Part I. DNA and RNA sequencing

Your budget is limited and you decide to make a whole-genome shotgun sequencing with a next-generation sequencing technique. Since your goal is to achieve a high qualityassembly, equivalent to the quality of the human reference genome, which sequencingtechnique do you recommend to your group?

□ 454/Roche
□ Oxford Nanopore
□ A combination of the two above
☑ The objective that you propose is not a realistic goal

Finally, your group invests a large part of the budget in generating the genomic libraries and sequencing. Now you have in your hand billions of Illumina and Pacific Biosciences sequencing reads. Explain which will be the strengths of Illumina and Pacific Biosciences data, and why it is a good idea to combine both:

- **Illumina**: Illumina will provide short reads (~200 bp – 1500 bp) of better quality than a third generation technique and they are of high throughput.
- **Pacific Biosciences:** Pacific Biosciences will provide long reads. These reads can be produced very fast since they are sequenced in real time.
- **Why combine both:** The combination of both approaches will help to identify long regions and therefore, more resolution (thanks to Pacific Biosciences) while checking small reads that are not taken into consideration and ensuring a better quality (thanks to Illumina).

It is time for assembly and you are still discussing which assembly software you are going to use. What is clear to you is the assembly strategy you are going to follow:

□ Mapping against a reference
☑ De novo assembly
□ Any of the two above
□ Expression profiling

Since we are going to sequence a genome for the first time, we have no reference genome to compare it to (so we cannot do mapping against a reference). Also, we want to assemble a genome so there is no need to assess the expression now.

You choose to try with three assembly programs for which you have good references. But you get soon disappointed... You cannot make the first work after two weeks trying to install it. The second launch a Segmentation fault message (the computer does not have enough memory). The third takes four weeks working without giving you the results.

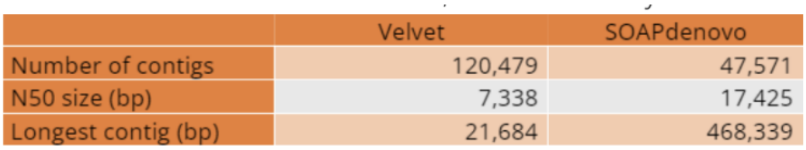
□ The problem is the data. It must have a problem, since a genome of these characteristics must be able to get assembled in a desktop computer.
□ The problem is the version of the program you installed. The lab next to yours has assembled a bacterial genome with older versions of the same program, and it just took few minutes.
□ The problem is your computer. Surely, your colleague’s computer, which is newer, will be able to assemble the genome without problems.
☑ You have underestimated the computing requirements for this project.

Finally, you get two separate assemblies of your sequencing data, made by two different assembly software. The first thing you do is compare basic metrics between the two. According to the values shown in the table below, which assembly looks best? Why?

SOAPdenovo

Why? SOAPdenovo presents less number of contigs (so there must be less sequencing error, because more contigs means more sequencing and therefore more errors). N50 is larger (so half of the contigs will even be larger and we will have more resolution). It also has the longest contig (so there is more resolution).

*** N50 means that 50% of the bases of the assembly are in contigs of length >= L (L=contig length)**



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

Considering that you have assembled 132.13 Gb of sequencing data and that the estimated genome size of the leafy seadragon is 695 Mb, calculate the redundancy (coverage):

Redundancy = (N * R)/G, where N = number of reads, R = average read length and
G = genome size. So, redundancy = 132130 Mb/695 Mb = 190.12

You decide to continue with one of the two previous assemblies. Now, to complete the assembly and form scaffolds it is essential to:

- ☑ sequence paired-end reads.
- ☐ eliminate repetitive regions.
- ☐ sequence the transcriptome by RNA-seq.
- ☐ compare contigs with a database of proteins of a nearby species.

The sequencing of a diploid species such as the leafy seadragon reveals sites in the genome where the individual has two different alleles in the form of a polymorphism. How do you think these sites can be detected?

- ☐ In the Illumina reads, heterozygous sites have an intermediate coloration between the two nucleotides corresponding to the two alleles.
- ☑ In the assembly, heterozygous sites have approximately half of the reads with one allele and the other half of the reads with the other allele.
- ☐ In the assembly, heterozygous sites have double the redundancy (coverage) thanthe rest
- ☐ The sequencing and assembly of a diploid individual results finally in 2n chromosomes assembled separately, so that heterozygous positions correspond to the differences between the two chromosomes.

The project discussed in this test corresponds to one of the applications of DNA sequencing: obtaining the genome sequence of a species for the first time. Mention an describe another application of DNA sequencing.

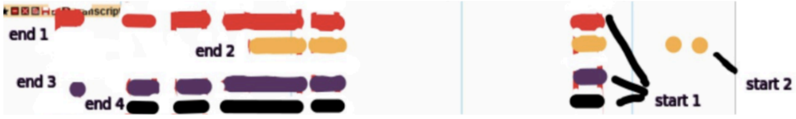
Another application for DNA sequencing is targeted genome resequencing. This technique is aimed for when we want to do metagenomics. It allows cheap and efficient sequencing of genes or regions of interest, such as exons, candidate regions for diseases, tumor genes and structural variants. First we sequence the library. Then we select the molecules that contain exonic sequences (so they are transcribed). Next, we sequence the selected molecules and identify any exonic variants from those produced reads. After that, we identify the variants that are more present in the reads and finally we select candidate genes.

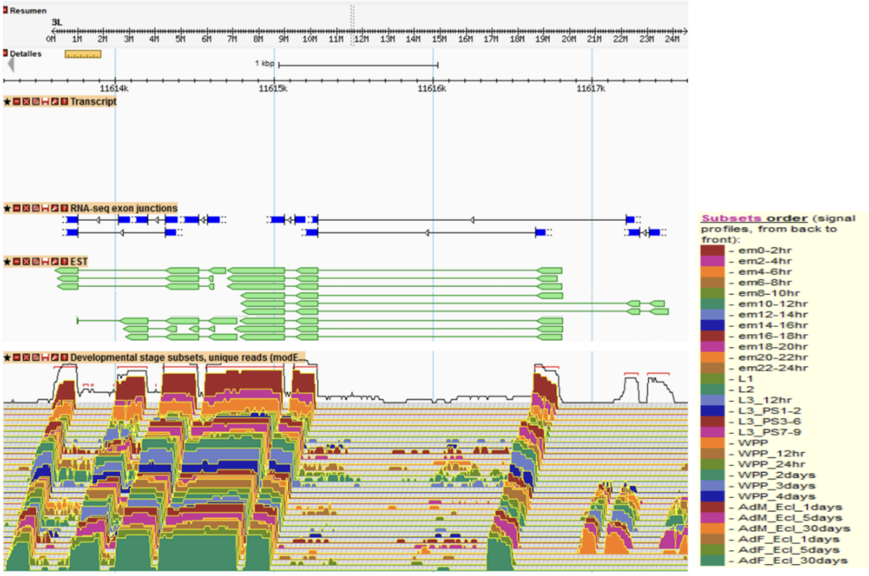
At a later/second stage of the genome sequencing of the leafy seadragon, you invest money in RNA-seq. Explain how will you process RNA-seq reads. Explain what information does transcriptomic data provide you.

To get RNA-seq reads, we need to first get the mRNA transcripts and convert them to cDNA in order to sequence them (or we could directly just use Oxford Nanopore). Then, these reads should be mapped against the genome, however, they are spliced so we need to map the different parts of the transcript against the genome. With the help of junction reads we can infer where the introns are and therefore to establish the boundaries of these exons. Taking into account these RNA reads, we will be able to determine the levels of expression of the reads. Transcriptomic data is useful to assess the expression of different cell types and to understand which genes seem to be tissue-specific and which ones are not.

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

- **How many genes does the genomic region contain?** 1 gene
  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?** Alternative transcription start-site, skipped exon
  Enumerate them and mark the place where they occur in the figure.
- **Do the different transcripts show differential gene expression throughout development?** Yes. Some transcripts are expressed only during stages of L3 until the initial AdM phase. Also, notice in some early embryo phases, the expression levels are reduced among all transcripts.
- **Are all the proteins encoded by the different transcripts identical?** No, they are not since they are composed of different exons.
  Mark in the figure the beginning and the end of the translation of each transcript.
  We cannot know the beginning and end of translation of the transcripts.
  That is because transcripts contain 5'UTR and 3'UTR regions (untranslated regions) and they are not indicated in the EST data therefore, we are unable to identify them.





Mention one “Pro” and one “Cons” for each of the six DNA-seq techniques below:

Sequencing Technique	Pros	Cons
Sanger	- Low error rate (reliable) - Long reads (up to 1000bp) - High accuracy (99.99%)	- Expensive - Low throughput - Time Consuming
454 / Roche	- Best quality among NGS technologies - Long reads (up to 700 bp)	- Expensive - Errors in homohaploid reading - Slow because of optical reading - Worse quality among NGS technologies - Homopolymer errors
Illumina	- Cheap - High throughput - High accuracy - Widely used in the market (highly accessible and cheap)	- Slow because of optical reading - Higher error rate than 454/Roche - Short reads - PCR amplification bias - Requires complex library preparation
Ion Torrent	- No need for optical reading, No need of fluorescence or other technology devices - Similar quality as 454/Roche	- Errors in homohaploid reading (Homopolymer errors) - Expensive - Medium/Bad quality - Lower accuracy compared to Illumina
Pacific Biosciences	- Long reads (up to tens of kb) - Single-molecule sequencing - High throughput - High consensus accuracy	- High error rate in raw reads - Requieres large amount of input DNA - Expensive
Oxford Nanopore	- No need for amplification - Portable and scalable - Very long reads (up to 2 Mb) - No optical reading - Real-time sequencing - Can be run with RNA	- High error rate - Still testing - Lower throughput compared to Illumina - Can be affected by DNA quality

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.

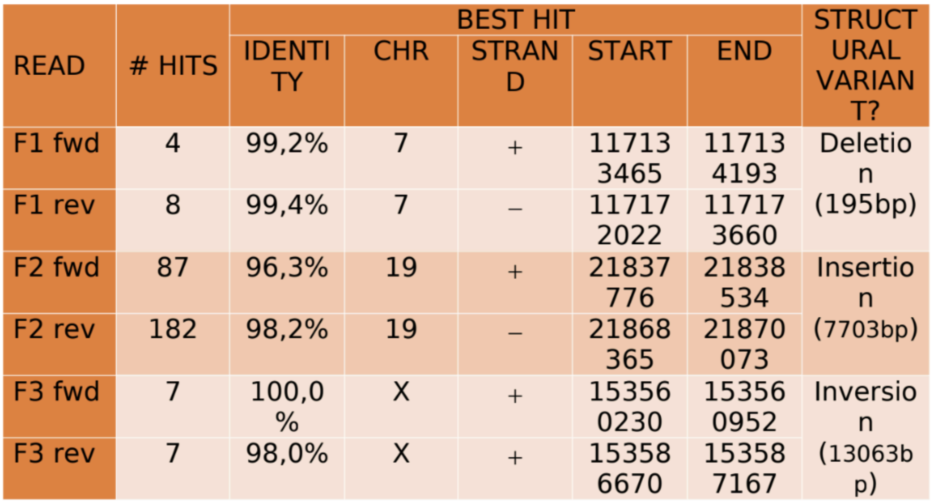
What do you need to form scaffolds? Briefly explain the process.

We need contigs in order to form scaffolds and reads to form contigs. (Contigs are a set of reads that overlap in their extremes to generate longer contiguous sequences and scaffolds are oriented and ordered contigs based on the information from PEM.) So, first we have the reads which create contigs overlapping within them, and based on information from Paired-end reads we can order and orient those contigs forming the scaffolds.

Paired-end mapping (PEM) is another application of DNA sequencing. Describe the aim and procedure of the PEM technique.

Paired-end mapping is a technique that allows us to detect structural variants in the DNA of an individual of interest by obtaining paired-end reads and the comparison of their positions in a reference genome.

I am providing paired-end mapping data for three fosmid sequences. Do they reveal the presence of structural variants in any of the regions? Specify the type and approximate size



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	

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

☑ Mapping against a reference

☑ 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.

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.

Paired-end mapping (PEM) is another application of DNA sequencing. Describe the aim and procedure of the PEM technique.

a. Constuction of a genomic library of DNA fragments of a certain size.
b. Pair end sequencing
c. Mapping to a reference genome.

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

Part II. Bioconductor, Experimental design for omics studies and differential expression analysis

Describe the structure of the summarizedExperiment Bioconductor class. Which are its main differences with respect to the expressionSet class?

A researcher wants to perform an RNA-seq experiment, followed by a differential gene expression (DGE) analysis, to compare gene expression between lung cancer patients and healthy controls. Draw a possible workflow with the steps to carry out such study, starting from the biological question, until the interpretation of the results.
First, the researcher should ask them self the biological question: “Are there any genes which are differentially expressed between lung cancer patients with respect to healthy ones?”. After that, the researcher should carry out and RNA-seq experiment, where it would be necessary to recollect samples from both types of patients, sequence them (having previously converted them to cDNA since where are collected RNA), applying some quality control, mapping them against the genome, counting reads and applying any correction regarding normalization, some duplicated information, outliers, etc. Then, we should get the matrix of TPM and try to find which genes are significantly expressed. We should calculate the logFC and finally, plot them in a Volcano plot. Genes with a very low p-value and with a logFC whose absolute value is large enough will be the genes that will be differentially expressed among patients.

Among the following metrics: read counts, CPM, RPKM and TPM, select the most appropriate one to compare the expression of a given gene between two technical replicates. Justify your answer.
CPM. CPM normalizes library size but not length. RPKM and TKM work with the length and knowing we are working with the same sequence is not useful.

Explain why raw read count data should not be directly modeled using standard (i.e. Normal) linear models. Enumerate two alternative strategies for this purpose.
Raw read count data is not normalized. That means that some genes were sequenced more than others and show more counts and others are just longer and will obviously have more reads. Therefore, we should apply some corrections. We can use scaling factors such as FPKM (fragments per kilobase per million) or CPM (counts per million).

Reason why lowly expressed genes across all samples should be removed prior to differential gene expression analysis.
Lowly expressed genes suppose some memory space for the computer to have, so for our computer in order to better work efficiently we should remove all genes that have lowly expressed and save memory space. Moreover, a differential gene expression analysis is centered on studying genes that are differentially expressed among samples. If some gene is lowly expressed between all individuals in the experiment, it is not making a difference and that means that the gene is just not involved in the process we are studying.

Describe the problem of overdispersion of read count data.
Read count data presents the problem that there is some mean-variance relationship and sometimes there is overdispersion (the variability among biological replicates is larger or equal than the mean). That supposes a problem because we would not be properly estimating in our model. That is why it is useful to use Negative Binomial models to avoid this and even have an overdispersion parameter (though sometimes it could be worse to have to estimate a parameter).

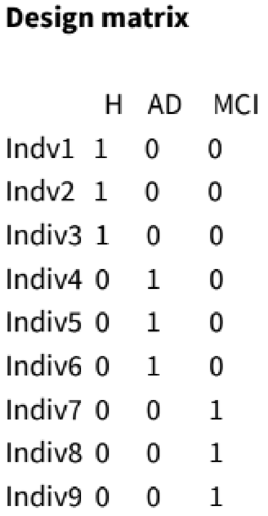
Which are the main differences between overrepresentation analysis (e.g. GO enrichment) and Gene Set Enrichment Analysis (GSEA)?

GO receives a set of genes and results the % of genes in the set involved in the pathway, while GSEA should receive all the genes and by comparing those genes with the ones involved in the pathway, the enrichment number/index will be increasing or decreasing and will be finally resulting in a number which will determine the significance of the genes related to the pathway.

Describe what you could do to identify a batch effect in your expression data.

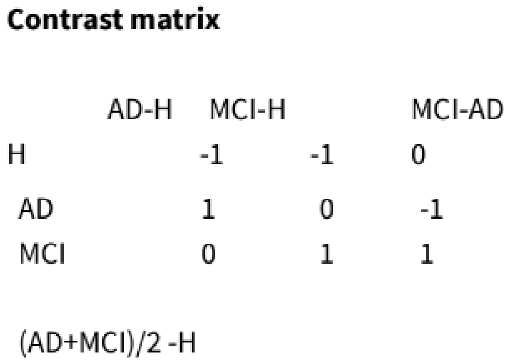
To identify batch effect in our expression data, we could use PCA and heatmap representations. These approaches offer the possibility to check how is our data aggroupated and how it is related to the different independent variables that we may be considering in our study. If the heatmap presents a certain variable that seems to associate with our explanatory variables, it is important to consider these problematic variables as confounding ones and include them in the study.

A researcher wants to study gene expression in Alzheimer's disease (AD), mild cognitive impairment (MCI) and healthy (H) conditions. He performs RNA-seq on three individuals per condition, followed by a DGE analysis (voom + limma, models without intercept term) to compare gene expression between all the conditions pairwise. Draw the corresponding design and contrast matrices.



Design matrix

	H	AD	MCI
Indv1	1	0	0
Indv2	1	0	0
Indiv3	1	0	0
Indiv4	0	1	0
Indiv5	0	1	0
Indiv6	0	1	0
Indiv7	0	0	1
Indiv8	0	0	1
Indiv9	0	0	1

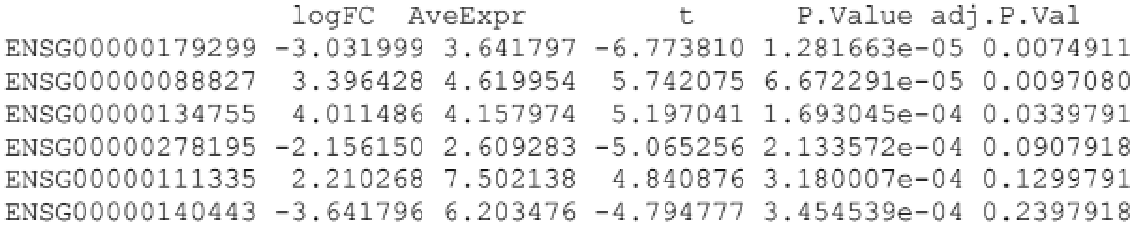


Contrast matrix

	AD-H	MCI-H	MCI-AD
H	-1	-1	0
AD	1	0	-1
MCI	0	1	1

(AD+MCI)/2 -H

In the previous study, after performing DGE analysis and adjusting for multiple testing via FDR, for the contrast AD – H, the researcher got the following results (only the top 6 genes are shown):



```
                    logFC  AveExpr         t      P.Value adj.P.Val
ENSG00000179299 -3.031999 3.641797 -6.773810 1.281663e-05 0.0074911
ENSG00000088827  3.396428 4.619954  5.742075 6.672291e-05 0.0097080
ENSG00000134755  4.011486 4.157974  5.197041 1.693045e-04 0.0339791
ENSG00000278195 -2.156150 2.609283 -5.065256 2.133572e-04 0.0907918
ENSG00000111335  2.210268 7.502138  4.840876 3.180007e-04 0.1299791
ENSG00000140443 -3.641796 6.203476 -4.794777 3.454539e-04 0.2397918
```

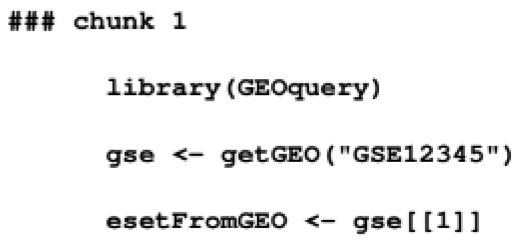
How many significant genes are there at 5% FDR? How many significant genes are over- and under-expressed in AD with respect to H? Which plot would you use to summarize the information contained in this table? Justify your answers.

There are 3 genes that are significant (we need to select those whose adj.P.val < 0.05%, in this case, the first 3 ones).

To know if the genes are over or under expressed in AD with respect to H, we must check the logFC. If it is positive, they are over-expressed and if not, they are under- expressed. In this case, from the 3 selected significant genes, 2 are over-expressed (ENSG00000088827 and ENSG00000134755) and the other one (ENSG00000179299) is under-expressed. (If we were considering all 6 genes, then there would be 3 under-expressed and 3 over-expressed).

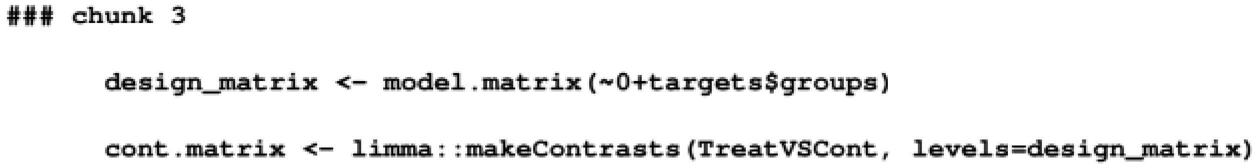
The ideal plot to summarize this information would be a Volcano-plot. In the X-axis we would display the logFC and in the Y-axis we would represent the p-values. Values on the top and far away from the center would indicate significant genes that are over/under-expressed.

The code below illustrates a typical use of Bioconductor packages for microarray analysis. Indeed it is something you have used in class. Explain briefly (not more than 2-4 lines) what does each code chunk do



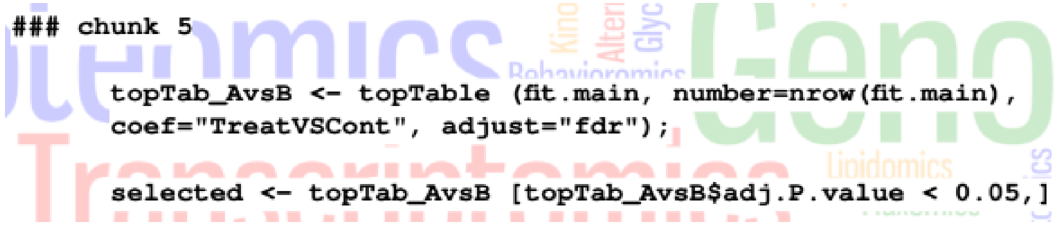
```
### chunk 1

      library(GEOquery)

      gse <- getGEO("GSE12345")

      esetFromGEO <- gse[[1]]
```

Call the GEOquery library.
Assign to the gse varialbe the dataset obtained with the getGEO function from the GEOquery library.
Assign to the esetFromGEO variable the previously obtained dataset but removing its first line, which contained the zip name of the file and can cause problems in the later analysis.



```
### chunk 3

      design_matrix <- model.matrix(~0+targets$groups)

      cont.matrix <- limma::makeContrasts(TreatVSCont, levels=design_matrix)
```

With the model.matrix function, create a design matrix taking in account our targets and the groups we have assigned. From the limma package, call the makeConstast function and create a contrast matrix for the comparison of TreatVSCont using values from the previously created design matrix.



```
### chunk 5

      topTab_AvsB <- topTable (fit.main, number=nrow(fit.main),
      coef="TreatVSCont", adjust="fdr");

      selected <- topTab_AvsB [topTab_AvsB$adj.P.value < 0.05,]
```

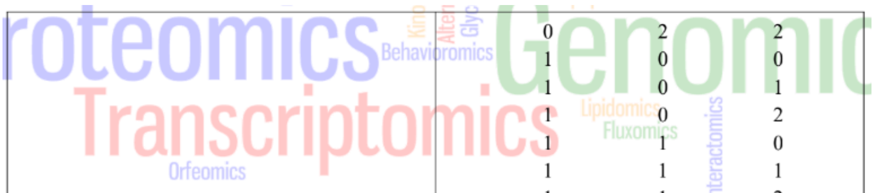
Create a toptab with the our fitted values from the TreatVSCont comparisson and only select those with a Pvalue higher than 0.05.

A researcher has done a microarray analysis to compare gene expression between healthy donors with patients affected from endometriosis in three types of populations. She has done 12 microarrays (or RN- seq runs, the design would be the same) 6 from patients with endometriosis and 6 from healthy donors.

In each group there are 3 different cell populations call them A, B, and C.
The researcher wishes to make the following comparisons: (i) Compare healthy vs affected in each population. (ii) Compare population A vs B in affected patients.

1. **Imagine Samples are named as [H, A][A,B,D][1,2,3]. Write down an appropriate "targets" table to describe the experimental layout for this study**
2. **Write down the design matrix associated with this study design. HINT: Do as we have done in class and combine the two main factors into a single one**
3. **Build the contrast matrix needed to do the comparisons described above.**

TARGETS			DESIGN MATRIX		
			H=0; A=1	A=0; B=1;	1=0; 2=1; 3=2
HA1	HA2	HA3			
HB1	HB2	HB3		C=2	
HC1	HC2	HC3	0	0	0
			0	0	1
			0	0	2
AA1	AA2	AA3	0	1	0
AB1	AB2	AB3	0	1	1
AC1	AC2	AC3	0	1	2
			0	2	0
			0	2	1
			0	2	2
			1	0	0
			1	0	1
			1	0	2
			1	1	0
			1	1	1
			1	1	2
			1	2	0
			1	2	1
			1	2	2



CONTRAST MATRIX

(doing everything in the same population)

α1 = E(logHA1) ; α4 = E(logAA1)

α2 = E(logHB1) ; α5= E(logAB1)

α3 = E(logHC1) ; α6 = E(logAC1)

COMPARISSONS 1: Healthy vs Affected

α1 vs α4
α2 vs α5
α3 vs α6

$$\begin{pmatrix}\beta_1^1\\ \beta_2^1\\ \beta_3^1\end{pmatrix} = \begin{pmatrix}-1 & 0 & 0 & 1 & 0 & 0\\ 0 & -1 & 0 & 0 & 1 & 0\\ 0 & 0 & -1 & 0 & 0 & 1\end{pmatrix}\begin{pmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\\ \alpha_4\\ \alpha_5\\ \alpha_6\end{pmatrix} + \begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \varepsilon_4\\ \varepsilon_5\\ \varepsilon_6\end{pmatrix}$$

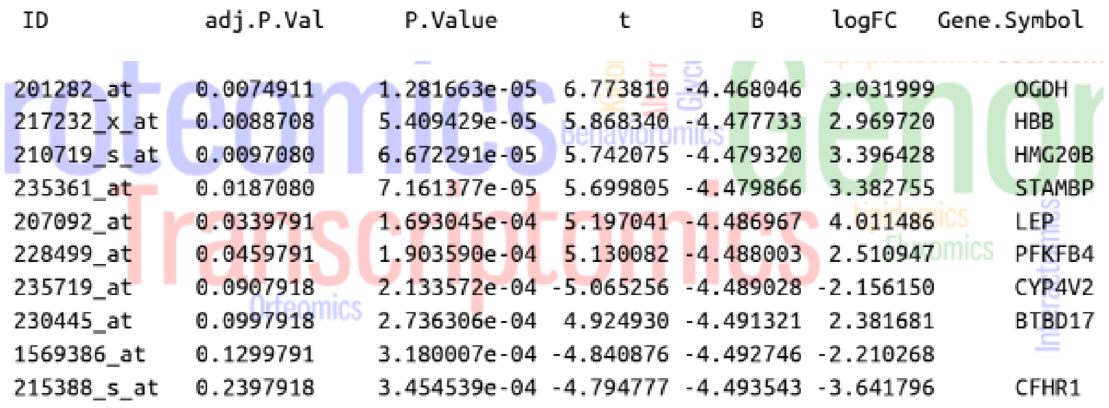
COMPARISSONS 2: A vs B in affected patients

λ1 = E(logAA1)

λ2= E(logAB1)

$$\left(\beta^{2}_{1}\right) = \begin{pmatrix} -1 & 1 \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

Using the previously defined matrices we have fitted a linear model to the data and have obtained one top tables (one for each comparison). Table below shows the results for a few genes arbitrarily selected



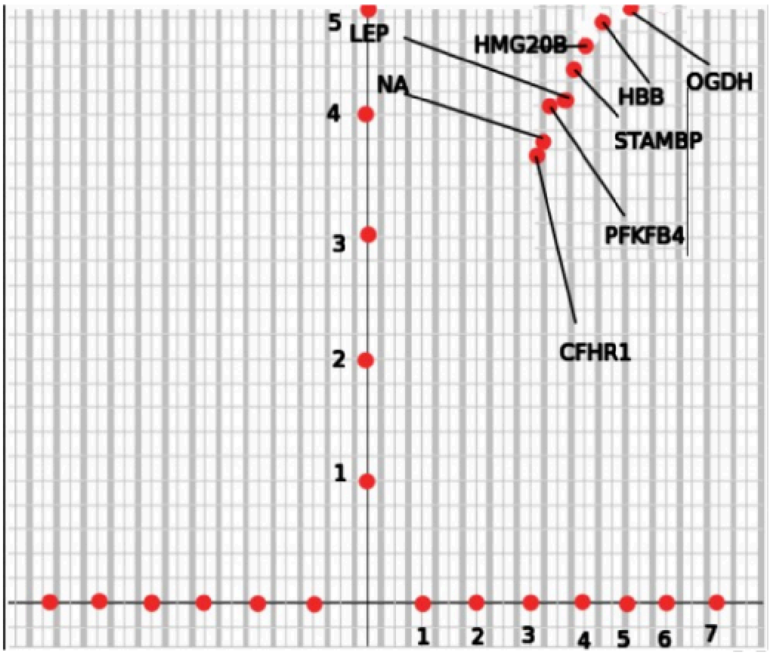
```
ID              adj.P.Val      P.Value          t          B      logFC   Gene.Symbol

201282_at      0.0074911     1.281663e-05  6.773810 -4.468046  3.031999    OGDH
217232_x_at    0.0088708     5.409429e-05  5.868340 -4.477733  2.969720    HBB
210719_s_at    0.0097080     6.672291e-05  5.742075 -4.479320  3.396428    HMG20B
235361_at      0.0187080     7.161377e-05  5.699805 -4.479866  3.382755    STAMBP
207092_at      0.0339791     1.693045e-04  5.197041 -4.486967  4.011486    LEP
228499_at      0.0459791     1.903590e-04  5.130082 -4.488003  2.510947    PFKFB4
235719_at      0.0907918     2.133572e-04 -5.065256 -4.489028 -2.156150    CYP4V2
230445_at      0.0997918     2.736306e-04  4.924930 -4.491321  2.381681    BTBD17
1569386_at     0.1299791     3.180007e-04 -4.840876 -4.492746 -2.210268
215388_s_at    0.2397918     3.454539e-04 -4.794777 -4.493543 -3.641796    CFHR1
```

a. Which genes would you call differentially expressed? Explain the criteria used to make this.

The most differentially expressed genes are the one without name and CFHR1 because are the ones with smaller P-value.

b. Draft an approximate Volcano Plot depicting all the genes. Explain what does the Volcano plot represent and why is it interesting (HINT: B ~ -log(P.Value))

Volcano plot represent the P-values for the expression of our genes. It is interesting because it is a very visual way for knowing which are the most expressed genes in our study.



Part III. Epigenomics, single-cell and proteomics

Write a definition of epigenetics.
Epigenetics is the study of the epigenetic variations in the genome of any organism. These genetic variations in DNA or in Histones can affect the expression of certain genes without inferring in the DNA sequence.

How would you expect to find the promoter region of a highly transcribed genes in terms of nucleosome positioning, DNA methylation and histone modifications?
Highly transcribed genes correspond to euchromatin. Nucleosomes will have low occupancy, DNA demethylated and histone acetylated.

Which is/are the chromatin state/s most highly associated with the following histone modifications:
H3K4me3 - Active promoters
H3K27ac - Active promoters and enhancers
H3K36me3 - Transcription (Tx) elongation
H3K9me3 - Heterochromatin
H3K27me3 - Repression state

Which genomics elements contain mostly methylated CpG sites?
CG islands contain methylated CpG sites. 70 % of proximal promoters are in CG islands. 60 % of genes have GC islands

What is a TAD?
A topologically associating domain (TAD) is a self-interacting genomic region, meaning that DNA sequences within a TAD physically interact with each other more frequently than with sequences outside the TAD.

Describe one technique to study chromatin interactions (3-C, 4-C, 5-C, Hi-C or GAM).
ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.

Which region of the genome is particularly useful and used for characterizing microbiomes richness and why?

What is an OTU?
Operational taxonomic unit Clusters of (uncultivated or unknown) organisms, grouped by DNA sequence similarity of a specific taxonomic marker gene

What is the G-value / C-value enigma and what explanation would you provide to solve it?
Lack of correlation between the number of protein-coding genes among eukaryotes and their relative biological complexity. The reason is found in the RNA world, which is more complex and it is related to gene regulation.

Which protein property is used to separate each dimension in a 2D-SDS-PAGE electrophoresis?
Isoelectric point (pH) and weight.

Pair with arrows:
Ion Source → Electrospray, MALDI
Mass filter → TOF, Ion Trap, Quadrupole

Describe the purpose of the second MS step in an MS/MS applied to protein identification.
Ions from the MS1 spectra are selectively fragmented and analyzed by a second MS stage to generate the spectra for the ion fragments.

Name three categories of epigenetic modifications.
Methylation at CpG sites:

- ☐ occurs at similar extent in all organisms.
- ☐ is always associated to silencing of gene expression.
- ☐ is irreversible
- ☑ none of the above

In a 16S rRNA next generation sequencing effort to determine microbial richness from a sample, can we consider a difference in one single nucleotide position between two sequences as evidence of the presence of two species? Give reasons for you answer.

How would you construct and interpret a rarefaction curve from next generation sequencing metagenomics data?
Why do we digest proteins to peptides before MS instead of running the whole molecule?
The problem of MS is: the higher the mass, the higher the error. So, when running the full molecule with MS, we will get a single peak (the mass of the whole protein), but that does not mean that it is trustworthy. In order to have more reliable results, it is highly recommendable to break the protein into fragments, run MS and even break those peptides and run MS again to better analyze our samples.

Digesting proteins into peptides before mass spectrometry (MS) is a crucial step in proteomics for several reasons:

1. Improved Ionization: Peptides ionize more efficiently than whole proteins, resulting in better sensitivity and more reliable detection in MS.
2. Increased Coverage: Digestion into smaller peptides allows for more extensive and comprehensive coverage of the protein's sequence.
3. Resolution and Accuracy: MS analysis of peptides provides higher resolution and accuracy, enabling the identification and quantification of proteins with greater confidence.
4. Complexity Management: The complexity of a protein sample is reduced by breaking down large proteins into smaller, more manageable peptides, facilitating more effective separation and analysis.
5. Database Matching: Peptide mass fingerprints and sequences are easier to match against protein databases, aiding in the identification of proteins.

In summary, digesting proteins into peptides enhances the analytical capabilities of MS, allowing for more precise and comprehensive proteomic studies.

Explain why and how isotope labelling can be used to quantify relative protein amounts among conditions.

What is the Jaccard similarity and how do we calculate it using genomic coordinates?
The Jaccard similarity index (sometimes called the Jaccard similarity coefficient) compares members for two sets to see which members are shared and which are distinct. It’s a measure of similarity for the two sets of data, with a range from 0% to 100%. We can use it to calculate two sets of binding sites.

What solution do you know it is used in single-cell genomics to deal with the problem of PCR duplicates?
Using UMI, unique molecular identifiers

If you can only work with frozen tissue, would you use single-cell or single-nuclei RNA-seq? Why?
Single - nuclei RNA-seq, You can extract nuclei from frozen tissue, but not entire cells whose membranes break when frozen.

Mention three advantages of using single-cell genomics versus bulk genomics.

- Study Rare Cell Types Obscured In bulk tissue
- Determine trajectory of differentiation
- Identify Cell Type Specific Effects Under a Comparison(treatments, diseases, evolution, etc)

What kind of information is provided by ATAC-seq?
ATAC-seq can be used to get some mixed samples directly from the environment and study the expression of the genes in it. With it, we can evaluate which genes are expressed in a certain tissue-specific cell type and use clustering approaches to understand the expression resemblances among different types.

What is the Louvain algorithm designed for in the context of single-cell?
Louvain algorithm helps to identify “neighboring” cells and therefore, cells that may form a cluster (in other words, cells that share some gene expression characteristics and consequently have similar properties and functions). Some cells interact more with some other cells than others. Therefore, it is ideal to identify which are more associated between them and which are not. Louvain algorithm calculates the QC and in each step it checks that this value is improving or not:

- $\Delta QC > 1$ → the last change improves the network, it should be considered
- $\Delta QC < 1$ → the last change does not improve the network, it should not be considered

What is the Louvain algorithm designed for in the context of single-cell?

In single-cell genomics, the Louvain algorithm is used for **community detection in graphs**. It is specifically designed to:

1. Cluster Cells: By detecting communities in the graph where nodes represent individual cells and edges represent similarity between cells, the Louvain algorithm clusters cells into distinct groups based on their gene expression profiles.
2. Identify Cell Types: These clusters often correspond to different cell types or states, which helps in the characterization of cellular heterogeneity within a sample.
3. Scale Efficiently: The Louvain algorithm is efficient and scalable, making it suitable for large single-cell RNA-seq datasets where the number of cells can be very high.

Explain all you know about the specific omics technique you presented.
The technique I presented was Methyl-Seq. This technique is based on bisulfite sequencing, a sequencing technique where we can know where are methylations located in any given genome. This bisulfite technique is based on a discovery from the 70s. It was discovered that the addition of Sodium Bisulfide to Cytosines made them turn into Uracil. Later in the 90s it was discovered that this conversion wasn't produced if the cytosine was methylated.
So basically, in this method we want to add Sodium bisulfite to every Cytosine and see if they turn into Uracil or not. Later, we would have to compare our reads from NGS with normal reads of that specie and compare which Guanines are real and which are from methylated Cytosines that did not turned into Uracil. This technique is quite reliable because 70-80% of CpG sites are methylated, so Cytosines are a good target if we are looking for methylations in the genome. This technique is harder and more slow than others because here we have to work with a whole genome, while other techniques are specific for a certain region.

What are UMIs, and what purpose do they serve?
UMIs, or Unique Molecular Identifiers, are short sequences of random nucleotides added to DNA or RNA molecules during sequencing library preparation. They serve the purpose of uniquely tagging each original molecule, allowing researchers to distinguish between true biological duplicates and PCR duplicates. This helps improve the accuracy of quantifying gene expression levels and detecting mutations by reducing biases introduced during amplification steps in high-throughput sequencing.

What are the advantages of using nuclei instead of cells in single- nuclei RNA-seq?

- **Accessibility**: Nuclei are more easily isolated from challenging tissues, such as frozen or hard-to-dissociate samples.
- **Integrity**: The technique minimizes RNA degradation, preserving the transcriptome's integrity.
- **Complexity**: It reduces technical noise and artifacts from cytoplasmic RNA, providing a clearer view of the nuclear transcriptome.
- **Specificity**: Allows for the study of gene expression in specific cell types, even in heterogeneous tissues.

Methylation at CpG sites:

a. occurs at similar extent in all organisms.
b. is always associated to silencing of gene expression.
c. is irreversible
d. none of the above

Rewrite the following terms in hierarchical order:
Nucleosomes, A/B compartments, FIREs, TADs, chromosome territories.
Chromosome territories > A/B compartments > TADs > FIREs > Nucleosomes

Nucleosomes: The basic unit of DNA packaging, consisting of a segment of DNA wound around a core of histone proteins.
FIREs (Frequently Interacting Regions): Specific regions within the genome that frequently interact with each other, often contributing to gene regulation.
TADs (Topologically Associating Domains): Regions of the genome that interact more frequently with each other than with other regions, playing a role in the regulation of gene expression.
A/B compartments: Larger structural domains within chromosomes that are either gene-rich and actively transcribed (A compartments) or gene-poor and less active (B compartments).
Chromosome territories: Distinct regions within the nucleus occupied by individual chromosomes, maintaining a specific spatial organization.

What kind of information is provided by ATAC-seq?

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a technique used to study chromatin accessibility. The key information provided by ATAC-seq includes:

1. Open Chromatin Regions: It identifies regions of the genome that are accessible to transcription factors and other DNA-binding proteins, which are often indicative of regulatory elements such as promoters, enhancers, and insulators.
2. Nucleosome Positioning: ATAC-seq can reveal the positioning of nucleosomes across the genome.
3. Transcription Factor Binding Sites: By identifying footprints of transcription factors within accessible chromatin, ATAC-seq helps in understanding gene regulation networks.
4. Chromatin Dynamics: It provides insights into how chromatin accessibility changes under different conditions or between different cell types, which is crucial for understanding developmental processes and disease mechanisms.

What solution do you know that is used in single-cell genomics to deal with the problem of PCR duplicates?

In single-cell genomics, particularly in single-cell RNA sequencing (scRNA-seq), **Unique Molecular Identifiers (UMIs)**are used to address the problem of PCR duplicates. UMIs are short, random nucleotide sequences added to each RNA molecule before amplification. This approach offers several advantages:

1. Quantification Accuracy: UMIs allow for accurate quantification of RNA molecules by distinguishing between original molecules and PCR duplicates.
2. Error Correction: They help in correcting amplification biases and sequencing errors, leading to more reliable and reproducible data.
3. Data Robustness: By mitigating the impact of PCR duplicates, UMIs enhance the robustness of single-cell transcriptomic data, enabling better downstream analysis and interpretation.

Which protein property is used to separate each dimension in a 2D-SDS-PAGE electrophoresis?

1. First Dimension - Isoelectric Focusing (IEF):
   - Property: Isoelectric Point (pI)
   - Explanation: Proteins are separated based on their isoelectric point, which is the pH at which the protein has no net charge. During isoelectric focusing, proteins migrate in a pH gradient until they reach the pH that matches their pI.
2. Second Dimension - SDS-PAGE:
   - Property: Molecular Weight
   - Explanation: In the second dimension, proteins are separated based on their molecular weight. SDS (sodium dodecyl sulfate) is used to denature the proteins and impart a uniform negative charge relative to their length. When subjected to an electric field in the polyacrylamide gel, proteins are separated according to their size, with smaller proteins migrating faster and farther through the gel matrix.

By combining these two separation methods, 2D-SDS-PAGE provides a high-resolution technique for analysing complex protein mixtures, as illustrated in the image provided.