Gene Set Enrichment Analysis of Human Microglia in Unique Extracellular Matrix Networks

Autumn Davis
davis33@wwu.edu
CS Graduate
Western Washington University
Bellingham, Washington, USA

Jed Pagcaliwagan
pagcalj@wwu.edu
CS Undergraduate
Western Washington University
Bellingham, Washington, USA

Jasdeep Singh
singhj9@wwu.edu
CS Undergraduate
Western Washington University
Bellingham, Washington, USA

Vivian White
whitev4@wwu.edu
CS Graduate
Western Washington University
Bellingham, Washington, USA

Abstract

Microglia are immune cells in the brain that, among other functions, “eat" and kill tumor cells through phagocytosis [4]. In an experiment conducted by Dr. Annelise Snyder (2024), certain extracellular matrix (ECM) proteins were found to inhibit the microglial uptake of tumor cells in culture; that is, phagocytosis of tumor cells was hindered.

The Snyder lab hypothesized that “certain brain ECM proteins will activate inhibitory pathways in microglia that explain how these molecules block microglia’s ability to eat tumor cells." As one test of this hypothesis, we perform gene set enrichment analysis (GSEA) on RNA sequence data collected from human microglia cultures grown on different plates, each coated with one of six ECM proteins. GSEA identifies differences in gene expression between certain conditions (e.g., different ECM coatings) [13], which can potentially provide insights into associated biological processes.

Our results suggest a possibility of the experimental ECM coatings having an effect on microglial genes. Specifically, the agrin coating shows evidence of downregulating the synthesis of ketone bodies and fatty acids. However, with all resulting false discovery rate adjusted p-values being above 0.15, the statistical significance is not strong enough to be conclusive. These results can potentially serve to guide future research on medicinal applicability.

ACM Reference Format:
Autumn Davis, Jed Pagcaliwagan, Jasdeep Singh, and Vivian White. 2025. Gene Set Enrichment Analysis of Human Microglia in Unique Extracellular Matrix Networks. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Bioinformatics, Fall 2024, Western Washington University

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Immunosurveillance is the process by which the immune system suppresses cancer development by deploying immune responses to compromised cells and restoring tissue homeostasis [14]. However, cancer cells have evolved numerous survival mechanisms to evade immune targeting.

The Snyder lab studies how macrophages—immune cells that engulf and break down other cells/debris—interact with tumor cells during brain metastasis in breast cancer. This research aims to characterize cellular mechanisms that control the function of microglia to help explain why immunosurveillance is inhibited during brain metastasis.

We study the effects of ECM proteins on microglia function using RNA sequence data to analyze which gene groups seem to be increased or decreased in expression based on “treatment" ECM coatings (Agrin, Pan-Laminin, Laminin-211, and Collagen I) compared to “control" coatings (Matrigel, Poly-L-lysine). The insights from this work could lead to the development of therapies targeting microglia to treat brain metastases.

2 Motivation

The ECM is a network of molecules and proteins that surround and shape itself to support cells and tissues in the body. Beyond structural support, ECM proteins have also been shown to affect immune cell function and tumor cell ability to evade immunosurveillance [7, 11]. However, it is not yet clearly understood how different brain ECM proteins affect microglia function and how they alter these cells on the genetic level. To explore this, the Snyder lab cultured immortalized human microglia on different ECM proteins and found that certain ones decreased the amount of tumor cells destroyed by the microglia.

Fig 1(A) shows the process of the ECM experiment: plates are coated with ECM proteins, then tumor and microglia cells are co-cultured on the plates, and finally, the uptake of tumor cells by the microglia is calculated. The tumor uptake results in Fig 1(B) show that Agrin inhibited tumor uptake the most, followed by Laminin-211, Laminin-511, and finally Laminin-411. This shows that certain ECM proteins can inhibit phagocytosis in the brain Fig 1(C).

An essential goal of the Snyder lab is to explain the underlying mechanisms of how these ECM proteins inhibit microglia function,

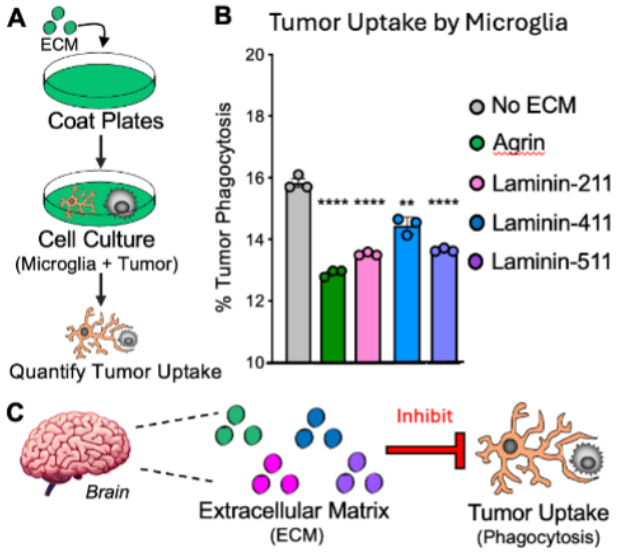


Figure 1: (A) Snyder's ECM—microglial tumor uptake experiment. (B) Agrin and laminin ECM coatings on plates decreased tumor uptake by microglia. (C) Implications: ECM proteins inhibit tumor uptake in the brain. Image from Snyder, 2024.

which can ultimately lead to potential methods of mitigating brain metastasis. To this end, we aim to analyze which gene groups appear often in the RNA sequences of microglia cultured on ECM-coated plates to assist in determining the inhibitory signaling pathways that these ECM molecules activate in microglia.

3 Related Work

The RNA sequencing data used in this study was provided by Dr. Annelise Snyder of Western Washington University. The data is part of ongoing research and is not yet published.

Prior to our work, Snyder's lab coated plates of immortalized human microglia with extracellular matrix (ECM) proteins, including agrin, pan-laminin, collagen I, and laminin-211. Poly-L-Lysine and Matrigel coatings were used as controls. These cells were cultured together, after which the microglial RNA was extracted and sequenced into FASTQ format [3].

The ECM coating components were previously found to be specifically increased in the brain, as opposed to other tissues [5]. Snyder's pilot study found that the proteins reduced the update of tumor cells by microglia, indicating that ECM molecules may take part in the inhibition of immunosurveillance by microglia during brain metastasis. This highlights the significance of ECM molecules on microglial functions and motivates ongoing efforts to identify the specific signaling pathways activated in microglia.

The FASTQ data produced by Snyder is presently also being used in other research concerning differentially expressed genes and transcription factor analysis.

4 Methods

At the start of our project, we received FASTQ files of RNA sequence data from Snyder's graduate students to perform GSEA on. Our tool pipeline is visualized in Fig 2. We aligned the reads to a human reference genome [8], ran high-throughput sequencing tools, ran differential gene expression analysis, and finally performed gene set enrichment analysis and plotted the results.

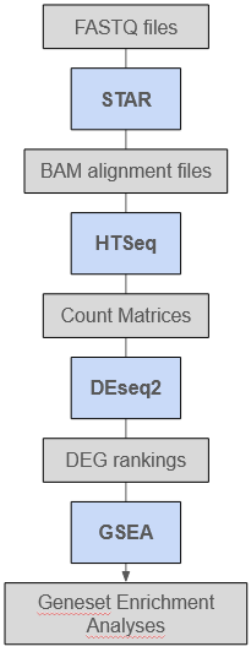


Figure 2: Tool pipeline. Gray boxes represent data inputs and outputs, and blue boxes represent software tools.

4.1 RNA Sequence Alignment

We used STAR (Spliced Transcripts Alignment to a Reference), a fast RNA sequence alignment tool, to generate genome indexes from a human reference genome and map our RNA sequences to the indexes [6] to determine where in the genome the reads originated from. STAR is highly accurate but memory intensive; running this software on the CSE cluster was an initial bottleneck in our project. We generated 46 SAM files, which we converted to BAM files for compression to speed up the following step.

4.2 Gene Read Counting

Once our RNA-seq reads were successfully mapped to the reference genome, we used htseq-count [1] to compute gene expression—the process by which information in a gene is used to create a functional product such as a protein—by counting how many reads were mapped to each gene. This tool outputs a CSV file for each sample. The rows in the CSV file correspond to individual genes, and the columns represent the read counts for each gene across the different ECM samples. These read counts are used for further analysis, in this case, differential gene expression analysis, to identify genes that are upregulated (increased in gene expression) or downregulated (decreased) in response to the treatment.

4.3 DEG Identification

Following the generation of read count matrices, the high-throughput requirement became less necessary, which allowed subsequent steps to be consolidated into a Python pipeline.

Identification of DEGs was accomplished with DEseq2 [10]. A Python implementation (pyDESeq2) was leveraged for pipeline cohesion [12]. For each experimental group, the htseq-generated gene count matrix of all trials from the experimental group was contrasted with the htseq-generated gene count matrices of all trials from both control groups. The result was a single differentially expressed gene analysis for each experimental group.

Each DEG analysis consisted of a list of genes with information about their changed expression. For each gene, there included statistical information such as the p-value and adjusted p-value, and log2 Fold-Change, which measures the extent to which the gene expression is downregulated or upregulated relative to the controls. These statistics are defined as follows:

- **p-value**:
  The p-value represents the probability that the observed changes involving gene expression occurred by random chance, under a null hypothesis that there was no difference between the experimental and control groups, respectively. A smaller p-value provides stronger evidence of true differential expression of a given gene.

- **Adjusted p-value**:
  Calculated with the Benjamini-Hochberg [2] procedure, the adjusted p-value corrects for multiple hypothesis testing as to control the false discovery rate. By utilizing this adjustment, further false positives can be prevented. A gene with an adjusted p-value of < 0.05 are typically considered significantly differentially expressed.

- **log2 Fold-Change**:
  The log2 fold-change is the measure of the magnitude and direction in which gene expression changes from experimental and control conditions. This is calculated as the log base 2 of the ratio between the expression levels of each respective group. A positive value or log2 fold-change indicates upregulation, whereas a negative value or log2 fold-change indicates downregulation.

In addition to these statistics, DESeq2 also contains base mean statistics (the average normalized count across all samples for each respective gene) and the log2 fold-change standard error (the estimate of the standard error of the log2 fold-change for each gene).

The DEG analyses gave relatively few genes with adjusted p-values below the 0.05 threshold, with Collagen I having the most at 53, and Agrin having the least at 3.

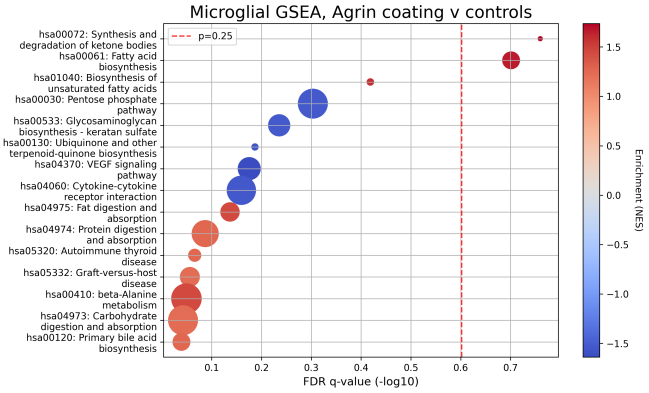


Figure 3: Bubble plot of Agrin, with the negative log10 of the FDR (False Discovery Rate) q-value on the x-axis. The red dotted line represents an FDR q-value of 0.25 (-log10(0.25)), the threshold at which the value becomes statistically significant. See Fig 4 for an explanation of the rest of the bubble plot.

4.4 Enrichment Analysis

Each experimental ECM coating's list of differentially expressed genes with their respective log2 Fold-Change scores was analyzed using GSEA [13] to identify enriched gene pathways. A Python implementation (gseapy) was used for pipeline integration[1]. The KEGG 2016 database was selected for enrichment analysis because of its widespread use for pathway identification in human disease research [9]. For reproducibility, 1000 permutations were performed and a seed of 6 was used. To eliminate noise and reduce bias, DEGs with counts below 5 or above 1000 were removed before this step, as well as novel genes that were not represented in the KEGG 2016 database

Due to the high adjusted p-values in the DEG analyses, filtering DEGs by p-values prior to the enrichment analysis proved impractical. Standard thresholds such as 0.05 and 0.1 failed to process, and usable results only started becoming visible close to a threshold of 1.0. For this reason, DEGs were not filtered by p-value.

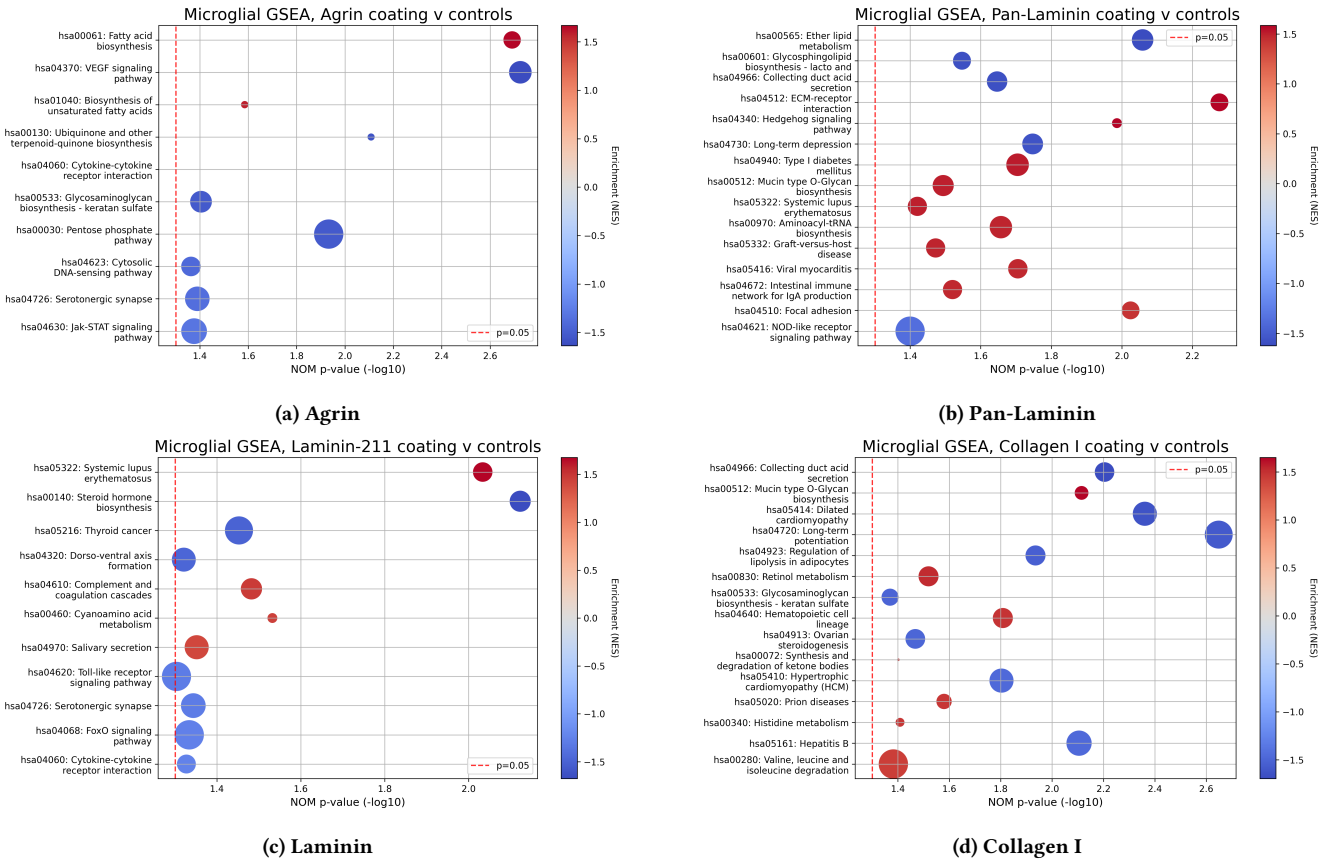
In the resulting enrichment analysis, the statistical significance of each pathway enrichment was initially defined as a pathway with an FDR q-val[2] below 0.25. Only two pathways (both in Agrin) met the FDR criteria. For this reason, a separate filter was applied, defining statistically significant pathways as those with a nominative (unadjusted) p-value below 0.05.

For each ECM coating, the 15 most enriched pathways (measured by the absolute value of the normalized enrichment score) with p-values below 0.05 were selected to be included in a bubble plot.

An initial intention to supplement the results with gene ontology information using DAVID was abandoned because the high p- and q-values made curating the list impractical.

[1]The DEGs list output by DEseq2 uses Ensembl IDs, so the Python module MyGene was used to convert them into HGNC Gene Symbols for KEGG compatibility

[2]FDR q-value (False Discovery Rate q-value): the statistical value under the consideration of a set of multiple hypotheses, that is the proportion of false positives among all significant findings.



(a) Agrin **(b) Pan-Laminin**

(c) Laminin **(d) Collagen I**

Figure 4: Gene set enrichment analysis results for each experimental coating: (a) agrin, (b) pan-laminin, (c) laminin-211, and (d) collagen I. The description of each pathway is on the y-axis. The x-axis shows the negative log10 of the nominative p-value. A higher (further right) -log10(p-val) means a lower p-val, which represents a higher statistical significance. Bubble sizes represent the proportion of genes in the pathway contributing to the enrichment signal. Bubble colors represent the direction. Pathways with bluer bubbles are more downregulated, and ones with redder bubbles are more upregulated. The red dotted line represents a p-value of 0.05.

5 Results

FDR q-values of each gene set enrichment analysis showed only tentatively statistically significant evidence of any of the experimental extracellular matrix coatings affecting gene expression in human microglial cells (using 0.25 as a threshold), but the less strict approach using nominative p-values (with a threshold of 0.05) gave significantly better results.

The most promising results were present in the Agrin coating (Fig 4a). Agrin had the only 2 pathways below an FDR q-val of 0.25: Pathway hsa00072, involved in synthesis and degradation of ketone bodies, had the lowest q-value at 0.174. Pathway hsa00061, involved in fatty acid biosynthesis, was a close second with a q-value of 0.199. Both of these pathways are metabolic and were shown to be upregulated (with an NES of 1.736 and 1.584, respectively). Table 1 displays the names of the Agrin pathways and whether they are involved in metabolism or inflammation. In total, six pathways have metabolic function and three are related to inflammation.

When defining statistically significant as having a nominative p-value below 0.05, we find the following results: The Agrin coating showed 11 statistically significant pathways (Fig 4a), the Pan-Laminin coating showed 17 (Fig 4b), the Laminin-211 coating showed 11 (Fig 4c), and the Collagen-I showed 21 (Fig 4d).

Some notable patterns include the downregulation of inflammation-related pathways, including hsa04060 (downregulated in all experimental coatings except for Collagen-I), hsa04750 (downregulated in Collagen-I), and hsa04623 (downregulated in Agrin). This is significant because inflammation regulated immune responses, so downregulation of inflammatory pathways may inhibit immune activity.

Several affected pathways across all experimental coatings are also involved in metabolic processes (metabolism, synthesis/biosynthesis, digestion, etc). These pathways are variously regulated up or down. This is of note because energy for cell functioning is dependent on metabolic processes.

These results suggest noteworthy affects of ECM coatings on gene expression in microglial cells that may inhibit or other affect

Table 1: Agrin Pathways

Pathway ID (KEGG)	Name	Function
hsa00061	Fatty acid biosynthesis	Metabolic
hsa01040	Biosynthesis of unsaturated fatty acids	Metabolic
hsa00130	Ubiquinone and other terpenoid-quinone biosynthesis	Metabolic
hsa00533	Glycosaminoglycan biosynthesis - keratan sulfate	Metabolic
hsa00030	Pentose phosphate pathway	Metabolic
hsa00072	Synthesis and degradation of ketone bodies	Metabolic
hsa04060	Cytokine-cytokine receptor interaction	Inflammatory
hsa04623	Cytosolic DNA-sensing pathway	Inflammatory
hsa04630	Jak-STAT signaling pathway	Inflammatory
hsa04370	VEGF signaling pathway	Other
hsa04726	Serotonergic synapse	Other

immune response. Further cross-referencing them with existing expert knowledge and literature of their function may provide insights into which can potentially be promising for future study.

6 Discussion and Future Work

The gene set enrichment analysis performed here shows enriched pathways falling below the standard 0.25 False Discovery Rate (FDR) q-value only in two metabolic pathways from the Agrin coating experiment (Fig 3). Pathways with a higher FDR indicate a 25% or more chance of being a false positive resulting from random variations or other factors not controlled for in the experiment. By convention, this suggests insufficient proof of an effect of the ECM treatment on gene expression in microglial cells. This analysis had to resort to using a less strict value (nominative p-value) to produce enough results to analyze.

The authors suggest some improvements or variations on the experiment that future studies may be able to employ to produce more statistically significant results. Firstly, sampling results intermittently throughout the culturing process might not only provide a more robust sample size, but be able to capture more complicated processes resulting from the exposure. Additionally, increasing the sample size by including more technical replicates of each ECM coating sample may improve statistical power. In the computational step, mistakes may be avoided by validating intermediate results (i.e. alignment results, gene reads, differential gene expression data) through comparison using other tools or methods could ensure that results are reproducible and thus reliable. Lastly, toolchain parameters can be refined, including choice of tools, input formats, databases and references employed. By experimenting with a variety of parameter and tool configurations, more promising results may be attainable.

Acknowledgments

We thank Dr. Annelise Snyder and her graduate students, Paige Edwards and Nevada Nelson, for guidance, data, and resources. Thank you to Zach McGrew for his help installing RNA STAR onto the CSE cluster and submitting job scripts. Thank you to Filip for organizing this project for us and getting us connected with our collaborators.

References

[1] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. 2014. HTSeq – A Python framework to work with high-throughput sequencing data. *bioRxiv* (2014). https://doi.org/10.1101/002824

[2] Yoav Benjamini and Yosef Hochberg. 2018. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (12 2018), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[3] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. 2009. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* 38, 6 (12 2009), 1767–1771. https://doi.org/10.1093/nar/gkp1137

[4] Marco Colonna and Oleg Butovsky. 2017. Microglia Function in the Central Nervous System During Health and Neurodegeneration. *Annual Review of Immunology* 35, Volume 35, 2017 (2017), 441–468. https://doi.org/10.1146/annurev-immunol-051116-052358

[5] Jinxiang Dai, Patrick Cimino, Kenneth Gouin, Candice Grzelak, Alex Barrett, Andrea Lim, Annalyssa Long, Stephanie Weaver, Lindsey Saldin, Aiyedun Uzamere, Vera Schulte, Nigel Clegg, Laura Pisarsky, David Lyden, Mina Bissell, Simon Knott, Alana Welm, Jason Bielas, and Cyrus Ghajar. 2022. Astrocytic laminin-211 drives disseminated breast tumor cell dormancy in brain. *Nature Cancer* 3 (01 2022). https://doi.org/10.1038/s43018-021-00297-3

[6] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2012. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 1 (10 2012), 15–21. https://doi.org/10.1093/bioinformatics/bts635

[7] Alex Gordon-Weeks and Arseniy E. Yuzhalin. 2020. Cancer Extracellular Matrix Proteins Regulate Tumour Immunity. *Cancers* 12, 11 (2020). https://doi.org/10.3390/cancers12113331

[8] Peter W Harrison, M Ridwan Amode, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, Simarpreet Kaur Bhurji, Sanjay Boddu, Paulo R Branco Lins, Lucy Brooks, Shashank Budhanuru Ramaraju, Lahcen I Campbell, Manuel Carbajo Martinez, Mehrnaz Charkhchi, Kapeel Chougule, Alexander Cockburn, Claire Davidson, Nishadi H De Silva, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Dionysios Grigoriadis, Gurpreet S Ghattaoraya, Jose Gonzalez Martinez, Tatiana A Gurbich, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Mike Kay, Vinay Kaykala, Tuan Le, Diana Lemos, Disha Lodha, Diego Marques-Coelho, Gareth Maslen, Gabriela Alejandra Merino, Louisse Paola Mirabueno, Aleena Mushtaq, Syed Nakib Hossain, Denye N Ogeh, Manoj Pandian Sakthivel, Anne Parker, Malcolm Perry, Ivana Piližota, Daniel Poppleton, Irina Prosovetskaia, Shriya Raj, José G Pérez-Silva, Ahamed Imran Abdul Salam, Shradha Saraf, Nuno Saraiva-Agostinho, Dan Sheppard, Swati Sinha, Botond Sipos, Vasily Sitnik, William Stark, Emily Steed, Marie-Marthe Suner, Likhitha Surapaneni, Kyösti Sutinen, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veidenberg, Thomas A Walsh, Doreen Ware, Elizabeth Wass, Natalie L Willhoft, Jamie Allen, Jorge Alvarez-Jarreta, Marc Chakiachvili, Bethany Flint, Stefano Giorgetti, Leanne Haggerty, Garth R Ilsley, Jon Keatley, Jane E Loveland, Benjamin Moore, Jonathan M Mudge, Guy Naamati, John Tate, Stephen J Trevanion, Andrea Winterbottom, Adam Frankish, Sarah E Hunt, Fiona Cunningham, Sarah Dyer, Robert D Finn, Fergal J Martin, and Andrew D Yates. 2023. Ensembl 2024. *Nucleic Acids Research* 52, D1 (11 2023), D891–D899. https://doi.org/10.1093/nar/gkad1049

[9] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. 2016. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic*

Acids Research 45, D1 (11 2016), D353–D361. https://doi.org/10.1093/nar/gkw1092

[10] Michael I Love, Wolfgang Huber, and Simon Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15 (May 2014), 550. https://doi.org/10.1186/s13059-014-0550-8

[11] Ding Ma, Senquan Liu, Bachchu Lal, Shuang Wei, Shuyan Wang, Daqian Zhan, Hao Zhang, Richard S. Lee, Peisong Gao, Hernando Lopez-Bertoni, Mingyao Ying, Jian Jian Li, John Laterra, Mary Ann Wilson, and Shuli Xia. 2019. Extracellular Matrix Protein Tenascin C Increases Phagocytosis Mediated by CD47 Loss of Function in Glioblastoma. *Cancer Research* 79, 10 (05 2019), 2697–2708. https://doi.org/10.1158/0008-5472.CAN-18-3125

[12] Boris Muzellec, Maria Teleńczuk, Vincent Cabeli, and Mathieu Andreux. 2023. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *Bioinformatics* 39, 9 (09 2023), btad547. https://doi.org/10.1093/bioinformatics/btad547

[13] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102, 43 (2005), 15545–15550. https://doi.org/10.1073/pnas.0506580102

[14] Jeremy B. Swann and Mark J. Smyth. 2007. Immune surveillance of tumors. *The Journal of Clinical Investigation* 117, 5 (5 2007), 1137–1146. https://doi.org/10.1172/JCI31405