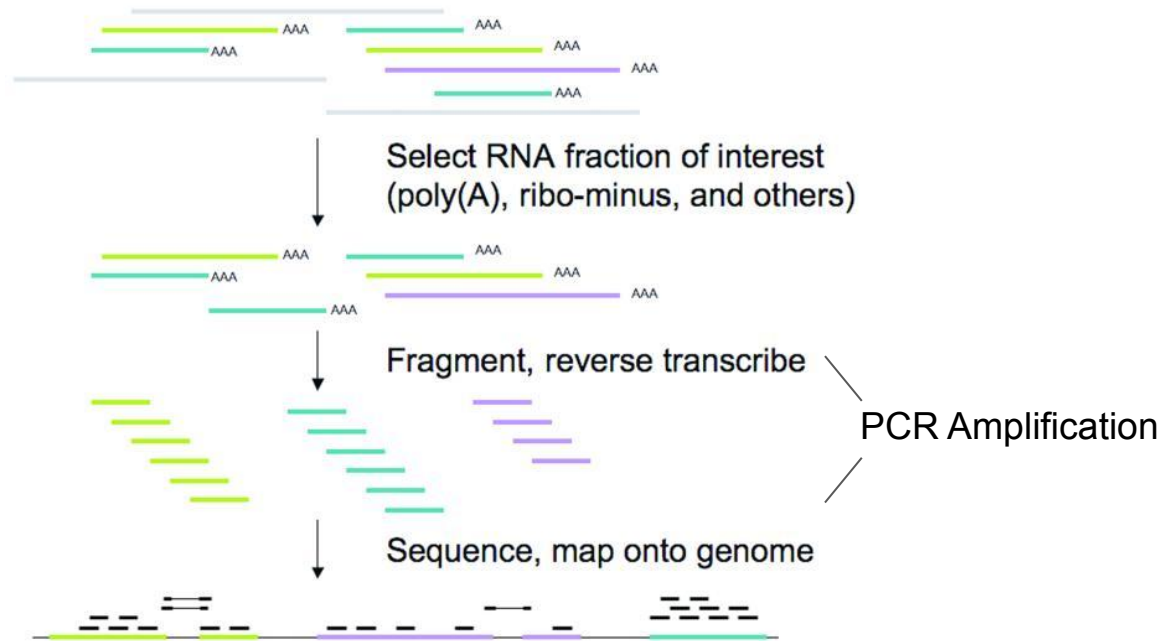# Introduction to RNA-seq

The CCDL

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.
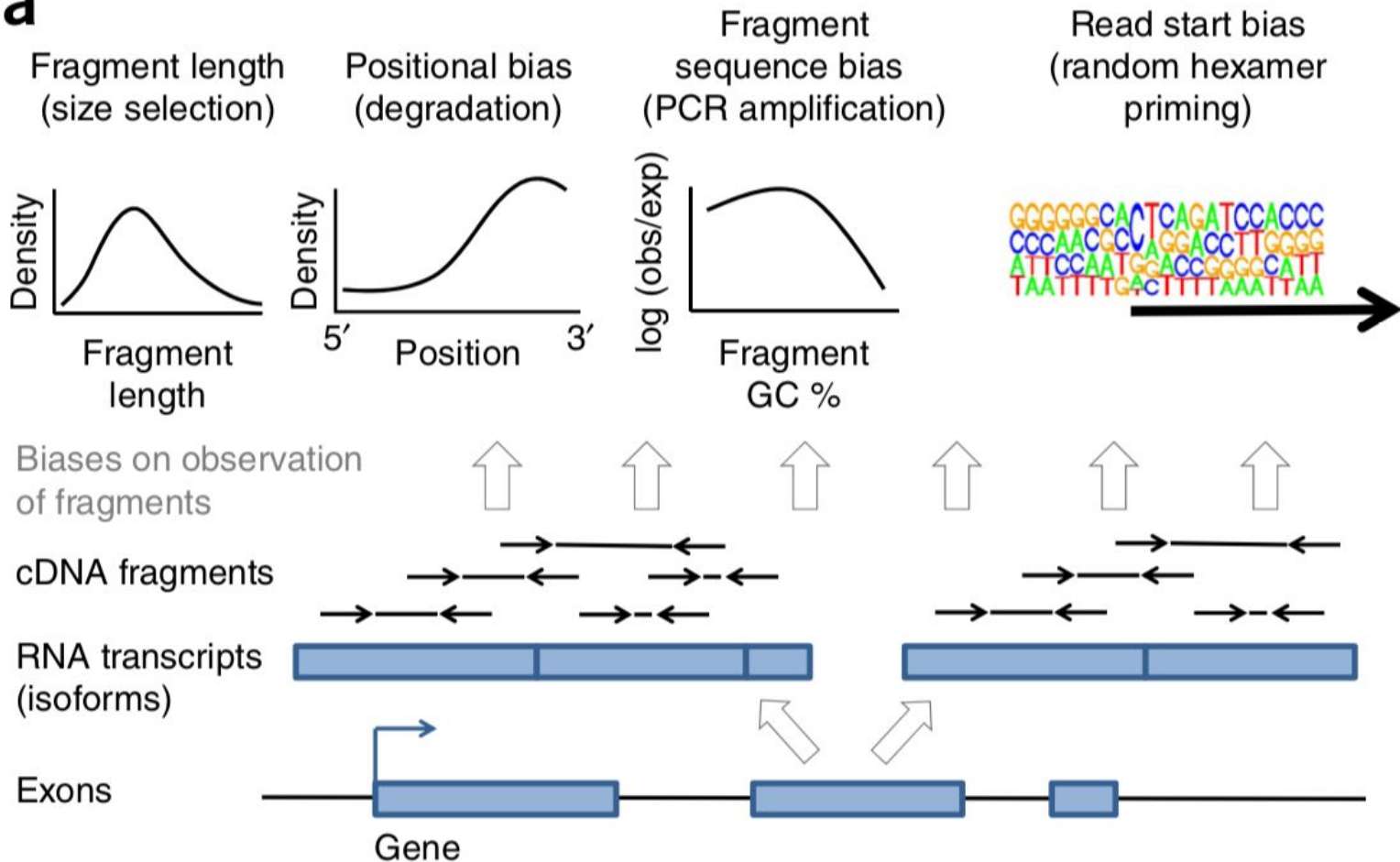
- Conesa et al. *Genome Biology. 2016.*

Select RNA fraction of interest (poly(A), ribo-minus, and others)

Fragment, reverse transcribe

PCR Amplification

Sequence, map onto genome
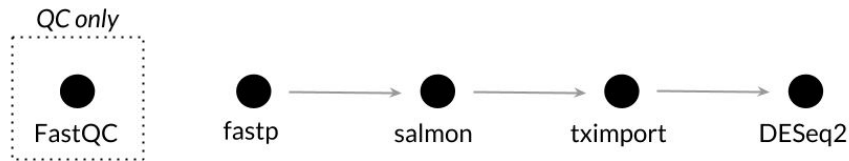
Pepke et al. *Nature Methods.* 2009.

Biases to be aware of

- Library size or sequencing depth - the total number of reads is not always equivalent between samples

- Gene length - longer genes are more likely to be observed

Abundance measures like TPM (Transcripts Per Million) take this into account

StatsQuest. RPKM, FPKM and TPM, clearly explained.

a

Fragment length (size selection) · Positional bias (degradation) · Fragment sequence bias (PCR amplification) · Read start bias (random hexamer priming)

Biases on observation of fragments

cDNA fragments

RNA transcripts (isoforms)

Exons

Gene

Overview of pipeline

TOOL	fastp ➡	Salmon ➡	tximport ➡	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, tx to gene identifier listing	tximport R object which contains unnormalized counts and length information

What you'll learn to do in this module

- Perform quality control checks with FastQC (Andrews.)

- Perform FASTQ preprocessing with fastp (Chen et al. *Bioinformatics. 2018.*)

- Quantify transcripts with Salmon (Patro et al. *Nature Methods. 2017.*)

- Summarize estimates to the gene-level with tximport (Soneson et al. *F1000Research. 2015.*)

- Perform exploratory data analysis with DESeq2 (Love et al. *Genome Biology. 2014.*)

- Perform differential expression analysis with DESeq2

Tool-specific tutorials

Getting Started with Salmon

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"