Differential gene expression analysis between triple-negative breast cancer (TNBC) tumours and healthy tissues.

Saskia Perret-Gentil

September 2021

Abstract

The aim is to find genes differentially expressed in triple-negative breast cancer (TNBC) when compared to healthy breast tissues and to identify gene ontology (GO) terms enriched for those genes. To do so, the reads were mapped to the reference genome and the number of reads per gene were count. The counts were used to perform a differential expression analysis and a overrepresentation analysis to identify the main GO terms.

1 Introduction

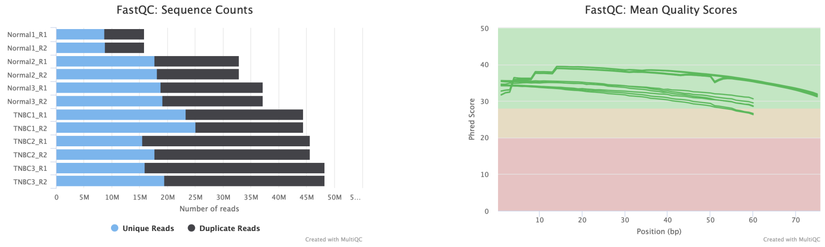
TNBC is a subtype of breast cancer were tumours lack expression of the Estrogen receptor (ER), the progesterone-receptor (PR) and the Human Epidermal Growth Factor Receptor 2 (HER2). This subtype, which can be differentiated from other breast cancer subtypes immunohistochemically, arrises more frequently to younger patients and is characterised by its increased aggressiveness with a shorter survival periods and a higher recurrence rates. [1]

2 Material and methods

2.1 Data

Data used is a subset from Eswaran et al. 2012 [1], composed of fastq files which were downloaded through the Gene Expression Omnibus (GEO), accession GSE52194. The libraries were sequenced on an Illumina HiSeq 2000 in paired-end mode.

The subset includes 3 replicates from TNBC human breast tumors and 3 healthy samples. The quality of the subset were assess with FastQC [8] (v. 0.11.9). The reference genome (assembly GRCh38) and associated annotation



(a) (b)

Figure 1: MultiQC summaries of FASTQC reports on the quality of samples.

were download from the Ensembl ftp site. Checksums were computed and compared to the values in the CHECKSUMS file on the ftp server.

2.2 Mapping reads to the reference genome.

The reference genome was indexed using Hisat2 [3] (v. 2.2.1). Then the reads were mapped to the reference genome also using Hisat2. SAMtools [4] (v. 1.10) was then used to convert the resulting sam files to bam format. Bam files were then sorted and indexed, also using SAMtools. A table of counts containing the number of reads per gene was then produced using featureCounts [5], the bam files and the annotation file.

2.3 Analysis

The table of counts was load in R [7] (v. 4.1.0) and the package DESeq2 [6] (v. 1.32.0) was used to perform the differential expression analysis.

3 Results

3.1 Data quality

MultiQC [2] (v. 1.8) was used to summarize the FastQC reports (see figure 1). The average base quality was good but with a tendency to decrease a little near the end of the reads (see figure 1b).

3.2 Mapping reads to the reference genome.

When mapping to the reference genome, further differences were noticed between the `Normal` and the `TNBC` samples. First the alignment rates was $\sim 96.4\%$ for the `Normal` samples, and $\sim 87.5\%$ for the `TNBC` samples. For the `Normal` samples, $\sim 87\%$ of the reads aligned concordantly exactly once and $\sim 5\%$ aligned concordantly more than one time. For the `TNBC1` sample 50.38% of the reads aligned concordantly exactly one time, and 28.20% aligned concordantly more than one time. For TNBC2 and TNBC3, it is only ~ 32% of the reads that aligned concordantly exactly once and ~ 41% that aligned concordantly more than one time (see output_mapping_hisat2.md in the Supplementary materials). So there is a strong evidence of multimapped reads for the TNBC samples.

When the number of reads per gene were count with featureCounts, the average proportion of reads overlapping with annotated genes for Normal samples is around 70%, whereas the successfully assigned alignments drop to ~ 8% for the TNBC samples, and almost exclusively because multimapping reads were unassigned (~ 80% unassigned due to multimapping for TNBC samples against only ~ 18% for Normal samples, see featureCounts_summary.md in the Supplementary materials).

3.3 Exploratory data analysis

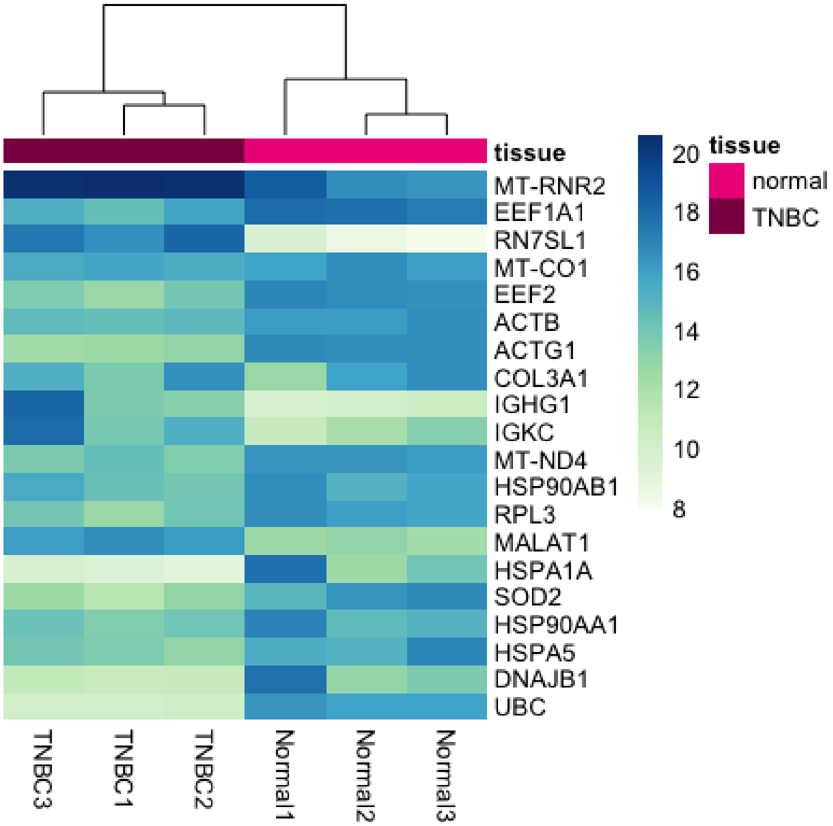


Figure 2: Heatmap of the more expressed genes across all samples.

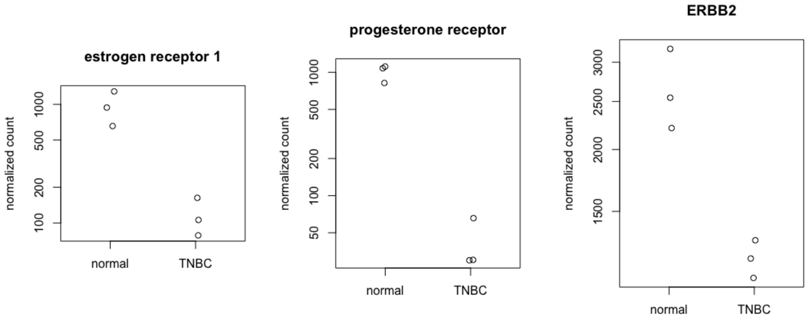


Figure 3: Expression levels.

3.4 Differential expression analysis

3.5 Overrepresentation analysis

4 Discussion

Supplementary materials

All scripts used for this project and other supplementary materials can be found here: `https://github.com/saskia-droid/summer_breast_de`.

References

[1] Jeyanthy Eswaran et al. "Transcriptomic landscape of breast cancers through mRNA sequencing". In: *Scientific Reports* 2.1 (Feb. 2012), p. 264. ISSN: 2045-2322. DOI: `10.1038/srep00264`. URL: `https://doi.org/10.1038/srep00264`.

[2] Philip Ewels et al. "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19 (June 2016), pp. 3047–3048. DOI: `10.1093/bioinformatics/btw354`. URL: `https://doi.org/10.1093/bioinformatics/btw354`.

[3] Daehwan Kim, Ben Langmead, and Steven L Salzberg. "HISAT: a fast spliced aligner with low memory requirements". In: *Nature methods* 12.4 (2015), pp. 357–360.

[4] H Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079. DOI: `10.1093/bioinformatics/btp352`. URL: `https://www.ncbi.nlm.nih.gov/pubmed/19505943`.

[5] Y. Liao, G. K. Smyth, and W. Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 30.7 (Nov. 2013), pp. 923–930. DOI: 10.1093/bioinformatics/btt656. URL: https://doi.org/10.1093/bioinformatics/btt656.

[6] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15 (12 2014), p. 550. DOI: 10.1186/s13059-014-0550-8.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[8] Andrews S. *FASTQC. A quality control tool for high throughput sequence data.* URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.