Lab 6: Colorectal Cancer Analysis

L. Ganser[1], M. Fernandes[2], R. Moreira[2], and V. Dalovic[2]

[1]École Polytechnique Fédérale de Lausanne, Switzerland
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal

4 November 2022

This report describes the work done for lab 6 by group 24. The dataset used comes from The Cancer Genome Atlas (TCGA)[1] and contains gene expression data from 623 patients with colorectal cancer (CRC). The goal of this lab is to perform a differential expression analysis of the data and to identify genes that are differentially expressed between the two groups of patients: those with a high risk of recurrence and those with a low risk of recurrence. We use the R package DESeq2 to perform the analysis and we use the R package ggplot2 to visualize the results. We also use the R package limma to perform pathway analysis of the differentially expressed genes. Finally, we use the R package clusterProfiler to perform a gene ontology analysis of the differentially expressed genes.

1 Quality Control

The quality control was done using FastQC. For the first Fastq file, we get warnings in the *Per tile sequence quality* and *Overrepresented sequences* and failures for *Per base sequence content* and *Sequence Duplication Levels*.

Looking at *Per tile sequence quality*, we see a pattern of lines emerge in both files. Because these samples were sequenced in Illumina, the encoding retains the original sequence identifiers, which allows us to look at the quality for each tile position in the flow cell.[2]. The pattern of lines in the graph indicates that the quality of the reads is not uniform across the flow cell. This

is a common problem in Illumina sequencing and is caused by the fact that the flow cell is not perfectly flat. [2]. This problem can be solved by using a different flow cell or by using different sequencing technology. However, this problem is usually negligible. LACKS CITATION

2 Alignment

The alignment was done using Kallisto. The reference genome used was the human genome Ch38. The results are shown in the following table.

References

[1] The Cancer Genome Atlas. The cancer genome atlas. http://cancergenome.nih.gov/, 2012.

[2] Babraham Institute. Fastqc. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, 2012.