# Title: *In silico* analysis of long non-coding RNAs in medulloblastoma and its subgroups

**Short title:** LncRNAs in medulloblastoma

Piyush Joshi[1, 2] and Ranjan J. Perera* [1, 2, 3]

[1] Cancer and Blood Disorder Institute, Johns Hopkins All Children's Hospital, 600 5th St. South, St. Petersburg, FL 33701 USA

[2] Department of Oncology, Sidney Kimmel Comprehensive Cancer Center ,School of Medicine, Johns Hopkins University, 1650 Orleans St., Baltimore, MD 21231 USA

[3] Sanford Burnham Prebys Medical Discovery Institute, 10901 N Torrey Pines Rd, La Jolla, CA 92037 USA

*Corresponding author email: jperera2@jhmi.edu

**Running head:** Long noncoding RNA expression profiling in medulloblastoma

## Abstract

**Background:** Medulloblastoma is the most common malignant pediatric brain tumor with high fatality rate. Recent large-scale studies utilizing genome-wide technologies have sub-grouped medulloblastomas into four major subgroups: wingless (WNT), sonic hedgehog (SHH), group 3, and group 4. However, there has yet to be a global analysis of long non-coding RNAs, a crucial part of the regulatory transcriptome, in medulloblastoma.

**Methods and findings:** We performed bioinformatic analysis of RNA-seq data from 175 medulloblastoma patients. Differential lncRNA expression sub-grouped medulloblastomas into the four main molecular subgroups. Some of these lncRNAs were subgroup-specific, with a random forest-based machine-learning algorithm identifying an 11-lncRNA diagnostic signature. We also validated the diagnostic signature in patient derived xenograft (PDX) models. We further identified a 17-lncRNA prognostic model using LASSO based penalized Cox' PH model (low risk group HR= 0.207, 95% CI= 0.133-0.323, p-value= 2e-14).

**Conclusions:** Our analysis represents the first global lncRNA analysis in medulloblastoma. Our results identify putative candidate lncRNAs that could be evaluated for their functional role in medulloblastoma genesis and progression or as diagnostic and prognostic biomarkers.

## Introduction

Medulloblastoma (MB), characterized as WHO group IV, represents the most common malignant pediatric central nervous system (CNS) tumor [1-4], representing 9.2% of all pediatric brain tumor cases [1, 5] and roughly 500 new cases each year in the US. MBs originate in the cerebellum and share molecular signatures with embryonic cerebellar lineages, with metastasis sites commonly include parts of the brain, spinal cord, and, rarely, to extraneural sites [6-8].

Commonly used treatment strategies for MB include maximal safe surgical resection, radiotherapy, and chemotherapy, which are poorly tolerated by pediatric patients who are usually under seven years of age [9]. Appropriate treatment selection depends upon the clinical subgroup, stage, extent of resection, location, and the patient's ability to withstand treatment [10]. In efforts to improve therapeutic outcomes, combined genetic and epigenetic approaches have refined MB classification into four clinically and molecularly distinct subgroups: wingless (WNT), sonic hedgehog (SHH), group 3, and group 4 [11]. Despite these significant advances, MB remains deadly for many patients, with a ~30% fatality rate. Further, even successful eradication of the tumor often results in a deteriorated overall quality of life due to side effects including organ dysfunction, neurocognitive impairment, endocrine disabilities, and secondary tumors [10-13]. Even with these advances in molecular classification, group 3 and group 4 tumors are heterogeneous groups that continue to make management challenging. There is an urgent need to identify the underlying molecular mechanisms in these subgroups to drive precision medicine-based approaches, improve quality of life, and increase our understanding of MB in general.

Long non-coding RNAs (lncRNAs) represent a major part of the transcribed genome that do not code for functional proteins. LncRNAs are more than 200 nucleotides in length and are transcribed by RNA polymerase II. While previously labelled as transcriptional “noise”, it is now understood that lncRNAs are functional and play important roles in cellular physiology, development, and disease progression. In humans, there are at least three times as many lncRNAs as protein-coding genes [14]. Although the precise roles of the vast majority of identified or predicted lncRNAs remain unknown [14], they are increasingly recognized as being involved in *cis* or *trans* interactions regulating gene expression in the nucleus and protein interactions in the nucleus and cytoplasm. Some of the functionally diverse roles of lncRNAs include transcriptional silencing (e.g., *XIST* [15]), enhancers by regulating three-dimensional chromosomal structure to strengthen interactions between enhancers and promoters (e.g., *LUNAR1* [16]), and as microRNA sponges that sequester microRNAs from their target sites (e.g., *SNHG7* [17]). LncRNAs can also act as scaffolds for protein-protein and protein-nucleic acid interactions [18]. They are potential biomarkers and therapeutic targets in cancer, with several lncRNAs now studied for their oncogenic or tumor suppressor potential in several cancer types through their regulation of the cell cycle, cell death, senescence, metastasis, immunity, and cancer cell metabolism [19].

LncRNAs are also implicated in CNS tumors including glioblastoma and neuroblastoma [20, 21]. However, there has yet to be a genome-wide study of MB to identify dysregulated lncRNAs. With this aim, we analyzed the transcriptomic profiles of 175 MB patients to map lncRNA expression profiles and identify subgroup-specific lncRNAs. We show that the MB lncRNAome exhibits significant heterogeneity that corresponds to the molecular subtypes. Using a random forest-based machine-learning algorithm, we identify lncRNA signatures that could improve on present diagnostic approaches, while penalized Cox-PH regression identifies prognostic lncRNAs. Taken together, our analysis identifies candidate lncRNAs with subgroup-specific activity in MB and with diagnostic and prognostic value.

## Materials and methods

### Datasets

Raw FASTQ files for RNA-seq data corresponding to 175 MB patients (referred to as the ICGC dataset) belonging to four subgroups (accession number EGAS00001000215) were downloaded from the European Genome-Phenome Archive (EGA, http://www.ebi.ac.uk/ega/) after obtaining approval from the Institutional Review Board (IRB) (**Table S1**)[22]. Pre-analyzed microarray expression datasets from 763 patients belonging to the four medulloblastoma subgroups were obtained from the study published by Cavalli et al. (referred to as the MAGIC dataset) [23].

### RNA-seq library preparation

RNA sequencing for patient derived xenograft (PDX) samples was undertaken at the Genetic Resources Core Facility at the Johns Hopkins University, School of Medicine, Baltimore, MD. Before sequencing, total RNA was extracted from PDX cell pellets using the Direct-zol RNA miniprep kit (R2060, Zymo Research, Irvine, CA), with subsequent quantification using Nanodrop (Thermo Fisher Scientific, Waltham, MA) and quality assessment with the Agilent Bioanalyzer Nano Assay (Agilent Technologies, Santa Clara, CA). RNA-seq libraries were constructed using the Illumina TruSeq Stranded Total RNA Library preparation Gold kit (20020598, Illumina Inc., San Diego, CA) as per the instructions. The quality and quantity of the libraries were analyzed using the Agilent Bioanalyzer and Kapa Biosystems qPCR (Sigma

Aldrich, St. Louis, MO). Multiplexed libraries were pooled, and paired-end 50 base-pair sequencing was performed on a NovaSeq6000. RNA-seq data are available at the Gene Expression Omnibus (GEO) Accession Number GSE134248.

### RNA-seq alignment, quantification, and differential gene expression analysis

Raw FASTQ files were quality checked for adapter contamination using FASTQC. FASTQ files containing adapter sequences were trimmed by running through trim_galore in default mode. The reads were mapped to the GRCh38/hg38 human genome assembly p12 (release 28, www.gencodegenes.org) using HISAT2 and annotated using the corresponding release version GENCODE comprehensive annotation file and LNCipedia 5.2 high confidence set annotation file. Mapped reads were quantified using StringTie to obtain FPKM values, which were converted to read counts using the prepDE.py script (provided in StringTie manual). For variance-stabilized normalized reads and differential gene expression analysis, reads counts were processed with *DESeq2* in R [24].

### Consensus clustering

Variance-stabilized expression levels of the top 10,000 variant lncRNAs determined from standard deviations of read counts normalized to library size were used as input to perform 1000 permutations of k-means-based consensus clustering using *ConsensusClusterPlus* R package [25].

### Co-expression module detection and trait correlation analysis

Variance-stabilized expression of top 5000 variant lncRNAs was used to obtain a weighted correlation network using the *WGCNA* R package [26]. The correlated lncRNA cohorts were associated with MB subgroups using the module-trait correlation algorithm as described [26].

### Random forest model

Subgroup-specific diagnostic models were obtained by performing variable selection using expression of differentially expressed lncRNAs/ protein coding genes (PCGs) and the *randomForest* R package (as described in [27]). For all models, variance-stabilized expression of differentially expressed lncRNA/PCG genes were used as variables to obtained models to classify patients into known subgroups. For the lncRNA model distinguishing SHH, group 3, and group 4, patient samples were divided into a 60% training set and 40% tuning set. Only differentially expressed (|logFC| >1.5, padj <0.01) lncRNAs genes between SHH, group 3, and group 4 were used to classify patients into known subgroups. The training model was used to find important genes ranked based on the “mean decrease accuracy” parameter. Low ranking genes with high expression correlation (>0.80) to high ranked genes were discarded. Gene combinations based on the final ranked list were used in the tuning model to find the minimum number of genes resulting in the minimum or comparatively lower error rate in the tuning set. A similar approach was used to distinguish WNT from the other subgroups. For the training set, all WNT samples were combined with 60% samples from SHH, group 3, and group 4 samples. For the tuning set, all WNT samples were combined with 40% training set the remaining subgroups. To distinguish group 3 and group 4, a similar 60%-40% training-tuning model was adapted for classification.

A random forest-based model for PCGs distinguishing WNT, SHH, group 3, and group 4 was obtained as described above for lncRNAs. To validate the protein coding model, expression levels of the obtained signature genes were used to classify patient samples from the MAGIC dataset using the random forest model and tSNE plots.

Receiver operating characteristics (ROC) curves and area under the ROC curve (AUC) values for one versus rest comparisons were computed based on a generalized linear model-based fit of subgroup identity with normalized gene expression levels of signature genes as the variable using the *pROC* R package.

**Transcriptional network inference**

A transcriptional inference network for putative regulation between candidate lncRNAs and transcription factors was obtained using *minet* R package [28]. Regulatory interaction measures were obtained based on the network obtained from CLR-, arcane-, and mrnet-based models. Only edge connections predicted in all three approaches were analyzed further for first neighbor connections of transcription factors with each candidate lncRNA.

**Survival analysis**

Expression of 621 lncRNAs included in the MAGIC microarray dataset for 612 patients with survival information belonging to the four subgroups was used as input to find prognostic signatures using penalized Cox's proportional hazard (Cox-PH) model with the LASSO ($\alpha$ =1) penalty using *glment* R package [29]. Smallest mean squared error and *lambda.min* was selected from 100 random runs of the model fitting with 10 fold cross validation in each run. Variables (lncRNAs) with non-zero coefficients associated with the *lambda.min* were selected as the prognostic model. Stability of the obtained model was verified by performing 1000 bootstraps on the data using *BootValidation* R package. A risk score was calculated for each patient by summing the value of the product of normalized expression and penalized Cox-PH coefficient of a candidate lncRNA with that of all other candidate lncRNAs. Kaplan-Meier analysis was conducted using the obtained risk score of a 17-lncRNA prognostic model for the MAGIC dataset with survival information and 74 patients belonging to four subgroups with survival information in the ICGC RNA-seq dataset.

## Results

### Long non-coding RNA signatures of medulloblastoma and subgroup-specific lncRNA enrichment

To determine genome-wide expression profiles of lncRNAs in MB, we analyzed RNA-seq data from 175 patient samples obtained from the ICGC PedBrain dataset. For comprehensive lncRNA annotation, we chose a combination of the GENCODE and LNCipedia datasets [30]: GENCODE represents the largest manually curated lncRNA dataset and LNCipedia contains the maximum number of high fidelity predicted lncRNA genes [30]. The expression of 52,128 unique lncRNAs and 19,033 protein-coding genes (PCGs) were quantified (**Fig 1A**) in 18 WNT MBs, 45 SHH MBs, 46 group 3 MBs, and 66 group 4 MBs. To better understand the role of lncRNAs in different MB clinical and molecular subgroups, we investigated correlations between lncRNA expression and subgroup type. First, we performed consensus clustering using the top 10,000 highly variant lncRNAs, which clustered the MBs into four different groups that highly overlapped with the known molecular subgroups (**Fig 1B, Fig S1**), suggesting that lncRNAs could contribute to subgroup-specific traits.

**Fig 1. Long non-coding RNA profiles of medulloblastoma.** (A) Schematic of raw data processing and analysis for medulloblastoma (MB) patients belonging to four subgroups: WNT, SHH, group 3 and group 4. (B) Heatmap showing cluster stability obtained from 1000 permutations of k-mean based clustering of 175 MB patients with top 10,000 variably expressed lncRNAs as the variable. Color range depicts samples never clustered together (0, blue) to always clustered together (1, red). (C) Heatmap showing correlations of identified lncRNA cohorts (y-axis) from highly variable 5000 lncRNAs with MB subgroup pehnotype (x-axis). Values in a cell show correlation level (above) and significance p-value (below in brackets). (D) Heatmap showing scaled expression level of lncRNAs in the identified cohorts (y-axis) across samples belonging to MB subgroups (x-axis).

With the objective of identifying highly variable lncRNAs specifically enriched in each subgroup, we performed weighted correlation analysis using WGCNA to find expressional co-related lncRNAs and their subgroup specificity. Weighted co-expression analysis of the top 5000 highly variable lncRNAs identified nine distinct cohorts after merging modules below the threshold. We next determined subgroup-specific module expression by performing module-trait association to obtain each module’s correlation and significance value (**Fig 1C, Table S2**). Module cohorts were significantly positively correlated with WNT (467 lncRNAs, A3), SHH (452 lncRNAs, A4), group 3 (629 lncRNAs, A9), and group 4 (760 lncRNAs, A8) MBs; respectively. Gene expression in each of the identified modules also showed that these genes are highly co-expressed in their respective groups compared to other groups (**Fig 1D**). In addition, cohorts enriched in group 3 and group 4 were more correlated than WNT and SHH MBs, and vice versa. This suggests that similar to protein-coding gene expression and DNA methylation patterns, group 3 and group 4 patients also share similarities based on lncRNAs' expression.

**A candidate diagnostic lncRNA signature for medulloblastoma subgroup classification**

WGCNA analysis suggested a number of lncRNAs with subgroup-specific expression. We therefore proceeded to identify the minimum number of lncRNAs that could faithfully classify MB subgroups. To achieve our objective, we used random forest based machine learning approach that has been shown to be a robust method for such classification objectives [27]. As patients were not evenly distributed between the four subgroups, we adopted a two-step approach: first, we identified a signature for groups with similar patient distributions i.e., SHH, group 3, and group 4; second, we identified a signature distinguishing WNT from the other subgroups (**Fig 2A**). Using this two-step approach, an 11-lncRNA signature was identified with an average <7% class error rate. Using the 11-lncRNA model, the 175 samples were re-classified into the already known subgroups with few misclassifications (**Fig 2B**), with individual lncRNA showing highly subgroup specific up/down expression (**Fig 2C**). Importantly, patient ICGC_MB23, which was classified as SHH in our random forest model but labeled WNT in the obtained dataset, was originally considered as an SHH MB in Kool et al. [31] and lacked the signature mutation in β-catenin. We also validated the 11-lncRNA based patient grouping using t-SNE based clustering (**Fig 2D**, that did not classify ICGC_MB23 as SHH) and specificity and sensitivity of the model using ROC/AUC analysis performing one versus rest comparison (**Fig 2E**).

**Fig 2. Random forest-based approach identifies an 11-lncRNA model to classify medulloblastoma subgroups.** (A) Schematic depiction of the modeling process. First, a 9-lncRNA model distinguishing SHH, group 3, and group 4 patients was obtained using a 60%-40% training-tuning partition. Then, a 2-lncRNA model distinguishing WNT from the rest of the group was obtained by combining all WNT samples with a 60%-40% partition of the other subgroup patients in a training-tuning model. (B) MB patient subgroups as identified from the random forest model using only 11-lncRNA expression as variables. Dendrogram representing hierarchical clustering of dissimilarity values obtained from random forest-based classification. ICGC_MB23 is the sole WNT MB patient that clusters with the rest of the SHH MBs. Bottom color bars represent known clinical groupings (blue=WNT, green=SHH, black=group 3, red=group 4). (C) Boxplot showing distribution of normalized expression values of identified 11-lncRNAs in each patient subgroup. Purple dots represent normalized expression values for a patient. (D) tSNE plot showing clustering of patients into four subgroups based on normalized expression level of identified 11-lncRNAs (blue=WNT, green=SHH, black=group 3, red=group 4). (E) ROC analysis of the linear model based on normalized expression of the identified 11-lncRNAs comparing one versus all (rest) classifications for each of the subgroups.

In the absence of an independent dataset containing expression levels of the 11 candidate lncRNAs to validate the model, we instead validated our random forest model building process. We performed a similar classification of 175 MB samples using protein-coding genes to produce a 14-PCG model with equivalent success to the lncRNA model in classifying patient samples into the four subgroups (**Fig S2**). We then validated the 14-PCG model using the independent MAGIC microarray dataset of 763 patient belonging to the four subgroups. As expected, the 14-

PCG model performed well with an overall <4% class error rate, validating our random forest model building process (**Fig S3**).

Group 3 and group 4 represent the two most heterogeneous yet closely related and difficult to distinguish MB subgroups, a pattern evident from diagnostic signature-based clustering (**Fig 2D**). To identify lncRNAs that could distinguish group 3 and group 4 tumors, we again used our random forest model approach to select highly differentially regulated and discriminative genes (**Table S3**). Our analysis yielded an 8-lncRNA model that did not improve the overall efficiency of group 3 versus group 4 classification (**Fig S4**) compared to the 11-lncRNA model distinguishing all subgroups (**Fig 2**). Nevertheless, the analysis did reveal some lncRNAs with potential functional roles in group 3 or group 4 MBs (**Fig S4B**), some of which overlapped with the 11-lncRNA model (i.e., *MIR100HG*, *USP2-AS1,* and *lnc-CFAP100-4*). However, we also identified other candidate lncRNAs including *ARHGEF7-AS2*, *lnc-HLX-1, lnc-EXPH5-2, lnc-CH25H-2 and lnc-TDRP-3* that showed group-specific differential expression in group 3 or group 4 patients (**Fig S4B**).

We further validated our random forest-based model in patient derived xenograft (PDX) samples derived from SHH (BT084, DMB012, RCMB32, and MED1712FH), group 3 (RCMB28, MB002, MB511H, and RCMB40), and group 4 (RCMB51, DMB006, RCMB45 and RCMB38) patients using the 9-lncRNA signature to classify SHH, group 3, and group 4 patients (**Fig 2A**), as WNT PDXs were not available for analysis. SHH, group 3, and group 4 samples were successfully identified using k-mean based clustering, principal component analysis (PCA) (**Fig 3A**) and hierarchical clustering using normalized RNA-seq expression levels (**Fig 3B**), with the exception of RCMB28 that was found to be more related to SHH PDXs. Quantitative expression of signature genes was validated by qPCR (**Fig 3C and D**) and closely resembled expression in patient RNA-seq data (**Fig 2C**).

**Fig 3. Candidate lncRNAs successfully classify PDX samples into medulloblastoma subgroups.** PDX sample clustering obtained using normalized expression (RNA-seq) of 9-lncRNAs (9-lncRNA model distinguishing SHH, group 3, and group 4) in each PDX sample as the variable via (A) k-means clustering superimposed on principal component analysis (PCA), and (B) hierarchical clustering. Boxplot distributions of expression levels of the identified 9-lncRNAs from (C) RNA-seq and (D) qPCR analysis (-dCt = Ct (*candidate*) – Ct (*ACTB*)). Purple dots represent the expression level in a sample belonging to the known MB subgroup.

In order to infer the physiological importance of the identified 11-lncRNA candidates, we used a combination of CLR, arcane, and mrnet transcriptional inference algorithms to identify potential interactions between the identified lncRNAs with the expressed transcription factors in MB patients. The identified lncRNAs could potentially interact with a number of transcription factors in a complex interconnected network with both highly positively and negatively correlated associations, suggesting putative biological cross-regulation (**Fig S5**).

## Prognostic lncRNAs in medulloblastoma

To identify prognostic lncRNAs, we used the MAGIC microarray array dataset containing survival data for 612 (out of 763) patients. The microarray expression data contained expression levels of 621 lncRNAs that we used for multivariate Cox proportional hazards (Cox-PH) regression analysis. For feature selection, we utilized a penalized multivariate Cox-PH model using the LASSO penalty ($\alpha = 1$). We first used the entire dataset to select 17 lncRNAs as prognostic markers and their associated penalized coefficients (**Table 1**). Of these 17 lncRNAs, 10 were markers of good prognosis and 7 were associated with poor prognosis. The 17-lncRNAs model was validated on 1000 bootstraps of the MAGIC dataset, that showed predictive stability of the proposed prognostic model. Using the penalized coefficients and log-normalized expression values for each lncRNA, we assigned a risk score to each patient. Kaplan-Meier analysis of 612 patients using the risk score as the input variable suggested that our risk score signature was a highly significant in prognostic value (Low risk HR= 0.207, 95% CI= 0.133-0.323, logrank p-value= 2e-14) (**Fig 4A**). To validate our 17-lncRNA prognostic model in an independent dataset, we used ICGC RNA-seq data of 74 patients with survival information and used the penalized coefficients obtained from the MAGIC dataset analysis and variance stabilized expression from RNA-seq data to obtain an equivalent risk score. Again, Kaplan-Meier analysis showed that prognostic significance of our 17-lncRNA model (Low risk HR= 0.135 , 95% CI= 0.017-1.08, logrank p-value= 0.026) (**Fig 4B**). These candidate lncRNAs could potentially be involved in MB development as pro or anti-tumorigenic factors.

**Table 1. Prognostic long non-coding RNA signature genes and associated penalized Cox-PH coefficient**

| **Ensembl Gene ID** | **Gene Symbol** | **Penalized coefficient** |
|---|---|---|
| ENSG00000124915 | lnc-TMEM258-3 | -0.47771 |
| ENSG00000130600 | H19 | 0.102589 |

| | | |
|---|---|---|
| ENSG00000163009 | lnc-RRM2-3 | 0.107783 |
| ENSG00000177993 | ZNRF3-AS1 | -0.24098 |
| ENSG00000186960 | LINC01551 | 0.073539 |
| ENSG00000197251 | LINC00336 | 0.042296 |
| ENSG00000205444 | lnc-CDYL-1 | 0.198379 |
| ENSG00000231010 | lnc-PRR34-1 | -0.0379 |
| ENSG00000231160 | KLF3-AS1 | -0.05371 |
| ENSG00000234665 | lnc-FOXD4L5-25 | -0.03664 |
| ENSG00000235954 | TTC28-AS1 | -0.00891 |
| ENSG00000244620 | lnc-TMEM121-3 | -0.17041 |
| ENSG00000255650 | FAM222A-AS1 | 0.007403 |
| ENSG00000256124 | LINC01152 | -0.0675 |
| ENSG00000267278 | MAP3K14-AS1 | -0.07358 |
| ENSG00000272841 | AL139393.2 | 0.231787 |
| ENSG00000276399 | AC209154.1 | -0.01405 |

**Fig 4. Penalized Cox proportional hazards-based lncRNA model classifies medulloblastoma patients into high and low risk groups.** (A) Patients (612 MAGIC dataset) were grouped into two groups based on risk score (above and below median risk score) derived from expression of

candidate prognostic lncRNAs significantly differing in their survival probability. (B) Risk score derived from the expression of the same candidate lncRNAs (except *lnc-TMEM123-3*, not detected in RNA-seq) in an independent patient cohort (74 patients in ICGC RNA-seq dataset).

## Discussion

Long non-coding RNAs are increasingly recognized as important players in cancer research [19], particularly as biomarkers and/or therapeutic targets [32-36], including in brain tumors [20, 37-40]. However, there is a lack of knowledge of lncRNAs' involvement in MBs. Here we bridged this knowledge gap by proposing diagnostic and prognostic biomarkers candidates for further study *in vitro* and *in vivo* systems to understand their potential function in MB genesis/progression.

Our study is the first genome-wide analysis of lncRNAs' expression profile in MB and its subgroups. Overall lncRNAs' expression dynamics mirrors the well-known MB heterogeneity seen in genetic and epigenetic analyses [23, 41]. MB subgroup clusters obtained using highly variable lncRNAs overlapped with existing clinical and molecular subgroups. Using variantly expressed lncRNAs and weighted correlations, we further identified subgroup-specific lncRNAs. These upregulated lncRNAs might represent functionally relevant genes and require further validation. The obtained lncRNA signature could be curated using transcriptional inference algorithms and proximity to or co-relation with known MB relevant protein coding genes for further functional validation *in vitro* and *in vivo* studies.

Presently, very few lncRNAs have been studied for their putative roles in MB or its subtypes. *NKX2-2AS* was shown *in vitro* to modulate SHH-potentiated MB development by acting as a miRNA sponge for miR-103 and miR-107, thereby depressing their tumor suppressive targets BTG2 and LATS1 and inhibiting proliferation and migration [42]. *CDKN2B-AS1 (ANRIL)* has been shown to promote proliferation *in vitro* studies by sponging miR-323 and activating BRI3 dependent p38-MAPK, AKT and WNT signaling [43], and in current analysis, it was found to be upregulated in group 4 patients compared to other MBs. *PVT1* is prevalently found fused to MYC and NDRG1 genes in group 3 tumors, leading to oncogenic transformation of these genes [44]. *lnc-IRX3-80 (CRNDE)* was also reported as an oncogenic lncRNA *in vitro* and *in vivo* studies [45]. Both *PVT1* and *lnc-IRX3-80* were upregulated in WNT and SHH MBs in our current analysis. *Lnc-FAM84B-15 (CCAT1),* which was found upregulated in WNT and group 3 MBs, has also been shown to be involved in promoting tumor proliferation and metastasis by activating MAPK pathway [46]. *MIR100HG (lnc-NeD125)* has been shown to be overexpressed in group 4 MBs, again acting as an miRNA sponge for miR-19a-3p, miR-19b-3p and miR-106a-5p, exerting an oncogenic function by de-repressing cell cycle target genes [47]. *MIR100HG* is also oncogenic in gastric cancer [48], breast cancer [49], and leukemia [50]. In our analysis, only *MIR100HG (lnc-NeD125)* was selected in our diagnostic signature, being highly expressed in all MBs but group 3 (**Fig 2**). In addition, our 11-lncRNAs model could complement existing molecular and clinical-based diagnostic approaches, particularly for group 3 and group 4 MBs. Some of the identified signature lncRNAs are highly subgroup-specific, such as: *lnc-CCL2-2* (WNT), *lnc-ABCE1-5* (SHH), *USP2-AS1* (group 3), and *lnc-TBC1D16-3* (group 4). Mutual information-based network analysis also identified putative interacting transcription factors involved in medulloblastoma and other cancers (**Fig S5**); for example, FOXO1 [51, 52], OTX2 [53], NRL, CRX [54]  and TET3 [55].

Our 17-lncRNAs prognostic model represents another set of putative functionally important lncRNAs. Of the 17 lncRNAs, seven were associated with poor prognosis, including *H19 (***Table 1**). None of the candidate poor prognostic marker were specifically expressed in a particular subgroup of patients, suggesting independent prognostic value, although patients with high expression of the signature tended to be group 3 (**Fig S6**). *H19* is a well-studied oncogenic lncRNA in various cancer systems including glioblastoma, where it has been shown to be promote cellular proliferation and metastasis [56-59]. LncRNA *LINC01551* has been found to upregulate cellular proliferation and migration in non-CNS cancers such as hepatocellular carcinoma by interacting with the miR122-ADAM10 axis [60]. *LINC00336* promoted lung cancer progression by inhibiting regulated cell death by knocking down miR-6852 function [61].

Overall, our analysis proposes new lncRNAs candidates in MB with functional, diagnostic, and prognostic significance that warrant further investigation and validation. This is the first global analysis of lncRNAs in MB that will provide an invaluable resource for those working in the field to prioritize for further study.

## Acknowledgements

This work was supported by National Institutes of Health grants NCI 5P30CA030199 (SBP), P30 CA006973 (JHU SKCCC) and Florida Department of Health, Bankhead-Coley Cancer Research Program 5BC08 to R.J.P.

## Competing interests

The authors declare that they have no competing interests.

## References

1. Ostrom QT, Gittleman H, Liao P, Vecchione-Koval T, Wolinsky Y, Kruchko C, et al. CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010-2014. Neuro-oncology. 2017;19(suppl_5):v1-v88. Epub 2017/11/09. doi: 10.1093/neuonc/nox158. PubMed PMID: 29117289; PubMed Central PMCID: PMCPMC5693142.

2. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, et al. The 2007 WHO classification of tumours of the central nervous system. Acta neuropathologica. 2007;114(2):97-109. Epub 2007/07/10. doi: 10.1007/s00401-007-0243-4. PubMed PMID: 17618441; PubMed Central PMCID: PMCPMC1929165.

3. Diamandis P, Aldape K. World Health Organization 2016 Classification of Central Nervous System Tumors. Neurologic clinics. 2018;36(3):439-47. Epub 2018/08/04. doi: 10.1016/j.ncl.2018.04.003. PubMed PMID: 30072064.

4. Northcott PA, Robinson GW, Kratz CP, Mabbott DJ, Pomeroy SL, Clifford SC, et al. Medulloblastoma. Nature reviews Disease primers. 2019;5(1):11. Epub 2019/02/16. doi: 10.1038/s41572-019-0063-6. PubMed PMID: 30765705.

5. Millard NE, De Braganca KC. Medulloblastoma. Journal of child neurology. 2016;31(12):1341-53. Epub 2015/09/04. doi: 10.1177/0883073815600866. PubMed PMID: 26336203; PubMed Central PMCID: PMCPMC4995146.

6. Kondoff SI, Milev MD, Laleva LN, Tzekov CC, Kostadinova CP, Kirova-Nedyalkova GI, et al. A case of early extraneural medulloblastoma metastases in a young adult. Asian journal of neurosurgery. 2015;10(4):331-3. Epub 2015/10/02. doi: 10.4103/1793-5482.162723. PubMed PMID: 26425169; PubMed Central PMCID: PMCPMC4558816.

7. Dufour C, Beaugrand A, Pizer B, Micheli J, Aubelle MS, Fourcade A, et al. Metastatic Medulloblastoma in Childhood: Chang's Classification Revisited. International journal of surgical oncology. 2012;2012:245385. Epub 2012/02/09. doi: 10.1155/2012/245385. PubMed PMID: 22312539; PubMed Central PMCID: PMCPMC3265270.

8. Vladoiu MC, El-Hamamy I, Donovan LK, Farooq H, Holgado BL, Sundaravadanam Y, et al. Childhood cerebellar tumours mirror conserved fetal transcriptional programs. Nature. 2019;572(7767):67-73. Epub 2019/05/03. doi: 10.1038/s41586-019-1158-7. PubMed PMID: 31043743; PubMed Central PMCID: PMCPMC6675628.

9. Smee RI, Williams JR, De-Loyde KJ, Meagher NS, Cohn R. Medulloblastoma: progress over time. Journal of medical imaging and radiation oncology. 2012;56(2):227-34. Epub 2012/04/14. doi: 10.1111/j.1754-9485.2012.02349.x. PubMed PMID: 22498198.

10. De Braganca KC, Packer RJ. Treatment Options for Medulloblastoma and CNS Primitive Neuroectodermal Tumor (PNET). Current treatment options in neurology. 2013;15(5):593-606.

Epub 2013/08/28. doi: 10.1007/s11940-013-0255-4. PubMed PMID: 23979905; PubMed Central PMCID: PMCPMC5026188.

11. Wang J, Garancher A, Ramaswamy V, Wechsler-Reya RJ. Medulloblastoma: From Molecular Subgroups to Molecular Targeted Therapies. Annu Rev Neurosci. 2018;41:207-32. Epub 2018/04/12. doi: 10.1146/annurev-neuro-070815-013838. PubMed PMID: 29641939.

12. Palmer SL, Reddick WE, Gajjar A. Understanding the cognitive impact on children who are treated for medulloblastoma. Journal of pediatric psychology. 2007;32(9):1040-9. Epub 2007/03/03. doi: 10.1093/jpepsy/jsl056. PubMed PMID: 17329318.

13. Martin AM, Raabe E, Eberhart C, Cohen KJ. Management of pediatric and adult patients with medulloblastoma. Current treatment options in oncology. 2014;15(4):581-94. Epub 2014/09/10. doi: 10.1007/s11864-014-0306-4. PubMed PMID: 25194927; PubMed Central PMCID: PMCPMC4216607.

14. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nature genetics. 2015;47(3):199-208. Epub 2015/01/20. doi: 10.1038/ng.3192. PubMed PMID: 25599403; PubMed Central PMCID: PMCPMC4417758.

15. Sahakyan A, Yang Y, Plath K. The Role of Xist in X-Chromosome Dosage Compensation. Trends in cell biology. 2018;28(12):999-1013. Epub 2018/06/19. doi: 10.1016/j.tcb.2018.05.005. PubMed PMID: 29910081; PubMed Central PMCID: PMCPMC6249047.

16. Trimarchi T, Bilal E, Ntziachristos P, Fabbri G, Dalla-Favera R, Tsirigos A, et al. Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute

leukemia. Cell. 2014;158(3):593-606. Epub 2014/08/02. doi: 10.1016/j.cell.2014.05.049. PubMed PMID: 25083870; PubMed Central PMCID: PMCPMC4131209.

17. Shan Y, Ma J, Pan Y, Hu J, Liu B, Jia L. LncRNA SNHG7 sponges miR-216b to promote proliferation and liver metastasis of colorectal cancer through upregulating GALNT1. Cell death & disease. 2018;9(7):722. Epub 2018/06/20. doi: 10.1038/s41419-018-0759-7. PubMed PMID: 29915311; PubMed Central PMCID: PMCPMC6006356.

18. Long Y, Wang X, Youmans DT, Cech TR. How do lncRNAs regulate transcription? Science advances. 2017;3(9):eaao2110. Epub 2017/09/30. doi: 10.1126/sciadv.aao2110. PubMed PMID: 28959731; PubMed Central PMCID: PMCPMC5617379.

19. Huarte M. The emerging role of lncRNAs in cancer. Nature medicine. 2015;21(11):1253-61. Epub 2015/11/06. doi: 10.1038/nm.3981. PubMed PMID: 26540387.

20. Pop S, Enciu AM, Necula LG, Tanase C. Long non-coding RNAs in brain tumours: Focus on recent epigenetic findings in glioma. Journal of cellular and molecular medicine. 2018;22(10):4597-610. Epub 2018/08/18. doi: 10.1111/jcmm.13781. PubMed PMID: 30117678; PubMed Central PMCID: PMCPMC6156469.

21. Pandey GK, Kanduri C. Long noncoding RNAs and neuroblastoma. Oncotarget. 2015;6(21):18265-75. Epub 2015/06/19. doi: 10.18632/oncotarget.4251. PubMed PMID: 26087192; PubMed Central PMCID: PMCPMC4621889.

22. Lin CY, Erkek S, Tong Y, Yin L, Federation AJ, Zapatka M, et al. Active medulloblastoma enhancers reveal subgroup-specific cellular origins. Nature. 2016;530(7588):57-62. Epub 2016/01/28. doi: 10.1038/nature16546. PubMed PMID: 26814967; PubMed Central PMCID: PMCPMC5168934.

23. Cavalli FMG, Remke M, Rampasek L, Peacock J, Shih DJH, Luu B, et al. Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell. 2017;31(6):737-54.e6. Epub 2017/06/14. doi: 10.1016/j.ccell.2017.05.005. PubMed PMID: 28609654; PubMed Central PMCID: PMCPMC6163053.

24. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550. Epub 2014/12/18. doi: 10.1186/s13059-014-0550-8. PubMed PMID: 25516281; PubMed Central PMCID: PMCPMC4302049.

25. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics (Oxford, England). 2010;26(12):1572-3. Epub 2010/04/30. doi: 10.1093/bioinformatics/btq170. PubMed PMID: 20427518; PubMed Central PMCID: PMCPMC2881355.

26. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008;9:559. Epub 2008/12/31. doi: 10.1186/1471-2105-9-559. PubMed PMID: 19114008; PubMed Central PMCID: PMCPMC2631488.

27. Mehrian-Shai R, Chen CD, Shi T, Horvath S, Nelson SF, Reichardt JK, et al. Insulin growth factor-binding protein 2 is a candidate biomarker for PTEN status and PI3K/Akt pathway activation in glioblastoma and prostate cancer. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(13):5563-8. Epub 2007/03/21. doi: 10.1073/pnas.0609139104. PubMed PMID: 17372210; PubMed Central PMCID: PMCPMC1838515.

28. Meyer PE, Lafitte F, Bontempi G. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC bioinformatics. 2008;9:461. Epub

2008/10/31. doi: 10.1186/1471-2105-9-461. PubMed PMID: 18959772; PubMed Central PMCID: PMCPMC2630331.

29. Tibshirani R. The lasso method for variable selection in the cox model. Statistics in Medicine. 1997;16(4):385-95. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3.

30. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigo R, Johnson R. Towards a complete map of the human long non-coding RNA transcriptome. Nature reviews Genetics. 2018;19(9):535-48. Epub 2018/05/26. doi: 10.1038/s41576-018-0017-y. PubMed PMID: 29795125; PubMed Central PMCID: PMCPMC6451964.

31. Kool M, Jones DT, Jager N, Northcott PA, Pugh TJ, Hovestadt V, et al. Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. Cancer Cell. 2014;25(3):393-405. Epub 2014/03/22. doi: 10.1016/j.ccr.2014.02.004. PubMed PMID: 24651015; PubMed Central PMCID: PMCPMC4493053.

32. Bhan A, Soleimani M, Mandal SS. Long Noncoding RNA and Cancer: A New Paradigm. Cancer research. 2017;77(15):3965-81. Epub 2017/07/14. doi: 10.1158/0008-5472.can-16-2634. PubMed PMID: 28701486.

33. Arriaga-Canon C, De La Rosa-Velazquez IA, Gonzalez-Barrios R, Montiel-Manriquez R, Oliva-Rico D, Jimenez-Trejo F, et al. The use of long non-coding RNAs as prognostic biomarkers and therapeutic targets in prostate cancer. Oncotarget. 2018;9(29):20872-90. Epub 2018/05/15. doi: 10.18632/oncotarget.25038. PubMed PMID: 29755696; PubMed Central PMCID: PMCPMC5945524.

34. Roy S, Trautwein C, Luedde T, Roderburg C. A General Overview on Non-coding RNA-Based Diagnostic and Therapeutic Approaches for Liver Diseases. Frontiers in pharmacology.

2018;9:805. Epub 2018/08/31. doi: 10.3389/fphar.2018.00805. PubMed PMID: 30158867; PubMed Central PMCID: PMCPMC6104154.

35. Qi P, Du X. The long non-coding RNAs, a new cancer diagnostic and therapeutic gold mine. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc. 2013;26(2):155-65. Epub 2012/09/22. doi: 10.1038/modpathol.2012.160. PubMed PMID: 22996375.

36. Sarfi M, Abbastabar M, Khalili E. Long noncoding RNAs biomarker-based cancer assessment. Journal of cellular physiology. 2019. Epub 2019/03/06. doi: 10.1002/jcp.28417. PubMed PMID: 30835829.

37. Reon BJ, Anaya J, Zhang Y, Mandell J, Purow B, Abounader R, et al. Expression of lncRNAs in Low-Grade Gliomas and Glioblastoma Multiforme: An In Silico Analysis. PLoS medicine. 2016;13(12):e1002192. Epub 2016/12/07. doi: 10.1371/journal.pmed.1002192. PubMed PMID: 27923049; PubMed Central PMCID: PMCPMC5140055.

38. Liang R, Zhi Y, Zheng G, Zhang B, Zhu H, Wang M. Analysis of long non-coding RNAs in glioblastoma for prognosis prediction using weighted gene co-expression network analysis, Cox regression, and L1-LASSO penalization. OncoTargets and therapy. 2019;12:157-68. Epub 2019/01/08. doi: 10.2147/ott.s171957. PubMed PMID: 30613154; PubMed Central PMCID: PMCPMC6306053.

39. Zeng T, Li L, Zhou Y, Gao L. Exploring Long Noncoding RNAs in Glioblastoma: Regulatory Mechanisms and Clinical Potentials. International journal of genomics. 2018;2018:2895958. Epub 2018/08/18. doi: 10.1155/2018/2895958. PubMed PMID: 30116729; PubMed Central PMCID: PMCPMC6079499.

40. Li J, Zhu Y, Wang H, Ji X. Targeting Long Noncoding RNA in Glioma: A Pathway Perspective. Molecular therapy Nucleic acids. 2018;13:431-41. Epub 2018/11/06. doi: 10.1016/j.omtn.2018.09.023. PubMed PMID: 30388617; PubMed Central PMCID: PMCPMC6202792.

41. Northcott PA, Buchhalter I, Morrissy AS, Hovestadt V, Weischenfeldt J, Ehrenberger T, et al. The whole-genome landscape of medulloblastoma subtypes. Nature. 2017;547(7663):311-7. Epub 2017/07/21. doi: 10.1038/nature22973. PubMed PMID: 28726821; PubMed Central PMCID: PMCPMC5905700.

42. Zhang Y, Wang T, Wang S, Xiong Y, Zhang R, Zhang X, et al. Nkx2-2as Suppression Contributes to the Pathogenesis of Sonic Hedgehog Medulloblastoma. Cancer research. 2018;78(4):962-73. Epub 2017/12/13. doi: 10.1158/0008-5472.Can-17-1631. PubMed PMID: 29229597.

43. Zhang H, Wang X, Chen X. Potential Role of Long Non-Coding RNA ANRIL in Pediatric Medulloblastoma Through Promotion on Proliferation and Migration by Targeting miR-323. Journal of cellular biochemistry. 2017;118(12):4735-44. Epub 2017/05/18. doi: 10.1002/jcb.26141. PubMed PMID: 28513871.

44. Northcott PA, Shih DJ, Peacock J, Garzia L, Morrissy AS, Zichner T, et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. Nature. 2012;488(7409):49-56. Epub 2012/07/27. doi: 10.1038/nature11327. PubMed PMID: 22832581; PubMed Central PMCID: PMCPMC3683624.

45. Song H, Han LM, Gao Q, Sun Y. Long non-coding RNA CRNDE promotes tumor growth in medulloblastoma. European review for medical and pharmacological sciences. 2016;20(12):2588-97. Epub 2016/07/08. PubMed PMID: 27383309.

46. Gao R, Zhang R, Zhang C, Zhao L, Zhang Y. Long noncoding RNA CCAT1 promotes cell proliferation and metastasis in human medulloblastoma via MAPK pathway. Tumori. 2018;104(1):43-50. Epub 2017/08/05. doi: 10.5301/tj.5000662. PubMed PMID: 28777430.

47. Laneve P, Po A, Favia A, Legnini I, Alfano V, Rea J, et al. The long noncoding RNA linc-NeD125 controls the expression of medulloblastoma driver genes by microRNA sponge activity. Oncotarget. 2017;8(19):31003-15. Epub 2017/04/19. doi: 10.18632/oncotarget.16049. PubMed PMID: 28415684; PubMed Central PMCID: PMCPMC5458184.

48. Li J, Xu Q, Wang W, Sun S. MIR100HG: a credible prognostic biomarker and an oncogenic lncRNA in gastric cancer. Bioscience reports. 2019;39(4). Epub 2019/03/20. doi: 10.1042/bsr20190171. PubMed PMID: 30886062; PubMed Central PMCID: PMCPMC6449568.

49. Wang S, Ke H, Zhang H, Ma Y, Ao L, Zou L, et al. LncRNA MIR100HG promotes cell proliferation in triple-negative breast cancer through triplex formation with p27 loci. Cell death & disease. 2018;9(8):805. Epub 2018/07/26. doi: 10.1038/s41419-018-0869-2. PubMed PMID: 30042378; PubMed Central PMCID: PMCPMC6057987.

50. Emmrich S, Streltsov A, Schmidt F, Thangapandi VR, Reinhardt D, Klusmann JH. LincRNAs MONC and MIR100HG act as oncogenes in acute megakaryoblastic leukemia. Molecular cancer. 2014;13:171. Epub 2014/07/17. doi: 10.1186/1476-4598-13-171. PubMed PMID: 25027842; PubMed Central PMCID: PMCPMC4118279.

51. Srivastava VK, Yasruel Z, Nalbantoglu J. Impaired medulloblastoma cell survival following activation of the FOXO1 transcription factor. International journal of oncology. 2009;35(5):1045-51. Epub 2009/09/30. doi: 10.3892/ijo_00000419. PubMed PMID: 19787258.

52. Pei Y, Liu KW, Wang J, Garancher A, Tao R, Esparza LA, et al. HDAC and PI3K Antagonists Cooperate to Inhibit Growth of MYC-Driven Medulloblastoma. Cancer cell.

2016;29(3):311-23. Epub 2016/03/16. doi: 10.1016/j.ccell.2016.02.011. PubMed PMID: 26977882; PubMed Central PMCID: PMCPMC4794752.

53. Lu Y, Labak CM, Jain N, Purvis IJ, Guda MR, Bach SE, et al. OTX2 expression contributes to proliferation and progression in Myc-amplified medulloblastoma. American journal of cancer research. 2017;7(3):647-56. Epub 2017/04/13. PubMed PMID: 28401018; PubMed Central PMCID: PMCPMC5385649.

54. Garancher A, Lin CY, Morabito M, Richer W, Rocques N, Larcher M, et al. NRL and CRX Define Photoreceptor Identity and Reveal Subgroup-Specific Dependencies in Medulloblastoma. Cancer cell. 2018;33(3):435-49.e6. Epub 2018/03/14. doi: 10.1016/j.ccell.2018.02.006. PubMed PMID: 29533784; PubMed Central PMCID: PMCPMC6368680.

55. Bezerra Salomao K, Cruzeiro GAV, Bonfim-Silva R, Geron L, Ramalho F, Pinto Saggioro F, et al. Reduced hydroxymethylation characterizes medulloblastoma while TET and IDH genes are differentially expressed within molecular subgroups. Journal of neuro-oncology. 2018;139(1):33-42. Epub 2018/03/28. doi: 10.1007/s11060-018-2845-1. PubMed PMID: 29582271.

56. Jiang X, Yan Y, Hu M, Chen X, Wang Y, Dai Y, et al. Increased level of H19 long noncoding RNA promotes invasion, angiogenesis, and stemness of glioblastoma cells. Journal of neurosurgery. 2016;124(1):129-36. Epub 2015/08/15. doi: 10.3171/2014.12.Jns1426. PubMed PMID: 26274999.

57. Fazi B, Garbo S, Toschi N, Mangiola A, Lombari M, Sicari D, et al. The lncRNA H19 positively affects the tumorigenic properties of glioblastoma cells and contributes to NKD1 repression through the recruitment of EZH2 on its promoter. Oncotarget. 2018;9(21):15512-25.

Epub 2018/04/13. doi: 10.18632/oncotarget.24496. PubMed PMID: 29643989; PubMed Central PMCID: PMCPMC5884644.

58. Raveh E, Matouk IJ, Gilon M, Hochberg A. The H19 Long non-coding RNA in cancer initiation, progression and metastasis - a proposed unifying theory. Molecular cancer. 2015;14:184. Epub 2015/11/06. doi: 10.1186/s12943-015-0458-2. PubMed PMID: 26536864; PubMed Central PMCID: PMCPMC4632688.

59. Zhou W, Ye XL, Xu J, Cao MG, Fang ZY, Li LY, et al. The lncRNA H19 mediates breast cancer cell plasticity during EMT and MET plasticity by differentially sponging miR-200b/c and let-7b. Science signaling. 2017;10(483). Epub 2017/06/15. doi: 10.1126/scisignal.aak9557. PubMed PMID: 28611183.

60. Gao J, Yin X, Yu X, Dai C, Zhou F. Long noncoding LINC01551 promotes hepatocellular carcinoma cell proliferation, migration, and invasion by acting as a competing endogenous RNA of microRNA-122-5p to regulate ADAM10 expression. Journal of cellular biochemistry. 2019. Epub 2019/07/05. doi: 10.1002/jcb.28549. PubMed PMID: 31270840.

61. Wang M, Mao C, Ouyang L, Liu Y, Lai W, Liu N, et al. Long noncoding RNA LINC00336 inhibits ferroptosis in lung cancer by functioning as a competing endogenous RNA. Cell death and differentiation. 2019. Epub 2019/02/23. doi: 10.1038/s41418-019-0304-y. PubMed PMID: 30787392.

# SUPPLEMENTAL INFORMATION

## Supplementary Tables

**Table S1.** List of 175 medulloblastoma patient samples and clinical information.

**Table S2.** List of subgroup specific lncRNA cohorts identified by WGCNA analysis.

**Table S3.** List of differentially expressed lncRNA between group 3 and group 4 patients.

## Supplementary Figures

**Fig S1. Medulloblastoma patients can be optimally subgrouped into 4 clusters based on long non-coding RNA expression.** (A) Heatmap depicting stability of k-means clusters, 2 to 6 (k=4, in Fig 1C), based on consensus clustering. Color range depicts samples never clustered together (0, blue) to always clustered together (1, red). (B) Colored line depicting relationship between cumulative distribution function (CDF) and consensus index for each of the k-means values 2 to 6. (C) Graph showing relative change in area under the CDF curve (in B) comparing k to k-1 for k from 3 to 6. For k=2, the value is total area under the CDF curve in B. (D) Cluster consensus plot showing mean of pairwise consensus value for all cluster members. For k=4, the graph shows that each of the obtained clusters are of similar stability.

**Fig S2. Random forest-based approach identifies a 14-PCG model to classify medulloblastoma subgroups.** (A) Schematic depiction of the modeling process. First, an 11-PCG model distinguishing SHH, group 3, and group 4 patients was obtained using a 60%-40% training-tuning partition. Then, a 3-PCG model distinguishing WNT from the rest of the group was obtained by combining all WNT samples with a 60%-40% partition of the other subgroup patients in a training-tuning model. (B) MB patient subgroups as identified from the random forest model using only 14-PCG expression as variables. Dendrogram represents hierarchical clustering of dissimilarity values obtained from random forest-based

classification. Bottom color bars represent known clinical groupings (blue=WNT, green=SHH, black=group 3, red=group 4). The obtained clustering and misclassification are similar to that in Fig 2B. (C) Boxplot showing distribution of normalized expression values of identified 14-PCGs in each patient subgroup. Purple dots represent normalized expression value for a patient. D) tSNE plot showing clustering of patients into four subgroups based on normalized expression level of identified 14-PCGs (blue=WNT, green=SHH, black=group 3, red=group 4). The obtained clustering is similar to that in Fig 3D, except for sample ICGC_MB23 that cluster with SHH subgroup. (E) ROC analysis of linear model based on normalized expression of identified 14-PCGs comparing one versus all (rest) classifications for each of the subgroups. The AUC values for 11-lncRNAs model and 14-PCGs model are comparable.

**Fig S3. 14-PCG model classifies independent medulloblastoma patient samples with high efficiency.** (A) MB patient subgroups as identified in the Cavalli17 dataset on applying a random forest model using only 14-PCG expression as variables. Dendrogram represents hierarchical clustering of dissimilarity values obtained from random forest-based classification. Bottom color bars represent known clinical grouping (blue=WNT, green=SHH, black=group 3, red=group 4). (B) Boxplot showing distribution of normalized expression values of identified 14-PCGs in each patient subgroup. Purple dots represent normalized expression value for a patient. The normalized expression distribution for the 14-PCGs are similar to that in RNA-seq analysis (Fig S2C).  (D) tSNE plot showing clustering of patients into four subgroup based on normalized expression level of identified 14-PCGs (blue=WNT, green=SHH, black=group 3, red=group 4). (D) ROC analysis of linear model based on normalized expression of identified 14-PCGs comparing one versus all (rest) classifications for each of the subgroups.

**Fig S4. Random forest-based approach identifies an 8-lncRNA model to classify group 3 and group 4 patents.** (A) MB patient subgroups as identified from a random forest model using 8-lncRNA expression as variables. Dendrogram represents hierarchical clustering of dissimilarity values obtained from random forest-based classification. Bottom color bars represent known clinical groupings

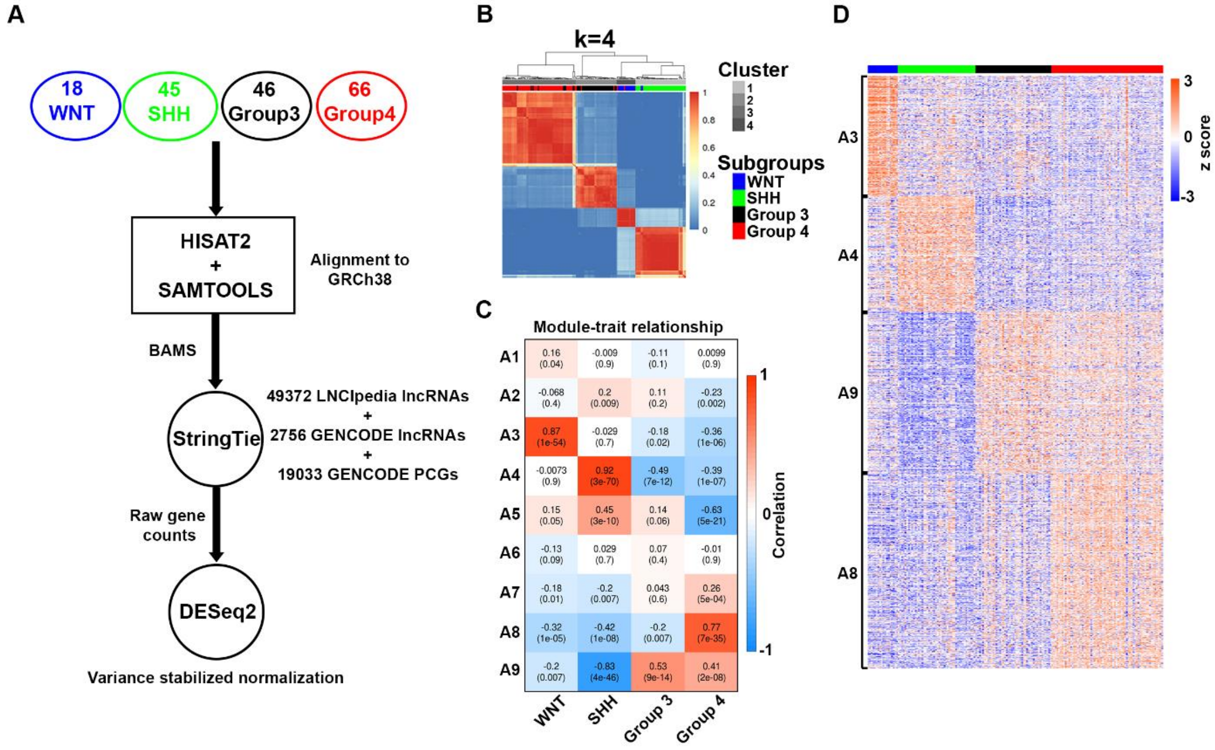
(black=group 3, red=group 4). (B) Boxplot showing distribution of normalized expression values of identified 8-lncRNAs in group 3 and group 4 MBs. Purple dots represent normalized expression value for a patient. (D) tSNE plot showing clustering of group 3 and group 4 patients into two heterogeneous clusters based on normalized expression levels of identified 8-lncRNAs (black=group 3, red=group 4). (E) ROC analysis of linear model based on normalized expression of identified 8-lncRNAs comparing group 3 versus group 4 samples.

**Fig S5. Putative interaction network of identified 11-lncRNAs with transcription factors.** Interaction network between the 11-lncRNAs and transcription factors based on a consensus of mutual information analysis using CLR, arcane, and mrnet-based approaches.

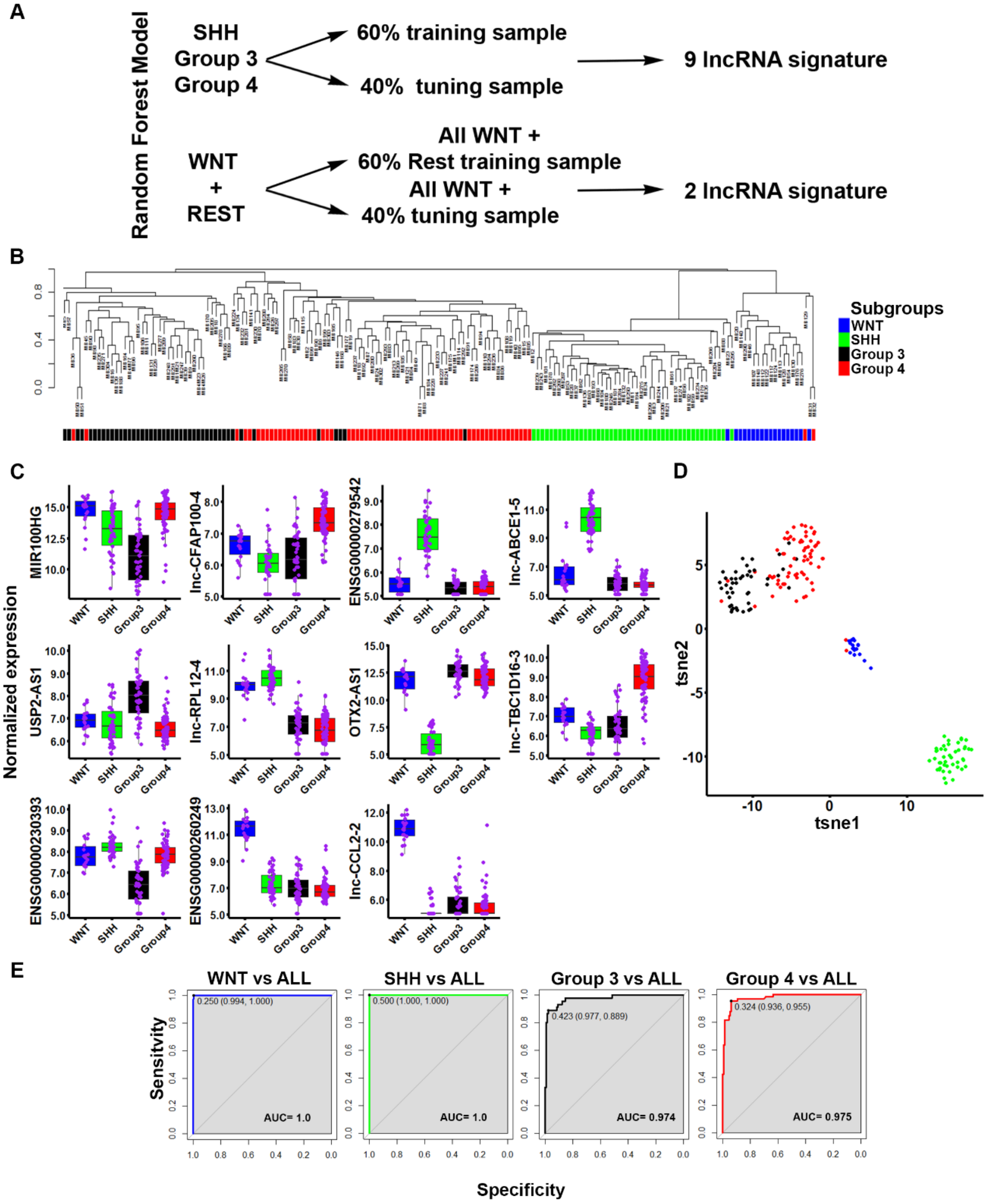
**Fig S6. Expression distribution of identified bad prognostic markers**. Box plot representing expression distribution of candidate bad prognostic markers in (A) MAGIC microarray dataset and (B) ICGC RNA-seq dataset in each subgroup. None of the identified lncRNAs is expressed specifically in a subgroup; however, patients expressing high levels were primarily in group 3.

# FIGURES

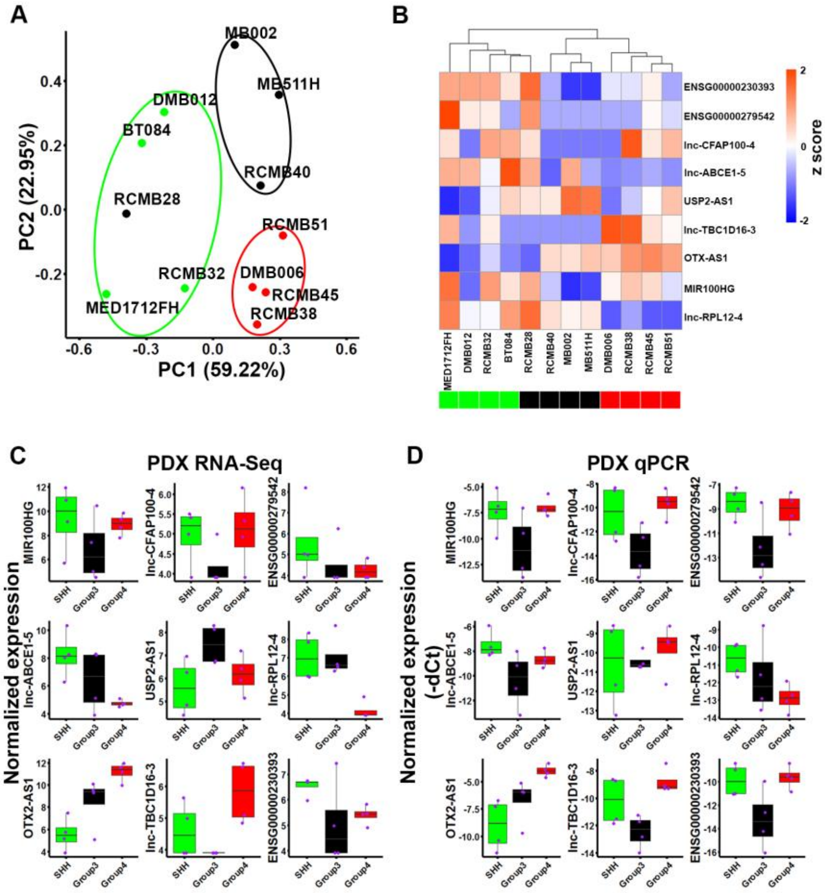
## Figure 1



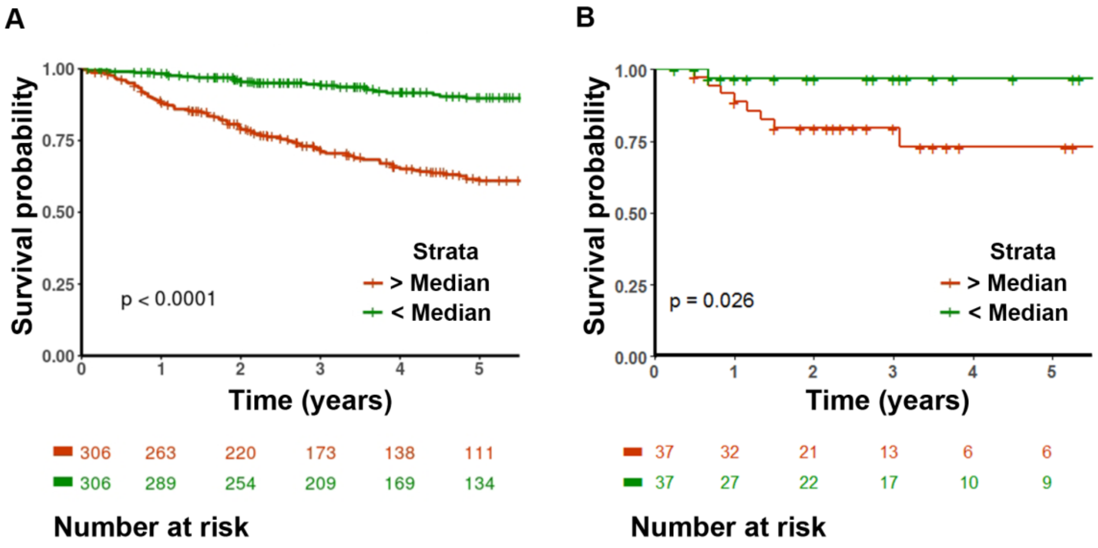
## Figure 2


A
Random Forest Model
SHH
Group 3
Group 4
60% training sample
40% tuning sample
9 lncRNA signature
WNT
+
REST
All WNT +
60% Rest training sample
All WNT +
40% tuning sample
2 lncRNA signature
B
0.0
0.4
0.8
Subgroups
WNT
SHH
Group 3
Group 4
C
Normalized expression
MIR100HG
lnc-CFAP100-4
ENSG00000279542
lnc-ABCE1-5
USP2-AS1
lnc-RPL12-4
OTX2-AS1
lnc-TBC1D16-3
ENSG00000230393
ENSG00000260249
lnc-CCL2-2
WNT
SHH
Group3
Group4
D
tsne2
tsne1
5
0
-5
-10
-10
0
10
E
WNT vs ALL
SHH vs ALL
Group 3 vs ALL
Group 4 vs ALL
0.250 (0.994, 1.000)
0.500 (1.000, 1.000)
0.423 (0.977, 0.889)
0.324 (0.936, 0.955)
AUC= 1.0
AUC= 1.0
AUC= 0.974
AUC= 0.975
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0

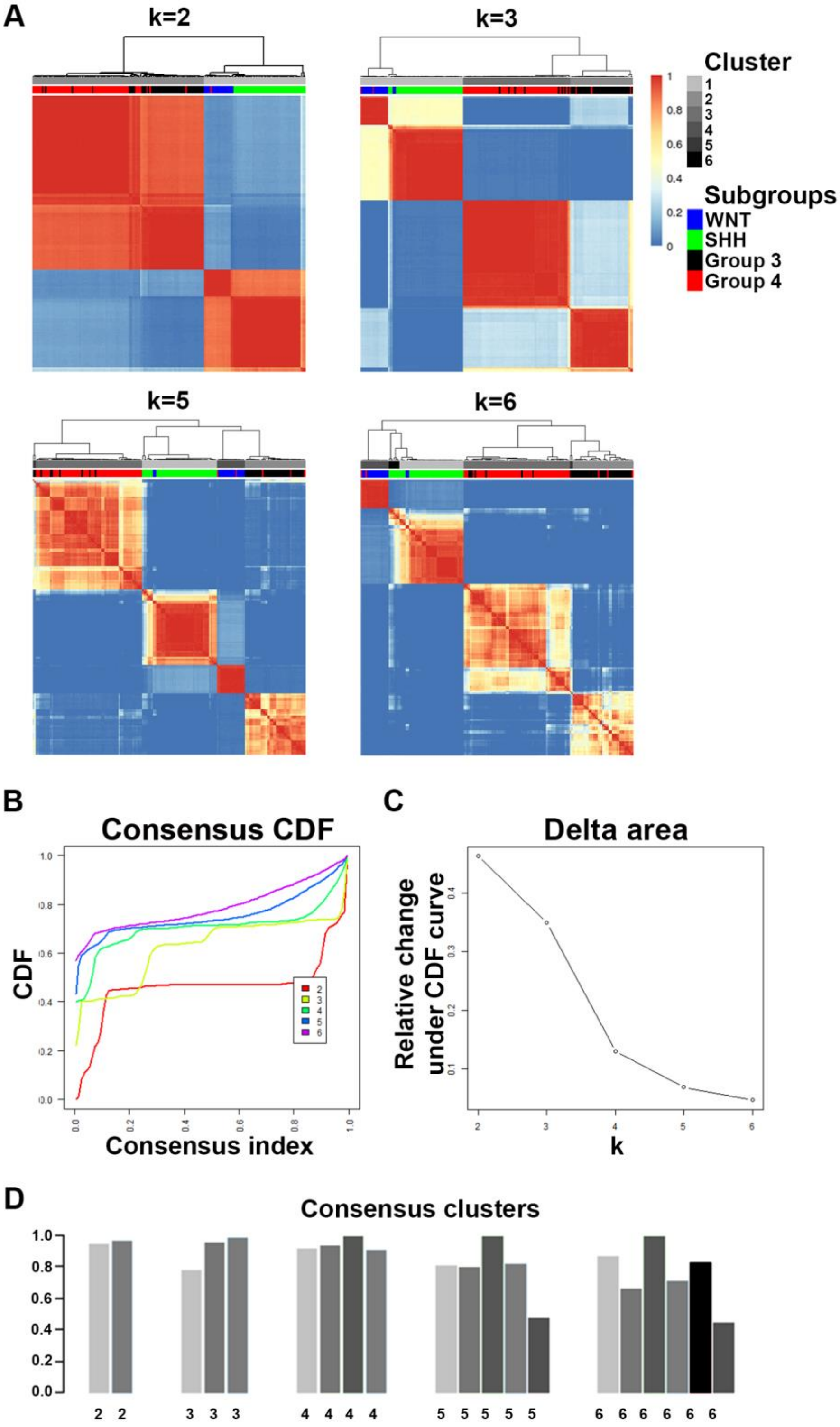
# Figure 3


A
PC1 (59.22%)
PC2 (22.95%)
MB002
MB511H
RCMB40
RCMB28
DMB012
BT084
RCMB32
MED1712FH
RCMB51
DMB006
RCMB45
RCMB38
B
ENSG00000230393
ENSG00000279542
lnc-CFAP100-4
lnc-ABCE1-5
USP2-AS1
lnc-TBC1D16-3
OTX-AS1
MIR100HG
lnc-RPL12-4
z score
MED1712FH
DMB012
RCMB32
BT084
RCMB28
RCMB40
MB002
MB511H
DMB006
RCMB38
RCMB45
RCMB51
C
PDX RNA-Seq
Normalized expression
D
PDX qPCR
Normalized expression (-dCt)
MIR100HG
lnc-CFAP100-4
ENSG00000279542
lnc-ABCE1-5
USP2-AS1
lnc-RPL12-4
OTX2-AS1
lnc-TBC1D16-3
ENSG00000230393
SHH
Group3
Group4

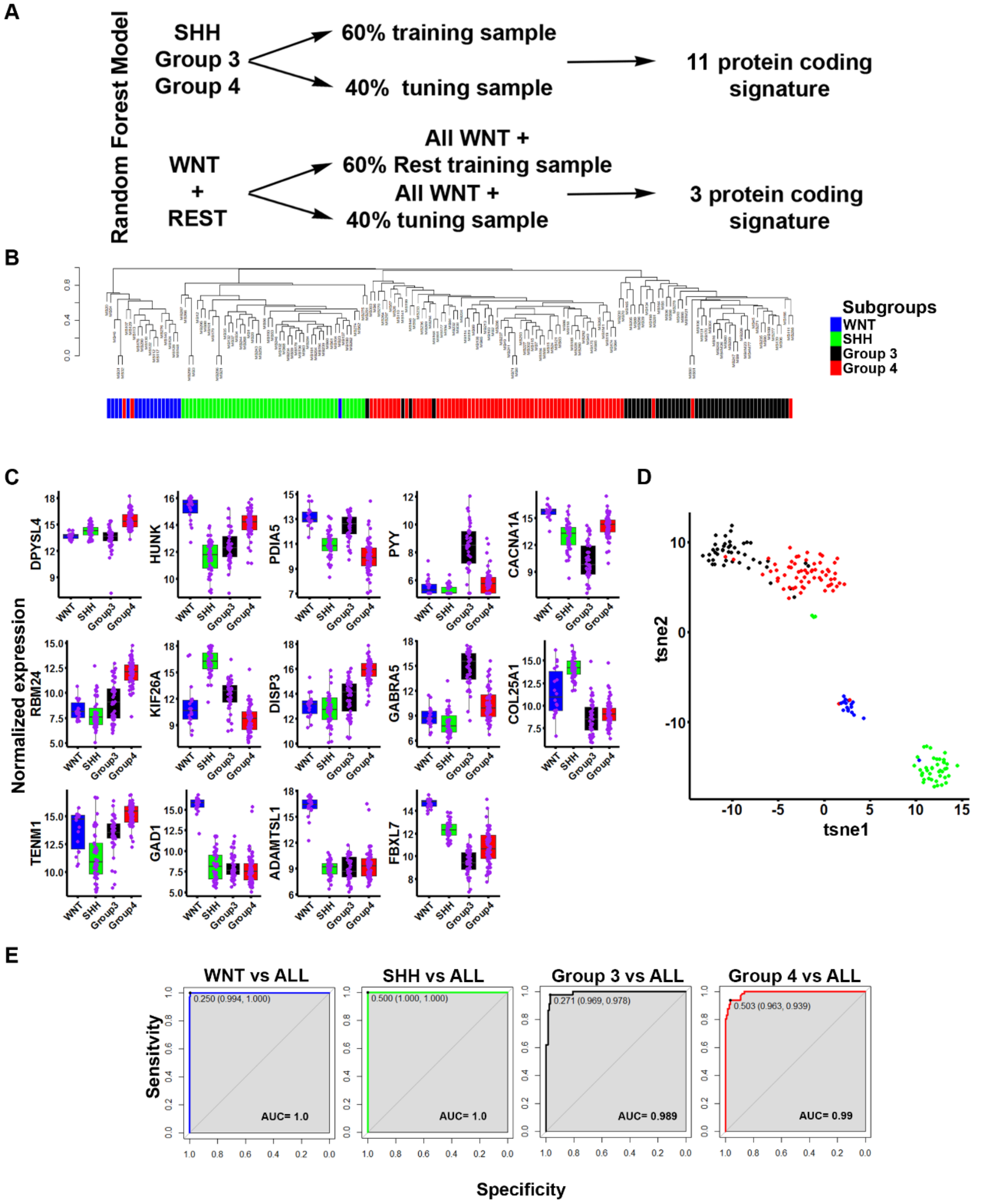
## Figure 4


A
Survival probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
Strata
> Median
< Median
0
1
2
3
4
5
Time (years)
306 263 220 173 138 111
306 289 254 209 169 134
Number at risk
B
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.026
Strata
> Median
< Median
0
1
2
3
4
5
Time (years)
37 32 21 13 6 6
37 27 22 17 10 9
Number at risk

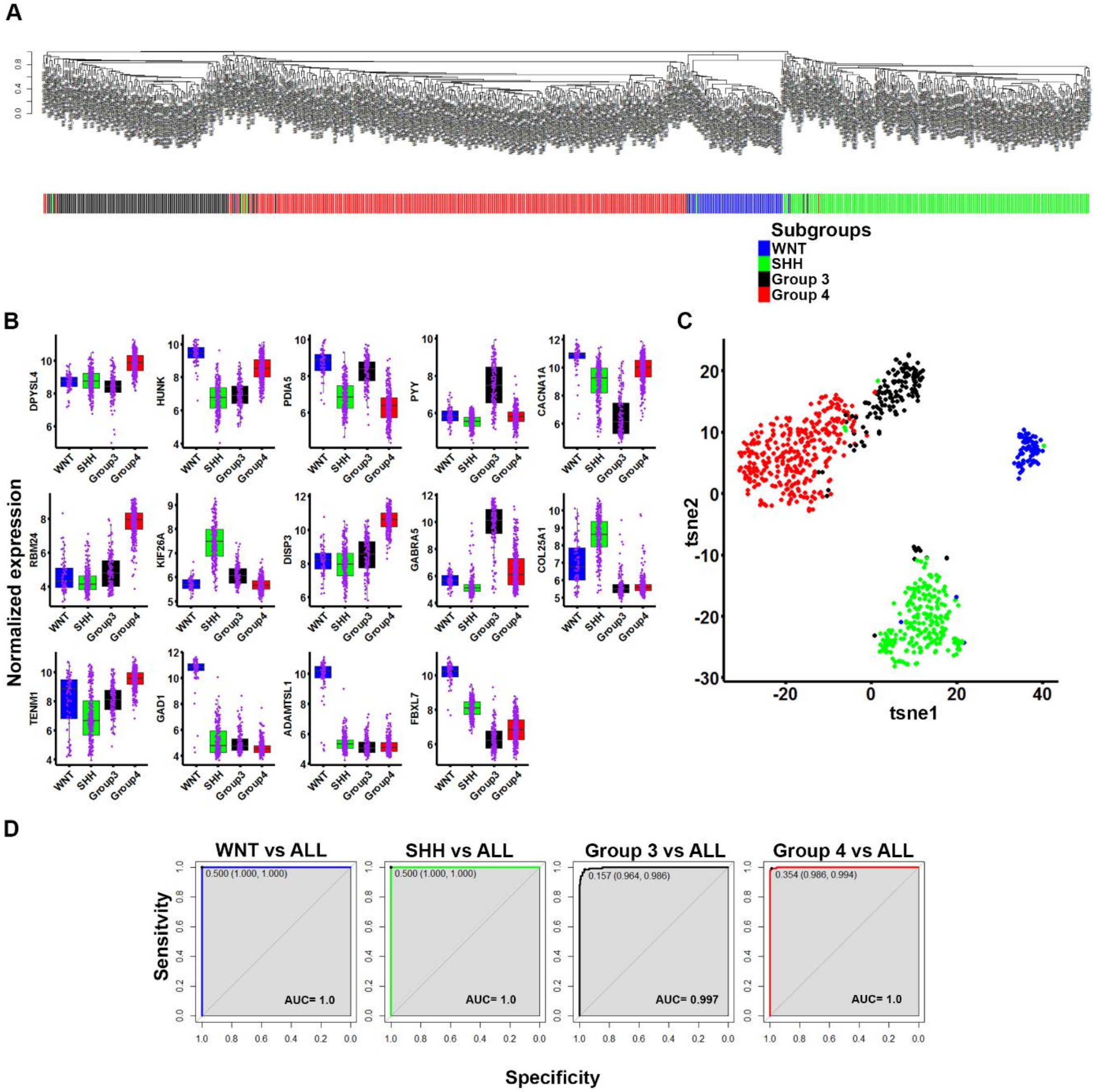
## Figure S1


A
k=2
k=3
Cluster
1
2
3
4
5
6
1
0.8
0.6
0.4
0.2
0
Subgroups
WNT
SHH
Group 3
Group 4
k=5
k=6
B
Consensus CDF
CDF
1.0
0.8
0.6
0.4
0.2
0.0
Consensus index
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6
C
Delta area
Relative change under CDF curve
0.4
0.3
0.2
0.1
k
2
3
4
5
6
D
Consensus clusters
1.0
0.8
0.6
0.4
0.2
0.0
2 2
3 3 3
4 4 4 4
5 5 5 5 5
6 6 6 6 6 6

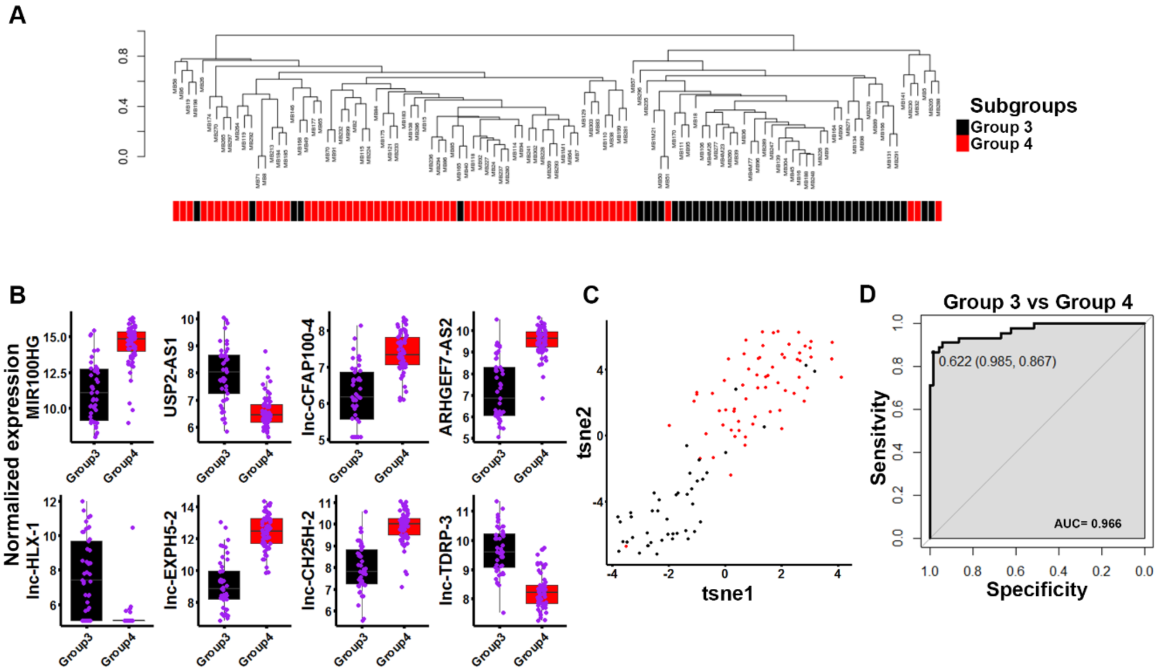
## Figure S2


A
Random Forest Model
SHH
Group 3
Group 4
60% training sample
40% tuning sample
11 protein coding signature
WNT
+
REST
All WNT +
60% Rest training sample
All WNT +
40% tuning sample
3 protein coding signature
B
0.8
0.4
0.0
Subgroups
WNT
SHH
Group 3
Group 4
C
Normalized expression
DPYSL4
HUNK
PDIA5
PYY
CACNA1A
RBM24
KIF26A
DISP3
GABRA5
COL25A1
TENM1
GAD1
ADAMTSL1
FBXL7
WNT
SHH
Group3
Group4
D
tsne2
tsne1
10
0
-10
-5
5
15
E
WNT vs ALL
SHH vs ALL
Group 3 vs ALL
Group 4 vs ALL
0.250 (0.994, 1.000)
0.500 (1.000, 1.000)
0.271 (0.969, 0.978)
0.503 (0.963, 0.939)
AUC= 1.0
AUC= 1.0
AUC= 0.989
AUC= 0.99
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0

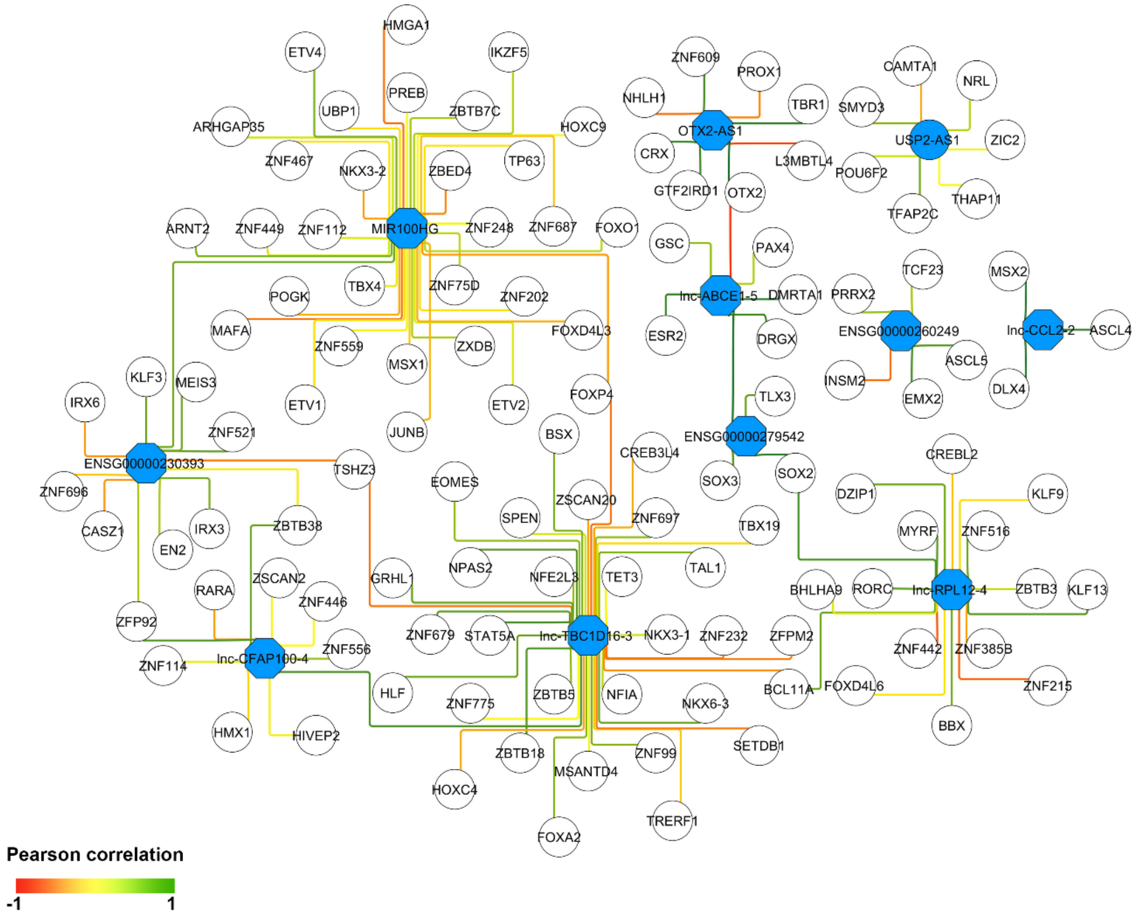
# Figure S3


A
0.0
0.4
0.8
Subgroups
WNT
SHH
Group 3
Group 4
B
Normalized expression
DPYSL4
HUNK
PDIA5
PYY
CACNA1A
RBM24
KIF26A
DISP3
GABRA5
COL25A1
TENM1
GAD1
ADAMTSL1
FBXL7
WNT
SHH
Group3
Group4
C
tsne2
tsne1
-20
0
20
40
20
10
0
-10
-20
-30
D
WNT vs ALL
SHH vs ALL
Group 3 vs ALL
Group 4 vs ALL
Sensitivity
Specificity
0.500 (1.000, 1.000)
0.500 (1.000, 1.000)
0.157 (0.964, 0.986)
0.354 (0.986, 0.994)
AUC= 1.0
AUC= 1.0
AUC= 0.997
AUC= 1.0
1.0
0.8
0.6
0.4
0.2
0.0

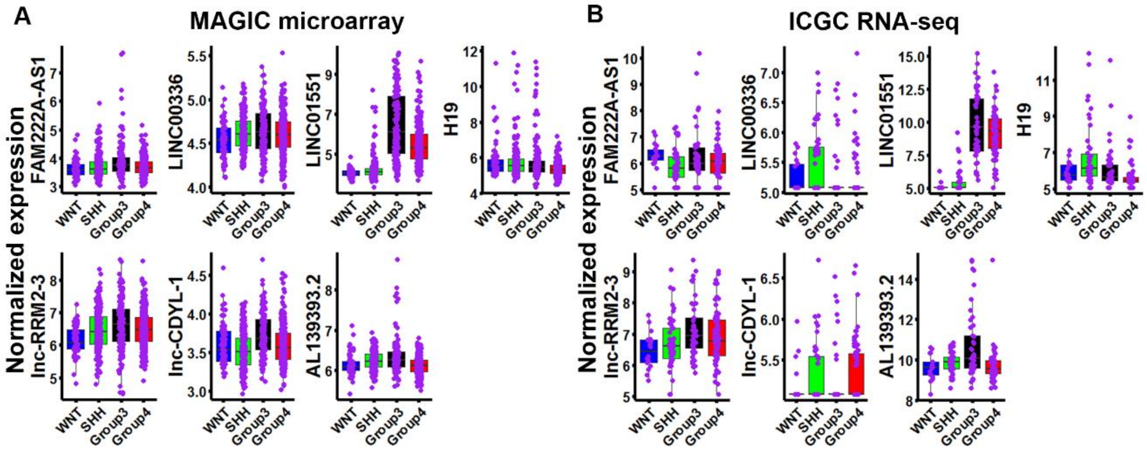
# Figure S4


A
0.0
0.4
0.8
Subgroups
Group 3
Group 4
B
Normalized expression
MIR100HG
USP2-AS1
lnc-CFAP100-4
ARHGEF7-AS2
lnc-HLX-1
lnc-EXPH5-2
lnc-CH25H-2
lnc-TDRP-3
Group3
Group4
C
tsne1
tsne2
D
Group 3 vs Group 4
Sensitivity
Specificity
0.622 (0.985, 0.867)
AUC= 0.966

## Figure S5


HMGA1
ETV4
IKZF5
PREB
UBP1
ZBTB7C
ARHGAP35
HOXC9
ZNF467
NKX3-2
ZBED4
TP63
ARNT2
ZNF449
ZNF112
MIR100HG
ZNF248
ZNF687
FOXO1
POGK
TBX4
ZNF75D
ZNF202
MAFA
FOXD4L3
ZNF559
ZXDB
MSX1
KLF3
MEIS3
IRX6
ETV1
ETV2
FOXP4
ZNF521
JUNB
BSX
ENSG00000230393
TSHZ3
ZNF696
CASZ1
IRX3
ZBTB38
EN2
EOMES
SPEN
ZSCAN20
ZNF697
TBX19
CREB3L4
RARA
ZSCAN2
ZNF446
GRHL1
NPAS2
NFE2L3
TET3
TAL1
ZFP92
ZNF679
STAT5A
lnc-TBC1D16-3
NKX3-1
ZNF232
ZFPM2
ZNF114
lnc-CFAP100-4
ZNF556
HLF
ZNF775
ZBTB5
NFIA
NKX6-3
BCL11A
FOXD4L6
HMX1
HIVEP2
ZBTB18
MSANTD4
ZNF99
SETDB1
HOXC4
TRERF1
FOXA2
ZNF609
PROX1
NHLH1
TBR1
OTX2-AS1
CRX
L3MBTL4
GTF2IRD1
OTX2
GSC
PAX4
lnc-ABCE1-5
DMRTA1
ESR2
DRGX
TLX3
ENSG00000279542
SOX3
SOX2
CAMTA1
NRL
SMYD3
USP2-AS1
ZIC2
POU6F2
THAP11
TFAP2C
TCF23
MSX2
PRRX2
ENSG00000260249
lnc-CCL2-2
ASCL4
ASCL5
INSM2
EMX2
DLX4
CREBL2
DZIP1
KLF9
MYRF
ZNF516
BHLHA9
RORC
lnc-RPL12-4
ZBTB3
KLF13
ZNF442
ZNF385B
ZNF215
BBX
Pearson correlation
-1
1

# Figure S6


A
MAGIC microarray
Normalized expression
FAM222A-AS1
LINC00336
LINC01551
H19
lnc-RRM2-3
lnc-CDYL-1
AL139393.2
WNT
SHH
Group3
Group4
B
ICGC RNA-seq
Normalized expression
FAM222A-AS1
LINC00336
LINC01551
H19
lnc-RRM2-3
lnc-CDYL-1
AL139393.2