Accurate identification of medulloblastoma subtypes from diverse data sources with severe batch effects by RaMBat

Mengtao Sun[1], Jieqiong Wang[2] and Shibiao Wan[1]*

[1]Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, USA

[2]Department of Neurological Sciences, University of Nebraska Medical Center, Omaha, NE, USA.

*Correspondence: Shibiao Wan, swan@unmc.edu

Abstract

As the most common malignant pediatric brain cancer, medulloblastoma (MB) accounts for around 20% of all pediatric central nervous system (CNS) neoplasms. MB includes a complex array of distinct molecular subtypes, mainly including SHH, WNT, Group 3 and Group 4. Accurate identification of MB subtypes enables improved downstream risk stratification and tailored therapeutic treatment design. Existing methods demonstrated the feasibility of leveraging transcriptomics data for identifying MB subtypes. However, their performance may be poor due to limited cohorts and severe batch effects when integrating various MB data sources. To overcome these limitations, we propose a novel and accurate approach called RaMBat for accurate MB subtype identification from diverse transcriptomics data sources with severe batch effects. Specifically, RaMBat leverages intra-sample gene expression ranking information instead of absolute gene expression levels, which can efficiently tackle batch effects across diverse transcriptomics data cohorts. By intra-sample gene rank analysis, reversal ratio analysis, and feature selection, RaMBat can select MB subtype-specific gene features and finally accurately identify MB subtypes. Benchmarking tests based on 13 datasets with severe batch effects suggested that RaMBat achieved a median accuracy of 99%, significantly outperforming other state-of-the-art MB subtyping approaches and conventional machine learning classifiers. In

addition, in terms of visualization, RaMBat could efficiently remove the batch effects and clearly separate samples from diverse data sources according to MB subtypes, whereas conventional visualization methods like tSNE suffered from severe batch effects. We believe that RaMBat is a promising MB subtyping tool that would have direct positive impacts on downstream MB risk stratification and tailored treatment design. To facilitate the use of RaMBat, we have developed an R package which is freely available at https://github.com/wan-mlab/RaMBat.

Keywords: Medulloblastoma; Transcriptome; intra-sample gene ranking; Cancer subtype identification; Batch effect.

Introduction

As the most common malignant pediatric cancer arising in the cerebellar vermis, medulloblastoma (MB) accounts for around 20% of all pediatric central nervous system (CNS) neoplasms and over 70% of patients diagnosed are children under the age of 10 [1,2]. Based on the genetic heterogeneity and phenotypic diversity [3], four major MB subgroups with distinctive molecular landscapes, somatic mutations and clinical outcomes are defined by the World Health Organization (WHO) [4], including Sonic HedgeHog (SHH) activated, Wingless-Type (WNT) activated, and the numerically designated Group 3 and Group 4 [5]. Some studies [6] further categorized SHH into two subgroups, namely SHH TP53 mutant and SHH TP53 wide-type. WNT and SHH are named after signaling pathway disturbances in the WNT (*CTNNB1* mutation [7,8]) and SHH (*PTCH*, *SUFU*, *SMO* mutation, or *GLI2* amplification [9]) respectively, which are clearly separable with other two subtypes. Currently, the long-term survival rate of MB has reached 70% based on the standard treatment, however, there are wide disparities among patient outcomes due to the significant difference in histologic characteristics of the tumor, age, presentation of metastases, residual disease and other factors [10,11]. WNT can be almost completely cured under current therapy schemes (5-year event-free survival greater than 90% [12]), whereas SHH-activated MB occurs more frequently in infants and largely dependent on the specific genetic features where *TP53* mutation caused tumor has poorer prognosis [9]. Group 3 and Group 4 do not exhibit subgroup-defining mutations [13], but still have distinct clinic-biological features where Group 3 patients have the most unfavorable prognosis due to a higher incidence of "high-risk" features, such as LCA (large cell/anaplastic) histology and *MYC* amplification [5,14–16], and Group 4 has an intermediate prognosis with the feature of isochromosome 17q (i17q) [17]. In other words, the survival rates vary significantly with different MB subtypes. Conventional radiation and chemotherapy have been demonstrated to have long-term sequelae. For example, craniospinal irradiation can cause endocrinopathies as well as neurocognitive and neurosensory impairment [18,19]. Consequently, this has driven the research communities to seek more targeted treatment approaches, with a particular focus on the molecular characteristics of these tumors. The current consensus agreed that different subtypes exhibit significant variations in treatment response and prognosis [5]. For instance, the WNT subtype of MB typically has a favorable prognosis, allowing for a more conservative treatment approach, whereas the SHH subtype may require more aggressive interventions. In contrast, Group 3 and Group 4 subtypes tend to be more aggressive and have higher recurrence

rates, often necessitating more intensive treatment strategies [3]. Given the heterogeneity of these molecular subtypes, tailoring individualized treatment plans not only enhances therapeutic efficacy but also minimizes the long-term side effects associated with conventional therapies [20]. Therefore, accurately identifying MB subtypes is crucial for guiding effective treatment strategies.

Conventional methods like morphological analysis, cytogenetic analysis, immunophenotyping, or molecular profiling for identifying MB subtypes are costly, time-consuming, labor-intensive or inaccurate. To overcome these problems, recent studies demonstrated the feasibility of using next generation sequencing (NGS) data including transcriptomics (e.g., RNA-seq [21]) and proteomics [22] for categorizing MB and developed in-silico predictive methods to facilitate a more efficient and convenient prediction process. Previous studies indicate that identifying MB subtypes based on transcriptomic data is an efficient and feasible way [23] as it is an essential tool for expression profiling and provides a way for comprehensive understanding of biological problems [24]. Most recently, medulloPackage [6] was constructed for rapid and accurate MB subtype identification based on the RNA-Seq and microarray datasets. medulloPackage identified discriminating features through gene ratio analysis and an unweighted mean of normalized scores were employed for sample classification. However, this approach would introduce bias as the data did not necessarily conform to normal distribution. In addition, another method named Medullo-Model ToSubtype (MM2S) [25] was developed to enable classification of individual gene expression profiles from MB samples, including patient samples, mouse models, and cell lines, against well-established molecular subtypes. MM2S achieves high accuracy for well-characterized subtypes like WNT and SHH. However, its performance is limited for Group 3 and Group 4, which remain poorly characterized and more heterogeneous, leading to challenges in accurate classification. In summary, existing MB subtyping methods directly use absolute gene expression levels without accounting for batch effects across different cohorts, which may lead to inaccurate classifications. Although a multitude of transcriptomics datasets with varying compositions and sample sizes have been made publicly available, effectively utilizing diverse data sources remains challenging due to the severe batch effects that are derived from systematic technical noises rather than underlying biological signals among MB samples. Various methods for batch effect adjustment, such as SVA [26], svaseq [27], DWD [28], XPN [29], and RUV [30], have been proposed, while normalizing across different batches can still

distort true biological signals, particularly when sample distribution between batches is uneven [31–34].

Instead of relying on absolute gene expression levels, strategies like pair-wise analysis of gene expression (PAGE) [35] leverages the gene expression ratio (GER) [36] which exhibits consistent patterns within samples to construct the model [37] for batch effect correction [38,39]. The GER strategy has been widely utilized in various biomedical domains [37,40,41], as its stability allows GER to be applied across different technological platforms, enhancing its utility in comparative omics studies. For instance, Zhao et al. [39] developed DRGpair which based on intra-sample paired gene ratio for ovarian cancer detection to overcome the limitations of absolute gene expression values including batch effects and biological heterogeneity. Rathi et al. [6] developed medulloPackage, a gene ratio based classifier for MB subtyping, addressing the limitations of absolute gene expression values, including batch effects and biological heterogeneity. Moreover, recent studies [40–42] suggested that ranking gene expression, unlike absolute values which are sensitive to batch effects, provided a more stable representation of gene expression patterns. Additionally, ranking is scale-independent and it proceeds multi-cohort integration regardless of the difference among profile measurements.

To address afore-mentioned concerns, herein we propose an accurate and novel computational approach called RaMBat for accurate MB subtyping across diverse transcriptomics datasets with severe batch effects. Instead of relying on absolute gene expression levels, RaMBat utilizes intra-sample gene expression ranking information as well as a series of steps for gene-rank based information processing, enabling effective batch effect correction and robust integration of multiple bulk transcriptomics datasets. Based on 13 heterogeneous transcriptomics data, we demonstrate the accuracy and efficiency of RaMBat by comparing it with state-of-the-art methods like medulloPackage [6] and MM2S [25] as well as other machine learning (ML) methods equipped with the conventional batch correction method ComBat [43]. Additionally, we have developed an R package RaMBat that is available on GitHub at https://github.com/wan-mlab/RaMBat.

Results

Study design for MB identification

In order to get subtype specific biosignatures for MB samples across different cohorts with heavy batch effects, we propose RaMBat to identify MB subtype accurately from diverse

transcriptomics data sources with severe batch effects based on the gene expression ranking. The overall workflow of this study was illustrated in **Fig. 1**. Three key steps were designed to efficiently capture both global and local dependencies within high-dimensional intra-sample gene expression data from different cohorts, including intra-sample gene expression analysis, reversed expression pattern analysis and informative GERs selection. As input, typical normalized gene expression measures like TPM or FPKM were firstly converted into gene expression ranking followed with intra-sample gene expression analysis which was a two-step process including rank difference calculation and two-sided t-test (See Methods). Gene rank analysis was performed between each pair of subgroups to identify genes that are specifically and significantly associated with each subtype. This approach allowed us to retain only those differentially ranked genes, significantly reducing the computational resources required for subsequent analyses. Next, the reversal ratio analysis was performed as a two-step process including rank level difference calculation and Fisher's exact test. Rank level difference leveraged rank levels between every possible pair of genes to retrieve subtype-specific GERs, combining each two genes together to construct the GER matrix. GER contained in each subtype was compared exhaustively to find the reversed gene expression pattern between each two subtypes cross samples. For example, to select these specific GERs within SHH subtype, a one-vs-one comparison strategy was employed, where the gene expression patterns of SHH were analyzed against each of the three other subtypes using reversal ratio analysis and Fisher's exact test. The final SHH-specific GERs were determined by identifying the intersection of GERs filtered from all three pairwise comparisons. To enhance the accuracy for MB subtyping, LASSO [44] was applied to identify more informative GERs that are strongly associated with subgroups. Under the penalty mechanism of LASSO, GERs assigned with non-zero coefficients were selected. Similarly, as described above, comparisons between each pair of subgroups were performed and the identified GERs form the foundation of RaMBat.

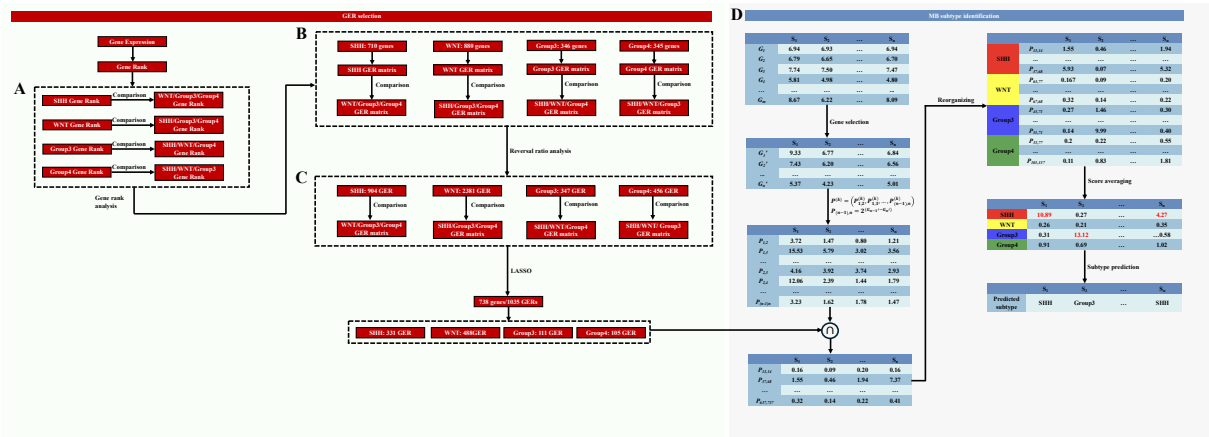

Fig. 1 The workflow of RaMBat based on gene expression ratios for MB subtyping. RaMBat consists of four steps for GERs selection and MB sample subtyping. The first step is to calculate the rank difference for each gene among samples (**A**). The second step is to calculate reversal ratio value for GERs filtering (**B**). The third step is to select GERs using LASSO (**C**). The fourth step is to identify MB subtypes based on GERs selected from previous three steps (**D**).

In this study, RaMBat was developed and trained on a large-scale microarray dataset (GSE85217), comprising 763 MB samples across four molecular subtypes. To ensure robust performance evaluation, RaMBat was tested on 13 independent datasets obtained from the GEO database, covering a total of 508 MB samples (**Supplementary Table 1**). These datasets were generated using different platforms and exhibited significant scale disparities, necessitating standardized normalization to enhance comparability.

As shown in **Fig. 1A-C**, 1035 GERs, corresponding to 738 unique genes, were identified. WNT had the highest number of GERs, totaling 488, followed by 331, 111, 105 GERs for SHH, Group 3 and Group 4, respectively. Subtype information was determined by calculating the average GER score for each subtype within each sample, with the subtype corresponding to the highest average GER score being assigned as the subtype for the sample (**Fig. 1D**).

Benchmarking with the state-of-the-art methods for identifying MB subtypes

Transcriptomics data provides comprehensive biological features for MB subtyping. Leveraging this, medulloPackage and MM2S were developed, both of which classify MB subtypes based on gene expression data. Additionally, we trained and optimized seven ML classifiers, including SVM, LR, RF, XGBoost, KNN, NB and MLP, as benchmarks for comparison. To assess the performance of RaMBat, we evaluated its identification accuracy against medulloPackage, MM2S and the seven ML classifiers across 13 independent test datasets. As shown in **Fig. 2**, RaMBat outperformed medulloPackage and MM2S in eight and eleven datasets respectively. For the remaining five and two datasets, respectively, RaMBat achieved the same accuracy as these methods. The overall accuracy for RaMBat across all datasets was 99.02%, where three Group 3 samples in GSE21140 and one Group 3 sample in GSE10327 were misclassified as Group 4, and one Group 4 sample in GSE74195 was misclassified as Group 3. As previously noted, the high similarity in transcriptomic and genomic profiles between Group 3 and Group 4 results in the molecular ambiguity and histological misclassification. However, the overall accuracy was still 11.8% higher than MM2S and 7.9% higher than medulloPackage. Additionally, the overall subtyping performance of RaMBat was better than all seven ML classifiers (**Fig. 2A**). The lowest accuracy achieved by RaMBat was 95% in GSE74195, and the highest accuracy of 100% was recorded in ten datasets, including GSE12992, GSE30074, GSE37382, GSE41842, GSE49243, GSE50161, GSE50765, GSE62803, GSE67850, GSE73038. When evaluating performance by subtype, RaMBat always outperformed medulloPackage and MM2S on all four subtypes across all independent datasets in terms of accuracy, precision, sensitivity, F1-score, MCC, G-measure, AUC, specificity, and Jaccard index (**Fig. 3**).

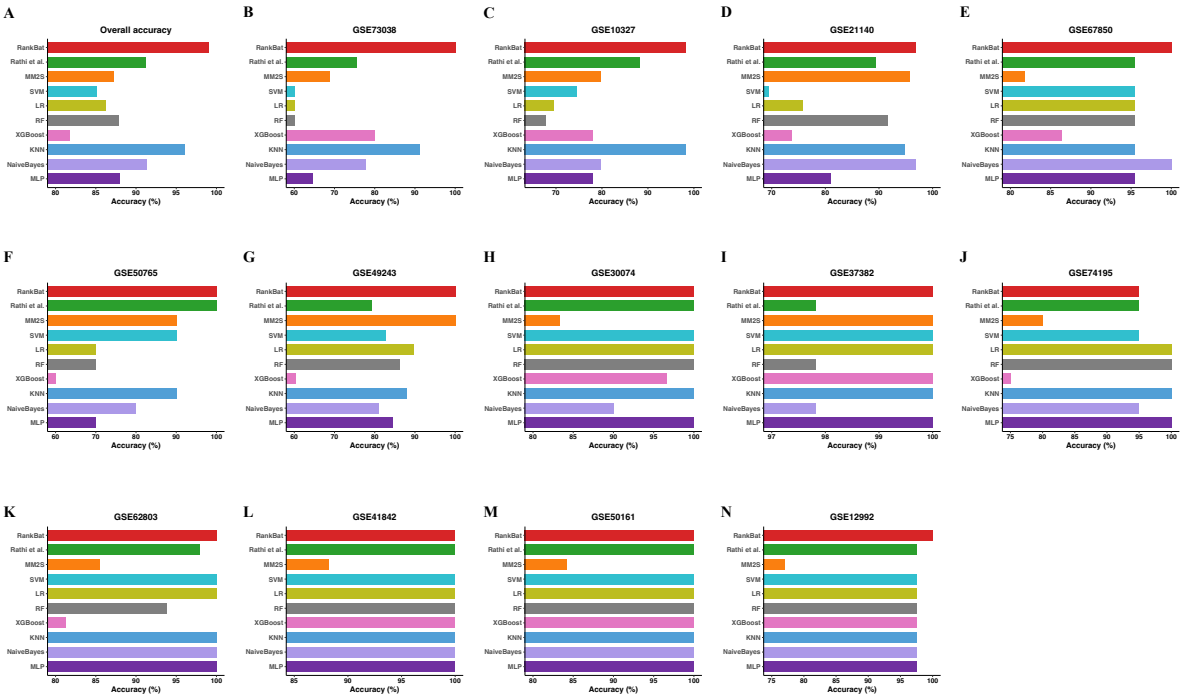

Fig. 2 Comparing RaMBat with the state-of-the-art method on independent test datasets. Overall accuracy of RaMBat, medulloPackage, MM2S and other 7 ML classifiers (**A**) were compared across all 13 independent datasets. Comparative performance of RaMBat, medulloPackage, MM2S and other 7 ML classifiers based on GSE73038, GSE10327, GSE21140, GSE67850, GSE50765, GSE49243, GSE30074, GSE37382, GSE74195, GSE62803, GSE41842, GSE50161, GSE12992 in terms of accuracy (**B-N**), respectively. MLP, multilayer perceptron; KNN, k-nearest neighbors; XGBoost, eXtreme Gradient Boosting; RF, random forest; LR, logistic regression; SVM, support vector machine; MM2S, medullo-model to subtypes.

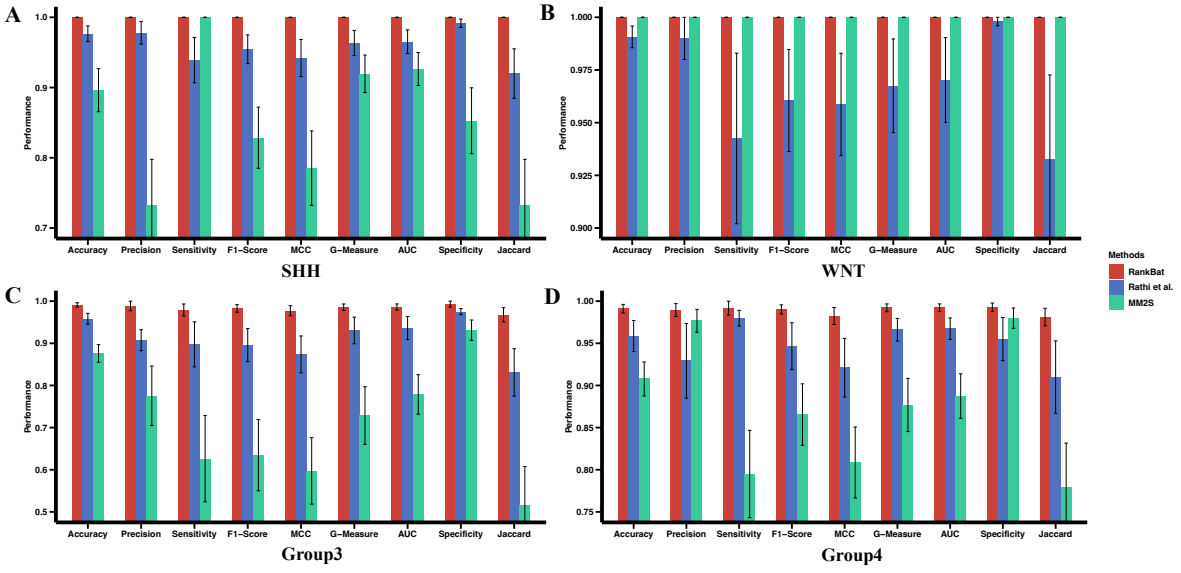

Fig. 3 Comparing RaMBat with state-of-the-art methods on independent test datasets. Comparing performance of RaMBat with medulloPackage and MM2S across datasets in terms of accuracy, precision, sensitivity, F1-Score, MCC, G-measure, AUC, specificity and Jaccard index for SHH (**A**), WNT (**B**), Group 3 (**C**) and Group 4 (**D**) subtypes. Metrics were obtained based on the prediction result of 508 samples in all 13 independent test datasets. MCC, Matthews correlation coefficient; AUC, area under the curve; Jaccard, Jaccard index.

To further evaluate the impact of batch effects and the ability of RaMBat to integrate multi-source data, we combined all 13 independent test datasets and visualized the sample distribution using t-SNE (**Fig. 4**). When applied to raw gene expression data, conventional t-SNE visualization exhibited severe batch effects, with samples clustering predominantly by dataset rather than by biological subtypes (**Fig. 4A**). However, **Fig. 4B** further illustrates that these technical variations obscured the biologically meaningful clustering of MB subtypes as the data did not cluster together according to major MB subtypes. In contrast, **Fig. 4C-D** demonstrate RaMBat effectively removed batch effects, enabling a clear separation of MB subtypes across diverse datasets. **Fig. 4C** shows that after integrating predicted subtype information from RaMBat into visualization, samples were well-integrated across different datasets, indicating that the batch effects were successfully removed. More importantly, **Fig. 4D** presents a clear visualization global landscape of all MB samples with all major subtypes distinctly separable. These results indicate that RaMBat can

accurately integrate diverse bulk transcriptomics datasets, effectively removing batch effects while maintaining the underlying biological signals crucial for MB subtype classification.

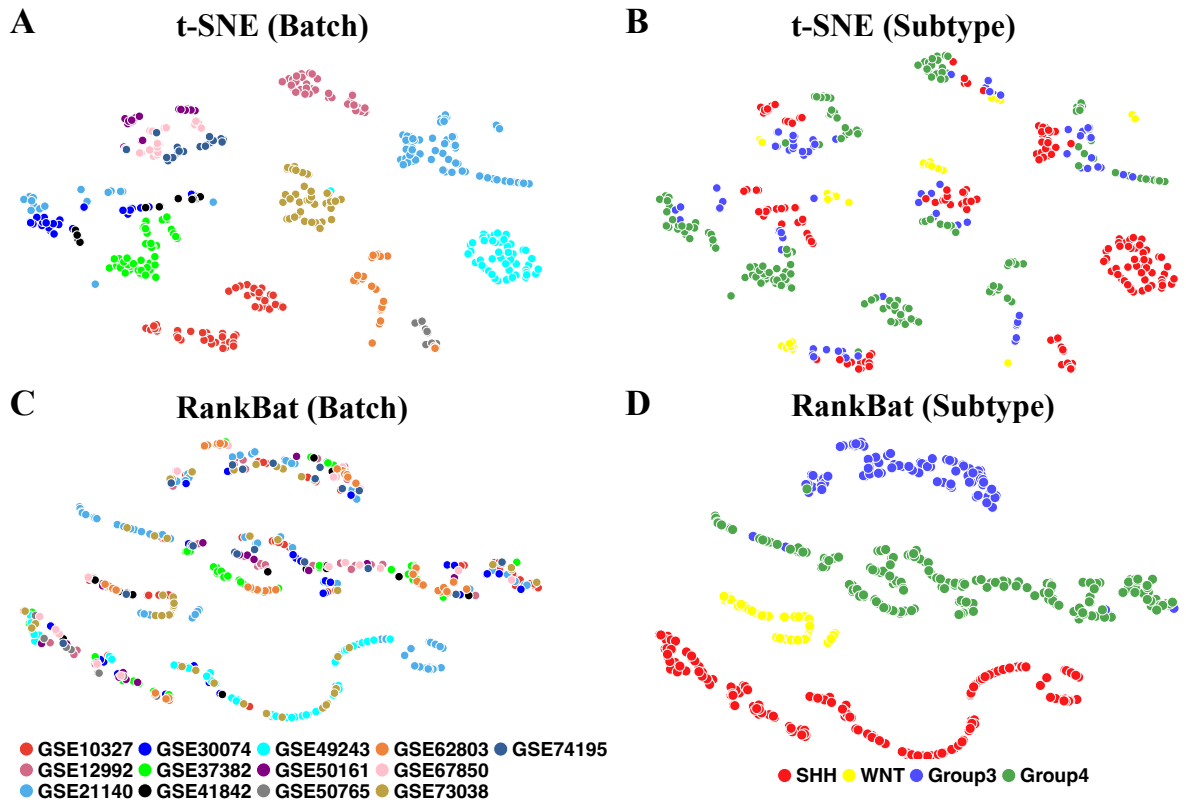

Fig. 4 Comparative visualization of subtype clustering by RaMBat using selected differentially ranked genes. (**A**) The t-SNE plot in which samples were labeled by batch based on all samples in independent test datasets without normalization. (**B**) The t-SNE Plot in which samples were labeled by subtype based on all samples in independent test datasets without normalization. (**C**) RaMBat visualization in which samples were labeled by batch based on all samples in normalized independent test datasets. (**D**) RaMBat visualization in which samples were labeld by subtype based on all samples in normalized independent test datasets.

Subtype-specific GERs selection

To identify differentially ranked genes on a broad scale, we used the training dataset, which included 763 samples, as the basis for GER selection. The absolute expression levels were converted into intra sample rankings, followed by gene rank analysis and two-sided t-test. We here selected 11% of the total number of all genes (2314) and set the p value < 0.05 as the threshold to filter significant differentially ranked genes between two subtypes. There were 1055 ranking down genes and 1257 ranking up genes between SHH and Group 3, 1123 ranking down genes and 1233 ranking up genes between SHH and Group 4, 1220 ranking down genes and 1201 ranking up genes between SHH and WNT, 1221 ranking down genes and 1325 ranking up genes between WNT and Group 3, 1476 ranking down genes and 1522 ranking up genes between WNT and Group 4, 632 ranking down genes and 561 ranking up genes between Group 3 and Group 4 (**Supplementary Fig. S1A-F**). After intersecting these ranked up and down genes, 710, 880, 346, and 345 differentially ranked genes were identified for the SHH, WNT, Group 3, and Group 4, respectively.

Supplementary Fig. S1 Differential rank gene analysis within RaMBat. Volcano plots (**A-F**) illustrate differentially ranked gene for specific medulloblastoma subtypes. The x-axis represents the gene rank difference where y-axis shows -log10(p-value). Red dots indicate significantly up-ranked genes and blue dots represent significantly down-ranked genes. Triangles represent top 5 up/down-ranked genes. Differentially ranked genes between SHH and WNT (**A**), SHH and Group 3 (**B**), SHH and Group 4 (**C**), WNT and Group 3 (**D**), WNT and Group 4 (**E**), Group 3 and Group 4 (**F**).

Based on the differentially ranked genes in each subtype, it results in 251,695, 386,760, 59,685 and 59,340 GERs for SHH, WNT, Group 3 and Group 4, respectively. For each subtype, we performed reversal ratio analysis three times. In the case of SHH, specific GERs were selected by the comparison between SHH and WNT, Group 3, Group 4 respectively. For each comparison, a contingency table was constructed for the Fisher's exact test, i.e., the number of samples where gene a > gene b and gene a < gene b within SHH, as well as in the respective comparison subtype. As shown in **Fig. 1C**, a total of 904, 2,381, 347 and 456 GERs were screened for SHH, WNT, Group 3 and Group 4 respectively based on the threshold of p value (0.05) after reversal ration analysis. Based on the reversed GERs, LASSO feature selection was performed to select the most

important features which results in a final list of 1,035 GERs, corresponding to 738 unique genes (**Fig. 1C**). To comprehensively analyze the subtype-specific gene expression landscape in MB, the gene frequency across all 1,035 GERs for each subtype was also analyzed, as shown in **Fig. 5A-D**. Specifically, the top 10 most frequent genes in each subtype were labeled: SHH included RPGRIP1, CACNA1F, NEUROG1, PDLIM3, ATOH1, SFRP1, FBXL21P, PPP2R2C, NDST3, ZDHHC22; WNT included TMEM51, PAX3, RSPO1, PDE11A, TMEM51-AS1, PDLIM4, ADGRB3, ARRB1, FZD10, TMEM132C; Group 3 included NPR3, PPP2R2B, PAX6, NXPH4, RALGPS2, CACNA1A, PCARE, TSC22D3, ZDHHC15, GABRA5; and Group 4 included RBM24, DIAPH2, CARMIL3, GPR12, GRM8, ROBO1, MFAP4, CAMKK1, DISP3, FIGNL2. To further investigate the gene expression patterns underlying MB subtype differentiation, differential gene expression (DEG) analysis was performed. **Fig. 5E-H** illustrates the DEGs between specific MB subtype and the rest subtypes. To assess the overlap between RaMBat selected genes and DEGs, the identified DEGs with 738 unique genes were intersected (**Fig. 5I-L**). Notably, all of the top 10 most frequently ranked genes (**Fig. 5A-D**) were present within the intersected genes, with the exception of *ZDHHC15* in Group 3, demonstrating strong consistency between RaMBat's feature selection and conventional DEG analysis. This overlap suggests that RaMBat effectively captures subtype-specific molecular signatures, boosting its reliability in MB classification. To further validate the robustness of these findings, the expression patterns of the intersected genes were tested using heatmaps. The clustered heatmap (**Fig. 6A)** of these 738 differentially ranked genes evidenced their distinctly different expression patterns, effectively distinguishing each MB subtype from others and enabling clear separation into specific subgroups. The hierarchical clustering results further confirm the biological relevance of the selected genes in MB subtyping. Specifically, **Fig. 6B-E** displays the expression profiles of the top 10 most frequently ranked intersected genes across MB subtypes, based on data from GSE85217. The results indicate distinct expression patterns that effectively distinguish MB subtypes. Additionally, the expression plots of the upregulated DEG PDLIM3, TMEM132C, GABRA5, RBM24 across all 13 independent samples are shown in **Fig. 6F-I**, highlighting their specific overexpression in SHH, WNT, Group 3 and Group 4 respectively.

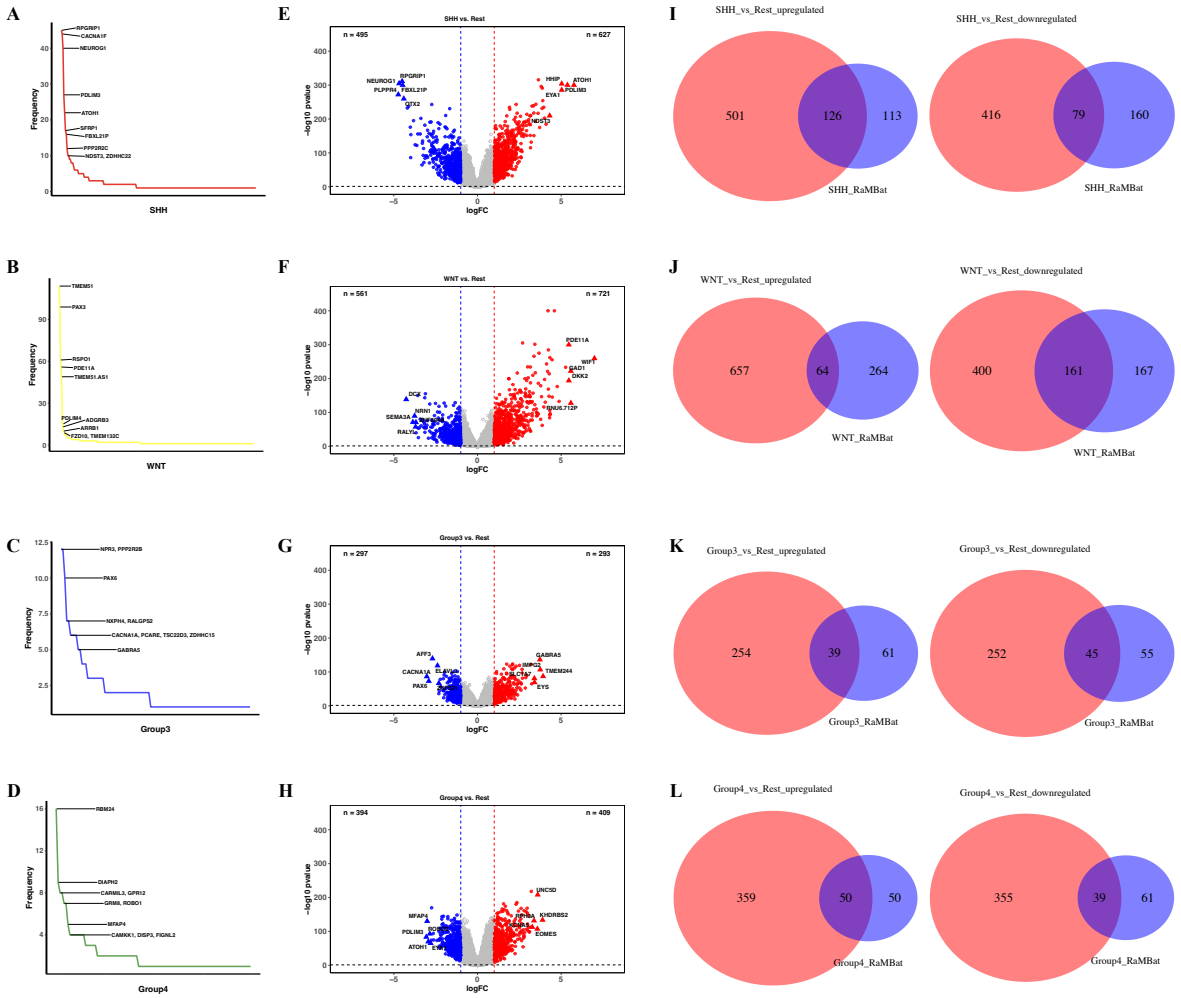

Fig. 5 Differential expression analysis within MB subtypes. (A-D) Line plots display the **frequency** of all unique genes included in GERs for SHH, WNT, Group 3, and Group 4, respectively. The top 10 most frequently ranked genes in SHH (**A**), WNT (**B**), Group 3 (**C**), and Group 4 (**D**) were labeled. (**E-H**) Volcano plots illustrate differential gene expression analysis for each MB subtype. The x-axis represents the log2 fold change while y-axis shows -log10(p-value). Red and blue dots indicate significantly upregulated and downregulated genes respectively. Triangles highlight top 5 upregulated and downregulated genes in each comparison. (**E**) SHH vs rest, (**F**) WNT vs rest, (**G**) Group3 vs rest, (**H**) Group 4 vs rest. (**I-L**) Venn diagrams display common genes between DEGs and unique genes selected by RaMBat. The intersections represent genes identified by both DEG analysis and differential gene ranking analysis for SHH (**I**), WNT (**J**), Group3 (**K**), Group 4 (**L**).

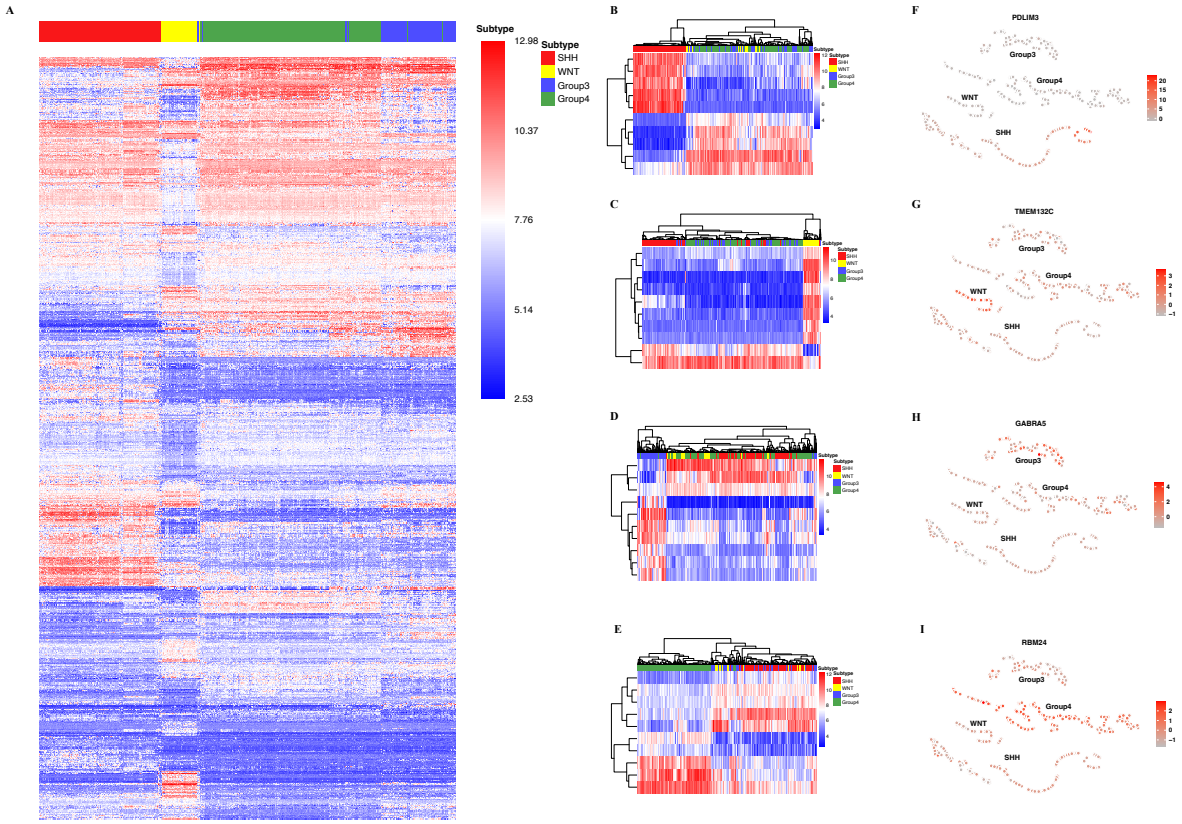

Fig. 6 Identified differentially ranked genes showed distinct subtype-specific gene expression patterns. The large heatmap (**A**) displays the expression pattern of 738 unique genes associated with 4 molecular subtypes across samples in GSE85217. Small heatmaps (**B-E**) display expression patterns of the top 10 frequently ranked genes within the intersected genes from venn diagrams for each MB subtype. For all heatmaps, rows correspond to genes, and columns correspond to samples in GSE85217. Color scale ranges from blue (low expression) to red (high expression). (**B**) SHH vs rest, (**C**) WNT vs rest, (**D**) Group3 vs rest, (**E**) Group 4 vs rest. (**F-I**) The expression plot of the upregulated DEG for each MB subtype across 13 independent datasets with normalization. Each point represents a sample, colored by expression intensity (red: high, grey: low). (**F**) DEG for SHH, (**G**) DEG for WNT, (**H**) DEG for Group3, (**I**) DEG for Group 4.

Discussion

Medulloblastoma is a highly malignant pediatric brain cancer in cerebellum, the region responsible for controls coordination, balance and other significant motor functions. It accounts for 20% around of pediatric brain tumors, with peak incidence between ages 3-7 and a slight male predominance [3]. Based on the molecular characteristics, it is mainly classified into four subtypes: SHH, WNT, Group 3 and Group 4, each with specific genetic profiles and prognostic outcomes. WNT subgroup typically has the best prognosis with mutations in the WNT signaling pathway, while the SHH subgroup has an intermediate prognosis, associated with alterations in the Sonic Hedgehog pathway. Generally, Group 3 is correlated with the worst prognosis due to the MYC amplification, whereas Group 4 carries an intermediate prognosis and is often associated with chromosomal abnormalities like isochromosome 17q.

These four subtypes have varied symptoms to each other and exhibit significant variations in treatment response and prognosis, leading to tailored treatment strategies based on age, tumor characteristics, and risk factors. Therefore, classifying patients into accurate subtype plays significant roles in the following therapy. Traditional methods like morphological analysis, immunophenotyping, or molecular profiling have been used for MB molecular characterization, are usually time-consuming, costly, labor-intensive, or inaccurate for MB subtyping whereas computational methods are designed to efficiently predict accurate subtype information. However, the absolute gene expression level deviated due to batch effect and biological heterogeneity even if using appropriate normalization methods. Previous studies [45,46], like iPAGE [35], has evidenced the analysis of the intra-sample relative expression level between genes is robust and generalizable in detecting biological signals. Furthermore, gene pair-based analysis is able to perform cross-platform comparison by integrating absolute expression from different cohorts and consequently stable across experimental assays and platforms [35,47]. One main concern for this method is the curse of dimensionality, as the number of possible pairs grows rapidly with the size of the dataset. Additionally, one of the genes in a pair may consistently serve as a reference across multiple samples in the pairwise analysis, leading to repetitive information and limited changes in ranking. To address these concerns, the gene rank analysis was introduced which screened the significantly varied intra-sample differentially ranked genes while filtering out unchanged ones. The number of genes were narrowed down, enhancing the efficiency and effectiveness of the subsequent gene pair matrix construction. Here, we presented RaMBat, a transcriptomic GERs based classifier for MB subtyping which achieved better prediction performance than existing state-of-the-art methods.

The benchmark comparison among RaMBat, medulloPackage, MM2S and other optimized ML classifiers demonstrated that analyzing differentially ranked genes provides robust and reliable insights for MB subtyping, complementing traditional differential expression-based approaches.

RaMBat presented high accuracy across 13 independent test datasets, including GSE10327, GSE12992, GSE21140, GSE30074, GSE37382, GSE41842, GSE49243, GSE50161, GSE50765, GSE62803, GSE67850, GSE73038, GSE74195, with the most misclassifications occurring between Group 3 and Group 4. Among all test samples, one Group 4 sample was misclassified as Group 3, and four Group 3 samples were misclassified as Group 4. Previous studies have noted that these two subtypes exhibit highly similarity in transcriptomic and genomic profiles, contributing to the ambiguity between these two subtypes. Although recent studies have uncovered the heterogeneity within Group 3 and Group 4 and even suggesting their division into eight subtypes [17], patients with Group 3 or Group 4 are currently still treated with same therapeutic regimen [1,3]. However, there are ongoing studies attempting to stratify treatment by molecular subtype to improve outcomes.

RaMBat effectively identifies subtype-specific marker genes, demonstrating its robustness in selecting biologically relevant features for MB subtyping. Specifically, selected GERs contained *NEUROG1, GLI1, ATOH1* associated with the SHH signature. *NEUROG1* and *ATOH1* were the top-selected genes with 40 and 22 GERs respectively, whereas *GLI1* was only involved in one GER. This aligns with the study published in 2007 which indicated *ATOH1, NEUROG1* and *GLI1* might be relevant biomarkers for SHH subtype [48]. *GLI1* is a transcription factor and a key downstream effector of SHH signaling, directly linking it to SHH-driven tumorigenesis. ATOH1 is a well-established marker of cerebellar granule precursor cells, and its expression is strongly associated with medulloblastomas that arise from these cells. In contrast, NEUROG1 highly expressed tumors are suggested to originate from cerebellar ventricular zone progenitors rather than granule cell precursors. Additionally, canonical over-expressed genes *HHIP, BOC* and *SFRP1* were also included in the GERs [49]. The WNT subtype of medulloblastoma is closely linked to the activation of the WNT signaling pathway, primarily through mutations that stabilize beta-catenin and drive oncogenesis. Five WNT signaling genes were observed within GERs associated with the WNT subtype, including *SP5, WIF1, FZD10, CER1,* and *PRKCB*, which play various roles in the WNT signaling pathway. For example, the activation of the WNT pathway leads to the

upregulation of *SP5*, which in turn promotes the expression of genes that contribute to tumorigenesis [50,51]. Its presence indicates active WNT signaling, which is characteristic of the WNT subtype. The *WIF1* plays a significant role in medulloblastoma (MB) by acting as a negative regulator of the WNT signaling pathway [52]. As a receptor in the Frizzled family, FZD10 mediates WNT ligand binding, activating the canonical WNT/β-catenin pathway [53]. Although less is known about the specific drivers, pathways, and biomarkers that discriminate Group 3 and Group 4 subtypes, it is established that these tumors share certain common pathways. For example, *UNC5D* has been identified as a key gene associated with neuronal migration in MB, showing specificity for certain subtypes and being present in GERs [52]. Additionally, specific chromosomal aberrations affecting *AFF3* may contribute to tumorigenesis by disrupting regulatory pathways. The *AFF3* gene could potentially discriminate these two subtypes, as it was downregulated in Group 3, but upregulated in Group 4, as expected, it only present in GERs for Group 3 [52]. Thus, the detailed function of *AFF3* may worth more investigation.

Gene expression ranking information, unlike absolute gene expression levels which are sensitive to batch effects, is scale-independent and proceeds multi-cohort integration regardless of the difference among profile measurements. However, a major limitation of gene ranking-based GERs construction is its high computational cost and time-consuming nature, particularly when handling large-scale datasets, which significantly reduces efficiency. Overall, this study proposes RaMBat, a novel approach leveraging ranking gene expression to select robust and subtype-specific GERs for accurate MB subtyping across diverse transcriptomics datasets with severe batch effects. Benchmarking across 13 datasets demonstrated that RaMBat outperforms state-of-the-art MB subtyping methods, achieving a high median accuracy of 99.02%, while also providing effective batch effect correction. RaMBat visualization further confirmed that RaMBat efficiently eliminates batch effects and clearly separates MB subtypes, unlike conventional approaches such as tSNE, which are susceptible to batch-induced variations. We believe that RaMBat will significantly enhance downstream risk stratification for MB and contribute to the design of more personalized treatment strategies. Additionally, we also expect that RaMBat will be highly effective in addressing many biomedical problems with batch effects for various types of pediatric, adolescent, and young adult (AYA) cancers.

Methods

Datasets

RaMBat was trained and developed on a transcriptomics dataset, GSE85217, characterized using Affymetrix Human Gene 1.1 ST Array and downloaded from the Gene Expression Omnibus [54] database (GEO, https://www.ncbi.nlm.nih.gov/geo). For model performance evaluation, 13 datasets were downloaded through the *getGEO* function in R with following accession numbers: GSE10327, GSE12992, GSE21140, GSE30074, GSE37382, GSE41842, GSE49243, GSE50161, GSE50765, GSE62803, GSE67850, GSE73038, GSE74195. Specifically, these downloaded datasets were in different scales. For example, the maximum value in GSE12992 was smaller than 20, whereas the maximum value in GSE21140 was larger than 70,000. To address scale disparities, datasets were normalized as $(x - \bar{x})/s$, where each gene expression value x was subtracted by the mean value $\bar{x}$ of the column and dividing it by the standard deviation s of the column. In both the training and independent test datasets, Group 4 samples represent the largest proportion, followed by SHH, Group 3, and WNT. Additionally, the number of samples for each subtype is consistently higher in the training dataset compared to the independent test dataset. The training dataset, GSE85217, included 763 MB samples (223 SHH, 70 WNT, 144 Group 3, and 326 Group 4), while the independent test datasets contained 508 samples (177 SHH, 42 WNT, 92 Group 3, and 192 Group 4) (**Supplementary Table 1**).

To identify MB subtypes based on multiple bulk transcriptomics data, an accurate MB subtyping approach based on gene ranking for heterogeneous data with severe batch effects, named RaMBat, was developed. As shown in **Fig. 1A**, RaMBat contains four key steps, including gene rank analysis, reversal ratio analysis, LASSO feature selection, and finally MB subtyping. It leveraged gene expression rankings instead of absolute gene expression levels to address batch effects of different data sources. Based on 13 transcriptomics datasets with severe batch effects, we demonstrated that RaMBat significantly outperformed state-of-the-art methods like medulloPackage and MM2S as well as 7 optimized ML classifiers equipped with the conventional batch correction method ComBat in terms of prediction accuracy for identifying MB subtypes for these 13 datasets.

Intra-sample gene expression analysis

Based on the ascending order, from the smallest to the largest values, we ordered all genes inside the dataset. Subsequently, we transformed the absolute gene expression levels

$$[G_1, G_2, \dots, G_n]$$

into rank values,

$$[T_1, T_2, \dots, T_n]$$

where G is the absolute gene expression value of gene expression, n is the total number of genes, T_i is the rank value of gene i, where ranking is performed in ascending order. The gene with the lowest expression value is assigned a rank of 1, while the gene with the highest expression value is assigned a rank of n. We then divided the rank matrix of GSE85217 into four submatrices corresponding to the SHH, WNT, Group 3, and Group 4, based on sample labels. For each gene, we calculated the gene rank value and a p value between each combination of submatrices, resulting in six different comparisons: SHH vs. WNT, SHH vs. Group 3, SHH vs. Group 4, WNT vs. Group 3, WNT vs. Group 4, and Group 3 vs. Group 4. The mean rank of each gene across all samples within one subtype was calculated and compared to its mean rank in another subtype, with the gene rank representing the difference between these two mean rank scores. A two-sided t-test was then performed to assess the statistical significance of the difference between subtypes. We applied thresholds of $p < 0.05$ and a gene rank > 2314 (representing 11% of the total number of genes) to identify significant differentially ranked genes.

Reversed expression pattern analysis

To construct gene pairs based on the differentially ranked genes, we followed iPAGE [35] protocols to construct GERs. Specifically, for each sample, the rank values of all possible gene pairs were compared, encoding their relationships as binary indicators, represented by

$$R^{(k)} = \left(R_{1,2}^{(k)}, R_{1,3}^{(k)}, \dots, R_{1,n}^{(k)}, R_{2,3}^{(k)}, R_{2,4}^{(k)}, \dots, R_{2,n}^{(k)}, \dots, R_{(n-1)n}^{(k)}\right) \quad (1)$$

$$= \left(s\left(T_1^{(k)} - T_2^{(k)}\right), s\left(T_1^{(k)} - T_3^{(k)}\right), \dots, s\left(T_1^{(k)} - T_n^{(k)}\right), \dots, s\left(T_2^{(k)} - T_3^{(k)}\right), s\left(T_2^{(k)} - T_4^{(k)}\right), \dots, s\left(T_2^{(k)} - T_n^{(k)}\right), \dots, s\left(T_{n-1}^{(k)} - T_n^{(k)}\right)\right)$$

where GERs (R) was constructed by gene rankings (T) in a sample k and function s denotes the process of converting gene rankings into GER values as following:

$$R_{i,j} = s\left(T_i - T_j\right) = \begin{cases} 1, if\ T_i > T_j \\ 0, if\ T_i \leq T_j \end{cases} \tag{2}$$

Every two genes within each sample were subtracted to convert the rank matrix into GER matrix. Then, GERs with a significant difference in their relative expression $R_{i,j}$ between each two subtypes were extracted.

Two situations were taken into consideration, where $R_{i,j} = 1$ and $R_{i,j} = 0$. We counted the number of samples with $R_{i,j} = 1$ and $R_{i,j} = 0$ in the first subtype group as $f_{1,1}$ and $f_{1,2}$ respectively, then in the second subtype group, the number of samples with $R_{i,j} = 1$ and $R_{i,j} = 0$ as $f_{2,1}$ and $f_{2,2}$ respectively. Finally, the contingency table was shown as follows:

	$T_i > T_j$	$T_i \leq T_j$
Selected subtype	$f_{1,1}$	$f_{1,2}$
Rest subtype	$f_{2,1}$	$f_{2,2}$

For example, for the comparison between SHH and WNT, $f_{1,1}$ and $f_{1,2}$ are the number of samples with $R_{i,j} = 1$ and $R_{i,j} = 0$ respectively in SHH, and $f_{2,1}$ and $f_{2,2}$ are the number of samples with $R_{i,j} = 1$ and $R_{i,j} = 0$ respectively in WNT.

$$reversal\ ratio = \frac{f_{1,1} + f_{2,2}}{f_{1,1} + f_{1,2} + f_{2,1} + f_{2,2}} \tag{3}$$

To calculate the reversal ratio, we considered 12 different comparisons, with each subtype having three pairwise comparisons. For example, when select specific GERs for SHH, the comparisons were with WNT, Group 3, and Group 4, respectively. We established different thresholds for the

reversal ratio based on the subtype comparisons. A threshold of 0.95 was set for the six pairwise comparisons within the SHH and WNT subtypes. For comparisons between Group 3 and SHH, Group 4 and SHH, we set the threshold at 0.9. A threshold of 0.85 was used for comparisons between Group 3 and WNT, Group 4 and WNT, Group 4 and Group 3. Lastly, a threshold of 0.8 was applied specifically for the comparison between Group 3 and Group 4. At the same time, we performed Fisher's exact test to obtain the p value for each of the twelve comparisons. The p value was calculated by

$$p = \frac{(f_{1,1} + f_{2,2})!\,(f_{2,1} + f_{2,2})!\,(f_{1,1} + f_{2,1})!\,(f_{2,2} + f_{2,2})!}{f_{1,1}!\,f_{2,2}!\,f_{2,1}!\,f_{2,2}!\,t!} \quad (4)$$

, where $t = f_{1,1} + f_{1,2} + f_{2,1} + f_{2,2}$ and the threshold for p value was set as 0.05. For each subtype, three sets of selected GERs were obtained which were then intersected to identify GERs consistently presenting across all three pairwise comparisons within that subtype. At the end of this step, subtype-specific GERs were selected for SHH, WNT, Group 3 and Group 4, respectively, based on a reversal ratio exceeding the threshold and a p value > 0.05 (**Fig. 1B**).

Selecting informative GERs

To reduce the computational complexity of the subsequent MB sample subtyping step based on GERs and boost the stability of this method, we furthermore refine the number of GERs using LASSO (Least Absolute Shrinkage and Selection Operator) [55]. LASSO naturally performs feature selection by shrinking the coefficients of less important features to zero, effectively eliminating them from the model. We conducted LASSO 12 times to select the most important GERs corresponding to three pairwise comparisons for each of the four subtypes (**Figure. 2C**). Similarly, we intersected three sets of selected GERs for each subtype. In the end, 1035 GERs, corresponding to 738 unique genes, with non-zero coefficients were filtered which were then used as the input for sample subtyping. LASSO was performed using R package "glmnet" [56]. Additionally, based on these 1035 GERs, heatmap was plotted using the R package "pheatmap" [57].

RaMBat Visualization

RaMBat utilizes a combination of two key matrices: a subtype score matrix and a sample-to-subtype matrix derived from prediction results. The subtype score matrix is the output of RaMBat and the prediction subtype for each sample is encoded by one-hot encoding to create a sample-to-subtype matrix, where each row corresponds to a sample, and each column represents a subtype. These two matrices are then combined directly to formulate the final visualization matrix. This combined matrix serves as the input for t-SNE visualization, allowing for a more informative and potentially more biologically relevant representation of the data.

Benchmarking with the state-of-the-art methods

To assess the prediction performance of RaMBat, we initially downloaded 13 independent test datasets with clear subtype information, containing 508 samples (**Table 1**). The subtyping process for a given sample was composed of three stages, 1) convert the absolute gene expression into GERs based on the selected unique genes, 2) filter to the selected GERs in the model and calculate the mean score of GERs for each subtype, 3) determine the subtype with the highest score and assign it to the sample (**Fig. 1D**). Finally, we compared the performance of RaMBat with state-of-the-art MB classifiers, including the medulloPackage [6] and MM2S [25,58]. medulloPackage was also developed using GERs but based on the differentially expressed genes (DEGs) selected by limma package in R. In contrast, MM2S employed a k-nearest neighbor (KNN) [59] classifier based on the gene-set enrichment analysis ranked matrix. Additionally, to compare the performance of RaMBat with machine learning methods, 7 classifiers, including support vector machine (SVM) [60], logistic regression (LR) [61], random forest (RF) [62], XGBoost [63], KNN, naïve bayes (NB) , multilayer perceptron (MLP) [54] were trained and optimized. The training dataset and independent test datasets which were normalized and then batch effect corrected by the *combat* function in R were used for training and testing of these ML classifiers. For performance evaluation, we considered nine major metrics: accuracy, precision, sensitivity, specificity, area under the curve (AUC), Jaccard index (Jaccard), Matthews correlation coefficient (MCC), G-measure and F1-score.

Competing interests.

The authors declare no competing interests.

Funding

Research reported in this publication was supported by the Office Of The Director, National Institutes Of Health of the National Institutes of Health under Award Number R03OD038391, and by the National Cancer Institute of the National Institutes of Health under Award Number P30CA036727. This work was supported by the American Cancer Society under award number IRG-22-146-07-IRG, and by the Buffett Cancer Center, which is supported by the National Cancer Institute under award number CA036727. This work was also partially supported by the National Institute of General Medical Sciences under Award Numbers P20GM103427. This study was in part financially supported by the Child Health Research Institute at UNMC/Children's Nebraska. This work was also partially supported by the University of Nebraska Collaboration Initiative Grant from the Nebraska Research Initiative (NRI). The content is solely the responsibility of the authors and does not necessarily represent the official views from the funding organizations.

Author contributions

SW conceived and designed the study. SW and MS developed the method, performed the experiments and analyzed the data. All authors participated in writing the paper. SW and MS revised the paper. The manuscript was approved by all authors.

Data availability

All the data used in this manuscript are publicly available in the corresponding references.

Code availability

RaMBat is available as an R package in GitHub at https://github.com/wan-mlab/RaMBat.

Reference

1. Juraschka, K. & Taylor, M. D. Medulloblastoma in the age of molecular subgroups: a review. *J Neurosurg Pediatr* **24**, 353–363 (2019).
2. Martin, A. M., Raabe, E., Eberhart, C. & Cohen, K. J. Management of Pediatric and Adult Patients with Medulloblastoma. *Curr. Treat. Options in Oncol.* **15**, 581–594 (2014).
3. Northcott, P. A. *et al.* Medulloblastoma. *Nat Rev Dis Primers* **5**, 11 (2019).
4. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* **131**, 803–820 (2016).
5. Taylor, M. D. *et al.* Molecular subgroups of medulloblastoma: the current consensus. *Acta Neuropathol* **123**, 465–472 (2012).
6. Rathi, K. S. *et al.* A transcriptome-based classifier to determine molecular subtypes in medulloblastoma. *PLoS Comput Biol* **16**, e1008263 (2020).
7. Clifford, S. C. *et al.* Wnt/Wingless pathway activation and chromosome 6 loss characterize a distinct molecular sub-group of medulloblastomas associated with a favorable prognosis. *Cell Cycle* **5**, 2666–2670 (2006).
8. Ellison, D. W. *et al.* beta-Catenin status predicts a favorable outcome in childhood medulloblastoma: the United Kingdom Children's Cancer Study Group Brain Tumour Committee. *J Clin Oncol* **23**, 7951–7957 (2005).
9. Kool, M. *et al.* Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. *Cancer Cell* **25**, 393–405 (2014).
10. Thomas, A. & Noël, G. Medulloblastoma: optimizing care with a multidisciplinary approach. *JMDH* **Volume 12**, 335–347 (2019).

11. Fellay, C. N. *et al.* Medulloblastomas in adults: prognostic factors and lessons from paediatrics. *Curr Opin Neurol* **24**, 626–632 (2011).
12. Choi, J. Y. Medulloblastoma: Current Perspectives and Recent Advances. *Brain Tumor Res Treat* **11**, 28–38 (2023).
13. Northcott, P. A. *et al.* The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017).
14. Kool, M. *et al.* Molecular subgroups of medulloblastoma: an international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas. *Acta Neuropathol* **123**, 473–484 (2012).
15. Northcott, P. A. *et al.* Medulloblastomics: the end of the beginning. *Nat Rev Cancer* **12**, 818–834 (2012).
16. Ryan, S. L. *et al.* MYC family amplification and clinical risk-factors interact to predict an extremely poor prognosis in childhood medulloblastoma. *Acta Neuropathol* **123**, 501–513 (2012).
17. Sharma, T. *et al.* Second-generation molecular subgrouping of medulloblastoma: an international meta-analysis of Group 3 and Group 4 subtypes. *Acta Neuropathol* **138**, 309–326 (2019).
18. Gurney, J. G. *et al.* Endocrine and cardiovascular late effects among adult survivors of childhood brain tumors: Childhood Cancer Survivor Study. *Cancer* **97**, 663–673 (2003).
19. Packer, R. J. *et al.* Long-term neurologic and neurosensory sequelae in adult survivors of a childhood brain tumor: childhood cancer survivor study. *J Clin Oncol* **21**, 3255–3261 (2003).
20. Ramaswamy, V. *et al.* Risk stratification of childhood medulloblastoma in the molecular era: the current consensus. *Acta Neuropathol* **131**, 821–831 (2016).

21. Gibson, P. *et al.* Subtypes of medulloblastoma have distinct developmental origins. *Nature* **468**, 1095–1099 (2010).
22. Godbole, S. *et al.* Multiomic profiling of medulloblastoma reveals subtype-specific targetable alterations at the proteome and N-glycan level. *Nat Commun* **15**, 6237 (2024).
23. Shih, D. J. H. *et al.* Cytogenetic prognostication within medulloblastoma subgroups. *J Clin Oncol* **32**, 886–896 (2014).
24. Rani, B. & Sharma, V. K. Transcriptome profiling: methods and applications- A review. *AG* **38**, (2017).
25. Gendoo, D. M. A. & Haibe-Kains, B. MM2S: personalized diagnosis of medulloblastoma patients and model systems. *Source Code Biol Med* **11**, 6 (2016).
26. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
27. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research* **42**, e161–e161 (2014).
28. Benito, M. *et al.* Adjustment of systematic microarray data biases. *Bioinformatics* **20**, 105–114 (2004).
29. Shabalin, A. A., Tjelmeland, H., Fan, C., Perou, C. M. & Nobel, A. B. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* **24**, 1154–1160 (2008).
30. Gagnon-Bartsch, J. A. & Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539–552 (2012).

31. Weishaupt, H. *et al.* Batch-normalization of cerebellar and medulloblastoma gene expression datasets utilizing empirically defined negative control genes. *Bioinformatics* **35**, 3357–3364 (2019).

32. Wang, D. *et al.* Extensive up-regulation of gene expression in cancer: the normalised use of microarray data. *Mol Biosyst* **8**, 818–827 (2012).

33. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29–39 (2016).

34. Parker, H. S. *et al.* Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. *Bioinformatics* **30**, 2757–2763 (2014).

35. Zheng, X., Leung, K.-S., Wong, M.-H. & Cheng, L. Long non-coding RNA pairs to assist in diagnosing sepsis. *BMC Genomics* **22**, 275 (2021).

36. Reddy, A. *et al.* Gene Expression Ratios Lead to Accurate and Translatable Predictors of DR5 Agonism across Multiple Tumor Lineages. *PLoS ONE* **10**, e0138486 (2015).

37. Li, Q. *et al.* bvnGPS: a generalizable diagnostic model for acute bacterial and viral infection using integrative host transcriptomics and pretrained neural networks. *Bioinformatics* **39**, btad109 (2023).

38. Jin, N. *et al.* PAGE-based transfer learning from single-cell to bulk sequencing enhances model generalization for sepsis diagnosis. *Briefings in Bioinformatics* **26**, bbae661 (2024).

39. Zhao, P. *et al.* Intra-sample reversed pairs based on differentially ranked genes reveal biosignature for ovarian cancer. *Computers in Biology and Medicine* **172**, 108208 (2024).

40. Wang, R. *et al.* Improving bulk RNA-seq classification by transferring gene signature from single cells in acute myeloid leukemia. *Briefings in Bioinformatics* **23**, bbac002 (2022).

41. Wu, Q. *et al.* meGPS: a multi-omics signature for hepatocellular carcinoma detection integrating methylome and transcriptome data. *Bioinformatics* **38**, 3513–3522 (2022).

42. Li, H. *et al.* Whole transcriptome analysis reveals non-coding RNA's competing endogenous gene pairs as novel form of motifs in serous ovarian cancer. *Comput Biol Med* **148**, 105881 (2022).

43. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

44. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**, 267–288 (1996).

45. Heinäniemi, M. *et al.* Gene-pair expression signatures reveal lineage control. *Nat Methods* **10**, 577–583 (2013).

46. Wang, H. *et al.* Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics* **31**, 62–68 (2015).

47. Li, N. *et al.* Pairwise analysis of gene expression for oral squamous cell carcinoma via a large-scale transcriptome integration. *J Cellular Molecular Medi* **28**, e70153 (2024).

48. Salsano, E. *et al.* Expression of the neurogenic basic helix-loop-helix transcription factor NEUROG1 identifies a subgroup of medulloblastomas not expressing ATOH1. *Neuro Oncol* **9**, 298–307 (2007).

49. Min, H. S., Lee, Y. J., Park, K., Cho, B.-K. & Park, S.-H. Medulloblastoma: histopathologic and molecular markers of anaplasia and biologic behavior. *Acta Neuropathol* **112**, 13–20 (2006).

50. Weidinger, G., Thorpe, C. J., Wuennenberg-Stapleton, K., Ngai, J. & Moon, R. T. The Sp1-related transcription factors sp5 and sp5-like act downstream of Wnt/beta-catenin signaling in mesoderm and neuroectoderm patterning. *Curr Biol* **15**, 489–500 (2005).
51. Fujimura, N. *et al.* Wnt-mediated down-regulation of Sp1 target genes by a transcriptional repressor Sp5. *J Biol Chem* **282**, 1225–1237 (2007).
52. Kool, M. *et al.* Integrated Genomics Identifies Five Medulloblastoma Subtypes with Distinct Genetic Profiles, Pathway Signatures and Clinicopathological Features. *PLoS ONE* **3**, e3088 (2008).
53. Northcott, P. A. *et al.* Medulloblastoma Comprises Four Distinct Molecular Variants. *JCO* **29**, 1408–1414 (2011).
54. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41**, D991–D995 (2012).
55. Kukreja, S. L., Löfberg, J. & Brenner, M. J. A LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO) FOR NONLINEAR SYSTEM IDENTIFICATION. *IFAC Proceedings Volumes* **39**, 814–819 (2006).
56. Friedman, J. D., Reece, G. R. & Eldor, L. The Utility of the Posterior Thigh Flap for Complex Pelvic and Perineal Reconstruction: *Plastic and Reconstructive Surgery* **126**, 146–155 (2010).
57. Raivo Kolde. pheatmap: Pretty Heatmaps.
58. Gendoo, D. M. A., Smirnov, P., Lupien, M. & Haibe-Kains, B. Personalized diagnosis of medulloblastoma subtypes across patients and model systems. *Genomics* **106**, 96–106 (2015).

59. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13**, 21–27 (1967).

60. Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20**, 273–297 (1995).

61. Tolles, J. & Meurer, W. J. Logistic regression: relating patient characteristics to outcomes. *Jama* **316**, 533–534 (2016).

62. Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Machine Intell.* **20**, 832–844 (1998).

63. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* 785–794 (2016).

64. Hand, D. J. & Yu, K. Idiot's Bayes: Not So Stupid after All? *International Statistical Review / Revue Internationale de Statistique* **69**, 385 (2001).