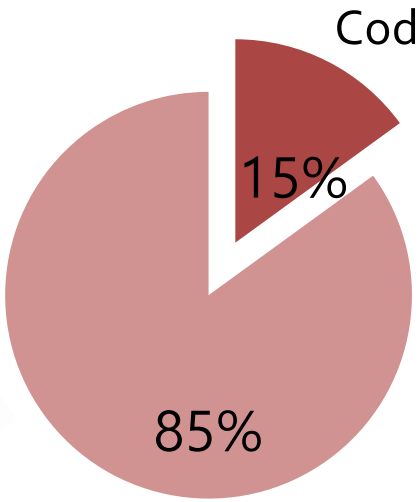
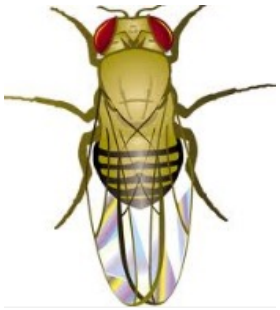
Transcriptome analyses

Session 14

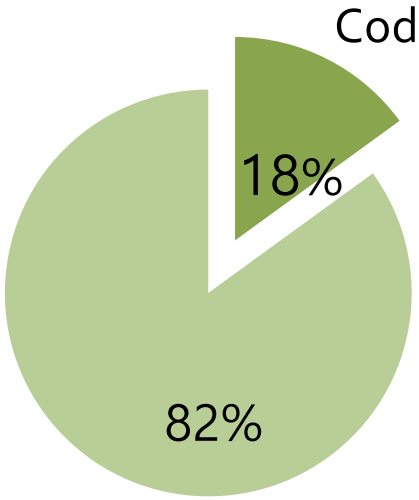
Animal genomes are largely non-coding

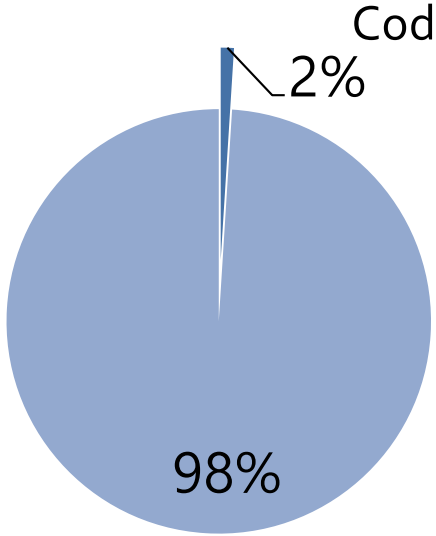
100Mb

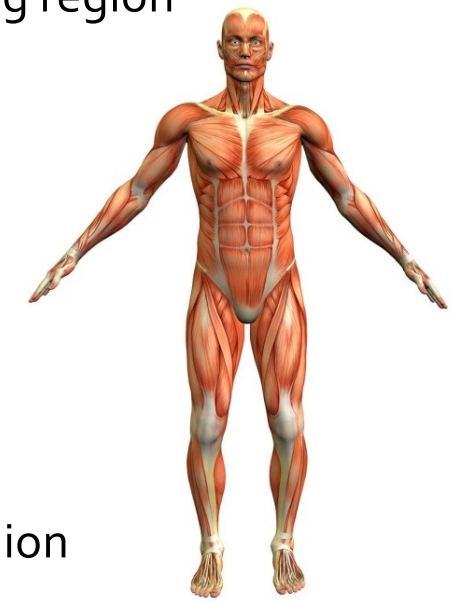




140Mb







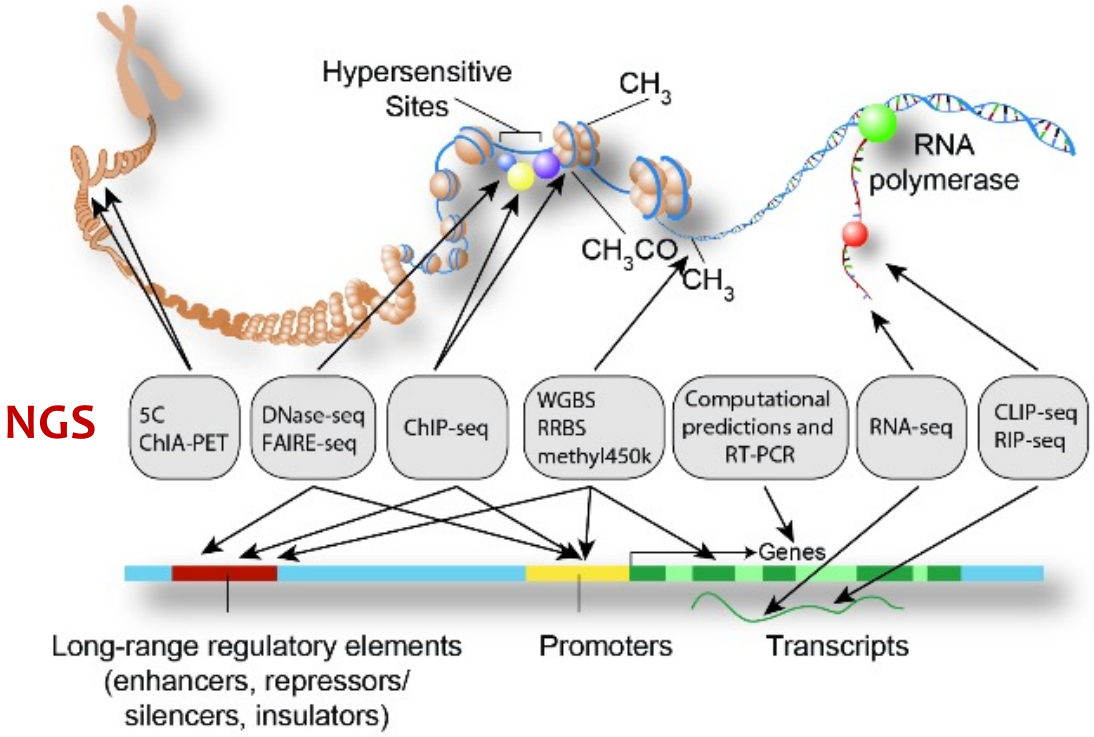
3Gb

NGS facilitated to decode noncoding genomes

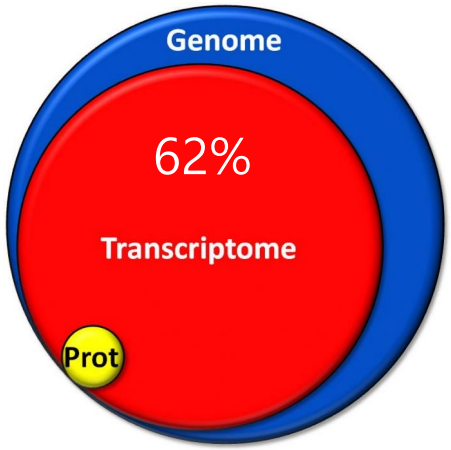
The Encyclopedia of DNA Elements (ENCODE)

The ENCODE Project Consortium
Nature 489, 57–74 (2012).

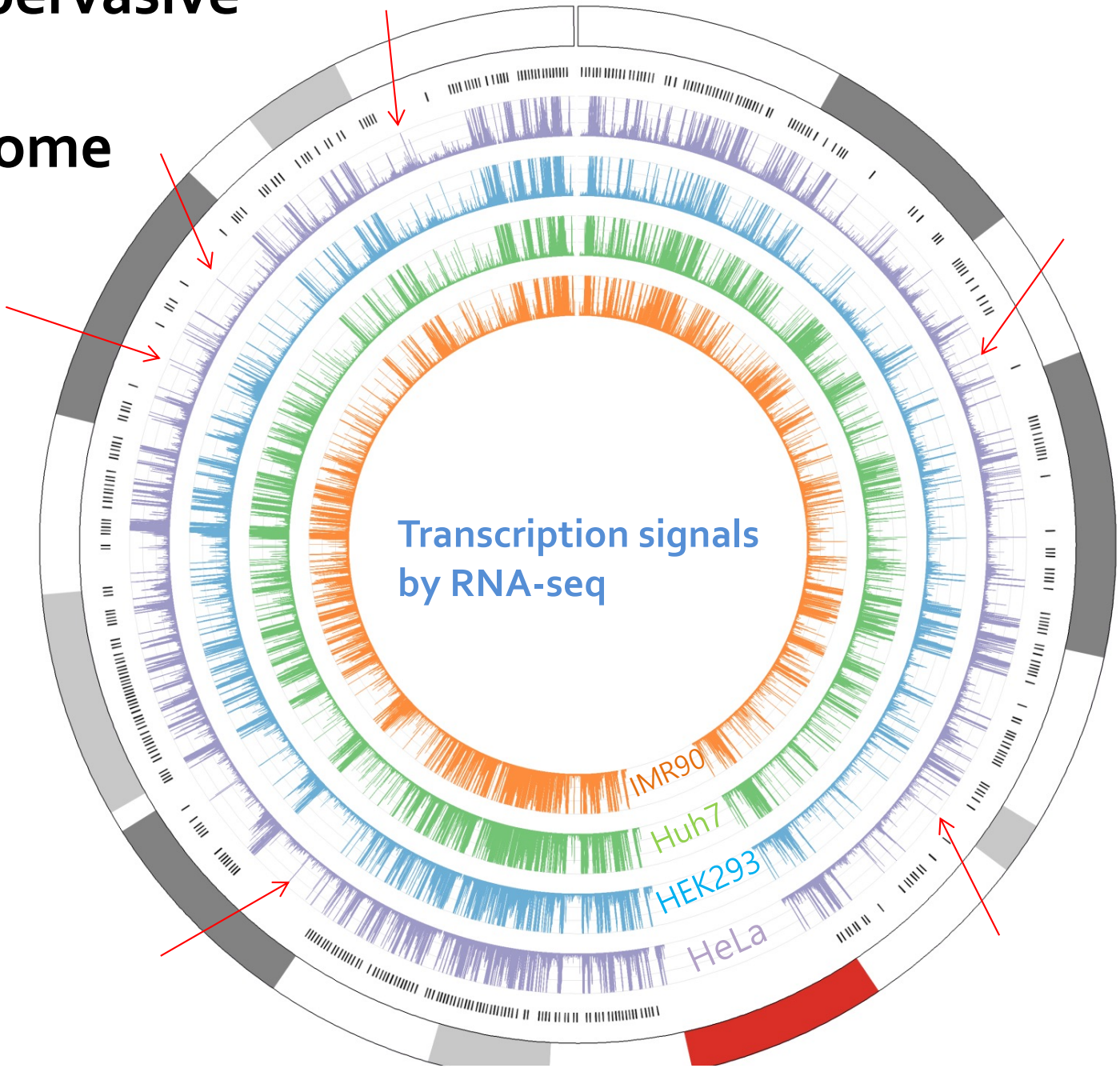
The modENCODE Project.
Science 330, 1775–1787 (2010).



The ENCODE projected announced that ~80% of genomes are associated with at least one biochemical signal in cells.



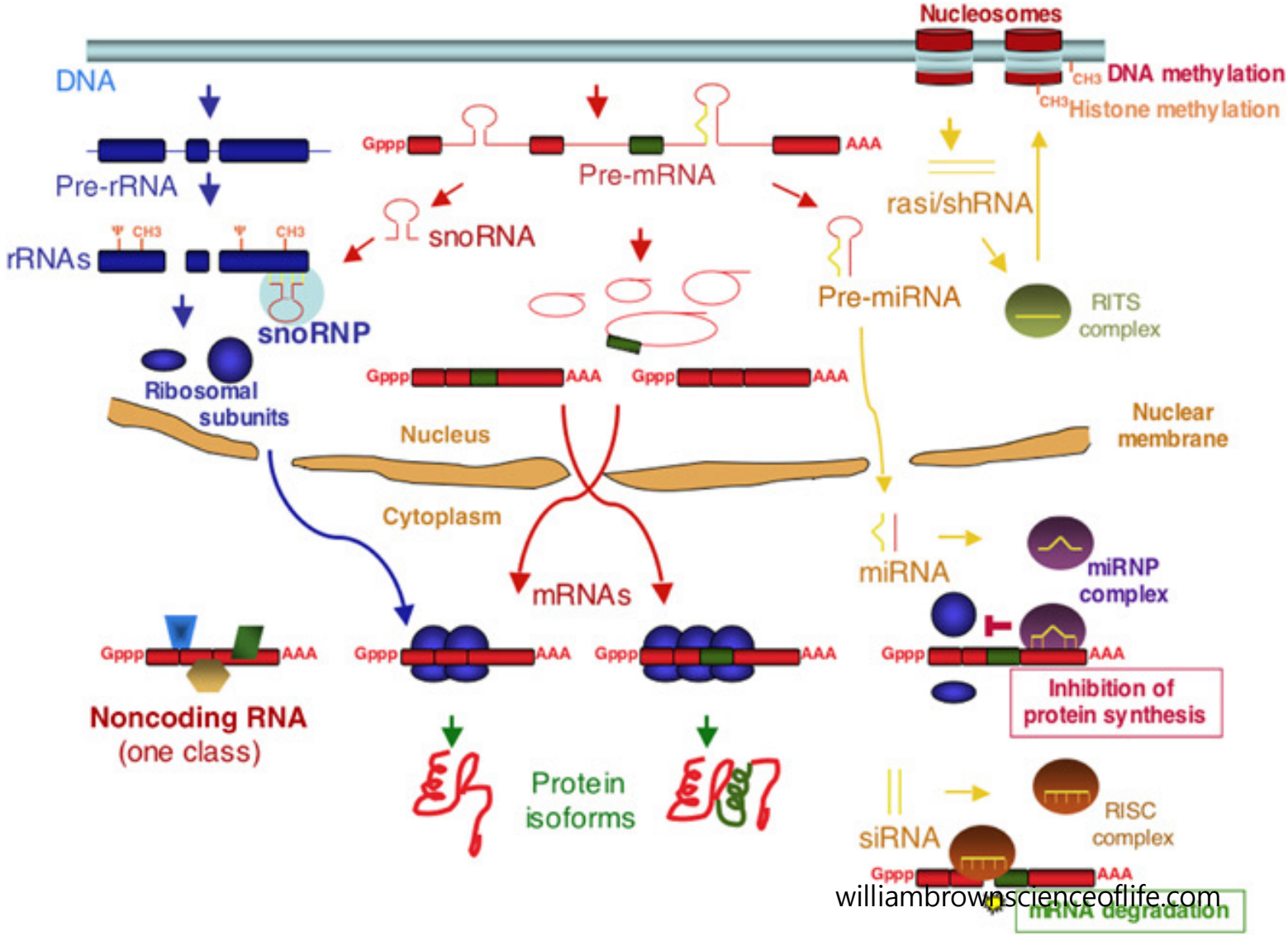
Reproducible, pervasive transcription in noncoding genome



Circos plot of transcription signals over chr 20

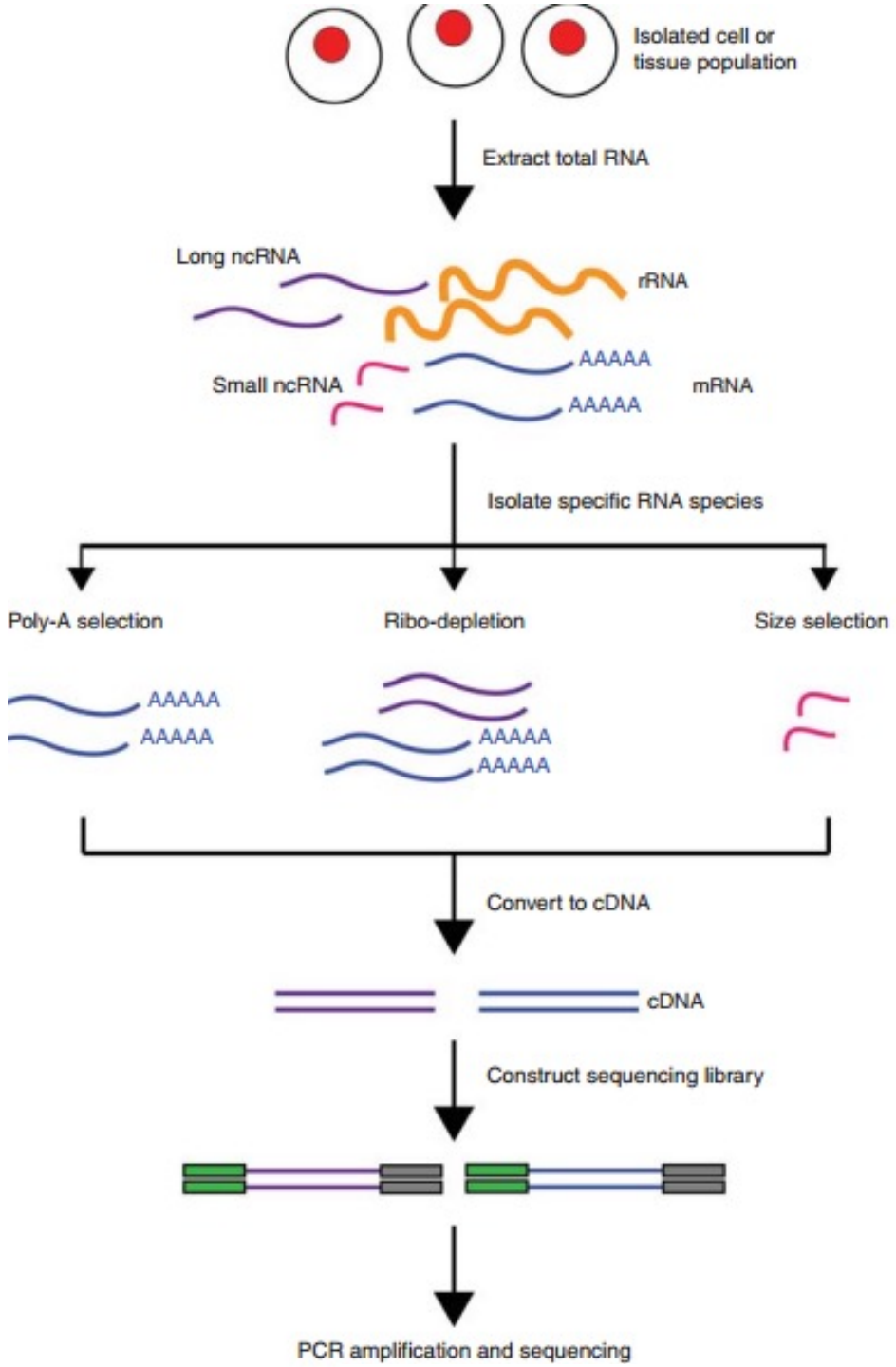
Transcriptome

The **transcriptome** is the entire set of all RNA molecules



Whole Transcriptome Sequencing (WTS)

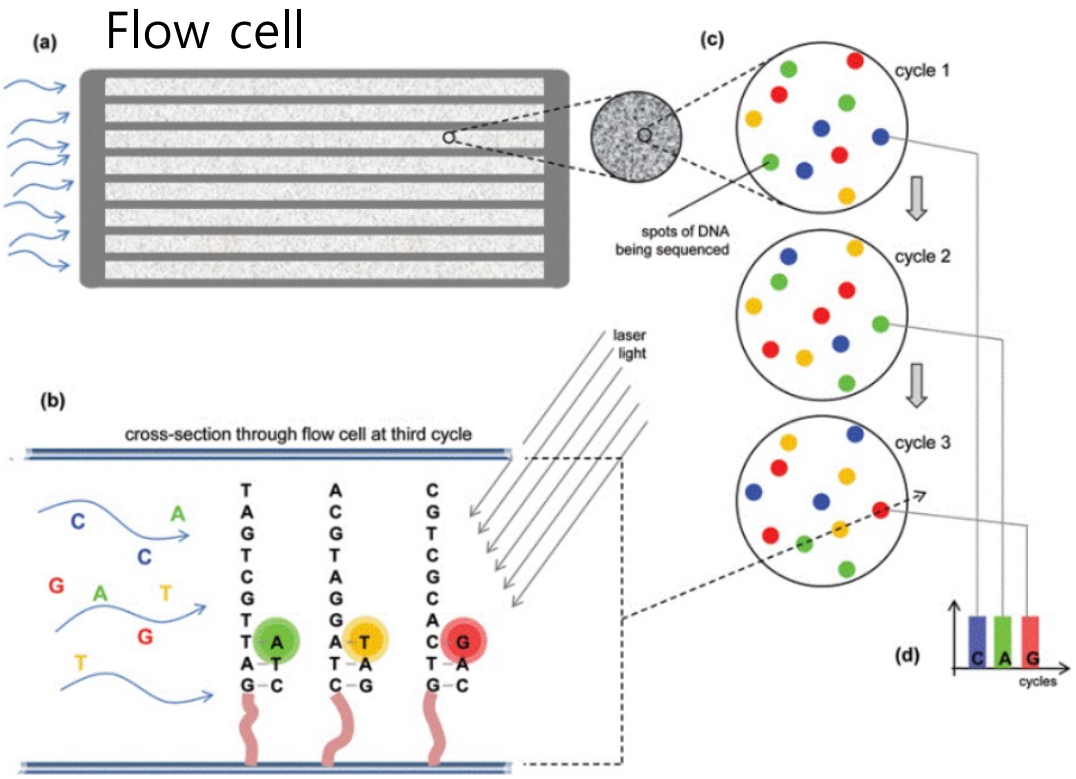
I. library construction & sequencing



Cold Spring Harb Protoc; doi:10.1101/pdb.top084970

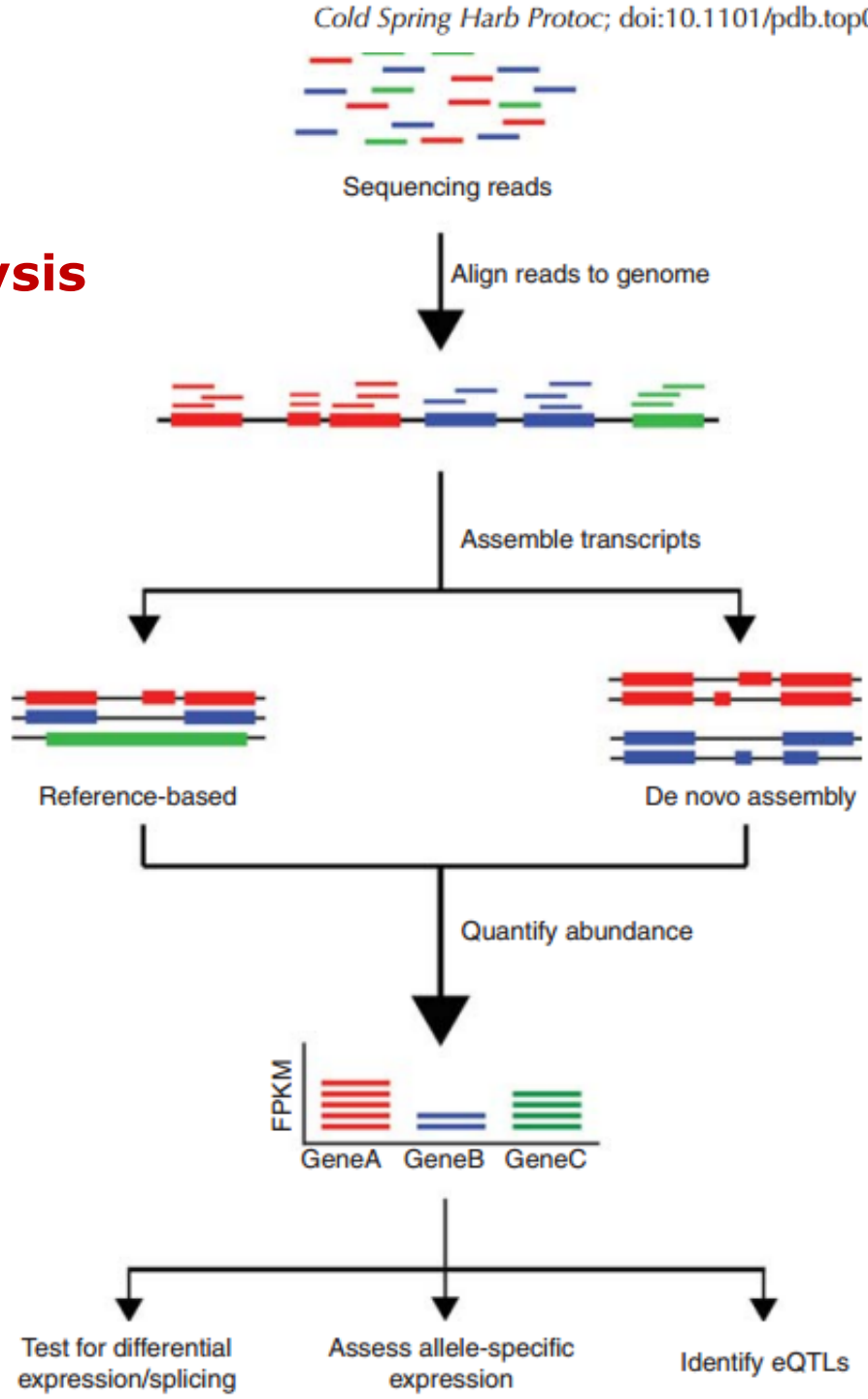
Next-Generation Sequencing (NGS)

The high demand for low-cost sequencing has driven the development of **high-throughput sequencing** (**also termed next-generation sequencing**) technologies that **parallelize the sequencing process, producing thousands or millions of sequences at once**.



Whole Transcriptome Sequencing (WTS)

II. Mapping, Assembly, & Analysis



General Transcriptome Analysis Procedure

1. Quality assessment

- **Checking biases throughout the pipeline**, RNA extraction, sample prep, library construction, sequencing and read mapping
- Sequencing quality checker: FASTQ-based FASTX-toolkit, FastQC
- Trimming by Sickle

2. Read alignment

- DNA mapping programs (not considering splice junction): BWA, Bowtie
- RNA mapping programs
  - To genome (considering splice junction): Tophat, STAR, GSNAP, MapSplice
  - To transcriptome (do not consider SJ but should include all isoforms): bowtie2

3. Transcriptome Assembly

- Reference-based assemblers: Cufflinks, StringTie, Scripture
- *De novo* assemblers: Trinity, Velvet
- Meta assemblers: CAFÉ, TACO

General Transcriptome Analysis Procedure

4. Quantification & other analysis

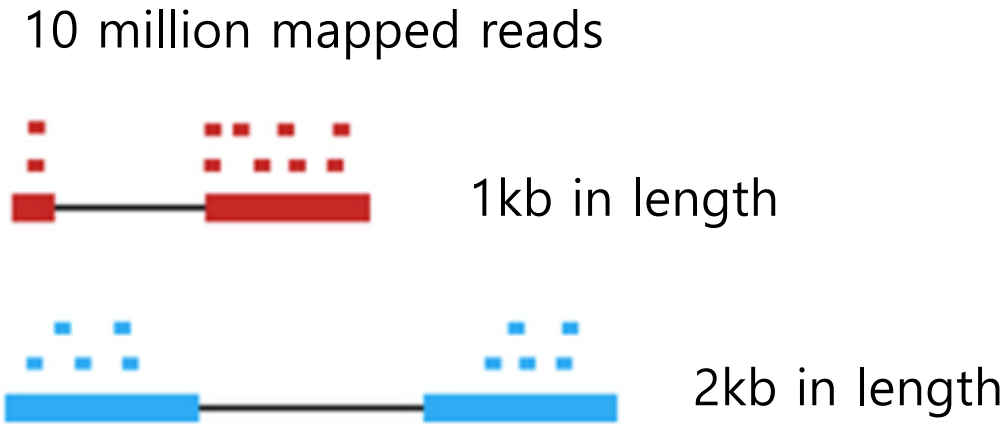
- **Normalization issues**: gene length, sequencing depth, complexity
- **Quantification metrics**: RPKM, FPKM, TPM, RPM,…
- **Differential gene expression**: read count-based methods → Negative binomial distribution models
- Cuffdiff, DESeq, edgeR, DEGseq, BitSeq, baySeq, …

Quantification of transcripts

- RPKM (~FPKM) = reads (fragment) per kilobase of exons per million mapped reads.
- 1 RPKM ~ 1 copy in a cell.

RPKM vs FPKM

Quantification of transcripts



Quantification

RPKM, FPKM, RPK, RPM, TPM

Popular gene/isoform quantification metrics

RPKM / FPKM (Reads or fragments per kilobase of exons per million mapped reads)

RPK (Reads per kilobase of exons)

RPM (reads per million mapped reads)

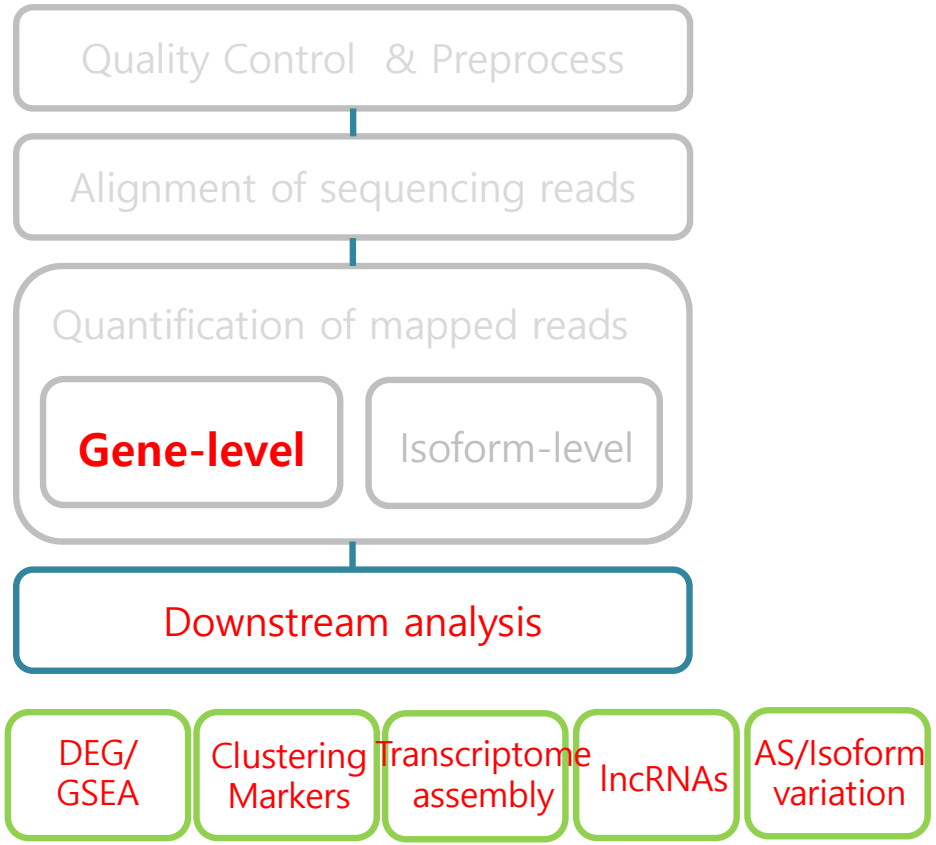
TPM (transcripts per kilobase of exons per million mapped reads)

Vs Single-cell RNAs: UMI counts

Bulk-sample RNA-seq analysis

Downstream analyses

- There are variety of options in RNA-seq downstream analysis
- Here, we are going to briefly review what we can do using gene expression values

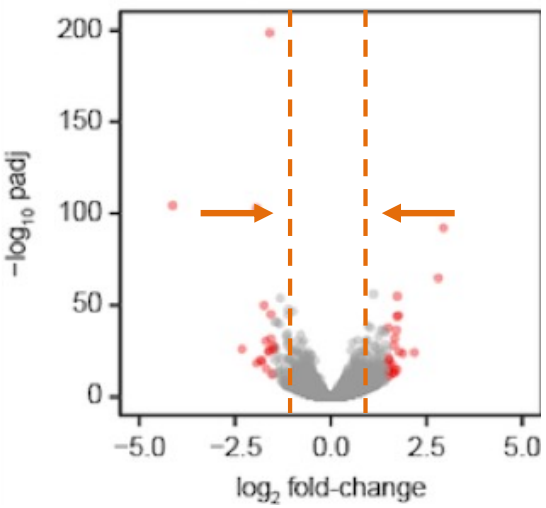


Bulk-sample RNA-seq analysis

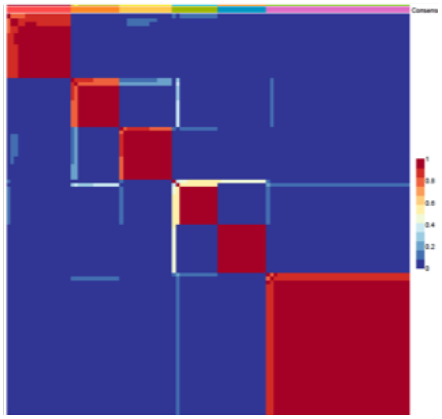
DEG analysis

- Conventional cutoff values used in the DEG analysis are "$|\log_2 FC| \geq 2$ & Adjusted P-value ≤ 0.05"
- We sometimes miss a lot of informative DEGs in that criteria we use in the analysis is too stringent
- What if we cannot find enough DEGs for the GO term analysis or experimental validation?

1. Adjust fold-change cutoffs



2. Search for gene signatures

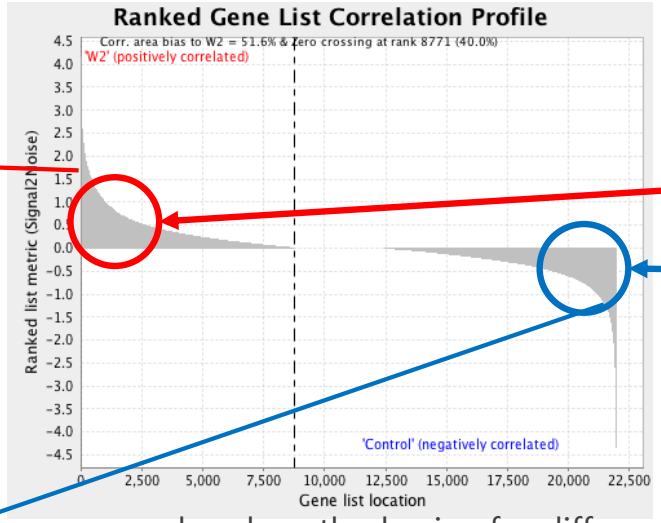


Bulk-sample RNA-seq analysis

Gene Set Enrichment Analysis

- GSEA; gene set enrichment analysis is a long standing approach estimating enrichment of certain gene sets or pathways in RNA-seq analysis

VK Mootha *et al., Nature genetics* (2003)

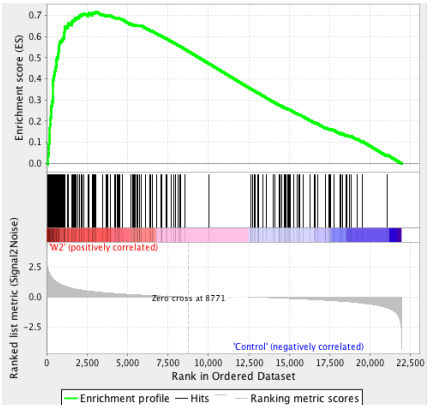


Genes are ordered on the basis of a difference metric

?

Gene set 1

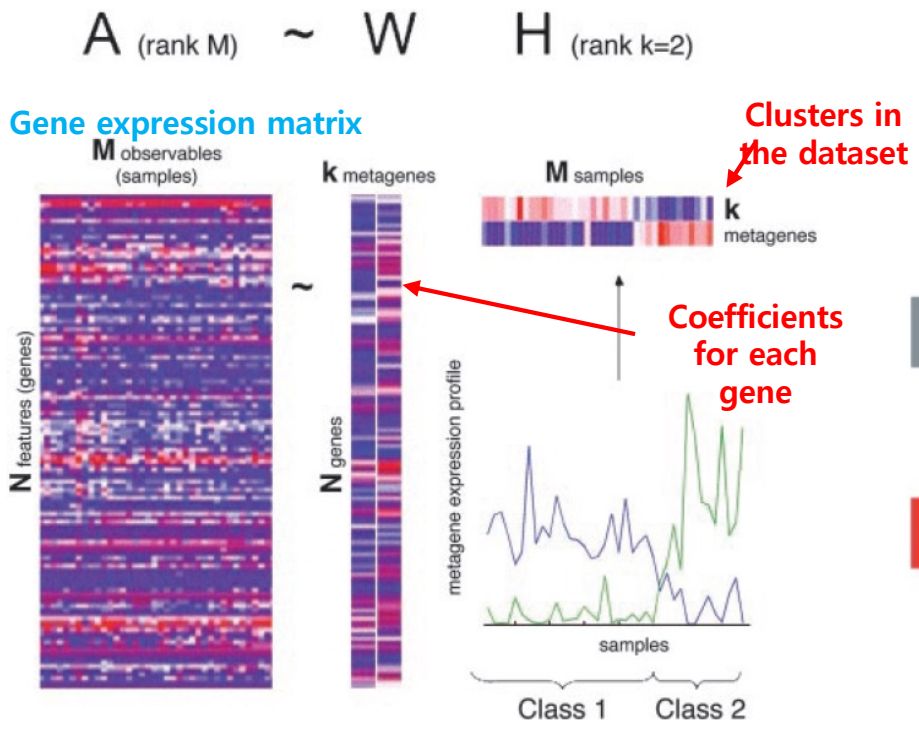
Gene a
Gene b
Gene e
Gene f
Gene g
Gene h
Gene j

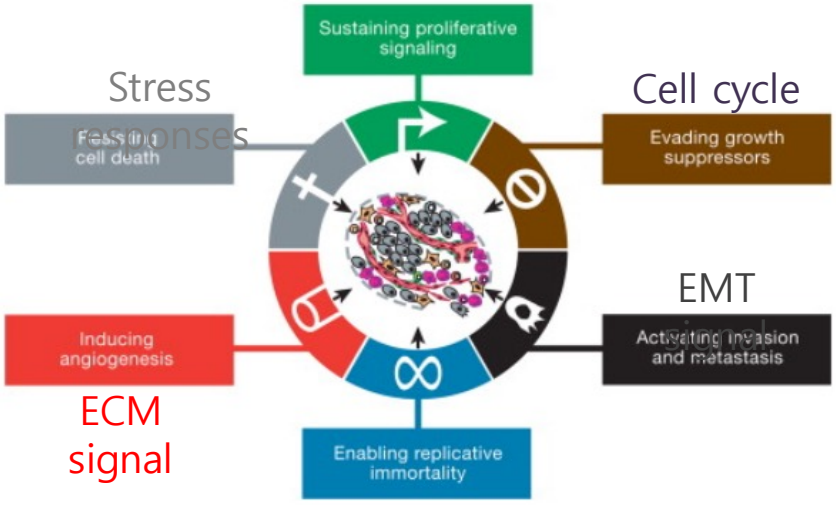


Bulk-sample RNA-seq analysis

Clustering & decomposition

- Non-negative matrix factorization (NMF)
- NMF is a decomposition technique breaking down a matrix (A) into two smaller matrices (W and H)





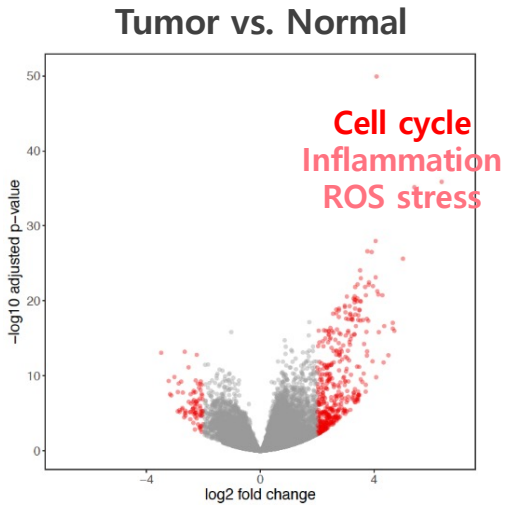
D Hanahan and AW Robert *Cell* (2011)

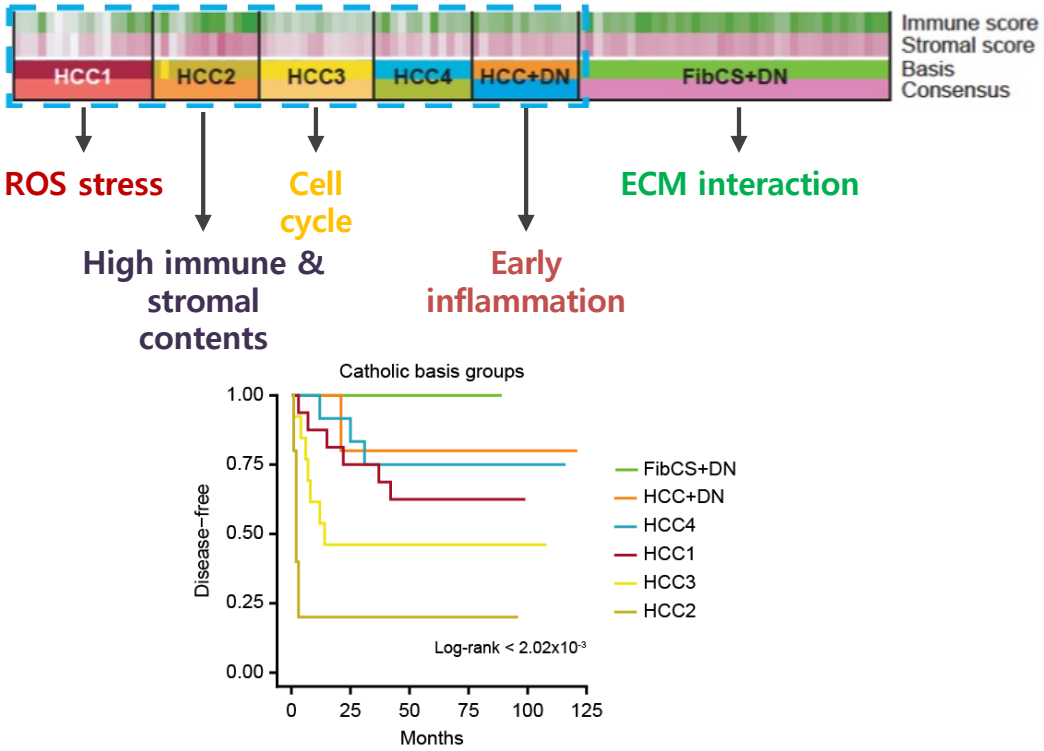
JP Brunet *et al., Proceedings of the national academy of sciences* (2004)

Bulk-sample RNA-seq analysis

Liver cancer study

- Molecular subtypes in Korean liver hepatocellular carcinoma patients
- In the datasets, premalignant (fibrosis~) and malignant samples were sequenced

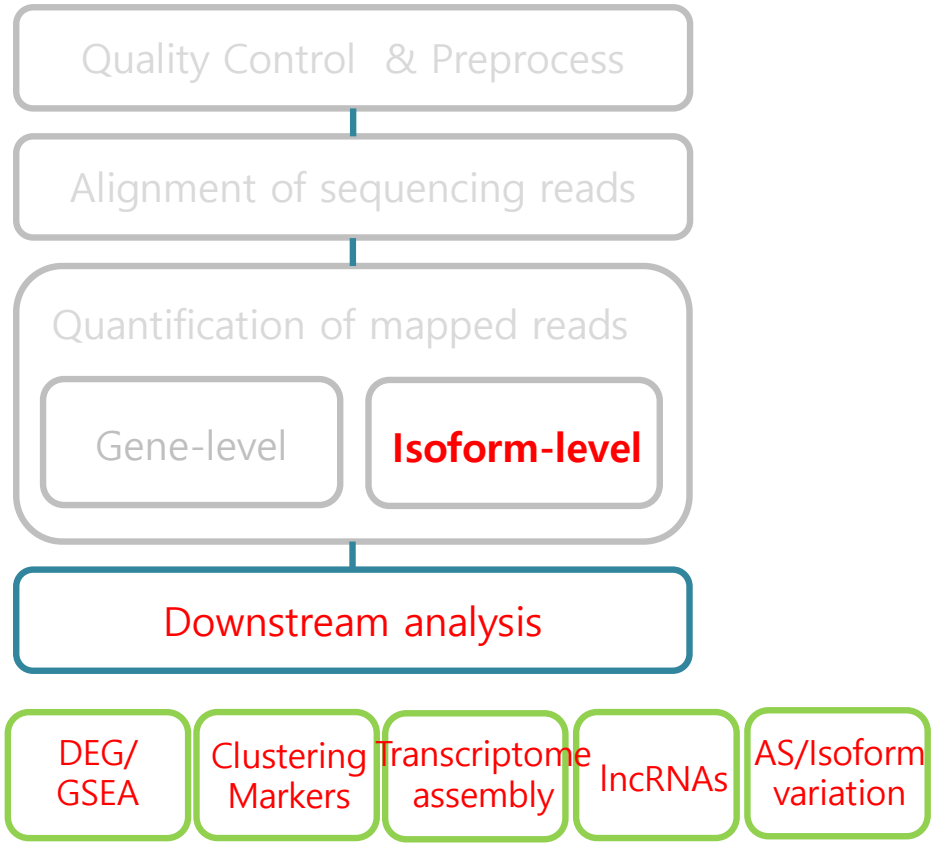




Bulk-sample RNA-seq analysis

Downstream analyses

- There are variety of options in RNA-seq downstream analysis
- Here, we are going to briefly review what we can do using gene expression values

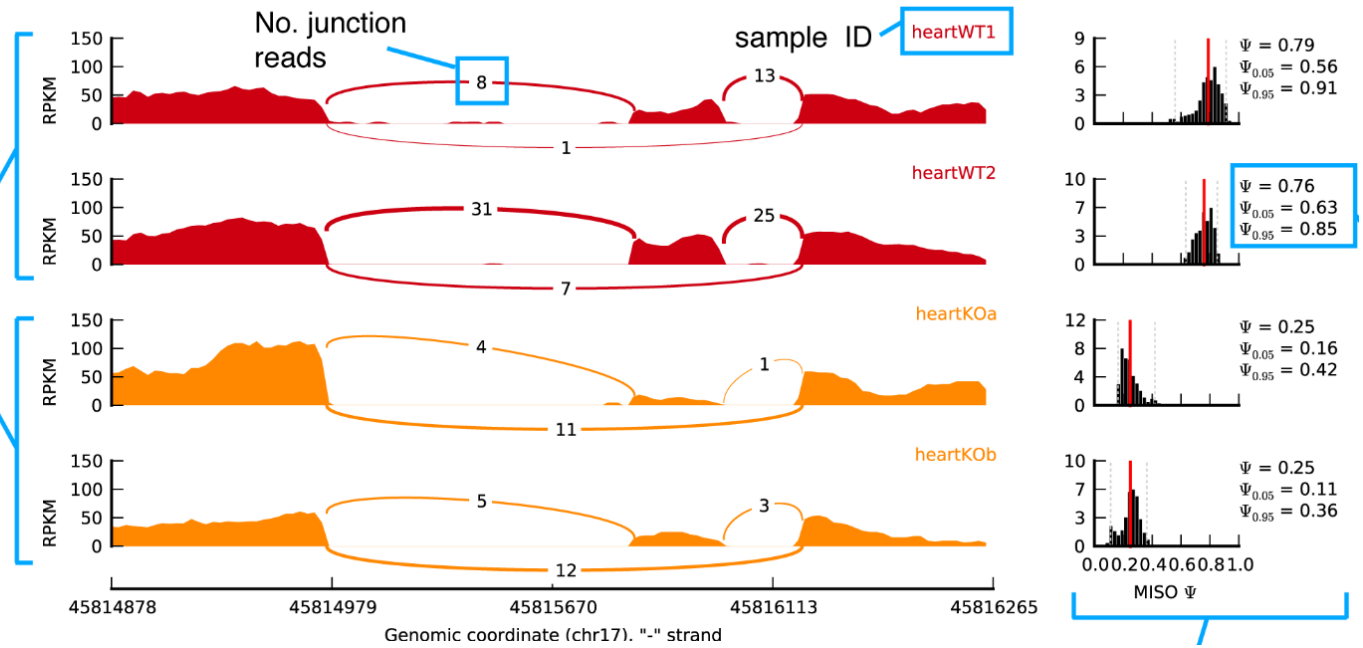


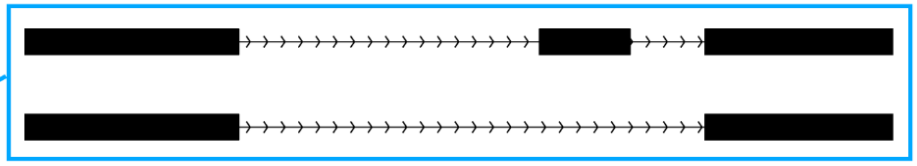
Isoform analysis

Event name

chr17:45816186:45816265:-@chr17:45815912:45815950:-@chr17:45814875:45814965:-

RNA-Seq data for all samples, color-coded by condition

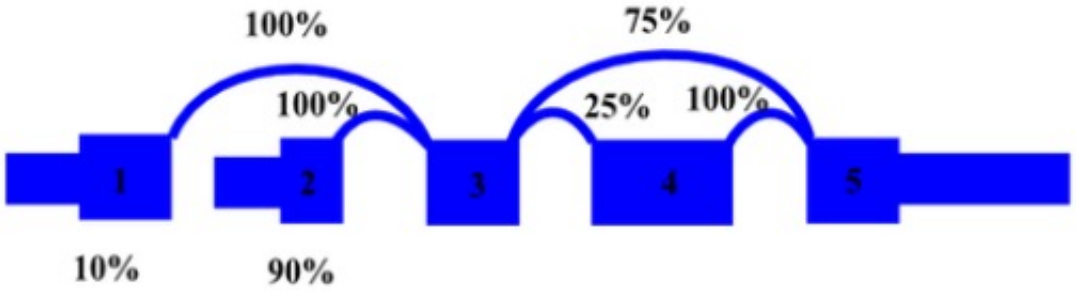




Alternative isoforms from GFF file

Katz Y, Wang ET, Airoldi EM, Burge CB. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods 7, 1009-1015. [abstract]

Expression profiles of isoforms using NGS



Exons 2-3-5 Isoform Abundance: 90% x 100% x 75% = 68%

Exons 2-3-4-5 Isoform Abundance: 90% x 100% x 25% x 100% = 23%

Exons 1-3-5 Isoform Abundance: 10% x 100% x 75% = 7%

Exons 1-3-4-5 Isoform Abundance: 10% x 100% x 25% x 100% = 2%

Ryan et al., Bioinformatics. 2012 Sep 15; 28(18): 2385–2387.

RNA-seq procedure

Input: FASTQ (reads)

Output: read counts (or RPKM) per gene

1. **Preprocessing of raw reads (fastq)**
2. **Mapping to genome (STAR or RSEM)**
3. **Read counts per gene (or calculating RPKM)**
4. **Differential gene expression**