RNAseq Analyses

POST-SEQUENCING STEPS

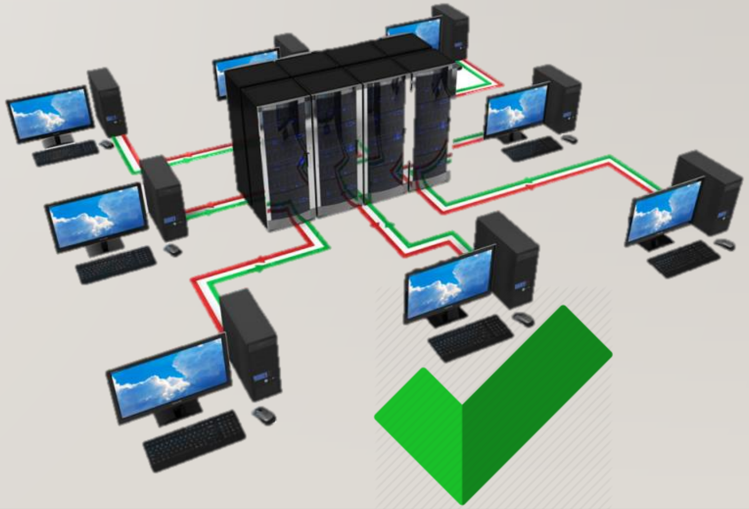
- ERNEST ALICHE

WHY RNAseq?

- Differential gene expression between conditions
- Alternative splice variants
- *De Novo* discovery of putative gene(s)
- Transcriptome-wide sampling of all genes
- Absence of house-keeping gene biases

THINGS TO TAKE INTO ACCOUNT

- Highly expressed genes could mask lowly expressed genes
  - Sufficient sequencing depth required to detect lowly expressed genes
  - Genome size and extent of sequence repeats (e.g. transposable elements) affect depth
- Batch effects
  - May affect different sequencing runs
  - May be factored in the analyses model
- Computing power challenge



HANDLING YOUR RNAseq BIG DATASET

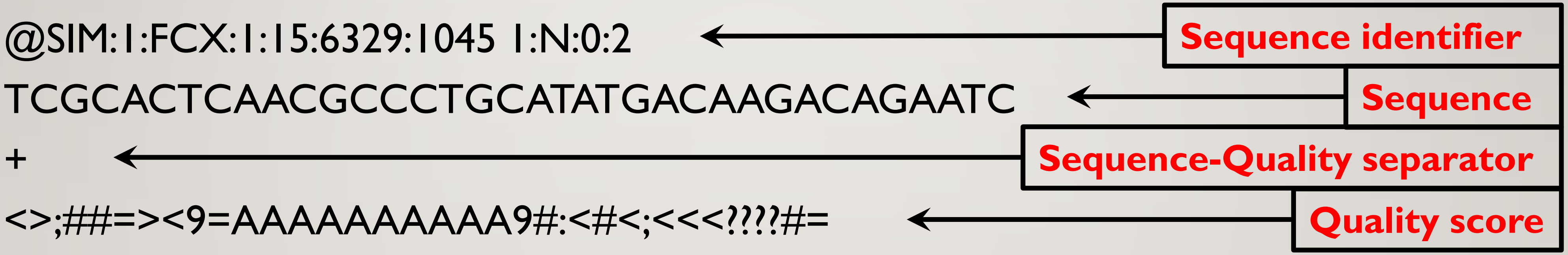
- Knowledge of your data structure (variables, factors, covariables…)
  - Psychological control
  - Confidence to go deeper

- genseq-h0.science.uva.nl/download/MAD1249

HOW'S THE BIG DATA PROJECT COMING ALONG, HOSKINS?
© D.Fletcher for CloudTweaks.com

THE RAW SEQUENCE DATA

- **Fastq_sequence**

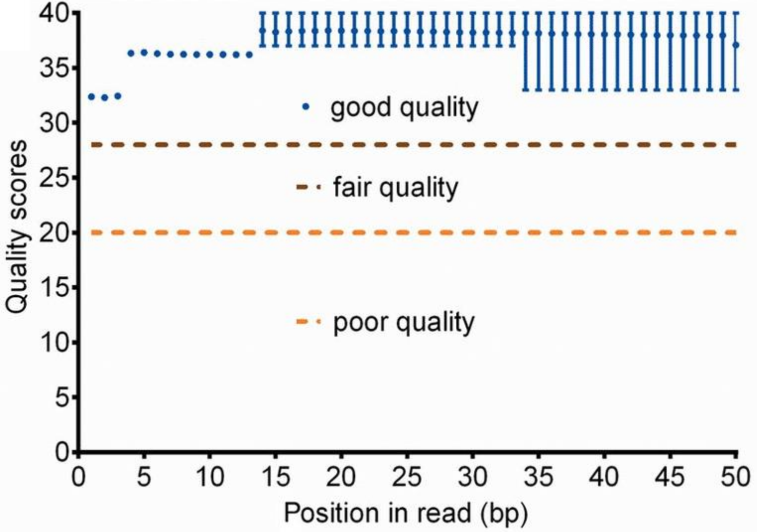


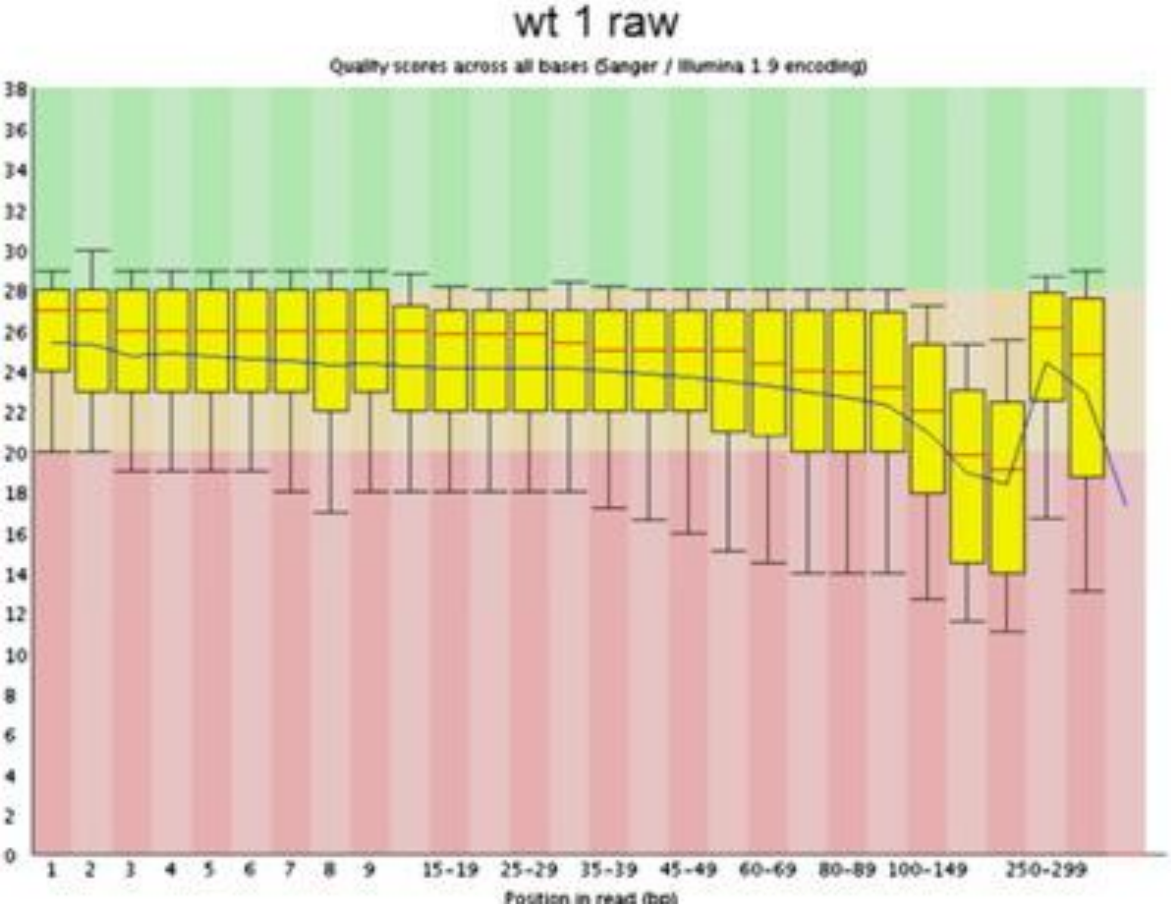
```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

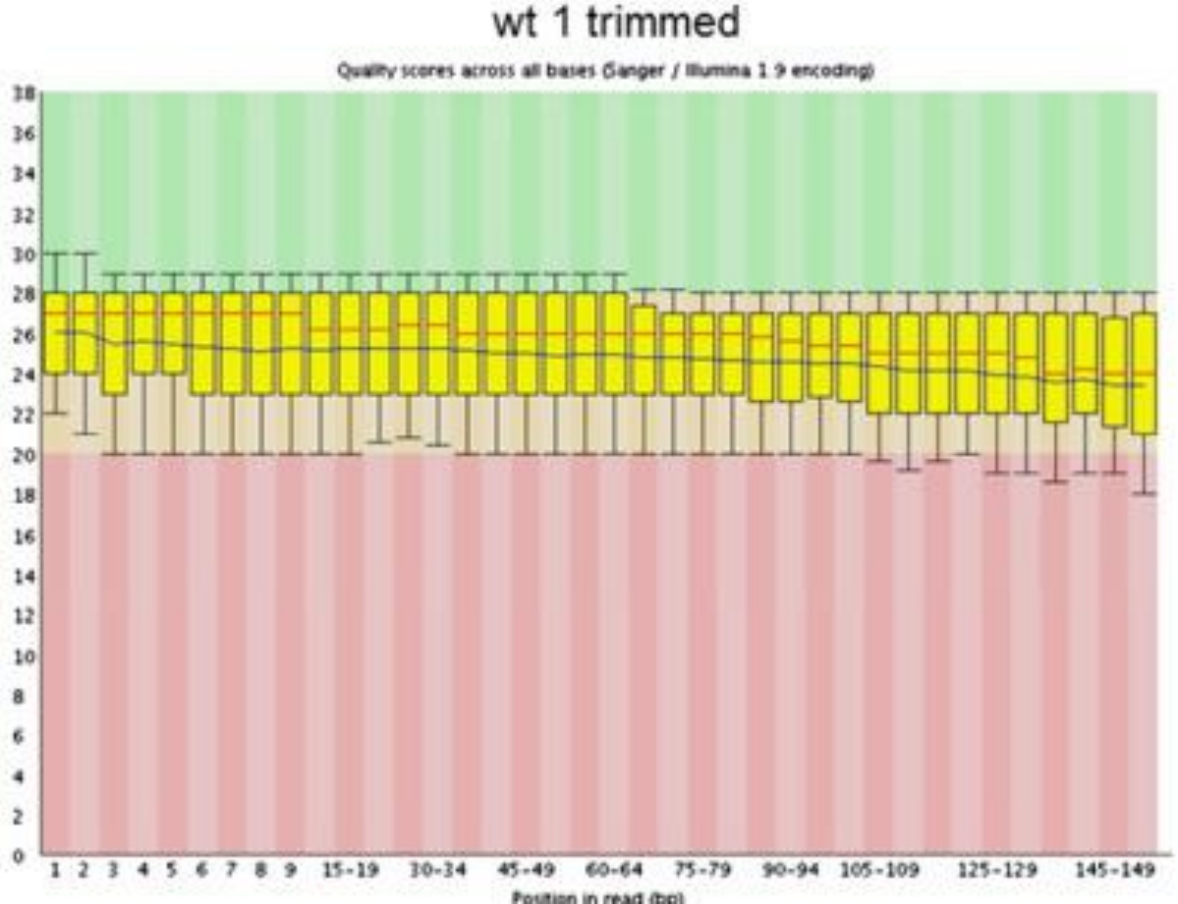
- **Adapter sequences**
    - Used in binding barcoding sequences and for immobilizing the fragments to the flow-cell
    - Removal: Cutadapt, Flexible Adapter Remover (FAR), Adapterremoval, Trimmomatic, FASTQ Clipper, PRINSEQ, ea-utils

QUALITY CONTROL (QC)

- NGS QC Toolkit, FastQC, Trim Galore!







PREPARING YOUR REFERENCE SEQUENCE

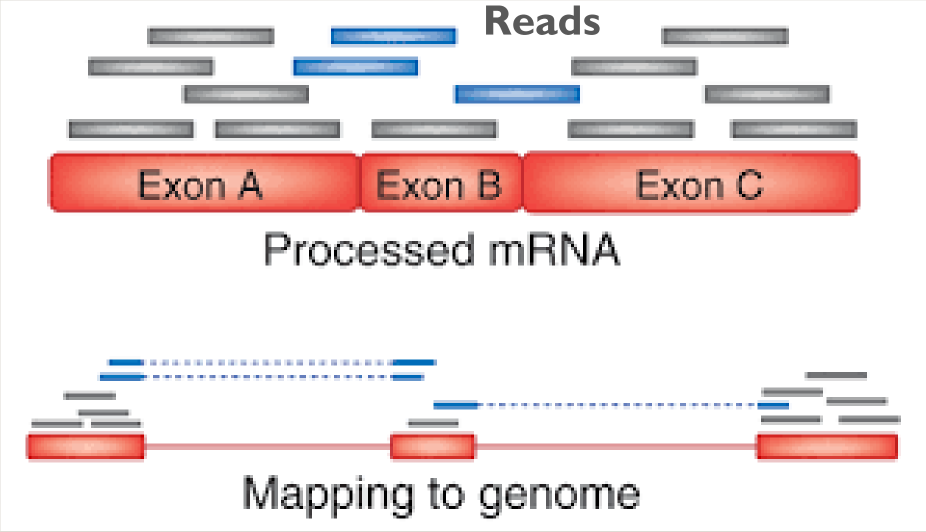
- Download reference genome (.fa and .gtf/.gff)
  - Confirm what genome version to download…
  - Ensembl plants, Gramene, etc.
- Build genome index files

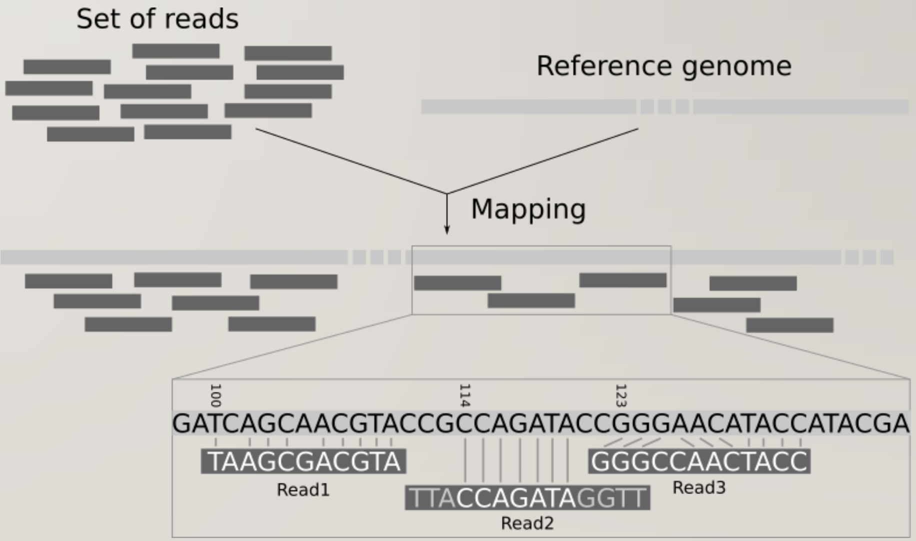
e!EnsemblGenomes

Gramene

READ ALIGNMENT TO REFERENCE

- Mapping reads to the reference genome
- Read mapping at exon-exon junctions





READS COUNT & DIFFERENTIAL EXPRESSION

- Various approaches depending on package used: HISAT2, DESeq2, edgeR, …

  Normalization for sequencing depth

  Normalization for Gene Length

  - RPKM (single-end sequencing) → $RPM = \frac{ReadCounts\ per\ sample}{1{,}000{,}000}$ → $RPKM = \frac{RPM}{GeneLenght\ in\ Kb}$
  - FPKM (paired-end sequencing) → $FPM = \frac{FragmentAbundance}{1{,}000{,}000}$ → $FPKM = \frac{FPM}{GeneLength\ in\ Kb}$

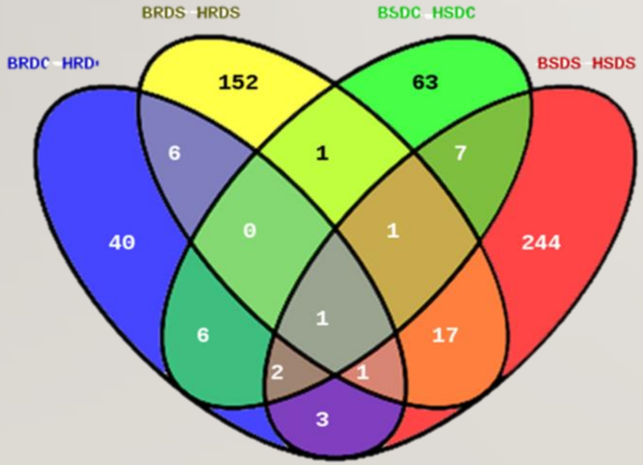
  Normalization for Gene Length

  Normalization for sequencing depth

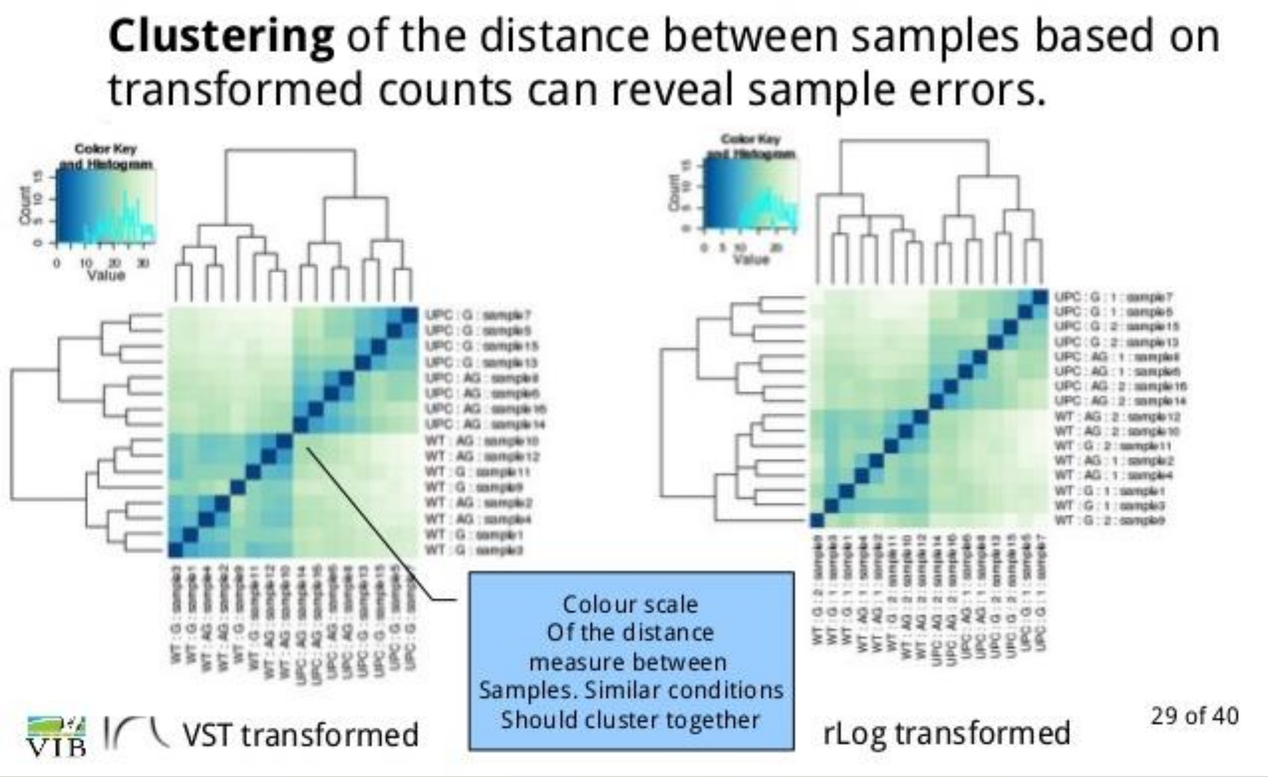
  - TPM (transcripts per Kilobase million…) → $RPK = \frac{ReadCounts}{GeneLength}$ → $TPM = \frac{RPK}{1{,}000{,}000}$
  - Raw counts (from e.g. Rsubread)

DATA & RESULTS VISUALIZATION

- Heat maps & PCAs
- Volcano plots
- Venn diagrams



- http://bioinformatics.psb.ugent.be/webtools/Venn/


Clustering of the distance between samples based on transformed counts can reveal sample errors.

Colour scale
Of the distance
measure between
Samples. Similar conditions
Should cluster together

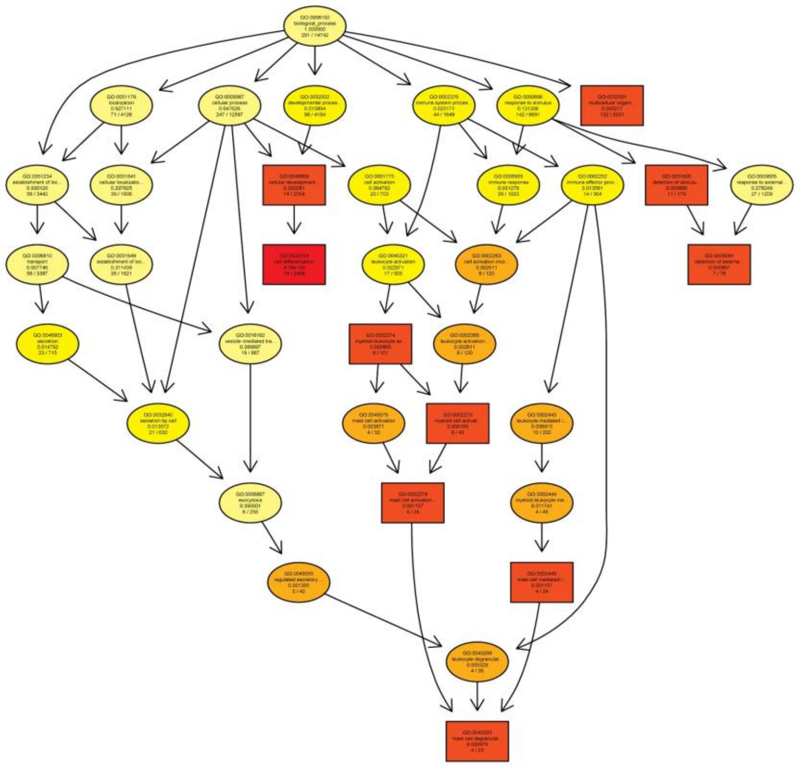
VST transformed

rLog transformed

29 of 40

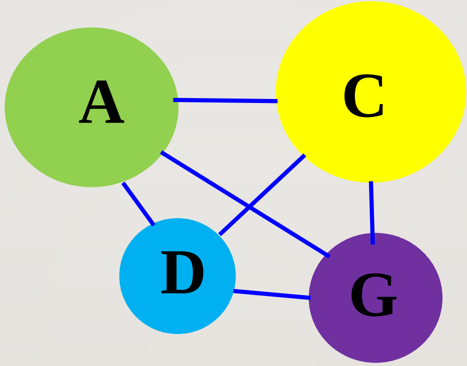
WHAT NEXT???

- Depends on your research questions (curiosity), number of DEGs, available time…
  - **Functional Annotation**
    - **agriGO** **http://bioinfo.cau.edu.cn/agriGO/analysis.php**
    - **PlantRegMap** **http://plantregmap.cbi.pku.edu.cn/go.php**
    - **Goseq** **https://bioconductor.org/packages/release/bioc/html/goseq.html**
    - …

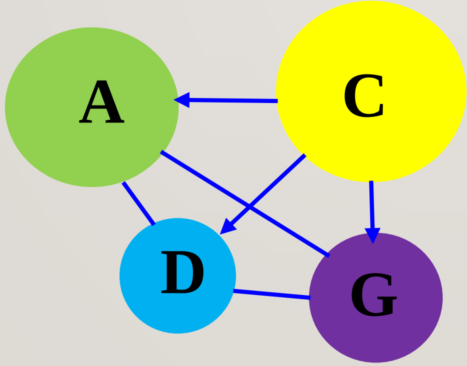


WHAT NEXT???

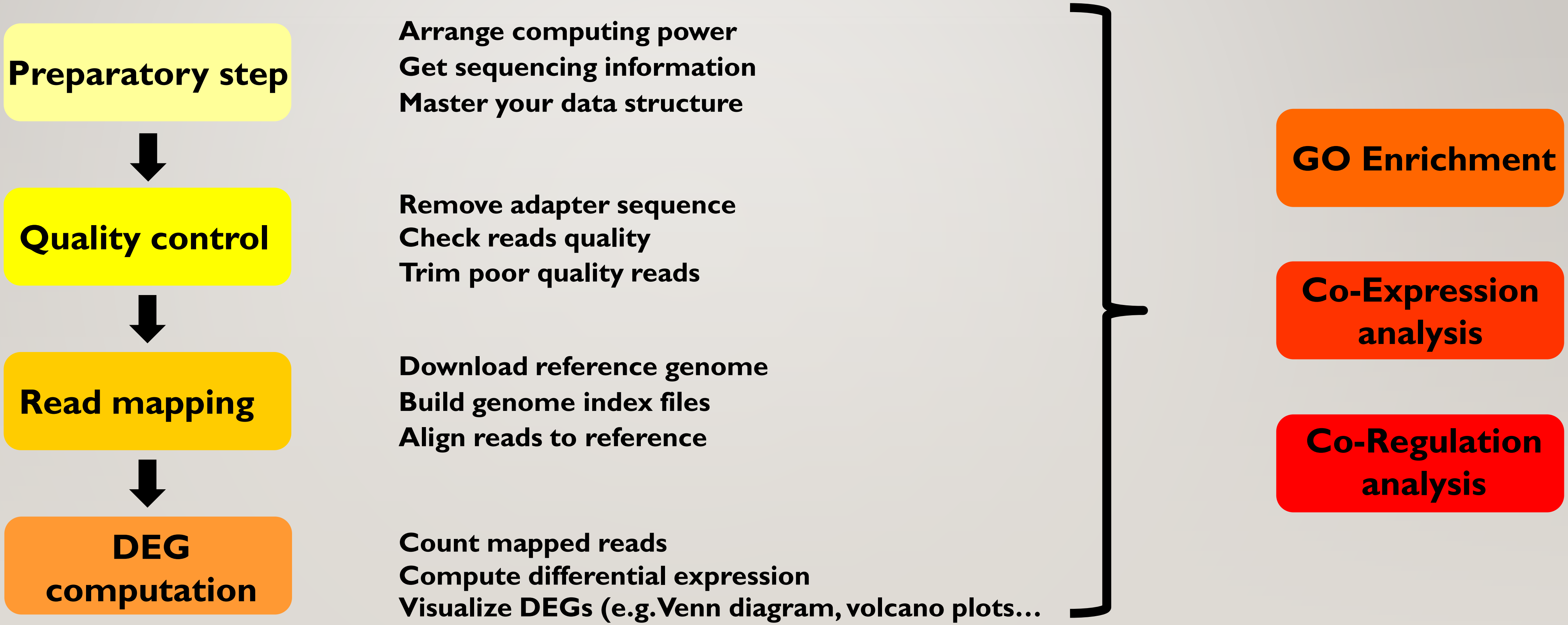
- Co-Expression analyses



- Regulatory network analyses



WORKFLOW SUMMARY



Preparatory step

Arrange computing power
Get sequencing information
Master your data structure

Quality control

Remove adapter sequence
Check reads quality
Trim poor quality reads

Read mapping

Download reference genome
Build genome index files
Align reads to reference

DEG computation

Count mapped reads
Compute differential expression
Visualize DEGs (e.g.Venn diagram, volcano plots…

GO Enrichment

Co-Expression analysis

Co-Regulation analysis

QUESTIONS