# Answers to Deliverable V: Class Room Quiz: NGS I

1. What is the difference between the SAM and BAM file format standard? Which one should we generally use and why?

Sequence Alignment Map (SAM) is a text-based format originally for storing biological sequences aligned to a reference sequence. The binary equivalent of a SAM file is a Binary Alignment Map (BAM) file, which stores the same data in a compressed binary representation.

We should use the BAM file format when possible.

Often the SAM/BAM files contain a lot of information and end up as quite large files. Since the BAM file is compressed it takes up less storage.

1. How many lines is in a fastq file? What does each of the four lines contain? (2 point)

Four lines...I literally gave you the answer in the next line

Line 1: @sequence identifier

Line2: raw sequence

Line3: + (seldomly also the sequence identifier

Line4: Sequence quality score. Must (obviously) contain the same number of scores as letters in the raw sequence

1. Why is it so important to know the exact type of encoding (Sanger, Solexa etc.)? What does a Phred quality score of 20 mean? (2 point)

Same symbol can mean several different quality scores depending on the encoding used.

Phred Quality Score Probability of incorrect base call Base call accuracy

10 1 in 10 90%

20 1 in 100 99%

30 1 in 1000 99.9%

40 1 in 10,000 99.99%

50 1 in 100,000 99.999%

60 1 in 1,000,000 99.9999%

1. Mention two reasons for trimming reads (2 point)

Remove adapters and/or barcodes

non-biological signal at 5’ end of reads caused by library creation kit or sequencing artifact

Bad quality is detrimental to both read mapping and *de novo* assembly

1. How does one check a 16s rRNA amplicon sample for adequate sequencing depth? (1 point)

Rarefaction plot

1. How does one check a shotgun metagenome sample for sequencing depth? (1 point)

Nonpareil. Nonpareil uses the redundancy of the reads in a metagenomic dataset to estimate the average coverage and predict the ammount of sequences that will be required to achieve "nearly complete coverage".

1. Why should we check for sequencing depth in a metagenomic study? (2 point)

We can see how much sequence is needed to describe an entire microbiome, thus avoiding over-sequencing. We can also give an honest estimate for how descriptive our dataset really is.

1. Name 2 of the 3 major types of sequencing read categories that one can encounter? Additionally, write down the read directions of those two (directions drawings are sufficient): (3 points)

Single-end reads

Paired-end reads

Mate pairs




1. Write how you would make a directory called Yucky\_samples and copy the files yucky\_biofilm\_R1.fastq & yucky\_biofilm\_R2.fastq to this directory (2 point)

mkdir Yucky\_samples

scp yucky\_biofilm\_R\*.fastq Yucky\_samples

1. Mention a program one could use for having a first look at raw sequencing reads (1 point)

fastqc gives the most information

head gives you speed

less is maybe overkill

Word would crash and burn...these files are large

1. Mention programs one could use to trim adapters from raw reads (1 point)

AdapterRemoval, cutadapt or bbduk