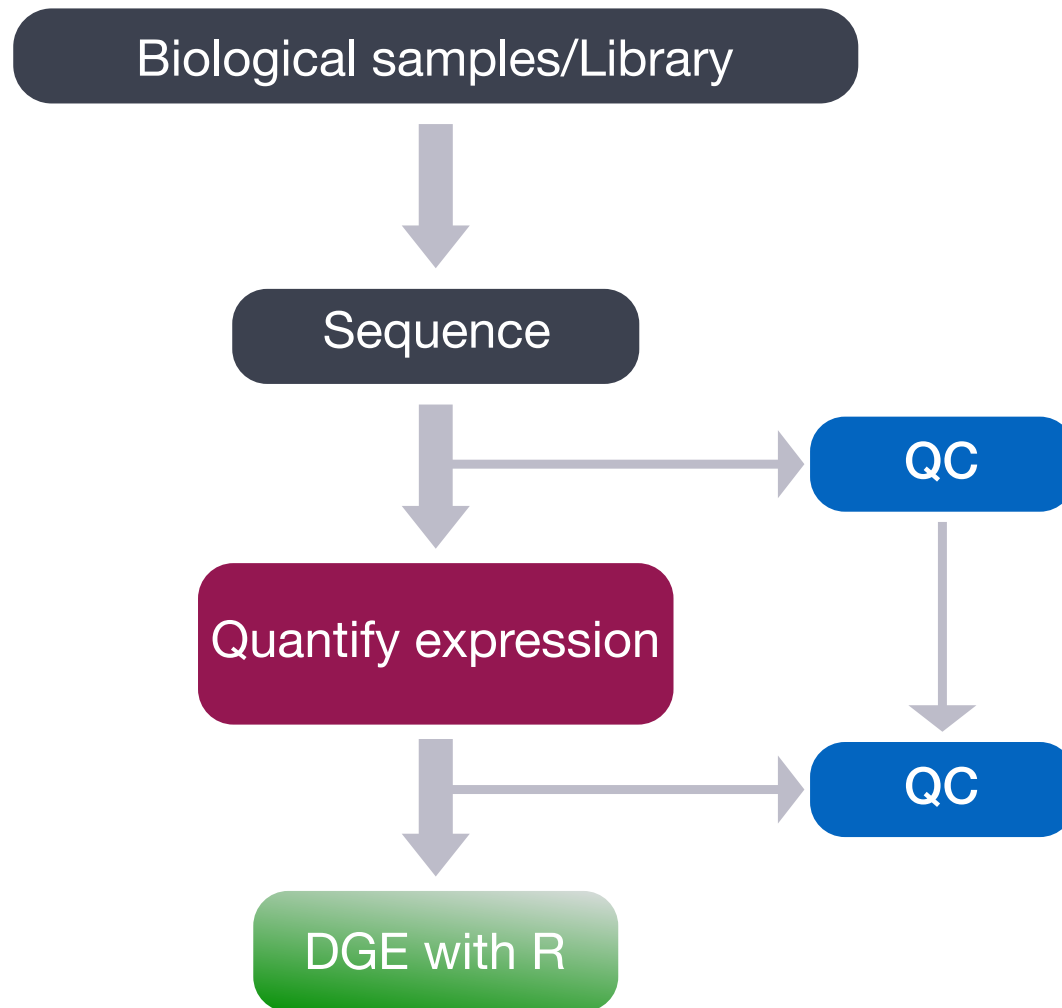
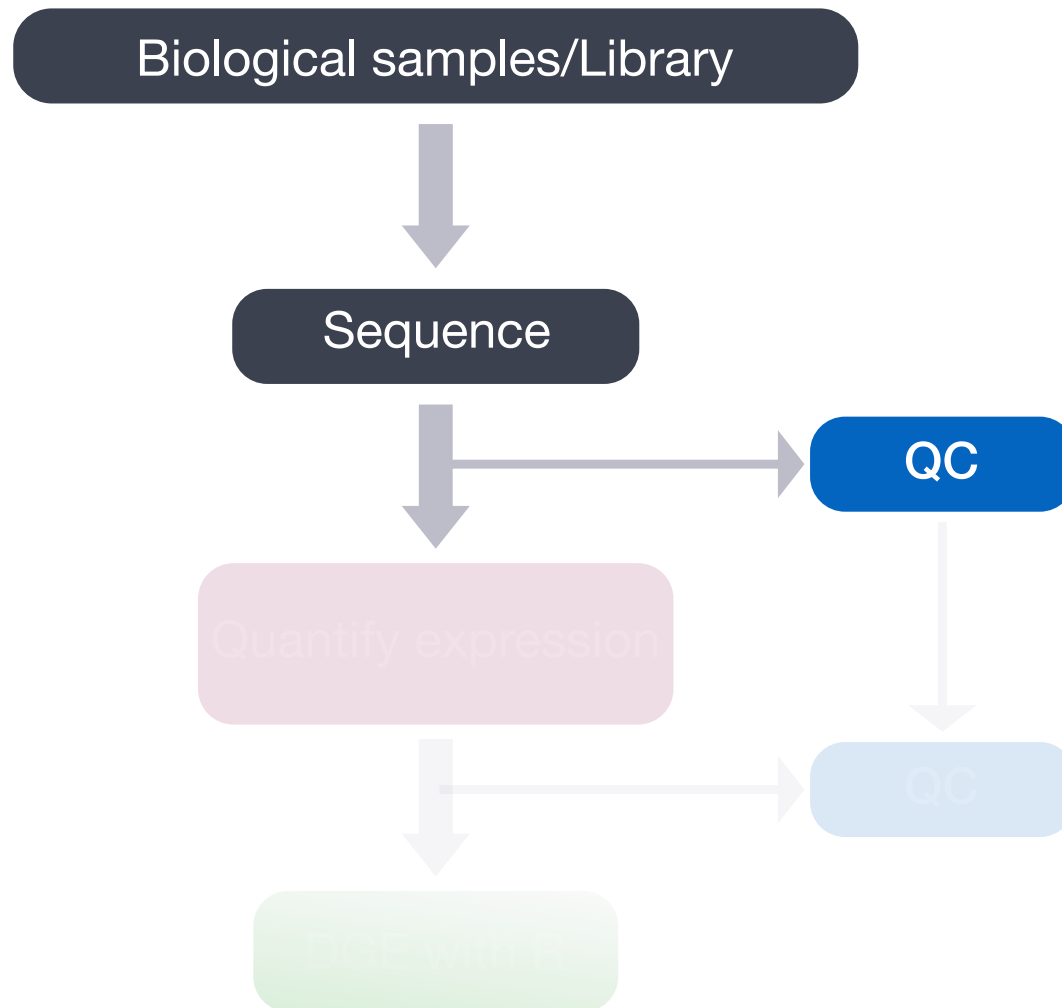


RNA-Seq Analysis Troubleshooting

RNA-seq Workflow



Quality Checks: Raw Data



Quality Checks: Raw Data

All NGS analyses require that the **quality of the raw data** is assessed prior to any downstream analysis.

The quality checks at this stage in the workflow include:

1. Checking the **quality of the base calls** to ensure that there were no issues during sequencing
2. Examining the reads to ensure their **quality metrics adhere to our expectations** for our experiment
3. Exploring reads for **contamination**

[illegible]

Quality Checks: Raw Data

Troubleshooting raw data quality problems:

- Poor quality data (problems at sequencing facility)
 - Poor quality across sequence
 - Drop in quality in the middle
 - Large percentage of sequences with low mean quality scores

Quality Checks: Raw Data

Quality Checks: Raw Data

Troubleshooting raw data quality problems:

- Issues based on read sequence expectations
 - Unexpected %GC for organism or % of each nucleotide does not remain similar across the read (except for first 10-12 bases)

Quality Checks: Raw Data

Troubleshooting raw data quality problems:

- Issues based on read sequence expectations
 - Unexpected %GC for organism or % of each nucleotide does not remain similar across the read (except for first 10-12 bases)
 - Contaminating sequences: different species, adapters, vector, mitochondrial/rRNA

Quality Checks: Raw Data

Troubleshooting raw data quality problems:

- Issues based on read sequence expectations
 - Unexpected %GC for organism or % of each nucleotide does not remain similar across the read (except for first 10-12 bases)
 - Contaminating sequences: different species, adapters, vector, mitochondrial/rRNA
 - Over-represented sequences: could be due to the experiment and not a problem

Quality Checks: Raw Data

Troubleshooting raw data quality problems:

- Issues based on read sequence expectations
 - Unexpected %GC for organism or % of each nucleotide does not remain similar across the read (except for first 10-12 bases)
 - Contaminating sequences: different species, adapters, vector, mitochondrial/rRNA
 - Over-represented sequences: could be due to the experiment and not a problem
 - High level of sequence duplications

Quality Checks: Raw Data

Troubleshooting raw data quality problems:

- Issues based on read sequence expectations
 - Unexpected %GC for organism or % of each nucleotide does not remain similar across the read (except for first 10-12 bases)
 - Contaminating sequences: different species, adapters, vector, mitochondrial/ rRNA
 - Over-represented sequences: could be due to the experiment and not a problem
 - High level of sequence duplications
 - low complexity library, too many cycles of PCR amplification / too little starting material

Quality Checks: Raw Data

Troubleshooting raw data quality problems:

- Issues based on read sequence expectations
 - Unexpected %GC for organism or % of each nucleotide does not remain similar across the read (except for first 10-12 bases)
 - Contaminating sequences: different species, adapters, vector, mitochondrial/ rRNA
 - Over-represented sequences: could be due to the experiment and not a problem
 - High level of sequence duplications
 - low complexity library, too many cycles of PCR amplification / too little starting material
 - Over-represented sequences more than 1-2%, unless expected based on experimental design

Quality Checks: Raw Data

Troubleshooting raw data quality problems:

- Issues based on read sequence expectations
 - Unexpected %GC for organism or % of each nucleotide does not remain similar across the read (except for first 10-12 bases)
 - Contaminating sequences: different species, adapters, vector, mitochondrial/rRNA
 - Over-represented sequences: could be due to the experiment and not a problem
 - High level of sequence duplications
 - low complexity library, too many cycles of PCR amplification / too little starting material
 - Over-represented sequences more than 1-2%, unless expected based on experimental design
 - contaminating sequences: adapters, vector, mitochondrial/rRNA

Quality Checks: Raw Data

Raw Data QC Goals:

- Identify sequencing problems and determine whether there is a need to contact the sequencing facility
- Identify over-represented contaminating sequences
- Gain insight into library complexity (rRNA contamination, duplications)
- Ensure organism is properly represented by %GC content

Quality Checks: Raw Data

Quality Checks: Raw Data

Raw Data QC Goals:

Can we identify a degraded RNA-Seq sample (low RIN #) using these raw data QC metrics?

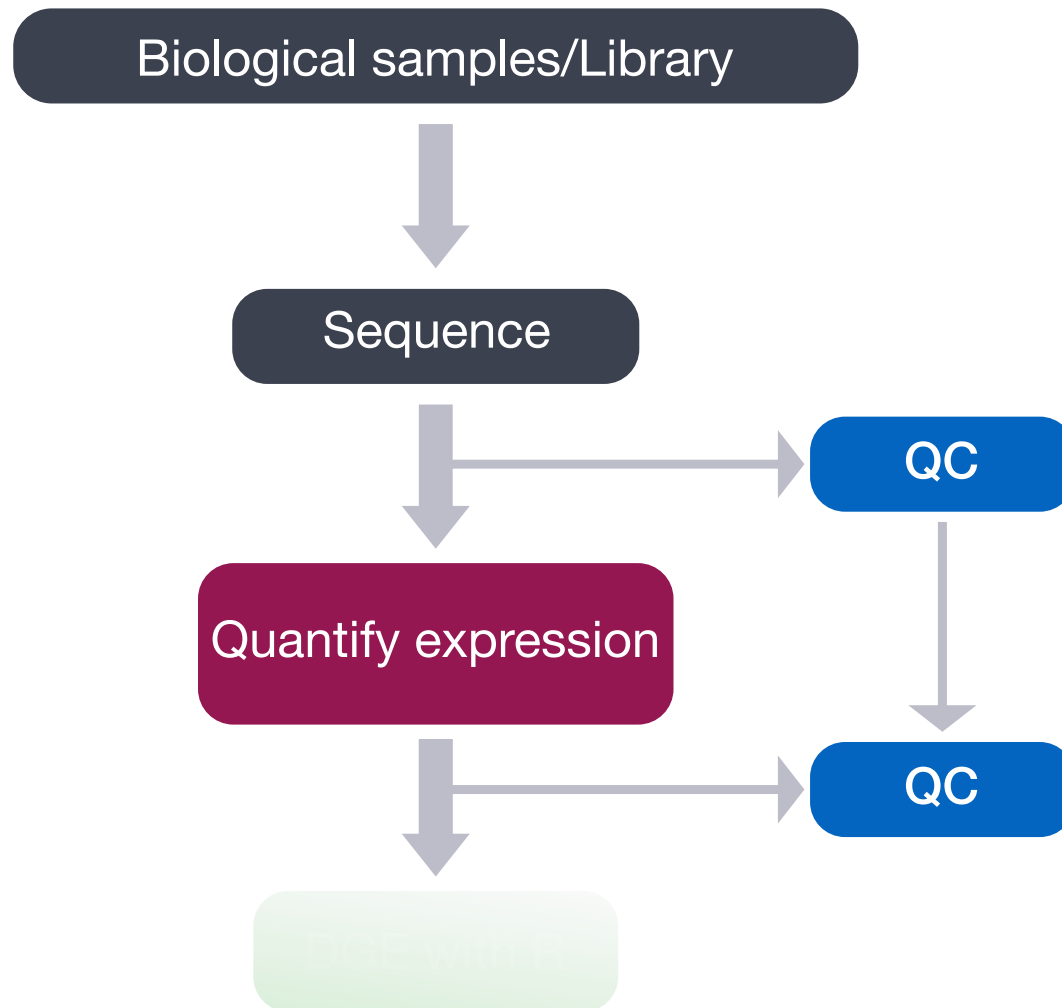
Quality Checks: Raw Data

Raw Data QC Goals:

Can we identify a degraded RNA-Seq sample (low RIN #) using these raw data QC metrics?

Since reads from degraded samples are generally just shorter, the quality of the sequenced nucleotides should be fine. At this step, **degraded libraries will not likely affect the quality metrics.**

Quality Checks: Aligned Data



Quality Checks: Aligned Data

Evaluating the **quality of the aligned data** can give important information about the quality of the library. The quality checks at this stage in the workflow include:

1. Checking the total percent of reads aligning to the genome
2. Determining the percent uniquely mapping reads
3. Examining the total number of reads aligning to each sample
4. Checking percent of paired-end reads that are properly paired

Quality Checks: Aligned Data

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low percentage ($< 80\%$) of reads aligned

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low percentage ($< 80\%$) of reads aligned
 - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome chosen, poor quality reference genome

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low percentage (< 80%) of reads aligned
 - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome chosen, poor quality reference genome
- Low percentage (< 60%) of **uniquely** aligning reads

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low percentage (< 80%) of reads aligned
 - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome chosen, poor quality reference genome
- Low percentage (< 60%) of **uniquely** aligning reads
 - low number of total reads aligning, organism has high number of paralogous genes, very short read length, low quality bases

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low percentage ($< 80\%$) of reads aligned
 - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome chosen, poor quality reference genome
- Low percentage ($< 60\%$) of **uniquely** aligning reads
 - low number of total reads aligning, organism has high number of paralogous genes, very short read length, low quality bases
- Large differences in sequencing depth between samples

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low percentage (< 80%) of reads aligned
 - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome chosen, poor quality reference genome
- Low percentage (< 60%) of **uniquely** aligning reads
 - low number of total reads aligning, organism has high number of paralogous genes, very short read length, low quality bases
- Large differences in sequencing depth between samples
 - library prep / sequencing

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low percentage ($< 80\%$) of reads aligned
 - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome chosen, poor quality reference genome
- Low percentage ($< 60\%$) of **uniquely** aligning reads
 - low number of total reads aligning, organism has high number of paralogous genes, very short read length, low quality bases
- Large differences in sequencing depth between samples
 - library prep / sequencing
- For paired-end data: large number of reads not properly paired

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

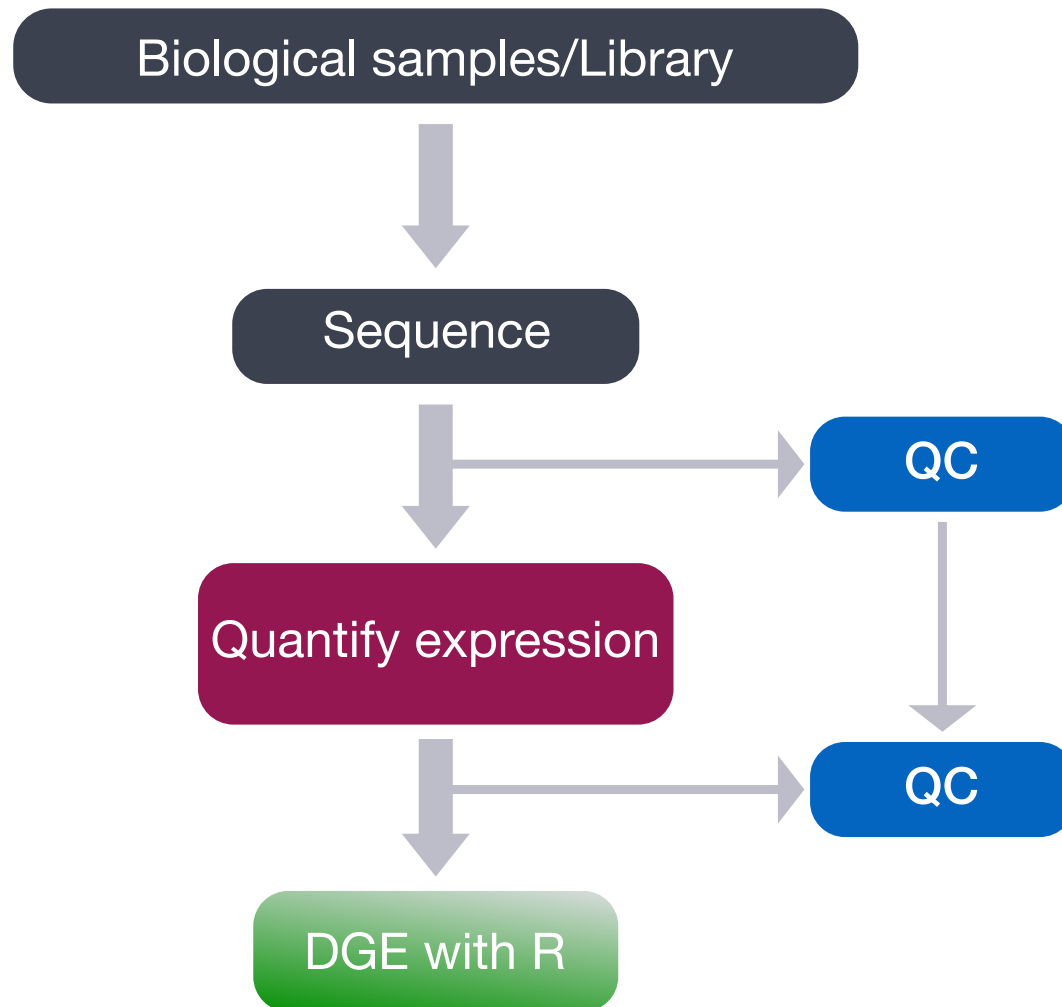
- Low percentage ($< 80\%$) of reads aligned
 - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome chosen, poor quality reference genome
- Low percentage ($< 60\%$) of **uniquely** aligning reads
 - low number of total reads aligning, organism has high number of paralogous genes, very short read length, low quality bases
- Large differences in sequencing depth between samples
 - library prep / sequencing
- For paired-end data: large number of reads not properly paired
 - poor quality reads

Quality Checks: Aligned Data

Aligned Data QC Goals:

- Ensure the library depth and percentage of reads mapping to each sample is similar
- Identify poor alignment parameters or low quality library
- Discover contamination from another organism

Quality Checks: Quantified Data



These materials have been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

