Lab 2b: Interpreting an RNASeq MultiQC Report

Tools

- fastQC
- DupRadar
- Picard
- Qualimap
- RSeqC
- samtools

Tools: FastQC

- Quality control of raw sequencing reads
- Generates quick metrics (e.g., per-base sequence quality, GC content, adapter content, overrepresented sequences)
- Visual reports help identify potential issues (e.g., poor-quality cycles, contamination)

When do you use this?

- After you receive the raw FASTQ files from a sequencing run, you use FastQC to assess read quality before further analysis.

Tools: FastQC

Are my raw reads of acceptable quality?
Per-base quality scores and distribution help you see if your reads are generally high-quality or if there are low-quality bases at read ends.

Do my reads have adapter contamination or other overrepresented sequences?
Overrepresented sequences or known adapter contaminants are flagged so you can decide if trimming is needed.

Is the GC content what I expect for my organism or library type?
GC bias plots can highlight deviations that might indicate contamination or PCR bias.

Do I see any odd patterns or biases in the sequence?
FastQC highlights unusual k-mer overrepresentations or sequence duplication levels. You may cross-reference duplication issues in FastQC results with the results from DupRadar and Picard-Mark Duplicates to understand the mechanisms of the pattern(s).

Tools: DupRadar

- Duplicate Reads Analysis in RNA-Seq
- Estimates the dependency between expression levels and duplication rates
- Helps determine how many duplication events are technical artifacts versus biologically relevant

When do you use this?

When you want to check if high duplication in your RNA-seq data might be affecting downstream analysis and interpret the nature of those duplicates.

Tools: DupRadar

How does duplication vary across different expression levels?

Sometimes highly expressed genes show higher duplication because of biological abundance; DupRadar helps tease out which duplicates may be “true” (biological) vs. technical artifacts.

Are the duplicate reads I'm seeing due to high-expression genes or technical artifacts?

DupRadar correlates duplication rates with gene expression levels (in RNA-seq data).

Tools: Picard

- A toolkit that works primarily with SAM/BAM/CRAM files
- Common utlity processes include:
    - MarkDuplicates (marks or removes PCR duplicates)
    - CollectInsertSizeMetrics (collects insert size statistics)

When do you use this?

When you want to remove or mark duplicate reads and generate alignment metrics for quality checks.

Tools: Picard

How many reads are duplicates, and should I mark or remove them?

MarkDuplicates identifies PCR or optical duplicates in SAM/BAM files.

What is the typical insert size for your sequenced reads?

If the distribution of fragment lengths is not similar to the expected length of reads, it could be due to library prep or library complexity issues.

Tool: Qualimap

- Quality control and analysis of genomic alignments
- Generates coverage statistics, alignment quality metrics, and graphical summaries for both DNA-Seq and RNA-Seq data
- RNA-Seq-specific module can analyze strand specificity, gene coverage profiles,

Tools: Qualimap

How uniform is my coverage across the genome or transcripts?

Coverage histograms and metrics show if coverage is even or if there are spikes or drop-offs.

- 3' or 5' coverage skew can be caused by fragmentation biases.
- Coverage spikes can be caused by library prep issues.

Tools: RSeQC (RNA-Seq Quality Control)

- RNA-Seq quality assessment to evaluate:
  - Read distribution (e.g., exonic, intronic, intergenic)
  - Junction annotation
  - Gene body coverage
  - GC content and bias
  - Inner distance: insertion profiles for paired-end data
  - Read Duplication

When do you use this?

Used to identify potential issues in RNAseq experiments, such as biases in coverage along transcripts, strand specificity problems, or anomalies in read distribution.