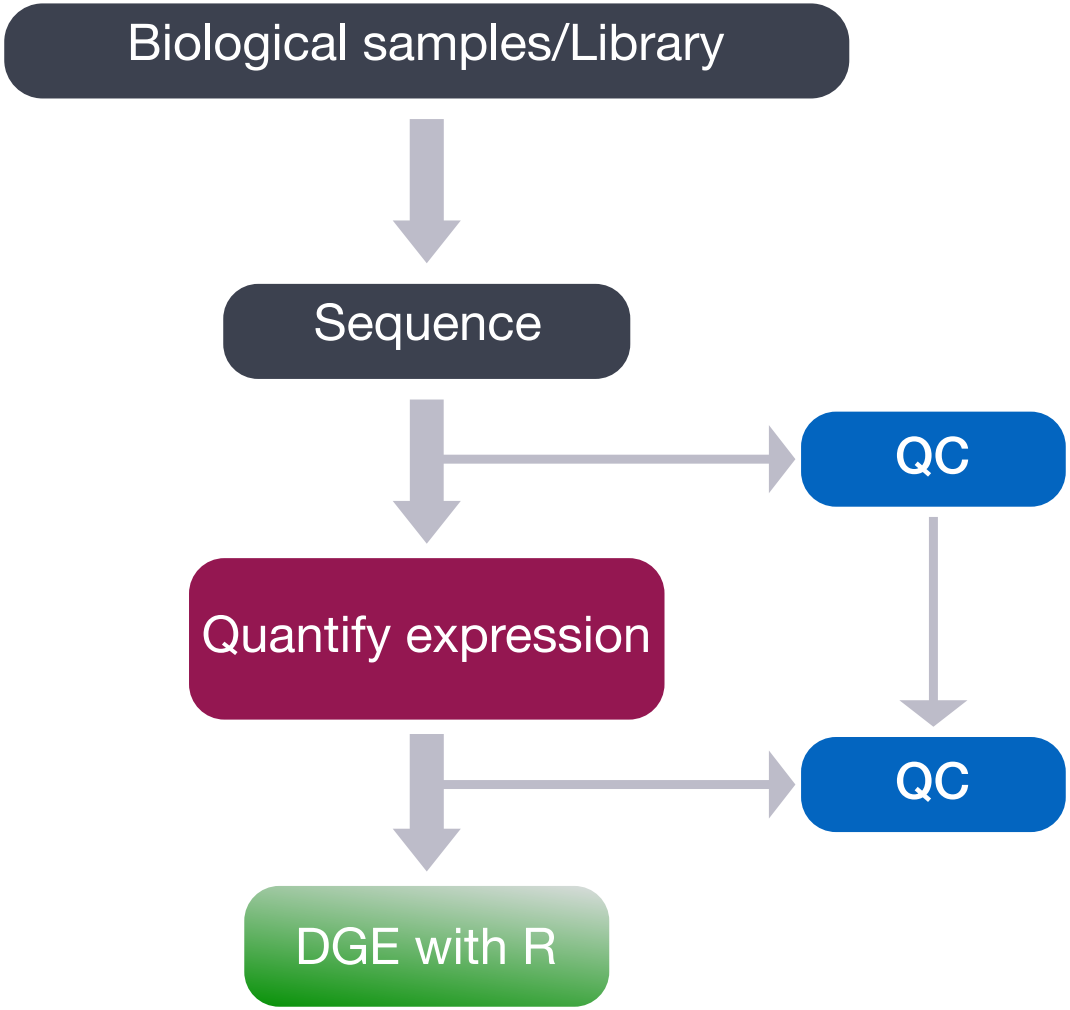
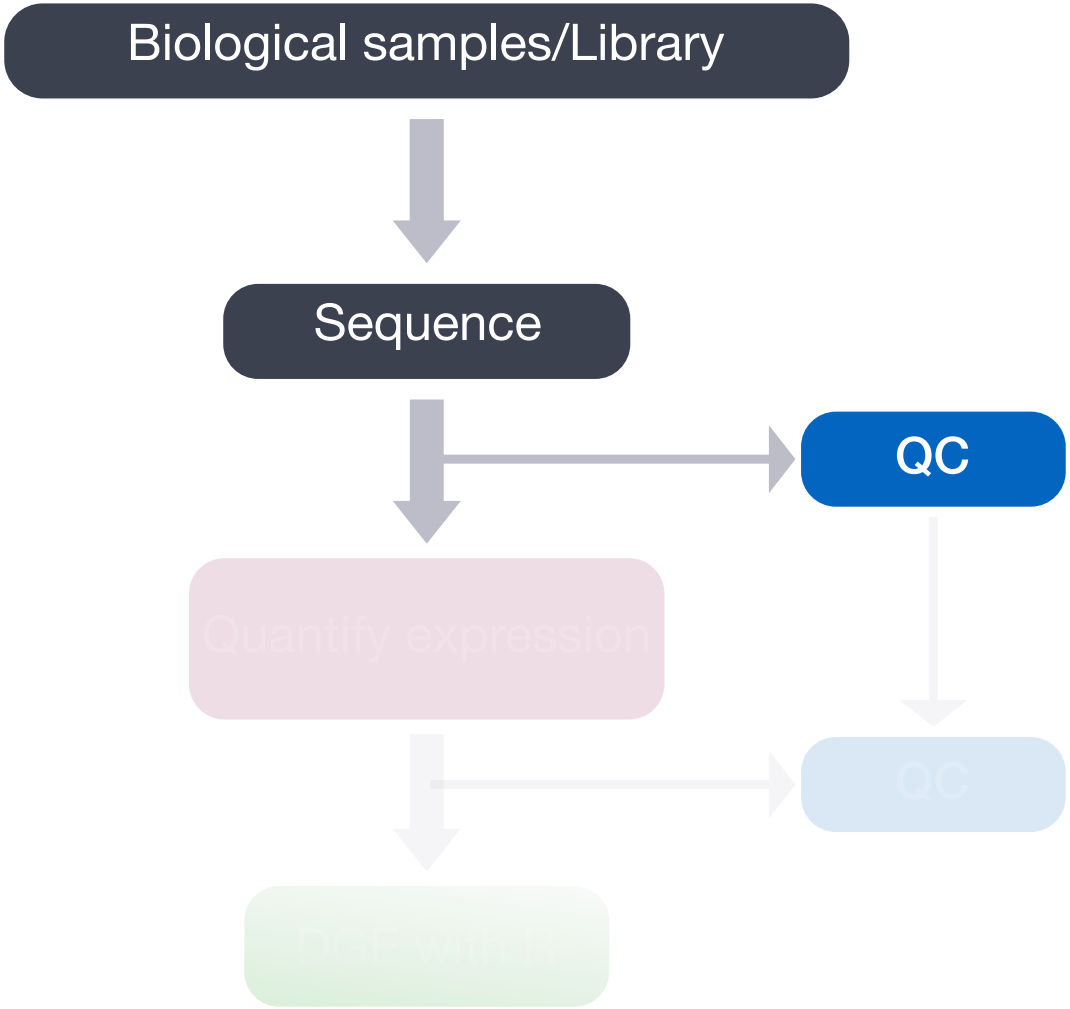
RNA-Seq Analysis Troubleshooting

RNA-seq Workflow



Quality Checks: Raw Data



Quality Checks: Raw Data

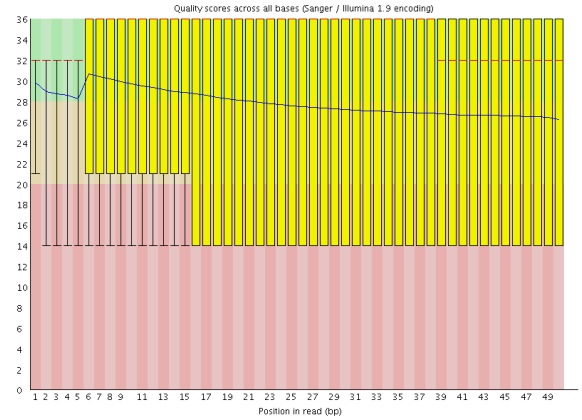
All NGS analyses require that the **quality of the raw data** is assessed prior to any downstream analysis.

The quality checks at this stage in the workflow include:

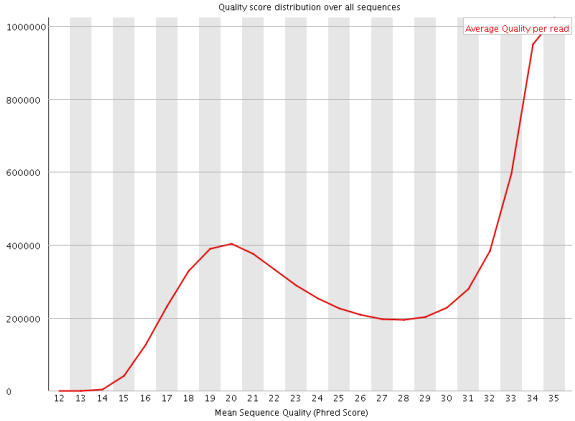
1. Checking the **quality of the base calls** to ensure that there were no issues during sequencing
2. Examining the reads to ensure their **quality metrics adhere to our expectations** for our experiment
3. Exploring reads for **contamination**

Quality Checks: Raw Data

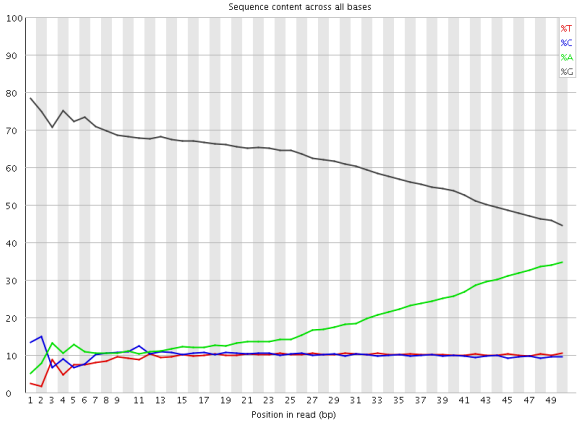
Per base sequence quality



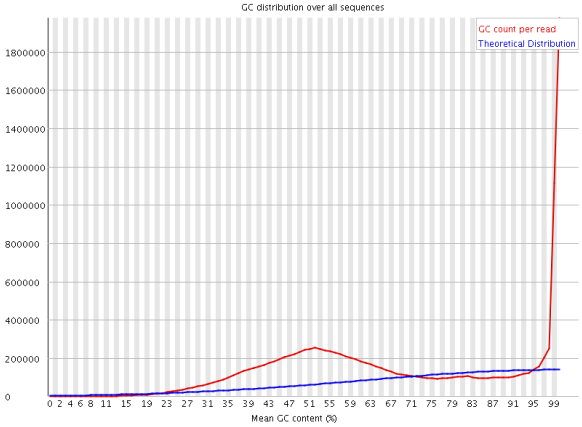
Per sequence quality scores



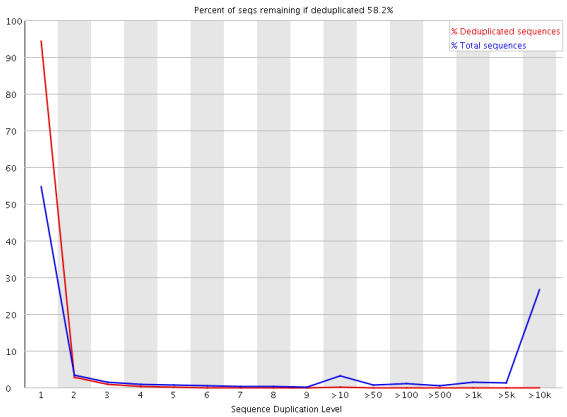
Per base sequence content



Per sequence GC content



Sequence Duplication Levels



Overrepresented sequences

Sequence	Count	Percentage	Possible Sou
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	1838931	25.270618073251 26	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCA	22246	0.3057048740042707	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAG	19143	0.26306340029954844	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAGG	14083	0.1935288025084125	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAAAAAAAAAA	13303	0.18281003051689354	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAAA	12912	0.17743690250576033	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAA	11561	0.1588714397358345	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCG	11428	0.15704375169112678	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAA	11152	0.1532509554479739	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCG	9922	0.1363482765382709	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAAAAAAA	9693	0.1332013550176839	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAA	9340	0.12835042359075288	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAAAA	8557	0.11759042555311268	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAA	8156	0.11207987738824203	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAGCCCG	7471	0.10266659685722856	No Hit
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAGCCCG	7294	0.10023426013607618	No Hit

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?

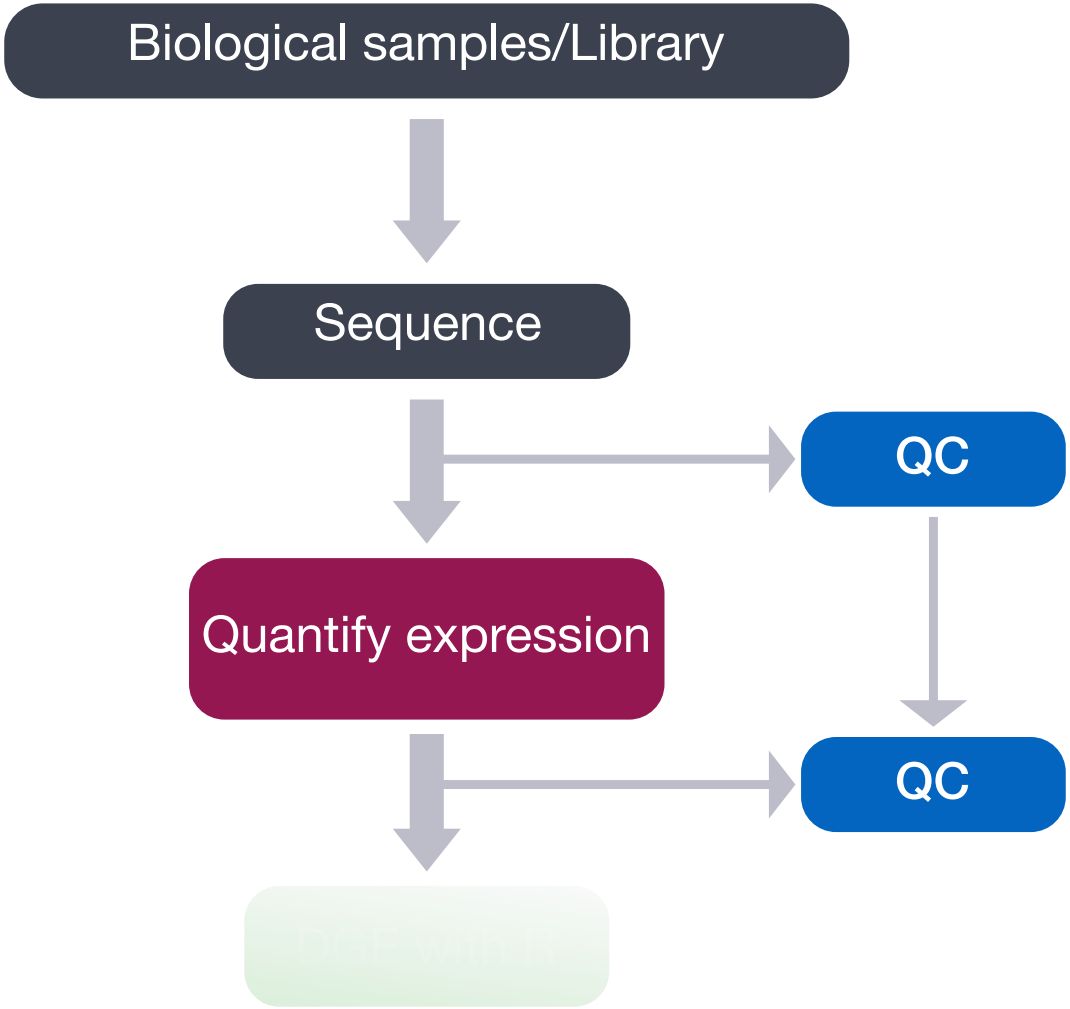
Quality Checks: Raw Data

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 (< 70%) of reads aligned

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low percentage (< 70%) 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 (< 70%) 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 (< 70%) 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 (< 70%) 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 (< 70%) 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 (< 70%) 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 (< 70%) 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

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- 5’ - 3’ coverage biases

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- 5’ - 3’ coverage biases

  - poor quality RNA samples (low RIN), library preparation method

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- 5’ - 3’ coverage biases
  - poor quality RNA samples (low RIN), library preparation method
- GC biases

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- 5’ - 3’ coverage biases
  - poor quality RNA samples (low RIN), library preparation method
- GC biases
  - PCR amplification

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
- Identify biases present in the data and correct for it

Quality Checks: Quantified Data