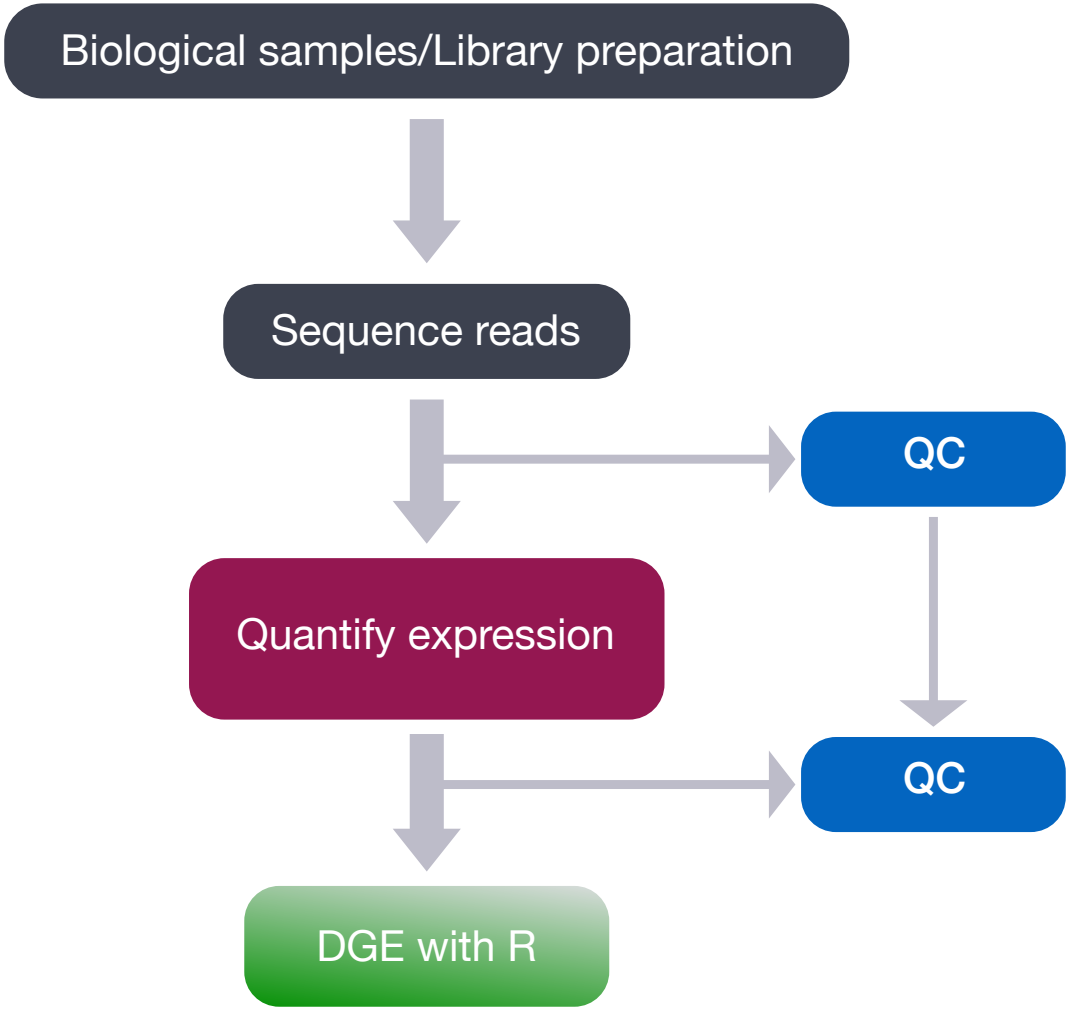
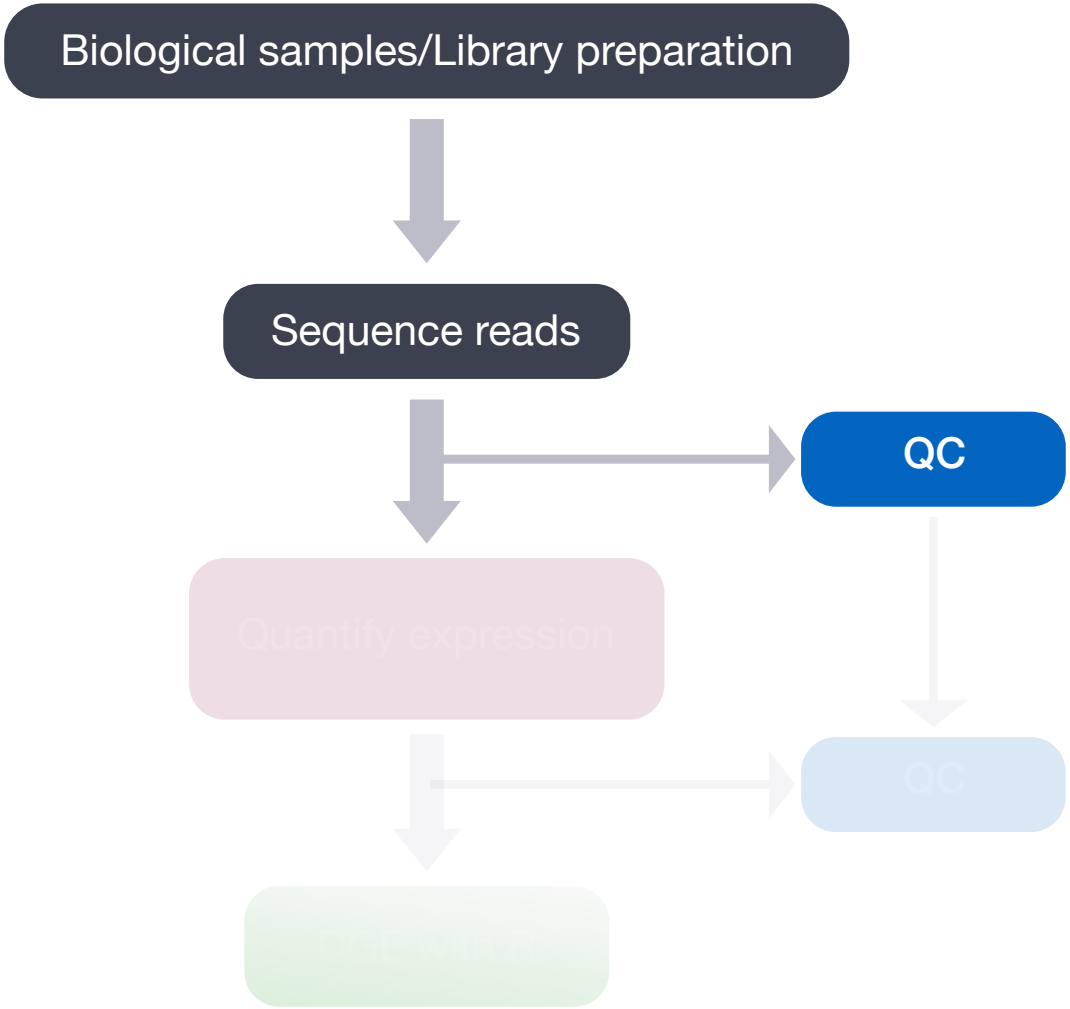
RNA-Seq Analysis Troubleshooting

RNA-seq Workflow



Quality Checks: Raw Data



Quality Checks: Raw Data

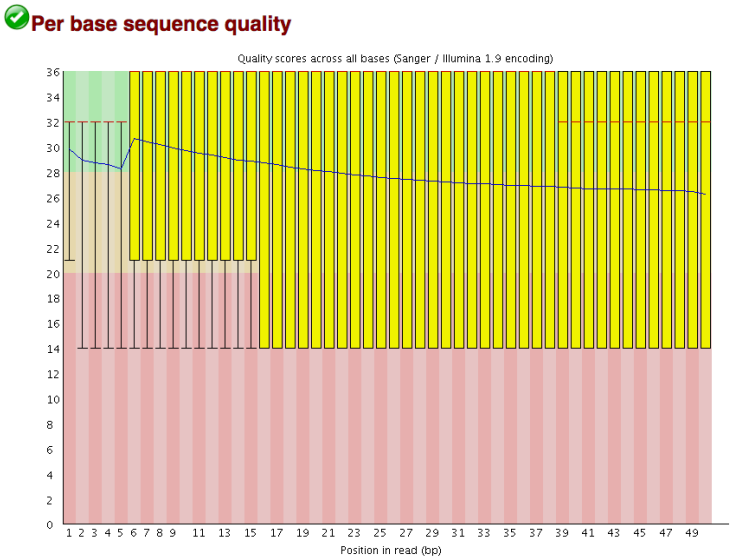
Raw Data QC Goals:

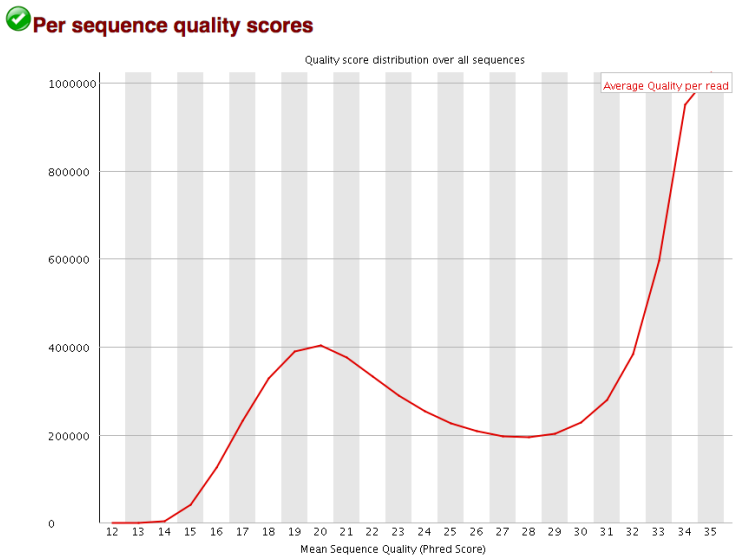
- **Identify sequencing problems** and determine whether there is a need to contact the sequencing facility
- Identify **contaminating** sequences
- Gain insight into **library complexity** (rRNA contamination, duplications)

Quality Checks: Raw Data

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



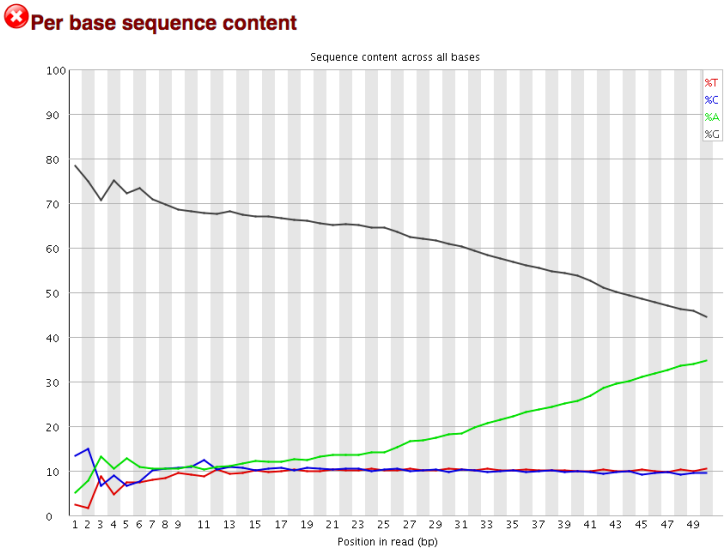


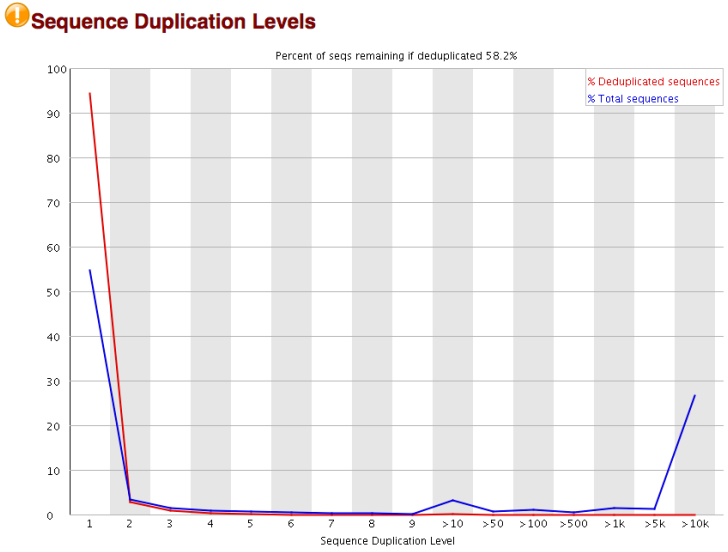
Quality Checks: Raw Data

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





Quality Checks: Raw Data

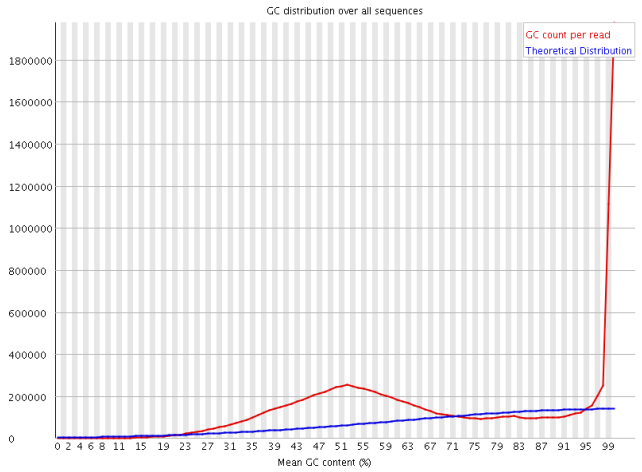
The quality checks at this stage in the workflow include:

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**

Per sequence GC content



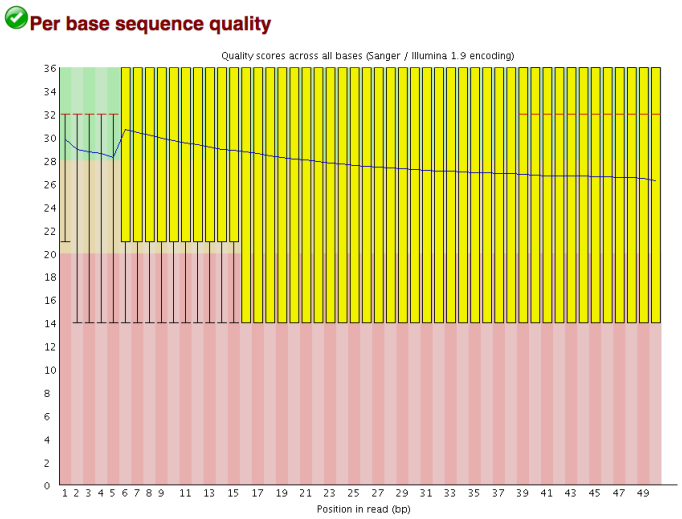
Overrepresented sequences

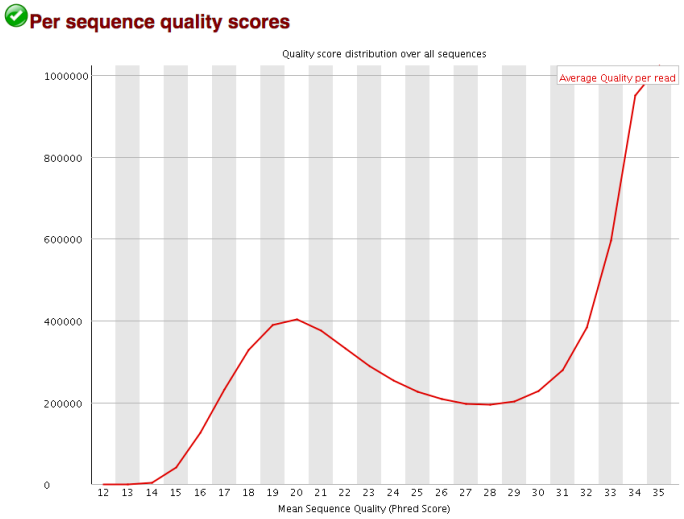
Sequence	Count	Percentage	Possible Sour
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	1838931	25.27061807325126	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGA	22246	0.3057048740042707	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAG	19143	0.26306340029954844	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGG	14083	0.1935288025084125	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAAAAAAAAA	13303	0.18281003051689354	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAA	12912	0.17743690250576033	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAA	11561	0.1588714397358345	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGG	11428	0.15704375169112678	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAAAA	11152	0.1532509554479739	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGG	9922	0.1363482765382709	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAAAAAAAAAA	9693	0.1332013550176839	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAA	9340	0.12835042359075288	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAA	8557	0.11759042555311268	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAA	8156	0.11207987738824203	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGGG	7471	0.10266659685722856	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGGGG	7294	0.10023426013607618	No Hit

Quality Checks: Raw Data

Troubleshooting low quality base calls

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



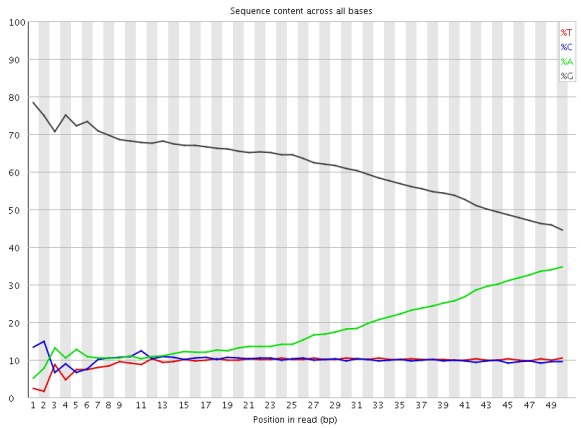


Quality Checks: Raw Data

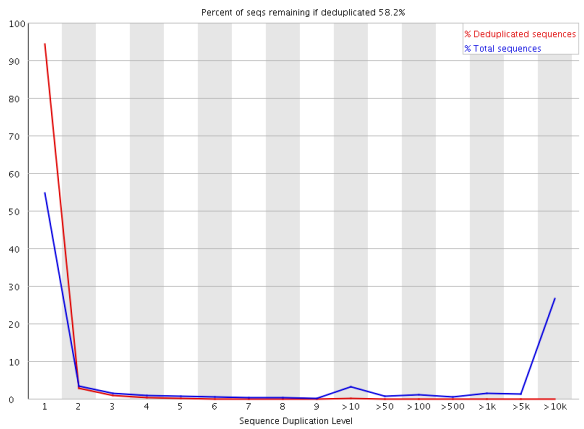
Troubleshooting unusual quality metrics

- Biased sequence composition
  - Contaminating sequences (mitochondrial/rRNA, adapters) or over-represented sequences
- High level of sequence duplications
  - Low complexity library, too many cycles of PCR amplification / too little starting material
  - S[illegible]script[illegible]

Per base sequence content



Sequence Duplication Levels

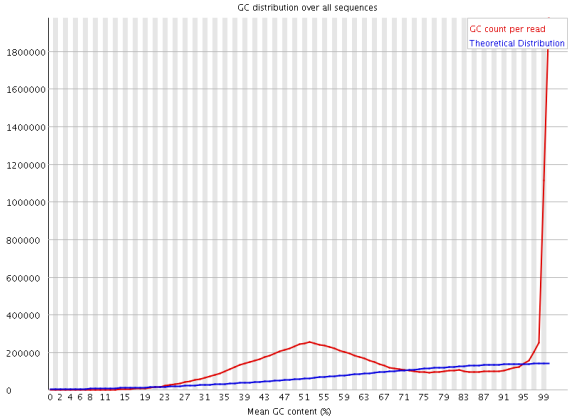


Quality Checks: Raw Data

Troubleshooting possible contamination:

- Unexpected %GC for organism
  - Contaminating sequences: different species, adapters, vector, mitochondrial/rRNA
  - Over-represented sequences: could be due to the experiment and not a problem
- **Over-represented sequences > 2%** (unless expected for the experiment)
  - Contaminating sequences: adapters, vector, mitochondrial/rRNA, over-expression

Per sequence GC content



Overrepresented sequences

Sequence	Count	Percentage	Possible Sour
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	1838931	25.27061807325126	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGA	22246	0.3057048740042707	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAG	19143	0.2630634002995484	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGG	14083	0.1935288025084125	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAAAAA	13303	0.1828100305168935 4	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAA	12912	0.1774369025057603 3	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAA	11561	0.158871439735834 5	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGG	11428	0.1570437516911267 8	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAAA	11152	0.153250955447973 9	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGG	9922	0.136348276538270 9	No Hit
GGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAAAAAAAAAA	9693	0.133201355017683 9	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAA	9340	0.128350423590752 8 8	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAA	8557	0.117590425553112 6 8	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAA	8156	0.112079873882420 3	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGGG	7471	0.102666596857228 5 6	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGGG	7294	0.100234260136076 1 8	No Hit

Quality Checks: Raw Data

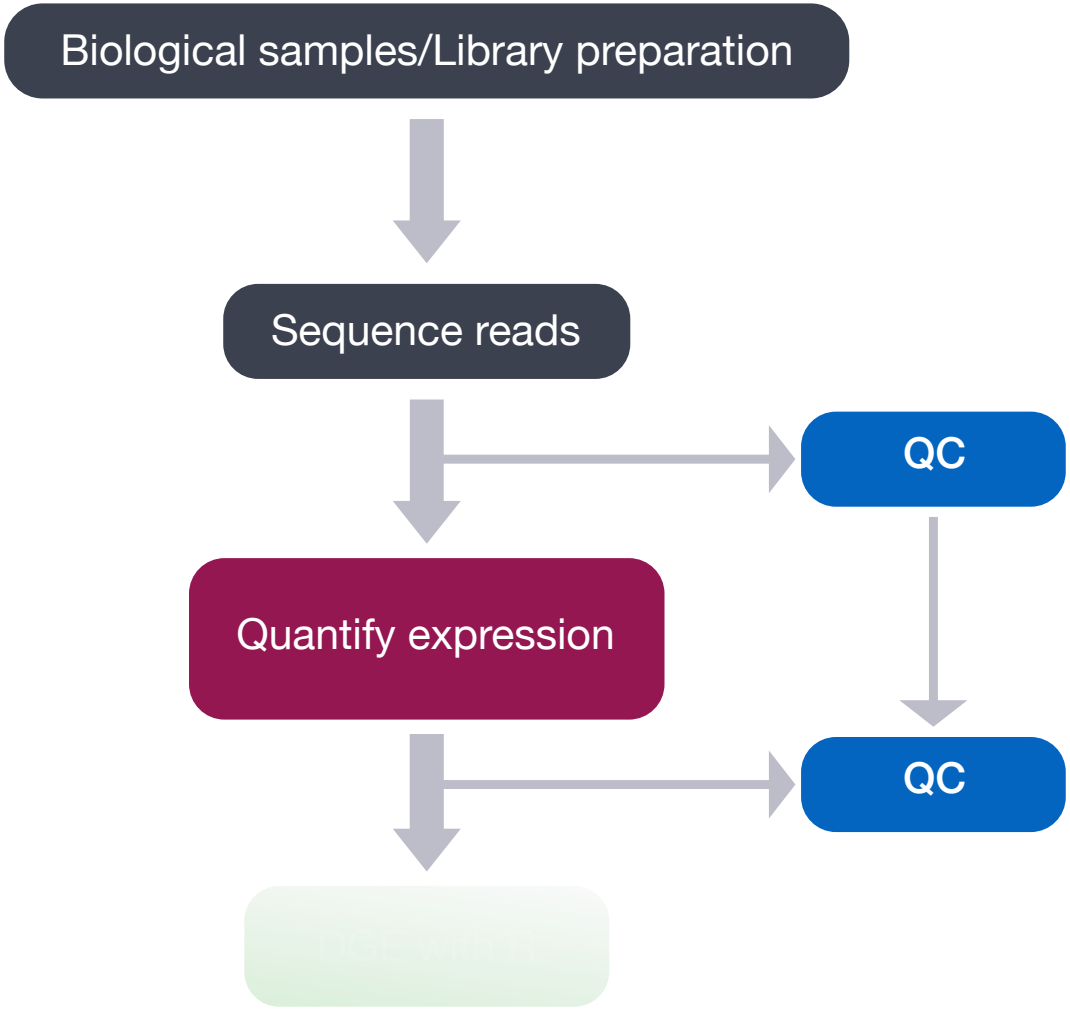
FAQ: Can we identify a degraded RNA-Seq sample (low RIN #) using these raw data QC metrics?

Quality Checks: Raw Data

FAQ: 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

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** libraries
- **Discover contamination** from another organism or from DNA
- **Identify biases** present in the data and **correct** for it
- Ensure the experiment generated the **expected data** (% intronic reads, etc.)

Quality Checks: Aligned Data

The quality checks at this stage in the workflow include:

1. Checking the total **percent of reads aligning** to the genome and transcriptome

General Statistics

Copy table | Configure Columns | Sort by highlight | Plot | Showing $^{8}/_{8}$ rows and $^{9}/_{11}$ columns.

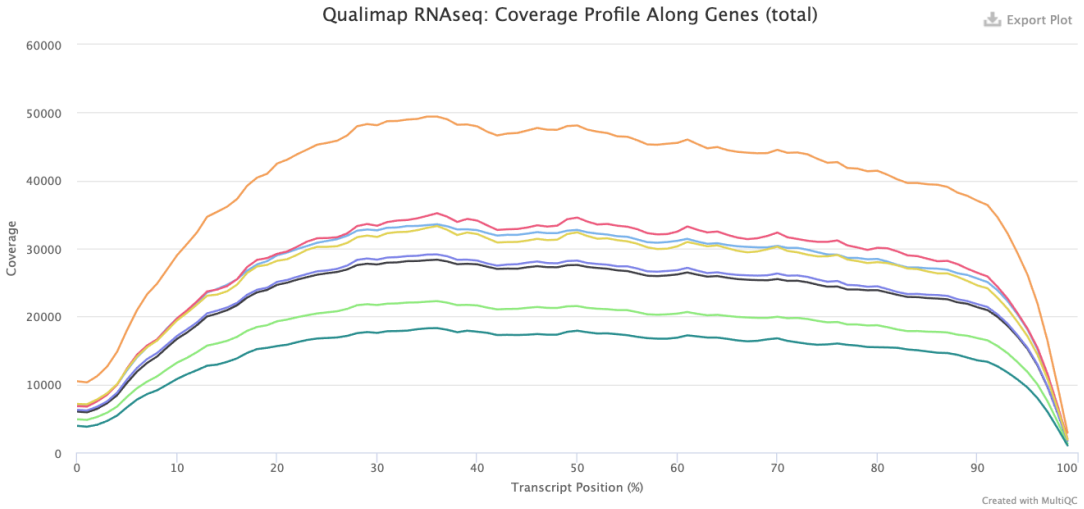
Sample Name	5'-3' bias	M Aligned	% Aligned	M Aligned	% Aligned	M Aligned	% Dups	% GC	M Seqs
Irrel_kd_1	1.18	35.6	86.4%	31.2	92.1%	33.2	55.9%	47%	36.1
Irrel_kd_2	1.14	30.4	86.0%	26.5	92.2%	28.4	53.6%	47%	30.8
Irrel_kd_3	1.19	23.6	85.7%	20.5	92.0%	22.0	50.1%	48%	23.9
Mov10_kd_2	1.13	51.9	86.0%	45.3	91.6%	48.3	60.5%	48%	52.7
Mov10_kd_3	1.13	30.7	86.0%	26.8	91.6%	28.5	54.6%	47%	31.1
Mov10_oe_1	1.09	38.1	80.2%	32.1	88.9%	35.5	56.5%	47%	40.0
Mov10_oe_2	1.18	35.4	81.0%	30.0	88.8%	33.0	55.9%	48%	37.1
Mov10_oe_3		20.3	81.5%	17.3	90.0%	19.1	50.1%	47%	21.2

Quality Checks: Aligned Data

The quality checks at this stage in the workflow include:

1. Checking the total **percent of reads aligning** to the genome and transcriptome

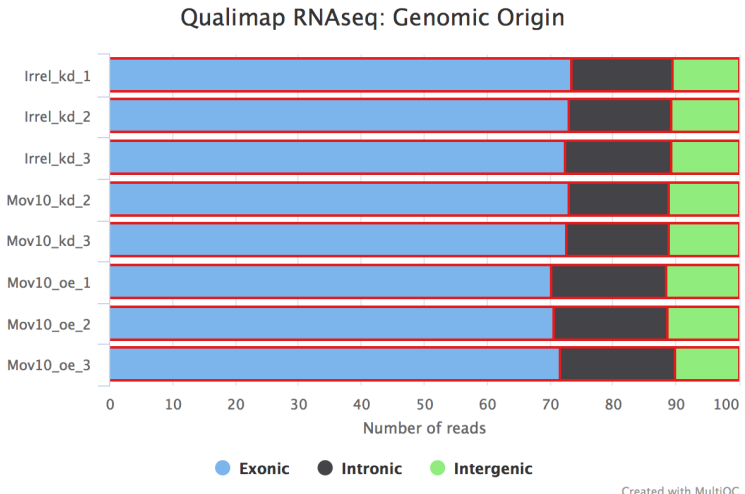
2. Check for any **biases in the data,** including positional coverage, GC bias and sequence biases at the 5’ and 3’ ends



Quality Checks: Aligned Data

The quality checks at this stage in the workflow include:

1. Checking the total **percent of reads aligning** to the genome and transcriptome
2. Check for any **biases in the data,** including positional coverage, GC bias and sequence biases at the 5’ and 3’ ends
3. Determine the **presence of any contamination**, by evaluating reads aligning to specific genomic features



Quality Checks: Aligned Data

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low read mapping rate (< 70% to the genome / 60% to the transcriptome)

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low read mapping rate (< 70% to the genome / 60% to the transcriptome)

  - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome/transcriptome chosen, poor quality reference genome/ transcriptome

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low read mapping rate (< 70% to the genome / 60% to the transcriptome)

  - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome/transcriptome chosen, poor quality reference genome/ transcriptome

- 5’ - 3’ coverage biases

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low read mapping rate (< 70% to the genome / 60% to the transcriptome)

  - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome/transcriptome chosen, poor quality reference genome/transcriptome

- 5’ - 3’ coverage biases

  - poor quality RNA samples (low RIN), library preparation method (3’ bias common with polyA selection, 5’ bias with rRNA depletion)

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low read mapping rate (< 70% to the genome / 60% to the transcriptome)

  - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome/transcriptome chosen, poor quality reference genome/transcriptome

- 5’ - 3’ coverage biases

  - poor quality RNA samples (low RIN), library preparation method (3’ bias common with polyA selection, 5’ bias with rRNA depletion)

- GC biases

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low read mapping rate (< 70% to the genome / 60% to the transcriptome)

  - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome/transcriptome chosen, poor quality reference genome/transcriptome

- 5’ - 3’ coverage biases

  - poor quality RNA samples (low RIN), library preparation method (3’ bias common with polyA selection, 5’ bias with rRNA depletion)

- GC biases

  - PCR amplification of fragments during library preparation

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low read mapping rate (< 70% to the genome / 60% to the transcriptome)

  - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome/transcriptome chosen, poor quality reference genome/transcriptome

- 5’ - 3’ coverage biases

  - poor quality RNA samples (low RIN), library preparation method (3’ bias common with polyA selection, 5’ bias with rRNA depletion)

- GC biases

  - PCR amplification of fragments during library preparation

- Low exonic mapping rates

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low read mapping rate (< 70% to the genome / 60% to the transcriptome)

  - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome/transcriptome chosen, poor quality reference genome/ transcriptome

- 5' - 3' coverage biases

  - poor quality RNA samples (low RIN), library preparation method (3' bias common with polyA selection, 5' bias with rRNA depletion)

- GC biases

  - PCR amplification of fragments during library preparation

- Low exonic mapping rates

  - low percentage of reads aligning to exons (<50%), high percentage in introns or intergenic regions (>30%) or high percentage in rRNA (>2%)

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low read mapping rate (< 70% to the genome / 60% to the transcriptome)

  - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, inappropriate reference genome/transcriptome chosen, poor quality reference genome/ transcriptome

- 5' - 3' coverage biases

  - poor quality RNA samples (low RIN), library preparation method (3' bias common with polyA selection, 5' bias with rRNA depletion)

- GC biases

  - PCR amplification of fragments during library preparation

- Low exonic mapping rates

  - low percentage of reads aligning to exons (<50%), high percentage in introns or intergenic regions (>30%) or high percentage in rRNA (>2%)

  - genomic DNA contamination, pre-mRNA, unsuccessful ribo-depletion

RNA-seq Workflow