1. **Purpose**

This document provides quality control (QC) guidance for the analysis of nucleic acid next generation sequencing (NGS) data. Following the generation of a FASTQ file, this guidance should be utilized to assess the quality of sequence data prior to assembly and further analysis. The guidance takes into account specific QC checkpoints between computational processes to ensure each step is completed correctly, with high confidence, and to generate quality data metrics that yield an informative study.

QC checkpoints are necessary at several stages of bioinformatics analysis including raw read sequence filtering, each alignment and characterization stage. These steps ensure the sequence data meets standards for analysis, allows removal of low quality reads, and reduces false negatives and positives. This guidance also aims to promote standardized best practice measures in order to improve reproducibility of results.

1. **Scope**

This document provides information on post-sequencing, pre-analysis QC: quality control steps to be performed on NGS data in the form of a FASTQ file prior to assembly or further analysis.

1. **Related Documents**

|  |  |
| --- | --- |
| **Title** | **Document Control Number** |
| Bioinformatics QC Workflows |  |

1. **Responsibilities**

|  |  |
| --- | --- |
| **Position** | **Responsibility** |
| All Laboratory Staff | * Follow documented procedures |
| Team Lead | * Ensure documented procedures for data quality checks are established * Ensure documented procedures are followed |
| Quality Manager | * Ensure documented procedures are available to the end user * Review records of data quality checks as required |

1. **Definitions**

|  |  |
| --- | --- |
| **Term** | **Definition** |
| FASTQC | * A quality control tool for high throughput sequence data |
| PrinSeq | * A quality control software for filtering, reformatting and trimming sequence data. |
| Trimmomatic | * A flexible read trimming tool for Illumina |

1. **Sample Information / Processing**

Upon completion of the NGS run, transfer data to Isilon. *(Specify your laboratory data storage location here.)*

1. **FASTQC** 
   1. Once the sequencing run and initial Sequencing QC (SOP1) has been completed, a FASTQ file can be exported from Illumina’s RTA (Real-Time Analysis Software).
      1. A quality control software such as FASTQC should be used to assess the quality of the sequence data. While most sequencers will generate their own quality reports, these reports are generally more useful for identifying issues that originate with the sequencer. FASTQC can be effective in detecting problems with either the sequencer and/or the starting library material.
      2. The output from FASTQC will provide several statistics (in HTML format) including: per-base sequence quality and per-sequence quality scores, per-base N content, per-sequence GC content, overrepresented sequences, adapter content and K-mer content etc. (see table below). By default FASTQC will provide a green check validating these metrics a red X signifying a failed test or a yellow exclamation mark to indicate potential areas of concern. While these tests may appear to give a pass/fail indication, these should be taken in the context of what is expected from your library. Please review these test results below before continuing to further analysis. (Please see Figures A-1 and A-2 for examples of good and bad reports)

|  |  |
| --- | --- |
| **Metric** | **Description** |
| **Total Sequences** | A count of the total number of sequences processed. There are two values reported, actual and estimated. At the moment these will always be the same. In the future it may be possible to analyze just a subset of sequences and estimate the total number, to speed up the analysis, but since we have found that problematic sequences are not evenly distributed through a file we have disabled this for now. |
| **Filtered Sequences** | If running in Casava mode sequences flagged to be filtered will be removed from all analyses. The number of such sequences removed will be reported here. The total sequences count above will not include these filtered sequences and will include the number of sequences actually used for the rest of the analysis. |
| **Sequence Length** | Provides the length of the shortest and longest sequence in the set. If all sequences are the same length only one value is reported. |
| **% GC** | The overall %GC of all bases in all sequences. |
| **Per base sequence quality** | For each position a Box and Whisker type plot is drawn. The elements of the plot are as follows:   * The central red line is the median value * The yellow box represents the inter-quartile range (25-75%) * The upper and lower whiskers represent the 10% and 90% points * The blue line represents the mean quality |
| **Per tile sequence quality** | This graph will only appear in your analysis results if you're using an Illumina library which retains its original sequence identifiers. The plot shows the deviation from the average quality for each tile. The colors are on a cold to hot scale, with cold colors being positions where the quality was at or above the average for that base in the run, and hotter colors indicate that a tile had worse qualities than other tiles for that base. In the example below you can see that certain tiles show consistently poor quality. A good plot should be blue all over. |
| **Per sequence quality scores** | The per sequence quality score report allows you to see if a subset of your sequences have universally low quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged (on the edge of the field of view etc.), however these should represent only a small percentage of the total sequences. |
| **Per base sequence content** | Plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called |
| **Per sequence GC content** | This module 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 |
| **Per base N content** | If a sequencer is unable to make a base call with sufficient confidence, then it will normally substitute an N rather than a conventional base call.  This module plots out the percentage of base calls at each position for which an N was called. |
| **Sequence Length Distribution** | Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of wildly varying lengths. Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end.  This module generates a graph showing the distribution of fragment sizes in the file which was analyzed |
| **Sequence Duplication Levels** | In a diverse library most sequences will occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias (e.g. PCR over amplification).  This module 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. |
| **Overrepresented sequences** | A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected. |
| **Adapter Content** | The Kmer Content module will do a generic analysis of all of the Kmers in your library to find those which do not have even coverage through the length of your reads. This can find a number of different sources of bias in the library which can include the presence of read-through adapter sequences building up on the end of your sequences. |
| **Kmer Content** | The analysis of overrepresented sequences will spot an increase in any exactly duplicated sequences, but there are a different subset of problems where it will not work. |

* + 1. FASTQ files should then be processed through a read trimming and filtering software of your choice such as PrinSeq or Trimmomatic. The recommended initial trimming cutoff is Q=5. This value may vary but most bases below a quality score of 4 or lower have been shown to be erroneous. \*Note that NextSeq data usually need to be filtered with an increased quality cutoff of Q<15.
    2. Subsequent levels of trimming/filtering with increased stringency might be needed for your data. After each round of trimming, use your FASTQ QC software to determine whether this will be needed.
    3. Once reports indicate that a satisfactory level of trimming and filtering has been completed, proceed to SOP3 for assembly and further analysis.

1. **Method Performance Specifications**

N/A

1. **Calculations**

N/A

1. **Reference Values, Alert Values**

N/A

1. **Interpretation of Results**

Of the metrics shown above, key metrics that should be considered are listed below. Please keep in mind that these results are variable based on several factors including organism and workflows and thus should be interpreted within the context of expected values based on historical results.

**Per Base Sequence Quality –** This plot reflects the Q-score of raw reads as a box-plot for each cycle. Higher values are always better and generally a decay of quality can be observed in most runs.

**Per Base Sequence Content** – This plot reflects the proportion of each base at each cycle. Generally in a random fragment library from a genome you would expect to see all four bases equally represented. However, some genomes can be very GC biased and thus, this information should be compared against historical data.

**Duplicate Sequences –** This plot reflects the number of times the same sequence is seen in a 200,000 read subset of your sample data. Ideally, one should expect to see <10% duplicate reads. A high amount of duplicate sequences might suggest over-amplification or poor library-prep.

1. **Results Review and Approval**

Document the data quality metrics on the appropriate form or test record and obtain applicable reviews and approvals. *(Update this section to specify your laboratory’s applicable form/record and processes.)*

1. **Reporting Results; Guidelines for Notification**

N/A

1. **Sample Retention and Storage**

Store data in compliance with all applicable regulations, CDC records retention policy, and laboratory data storage procedures.*(Update to specify your laboratory’s data retention and storage policy)*

1. **References**
   1. Illumina Sequence Analysis Viewer v1.11 Part # 15066069 v03 February 2017
2. **Appendices** *(Include example screen shots of good and poor quality data applicable to your laboratory methods)*

FASTQC Screenshots:

Figure A-1. FASTQC (Sample Good Report)

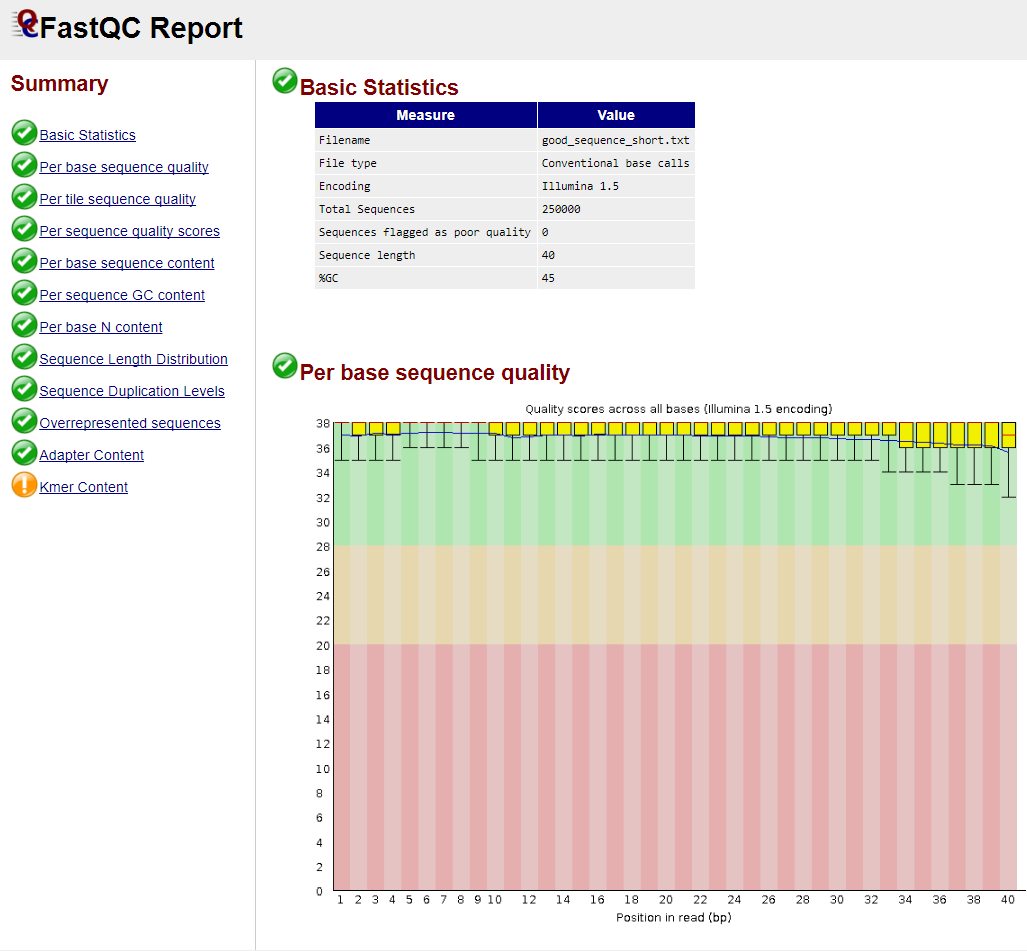
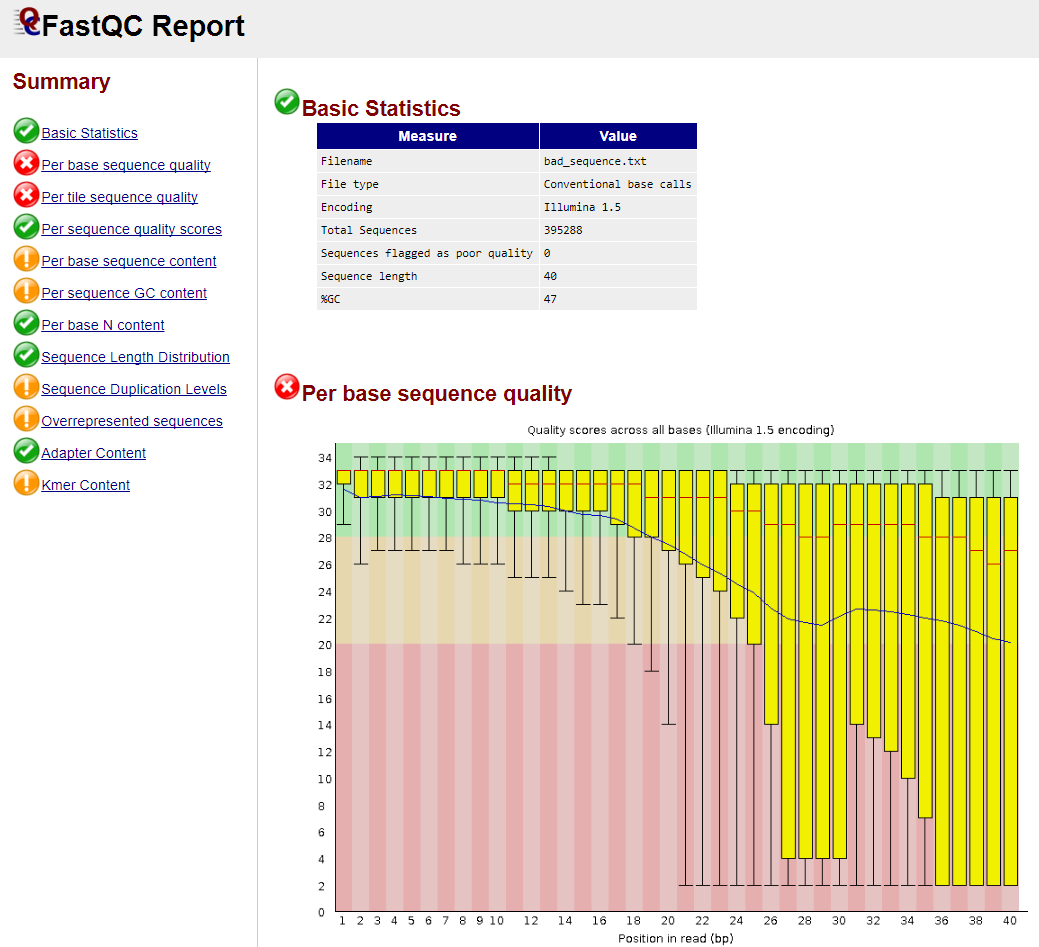


Figure A-2. FASTQC (Sample Bad Report)



1. **Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rev #** | **DCR #** | **Change Summary** | **Date** |
|  |  |  |  |

1. **Approval**

Approval Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_