The evaluation of data quality of the fatsq files is the first one of several Quality Control **(QC)** steps in the analysis of Next Generation Sequnecing (NGS) data. For such purpose we will use the software FastQC.

The process is quite simple:

1. Download and install FastQC in your local server following instructions. In detritus it can be found at `/opt/exoma/bin/fastqc`.
2. Create a directory where to save the outputs of FastQC, for example name it `fastqcRawdata`
3. Check the quality by typing: `$ fastqc -o fastqcRawdata *_1_sequence.fq.gz`
    1. the file does not need to be decompressed to run FastQc
4. This generates a folder for each file analyzed with several files:
    1. fastqc_data.txt - this contains the quality statistics in txt format.
    2. summary.txt - contains a summary of this file quality statistics in form of pass or not pass
    3. fastqc_report.html - same as before but it can be opened with `$ firefox fastqc_report.html` which allows viewing graphs
    4. Icons - folder with
    5. Images - folder with graphs as png

To understand the output, there is a nice explanatory video by Babraham Institute.

Example of running FastQC in one of our samples:

Bonn's fastq files are stored at directory: Bonn_0_fastq, under different folders according to its plate of origin, hence:

```
[vifehe@detritus bonn_data]$ ls Bonn_0_fastq/
P1_001-040     P1_041-080          P1_081-095
P2_001-040     P2_041-080          P2_081-095
P3_001-040     P3_041-080          P3_081-095
P4_001-040     P4_081-095          P4_041-080  P5_001-017
```

we create the directory where we will save FastQC output:

```
[vifehe@detritus Bonn_0_fastq]$ touch fastqcRawdata
```

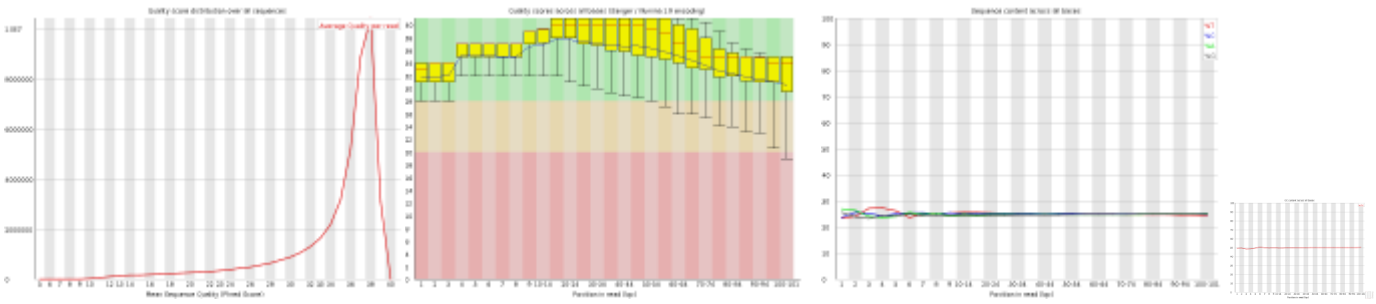
and we further create directories for each of the plates

```
[vifehe@detritus Bonn_0_fastq]$ cd fastqcRawdata
[vifehe@detritus fastqcRawdata]$ touch fastqcRawdata_P1 fastqcRawdata_P2
fastqcRawdata_P3 fastqcRawdata_P4

<code/>
# to run fastqc on a single file, return to folder where we have our vcf
files
[vifehe@detritus fastqcRawdata]$ cd ..
[vifehe@detritus Bonn_0_fastq]$ cd P1_001-040
#to see just the first two files
[vifehe@detritus P1_001-040]$ ls | head -n2
```

```
SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
SN7640211_14074_P1A01_MND1014_2_sequence.fq.gz
[vifehe@detritus P1_001-040]$ fastqc -o ../fastqcRawdata/fastqcRawdata_P1
Started analysis of SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Started analysis of SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz  #
started at 13:57
Approx 5% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 10% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 15% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 20% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 25% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 30% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 35% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 40% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 45% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 50% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 55% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 60% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 65% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 70% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 75% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 80% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 85% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 90% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 95% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Approx 100% complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
Analysis complete for SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz #
finished at 14:09
# the process per sample takes about 12 minutes
#this can be run in a loop
[vifehe@detritus P1_001-040]$  for x in P1_001-040/*.gz; do fastqc -o
../fastqcRawdata/fastqcRawdata_P1/ $x; done
# to examine the output
[vifehe@detritus P1_001-040]$ cd ../fastqcRawdata/fastqcRawdata_P1

#the program has created a folder named like the sequence and another
compressed folder
[vifehe@detritus fastqcRawdata_P1]$ ls | head -n2
SN7640211_14074_P1A01_MND1014_1_sequence.fq_fastqc
SN7640211_14074_P1A01_MND1014_1_sequence.fq_fastqc.zip

#list the contents of the folder created
[vifehe@detritus fastqcRawdata_P1]$ cd
SN7640211_14074_P1A01_MND1014_1_sequence.fq_fastqc
[vifehe@detritus SN7640211_14074_P1A01_MND1014_1_sequence.fq_fastqc]$ ls
fastqc_data.txt  fastqc_report.html  Icons  Images  summary.txt

#examine Summary.txt output
[vifehe@detritus SN7640211_14074_P1A01_MND1014_1_sequence.fq_fastqc]$ cat
summary.txt
```

```
PASS    Basic Statistics    SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
PASS    Per base sequence quality
SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
PASS    Per sequence quality scores
SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
PASS    Per base sequence content
SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
PASS    Per base GC content
SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
WARN    Per sequence GC content
SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
PASS    Per base N content    SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
PASS    Sequence Length Distribution
SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
WARN    Sequence Duplication Levels
SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
PASS    Overrepresented sequences
SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
PASS    Kmer Content    SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz

# examine fastqc_data.txt output
[vifehe@detritus SN7640211_14074_P1A01_MND1014_1_sequence.fq_fastqc]$ more -
n20  fastqc_data.txt
##FastQC    0.10.1
>>Basic Statistics    pass
#Measure    Value
Filename    SN7640211_14074_P1A01_MND1014_1_sequence.fq.gz
File type   Conventional base calls
Encoding    Sanger / Illumina 1.9
Total Sequences 44012752
Filtered Sequences  0
Sequence length 101
%GC 49
>>END_MODULE
>>Per base sequence quality   pass
#Base   Mean    Median    Lower Quartile    Upper Quartile    10th
Percentile    90th Percentile
1   31.64506284451379    33.0    31.0    34.0    28.0    34.0
2   31.880190722906853    34.0    31.0    34.0    28.0    34.0
3   31.972653289210363    34.0    31.0    34.0    28.0    34.0
4   35.39369340049448    37.0    35.0    37.0    32.0    37.0
5   35.09201710449735    37.0    35.0    37.0    32.0    37.0
6   35.08697933726116    37.0    35.0    37.0    32.0    37.0
7   35.06162818448617    37.0    35.0    37.0    32.0    37.0
....
....
....
>>Sequence Duplication Levels warn
#Total Duplicate Percentage 33.90859348891959
#Duplication Level  Relative count
1   100.0
```

```
2   29.769972680482393
3   10.634235430848415
4   4.525832792477321
5   1.99009856157457
6   1.1335335758380753
7   0.6905347682369101
8   0.447526904442063
9   0.296590343078804
10++    1.4482363062804704
>>END_MODULE
>>Overrepresented sequences   pass
>>END_MODULE
>>Kmer Content    pass
>>END_MODULE


# Explore html file
[vifehe@detritus SN7640211_14074_P1A01_MND1014_1_sequence.fq_fastqc]$
firefox fastqc_report.html
# this opens the file in firefox in which the following pictures can be seen
```



Because examining each file is time consuming, I've created a couple of scripts with which we can extract the information of our interest:

script1 - to summarize output from summary.txt

script2 - to summarize output from fastqc_data.txt