---

1. Mention one "Pro" and one "Cons" for each of the six DNA-seq techniques below:

	Pro	Cons
Sanger	-Low error rate -Long reads (800nt)	-Expensive -Low throughput
454/Roche	-Best quality among NGS technologies -Long reads (800nt)	-Expensive -Errors in homohaploid reading -Slow because of optical reading
Illumina	-Cheap -High throughput	-Slow because of optical reading -Higher error rate than 454/Roche -Short reads
Ion Torrent	-No need for optical reading -Similar quality as 454/Roche	-Errors in homohaploid reading -Expensive
Pacific Biosciences	-Long reads (50kb) -High throughput	-High error rate -Expensive
Oxford Nanopore	-No need for amplification -Portable -Very long reads -No optical reading	-High error rate -Still testing

2. You want to sequence an eukaryotic genome never sequenced before. Your budget is limited and you decide to make a whole-genome shotgun sequencing with a next-generation sequencing technique. If you could choose one sequencing technique, which one would you recommend? Why?

Since my budget is limited, I would probably recommend Illumina because it is a cheap technology and it is the most used. Also, if we are doing WGS sequencing, we will have a big amount of reads that we could use for detecting and fixing any possible errors that Illumina reads may have.

Would it be a good idea to combine two sequencing techniques? Which ones would you combine? Why?

If our budeget is limites, I don't know if it would be a good idea to combine two sequencing techniques.

But if the budget allow us, I think it could be a good idea to maybe combine the quality of 454/Roche with the chap reads from Illumina in order to use both to obtain a final read with a very high quality.

3. It is time for assembly. Which assembly strategy are you going to follow? Why?

- ☐ Mapping against a reference
- ☐ (This)De novo assembly

I would follow a De novo assembly because it is the first time we are sequencing this eukariotic genome, so we don't have any reference genome in order to do the mapping.

4. You get two separate assemblies of your sequencing data, made by two different assembly software. According to the metrics shown in the table below, which assembly looks best?

	ABySS	Trinity
Number of contigs	120,479	47,571
N50 size (bp)	7,338	17,425
Longest contig (bp)	21,684	468,339

- ☐ (This)ABySS
- ☐ Trinity

Why?

I would choose AbySS because of the N50 size value. We know that we will have more than 50% (60000 aprox) of the contigs with size 7338 bp, which is relatively close to the size of the longest contig if we compare this same values from Trnity. There is less variation of size from 7338bp to 21684bp than from 17425bp to 468339bp.

5. What do you need to form scaffolds? Briefly explain the process.

In order to form scaffolds, fist of all we will need several reads that will form contigs. Then this contigs will be joined by paired-end and we will have our scaffold.

6. Paired-end mapping (PEM) is another application of DNA

sequencing. Describe the aim and procedure of the PEM technique.

1. Constuction of a genomic library of DNA fragments of a certain size.

2. Pair end sequencing

3. Mapping to a reference genome.

7. I am providing paired-end mapping data for three fosmid sequences (average insert fragment length of ~40 Kb). Do they reveal the presence of structural variants in any of the regions? Specify the type and approximate size (if possible).

READ	# HITS	BEST HIT					STRUCTURAL VARIANT?
		IDENTITY	CHR	STRAND	START	END	
F1 fwd	4	99,2%	7	+	117133465	117134193	Deletion (195bp)
F1 rev	8	99,4%	7	–	117172022	117173660	
F2 fwd	87	96,3%	19	+	21837776	21838534	Insertion (7703bp)
F2 rev	182	98,2%	19	–	21868365	21870073	
F3 fwd	7	100,0%	X	+	153560230	153560952	Inversion (13063bp)
F3 rev	7	98,0%	X	+	153586670	153587167	

8. Explain what information does RNA-seq transcriptomic data provide you.

From RNA-seq transcriptomic data we can:

-Catalog the different types of RNA: mRNA, non-codingRNA (ncRNA), smallRNA (sRNA)

-Study the Structure of the genome: transcription start and end, 5' and 3' UTR, splicing patterns, Post Transnational Modifications

-Analyze the transcript expression over development or under certain physiological conditions.

9. The figure below displays EST and RNA-seq data mapped to a given genomic region.

- ☐ How many genes does the genomic region contain? _____1_____
- ☐ Do/does the gene(s) show(s) alternative splicing? _____Yes_____ Draw all the transcripts in the reserved space within the figure.
- ☐ What alternative splicing mechanisms are used to generate the different transcripts? Enumerate them and mark the place where they occur in the figure.

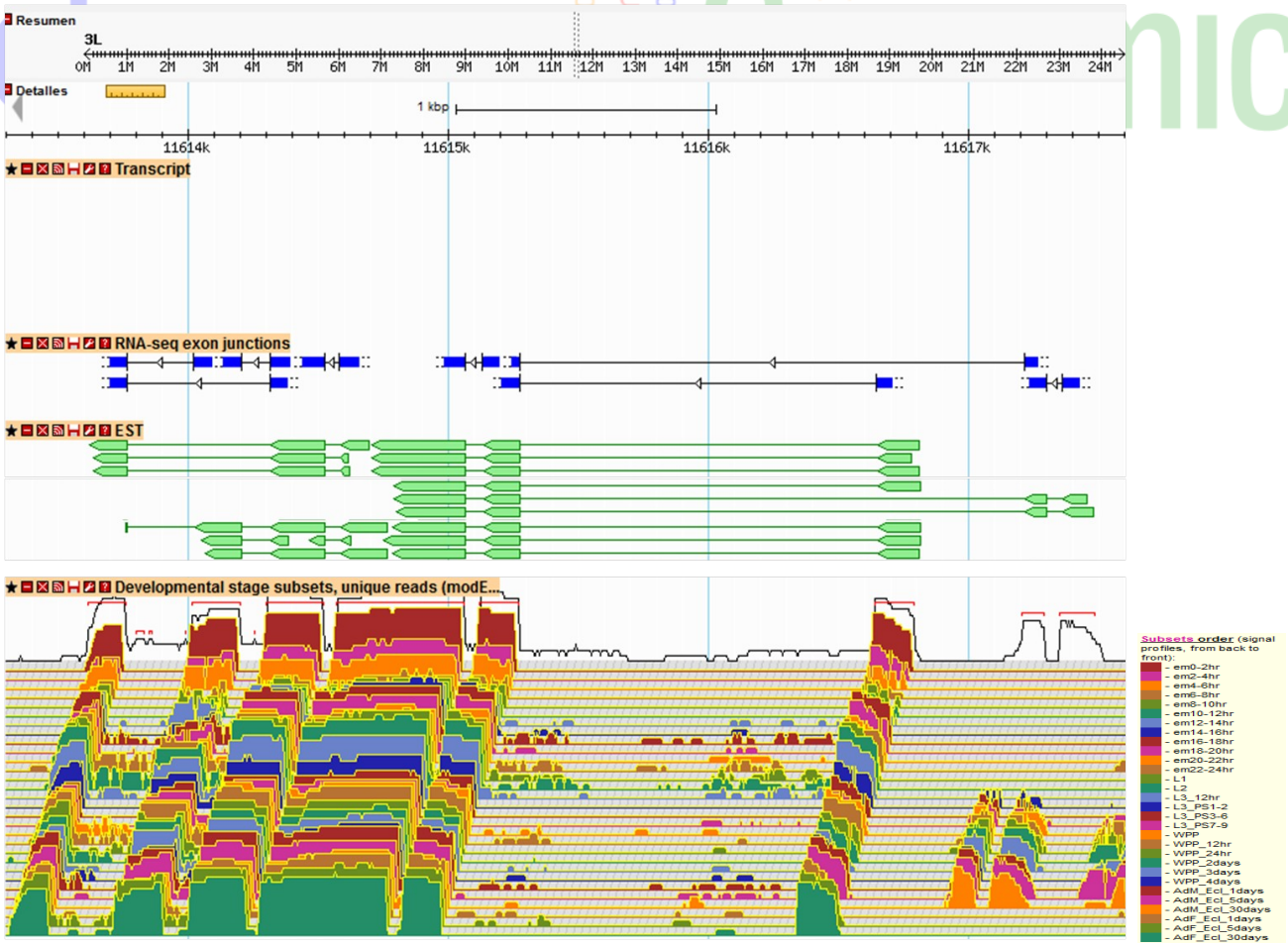
Here we only have alternative transcription start and alternative transcription end.

- ☐ Do the different transcripts show differential gene expression throughout development?

Yes, transcript 2 starts its expression earlier in the development.

- ☐ Are all the proteins encoded by the different transcripts identical? _____ Mark in the figure the beginning and the end of the translation of each transcript.

From this data we cannot know the proteins this gene codify nor the translation start and end.


Resumen
3L
0M 1M 2M 3M 4M 5M 6M 7M 8M 9M 10M 11M 12M 13M 14M 15M 16M 17M 18M 19M 20M 21M 22M 23M 24M
Detalles
1 kbp
11614k
11615k
11616k
11617k
Transcript
RNA-seq exon junctions
EST
Developmental stage subsets, unique reads (modE...
Subsets order (signal profiles, from back to front):
em0-2hr
em2-4hr
em4-6hr
em6-8hr
em8-10hr
em10-12hr
em12-14hr
em14-16hr
em16-18hr
em18-20hr
em20-22hr
em22-24hr
L1
L2
L3_12hr
L3_PS1-2
L3_PS3-6
L3_PS7-9
WPP
WPP_12hr
WPP_24hr
WPP_2days
WPP_3days
WPP_4days
AdM_Ecl_1days
AdM_Ecl_5days
AdM_Ecl_30days
AdF_Ecl_1days
AdF_Ecl_5days
AdF_Ecl_30days