Block 1. Exit ticket 1

1. **Sanger sequencing**

1.1. Regarding to template amplification in Sanger sequencing it is true:

a- It can be done by cloning

1.2 How does the Sanger sequencing reaction works?

b- Nucleotide addition

1.3 Most representative features of Sanger sequencing is/are:

d- A and B → (a- Produce high quality reads; b- Low throughput)

1.4 Read the descriptions and figures below. What description and figure belong to Sanger sequencing reaction?

b- description F and figure 7

2. **Illumina sequencing**

2.1. Regarding to template amplification in Illumina sequencing it is true:

c- it is carried out by solid-phase bridge amplification

2.2 How does the Illumina sequencing reaction works?

a- Cyclic reversible termination

2.3 The most representative features of Illumina sequencing is/are:

e- All are correct → (a- High quality reads; b- High throughput; c- Short reads; d- A and C)

2.4. Read the descriptions and figures below. What description and figure belong to Illumina sequencing reaction?

d- description H and figure 1

3. **Nanopore sequencing**

3.1. Regarding to template amplification in Nanopore sequencing it is true:

d- None of the above → (a- It can be done by cloning; b- It is carried out by emulsion PCR; c- it is carried out by solid-phase bridge amplification)

3.2 How does the Nanopore sequencing reaction works?

c- Real time long read sequencing

3.3 Most representative features of Nanopore sequencing are:

c- Real time and portable long read sequencing

3.4. Read the descriptions and figures below. What description and figure belong to Nanopore sequencing reaction?

a- description E and figure 4

4. **PacBio sequencing**

4.1. Regarding to template amplification in PacBio sequencing it is true:

d- None of the above → (a- It can be done by cloning; b- It is carried out by emulsion PCR; c- it is carried out by solid-phase bridge amplification)

4.2 How does the PacBio sequencing reaction works?

c- Real time long read sequencing

4.3 Most representative features of PacBio sequencing is/are:

d- A and C → (a- Real time sequencing; c- Long reads)

4.4 Read the descriptions and figures below. What description and figure belong to PacBio sequencing reaction?

a- description G and figure 6

Descriptions

A. (454/Roche — Pyrosequencing, Fig 5) After bead-based template enrichment, the beads are arrayed onto a microtitre plate along with primers and different beads that contain an enzyme cocktail. During the first cycle, a single nucleotide species is added to the plate and each complementary base is incorporated into a newly synthesized strand by a DNA polymerase. The by-product of this reaction is a pyrophosphate molecule (PPi). The PPi molecule, along with ATP sulfurylase, transforms adenosine 5' phosphosulfate (APS) into ATP. ATP, in turn, is a cofactor for the conversion of luciferin to oxyluciferin by luciferase, for which the by-product is light. Finally, apyrase is used to degrade any unincorporated bases and the next base is added to the wells. Each burst of light, detected by a charge-coupled device (CCD) camera, can be attributed to the incorporation of one or more bases at a particular bead.

B. (ePCR, Fig 8: 454, SOLiD, Ion Torrent) Fragmented DNA templates are ligated to adapter sequences and are captured in an aqueous droplet (micelle) along with a bead covered with complementary adapters, deoxynucleotides (dNTPs), primers and DNA polymerase. PCR is carried out within the micelle, covering each bead with thousands of copies of the same DNA sequence.

C. (DNA preparation, Fig 9) Cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented DNA may then be cloned into a DNA vector and amplified in a bacterial host such as Escherichia coli.

D. (Solid-phase bridge amplification, Fig 3: Illumina) Fragmented DNA is ligated to adapter sequences and bound to a primer immobilized on a solid support (patterned flow cell). The free end can interact with other nearby primers, forming a bridge structure. PCR is used to create a strand from the immobilized primers, and unbound DNA is removed.

E. (NANOPORE SEQUENCING, Fig 4) DNA is initially fragmented to 8–10 kb. Two different adapters, a leader and a hairpin, are ligated to either end of the fragmented dsDNA. There is no method to direct the adapters to a particular end of the DNA molecule, so there are 3 possible library conformations: leader–leader, leader–hairpin and hairpin–hairpin. The leader adapter is a ds adapter containing a sequence required to direct the DNA into the pore and a tether sequence to help direct the DNA to the membrane surface. Without leader adapter, there is minimal interaction of the DNA with the pore, which prevents any hairpin–hairpin fragments from being sequenced. The ideal library conformation is the leader–hairpin. In this conformation the leader sequence directs the DNA fragment to the pore with current passing through. As the DNA translocates through the pore, a characteristic shift in voltage through the pore is observed. Various parameters (magnitude and duration of the shift) are recorded and can be interpreted as a particular k-mer sequence. As the next base passes into the pore, a new k-mer modulates the voltage and is identified. At the hairpin, the DNA continues to be translocated through the pore adapter and onto the complement strand. This allows the forward and reverse strands to be used to create a consensus sequence called '2D' read.

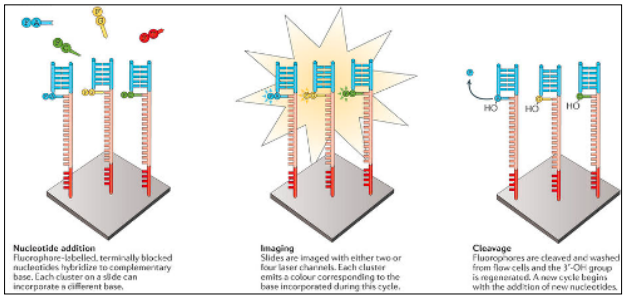
F. (SANGER SEQUENCING, Fig 7) Sequencing uses ddNTPs (dideoxynucleotide triphosphates) which do not have a free 3’ OH mixed in with dNTPs. Whenever the DNA polymerase incorporates a ddNTP it won’t be able to add any other nucleotides. Then gel electrophoresis is used to separate the DNAvv.

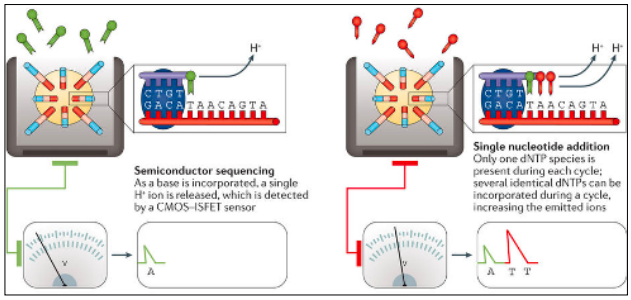
G. (PACBIO SEQUENCING, Fig 6) Template fragments are processed and ligated to hairpin adapters at each end, resulting in a circular DNA molecule with constant ssDNA regions at each end with the dsDNA template in the middle. The resulting 'SMRTbell' template undergoes a size-selection protocol in which fragments that are too large or too small are removed to ensure efficient sequencing. Primers and an efficient ϕ29 DNA polymerase are attached to the ssDNA regions of the SMRTbell. The prepared library is then added to the zero-mode waveguide (ZMW) SMRT cell, where sequencing can take place. To visualize sequencing, a mixture of labelled nts is added; as the polymerase-bound DNA library sits in one of the wells in the SMRT cell, the polymerase incorporates a fluorophore-labelled nt into an elongating DNA strand. During incorporation, the nt momentarily pauses through the activity of the polymerase at the bottom of the ZMW, which is being monitored by a camera.

H. (ILLUMINA SEQUENCING, Fig 1) After solid-phase template enrichment, a mixture of primers, DNA polymerase and modified nts are added to the flow cell. Each nt is blocked by a 3’-O-azidomethyl group and is labelled with a base-specific, cleavable fluorophore (F). During each cycle, fragments in each cluster will incorporate just one nt as the blocked 3’ group prevents additional incorporations. After base incorporation, unincorporated bases are washed away and the slide is imaged by total internal reflection fluorescence (TIRF) microscopy using either two or four laser channels; the colour (or the lack or mixing of colours in the two-channel system used by NextSeq) identifies which base was incorporated in each cluster. The dye is then cleaved and the 3’-OH is regenerated with the reducing agent tris(2-carboxyethyl)phosphine (TCEP). The cycle of nt addition, elongation and cleavage can then begin again.

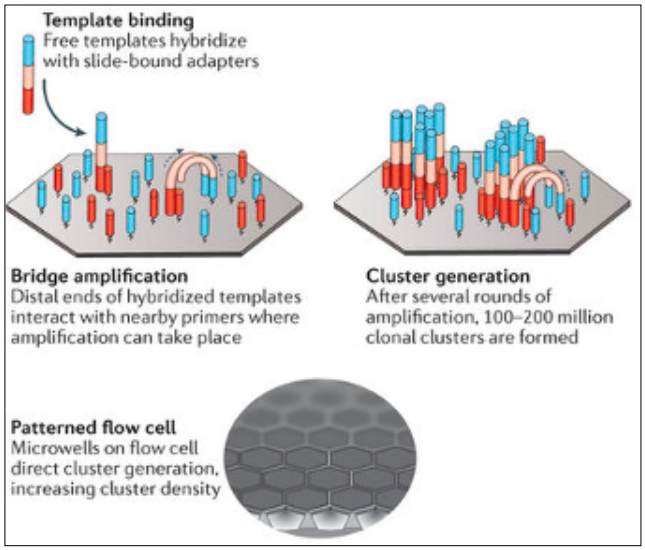
I. (Ion Torrent sequencing, Fig 2) After bead-based template enrichment, beads are carefully arrayed into a microtitre plate where one bead occupies a single reaction well. Nucleotide species are added to the wells one at a time and a standard elongation reaction is performed. As each base is incorporated, a single H+ ion is generated as a by-product. The H+ release results in a 0.02 unit change in pH, detected by an integrated complementary metal-oxide semiconductor (CMOS) and an ion-sensitive field-effect transistor (ISFET) device. After the introduction of a single nucleotide species, the unincorporated bases are washed away and the next is added.

Images Illumina sequencing

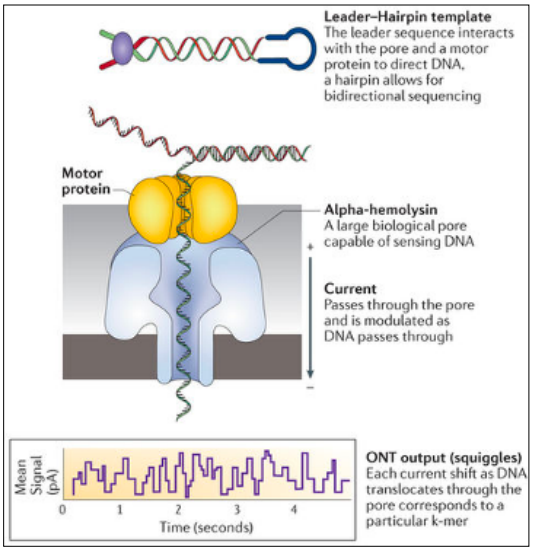


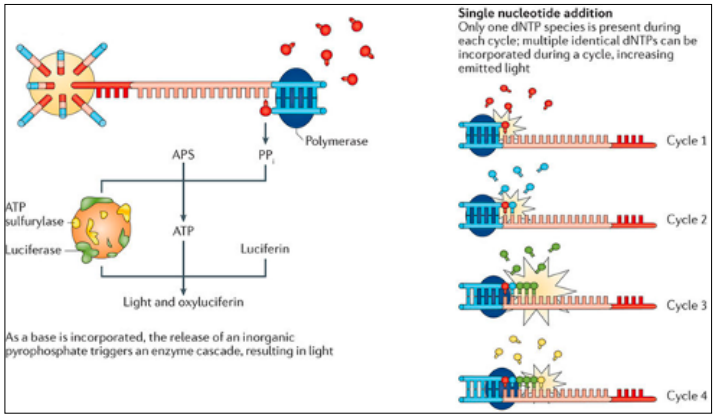


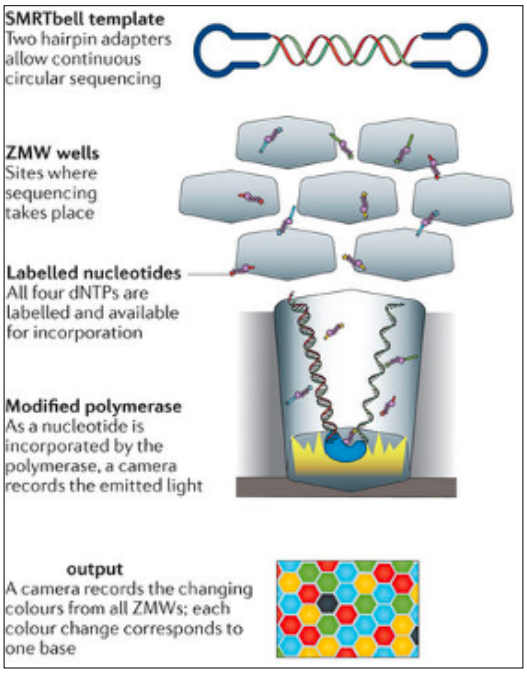
Oxford Nanopore sequencing



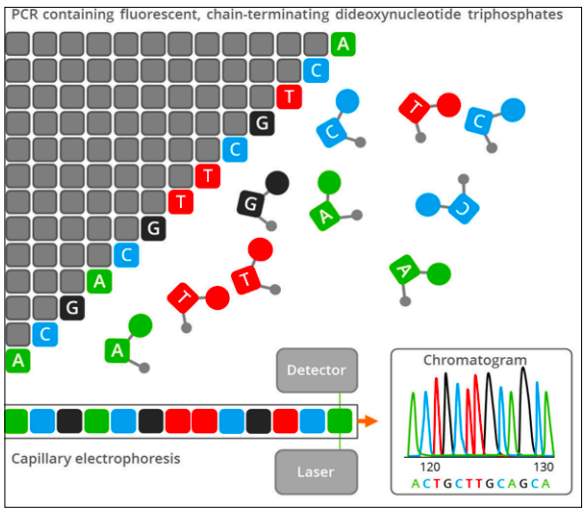
D



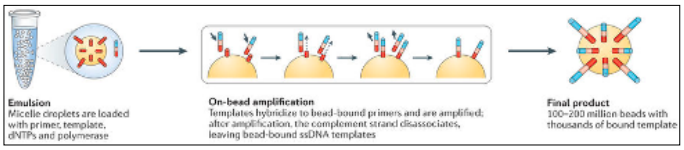




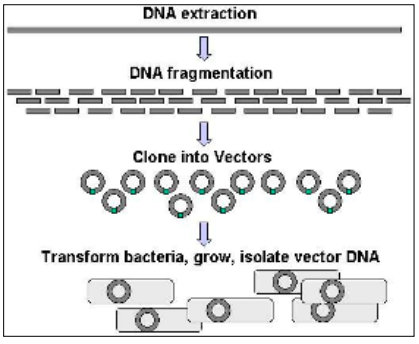
PacBio sequencing



Sanger sequencing



B



C

When reads are wanted, convert with 10^9

Practical Exit ticket 1

Your work team has the mission of sequencing a collection of 30 bacterial strains that have caused an outbreak of nosocomial infections. The sequencing of these strains aims to taxonomically identify the strains, study their virulence profiles and antibiotic resistance mechanisms.

Based on the above, answer the following questions:

1. Indicate and justify the sequencing platform you would propose as the best option to sequence these strains.

d- Illumina MiSeq using a paired-end library and 250bp read length

2. If the genomes that need to be analyzed have an average genome lenght of 7.3Mb, what sequencing output do you need to sequence the 30 complete genomes with a depth of 100X? Indicate the result in Gb.

21.9 Gb

7.3 x 100 x 30= 21 900 Mb
21 900 Mb/1 000 = 21.9 Gb

Sequencing output = Genome size x Coverage depth x # of genomes

3. What number of reads do you need to obtain if you sequence <u>all the 30 genomes</u> at 100X depth? Consider a read size of 150bp and a genome size of 7.3Mb.

146,000,000 reads

(((7.3/1000) x 100 x 30) / 150) x 10^9 = 146,000,000 (million) reads

1 billion (10^9) = 1 Gb

Number of reads = (Genome size (Gb) x Coverage depth x Number of genomes) / Read length

4. What number of reads are needed to sequence a <u>single complete genome</u> at 100X? Consider a reads size of 125bp and a genome size of 7.3Mb.

5,840,000 reads

(((7.3/1000) x 100 x 1) / 125) x 10^9 = 5,840,000 (million) reads

1 billion (10^9) = 1 Gb

Number of reads = (Genome size (Gb) x Coverage depth x Number of genomes) / Read length

5. a) If the sequencer can generate a maximum of 20Gb of information per sequencing run. What is the maximum number of 5Mb genomes that can be sequenced at 100X depth?

40 genomes

(20) / ((5/1000) x 100) = 40 genomes

Number of genomes = (Sequencing output per run) / (Genome size (Gb) x Coverage depth)

b) Considering a sequencing output of 20Gb, how many reads of 100 bp in <u>paired-end</u> format will be produced?

100,000,000 reads

(20 x 10^9) / (100 x 2) = 100,000,000 reads

1 billion (10^9) = 1 Gb

Number of reads (paired-end) = (Sequencing output x 10^9) / (Read length x 2)

6. a) If the sequencer can generate a maximum of 20Gb of information per sequencing run. How many genomes could be sequenced if the depth is lowered to 10X? Consider a genome size of 7.5Mb and include one decimal.

266.7 genomes

(20) / ((7.5/1000) x 10) = 266.7 genomes

Number of genomes = (Sequencing output per run) / (Genome size (Gb) x Coverage depth)

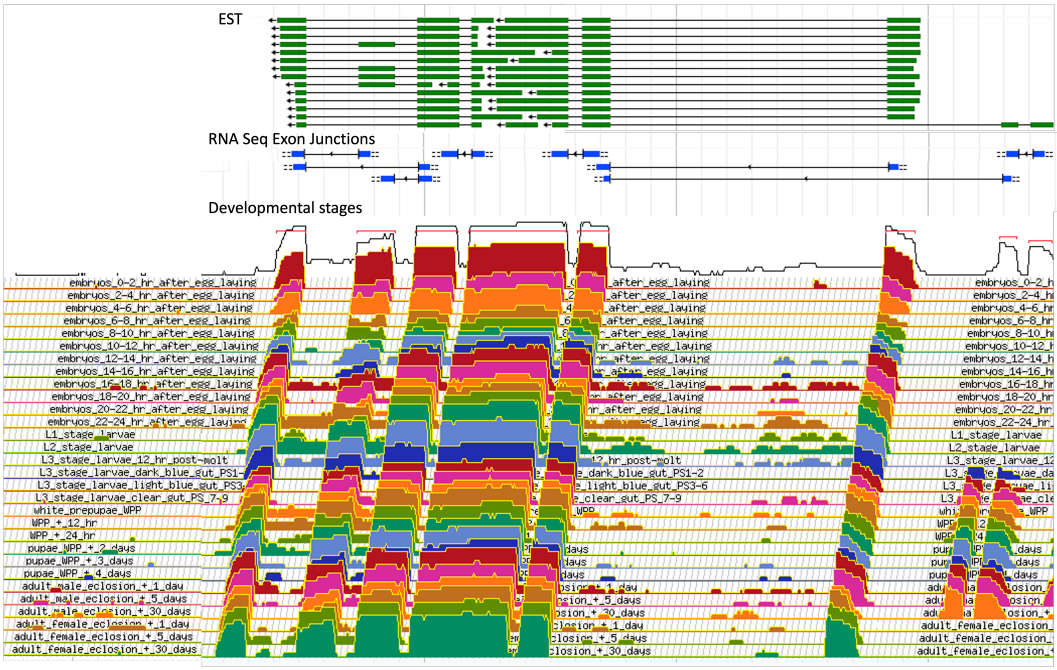
b) What effects could a lower sequencing depth bring to the assembly of sequences?

a-The quality will decrease due to the low number of sequences produced per genome, which will make it difficult to assemble contigs and scaffolds.

Exit ticket 2

For a region of the Drosophila melanogaster genome you have the following expression data (https://aula.esci.upf.edu/pluginfile.php/276605/question/questiontext/389805/1/4590676/Picture%201%20%283%29.png) as a result of EST sequencing (from a cDNA library made with embryonic mRNA of all ages) and experiments of RNA-seq across fly development . For RNA-seq data you have the information in two different tracks: RNA-seq exon junctions (RNA-seq reads that mapped to two different places of the genome, thus indicating the end of one exon and the beginning of the next exon), and expression profiles in different developmental stages from embryos to adult individuals (each phase is indicated to the right in the color corresponding to the graph).

1. Draw the identified transcript and mark the alternative splicing mechanisms in the image and indicate the site in which they occur within the image. Can you indicate the transcription start and end of the different transcripts? If so, indicate the sites within the image. Upload your file in the "RNA Seq exit ticket assignment" section in block 1-3.



2. How many genes can you identify in this region?

b- 1 gene

3. How many transcripts can you predict?

a- 3 transcripts

4. What alternative splicing mechanisms are used to generate the different transcripts?

d- Intron retention and exon skipping

5. At which moments of fly development is the gene expressed?

b- During all the development stages

6. Do the different transcripts show differential expression throughout development?

d- Yes there is one transcript expressed from L3 to adult male eclosion

7. Do the data of ESTs and those of RNA-seq match?

a- Yes, EST and RNA-seq data match

8. What is the relevance of junction reads of RNA-seq?

c- Junction reads of RNA-seq show all exon connections and can putatively reveal all transcripts.

9. Is this sentence true or false?

RNA-seq coverage (expression profiles) show length of exons and levels of expression along development. ESTs emphasize structure information, although they do not add any additional information on top of what RNA-seq reveals.

a- True

Practical Exit ticket 2 → Practical_exit_ticket_solved.pdf

Exit ticket 3

1. Select the right option:

a- The best approach for bacterial diversity studies is 16S metabarcoding only

2. Regarding to the sequencing technologies that can be used for 16S metabarcoding:

c- Illumina sequencing allows the analysis of only few variable regions of the 16S gene

3. 16S metabarcoding based on Nanopore sequencing offer advantages such as:

c- Allows to sequence the entire 16S gene

4. Metagenomic binning is:

d- Reconstruction of complete or near to complete genomes from a metagenome

5. Metagenomics allows:

d- To determine functionality based on genetic content

6. Metagenomics can be carried out using:

c- Illumina, Oxford Nanopore and PacBio

Exit ticket 4

1. Culturable pathogen with high infectivity and multi drug-resistance which is circulating in a hospital:

a- Genomics

A culturable pathogen with high infectivity and multi-drug resistance that is circulating in a hospital requires a genomic approach, specifically whole-genome sequencing (WGS) of the pathogen, to provide **detailed information** about the pathogen's genetic makeup, including specific mutations that may contribute to its virulence or resistance to antibiotics.

2. High mortality caused by an outbreak of unknown pathogens after a heat wave in an aquaculture farm

b- Metagenomics

In the case of an outbreak of unknown pathogens after a heat wave in an aquaculture farm, metagenomics is an appropriate approach for identifying the pathogen(s) responsible. Metagenomics can analyze multiple samples simultaneously, providing a broad overview of the microbial community present in the samples, including potential pathogens. This approach does **not require** the **isolation** of individual pathogens, making it useful for identifying non-culturable or difficult-to-culture microorganisms.

3. Pathogen dissemination along hospital units:

c- Metabarcoding.

Metabarcoding is an appropriate approach for identifying the **presence and diversity** of pathogens in environmental samples from different areas of the hospital, including surfaces, air, and water. This approach can provide valuable information about the types of pathogens that are present in the hospital and how they are spreading between units.

4. Functional characterization of a bacterial consortia (community of different bacterial species) that degrade micro plastics:

b- Metagenomics

Metagenomics is an appropriate approach for characterizing the functional potential of a microbial community, including a bacterial consortia that degrade microplastics. Metagenomics can provide information about the metabolic pathways and genes involved in the degradation process, as well as potential factors that influence the efficiency of degradation.

5. Identification of antibiotic resistance mechanisms of a new isolated pathogen serotype:

a- Genomics

For the identification of antibiotic resistance mechanisms of a new isolated pathogen serotype, a genomic approach, specifically whole-genome sequencing (WGS) of the pathogen, is appropriate. WGS can provide information about the pathogen's genetic makeup, including its antimicrobial resistance genes, that can guide treatment strategies for patients infected with the pathogen.

6. Taxonomic identification of a new non-culturable bacterial pathogen:

b- Metagenomics

Metagenomics is an appropriate approach for identifying the taxonomic identity of a new non-culturable bacterial pathogen, as it can provide a broad overview of the microbial community present in the sample, including potential new or uncharacterized microorganisms.

7. Surveillance of pathogens in production lines of a food industry:

c- Metabarcoding

Metabarcoding is an appropriate approach for the surveillance of pathogens in production lines of a food industry. This approach can identify the presence and diversity of microorganisms, including potential pathogens, in food and environmental samples, providing valuable information for food safety and quality control measures.

Genomics: study of an **organism's entire** genetic material, typically through the **sequencing** and **analysis** of its DNA or RNA. It is used to understand the genetic basis of traits, including **pathogenicity** and **antibiotic resistance**, among others.
Metagenomics: study of genetic material from complex **microbial communities**, typically in environmental or clinical samples. It allows for the analysis of all the genes present in a sample, including those from non-culturable microorganisms, and can provide insights into the functional potential of the microbial community.
Metabarcoding: subset of metagenomics that focuses on sequencing a **specific gene/region** of DNA, typically a marker gene like 16S rRNA, to identify the taxonomic composition of a microbial community. It can provide a rapid and cost-effective method for characterizing microbial diversity.

Block 2. Quiz 1

1. Which of the following is not a common visualization and exploratory data analysis tool for gene expression data?
   b. Generalized Linear Models (GLM)

2. Which of the following is not true about normalization?:
   b. Within-sample normalization is required for differential gene expression analysis

3. Which of the following is false about RNA-seq data?
   c. Continuous data, with a large dynamic range and presence of 0 counts

4. Which of the following is false about Bioconductor?
   a. Bioconductor packages can be installed using the R function 'install.packages'

5. Which of the following metrics will be the most appropriate to compare the expression of a gene between two technical replicates?
   b. CPM

6. Which of the following normalization methods assumes that read counts are proportional to gene length?
   b. Both RPKM and TPM

7. Regarding the 'ExpressionSet' class:
   a. The 'GEOquery' package can be used to create an 'ExpressionSet' from a GEO dataset

8. Which of the following is true about the 'SummarizedExperiment' class:
   b. The rows of a 'SummarizedExperiment' object can correspond to a 'GRanges' object

9. Which of the following is a variance stabilizing transformation?
   e. None of the other answers is correct

10. Which of the following is a typical example of a "class comparison" task with RNA-seq data?
    c. Differential gene expression analysis

Quiz 2

1. Which of the following is false regarding design and contrast matrices?
   a. A contrast matrix should define more than one contrast

2. When applied to differential gene expression, linear models…
   d. Often assess significance of contrasts built upon model parameters

3. Select the right answer regarding the main assumptions of linear models:
   e. All the answers are correct (a. Errors are independently distributed, b. Errors have constant variance, d. Errors are normally distributed with mean 0)

4. Which of the following is the most appropiate model for overdispersed RNA-seq data?
   e. Negative Binomial

5. Which of the following statements is false?
   c. Two RNA libraries prepared from the same lung cancer sample and sequenced separatedly are biological replicates

6. Select the wrong answer regarding TMM normalization:
   b. It ensures identical gene expression distributions across samples

7. Prior to differential gene expression analysis…
   e. All the answers are correct (a. Exploratory data analysis should be performed, b. Raw read counts should be normalized, c. Genes that are lowly expressed should be removed)

8. You study the effect of a drug on the expression of a gene using linear models. You assign mice randomly to a group and treat them accordingly: treated (n = 5) and untreated (n = 5). You fit a model with a single explanatory variable: "group". Your null hypothesis is: "The drug has no effecton on gene expression". Which of the following is correct?
   c. All the answers are correct (a. If the design matrix contains an intercept term, it corresponds to the mean of one of the groups, b. The response variable is the gene expression, d. If the design matrix does not contain an intercept term, a possible contrast matrix is (-1, 1))

9. You study the effect of 4 different drugs on the expression of a gene using linear models. You assign mice randomly

to a group and treat them accordingly: control/no drug (n = 2), drug A (n = 2), drug B (n = 2), drug C (n = 2), drug D (n = 2). You fit a model, without intercept, with a single explanatory variable ("treatment"). Your alternative hypothesis is: "The mean gene expression of all treatment groups is different from the control group"? Which of the following is an appropriate contrast matrix?

d. (−1, 0.25, 0.25, 0.25, 0.25)

10. Which of the following is false about linear models?

c. Linear models require a large number of samples to be fit

Quiz 3

1. Modelling the read counts of a gene across replicates using a Negative Binomial model assumes that:

b. The variance is larger or equal than the mean.

2. Which of the following is true about REVIGO?

e. It enables similarity-based reduction of GO term lists and representation in 2D.

3. Which of the following is false regarding usual representations of DGE analysis results?

e. The best way to summarize the results of a DGE analysis is representing boxplots for each gene

4. Which of the following is true regarding the voom-limma approach?

c. None of the answers is correct. (a. limma learns the mean-variance trend and adjusts it in the context of standard linear models (implemented in voom), d. It tipically requires as input TPMs or RPKMs, e. It does not require the definition of design and contrast matrices.)

5. In a principal component analysis (PCA) of gene expression data from RNA-seq of cancer patients and healthy controls:

a. All the answers are correct (b. Points represent samples and principal components are linear combinations of gene expression values, d. The distance betweeen two points represents the overall difference in gene expression values between two samples, e. The information about sample groups (cancer, controls) is not used to calculate the principal components.)

6. Overdispersion of read counts is due to the fact that:

c. The variability among biological replicates is larger than in a Poisson model

7. Which of the following is false regarding Generalized Linear Models (GLMs):

d. voom+limma implements a GLM for DGE analysis

8. Which of the following is false regarding biological significance assessment?

e. Both GO enrichment and GSEA require a ranked gene list

9. After a clustering analysis of gene expression measured via RNA-seq on 20 individuals (10 treated with drug A, 10 treated with drug B), you observe two major clusters. The first cluster contains mostly male, drug A-treated patients, the second contains mostly female, drug B-treatedpatients. Which of the following is correct?

d. Sex should be included as a potential confounder in the models for DGE analysis between treatments

10. Select the right answer regarding multiple testing correction

a. Bonferroni correction is tipically more stringent than FDR