Transcriptomics Landscape of Medulloblastoma: Identifying the Significant Molecular Signals in Subgroup Classification.

Sayan Patra

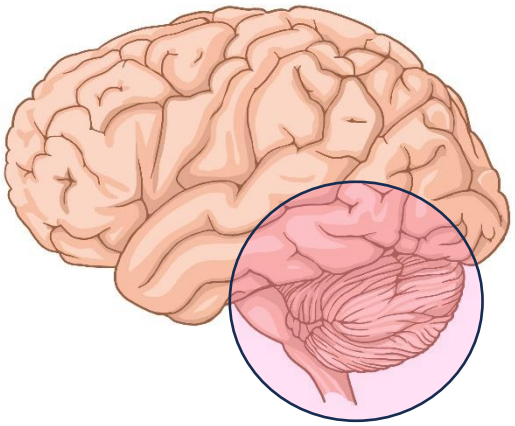
Department of Biosciences and Bioengineering, Indian Institute of Technology, Bombay

Abstract: Medulloblastoma is a prevalent type of tumor among juveniles. According to WHO, it is classified as Grade IV. It mainly occurs in children and is a development-related cancer. Although adults are also affected by this cancer, the ratio is minimal. Medulloblastoma could be classified into four subgroups, e.g., WNT, SHH, Group 3, and Group 4. The subgroups differ in their molecular characteristics and mechanisms. In this study, the molecular signatures for each of the subtypes are examined. This study also showed the similarity of the number of significant genes within the subgroups. All the subgroups form different clusters, except there could be a common link between Group 3 and Group 4 Medulloblastoma. The sample-wise correlation also shows a good correlation among samples. This study also showed highly upregulated group-specific genes in different subgroups that could be very good signatures of this cancer. Next, using chord plots and a database search, the chromosomes for the highly upregulated genes were found. The immune cell enrichment showed a little change in the specific immune cells in Medulloblastoma. Overall, the medulloblastoma subgroups showed a similar type of immune cell enrichment. The logistic regression-based classifier using the upregulated subgroup-specific genes showed an overall good classification of the subgroups, with an accuracy of about 82%. Finally, the study has shown that those molecular signatures could be used to classify the subgroups of medulloblastoma, mainly Group 3 and Group 4 Medulloblastoma.

Keywords: Medulloblastoma | WNT |SHH | Group 3 | Group 4 | Immune Cell Enrichment | Logistic Regression | WHO | NIH

Introduction:

Medulloblastoma (MB) is one type of CNS tumor. It is very common among juveniles, and the chance of its occurrence decreases with age. It is highly aggressive and malignant in nature, and due to this type of characteristics, WHO (World Health Organization) has put its name in Grade IV cancer type. The diagnosis has found that, generally, it occurs within the Posterior fossa of the brain (**Figure 1**), mainly in the Cerebellum (1). The occurrence of this tumor often blocks the path of the CSF (Cerebrospinal Fluid), and the storing of CSF causes high pressure over the skull (NIH NCI, https://www.cancer.gov/). Until now, proper genetic alterations have yet to be found to generate the Medulloblastoma. As it occurs during development, it is also known as an embryonal neuroepithelial tumor (NIH NCI https://www.cancer.gov/).



Posterior Fossa

Figure 1: The color-coded region near the cerebellum and brainstem is common for medulloblastoma occurrence.

According to WHO CNS classification, Medulloblastoma was classified as WNT, SHH, and non-WNT/non-SHH. As per the current WHO classification (2), the non-WNT/non-SHH subgroup is differentiated into two subgroups – Group 3 and Group 4. Hence, currently, there are four subgroups – 1. WNT, 2. SHH, 3. Group 3, 4. Group 4. Among the four subgroups of Medulloblastoma, the WNT subgroups followed the ß-catenin-mediated WNT pathway; the SHH subgroup followed the Sonic Hedgehog pathway; Group 3 and Group 4 Medulloblastoma don't have any proper pathways, but the Group 3 Medulloblastoma shows upregulation of Myc and Group 4 Medulloblastoma shows neuronal developmental related gene upregulations (3).

Among these four subgroups, the occurrence of the WNT is ~10%, and the occurrence of the SHH subgroup is ~30 %, whereas for Group 3 and Group 4, the occurrence is ~25 % and ~35 %, respectively (4).

According to the region of occurrence in the brain, the four subgroups also differ. All of them have different occurrences in the brain. The WNT subgroup mainly occurs in the lower rhombic lip (LRL) and embryonal dorsal brainstem. The SHH subgroup proliferates in the cerebellar hemispheres. Group 3 Medulloblastoma is often found within the vermis of the cerebellum and is most common among children. The midline/ front ventricle of the brain is the predominant site of occurrence of Group 4 Medulloblastoma (5).

The SHH is highly abundant among these four subgroups, and Group 3 Medulloblastoma is the most aggressive and often undergoes metastasis. Thus, the survival is high for other subtypes, but for Group 3 Medulloblastoma, it is deficient.

In this study, the subgroup-specific gene upregulation has been analyzed to find the differentiation among each group. The upregulated pathways in each subgroup have been analyzed using RNA sequencing data analysis. The upregulated genes and their chromosomes have been studied by database search to find the predominant chromosomes for each subset. Next, the immune cell enrichment analysis has been done to find the abundance of types of immune cells in different subgroups. Moreover, using Machine learning classification, our study shows the confidence of the upregulated genes for subgroup classification.

Methods:

1. Transcriptomics Data Collection:

The RNA sequencing data count matrix was taken from the Gene Expression Omnibus database (GEO https://www.ncbi.nlm.nih.gov/geo/). The GSE ID of the RNA sequencing project is GSE164677. The data contains 59 different group and control samples of Medulloblastoma from the Asian cohort. The disease samples were from Medulloblastoma tumors, but the Normal or Control samples were taken from para-tumor (6). The metadata showed that the whole count matrix contains 4 Normal, 6 WNT, 19 SHH, 15 Group 3, and 19 Group 4 Medulloblastoma samples. The count matrix has 35265 features or genes.

2. Differential Gene Expression Analysis:

After quality check and removal of outliers, the differential gene expression analysis for each subgroup was performed using the DESeq2 library (Version 1.44.0). From the normalized data of the DESeq2 analysis, the PCA plot was plotted. Next, the statistical analysis was performed for each Subgroup and Normal condition. The cut-off for the p-adj value was set to less than 0.05, and the log2 Fold Change was set to $-1.5 > \log2FC < 1.5$. The differentially expressed genes were taken for each subgroup.

3. Common Significant Genes for Subgroups:

An Upset plot was plotted to find the number of common significant genes among the groups. The Upset plot showed the size of the differentially expressed genes

as set size, followed by intersection genes for the groups by plotting a bar plot.

4. Enrichment Analysis:

The upregulated and downregulated genes for each subgroup were taken, and using the enrichR library (Version 3.2) of R, the top 20 upregulated and downregulated Gene ontology Biological Processes were analyzed to get insight into different processes among groups.

5. Subgroup-Specific Upregulated and Downregulated Genes:

Using R and Python, the upregulated and downregulated genes were filtered. The genes that are upregulated and downregulated only in one group and show opposite regulation in other subgroups were only taken to find the group-specific genes. Using the log 2-Fold Change values, the top 10 up and down-regulated genes were plotted into heatmaps by using the seaborn library.

6. Identification of Chromosomes of the Upregulated Genes:

After taking the top 20 upregulated genes from the group-specific genes for each group, a database search within the NCBI website (https://www.ncbi.nlm.nih.gov/) was to get the chromosome from where it was derived. Then, using an cord plot, the genes and their originated chromosomes were plotted to find the chromosome that showed high gene synthesis during the disease state.

7. Immune Cell Enrichment Analysis:

The raw count matrix of the RNA sequencing was uploaded into xCell, which was developed by UCSF (https://comphealth.ucsf.edu/app/xcell) to get immune cell type enrichment data. The output matrix was taken and standardized using the scikit learn library of Python, and using Seaborn, a heatmap was generated.

8. Logistics Regression-Based Classification Using the Top 50 Upregulated Genes from Each Subgroup

To find whether the selected genes could classify the four subgroups of Medulloblastoma, a machine learning-based classifier was created based on logistic regression. The Scikit-Learn library was used to access the Logistic Regression Model.

The top 50 subgroup-specific genes from different subgroups were taken, and from the count matrix, only those 200 features were selected. Using the metadata, the target was set to the Subgroups. Using the train test split of Scikit-Learn, the data was separated into 80% Train and 20% test set. The model was run, followed by plotting the ROC curve, Confusion Matrix, and Classification report.

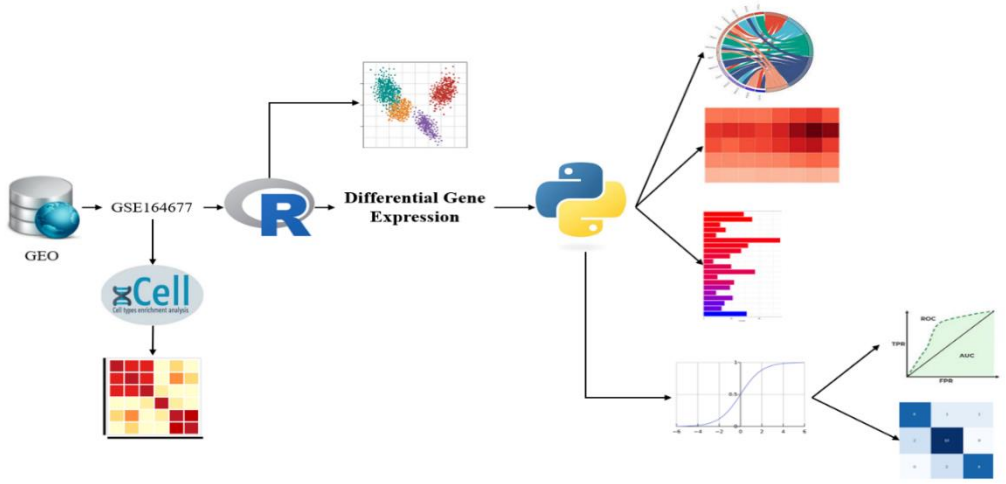


Figure 2: Overall workflow of the RNA seq data analysis of four subgroups of Medulloblastoma.

Result:

1. Dimensionality Reduction Showing Different Clusters of Control vs Different Subgroups

The PCA plot showed separated clusters for all the samples. The WNT and SHH subgroups were separated clearly, but Group 3 and Group 4 Samples were overlapping, possibly due to the similarity of these non-WNT / non-SHH groups **(Figure 3)**.

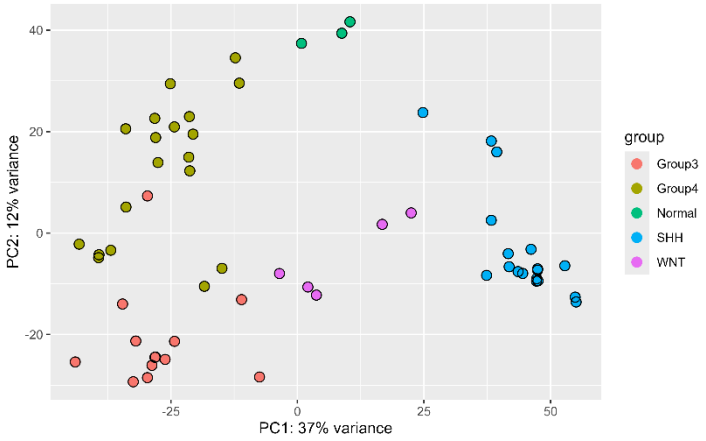


Figure 3: PCA plot for Control vs other Medulloblastoma subgroups.

2. Common Genes Among Different Subgroups of Medulloblastoma

The upset plot showed that more than 6000 genes were differentially expressed in SHH, and approximately a similar number of genes were differentially expressed in Group 3. Around 4500 genes were differentially expressed in WNT, and about 6000 genes were differentially expressed for Group 4. Among the differentially expressed genes, only 1938 genes were specific for SHH, 1141 genes were specific for Group 3, 1075, and 930 genes were specific for Group 4 and WNT, respectively. Group 3 and Group 4 share about 920 common genes. 700 genes were common for SHH and Group 3, and543 genes were the same for SHH and Group 4, 433 genes were similar for SHH and Group 4. 346 genes were similar in Group 3 and WNT, 132 genes were similar for Group 4 and WNT (**Figure 4**).

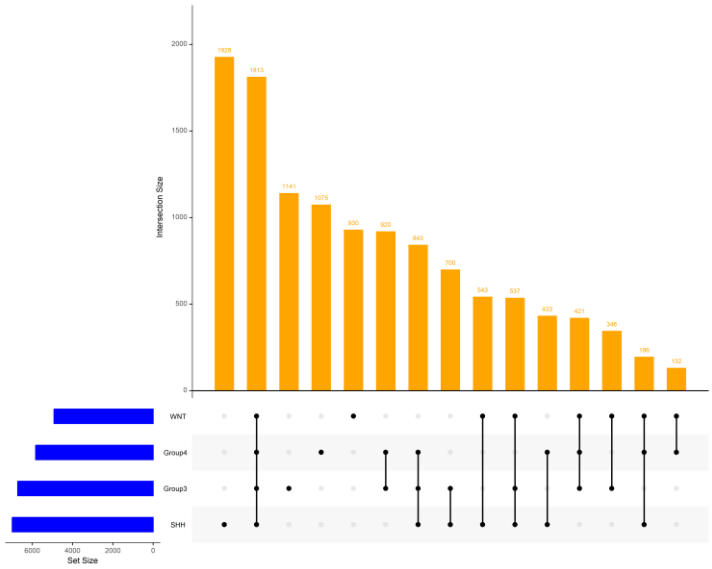


Figure 4: Upset plot shows the common proteins in different subgroups

3. Medulloblastoma Subgroups are Enriched with Different Biological Pathways

The WNT subgroup showed that the P53 signaling pathway, Hippo Signaling Pathway, WNT signaling Pathway, Transcriptional misregulation, TGF-ß signaling pathway, etc., were upregulated. Overall, the cell cycle-related pathways were very high in WNT. It was also found that the WNT pathway controls the PCP pathway (Planner Cell Polarity), which causes a change in the cytoskeleton element, and that may be why the cell cycle is upregulated. Besides, Glutamatergic synapse, Circadian entrainment, Synaptic vesicle cycle, GABAergic synapse, Dopaminergic synapse, Aldosterone synthesis, and Neuroactive ligand-receptor interaction were downregulated in WNT (**Figure 5 (A) (B)**).

The SHH subgroups showed upregulation of the Hedgehog signaling pathway, P53 signaling, ECM receptor interaction, Mismatch repair, Galactose metabolism, ribosome etc. The SHH pathway actually controls embryonic patterning and cell proliferation and growth, and that's why protein biosynthesis and transport are highly upregulated in this cancer by upregulation of ribosomes. Although similar to the WNT

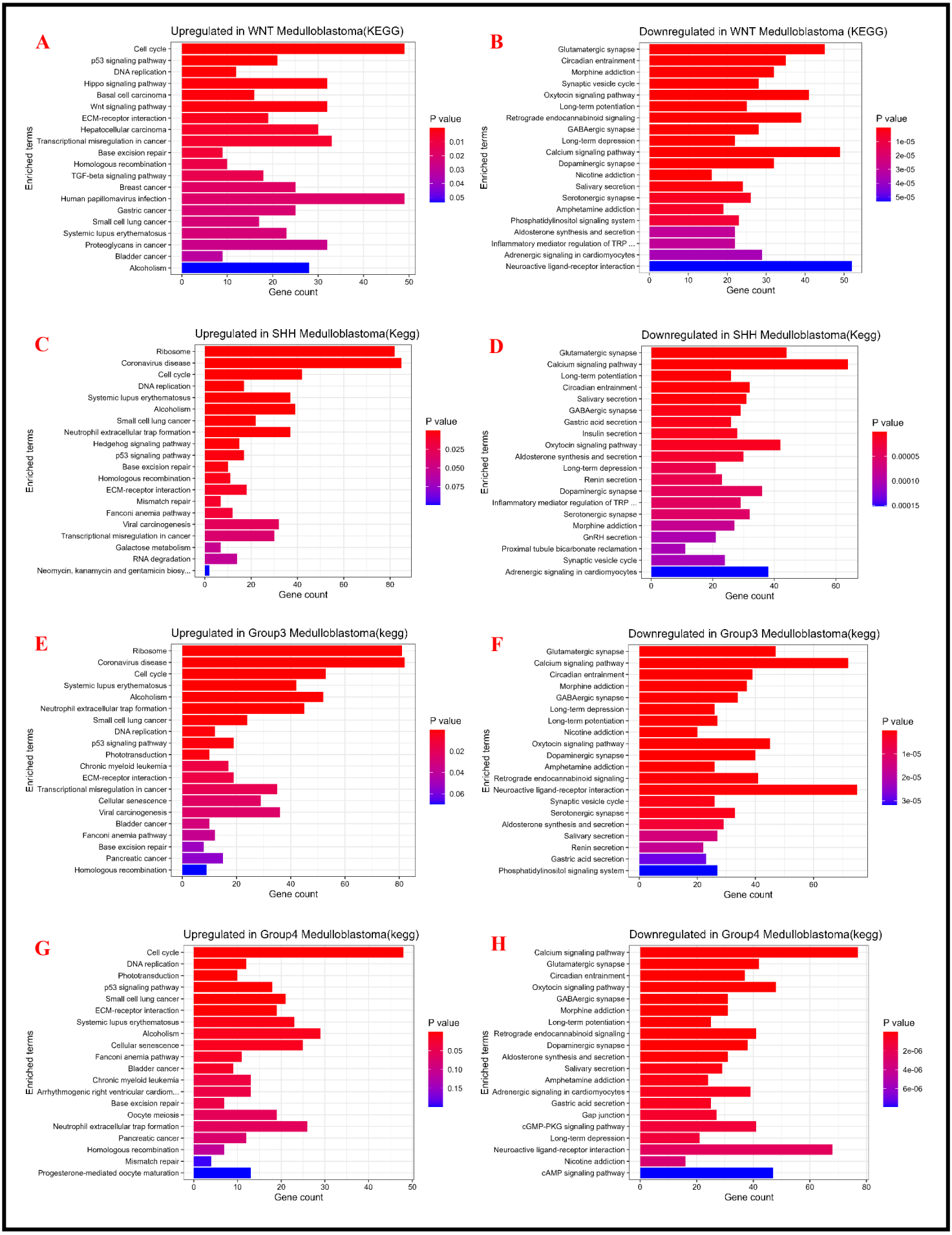


Figure 5: KEGG pathway enrichment analysis showing different pathway enrichment in Medulloblastoma subgroups. A: Upregulated pathways in WNT. B: Downregulated pathways in WNT. C: Upregulated pathways in SHH. D: Downregulated pathways in SHH. E: Upregulated pathways in Group 3 Medulloblastoma. F: Downregulated pathways in Group 3 Medulloblastoma. G: Upregulated pathways in Group 4 Medulloblastoma. H: Downregulated in Group 4 medulloblastoma.

subgroup, the SHH subgroups showed downregulation of Glutamatergic synapse, GABAergic synapse, Dopaminergic synapse, etc. The unique downregulated pathways were calcium signaling pathway, Long-term depression, Serotonergic pathway, GnRH secretion, etc (**Figure 5 (C) (D)**).

Group 3 Medulloblastoma showed the upregulation of pathways similar to SHH. In Group 3, Ribosome, Cell cycle, Neutrophil extracellular trap formation, Transcriptional misregulation, Base excision repair, and homologous recombination were highly upregulated. Even in Group 3, the downregulated enriched processes were common in WNT and SHH, like- Glutamatergic synapse, Calcium signaling, GABAergic synapse, Dopaminergic synapse, serotonergic synapse, Phosphatidylinositol signaling system, etc (**Figure 5 (E) (F)**).

Group 4 Medulloblastoma showed Cell cycle, DNA replication, p53 signaling pathway, ECM receptor interaction, Cellular senescence, Base excision repair, Homologous recombination, Mismatch repair, etc, to be upregulated. Meanwhile, the downregulated pathways were the cGMP-PKG signaling pathway, cAMP signaling Pathway, Gap junction, etc. (**Figure 5 (G) (H)**).

4. Subgroup Specific Upregulate and Downregulate Genes for Medulloblastoma

The subgroup-specific upregulated genes are given in **Table 1**. These genes could be signatures for the subgroup classifications. The subgroups-specific gene classification showed that the long ncRNAs were highly upregulated in Medulloblastoma (**Figure 6 (A)**). All the subgroups were showing similar results. Similarly, the subgroup-specific downregulated genes are given in **Table 2**. These genes are highly downregulated in each subgroup of Medulloblastoma and could be a potential marker to find the downregulated mechanism of this cancer (**Figure 6 (B)**).

Table 1: Subgroup Specific Upregulated Genes

Subgroup of MB	Genes
WNT	LINC02178, HOXD13, HOXD11, ISL1-DT, HOXD10, LINC02898, HOXC-AS2 DDC-AS1, TRDD3, GPR50
SHH	IGHEP1, LOC107987211, MIR8071-1, LOC105378184, IGHE, LOC105378187, IGHJ3, KRT14, LOC107986551, IGHJ2P
Group 3	LOC102724419, LINC01467, GALP, LOC105372932, TYR, CLC, BPIFB4, LOC107984408, LOC107985462, LOC107986151
Group 4	LINC01419, FOXA2, LINC00261, HMX3, HOXC13-AS, HOXB13, LINC03017 LOC105378998, LOC102724749, LINC00879

Table 2: Subgroup-Specific Downregulated Genes

Subgroup of MB	Genes
WNT	LINC02770, LOC105371241, FADS6, LOC107984456, LINC02470, LOC105379085, FREM3, FOXG1-AS1, LOC101929617
SHH	LOC105372645, LOC107986395, CPNE6, TAB2-AS1, SST, TFAP2B, GJB1, SLC39A12, LOC105371735
Group 3	LINC01726, LINC01727, SYCE1, GABRA3, LOC101927461, KLHL1, LOC105373504, TCERG1L, ANKRD18B
Group 4	DRD5, DPP10-AS1, LINC01885, LOC105379383, LOC105377412, LOC105369468, TMEM252, LINC01322, LOC107985937

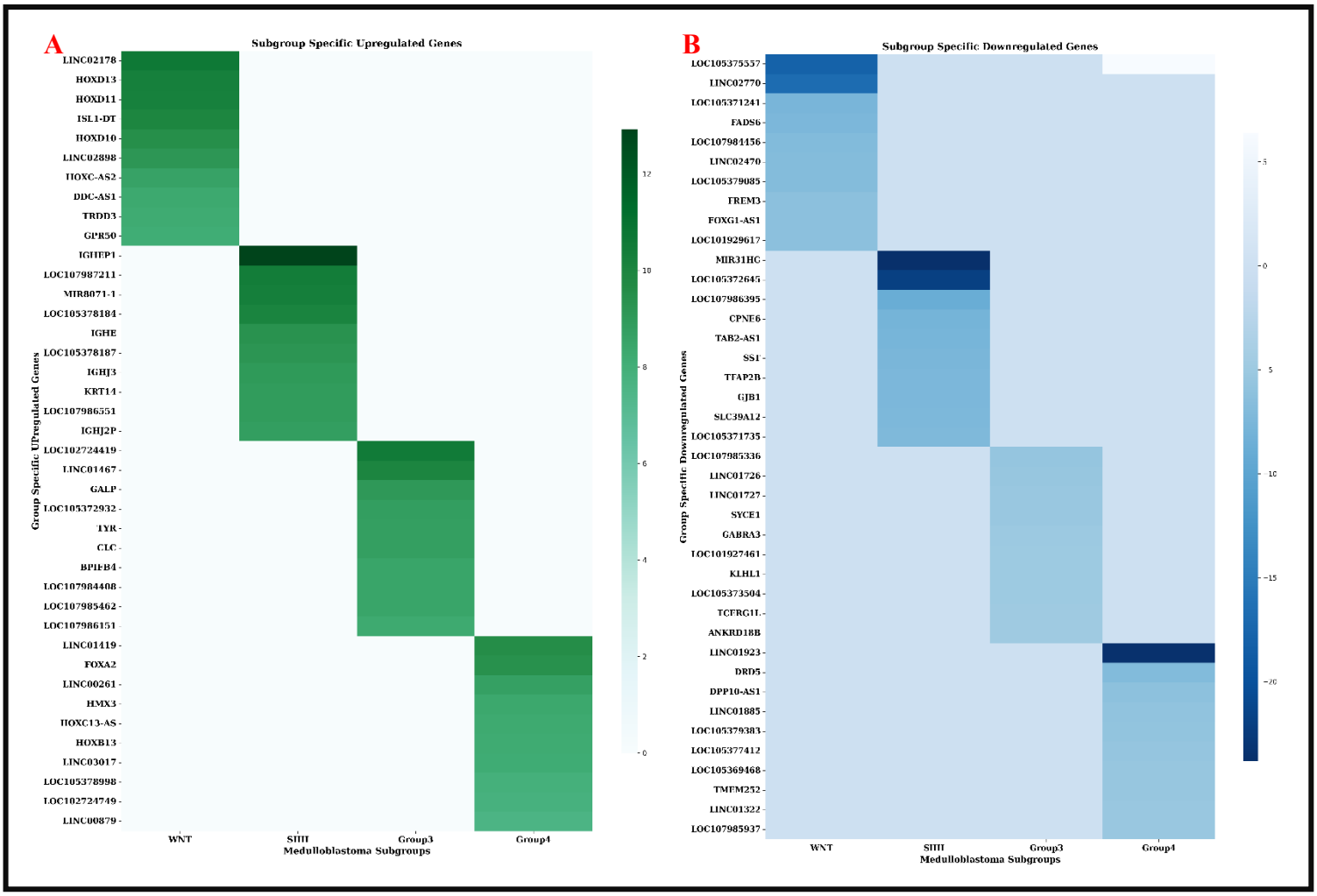


Figure 6: A: The image showing the upregulation of genes which are specific for each subgroup. B: The image is showing downregulated genes that are specific for Medulloblastoma groups

5. Group-specific Upregulated Genes Showed Different Chromosome Origination

The WNT-specific upregulated genes showed high origination from Chromosome 2 and Chromosome 7, followed by Chromosome 5, Chromosome 12, and X Chromosome (**Figure 7 (A)**).

The SHH-specific overexpressed genes originated mainly from Chromosome 14. Chromosome 6, Chromosome 15, and Chromosome 10 were shown to be moderately active (**Figure 7 (B)**).

Chromosome 1 showed the origination of more group-specific upregulated genes in Group 3, followed by Chromosome 3, Chromosome 19, Chromosome 11, Chromosome 9, and Chromosome 2 (**Figure 7 (C)**).

Group 4 has not presented any Chromosome specific overexpression rather a set of chromosomes shown to be the location of origination of group 4 specific genes. The genes originated from Chromosome 8, Chromosome 20, Chromosome 7, Chromosome 5, Chromosome 6, chromosome 4, etc. (**Figure 7 (D)**)

6. Immune Cell Enrichment Analysis for Different Medulloblastoma Subgroups.

The Immune cell enrichment analysis showed downregulation of HSC in Group 3 and Group 4. Macrophages M2 also showed downregulation in all the Medulloblastoma subgroups. The Mesangial cells are slightly upregulated in WNT, SHH, and Group 4. The Neutrophiles were depleted in WNT, SHH, and Group 3 Medulloblastoma. Th1 and Th2 cells were upregulated in all Medulloblastoma subgroups. The Tregs were downregulated in Medulloblastoma. The pDC cells were

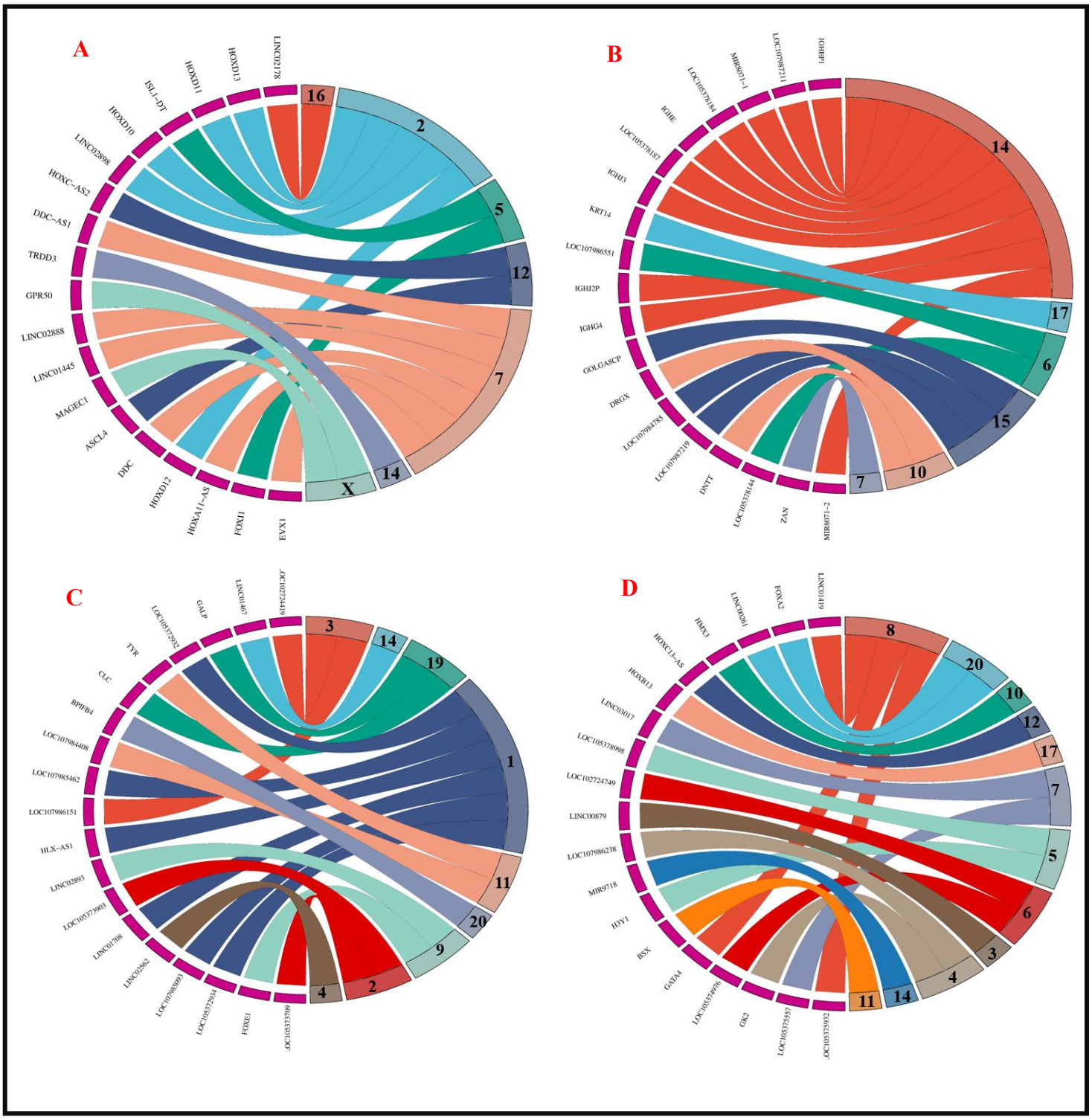


Figure 7: A: Upregulated WNT-specific genes and originated Chromosome, B: Upregulated SHH-specific genes and originated Chromosome, C: Upregulated Group 3-specific genes and originated Chromosome, D: Upregulated Group 4-specific genes and originated Chromosome

upregulated in Group 4 Medulloblastoma. This immune landscape could be a signature of subgroup variation in Medulloblastoma (**Figure 8 and Figure 9**).

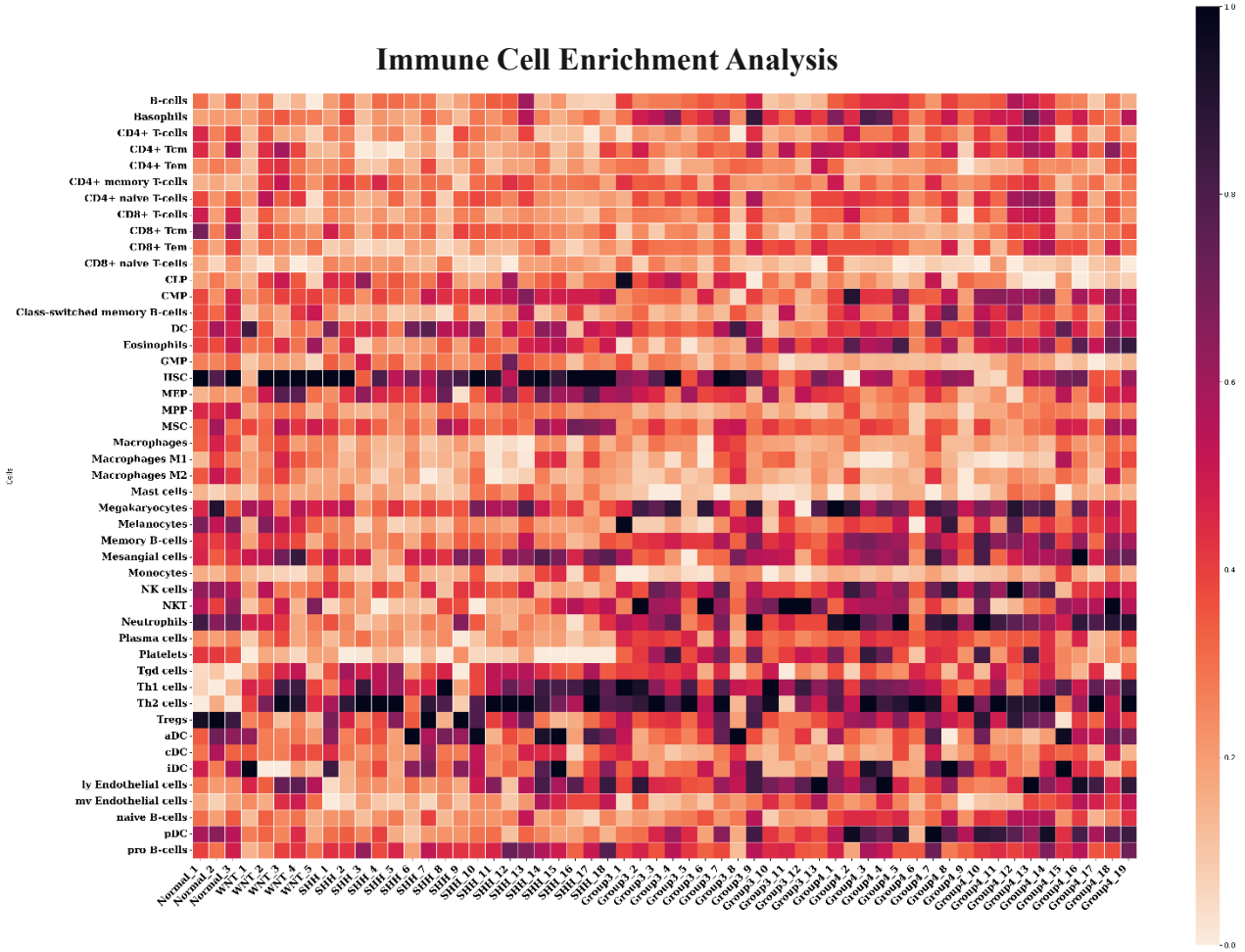


Figure 8: Immune-specific cell abundance in different samples of Medulloblastoma

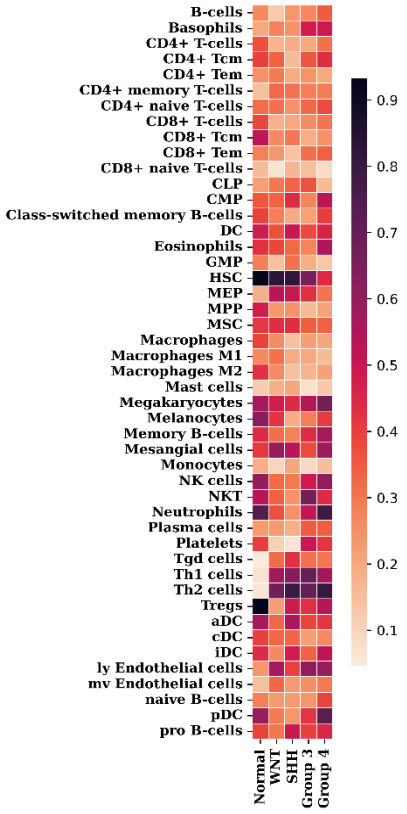


Figure 9: Immune Cell abundance in Different Group of Medulloblastoma

7. The Logistics Regression-Based Model Showed Group Specific Molecular Signatures are Able to Classify the Subgroups from Transcriptomics Profile

The Receiver Operating Characteristic (ROC) curve showed AUC (Area Under the Curve) to be 1. Hence, the classification model was able to distinguish the classes properly (**Figure 10 (A)**).

The Confusion matrix also showed that for the classification of the SHH, Group 3, and Group 4, the prediction was perfect, but for the WNT subgroup, two out of three were predicted as SHH, and only one prediction was right. This could be due to a low number of samples of the WNT subgroup (**Figure 10 (B)**).

The classification report of the logistics regression-based model showed that overall, for

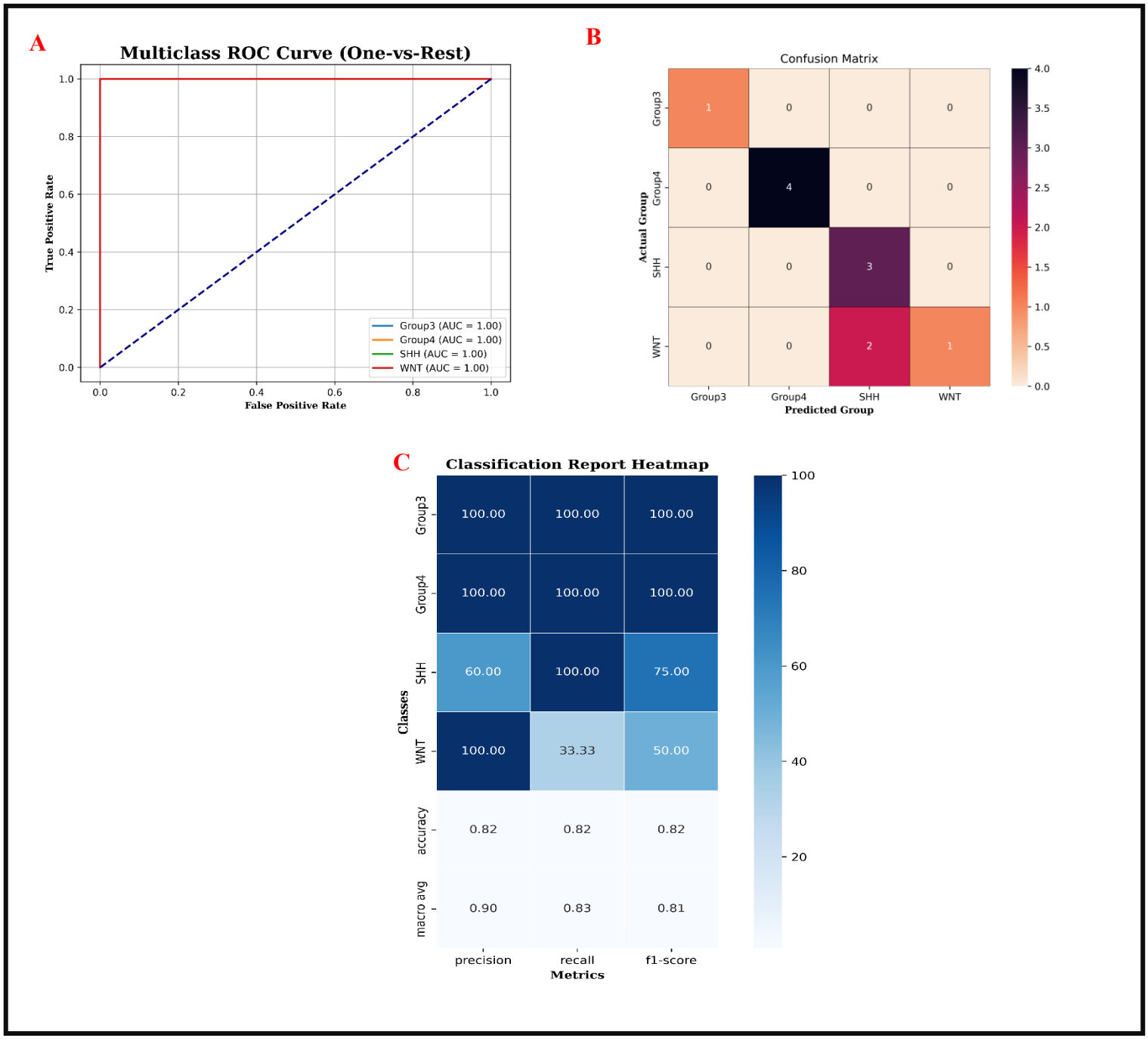


Figure 10: A: ROC curve showing AUC 1, and its a sign of good claasifier. B: Confusion matrix showing the actual vs predicted subgroup of Medulloblastoma. C: Classification report of the logistics regression showing that the top abundance gene are capabale of predicting the subgroups from the group specific genes.

subgroup classification, Group 3 and Group 4 Medulloblastoma were classified at 100 % Precision, Recall, and F1-Score. The Precision of the SHH subgroup classification was 60 % with a 75 % f1-score. The Recall for the WNT subgroup was 33.33 % with 50% f1-score (**Figure 10 (C)**).

Overall, the accuracy of the model was 82 %.

Discussion:

The vital pediatric cancer, Medulloblastoma, showed different levels of gene expression in different subgroups. Because of the subgroup-specific gene expression, different clusters in PCA were shown. Although the subgroups have similarities in gene expression levels, the abundance and expression level of lots of genes are still specific for each of the subgroups.

The pathway enrichment also showed that the WNT subgroup actually upregulated the WNT pathway, whereas SHH overexpressed the hedgehog pathway. Though, Group 3 and Group 4 show very common pathway enrichment, the downregulation of the cGMP-PKG pathway and cAMP signaling is actually significant for Group 4.

The subgroup-specific upregulation and downregulation showed different non-coding RNAs to be specific for each group, and it could have a specific role in cancer progression and tumor invasion.

The origination of genes from the Chromosome showed that the upregulated genes for WNT are located in Chromosome 2 and Chromosome 7. The SHH-specific genes showed mostly from Chromosome 14. The Group 3-specific genes showed mostly from Chromosome 1. However, Group 4 Medulloblastoma did not show any chromosome specificity. These could be specific signatures for the subgroups and their origination.

The immune cell enrichment showed that Th1 and Th2 were very high in medulloblastoma, and the Tregs were low in medulloblastoma. The hematopoietic stem cells showed that Group 4 had a very small number of them. A smaller number of neutrophils could be a sign of molecular mechanism in WNT, SHH, and Group 3 Medulloblastoma. Overall, the immune cell profiling shows a lesser number of variations among the subgroups, but few specific cell types are specific for each subgroup.

The logistic regression-based classifier was able to predict the subgroups from the signature genes. The lower recall for WNT could be due to sample size. Overall, the model can classify subgroups. A large number of datasets could be a good choice to train the model with those signature genes for better precision and recall.

Overall, the selected signature group-specific genes are showing good targets for the classification of Medulloblastoma subgroups.

Availability of materials and data

The data files, R script, Python script, and plots are available in the GitHub repository: https://github.com/Sayan-InSilicoLife/Transcriptomics_Landscape_of_Medulloblastoma.git.

Acknowledgment

The author would like to thank Dr. Saket Chowdhary for his invaluable guidance and support throughout the course of this research work.

References:

1. KP M, Kumar A, Biswas D, Moiyadi A, Shetty P, Gupta T, et al. The proteomic analysis shows enrichment of RNA surveillance pathways in adult SHH and extensive metabolic reprogramming in Group 3 medulloblastomas. Brain Tumor Pathol. 2021 Apr 1;38(2):96–108.

2. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro-Oncology. 2021 Aug 1;23(8):1231–51.

3. Northcott PA, Robinson GW, Kratz CP, Mabbott DJ, Pomeroy SL, Clifford SC, et al. Medulloblastoma. Nat Rev Dis Primers. 2019 Feb 14;5(1):1–20.

4. Northcott PA, Dubuc AM, Pfister S, Taylor MD. Molecular subgroups of medulloblastoma. Expert review of neurotherapeutics. 2012 Jul;12(7):871.

5. Gibson P, Tong Y, Robinson G, Thompson MC, Currle DS, Eden C, et al. Subtypes of medulloblastoma have distinct developmental origins. Nature. 2010 Dec;468(7327):1095–9.

6. Luo Z, Dong X, Yu J, Xia Y, Berry KP, Rao R, et al. Genomic and Transcriptomic Analyses Reveals ZNF124 as a Critical Regulator in Highly Aggressive Medulloblastomas. Frontiers in Cell and Developmental Biology. 2021 Feb 18;9:634056.