Report for LAB6: Omics Applications
Bioinformatics

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

16th December 2019

Group I.

a)

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.

b)

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 had to download 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. Analysing this graph, we can see that most genes are linearly correlated. Most outliers are from read counts of 0 which most likely represents a mismatch between the transcript identifiers and the gene names.

[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`

Group II.

a) Read coverage and library complexity

To analyse the read coverage, we created the histogram in figure 3. We can see that most samples have more than 6×10^7 reads, although a very small number has only around 4×10^7. This might mean that in these samples lowly expressed genes are below the detection threshold.

To compute library complexity, we used a sampling-based approach. For each patient, 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, samples from normal tissues has a

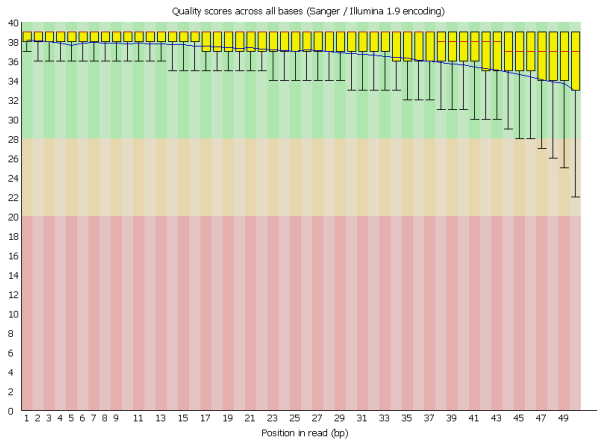
b)

To normalise the data, we used the Trimmed Mean of M (TMM) method. To analyse our data after normalisation we produced a plot where the distance between samples correponds 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 tumors. It also appears that there are two distinct groups of samples from tumors.

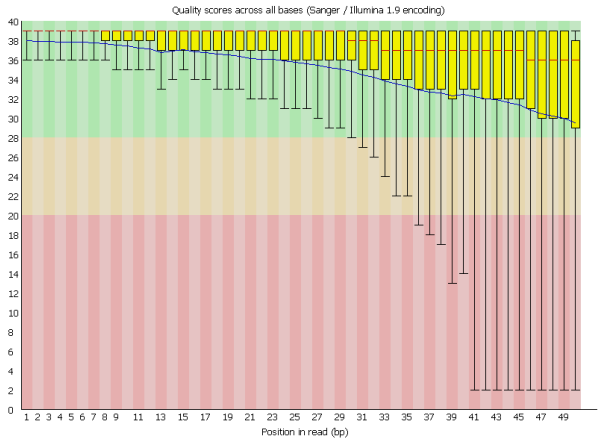
Group III.

a)

We need to show how the gene expression for those genes is related to the binary classification for the same tests in the patient data.

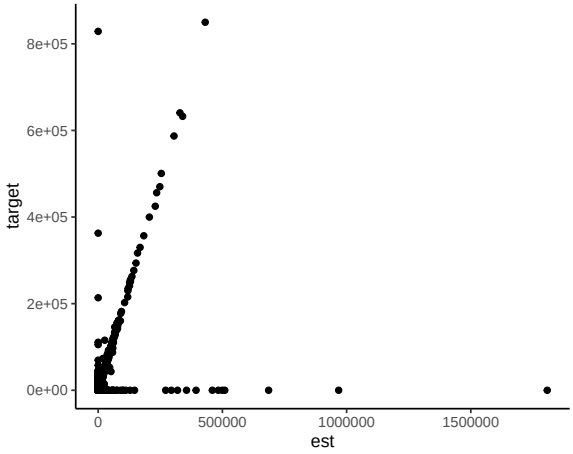


(a) Quality graph for raw sequence 1.

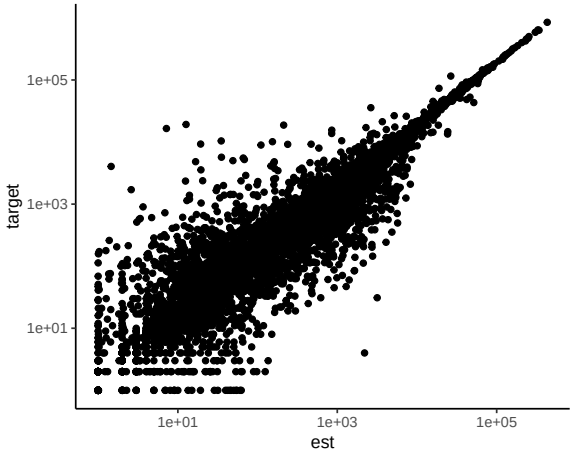


(b) Quality graph for raw sequence 2.

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



(a)



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

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

b)

Just do the survival analysis according to the PAM50 groups.

c)

	logFC	AveExpr	t	P.Value	adj.P.Val	B
FOXC1	-3.982	3.321	-15.80	0	0	101.94
SFRS13B	-1.941	0.854	-15.29	0	0	95.74
MLPH	4.168	7.127	15.03	0	0	92.64
SIDT1	3.183	4.731	14.59	0	0	87.45
FOXA1	5.231	7.212	14.48	0	0	86.20
PRR15	4.545	3.818	14.46	0	0	85.99
AR	4.472	4.215	14.43	0	0	85.65
ERBB2	3.179	8.391	14.18	0	0	82.80
PGAP3	2.555	5.432	14.10	0	0	81.89
ABCC11	6.595	3.436	14.06	0	0	81.46

Group IV.

Group V. One more question

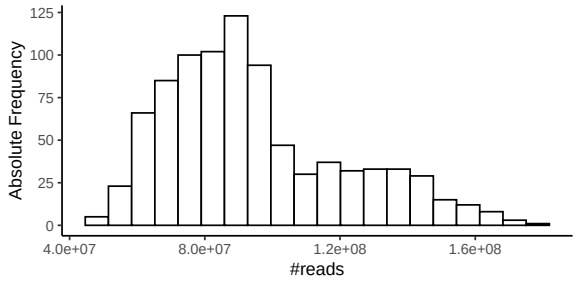


Figure 3: Read counts histogram.

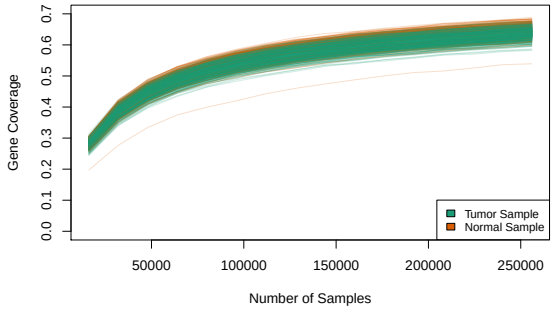


Figure 4: Library complexity plot.

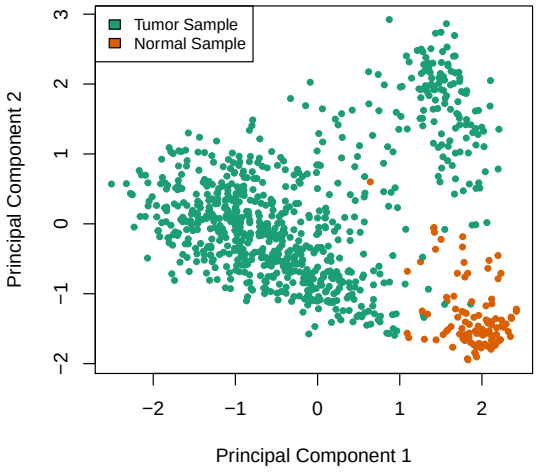
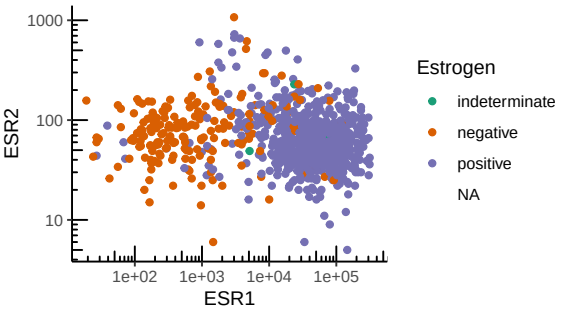
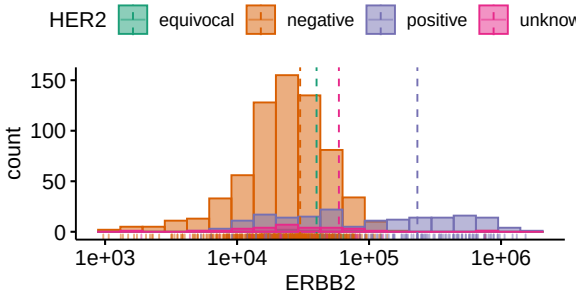


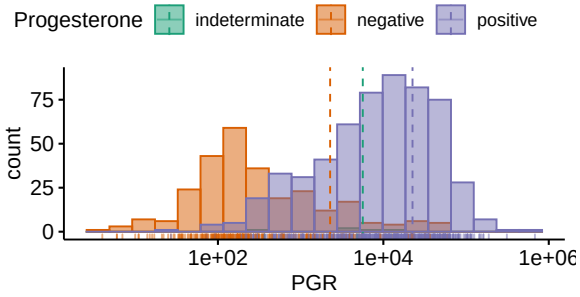
Figure 5: Multidimensional Scaling plot of the data after normalizing.



(a) a



(b) b



(c) c

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

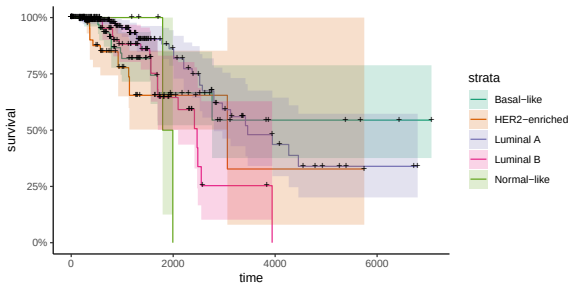


Figure 7: Hello

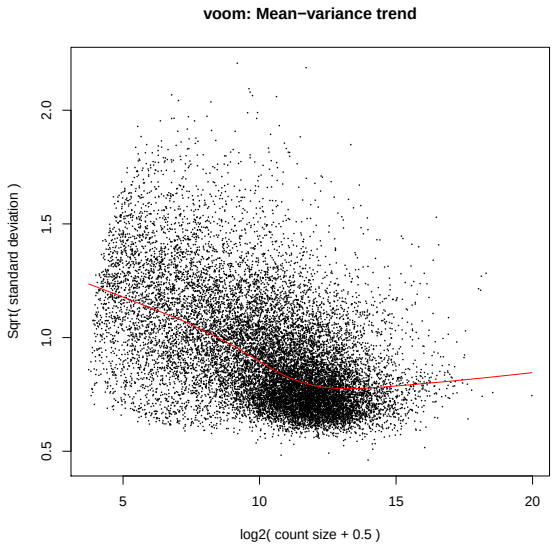


Figure 8: Hello

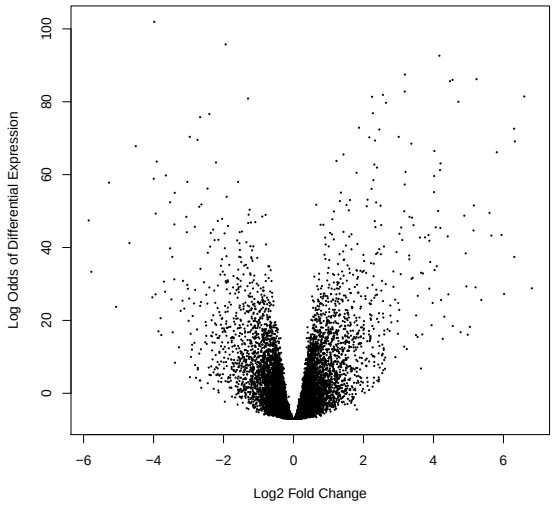


Figure 9: Hello