Introduction to RNA-seq

Childhood Cancer Data Lab

There is no optimal pipeline for the variety of different applications and analysis scenarios in which RNA-seq can be used. Scientists plan experiments and adopt different analysis strategies depending on the organism being studied and their research goals.

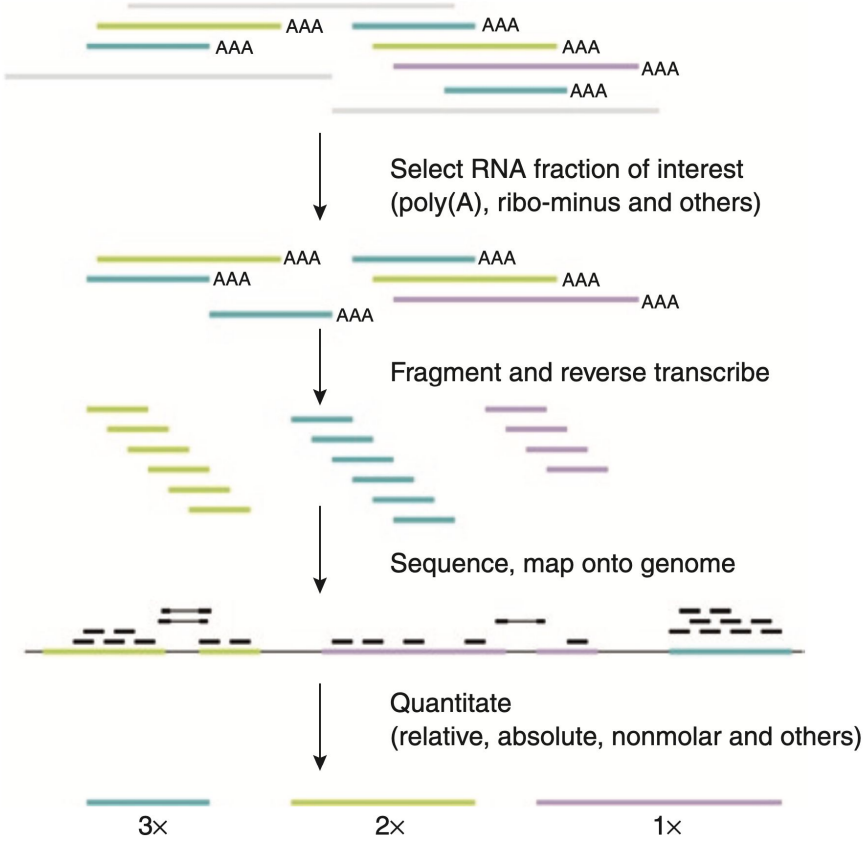
- Conesa *et al.* 2016

Today’s Objectives

- Learn to navigate the terminal interface

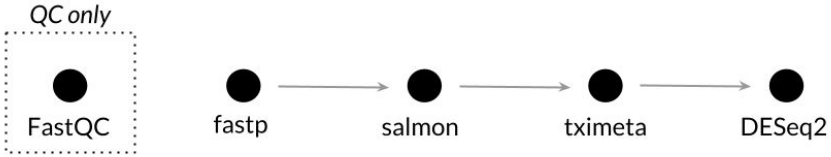
- Demonstrate how to:
  - Perform quality control checks using **FastQC** and preprocess reads with **fastp**
  - Quantify RNA-seq expression with **Salmon**
  - Summarize transcript-level Salmon output to the gene-level with **tximeta**
  - Perform exploratory data analysis with **DESeq2**

RNA-Seq Overview



Pepke *et al.* 2009

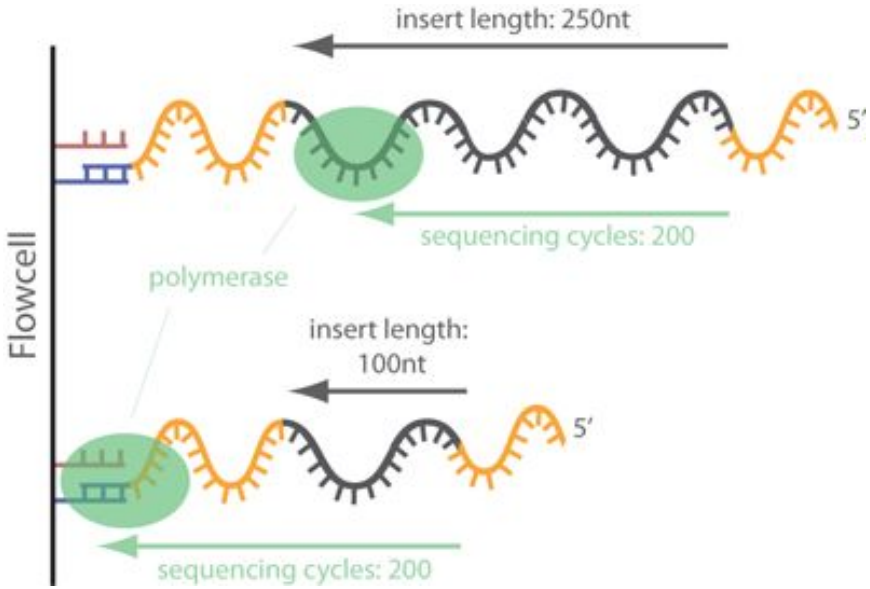
Overview of pipeline



TOOL	**fastp** →	**Salmon** →	**tximeta** →	**DESeq2**
PURPOSE	Adapter trimming, quality filtering, length filtering	Quantification of transcripts via lightweight mapping to *transcriptome*, GC-bias correction	Import of transcript (tx) abundances and counts from Salmon and summary to the gene-level for more robust statistics, accounts for gene length changes across samples due to differential isoform usage	Library size and composition normalization, transformation for visualization and clustering, testing for differential gene expression
INPUT FILES	FASTQ	FASTQ that have been preprocessed with fastp, transcriptome to map against	Estimated counts and abundances from Salmon (transcript metadata is added automatically)	SummarizedExperiment R object which contains unnormalized counts and length information

fastp Adapter trimming, quality filtering, length filtering

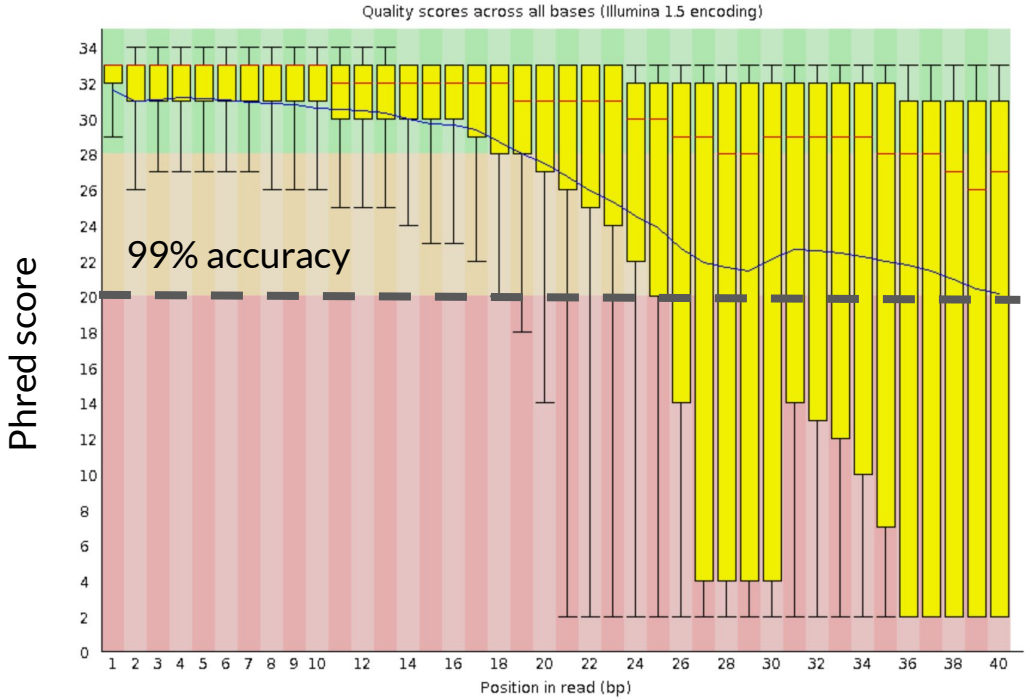
- Performs adapter trimming, quality control, and quality filtering all in one
- Automatically detects and removes adapter sequences
- Removes reads with low quality bases
- Removes reads below minimum read length
- Outputs QC and filtering results into a single HTML file



[Image from ECSEQ Bioinformatics]

fastp Adapter trimming, quality filtering, length filtering

fastp uses a Phred score to determine base quality, reads with a high percentage of low quality bases are removed.



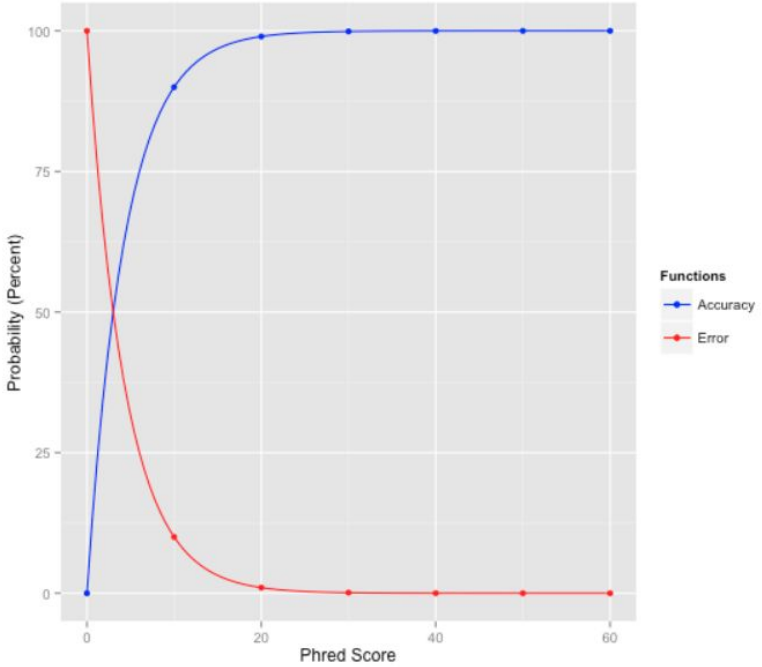
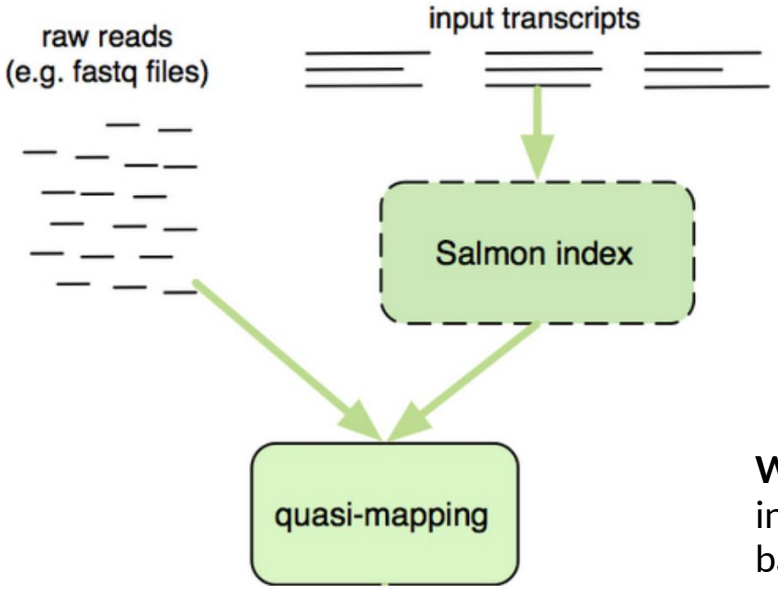


Image from FastQC bad sequence example

Image from GATK Technical Documentation

Salmon lightweight mapping to *transcriptome*



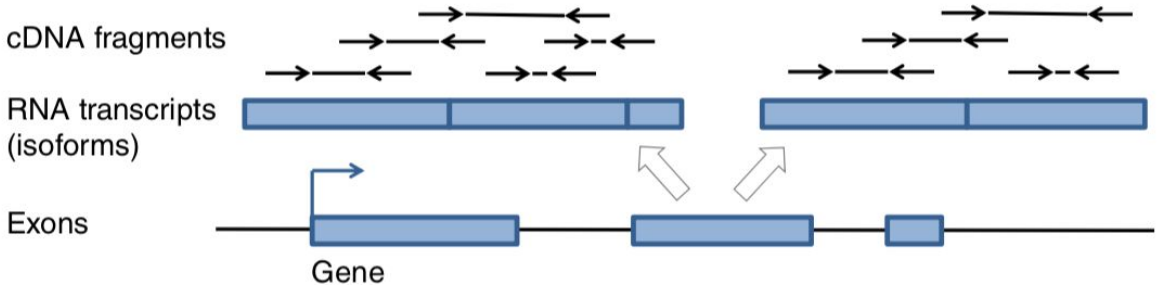
Reference that contains all transcript sequences ("transcriptome")

- Can not identify anything that's not in the transcriptome (e.g., novel isoforms)
- Requires a well-characterized reference transcriptome

Where do the raw reads best map? Identify where informative sequences in the read map without performing base-by-base alignment.

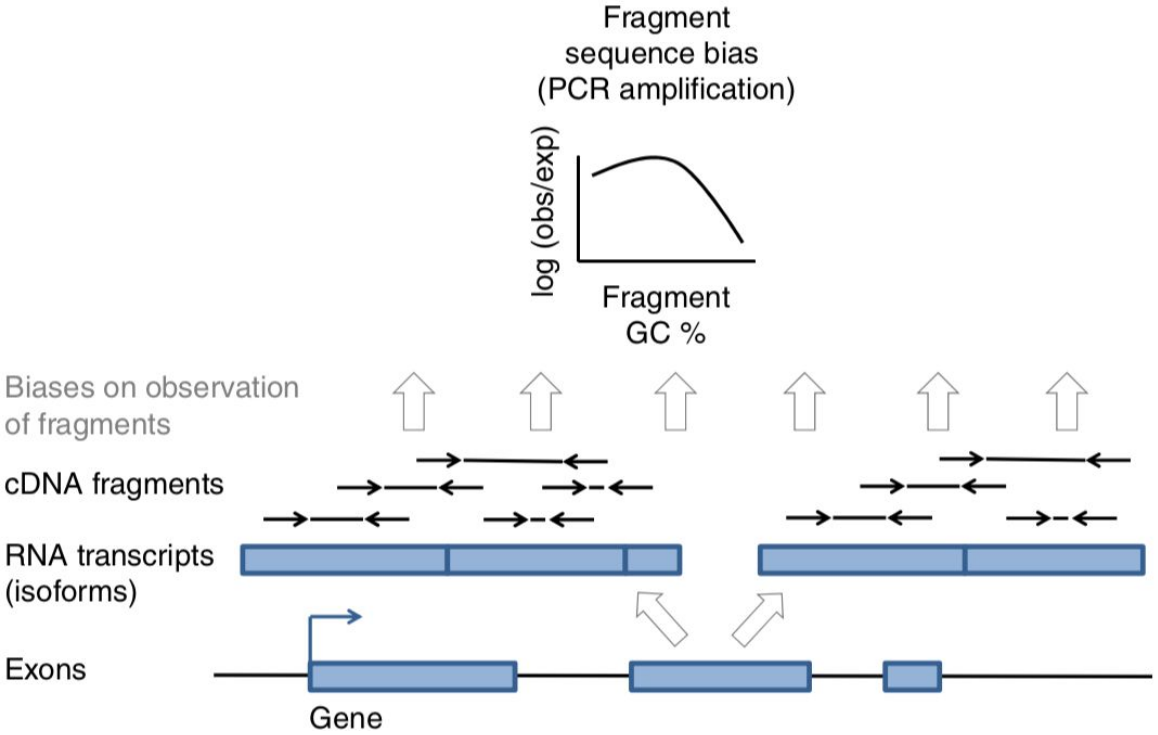
Image from RNA-seq blog via HBC Training

Salmon learning sample-specific biases



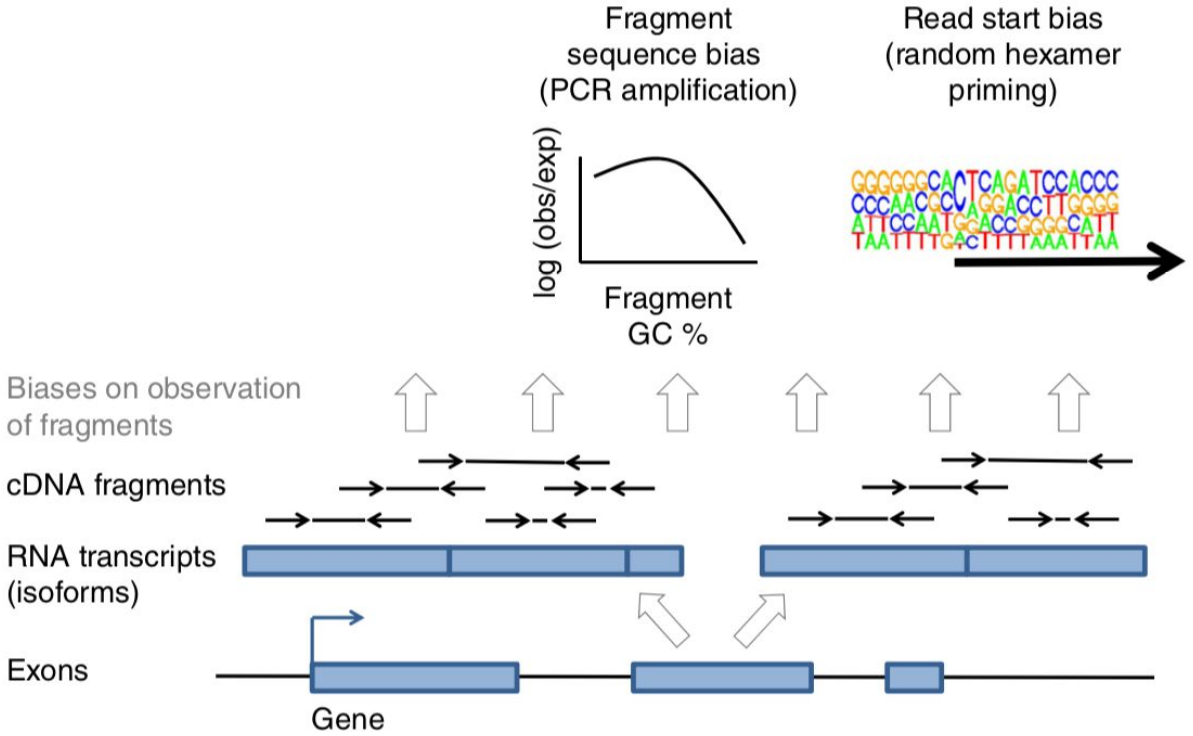
Love *et al.* 2016

Salmon learning sample-specific biases



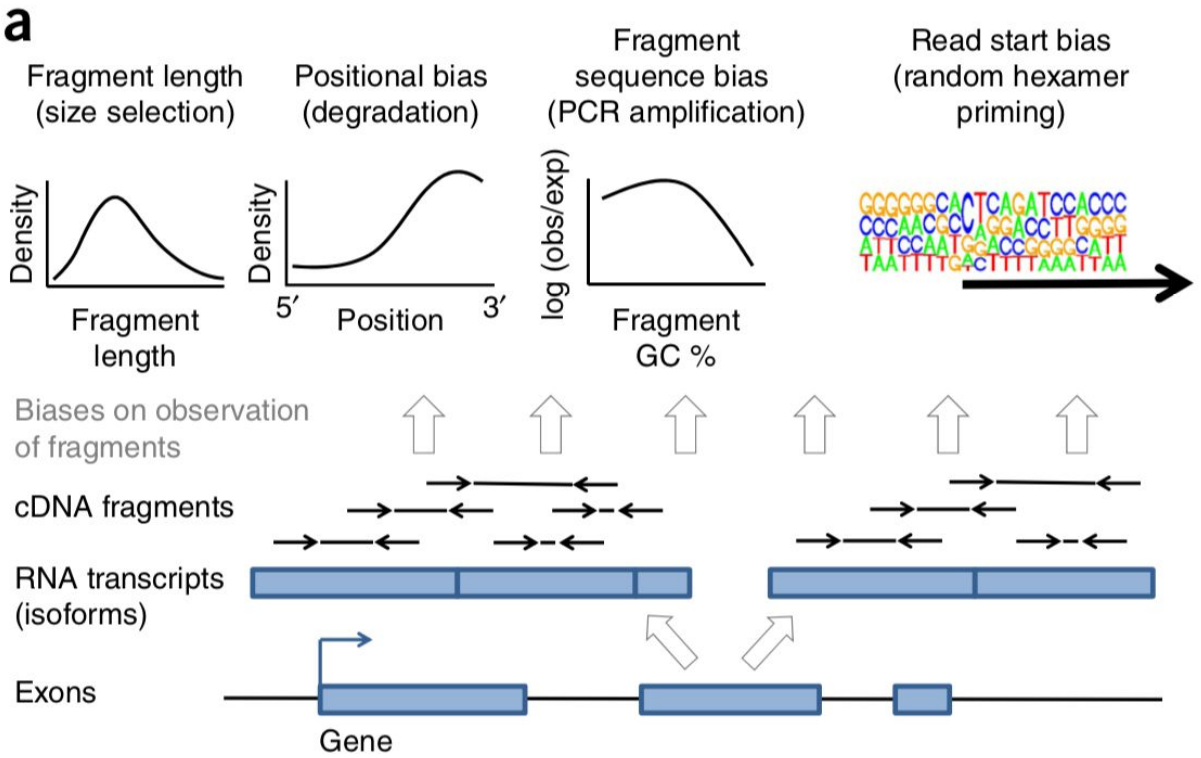
Love *et al.* 2016

Salmon learning sample-specific biases



Love *et al.* 2016

Salmon learning sample-specific biases



Love *et al.* 2016

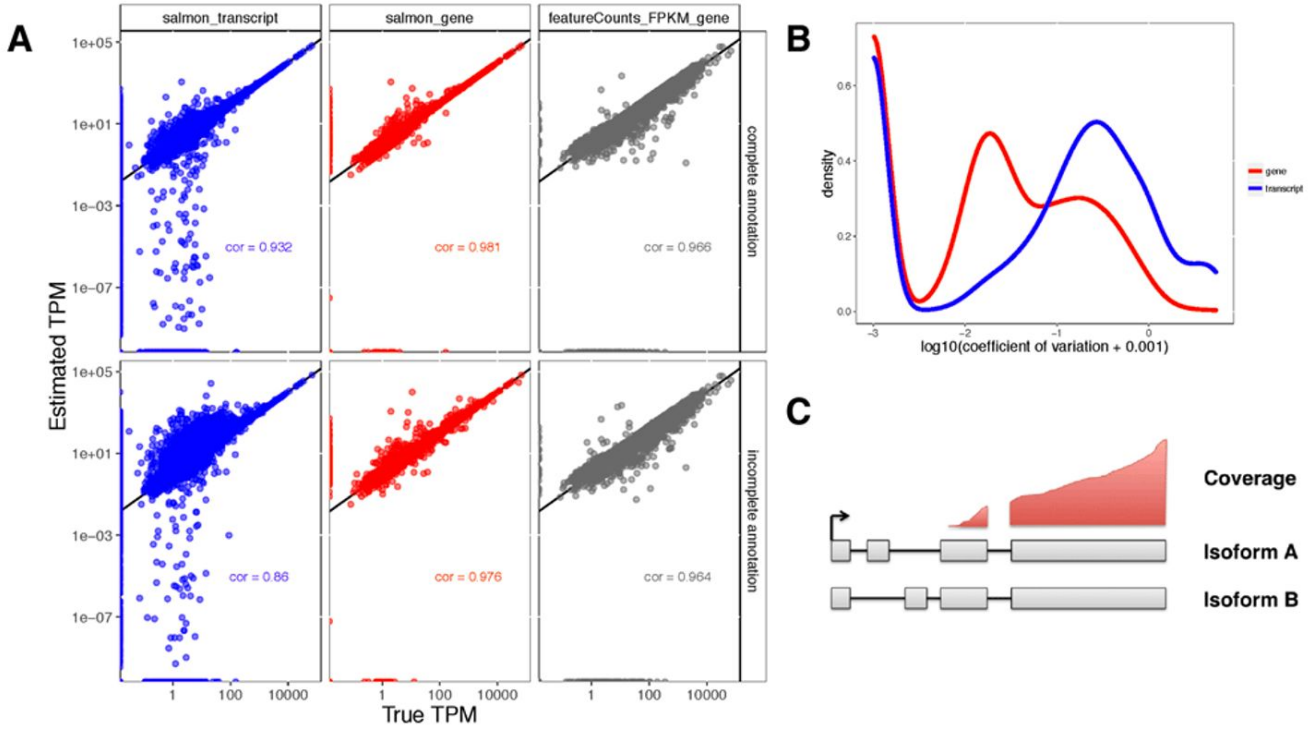
Salmon abundance measures

Salmon output includes the following for each transcript quantified:

- *Effective length* - the computed effective length of the transcript taking into account all factors that affect the probability of sampling fragments from this transcript
- *Transcript length* - longer genes are more likely to be observed
- *Read Counts* - estimate of the number of reads mapping to each transcript, used as input for downstream analysis like differential expression across samples with DESeq2
- *Transcripts per Million (TPM)* - relative abundance of the transcript taking into account the read counts and transcript length, used to compare gene expression within a sample

StatsQuest. RPKM, FPKM and TPM, clearly explained.
HBC Training. Common normalization methods for RNA-seq data.

tximeta import transcript-level and summarize to gene-level



tximeta (a wrapper around ***tximport***) also allows for automatic attachment of metadata during data import.

Soneson *et al.* 2016.

DESeq2 transform data and identify differentially expressed genes

Data Transformation

- Transform data across all samples being compared
- Needed to minimize the amount of variance in the data explained by technical bias (i.e., eliminating the dependence of variation on gene expression)

Differential gene expression

- Requires raw un-normalized counts data as input and applies an internal statistical model to correct for library size across all samples
  - Estimates size factors - accounts for differences in sequencing depth across all samples
  - Estimates dispersion - accounts for variability between replicates
- Returns a log2 fold change, p value (calculated using Wald test), and adjusted p value for each gene for the given comparisons

Love, Huber, and Anders *et al.* 2014.

What you'll learn to do in this module

- Perform quality control checks with FastQC (Andrews)
- Perform FASTQ preprocessing with fastp (Chen *et al.* 2018)
- Quantify transcripts with Salmon (Patro *et al.* 2017)
- Import quantification estimates with tximeta and summarize to the gene level (Love *et al.* 2020; Soneson *et al.* 2015)
- Perform exploratory data analysis with DESeq2 (Love *et al.* 2014)
- Perform differential expression analysis with DESeq2
- Make fancy volcano plots and fancy heatmaps (Blighe *et al.*; Gu 2016)

Tool-specific tutorials

Getting Started with Salmon

Tximeta: transcript quantification import with automatic metadata

Note: if you are not using Salmon, you can't use tximeta (for now) so you will want to look at tximport: Importing transcript abundance datasets with tximport

Analyzing RNA-seq data with DESeq2

Links to follow-up information

StatQuest Video: A Gentle Guide to RNA-seq

StatQuest Video: RPKM, FPKM, and TPM

StatQuest Video: DESeq2, part 1, Library Normalization

Hansen et al. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acid Research*. 2010.

Michigan State University Research Technology Support Facility “FastQC Tutorial & FAQ”