Differential Expression Analysis

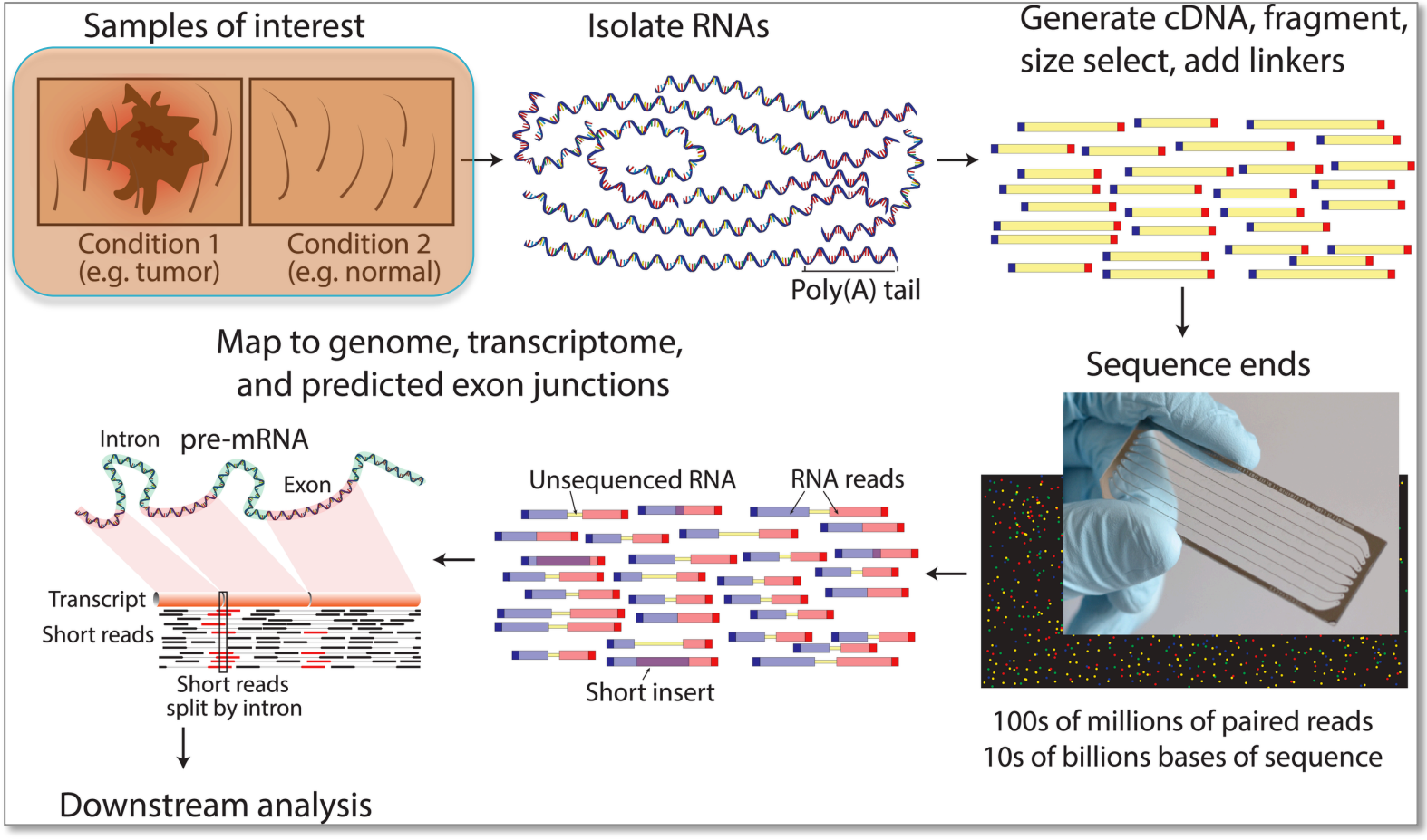
657.001 Transcriptomics (WS 2018/19)

Main question of differential expression (DE) analysis

- Which transcripts are expressed more in which sample?
  - Difference should be statistically significant
  - GR vs GS in our example RNA-seq data
  - Can compare any conditions of interest

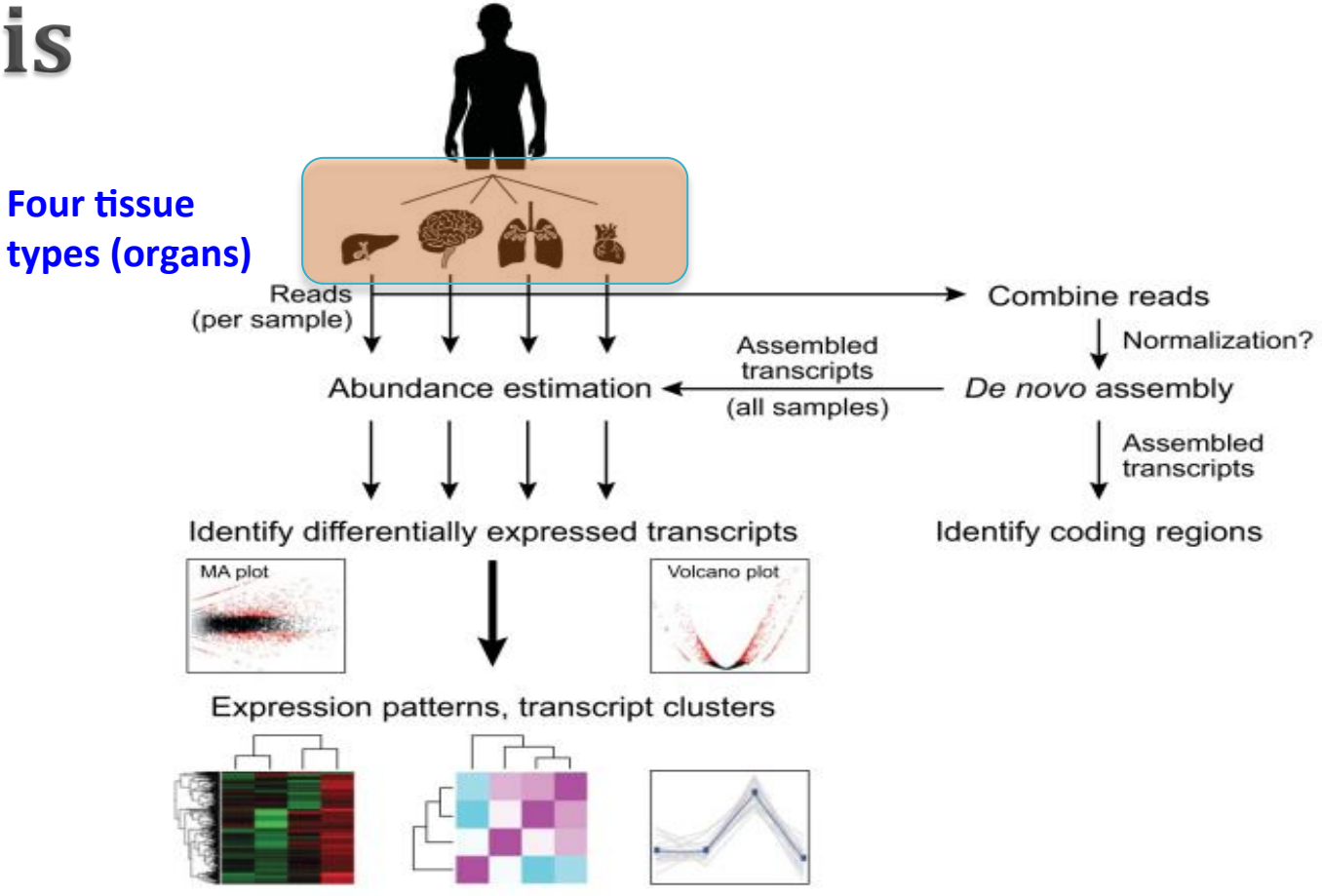
Overview of RNA-seq

Two conditions



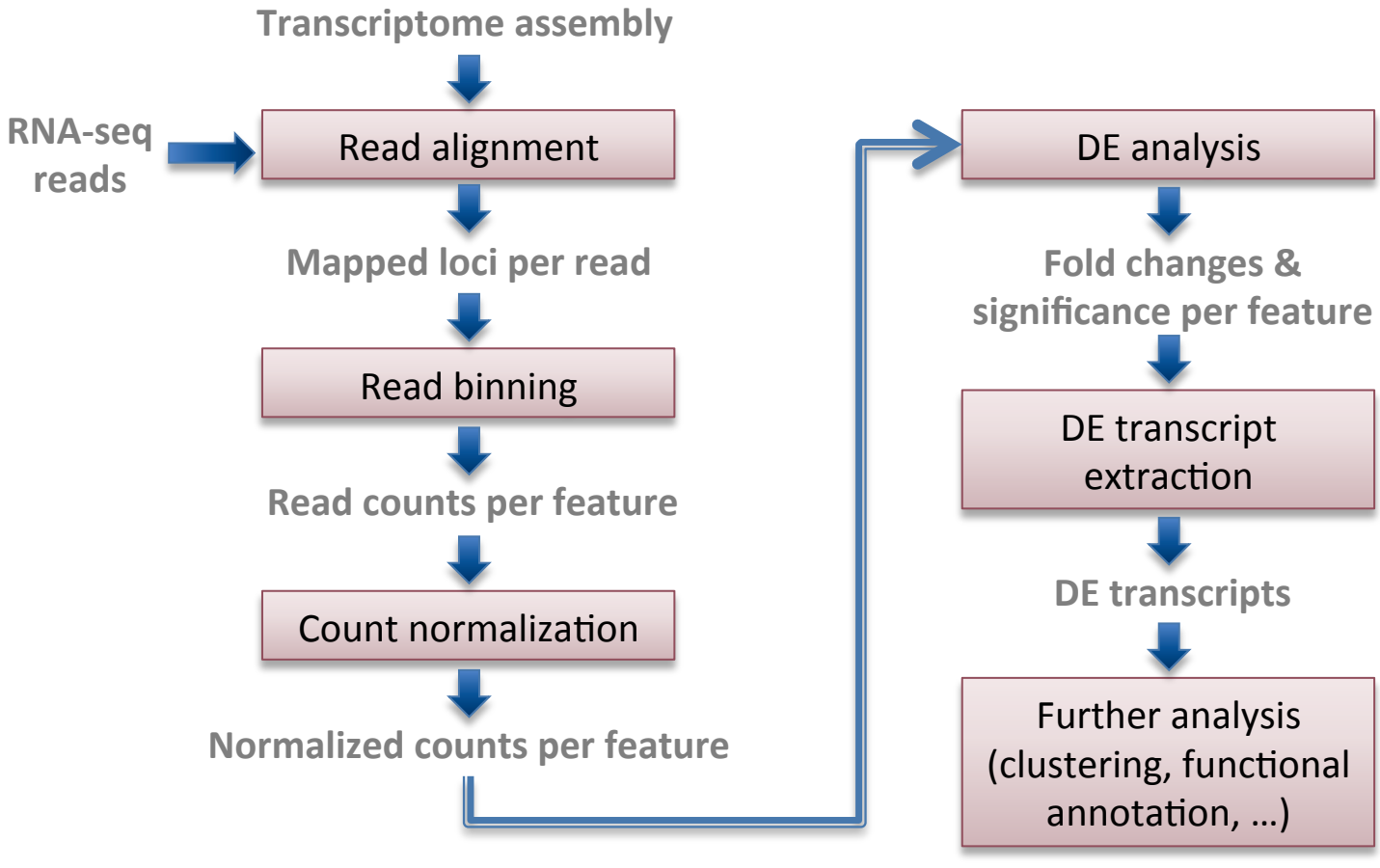
https://commons.wikimedia.org/wiki/File:Journal.pcbi.1004393.g002.png

De novo transcriptome assembly and analysis



Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature Protocols*, *8*(8)

DE analysis workflow



Transcript abundance estimation

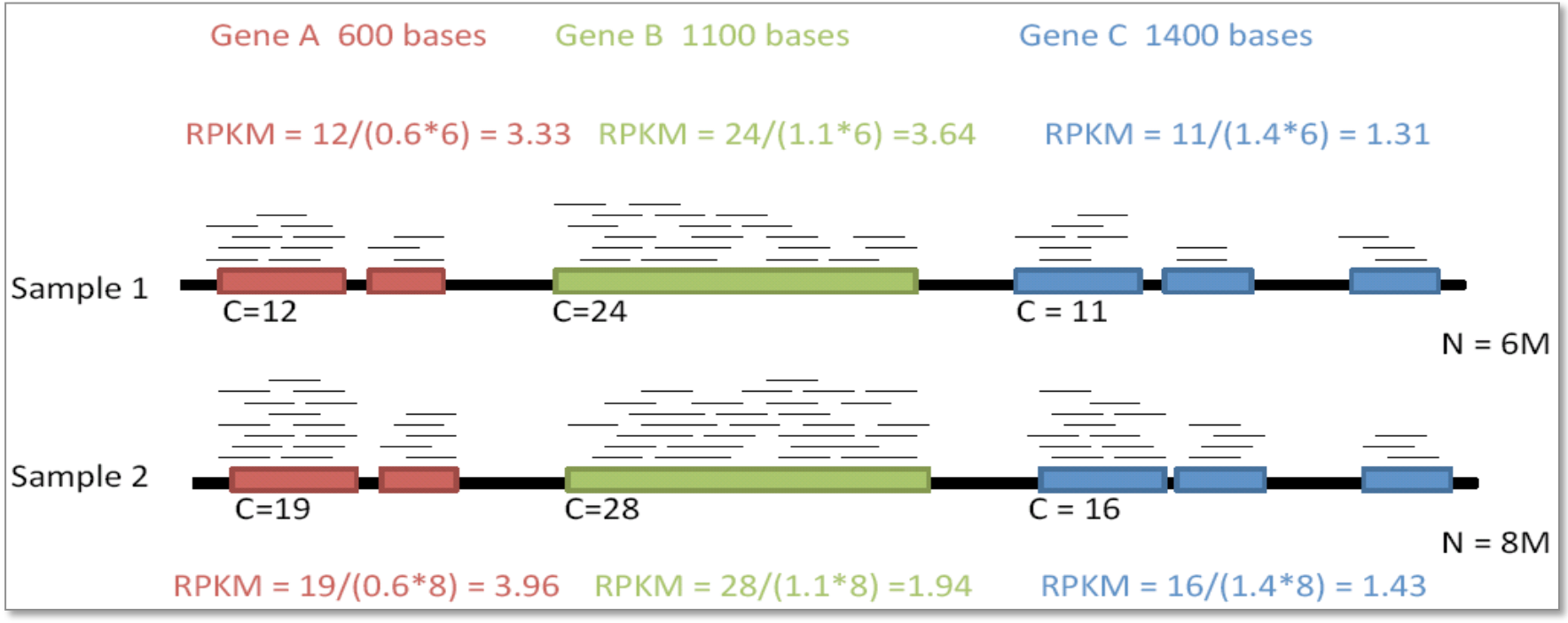
- For each transcript, count total number of reads mapped
  - Also called “binning” the reads
  - Counts are not directly comparable across features or samples (yet)
  - Need normalization into comparable expression values

Normalizing counts

- Why normalize?
  - Longer features (naturally) can have more reads mapped
  - Deeper sequencing produces more reads

- RPKM (or FPKM) most commonly used
  - Reads (Fragments) per Kilobase per Million reads
  - Defined as $C/(LN)$
    - C = number of reads mapped to a feature
    - L = length of the feature (in kilobases)
    - N = total number of reads from the sample (in millions)

RPKM examples



http://jura.wi.mit.edu/bio/education/hot_topics/RNAseq/RNA_Seq.pdf

DE analysis

- Compare quantification values across samples or across features
  - The goal is to find differentially expressed (DE) features
- Many tools summarize/normalize counts and suggest DE features
  - Cufflinks/Cuffdiff, R packages (DESeq, edgeR, baySeq, TSPM), Samtools
- Determination of DE features depends on
  - Fold changes (FC)
  - Statistical significance of FC (FDR, p-value)

Lab overview