Report for LAB6: Omics Applications

Bioinformatics

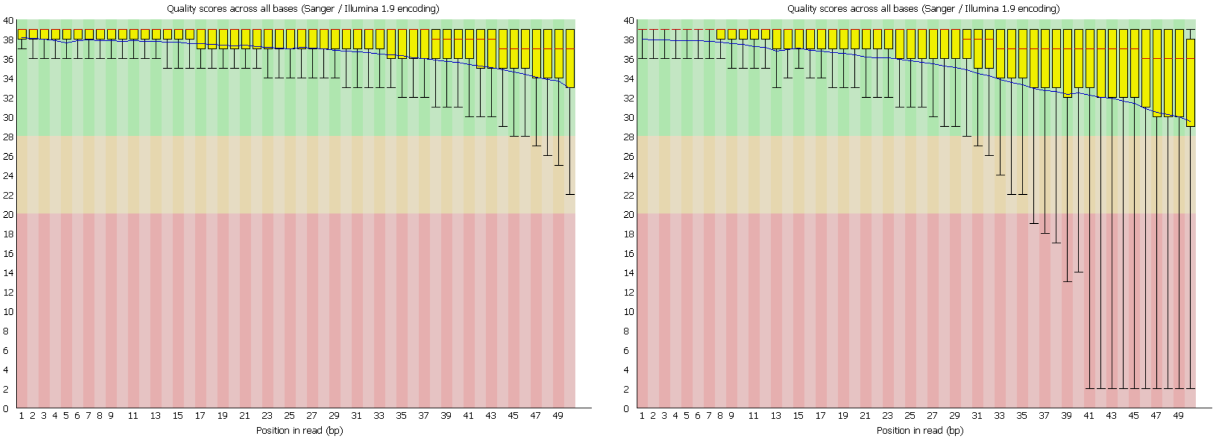
Ricardo Brancas 83557 | Margarida Ferreira 80832 | Felipe Gorostiaga 95383 | Benedict Schubert 95034

14th 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.



(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.

b)

We chose to use 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 transfer 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.

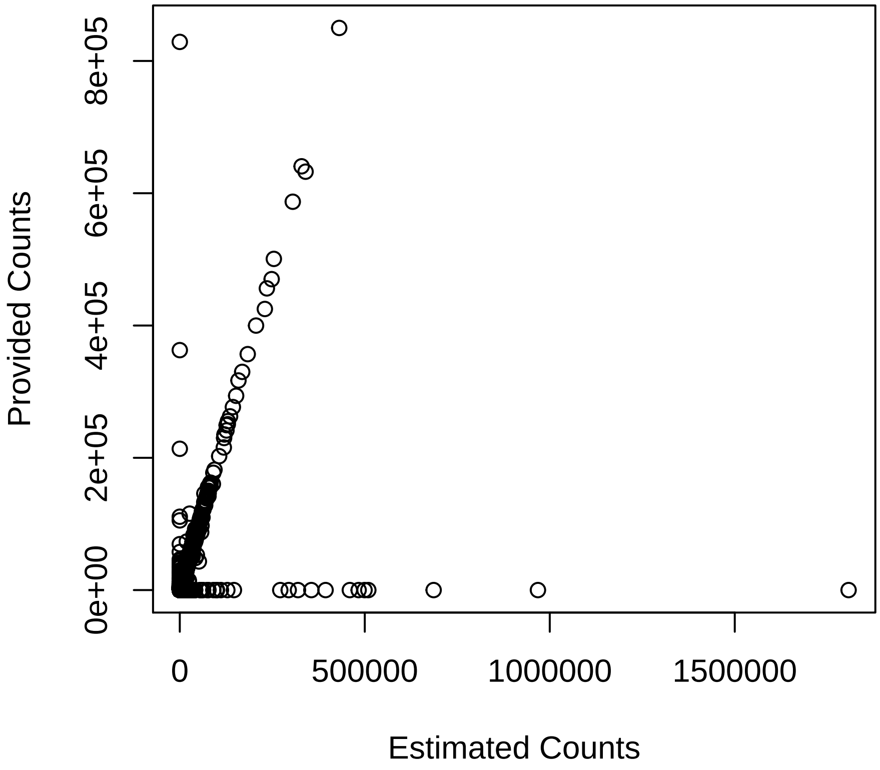


Figure 2: Count comparison.

[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.

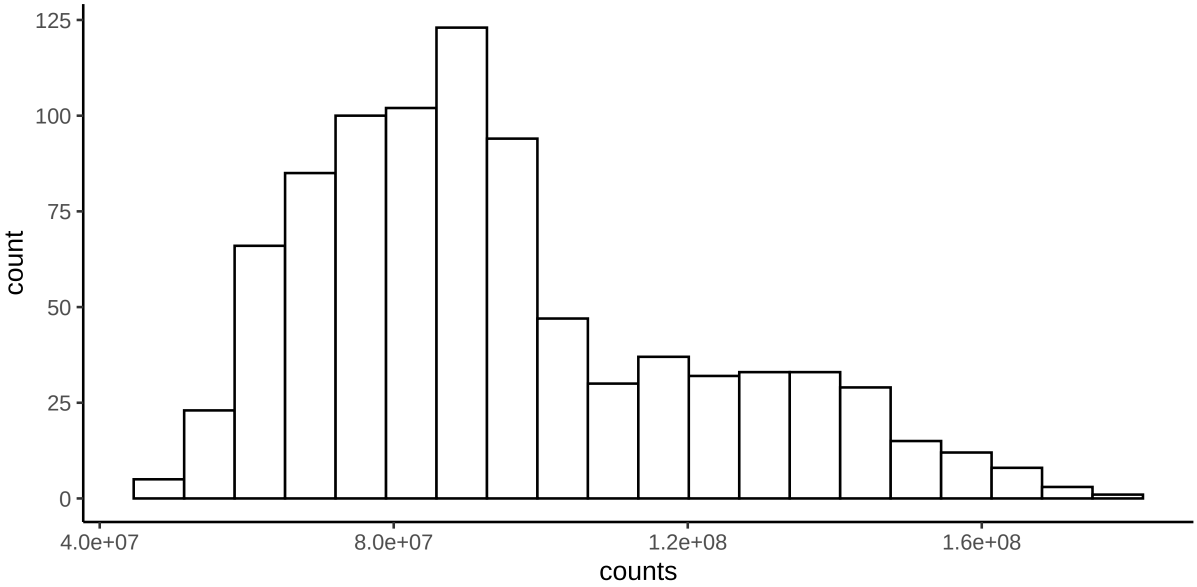


Figure 3: Hello

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 (presented in red) and the samples taken from tumors (in blue). It also appears that there are two distinct groups of samples from tumors.

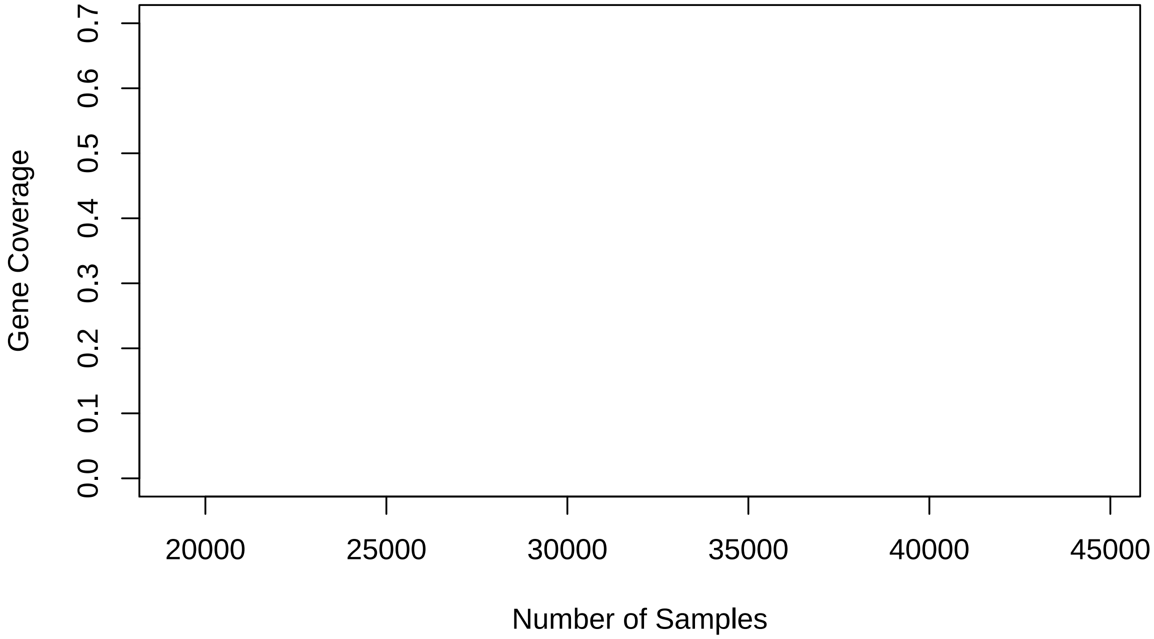


Figure 4: Hello

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.

b)

Just do the survival analysis according to the PAM50 groups.

Group IV.

Group V. One more question

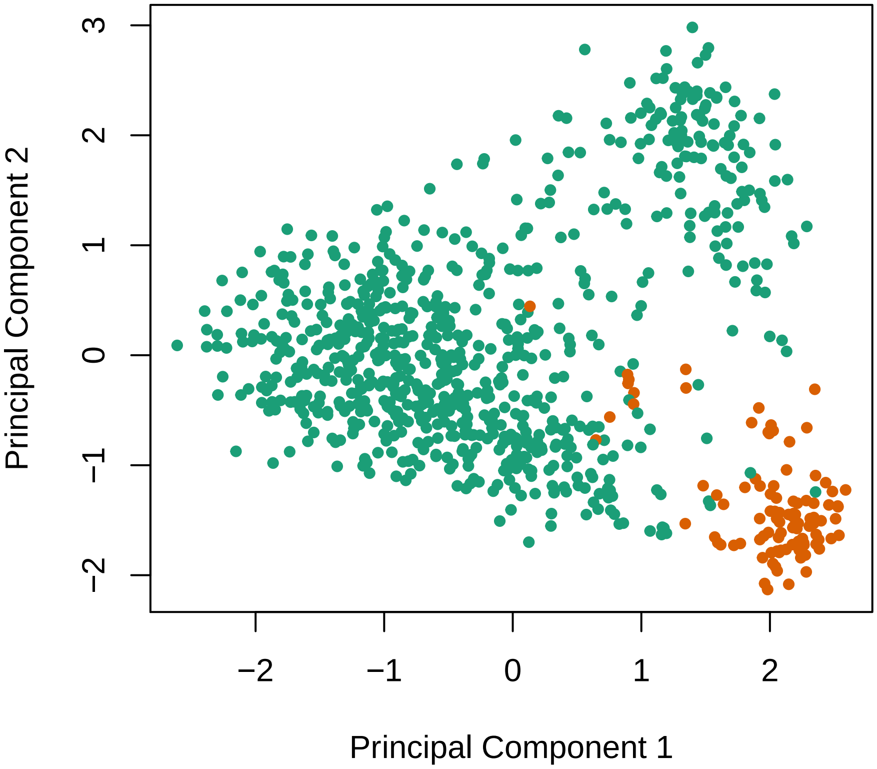
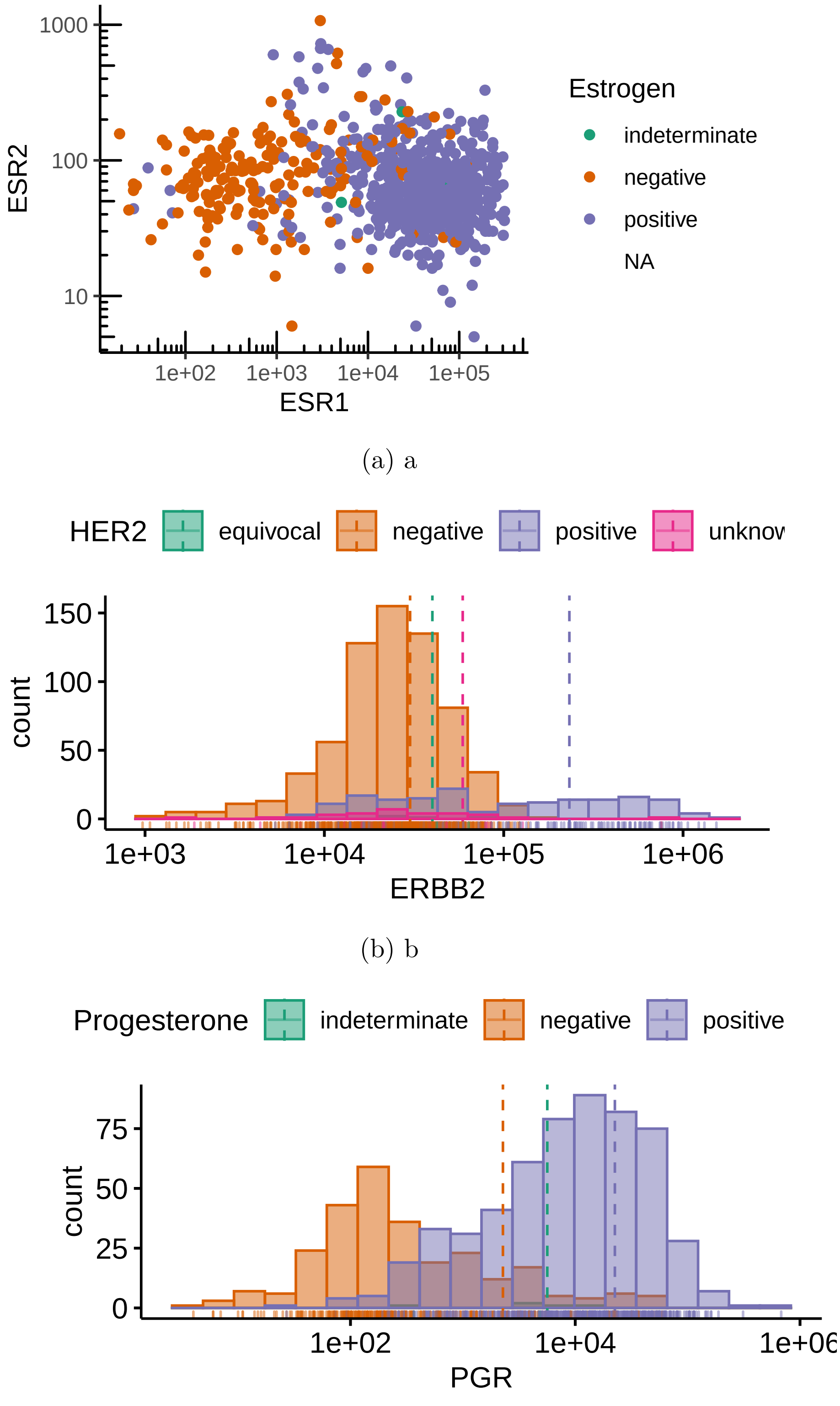


Figure 5: Hello



(a) a

(b) b

(c) c

Figure 6: Count comparison.

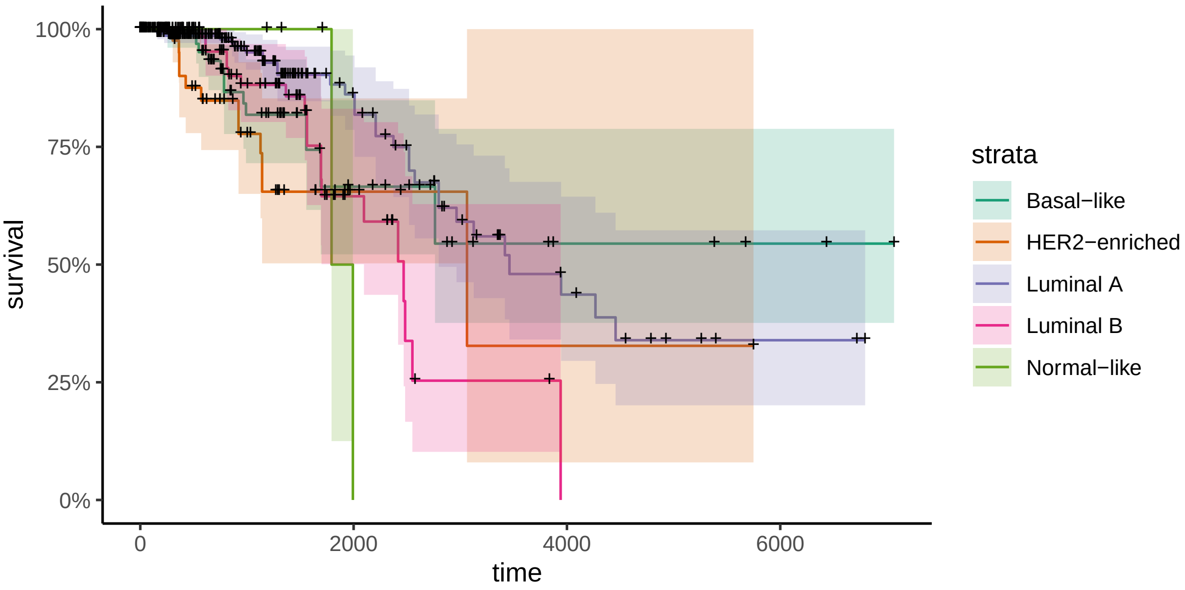


Figure 7: Hello