RNA-Seq Assembly and Analysis

Dr. Jung Soh
MOL.923 SS 2018

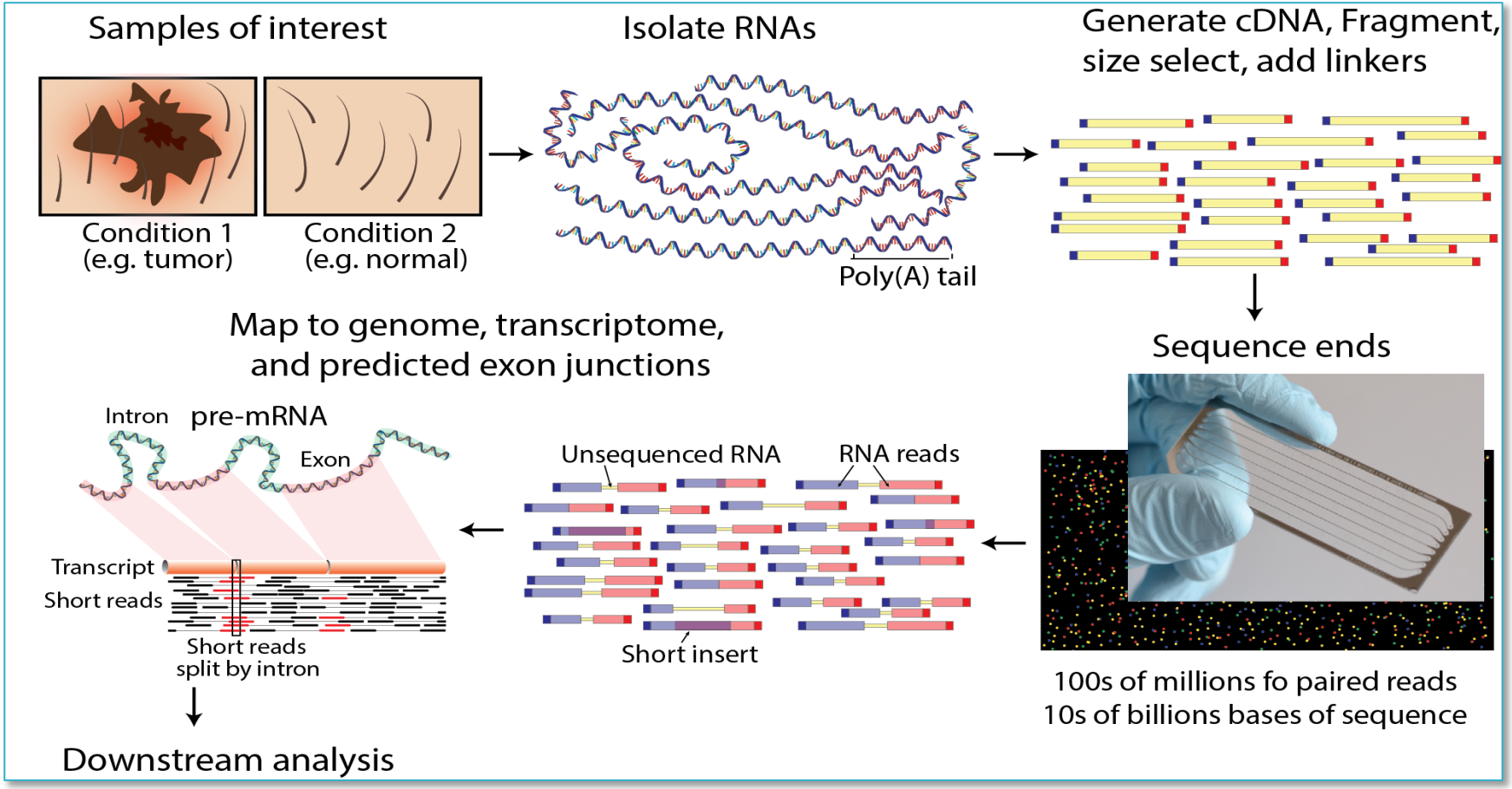
What is RNA-seq?

- An experimental protocol that uses next-generation sequencing technologies to sequence the RNA molecules within a biological sample in an effort to determine the primary sequence and relative abundance of each RNA
  - Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet. 12(10):671-682

Why RNA-seq?

- Some studies not possible with DNA sequences
  - Novel transcript discovery
- To study functions based on gene expression changes
  - Drug treatment vs. no treatment
  - Patients vs. healthy people
  - Wild type vs. knock-out
- Some features available only at RNA level
  - Alternative isoforms, transcript fusion

RNA sequencing

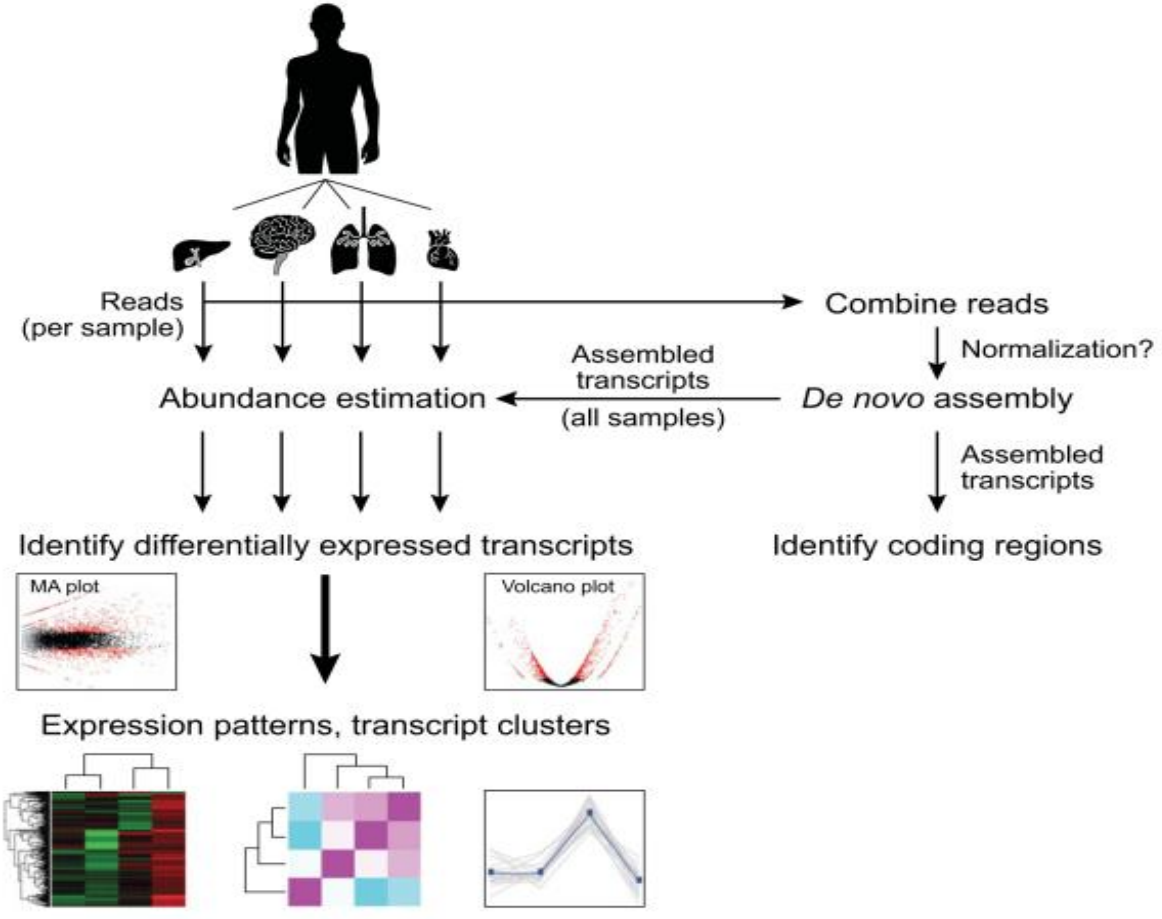


https://bioinformatics.ca/rnaseq2016module-2-ppt

Challenges of RNA-seq

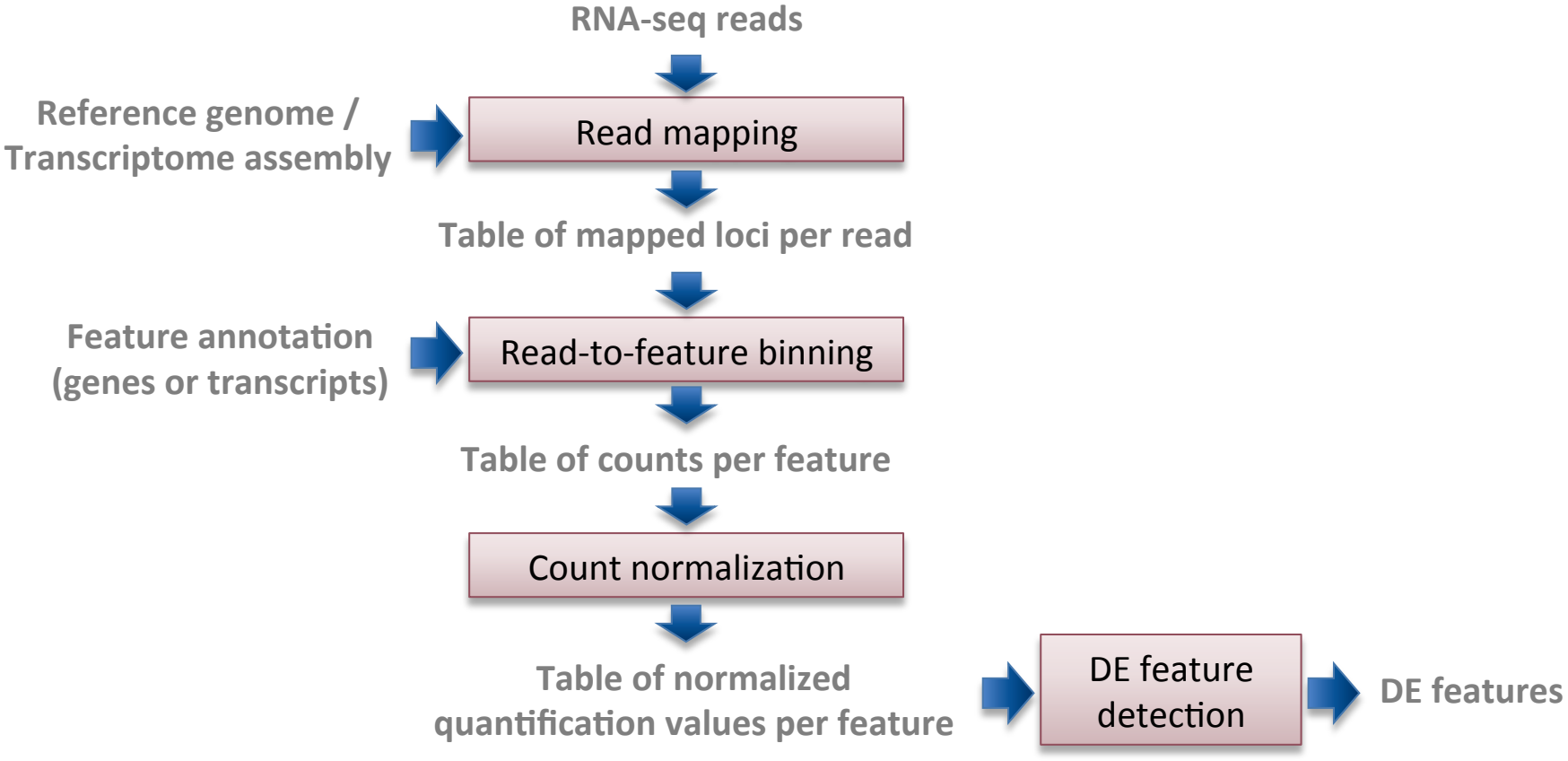
- Difficulty with sampling
  - More fragile than DNA
  - Different sizes of RNA
- Relative abundance of RNAs hard to control
  - Can vary by orders of magnitude
  - Uneven coverage
    - Many reads from a small number of highly expressed genes
- Computational analysis challenges

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)

Differential expression (DE) analysis

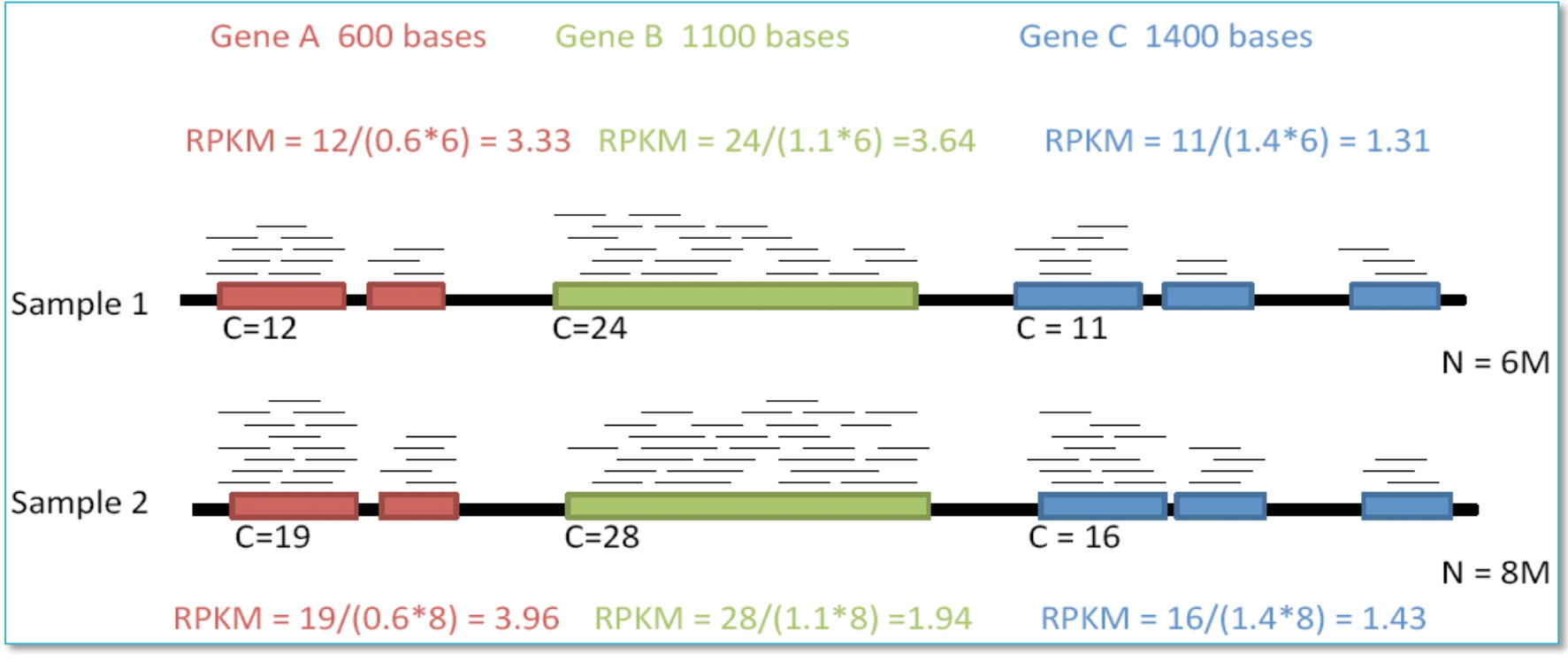


Normalizing counts

- Why normalize?
  - Longer features 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

Detecting DE features

- Compare quantification values across samples or across features
  - Which features are more expressed (up-regulated) or less expressed (down-regulated) under one condition vs another?
- DE analysis tools summarize/normalize counts and suggest DE features
  - Cufflinks/Cuffdiff, R packages (DESeq, edgeR, baySeq, TSPM), Trinity
- Downstream analysis of DE features
  - Cluster analysis
  - Correlation analysis

Lab overview