Report for LAB6: Omics Applications
Bioinformatics

Ricardo Brancas 83557
Margarida Ferreira 80832
Felipe Gorostiaga 95383
Benedict Schubert 95034

17th December 2019

Group I.

I. a) Quality assessment of FASTQ

In figure 1 we present the per base sequence quality graphs obtained using the tool FastQC. The sequences have reasonably good quality scores, although there is some disparity between them. In particular the first sequence has slightly better quality than the second one, which is something one should be aware of.

I. b) Estimated gene expression vs. provided read counts

We chose to use the aligner Kallisto, together with the annotated human transcriptome from Ensembl [1]. We chose the cDNA sequences, as advised in the Kallisto FAQ [2].

To compare the transcript expression with the gene expression counts given, we downloaded a mapping from the Ensembl transcript identifiers to gene names. We used Ensembl Biomart [3] to get this mapping. We then took the transcript count estimates and the mapping and created a summarised table containing the total estimated number of reads for each gene. Finally, comparing this table with the read count table we were given, results in the graph in figure 2. Figure 2a shows all the estimated and provided read counts plotted against each other: each black mark represents a gene, the value on the y-axis is the provided read count, while the value on the x axis is the estimated read count; therefore, the closer a mark is to the $y = x$ line, the better the estimation. Analysing this graph, we can see that most genes are correctly estimated, as most marks lie within the $y = x$ line. Most outliers are from read counts of 0 which most likely represents a mismatch between the transcript identifiers and the gene names. In Figure 2b we show the same data, this time plotted in logarithmic scale so the values are more easily seen. Note that, in order to plot in a logarithmic scale, all the zero-valued read counts must be removed. In this plot we can again see that most of the estimates are accurate, even though there are a few outliers. We can also see that it appears that the higher the read count the better the estimation. This however is no true: what we are seeing is a result of the logarithmic scale: the errors are of the same order of magnitude, but since the scale is much larger, they seem smaller when compared to the actual read count values.

[TODO: MACF: Add y=x line to the plots?]

Group II.

II. a) Read coverage and library complexity

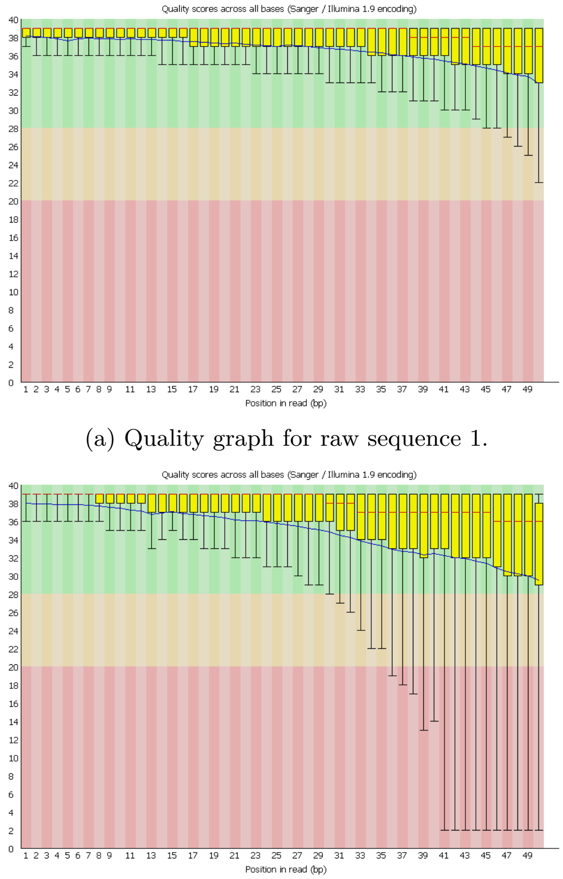
To analyse the read coverage, we created the histogram in figure 3. We can see that the count values range from roughly 40 million to 180 million. We can also see that most samples have more than 60 million reads, although a very small number has only around 40 million. This might mean that in these samples lowly expressed genes are below the detection threshold. We can also see that there are a lot more genes with read counts in the 60 to 100 million range than in the rest of the domain, which indicates that we will probably need some kind of normalisation of the data.

Library complexity refers to the number of unique DNA fragments present in a given library. To compute library complexity, we used a sampling-based approach. For tissue sample, we sampled the genes with probability proportional to the read count of each gene for that patient. In figure 4 we plot the library complexities for all samples. Analysing the figure we can see that, in general, in samples from normal tissues we hit more genes with fewer samples, which means that the genes are more uniformly represented. On the other hand, in deseased tissues, we get a lower

[1] `ftp://ftp.ensembl.org/pub/release-98/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz`

[2] `https://pachterlab.github.io/kallisto/faq`

[3] `http://www.ensembl.org/biomart/martview/a797838aa8255de1efa6fb6d11322eb5`



(a) Quality graph for raw sequence 1.

(b) Quality graph for raw sequence 2.

Figure 1: Quality graphs for the raw sequences, as obtained in FastQC.

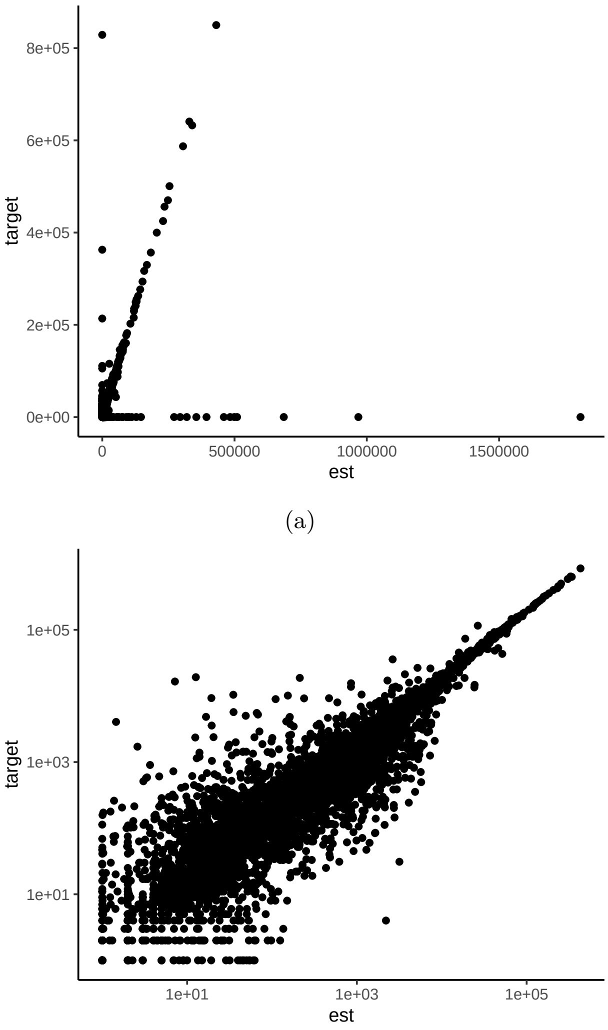
percentage of all the detected genes with the same number of samples, showing that the read counts for the genes in deseased tissues is farther from a uniform distribution: there are a few genes with a much higher read count than the others.

II. b) Normalisation of the data

To normalise the data, we used the Trimmed Mean of M (TMM) method. In order to evaluate our normalisation, we produced a series of plots which help us visualise how the data changed after the normalisation.

First, we plotted the histogram of the normalised read counts. When compared with the histogram resulting of the raw read counts (in figure 3), we can se that ..... **[TODO: MACF: plot histogram of the normalised read counts (hopefully it is flatter)]**

We produced a plot where the distance between samples corresponds to the leading Biological Coefficient of Variation (BCV), using the function `plotMDS`. The result can be seen in figure 5. There is a clear separation between the samples of normal tissues and the samples taken from tumours. It also appears that there are two distinct groups of samples from tumours.



(a)

(b) Log-log plot after removing 0 values.

Figure 2: Comparison between the estimated counts and the provided values.

Finally, we wanted to ensure that after normalisation there was no difference between the highly expressed and lowly expressed genes. To so so, we coloured the marks according to their read count.........

[TODO: MACF: different colours for 2 groups of genes with varying number read counts: produce histogram with the columns on the right in one color and the columns on the left another column (separate by mode/median=?). Then plotMDS using the same colours for the respetive marks.]

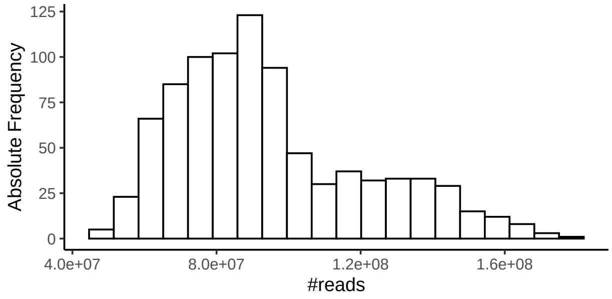


Figure 3: Read counts histogram.

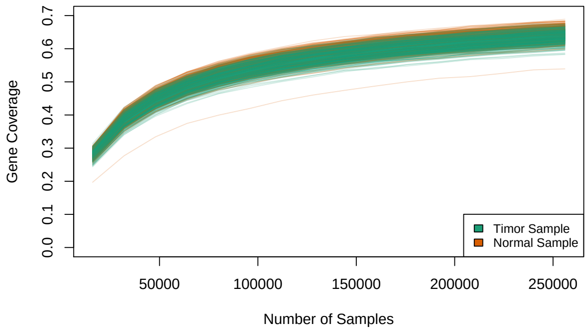


Figure 4: Library complexity plot.

II. c) Phenotypic traits and genes that dominate variance

[TODO: MACF: try to identify what separates the 2 groups of tumours in the MDS plot.]

II. d)

Group III.

III. a)

We want to show how the gene expression for some genes relates to the binary classification for the related immunohistochemistry-based tests in the patient data. To do so, we will produce 3 plots.

First, in figure 6a we have a scatter tissue sample the read count of ESR1 on the x-axis and the read count of ERS2 on the y-axis. Each mark is colored according to the presence of estrogen receptors in the respective tissue sample. We can see by this graph that the expression of ESR1 is highly related to the presence of estrogen receptors: to be continued.....

[TODO: MACF: try to add ROC curves to these plots, or alternatively create another plot with ROC curves.]

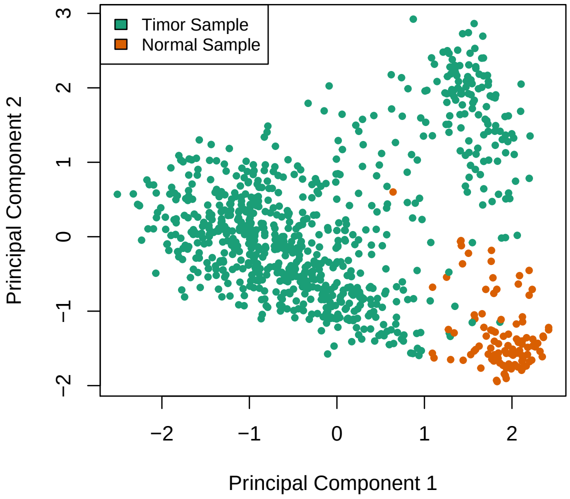


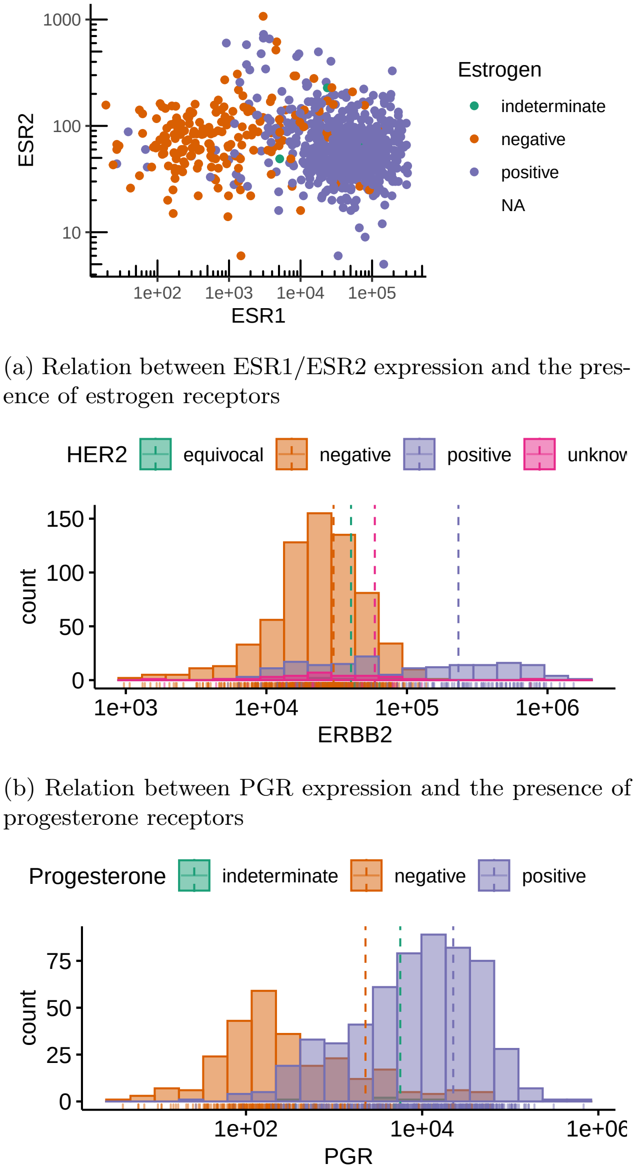
Figure 5: Multidimensional Scaling plot of the data after normalizing.

III. b)

Just do the survival analysis according to the PAM50 groups.

III. c)

	logFC	AveExpr	t	P.Value	adj.P.Val	
FOXC1	-3.982	3.321	-15.80	0	0	101
SFRS13B	-1.941	0.854	-15.29	0	0	95
MLPH	4.168	7.127	15.03	0	0	92
SIDT1	3.183	4.731	14.59	0	0	87
FOXA1	5.231	7.212	14.48	0	0	86
PRR15	4.545	3.818	14.46	0	0	85
AR	4.472	4.215	14.43	0	0	85
ERBB2	3.179	8.391	14.18	0	0	82
PGAP3	2.555	5.432	14.10	0	0	81
ABCC11	6.595	3.436	14.06	0	0	81



(a) Relation between ESR1/ESR2 expression and the presence of estrogen receptors

(b) Relation between PGR expression and the presence of progesterone receptors

(c) Relation between ERBB2 expression and the presence of HER2 protein

Figure 6: Relation between the cognate genes and the immunohistochemistry-based tests.

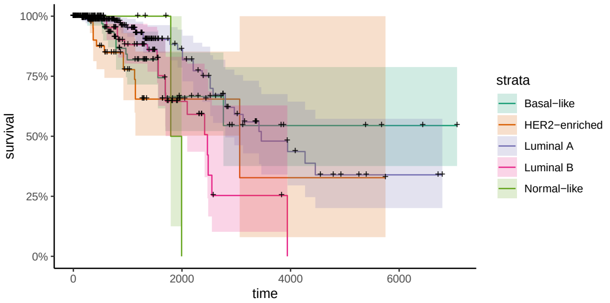


Figure 7: Hello

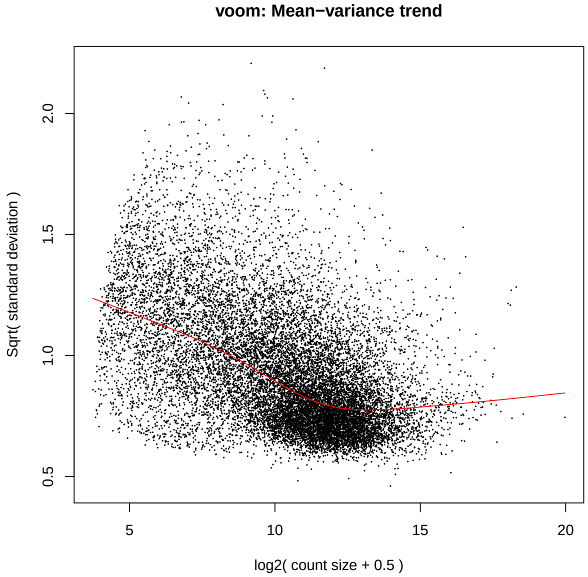


Figure 8: Hello

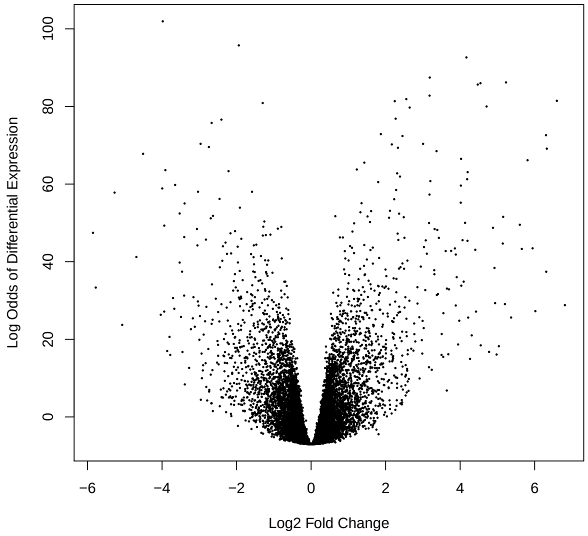


Figure 9: Hello