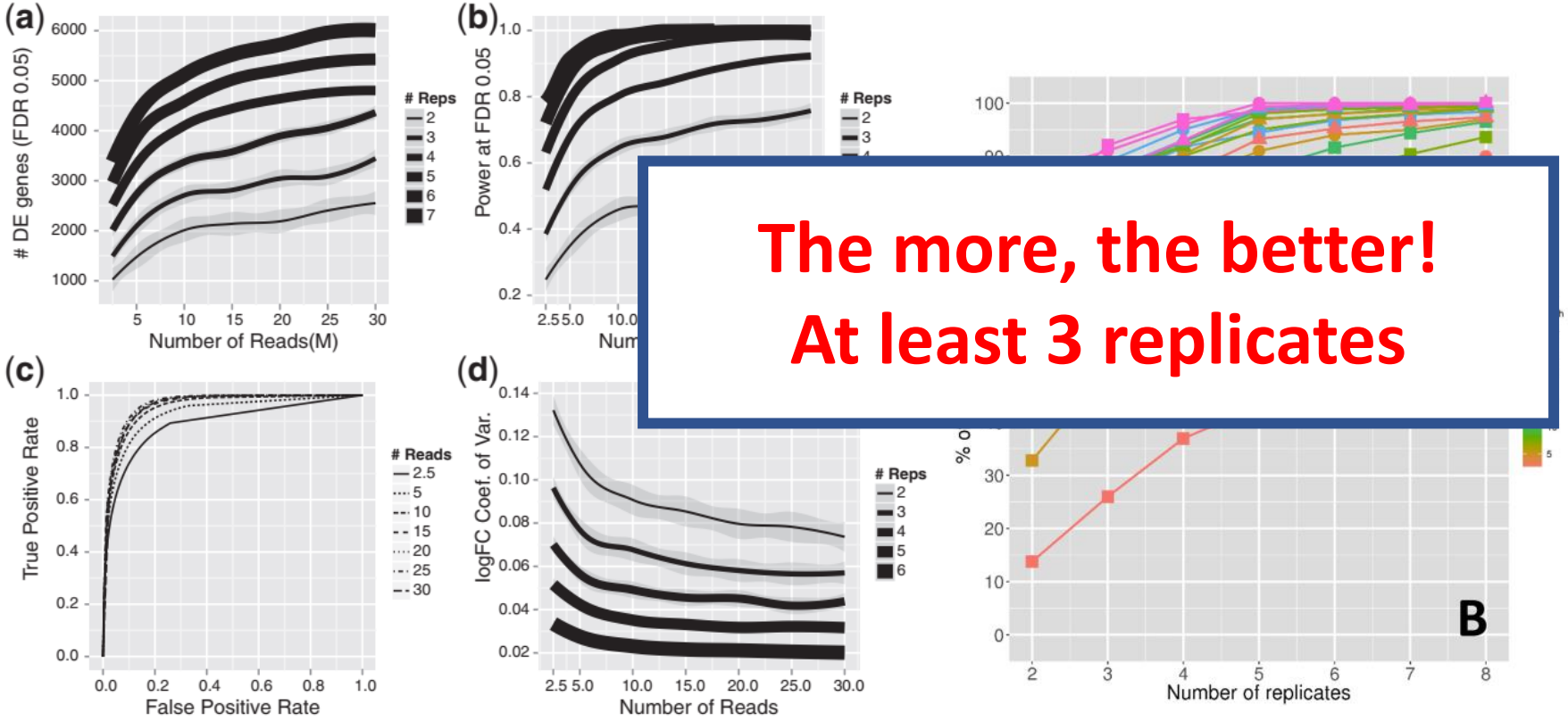
Removal/Enrichment Methods:

- rRNA depletion
- Size selection
- poly-A selection (eukaryotes only)

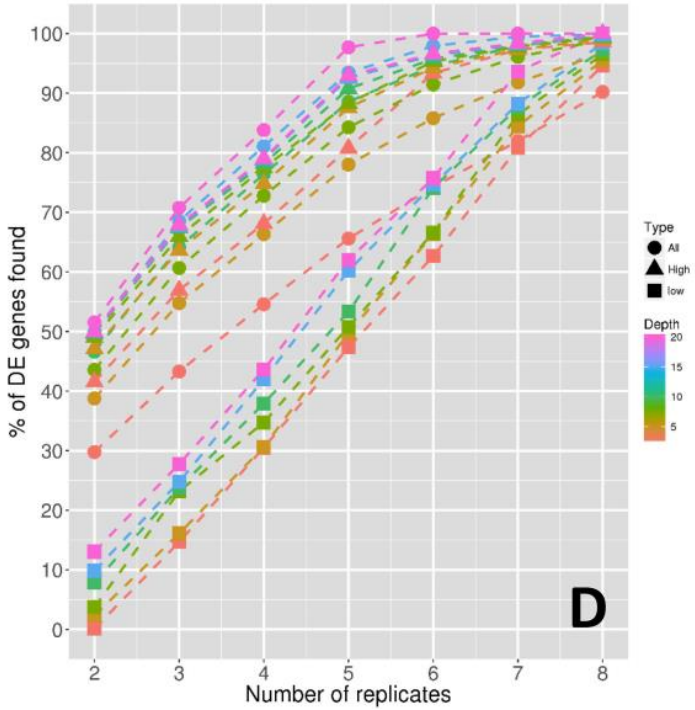


Increases in the Biological replicates number

Increasing the number of biological replications consistently increases the power of detecting DE genes significantly, regardless of sequencing depth.



The more, the better!
At least 3 replicates



Bioinformatics. 2014 Feb 1;30(3):301-4.

Front Plant Sci. 2018 Feb 14;9:108.

How many reads should be enough? (Coverage)

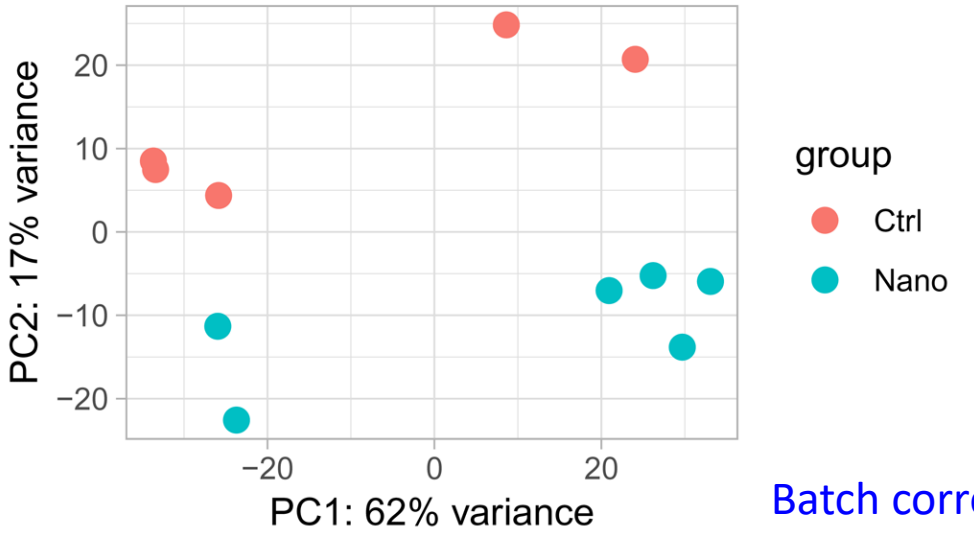
Experiments	Reads
For a quick snapshot of highly expressed genes	5–25 million reads per sample
For a more global view of gene expression or alternative splicing (isoforms)	30–60 million reads per sample
For in-depth view of the transcriptome or novel transcriptome assembly	100–200 million reads per sample
For miRNA-Seq or small RNA Analysis	1–5 million reads per sample

How long should the reads be?

Analysis	Read length
Gene expression Profiling	1 x 75–100 bp
Transcript expression Profiling	2 x 75–100 bp
Transcriptome Analysis	2 x 100–150 bp
Small RNA Analysis	1 x 50 bp

Beware confounding factors (batch effects)

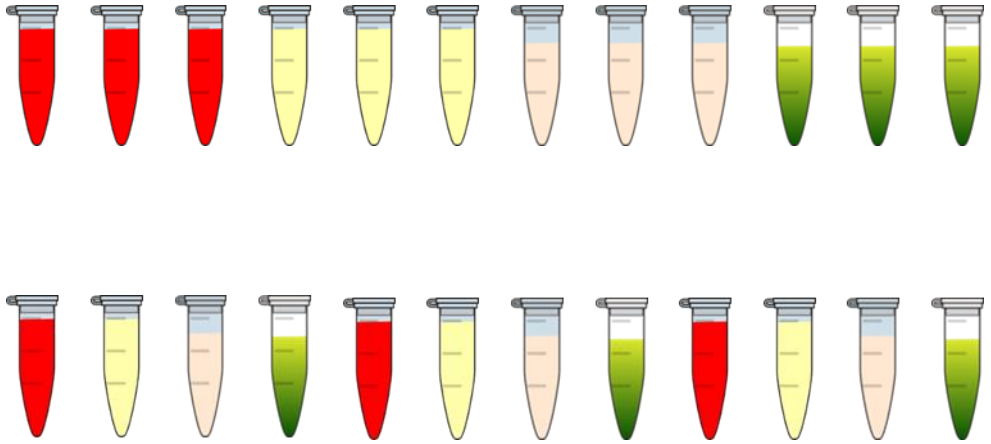
- The Ideally experimental design is to compare two groups that **only differ in one factor.**
- Batch effect can occur when subsets of the replicates are handled separately at any stage of the process -- handling group becomes in effect another factor.
- Avoid processing all samples from single group if you can’t do all the samples at once.



Batch correction methods: SVA, PVCA, BatchQC

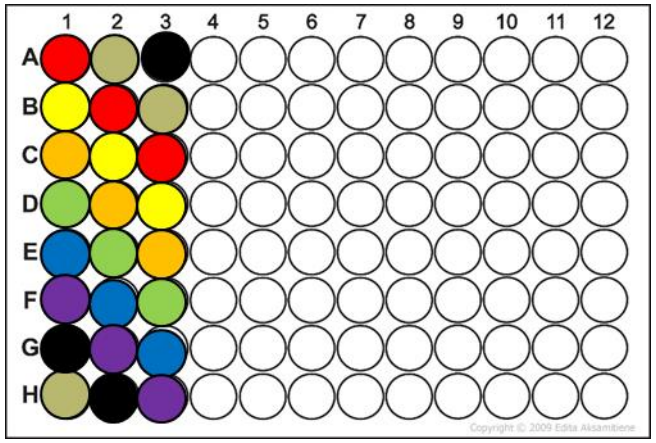
Beware systematic biases (randomization)

- Avoid systematic biases in the arrangement of replicates.
  - Don’t arrange replicate sample sets in the same order



1

2



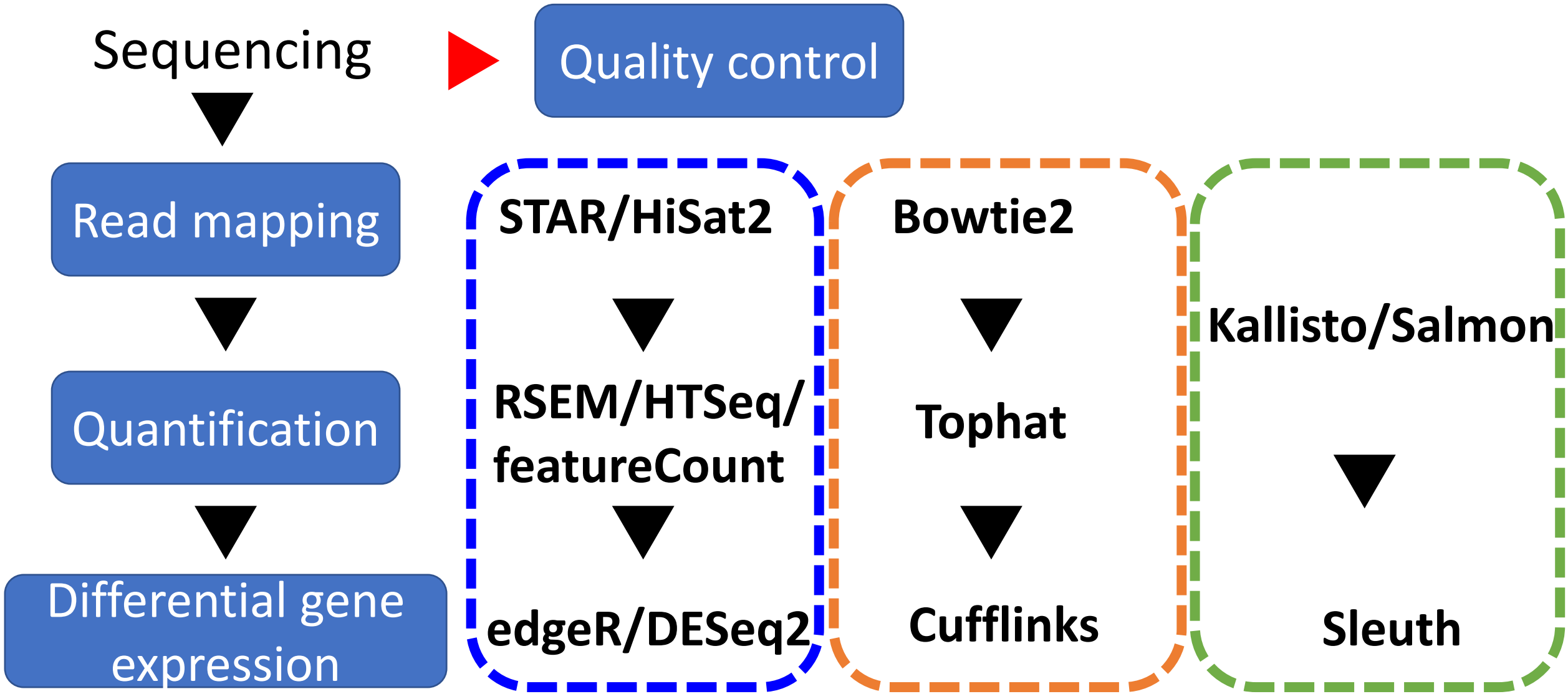
Outline

1. Experimental design and practical considerations

2. Differential gene expression analysis pipeline

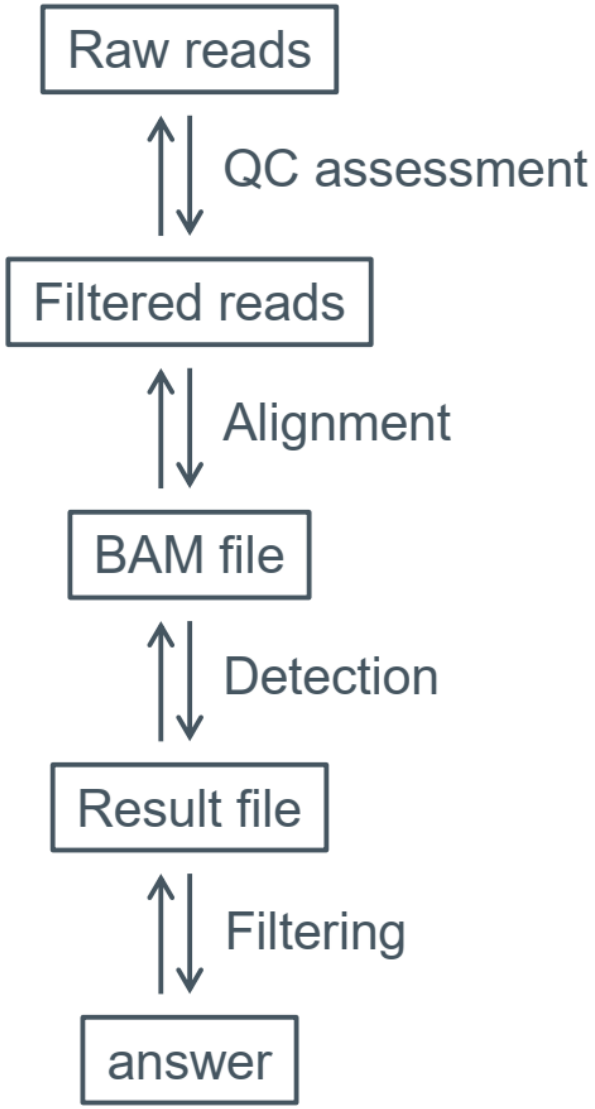
3. IMB Bioinformatics Core analysis tools (DEMO)

Data analysis pipeline (DEG)



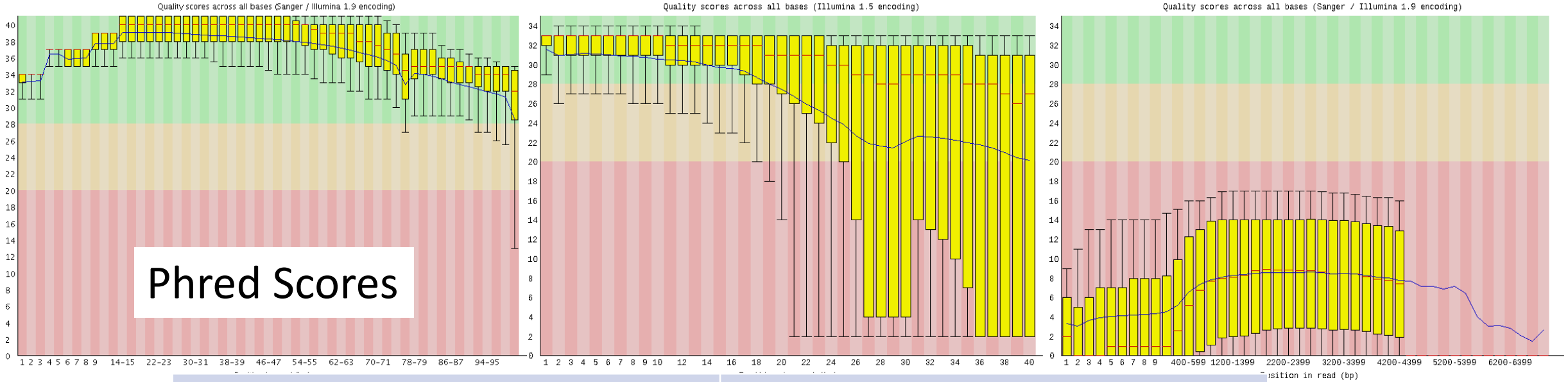
File formats

- Raw data (FASTQ/FAST5)
- Alignment (SAM/BAM)
- Reference genomes (FASTA)
- Annotation files (GTF/GFF)
- Result files (TXT)



Read QC– always check your data first

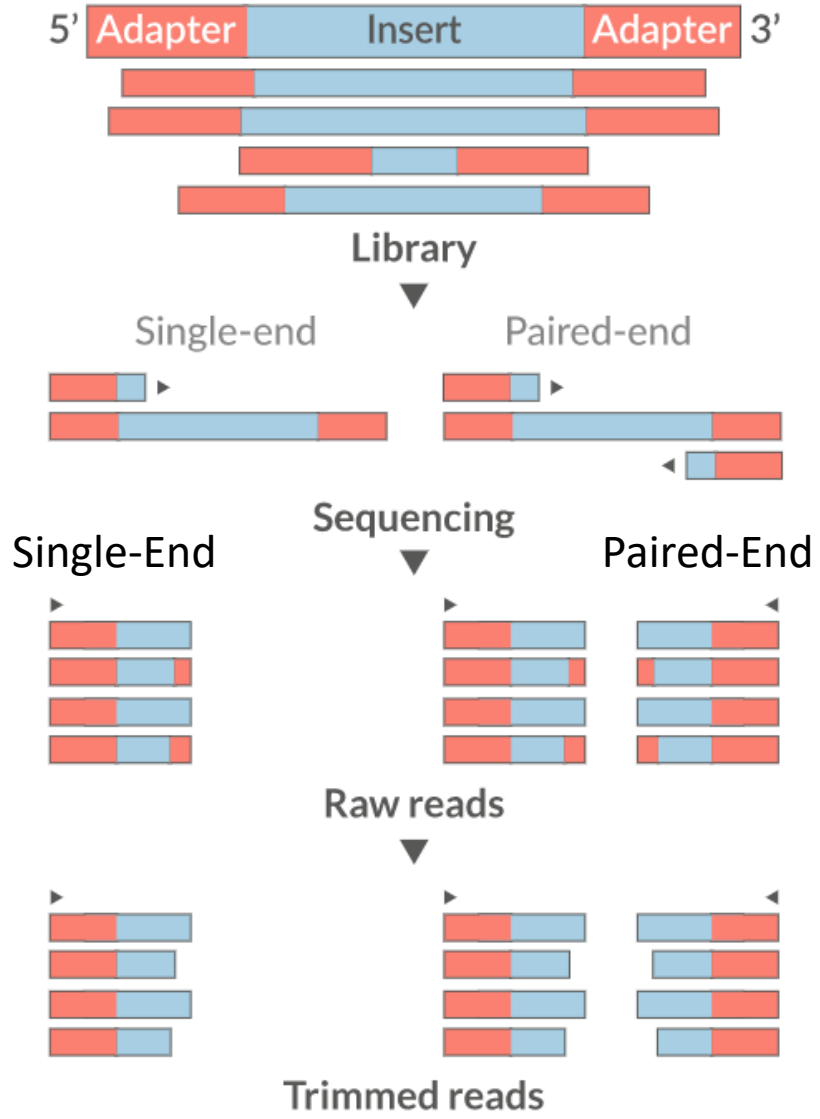
FastQC



Number of reads	Sequence length distribution
Per base sequence quality	Sequence duplication levels
Per sequence quality score	Overrepresented sequences
Per base sequence content	Adapter content
Per sequence GC content	Kmer content
Per base N content	

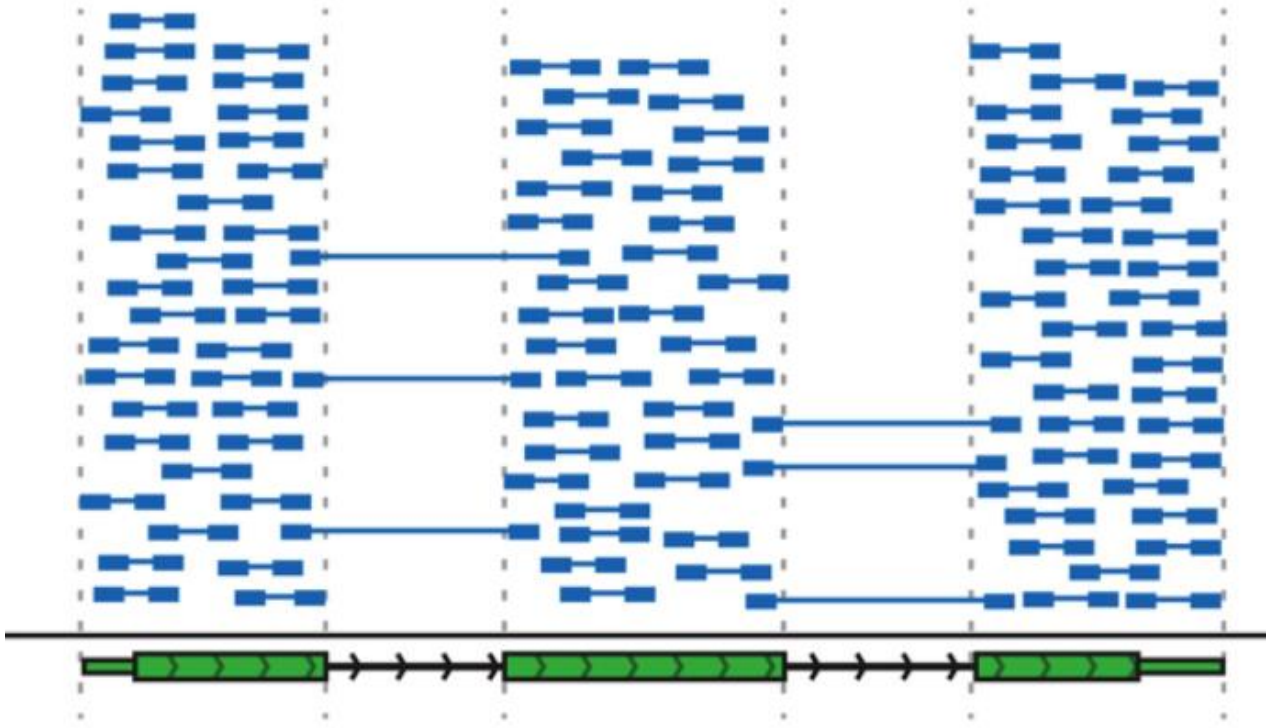
Read trimming and filtering (Optional) - Cutadapt

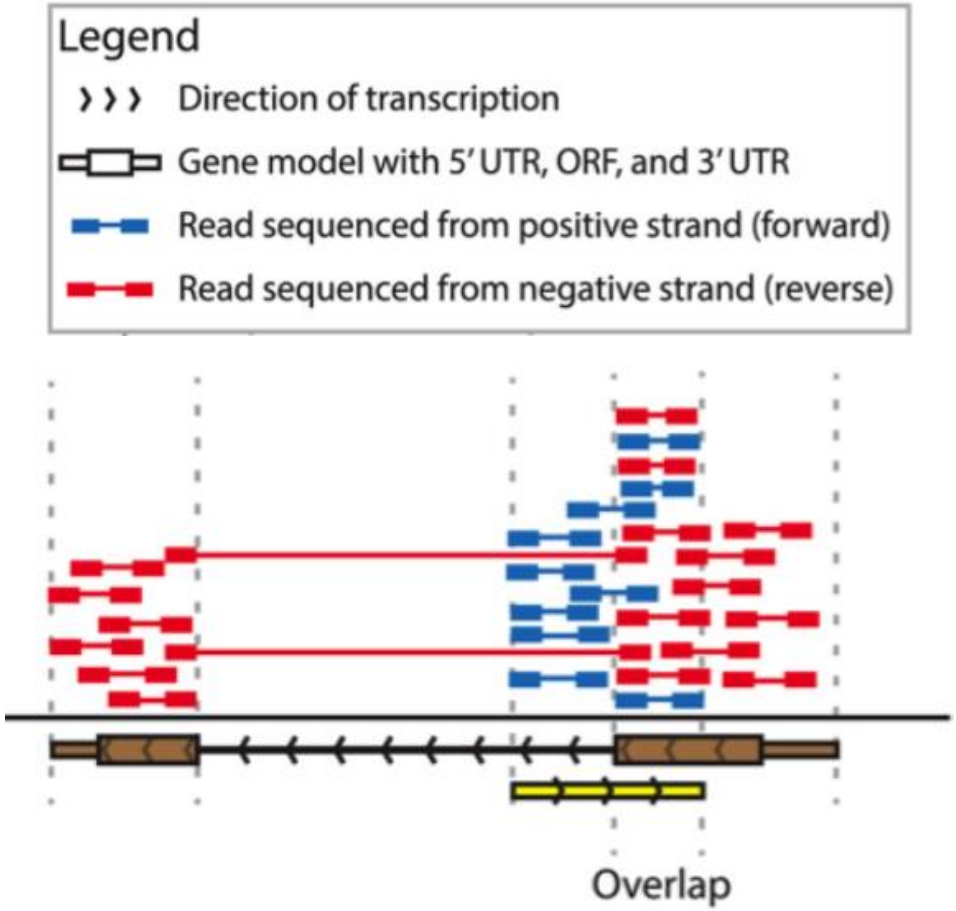
- Remove adapter sequences
- Trim reads by quality
- Filter by min/max read length



Read alignment - STAR

1. For a sequencing read, to determine the origin location within the reference genome
2. Reference include complete genome and transcriptome sequence
3. Splice-aware alignment

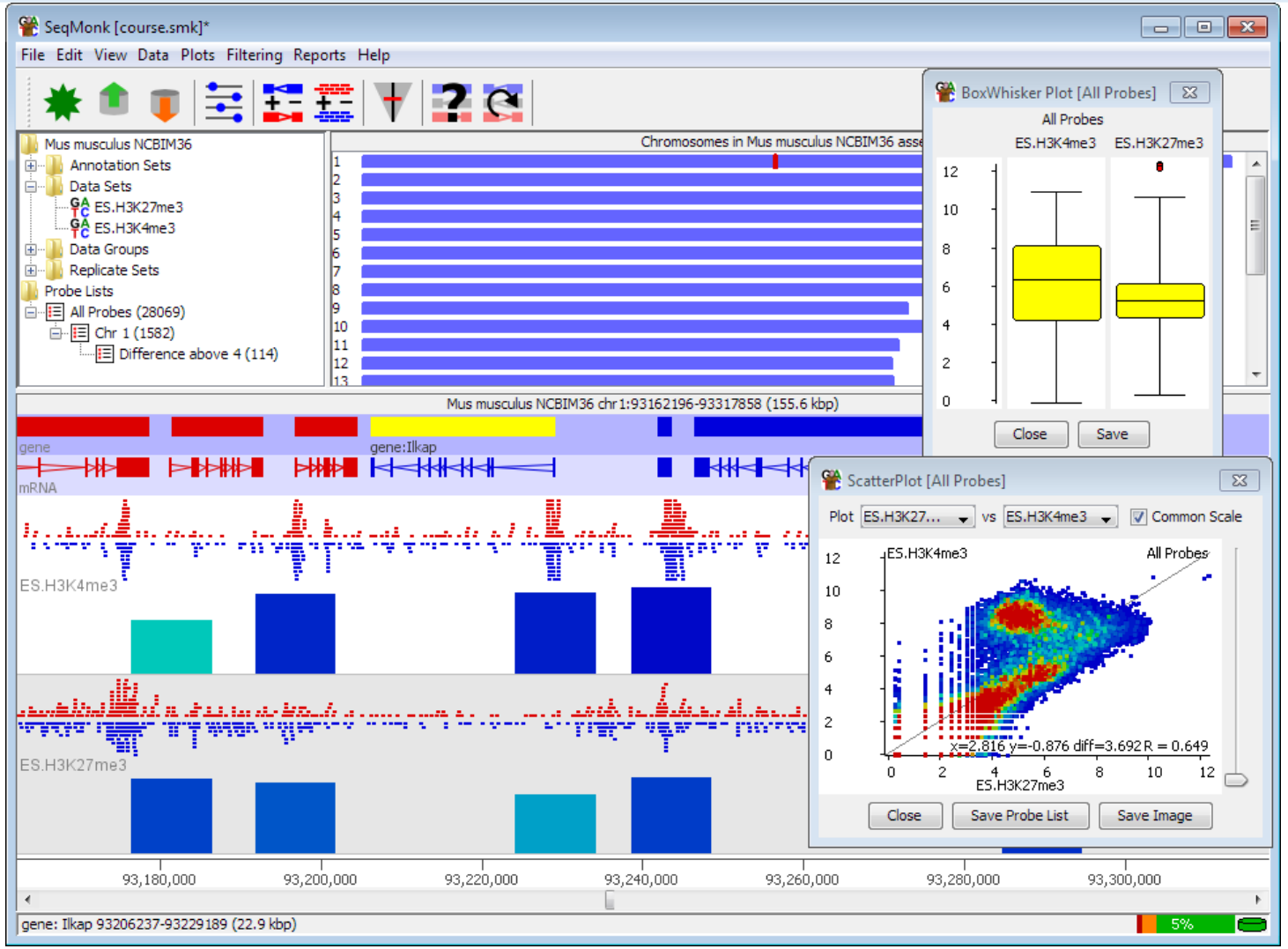




Visualization of read alignments - IGV

https://software.broadinstitute.org/software/igv/

Visualization of read alignments - SeqMonk



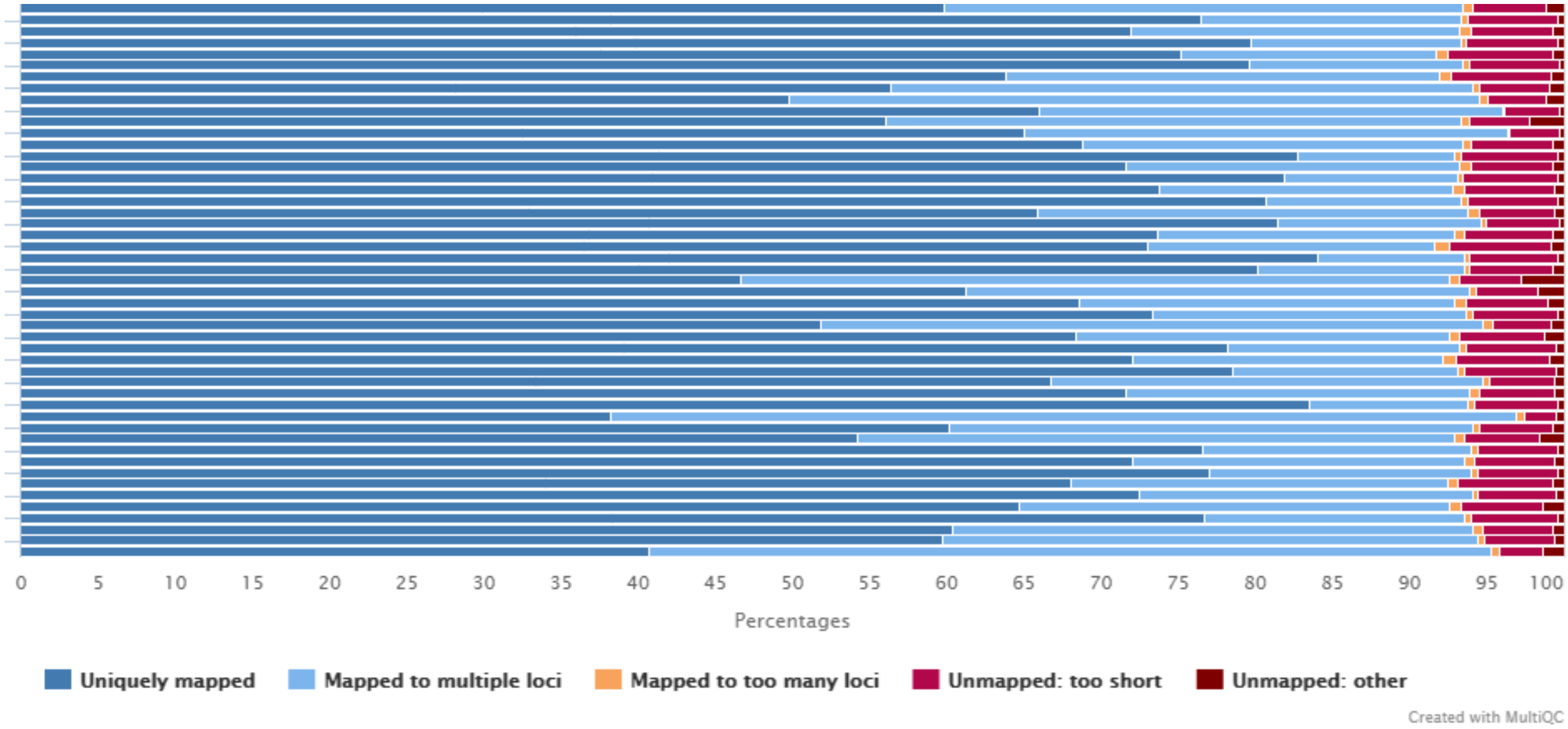
https://www.bioinformatics.babraham.ac.uk/projects/seqmonk

Alignment QC

- Number of reads mapped/unmapped/paired etc
- Uniquely mapped
- Insert size distribution
- Coverage
- Gene body coverage
- Chromosome counts
- Counts by region: gene/intron/non-genic
- Sequencing saturation
- Strand specificity

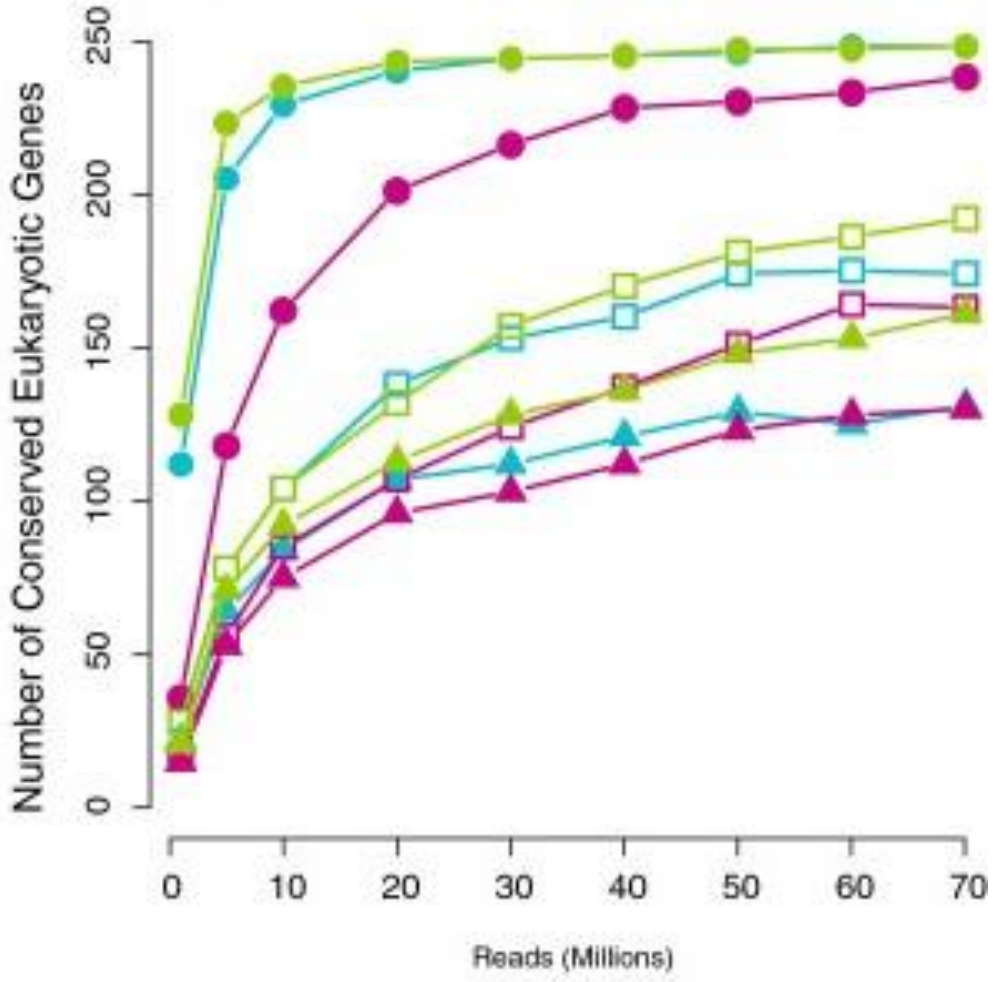
STAR (log file)
QoRTs
RSeQC
MultiQC

Alignment QC

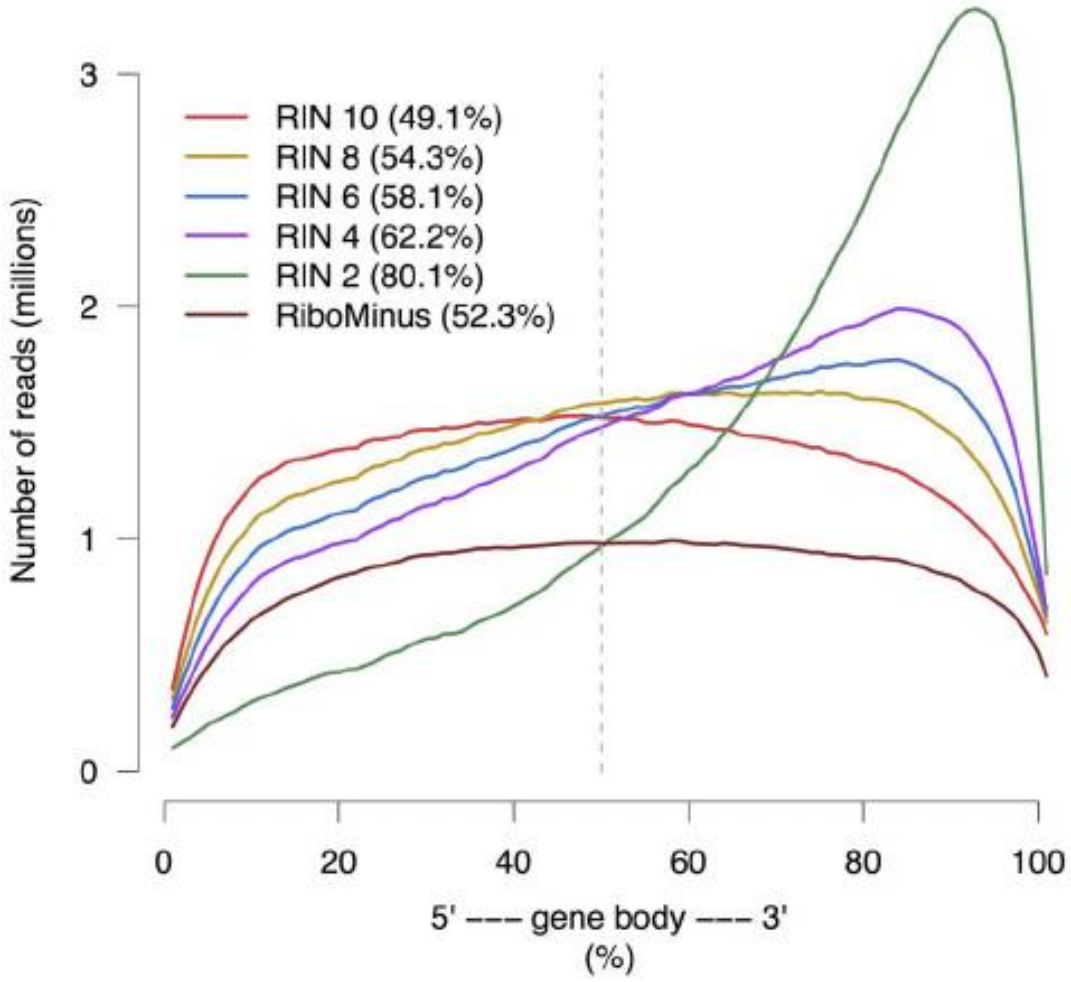


Alignment QC

Saturation curve (gene diversity)

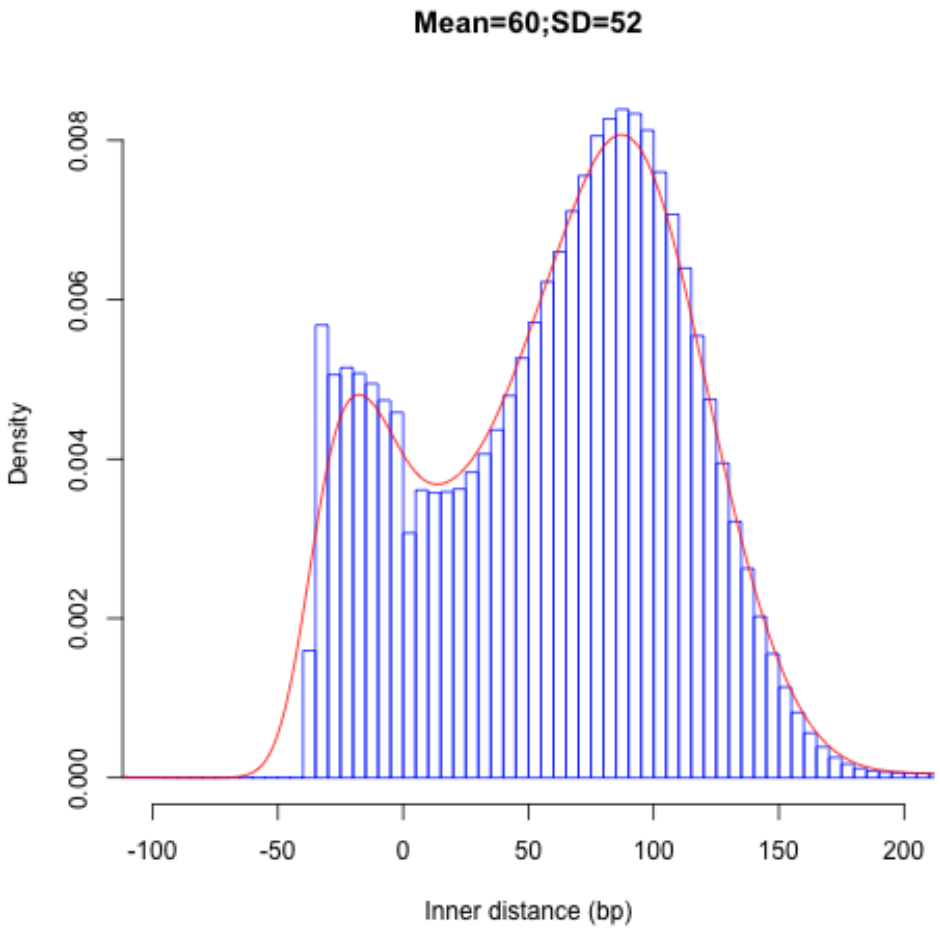


Gene body coverage

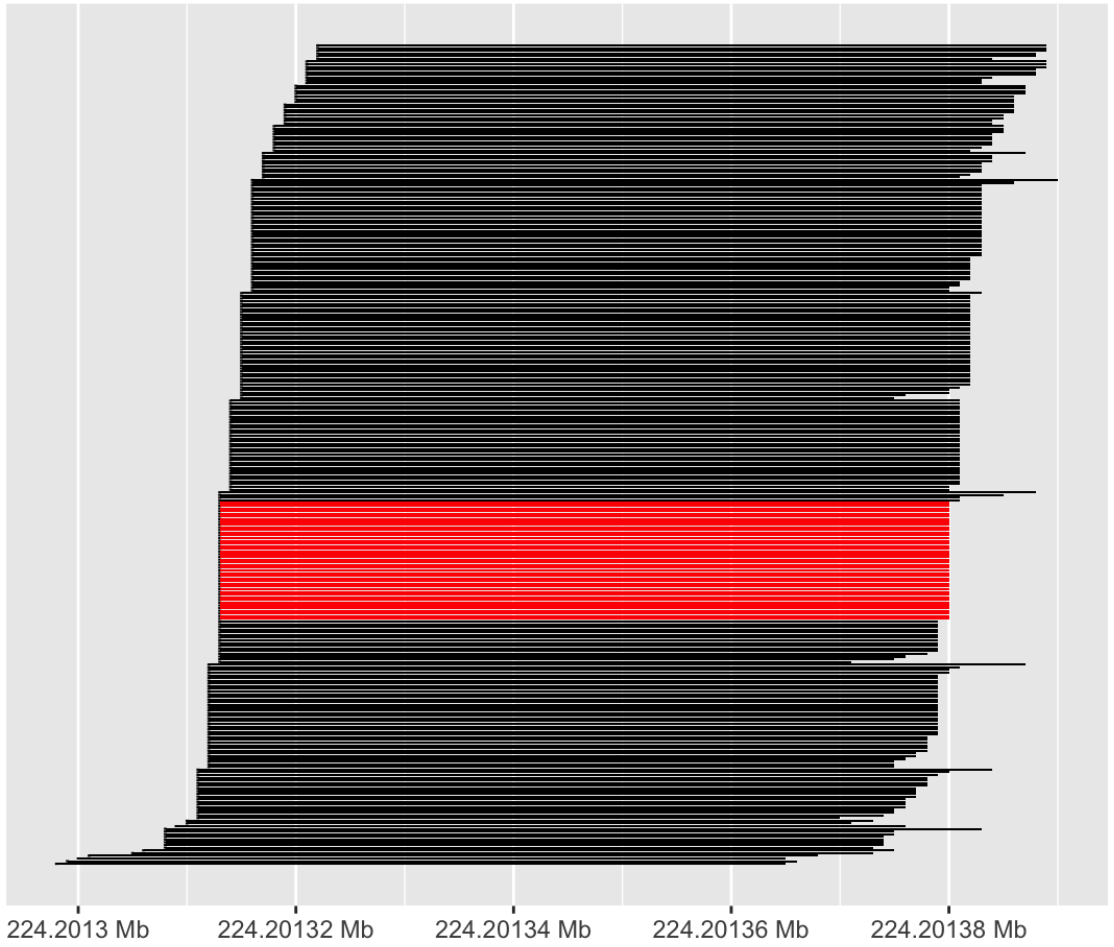


Alignment QC

Insert size

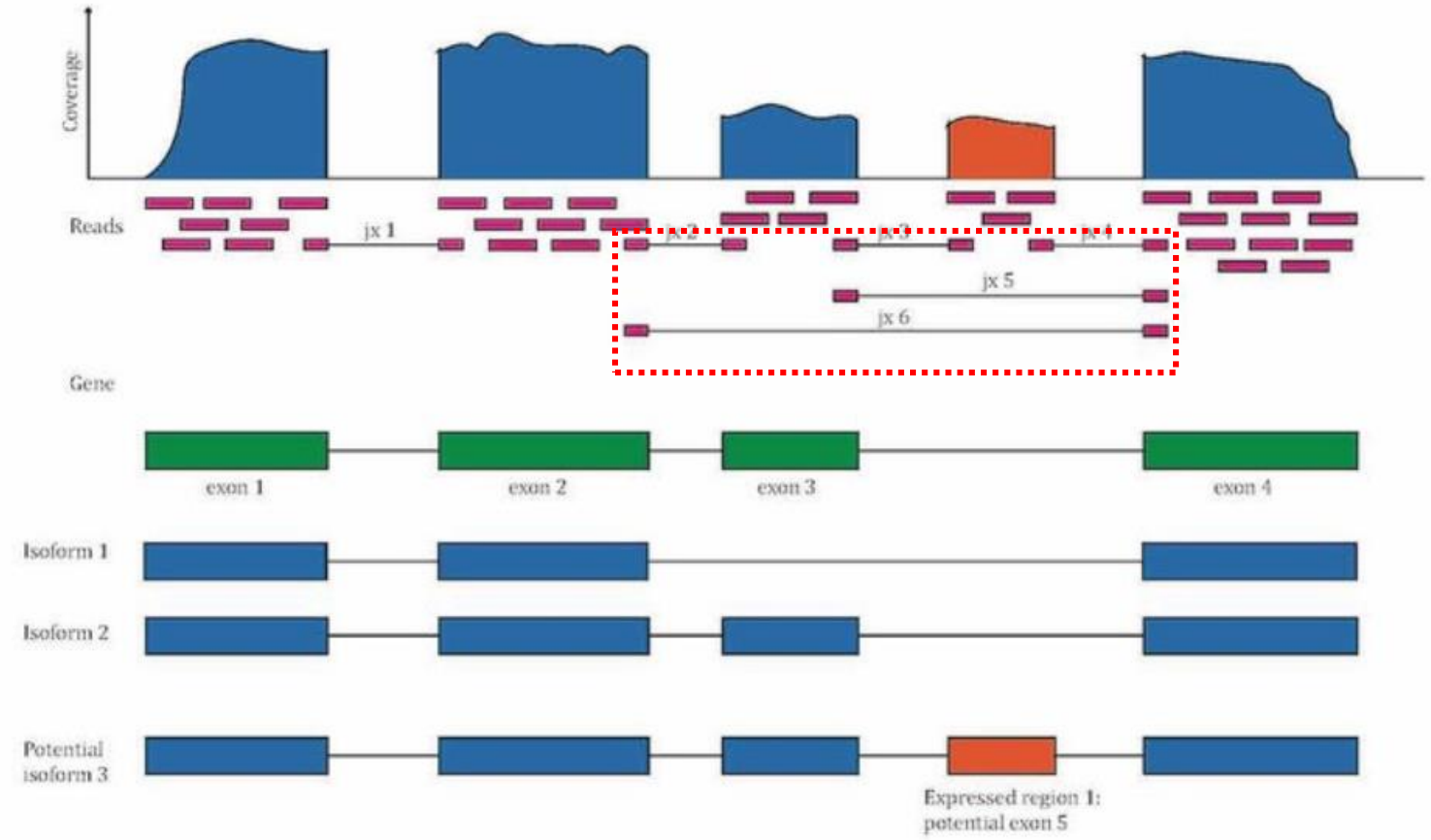


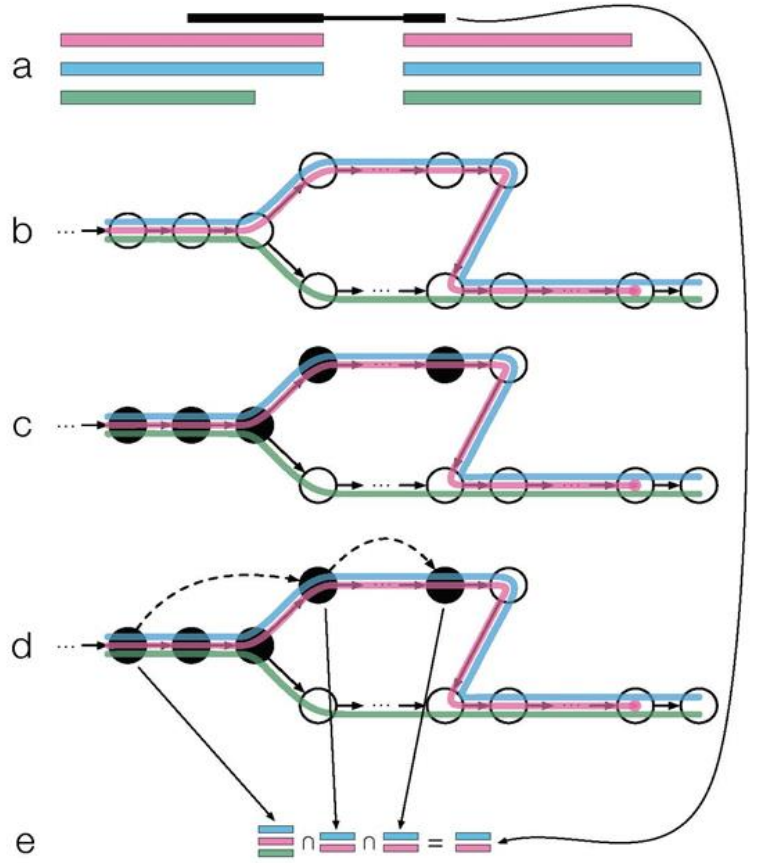
PCR duplications



Quantification of Gene/Transcript expression

- Conventional methods: RSEM; featureCounts
- Novel methods: Salmon; Kallisto (based on pseudo-alignments)





Statistical analysis

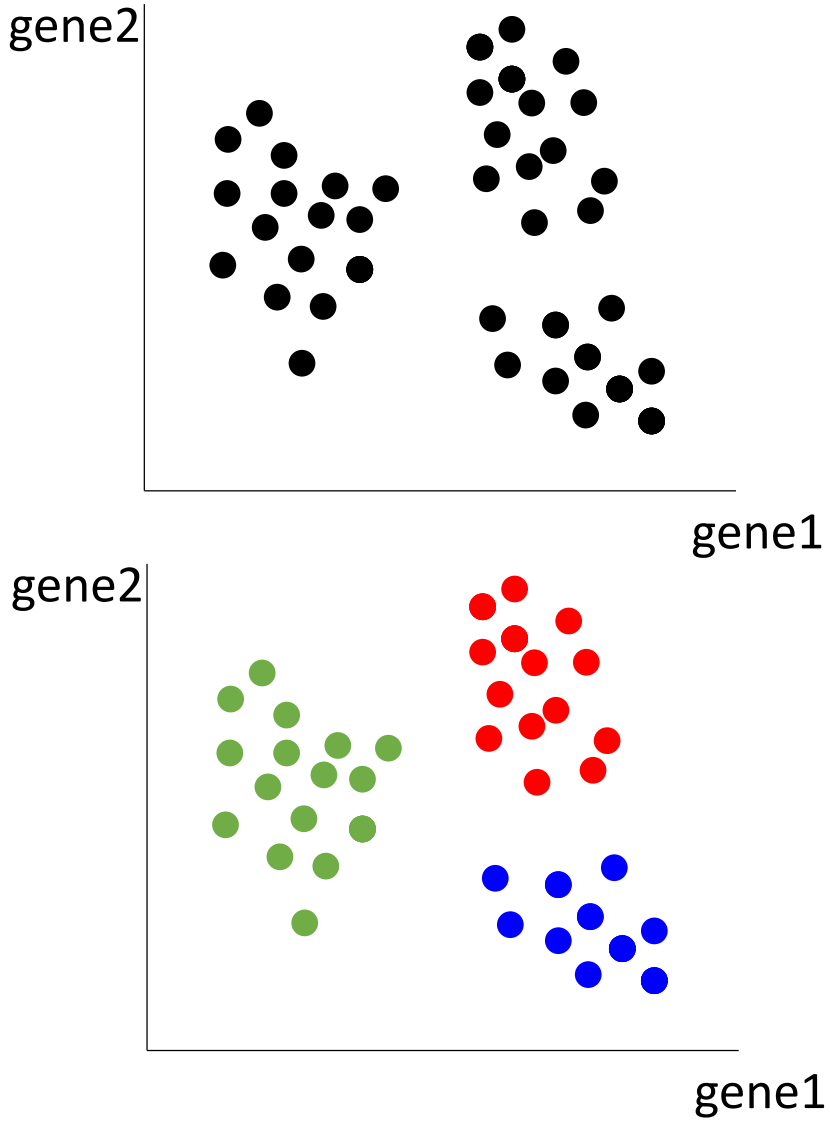
1. Normalization of gene counts across all samples
2. Clustering of samples based on all gene expression profiles
3. Identification of differential expression genes (DEGs)
4. Functional annotation
5. Gene Set Enrichment Analysis (GSEA)

Normalization of expression data

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	**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	**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	**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**

Visualize high-dimensional data

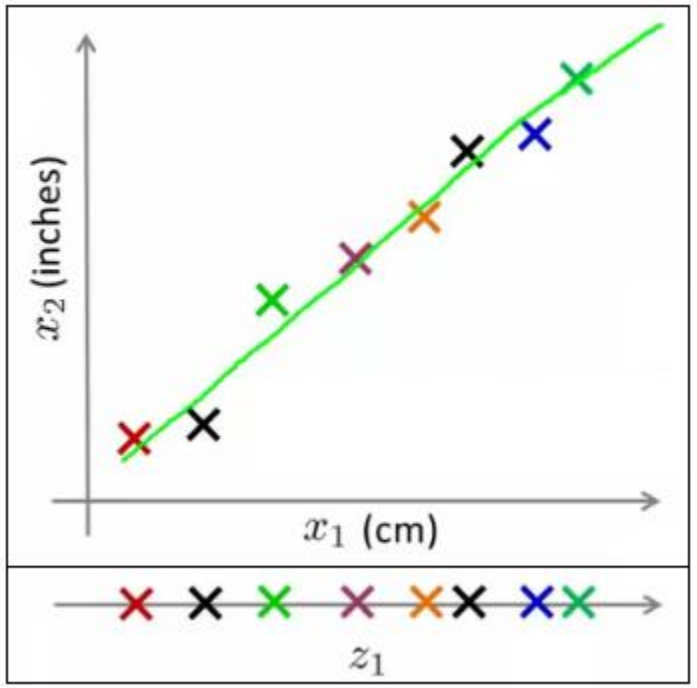
- Dimensional Reduction:
  - Principal component analysis (PCA)
  - Multidimensional scaling plot (MDS)
  - t-SNE
  - UMAP
- Clustering:
  - Hierarchical Clustering
  - K-means Clustering



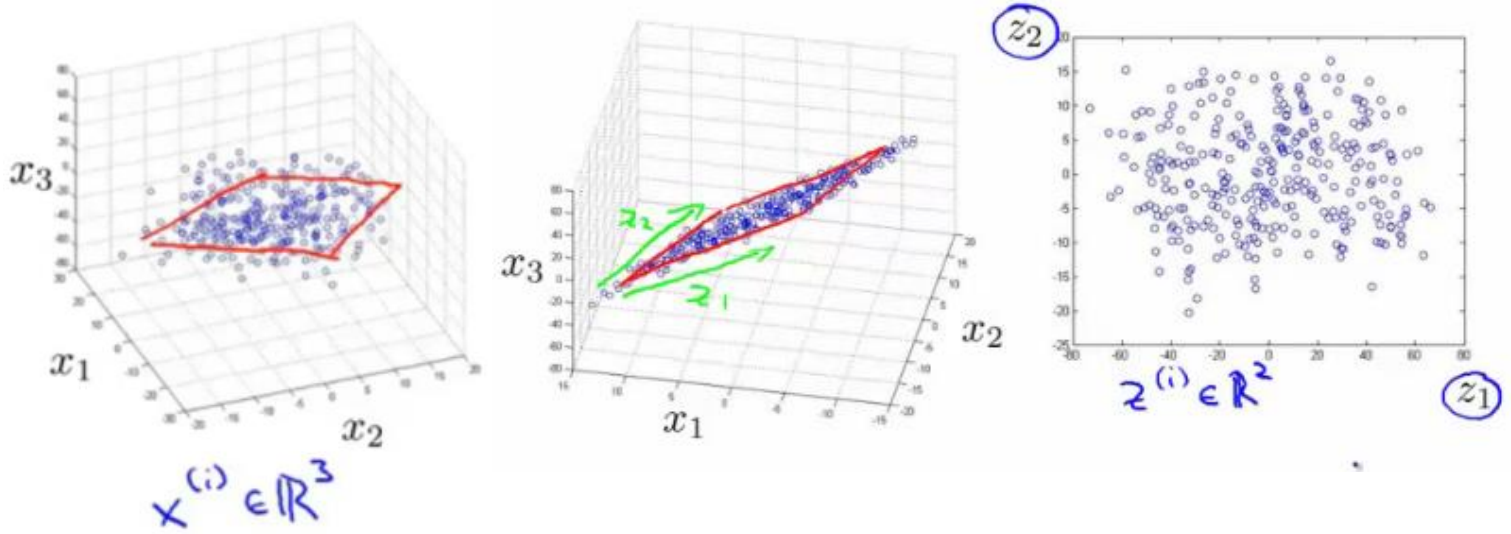
PCA

- An unsupervised, non-parametric statistical procedure that uses an orthogonal linear transformation that converts a set of correlated variables to a set of uncorrelated variables (coordinates).

Reduce data from 2D to 1D



Reduce data from 3D to 2D



Find the best fitting line by maximizing the sum of the squared distances from the projected points to the origin.

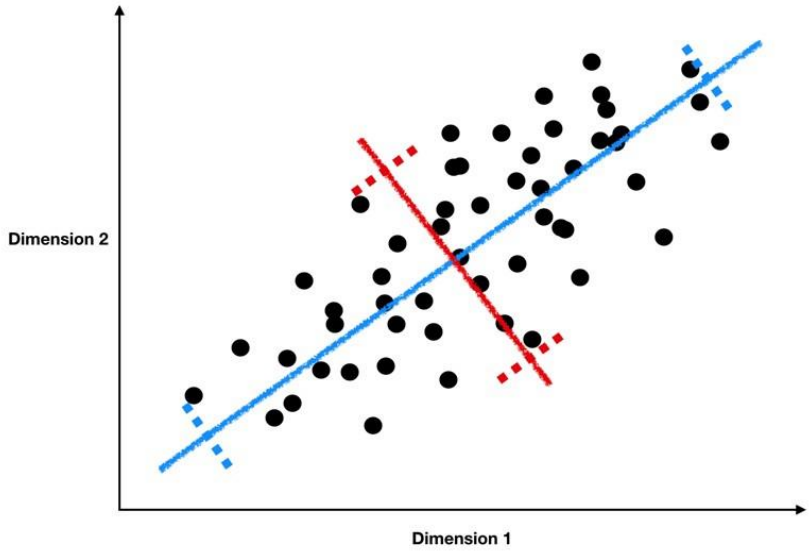
PCA

- Principle Component – a linear combination of variables
- Eigenvector of PC – the sum of squared distance ➔ to determine the proportion of the total variation that each PC accounts for
- Loading score for each variant in each PC – the proportions of each gene in PC ➔ to determine each gene contribute to the principal components

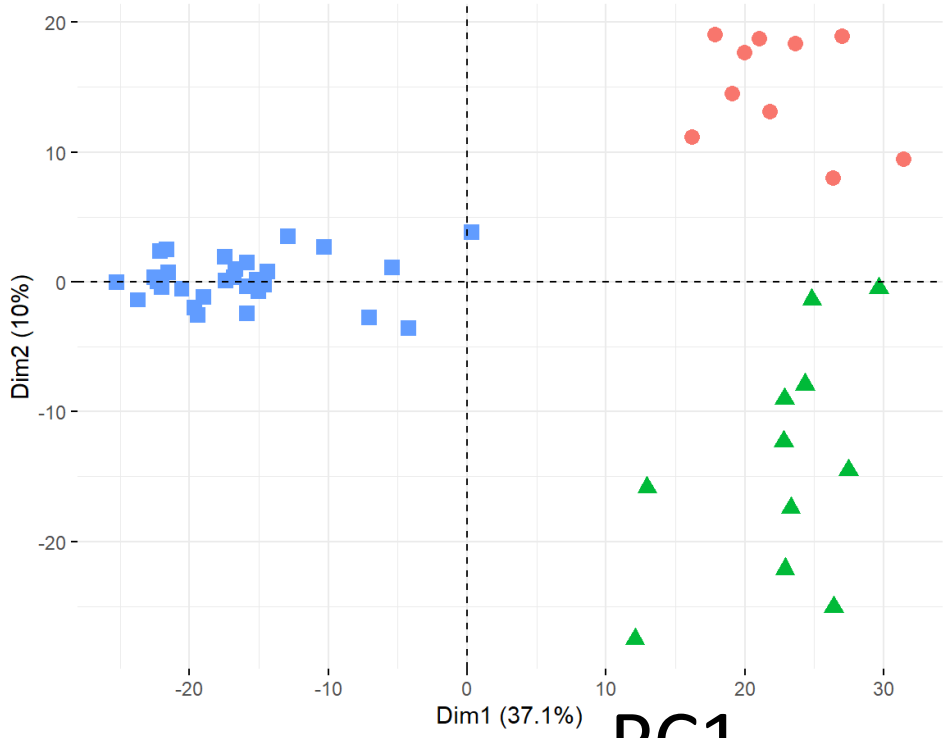
Ex:

PC1 = s1*gene1 + s2*gene2 + s3*gene3 …

The proportion of variation : PC1 > PC2 > PC3…



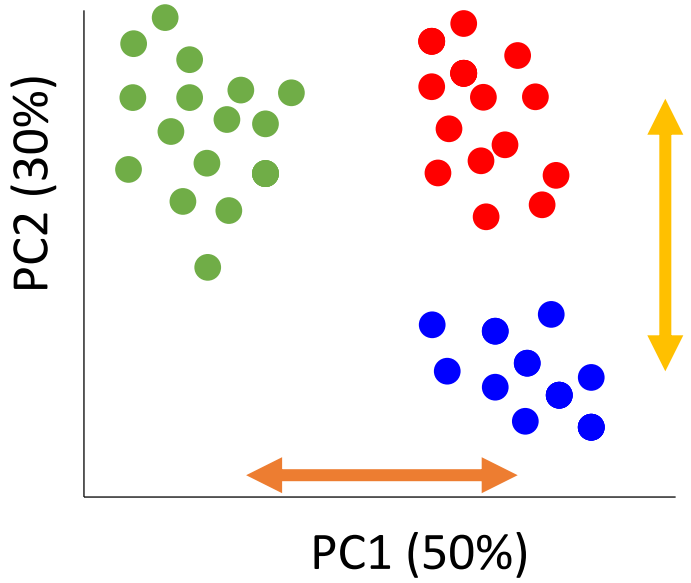
PC2



PC1

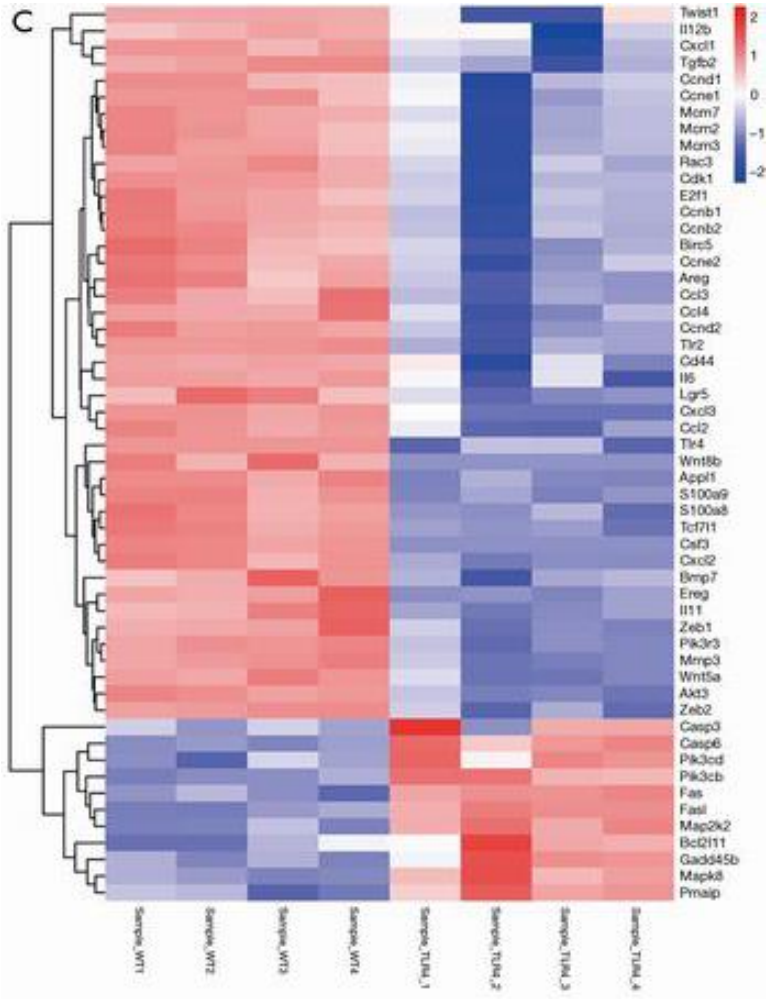
PCA

- PCA is for clustering/grouping all samples based on all gene expressions
- Principle component 1 (PC1) is the dimension which is accounts for the most of the total variation. PC2 is the second most. etc
- To find out which genes are important in PC -- The loading scores of each gene in PC.

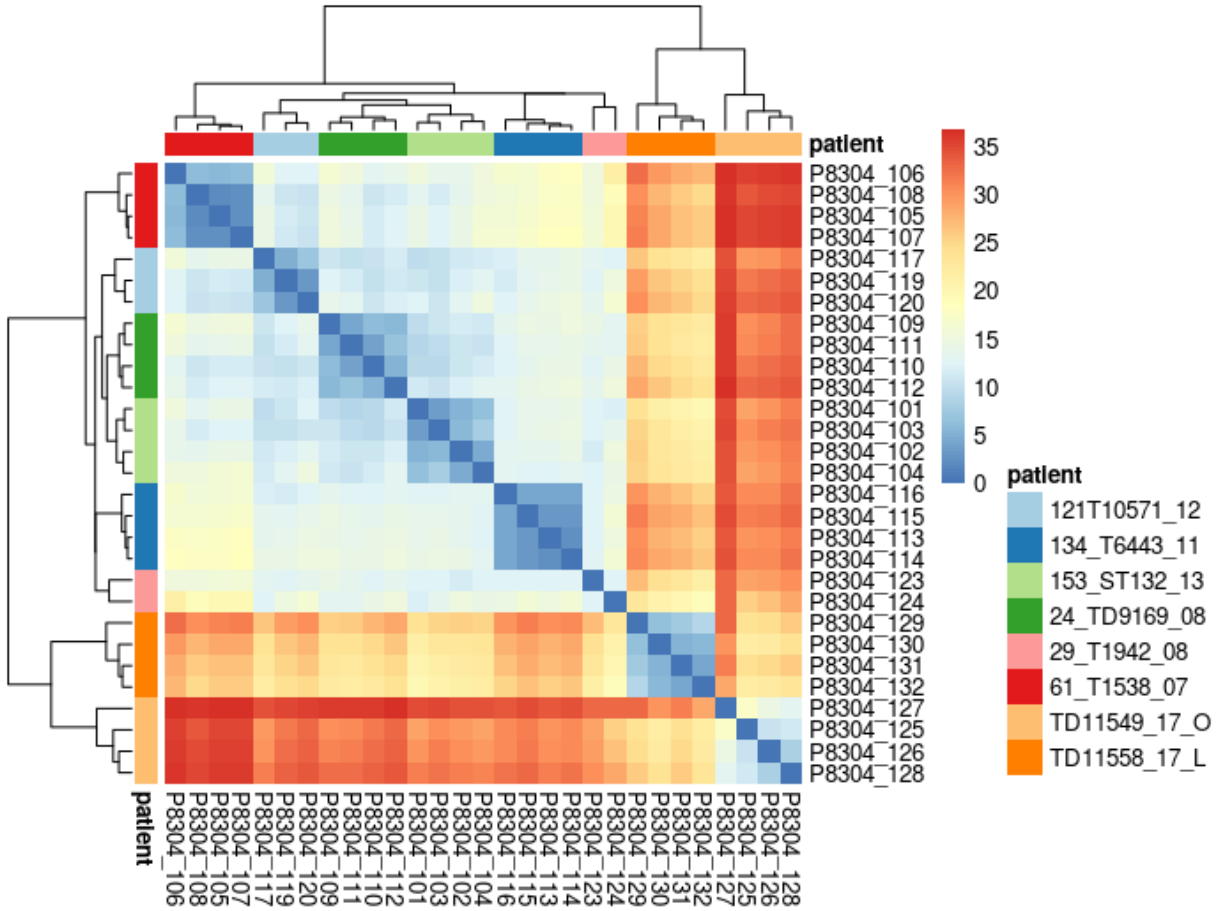


Heatmap

Hierarchical clustering based on expression profiles

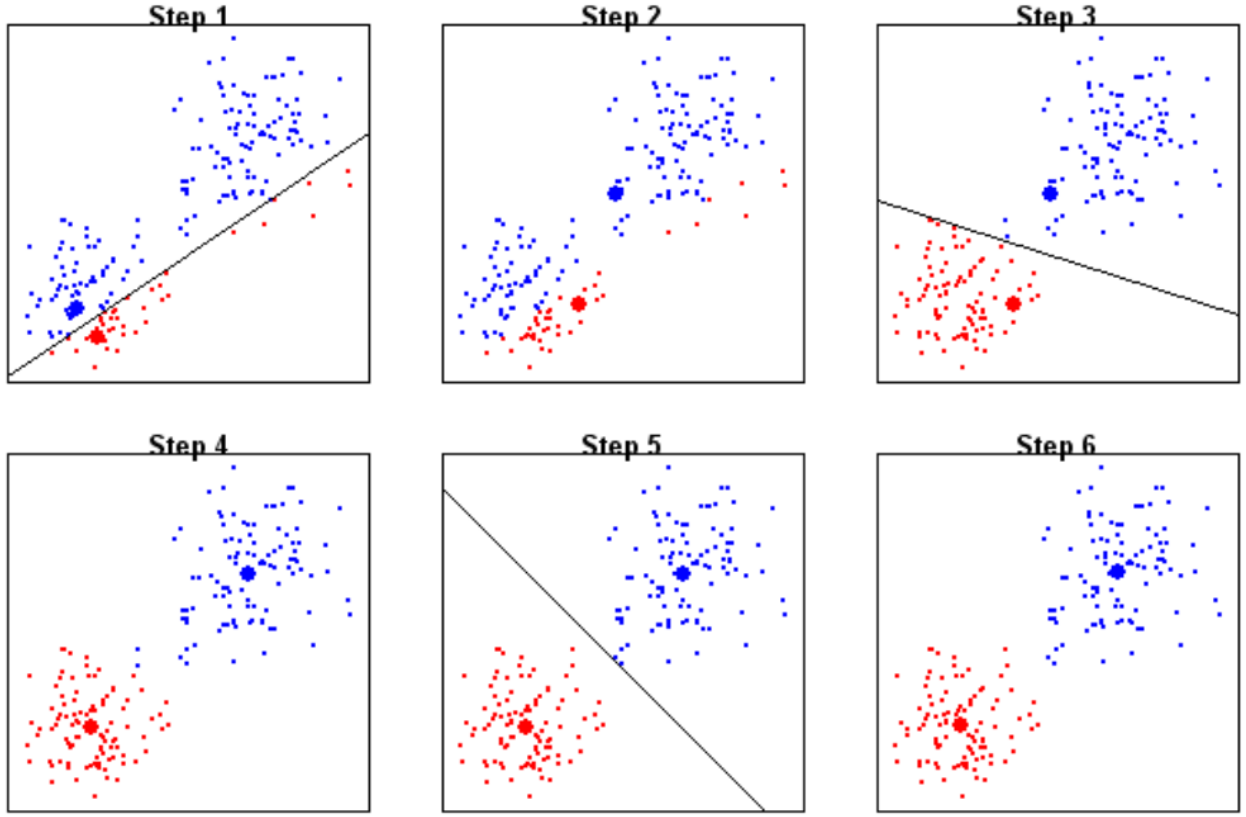


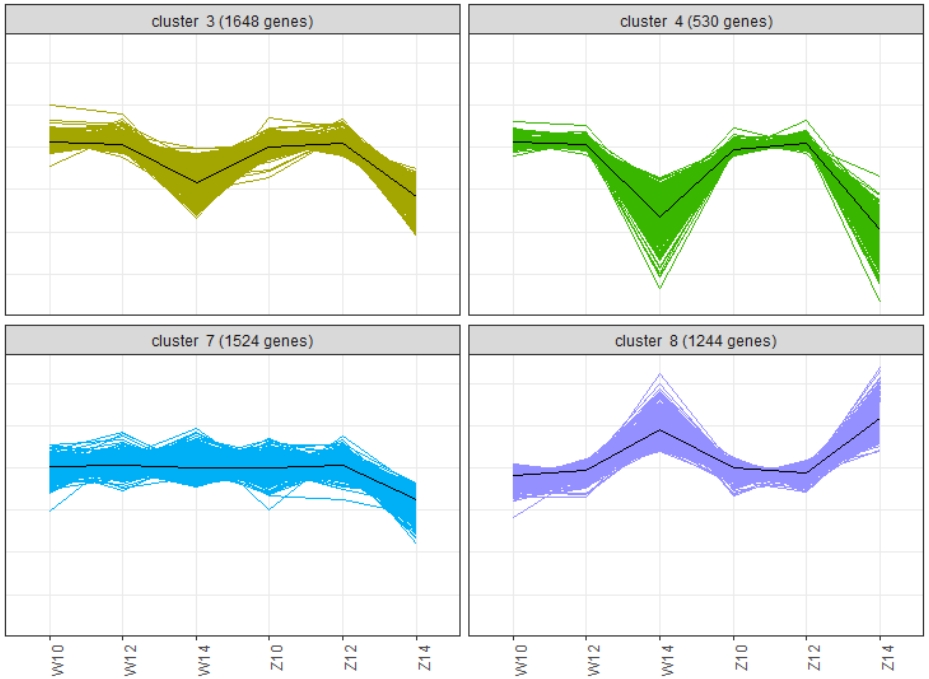
Correlations between samples



K-means clustering

an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters





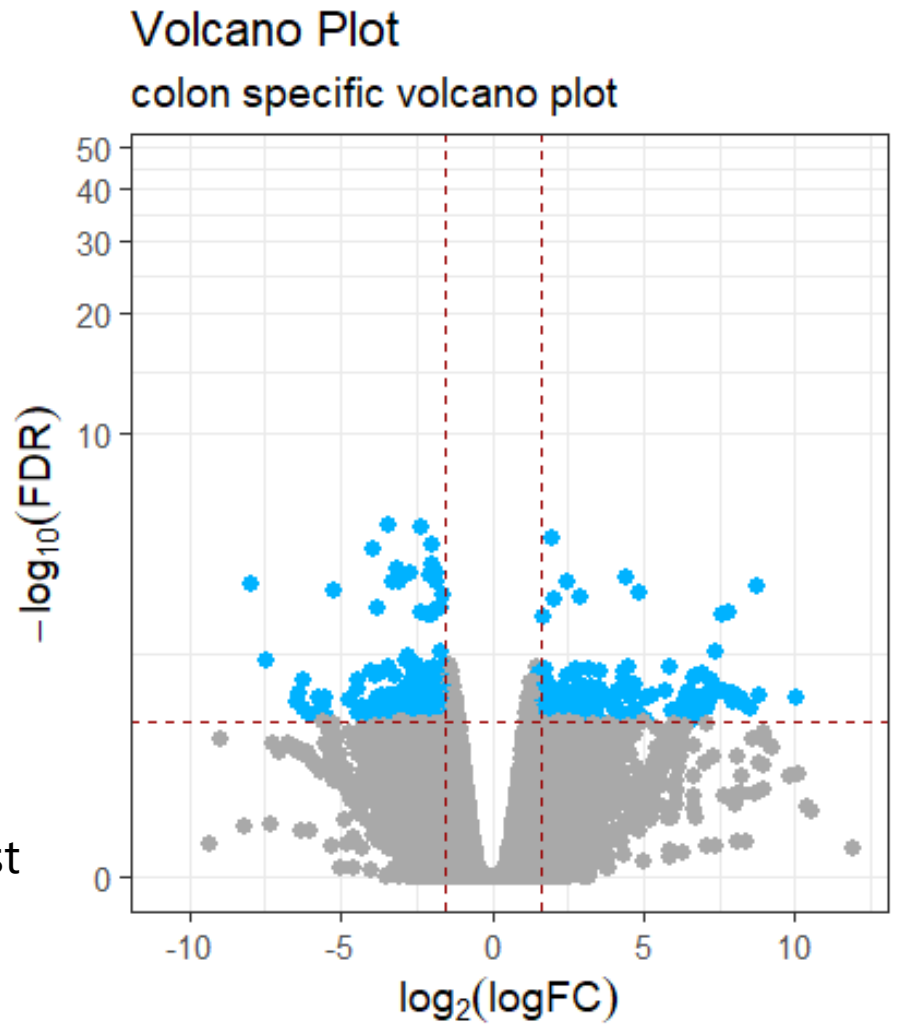
Differential expression genes (DEG)

1. Pairwise comparison between two groups
2. Statistical package: edgeR; DESeq2; limma

- edgeR - exactTest

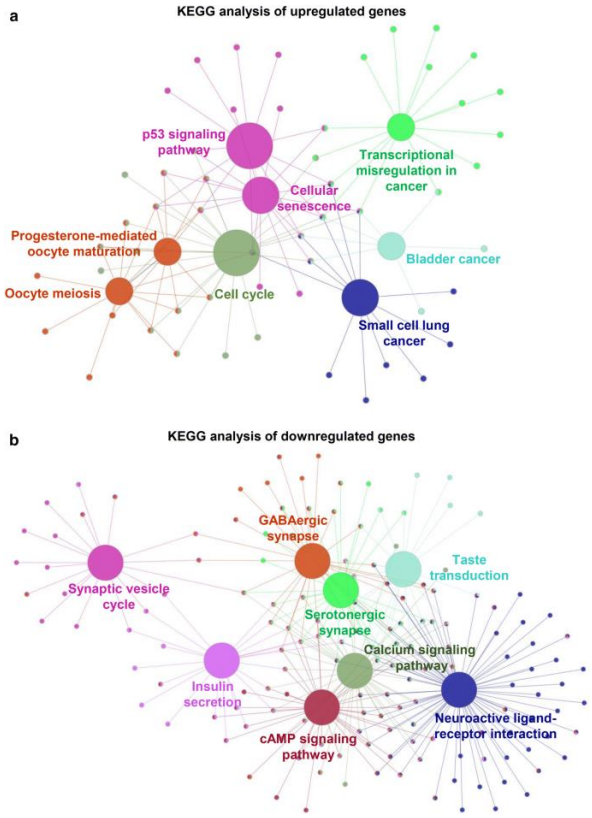
Compute genewise exact tests for differences in the means between two groups of negative-binomially distributed counts.

- DESeq2 - DESeq
  1. estimation of size factors: estimateSizeFactors
  2. estimation of dispersion: estimateDispersions
  3. Negative Binomial GLM fitting and Wald statistics: nbinomWaldTest



Functional annotations / pathway analysis

1. Gene Ontology (GO) enrichment
2. Pathway (KEGG) enrichment
3. DAVID
4. GSEA
5. STRING
6. Enrichr
7. WGCNA
8. …



GSEA (Gene Set Enrichment Analysis)

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

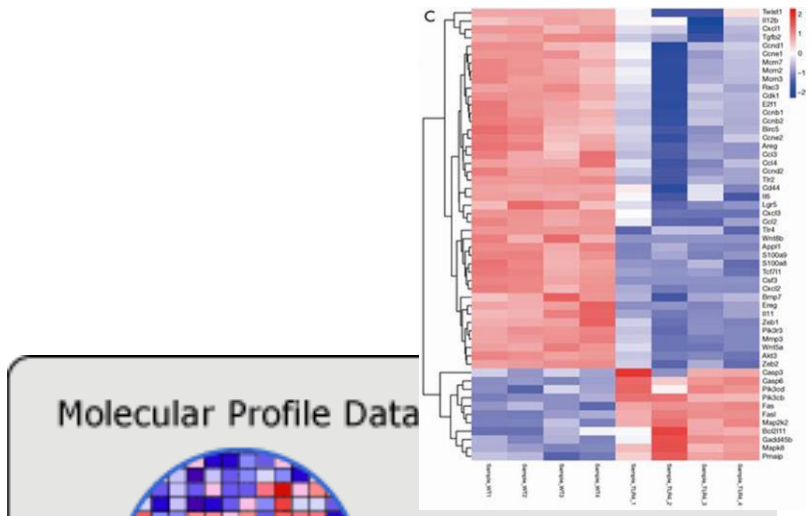
C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

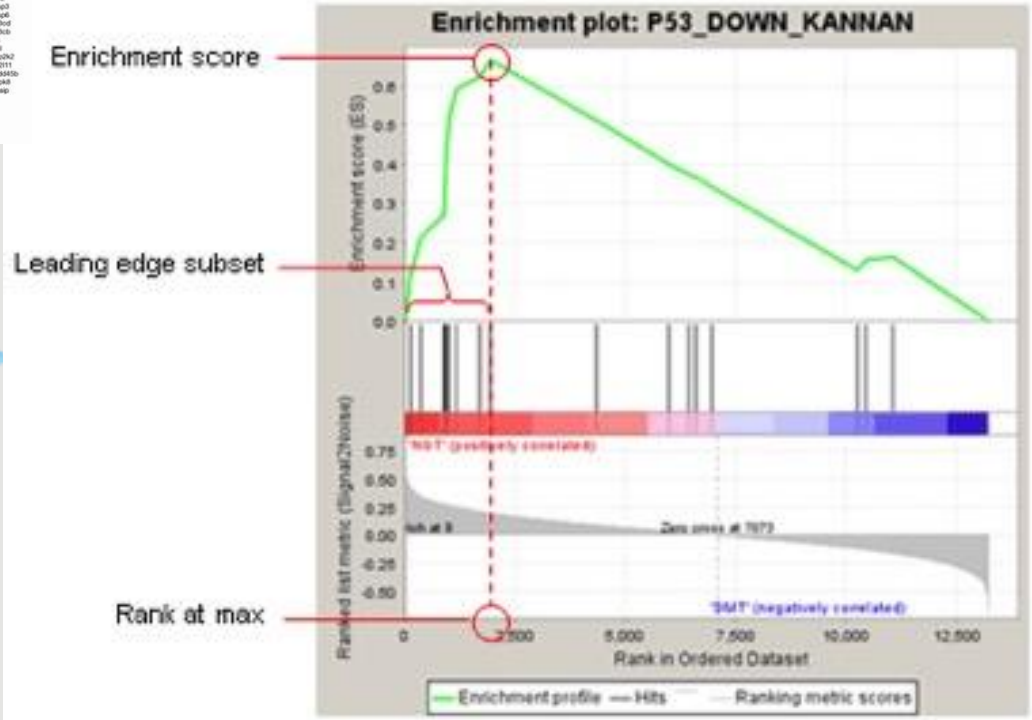
C5 **ontology gene sets** consist of genes annotated by the same ontology term.

C6 **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.





Let's start to run the codes…

```
# STAR
STAR --twopassMode Basic --outSAMtype BAM Unsorted --outFilterMultimapNmax 1 --chimSegmentMin 20 --quantMode TranscriptomeSAM --genomeDir /path_to_reference_star_index
--sjdbGTFfile /path_to_reference_gtf --outFileNamePrefix sample_name --readFilesIn /input_sample_R1.fastq.gz /input_sample_R2.fastq.gz --runThreadN 32 --readFilesCommand zcat

# RSEM
rsem-calculate-expression --quiet -p 32 --bam --no-bam-output --paired-end <sample_name_Aligned.toTranscriptome.out.bam> <path_to_reference_rsem_index> <sample_name>

# SORT BAM file -- for IGV or further analysis
samtools sort -@ 32 -o sample_name_Aligned.out .sorted.bam sample_name_Aligned.out.bam
samtools index sample_name_Aligned.out .sorted.bam

# edgeR – for DEG
read.table(input_data, row.names = 1, header = TRUE, sep = ",", quote = "", check.names=FALSE)
DGEList(counts= exp[,sample_meta], group=sample_group)
calcNormFactors(alldata)
design <- model.matrix(~compgrp)
exactTest(all.deg.data)
decideTestsDGE(degresult, p.value = degpv, lfc = log2(degfc), adjust.method = padjust)

# PCA
PCA(t(lcpm), graph = FALSE)

# Heatmap
pheatmap(lcpm, cex=1, scale = "row", show_rownames = FALSE)
```

```
# volcano plot
# volcanoplot
ggplot(data = degalltable(), aes(x = logFC, y= -log10(Pvalue), col = DGEtest)) +
geom_point() +
   theme_minimal() + scale_color_manual(values = color) +
   theme(axis.title.x = element_text(size = 18),
         axis.text.x = element_text(size = 15),
         axis.title.y = element_text(size = 18),
         axis.text.y = element_text(size = 15),
         legend.text = element_text(size = 15),
         legend.title = element_text(size = 18))

# K-means
kmeans(dt, kvalue, iter.max = 50)

# Enrichment
library(clusterProfiler)
enrichGO(glist, OrgDb = 'org.Mm.eg.db', ont="BP", keyType = "ENTREZID",
pAdjustMethod = "none", pvalueCutoff = 0.001, qvalueCutoff = 0.05)
enrichKEGG(glist, organism = "mmu", keyType = "ncbi-geneid", pAdjustMethod =
"none", pvalueCutoff = 0.05, qvalueCutoff = 0.05)
```

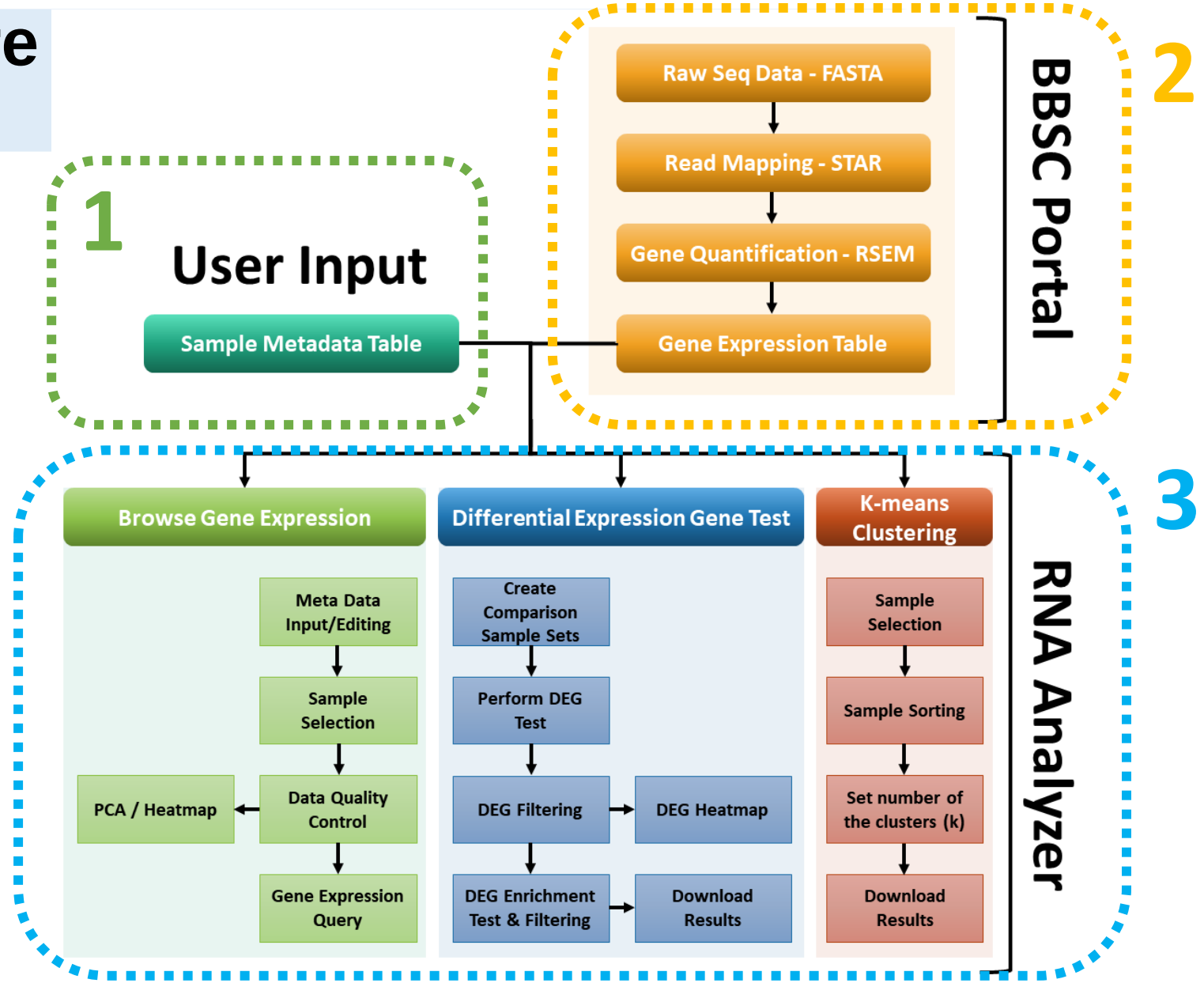
Outline

1. Experimental design and practical considerations

2. Differential gene expression analysis pipeline

3. IMB Bioinformatics Core analysis tools (DEMO)

IMB Bioinformatics Core
Analysis pipeline/tool



Samples metadata

Example:

sample	group	time	replicate	genotype
wt_0_rep1	wt_0	0	rep1	wt
wt_0_rep2	wt_0	0	rep2	wt
wt_1_rep1	wt_1	1	rep1	wt
wt_1_rep2	wt_1	1	rep2	wt
ko_0_rep1	ko_0	0	rep1	ko
ko_0_rep2	ko_0	0	rep2	ko
ko_1_rep1	ko_1	1	rep1	ko
ko_1_rep2	ko_1	1	rep2	ko

Access to BBSC servers

useful tools:

PC: command prompt / MobaXterm / Putty

Linux Terminal

Mac Terminal

Command: **ssh N123@bc.imb.sinica.edu.tw**

Command: **ssh bbsc200**

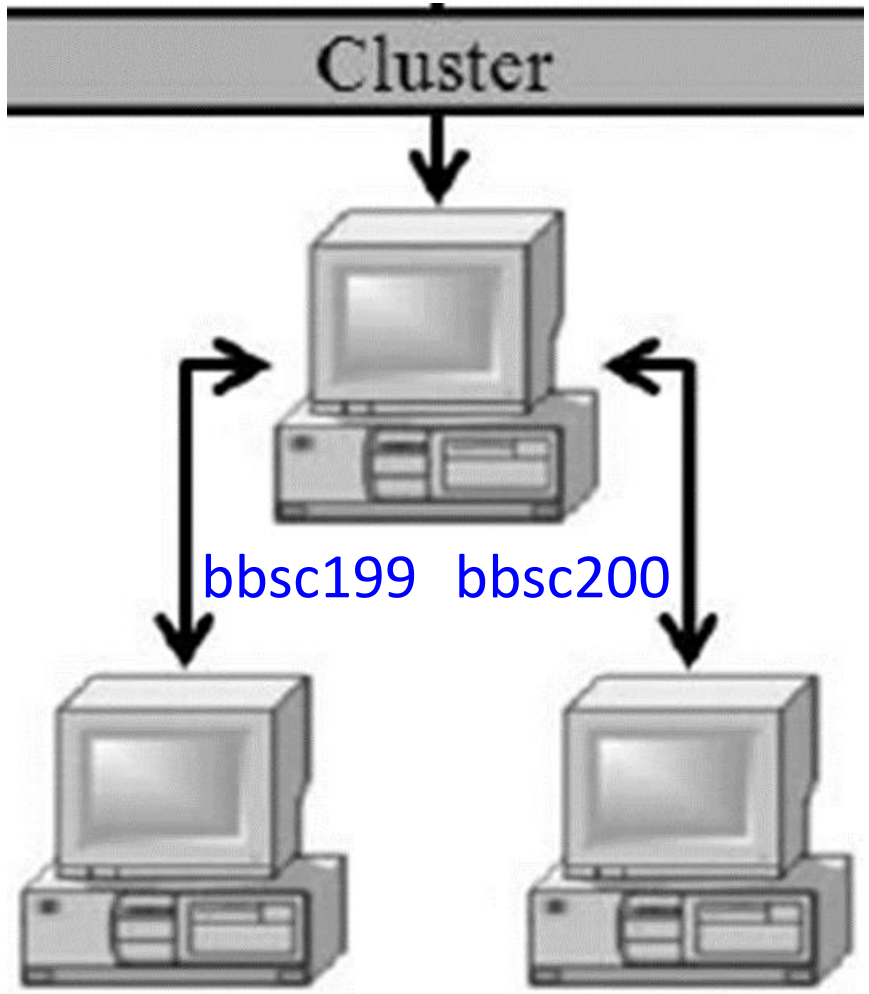
Port: 22

Login: IMB username and password (Contact Bioinformatics Core or PI if you don't know it)

Access to BBSC servers

```
                    • MobaXterm 20.1 •
            (SSH client, X-server and networking tools)

 ➤ SSH session to bbsc@192.168.100.158
   • SSH compression : ✔
   • SSH-browser     : ✔
   • X11-forwarding  : ✔  (remote display is forwarded through SSH)
   • DISPLAY         : ✔  (automatically set on remote server)

 ➤ For more info, ctrl+click on help or visit our website


Welcome!

We provide the following computer servers for data analysis:

bbsc197 (AMD Ryzen 128 cpus, 256 GB memory, GTX 3080 Ti 12GB)
bbsc198 (AM(AMD Ryzen 256 cpus, 512 GB memory, GTX 3090 24GB) **NeuroScience-Labs first**
bbsc199 (Intel Xeon 120 cpuss, 512 GB memory)
bbsc200 (AMD Ryzen 128 cpus, 256 GB memory, GTX 2070 Super 8GB)

1. Please use the command 'ssh bbscXXX' to login to a server and run your tasks on that server.
2. The maximum number of threads (CPUs) for each command is 16 (with total memory 64GB).
3. If you need more CPUs, please contact bioinformatics core (ext.9965/9967) first.
4. Please DO NOT run any time/space consuming commands on bbsc158 (bc) or bbsc159 (bc2). Thank you!!

Have a nice day!

You have new mail.
Last login: Tue Nov  2 14:41:21 2021 from 192.168.100.206
bbsc@bbsc158:~$ ssh bbsc200
Last login: Tue Nov  2 12:25:30 2021 from 192.168.100.159
bbsc@bbsc200:~$
```

ssh bc.imb.sinica.edu.tw



Run BBSC Portal

Command: **bbsc_portal**

```
****** Welcome to BioInformatics Core Portal ******
    1. RNA-Seq data analysis.
    2. De novo sequence assembly.
    3. Nanopore data manipulation.
    4. Genomic variant calling.
    5. 10x data analysis.
    6. protein data analysis.
    X. Exit.

Your choice:
```

Run BBSC Portal

```
******* RNA-Seq data analysis *******
        1. STAR + RSEM (single sample).
        2. STAR + RSEM (multiple samples).
        3. DESeq2 analysis.
        b. Back to main menu.
        x. Exit.

Your choice:
```

Run BBSC Portal

```
Your choice:2
Reference options: hg38, hg38_lncRNA, mm10, mm39, ecoli, drosophila, rice-MSU7,
rice-RAP, arabadosis11, c_elegans, yeast, danio_rerio_ncbi, danio_rerio_ensembl
Select a reference:mm39
read folder:raw_data
quantify gene expression levels (y/n)?y
run geneBody_coverage (y/n)?n
Number of threads:2
Machine options: bbsc198, bbsc199, bbsc200, localhost, auto, cmd
Select a machine:bbsc200
```

Enter the path of your raw seq data – FASTA files

Choose the number of nodes to run the analysis

Outputs

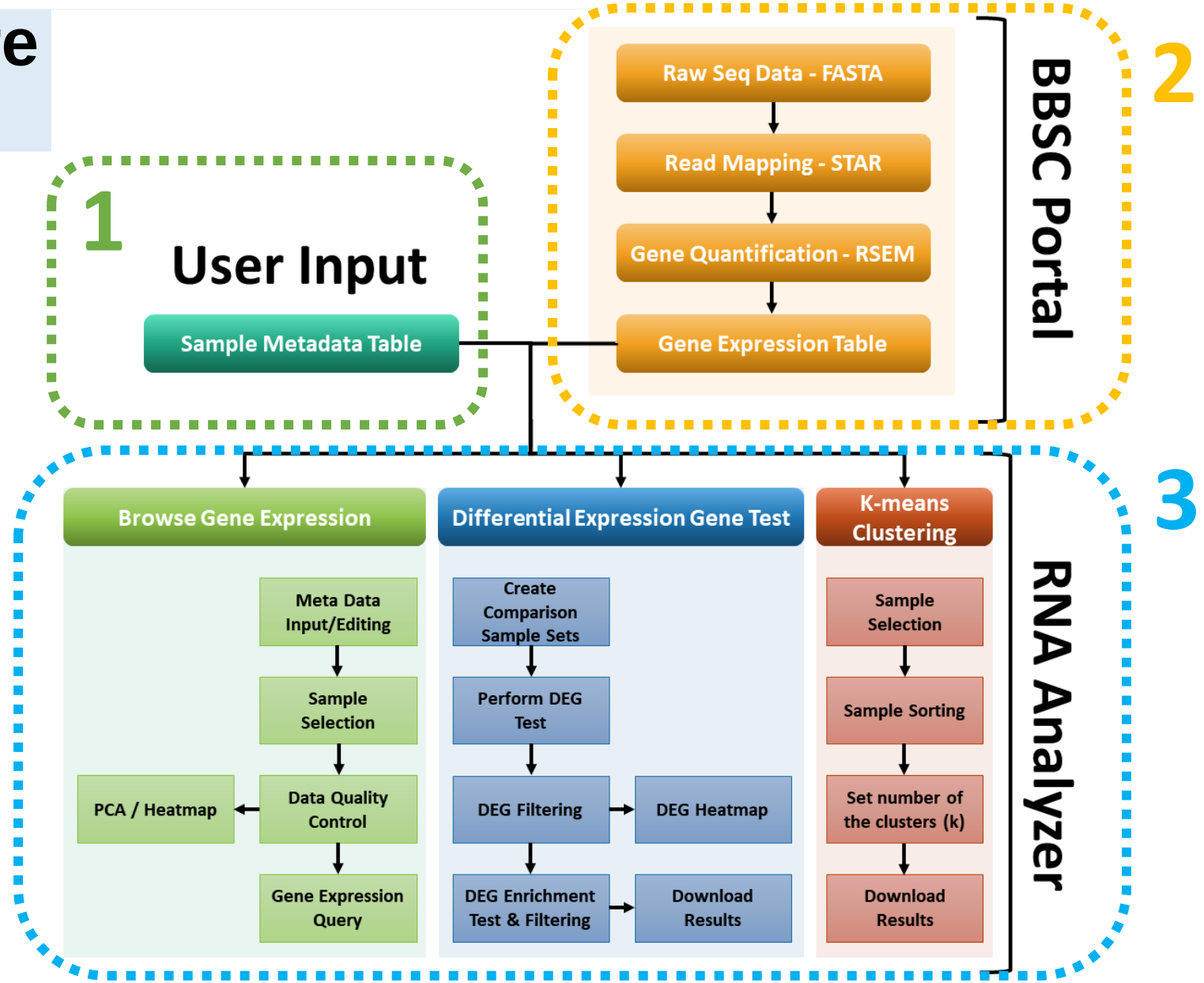
sample_expression.tsv (expected raw read count)

gene	ko_0_rep1	ko_0_rep2	ko_1_rep1	ko_1_rep2	wt_0_rep1	wt_0_rep2	wt_1_rep1	wt_1_rep2
gene1	292	210	197	209	443	304	287	242
gene2	25	40	15	17	28	13	21	24
gene3	827	752	807	691	484	435	364	406
…	117	381	153	107	179	134	99	82

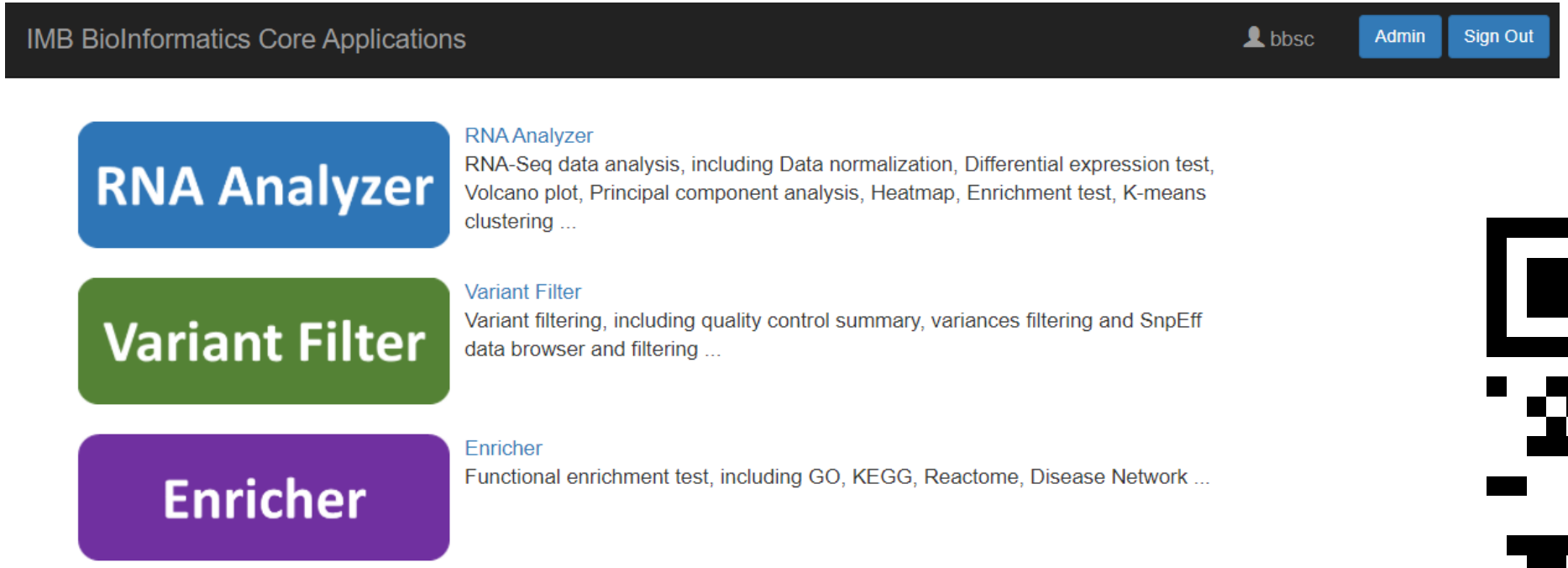
sample_metadata.tsv

sample	group	time	rep	genotype
wt_0_rep1	wt_0	0	rep1	wt
wt_0_rep2	wt_0	0	rep2	wt
wt_1_rep1	wt_1	1	rep1	wt
wt_1_rep2	wt_1	1	rep2	wt
ko_0_rep1	ko_0	0	rep1	ko
ko_0_rep2	ko_0	0	rep2	ko
ko_1_rep1	ko_1	1	rep1	ko
ko_1_rep2	ko_1	1	rep2	ko

IMB Bioinformatics Core
Analysis pipeline/tool



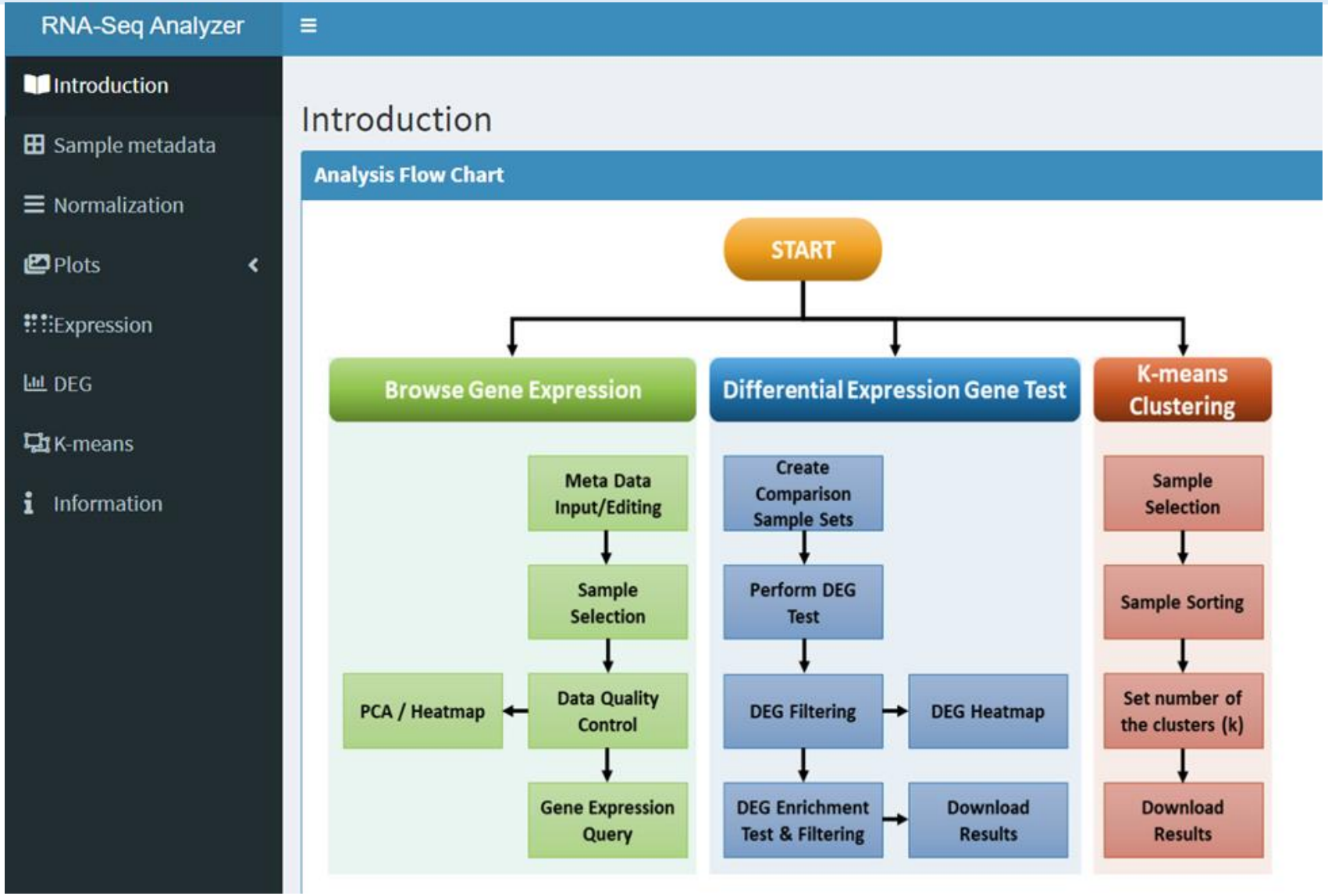
IMB Bioinformatics Core APPs



Bioinformatics Apps

http://bc.imb.sinica.edu.tw:8080

RNA Analyzer (DEMO)



References – RNA-Seq Analysis Tutorials

Harvard Chan Bioinformatics Core:
https://hbctraining.github.io/Training-modules/planning_successful_rnaseq

Babraham Bioinformatics:
https://www.bioinformatics.babraham.ac.uk/training.html#rnaseq

F1000Research
https://f1000research.com/articles/5-1408

Cancer Research UK Cambridge Institute
https://bioinformatics-core-shared-training.github.io/RNAseq-R/

Thanks for your attention!!

https://bc.imb.sinica.edu.tw

If you have any problems in Bioinformatics, please feel free to stop by N419