LOCC: a novel visualization and scoring of cutoffs for continuous variables

George Luo[1] and John J. Letterio[2,3,4]

1 Department of Pathology, Case Western Reserve University School of Medicine, Cleveland, Ohio

2 The Angie Fowler Adolescent and Young Adult Cancer Institute, University Hospitals Rainbow Babies & Children's Hospital, Cleveland, Ohio

3 The Case Comprehensive Cancer Center, Cleveland, Ohio

4 Department of Pediatrics, Case Western Reserve University, Cleveland, Ohio

Corresponding Authors:

George Luo, Case Western Reserve University School of Medicine, 2103 Cornell Rd., Wolstein Research Bldg. Rm 3501, Cleveland, Ohio 44106; gxl263@case.edu

John J. Letterio, M.D., The Angie Fowler Adolescent and Young Adult Cancer Institute, University Hospitals Rainbow Babies & Children's Hospital, 11100 Euclid Avenue, Cleveland, Ohio 44106; jjl26@case.edu

Running Title: 'Novel visualization tool and scoring for cutoff selection'

Key Words: LOCC, Cutoff, Prognosis, Gene Expression, Tool, Score

Abstract

Background: There is a need for new methods to select and analyze cutoffs employed to define genes that are most prognostic significant and impactful. We designed LOCC, a novel tool to visualize and score continuous variables for a dichotomous outcome.

Methods: We analyzed TCGA hepatocellular carcinoma gene expression and patient data using LOCC. Analysis of *E2F1*, *TP53*, and a previously published gene signature were performed to demonstrate the utility of LOCC.

Results: LOCC demonstrated that high *E2F1* expression and low *TP53* expression were associated with worse prognosis in hepatocellular carcinoma. Analysis of a previously published gene signature showed large differences in LOCC scoring. Optimization of the gene signature by selecting a subset of genes shows similar significance and hazard ratio stratification of the risk groups.

Conclusions: LOCC is a novel tool for defining prognostic significance that aids our understanding and selection of cutoffs, particularly for gene expression analysis in cancer.

Impact: LOCC can be used in prognosis studies to select and comprehend variables and cutoffs.

Introduction

In medicine, continuous markers refer to those that can take on a range of continuous values, rather than discrete categories. These markers can provide important information about the health status of an individual and can be used in a variety of medical applications, including diagnosis, prognosis, and treatment planning (1). Examples of

continuous markers in medicine include laboratory test results such as blood glucose levels, cholesterol levels, and white blood cell count. For cancer, examples include tumor size, gene expression, and many other biomarkers such as prostate specific antigen (1,2). Use of these markers can be helpful for both the patient and physician.

The interpretation of continuous markers can also be challenging, and this is especially true for new biomarkers that are just being tested. The most important question for new biomarkers is where to draw the line to divide groups (3,4). While traditional methods have used median or quartiles, more recent computational methods choose the cutoff that has the most significant p-value (1,3). However, in both cases, there is a significant loss of information as all other possible cutoffs are ignored and only information using the specific cutoff is relayed to the user. There are other computational tools which also try to convey information about the p-value or hazard ratio over the range of the continuous variable yet we have not found a comprehensive method that allows visualization of the full picture (3,5).

Current calculation methods employed to evaluate models for prognosis and outcome are also challenged by limitations. A popular method is use of receiver operator characteristic (ROC) curve, a plot of sensitivity and specificity (6,7). The area under the curve (AUC) provides information whether the variable has good performance at distinguishing the outcome. While ROC is appropriate for definitive diagnosis, it lacks discrimination and calibration in most prognostic studies (6-9). For example, a PCR test for HIV would be appropriate for ROC evaluation but single gene expression in cancer biopsy is too little information to determine the consequences. Hence, there is a need for better calculations for continuous variables relating to outcomes.

Therefore, we developed Luo's Optimization Categorization Curves (LOCC) to help visualize more information for better cutoff selection and understanding of the importance of the continuous variable against the measured outcome. We describe the process of understanding and generating LOCC with practical survival curve examples using real data from The Cancer Genome Atlas (TCGA, 10). We also reanalyzed published data and optimize gene signatures using a LOCC score to evaluate the gene expression significance and impact on prognosis. We believe LOCC will lead to improved analysis and understanding of continuous variables in many biological settings.

Methods

Data Sources

We used TCGA data from cBioportal.com (11) and LIRI-JP (Liver hepatocellular carcinoma – Japan) data from International Cancer Genome Consortium (ICGC). Z-scores of gene expression analyzed by RNA-Seq by Expected Maximization (RSEM) for TCGA was used while RNA-sequencing data was acquired in normalized read counts for LIRI-JP data. Normalization of read counts to z-scores was performed for each sample by subtracting the individual expression by the mean expression and dividing by standard deviation. Mutation data was acquired for TCGA data and considered mutant for any non-conservative mutation.

LOCC Visualization

To generate the LOCC ranking graphs, data was processed in R. Samples were ordered by expression z-scores and then graphed using ggplot. A horizontal line representing the ideal or selected cutoff was added as needed.

To generate the LOCC cutoff selection, we analyzed possible categorization by analyzing every cutoff according to the gene expression that would result in a different grouping of patient samples and calculating the corresponding hazard ratio (HR) and p-values. R package survival was used to calculate the hazard ratios and p-values. The HR was calculated with a cox proportional hazard regression model. A log-rank test was used to evaluate p-values. For n number of patients, we would perform n-1 number of cuts to calculate corresponding HR and p-values. We graph the HR and p-values for each cutoff that has a minimum of 5% of the total population in each group. For the optimal cutoff, the cutoff with the lowest p-value in which each group has at least 10% of the population is selected. The cutoff was confirmed with cutpointR (3). Code for LOCC algorithm will be available on publication.

LOCC Score

The LOCC score is made of three numeric components: a significance aspect, a range aspect, and an impact aspect. The significance is represented by the -log (p-value). The range is represented by the percentage of cutoffs that has a p-value below 0.01. The impact is represented by the highest HR, which includes ranking expression from highest to lowest as well as lowest to highest. For significance and impact, the numbers are restricted to cutoffs such that at least 10% of the population is in each group. The three numbers are multiplied together for the LOCC score.

RISK Gene Signature

The risk score analysis for hepatocellular was used from a previous publication (12) using the same gene expression and weight coefficients. Cox regression analysis was

calculated with Cox proportional hazard package in R. ROC analysis was conducted with ROCR package. Gene expression correlations were performed from cBioportal.com (11).

Results

LOCC demonstrates *E2F1* expression is associated with a poor prognosis in hepatocellular carcinoma

To demonstrate the utility of LOCC, we used TCGA data to showcase how LOCC can help analyze the role of transcription factor *E2F1* in liver hepatocellular carcinoma (TCGA LIHC). *E2F1* is an important transcription factor that has roles in cell cycle, DNA repair, and even apoptosis (13,14). *E2F1* can bind p53 to induce apoptosis and can also be inhibited by the retinoblastoma protein (Rb) to arrest the cell at the G1/S checkpoint (13,14). As such, *E2F1* is an important target in cancer where it is often overexpressed.

LOCC is a comprehensive visualization of multiple parameters for continuous variables (See Supplemental Figure 1 for detailed labeling of LOCC). The first graph, the LOCC ranking, plots the values of the continuous variable against the ranking of all samples. In our case, we plot *E2F1* z-score expression in the tumor samples on the y-axis while we show the ranking of the samples on the x-axis (Fig 1A). We use the z-score because many datasets have different methods of standardizing RNA-Seq data, but we use a normalization method to approximate the distribution. We also plot mutations of this gene with different colors to see if mutations affect gene expression. Next, we graph the LOCC cutoff selection, which plots the hazard ratio at every single cutoff (Fig 1B). We observe the hazard ratio (black line) is almost always above 1 (red line) which means that a higher *E2F1* expression is almost always associated with a worse prognosis in LIHC. However, an important question is whether *E2F1* expression is truly significant at any

point. To answer this question, we add a second line defined by its own y-axis scale, -log (p-value), on the hazard ratio graph (Fig 1C). The key values for the yellow – log (p-value) line is $p = 0.01$ (red horizontal line) and $p = 0.05$ (green horizontal line). Therefore, whenever the yellow line is above the green line, it is statistically significant ($p < 0.05$) and above the red line is very significant ($p < 0.01$). In this case, the ideal cutoff should be between the top 50% and top 75% of *E2F1* expression. Using a lowest p-value approach, a cutoff at 61.6% yields the lowest p value and seems appropriate. At the same time, the exact *E2F1* expression at this cutoff can be visualized using the first graph shown in Fig. 1C. We also create the Kaplan-Meier curve at that cutoff to see its appropriateness and calculate a median survival (Fig 1D). Previous literature has demonstrated *E2F1* gene expression is a prognostic marker for hepatocellular carcinoma using a median expression cutoff (15).

Validation of *E2F1* as prognostic biomarker in the LIRI-JP hepatocellular dataset

After finding a very significant cutoff and setting the cutoff, we need to validate this result in another cohort of hepatocellular carcinoma. For this, we used data from the International Cancer Genome Consortium (ICGC) which had a large Japanese cohort of liver hepatocellular carcinoma (LIRI-JP). The LIRI-JP data gives normalized read counts which is different from the RNA-Seq by expectation maximization (RSEM) used by TCGA. While it is possible to reprocess the raw data such that both groups are using the same expression format, that would require tremendous effort. Instead, for this example, we will use the normalized z-score to estimate the cutoff which should approximately judge the relative expression between patients, though it is imperfect. While the z-scores range and distribution may vary significantly between populations, we hope to capture a similar

categorization of the population in which we believe there are at least two groups that can be stratified by *E2F1* expression.

We used the *E2F1* cutoff from the TCGA training of -0.305 z-score and set that as the cutoff for the validation cohort (Fig 2A). Using LOCC, we can analyze how significant and appropriate that cutoff is (Fig 2B). Even though the cutoff is not quite the highest peak of significance, it is still highly significant so the cutoff is still adequate. Using the cutoff, we generate a Kaplan-Meier plot to visualize the survival curves between the two groups (Fig 2C). With significant survival differences in both TCGA and LIRI-JP cohorts using the same cutoff, it would be appropriate to suggest that *E2F1* is a prognostic biomarker for LIHC associated with a poor prognosis.

Low expression and mutations of *TP53* are associated with a poor prognosis

Next, we demonstrate LOCC utility in helping evaluate gene expression in genes with prognostic significant mutations. For TCGA LIHC, *TP53* mutation is prognostic significant for survival as patients with mutations have a worse prognosis (p – 0.04, Fig 3A). However, there are other mechanisms leading to disruption of the p53 pathway besides mutations and LOCC can help evaluate them. Low *TP53* expression is also significantly associated with a poor prognosis (p – 0.0005, Fig 3B). A z-score *TP53* expression plot demonstrates that *TP53* mutations are associated with low expression of *TP53* (Fig 3C). We also ranked *TP53* expression from low to high so the LOCC graph will show a significant HR above 1, reducing overlapping lines and making it easier to read the graph (Fig 3D). Therefore, both expression and mutation of TP53 can be used in combination to help determine the prognosis of patients with LIHC, which is in line with previous studies (16,17).

LOCC score helps rank prognostic importance of predictors

We have also developed a LOCC score to help judge significance and overall impact of any single or group gene expression. Three factors are multiplied together to get the LOCC score: the significance, the range, and the impact. The significance is calculated by taking the -log (lowest p-value). The range is the percentage of LOCC significance line (yellow line) that is above the red line ($p < 0.01$). Finally, the impact is the highest HR on the LOCC graph. We also limit group sizes for significance and impact to be at least 10% of the population to minimize extreme effects from small samples. Thus, the higher the LOCC score, the more critical and predictive the expression is for prognosis. We use the LOCC score to rank the 12 individual genes from a previously published prognostic gene signature (RISK) in hepatocellular carcinoma (Table 1, (12).

We found the top eight gene expression by LOCC score (*KIF20A, TTK, TPX2, LCAT, SPP1, HMMR, CYP2C9, ANXA10*) were all associated with a worse prognosis and had a significantly higher LOCC score than the other 4 genes ($\log_{10}$ LOCC p-value < 0.01, Table 1). This was different from ROC analysis in the article where the AUC of the 12 genes were all between 0.57 and 0.63 and 10 genes between 0.60 and 0.63 (12). Meanwhile, cox regression analysis also found a significant difference in the -log (p-value) between the 8 genes from LOCC analysis and the other 4 genes ($p = 0.0007$, Table 1). We tested if this new gene signature (8-gene RISK) would be as useful as the original 12 as it contained the most significant genes according to LOCC scoring. To do this, we used the same weight coefficient as the original article and used it to calculate the 8-gene RISK score (12). Indeed, we found that the optimized gene signature had a similar p-value and HR between the high and low risk groups compared to the original 12 gene signature (Fig

4A-D, Supplemental Figure 2A-C). Furthermore, the LOCC score, AUC of the ROC curves, and cox regression analysis of the 8-gene RISK and original 12-gene RISK produced very similar numbers. (Fig 4D and Supplemental Fig 2D-G). Overall, from the extensive analysis and comparisons, we believe the gene signature from the top 8 genes by LOCC score is non-inferior to the original 12 gene original signature demonstrating LOCC score ability to decipher the key predictors.

Finally, we show that it is possible to use multiple cutoffs if there are multiple peaks on the yellow line above the red line. We chose two peaks on the LOCC cutoff selection and as such, separated patients into three well-stratified groups of high, middle, and low risk (Fig 4B and E).

Discussion

In this study, we demonstrate a novel visualization tool (LOCC) that presents the distribution, significance, and hazard ratios related to a continuous variable. Using this information, either an optimized cutoff or multiple cutoffs can be chosen while seeing the whole distribution. There are a number of advantages to LOCC over previous methods. By combining hazard ratio and p-value into a single graph, one can simultaneously understand both the significance and effect of a variable. The LOCC ranking can show the distribution and mutation status of the samples. These characteristics of LOCC help analysis of the variable in relationship to the outcome and mutation status such as the case with *TP53*. Therefore, we believe this LOCC visualization tool can be useful to many continuous variables to outcome analyses (9).

We proposed a LOCC score that is an interesting metric for ranking variables by significance, range, and impact. Using a combination of three variables, we can generate

a score to help determine the most valuable variables related to the outcome. The LOCC score appears to be better for prognostic studies than ROC as the AUC cannot discriminate between variables if only modest differences exists. While the LOCC score is correlated with the cox regression coefficients and -log(p-values) (Sup Fig S3A), we believe LOCC score provides more information particularly for prognosis. One example of the difference between the cox regression p-value and LOCC score is the rankings of *KIF20A*, *TTK*, and *TPX20,* where the LOCC score values having very significant cutoffs while cox regression favors having a more consistent HR (Table 1). While the LOCC score does favor significant cutoffs, it is not as singular as using only the most significant cutoff as required by some established methods (3,5). Instead as a hybrid score, LOCC is more holistic and can function without adjustment, though future studies will be needed to evaluate LOCC score utility in full gene expression prognosis analysis.

There are still many limitations and issues to be solved in the field of bioinformatics with regard to categorization of continuous variables. Focusing on cancer and gene expression, LOCC can help with ranking variables and visualizing the relevance of the variable and outcome, but needs better multi-variable integration. Currently, we are using weighted coefficients from the original study which used LASSO (Least Absolute Shrinkage Selector Operator) regression. While many models like LASSO use single weight approaches, new methods may be needed for optimizing LOCC score and prognosis(18). This is true because the LASSO regression discards variables that may be active predictors especially if predictors are correlated or if the number of predictors is not significantly larger than the number of samples (18,19). Many highly prognostic genes are strongly correlated: for example, *KIF20A*, *TTK*, and *TPX20* in hepatocellular

carcinoma all have correlation coefficients above 0.85 (Sup Fig S3B). LASSO regression may discard one or more of these variables during its regression process. Furthermore, it is clear from LOCC visualization that gene expression may be significant throughout the entire spectrum or only for a certain range. Thus, a continuous weighted score is appropriate for some variables but unsuitable for others. Biologically, this is understandable because a certain level of protein or gene expression may be required for or representative of activation while lower levels do not signify any meaningful biological difference. Thus, a better method of integrating these variables is anticipated and remains a challenge in research (20).

One valuable aspect of LOCC that is not described in this study is the potential for this tool to find similar patterns between variables. Genes regulated by the same pathway and transcription factor may have similar LOCC scores as expression patterns are similar. Thus, it is possible to evaluate a pathway by choosing downstream target genes and determining the effect of the pathway on the outcome. This similarity between variables is not easily evaluated by cox regression or ROC curves, but LOCC visualization and score can help in this process. However, this requires many basic science studies to understand the pathway, and the impact of genes that are often regulated by multiple pathways, leading to overlapping signals.

Limitations

LOCC and other cutoff selection methods are prone to overfitting and the p-values selected at the best cutoff should be not be interpreted as typical p-values. Nevertheless, we believe LOCC and its scoring can help alleviate simple overfitting by single point cutoff

because it visualizes the entire distribution of p-values and hazard ratios. Additionally, the LOCC score incorporates a range component which helps combat overfitting at a specific cutoff. Ultimately, the best way to verify a result is through external validation with another dataset.

Conclusion

LOCC is a visualization tool that can appropriately display the vast information of continuous variables with regard to survival or prognosis. LOCC can help with cutoff selection and analyze the extent in which a variable is significantly associated with an outcome. The LOCC score can help identify variables that best associate with survival throughout the entire range and rank predictors that allows for discrimination of good and poor predictors.

Data Availability

Gene expression and mutation data was accessed from cBioportal.com for TCGA LIHC and https://dcc.icgc.org/projects/LIRI-JP for LIKI-JP. Code for LOCC will be available at the time of publication.

Author Disclosures

The authors have no disclosures.

Acknowledgments

G.L is supported by NIH MSTP training grant 5T32GM007250. J.J.L. is supported by the Jane and Lee Seidman Chair in Pediatric Cancer Innovation.

References:

1. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol **2012**;12:21 doi 10.1186/1471-2288-12-21.
2. Eng KH, Schiller E, Morrell K. On representing the prognostic value of continuous gene expression biomarkers with the restricted mean survival curve. Oncotarget **2015**;6(34):36308-18 doi 10.18632/oncotarget.6121.
3. Budczies J, Klauschen F, Sinn BV, Győrffy B, Schmitt WD, Darb-Esfahani S*, et al.* Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. PLoS One **2012**;7(12):e51862 doi 10.1371/journal.pone.0051862.
4. Altman DG, Royston P. The cost of dichotomising continuous variables. Bmj **2006**;332(7549):1080 doi 10.1136/bmj.332.7549.1080.
5. Nagy Á, Munkácsy G, Győrffy B. Pancancer survival analysis of cancer hallmark genes. Sci Rep **2021**;11(1):6047 doi 10.1038/s41598-021-84787-5.
6. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation **2007**;115(5):654-7 doi 10.1161/circulationaha.105.594929.
7. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem **2008**;54(1):17-23 doi 10.1373/clinchem.2007.096529.
8. Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Brief Bioinform **2012**;13(1):83-97 doi 10.1093/bib/bbr008.
9. Grund B, Sabin C. Analysis of biomarker data: logs, odds ratios, and receiver operating characteristic curves. Curr Opin HIV AIDS **2010**;5(6):473-9 doi 10.1097/COH.0b013e32833ed742.
10. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell **2017**;169(7):1327-41.e23 doi 10.1016/j.cell.2017.05.046.
11. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA*, et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov **2012**;2(5):401-4 doi 10.1158/2159-8290.cd-12-0095.
12. Ouyang G, Yi B, Pan G, Chen X. A robust twelve-gene signature for prognosis prediction of hepatocellular carcinoma. Cancer Cell Int **2020**;20:207 doi 10.1186/s12935-020-01294-9.
13. Kent LN, Leone G. The broken cycle: E2F dysfunction in cancer. Nat Rev Cancer **2019**;19(6):326-38 doi 10.1038/s41568-019-0143-7.
14. Meng P, Ghosh R. Transcription addiction: can we garner the Yin and Yang functions of E2F1 for cancer therapy? Cell Death Dis **2014**;5(8):e1360 doi 10.1038/cddis.2014.326.
15. Huang YL, Ning G, Chen LB, Lian YF, Gu YR, Wang JL*, et al.* Promising diagnostic and prognostic value of E2Fs in human hepatocellular carcinoma. Cancer Manag Res **2019**;11:1725-40 doi 10.2147/cmar.s182001.
16. Donehower LA, Soussi T, Korkut A, Liu Y, Schultz A, Cardenas M*, et al.* Integrated Analysis of TP53 Gene and Pathway Alterations in The Cancer

Genome Atlas. Cell Rep **2019**;28(5):1370-84.e5 doi 10.1016/j.celrep.2019.07.001.
17. Yang Y, Qu Y, Li Z, Tan Z, Lei Y, Bai S. Identification of Novel Characteristics in TP53-Mutant Hepatocellular Carcinoma Using Bioinformatics. Front Genet **2022**;13:874805 doi 10.3389/fgene.2022.874805.
18. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw **2010**;33(1):1-22.
19. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J*, et al.* Strong rules for discarding predictors in lasso-type problems. J R Stat Soc Series B Stat Methodol **2012**;74(2):245-66 doi 10.1111/j.1467-9868.2011.01004.x.
20. McDermott JE, Wang J, Mitchell H, Webb-Robertson BJ, Hafen R, Ramey J*, et al.* Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. Expert Opin Med Diagn **2013**;7(1):37-51 doi 10.1517/17530059.2012.718329.

FIGURE LEGENDS

Figure 1. LOCC demonstrates *E2F1* is associated with a poor prognosis in TCGA hepatocellular carcinoma. (A) The z-score expression of *E2F1* from TCGA hepatocellular carcinoma patient samples was ordered in descending order and plotted against the ranking of the samples. Samples with non-conservative mutations of *E2F1* are colored in orange while samples with wildtype *E2F1* are colored turquoise. **(B)** A black line depicting the hazard ratio (1.0) is plotted for every cutoff for *E2F1* expression. A red horizontal line is place at HR = 1.0. (**C**) A yellow line depicting the -log (p-value) is added to the graph to display the significance of each cutoff. The red horizontal line is also aligned with p = 0.01 while the green horizontal line is aligned with p = 0.05. The cutoff with the lowest p-value is selected to be the ideal cutoff, indicated by the arrow which corresponds to a z-score of -0.3. (**D**) A Kaplan-Meier overall survival curve is plotted at the ideal cutoff to separate patients into high or low *E2F1*. A data table of the details of the groups is shown. Patients' survival times are expressed in months. P-

values are calculated using log-rank test. HR are calculated using Cox proportional hazard regression.

Figure 2. Validation of *E2F1* as prognostic biomarker in the LIRI-JP hepatocellular dataset (A) The Z-score expression of *E2F1* from LIRI-JP hepatocellular carcinoma patient samples was ordered in descending order and plotted against the ranking of the samples. A horizontal line is graphed at -0.3 to separate patients into high and low *E2F1* groups using the TCGA dataset cutoff. **(B)** The LOCC cutoff selection was graphed for LIRI-JP samples. The cutoff from TCGA data was used to separate patients into high and low *E2F1* groups. (**C**) A Kaplan-Meier overall survival curve is plotted at the validation cutoff to separate patients into high or low *E2F1*. A data table of the details of the groups is shown. Patients' survival times are expressed in days. P-values are calculated using log-rank test. HR are calculated using Cox proportional hazard regression.

Figure 3. Low expression and mutations of *TP53* are associated with a poor prognosis in TCGA hepatocellular carcinoma (A) The Kaplan-Meier survival curve is plotted for *TP53* mutations. **(B)** The Kaplan-Meier survival curve is plotted for *TP53* expression with groups determined by the ideal cutoff in LOCC visualization. (**C**) The LOCC ranking is graphed for *TP53* expression z-score from lowest expression to highest expression. Samples with *TP53* mutations are colored in orange while samples with wildtype *TP53* are colored in turquoise. (**D**) The LOCC cutoff selection is graphed for TCGA hepatocellular *TP53* expression. The ideal cutoff is selected and this

corresponds to the *TP53* z-score of -0.78. Patients' survival times are expressed in days. P-values are calculated using log-rank test. HR are calculated using Cox proportional hazard regression.

Figure 4. LOCC score helps rank prognostic importance of predictors (A) The 8-gene modified RISK score was ordered in descending order and plotted against the ranking of the samples. A horizontal line is graphed at 2.00 to separate patients into high and low risk groups using ideal LOCC cutoff. **(B)** The LOCC cutoff selection was graphed for the 8-gene RISK score using TCGA hepatocellular data. The most significant cutoff was chosen to separate patients into two groups. In addition, another cutoff at 1.077 was selected for dividing patients into three groups. (**C**) A Kaplan-Meier overall survival curve is plotted at the validation cutoff to separate patients into high or low risk score. (**D**) A comparison of the RISK score of 8 leading genes by LOCC score and the original 12 gene RISK score is shown. The lowest p-value and hazard ratio (HR) are selected with a minimum of at least 10% of the total samples in each group. (**E**) Low risk patients are further stratified using the LOCC cutoff selector into middle and low risk. A corresponding Kaplan-Meier graph is plotted for the three different risk group. Patients' survival times are expressed in months. P-values are calculated using log-rank test. HR are calculated using Cox proportional hazard regression.

Table 1. LOCC score of RISK gene signature

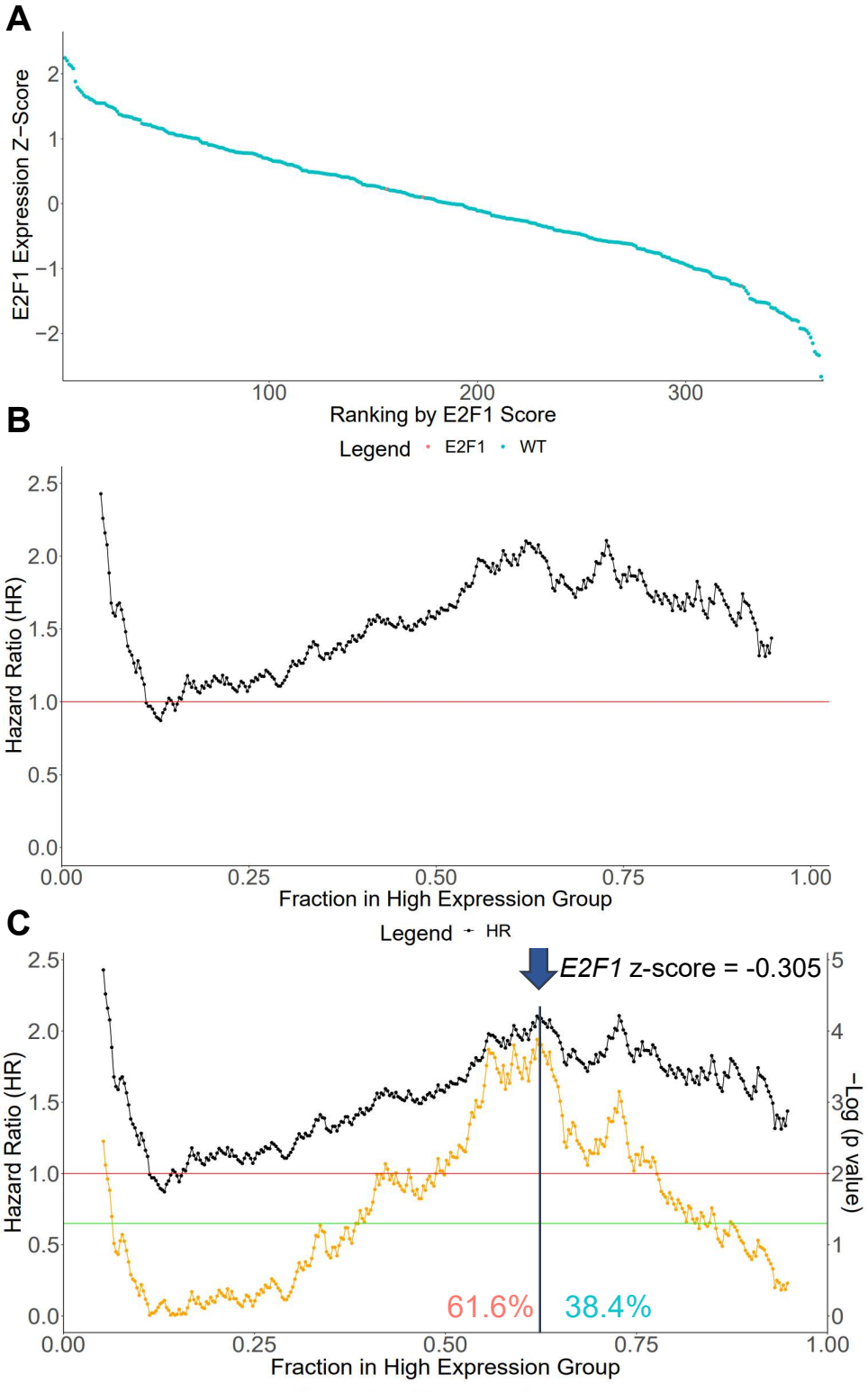
SUPPLEMENTAL FIGURE LEGENDS

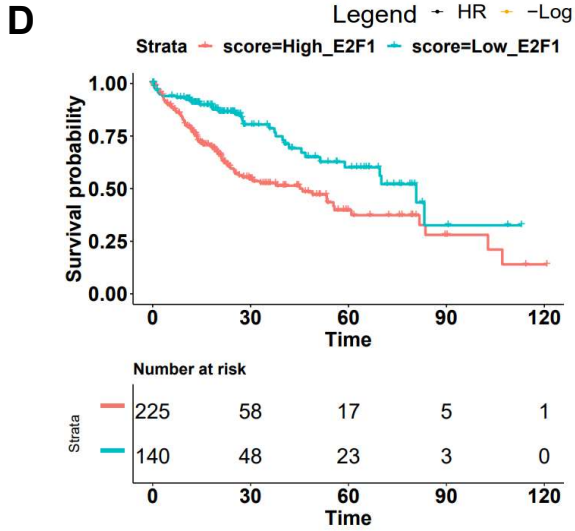
Supplemental Figure 1. Detailed description of LOCC graphs. The LOCC ranking is a graph that plots individual samples values and their respective ranking across all samples. The samples are color coded according to mutation status. The x-axis of ranking directly corresponds to the x-axis of the LOCC cutoff selection graph such that each sample are the same. Therefore, a vertical line from the cutoff selection matches the sample expression and ranking on the LOCC ranking. The cutoff selection has many lines labeled and color coded for ease of visualization. Ideally, the LOCC ranking and cutoff selection are set up so that the hazard ratio is above the red line (HR = 1 as it is easier to visualize HR above 1 than HR below 1. The best cutoff is usually the one with the most significance which is the lowest p-value, visually shown by a peak in the yellow line. However, exceptions can occur if one peak is very close to the edge while comparable peaks exist closer to the middle of the graph. After the cutoff is selected, a corresponding Kaplan-Meier curve is graphed to match the cutoff. The two groups from the Kaplan-Meier curve correspond to the two groups separated by the cutoff on the LOCC ranking. Theoretically, a Kaplan-Meier graph can be plotted for every point on the LOCC graphs which is why there is a large amount of information contained in LOCC.

Supplemental Figure 2. Analysis of 12-gene original RISK score using LOCC (A) The 12-gene original RISK score was ordered in descending order and plotted against the ranking of the samples. A horizontal line is graphed at 2.05 to separate patients into high and low risk groups using ideal LOCC cutoff. **(B)** The LOCC cutoff selection was

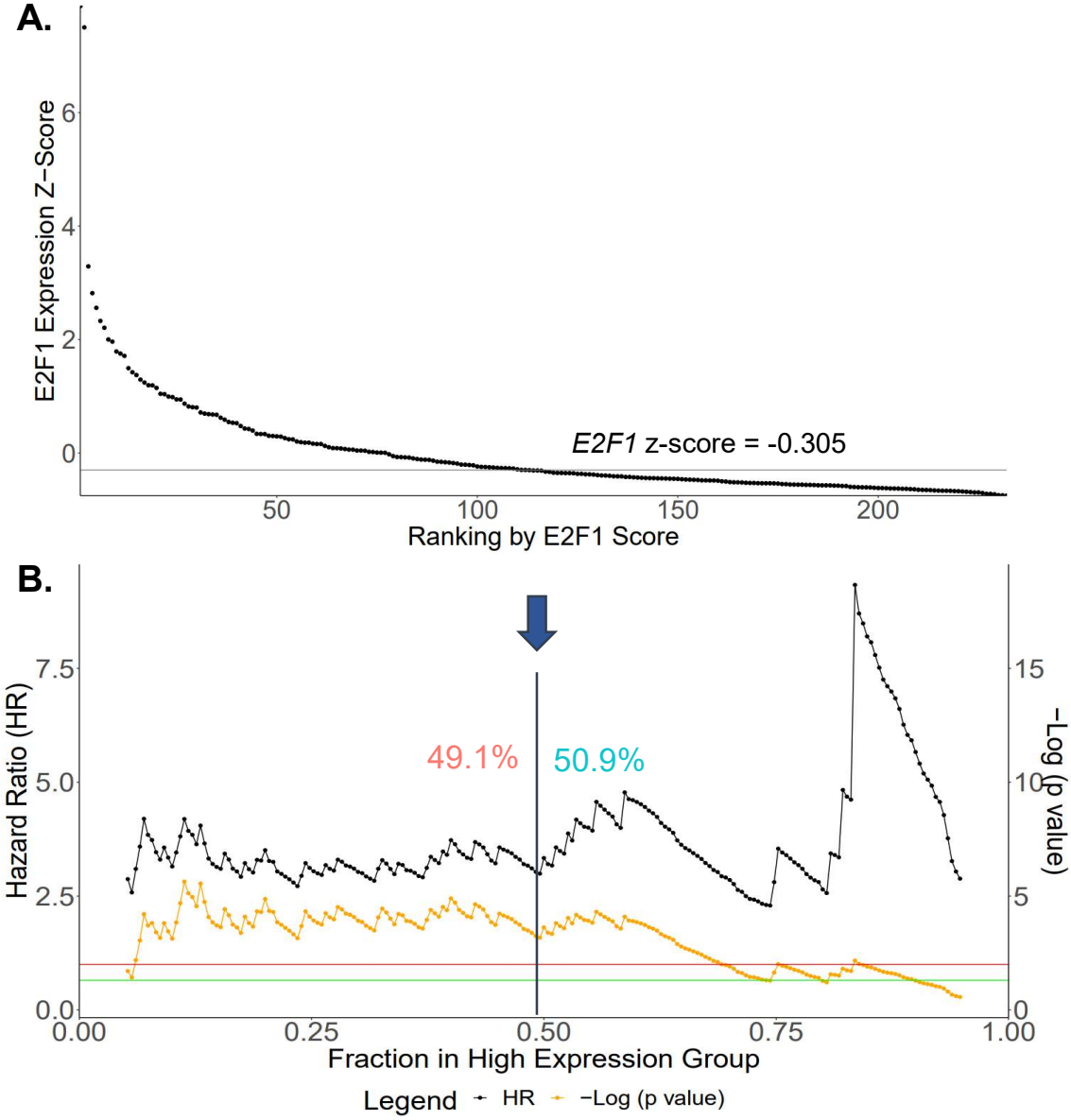
graphed for the 12-gene RISK score using TCGA hepatocellular data. The most significant cutoff was chosen to separate patients into two groups. (**C**) A Kaplan-Meier overall survival curve is plotted at the validation cutoff to separate patients into high or low risk score. (**D**- **E**) Time-dependent ROC curve of risk score model for 1- and 3-year overall survival predictions were plotted for the original and modified 8-gene RISK scores. (**F**) A full LOCC score analysis is shown comparing the original 12-gene RISK score and the modified 8-gene RISK score. Patients' survival times are expressed in months. P-values are calculated using log-rank test. HR are calculated using Cox proportional hazard regression.

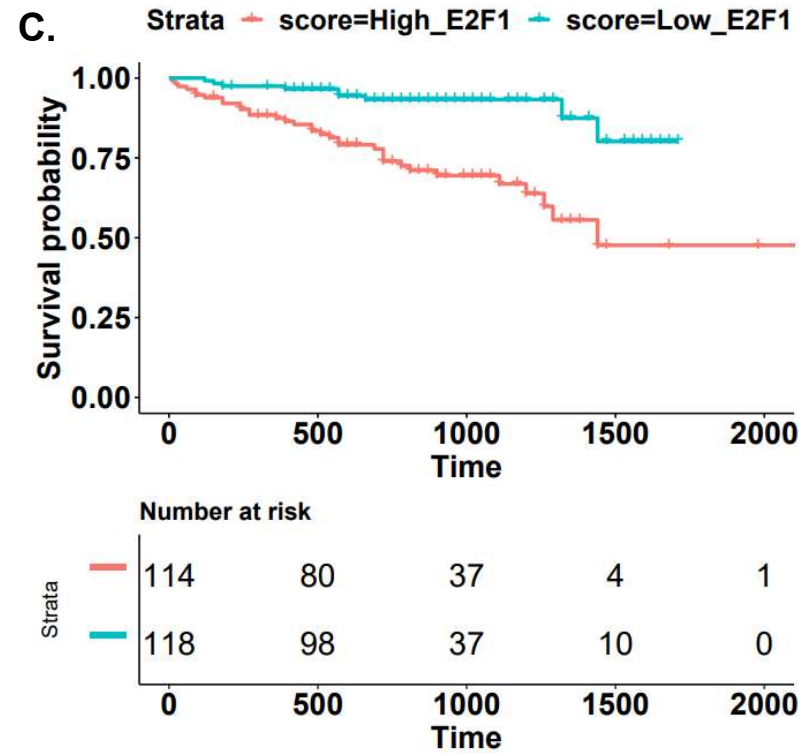
Supplemental Figure 3. Correlations between LOCC score and cox regression (A) A scatterplot is plotted between the Log_2(LOCC Score) and -Log_{10} (cox p-value). There is a strong correlation (r = 0.705) between the two markers. **B-D.** Scatterplots of gene expression are plotted between the following genes: (**B**) TTK vs TPX2, (**C**) TTK vs KIF20A, (**D**) KIF20A vs TPX2.



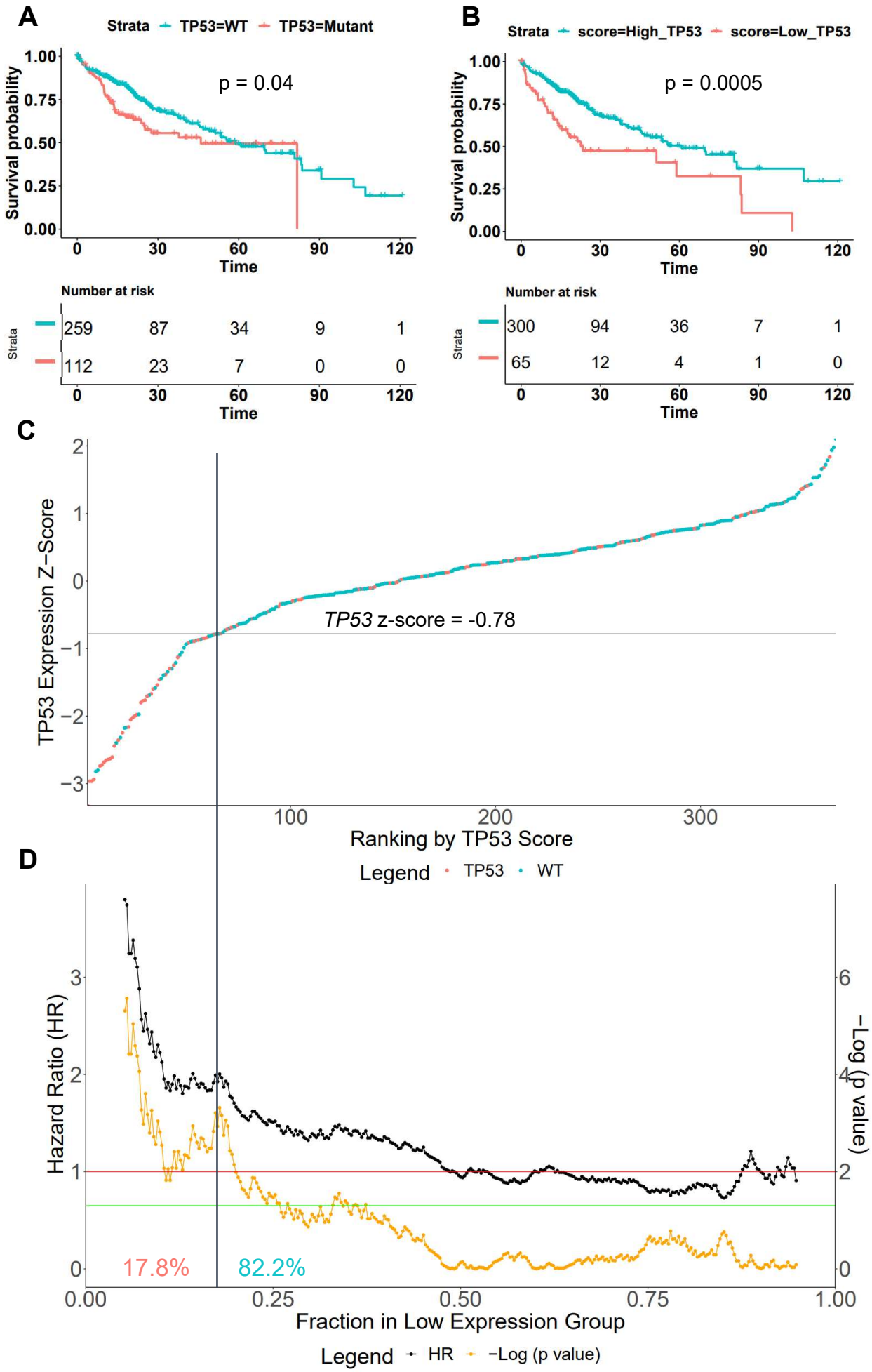


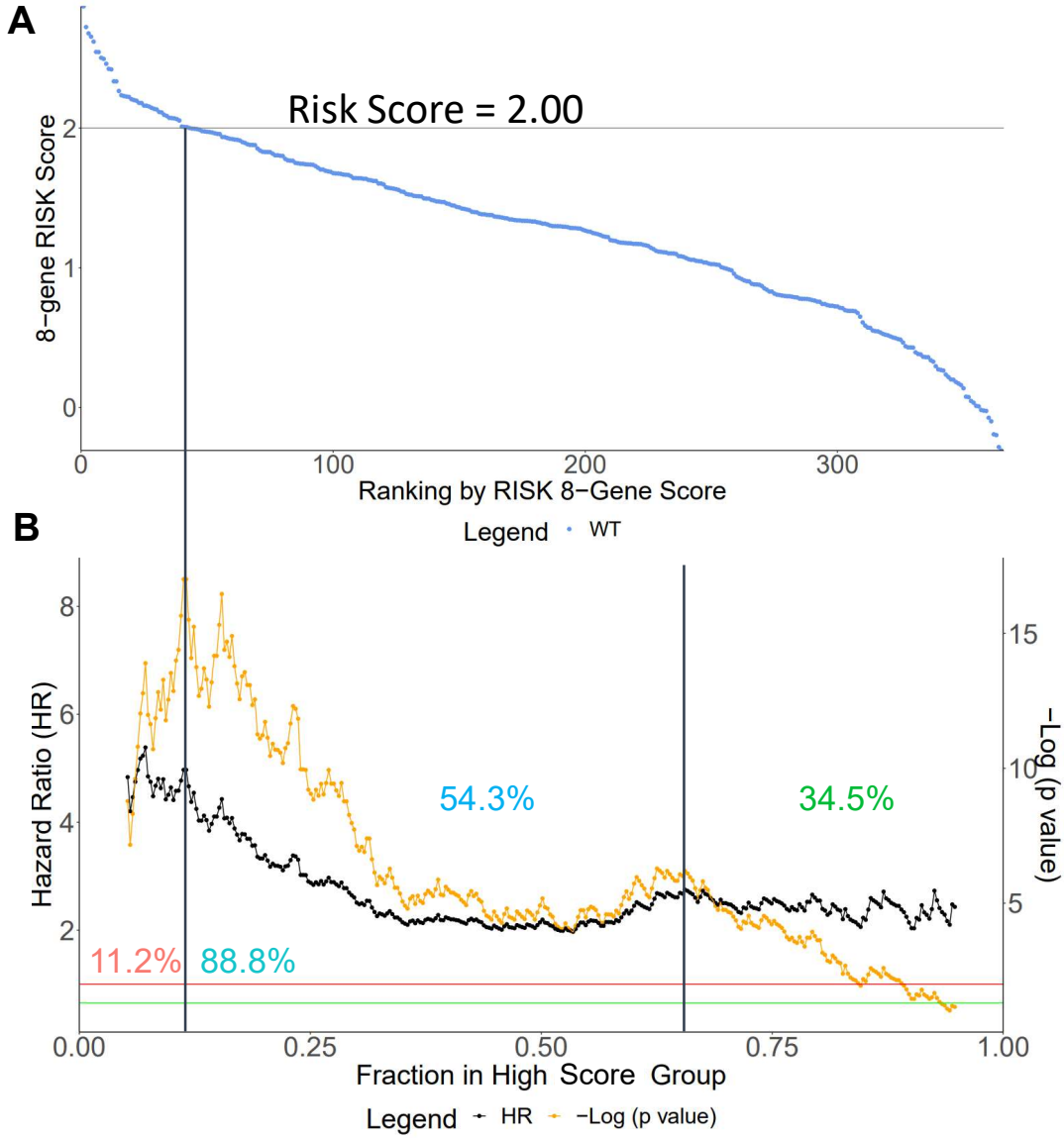
p = 0.0001	n	events	median
score=High_E2F1	225	95	45.9
score=Low_E2F1	140	34	80.7

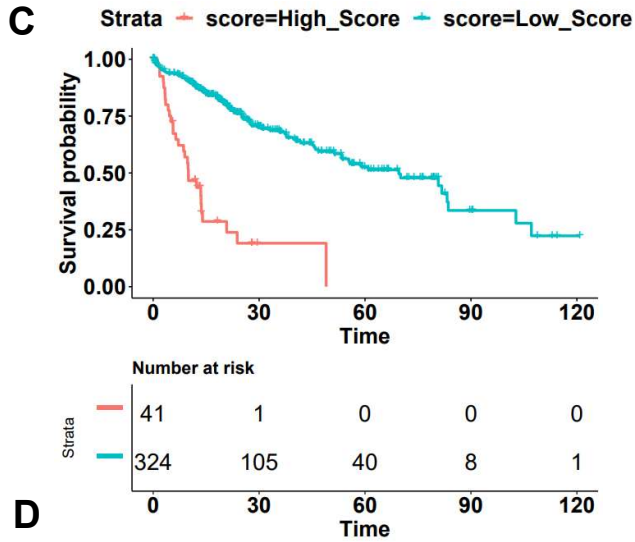


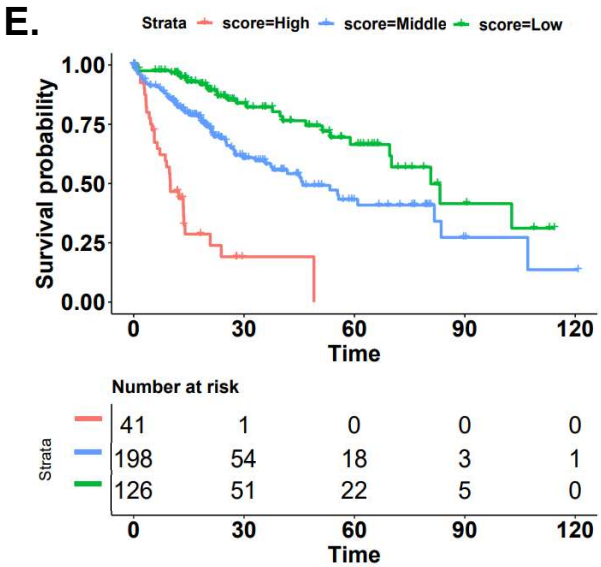


p = 0.0001	n	events	median
score=High_E2F1	114	34	NA
score=Low_E2F1	118	9	NA


A
Strata TP53=WT TP53=Mutant
p = 0.04
Survival probability
Time
Number at risk
Strata
259 87 34 9 1
112 23 7 0 0
B
Strata score=High_TP53 score=Low_TP53
p = 0.0005
Survival probability
Time
Number at risk
Strata
300 94 36 7 1
65 12 4 1 0
C
TP53 Expression Z−Score
TP53 z-score = -0.78
Ranking by TP53 Score
Legend TP53 WT
D
Hazard Ratio (HR)
−Log (p value)
17.8%
82.2%
Fraction in Low Expression Group
Legend HR −Log (p value)





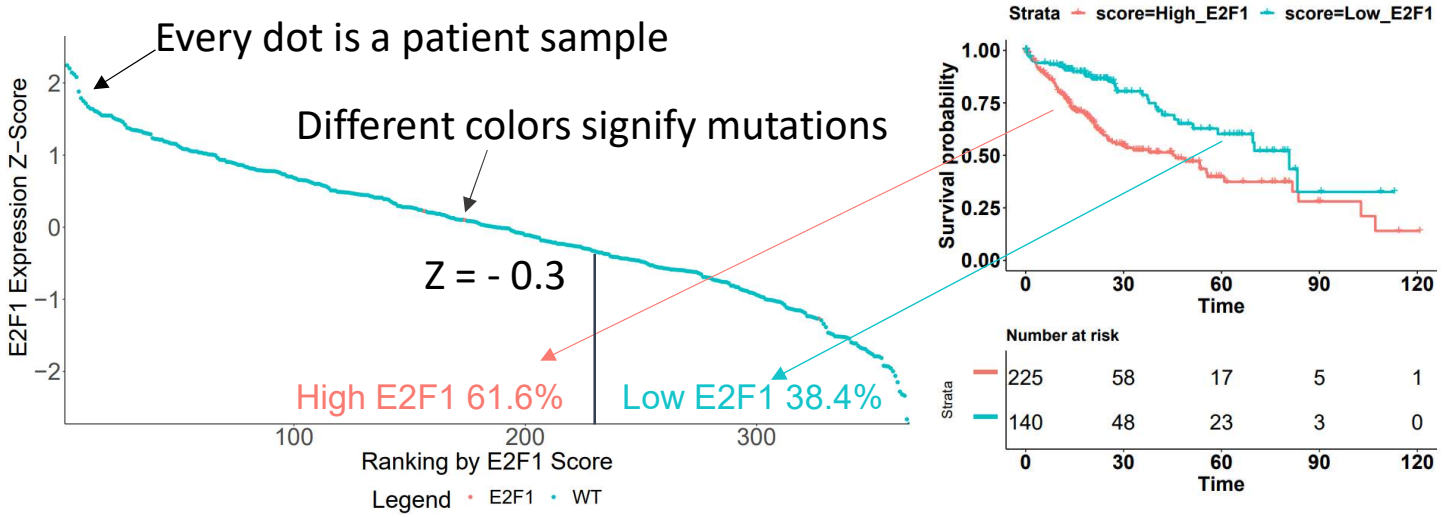


D

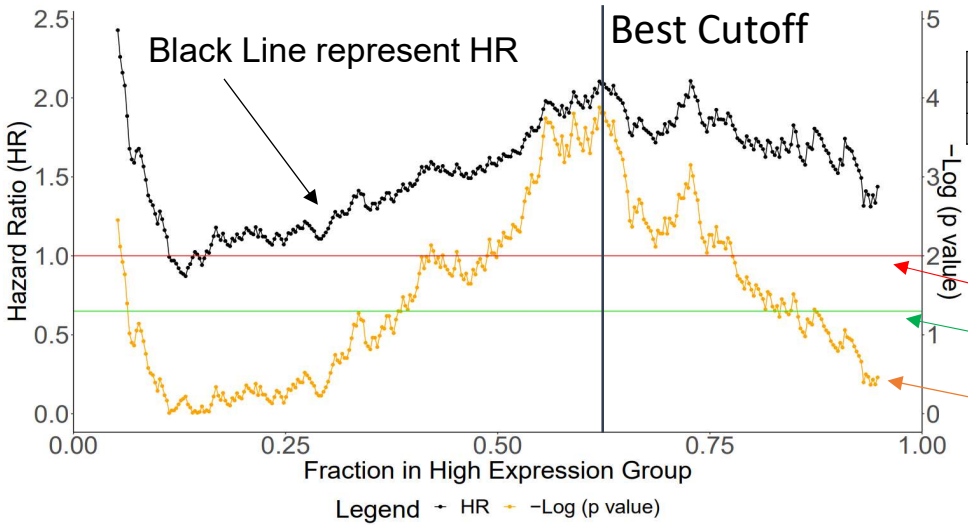
Gene Set	Lowest p value	Estimated HR at best cutoff	LOCC Score
8 Leading Genes	<2E-16	5.0	73.6
12 Original Genes	<2E-16	4.7	67.6

Gene	-log (lowest p value)	Highly significantly Range (p <0.01)	Highest HR	Lowest HR	LOCC Score	Log_2 LOCC Score	Good or Poor Prognosis	Cox Regression exp(β)	Absolute Value (β)	Cox Regression p value	-log(cox p value)
KIF20A	16.097	0.711	4.094	1.470	46.842	5.550	Poor	1.320	0.278	0.000006	5.22
TTK	9.874	0.694	3.291	1.232	22.558	4.496	Poor	1.260	0.231	0.00002	4.70
TPX2	5.650	0.818	2.761	1.661	12.763	3.674	Poor	1.375	0.318	0.000004	5.40
LCAT	8.077	0.598	0.714	0.381	12.665	3.663	Good	0.795	0.229	0.000001	6.00
SPP1	5.692	0.799	2.415	1.740	10.983	3.457	Poor	1.125	0.118	0.000004	5.40
HMMR	5.363	0.780	2.480	1.417	10.369	3.374	Poor	1.299	0.262	0.00002	4.70
CYP2C9	5.599	0.658	0.677	0.410	8.983	3.167	Good	0.900	0.105	0.0001	4.00
ANXA10	4.993	0.738	0.781	0.420	8.779	3.134	Good	0.887	0.120	0.0001	4.00
LAPTM4B	4.836	0.543	2.532	1.444	6.646	2.733	Poor	1.270	0.239	0.0003	3.52
RDH16	5.332	0.562	0.689	0.452	6.634	2.730	Good	0.897	0.109	0.0002	3.70
LECT2	5.041	0.501	0.847	0.430	5.877	2.555	Good	0.920	0.083	0.0002	3.70
MAGEA6	3.622	0.394	2.518	0.874	3.593	1.845	Poor	1.060	0.058	0.001	3.00
Top 8 vs Bot 4 (t test p value)	0.066	0.003	0.482	0.385	0.047	0.004		0.499	0.135	0.134	0.00070

LOCC Ranking



LOCC Cutoff Selection



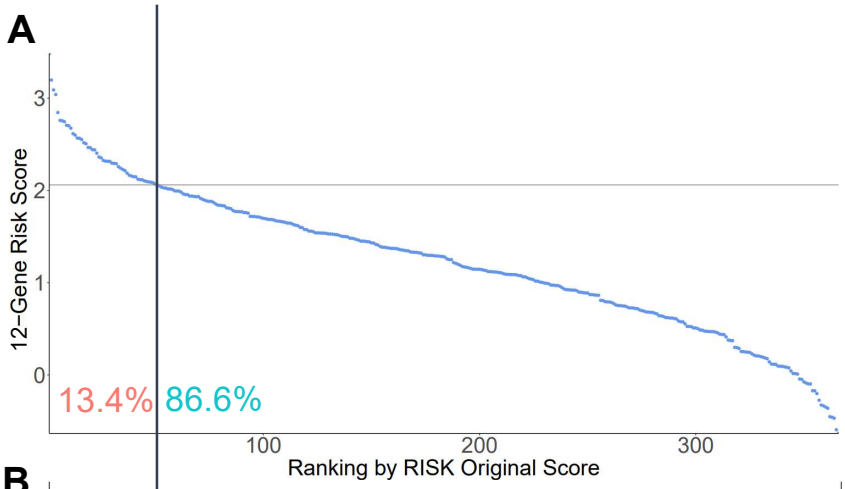
p = 0.0001	n	events	median
score=High_E2F1	225	95	45.9
score=Low_E2F1	140	34	80.7

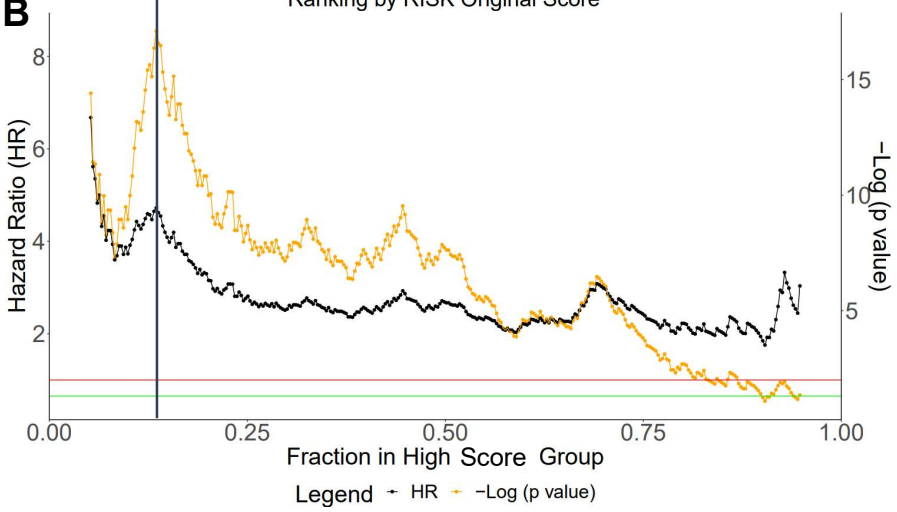
Red line represent p = 0.01
And HR = 1

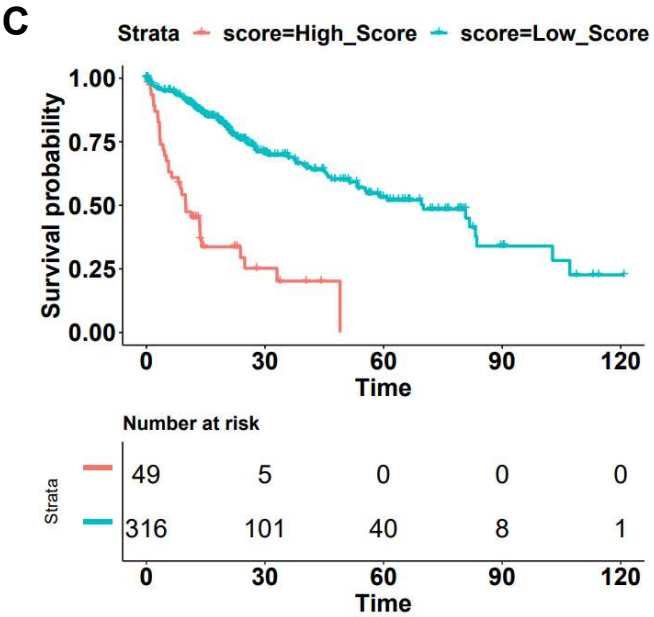
Green line represent p = 0.05

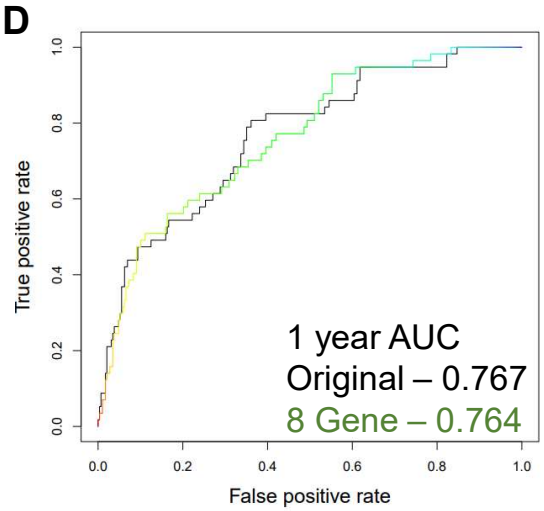
Yellow line represent significance

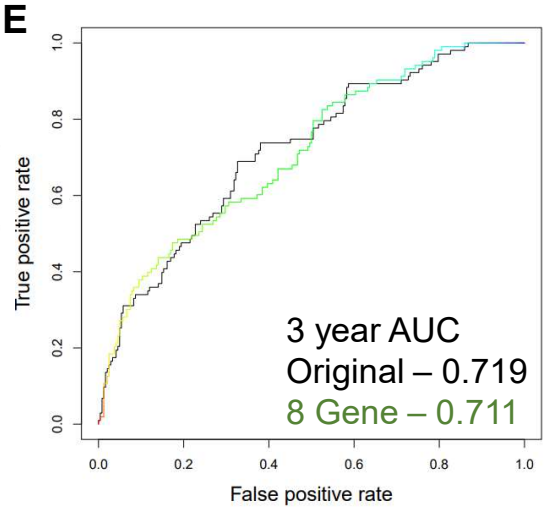
All tumor samples are ranked by E2F1 expression
Both graphs' samples and x-axis are lined up









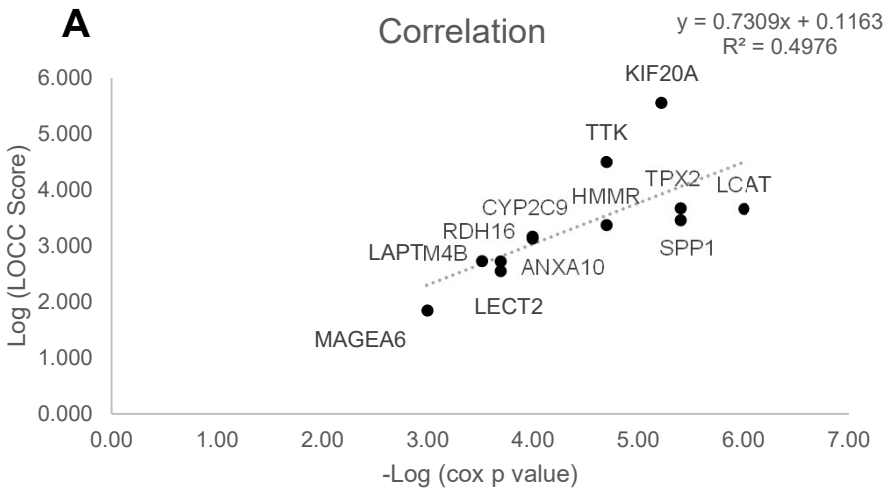


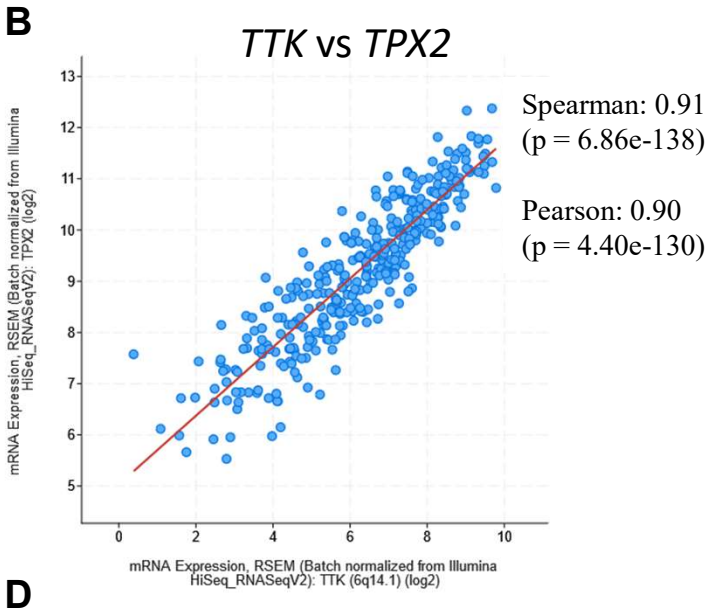
F

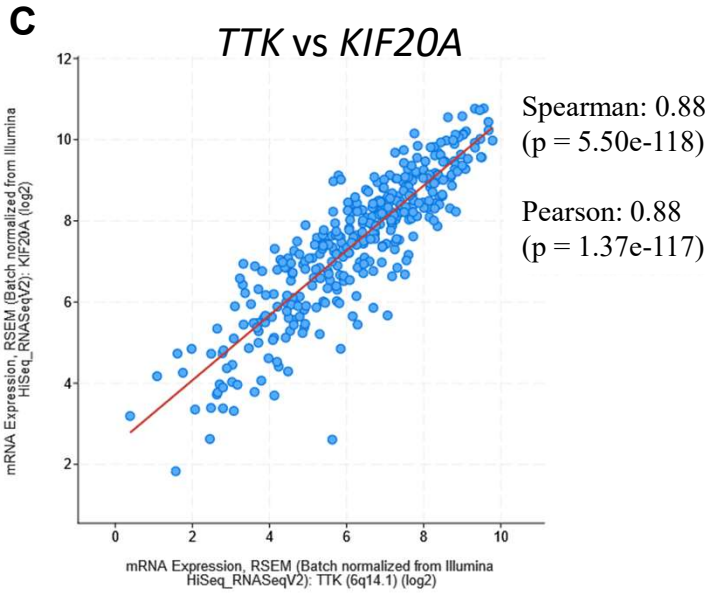
LOCC Analysis	-log (p value)	Percentage (<0.01)	Highest HR	Lowest HR	LOCC Score
Original Risk	17.093	0.837	4.723	1.942	67.6
8-Gene Risk	17.008	0.871	4.971	1.972	73.6

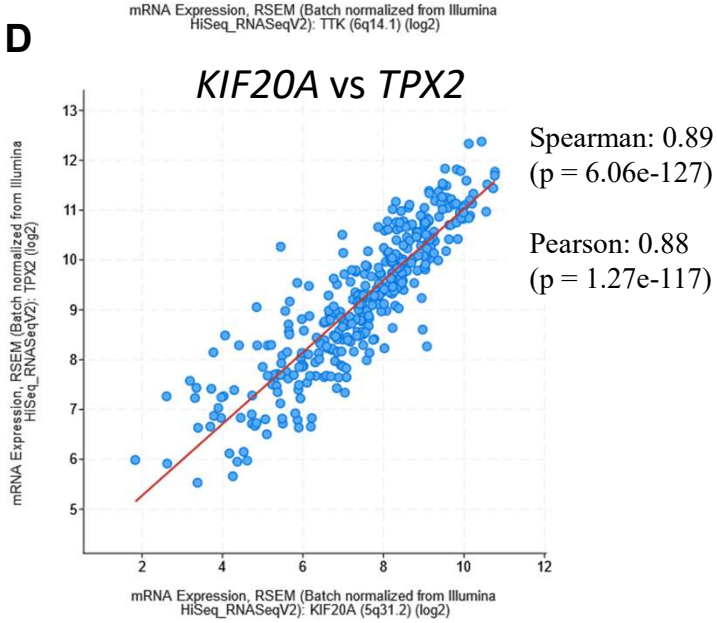
G

Cox Regression	β Coefficient	Exp (β)	Concordance	Log rank test (p value)
Original Risk	0.893	2.44	0.699	1.0E-11
8-Gene Risk	1.04	2.82	0.696	2.0E-10


A
Correlation
y = 0.7309x + 0.1163
R² = 0.4976
KIF20A
TTK
TPX2
LCAT
HMMR
CYP2C9
RDH16
SPP1
LAPTM4B
ANXA10
LECT2
MAGEA6
Log (LOCC Score)
-Log (cox p value)


B
TTK vs TPX2
Spearman: 0.91
(p = 6.86e-138)
Pearson: 0.90
(p = 4.40e-130)
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): TPX2 (log2)
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): TTK (6q14.1) (log2)


C
TTK vs KIF20A
Spearman: 0.88
(p = 5.50e-118)
Pearson: 0.88
(p = 1.37e-117)
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): KIF20A (log2)
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): TTK (6q14.1) (log2)


D
KIF20A vs TPX2
Spearman: 0.89
(p = 6.06e-127)
Pearson: 0.88
(p = 1.27e-117)
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): TPX2 (log2)
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): KIF20A (5q31.2) (log2)