Quality Control Report Inference

Summary: W - Warning, F - Fail

	Per base sequence content	**Per sequence GC content**	**Sequence Duplication Levels**	**Overrepresented sequences**
KO_R1	Mostly, parallel - high deviation at the very beginning (F)	Sharp peak from 27 to 45% - crosses 2500000 (F)	Unique - 25.34% (F)	GTTTCGTTTACCTTCTATAAGGCTATGATGAGCTCATGTAATTGAAACAC (0.537%) (W)
KO_R2	Mostly, parallel - some deviation at the very beginning (F)	Sharp peak from 27 to 45% - crosses 2000000 (F)	Unique - 28.68% (F)	CGGTGGCGCACGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGACAGGAGG (0.3129%) (W)
WT_R1	Mostly, parallel - high deviation at the very beginning (F)	Sharp peak from 27 to 45% - crosses 1800000 (F)	Unique - 20.73%% (F)	GTTTCGTTTACCTTCTATAAGGCTATGATGAGCTCATGTAATTGAAACAC (0.5408%) (W)
WT_R2	Mostly, parallel - some deviation at the very beginning (F)	Sharp peak from 27 to 45% - crosses 1600000 (F)	Unique - 23.51% (F)	GGGGCGCGAAGCGGGGCTGGGCGCGCCGCGGCTGGACGAGGCGCCGCC (0.2347%) (W)

In Detail:

Problem areas and inferences: for **KO_R1**

Warnings:

1. Per tile sequence quality: Some tile may have shown a mean Phred score more than 2 less than the mean for that base across all tiles. Our plot is blue all over, with the exception of 4-5 red spots i.e. regions of low quality.

4.Sequence Length Distribution: All sequences are not the same length.

6. Overrepresented sequences: The sequence - "GTTTCGTTTACCTTCTATAAGGCTATGATGAGCTCATGTAATTGAAACAC" has the highest percentage (0.537%).

Failures:

2. Per base sequence content: The module failed because the difference between A and T, or G and C was greater than 20% in any position. The lines in our plot are mostly parallel, with the exception of high deviation at the very beginning.

3. Per sequence GC content: The sum of the deviations from the normal distribution has exceeded 30% of the reads. Our graph has a steep rise and sharp peak (indicative of over-represented sequences) from 27 to 45% of mean GC Content (that crosses 2500000)

5. Sequence Duplication Levels: With the percentage of sequences remaining if deduplicated being 25.34%, which means non-unique sequences make up 74.66% of this library.

Problem areas and inferences: for **KO_R2**

Warnings:

1. Per tile sequence quality: Plot is blue all over, with the exception of 4-5 red spots i.e. regions of low quality.

2. Per base sequence content: The lines in the plot are mostly parallel, with the exception of some deviation at the very beginning.

4.Sequence Length Distribution: All sequences are not the same length.

6. Overrepresented sequences: The sequence - "CGGTGGCGCACGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGACAGGAGG" has the highest percentage (0.3129%).

Failures:

3. Per sequence GC content: Similar observation as before - peak crosses 2000000.

5. Sequence Duplication Levels: Mostly similar observation - the percentage of sequences remaining if deduplicated is 28.68%.

Problem areas and inferences: for **WT_R1**

Warnings:

1. Per tile sequence quality: Plot is blue all over, with the exception of 4-5 red spots i.e. regions of low quality.

4.Sequence Length Distribution: All sequences are not the same length.

6. Overrepresented sequences: The sequence - "GTTTCGTTTACCTTCTATAAGGCTATGATGAGCTCATGTAATTGAAACAC" has the highest percentage (0.5408%).

Failures:

2. Per base sequence content: The lines in the plot are mostly parallel, with the exception of high deviation at the very beginning.

3. Per sequence GC content: Mostly similar observation as before, but apart from the peak the graph is a bit spread out as well - peak crosses 1800000.

5. Sequence Duplication Levels: Mostly similar observation - the percentage of sequences remaining if deduplicated is 20.73%.

Problem areas and inferences: for **WT_R2**

Warnings:

1. Per tile sequence quality: Plot is blue all over, with the exception of 4-5 red spots i.e. regions of low quality.

4.Sequence Length Distribution: All sequences are not the same length.

6. Overrepresented sequences: The sequence - "GGGGCGCGAAGCGGGGCTGGGCGCGCCGCGGCTGGACGAGGCGCCGCC" has the highest percentage (0.2347%).

Failures:

2. Per base sequence content: The lines in the plot are mostly parallel, with the exception of some deviation at the very beginning.

3. Per sequence GC content: Similar observation as before, but with the presence of minor peaks as well - peak crosses 1600000.

5. Sequence Duplication Levels: Mostly similar observation - the percentage of sequences remaining if deduplicated is 23.51%.

---

1. Per tile sequence quality: Depicts the quality scores from each tile across all of our bases to see if there was a loss in quality associated with only one part of the flowcell. Ideally, a good plot should be blue all over.

2. Per base sequence content: It plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called. Ideally in a random library the lines in a plot should run parallel with each other.

3. Per sequence GC content: It measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content.

4.Sequence Length Distribution: Plots the distribution of fragment sizes in the file which was analysed. Ideally, all sequences in a library should be of the same length.

5. Sequence Duplication Levels: Counts the degree of duplication for every sequence in a library and creates a plot showing the relative number of sequences with different degrees of duplication. A high level of duplication is more likely to indicate some kind of enrichment bias (eg PCR over amplification).

6. Overrepresented sequences: List of sequences which appear more than expected in the file. A sequence is considered overrepresented if it accounts for ≥ 0.1% of the total reads.