QC of RNA-sequencing data

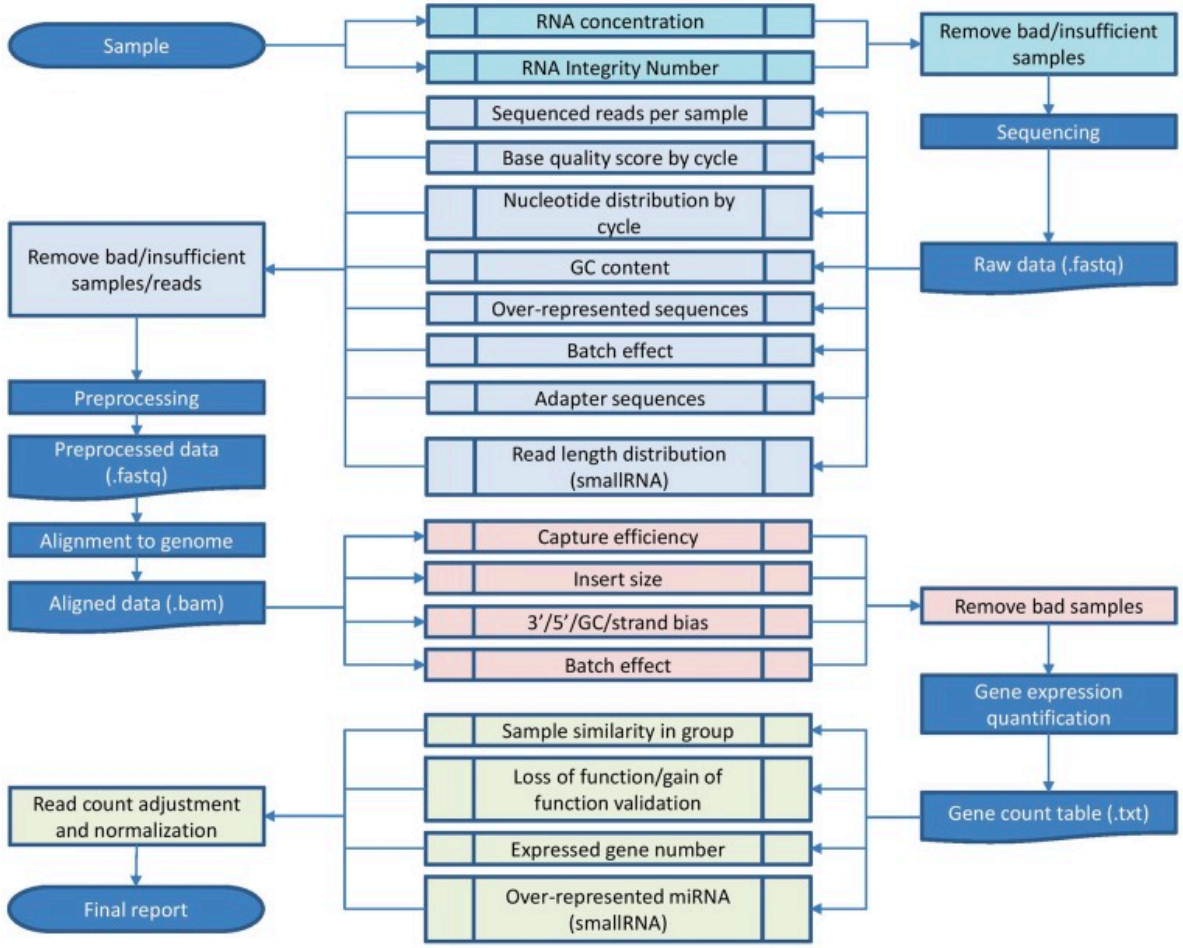
Clinical Cancer Genomics
April 2023

Learning objectives and lecture agenda

- Learn about the quality control (QC) steps required for RNA-seq data and why these steps are necessary.
- Learn about bioinformatics tools used in RNA-seq QC.

Stages of RNA-seq QC

1. Checking RNA quality in samples
2. Checking quality of raw read data in FASTQ files
3. Checking RNA-seq alignment quality
4. Checking quality of gene expression estimates



Sheng et al., Briefings in Functional Genomics, 2017

Tools for RNA-seq QC

- Many tools exist. Some are able to perform QC over several of the stages mentioned on the previous slide, while others focus on a single stage.
- Some tools work for both DNA and RNA data, others are specific to RNA.
- Some example tools:
  - FastQC
  - Qualimap
  - RNA-SeQC
  - RSeQC
  - RNA-QC-chain

1. Checking RNA quality in samples

- Is the RNA in the sample intact (and good for downstream analysis) or degraded?
- RNA can be digested by RNase enzymes also present in the sample
- Generally, RIN > 7 is good
- Formalin-fixed samples will have lower RIN values (range ~2-5)

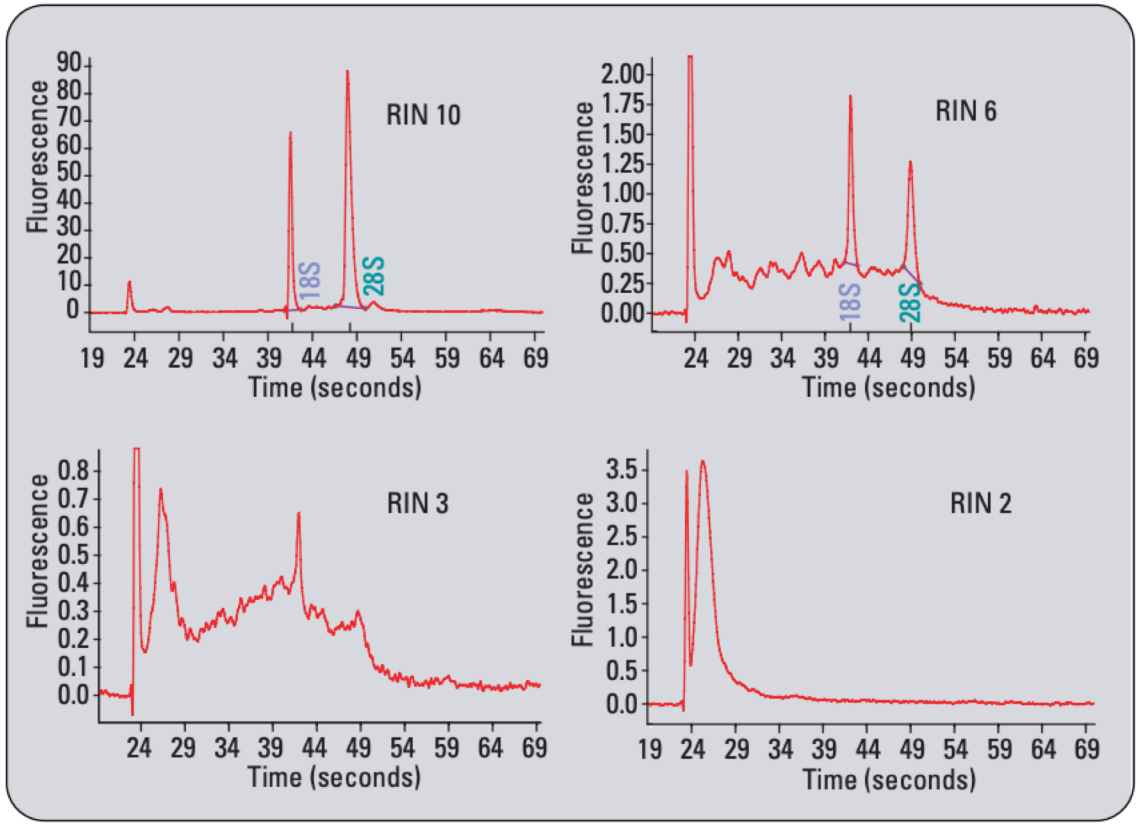
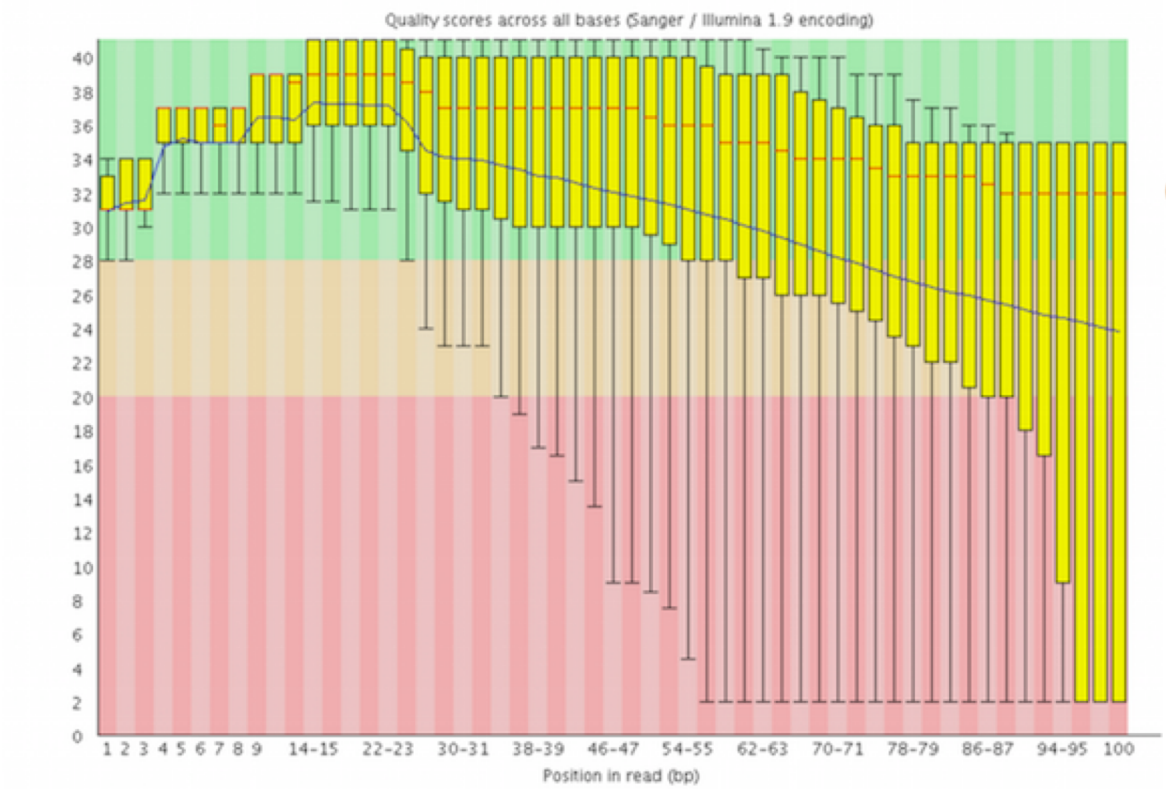


Figure 2
Sample electropherograms used to train the RNA Integrity Number (RIN) software. Samples range from intact (RIN 10), to degraded (RIN 2).

Source: Agilent

2. Checking quality of raw read data in FASTQ files

Per base sequence quality



Are these low quality bases in the sequencing reads?

Metrics from FASTQC tool

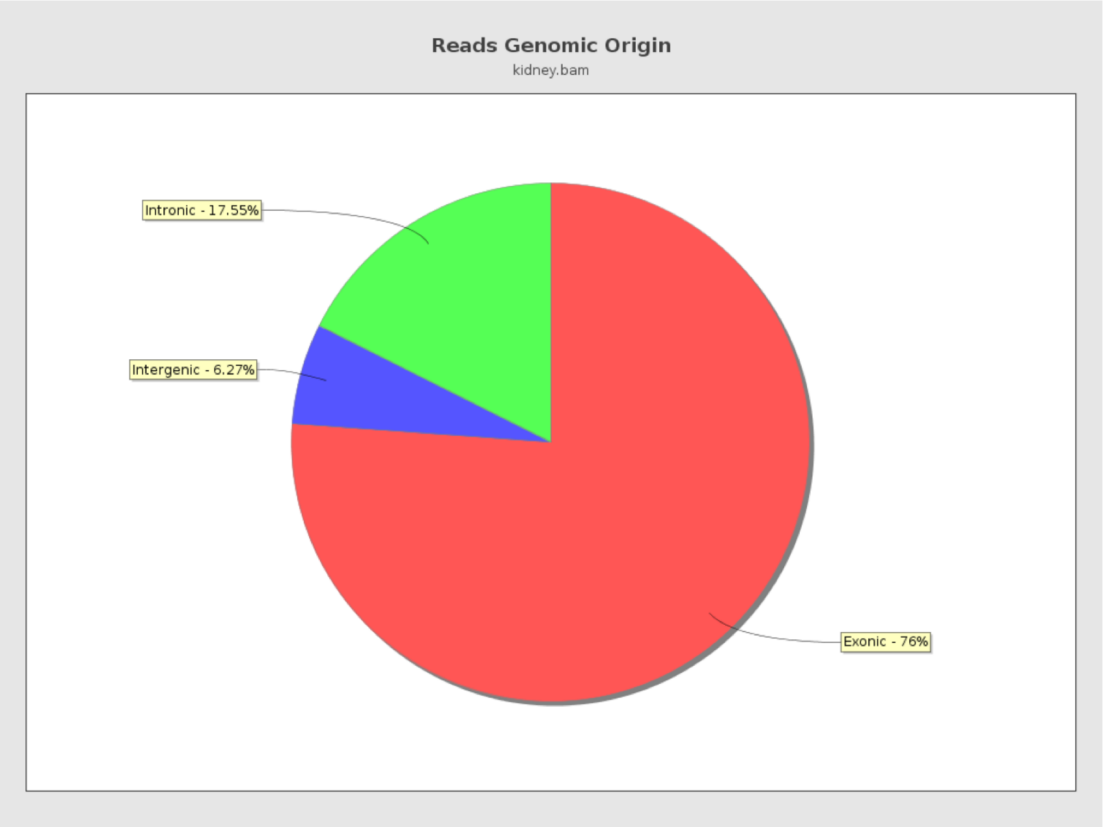
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGCTATGGCCACCAGACTCTCAGGCTCCATGCAGTGGCCAGCCTCATCG	2554	0.8349133703824779	No Hit
CAGCGGTCTAGTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAG	2463	0.8051650866296176	No Hit
GTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATC	1920	0.6276560967636483	No Hit
CCACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATG	1219	0.39849624060150374	No Hit
GAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATCTTCA	1186	0.3877084014383786	No Hit
GGCAGGTGGACCCGGAGCCGCTGACAGAGGAGGTCAGCCCCTGAGTTGGA	1111	0.3631905851585486	No Hit
CACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATGT	1079	0.35272965021248776	No Hit
GTCCCTGCTGGGGGCCAGGAGACGGTAGATGGAGCTGGGCAGGTGGACCC	1036	0.3386727688787185	No Hit

Identification of contamination, including adapter sequences in the RNA-seq data.

3. Checking RNA-seq alignment quality

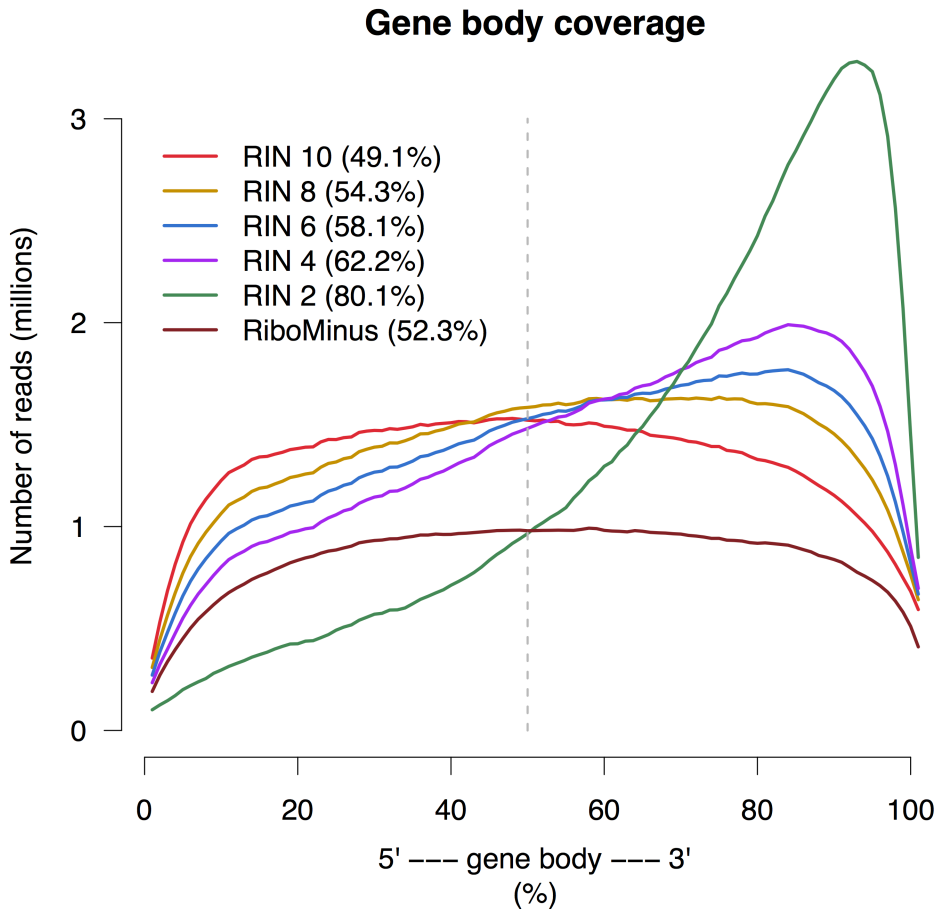
- **Capture efficiency:** percentage of total sequenced reads mapped to the intended target region (e.g. coding regions).



From Qualimap tool, expect > 60% exonic reads

Checking RNA-seq alignment quality

- **5’-3’ bias:** nonuniform coverage of transcripts in the sample
- If some bias is observed, important that it is similar across samples/groups.
- Can be reflective of sample quality, but bias can also be introduced by library preparation methods.



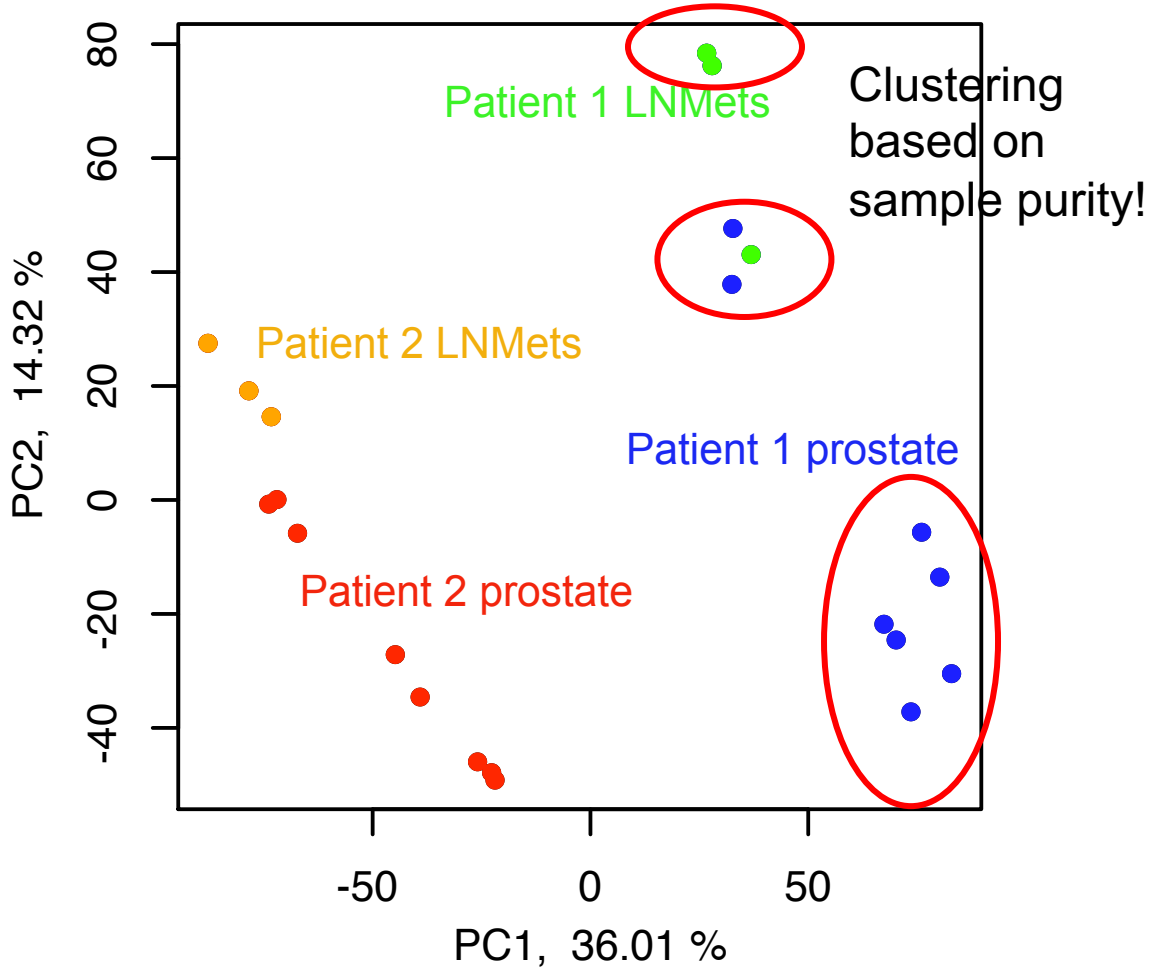
RIN = RNA integrity number

Checking RNA-seq alignment quality

- **Ribosomal RNA fraction**
- Ribosomal RNA (rRNA) is the most abundant RNA species (~80%), so it can prevent the sequencing of other RNA species if not properly eliminated from the sequencing library.

4. Checking quality of gene expression estimates

- Clustering can be used to identify which samples are closely related based on gene expression, but can also identify **outlier** samples or **batch effects** (e.g. sample purity, see Aran et al., Nature Communications, 2015).
- Samples might have to be excluded or re-sequenced.
- Can identify variables that have to be accounted for in downstream analyses.



Karolinska
Institutet