Assessing Illumina Data Quality

It is always good practice to assess the quality of your Illumina reads before using the data in downstream analyses. This is true for both single or paired-end reads, no matter the run length. At GGBC we use the open source software package **FastQC**, an easy to use tool to accomplish this task.

FastQC was developed at Babraham Bioinformatics and is available for free download here https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ FastQC is Java based so it will run on PC, Mac or Linux operating systems, if you want to run it on your local computer. Otherwise, is also installed on the Sapelo2 cluster.

Illumina data downloaded from BaseSpace will be in .fastq.gz format. There is no need to decompress the file (gunzip) prior to running FastQC, since it will accept .fastq or .fastq.gz inputs.

Running FastQC interactively on Sapelo2 is done as follows:
First, navigate to the directory containing your raw data files.
Next, create a directory where the results will be written, e.g. **FQC_out**
Finally, load the required module and run the following script.

```
ml FastQC
fastqc *.fastq.gz -o FQC_out
```

After completion, the FQC_out dir. will contain .html and zip files prefixed with each sample name. The .zip file contains the same results found in the .html file, but in text format. The .html files can be imported to your local computer for easy viewing in a browser.

The FastQC website documents the multiple graphical outputs found in the results with examples of good and bad data. A few of the more important ones to pay attention to are listed below:

1) **Basic statistics** gives you the total number of reads in the samples and basic sequence metrics.

2) **Per base sequence quality** shows the average Phred quality score over the total length of the read. In most cases the Phred score is typically around 30. Some drop off is expected at the beginning and end of the reads with the latter being more pronounced in longer reads e.g. PE 150. Reverse read quality in PE data is typically slightly less that the forward read. Note that the Phred score is logarithmic so, for example, a score of 20 means that there is a 1/100 chance that the base call is in error. Average quality scores for short read data generated at GGBC are typically > 30.

3) **Per sequence quality scores** shows the mean quality distribution across all reads.

4) **Per sequence GC** should be a normal distribution with the peak mean GC content being what is expected for the organism being sampled. If there is a shoulder or multiple peaks, that is a good indication that the sample may be contaminated with RNA/DNA from more than one source.

5) **Overrepresented sequences**. This will often be blank. If not, the sequence of the overrepresented read will be displayed along with it's count and percentage. If the sequence is a known adapter or index, FastQC will identified it as such. Unknown sequences will be marked as "No Hit." A quick way to identify these unknown reads is to simply copy and paste the sequence string into the nucleotide BlastN search engine. In RNA-Seq data, these will often be abundant "house-keeping" transcripts or,

in libraries not synthesized with oligo-dT priming, they may be ribosomal, mitochondrial or chloroplastic sequences.

Quality Trimming Raw Data

Once the quality of the data has en assessed with FastQC, there are a number of quality trimming programs available for download, and some are already installed on the Sapelo2 cluster. For example, the Java-based Trimmomatic and Python-based Cutadapt are two of the most commonly used programs currently installed on our cluster.

Trimmomatic is very easy to use software. For most cases, the script below can be used to efficiently clean paired end, short read data generated by GGBC with TruSeq3 chemistry where MINLEN is the minimum length of a trimmed read that will be retained. Descriptions of the other flags/settings used can be found in the Trimmomatic manual and only slight modification for single read data is required.

```
ml Trimmomatic
java -jar /apps/eb/Trimmomatic/0.39-Java-1.8.0_144/trimmomatic-0.39.jar PE -threads 2
samplename_R1.fq.gz samplename_R2.fq.gz \
samplename_R1.paired.fq samplename_R1.unpaired.fq \
samplename_R2.paired.fq samplename_R2.unpaired.fq \
ILLUMINACLIP:/apps/eb/Trimmomatic/0.39-Java-1.8.0_144/adapters/TruSeq3-PE-2.fa:2:30:10
LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:30
```

In the example above, there will be four output files generated for each PE sample. Here, they are given **.paired** and **.unpaired** extensions, but naming conventions are up to the user. Almost all trimmed reads will usually be paired outputs, which can then be used for downstream mapping, RNA-Seq, assembly etc… The unpaired read data can be also be used used in conjunction with paired data, where its use is required and/or allowed based on the downstream software being employed.