Sorbonne University

Master 2 BMC - Systems Immunology

Characterization of osteosarcomas and association with the tumor's inflammatory status

By

HUYNH Minh-Anh

September 2022

Table of Contents

List of Abbreviations

- DEG : Differentially Expressed Genes
- MDS : MultiDimensional Scaling
- scRNA-seq : single-cell RNA sequencing
- TANs: Tumor-associated Neutrophils
- TME: Tumor MicroEnvironment
- TPM : Transcripts Per Million
- UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction

Acknowledgements

Thank you for following this tutorial!

I hope you'll find it useful to write a very professional dissertation.

1 Introduction

- introduce the reader to the subject area and clarify the knowledge gap that the dissertation research will fill.
- set the context for the dissertation by reviewing the relevant literature.
- include relevant references to general (theoretical papers and reviews) and specific (specific to the particular question addressed) literature, to justify the research that has been undertaken and define the questions being addressed.
- state the primary research questions and hypotheses in the final paragraph.
- follow an 'inverted triangle' format, progressing from general scientific ideas and why they matter to the specific research questions addressed in the dissertation project.

The introduction should not be just a 'Literature Review.'

My 4.5 month research internship was dole out under the supervision of Dr. Dominique MODROWSKI at INSERM U1132 "BIOSCAR" directed by Dr. Martine COHEN-SOLAL, at the Lariboisière Hospital, a look unit that has been dedicated to the physiopathology of bone and cartilage diseases. My work has focused on elucidating the mechanisms associated with osteosarcoma and their inflammatory status, with particular emphasis on neutrophils.

Osteosarcomas are one of the most aggressive diseases, characterized by high tumor heterogeneity and numerous genomic instabilities. They usually occur in long bones and represent one-fifth of all primary malignant bone tumors and 2.4% of pediatric cancers (Fu et al., 2022). Osteosarcomas occur most frequently in children aged 14-18 years, and more than 30% of cases, within adults over 40 years of age (**Mirabello.2009?**). The combination of surgery and chemotherapy has increased the 5-year survival rate of patients to over 70%, however metastatic osteosarcoma, frequently occurring in the lungs, drops the survival rate to less than 20%. (Fu et al., 2022).

The difficulty in treating these tumors is largely due to the genomic instability that characterizes them and that leads to a great inter and intra tumor heterogeneity. Thus, the cancer cells can have various phenotypes. Some tumors are osteoblastic, chondroblastic or fibroblastic dominant.

It is established that tumor cells recruit numerous accessory cell types that will form a stroma that supports tumor growth, metastatic progression and resistance to treatment. Cellular (immune cells) and molecular (cytokines, metabolites) inflammatory mediators play a key role in the tumor stroma. The main immune component has been shown to be M2 macrophages. However, scRNA-

seq analyses have revealed other immune cell populations such as neutrophils that are also important mediators of inflammation. Bone tumors are a source of inflammatory factors that serve to recruit and educate the stroma.

An issue that remains poorly documented concerns the heterogeneity of the inflammatory landscape of osteosarcomas and the possible relationships between tumor inflammatory status (TIS) and phenotypic and/or clinical characteristics of these tumors. Similarly, the characteristics of neutrophils associated with bone tumors and their role in tumor inflammation have not been explored to date.

Recently, immunotherapy has proven to be particularly effective in treating previously difficult and deadly cancers, and has been promoted as a medical care therapy in many of them. However, osteosarcomas have not withstood immunotherapy and have instead turned to conventional chemotherapy, which is not very effective and has a plethora of side effects. In order to find a cure for these difficult-to-treat cancers, a better understanding of the tumor microenvironment (TME) could potentially lead to effective and novel targeted TME therapies, especially since osteosarcomas exhibit high heterogeneity and the immunological mechanisms of immune resistance are still unknown. Current advances in bulk RNA sequencing (RNA-Seq) and single-cell RNA sequencing (scRNA-seq) have shown their potential in exploring the tumor microenvironment (TME) to explore intra-tumor heterogeneity and cellular dialogue between tumor and inflammatory cells.

Neutrophils are the primary innate immune cells recruited during an inflammation and multiple papers have already elucidated that tumor-associated neutrophils (TANs) or circulating neutrophils are related to worse patient survival therapy and chemoresistance (Faget et al., 2021; Long et al., 2021). Neutrophils support cancer development through 3 pathways : they’re able to promote cancer initiation, assistance of metastasis and increase of tumor growth (Faget et al., 2021).

2 Objective and experimental principle

Previously, the lab has demonstrated that tumor stem cells have specific properties inside the tumor, characterized by the calpaïne 6 biomarker. This cell is capable of coordinating invasion of distant tissues and confers to the whole tumor a specific phenotype.

The study aims to analyze public dataset and to make a custom analysis using bioinformatic tools in order to understand the cellular crosstalk and interactions in the TME, between neutrophils

and tumor cells and their inflammatory status.

Collaborating with Dr Jean-Marc Schwartz's team in Manchester and Sophie Khakoo, a Master student in Systems Biology, collecting data from a scRNA-Seq dataset, a customized downstream analysis will be performed on GSE87686 and TargetOS, and a scRNAseq dataset in order to identify the relationships between neutrophils and osteosarcoma through the use of genetic signatures available from MSigDB and customized gene lists.

3 Methods

3.1 Data collection

Analyses were performed on two osteosarcoma bulk RNA-seq datasets, TARGET-OS Osteosarcoma and GSE87686. Data was collected from TARGET-OS using GDC Data Transfer Tool UI (v1.0.0), returning 19493 protein coding genes and 88 samples for TARGET-OS, containing both raw data and TPM data. GSE87686 data was obtained through the lab's previously pre-processed kallisto files, downloaded through SRA Run Selector to obtain SRA run files. Data was then imported from kallisto files via *tximport (v1.22.0)* R package.

Genes were converted to ENST and ENSG and finally HUGO gene symbols through *biomaRt (v2.50.3)*.

3.2 Normalization

Raw data was converted to TPM as it is the best performing normalization method than FPKM or RPKM, based on its perservation of biological signal as compared to other methods (Abrams et al., 2019). Calculation was performed using counts and lengths for each gene, returned from the tximport to kallisto process.

Z-score was used to normalize the data for each gene, in order to be able to visualize the data in corresponding heatmaps.

Z-score of the mean of tpm data was used to construct grouped heatmaps for a given gene or gene signature.

Inflammatory groups characterizing the intensity of inflammatory status in tumors were created by choosing the lowest and highest mean of Z-score of the Hallmark Inflammatory Response signature from MSigDB, containing 200 genes. ICAM4 was notably not present in the dataset in TARGET-OS cohort. The groups were cut off evenly using the *ntile* function in *dplyr (v1.0.8)* R package.

3.3 Construction of gene signatures

Gene signatures relevant to the research topic were obtained from MSigDB via *msigdbr* (v7.5.1)
HAY_BONE_MARROW_NEUTROPHIL.v7.5.1

Canonical markers for osteosarcoma markers were adapted from Zhou et al. (2020) 's analysis and their canonical markers generated from the literature in their Supplementary Table 2.

3.4 MDS Clustering

3.5 Determination of cell population abundance

3.5.1 MCP Counter

Estimation of tumor immune infiltration using bulk RNA-seq can be estimated using Microenvironment Cell Populations-counter (MCP-counter) (Becht et al., 2016). Non-metric Kruskal-Wallis testing was used to determine significant differences between cell populations between Low, Medium and High groups with $P < 0.05$ were considered as statistically significant.

3.5.2 CIBERSORTx.

CIBERSORTx algorithm was also used for immune cell deconvolution, from the CIBERSORTx platform (https://cibersortx.stanford.edu/) to generate abundance for 22 immune cells.

3.6 Differential Expression Gene analysis

3.6.1 Differentially Expressed Genes analysis (DEG)

Using *DESeq2 (v1.34.0)* (Love et al., 2014), standard DEG pipeline using raw data was performed between inflammatory groups (Low, Medium, High). DEGs were identified with *adjusted p.value (or FDR)* ≤ 0.05 and _log2FoldChange ≥ 1 or *log2FoldChange* ≤ -1. An interpretation of the FDR/Benjamini-Hochberg method for controlling the FDR is implemented in DESeq2 in which we rank the genes by p-value, then multiply each ranked p-value by m/rank (m = total number of tests).

3.6.2 Over Representation Analysis (ORA)

enrichR v(3.0) was used (Xie et al., 2021; **R-enrichR?**) for functional gene enrichment/pathway analysis. Following database were queried : *"GO_Molecular_Function_2021," "Human_Gene_Atlas," "BioPlanet_2019," "GO_Biological_Process_2021," "GO_Cellular_Component_2021"*

3.7 Statistical Testing

Statistical testing was performed using *R software 4.1.2 (2021-11-01)*. Kruskal-Wallis testing was performed for non-metric comparative analysis between groups. Post-hoc analysis was performed using Dunn's test as opposed to Wilcoxon due to the test taking into account Kruskal-Wallis's rank. $P < 0.05$ and $P.adj < 0.05$ was considered statistically significant.

3.8 Gene Set Enrichment Analysis (GSEA)

4 Results

Some more guidelines from the School of Geosciences.

This section should summarize the findings of the research referring to all figures, tables and statistical results (some of which may be placed in appendices).

- include the primary results, ordered logically - it is often useful to follow the same order as presented in the methods.
- alternatively, you may find that ordering the results from the most important to the least important works better for your project.
- data should only be presented in the main text once, either in tables or figures; if presented in figures, data can be tabulated in appendices and referred to at the appropriate point in the main text.

Often, it is recommended that you write the results section first, so that you can write the methods that are appropriate to describe the results presented. Then you can write the discussion next, then the introduction which includes the relevant literature for the scientific story that you are telling and finally the conclusions and abstract – this approach is called writing backwards.

4.1 Determination of inflammatory groups

4.1.1 MDS visualization and k-means clustering

Functional clusters of osteosarcoma samples were created using k-means clustering of from MDS visualization of the Hallmark Inflammatory Response signature.

Inflammatory groups characterizing the intensity of inflammatory status in tumors were created by choosing the lowest and highest mean of Z-score of the hallmark inflammatory response signature from MSigDB, containing 200 genes. ICAM4 was notably not present in the dataset in TARGET-OS cohort. The groups were cut off evenly using the *ntile* function in *dplyr* R package.

Those manually defined groups are relevant as they correspond fairly well to functional groups, defined by k-means clustering based on MDS visualization **(Figure ??)**. Each sample is thus attributed to its inflammatory status and this group will be subsequently used for the following results.

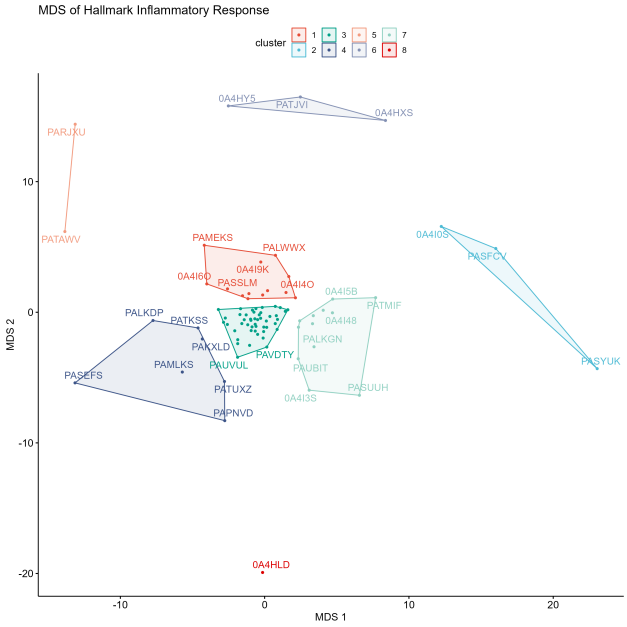


Figure 1: MDS of Hallmark Inflammatory Signature

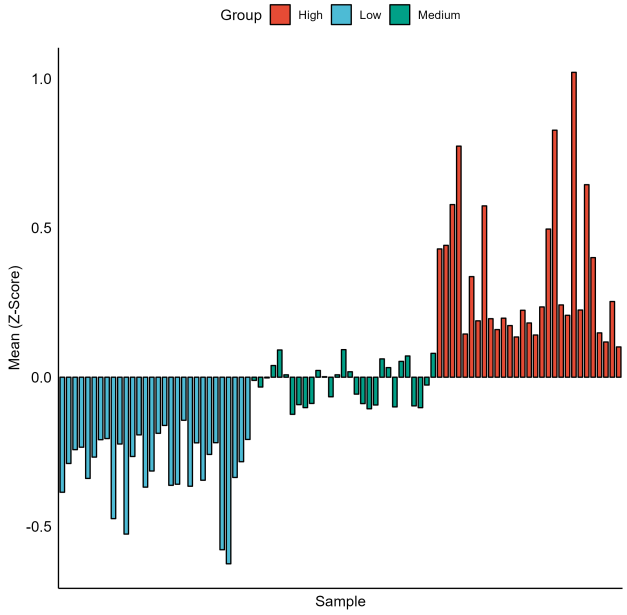


Figure 2: MDS of Hallmark Inflammatory Signature

4.2 Characterization of osteosarcomas associated to inflammatory status

Comparison of the mean of *Hp Osteosarcoma* gene signature and gene relating to types of osteosarcomas associated to inflammatory groups were performed, represented through a heatmap. Despite high heterogeneity between samples and inflammatory status, the Z-score of the mean of genes in Low versus High group is statistically significatively different.

However *Hp Osteosarcoma* in GSEA is not significantly differrent ($p = 0.163$).

Comparison of the expression of specific osteosarcoma markers relating to osteoblastic, chondroblastic, fibroblastic markers through a heatmap representation. Hierarchical clustering of the samples does not appear to be associated with corresponding inflammatory status. However it does reveal that there are groups of osteoblastic, chondroblastic and fibroblastic osteosarcomas which is expected.

The mean of markers of proliferation (MKI67, PCNA, TOP2A) associated to osteosarcomas have been compared to inflammatory status, along with the mean of the three markers. Kruskal-Wallis testing is significant ($p = 0.00968$) and post-hoc Dunn analysis reveals that the mean of the proliferation markers between low and high group is significantly different ($p = 0.016$). The data suggests that proliferation is hindered when inflammatory status is high in the osteosarcoma samples.

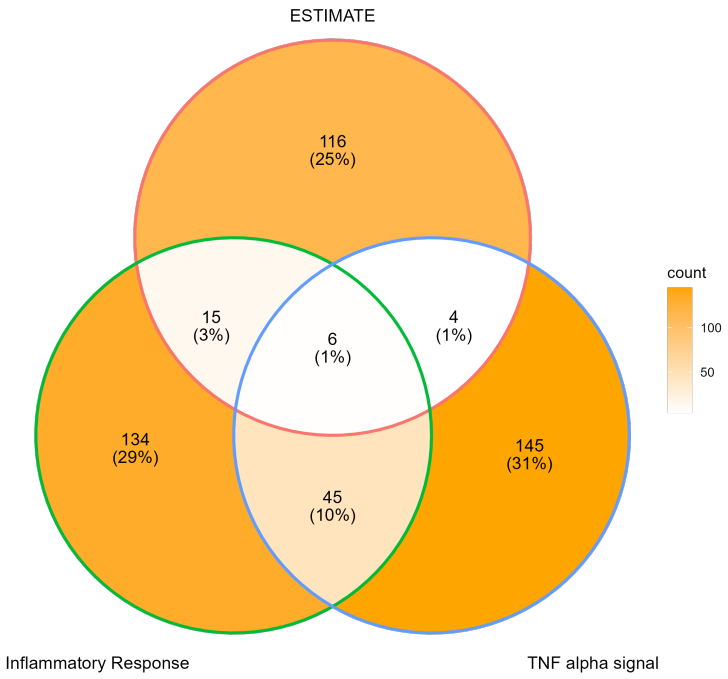
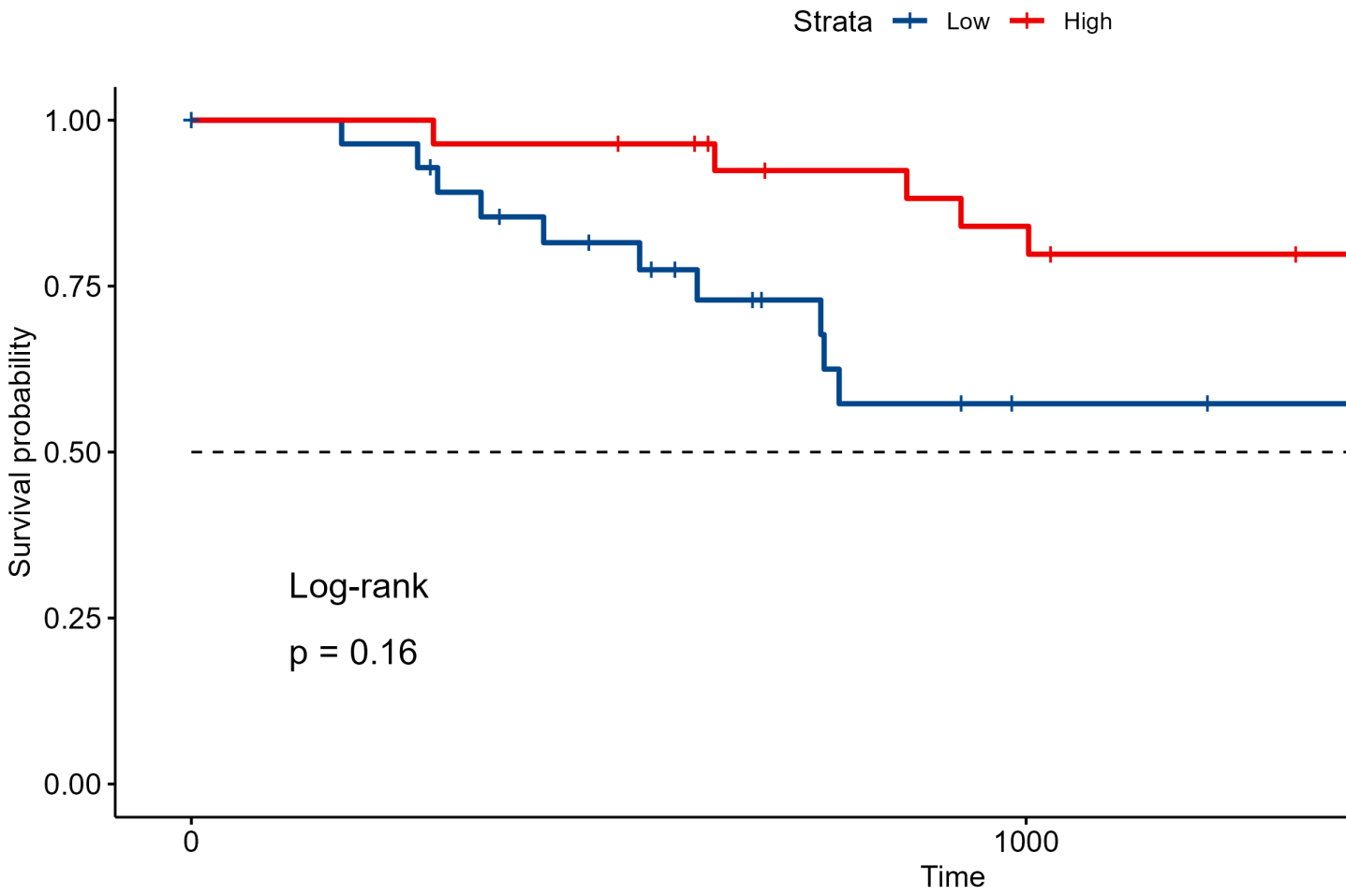


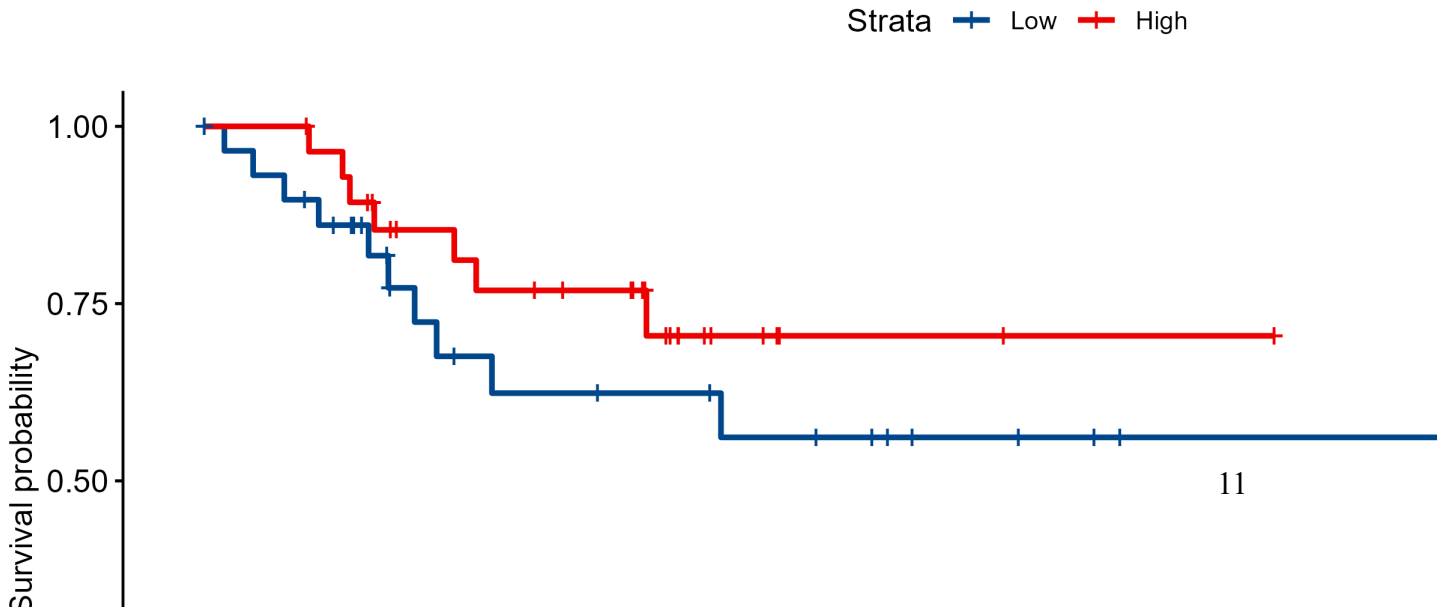
Figure 3: Venn diagram comparing three inflammatory signatures.

4.3 Characterization of intra-tumor inflammation associated to inflammatory status

4.4 Similarity of signatures

4.5 Survival





4.6 single cell RNA-Seq analysis

Using scRNA-Seq data from GSE152048, used by Zhou et al. (2020) and normalized by Sophie

5 Discussion

the purpose of the discussion is to summarise your major findings and place them in the context of the current state of knowledge in the literature. When you discuss your own work and that of others, back up your statements with evidence and citations.

- The first part of the discussion should contain a summary of your major findings (usually 2 – 4 points) and a brief summary of the implications of your findings. Ideally, it should make reference to whether you found support for your hypotheses or answered your questions that were placed at the end of the introduction.
- The following paragraphs will then usually describe each of these findings in greater detail, making reference to previous studies.
- Often the discussion will include one or a few paragraphs describing the limitations of your study and the potential for future research.
- Subheadings within the discussion can be useful for orienting the reader to the major themes that are addressed.

5.1 Limitations

- Immune deconvolutions methods do not give the same results, thus they are unclear. Only MCP Counter reports significantly different neutrophil abundance between low and high group.
- Immune deconv paper has 7 limitations (Sturm et al., 2019).
- Z-score and equal mean of groups may not be the best normalization method and way of making groups. Geometric mean (inspired from a package) could have been used but returned 0 value despite all values being positive.
- Literature analysis comparing normalization methods reveals library size normalization (TPM, FPKM, RPKM) methods perform worse than distribution normalization methods (DESeq2, TMM), which are recommended by Zhao et al. (2021). Notably, library size normalization methods assume that the total amount of mRNA per cell is identical in every condition.
- DESeq2's normalized count could have then been used, as it is possible to get it from raw-counts after performing DESeq2 analysis, using two functions, `estimateSizeFactors` and

`counts(normalized = TRUE)`. However some immune deconvolution methods still expect TPM as input.

- Using DESeq2's normalized raw count as input, it is inconsistent to then use TPM in order to perform other analysis.
- Perhaps some outliers could have been removed in order to better identify clusters.
- In single-cell, one potential reason that we do not find neutrophils is because chemotherapy induces aplasia which kills off all short lives immune cells.

6 Bibliography

Abrams, Z.B., Johnson, T.S., Huang, K., Payne, P.R.O., and Coombes, K. (2019). A protocol to evaluate RNA sequencing normalization methods. BMC Bioinformatics *20*, 679.

Becht, E., Giraldo, N.A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W.H., et al. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biology *17*, 218.

Faget, J., Peters, S., Quantin, X., Meylan, E., and Bonnefoy, N. (2021). Neutrophils in the era of immune checkpoint blockade. Journal for Immunotherapy of Cancer *9*, e002242.

Fu, Y., Jin, Z., Shen, Y., Zhang, Z., Li, M., Liu, Z., He, G., Wu, J., Wen, J., Bao, Q., et al. (2022). Construction and validation of a novel apoptosis-associated prognostic signature related to osteosarcoma metastasis and immune infiltration. Translational Oncology *22*, 101452.

Long, W., Chen, J., Gao, C., Lin, Z., Xie, X., and Dai, H. (2021). Brief review on the roles of neutrophils in cancer development. Journal of Leukocyte Biology *109*, 407–413.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Sturm, G., Finotello, F., Petitprez, F., Zhang, J.D., Baumbach, J., Fridman, W.H., List, M., and Aneichyk, T. (2019). Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics *35*, i436–i445.

Xie, Z., Bailey, A., Kuleshov, M.V., Clarke, D.J.B., Evangelista, J.E., Jenkins, S.L., Lachmann, A., Wojciechowicz, M.L., Kropiwnicki, E., Jagodnik, K.M., et al. (2021). Gene Set Knowledge Discovery with Enrichr. Current Protocols *1*, e90.

Zhao, Y., Li, M.-C., Konaté, M.M., Chen, L., Das, B., Karlovich, C., Williams, P.M., Evrard, Y.A., Doroshow, J.H., and McShane, L.M. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. Journal of Translational Medicine *19*, 269.

Zhou, Y., Yang, D., Yang, Q., Lv, X., Huang, W., Zhou, Z., Wang, Y., Zhang, Z., Yuan, T., Ding, X., et al. (2020). Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. Nature Communications *11*, 6322.

7 Appendix: code

Analyses were conducted using the R Statistical language (version 4.1.2; R Core Team, 2021) on Windows 10 x64 (build 22000). Code is available at _https://github.com/Minh-AnhHuynh/Osteosarcoma-Project_.

Code philosophy follows “The tidyverse style guide” written by Hadley Wickham. and usage of tidyverse functions, in order to ensure consistent code readability, usage and naming.

```
## - Session info ---------------------------------------------------------------
##  setting  value
##  version  R version 4.1.2 (2021-11-01)
##  os       Windows 10 x64 (build 22000)
##  system   x86_64, mingw32
##  ui       RTerm
##  language (EN)
##  collate  French_France.1252
##  ctype    French_France.1252
##  tz       Europe/Paris
##  date     2022-06-20
##  pandoc   2.14.0.3 @ C:/Program Files/RStudio/bin/pandoc/ (via rmarkdown)
##
## - Packages -------------------------------------------------------------------
##  package   * version date (UTC) lib source
##  bookdown  * 0.27    2022-06-14 [1] CRAN (R 4.1.2)
##  knitr     * 1.39    2022-04-26 [1] CRAN (R 4.1.3)
##  librarian * 1.8.1   2021-07-12 [1] CRAN (R 4.1.3)
##  rmarkdown * 2.14    2022-04-25 [1] CRAN (R 4.1.3)
##
##  [1] C:/Users/Minh-Anh/Documents/R/win-library/4.1
##  [2] C:/Program Files/R/R-4.1.2/library
##
## ------------------------------------------------------------------------------

## {paint} masked print.tbl_df
```

package	version	source
bookdown	0.27	CRAN (R 4.1.2)
knitr	1.39	CRAN (R 4.1.3)
librarian	1.8.1	CRAN (R 4.1.3)
rmarkdown	2.14	CRAN (R 4.1.3)

```
## Analyses were conducted using the R Statistical language (version 4.1.2; R Core Team,
##
```

```
## References
## ----------
##   - Desi Quintans (2021). librarian: Install, Update, Load Packages from CRAN, 'GitHu
##   - JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ush
##   - R Core Team (2021). R: A language and environment for statistical computing. R Fo
##   - Yihui Xie (2022). bookdown: Authoring Books and Technical Documents with R Markdo
##   - Yihui Xie (2022). knitr: A General-Purpose Package for Dynamic Report Generation
```

Summary

English Summary