Medulloblastoma: Human Specific Genes Contribution

Sara Baldinelli, Letizia De Pietri, Gaia Faggin, Huyen Pham, Roan Spadazzi

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 (MB-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. We will integrate the set of MB-associated HSGs with FANTOM causal relations and perform functional and network analyses on the resultant gene set. Our findings will lead to further investigation on the biological development of the disease, that may offer valuable insights into its evolution and implications for targeted therapies or treatment strategies.

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: MB-WNT, Sonic Hedgehog (SHH), Group 3 (Gp3) and Group 4 (Gp4). 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 primary driven by genomic alterations.
Previous studies showed that the least to most common subgroups are: WNT, Group 3, SHH and Group 4. Respectively, they have survival rates of: $\sim$95%, $\sim$50%, $\sim$60-80% and, $\sim$75%. [4, 5].
In our project we will investigate the contribution of Human Specific Genes (HSGs) with respect to non-HSGs in the context of this disease.
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.

It is worth noting that a substantial portion of HSGs plays a critical role in brain functions, alongside their roles in the immune system and metabolic processes [6].
Consequently, our efforts are directed towards identifying HSGs that may contribute to the development of MB.
To this purpose, the set of MB HSGs will be integrated with FANTOM [7] causal relations. A functional and a network analysis will be carried out on the resulting set of genes and hypotheses about the role of HSGs in the context of MB will be formulated based on the results.

1. Biological question

The question driving this project revolves around the role of Human Specific Genes in comparison to non-HSGs within the context of MB.
In order to address this question, we will assess the following points:

1. Identifying the existence of causal relationships among HSGs specific to MB and elucidating their significance.
2. Evaluating the functional significance of the subset of HSGs associated with MB.
3. Investigating the network structure constructed from the subset of HSGs.

Collectively, these objectives will help us to acquire the knowledge necessary to uncover the significance of HSGs in MB and determine the level at which these genes can serve as markers for subtyping MB.

2. Data

2.1 Cohort selection

2.1.1 GSE155446

Cohort GSE155446 [8] 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 subpopulations associated with mitotic, undifferentiated and neuronal profiles.

2.2 List of Human Specific Genes

A list of 856 genes specific to humans was extracted from the research conducted by Bitar et al. [6]. This list will serve as a reference for pinpointing HSGs among the set of genes retrieved.

2.3 FANTOM dataset

The FANTOM [7] dataset will be used to expand the set of MB HSGs so that genes involved in causal relations will also be considered.

3. Pipeline

The project workflow is described in **Figure 1**. Following it, it is possible to divide the work in six main steps.

1. Data pre-processing;
2. Differential expression analysis;
3. Intersection between MB differentially expressed genes and the list of Human Specific Genes;
4. Data integration between the filtered MB Human Specific Genes and the FANTOM data;
5. Functional Analysis;
6. Network Analysis.

3.1 Data pre-processing

In our project, we will leverage the power of *Scanpy* [9] to conduct data pre-processing for scRNA-seq datasets, a crucial initial step in our analysis pipeline. This pre-processing workflow is pivotal for handling the complexity of scRNA-seq data and ensuring the accuracy of downstream analyses. We will begin by importing the raw count matrix into *Scanpy*'s AnnData object, facilitating data organization and management. Quality control will follow, where we will apply stringent filters to eliminate low-quality cells and genes based on criteria such as minimum count per cell or gene expression levels. To reduce technical bias, we will normalize the data using library size scaling or log transformation. For dimensionality reduction, techniques like PCA or UMAP will be employed to uncover the underlying structure of the data. Clustering algorithms will help identifying distinct cell populations, and marker genes will be detected to label cell types accurately.

3.2 Differential expression analysis

Following the initial data pre-processing steps, the next crucial phase involves conducting differential expression analysis. This process is a fundamental component of scRNA-seq data analysis and it plays a key role in the identification of genes that exhibit significant unpregulation or downregulation within a specific cell population, under various conditions or in response to distinct treatments.
To accomplish this task, we will also use, during this step, the *Scanpy* library [9].

3.3 Intersection between MB differentially expressed genes and the list of Human Specific Genes

After obtaining our differentially expressed genes, we will intersect them with the list of Human Specific Genes.

3.4 Integrate the filtered MB Human Specific Genes with the FANTOM data

After obtaining our list of HSGs related to MB, we will integrate them with the FANTOM [7] dataset in order to retrieve causal relations.

3.5 Functional Analysis

Functional analysis involves leveraging various bioinformatics tools and databases such as Gene Ontology (GO) [10] enrichment analysis and Kyoto Encyclopedia of Genes and Genome (KEGG) [11] pathway analysis.
By identifying the biological processes, molecular pathways, and cellular functions that these genes are associated with, we can gain a comprehensive understanding of the underlying mechanisms at play.

3.6 Network Analysis

With the data obtained, we aim to construct a network of interactions. Once the network is established, we will embark on measuring various metrics, including clustering coefficient, diameter, and centrality measures to uncover its characteristics.
A key focus of our analysis will be the identification of hub genes, as well as the potential delineation of network communities. These analyses will shed light on the most influential components of the network and the existence of functional modules within it.
To accomplish this, we will employ *NetworkX* [12], a Python library specifically built for network analysis. However, it is worth noting that we approach this task without any a priori knowledge of the network's size. Given the potential computational challenges this may pose, we will evaluate the feasibility of each step during the analysis.

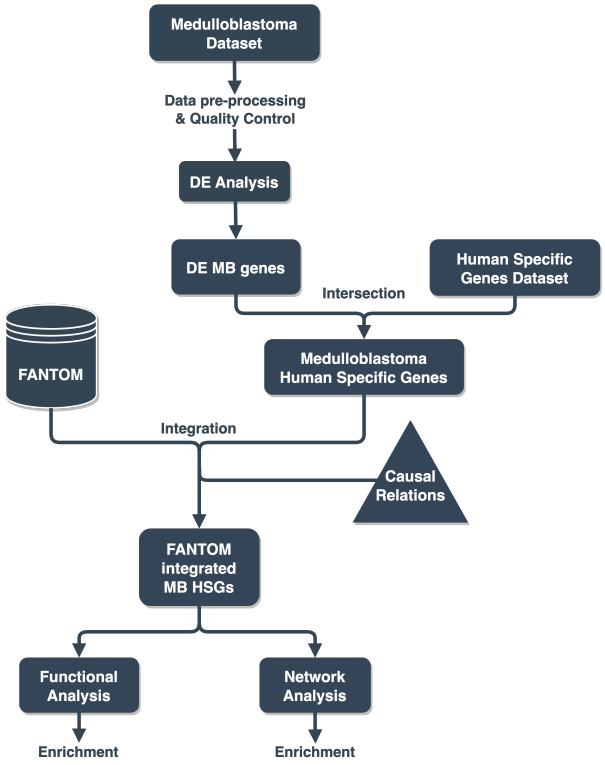

Figure 1. General workflow depicting the six main steps of the project.

4. Preliminary results

Through cBioPortal [13], we retrieved a list of genes that, due to mutations, are related to MB. By intersecting it with the list of HSGs [6], a set of 14 MB-related HSGs was obtained. Using Cytoscape [14], a network on the list of MB genes was built and MB-related HSGs were highlighted. We observed that 12 out of the 14 MB-related HSGs are taking part in many interactions **[S1]**. This suggests us that it is worth analyzing in greater detail the possible contributions of HSGs in MB development.

5. Expected results and contingency plans

By undertaking this analysis we expect to observe a functionally significant role played by Human Specific Genes with respect to non-HSGs. Should this functional significance be confirmed, further studies will be undertaken in order to evaluate the biological functions of the inferred genes. Otherwise, the absence of such functional significance could be as well considered a result. However, this would prompt us to possibly redirect our research focus towards another neurological disorder.

6. Project management

Foremost the project will be developed in a teamwork driven manner. In the initial stages (e.g. data selection, data preprocessing, etc.), the entire team will work together. After integrating the MB HSGs with the FANTOM dataset [7], the team will be ideally divided into two subgroups. The first one, consisting of Gaia Faggin and Huyen Pham, will be responsible for conducting the functional analysis. Meanwhile, the second subgroup, composed of Sara Baldinelli, Letizia De Pietri and Roan Spadazzi, will focus on network analysis. It is worth noting that both subgroups will maintain collaboration, allowing them to benefit from the diverse backgrounds and experiences of each member. Finally, the data interpretation and the conclusion will be discussed together.

7. Resources

It is important to note that the resources listed are not to be considered final. As this is merely a project proposal, during the analysis phase, we may require additional libraries or tools.

Scanpy

Since we are using scRNA-seq datasets, we will employ the *Scanpy* library [9], a powerful Python tool tailored for the analysis of single-cell gene expression data. *Scanpy* is purposefully designed in collaboration with *anndata* [15], encompassing a wide array of functionalities, including preprocessing, data visualization, clustering, trajectory inference, and differential expression analysis.

NetworkX

To analyze the properties of the whole network we will use *NetworkX*, a Python package specifically designed for exploring and analyzing networks [12]. *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.

Gene Ontology

Gene Ontology (GO) [10] is a comprehensive and widely-used bioinformatics resource that provides a standardized vocabulary and framework for describing the functions of genes and gene products. It categorizes genes into terms related to their molecular functions, biological processes, and cellular component, creating a structured, hierarchical system of annotations.

Kyoto Encyclopedia of Genes and Genome

The Kyoto Encyclopedia of Genes and Genome (KEGG) [11] is a broad and integrated bioinformatics database that serves as a valuable resource for understanding molecular-level information about biological pathways and functions. KEGG encompasses a wide range of data, including information on

metabolic pathways, genetic sequences, diseases, and various other molecular processes.

Cytoscape

Cytoscape [14] is a powerful and widely-used bioinformatics software platform designed for the visualization and analysis of biological networks. It allows to create, manipulate, and explore complex network representations of biological data, such as protein-protein interactions, gene regulatory networks, and metabolic pathways.

ggplot2

ggplot2 [16] is a powerful plotting package designed for creating intricate visualizations from data stored in data frames. With its user-friendly commands, *ggplot2* offers a programmable approach to specifying which variables to visualize, how they should be presented, and overall visual characteristics. It simplifies the process of generating complex plots and enhances the overall data visualization experience.

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] Functional annotation of the mammalian genome 5 (fantom5). `http://fantom.gsc.riken.jp/5/`, 2014. Accessed: October 16, 2023.

[8] Kent A Riemondy, 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.

[9] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[10] Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.

[11] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2017.

[12] 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.

[13] cbioportal for cancer genomics. `https://www.cbioportal.org/study/summary?id=mbl_pcgp`, 2014. Accessed: October 20, 2023.

[14] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[15] Isaac Virshup, Sergei Rybakov, Fabian J Theis, Philipp Angerer, and F Alexander Wolf. anndata: Annotated data. *BioRxiv*, pages 2021–12, 2021.

[16] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

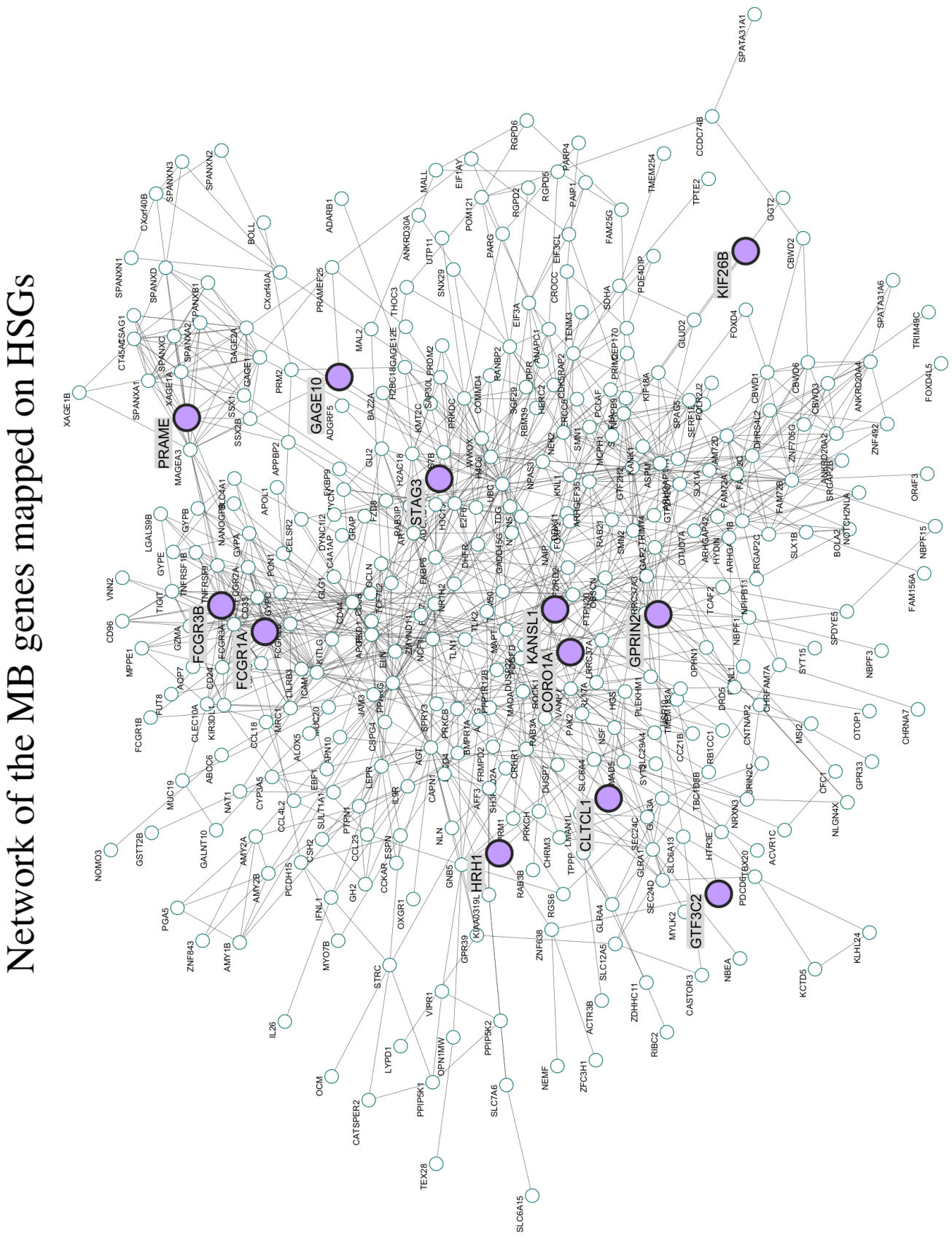
Network of the MB genes mapped on HSGs
PRAME
GAGE10
STAG3
FCGR3B
FCGR1A
KANSL1
CORO1A
GPRIN2
CLTCL1
HRH1
GTF3C2
KIF26B

Figure S1