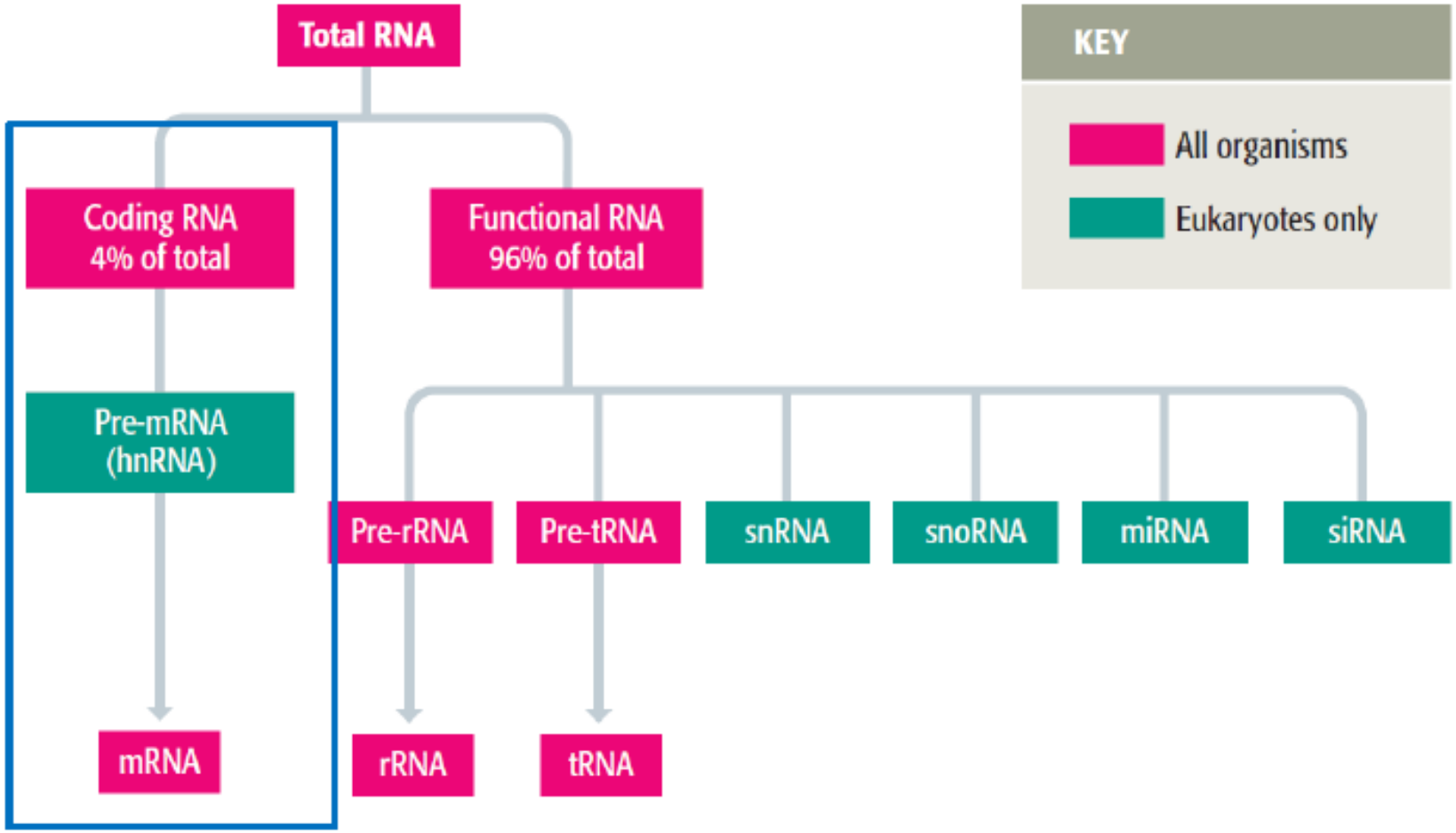
Differentially Gene Expression Analysis using Galaxy

Patipark Kueanjinda, Ph.D.

SIPM610 Systems Pharmacology Lab I

RNA contents of cell



rRNA: 80-90%, tRNA: 5-15%, mRNA: 2-4%, ncRNA: 1%

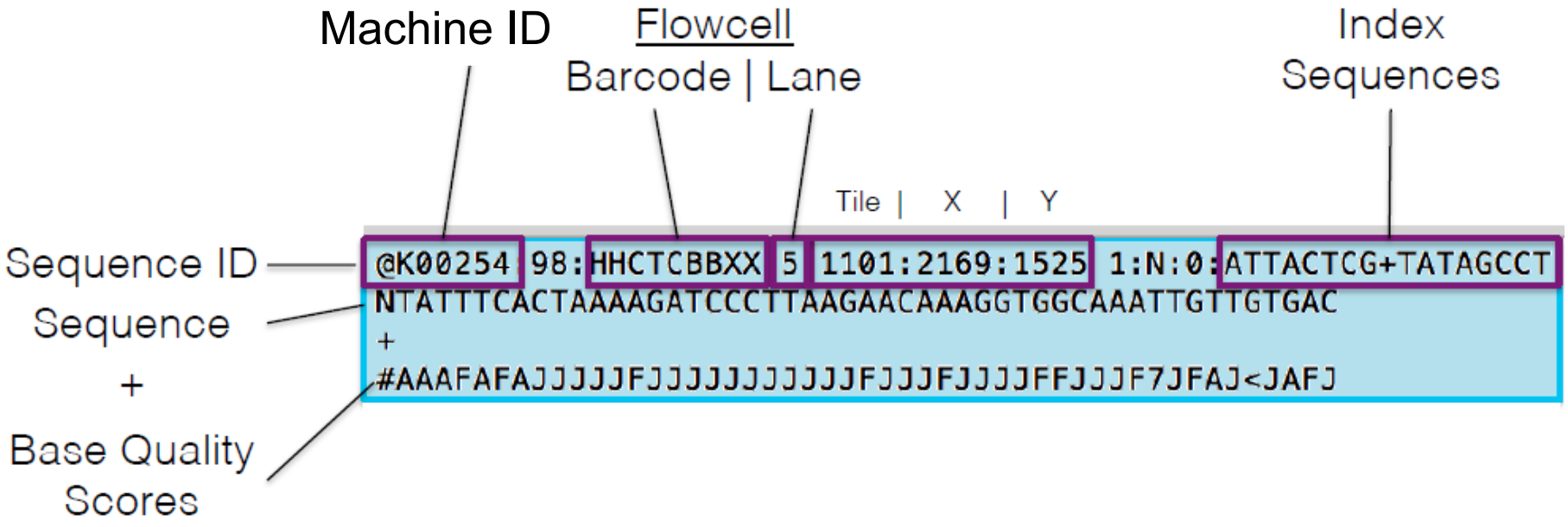
Sequencing technologies

RNA sequencing

- Use ultra high-throughput sequencing (next generation sequencing) technology.
- Many applications available.
  - Differential gene expression.
  - Transcript discovery.
  - Splice variants.
  - Allele-specific expression.

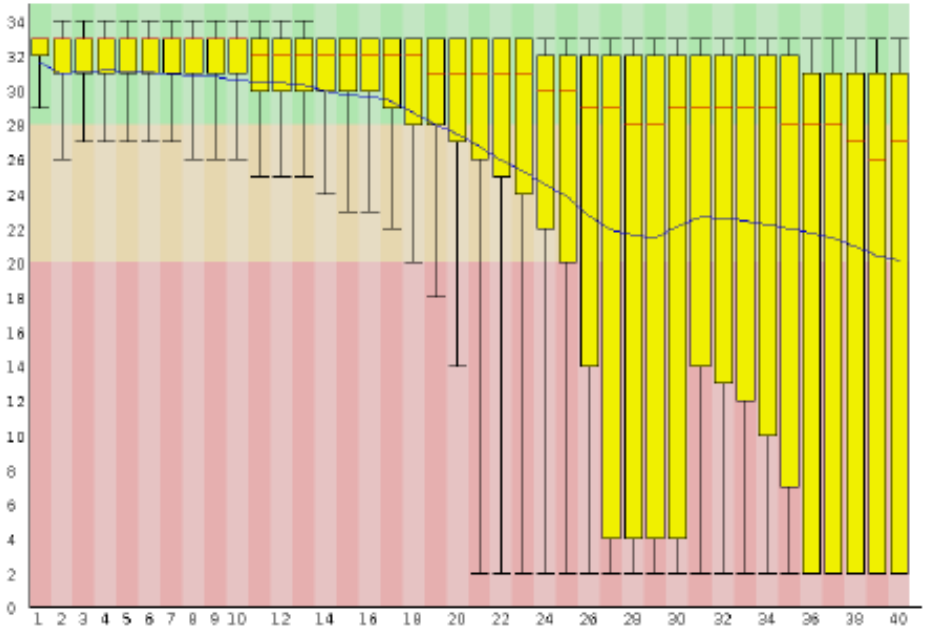
Raw sequence reads

- Large text file (millions of lines) with simple format.
- Most frequently used is FASTA format for storing the sequences or FASTQ format for storing both the sequence and quality scores.



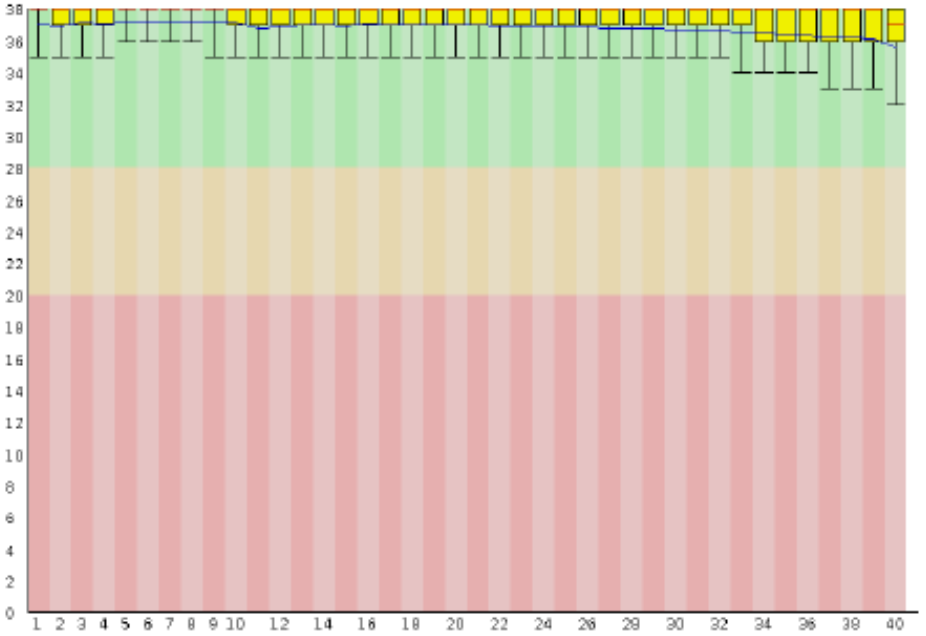
Quality control: per-base quality plot

Per base quality score across all bases



Position in read (bp)

Per base quality score across all bases

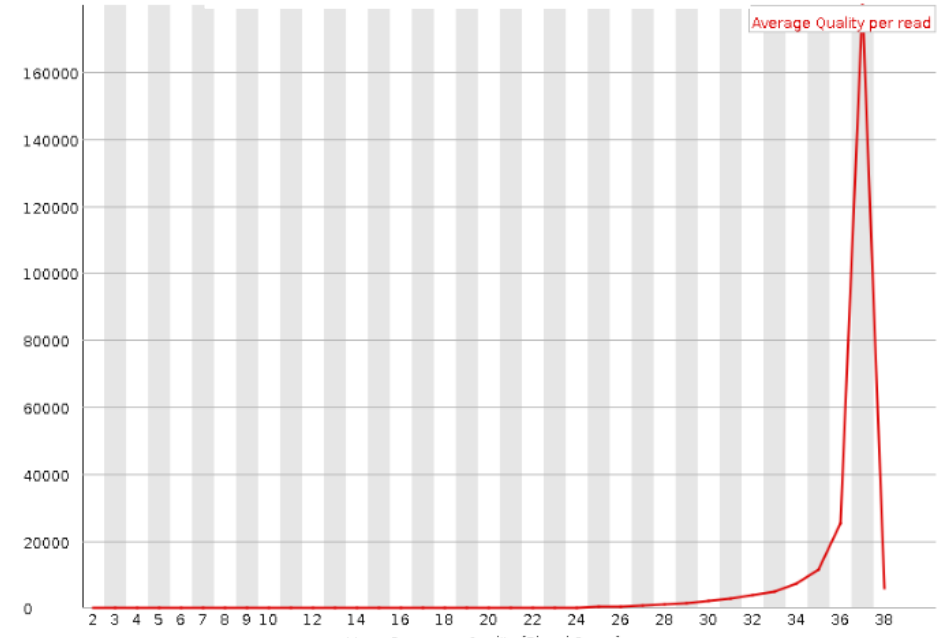


Position in read (bp)

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

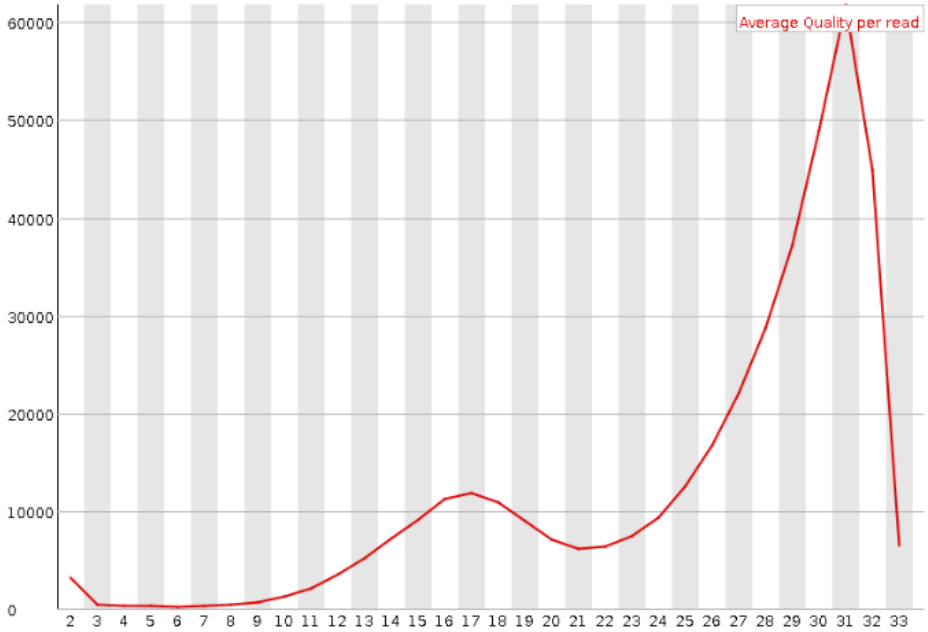
Quality control: per-sequence quality plot

Quality Score Distribution



Mean Sequence Quality (Phred Score)

Quality Score Distribution



Mean Sequence Quality (Phred Score)

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

Phred score

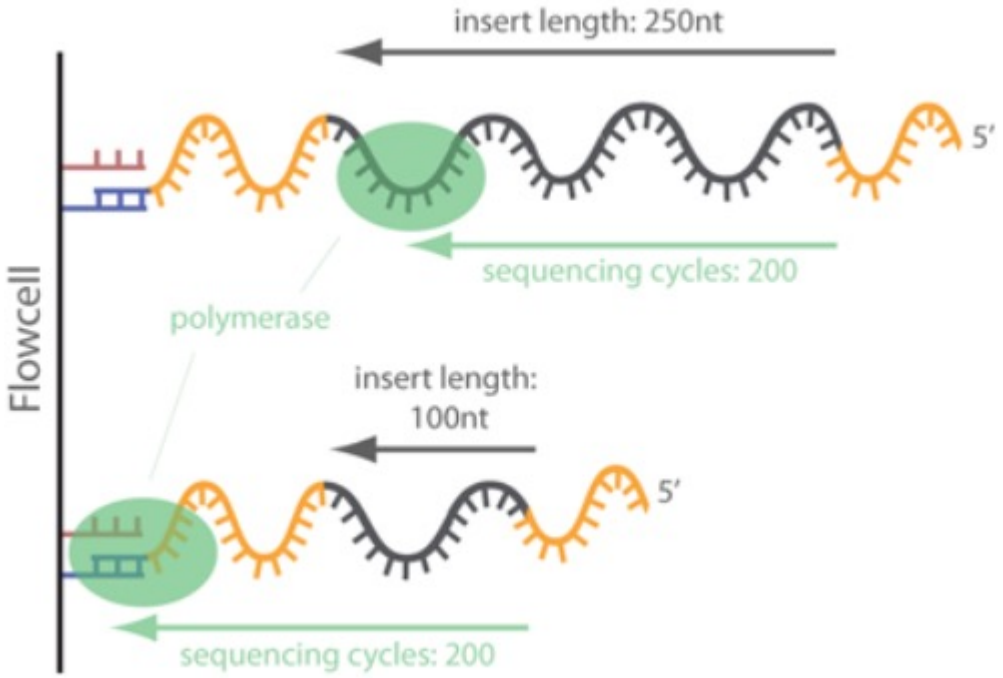
- Each base has certain error proability (p).
- The quality score for each base ranges from -5 to 40.
- ***Qphred = -10 log10(p)***
- Qphred of 20 corresponds to a 99 % probability of a correctly identified base.

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit

- This table allows user to look for adapter contamination.
- Trimming is required.



Summary

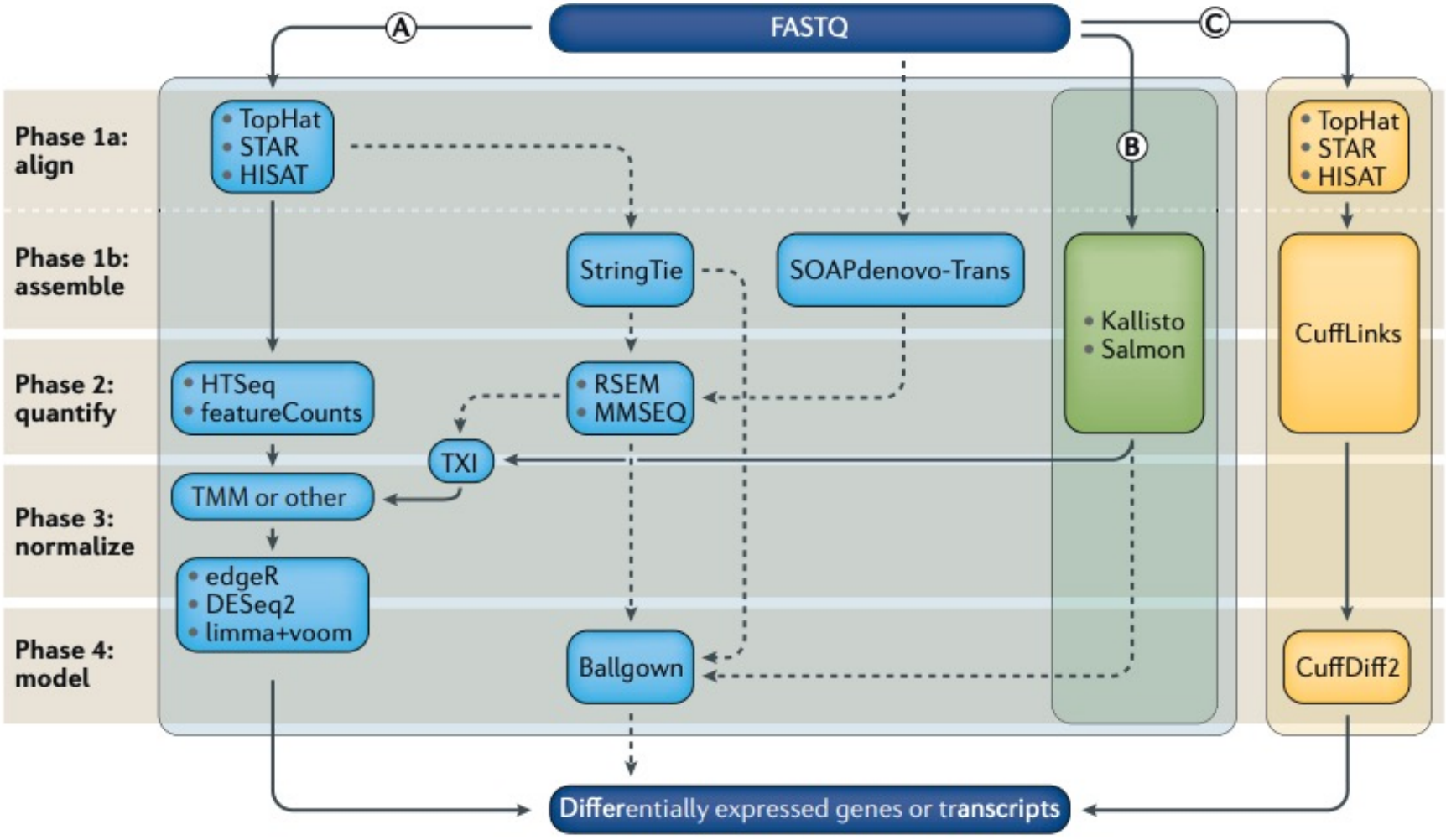
- RNA-Seq analysis is still evolving.
- Differential expression analysis of genes is mature and seems to become standard.
- No analysis tool can replace common sense and knowledge about the biology behind the experiment.
- Garbage in => garbage out.
- More replicates are better investments than more reads.

Bioinformatics: Mapping

Ask yourself these questions

- Do you have reference transcriptome?
  - Yes. You can align reads directly to transcriptome (no gap).
- Are you interested in discovering novel splice junctions?
  - Yes. You have to use reference genome instead.
- Do you have reference genome with gene annotation?
  - No. Your last hope is *de novo* RNA-seq assembler.

Analysis workflow for differential gene expression

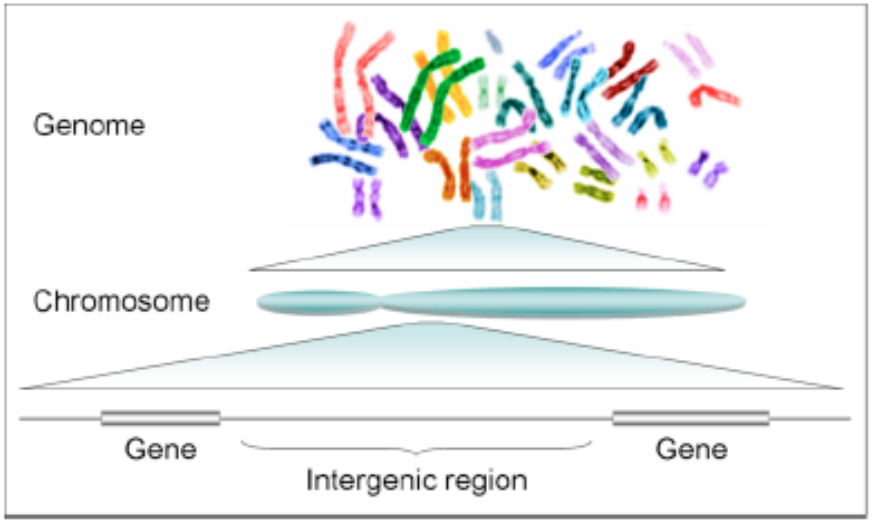


(Stark, Grzelak, and Hadfield, 2019, *Nature Genetics*)

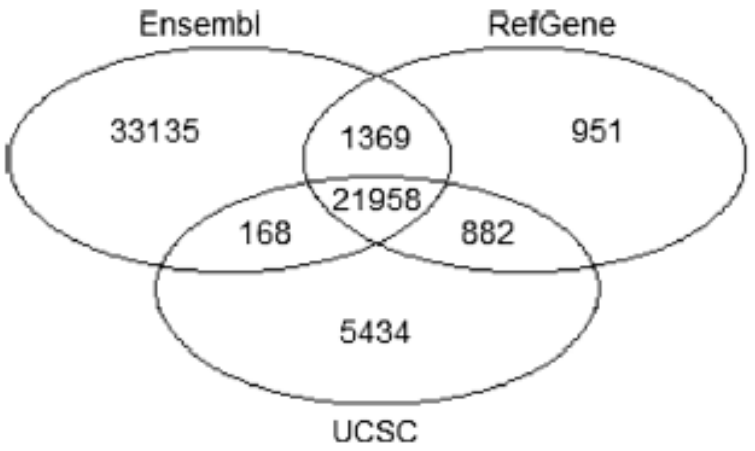
File formats

- RAW read data = FASTQ
- Reference genome or transcriptome = FASTA
- Aligned read data = SAM/BAM
  - Same data.
  - BAM allows sorted and indexted data.
- Genome annotation data = GTF/GFF

Annotation databases



https://commons.wikimedia.org/wiki/File:Human_genome_to_genes.png



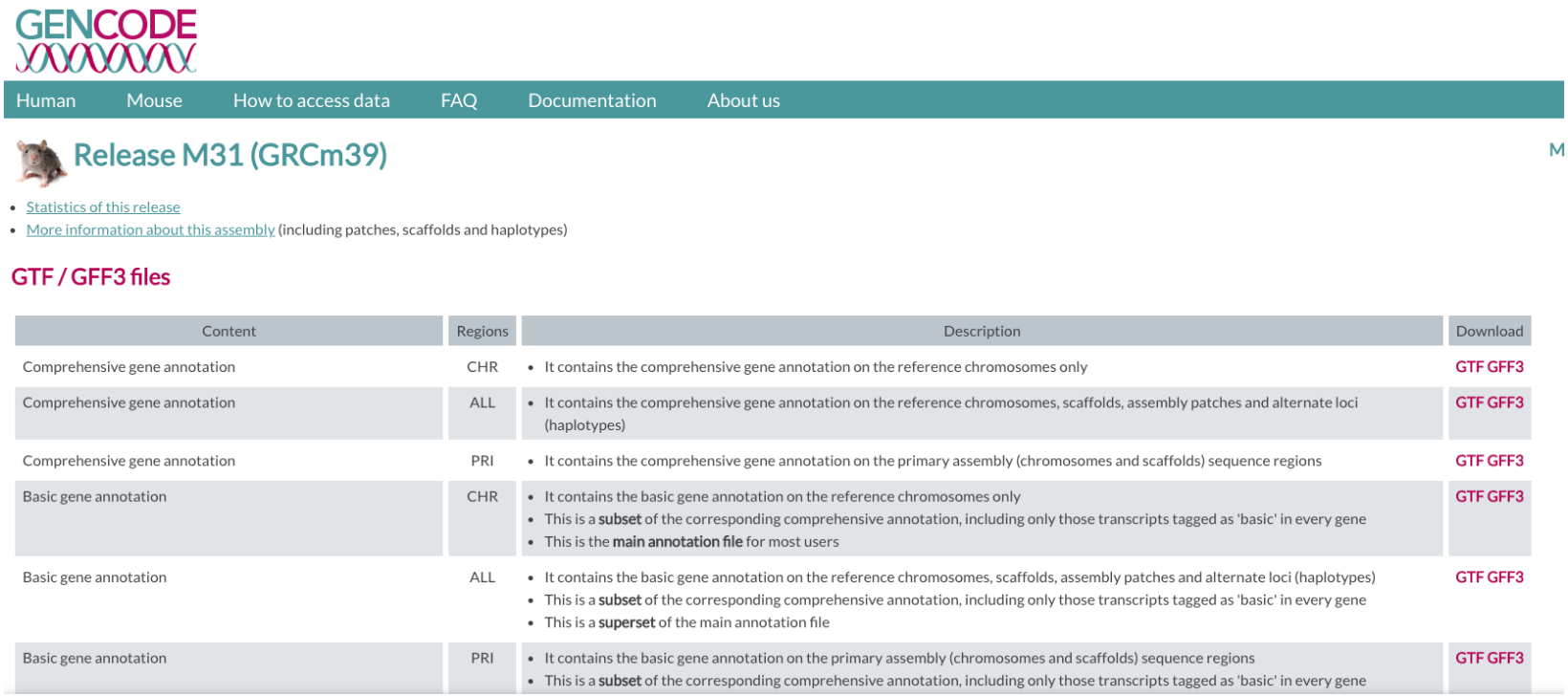
RefSeq ncbi.nlm.nih.gov/refseq

UCSC Known Genes genome.ucsc.edu

Ensembl/Gencode gencodegenes.org

1/3 protein-coding genes
> 17,000 non-coding RNAs
> 15,000 pseudogenes

Gencode databases



GENCODE

Human Mouse How to access data FAQ Documentation About us

Release M31 (GRCm39)

M

- Statistics of this release
- More information about this assembly (including patches, scaffolds and haplotypes)

GTF / GFF3 files

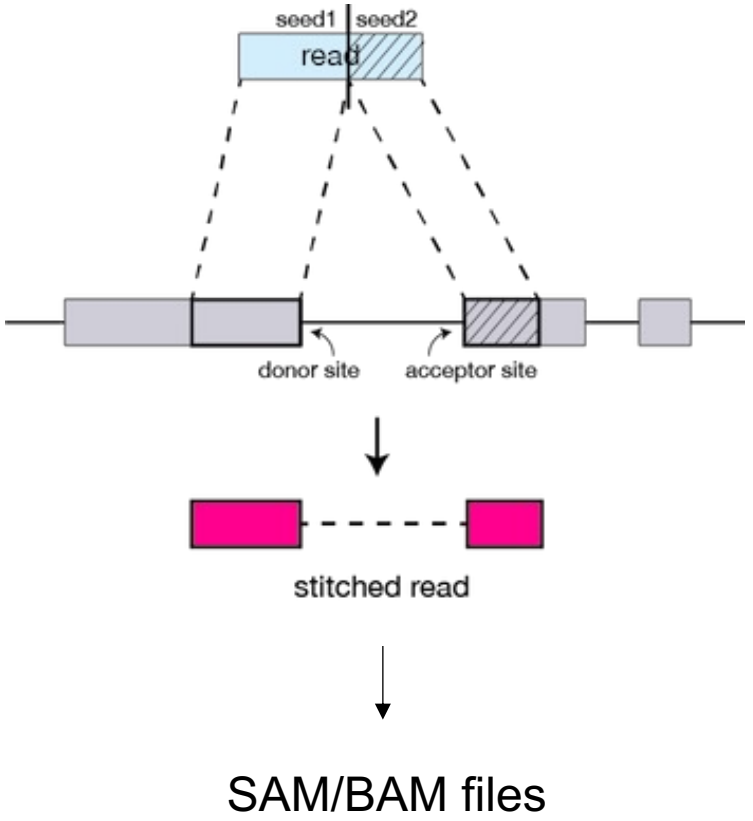
Content	Regions	Description	Download
Comprehensive gene annotation	CHR	• It contains the comprehensive gene annotation on the reference chromosomes only	GTF GFF3
Comprehensive gene annotation	ALL	• It contains the comprehensive gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes)	GTF GFF3
Comprehensive gene annotation	PRI	• It contains the comprehensive gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions	GTF GFF3
Basic gene annotation	CHR	• It contains the basic gene annotation on the reference chromosomes only • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene • This is the **main annotation file** for most users	GTF GFF3
Basic gene annotation	ALL	• It contains the basic gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene • This is a **superset** of the main annotation file	GTF GFF3
Basic gene annotation	PRI	• It contains the basic gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	GTF GFF3

Mouse reference genome file (GFT)
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M31/gencode.vM31.chr_patch_hapl_scaff.annotation.gtf.gz

Mouse reference transcriptome file (fasta)
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M31/gencode.vM31.transcripts.fa.gz

STAR (Spliced Transcriptome Alignment to Reference)

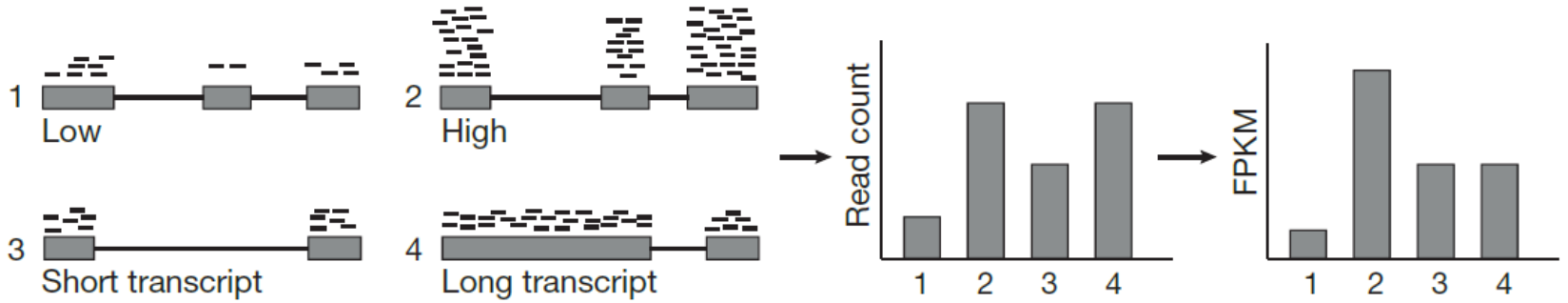
- Accurate and sensitive.
- Very fast (50X faster than HISAT).
- Memory intensive.
- Find maximum matching portion (MMP).
- Reads are split when a continuous alignment is not possible.
- Remaining unmappable portion is aligned again.
- Aligned portions of the original full-length reads are stitched together.



(10.1002/0471250953.bi1114s51)

Bioinformatics: Gene quantification

Gene expression quantification



Transcripts of different lengths with different read coverage levels.

Total read counts observed in each transcript.

FPKM-normalized read counts.

<u>Raw counts != expression levels</u>

Influences on raw counts include:

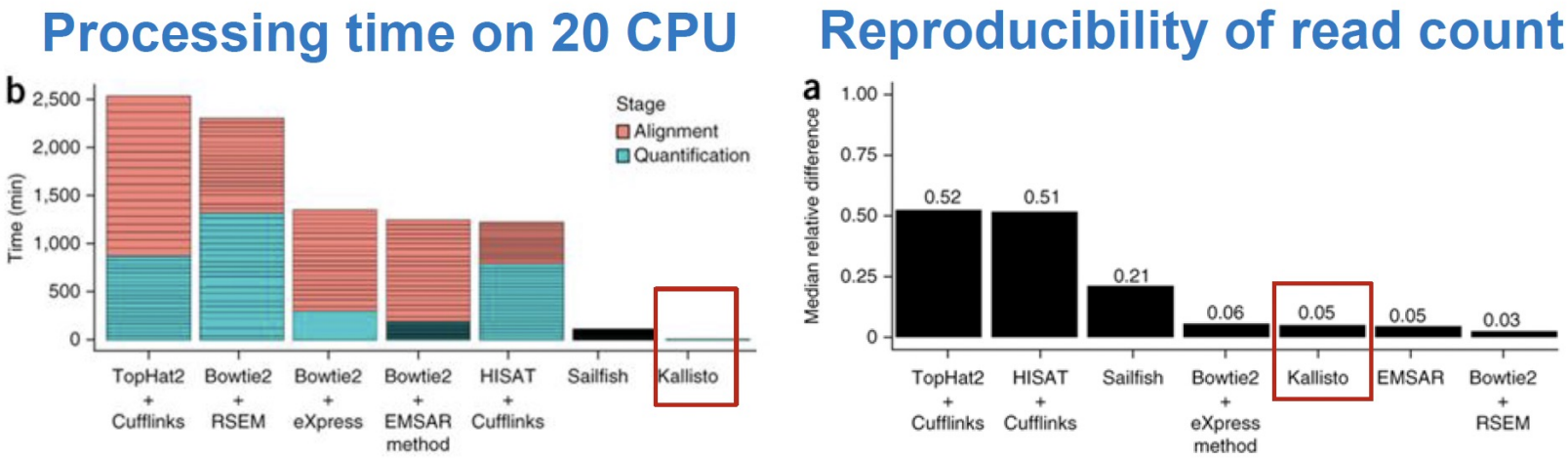
- Gene length
- % GC
- Sequencing depth
- Expression of all other genes in the same sample

} Variation for different genes expressed at the same level.

} Variation for the same gene in different samples.

(Garber et al., 2011)

Kallisto



Bray *et al.* Nat Biotech 34:525-527 (2016)

- Fast in both pseudoaligment and quantification.
- Low memory consumption.
- Generate reproducible transcript quantification.
- Run natively on Windows.

Different expression units

Raw counts: number of reads/fragments overlapping with the union of exons of a gene.

1. RPKM/FPKM (Mortazavi et al., 2008)

- differences in sequencing depth and transcript length.
- compare a gene across samples and diff genes within sample.

2. TMM (Robinson and Oshlack, 2010)

- differences in transcript pool composition and extreme outliers.
- provide better across-sample comparability.

3. TPM (Li et al., 2010; Wagner et al., 2012)

- transcript length distribution in RNA pool.
- provide better across-sample comparability.

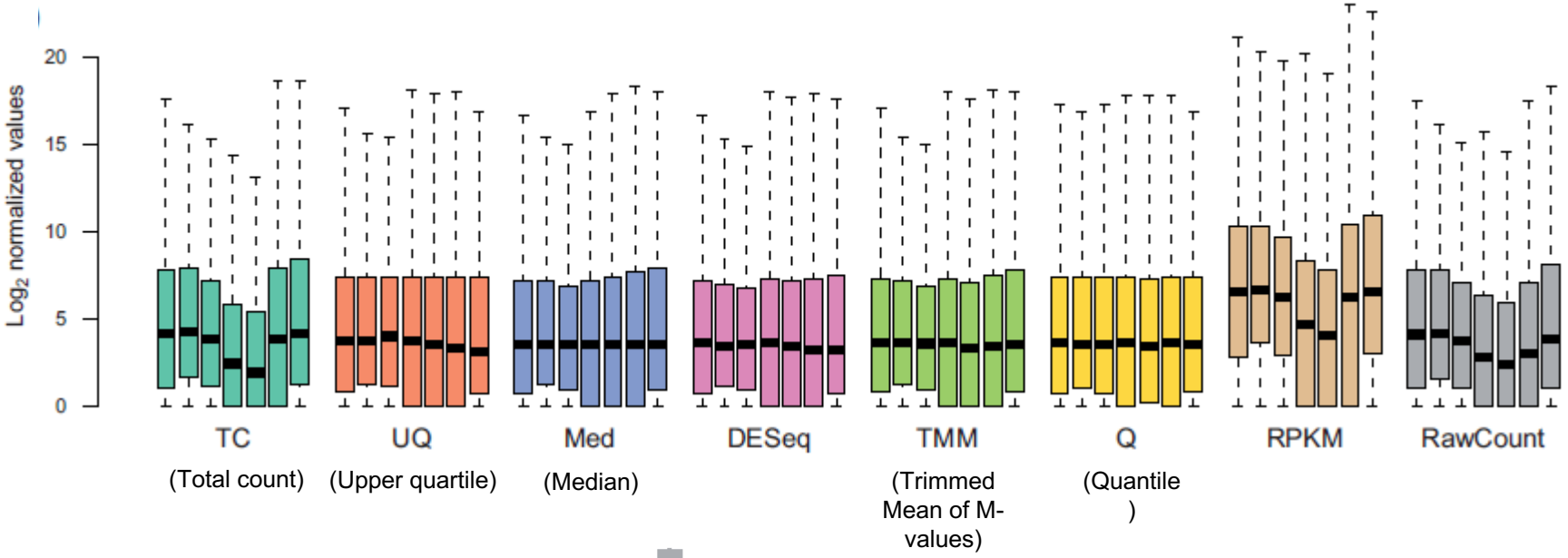
4. logCPM (Law et al., 2013)

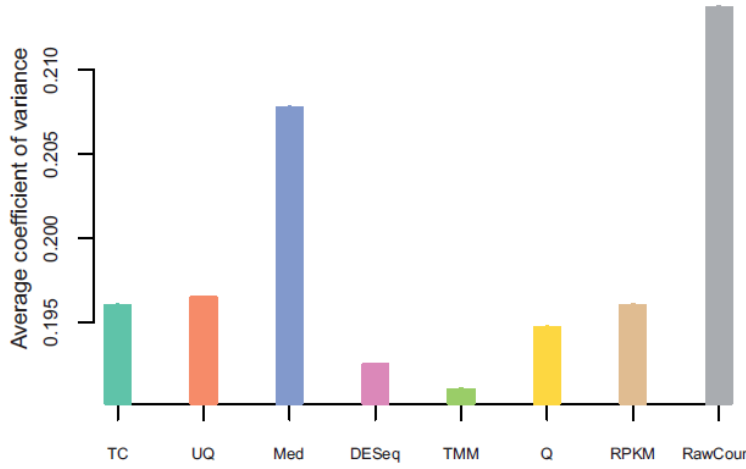
- generated from limma-voom.
- stabilize variance and remove dependence of variance or the mean.

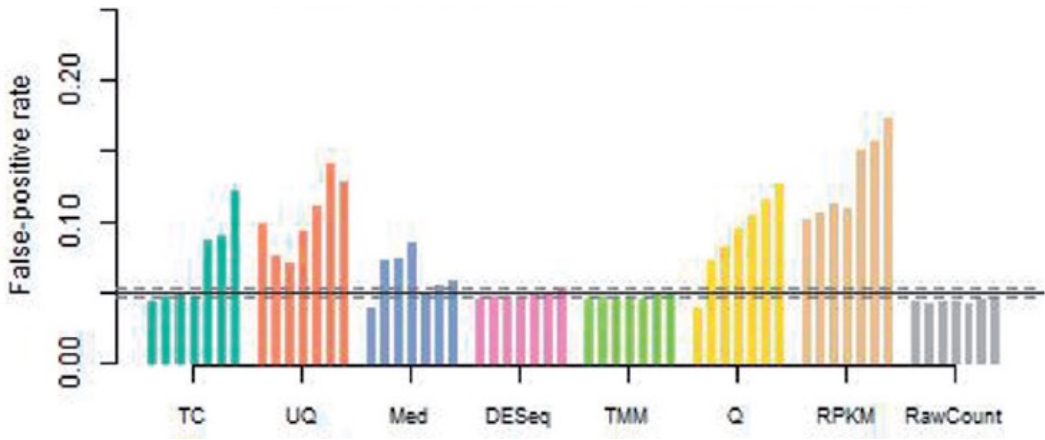
5. rlog

- small counts and library size (DESeq2).
- log2-transformed count data normalized.

Effects of normalization methods on fold change calculation and DGE analysis





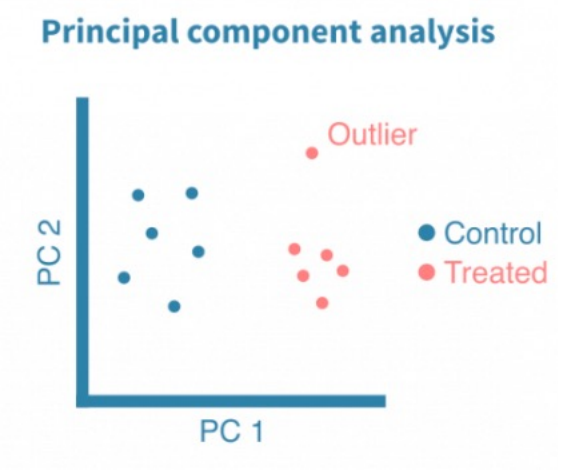


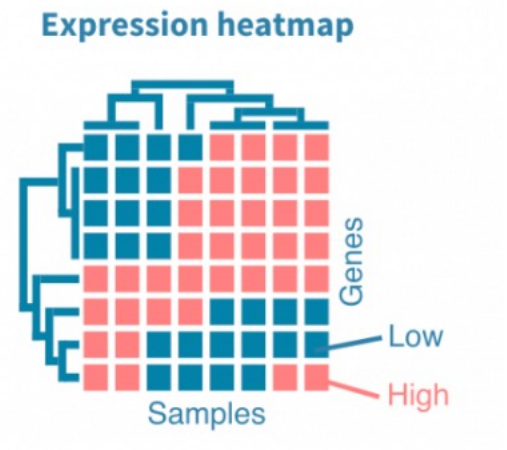
(doi:10.1093/bib/bbs046)

Bioinformatics: Exploratory gene expression analysis

Exploratory gene expression analysis

- After raw reads have been quality controlled and gene counts delivered, we can visualize general patterns of the samples using:
  - Principal component analysis (PCA)
  - Expression heatmaps
- To answer a few questions:
  - Do the biological replicates resemble each other?
  - Do distinct sample groups form separate clusters?
  - Are there outlier samples?

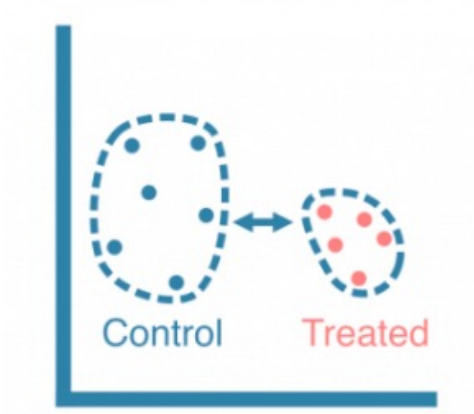


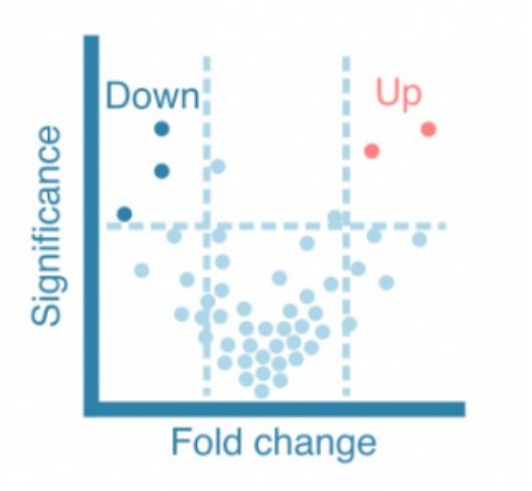


Bioinformatics: Diffferential expression analysis

Exploratory gene expression analysis

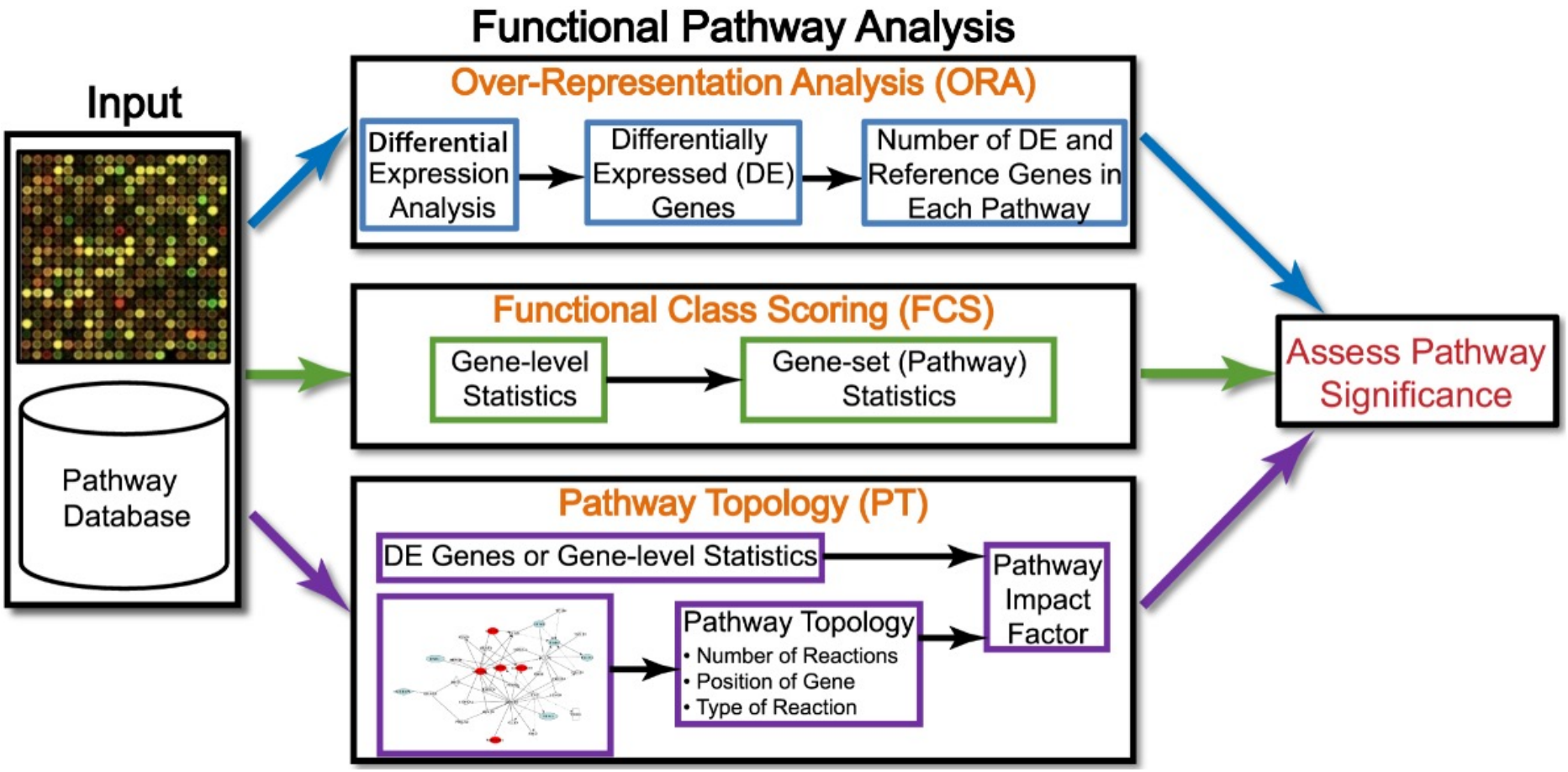
- A statistical comparison between 2 sample groups.
- Results are reported as fold-change values (e.g., log2FC) and statistical significance (P-value).
- Three biological replicates (n = 3) per condition is a “rule-of-thumb” minimum.
- More biological replicates, more subtle differences can be detected.
- Visualize using volcanoplot.





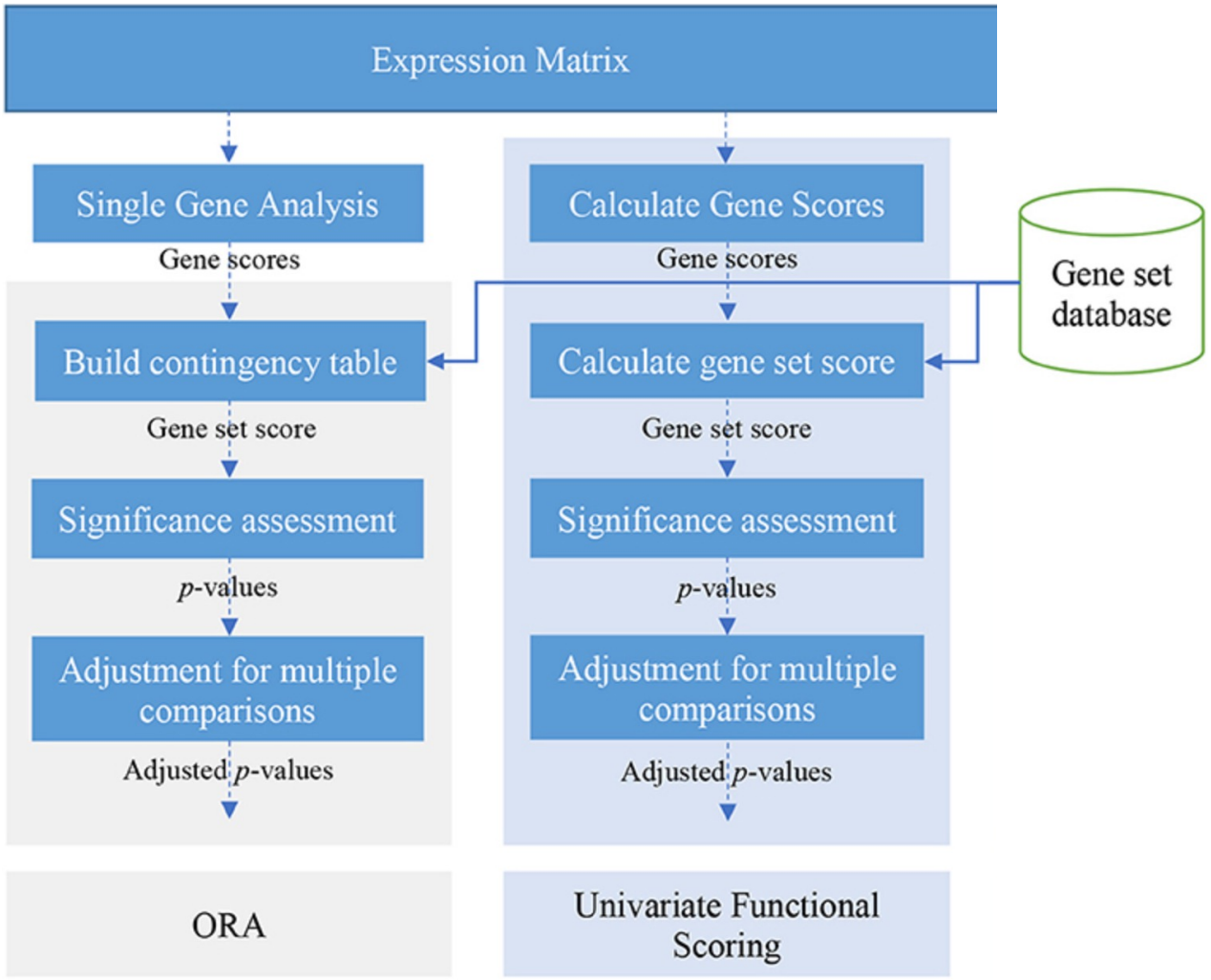
Bioinformatics: Functional pathway analysis

Gene set analysis methods



(Khatri, Sirota, and Butte, 2012, PLoS Comp Biol)

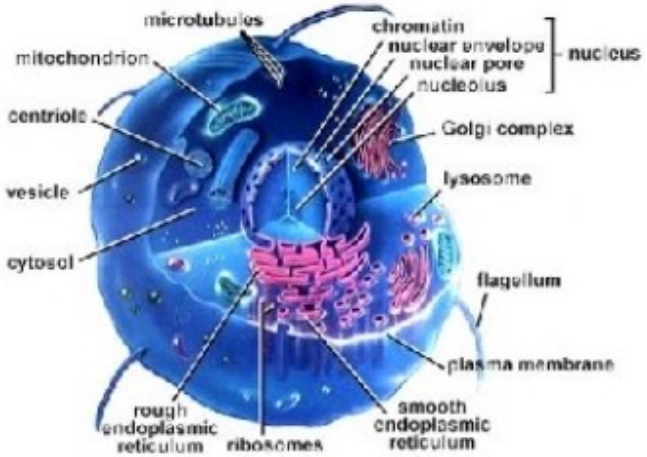
Functional pathway analysis



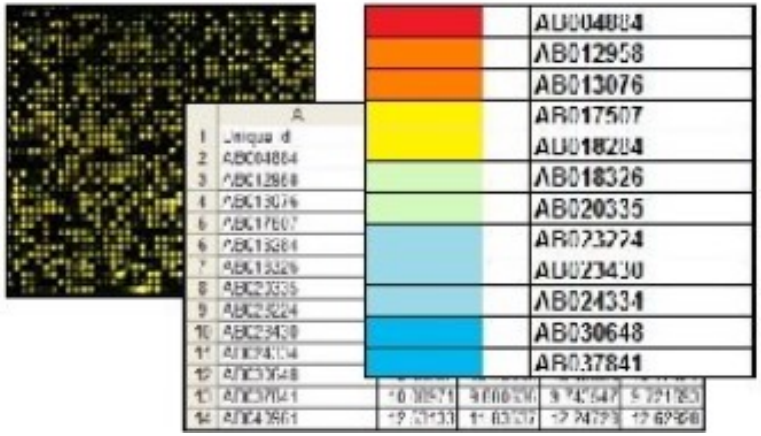
(Maleki, Ovens, Hogan, and Kusalik, 2020, Front. Genet.)

Gene sets overview

Cellular components/functions



Gene expression patters



Gene sets

Not significant

Nuclear Pore

Gene.AAA
Gene.ABA
Gene.ABC

Ribosome

Gene.RP1
Gene.RP2
Gene.RP3
Gene.RP4

Not significant

Up

Cell Cycle

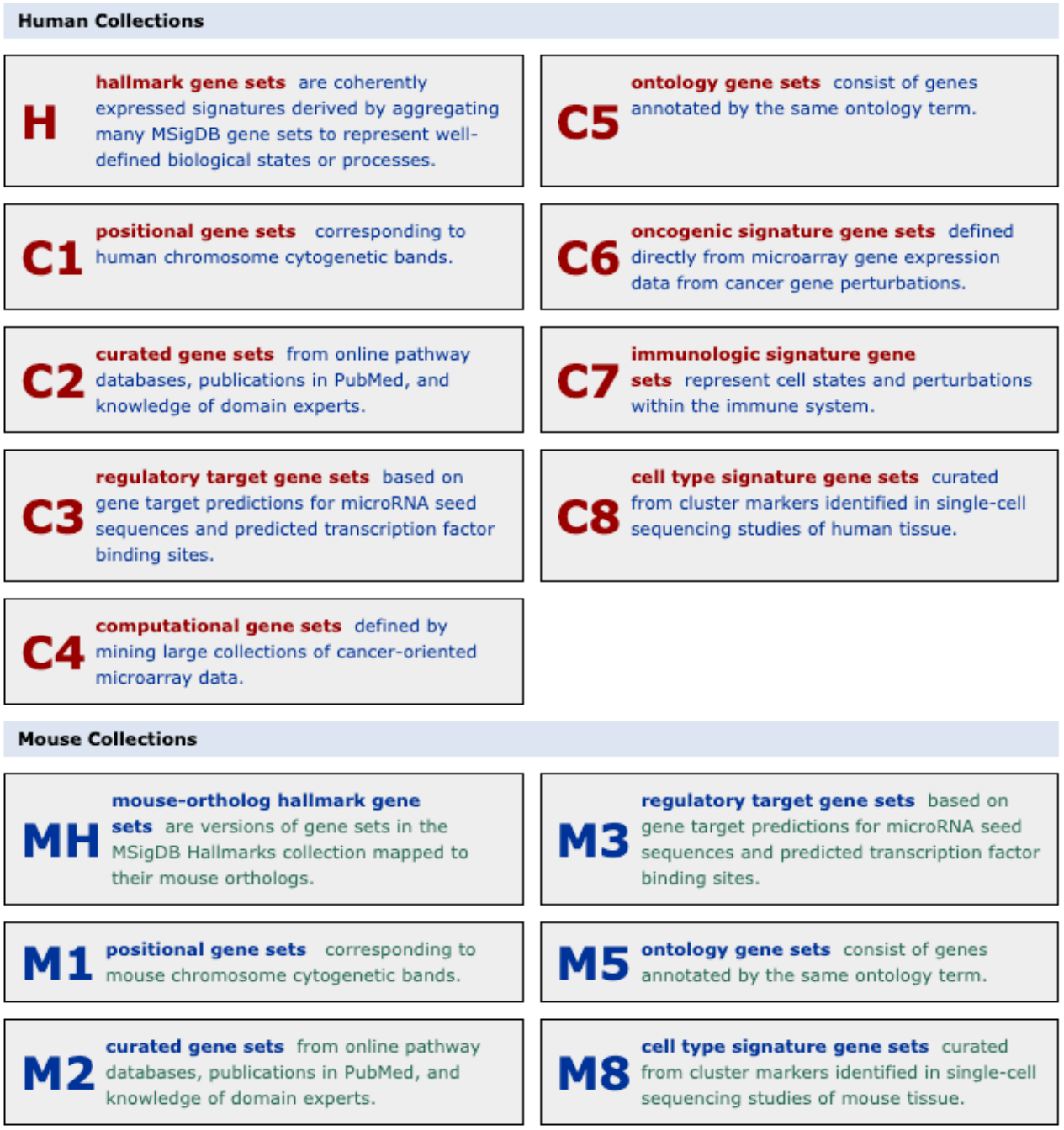
Gene.CC1
Gene.CC2
Gene.CC3
Gene.CC4
Gene.CC5

P53 signaling

Gene.CC1
Gene.CK1
Gene.PPP

Down

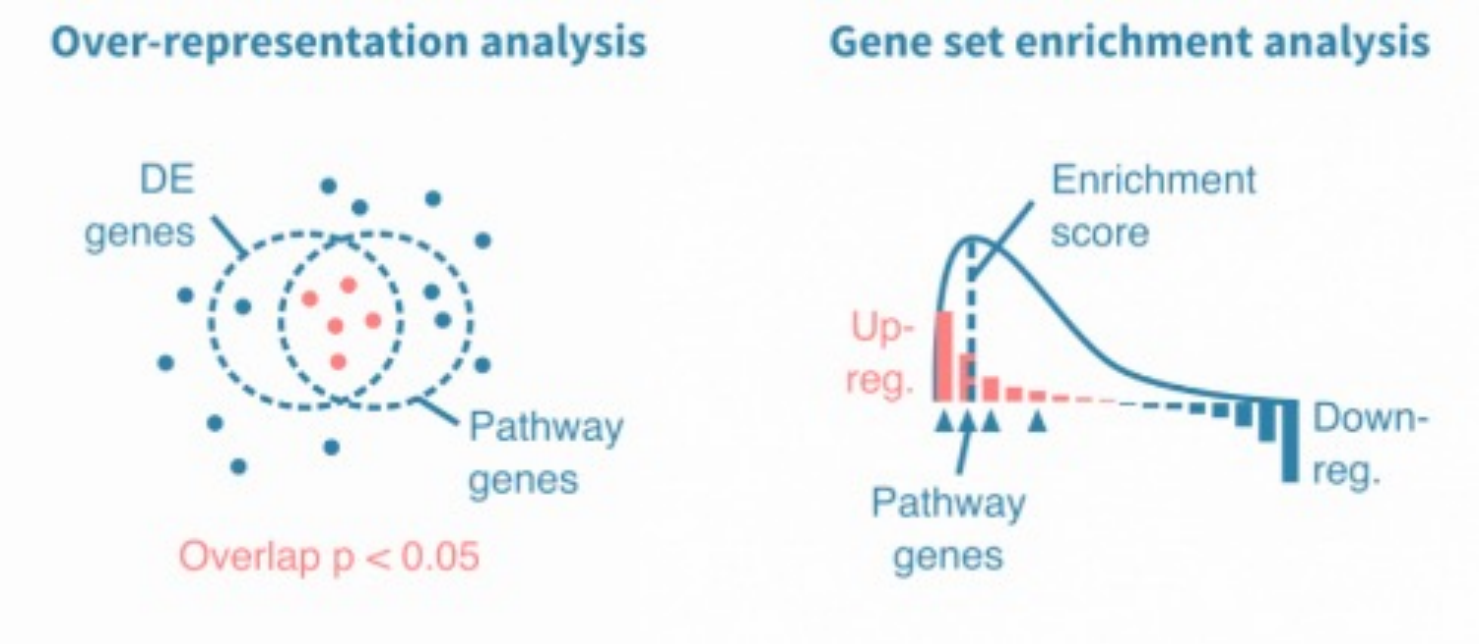
Gene sets databases



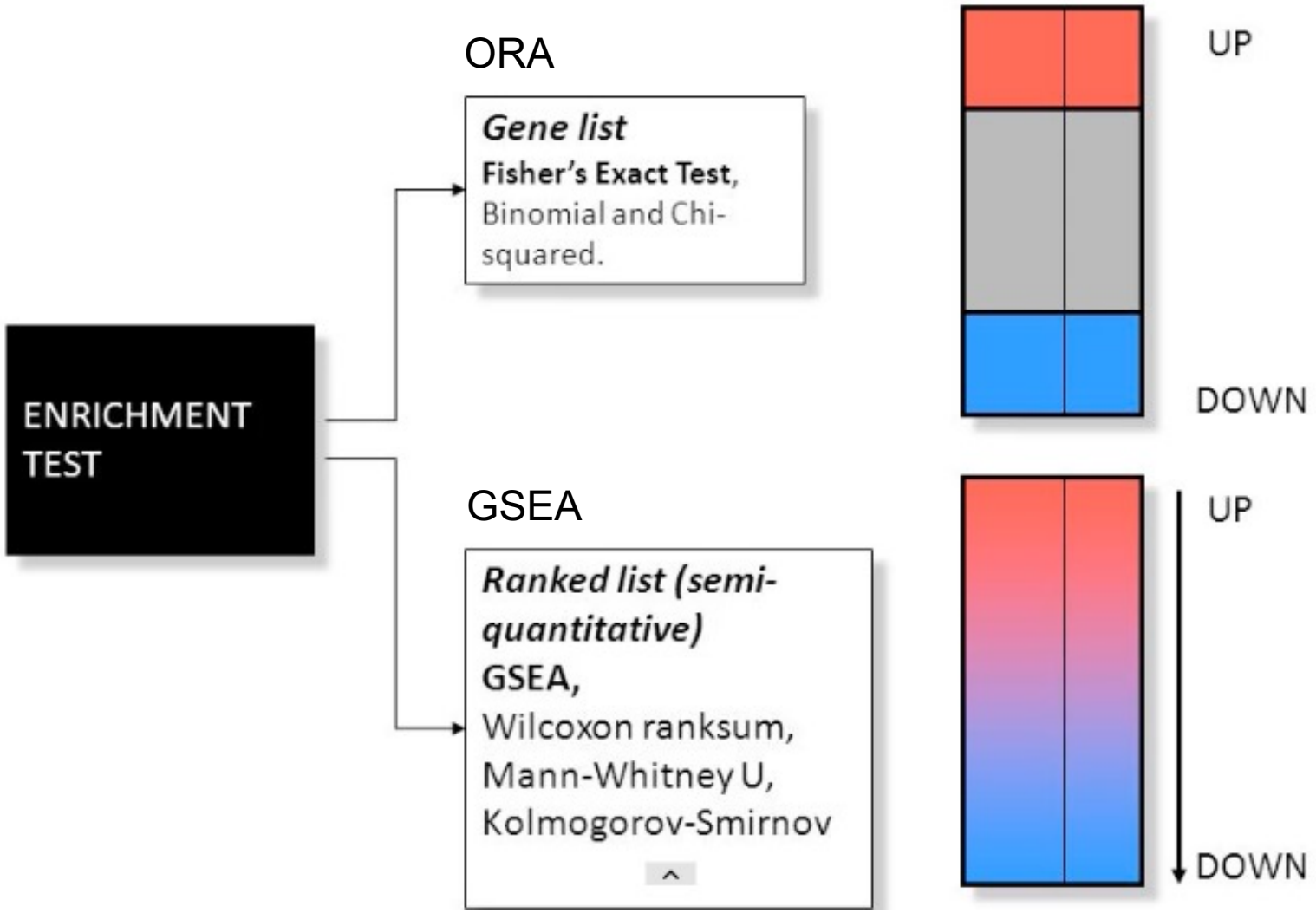
https://www.gsea-msigdb.org/gsea/msigdb/

ORA vs. GSEA

- Overrepresentation analysis (ORA)
  - determines a set of differentially experssed genes for the ones that could be part of a biological pathway.
- Gene set enrichment analysis (GSEA)
  - determines whether an *a priori* defined sets of genes shows statistical significant.



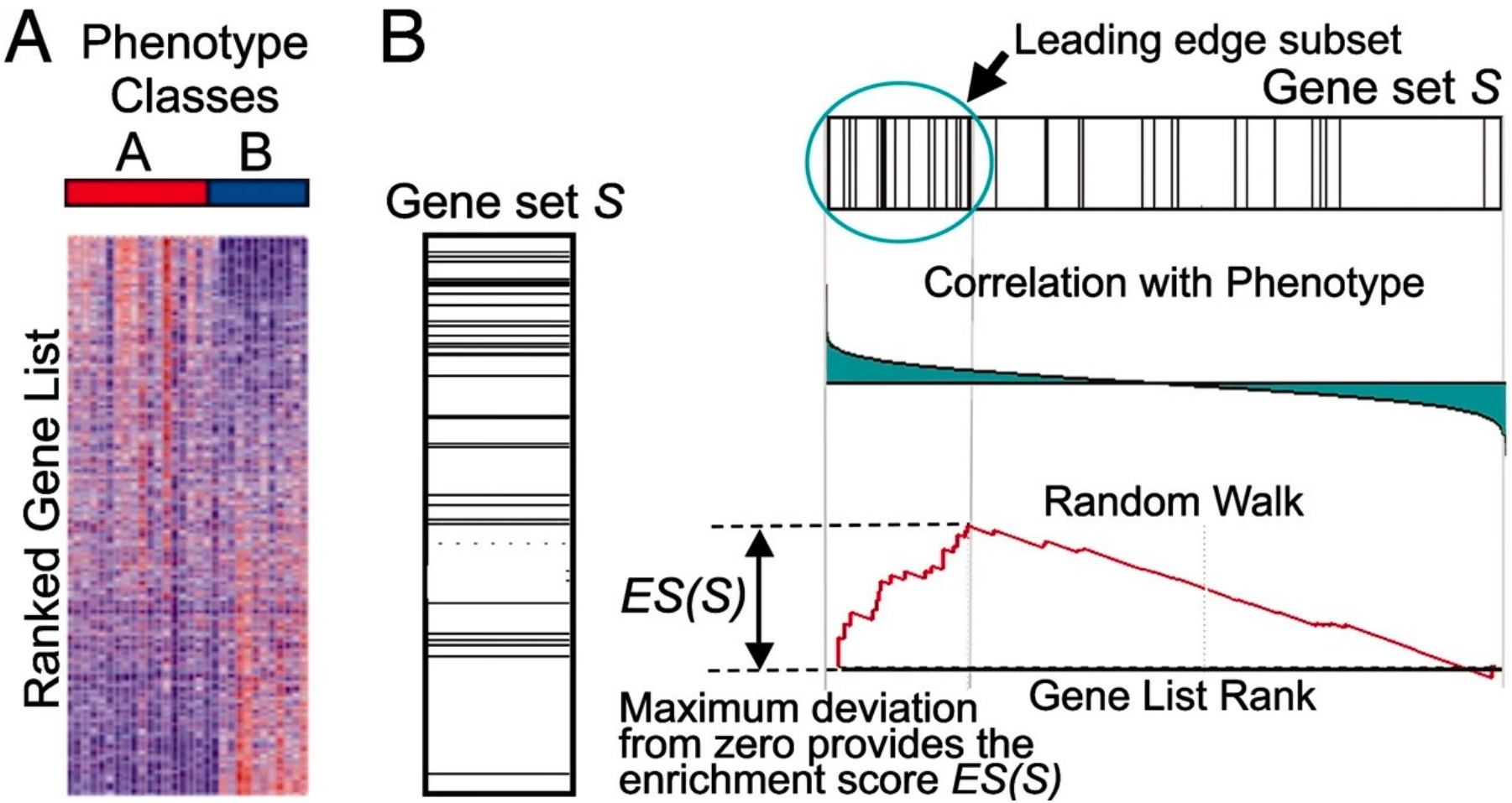
ORA vs. GSEA statistics



ORA vs. GSEA summary

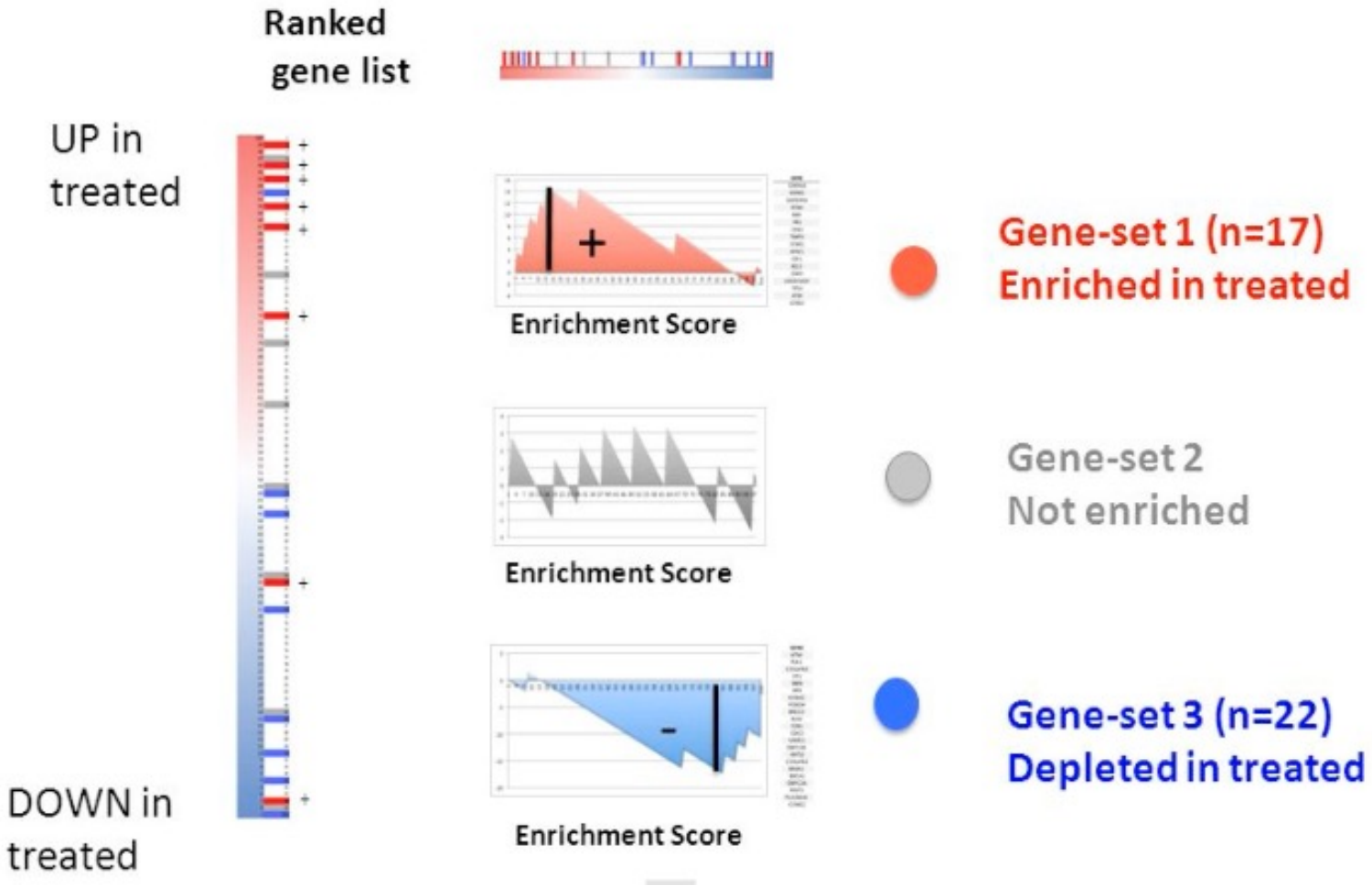
Analysis	What is required for input	What output looks like	✅ Pros	⚠️ Cons
ORA (Over-representation Analysis)	A list of gene IDs (no stats needed) (e.g., expression change > 2-fold)	A per-pathway hypergeometric test result	- Simple - Inexpensive computationally to calculate p-values	- Requires arbitrary thresholds and ignores any statistics associated with a gene - Assumes independence of genes and pathways
GSEA (Gene Set Enrichment Analysis)	A list of genes IDs with gene-level summary statistics (e.g., ranked by expression change)	A per-pathway enrichment score	- Includes all genes (no arbitrary threshold!) - Attempts to measure coordination of genes	- Permutations can be expensive - Does not account for pathway overlap - Two-group comparisons not always appropriate/feasible

GSEA interpretation



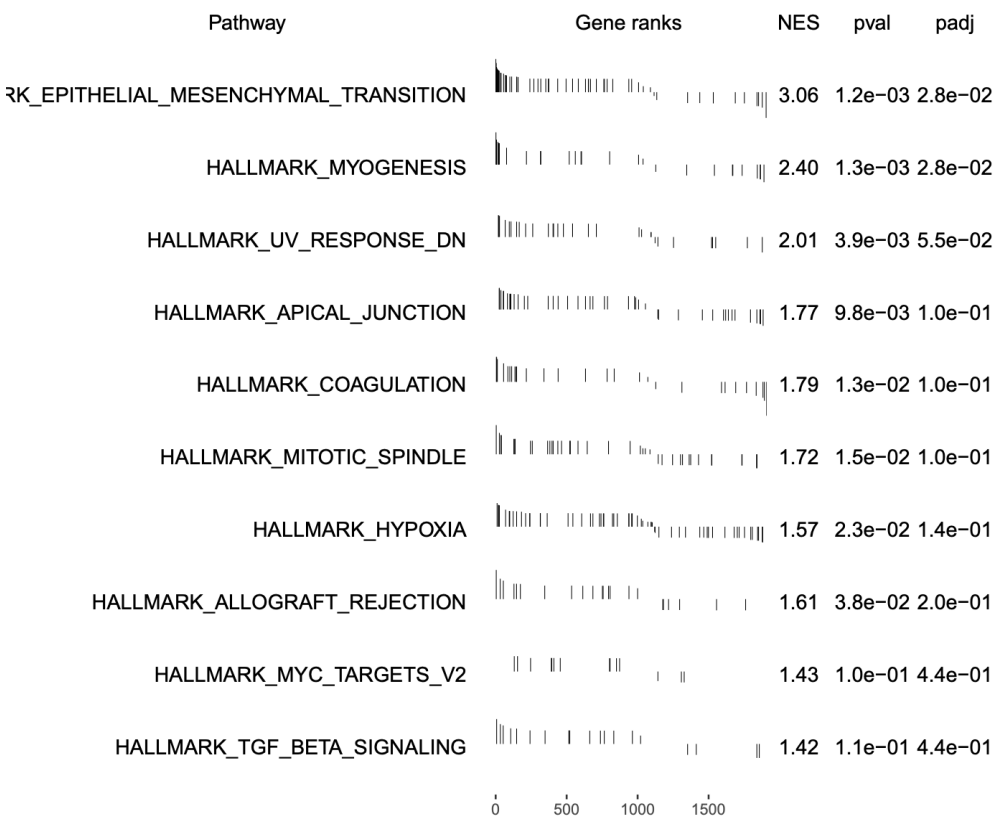
High ES score = high local enrichment

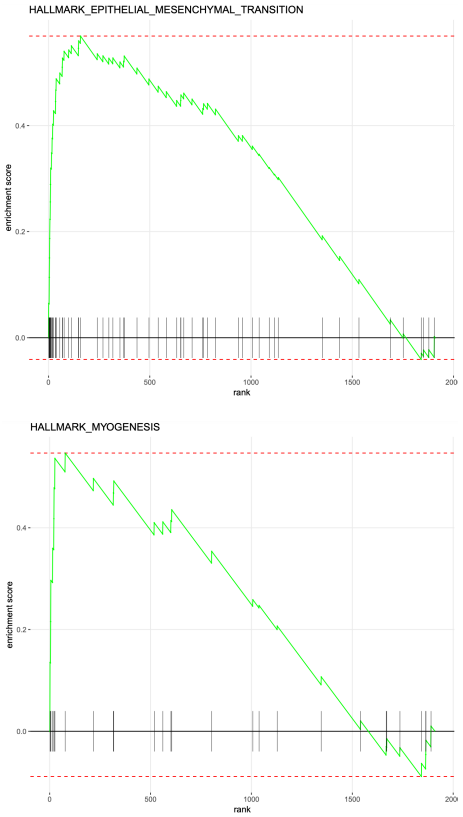
GSEA interpretation



GSEA outputs

pathway	pval	padj	ES	NES	nMoreExtreme	size	leadingEdge
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0.00116414435389988	0.0275590551181102	0.568819529672154	3.05758117225939	0	60	Acta2, Flna, Sparc, Mylk, Thbs1, Lama1, Ccn2, P
HALLMARK_MYOGENESIS	0.00131233595800525	0.0275590551181102	0.54691625732516	2.39989761993306	0	28	Myh11, Sparc, Mylk, Tagln, Col4a2, Bhlhe40, My
HALLMARK_UV_RESPONSE_DN	0.0038961038961039	0.0545454545454545	0.447708476093017	2.00517858415535	2	30	Dusp1, Bhlhe40, Igfbp5, Rbpms, Cav1, Plpp3, Se
HALLMARK_APICAL_JUNCTION	0.00982800982800983	0.103194103194103	0.35681472049475	1.77444548846055	7	44	Lama3, Myh9, Myl9, Actn4, Icam1, Mmp2, Tro, F
HALLMARK_COAGULATION	0.0131233595800525	0.103703703703704	0.407748460099051	1.7892219253254	9	28	Sparc, Thbs1, Csrp1, Mmp2, Arf4, Fbn1, Timp3,
HALLMARK_MITOTIC_SPINDLE	0.0148148148148148	0.103703703703704	0.356554980687282	1.7202522249314	11	39	Flna, Myh9, Actn4, Mapre1, Capzb, Csnk1d, Cttr
HALLMARK_HYPOXIA	0.0232288037166086	0.139372822299652	0.288750917285718	1.57353649357873	19	63	Ccn2, Dusp1, Bhlhe40, Fos, Myh9, Eno1b, Cav1,





Tutorial:
Galaxy project

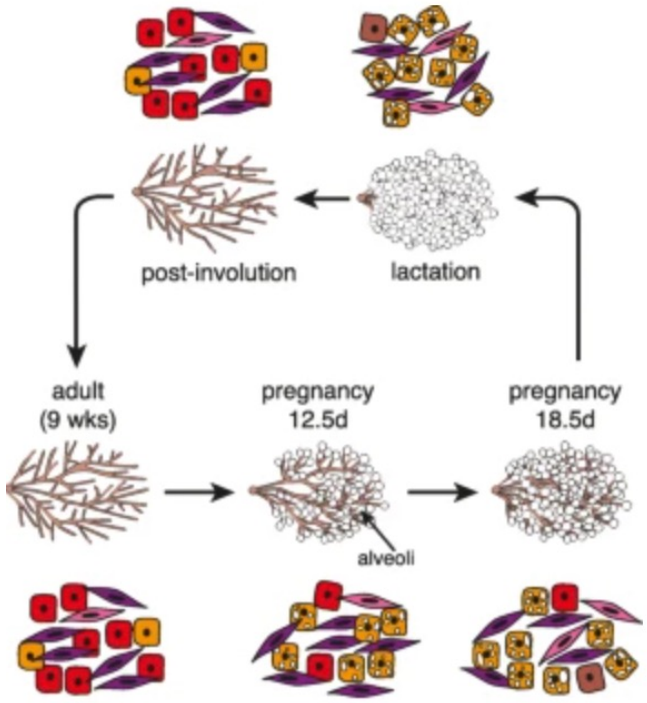
Download data

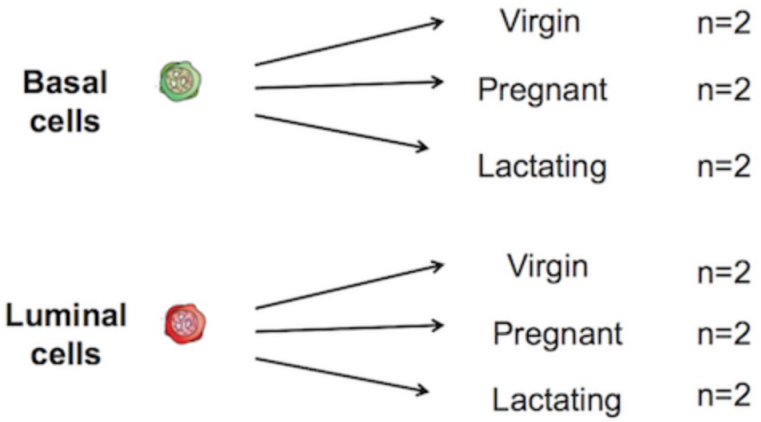
Mus musculus RNA-Seq data

- Two types, three conditions.
  - Basal, luminal cells.
  - Virgin, pregnant, lactate.
- Two replicates.
- Download at https://bit.ly/3OnLTCq

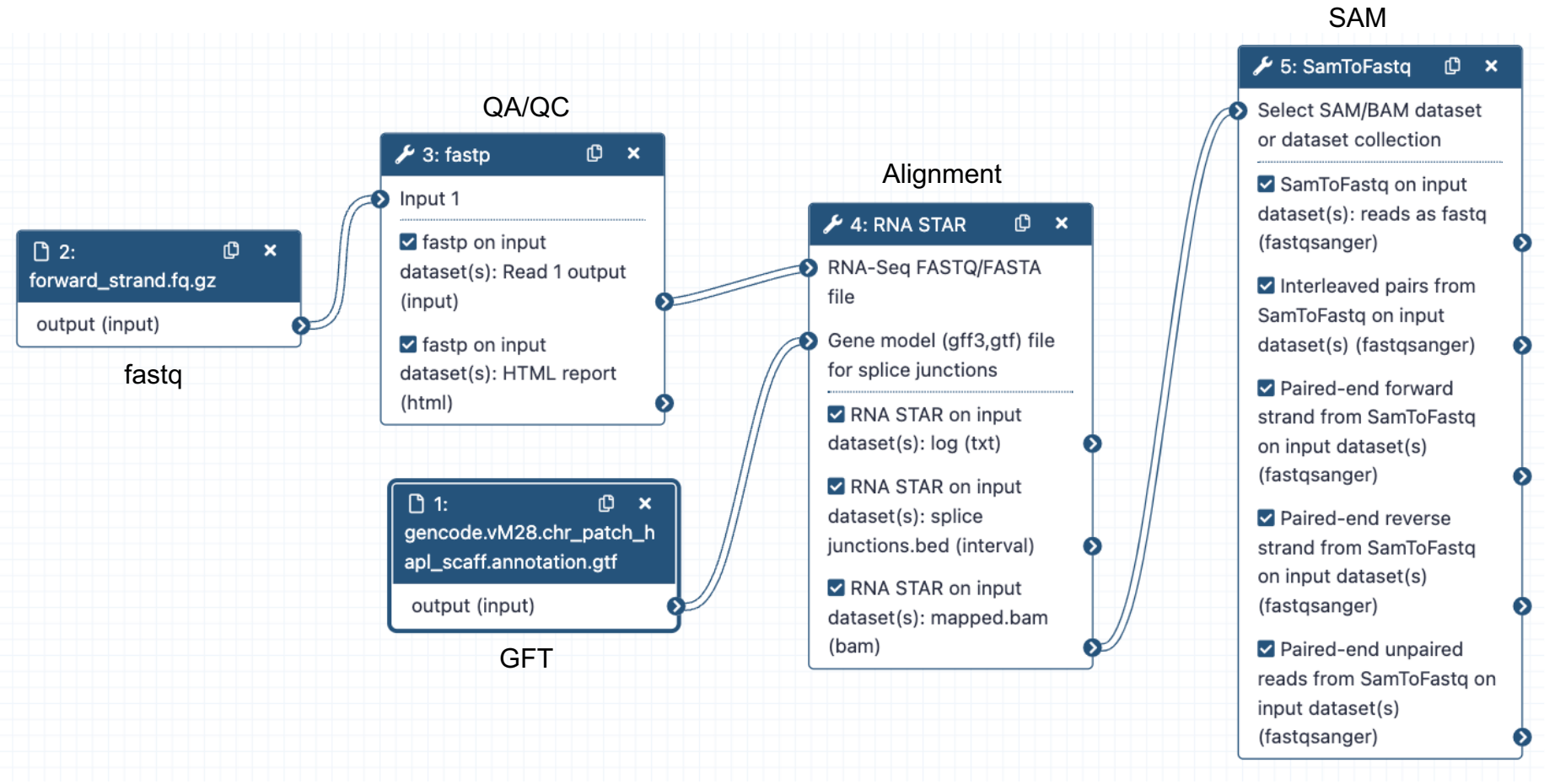
Reference database

- Mouse genome.
- Mouse transcriptome.
- Hallmark gene sets.
- Download at https://bit.ly/3UUCKUr

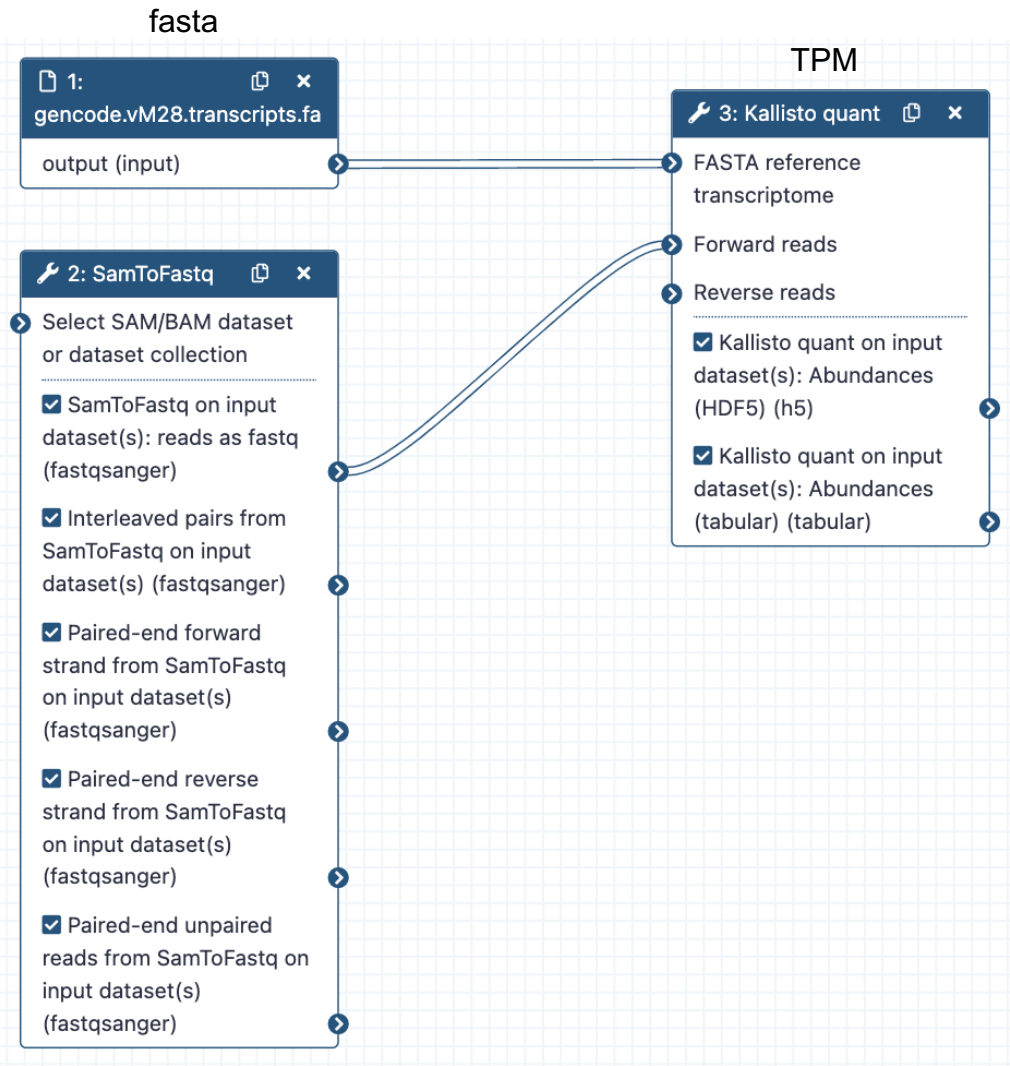




STAR alignment workflow



Kallisto quantification workflow

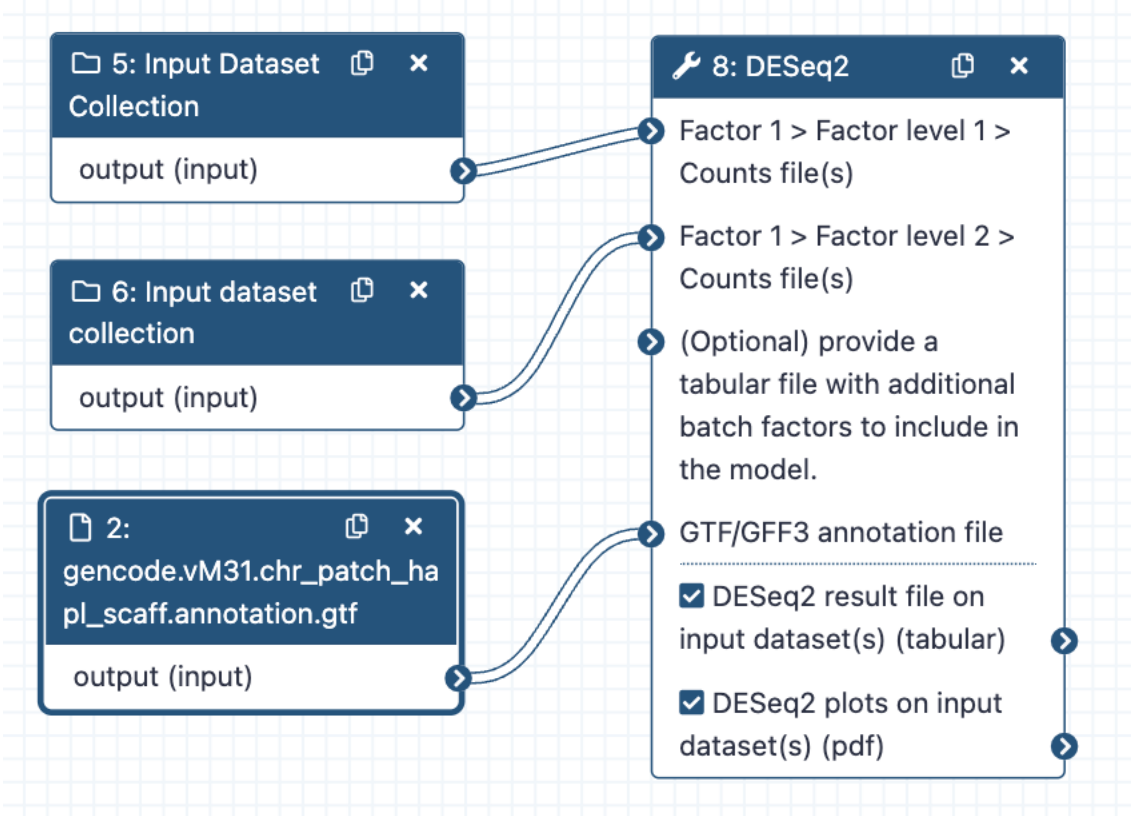


DEG using DESeq2 workflow

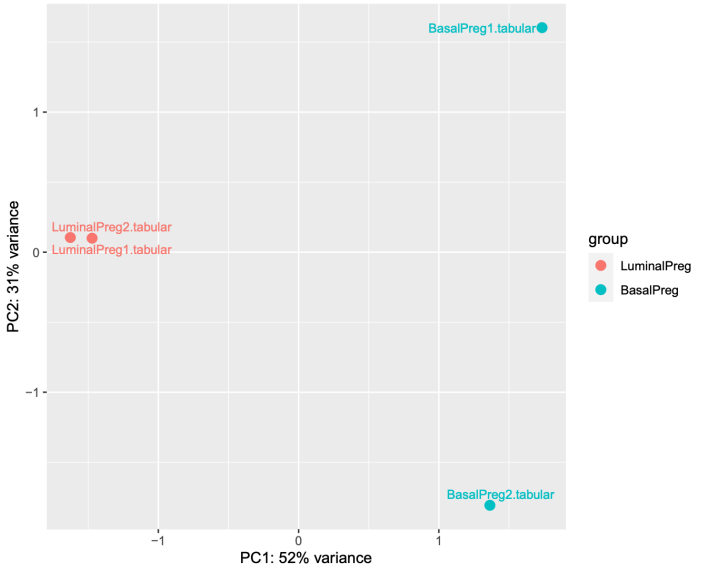
Data collection A
(output from Kallisto)

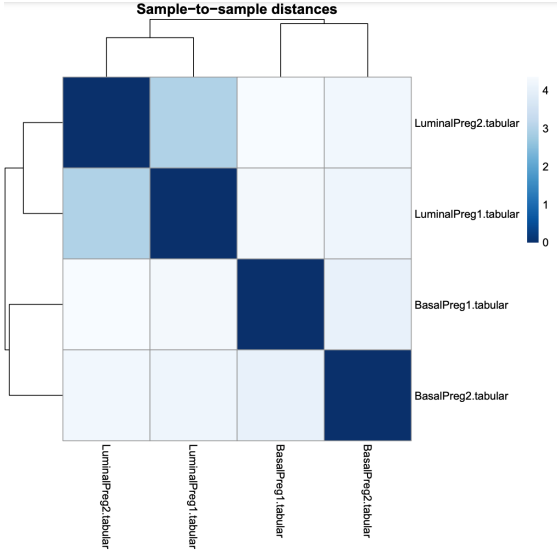
Data collection B
(output from Kallisto)

Reference genome
(GFT format)



DEG results: basal vs. luminal mammary cells in pregnant mice





GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value
ENSMUSG00000063157.10	22.7053166147197	-4.82006060543848	0.870801517979073	-5.53520004951842	3.10873323980544e-08
ENSMUSG00000061937.8	19.0770163172454	-4.53054003503599	0.879112516114901	-5.15353831504755	2.55616816979082e-07
ENSMUSG00000070702.10	12.5885026345625	-3.93134050235266	0.914268536648597	-4.29998446273088	1.70810084454306e-05
ENSMUSG00000032554.16	8.28597383794684	-5.37544553758729	1.32438171526543	-4.05883400203088	4.9318355241298e-05
ENSMUSG00000035783.10	6.09107459376949	5.15837631229398	1.41015102798059	3.65803109733646	0.000254160215059699
ENSMUSG00000018830.11	5.28307026982656	4.88146247165951	1.46694178029199	3.32764567567765	0.000875831842578717
ENSMUSG00000000001.5	0.275301258287342	0.856584944313586	1.85637240555842	0.461429474898877	0.644490508794769
ENSMUSG00000000058.7	0.291953014585935	0.829860348660042	1.85273328853974	0.447911393287541	0.65421714549303
ENSMUSG00000000085.17	0.291953014585935	0.829860348660042	1.85273328853974	0.447911393287541	0.65421714549303
ENSMUSG00000000303.13	0.441691506920243	-0.913681174625269	1.83380113775239	-0.498244414737974	0.618311780919906

GSEA workflow

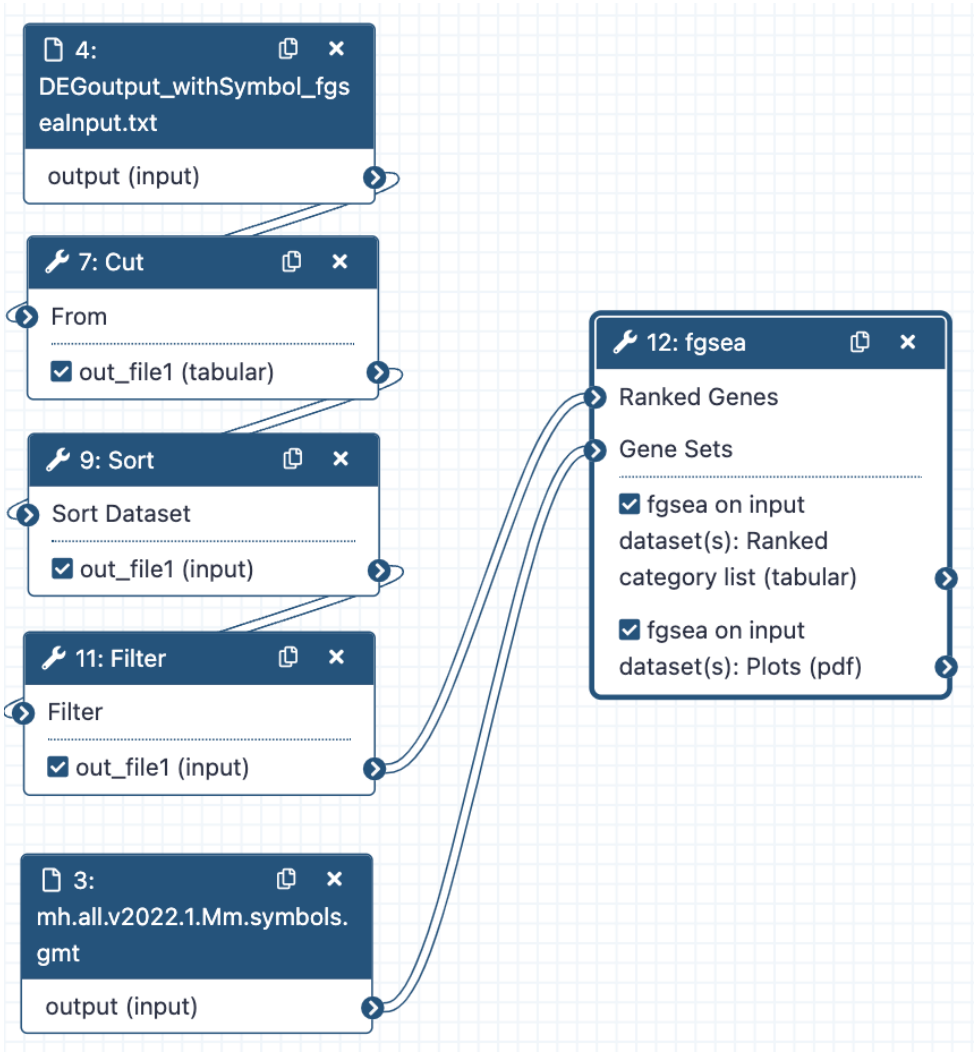
Output from DESeq2 (or featureCount)

Select columns with gene ID and log2(FC)

Sort by log2(FC) from high to low

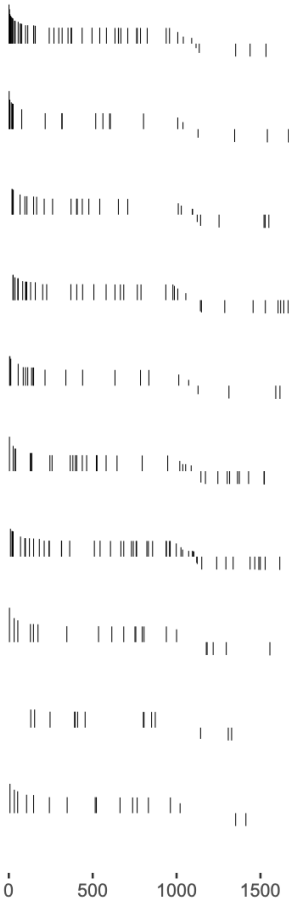
Filter unused genes

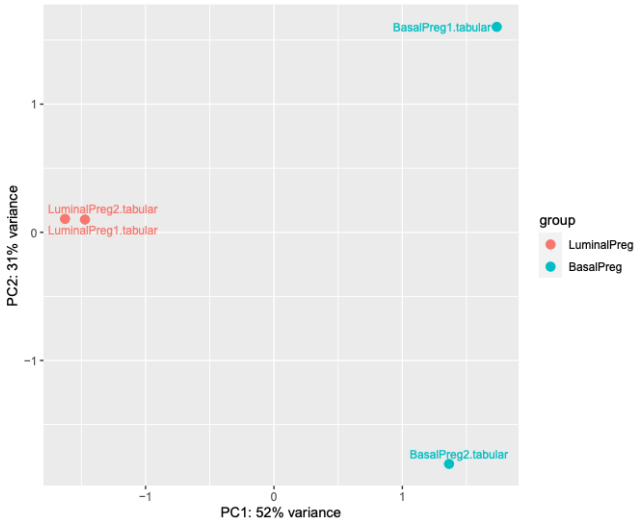
Mouse hallmark gene sets (GMT format)

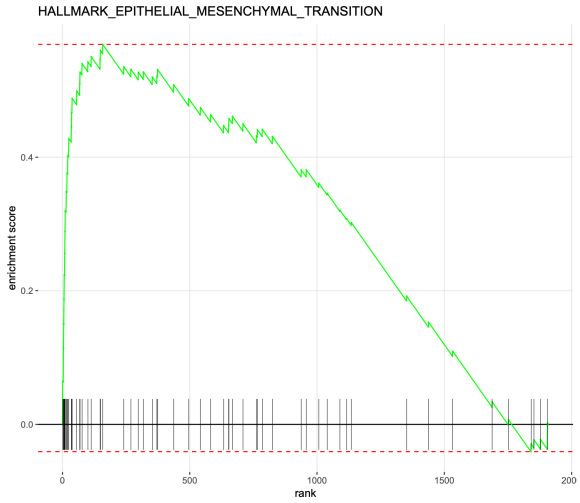


GSEA results: basal vs. luminal mammary cells in pregnant mice

Pathway	Gene ranks	NES	pval	padj
RK_EPITHELIAL_MESENCHYMAL_TRANSITION		3.06	1.2e−03	2.8e−02
HALLMARK_MYOGENESIS		2.40	1.3e−03	2.8e−02
HALLMARK_UV_RESPONSE_DN		2.01	3.9e−03	5.5e−02
HALLMARK_APICAL_JUNCTION		1.77	9.8e−03	1.0e−01
HALLMARK_COAGULATION		1.79	1.3e−02	1.0e−01
HALLMARK_MITOTIC_SPINDLE		1.72	1.5e−02	1.0e−01
HALLMARK_HYPOXIA		1.57	2.3e−02	1.4e−01
HALLMARK_ALLOGRAFT_REJECTION		1.61	3.8e−02	2.0e−01
HALLMARK_MYC_TARGETS_V2		1.43	1.0e−01	4.4e−01
HALLMARK_TGF_BETA_SIGNALING		1.42	1.1e−01	4.4e−01







For more information

GSEA user guide

- https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html

GSEA by Pathway Commons

- https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

RNA-seq genes to pathway using Galaxy

- https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-genes-to-pathways/tutorial.html

Thank you
Q&A

patipark.k@chula.ac.th