Title: A prognostic signature for lower-grade gliomas based on expression of long noncoding RNAs

Authors and affiliations: **Manjari Kiran**, Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Pinn Hall 1232, Charlottesville, Virginia 22908

Ajay Chatrath, Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Pinn Hall 1232, Charlottesville, Virginia 22908

Xiwei Tang, Department of Statistics, University of Virginia, Charlottesville, VA 22904

Daniel Macrae Keenan, Department of Statistics, University of Virginia, Charlottesville, VA 22904

Anindya Dutta, Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Pinn Hall 1232, Charlottesville, Virginia 22908

Corresponding Author: Anindya Dutta, Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Pinn Hall 1232, Charlottesville, Virginia 22908, Email: ad8q@virginia.edu

Running title: A long non-coding RNAs based prognostic signature for glioma

Abbreviation: lncRNA: Long Non-Coding RNAs, WHO: World Health Organization, LGG: Lower Grade Gliomas, GBM: Glioblastoma Multiforme, CNS: Central Nervous System, TCGA: The Cancer Genome Atlas, CGGA: Chinese Glioma Genome Atlas, HR: Hazard Ratio, PFS: Progression Free Survival, IFNG: Interferon Gamma, Cindex: Concordance Index, AUC: Area Under Curve, ROC: Receiver Operating Characteristics

Abstract

Diffuse low-grade and intermediate-grade gliomas (together known as lower-grade gliomas, WHO grade II and III) develop in the supporting glial cells of brain and are the most common types of primary brain tumor. Despite a better prognosis for lower-grade gliomas, 70% of patients undergo high-grade transformation within 10 years, stressing the importance of better prognosis. Long non-coding RNAs (lncRNAs) are gaining attention as potential biomarkers for cancer diagnosis and prognosis. We have developed a computational model, UVA8, for prognosis of lower-grade gliomas by combining lncRNA expression, Cox regression and L1-LASSO penalization. The model was trained on a subset of patients in TCGA. Patients in TCGA, as well as a completely independent validation set (CGGA) could be dichotomized based on their risk score, a linear combination of the level of each prognostic lncRNA weighted by its multivariable cox regression coefficient. UVA8 is an independent predictor of survival and outperforms standard epidemiological approaches and previous published lncRNA-based predictors as a survival model. Guilt-by-association studies of the lncRNAs in UVA8, all of which predict good outcome, suggest they have a role in suppressing interferon stimulated response and epithelial to mesenchymal transition. The expression levels of 8 lncRNAs can be combined to produce a prognostic tool applicable to diverse populations of glioma patients. The 8 lncRNA (UVA8) based score can identify grade II and grade III glioma patients with poor outcome and thus identify patients who should receive more aggressive therapy at the outset.

Keywords: long non-coding RNAs, gliomas, gene expression profiling, prognosis

Introduction

Over the past decade, high-throughput RNA-seq technology discovered many novel transcriptional units, which were otherwise missed by probe design based transcriptome profiling. Among these transcriptional units were many long non-coding RNAs (lncRNA), which are transcripts longer than 200 bases with almost no protein-coding potential or open reading frames of <50 amino acids. These lncRNAs are numerous in cells [1], are highly regulated and are more cell-type specific than protein-coding genes [2]. LncRNAs are involved in a broad spectrum of function and recent studies suggest they have specific roles in different diseases like cancer (reviewed in [3, 4]).

Gliomas are the most common form of primary malignant brain tumor, which originate in the supporting glial cells in the brain, including astrocytes, oligodendrocytes and ependymal cells. Based on WHO 2016 grading system, gliomas are classified into lower-grade and much aggressive high-grade gliomas. Grade I is mostly benign, whereas diffuse low-grade and intermediate-grade gliomas make up the WHO grade II and III lesions. Grade IV gliomas include secondary glioblastomas (derived from lower grade gliomas) and primary glioblastoma multiforme (GBM). Surgical resection of tumor is the most common initial treatment for gliomas followed by radiation therapy and chemotherapy, which can increase survival to 12 months [5, 6]. Molecular markers like 1p/19q co-deletion, MGMT promoter methylation and mutation in IDH1 gene are strong predictors of survival for gliomas [7]. Lower-grade gliomas have a better prognosis than high-grade gliomas. Despite a better prognosis for lower-grade gliomas than the grade IV tumors, 70% of patients from the former group undergo high-grade transformation within 10 years.

LncRNAs are widely expressed in the central nervous system (CNS) and are involved in several pathways related to CNS development [8–13]. LncRNA BRN1B is one of the critical lncRNAs for brain development [13]. LncRNA Sox2OT plays an important role in determining neural fate [14] . Dysregulation of many lncRNAs like DGCR5, NRON, H19, DISC2 have been associated with different CNS diseases [15–18]. Previous studies have shown that specific lncRNA expression patterns are also associated with different histological subtypes and grade in gliomas [19, 20]. For example, expression of MALAT1, POU3F3 and H19 are highly correlated with glioma malignancy. More recently, lncRNAs are also found to be of prognostic significance suggesting their role in glioma malignancies and as a potential therapeutic target and biomarker [19, 20]. Li et al, 2014 revealed three molecular

subtypes of gliomas based on lncRNAs expression that has a strong correlation with patient's survival [21]. Furthermore, analysis on previously published microarray data has explored lncRNA-based signature as a prognostic marker in gliomas ([20, 22–25]).

Many studies have highlighted the power of gene expression profiles to predict tumor classification, patient outcome and tumor response to therapy. Differentially expressed genes in cancer patients versus normal individuals are often the starting set to predict prognostic signature associated with survival. This strategy suffers from false negatives and from the fact that differentially expressed genes might not be associated with differences in survival at all. Another limitation of this method is the requirement of perfect matched normal to identify differentially expressed genes. This creates a major hurdle in case of brain cancer where getting a perfect matched normal tissue is not trivial. While high-throughput technologies have facilitated the search of biomarkers through multivariate data analyses, there still remain challenges with respect to meaningful statistical and biological information. Firstly, most of the biological datasets suffer with multicollinearity: the influence of one gene on expression of other genes. Secondly, there are more features (genes) than observations (patients), which leads to overfitting by most of existing learning algorithms and results in poor performance of the model in prediction in an unseen testing dataset. Thus, a more robust machine learning approach is required to find genes as prognostic signature from a multi-dimensional multivariate gene-expression data. Regression models like lasso, ridge and elastic net are some widely used approaches to penalize the effect of multicollinearity and are well suited for constructing models when there are large numbers of features.

In the present study, we develop an lncRNA-based prognostic signature in combination with Cox regression and L1-LASSO regularization to model survival of grade II and grade III glioma patients. This is the first study that combined Cox and Lasso regularization to select lncRNAs that can predict survival in glioma patients. After controlling for covariates associated with glioma survival (age, grade, IDH1 mutation status), we selected 8 lncRNAs UVA8, to calculate a risk-score, which successfully divides patients into high-risk and low-risk groups in both TCGA (461 patients) and CGGA (274 patients) dataset. The risk score calculated by these 8 lncRNAs is an independent and better prognostic marker for grade II and grade III glioma patient survival. The guilt-by-association analysis of lncRNAs in UVA8 indicated their role in suppressing interferon signaling pathway and epithelial to

mesenchymal transition. Besides their use as a biomarker, these lncRNAs need to be studied in detail to determine how they affect patient outcome.

Materials and Methods

Patients and samples

Aligned bam files and clinical information for 512 LGG patients (grade II and III) were retrieved from The Cancer Genome Atlas (TCGA) data portal https://portal.gdc.cancer.gov/. The study is performed on 461 patients for which both RNAseq and survival information were available. Most samples in TCGA are collected from patients from the US and also from other countries including Canada, Russia, and Italy. This dataset being the largest and most updated glioma dataset is used as training dataset in the present study. The raw sequencing data for 274 glioma patients from Chinese Glioma Genome Atlas (CGGA) as independent cohort was downloaded using accession no. SRP027383 [26]. The survival information for these Chinese patients was downloaded from CGGA http://www.cgga.org.cn/. IDH1 mutation data for all the LGG patients were retrieved from Tier 3 TCGA data accessed from the Broad GDAC Firehose; https://gdac.broadinstitute.org.

RNASeq data quantification and analysis

The most recent version of Gencode (GENCODE v 26) GTF file available at the time of this study was used for the gene quantification [27]. Gene abundance in FPKM was obtained for 58219 genes with 15787 genes annotated as lncRNA in GENCODE v26 using Stringtie v1.3.3 [28] . Out of 15787 lncRNAs, 1289 lncRNAs with a median expression of 1 FPKM in 512 LGG patients were finally considered for the survival model.

Survival model selection process

The gene-expression data for lncRNAs was Z-score transformed to avoid systematic error across different experiments. We first randomly selected 60% of TCGA patients for training set and remaining 40% of TCGA patients for testing set. Since, clinical information like age, gender, tumor grade or IDH mutation status can have an effect on survival (**Figure S1**), we assessed the prognostic potential of each lncRNA by multivariate Cox-regression controlling the effects from these other variables. We used FDR corrected p-value cutoff of 0.05 obtained after log-

likelihood test comparing restricted (Age, Gender, tumor grade and IDH mutation status) with unrestricted (lncRNA expression, Age, Gender, tumor grade and IDH mutation status) model to identify the significant association of an lncRNA with survival. We used Cox-Proportional Hazards model based on L1 – penalized (LASSO) estimation to select the best model comprising a subset of prognostic lncRNA [29–31]. We used LASSO because it is suited for constructing models when there is a large number of correlated covariates [30].

Risk Score calculation

Risk score for each patient was established by including each of the selected genes weighted by their estimated regression coefficients in the multivariable Cox regression analysis as discussed in previous studies [32, 33].

UVA8 Risk score = (-0.378 x expression value of RP11-266K4.14) + (-0.301 x expression value of FLJ37035)+ (-0.280 x expression value of LINC01561) + (-0.368 x expression value of RP11-118K6.3) + (-0.369 x expression value of DGCR9) + (-0.299 x expression value of RP11-142A22.3) + (-0.434 x expression value of LINC00641) + (-0.543 x expression value of RP11-96H19.1).

Coefficients are median cox-coefficient (after lasso selection and multivariate cox-regression) for each of the 8 lncRNAs from the successful models (models which can stratify patients in testing set).

Statistical Analysis

R package glmnet was used to perform L1-penalized cox regression [34] . R package survival and survminer were used for survival data analysis and generating Kaplan–Meier plots. Different survival models were compared by time-dependent concordance index (Cindex) [35]. Cindex is the most commonly used performance measure for survival models, which calculates the fraction of pairs whose predicted survival time is correctly ordered. R package pec::cindex is used to calculate time dependent cindex [36].

Results

Building the lncRNA based survival model

We developed an lncRNA based survival model for gliomas through the following steps **(Figure 1).**

1) We first randomly selected 60% (n=277) of the patients from TCGA as training set and reserved the remaining 40% (n=184) of patients as testing set. The results remain similar with 70% patients in training and 30% in testing set (**Figure S3 A**).
2) Cox multivariate regression was carried out in the training set on 1289 lncRNA controlling for effects from other covariates like age, gender, tumor grade and IDH1 mutation status.
3) LncRNAs significantly associated with survival after likelihood ratio test (FDR p□<□0.05) were retained for selecting lncRNAs by lasso regularization.
4) After lasso regularization and lncRNA selection, a risk score formula was established by including selected lncRNAs weighted by their estimated regression coefficients in the multivariable Cox regression analysis. Risk Score = $\sum_{i=1}^{n} \beta_i * x_i$ (where, □ is coefficient and *x* is expression level of lncRNA *i*)
5) Patients were classified into high-risk and low-risk group by using the median risk score as the cutoff in the training set. The coefficient for each lncRNA and cutoff of risk score obtained from training set was used to calculate risk score and stratify patients into two groups in testing set.
6) Survival differences between the low-risk and high-risk groups in the training and testing sets were assessed by the Kaplan–Meier estimate and compared using the log-rank test.

Steps 1-6 were repeated 100 times to obtain up to 100 different lncRNA subsets (models). Only those models that separated patients in the testing set such that those with low-risk score had significantly better survival than those with high-risk score were considered as successful models and retained.

The result obtained from one such survival model is shown in **FigureS2**. In ~20% of the trials the multivariate cox-regression and lasso regularization in the training set did not select any lncRNAs significantly associated with survival (NA in **Figure 2A**). The remaining 80% of the survival models contained different numbers of lncRNAs (x-axis of **Figure 2A**) that significantly stratify patients into low and high-risk groups in training set (**Figure 2A**). Among these 80% of survival models, 86% also significantly separated patients into high-risk and low-risk in the testing set and are referred to as successful survival models. In order to create a robust survival model we sorted the lncRNAs based on the number of times an lncRNA was selected by successful survival models (**Figure 2B**). Out of 167 total prognostic lncRNA in 69 successful survival models, we first ranked lncRNAs based on number of times a given RNA was selected by successful models and then from the top 20 selected 8 lncRNAs with the highest median

cox-coefficient (Absolute value > 0.2) and least variance in the successful models in the testing set (Absolute value < 0.10). 7 out of these 8 lncRNAs were also selected after 70%-30% split of training and testing patients (**Figure S3A**), after 1000 trials instead of 100 (**Figure S3B**) and all 8 lncRNAs were selected when we used Elastic net, instead of Lasso, for regularization and lncRNA selection (**Figure S3C**) suggesting the prognostic importance of these 8 lncRNAs in gliomas. For brevity, this set of 8 lncRNA as a prognostic signature of gliomas will be referred to as UVA8 in the manuscript.

UVA8 is predictive of survival in training and independent validation set

We assessed the predictive power of UVA8 by comparing overall survival of low and high risk patients in the entire TCGA dataset stratified based on median risk score obtained by UVA8 (risk score calculation discussed in methods). Patients in the low-risk group showed longer overall survival than the high-risk group in TCGA dataset **(Figure 3A**, median OS 741.5 vs 639 days; P = 3.1e-15, HR=5.8). The risk scores of the patients in the TCGA dataset range from -4 to 4 with median risk score of -0.023 (**Figure 3B,** top panel). Moreover, there are more patients alive in the low risk group than in the high-risk group (**Figure 3B,** middle panel). Interestingly, expression levels of all lncRNA in UVA8 are high in low risk patients than in high-risk patients indicating these lncRNAs as favorable prognostic genes **((Figure 3B,** bottom panel)**)**. These findings were further validated in an independent validation dataset comprising of 274 patients obtained from CGGA. Using the same median coefficient of UVA8 obtained from the successful survival models in TCGA, patients showed longer overall survival in low-risk than in high-risk group in CGGA (**Figure 3C**, median OS = 1120.5 vs 587 days; P = 0.0017, HR=1.68). Moreover, low-risk group in CGGA has also longer progression free survival (PFS) than the high-risk group (**Figure 3D,** median PFS 597.5 vs 411.5 days; P = 0.00088, HR=1.70). Thus, UVA8 can predict survival in both training and independent validation set.

Since, 32% of patients in CGGA are in grade IV, the difference in overall survival could be due to over-representation of grade IV patients in high-risk group. However, even when only lower-grade gliomas (grade II and III) were separately examined we found significantly longer survival for low-risk versus high-risk patients (**Figure S4A**). UVA8 fails to cluster grade IV patients from CGGA into two distinct groups highlighting the specificity of signature for lower-grade gliomas (**Figure S4B**).

8-lncRNA based risk score is an independent predictor of survival

Lower grade gliomas have poorer outcomes in older patients, in tumors of higher grade and tumors with wild type IDH1 status (**Figure S1**). Interestingly, the risk score derived from UVA8 is higher in patients older than 40 years, patients in grade III vs grade II and patients harboring wild-type IDH1 gene (**Figure S5**). It was therefore important to determine whether UVA8 derived risk score is an independent predictor of survival. We divided the patients into younger (Age < 40) and older (Age >= 40) groups and found that risk-score can still stratify the patients into low-risk and high risk in both groups (**Figure 4A**). Similarly, UVA8 based risk score can still separate the patients into low and high-risk groups in grade II or grade III gliomas (**Figure 4B**). Although, IDH mutation status is a widely used prognostic and predictive biomarker, the UVA8 based risk score can also separate patients into two risk groups in patients presorted based on IDH mutation status **(Figure 4C)**. UVA8 derived risk score can also stratify patients into two risk groups among male and female patients (**Figure 4D**).

Conversely we tested whether these standard clinically used parameters, age, gender, grade and IDH mutation status, continue to independently stratify patients even after they have been presorted into two groups by UVA8 risk score (**Figure S6**). In patients with high UVA8 risk score, age, grade and IDH mutations status can further separate the patients into two groups of better or worse outcome. In contrast, in patients with low UVA8 risk scores, none of the clinical factors could further stratify patients into two different survival groups with a pvalue<0.05 (**Figure S6**). Consistent with the previous observation (**Figure S1**), gender is ineffective in stratifying patients into two categories within patients with high- or low-risk score.

UVA-8 is a better predictor of glioma patients' survival

We assessed the accuracy of UVA8 in prediction of survival by comparing its time-dependent area under curve (AUC) with other clinical characteristics. For each prognostic factor (e.g. UVA8, IDH status etc.) we varied the cut-off so as to vary the false positive rate for five-year survival prediction from 0 to 1. For each cut-off the corresponding true positive rate for five-year survival was calculated (**Figure 5A**). Comparing the Area-under the curve (AUC) for these ROC curves suggested that UVA8 performs best in predicting survival of the glioma patients compared to the other criteria. This calculation was extended to predict survival of other durations (1-16 years) and the AUC plotted for each predictor (**Figure 5B**). UVA8 can predict survival better for all durations, particularly at

the very early years after diagnosis when the prediction is worse for most of the predictors. Since, gender is not associated with glioma patients' survival (**Figure S1**), the prediction of outcome was no better than random guess (AUC = 0.5) **(Figure 5A and 5B)**. We employed Cox multivariable probability hazard model to identify the impact of UVA8 and different clinicopathological characteristics in estimating hazard (**Figure 5C**). UVA8 is most significantly correlated with the survival information (p = 1.4e-07) and shows highest hazard ratio (HR = 4), indicating that the risk score performs better than any other currently used approaches for prognosis. Here, the hazard ratio of UVA8 is calculated by dichotomizing the risk score of > -0.023 (median risk score from TCGA) to 1 and < -0.023 to 0 to compare the hazard rates of high risk versus low risk patients. The hazard ratio of the 8 lncRNAs individually and combined as risk score is tabulated in **Supplementary Table S1**. The UVA8 Risk score is associated with more hazard (HR=2) than any of the individual lncRNA supporting the importance of a combinatorial signature than an individual RNA for predicting survival. The hazard ratio of UVA8 in **Supplementary Table S1** is different from that in **Figure 5C** because in the former the hazard ratio is calculated with the risk score as a continuous variable.

We then sought to compare the performance of UVA8 based survival model with published lncRNA based survival models by calculating Cindex (as discussed in Methods) for TCGA dataset for each of the models. We first calculated risk score for each patient by considering the expression level of the prognostic lncRNAs in each model weighted by their estimated regression coefficients retrieved from the respective studies (**Supplementary Table S2**). The patients were ordered based on their actual survival at a given time after diagnosis and based on their risk score in each model. The concordance of the two orders is measured in pairwise comparisons of the patients to calculate a single time-dependent concordance index for the model that is being evaluated. UVA8 outperforms all existing lncRNA based survival models at different times after diagnosis (**Figure 5D).** As expected, prognostic signatures that were specific to GBMs (Zhang6_2013 and Zhou6_2017) show poor concordance index when used to predict survival of lower-grade glioma patients.

Interferon signaling is the most enriched pathway in guilt by association with UVA8

Although many lncRNAs have been identified there has been very little functional annotation of the RNAs. We therefore applied guilt-by-association to infer functions of the lncRNAs associated with survival in UVA8. First we

interrogated whether protein-coding genes most correlated with an lncRNA in TCGA glioma cohort are themselves predictive of outcome. All the lncRNAs in UVA8 are associated with a negative cox coefficient (protective). Of the 8 mRNAs most correlated positively with these 8 lncRNAs, 5 also have a negative cox-coefficient with a significant p-value. Conversely, of the 8 mRNAs most anti-correlated with these lncRNAs, 5 have a positive cox coefficient with a significant p-value (**Figure 6A**). This result is consistent with the expectation that the expression of these protective lncRNAs will be positively correlated with expression of protective mRNAs and negatively correlated with the expression of harmful mRNAs.

GSEA analysis on protein-coding genes pre-ranked from most positively correlated to most negatively correlated to the lncRNA revealed several common pathways co-regulated with each of the 8 lncRNAs (**Figure 6B**). Interestingly, among the mRNAs that are negatively correlated with the lncRNAs, genes involved in immune and inflammatory response (IFNG, IFNA, allograft rejection, NFkB inflammatory response and JAK-STAT pathway) are highly enriched. Similarly genes involved in epithelial to mesenchymal transition and cell-cycle progressions are also most enriched. These gene-set enrichments suggest a conventional tumor suppressor phenotype associated with these 8 lncRNAs.

Many of the mRNAs are common in the IFNG, IFNA, allograft rejection, NFkB inflammatory response and JAK-STAT gene sets. The genes up-regulated in response to IFNG are mostly negatively correlated to lncRNAs in UVA8. To visualize this, the correlation coefficients were plotted for each lncRNA (columns) with individual mRNAs in the IFNG response pathway (rows) (**Figure 6C**). Out of 8, 6 lncRNAs (RP11-266K4.14, FLJ37035, RP11-118K6.3, RP11-142A22.3, LINC00641 and RP11-96H19.1) are clustered together because they are more negatively correlated with genes of interferon gamma response pathway (**Figure 6C**).

We found both NFKB and STAT3 genes as highly negatively correlated with the expression of the protective lncRNAs in UVA8. Genes involved in epithelial to mesenchymal transition and encoding cell cycle related targets of E2F transcription factors and involved in G2/M checkpoints were also negatively correlated with UVA8 expression. On the other hand, genes that are down regulated upon activation of the oncogenes KRAS are positively correlated with the expression of the protective lncRNAs of UVA8.

In order to check whether these lncRNAs can possibly act as eRNAs, we also checked the distance between lncRNAs and their correlated genes and found that these lncRNAs are correlated to several genes located in different location of genome suggesting a trans-regulation by these lncRNAs (data not shown). More experimental studies are required in future to decipher the role of these lncRNAs in regulating these genes and whether this regulation explains the effect of the lncRNAs on glioma tumor progression.

Discussion

Gene expression profile reflects the underlying biological processes of disease. Cox regression is a widely used approach to decipher correlation between gene expression profile and patient outcome. Previous analyses on microarray data explored protein coding genes that could predict the prognosis of gliomas, particularly focusing on high grade GBMs. LncRNAs are a class of RNA which can serve as a better prognostic marker than protein coding mRNAs because they are numerous and cell-type specific [2, 3]. Additionally, since lncRNAs do not encode protein, they are the ultimate effectors, and their expression levels more accurately predict the levels of their activity. Recent studies have detected tumor-specific lncRNAs in exosomes, apoptotic bodies and microparticles highlighting another advantage of considering lncRNAs in tumors, because they are expected to appear as fluid-based markers for the diagnosis of different cancers [37–39]. Among six published lncRNA-based prognostic signatures for gliomas two are for predicting outcome in GBMs and one specifically for anaplastic gliomas. Wang et al, 2016 and Chen et al, 2017 have shown that a set of only four lncRNAs could predict survival in gliomas [23, 25]. However, the sequence of one of the lncRNAs in Chen et al., 2017, CR613436, was removed by the submitter on NCBI. Recently, the role of immune-related genes in glioma malignancies is gaining attention leading to the discovery of immune-related lncRNA-based prognostic markers for GBMs and anaplastic gliomas [40, 41]. Remarkably, there is no overlap between the prognostic lncRNAs identified in the aforementioned studies. Moreover, these studies are based on microarray data raising concerns particular to hybridization-based approaches including reliance on current knowledge of expressed genes, problems of cross-hybridization and cross-experiment comparison. Another issue is that association of lncRNAs with survival using cox-regression was sometimes carried out without controlling for any dependent variables and without penalizing for the effect of large number of variables.

In the present study, we have used an approach to screen lncRNAs from high-dimensional TCGA RNA-Seq data, which is one of the largest and the most updated data for lower-grade gliomas. After controlling for effects like age, grade, gender and IDH mutation status, we applied regularization to penalize the effect of many dependent variables and select the lncRNAs based on 100 trials. We showed the robustness of eight-lncRNA based predictor in a completely independent cohort of Chinese glioma patients. The lncRNA prognostic signature identified in the present study, UVA8, is an independent predictor of survival in TCGA glioma patients. Since UVA8 is also a better predictor than the few patient and molecular characteristics currently used for prognosis in the clinic, a simple RNA quantification will aid the physician to decide whether to adopt more aggressive therapy at the outset.

The protective lncRNAs that constitute UVA8 are negatively correlated with protein coding genes involved in interferon gamma and inflammatory response highlighting the role of immune-response genes in glioma progression. Except LINC01561, all 7 lncRNAs (RP11-266K4.14, FLJ37035, RP11-118K6.3, DGCR9, RP11-142A22.3, LINC00641 and RP11-96H19.1) are negatively correlated to most of the protein-coding genes which are up-regulated in response to interferon gamma/alpha, genes regulated by NF-kB in response to TNF, inflammatory response, and genes up-regulated by IL6 via STAT3. This suggests that an active immune reaction perhaps in response to cytokines secreted from tumor and immune cells is predictive of poor outcome in gliomas. NF-κB and JAK/STAT pathways are known to be aberrantly up-regulated in GBMs. The level of NF-κB increases as the tumors progress in astrocytic tumors [42, 43] and STAT3 is constitutively active in GBMs [44, 45]. Immune related pathways are also known to be involved in glioma tumor cell proliferation [46], survival [40], invasion [47] and chemoresistance [48]. In addition, epithelial-mesenchymal transition (associated with invasion) and active cell proliferation are suppressed if UVA8 lncRNAs are high, and this leads to better outcome, consistent with our understanding of how invasion and cell proliferation negatively impact outcome. On the other hand, genes that were positively correlated with the expression of UVA8 are enriched in genes that are down regulated by activation of the oncogene KRAS.

There are reports of the same lncRNA being predictive of outcome in the same manner in multiple tumor types. For example, DRAIC expression predicts good outcome in gliomas, melanomas, and cancers of the prostate, stomach, liver, kidney and lung [49]. In contrast, expression of LINC00152/CYTOR is predictive of poor outcome in gliomas, and cancers of the head & neck, lung, kidney, liver and pancreas (our unpublished work). Such observations are particularly exciting because they imply that the lncRNA has an important role in tumor biology

that transcends tumor types, and these RNAs should be prioritized for cell- and molecular-biology studies to discern their function. It will thus be very interesting to explore whether any of the lncRNAs of UVA8 will be protective in other tumor types. Finally, future studies will address whether structural variation, copy number variations and sequence polymorphism of these lncRNAs contribute to the prognostic outcome. We are excited that UVA8 was also predictive of outcome in a completely different tumor cohort (CGGA) from a patient population that is from an entirely different geographical location with attendant differences in environment and population genotypes. It will be interesting to see if UVA8 is equally predictive of outcome in other patient populations from other parts of the world.

Acknowledgments

We thank Dutta lab members for helpful discussions. M.K. is supported by a DOD award PC151085. The work was supported by a V foundation award D2018-002 and R01 AR067712 from NIAMS.

References

1. Derrien T, Johnson R, Bussotti G, et al (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res 22:1775–1789 . doi: 10.1101/gr.132159.111
2. Cabili M, Trapnell C, Goff L, et al (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25:1915–1927 . doi: 10.1101/gad.17446611
3. Huarte M (2015) The emerging role of lncRNAs in cancer. Nat Med 21:1253–1261 . doi: 10.1038/nm.3981
4. Schmitt AM, Chang HY (2016) Long Noncoding RNAs in Cancer Pathways. Cancer Cell 29:452–463 . doi: 10.1016/j.ccell.2016.03.010
5. Stupp R, Mason WP, van den Bent MJ, et al (2005) Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. N Engl J Med 352:987–996 . doi: 10.1056/NEJMoa043330
6. Huang J, Samson P, Perkins SM, et al (2017) Impact of concurrent chemotherapy with radiation therapy for elderly patients with newly diagnosed glioblastoma: a review of the National Cancer Data Base. J Neurooncol 131:593–601 . doi: 10.1007/s11060-016-2331-6
7. Ducray F, Idbaih A, Wang X-W, et al (2011) Predictive and prognostic factors for gliomas. Expert Rev

Anticancer Ther 11:781–789 . doi: 10.1586/era.10.202

8. Carninci P, Kasukawa T, Katayama S, et al (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–63 . doi: 10.1126/science.1112014

9. Ravasi T, Suzuki H, Pang KC, et al (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res 16:11–19 . doi: 10.1101/gr.4200206

10. Mehler MF, Mattick JS (2007) Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. Physiol Rev 87:799–823 . doi: 10.1152/physrev.00036.2006

11. Taft RJ, Pang KC, Mercer TR, et al (2010) Non-coding RNAs: Regulators of disease. J. Pathol. 220:126–139

12. Qureshi IA, Mattick JS, Mehler MF (2010) Long non-coding RNAs in nervous system function and disease. Brain Res. 1338:20–35

13. Mercer TR, Dinger ME, Sunkin SM, et al (2008) Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci U S A 105:712–716 . doi: 0706729105 [pii] 10.1073/pnas.0706729105

14. Amaral PP, Neyt C, Wilkins SJ, et al (2009) Complex architecture and regulated expression of the Sox2ot locus during vertebrate development. RNA 15:2013–27 . doi: 10.1261/rna.1705309

15. Johnson R, Teh CH-L, Jia H, et al (2009) Regulation of neural macroRNAs by the transcriptional repressor REST. RNA 15:85–96 . doi: 10.1261/rna.1127009

16. Arron JR, Winslow MM, Polleri A, et al (2006) NFAT dysregulation by increased dosage of DSCR1 and DYRK1A on chromosome 21. Nature 441:595–600 . doi: 10.1038/nature04678

17. Wang J, Zhao H, Fan Z, et al (2017) Long Noncoding RNA H19 Promotes Neuroinflammation in Ischemic Stroke by Driving Histone Deacetylase 1-Dependent M1 Microglial Polarization. Stroke 48:2211–2221 . doi: 10.1161/STROKEAHA.117.017387

18. Chubb JE, Bradshaw NJ, Soares DC, et al (2008) The DISC locus in psychiatric illness. Mol Psychiatry 13:36–64 . doi: 10.1038/sj.mp.4002106

19. Zhang X, Sun S, Pu JKS, et al (2012) Long non-coding RNA expression profiles predict clinical phenotypes in glioma. Neurobiol Dis 48:1–8 . doi: 10.1016/J.NBD.2012.06.004

20. Reon BJ, Anaya J, Zhang Y, et al (2016) Expression of lncRNAs in Low-Grade Gliomas and Glioblastoma Multiforme: An In Silico Analysis. PLOS Med 13:e1002192 . doi: 10.1371/journal.pmed.1002192

21. Li R, Qian J, Wang Y-Y, et al (2014) Long Noncoding RNA Profiles Reveal Three Molecular Subtypes in Glioma. CNS Neurosci Ther 20:339–343 . doi: 10.1111/cns.12220

22. Wang W, Zhao Z, Yang F, et al (2018) An immune-related lncRNA signature for patients with anaplastic gliomas. J Neurooncol 136:263–271 . doi: 10.1007/s11060-017-2667-6

23. Wang W, Yang F, Zhang L, et al (2016) LncRNA profile study reveals four-lncRNA signature associated with the prognosis of patients with anaplastic gliomas. Oncotarget 7:77225–77236 . doi: 10.18632/oncotarget.12624

24. Zhang X-Q, Sun S, Lam K-F, et al (2013) A long non-coding RNA signature in glioblastoma multiforme predicts survival. Neurobiol Dis 58:123–131 . doi: 10.1016/J.NBD.2013.05.011

25. Chen G, Cao Y, Zhang L, et al (2017) Analysis of long non-coding RNA expression profiles identifies novel lncRNA biomarkers in the tumorigenesis and malignant progression of gliomas. Oncotarget 8:67744–67753 . doi: 10.18632/oncotarget.18832

26. Bao ZS, Chen HM, Yang MY, et al (2014) RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. Genome Res 24:1765–1773 . doi: 10.1101/gr.165126.113

27. Harrow J, Frankish A, Gonzalez JM, et al (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22:1760–74 . doi: 10.1101/gr.135350.111

28. Pertea M, Pertea GM, Antonescu CM, et al (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295 . doi: 10.1038/nbt.3122

29. Tibshirani R (1997) The lasso method for variable selection in the cox model. Stat Med 16:385–395 . doi: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

30. Tibshirani R (2011) Regression shrinkage and selection via the lasso: A retrospective. J R Stat Soc Ser B Stat Methodol 73:273–282 . doi: 10.1111/j.1467-9868.2011.00771.x

31. Goeman JJ (2010) L1 penalized estimation in the Cox proportional hazards model. Biometrical J 52:70–84 . doi: 10.1002/bimj.200900028

32. Alizadeh AA, Gentles AJ, Alencar AJ, et al (2011) Prediction of survival in diffuse large B-cell lymphoma based on the expression of 2 genes reflecting tumor and microenvironment. Blood 118:1350–8 . doi: 10.1182/blood-2011-03-345272

33. Lossos IS, Czerwinski DK, Alizadeh AA, et al (2004) Prediction of Survival in Diffuse Large-B-Cell Lymphoma Based on the Expression of Six Genes. N Engl J Med 350:1828–1837 . doi: 10.1056/NEJMoa032520

34. Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33: . doi: 10.18637/jss.v033.i01

35. Raykar VC, Steck H, Krishnapuram B, et al On Ranking in Survival Analysis: Bounds on the Concordance Index

36. Gerds TA, Kattan MW, Schumacher M, Yu C (2013) Estimating a time-dependentconcordance index for survival prediction models with covariate dependent censoring. Stat Med 32:2173–2184 . doi: 10.1002/sim.5681

37. Panzitt K, Tschernatsch MMO, Guelly C, et al (2007) Characterization of HULC, a Novel Gene With Striking Up-Regulation in Hepatocellular Carcinoma, as Noncoding RNA. Gastroenterology 132:330–342 . doi: 10.1053/J.GASTRO.2006.08.026

38. Du Z, Fei T, Verhaak RGW, et al (2013) Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat Struct Mol Biol 20:908–13 . doi: 10.1038/nsmb.2591

39. Mohankumar S, Patel T (2016) Extracellular vesicle long noncoding RNA as potential biomarkers of liver cancer. Brief Funct Genomics 15:249–56 . doi: 10.1093/bfgp/elv058

40. Zhou M, Zhang Z, Zhao H, et al (2017) An Immune-Related Six-lncRNA Signature to Improve Prognosis Prediction of Glioblastoma Multiforme. Mol. Neurobiol. 1–14

41. Wang W, Zhao Z, Yang F, et al (2018) An immune-related lncRNA signature for patients with anaplastic gliomas. J Neurooncol 136:263–271 . doi: 10.1007/s11060-017-2667-6

42. Angileri FF, Aguennouz M, Conti A, et al (2008) Nuclear factor-κB activation and differential expression of survivin and Bcl-2 in human grade 2-4 astrocytomas. Cancer 112:2258–2266 . doi: 10.1002/cncr.23407

43. Korkolopoulou P, Levidou G, Saetta AA, et al (2008) Expression of nuclear factor-κB in human astrocytomas: relation to pIκBa, vascular endothelial growth factor, Cox-2, microvascular characteristics, and survival. Hum Pathol 39:1143–1152 . doi: 10.1016/J.HUMPATH.2008.01.020

44. Schaefer LK, Ren Z, Fuller GN, Schaefer TS (2002) Constitutive activation of Stat3α in brain tumors: localization to tumor endothelial cells and activation by the endothelial tyrosine kinase receptor (VEGFR-2).

Oncogene 21:2058–2065 . doi: 10.1038/sj.onc.1205263

45. Abou-Ghazal M, Yang DS, Qiao W, et al (2008) The incidence, correlation with tumor-infiltrating inflammation, and prognosis of phosphorylated STAT3 expression in human gliomas. Clin Cancer Res 14:8228–35 . doi: 10.1158/1078-0432.CCR-08-1329

46. Puliyappadamba VT, Hatanpaa KJ, Chakraborty S, Habib AA (2014) The role of NF-κB in the pathogenesis of glioma. Mol Cell Oncol 1:e963478 . doi: 10.4161/23723548.2014.963478

47. Kesanakurti D, Chetty C, Rajasekhar Maddirela D, et al (2013) Essential role of cooperative NF-κB and Stat3 recruitment to ICAM-1 intronic consensus elements in the regulation of radiation-induced invasion and migration in glioma. Oncogene 32:5144–5155 . doi: 10.1038/onc.2012.546

48. Coupienne I, Bontems S, Dewaele M, et al (2011) NF-kappaB inhibition improves the sensitivity of human glioblastoma cells to 5-aminolevulinic acid-based photodynamic therapy. Biochem Pharmacol 81:606–616 . doi: 10.1016/J.BCP.2010.12.015

49. Sakurai K, Reon BJ, Anaya J, Dutta A (2015) The lncRNA DRAIC/PCAT29 Locus Constitutes a Tumor-Suppressive Nexus. Mol Cancer Res 13:828–38 . doi: 10.1158/1541-7786.MCR-15-0016-T

Figure Legend

Figure 1. Flowchart showing steps involved in identification of lncRNA based prognostic signature.

Figure 2. Selection of lncRNAs with best predictors of outcome. A) Barplot showing number of lncRNAs that predicted outcome in the training set in 100 trials. The successful models were those that also predicted outcome in the testing set. NA: no lncRNA predicted outcome in training set. **B)** Barplot showing number of times each of the top 20 lncRNAs (out of 167) were present in successful survival models (significant in testing set). The lower panel shows median Cox-coefficient (after lasso penalization) and the variance of the cox-coefficient for each of the above 20 lncRNAs from the successful models where they were selected. The arrow points towards lncRNAs selected for UVA8.

Figure 3. Survival analysis of the patients divided by the prognostic lncRNAs in two data sets. A) Patients in the entire TCGA dataset with risk score greater than median score of -0.023 show poor survival compared with patients with risk score less than median risk score. **B) Upper panel:** Plot showing patients sorted based on UVA8 risk score with black representing patient with risk score below median and red showing those with risk score above median. **Middle panel:** Number of days of survival indicated on Y-axis of patients sorted on the X-axis based on the risk scores in the top panel and alive/dead status indicated by color. **Bottom panel:** z-score transformed expression value of lncRNAs in UVA8 show higher expression in patients with low risk score. **C)** Kaplan Meier plot of overall survival of patients in CGGA dataset with risk score greater than (red) or less than (black) median risk score of TCGA dataset. **D)** Kaplan-Meier plot for progression free survival in CGGA dataset showed poor survival for patients with high-risk score. Rest as in C.

Figure 4. Stratification analysis by different clinical variables. Kaplan-Meier curve analysis of overall survival in high- and low-risk groups for **A)** younger (Age < 40) and older patients (Age >= 40). **B)** Grade II and Grade III patients **C)** IDH mutation status as WT and mutation (MUT) patients **D)** Male and Female patients. Black dashed line: patients with high risk score, Gray solid line: patients with low risk score. The tables on the right show log-rank p-value, hazard ratio and 95% confidence interval for each Kaplan-Meier plot.

Figure 5. Performance evaluation of the 8-lncRNA based risk score. A) Receiver operating characteristic curve for 5-year survival shows UVA8 has better Area-Under-Curve compared with other predictors. **B)** Area-Under-Curve plotted for different durations of survival for 8-lncRNA based risk score, tumor grade, Age, IDH mutation status and gender of patients in TCGA cohort. **C)** Cox multivariate regression with clinical information and risk

score calculated from UVA8 for survival in TCGA cohort. **D)** Concordance-index showing measure of concordance of predictor with survival of patients in TCGA.

Figure 6. Guilt-by-association analysis of the 8lncRNAs in UVA8. A) Correlation and Cox-regression coefficient for the mRNAs that are most correlated (positive and negatively) with each of the lncRNAs in UVA8. a, b and c defined below the table. **B)** List of pathways that are most enriched in protein-coding genes that are negatively correlated with the UVA8 lncRNAs. **C)** Heatmap showing correlation of different genes in the interferon-gamma response gene set (rows) to the lncRNAs in UVA8 (columns).

Figure 1

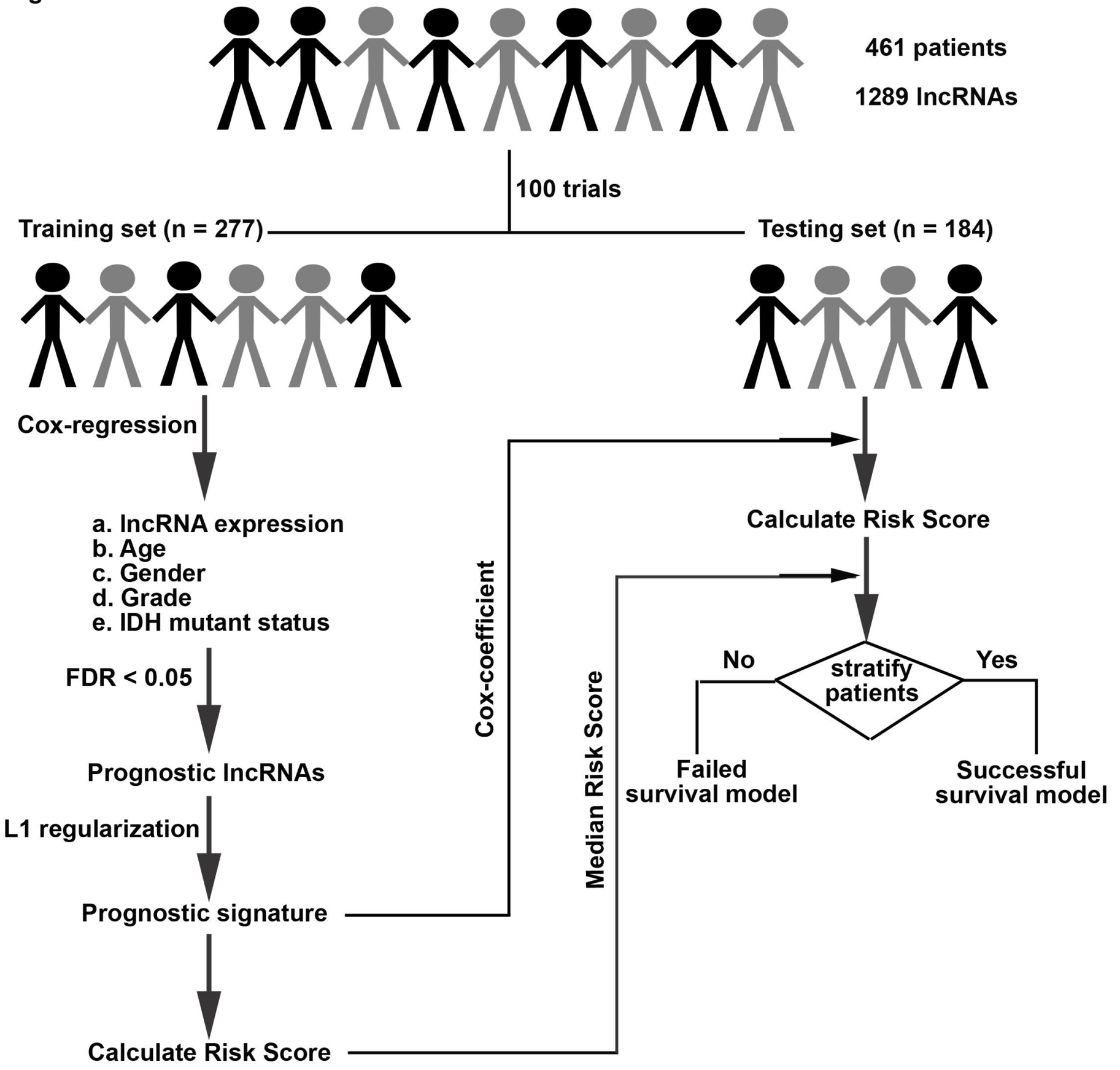
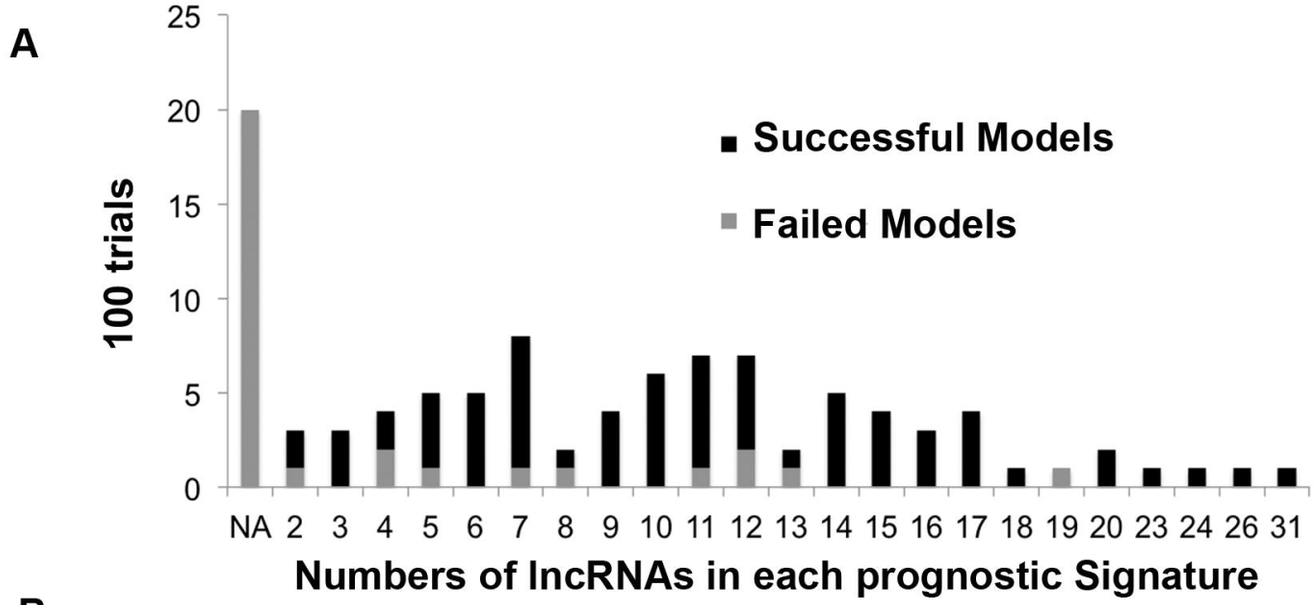

461 patients
1289 lncRNAs
100 trials
Training set (n = 277)
Testing set (n = 184)
Cox-regression
a. lncRNA expression
b. Age
c. Gender
d. Grade
e. IDH mutant status
FDR < 0.05
Prognostic lncRNAs
L1 regularization
Prognostic signature
Calculate Risk Score
Cox-coefficient
Median Risk Score
Calculate Risk Score
No
stratify patients
Yes
Failed survival model
Successful survival model

Figure 2


A
100 trials
Successful Models
Failed Models
NA 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 23 24 26 31
Numbers of lncRNAs in each prognostic Signature

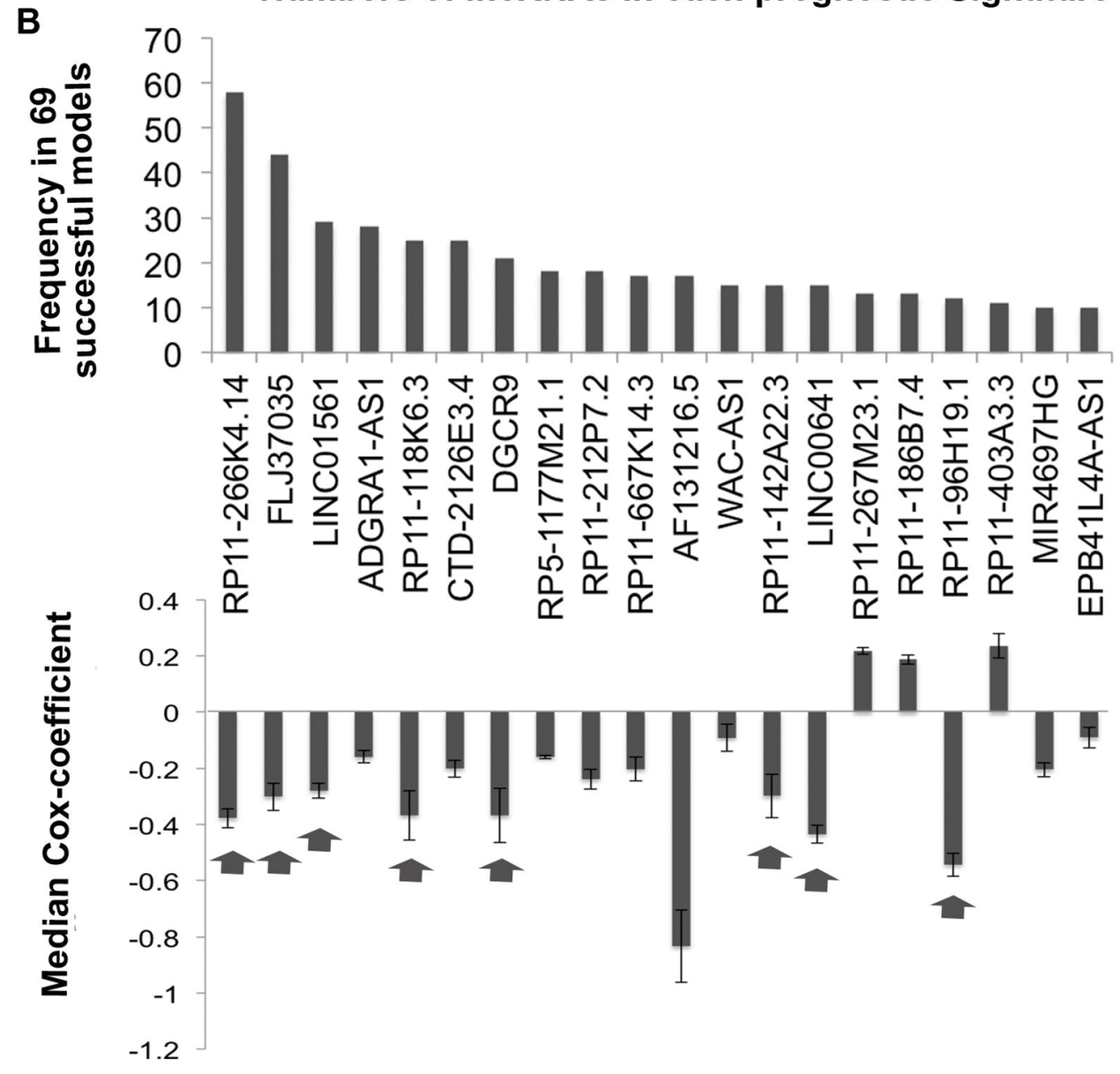

B
Frequency in 69 successful models
RP11-266K4.14
FLJ37035
LINC01561
ADGRA1-AS1
RP11-118K6.3
CTD-2126E3.4
DGCR9
RP5-1177M21.1
RP11-212P7.2
RP11-667K14.3
AF131216.5
WAC-AS1
RP11-142A22.3
LINC00641
RP11-267M23.1
RP11-186B7.4
RP11-96H19.1
RP11-403A3.3
MIR4697HG
EPB41L4A-AS1
Median Cox-coefficient

Figure 3

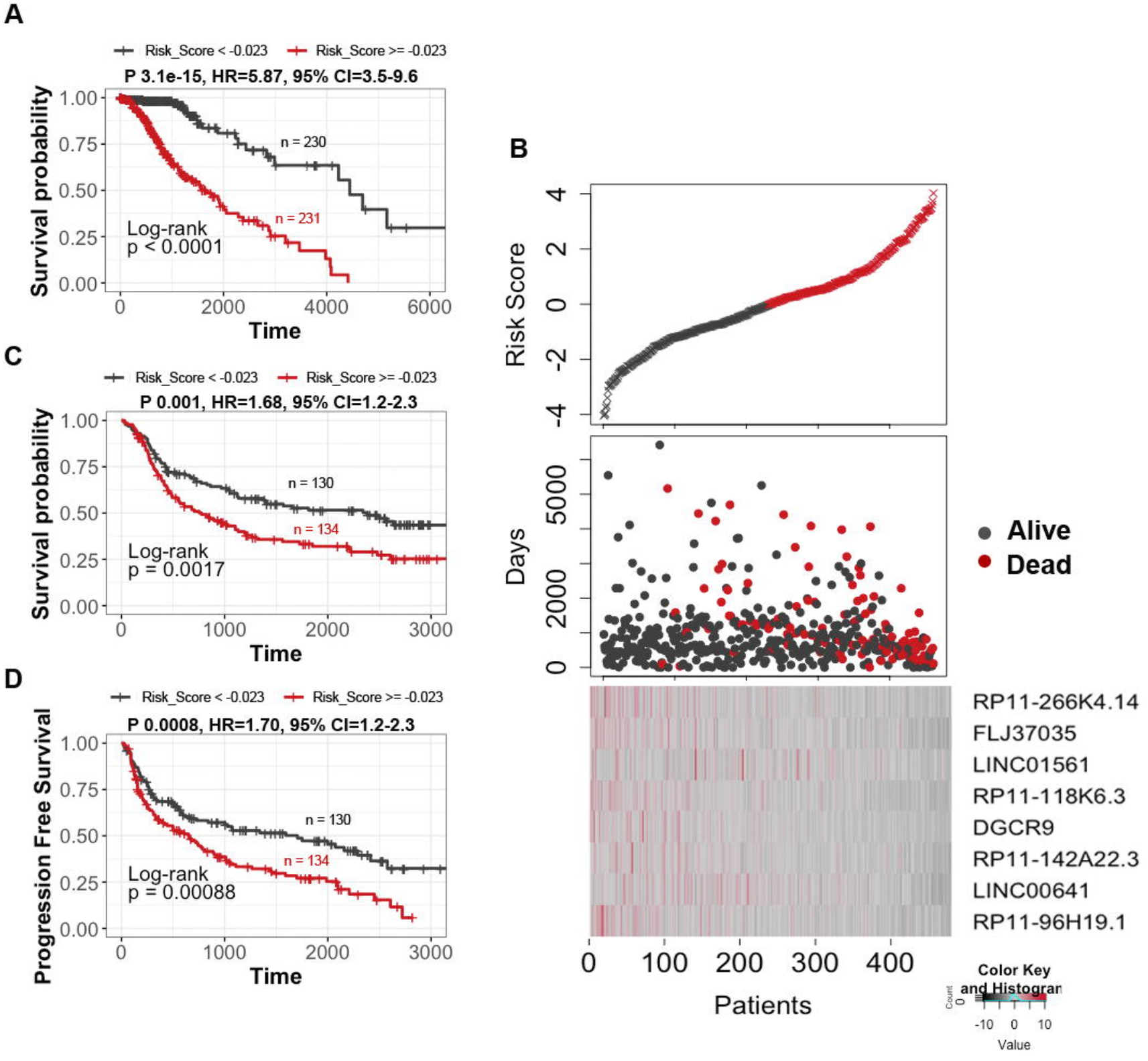
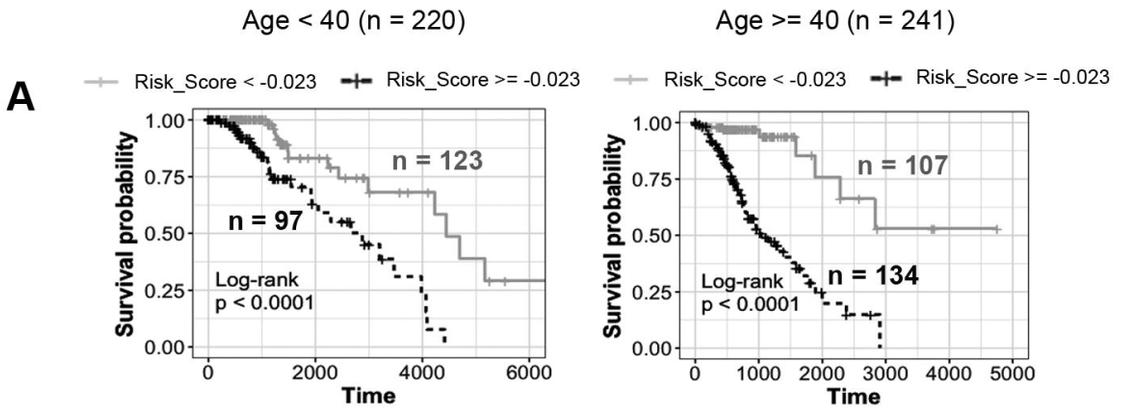
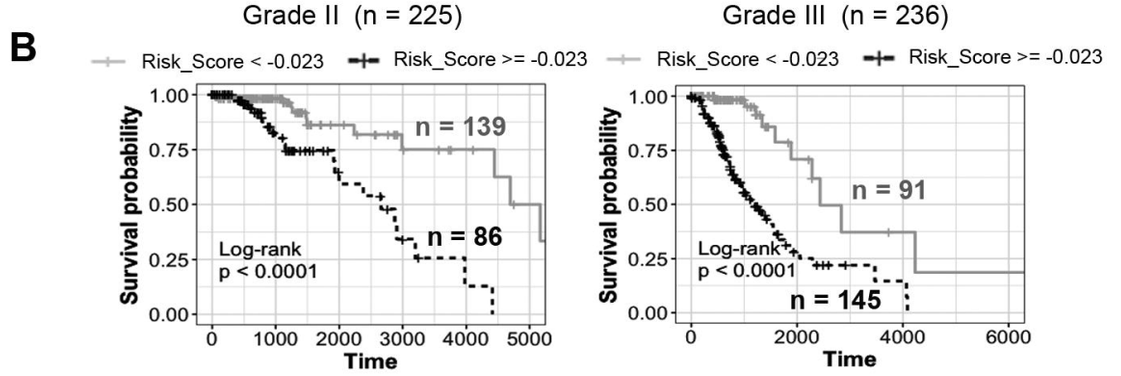

A
Risk_Score < -0.023
Risk_Score >= -0.023
P 3.1e-15, HR=5.87, 95% CI=3.5-9.6
Survival probability
1.00
0.75
0.50
0.25
0.00
n = 230
n = 231
Log-rank
p < 0.0001
0
2000
4000
6000
Time
B
Risk Score
4
2
0
-2
-4
Days
5000
2000
0
Alive
Dead
RP11-266K4.14
FLJ37035
LINC01561
RP11-118K6.3
DGCR9
RP11-142A22.3
LINC00641
RP11-96H19.1
0
100
200
300
400
Patients
Color Key and Histogram
Count
0
-10
0
10
Value
C
Risk_Score < -0.023
Risk_Score >= -0.023
P 0.001, HR=1.68, 95% CI=1.2-2.3
Survival probability
1.00
0.75
0.50
0.25
0.00
n = 130
n = 134
Log-rank
p = 0.0017
0
1000
2000
3000
Time
D
Risk_Score < -0.023
Risk_Score >= -0.023
P 0.0008, HR=1.70, 95% CI=1.2-2.3
Progression Free Survival
1.00
0.75
0.50
0.25
0.00
n = 130
n = 134
Log-rank
p = 0.00088
0
1000
2000
3000
Time

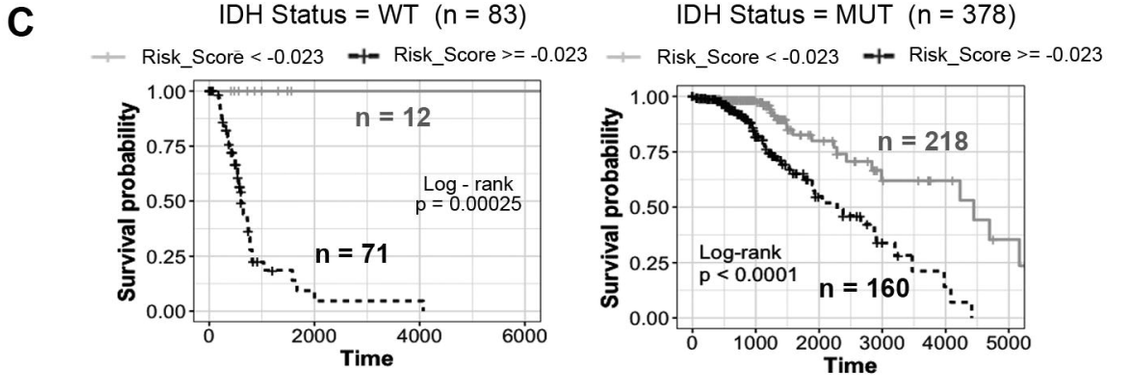
Figure 4



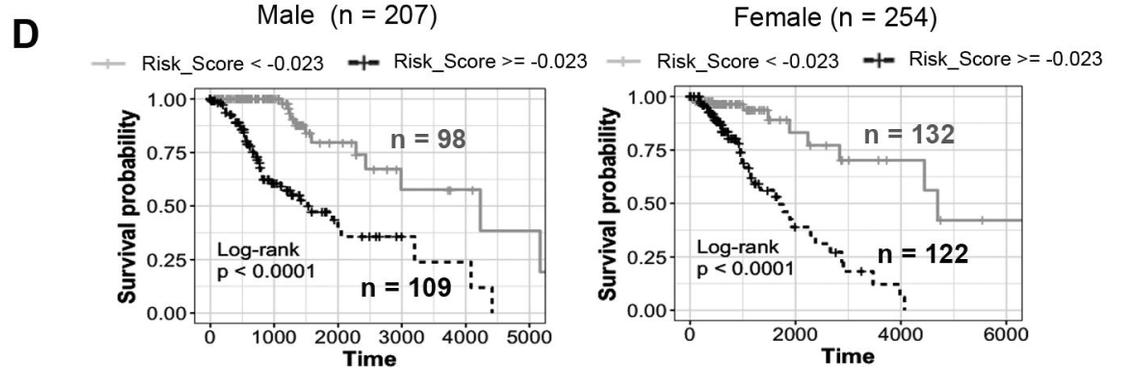
Age	P-val	HR	CI
<40	4.65e-5	3.99	2.0-8.2
>=40	1.83e-9	7.20	3.4-15.2



Grade	P-val	HR	CI
II	2.8e-5	4.60	2.1-10.1
III	2.5e-7	5.21	2.6-10.5



IDH	P-val	HR	CI
WT	0.00025	3e+8	0-Inf
MUT	4.7e-6	3.36	1.9-5.8



Gender	P-val	HR	CI
Male	1.3e-8	5.50	2.8-10.6
Female	5.6e-7	5.75	2.7-12.4

Figure 5

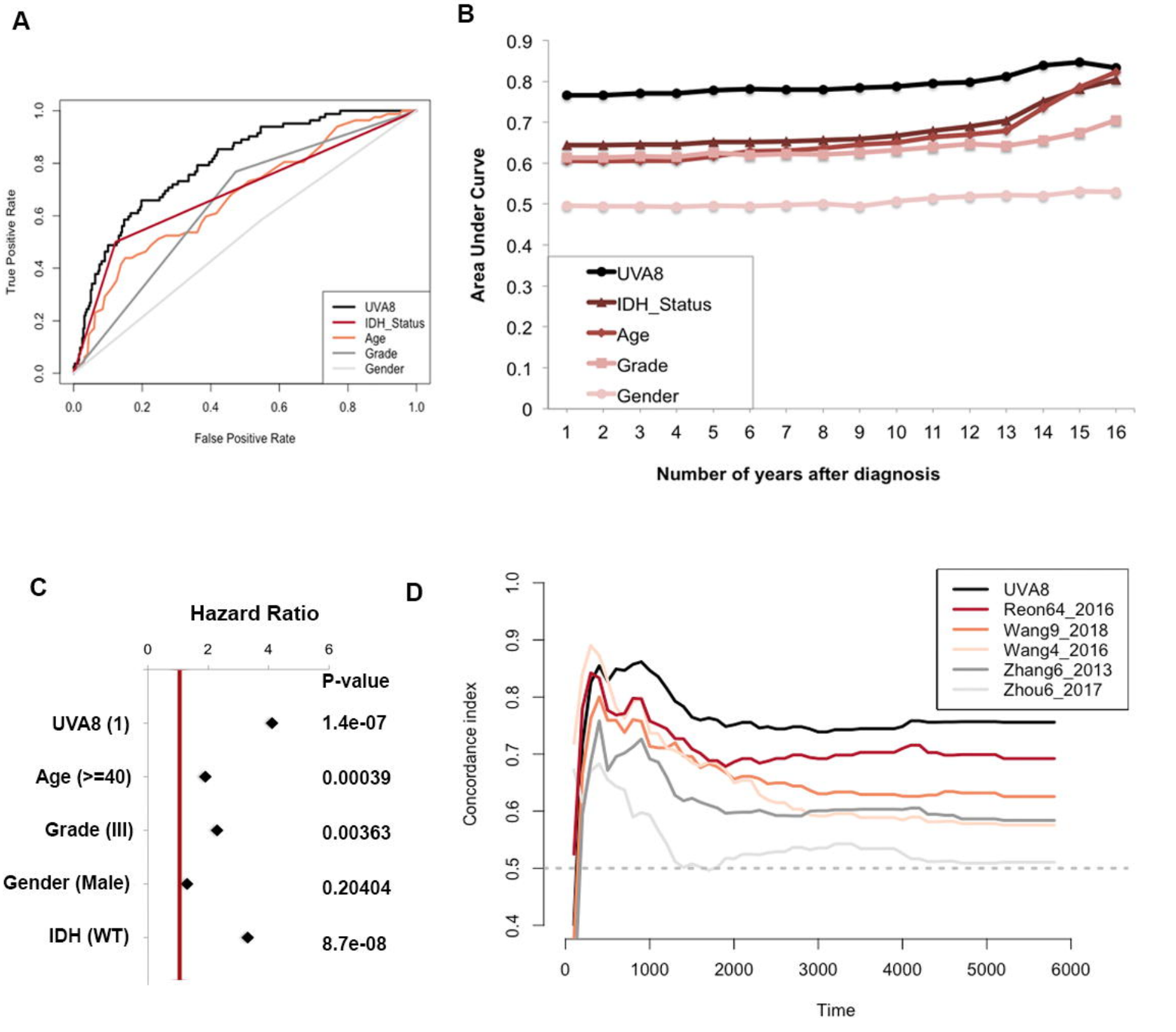

A
True Positive Rate
False Positive Rate
UVA8
IDH_Status
Age
Grade
Gender
B
Area Under Curve
Number of years after diagnosis
UVA8
IDH_Status
Age
Grade
Gender
C
Hazard Ratio
P-value
UVA8 (1)
1.4e-07
Age (>=40)
0.00039
Grade (III)
0.00363
Gender (Male)
0.20404
IDH (WT)
8.7e-08
D
Concordance index
Time
UVA8
Reon64_2016
Wang9_2018
Wang4_2016
Zhang6_2013
Zhou6_2017

Figure 6

A

lncRNAs	Positively correlated (a,b,c)	Negatively correlated (a,b,c)
RP11-266K4.14	FGFBP3 (0.57,-0.39,3.6e-5)	CYSTM1 (-0.50,0.20,0.13)
FLJ37035	ZNF32 (0.60,-0.53,0.001)	CDC42 (-0.55,0.22.0.03)
LINC01561	PLPP4 (0.78,-0.23,0.018)	NR2E1 (-0.53,0.004,0.92)
RP11-118K6.3	RTP5 (0.61,-0.19,0.174)	DDOST (-0.62,0.23,0.027)
DGCR9	BRSK2 (0.74,-0.44,0.00017)	GNG5 (-0.66,0.45,8.5e-06)
RP11-142A22.3	PRRT1 (0.70,-0.15,0.25)	SMC5 (-0.68,0.22,0.006)
LINC00641	AKAP6 (0.68,0.07,0.51)	EDEM2 (-0.72,0.07,0.53)
RP11-96H19.1	NUDT7 (0.63,-0.311,0.021)	RTCA (-0.58,0.28,0.02)

a spearman correlation
b Cox-coefficient
c log-rank p-value

B

Hallmark Gene Sets	Normalized Enrichment Score
HALLMARK_INTERFERON_GAMMA_RESPONSE	4.76
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	4.38
HALLMARK_E2F_TARGETS	4.22
HALLMARK_INTERFERON_ALPHA_RESPONSE	3.79
HALLMARK_G2M_CHECKPOINT	3.74
HALLMARK_ALLOGRAFT_REJECTION	3.56
HALLMARK_TNFA_SIGNALING_VIA_NFKB	3.52
HALLMARK_INFLAMMATORY_RESPONSE	3.31
HALLMARK_IL6_JAK_STAT3_SIGNALLING	3.13
HALLMARK_COMPLEMENT	2.96

C