

Medulloblastoma: Human Specific Genes Contributions

Sara Baldinelli¹, Letizia De Pietri¹, Gaia Faggin¹, Huyen Pham¹, Roan Spadazzi¹

Abstract

Medulloblastoma (MB) is the most prevalent malignant brain tumor in childhood. It poses a significant health challenge and impacts children at a rate tenfold higher than adults. The disease is molecularly heterogeneous, with distinct subgroups (WNT, SHH, GP3, and GP4) exhibiting variations in cytogenetics, mutational profiles, and gene expression signatures, shaping treatment strategies and outcomes. Since the disease origins are primarily genomic alterations, we aim to investigate if the MB driving genes are also Human Specific Genes (HSGs). As HSGs are unique to humans, they often underpin distinctive traits that evolved recently or undergo significant changes. As a result, they often play critical roles in brain functions, immune systems, and metabolic processes. Our research seeks to identify MB HSGs, their function and their potential for being bio-markers and drug targets. To this aim, we retrieved both a cases and a control MB datasets, we pre-processed them and we analyzed the outputs. In particular, we carried out network and functional analyses on sets of genes that were found to be differentially expressed between MB cases and controls or between one of the four MB subgroups and controls. The results of the analyses confirmed the robustness of our datasets, showing patterns expected in the MB gene environment. Also, we identified a set of genes that can be considered for further studies as they were found to be particularly significant in MB development.

¹University of Trento, Trento 38123, Italy

Introduction

Medulloblastoma (MB) stands as the most prevalent malignant brain tumor in childhood, constituting roughly 20% of pediatric central nervous system tumors [1], and impacting children at a rate tenfold higher than adults [2]. This condition is categorized into distinct molecular subgroups: WNT, Sonic Hedgehog (SHH), Group 3 (GP3) and Group 4 (GP4) (**Table 1**). These subgroups exhibit striking differences in cytogenetics, mutational profiles, and gene expression signatures, all of which significantly influence the treatment strategies and outcomes [3].

Given its prevalence in childhood, MB's origins cannot be attributed to environmental or age-related factors. Instead, it is primarily driven by genomic alterations. Previous studies showed that the least to most common sub-groups are: WNT, SHH, Group 3 and Group 4. Respectively, they have survival rates of: ~95%, ~60-80%, ~50% and, ~75% [4, 5].

MB Subtype	WNT	SHH	Group 3	Group 4
Prognosis	Very good	Good	Poor	Intermediate
Signaling Pathway	WNT	SHH, PI3K	TGF- β	NF- κ B
Survival Rates	~95%	~60-80%	~50%	~75%
HSGs	ALOX5, ARHGAP42	CNTNAP2, CHRM3	RAB3A, CHRM3	PP1/P5K2

Table 1. Table describing the main features of the four different subtypes of MB. The last row highlights the HSG we found, at the end of this study, to be related to each subtype.

Currently, an insufficiently accurate disease risk stratification does not allow for therapies that are specific to standard or high-risk patients. Identifying potential biomarkers specific to each of the four sub-groups and associating them to a specific risk-ratio could introduce the application of more personalized therapies.

A way to narrow down the set of genes to investigate, is to consider the set of Human Specific Genes (HSGs) since a substantial portion of them plays a critical role in brain functions, alongside their roles in the immune system and metabolic processes [6]. HSGs are genes that are unique to the human species and have no direct counterparts in the genome of other closely related species, such as our primate relatives. These genes are often considered to be responsible for traits or functions that are distinctive to humans; they may have evolved relatively recently in the human lineage or have undergone significant changes that make them distinct from the genes found in other species. Identification and study of HSGs could provide insight into what sets humans apart from other species.

Consequently, in our project we will investigate the contribution of HSGs with respect to non-HSGs in the context of this disease. More specifically, our efforts are directed towards identifying HSGs that may contribute to the development of MB. To this purpose, we retrieved case and control datasets respectively from GSE155446 [7] and GSE118068 [8], pre-processed them and performed Differential Expression (DE) Analyses to obtain a list of statistically significant Differentially Expressed Genes (DEGs).

The so obtained list was expanded through causal relationships identified by the GENE@HOME project [9]. This was followed by network and functional analyses, which allowed us to highlight the potential role of the imputed genes. These analyses, together with a literature search, led us to formulate hypotheses about the role of HSGs in the context of MB.

1. Materials

1.1 Datasets

1.1.1 GSE155446

Cohort GSE155446 [7] represents a comprehensive exploration of cellular heterogeneity within 28 childhood MB cases, classified into different subgroups, including 1 WNT, 9 SHH, 7 GP3 and 11 GP4 MB.

This study investigates cellular diversity in childhood MB using single-cell RNA sequencing, revealing distinct neoplastic cell sub-populations associated with mitotic, undifferentiated and neuronal profiles.

The cohort comprises a total of 39946 cells and explores the expression patterns of 26841 genes.

1.1.2 GSE118068

GSE118068 [8] cohort was chosen to serve as a reference point for normal cellular states.

This dataset comprises 9 hindbrains gathered from diverse developmental stages, encompassing both embryonic and postnatal phases. Given that MB predominantly emerges in childhood, our emphasis was directed specifically towards the 4 postnatal samples. By doing so we retrieved a total of 27997 genes and 28210 cells.

1.2 Homologous genes

We used a list of homologous genes downloaded from Ensembl (version 110) [10] to identify matches between the two datasets under investigation. Employing this resource provided us with a reliable foundation for pinpointing meaningful homologous relationships.

1.3 Human Specific Genes

A list of 856 of HSGs was retrieved from the study conducted by Bitar et al [6]. This comprehensive list was employed as a valuable reference for identifying HSGs within the broader set of genes after the DE analysis.

1.4 FANTOM & GENE@HOME

The GENE@HOME [9] initiative represents an implementation of the PC-IM algorithm, designed to enhance Gene Regulatory Networks (GRN). The project utilizes gene causal relations extracted from the FANTOM dataset [11] to unravel the complexities of these regulatory networks. Through its utilization we aim to expand the set of MB HSGs and identify causal relations to be further investigated during our analysis.

1.5 STRING

For an initial functional investigation the STRING database [12] with default parameters was used. STRING compiles data from various origins, incorporating experimental findings, computational predictions, and publicly available text collections. It additionally facilitates the identification of functional enrichments within user-submitted protein lists by leveraging various functional classification systems such as Gene Ontology (GO) [13], Pfam [14], and KEGG [15].

1.6 Enrichr

Enrichr [16] is a comprehensive web-based and mobile software application designed to enhance the integration of gene-set libraries. This platform introduces novel methods for ranking enriched terms, offering an alternative approach to traditional techniques and provides a range of interactive visualization tools that facilitate the display of enrichment results. From Enrichr website we have decided to investigate the enriched diseases and tissues respectively from the DisGeNET [17] and Jensen TISSUES [18] databases.

1.7 Reactome Pathway Database

The Reactome website [19] serves as an interactive platform that provides users with a visual representation of established biological processes and pathways. This graphical map also functions as an interface, allowing users to navigate through and access information about various components and their relationships. We aim to use this database to find the pathways in which the genes of our analysis may be involved.

1.8 Libraries

Scanpy

Since we were dealing with scRNA-seq datasets, we employed the *Scanpy* library [20], a powerful Python tool tailored for the analysis of single-cell gene expression data. *Scanpy* is purposefully designed in collaboration with *anndata* [21], encompassing a wide array of functionalities.

NumPy

Since the huge amount and complexity of data and the necessity to enhance our analysis, we used the *NumPy* [22] library in order to gain more efficiency during the normalization processes and the DE Analysis. Moreover, *NumPy* can be well integrated with *SciPy* and *Statsmodels* that were used to build the script for the DE analysis.

SciPy and Statsmodels

SciPy [23] and *Statsmodels* [24] are two Python libraries enabling users to delve into data exploration, conduct statistical model estimations, and execute statistical tests.

ggplot2

In order to visualize the results of the DE analysis and to get visual support in the choice of the thresholds, we generated volcano plots through the *ggplot2* library. *ggplot2* is an open-source data visualization package for creating graphics, based on the *Grammar of Graphics*, a comprehensive framework for data visualization that breaks up graphs into semantic components, including scales and layers.

NetworkX

To analyze the properties of the whole network we used *NetworkX*, a Python package specifically designed for exploring and analyzing networks [25]. *NetworkX* offers a comprehensive set of tools for network representation, allowing nodes to be any hashable Python object and edges to contain diverse data. The package provides various data structures to handle different types of networks, along with a wide range of implemented algorithms for measuring network properties.

2. Methods

2.1 Homologous Genes Filtering

To ensure direct comparability between human and murine datasets, we performed homology mapping using the annotation provided by Ensembl version 110 [10]. This approach allowed us to establish genes correspondences between the two organisms, facilitating significant convergence in the analysis of gene expression data.

Subsequently, we conducted meticulous filtering to identify and retain only genes common to both datasets. This step is crucial to eliminate any discrepancies in genomic annotation and ensure that comparative analyses are conducted on a consistent set of genes.

The outcome of this process is a set of shared homologous genes, forming the foundation for subsequent analyses.

2.2 Quality Control

GSE155446

In the GSE155466 dataset, the quality control (QC) process started with the removal of cells expressing less than 200 genes per cell, with the aim to eliminate cells having potentially low-quality.

Subsequently, genes with fewer than three occurrences in the matrix were filtered out, ensuring that only genes with a minimum

level of expression across cells were retained for further analysis.

Cells expressing an excessive number of genes (more than 4845) were further filtered to maintain data quality by eliminating cells that are potentially contaminated or damaged.

Finally, cells with a high percentage of mitochondrial (more than 30%) genes were filtered since this condition is an indicator of potential stress or data artifacts.

The threshold values for these filters were determined by plotting the distribution of occurrences for each cell and the percentage of mitochondrial genes. This data-driven approach allowed for a nuanced determination of thresholds, considering the dataset's specific characteristics.

GSE118068

Similar to the GSE155446 dataset, the QC process for GSE118068 involved the removal of cells with fewer than 200 of expressed genes per cell.

Genes with fewer than three occurrences in the matrix were then filtered out, ensuring a minimum level of gene expression across cells.

To further enhance data quality, cells with an excessive number of expressed genes (more than 3625) and those with a high percentage of mitochondrial genes (more than 10%) were filtered out. As with the GSE155446 dataset, the threshold values for these filters were determined by analyzing the distribution of occurrences and the percentage of mitochondrial genes for each cell. Cells with an excessive number of expressed genes and those with a high percentage of mitochondrial genes were filtered out as well.

2.3 Data Filtering for Common Genes

After initially identifying common homologous genes between human and murine datasets, an additional step of matrix re-filtering was performed. This step was essential to maintain consistency between the two matrices after the application of QC, which could have influenced the presence of certain genes.

The re-filtering was carried out to ensure that only genes that remained after QC in both datasets were retained during the normalization phase.

2.4 Data Normalization

We applied normalization to both the scRNA datasets separately. The primary goal of this step is to mitigate potential technical biases and ensure comparability of the data across different cells and samples.

2.4.1 Total reads count normalization

To achieve a comparable representation across individual cells, the read count was normalized so that the sum of counts for each cell reached a fixed value. This was done by normalizing the scRNA data such that the total count for each cell equaled **1e4**. This approach is widely employed to ensure a uniform representation of cells within each sample.

2.4.2 Logarithmic transformation

Following the normalization of total read counts, a logarithmic transformation was applied to the data. This transformation was executed using the *log1p* function, which adds 1 to read counts before applying the natural logarithm. This operation is commonly performed to handle zero values in gene expression data and stabilize variance, making the output more suitable for statistical analyses and visualizations.

2.5 Genes Collapsing

To address instances where multiple murine genes map to the same human gene after homology mapping and normalization, we implemented a gene collapse strategy. This involved aggregating values for murine genes that share the same human gene identifier by calculating their mean expression.

This gene collapse approach serves to consolidate the information from multiple murine genes that correspond to a single human gene, providing a more integrated representation of cross-species expression patterns. By taking the mean expression contribution of homologous murine genes, thereby reducing redundancy and simplifying downstream analyses.

The application of gene collapse promotes a more coherent interpretation of cross-species gene expression dynamics. This step is instrumental in refining the dataset and facilitating a nuanced exploration of conserved and divergent gene profiles across tumor and control datasets.

2.6 Datasets Integration

To achieve a comprehensive integration of the human and murine datasets, we employed the *ingest* method from *Scanpy* [20], a powerful technique designed for harmonizing heterogeneous scRNA data. Ingest leverages biological information shared between datasets to align and merge them into unified representation, fostering a more cohesive analysis of cross-species gene expression. We name the resulting matrix I .

2.7 Dataset Aggregation

Following the integration of the two datasets, we performed aggregation steps. Two aggregations were implemented:

- **Aggregation by Tumor and Control Samples:** The initial aggregation involved collapsing the data by calculating the mean expression values for tumor samples and control samples separately. This provided a comprehensive overview of the average expression patterns within each group, enabling initial explorations of differential expression trends between tumor and control conditions.
- **Subgroup-Specific Aggregation:** A subsequent aggregation focused on collapsing the data into four distinct tumor subgroups within MB and control samples. This finer-grained approach allowed us to assess subtype-specific expression changes and gain insights into potential variation within tumor samples.

Originally intended for DE Analysis, these aggregated tables were subsequently repurposed to evaluate fold changes resulting from our DE Analyses conducted on the entire integrated dataset.

This decision was motivated by the desire to capture comprehensive trends across diverse conditions, considering the interplay within the integrated dataset. The two aggregated matrices are referred to as respectively $A1$ and $A2$.

2.8 Differential Expression Analysis

To perform the DE Analysis we employed the Wilcoxon-Mann-Whitney test applied to the integrated matrix I , for all cases and for each MB subtype against controls. This choice was driven by the fact that we were handling normalized counts rather than raw ones that, on the contrary, are used by other methods such as *DESeq2* [26]. In addition, the Wilcoxon-Mann-Whitney test is non parametric, it does not make assumptions regarding the data distributions and it has advantages in terms of outlier robustness. This was complemented with Benjamini-Hochberg correction in order to account for multiple testing allowing to limit the false discovery rate. Those steps were implemented using the *SciPy* [23] and the *statsmodels* [24] Python libraries.

1. Cases VS Controls.

A first DE Analysis was performed on the integrated I table directly measuring the differences between the distributions of each gene in cases against controls.

2. MB Subgroups VS Controls.

The second DE Analysis was carried out on the same integrated matrix I by considering four pairwise comparisons: one for each subgroup against controls. We chose this procedure since the Wilcoxon-Mann-Whitney test only allows comparisons between pairs but is not limited by different sample sizes.

After saving each Benjamini-corrected p-value, we decided to visualize each set of DEGs. In order to achieve this, we built Volcano plots (**Figure 4**) through the *ggplot2* package for every DE analysis: cases VS controls, Gp3 VS controls, Gp4 VS controls, WNT VS controls and SHH VS controls. We computed the a FC function as the absolute value of the base 2 logarithm of the fraction of the means. This last parameter was retrieved respectively from the $A1$ and $A2$ aggregated matrices, one for cases against controls and one for subtypes against controls. Volcano plots represent on the x axis the fold change and on the y axis $10000^{-\text{adjusted-pvalue}}$; the basis was chosen as a stretching factor to better visualize the distributions of the p-values and accentuate the disparity between genes that are differentially expressed and genes that are not.

From the volcano plot we observed that many output adjusted p-values, being extremely low, were collapsed to 0. Thus, in order to get a reasonable number of significantly DE genes for the following steps, we set a p-value threshold of 0 and a FC threshold of $|(\log_2(FC))| > 2$. The only exception to this parameter setting is the volcano plot resulting from WNT vs controls DE analysis: p-values did not indeed collapse at the 0 value and so we set a p-value threshold of 10^{-7} .

2.9 HSGs intersection

The sets of DE genes that passed the filtering (colored in the volcano plots in **Figure 4**) were collected and then intersected with the HSGs, obtaining five lists of significantly differentially expressed HSGs, one for each DE Analysis test. The list of all differentially expressed case-related genes was also separately kept for further investigation.

2.10 FANTOM integration

The gene lists resulting from the DE Analysis and filtering were expanded using causal relations from the FANTOM dataset [11] made by the GENE@HOME Project [9]. For each gene, we retrieved their respective causal relations from the isoform 1 tables (*p1@gene.csv*) due to them having the highest number of causal relations. We kept a relative frequency threshold of 0.95, slightly more conservative than the 95th percentile, and merged them together into a unique edge list file.

2.11 Network Analysis

Network construction was performed via the *NetworkX* [25] Python library [27]. Network edges were inputted using a separate merged edge list file for both cases and subtypes HSGs. Following, we excluded from the network all the nodes without any connected component and we computed the betweenness centrality measure of each node. To all nodes we assigned a size based on the centrality measure and built the network accordingly using the *kamada_kawai_layout* to compute the node coordinates. Notably (**Figure 5 a-d**), the colors of the nodes correspond to their community index.

Another, differently-shaped, network (**Figure 5e**) was built by tuning the size parameters and colors in order to highlight the HSGs found among all case-related genes that passed the filtering procedure (2.8 Differential Expression Analysis). The edges were retrieved using STRING [12] by inputting the previously mentioned case-related gene list: out of the 8 HSGs differentially expressed in cases, 5 had edges connecting them to the rest of the network with a medium STRING confidence score of >0.400. The Python codes used to generate the networks are available on the GitHub repository (Supplementary Material).

2.12 Functional Analysis

A functional analysis was performed on the five FANTOM-integrated [11] lists of DEGs.

This step was brought forward using, as a starting point, the STRING [12] functional enrichment analysis to gain an overview regarding a possible involvement of our genes in specific pathways and diseases, presence in tissues and molecular functions. A first deeper investigation was performed on Enrichr [16] by retrieving information from DisGeNET [17] about human diseases associated to our genes of interest, and from Jensen TISSUES [18]. Moreover, through Reactome Pathway Database [19] we investigated the enriched pathways.

In the next page, a schematized image of the workflow is available (**Figure 1**).

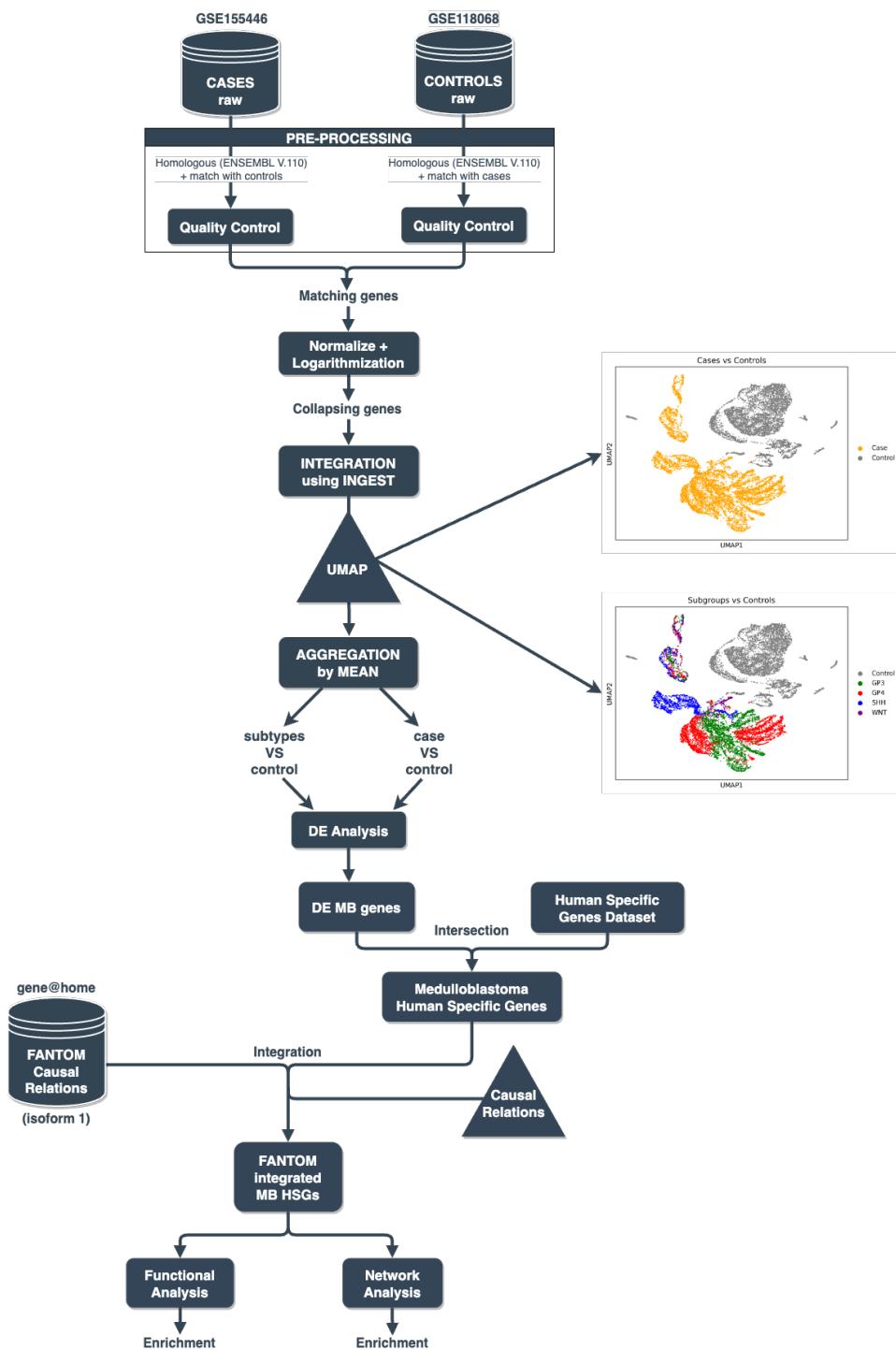


Figure 1. General workflow depicting the main steps of the project.

3. Results

3.1 Homologous Gene Filtering

In the initial phase of homologous gene identification, we sought to establish orthologous relationships between human and murine genes. Following the mapping using Ensembl (v.110) [10], the results revealed a total of 15180 homologous genes in the human dataset and 16866 genes in the murine dataset. Subsequently, to ensure consistency and compatibility between the two datasets, we performed a refined filtering step to identify and retain only the homologous genes present in both datasets. After applying this filtering criterion, 14524 homologous genes were identified as shared within the two matrices.

3.2 Quality Control

GSE155446

After the QC procedures, the GSE155446 dataset retained 38598 cells and 14219 genes. Following this process, an insightful observation emerged. Notably, the cells that did not pass the QC set by applying thresholds were consistent with those identified when initially considering the entire dataset.

GSE118068

As with the GSE155446 dataset, the threshold values for these filters were determined by analyzing the distribution of occurrences and the percentage of mitochondrial genes for each cell. Post-QC, the GSE118068 dataset retained 26914 cells and 14484 genes. As proof of concept, also in this case a parallel observation was made: the cells that did not meet the QC criteria established through threshold analysis were found to be the same as those identified when the dataset was considered in its entirety.

3.3 Data Filtering for Common Genes

Due to QC measures, a subset of genes in both datasets did not meet the defined criteria and were subsequently filtered out. To maintain consistency and comparability post-QC, we performed a secondary data filtering step specifically targeting common genes between the human and murine datasets.

As a result of refiltering step, 13477 genes were identified as common homologous genes shared within the two matrices. However, some genes from the murine dataset were found to map for more than one gene in the human cohort. For this reason, a genes collapsing strategy was employed after the normalization process.

3.4 Data Normalization

A data normalization step was applied to both the two datasets in order to mitigate technical biases and promoting data comparability.

The target sum for each cell was set to $1e4$, ensuring a comparable representation across individual cells. Subsequently, a logarithmic transformation was applied using \log_{10} function.

3.5 Genes Collapsing

In response to instances where multiple murine genes mapped to the same human gene, we implemented a gene collapse strategy. This involved aggregating expression values of these genes by calculating their mean expression.

As a result, we obtained two matrices, each with a consistent and unified set of genes. In total 13477 unique genes were retrieved for both the two datasets. In **Figure 2** the UMAP of the pre-processed dataset is shown.

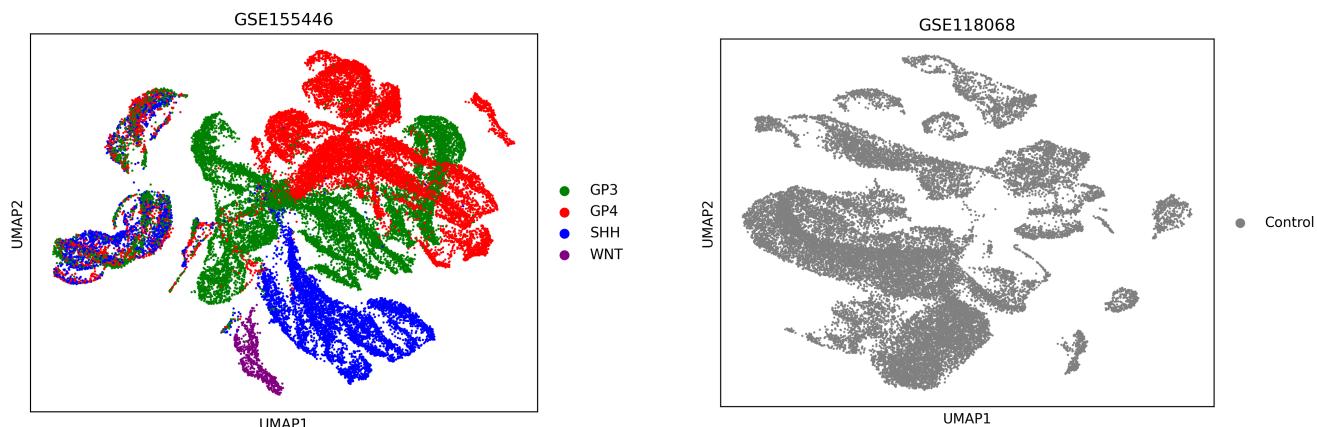


Figure 2. UMAP plots of pre-processed data from two distinct datasets: GSE155446 (MB - GP3, GP4, WNT, SHH) and GSE118068 (controls). These visualizations depict dimensional reduction through UMAP after pre-processing of gene expression data.

3.6 Datasets Integration

The integrated dataset is composed of 65506 cells and 13477 genes. In **Figure 3** the UMAP of the integrated dataset is shown.

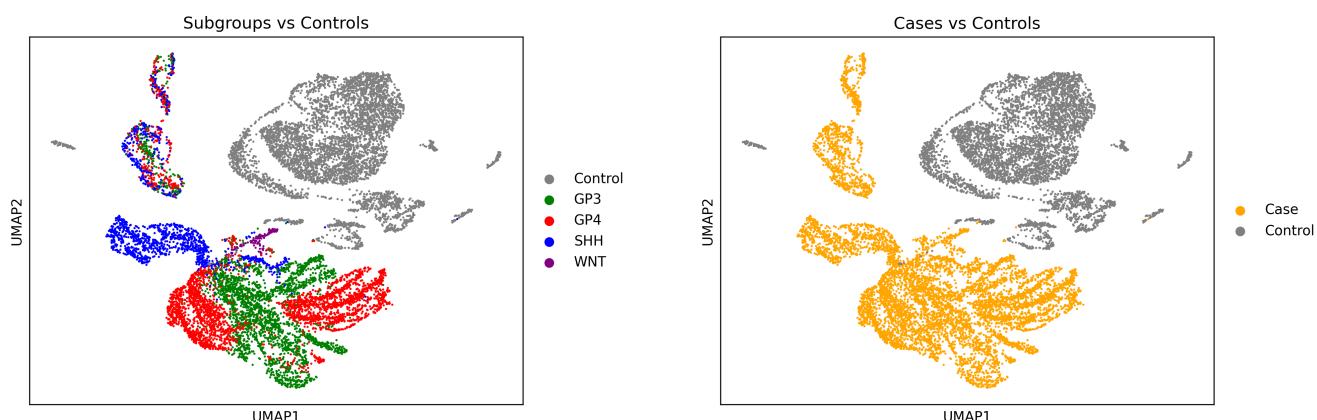


Figure 3. The UMAP plots of the integrated dataset were generated by sampling 80% of the cells from the entire matrix, while maintaining the proportions of composition.

3.7 DE Analysis

The DE Analysis outputted adjusted p-values and fold changes for all 13477 genes. Those were used to build the volcano plots (**Figure 4**): firstly by considering all genes and secondly by plotting only the human specific ones. The filtering procedure applied to cases and subgroups outputted what we ultimately refer to as hit genes, used in both Network and Functional analysis and highlighted in the volcano plots (**Figure 4**, second column). The difference between the quantity of up and down-regulated genes is distinguishable when comparing cases against the four subtypes.

3.8 Network Analysis

By expanding with FANTOM [11] causal relations our hit genes, we ended up building 4 small networks in order to gain further knowledge regarding whether our human specific hit genes were causally related to one another with bridge-like causal relations. The size of the nodes is directly proportional to their betweenness centrality measure and logically the hit genes are represented as community hubs surrounded by the cloud of FANTOM [11] genes and sometimes connected to one another either via intermediate genes or, less often, directly.

WNT and SHH subgroups (**Figure 5c,d**) have more complex networks due to them having a higher number of hit genes passing through the filtering and FANTOM [11] integration. Out of the three Gp3-HSGs (**Figure 5b**), two had the bridge gene *KCNAB2* connecting them. Note that the Gp4 MB subgroup network is missing due to the absence of causal relations connecting differentially expressed HSG.

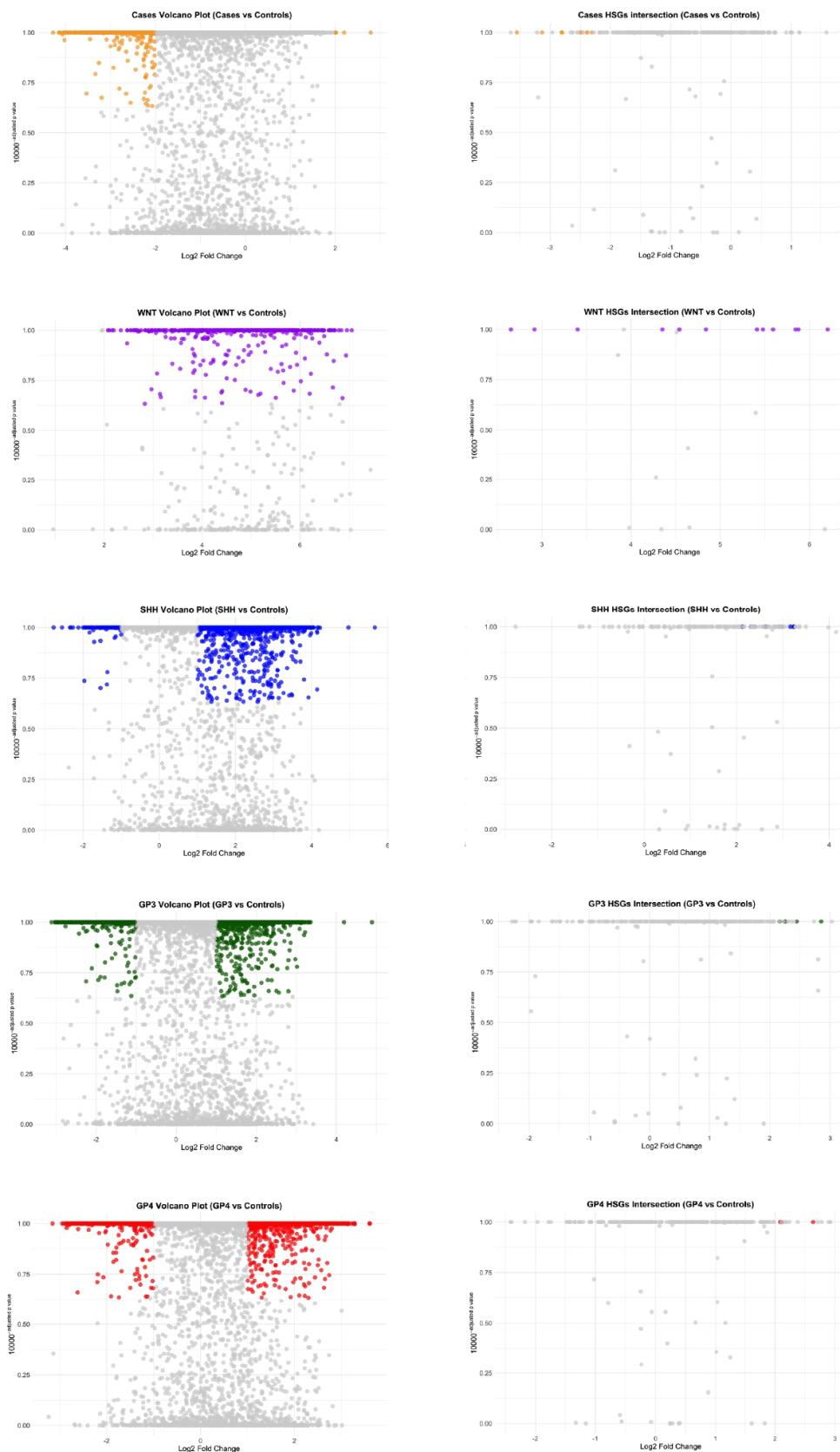


Figure 4. Panel displaying the 5 different volcano plots for the DE analysis. In order: Cases, WNT, SHH, Gp3, Gp4.

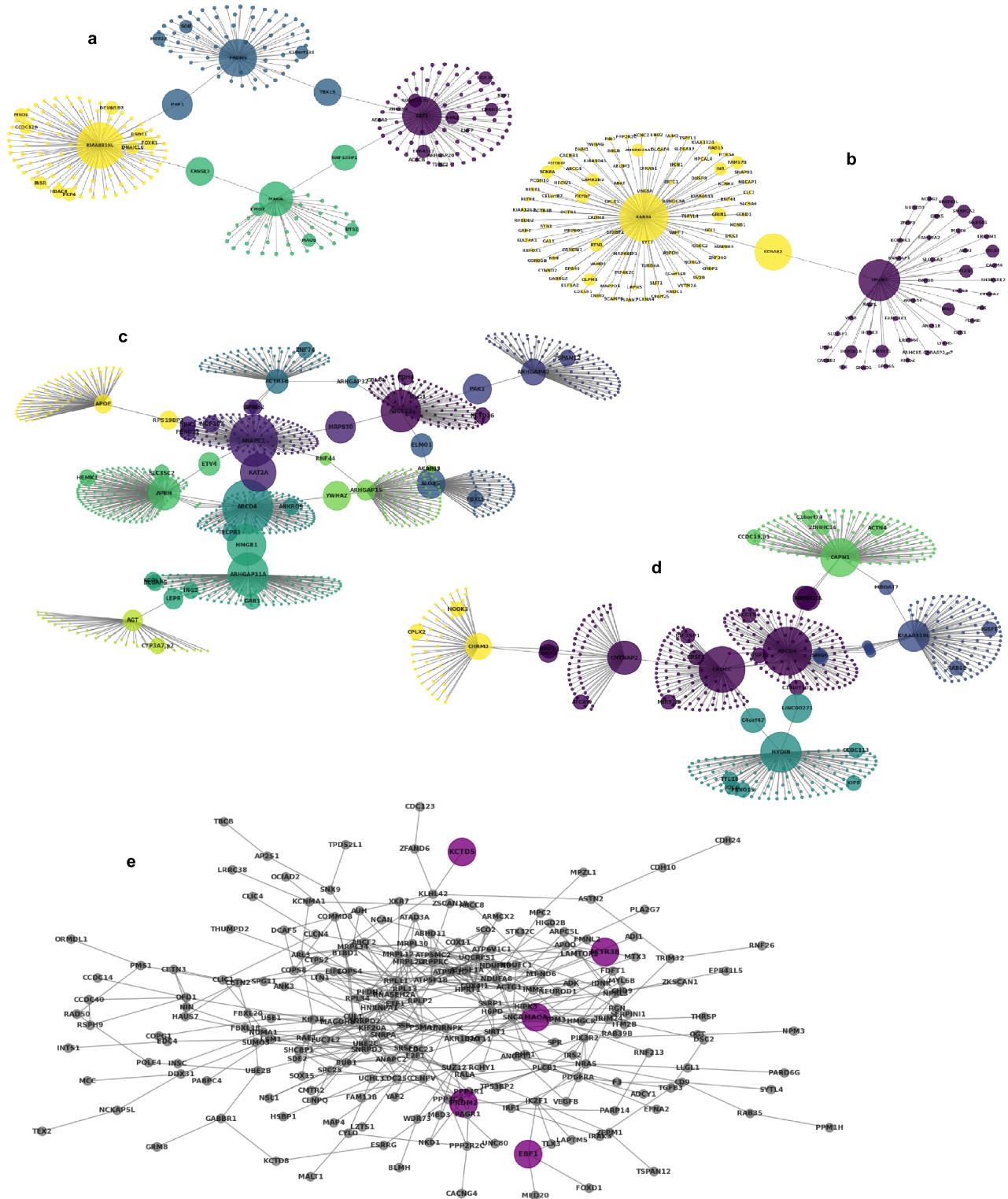


Figure 5. Panel depicting the 5 networks of our analysis. **a.** Network of the cases group intersected with HSGs; **b.** Network of the Gp3 subtype intersected with HSGs; **c.** Network of the WNT subtype intersected with HSGs; **d.** Network of the SHH subtype intersected with HSGs; **e.** Network representing all the cases, where the HSGs were highlighted. (High quality versions of the images are available in the Supplementary Material.)

Despite the fact that some HSGs did not have edges connecting them or other intermediate genes, they were kept for functional analysis and treated equally together with their FANTOM [11] causal relations. Using the complete list of case-DE genes we built a network (**Figure 5e**) by retrieving edges with STRING [12] and highlighting the presence of 5 HSGs out of the total of 8 previously found.

3.9 Functional Analysis

After gaining an overview on STRING [12] about which specific terms to look for in other databases, we looked for diseases associated to our genes on Enrichr [16]: all five gene sets showed a prevalence of neuro-related diseases such as epilepsy, mental retardation, intellectual disability and seizures. We then investigated the localization of the gene sets finding that all of them were significantly expressed in the brain and in the cerebellum. These results yielded highly significant with p-values of: 10^{-10} , 10^{-18} , 10^{-8} , 0.009, and 10^{-9} respectively for SHH, WNT, Gp3, Gp4 and cases.

Lastly, through the pathway enrichment analysis we observed a trend characterizing our genes. The cases and Gp3 sets showed involvement in pathways related to the neuronal system and brain functions. SHH, WNT, and Gp4 gene sets were found to participate in cell cycle pathways known to be, if altered, one of the major causes of tumors.

These analyses led us to acknowledge the importance of the involvement of the FANTOM [11] expanded DEGs in brain functions and localization.

Tables retrieved from Enrichr [16] and from Reactome Pathway Database [19] can be found in the Supplementary Material.

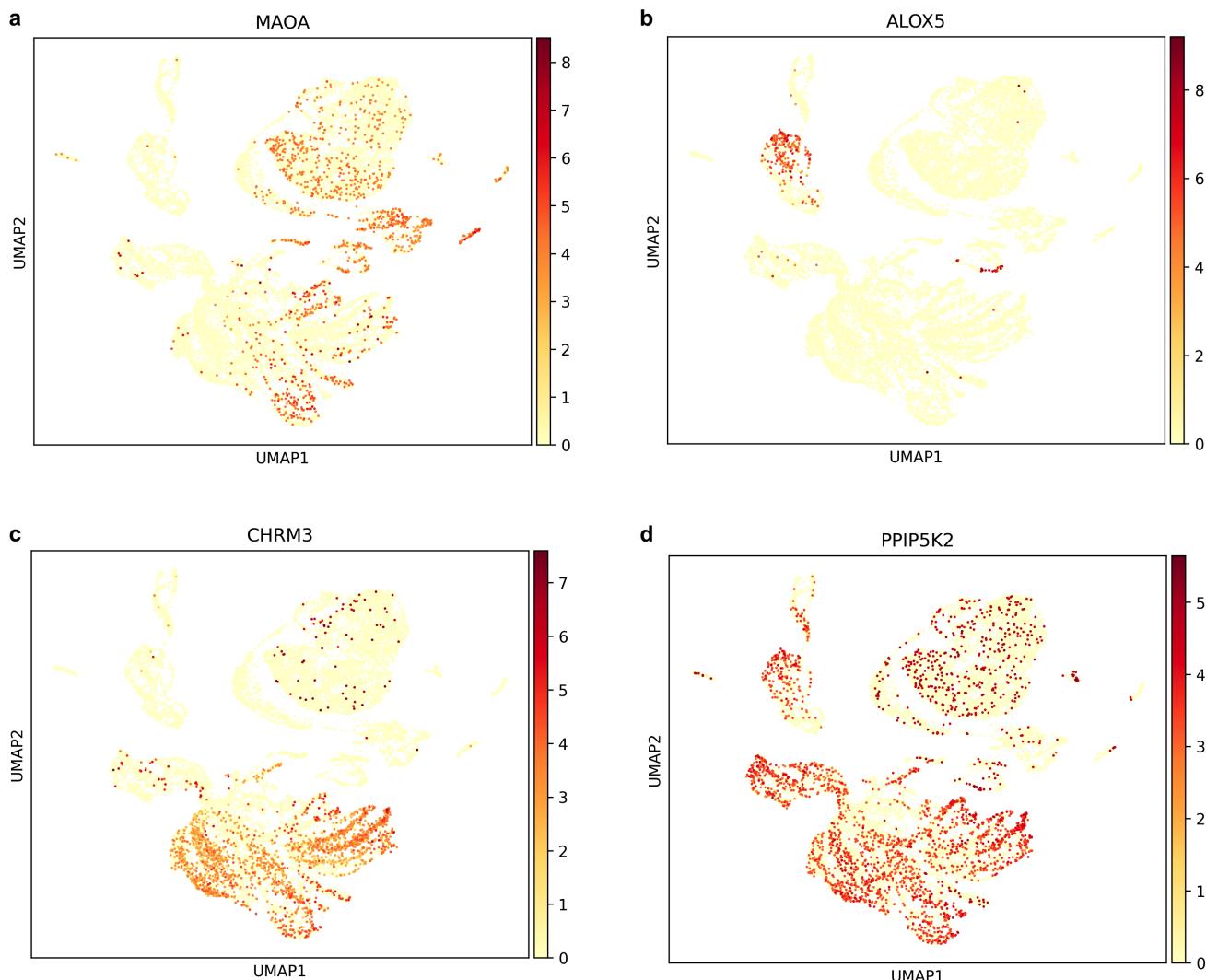


Figure 6. Panel displaying the UMAP-heatmaps of 4 hit genes: **a:** MAOA (for cases), **b:** ALOX5 (for WNT), **c:** CHRM3 (for Gp3 and SHH) and **d:** PPIP5K2 (for Gp4)

4. Discussion

Starting from a MB cases matrix and a control one, we performed a pre-processing pipeline and a DE Analysis. We obtained a set of HSGs that with a very high confidence were differentially expressed between MB cases and controls or among the four MB subgroups versus controls. Indeed, the Benjamini-corrected p-values turned out to be very low: the lowest ones were in the order of -300 or less and they were collapsed to the zero value. Although they might seem extremely low, it is reasonable to obtain such p-values since the distribution comparison that is performed during the DE analysis is applied to around 30.000 samples (**Table 2**).

DE Analysis	Sample Size
Cases vs Controls	38793 - 26719
WNT vs Controls	831 - 26719
SHH vs Controls	8123 - 26719
GP3 vs Controls	14222 - 26719
GP4 vs Controls	16750 - 26719

Table 2. Number of samples used for each DE Analysis: respectively cases and subtypes against controls.

Regarding the distribution of the p-values visible from the volcano plots (**Figure 4**), the crowding of points in the values 0 and 1 corresponds to robustness in our statistical test. Having many samples to compare one against the other, the Wilcoxon-Mann-Whitney test manages to separate with high degree of certainty the genes that are differentially expressed and those that are not, with fewer hits in the uncertain area.

We noticed a widespread down-regulation of genes in MB cases. There could be two possible reasons for this: first, it could be that these genes are naturally more active during pre-natal development rather than after birth like in our samples; second, there might be a technical problem affecting the data. On the other hand, the four subgroups show a prevalence of up-regulated genes suggesting that there could be a set of them expressed during the post-natal phase (1.1.2 GSE118068) specific to each MB subgroup.

From the UMAP plots (**Figure 3**), we can observe that cases and controls are well separated, with few overlaps. The distinction between the four MB subtypes is noticeable, with some intersection regions including all of them. These two observations constitute a solid basis to build our results on.

Based on the outcomes of the DE Analyses and on the topology of the networks (**Figure 5**), some HSGs and intermediate genes have been selected for subsequent in-depth investigations.

4.1 Cases DEGs

PRDM2, which was found to exhibit low expression in MB cases, is a histone 3 lysine 9 methyltransferase. The low expression level of this gene is in accordance with its role of tumor suppressor gene. It is usually low expressed in MB Group 3 compared to other subgroups [28]; unexpectedly the *PRDM2* gene is up-regulated in GP3 subgroup according to our data.

The *TBX15* gene, acting as a non-HSG bridge in the cases network and causally linked to *PRDM2* has also been related to a variety of tumors [29] when up-regulated, in agreement to our cases sample.

MAOA was also found to have a significantly low expression level in MB patients. *c-Myc*, a well known oncogene which is capable of promoting MB development, induces the expression of the *R1* gene that is able in turn to potentiate the effect of *c-Myc* on MB by increasing its stability. At the same time, the *R1* gene inhibits *MAOA* mRNA levels [30]. As a result, we described a pathway in which the *R1* gene is responsible both for enhancing cell proliferation and MB development through *c-Myc* stabilization and for inhibiting *MAOA* expression levels. This is coherent with our findings since the under-expression of the *MAOA* gene, also visible in our UMAP-heatmap (**Figure 6a**), has been demonstrated to be correlated with a high expression of the MB-related *c-Myc* oncogene.

4.2 WNT DEGs

The *ALOX5* gene, which resulted to be up-regulated in the MB WNT group, is responsible for Leukotrienes synthesis together with 5-lipoxygenase-activated protein (*FLAP*). Leukotrienes are inflammatory mediators that, besides having a wide range of biological functions including astrocyte proliferation, are also highly synthesized in MB cells. This happens through the production by Leukotrienes of Nestin, an intermediate filament protein required for MB progression [31]. Intuitively, blocking

ALOX5 expression could impair Nestin production and hence MB progression; it was indeed discovered that mutating the *ALOX5* gene, a markedly prolonged tumor latency was observed [31]. Hence, the discovery of *ALOX5* high expression level in the MB WNT subgroup (**Figure 6b**) confirms the role of this gene in MB progression through Leukotrienes production and suggests us to consider this gene as a candidate biomarker for the WNT group. To elucidate this hypothesis, further studies are needed. Moreover, the *ALOX5* gene could be experimented as a drug target for impairing MB advancement. Causally linked to *ALOX5*, *ELMO1*, a non-HS down-regulated gene, is also reported to be a MB-related gene and potential biomarker [32]. Following, the non-human specific *PAK1* gene, which resulted from the FANTOM [11] integration, was found to be up-regulated in the MB WNT subgroup and it plays an important role in MB cell migration [33]. Phosphorylation of Thr212 and Thr423 on *PAK1* by *ERK* in response to *PDGF* correlates with the functional activities of MB cells. The *PDGFR-ERK* signaling is relevant for MB growth and metastasis and it was also proven that their regulation towards the *PAK1-RAC1* signaling is linked to *PDGF*-mediated MB cell migration. Additionally, the *PDGFR* gene is also present in our cases DEGs, whose abnormal signaling can lead to disruptions in neuronal migration and normal cerebellar development [34][35]. These considerations underline the relevance of the *PAK1*, *RAC1*, *PDGFR* and *ERK* gene interactions in MB, making them robust candidate genes to take into consideration while profiling the WNT MB subtype. Lastly, we report that the *ARHGAP42* WNT-HSG is causally related to *PAK1* with a relative frequency above 0.95. They are both present in the WNT Network (**Figure 5c**) and are related to Rho-GTPases as respectively an activator and an effector [36].

4.3 SHH DEGs

The gene *CNTNAP2*, usually linked to a syndromic form of autism spectrum disorder and other neurodevelopmental disorders, was found to be up-regulated in SHH MB. Upon investigating its role in existing literature, it was interestingly discovered that individuals with a mutation in this gene manifest clinical features associated with brain overgrowth [37]. Also, *CNTNAP2* plays a crucial role in neuronal development and synapse formation [38], thus its over-expression suggests a potential abnormal growth of brain tissue, which could contribute to MB development. To validate this hypothesis, further analysis and experiments are necessary to explore whether *CNTNAP2* could serve as a biomarker for MB.

MAP2, a non human specific bridge gene resulting from FANTOM [11] integration of the *CNTNAP2* gene, is also important during brain development and it was found to be up-regulated in its initial phases, decreasing only during adulthood [39]. This gene was indeed found to be up-regulated in our SHH dataset with a FC of about 9.

4.4 Gp3 DEGs

The observed up-regulation of *RAB3A* in the Gp3 MB subtype, prompts an exploration of its potential implications in MB pathogenesis and in the context of brain-derived neurotrophic factor (*BDNF*), up-regulated in the comparison between Gp3 and controls, but not human specific. The functional interplay between *RAB3A* and *BDNF* has been elucidated in a study published in September 2001 by Thakker et al. [40]. In this work, the authors employed cells derived from *RAB3A* knock-out mice to investigate the role of *RAB3A* in *BDNF*-induced synaptic plasticity. Notably, their observations with mutant mouse neurons revealed that *RAB3A* was crucial for the early increase in synaptic charge induced by *BDNF*. Following these findings, it becomes necessary to consider the broader implications of *BDNF* in MB. Contrary to its well-established role in promoting survival and differentiation in hippocampal neurons, *BDNF* appears to exhibit complex and context-dependent effects in MB cells. A separate study [41] demonstrated that human recombinant *BDNF* treatment significantly inhibited the viability of MB cells expressing *BDNF* mRNA. This unexpected result challenges the conventional understanding of *BDNF* as a pro-survival factor in neuronal cells [42, 43]. Further investigations are needed to discover the precise mechanisms through which *RAB3A*, in conjunction with *BDNF*, influences cellular responses and contributes to the observed alterations in MB biology.

The *CHRM3* HSG is found to be up-regulated in both SHH and Gp3 subtypes. Interestingly, *CHRM3* has been linked in the literature to Gp4 MB as well [44]. Those findings are comparable to the *CHRM3* UMAP-heatmap (**Figure 6c**) showing prevalence in all subtypes but WNT.

4.5 GP4 DEGs

PPIP5K2 is found to be up-regulated in Gp4 MB subtype (**Figure 6d**). In a paper published on April 2021 [45] we can notice how its correlation with *BMI1* expression at bivalent domains, marked by *H3K27me3* and *H3K4me3*, may suggest a potential regulatory mechanism in MB cells. The fact that *BMI1* co-localizes at these domains, particularly at the promoter of the *PPIP5K2*, highlights a regulatory relationship that may have significant implications for inositol metabolism in MB cells. The connection between *PPIP5K2* and cell proliferation, as demonstrated by decreased cell proliferation in colon cancer cells upon *PPIP5K2* silencing, suggests that *PPIP5K2* may play a crucial role in MB metabolism and growth. This finding prompts

additional investigation into the specific functions of *PPIP5K2* in MB and its potential as a predictive biomarker, especially in the context of *BMI1*-High patients.

After having analyzed in greater detail some HSGs in relation to their subtype, we can infer that those genes could play a role in MB development. It is possible to study more in depth both the hit genes we reported to strengthen known facts about MB and confirm hypotheses, as well as take into consideration a lower threshold post-DE, outputting more genes to be examined.

5. Limitations & Future perspectives

The primary constraint arises from the substantial size of single-cell datasets, demanding significant computational power. Consequently, we were unable to explore the entire spectrum of DE analysis tests due to impractical running time. The pre-processing of datasets, involving such as cell filtering, data scaling, and integration, consumed a substantial portion of our time. Additionally, the sheer size of the matrices necessitated memory partitioning to prevent loading the entire dataframe simultaneously and risking memory overflow. This process was complemented by creating intermediate output file checkpoints, which were later merged.

A secondary limitation emerged from the integration of data originating from two distinct experiments (cases and controls), resulting in the handling of single-cell normalized counts. This compelled us to manually employ the Wilcoxon-Mann-Whitney test on the integrated matrix I , precluding the use of other libraries like *DESeq2* [26] and *edgeR* [46] that operate on raw counts. Despite their utility, these methods saturate the false-positive discovery rate, prompting us to consider alternative approaches even when working with raw counts [47].

Furthermore, dealing with MB as a pediatric tumor posed challenges in acquiring sufficient control data due to the inherent difficulty in obtaining a large sample size for pediatric cases. Many tumor-related datasets lacked comprehensive diagnostic information, patient data, therapy details, current health status, and concurrent diseases, introducing potential variability, bias or noise into our analysis.

Additionally, the genes identified as differentially expressed in MB may not necessarily be causative. Therefore, caution is required in asserting them as target for MB-related therapies, necessitating further studies for validation. However, they can serve as potential MB biomarkers, being differentially expressed between cases and controls or among MB subgroups and controls.

Issues may also arise from the presence of rare genes, where differential expression could result from sequencing or mapping problems. Deep sequencing experiments could potentially address this challenge.

Limitations were also encountered during FANTOM [11] integration, as certain genes such as *FAM156A*, *AREG*, and *EIF3CL* were not found in the database, hindering the expansion with causal relations. This step was also restricted to examine only isoform 1, which possesses the highest number of causal relations for each gene of interest. To enhance the robustness of future causal relations, the consideration of all isoforms and wet lab techniques for validating imputed MB-related genes are recommended.

Looking ahead, performing DE Analysis comparing the MB subgroups could unveil exclusive genes for each subtype, enabling a more detailed diagnosis and facilitating improvements in personalized medicine. Validation of our analysis is crucial not only in laboratory settings but also in evolutionary contexts, examining gene evolution across different species (human, monkey, mouse).

Finally gathering additional data on gene expression from each stage of the disease and each subtype could facilitate trajectory analysis, providing insights into disease progression and subtype differentiation.

6. Conclusion

In this study, starting from a cases dataset and a control one, we aimed at investigating the role of HSGs in the context of MB. To this purpose, after pre-processing data, a DE Analysis was performed between MB cases and controls and between each of the four MB subgroups and controls. The resulting sets of DEGs were then intersected with a set of HSGs and, after FANTOM [11] integration, network and functional analyses were carried out on the so obtained sets of human specific DEGs. The results of the functional analyses confirmed the coherence of the expanded gene sets with respect to the context of MB.

The results of the DE Analysis together with the topology of the networks led us to better investigate the role of some genes through a literature search. A considerable number of human specific DEGs were found to be in accordance with what is reported in literature. These genes, respectively *MAOA*, *ALOX5*, *ARHGAP42*, *CNTNAP2*, *RAB3A*, *CHRM3*, and *PPIP5K2* are highly significant in MB development and progression; they are potential candidate disease biomarkers and can be considered for further studies. The only gene whose expression level conflicted with the behaviour described in literature is *PRDM2*.

Moreover, some genes that resulted from the FANTOM [11] integration were investigated. Their relationship with the HSGs

listed above was explored and it was found to be coherent with the current literature knowledge. This improved the reliability of our results and their robustness, together with describing in greater detail MB-subtype-specific environments.

Acknowledgements

We want to express our appreciation to Professors Tebaldi and Blanzieri for their support during the "Laboratory of Biological Data Mining" course. A special thank you goes to Matteo Ganesello for his continuous and precious help throughout the development of our project.

Abbreviations

HSG: Human Specific Gene; **DE:** Differential Expression; **DEG:** Differentially Expressed Gene; **MB:** Medulloblastoma; **SHH:** Sonic Hedgehog; **Gp3:** Group 3; **Gp4:** Group 4; **FC:** fold change

Supplementary Material

Github: <https://github.com/Sara-Baldinelli/LBDM-project/tree/main>

References

- [1] Vinod Kumar, Virender Kumar, Timothy McGuire, Donald W Coulter, John G Sharp, and Ram I Mahato. Challenges and recent advances in medulloblastoma therapy. *Trends in pharmacological sciences*, 38(12):1061–1084, 2017.
- [2] Roberto Carta, Giada Del Baldo, Evelina Miele, Agnese Po, Zein Mersini Besharat, Francesca Nazio, Giovanna Stefania Colafati, Eleonora Piccirilli, Emanuele Agolini, Martina Rinelli, et al. Cancer predisposition syndromes and medulloblastoma in the molecular era. *Frontiers in oncology*, 10:566822, 2020.
- [3] Jinyi Chen, Zhuang Kang, Shenglan Li, Can Wang, Xiaohong Zheng, Zehao Cai, Lexin Pan, Feng Chen, and Wenbin Li. Molecular profile reveals immune-associated markers of medulloblastoma for different subtypes. *Frontiers in Immunology*, 13:911260, 2022.
- [4] Paul A Northcott, David TW Jones, Marcel Kool, Giles W Robinson, Richard J Gilbertson, Yoon-Jae Cho, Scott L Pomeroy, Andrey Korshunov, Peter Lichter, Michael D Taylor, et al. Medulloblastomics: the end of the beginning. *Nature Reviews Cancer*, 12(12):818–834, 2012.
- [5] Volker Hovestadt, Olivier Ayrault, Fredrik J Swartling, Giles W Robinson, Stefan M Pfister, and Paul A Northcott. Medulloblastomics revisited: biological and clinical insights from thousands of patients. *Nature Reviews Cancer*, 20(1):42–56, 2020.
- [6] Mainá Bitar, Stefanie Kuiper, Elizabeth A O'Brien, and Guy Barry. Genes with human-specific features are primarily involved with brain, immune and metabolic evolution. *BMC bioinformatics*, 20(9):1–12, 2019.
- [7] Kent A Riemonyd, Sujatha Venkataraman, Nicholas Willard, Anandani Nellan, Bridget Sanford, Andrea M Griesinger, Vladimir Amani, Siddhartha Mitra, Todd C Hankinson, Michael H Handler, et al. Neoplastic and immune single-cell transcriptomics define subgroup-specific intra-tumoral heterogeneity of childhood medulloblastoma. *Neuro-oncology*, 24(2):273–286, 2022.
- [8] Maria C Vladoiu, Ibrahim El-Hamamy, Laura K Donovan, Hamza Farooq, Borja L Holgado, Yogi Sundaravadanam, Vijay Ramaswamy, Liam D Hendrikse, Sachin Kumar, Stephen C Mack, et al. Childhood cerebellar tumours mirror conserved fetal transcriptional programs. *Nature*, 572(7767):67–73, 2019.
- [9] Francesco Asnicar, Nadir Sella, Luca Masera, Paolo Morettin, Thomas Tolio, Stanislau Semeniuta, Claudio Moser, Enrico Blanzieri, Valter Caveccchia, et al. Tn-grid and gene@ home project: Volunteer computing for bioinformatics. In *CEUR WORKSHOP PROCEEDINGS*, pages 1–15. Aachen University, 2015.
- [10] Biomart: Biological queries made easy. <https://www.ensembl.org/biomart/martview/>, 2023. Ensembl version 110.
- [11] Functional annotation of the mammalian genome 5 (fantom5). <http://fantom.gsc.riken.jp/5/>, 2014. Accessed: October 16, 2023.
- [12] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2019.

- [13] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
- [14] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021.
- [15] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.
- [16] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.
- [17] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, page gkw943, 2016.
- [18] Oana Palasca, Alberto Santos, Christian Stolte, Jan Gorodkin, and Lars Juhl Jensen. TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database*, 2018:bay003, 02 2018.
- [19] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. The reactome pathway knowledgebase 2022. *Nucleic acids research*, 50(D1):D687–D692, 2022.
- [20] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.
- [21] Isaac Virshup, Sergei Rybakov, Fabian J Theis, Philipp Angerer, and F Alexander Wolf. anndata: Annotated data. *BioRxiv*, pages 2021–12, 2021.
- [22] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.
- [23] Eric Jones, Travis Oliphant, Pearu Peterson, et al. Scipy: Open source scientific tools for python. 2001.
- [24] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, pages 10–25080. Austin, TX, 2010.
- [25] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
- [26] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.
- [27] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.
- [28] Jennifer Stripay, Marie Morfouace, Christopher Vakoc, and Martine Roussel. Medu-24. interrogating the role of epigenetic regulators in myc-driven medulloblastoma. *Neuro-Oncology*, 21(Suppl 2):ii108, 2019.
- [29] Dongming Yan, Yaping Yu, Qiongwei Ni, Qingwen Meng, Haolin Wu, Shun Ding, Xiaoqian Liu, Caiying Tang, Qibing Liu, and Kun Yang. The overexpression and clinical significance of tbx15 in human gliomas. *Scientific Reports*, 13(1):9771, 2023.
- [30] Xiao-Ming Ou, Kevin Chen, and Jean C Shih. Monoamine oxidase a and repressor r1 are involved in apoptotic signaling pathway. *Proceedings of the National Academy of Sciences*, 103(29):10923–10928, 2006.
- [31] Fang Du, Larra Yuelling, Eric H Lee, Yuan Wang, Shengyou Liao, Yan Cheng, Li Zhang, Chaonan Zheng, Suraj Peri, Kathy Q Cai, et al. Leukotriene synthesis is critical for medulloblastoma progression. *Clinical Cancer Research*, 25(21):6475–6486, 2019.
- [32] Zhendong Liu, Ruotian Zhang, Zhenying Sun, Jiawei Yao, Penglei Yao, Xin Chen, Xinzhuang Wang, Ming Gao, Jinzhao Wan, Yiming Du, et al. Identification of hub genes and small-molecule compounds in medulloblastoma by integrated bioinformatic analyses. *PeerJ*, 8:e8670, 2020.

- [33] Liangping Yuan, Mariarita Santi, Elisabeth J Rushing, Robert Cornelison, and Tobey J MacDonald. Erk activation of p21 activated kinase-1 (pak1) is critical for medulloblastoma cell migration. *Clinical & experimental metastasis*, 27:481–491, 2010.
- [34] Johanna Andrae, Gijs Afink, Xiao-Qun Zhang, Wolfgang Wurst, and Monica Nistér. Forced expression of platelet-derived growth factor b in the mouse cerebellar primordium changes cell migration during midline fusion and causes cerebellar ectopia. *Molecular and Cellular Neuroscience*, 26(2):308–321, 2004.
- [35] Tobey J MacDonald, Kevin M Brown, Bonnie LaFleur, Katia Peterson, Christopher Lawlor, Yidong Chen, Roger J Packer, Philip Cogen, and Dietrich A Stephan. Expression profiling of medulloblastoma: Pdgfra and the ras/mapk pathway as therapeutic targets for metastatic disease. *Nature genetics*, 29(2):143–152, 2001.
- [36] Edward Manser, Thomas Leung, Harfizah Salihuddin, Zhuo-shen Zhao, and Louis Lim. A brain serine/threonine protein kinase activated by cdc42 and rac1. *Nature*, 367(6458):40–46, 1994.
- [37] Job O de Jong, Ceyda Llapashtica, Matthieu Genestine, Kevin Strauss, Frank Provenzano, Yan Sun, Huixiang Zhu, Giuseppe P Cortese, Francesco Brundu, Karlla W Brigatti, et al. Cortical overgrowth in a preclinical forebrain organoid model of cntnap2-associated autism spectrum disorder. *Nature Communications*, 12(1):4087, 2021.
- [38] Frances St George-Hyslop, Toomas Kivisild, and Frederick J Livesey. The role of contactin-associated protein-like 2 in neurodevelopmental disease and human cerebral cortex evolution. *Frontiers in Molecular Neuroscience*, 15:1017144, 2022.
- [39] Roohangiz Safaei and Itzhak Fischer. Regulation of microtubule-associated protein 2 (map2) mrna expression during rat brain development. *Journal of Molecular Neuroscience*, 1:189–198, 1989.
- [40] Smita Thakker-Varia, Janet Alder, Robert A Crozier, Mark R Plummer, and Ira B Black. Rab3a is required for brain-derived neurotrophic factor-induced synaptic plasticity: transcriptional analysis at the population and single-cell levels. *Journal of Neuroscience*, 21(17):6782–6790, 2001.
- [41] Anna Laura Schmidt, Caroline Brunetto de Farias, Ana Lucia Abujamra, Flávio Kapczinski, Gilberto Schwartzmann, Algemir Lunardi Brunetto, and Rafael Roesler. Bdnf and pde4, but not the grpr, regulate viability of human medulloblastoma cells. *Journal of molecular neuroscience*, 40:303–310, 2010.
- [42] Kevin G Bath and Francis S Lee. Variant bdnf (val66met) impact on brain structure and function. *Cognitive, Affective, & Behavioral Neuroscience*, 6(1):79–85, 2006.
- [43] Eric J Huang and Louis F Reichardt. Trk receptors: roles in neuronal signal transduction. *Annual review of biochemistry*, 72(1):609–642, 2003.
- [44] Paul A Northcott, Andrey Korshunov, Hendrik Witt, Thomas Hielscher, Charles G Eberhart, Stephen Mack, Eric Bouffet, Steven C Clifford, Cynthia E Hawkins, Pim French, et al. Medulloblastoma comprises four distinct molecular variants. *Journal of clinical oncology*, 29(11):1408, 2011.
- [45] Sara Badodi, Nicola Pomella, Xinyu Zhang, Gabriel Rosser, John Whittingham, Maria Victoria Niklison-Chirou, Yau Mun Lim, Sebastian Brandner, Gillian Morrison, Steven M Pollard, et al. Inositol treatment inhibits medulloblastoma through suppression of epigenetic-driven metabolic adaptation. *Nature communications*, 12(1):2148, 2021.
- [46] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.
- [47] Yumei Li, Xinzhou Ge, Fanglue Peng, Wei Li, and Jingyi Jessica Li. Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome biology*, 23(1):79, 2022.