Quality Control Analysis of NGS Data

FastQC Report

- You should quality control check your reads before starting <u>any</u> analysis
- FastQC is a quality control pipeline for raw sequence data coming from high throughput sequencing machines
- http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- FastQC takes a sequence read file (or files) and runs a series of QC analyses and generates a comprehensive graphical QC report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Each QC analysis is flagged as a **pass**, **warning** or **fail**

NOTE: A **warning** or **failure** for a QC analysis **do not necessarily mean that there is a problem with your data**, only that the results of the QC analysis exceeds the thresholds set by the programmer

Running FastQC

A command line program, which will generate an HTML report for each file you process.

```
Default
ag5431pi@labh02 [~] % ls
Sorghum_bicolor_transcript.fasta  leaf.fastq  leaf_gene_FPKM.txt  qc
ag5431pi@labh02 [~] % fastqc -o qc -f fastq leaf.fastq
Started analysis of leaf.fastq
Approx 5% complete for leaf.fastq
Approx 10% complete for leaf.fastq
Approx 15% complete for leaf.fastq
Approx 20% complete for leaf.fastq
```

Basic Statistics

Measure	Value
Filename	Os_nb_RNAseq.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000000
Filtered Sequences	0
Sequence length	100
%GC	53

The Basic Statistics module generates some simple composition statistics for the file analyzed

Filename: The original filename of the file which was analyzed

File type: Says whether the file appeared to contain actual base calls or colorspace data which had to be converted to base calls

Encoding: Says which ASCII encoding of quality values was found in this file.

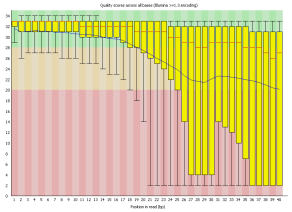
Total Sequences: A count of the total number of sequences processed. There are two values reported, actual and estimated. At the moment these will always be the same. In the future it may be possible to analyze just a subset of sequences and estimate the total number, to speed up the analysis, but since 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 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



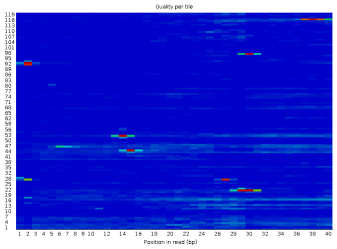
For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows:

- 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

Warning - a warning will be issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25

Failure - this module will raise a failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20

Per Tile Sequence Quality

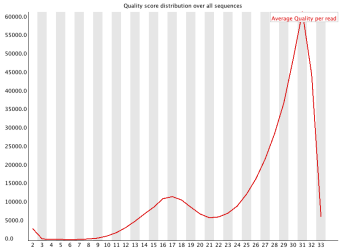


- This graph shows the quality scores from each tile across all of the bases in the read to see if there was a loss of quality associated with only part of the flowcell
- Issues can come from bubbles going through the flowcell, smudges on the flowcell, debris in the lane, etc.
- This graph is only available for Illumina data that retains original sequence identifiers.

Warning - a warning will be issued if any tile shows a mean Phred scoe more than 2 less than the mean of that base across all tiles

Failure - this module will raise a failure if any tile shows a mean Phred score more then 5 less than the mean of that base across all tiles

Per Sequence Quality Score

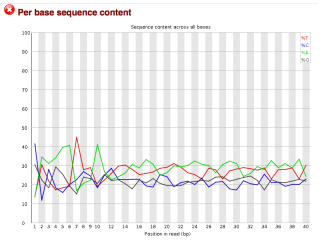


Allows you to see if a subset of your sequences have universally low quality values

Warning - a warning is raised if the most frequently observed mean quality is below 27 - this equates to a 0.2% error rate.

Failure - an error is raised if the most frequently observed mean quality is below 20 - this equates to a 1% error rate.

Per Base Sequence Content



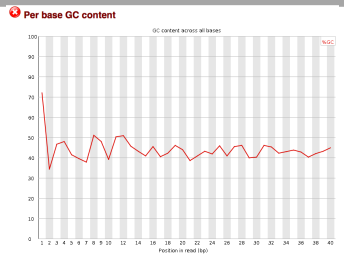
Proportion of each base position in a file for which each of the four normal DNA bases has been called

In a random library expect little to no difference between the different bases of a sequence run

Relative amount of each base should reflect the overall amount of these bases in your genome, transcriptome, etc

Warning - difference between A and T, or G and C is greater than 10% in any position
Failure - difference between A and T, or G and C is greater than 20% in any position

Per Base GC Content



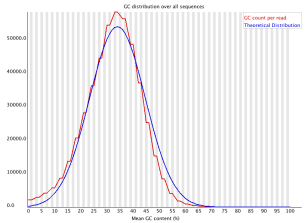
GC content of each base position in a file

Expect little to no difference between the different bases of a sequence run in a random library (horizontal line across the graph) and overall GC content should reflect the GC content of the underlying genome

Non horizontal line could indicate an overrepresented sequence which is contaminating your library

Warning - GC content of any base strays more than 5% from the mean GC content
Failure - GC content of any base strays more than 10% from the mean GC content

Per Sequence GC Content



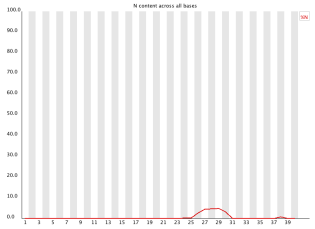
Measures the GC content across the whole length of each sequence in a file and compares it to a modeled normal distribution of GC content

Expect to see a roughly normal distribution of GC content in a random library

Unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset

Warning - sum of the deviations from the normal distribution represents more than 15% of the reads
Failure - sum of the deviations from the normal distribution represents more than 30% of the reads

Per Base N Content

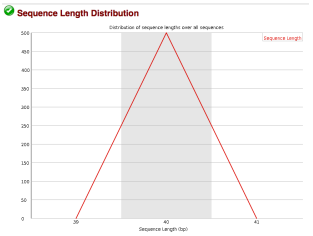


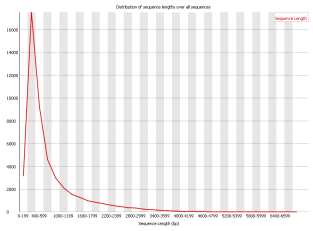
Percentage of base calls at each position for which an N was called

If sequencer is unable to make a base call with sufficient confidence it will normally substitute an N rather than a conventional base call

Warning - any position shows an N content of >5%
Failure - any position shows an N content of >20%

Sequence Length Distribution



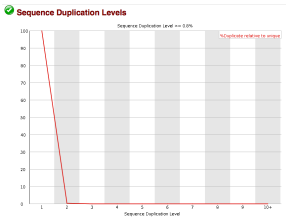


Distribution of fragment sizes in the sequence file

Warning - if all sequences are not the same length
Failure - any of the sequences have zero length

Which type of sequencing platforms would produce each of these graphs?

Duplicate Sequences



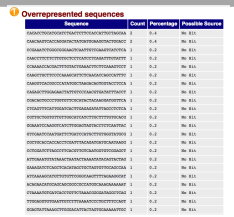
Counts the degree of duplication for every sequence in the set and creates a plot showing the relative number of sequences with different degrees of duplication

A high level of duplication indicate some kind of enrichment bias (eg PCR over amplification)

To reduce memory requirements the first 200,000 sequences are analyzed

Warning - non-unique sequences make up more than 20% of the total
Failure - non-unique sequences make up more than 50% of the total

Overrepresented Sequences



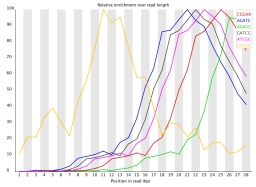
A single sequence that is very overrepresented in the set either means that it is highly biologically significant, or that the library is contaminated, or not as diverse as you expected

This module lists all of the sequence which make up more than 0.1% of the total. To reduce memory requirements the first 200,000 sequences are analyzed

Many adapter sequences are very similar to each other so you may get a hit reported which isn't technically correct, but which has very similar sequence to the actual match.

Warning – any sequence is found to represent more than 0.1% of the total
Failure – any sequence is found to represent more than 1% of the total

Overrepresented k-mers



Counts the enrichment of every 5-mer within the sequence library

Calculates an expected level at which this k-mer should have been seen based on the base content of the library as a whole and calculates an observed/expected ratio for that k-mer

To conserve time, only 20% of the whole library is analyzed

Warning - any k-mer is enriched more than 3 fold overall, or more than 5 fold at any individual position
Failure - any k-mer is enriched more than 10 fold at any individual base position

Read The Manual!

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

Index of /projects/fastqc/Help

Name	Last modified	Size	Description
Parent Directory		-	
1 Introduction/	2018-10-04 12:28	-	
2 Basic Operations/	2018-10-04 12:28	-	
3 Analysis Modules/	2018-10-04 12:28	-	

Running FastQC

Note % indicates terminal prompt

- What do our read files look like?
  % head sample1.fastq
  % head sample2.fastq

```
ssh
cnhirsch@ln0003 [/home/agro5431/cnhirsch/read_qc] % head sample1.fastq
@HISEQ06:196:C18Y1ACXX:6:1101:1542:2325 1:N:0:CTGCTG
TACTTGCTCATATGGCTTCCCGTTGTCCCATGTCTCCAAAGCAGCAAGCTCGTCATTGTGCTTCTCTATCAAATCAGCAA
ACCGCAGTAGGATACGGGACCTCTCATAGGCAGTCATCTTCGGCCATGGCCCTTCATCAAAAGCCTTGCG
+
@@@@FFFFFHHH<CEHHIJJJJIIHFGHHIJIIIJJJJCHIGJJGIJJJGHJJJJJJJJJJIJJIIIJJIFE?AC?;;?;
BAABB@BBDDDCDDC<?@D<BDDDDDECDCDDD<CC:ACCD(5>559:?<922<AC@##############
@HISEQ06:196:C18Y1ACXX:6:1101:1542:2325 2:N:0:CTGCTG
ATTAATGGGAACTTTGTTGATTCCGCATCCGGCAAGACCTTCCCAACTCTGGACCCTCGTACAGGGGAGGTGATTGCTCA
TGTGGCTGAGGGTGACGCAGAGGACATTAACCGTCCAGTAGCTGCGGCTCGCCAGGCTTTTGATGAAGGG
+
;;=DDDEFHHHHHJJJJJJJJJJJJJJJJJJJJJIIJJHHGIGIEIJIIJJIJIJJHEHDFFFBCDDD?BDDEDDDDD
CDCDDDDD?BDD2<ACBDDDDDB?B?CDDDD@@B&5B@CDCD@CDDD>BB>.&20<(29?C#########
@HISEQ06:196:C18Y1ACXX:6:1101:1540:2476 1:N:0:CTGCTG
CGCCAGAAGCGCCCCAGGAGGCCAGGGGAGGACACGACGCCGCCATCAGCCATCATCATCATCTCAGACTCAGAGCATGC
CATGGCACGCAACTGACTGACTGACTCGAAACACACACAGCGACGCCGCTATACTTTATTAAGCTTTATA
cnhirsch@ln0003 [/home/agro5431/cnhirsch/read_qc] %
```

Running FastQC

Note % indicates terminal prompt

- Looking at the help documentation for FastQC
  % fastqc -help
- Running FastQC from the command line
  % fastqc -f fastq sample1.fastq
  % fastqc -f fastq sample2.fastq

```
cnhirsch@ln1002 [/home/agro5431/cnhirsch/read_qc] % ls
fastqc_sample2.sh          sample1.fastq        sample2.fastq
fastqc_sample2.sh.e999084  sample1_fastqc.html  sample2_fastqc.html
fastqc_sample2.sh.o999084  sample1_fastqc.zip   sample2_fastqc.zip
cnhirsch@ln1002 [/home/agro5431/cnhirsch/read_qc] %
```

Interpreting FastQC

FastQC reports have been put on GitHub (see Day1_07 files)

- How many sequence reads are in each file?
- What are the lengths of the sequence reads?
- What is the quality of the reads from each sample?
- What overrepresented sequences are present?
- Do you notice any differences between the two samples?

Aggregating Results

Read QC Takeaways

- Perform QC analysis of your reads before starting any downstream analysis
- FastQC report provides valuable information to assess sequence quality
- Expectations are dependent on the sample and the assumptions used to assign pass/fail do not always apply
- Use a tool like Cutadapt or Trimmomatic to remove adapter contamination
- Take good notes of everything you do!!