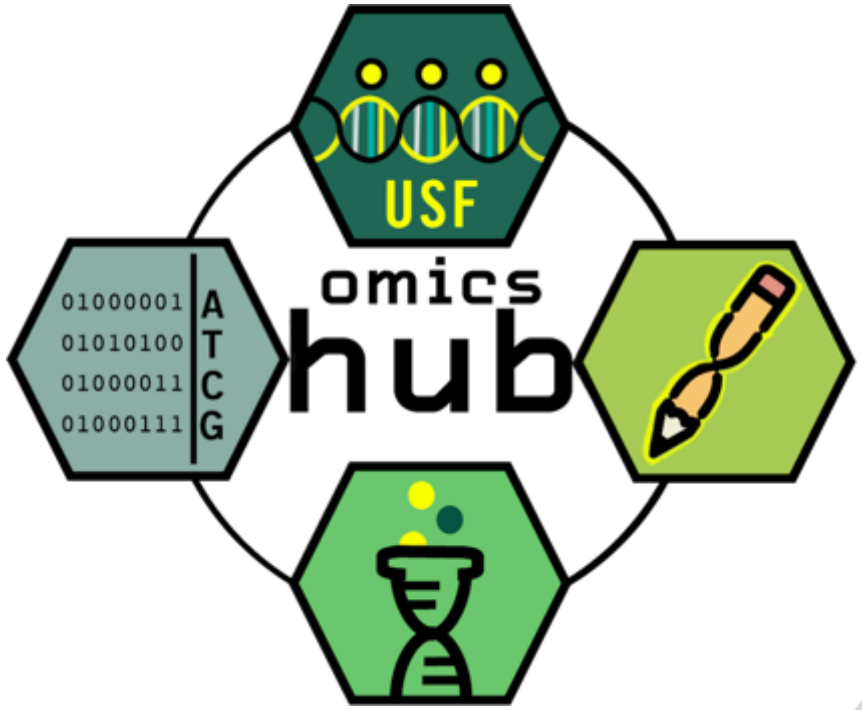
USF GENOMICS PROGRAM

RNA-seq Data-Analysis Workshop



Intro to RNA-seq

Justin Gibbons, PhD

Postdoc, USF Genomics Program

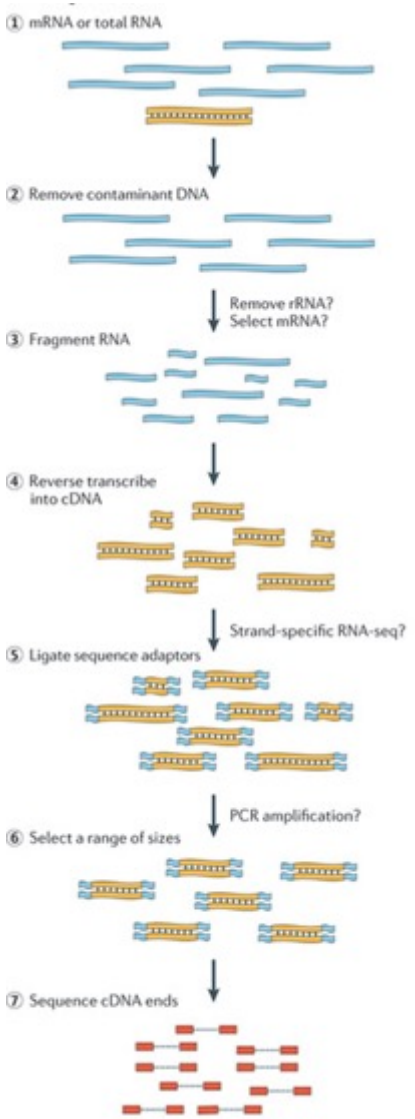
Consultant, USF Omics Hub

February 2020

RNA-seq

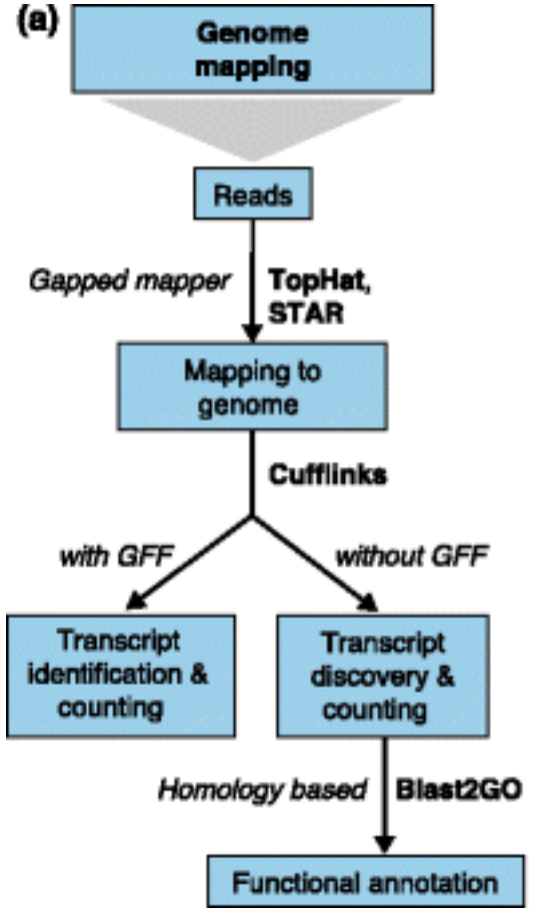
- A major breakthrough (replaced microarrays) in the late 00's and has been widely used since
- Measures the **average expression level** for each gene across a large population of input cells
- Useful for comparative transcriptomics, e.g. samples of the same tissue from different species
- Useful for quantifying expression signatures from ensembles, e.g. in disease studies

Data Generation



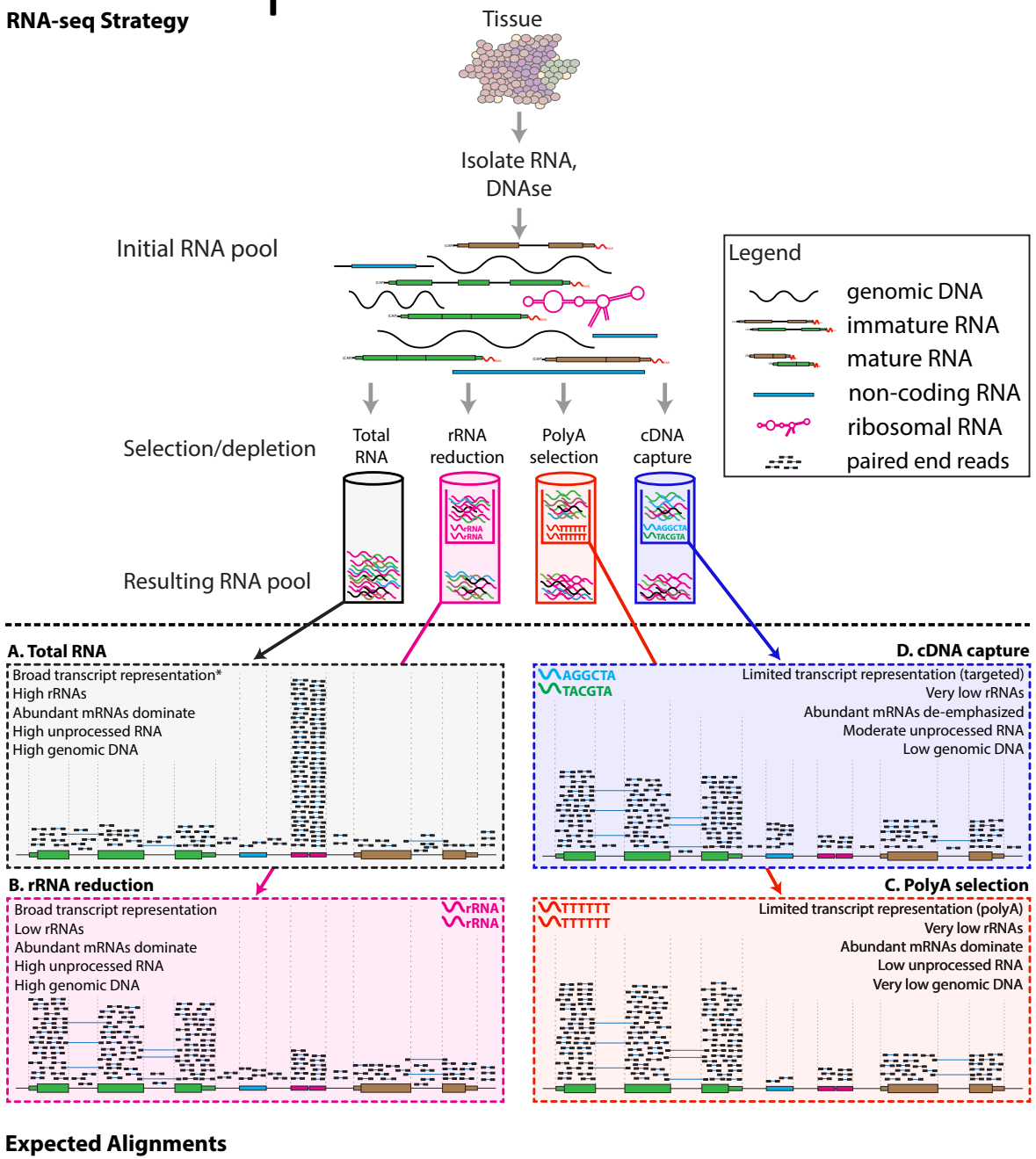
Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nature Rev. Genet.* **12**, 671–682 (2011).

Data Analysis



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

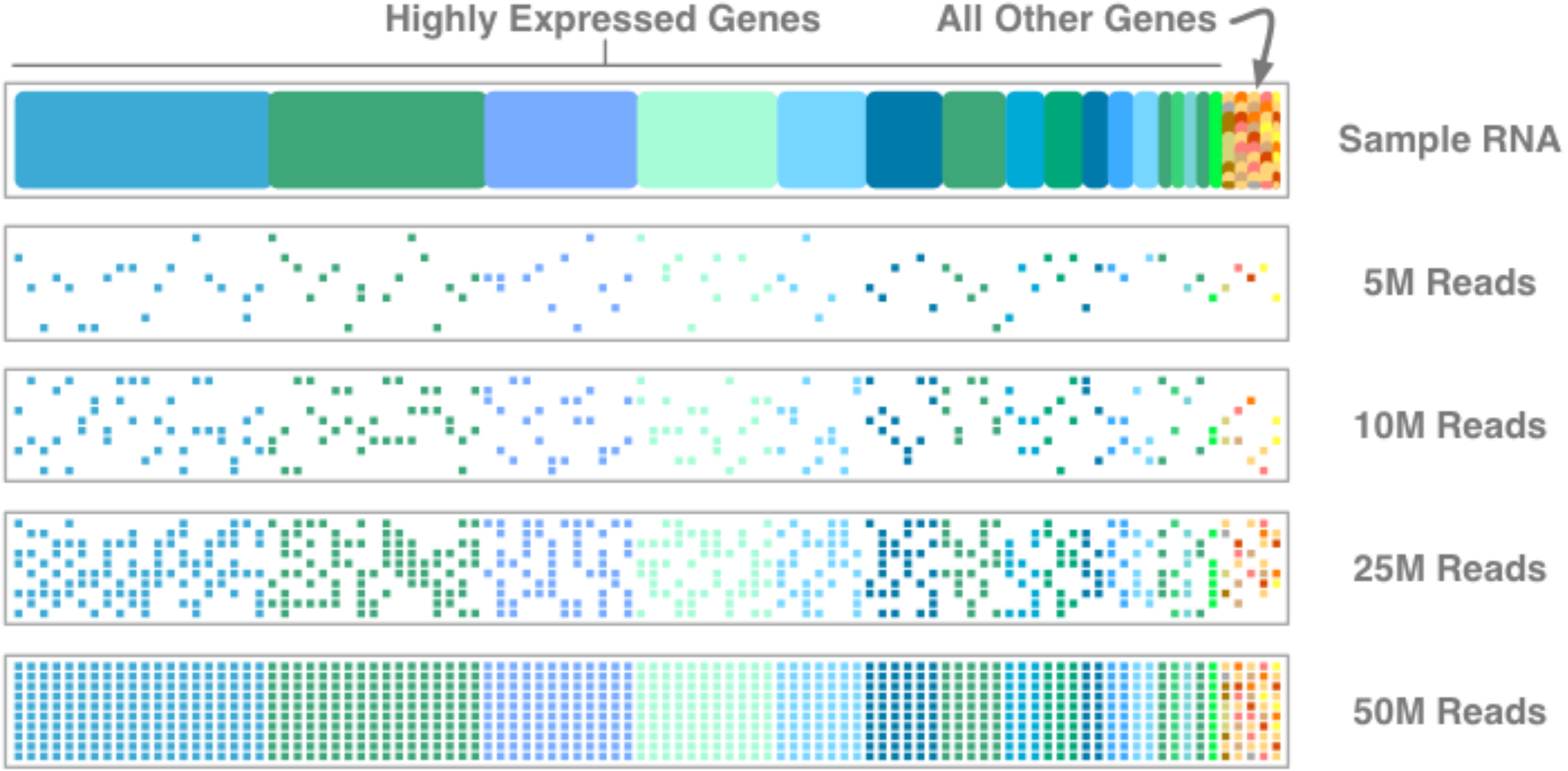
Types of RNA-seq



Experimental design: Number of replicates

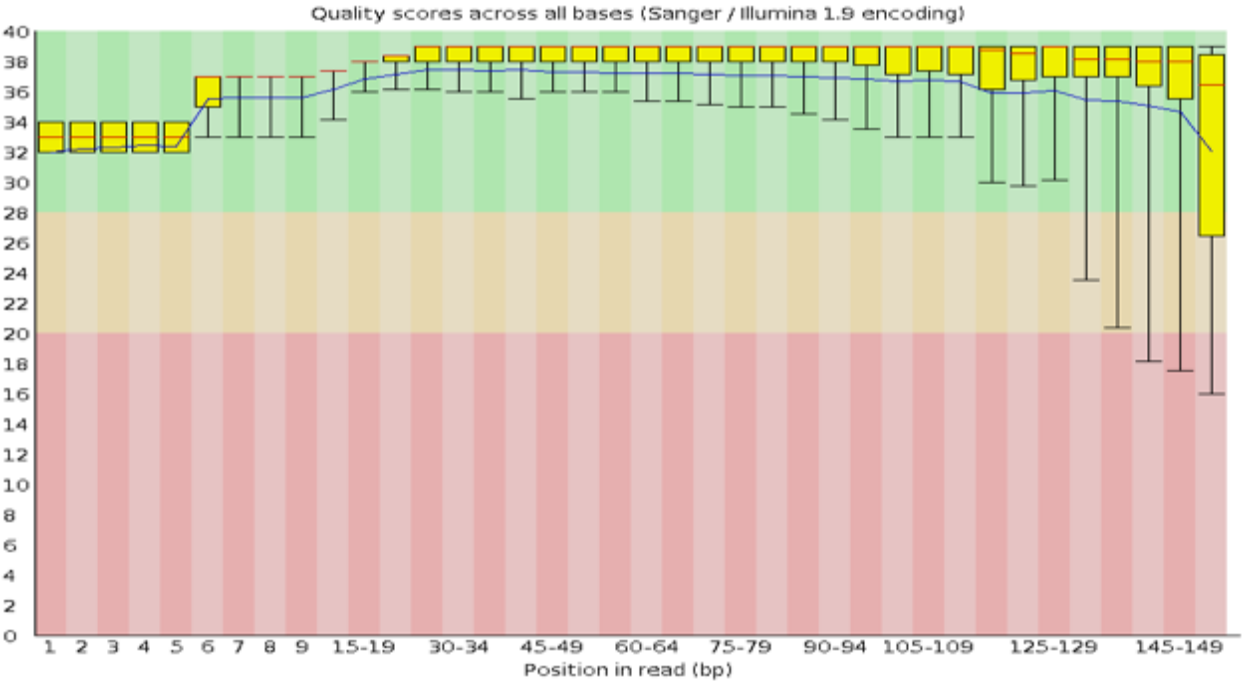
- Number of replicates more important than read depth or read length [93]
- Factors dictating sample size:
  - Effect size
  - Within-group variation
  - Acceptable false-positive and false-negative rates
  - Maximum sample size
- Tools for calculating sample size:
  - Scotty—Power Analysis for RNA Seq Experiments
    - Website
    - Uses a pilot run or publicly available to perform power calculations
    - Allows modeling of how much additional power costs ($$$)
  - PROPER—R package for RNA-seq power calculations
    - Creates simulations of RNA-seq data from provided data
    - Creates plots demonstrating how sample size and sequencing depth affect true discovery rate and false discovery rate

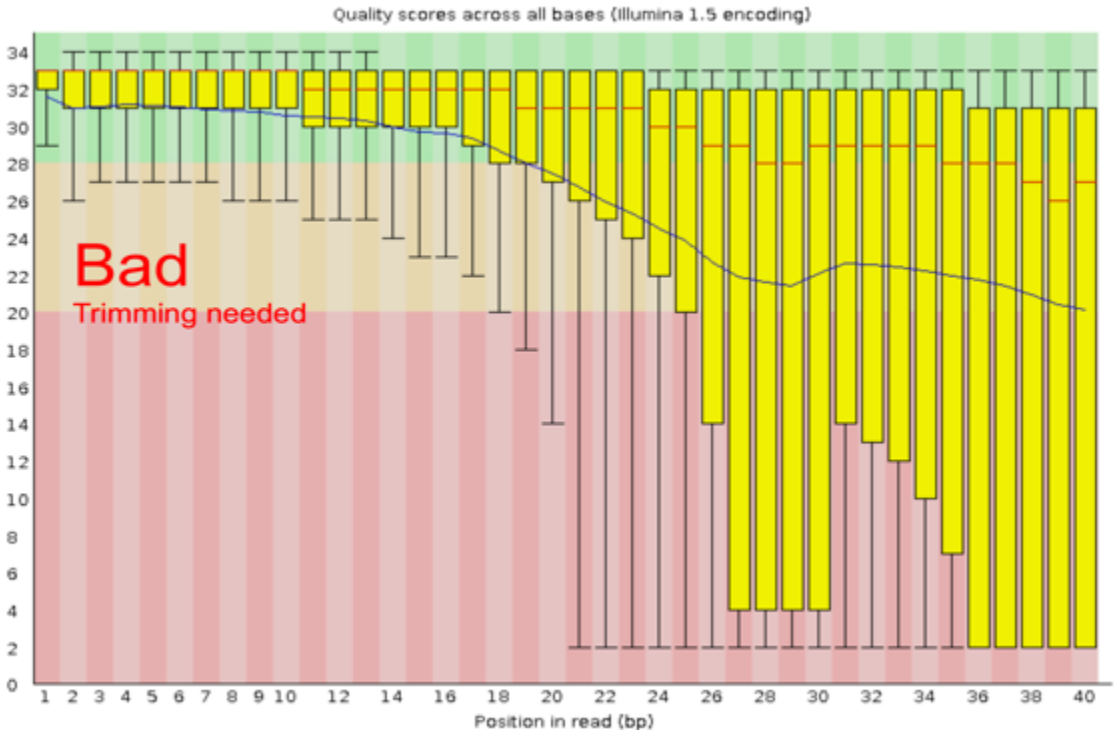
Experimental design: Sequencing depth



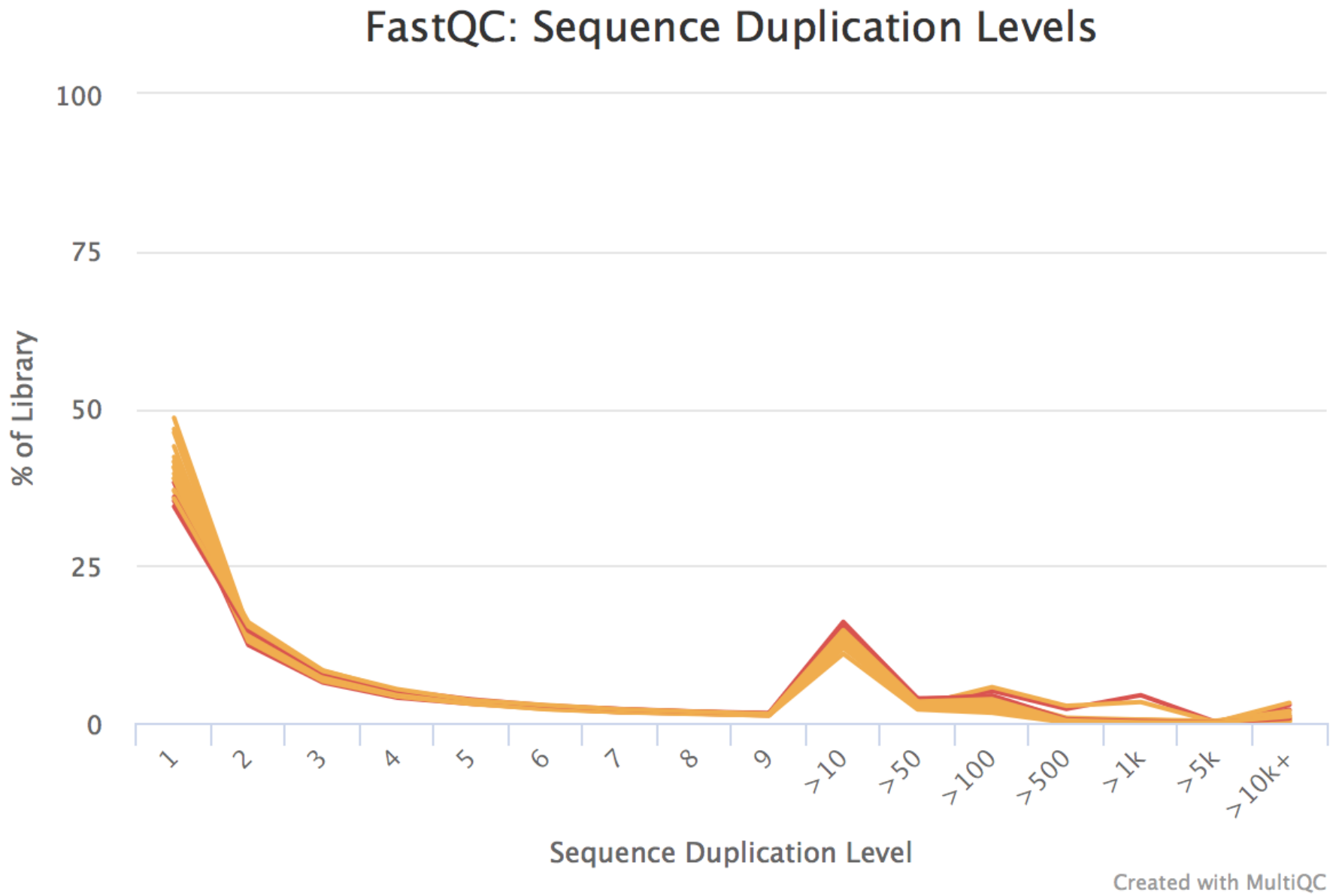
First analysis step: Quality control

Quality control: Base quality scores

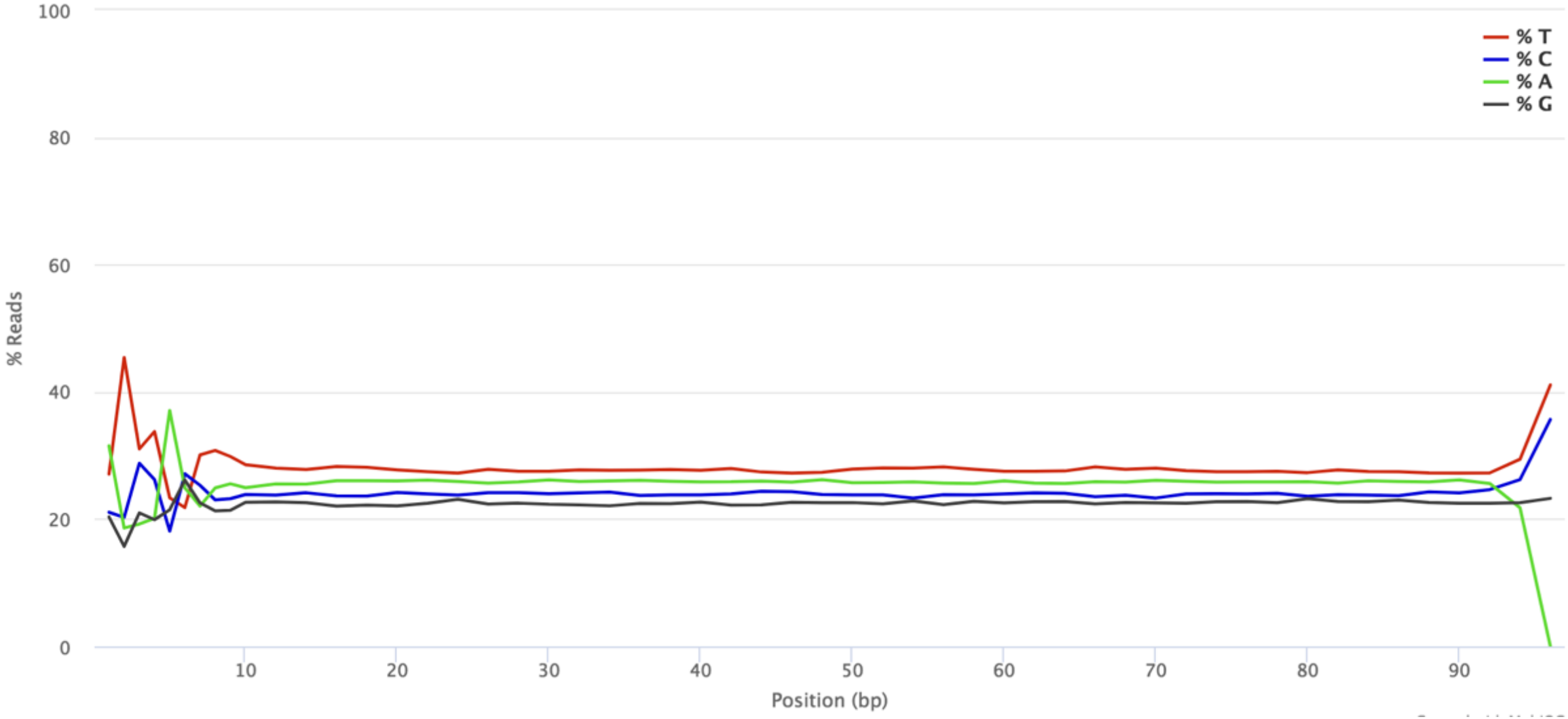




Quality Control: Sequence Duplication

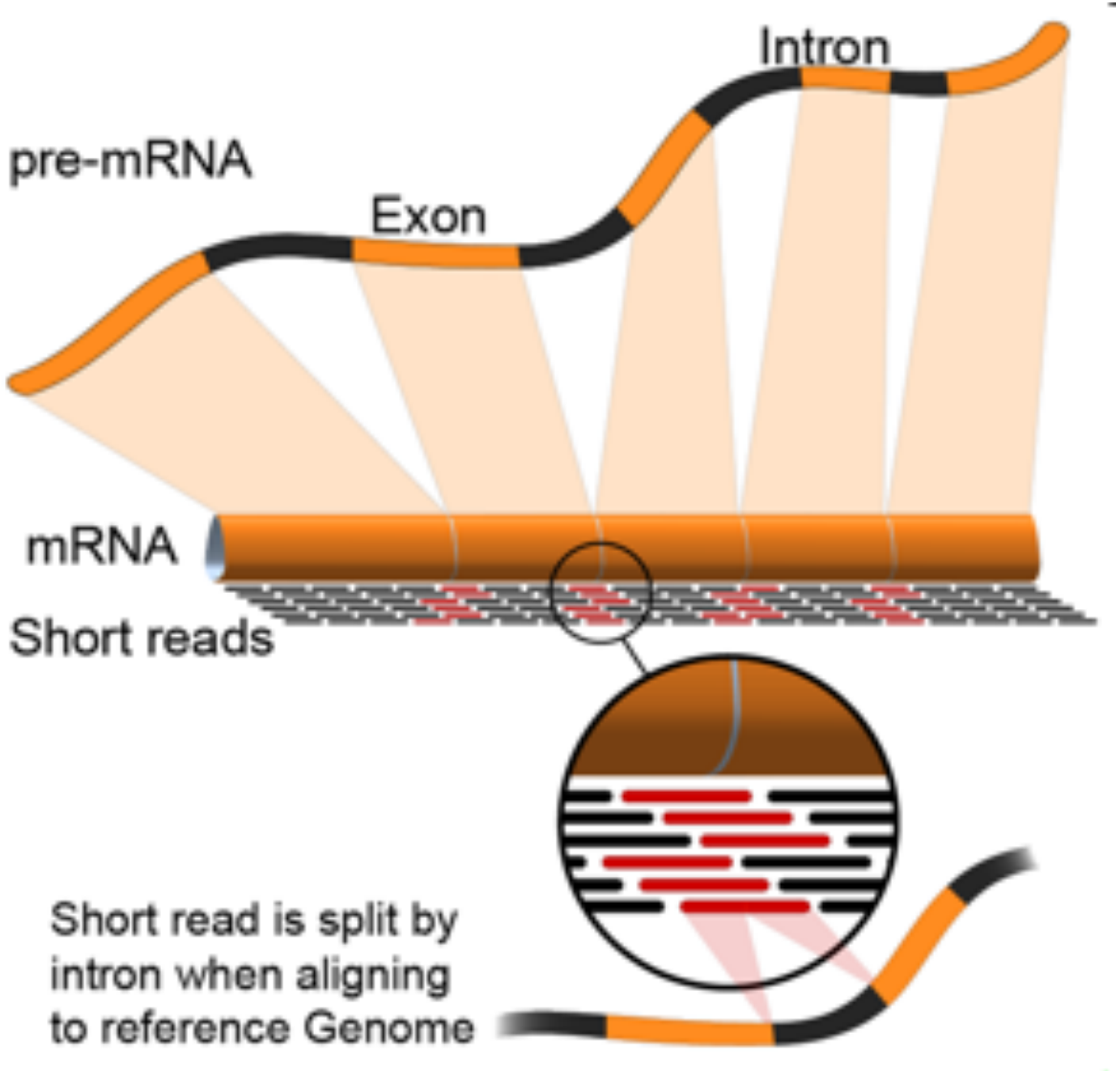


Quality Control: Composition bias



Second analysis step: Transcriptome reconstruction

Transcriptome reconstruction



What we will be doing:

1. HISAT2: Align reads to a reference genome
2. Cufflinks: Assemble reads into transcripts
3. Cuffnorm: Get normalized gene counts
4. featureCounts: Get raw gene counts