Margus Veanes
Uppsala University, 2019\
Genome Analysis, Final Report

FastQC (illumina QC)

- **What is the structure of a FASTQ file?**
  FASTQ files are text files that contain sequence in text form.
- **How is the quality of the data stored in the FASTQ files?**
  FASTQ files contain Phred score for each base. The range of the score values is typically between 2-40, with higher scores representing higher quality base reads.
- **How are paired reads identified?**
  Paired reads are identified within the 'id' field of the file denoting which file the reads are from. You will have multiple files for paired reads.

Trimmomatic (quality improvement)

- **How is the quality of your data after trimming?**
  The lower the data quality, the more gets trimmed.
- **What do the LEADING, TRAILING and SLIDINGWINDOW options do?**
  Leading and trailing are parameters which tell the software how many bases to cut when under a quality threshold, while the sliding window parameter identifies the length of the sequence the software will calculate average scores for. Once the quality within the window drops below 15 the software cuts the sequence.

CANU (genome assembly)

- **How many contigs do you expect? How many do you obtain?**
  8 contigs were expected and 8 were obtained.
- **What is the difference between a 'contig' and a 'unitig'?**
  A unitig is a high confidence contig.
- **What is the difference between a 'contig' and a 'scaffold'?**
  A scaffold is an ordered and oriented set of one or more contigs, with distances assigned to gaps between contigs.
- **What are the kmers? What kmer(s) should you use? What are the problems and benefits of choosing a small kmer? And a big kmer?**
  Kmers are nucleotides of a certain length. Shorter kmer means more contigs with lots of possible connections. Large kmers result in longer contigs with fewer connections. Depending on complexity of the assembled region, larger kmers can be more useful for resolving complex areas. Working with short reads or lower read depths, shorter kmers work better.

QUAST (quality control)

- **What do measures like N50, N90, etc. mean? How can they help you evaluate the quality of your assembly? Which measure is the best to summarize the**

quality of the assembly (N50, number of ORFs, completeness, total size, longest contig ...)

N50 represents the length for which the collection of all contigs of that length or longer cover at least half of the assembly. N90 represents the same thing as N50, but for the length of contigs that cover at least 90% of the assembly.

- **How does your assembly compare with the reference assembly? What can have caused the differences?**

  Differences are due to a mistake during initial genome assembly construction with CANU which was incorrectly provided a genome size parameter of 2.8m instead of 3.1m. Therefore, my assembly only contains the core genome (2.75m) and none of the plasmids (~300k).

PROKKA (annotation)

- **What types of features are detected by the software? Which ones are more reliable a priori?**

  Prokka finds and annotates features like protein coding regions and RNA genes. The more reliable features are those for which there exist prior annotation.
- **How many genes are annotated as 'hypothetical protein'? Why is that so? How would you tackle that problem?**

  Several hundred are hypothetical. This is due to these proteins lacking existing annotation.

BWA (mapping)

- **Do you see big differences between replicates?**

  There are small differences between replicates, but in general the coverage appears similar. These small differences are overshadowed by the differences observed across treatments.

SAMtools (format changes)

- **What is the structure of a SAM file, and how does it relate to a BAM file?**

  SAM files consist of a header and an alignment section. The BAM file is the equivalent of a SAM file, but in binary format.
- **What is the structure of vcf and bcf files?**

  VCF is a text file format that contains meta-information lines, a header line, and data lines which contain genomic positional information. BCF files are an open XML file format.

DESeq2 (expression analysis)

- **How do the different samples and replicates cluster together?**

  Replicate counts clustered together more than counts from different samples (treatments) and this makes logical sense as replicates from the same treatment would be expected to be roughly similar.

- **What effect and implications has the p-value selection in the expression results?**

  The p-value indicates the confidence that the observed differences in expression across treatments are caused by chance, therefore a lower value is desirable as the difference is unlikely to be caused by a random chance.

- **What is the q-value and how does it differ from the p-value? Which one should you use to determine if the result is statistically significant?**

  P-values indicate the likelihood of a false positive. Q-values are determined from p-values that are adjusted by an optimsed FDR approach. The FDR approach involves using characteristics of p-value distributions to produce a list of q-values.

- **What would you do to increase the statistical power of your expression analysis?**

  More replicates or more extensive input data from a source with greater read depth.