

Transcriptome-based Study Designs

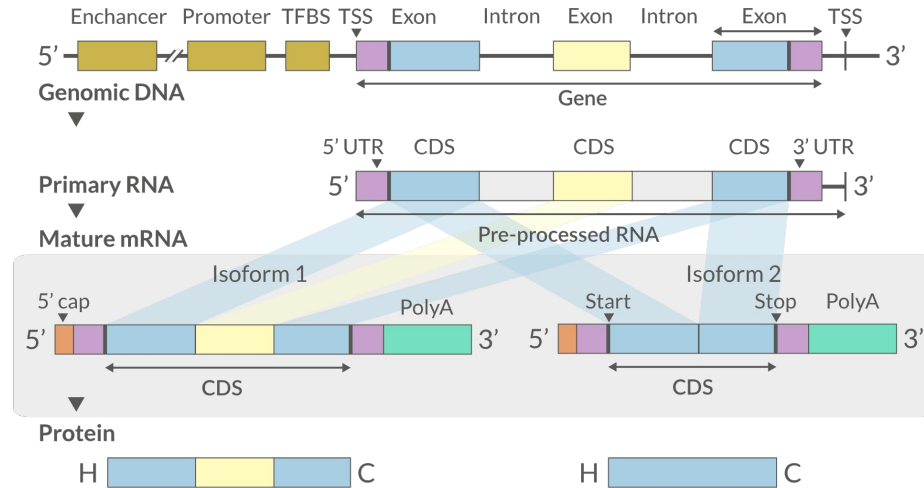
*Emerging Approaches For Tumor Analyses
in Epidemiological Studies*

Session overview

- Why RNA-seq
- Challenges
- Study design
 - Goal of study
 - Experimental strategy
- RNA-seq analysis pipeline
 - Alignment and Quality Control (QC)
 - Visualization
 - Quantification

Why RNA-seq

Why RNA-seq – RNA vs DNA-seq



The transcriptome
is dynamic.
represents only a small fraction of the genome (<5%).
is indicative of gene activity.

Why RNA-seq – RNA vs DNA-seq

- Evaluate the functional consequences of genomic changes
 - Difficult to infer from DNA sequences
 - ‘Regulatory’ mutations that do not directly affects the protein sequences
- Regulation at RNA level
 - Detection of alternative splicing, RNA editing, gene fusions
- Prioritize the genomic alterations that are more likely to be relevant
 - Mutations in expressed genes are more likely to be functionally relative
 - Allele-specific expression of the wild type or mutant allele.

Why RNA-seq – RNA vs DNA-seq

- Understand the interactions between genes/pathways
 - Gene-gene interaction could be indicative of the activity of transcription factors that orchestrate gene expression.
- Identify the consequence of experimental perturbation
 - E.g. measuring the changes in gene expression in response to drug treatment

Why RNA-seq – other related technologies/platforms

	RNA-seq	microarray	Nanostring nCounter (array-based)	Single-cell/nucleus RNA-seq	Spatial transcriptomics
Identify new RNA sequences?	Yes	No	No	Yes	No
Resolution	Non-single-cell Can be as low as 100 cells	Non-single-cell	Non-single-cell Less starting material than bulk RNA-seq	Single-cell	Ranges from single-cell level to small spots/regions of interest.
Sensitivity	High	Low	High	Low	Low
Dynamic range	High	Low	High	Low	Low
Sample types	Fresh, Fresh-frozen, FFPE	Fresh, Fresh-frozen, FFPE	Fresh, Fresh-frozen, FFPE	Fresh, Fresh-frozen (compromised performance)	Fresh, Fresh-frozen, FFPE
# transcripts/genes	Whole transcriptome	Thousands to 50,000	Up to 800	Whole transcriptome	Targeted or whole transcriptome
Output	Characterizes the sequences and measures the average gene expression of cell populations.	Measures the average gene expression of preselected genes of cell populations.	Measures the average gene expression of preselected genes of cell populations.	Measures the gene expression of individual cells.	Quantify the gene expression within the spatial context of tissues. Usually integrated with scRNA-seq analysis.

Challenges

Challenges

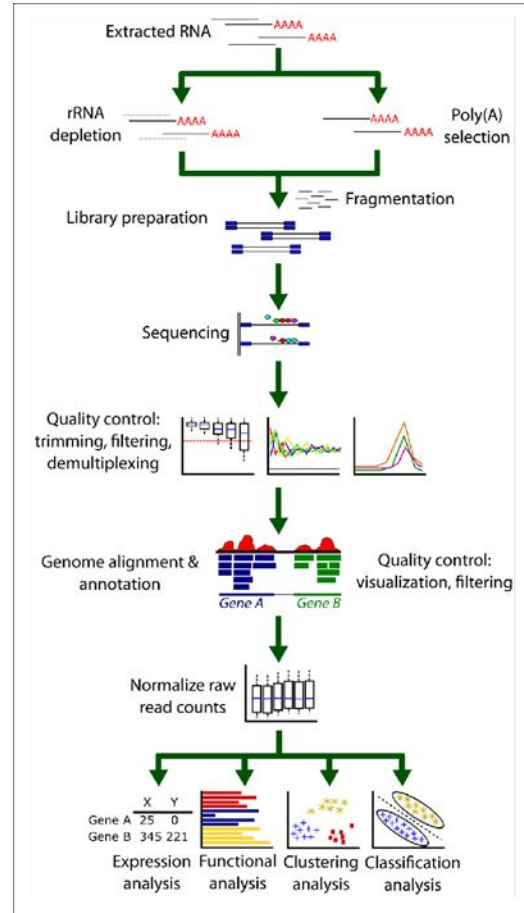
- Compared to DNA, RNAs
 - Are more fragile
 - Have tissue-specificity
 - Vary in orders of magnitudes in quantity
- Sample – quality and quantity of samples (purity, clinical sample)
- Types of RNA (rRNA, miRNA) should be removed or processed separately
- Technical bias in mapping and quantification due to degradation.
- Specific gene structure can be challenging (e.g. large intron)

Study design

Study design – Goal of study

- Characterization of the gene expression pattern
- Differential gene expression across samples
- Identification of novel transcripts
- Alternative splicing analysis
- Detection of gene fusion events
 - Explore the functional consequences of structural variants
- Discovery or validation of mutation
 - Allele-specific expression analysis
- Identification of RNA editing

Study design – Generic pipeline of RNA-seq studies



Study design – Experimental strategy

- General best practices for RNA-seq experiments
- Goal-driven experimental variables
- RNA Quality considerations

General best practices for RNA-Seq

- Use of replicates
- Importance of batching, randomization during processing
- RNA-sequencing best practices (short read)

Replicates

Experimental replicates can be performed as **technical replicates** or **biological replicates**.

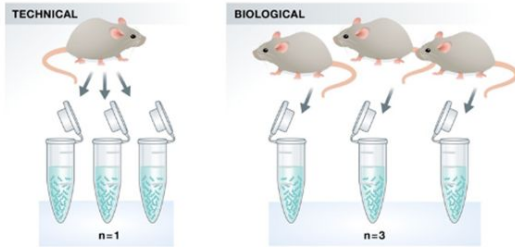


Image credit: Klaus B., *EMBO J* (2015) **34**: 2727-2730

- **Technical replicates:** use the same biological sample to repeat the technical or experimental steps in order to accurately measure technical variation and remove it during analysis.
- **Biological replicates** use different biological samples of the same condition to measure the biological variation between samples.

- What is the difference between technical, biological replicates?
- For RNA-seq, technical replicates are not generally necessary
- Biological replicates are essential; 6-12 biological replicates generally recommended

https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/experimental_planning_considerations.html#:~:text=Technical%20replicates%3A%20use%20the%20same,the%20biological%20variation%20between%20samples.

Batch effects

- More likely to impact RNA-seq than DNA sequencing
- When, where, method, reagents
- Were all study samples:
 - Extracted on the same day?
 - Libraries prepared on the same day?
 - Was the same method used for extraction, library prep, sequencing, at the same facility?
 - Was all sequencing performed on the same run?
 - If any of the answers are “no”, then you have batches....
- Batch effects can sometimes be addressed during analysis, but better to address during experimental design

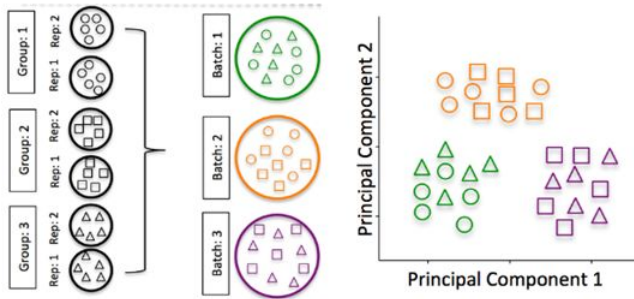
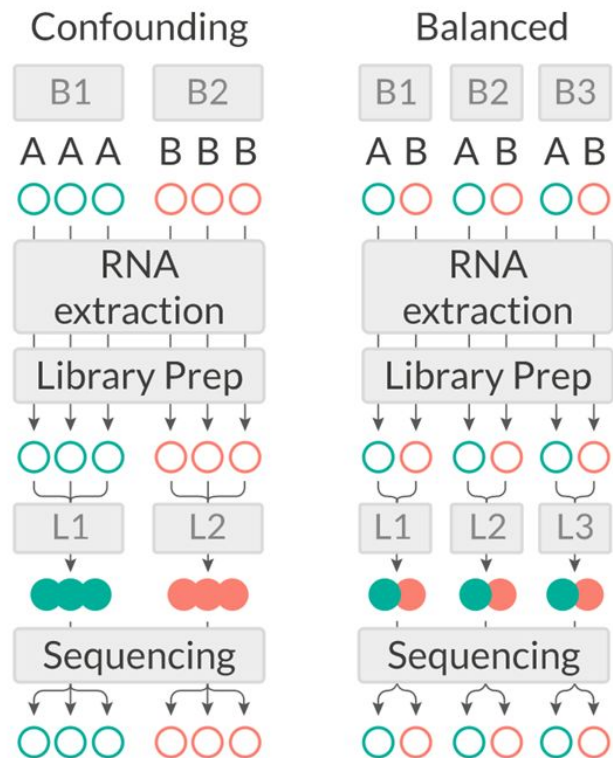


Image credit: Hicks SC, et al., bioRxiv (2015)

Are the differences in expression observed in the data due to biological differences in your treatment groups? Or due to processing variables between batches of samples?

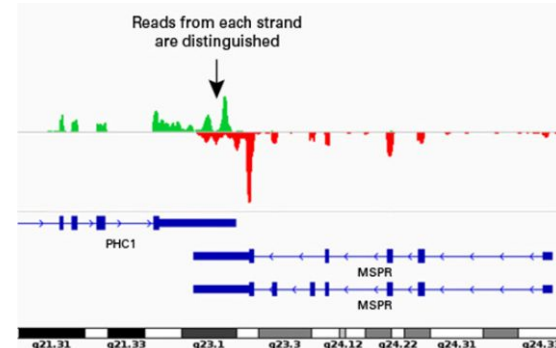
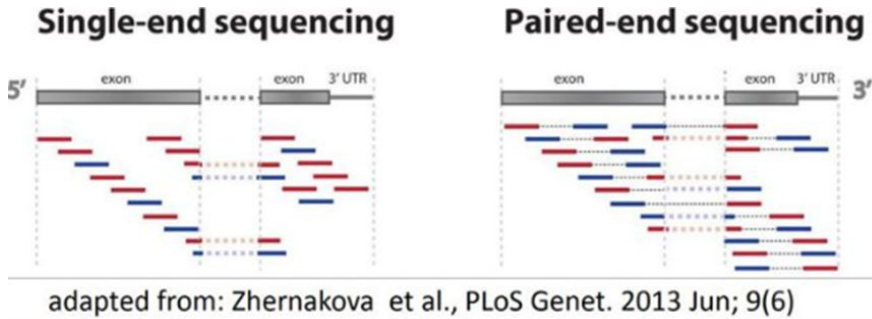
Best practices for batches



- Avoid confounding your experiment by batch
- Reduce processing variables between treatment groups (A, B at left)
- When batches are unavoidable, split biological replicates across batches
- Track batch metadata for downstream analysis

RNA-seq best practices (short read)

- Paired End (PE) sequencing generally preferable to Single End (SE) sequencing
 - Improved mappability
 - Read pairs more likely to span exons
- Stranded RNA-seq preferable to non-stranded
 - Better resolution for reads from overlapping genes transcribed on opposite strands
 - Better resolution for bi-directional transcription
- Both strategies improve the accuracy of gene expression quantification



<http://www.takara.co.kr/file/manual/634836.html>

Goal-driven experimental design variables

- Sequencing depth
- RNA Library prep methods
- Sequencing platform

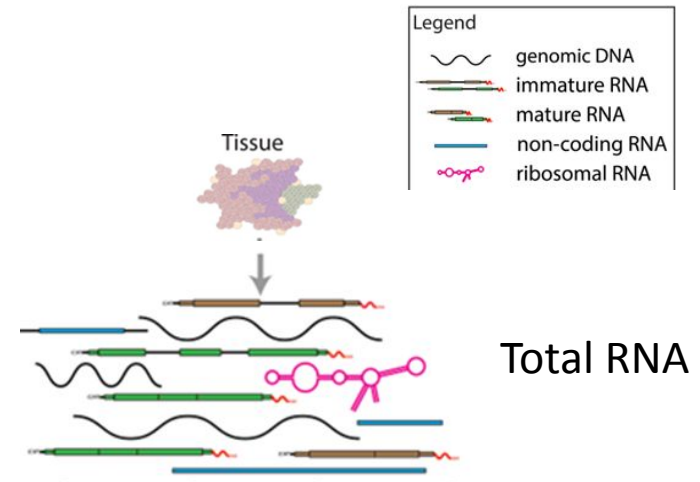
Sequencing depth

- Highly dependent on goals
- Why it's difficult to answer the question “How much does RNA-seq cost”?

Analysis Goal	Reads Recommended
Differential expression of highly expressed genes	5-25 million
Differential expression across transcriptome, isoform level	30-60 million
Alternative splicing	80-100 million
Assembly of new transcripts	100-200 million

RNA Library prep methods

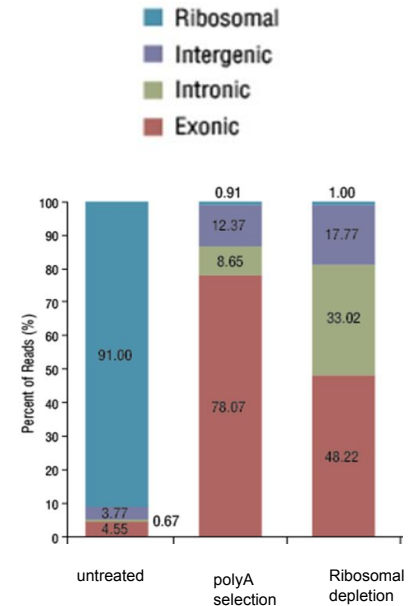
- Total RNA is 80-95% rRNA (ribosomal RNA)
- In order to focus on the transcriptome, we have to reduce the rRNA in total RNA during Library prep
- Two methods:
 - Poly A selection
 - Ribosomal depletion
 - Selection depends on goals, RNA quality



<https://journals.plos.org/ploscompbiol/article/figure?id=10.1371/journal.pcbi.1004393.g004>

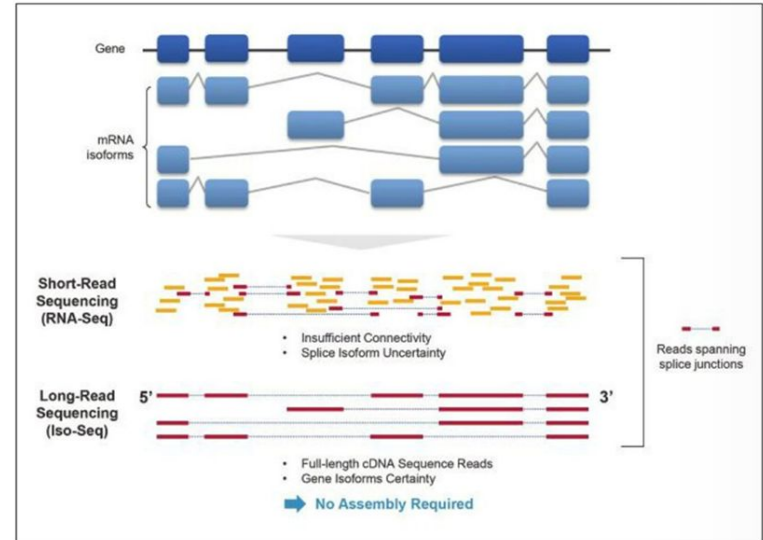
RNA Library prep methods

- polyA (mRNA) selection
 - Selects for only mRNAs with a polyA tail
 - Enriched for protein coding only reads
 - Very little pre-mRNA, lincRNA, etc.
 - Doesn't work well with low quality, degraded RNA (bias)
 - **Bottom line**: more efficient if all you're interested in is coding RNA, and your RNA is very high quality
- Ribosomal depletion
 - Removes rRNA with probes
 - Leaves everything else to be sequenced
 - Diverse RNA sequences
 - Not as enriched for exonic reads
 - Effective for degraded RNA
 - **Bottom line**: good for large studies where RNA quality might be variable; or non-coding RNAs are of interest



Sequencing platform

- Short-read vs Long-read
 - Short-read is by far the most common approach for RNA-seq, with broad analytical utility
 - RNA-seq can also be performed using long-read technologies (PacBio, Oxford Nanopore)
 - Long-read RNA-seq often called “Iso-Seq”
 - Not geared towards Differential Expression
 - Entire transcripts spanned by single reads
 - Isoform Discovery, resolution of complex splicing/fusion events
 - Disambiguates Isoforms that short reads can't resolve
 - Requires VERY high quality RNA



RNA Quality considerations

- RNA MUCH more susceptible to degradation than DNA
 - During processing, handling
 - Depending on storage temperature
 - Freeze/Thaw cycles
- How intact or degraded your RNA is can impact feasibility of
 - Experiments that can be performed
 - Reliability of analysis

RNA Quality considerations

- Quality of RNA should be assessed, reported after extraction, before library prep
- Often reported as the “RIN”, or RNA Integrity Number
- Quality score from 1-10
- Virtually all FFPE RNA will be “Low” quality

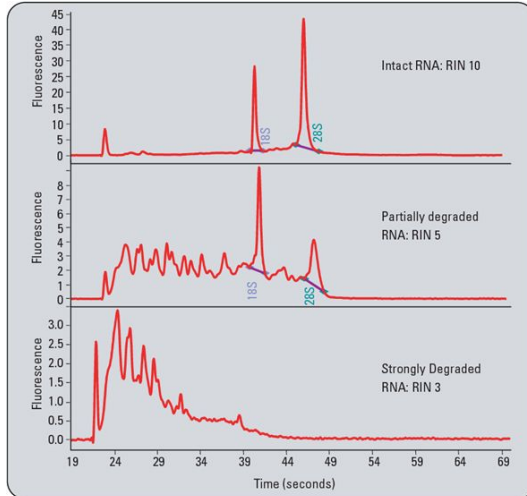


Figure 6

RNA Quality	RIN Score	Limitations
High	8-10	
Medium	4-7	No long-read, no polyA selection
Low	1-3	No long-read, no polyA selection, RNA-seq may be piloted to assess suitability

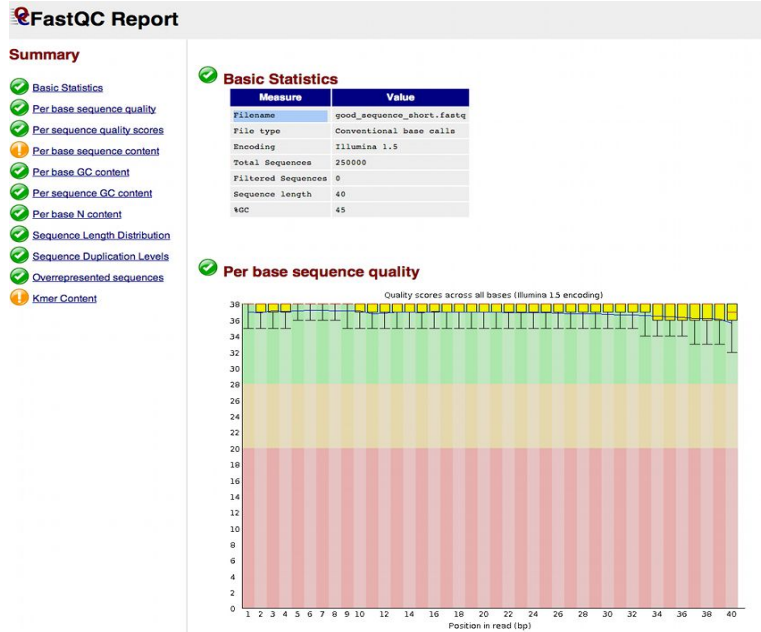
<https://blog.genohub.com/2017/12/24/rin-numbers-how-theyre-calculated-what-they-mean-and-why-theyre-important/>

RNA-seq analysis pipeline

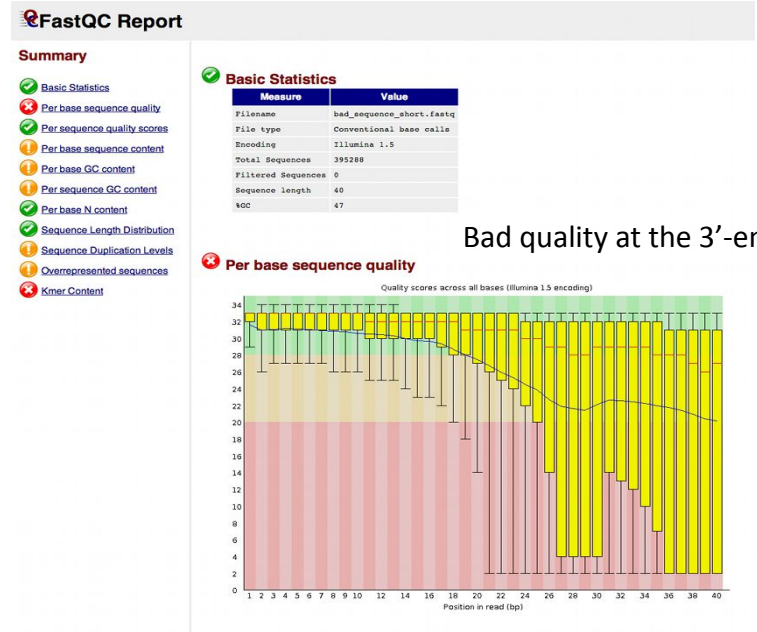
Quality Control (QC) of Raw Reads

■ FastQC

■ Good example



■ Bad example

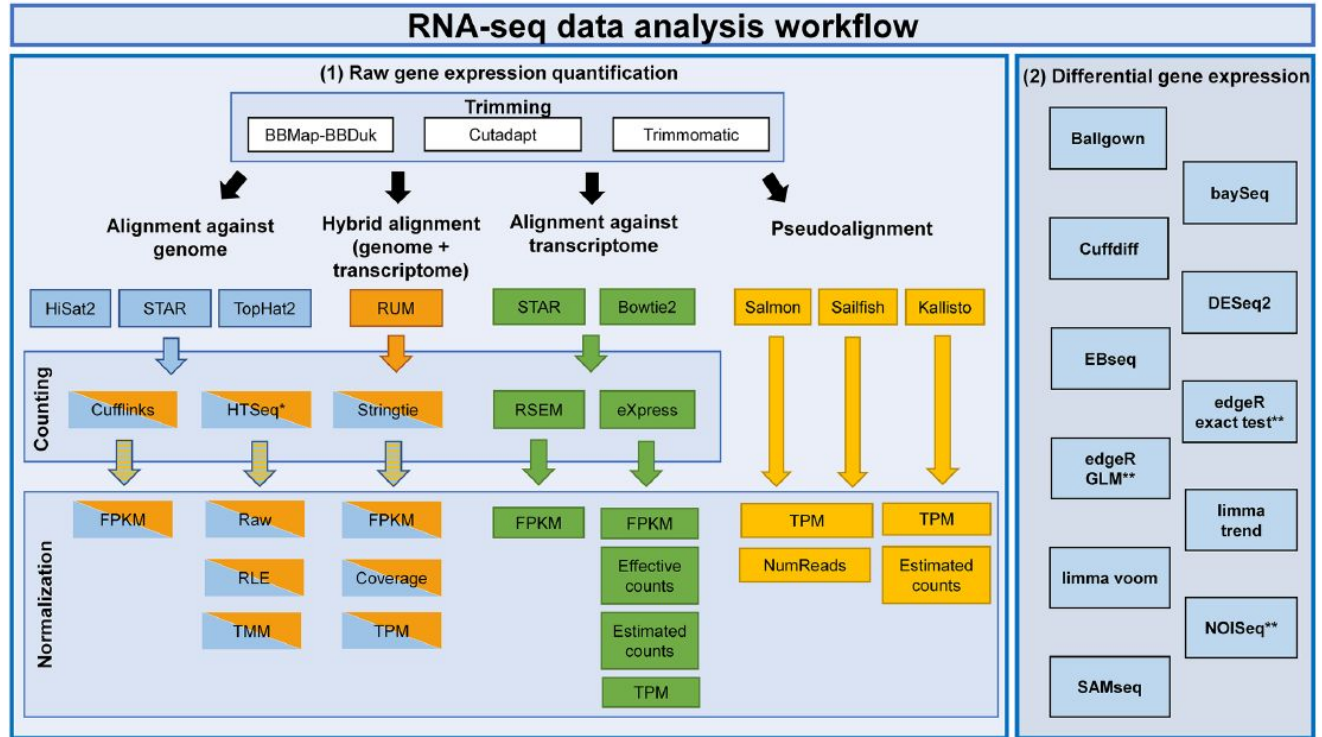


Bad quality at the 3'-end

RNA-seq analysis pipeline

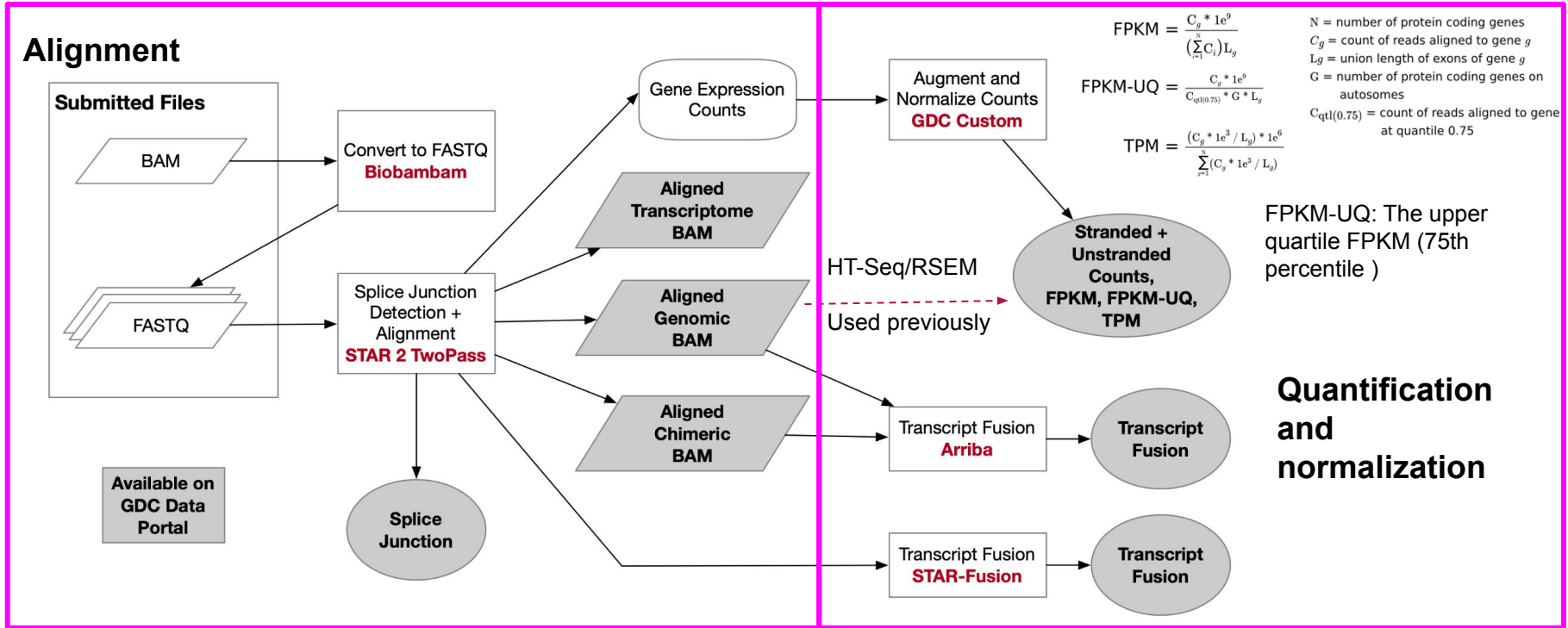
- No standard pipeline, many tool combinations, reference based

1. Alignment
2. Quantification and normalization
3. Downstream analysis



RNA-seq analysis pipeline

TCGA RNA-seq pipeline as an example



ICGC has a very similar pipeline.

Alignment

Splice-aware aligner

RPKM, FPKM, TPM etc

QC After Alignment

- %mapped reads/uniquely mapped reads (aka, sequencing depth)
- Gene body coverage (e.g. 5'-to-3' bias)
- Quality distribution
- Average insert size
- PCR duplicates
- Distribution across exon/intron/intergenic regions
- Strand-specificity
- Sequencing saturation

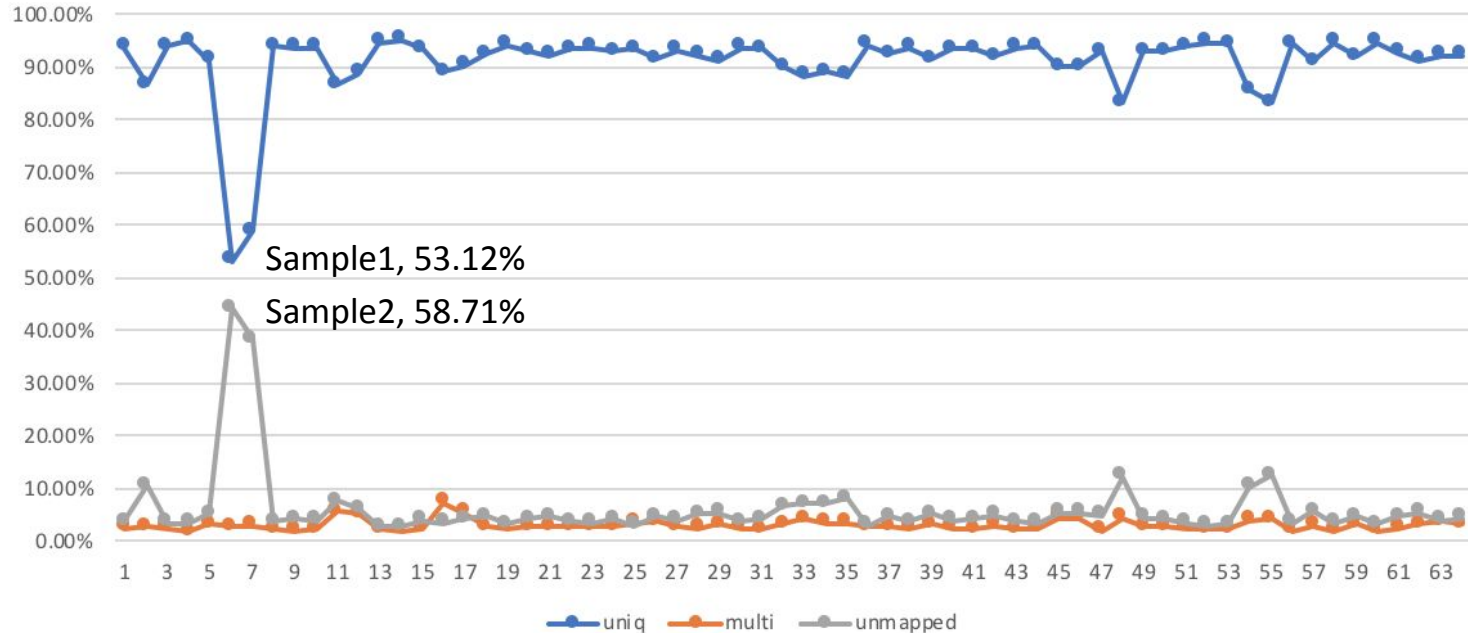
RSeQC and
RNA-SeQC2

Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics* (Oxford, England), 28(16), 2184–2185.

A. Graubert*, F. Aguet*, A. Ravi, K.G. Ardlie, Gad Getz, "RNA-SeQC 2: efficient RNA-seq quality control and quantification for large cohorts," *Bioinformatics*, 2021.

QC After Alignment

Mapping rate to the Human Genome (hg19): An example

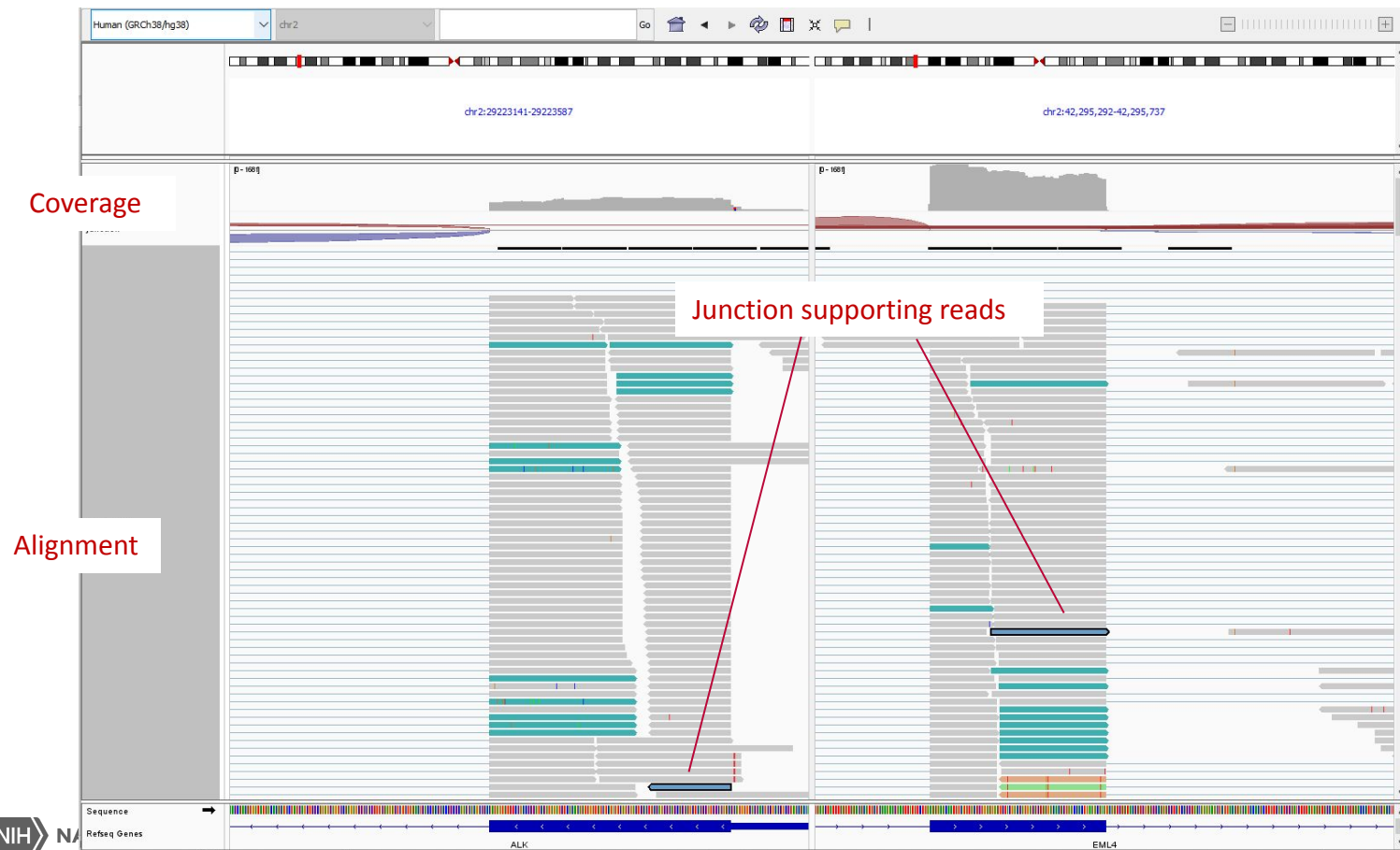


Sample1 and sample2 need to be removed for further analysis.

Visualization

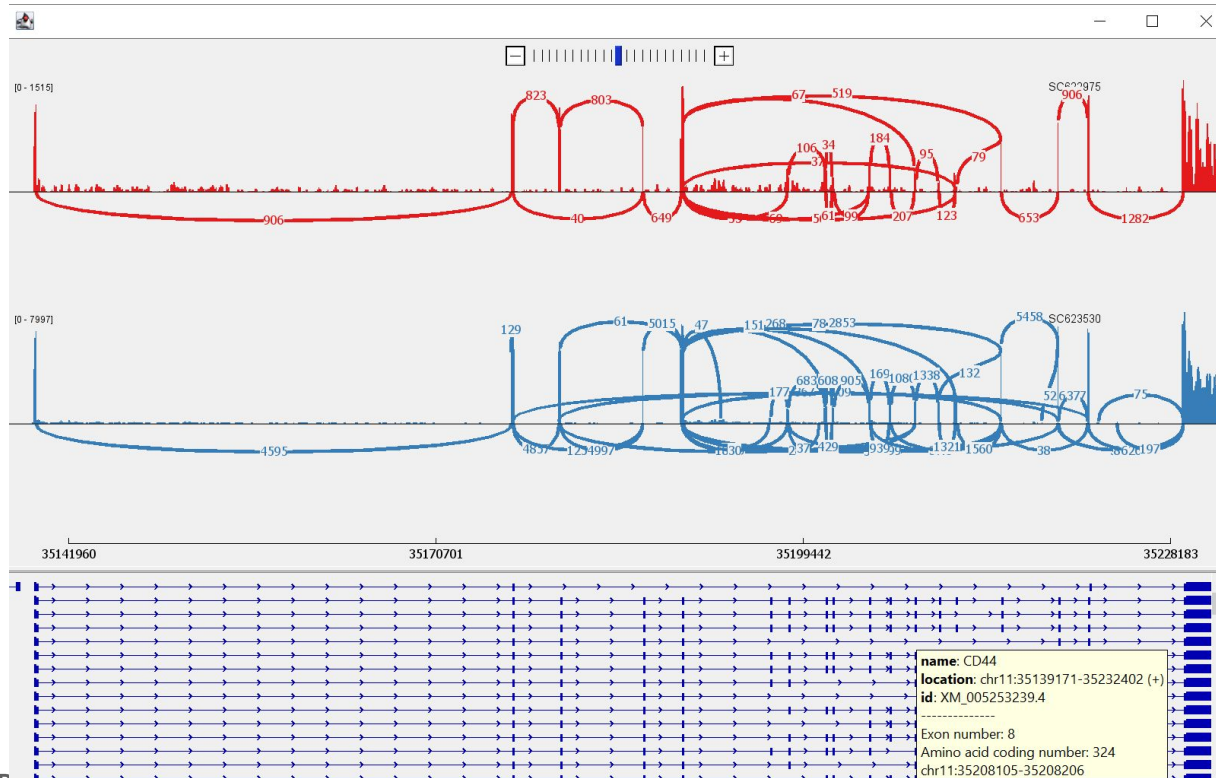
- Plot types
 - Pileup plot
 - Sashimi plot
- Tools
 - Best overall: [Integrative Genomics Viewer/IGV](#)
 - Alternative: [samtools tview](#) (for quick check), [Tablet](#) (support more format, high-quality graphs), [Integrated Genome Browser/IGB](#)
 - More information:
 - <https://github.com/cmdcolin/awesome-genome-visualization>
 - <http://jermdemo.blogspot.com/2010/08/ngs-viewers-reviewed.html>

Visualization – IGV



Visualization – IGV

- Sashimi plots visualize junctions from alignments.



[illegible]

```
samtools tview -p chr:pos -d H alignment.bam genome.fasta
```

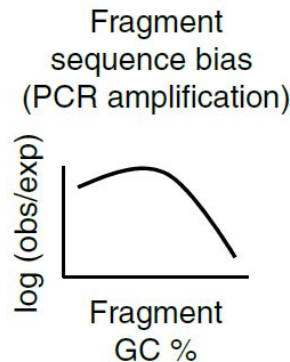
[illegible]

Quantification

- Counting mapped reads as a measure of expression
- Reads can be summarized and aggregated over any biologically meaningful features (e.g. genes, transcripts, exons, etc.)
- Intersection on gene models

Quantification – artifacts and biases

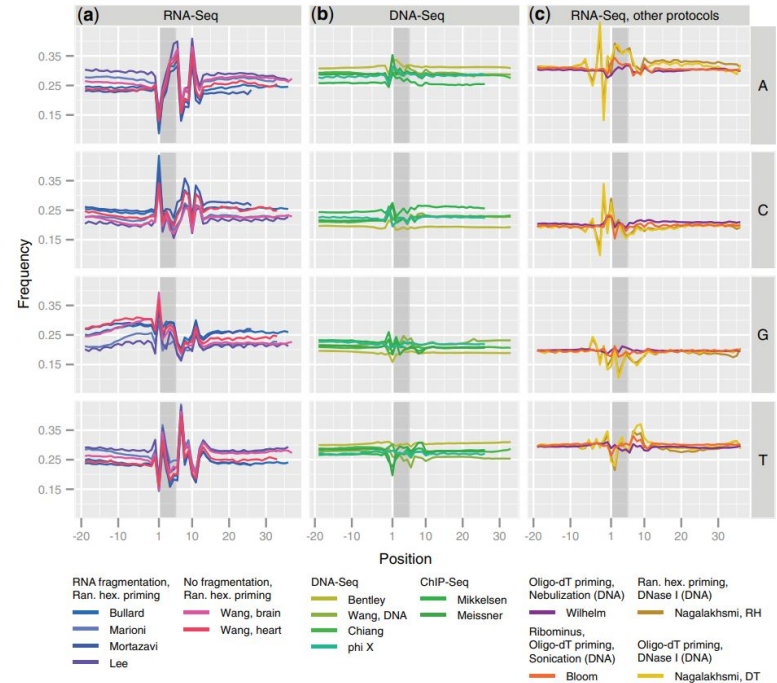
- GC bias
 - Samples processed in the same batch with the same protocol usually would show similar GC bias. Then the GC bias correction can be skipped for comparisons across samples.
 - Can be examined computationally using tools like FastQC.
 - If significant difference in GC content exist, can be corrected using tools like [EDASeq](#) or [alpine](#).



Quantification – artifacts and biases

- Random hexamer bias
 - a bias in the nucleotide composition at the start of sequencing reads
 - no significant pattern in DNA and Chip-seq
 - some tools (e.g. Cufflinks) would correct the bias. Usually doesn't need correction separately.

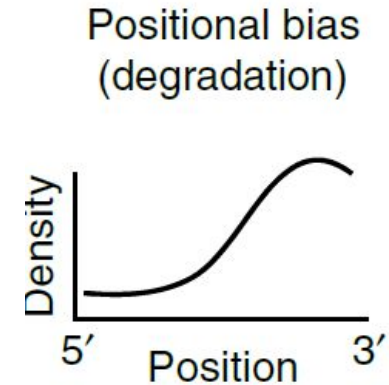
Read start bias
(random hexamer
priming)



Love, M., et. al. *Nat Biotechnol* **34**, 1287–1291 (2016).
Hansen, K, et. al. *Nucleic Acids Res.* **38**(12): e131 (2010)

Quantification – artifacts and biases

- Positional bias: 5'-to-3' bias
 - Mostly due to the poly-A selection and partial degradation of RNA
 - Correlated with RIN score.
 - If the bias is similar across samples, usually doesn't need to make adjustment.
 - Some mathematical models are developed to neutralize the effect of positional bias.
 - Alternatively, RIN scores can be used as a covariate.

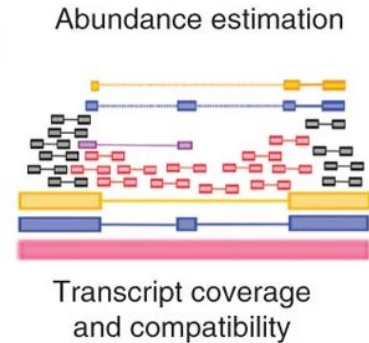
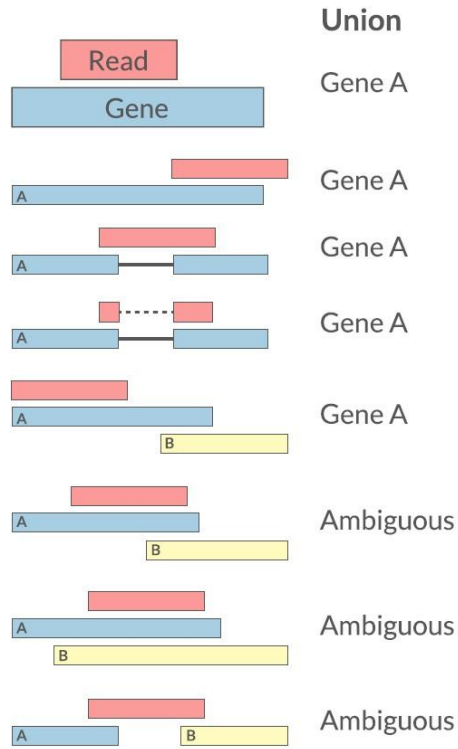


Quantification – artifacts and biases

- PCR duplicates
 - Most RNA-seq pipelines do NOT include computational deduplication.
 - De-duplication computationally is carried out by sequence of comparison or aligned coordinates.
 - Short transcripts and very highly expressed transcripts (common in some species) will contribute the last majority of biological ‘duplicates’.
 - The fraction of identified duplicates is correlated with the number of aligned reads.
 - Use UMIs in case of very low input sample or very deep sequencing library.
 - More information: <https://www.biostars.org/p/55648/>
- Mapping errors

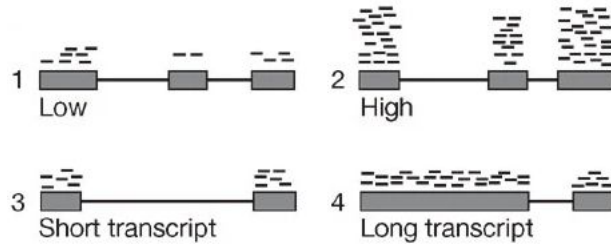
Quantification – artifacts and biases

- Ambiguous alignment
 - Overlapping genes
 - Alternate splicing



Quantification – artifacts and biases

- Sequencing depth
- Gene length bias



Quantification – Common software

Methods	Quantification features	Output	Ambiguous reads	Confidence level	Notes
HTSeq	Gene	Counts	Ignore	NA	
STAR geneCounts	Gene	Counts	Ignore	NA	
Cufflinks	Transcript	RPKM	Split equally, rescue	NA	Assembly
featureCounts	Gene	Counts	Ignore, count all, split equally	NA	
RSEM	Gene, transcript, Exon	Counts, RPKM, TPM	Expectation-Maximization (EM)	95% confidence intervals	
StringTie	Transcript	FPKM, TPM	Flow network	NA	Assembly
eXpress	Gene, transcript	Counts, FPKM, TPM	EM	95% confidence intervals	
Kallisto	Transcript	TPM	EM	Bootstrap	Pseudoalignment*
Salmon	Transcript	Counts, TPM	EM	Bootstrap	Pseudoalignment*

*Pseudoalignment: Reads are mapped to a reference transcriptome, and are judged on compatibility with transcripts, not aligned.

Resources

- NCI BTEP Bioinformatics for beginners:
<https://btep.ccr.cancer.gov/docs/b4b/>
- UC Davis Bioinformatics workshop RNA-seq analysis:
<https://ucdavis-bioinformatics-training.github.io/2021-September-RNA-Seq-Analysis/>



**NATIONAL
CANCER
INSTITUTE**

www.cancer.gov

www.cancer.gov/espanol