Practical session 1

Designing a sequencing experiment

An infection outbreak has been detected in a hospital. This has been caused by a bacterium accumulating 55 cases up to date, all affecting immunocompromised patients. Cases are increasing rapidly and the origin of the infection remains unknown. These **55 strains** have been isolated, but it has not been possible to identify it taxonomically. The organism is resistant to multiple antibiotics.

Given the severity of the situation, the medical team urgently requests the sequencing of these strains **(55)** for taxonomic identification, study of virulence profiles and characterization of antibiotic resistance genes.

You are the bioinformatician of your team, and you are in charge of designing the best sequencing strategy for these bacterial strains. **Note: The expected genome length is 5Mb**

1. Based on the above:

   a) What is the required output (Mb) for sequencing all these strains? **Note: Consider a 50X depth.**

   Answer: Number of genomes * Genome length * Sequencing depth = 13750 Mb or 13,75 Gb

   b) Indicate and justify the sequencing platform you would propose as the **fastest option** for sequencing these strains (See Fig 1 and 2).

   Answer: Illumina MiSeq because it meets the required output (15Gb), and we can sequence the 55 samples in just one sequencing run.

   c) Describe the technical characteristics of the suggested sequencing equipment mentioning elements such as maximum sequencing output, read length, analysis time, etc. **Is this sequencing technology the best option? Why?**

   Answer: The Illumina MiSeq has an output of 15Gb with a read length of 2x300 bp and the sequencing can be carried out in two days. The MiSeq could be the best option since the other option (considering de number of strains) is the Illumina HiSeq platform. It has a bigger output which will exceed the needed output for this sequencing experiment.

   d) What **sequencing depth** would be **necessary to achieve the required objectives** -considering the urgency to obtain the data?

Answer: Normally, the recommended sequencing depth per genome is 100X. However, in this case we just need a taxonomic classification and resistance genes identification, which can be achieved with lower sequencing depth (~50X). Using this sequencing depth, could allow us to obtain a genome draft usable for taxonomy and resistance genes identification.

2. The outbreak persists, and information on the origin of the pathogen and how it is spreading throughout the hospital is urgently needed. To this end, a phylogenetic and evolutionary study will be carried out to identify different genetic variants among the population and the informative mutation sites to try to reconstruct the dispersion of the populations through the different areas across the hospital. With this strategy, we seek to identify the common ancestor of the entire population and to infer the most probable place and date of introduction of the pathogen in the hospital.

a) Describe how the sequencing strategy should change in terms of depth and accuracy in relation to the type of analysis requested. **What sequencing depth should you use?**

Answer: In this case we need to carry out a genomic and phylogenetic analysis which require a very good assembly for bioinformatic analyses. For this, the sequencing depth should be increased to 100X to benefit the assembly process and decrease the gaps along the assembled genome.

b) For the proposed depth, choose one sequencing platform (Fig 1 and 2) and indicate the maximum number of strains that can be sequenced per assay. How many runs will you need to carry out to complete the sequencing of the 55 strains?

Answer: Sequencing output required for 100X sequencing depth is 27.5 Gb.

Sequencing output= Number of genomes * Genome length * Sequencing depth = 27.5 Gb.

Therefore, the best option could be Illumina HiSeq (Sequencing output: 45Gb) and we will need just one sequencing run to complete the analysis.

c) Suppose that you have chosen the Illumina MiSeq platform using a 150 bp single end library. Indicate the total number of reads that you will obtain per sequencing run.

Illumina MiSeq output: 15Gb
Number of reads = Sequencing output/read length = 100 Millions of reads.

d) What number of reads do you expect to obtain for each sequenced genome?

Answer: Reads per genome = 100 millions of reads / 55 genomes = 1,818,181 reads per genome

e) What amount of data do you expect to obtain for each sequenced genome? Express the amount in Mb (million bases).

Answer:

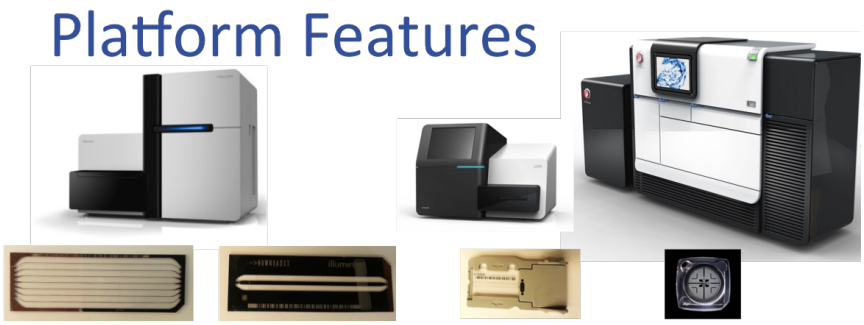
Total data per genome = number of reads * read length = 272,7Mb

2. Considering the sequencing proposal developed in point 2: how many genomes could be sequenced if the depth is lowered to half the value proposed in point 2a? What effects could this change bring in terms of bioinformatic analysis of the data? Explain and justify in detail.

   Answer: Considering the HiSeq output is 45Gb, and a that the proposed sequencing depth in question 2a was 100X the half is 50X. Thus, we could sequence a total number of 180 genomes in one run of a HiSeq sequencer.

   Number of genomes = Sequencer output / data needed per genome = 180 Genomes.

   This decrease on the sequencing depth will directly affect the assembly process which will probably produce a very fragmented assembly. This will also have an effect the downstream genomic and phylogenetic analysis.



Feature	HiSeq2500 - Highoutput	HiSeq2500 – Rapid mode	MiSeq	PacBio RSII
Number of reads	150-180M/lane	100-150M/lane	12-15M (v2) 20-25M (v3)	50-80K/SMRT cell
Read length	2 x 100 bp	2 x 150 bp	2 x 300 bp (v3)	~ 10-20 kb
Yield per lane (PF data)	up to 35 Gb	up to 45Gb	up to 15 Gb	up to 0.4 Gb
Instrument Time	~12-14 days	~2 days	~2 days	~2 hours
Pricing per Gb	$59 (PE100)	$53 (PE150)	$108 (PE300)	$697

Fig 1. Characteristic of Illumina and Pacbio sequencing platforms

Sequencing Platform	Read Length (bp)	Accuracy	Output	Sequencing Chemistry	Run Time	Advantages in Metabarcoding approaches
Sanger	400–900	99.999%	1.9–84 Kb	Dideoxy chain termination	20 min –3 h	Long read length, high quality
Illumina MiSeq	75–300	99.9%	13.2–20 Gb	Sequencing by Synthesis	21–56 h	High Throughput, read quality
MinION	>200,000	~95%	~50 Gb	Single Sequencing real time-long reads	1–48 h	High Throughput, Long read length, portability
PacBio	10–15 Kb	99.999	5–10 Gb	Single Sequencing real time-long reads	4 h	Long read length and quality

Fig 2. Comparison of Nanopore, illumina MiSeq and Pacbio sequencing platforms.