

RESEARCH

Open Access



LOCC: a novel visualization and scoring of cutoffs for continuous variables with hepatocellular carcinoma prognosis as an example

George Luo^{1*}, Toby Chen² and John J. Letterio^{3,4,5}

*Correspondence:
gxl263@case.edu

¹ Department of Pathology, Case Western Reserve University School of Medicine, 2103 Cornell Rd., Wolstein Research Bldg. Rm 3501, Cleveland, OH 44106, USA

² School of Medicine, University of Michigan, Ann Arbor, MI, USA

³ The Angie Fowler Adolescent and Young Adult Cancer Institute, University Hospitals Rainbow Babies & Children's Hospital, Cleveland, OH, USA

⁴ The Case Comprehensive Cancer Center, Cleveland, OH, USA

⁵ Department of Pediatrics, Case Western Reserve University, Cleveland, OH, USA

Abstract

Background: The interpretation of large datasets, such as The Cancer Genome Atlas (TCGA), for scientific and research purposes, remains challenging despite their public availability. In this study, we focused on identifying gene expression profiles most relevant to patient prognosis and aimed to develop a method and database to address this issue. To achieve this, we introduced Luo's Optimization Categorization Curve (LOCC), an innovative tool for visualizing and scoring continuous variables against dichotomous outcomes. To demonstrate the efficacy of LOCC using real-world data, we analyzed gene expression profiles and patient data from TCGA hepatocellular carcinoma samples.

Results: To showcase LOCC, we demonstrate an optimal cutoff for *E2F1* expression in hepatocellular carcinoma, which was subsequently validated in an independent cohort. Compared to ROC curves and their AUC, LOCC offered a superior description of the predictive value of *E2F1* expression across various cancer types. The LOCC score, comprised of factors representing significance, range, and impact of the biomarker, facilitated the ranking of all gene expression profiles in hepatocellular carcinoma, aiding in the evaluation and understanding of previously published prognostic gene signatures. We also demonstrate that LOCC does not have the same assumptions required of Cox proportional hazards modeling for accurate analysis. Repeated sampling demonstrated that LOCC scores outperformed ROC's AUC in discriminating predictors from non-predictors. Additionally, gene set enrichment analysis revealed significant associations between certain genes and prognosis, such as E2F target genes and G2M checkpoint with poor prognosis, and bile acid metabolism and oxidative phosphorylation with good prognosis.

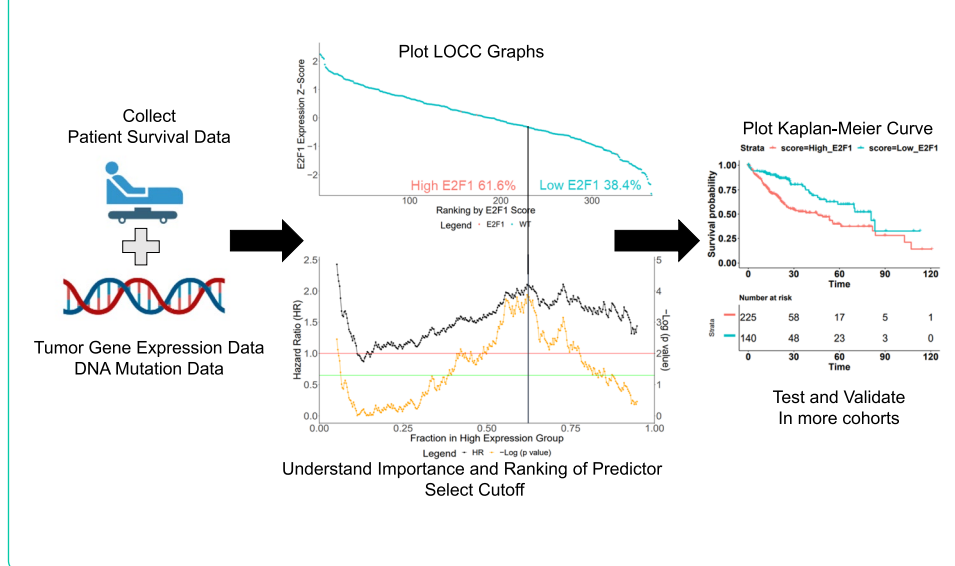
Conclusion: In summary, we present LOCC as a novel visualization tool for the analysis of gene expression in cancer, particularly for understanding and selecting cutoffs. Our findings suggest that LOCC scores, which effectively rank genes based on their prognostic potential, represent a more suitable approach than ROC curves and Cox proportional hazard for prognostic modeling and understanding in cancer gene expression analysis. LOCC holds promise as an invaluable tool for advancing precision medicine and furthering biomarker research. Further research regarding multivariable



integration and validation will help LOCC reach its full potential and establish its utility across diverse cancer types and clinical settings.

Keywords: LOCC, Cutoff, Prognosis, Gene expression, Score

Graphical abstract



Background

In the field of medicine, continuous markers play a crucial role in assessing an individual's health status, offering valuable information through a range of continuous values rather than discrete categories. These markers find extensive application in medical practice, aiding in diagnosis, prognosis, and treatment planning [1]. Examples of such markers include blood glucose levels, cholesterol levels, white blood cell count, and various biomarkers like gene expression and tumor size in cancer patients [1, 2].

Currently, there are clinical tests that utilize gene expression for cancer prognostic interpretation and treatment planning, such as Oncotype DX and MammaPrint [3, 4]. However, despite the excitement surrounding gene profiling for hepatocellular carcinoma, gene signatures face challenges related to their functional connections and validation in independent cohorts, preventing their adoption in clinical settings [5–10].

Interpreting continuous markers, especially new biomarkers, can be challenging, primarily when determining the optimal threshold to divide groups [11, 12]. In the clinical setting, knowing that variables affect outcomes is not enough as it is critical to identify relevant cut-offs that can be used for clinical decision-making. Traditional methods, like using median or quartiles, and more recent computational methods using significant p values, often result in a loss of information and fail to convey the full picture [1, 11, 13].

Current calculation methods, such as receiver operating characteristic (ROC) curves, have limitations in evaluating prognosis and outcome models effectively [14–17]. While ROC is useful for definitive diagnoses, it lacks discrimination and calibration in most prognostic studies, demanding better calculations for continuous variables and their outcomes [14, 15]. Moreover, interpreting the meaning of the ROC area under the curve (AUC) for prognosis, where values are typically around 0.6, remains unclear.

Another commonly referenced methodology for examining the relationship between continuous variables and their associated outcomes is the Cox proportional hazard (Cox PH) model [18]. The Cox PH model is a powerful tool to analyze survival data and establish the effects of a continuous variable across an entire distribution; however, it relies on assumptions of proportionality and linearity [19]. As a result, its applicability to prognostic studies needs to be validated as it may not always hold true [19, 20].

To address these challenges, we have developed Luo's Optimization Categorization Curves (LOCC), an innovative tool that visualizes more information for improved cut-off selection and understanding the significance of continuous variables in relation to measured outcomes. LOCC visualizes and uses information from the entire dataset to create a graph and score to explain the prognostic cutoffs and significance of the biomarker of interest. In this paper, we present the process of generating and interpreting LOCC using practical survival curve examples with real data from The Cancer Genome Atlas [21]. We compare LOCC with ROC curve analysis and demonstrate how LOCC scores better represent prognostic value in the context of gene expression, using *E2F1* as an example. Cox PH modeling was also included for comparison which highlighted the potential issues with assumptions of proportional hazard and linearity. Meanwhile, LOCC investigates the prognostic value of gene expression through categorically evaluating cut-offs within a continuous gene expression distribution with little reliance on proportionality hazards and better illustrated biomarker prognostic potential compared to Cox PH numbers. As a result, while the Cox PH model is often invaluable for survival analysis, its methodological constraints and assumption requirements make it a less direct comparison for LOCC and less accurate for cancer biomarker analysis. ROC AUC is a standard and widely accepted metric for assessing the performance of diagnostic tests and predictive models across various thresholds, making it a more relevant comparison for our LOCC approach but still we demonstrate LOCC has significant advantage in interpretability and robustness compared to ROC.

Additionally, we reanalyze several published gene signatures, illustrating how LOCC scores effectively rank each gene's prognostic value. Moreover, we showcase how optimizing a gene signature with LOCC scores can simplify the model without compromising predictive power. The consistency and robustness of LOCC scoring are further validated through various sampling of the TCGA dataset. Finally, we explore pathways associated with prognosis in hepatocellular carcinoma.

We firmly believe that LOCC can revolutionize the analysis and understanding of continuous variables in various biological settings, providing an invaluable contribution to medical research and biomarker analysis. Through LOCC's enhanced visualization and scoring capabilities, medical professionals can better understand biomarkers and form hypotheses to solve medical issues.

Methods

Data sources

Our study utilized two primary data sources: The Cancer Genome Atlas (TCGA) data from cBioportal.com [22] and LIRI-JP (Liver hepatocellular carcinoma – Japan) data from the International Cancer Genome Consortium (ICGC). The TCGA data was processed using z-scores of gene expression analyzed by RNA-Seq by Expected

Maximization (RSEM). On the other hand, the LIRI-JP data was processed using Fragments Per Kilobase of transcript per Million mapped reads (FPKM), which was then converted to transcripts per million (TPM). Subsequently, z-scores for each sample were computed by subtracting individual expression by mean expression and dividing by the standard deviation. For TCGA data, mutation data was considered as a mutant for any non-silent mutation, which is defined as a mutation that results in different amino acids.

LOCC visualization

To generate the LOCC ranking graphs, data was processed in R. We ordered samples by expression z-scores and graphed them using ggplot. As necessary, a line representing the selected or ideal cutoff was added.

To determine the LOCC cutoff, we considered every possible categorization that would result in a distinct grouping of patient samples. We then calculated the corresponding hazard ratio (HR) and p values using the survival package in R. The HR was calculated using a Cox proportional hazard regression model, with p values evaluated via a log-rank test in R using survdiff function. The HR and p values were computed for each cutoff which resulted in at least 10% of the total population in each group. We selected the optimal cutoff with the lowest p value where each group had at least 10% of the population. We restricted the cutoffs to be at least 10% of the dataset because we wanted to reduce bias on the extremes from impacting the LOCC score. The cutoff was further verified with cutpointR [11]. For evaluation, we considered genes that were expressed in at least 20% of tumor samples.

In instances where numerous samples exhibited no gene expression, we applied a decrement of 0.1 to both the HR value and the $-\log(p$ value) for every such sample, continuing this adjustment until the HR value reached 1 and the $-\log(p$ value) reached 0. This adjustment served as a penalty for genes with widespread lack of expression, ensuring a more balanced assessment. We used these penalty numbers as they applied some reduction to the LOCC scores to these genes but not to the worst case scenario. The LOCC algorithm code is accessible in the data availability section.

LOCC score

The LOCC score is composed of three numeric components: a significance aspect, a range aspect, and an impact aspect. The three factors were chosen as they illustrate important insights about the data set. The significance aspect is denoted by $-\log(p$ value), which illustrates the greatest statistical significance possible within the data; the range aspect is the percentage of cutoffs with a p value below 0.01 to exemplify the general strength of the relationship between the variables in the data set; the impact is the highest HR, which showcases the peak predictive power of the continuous variable. A p value of 0.01 was chosen for the range aspect as a p value less than 0.01 is a commonly used cut-off for significance in biomedical research. For significance and impact, the numbers are restricted to cutoffs such that at least 10% of the population is in each group. This restriction is to ensure that all cutoffs continue to include a substantial proportion of the population and to minimize the possibility of extremes and outliers portrayed by a nominal proportion of the sample, as these outliers are not representative of the population and can substantially skew HR_p and p_l values. Through extensive

observation of the data, we identified that 10% is the optimal restriction to ensure statistical robustness while also minimizing biases by including a substantial proportion of patients.

The LOCC incorporates significance, range, and impact and is calculated by multiplying these three components together, represented by the equation:

$$\text{LOCC Score} = p_l * R_s * HR_p$$

where p_l is the highest value of $-\log(p \text{ value})$, R_s is the percentage of cutoffs that have a highly significant p value ($p < 0.01$), and HR_p is the HR at the most significant point. The resulting LOCC score allows for ordering of variables from lowest to highest or vice versa, ensuring the most significant HR is above one, or its reciprocal if the most significant HR is below one.

For segregating poor prognostic markers from good prognostic markers, we assigned negative scores to all poor prognostic markers when ranking all gene expressions by LOCC score. Thus, the most positive LOCC scores were linked with good prognosis, while the most negative scores corresponded to poor prognosis. Genes expressed in at least 20% of tumor samples were evaluated. In the absence of gene expression in any tumor samples, we followed the same incremental decrease approach as mentioned in the LOCC visualization section.

LOCC cutoff estimated p value and q value

Traditional methods of estimating and interpreting p values would not work well for significance values since they are the lowest p value from a range of p values that are related to each other. To estimate a p value and q value for the most significant cutoffs for all the gene expression profiles, we applied Monte Carlo methods to estimate significance values, p_p , from randomly generated data. The randomly generated data used the TCGA LIHC data but instead of being ordered by gene expression values, it was ordered through a random number generator in R. Thus, although the data had the same samples and survival times, the order was random which made it comparable to the existing results. We then calculated the LOCC scores for this randomly generated data and recorded it. We did this 10,000 times to have a large enough dataset to compare to the existing results.

Using these 10,000 random simulations, we can estimate the p value according to the empirical p values calculations [23]. Afterward, we used the empirical p values generated to calculate q values using the `qvalue` package in R. These q values are False Discover Rate (FDR) adjusted p values.

Cox PH modeling

We employed Cox proportional hazards (Cox PH) regression modeling to assess the association between gene expression levels and overall survival in patients. The analysis was performed using the R programming language with the `survival` package.

We constructed Cox PH models for each gene by treating the gene expression level as the predictor variable.

The primary Cox PH model assumed a linear relationship between gene expression and the hazard of death. To test for non-linearity, we also fit a secondary Cox PH model

that included a quadratic term for gene expression. We then performed a likelihood ratio test to compare the linear and quadratic models, calculating a p values to evaluate the evidence for non-linearity.

To verify the proportional hazards assumption, we conducted a Schoenfeld residuals test using the `cox.zph` function, which provides a p values indicating whether the proportional hazards assumption holds for each gene. If the p value from this test was less than 0.05, it suggested a violation of the proportional hazards assumption.

For each gene, we stored the following statistics in a global summary table: the regression coefficient, the exponentiated coefficient, the standard error of the coefficient, the z -value, the p value for the gene's effect on survival, the p value from the non-linearity test, and the p value from the proportional hazards test. The analysis was performed for each gene across the entire dataset, resulting in a comprehensive evaluation of the potential prognostic significance of gene expression levels with respect to overall survival.

ROC curve

Upon processing the patient and tumor data, Receiver Operating Characteristic (ROC) curves were constructed using the `ROCR` package in R. The Area Under the Curve (AUC) was estimated using `ROCR`, which took into account overall survival, excluding patients with incomplete survival or expression data. For comparative purposes, a red line representing sensitivity = (1 - specificity) was included. To align with LOCC score ranking, we associated the highest AUC with good prognosis and the lowest AUC with poor prognosis. However, individual gene expression profiles or gene sets were required to have an AUC above 0.5. We evaluated genes that were expressed in at least 20% of tumor samples.

RISK gene signature

We applied the risk score analysis for hepatocellular carcinoma from a previous study [24] using the same gene expression and weight coefficients. Cox regression analysis was performed with the Cox proportional hazard package in R, and ROC analysis was conducted using the `ROCR` package. The proportionality assumption was verified using the function `cox.zph`. We selected patients who survived at least one month for ROC calculations. We used the Akaike information criterion (AIC) to assess the relative quality of models during Cox regression modeling, calculating AIC using `extractAIC`. In particular, for the original RISK gene signature and the 8-gene RISK gene signature, we used Cox regression proportional hazard modeling of patient survival with gene expression profiles of the original RISK or the 8-gene RISK. We then used `extractAIC` to find the AIC of this modeling. This AIC is the relative quality of the model with a lower number being higher quality relative to the number of variables. Gene expression correlations were derived from cBioportal.com [22].

2-Fold cross validation of TCGA dataset

For the twofold cross validation, the TCGA hepatocellular data set was randomly split into two random halves; one half was used to calculate the relevant cut-offs, whereas the other half was used for validation. The procedure is as follows: first, we randomly sampled half of the TCGA hepatocellular carcinoma data. With this half, we processed the

data and calculated LOCC scores, cutoffs, and ROC c-statistics. The ROC c-statistic is equivalent to the ROC AUC. We then examined the generalizability of our calculations by examining its validity within the other half of the dataset. This twofold cross validation procedure was replicated 100 times for each gene under evaluation. The cutoff with the lowest p value was chosen and tested in the validation set. The validation p value was recorded and deemed significant if it fell below 0.05.

Gene set enrichment analysis (GSEA)

After we ranked all genes by their LOCC score in hepatocellular carcinoma, we analyzed these genes and LOCC scores using pre-ranked GSEA to understand what pathways are associated with prognosis. Utilizing pre-ranked GSEA, we investigated hallmark gene sets and ranked them by their false discovery rate (FDR). FDR is a proven method to identify and minimize false positives; compared to its alternatives, FDR has been shown to be a more powerful and consistent approach, and its applicability to computational biology and genomics has been illustrated in prior research [25]. Additionally, the normalized enrichment score (NES) and p values were recorded.

To evaluate gene sets without overlapping genes, we identified and removed all overlapping genes between the gene sets. Then we used a custom gene set in pre-ranked GSEA to evaluate the FDR, NES, and p values. This will reduce the chance that a gene set is significant only due to other genes from another significant gene set.

We also evaluated gene sets of randomly selected genes to ensure p values followed a uniform distribution. We generated 100 gene sets of 200 randomly selected genes from the TCGA expression list. We used GSEA preranked with LOCC scores to evaluate the p values from these gene sets. We compiled the p values and graphed them in Microsoft Excel and used a line of best fit to evaluate uniformity.

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 [26, 27]. *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 [26, 27]. As such, *E2F1* is an important target in cancer where it is often overexpressed.

LOCC offers a comprehensive visualization of multiple parameters for continuous variables (refer to Supplemental Fig. S1 for detailed LOCC labeling). The initial graph, the LOCC ranking, presents the values of the continuous variable versus the ranking of all samples. In our instance, we plotted *E2F1* z-score expression in the tumor samples on the y-axis, and the sample rankings on the x-axis (Fig. 1A). We employed the z-score because various datasets adopt different standardization methods for RNA-Seq data, hence, we used a normalization method to approximate the distribution. We also highlighted gene mutations in different colors to observe if they affect gene expression.

The LOCC cutoff selection graph, which plots the hazard ratio at every single cutoff (Fig. 1B), reveals that the hazard ratio (black line) is nearly always above 1 (red line), suggesting that higher *E2F1* expression consistently correlates with a worse prognosis in LIHC.

The critical inquiry is whether *E2F1* expression is truly significant at any point. To address this, we supplemented a second line, defined by its own y-axis scale, $-\log(p)$ value, on the hazard ratio graph (Fig. 1C). When the yellow line (representing $-\log(p)$ value) lies above the green line, it is statistically significant ($p < 0.05$), and above the red line is very significant ($p < 0.01$). Hence, the ideal cutoff should be within the top 50% to 75% of *E2F1* expression. Using the lowest p value approach, a cutoff at 61.6% is calculated to produce the lowest p value. The appropriateness of this cut-off can be confirmed via visual inspection. Simultaneously, the precise *E2F1* expression (z-score = -0.305) at this cutoff can be visualized using the first graph shown in Fig. 1A. We also generated a Kaplan–Meier curve at that cutoff to evaluate its appropriateness and calculate median survival (Fig. 1D, E). Previous literature corroborates the prognostic marker role of *E2F1* gene expression in hepatocellular carcinoma using a median expression cutoff [28].

We have also developed a LOCC score to help judge significance and overall impact of any single or group of predictors. Three factors are multiplied together to get the LOCC score: the significance, the range, and the impact. The significance value 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 HR at the most significant cutoff (if HR is below 1, use the reciprocal). 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. In our evaluation of *E2F1*, the LOCC score was calculated to be 2.58 (Fig. 1C), signifying there is some prognostic value in this dataset. In the following sections, the scale of the LOCC score will be established.

We generated a ROC curve using *E2F1* expression and survival status (Fig. 1F). With an area under the curve (AUC) of 0.578, this implies a heightened risk of death with increasing *E2F1* expression. Interestingly, we observed that the red line for LOCC (HR = 1) and ROC (True Positive Rate = False Positive Rate) are derived from the same underlying equation (Supplemental Fig. S2A–E). If the HR line remains above the red line throughout the graph, then the ROC curve will also be above the red line of random

(See figure on next page.)

Fig. 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 mutations of *E2F1* that modify the characteristics of their corresponding amino acids (non-silent mutations) are colored in orange while samples with wildtype *E2F1* are colored turquoise. **B** A black line depicting the hazard ratio is plotted for every cutoff for *E2F1* expression. A red horizontal line is placed 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 -0.305 . **D** A Kaplan–Meier overall survival curve is plotted at the ideal cutoff to separate patients into high or low *E2F1*. **E** 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. HRs are calculated using Cox proportional hazard regression. **F** A ROC is plotted for *E2F1*, and the AUC was calculated

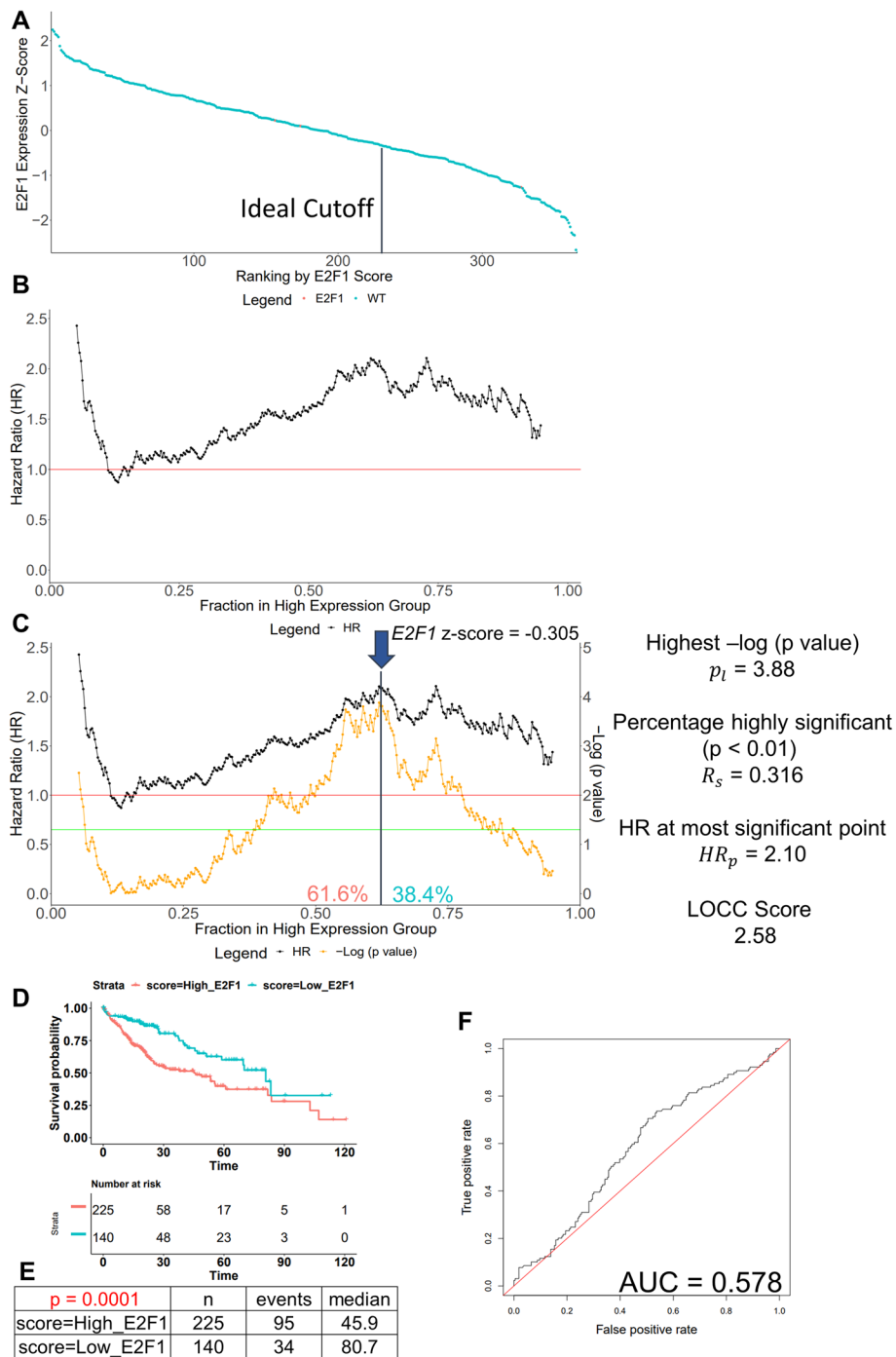


Fig. 1 (See legend on previous page.)

classifier. Thus, the relative positions (above/below) of the black and red lines should be similar in both LOCC and ROC curves.

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

Following the identification and establishment of a significant cutoff in the TCGA data, validation in another hepatocellular carcinoma cohort is necessary. To achieve this, we utilized data from the International Cancer Genome Consortium (ICGC), which includes a large Japanese cohort of liver hepatocellular carcinoma (LIRI-JP). The LIRI-JP data provides normalized read counts, represented as Fragments Per Kilobase of transcript per Million mapped reads (FPKM), which differ from the RNA-Seq by expectation maximization (RSEM) used by TCGA. While it is feasible to reprocess both dataset's raw data with the same programs to achieve a consistent expression format, this would be a highly demanding process. Therefore, we opted to use the normalized z-score to approximate the cutoff, despite its imperfections, which should give us a relative comparison of expression between patients.

Taking the *E2F1* cutoff of -0.305 z-score established in the TCGA training, we applied this to the validation cohort (Fig. 2A). Employing LOCC, we evaluated the significance and appropriateness of this cutoff (Fig. 2B). Despite the TCGA cutoff not being the lowest *p* value of this dataset, the cutoff remains highly significant, indicating the cutoff is still appropriate. We then generated a Kaplan–Meier plot using the TCGA cutoff, to compare the survival curves of the two groups (Fig. 2C). Given the significant survival differences observed in both TCGA and LIRI-JP cohorts using the same cutoff (Fig. 2D), it's reasonable to propose that *E2F1* is a prognostic biomarker for LIHC, associated with poor prognosis. The LOCC score, calculated to be 18.9 (Fig. 2B), implies strong prognostic potential of *E2F1* in this dataset.

The ROC curve indicates an AUC of 0.614, suggesting increased death risk associated with higher *E2F1* expression. However, determining the significance of this predictor or the optimal, significant cutoffs is challenging through the ROC curve alone. Therefore, the utility of ROC in evaluating *E2F1* expression as a prognostic biomarker is limited compared to LOCC.

Comparison of ROC and LOCC in evaluating prognostic biomarkers

The use of ROC curves and their corresponding Area Under the Curve (AUC) is ubiquitous in biomarker studies and prognostic modeling [14, 15]. However, the application and interpretation of ROC graphs and the c-statistic (or AUC) for prognostic purposes raise questions. In diagnostics, the c-statistic is akin to the probability that a randomly selected subject who experienced the event will have a higher test score than a subject who did not [29]. For prognosis, it represents the likelihood that a patient who succumbed to the disease had a higher test score than a patient who survived. Although appropriate for diagnosis, the interpretation of c-statistics is challenging for prognosis, as the binary classification of patients as alive or dead is oversimplistic for overall survival and unsuitable for censored and time-dependent data. Even the time-dependent ROC, designed to address this issue, still struggles with censoring [30].

LOCC, on the other hand, utilizes a cox regression model and examines hazard ratios between groups. Patients are classified into two categories based on the biomarker, enabling us to evaluate the hazard ratio and associated *p* values. Similar to ROC, LOCC is applied across all potential cutoffs and the resulting hazard ratios and *p* values are

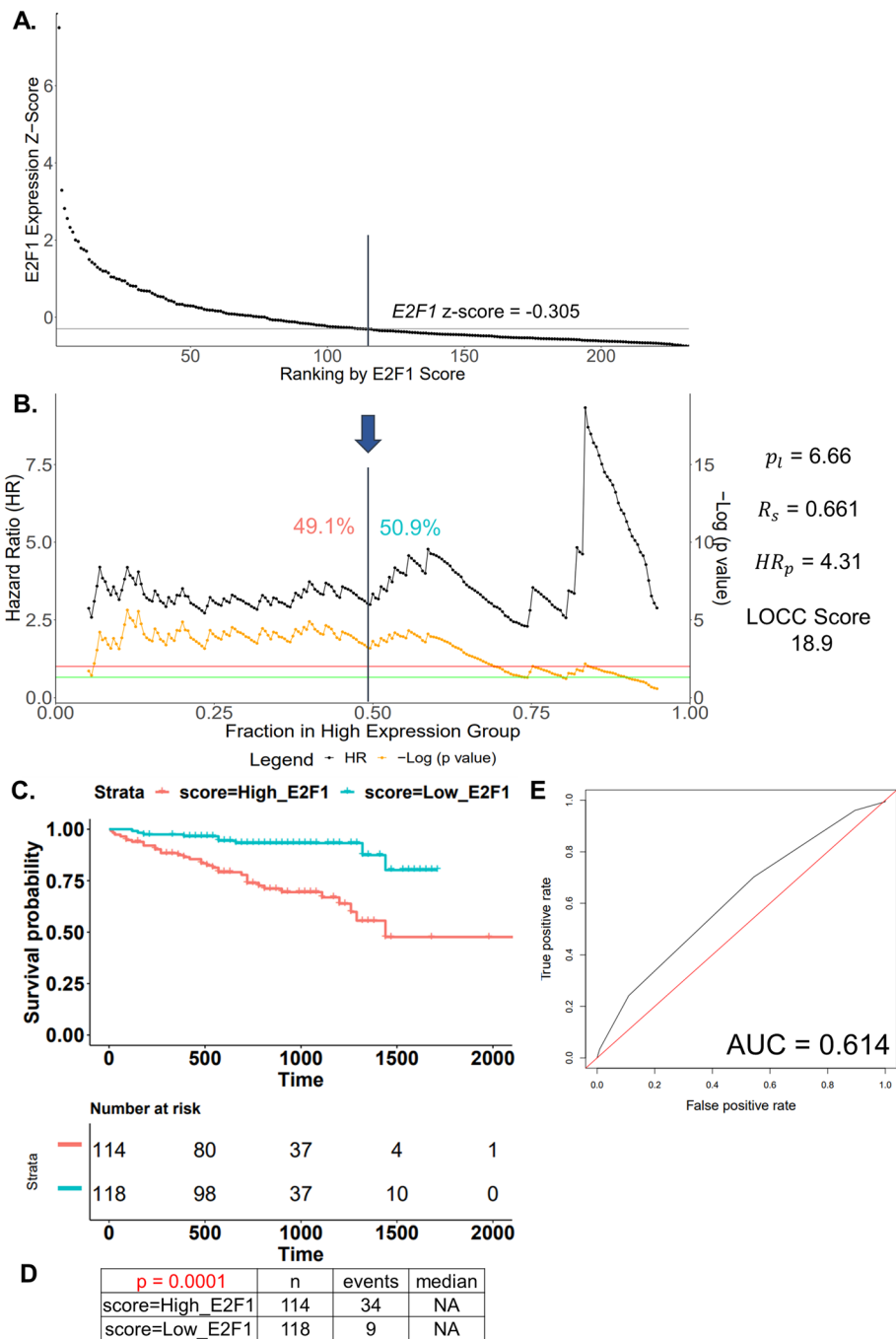


Fig. 2 Validation of *E2F1* as prognostic biomarker in the LIPI-JP hepatocellular dataset **A** The z-score expression of *E2F1* from LIPI-JP hepatocellular carcinoma patient samples was ordered in descending order and plotted against the ranking of the samples. A horizontal line is graphed at -0.305 to separate patients into high and low *E2F1* groups using the TCGA dataset cutoff. **B** The LOCC cutoff selection was graphed for LIPI-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*. **D** 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. HRs are calculated using cox proportional hazard regression. **E** A ROC is plotted for *E2F1* and the AUC was calculated

plotted to identify the optimal cutoff. This process aids not only in the selection of the cutoff but also enables us to assess changes in the hazard ratio across the biomarker's range.

To compare LOCC's and ROC's ability to evaluate prognostic biomarkers in various cancers, we investigate *E2F1* cut-offs within LIHC, Sarcoma (SARC), pancreatic ductal adenocarcinoma (PAAD), kidney renal papillary cancer (KIRP), and lower grade glioma (LGG). Different cancer types should presume different cut-offs since their gene expression and prognostic distributions differ; what is considered "high" or "low" expression for each cancer type is expected to vary based on intrinsic differences between the cancers.

The TCGA SARC dataset [31] presents an ROC curve for *E2F1* expression with an AUC of 0.576 (Fig. 3A). While the AUC for *E2F1* expression in LIHC appears similar, the LOCC graph reveals notable differences, with a LOCC score for sarcoma *E2F1* of only 0.11 (Fig. 3B), much lower than the 2.58 for *E2F1* in LIHC. The main difference lies in the p values and significant ranges, whereas the hazard ratios (HR) were comparable. Once the most significant cutoff was chosen, a Kaplan–Meier curve was plotted (Fig. 3C). Despite the low LOCC score, the prognosis value of *E2F1* in sarcoma is less robust than *E2F1* in hepatocellular carcinoma, in stark contrast to the similar ROC c-statistics.

In the TCGA PAAD dataset [32], we plotted the ROC curve for *E2F1*, yielding an AUC of 0.646 (Fig. 3D). With a comparably high AUC and a LOCC score of 3.00 (Fig. 3E), both indices increased from the SARC data. However, the HR for both cancers at the most significant cutoffs was approximately 2.0, with the difference in p values across the LOCCs leading to discrepancies in scores. Using the most significant cutoff, a Kaplan–Meier plot showed significant stratification using *E2F1* expression (Fig. 3F).

In the TCGA KIRP dataset [33], the ROC curve for *E2F1* revealed an AUC of 0.623 (Fig. 3G). While the KIRP ROC curve appears similar to the PAAD curve at first glance, LOCC provides a contrasting view, identifying a high-risk group among the top 20% of *E2F1* expression, while cutoffs near the median were marginally significant (Fig. 3H). With a LOCC score of 9.04, this indicates a high prognostic potential for *E2F1* in this dataset, if the correct cutoff is employed. The Kaplan–Meier curve confirms that a small proportion of KIRP patients exhibit significantly higher risk with increased *E2F1* expression (Fig. 3I).

Lastly, the TCGA LGG data [34], was evaluated with the ROC curve for *E2F1*, resulting in an AUC of 0.614 (Fig. 3J). Despite a lower AUC compared to PAAD and KIRP, the LGG LOCC score of 32.5 was significantly higher, demonstrating that most cutoffs were highly significant (Fig. 3K). This is evident in the Kaplan–Meier curve, which identifies two distinct patient groups based on *E2F1* expression (Fig. 3L). Consequently, *E2F1* expression is a significant prognostic predictor in LGG, a finding that has prompted further research into targeting this pathway [35].

While cancer types vary, ROC AUC values remain relatively similar, whereas LOCC values vary. These variabilities in LOCC correctly exemplify differences in the gene expression distributions between these cancer types. ROC AUC values seem to fluctuate minimally and somewhat randomly with no clearly defined meaning within its fluctuations, whereas LOCC scores vary based on prognostic potential. Notably, *E2F1* HR

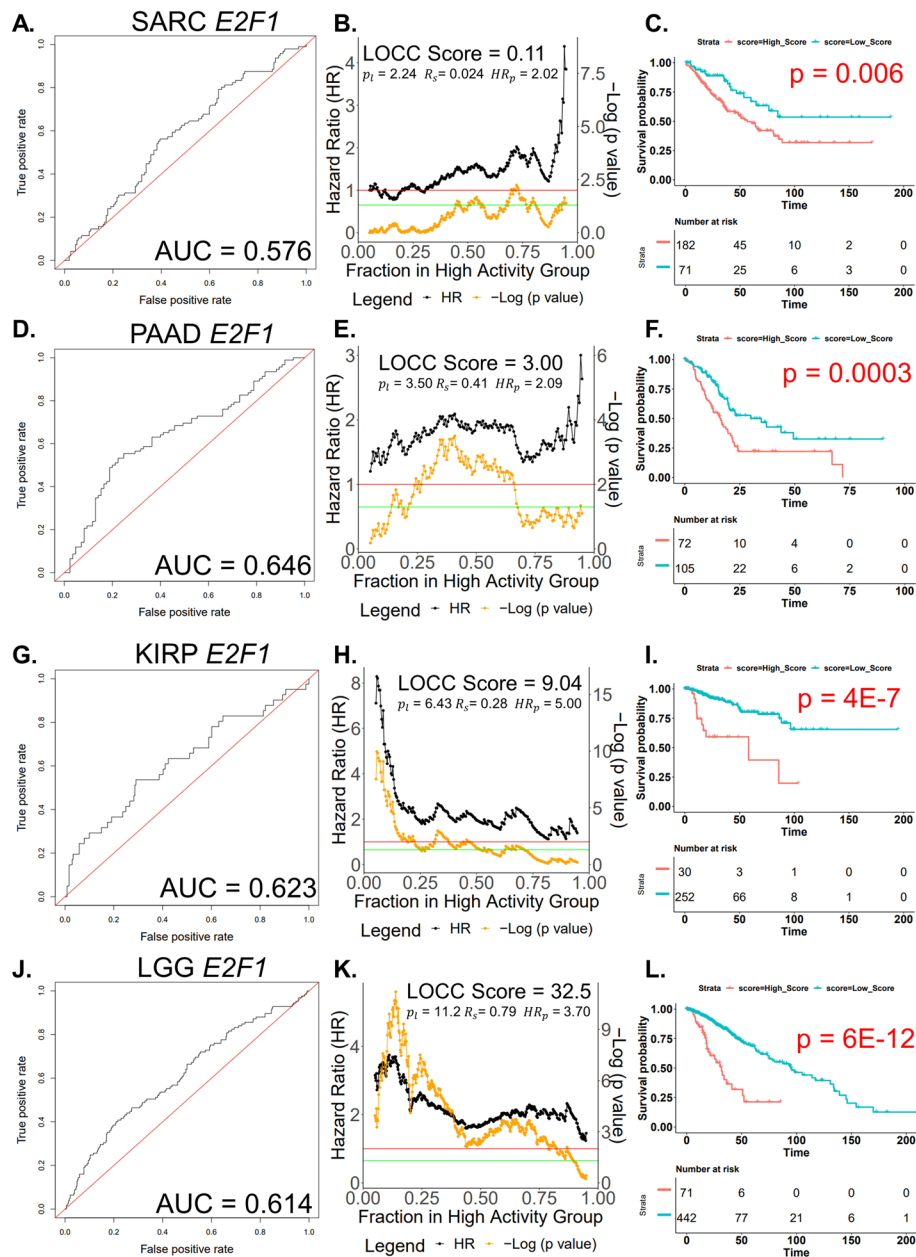


Fig. 3 Comparison of ROC and LOCC in evaluating prognostic biomarkers **A** A ROC curve is plotted and the AUC was calculated for TCGA SARC *E2F1*. **B** LOCC was plotted and scored for SARC *E2F1*. **C** The most significant cutoff was selected and a Kaplan–Meier plot was graphed to illustrate the best stratification according to *E2F1* in SARC. **D** A ROC curve is plotted and the AUC was calculated for TCGA PAAD *E2F1*. **E** LOCC was plotted and scored for PAAD *E2F1*. **F** The most significant cutoff was selected and a Kaplan–Meier plot was graphed to illustrate the best stratification according to *E2F1* in PAAD. **G** A ROC curve is plotted and the AUC was calculated for TCGA KIRP *E2F1*. **H** LOCC was plotted and scored for KIRP *E2F1*. **I** The most significant cutoff was selected and a Kaplan–Meier plot was graphed to illustrate the best stratification according to *E2F1* in KIRP. **J** A ROC curve is plotted and the AUC was calculated for TCGA LGG *E2F1*. **K** LOCC was plotted and scored for LGG *E2F1*. **L** The most significant cutoff was selected and a Kaplan–Meier plot was graphed to illustrate the best stratification according to *E2F1* in LGG. Abbreviations and symbols: TCGA – The Cancer Genome Atlas, SARC – Sarcoma, PAAD – Pancreatic ductal adenocarcinoma, KIRP – Kidney renal papillary cell carcinoma, LGG – low grade glioma, p_i is $-\log(p)$ value at most significant cutoff, R_s is percentage of cutoffs that have a highly significant p value ($p < 0.01$), HR_p is the HR at the most significant cutoff

values were all greater than 1 across the entire distribution of all 5 cancer types meaning that higher *E2F1* is associated with worse prognosis at all points though of different statistical significance and hazard ratios. As a result, LOCC not only correctly identifies that *E2F1* is associated with poor prognosis in all of these cancers but it also provides additional insight into the gene expression distribution and prognostic power through the absolute value of its score.

LOCC score helps rank prognostic importance of predictors

First, we applied LOCC analysis and scoring to all gene expression profiles in TCGA LIHC (Supplemental Table 1). Furthermore, we demonstrate that the top prognostic cut-offs identified by LOCC are statistically significant even after adjusting for multiple testing through Monte Carlo methods of simulation of random permutations of the dataset to estimate empirical p values and q values (Supplemental Table 1 and Supplemental Table 2).

Next, we demonstrate the Cox PH analysis in cancer biomarker analysis has a few critical flaws which limit its appropriateness in ranking predictors. We performed Cox PH analysis for all gene expression of LIHC (Supplemental Table 3). Some of the top ranked genes by Cox PH p value include GAGE family genes which are aberrant expressed genes in cancers [36]. However, only a minority of tumor samples have expression of GAGE genes and many other aberrant expressed genes and this is not clear on the Cox PH analysis summary (Supplemental Fig. S3A, B). LOCC analysis also rank many of these genes near the top of the list but not nearly as high due the limitation that it is only expressed in a minority of samples (Supplemental Table 1). Furthermore, LOCC analysis clearly show that these genes have duplicate gene express levels, a sign of non-expression.

The larger issue with Cox PH analysis is that it has two major assumptions which can lead to skewed results if they are violated [20, 37]. In the situation of cancer biomarkers, this is a significant issue as 52 of the top 100 significant genes by Cox PH p value for LIHC have violation of at least one assumption of proportional hazard or linearity (Supplemental Table 3). To demonstrate how this can affect accurate portrayal of the prognostic value of predictors, we compare two gene expression, *POLR2H* and *TBP*, which have similar Cox PH p values of 0.049 but differing LOCC scores of 0.0097 and 1.1, respectively.

For *POLR2H*, both assumptions of proportional hazard and linearity are not being violated ($p > 0.05$) and thus the Cox PH p value is appropriate (Supplemental Fig. S3C, D). However, using LOCC, we can see that this biomarker has little very significant cutoffs and a small portion in the significant range (Supplemental Fig. S3E). As such, its LOCC score is very low at 0.0097 suggesting this gene has little prognostic potential.

For *TBP*, both assumptions of proportional hazard and linearity are violated ($p < 0.05$) in this dataset which suggest the Cox PH p value may not be accurate (Supplemental Fig. S3F, G). Violation of the proportional hazard means that the relationship between the covariate and the hazard rate is not constant over time, which can lead to incorrect or misleading estimates of hazard ratios in the Cox PH model. Violation of the linearity assumption indicates that the relationship between a continuous covariate and the log hazard is non-linear which can lead to biased estimates of hazard ratios and incorrect conclusions about the effect of the covariate. However, it is not clear if this will lead to

overestimation or underestimation of the p value but simply that Cox PH analysis may be inaccurate. Using LOCC, we can visualize what is happening with TBP that is causing problems with Cox PH analysis. Violation of the proportional hazard leads to varying hazard ratios over time which may be seen in highly fluctuating hazard ratios on the LOCC graph seen in Supplemental Fig. S3H. Similarly, violation of linearity is visualized in LOCC when hazard ratios show significant inflection points. In this scenario, the LOCC is better in assessing the prognostic potential of the predictor since certain portions of the gene expression range are useful for prognostic evaluation while other portions are not. Cox PH uses the entire range which is not appropriate in this scenario when only a certain portion is prognostic significant. Biologically, this can occur for many reasons such as needing a threshold for pathway activation or where overexpression of a gene has diminishing effects. LOCC visualization show *TBP* is significantly associated with poor expression at high expression but that low expression is not associated with good prognosis. Cox PH modeling only show that the assumptions are violated and thus fails to provide accurate commentary regarding the prognostic potential of this biomarker.

LOCC scores also have better prognostic interpretability and understanding compared to ROC score. With the LOCC score, a marker scoring zero signifies a lack of prognostic value; an AUC of 0.5 in the ROC score system also implies no predictive value. Yet, the threshold for the ROC AUC is nebulous as almost all genes register an AUC above or below 0.5.

We utilize TCGA hepatocellular carcinoma data to compare the prognostic value of LOCC scores vs AUC scores and confirm their validity by comparing identified prognostic biomarkers with previous literature. The following thresholds were established for this analysis; LOCC scores >0.1 , >1 , and >8 . An LOCC score above 0.1 means there exists some highly significant prognostic cutoff; an LOCC score greater than 1 means that the corresponding cancer biomarker has a sizable range of highly significant cutoffs ($>10\%$) along with other appropriate significance values and hazard ratios for biomarker consideration; an LOCC score above 8 represents the top 1% of LOCC scores in this TCGA LIHC dataset and these biomarkers have the highest potential prognostic value.

We evaluated 18,789 genes within TCGA hepatocellular carcinoma. 5398 (28.7%) showed LOCC scores above 0.1 or below -0.1 , 2455 (13.1%) had LOCC scores above 1 or below -1 , and 214 (1.1%) had LOCC scores above 8 or below -8 (Fig. 4A, Supplemental Table 1, Table 1). With these percentages, we identified ROC AUC cutoffs that pinpointed a similar proportion of genes. An ROC AUC above 0.54 or below 0.46 encompassed 6301 (33.5%) genes; an ROC AUC above 0.56 or below 0.44 included 2731 (14.5%) genes; an ROC AUC above 0.60 or below 0.40 contained 203 (1.1%) genes (Fig. 4B, Supplemental Table 1, and Table 1). Absolute AUC cutoffs of 0.54, 0.56, and 0.60 corresponded most closely to absolute LOCC scores of 0.1, 1, and 8; these thresholds were utilized to compare the predictive power of LOCC scores vs AUC scores and assess the potential for enrichment.

Notably, LOCC has a clear correlation between absolute value and predictive power and a clear score cut-off for delineating prognostic potential. A LOCC score of zero suggests no cutoffs can achieve highly significant prognostic importance ($p < 0.01$). A LOCC score of 1 may indicate potential prognostic value, as this score corresponds to

Table 1 Analysis of hepatocellular carcinoma genes and prognostic gene signatures by LOCC and ROC conditions

Conditions	All genes	Percentage of all genes (%)	Ouyang et al	Liu et al	Chen et al	Zhou et al	Sum of gene predictors	Different gene predictors	Percentage of predictors (%)	Enrichment (%predictors/%All)
None	18,789	100	12	6	8	10	36	34	100	100
Abs LOCC score > 0.1	5398	28.7	12	6	8	10	34	32	94	328
Abs LOCC score > 1	2455	13.1	12	4	7	7	30	28	82	630
Abs LOCC score > 8	214	1.1	7	1	4	3	15	14	41	3615
Abs ROC AUC > 0.54	6301	34	12	5	8	8	33	31	91	272
Abs ROC AUC > 0.56	2731	14.5	12	2	7	7	30	28	82	567
Abs ROC AUC > 0.6	203	1.1	7	0	1	2	10	9	26	2450

a substantial proportion (~0.1–0.2) of highly significant cutoffs (Supplemental Table 1). In contrast, the literature lacks a clearly defined set AUC for prognosis, which is instinctively challenging to establish as evidenced in the example above.

Four recent publications describe prognostic gene signatures using the TCGA hepatocellular carcinoma dataset [24, 38–40]. These works identify 34 distinct genes potentially useful as hepatocellular carcinoma predictors. However, the predictor overlap between the four publications is minimal, with only two genes (*SPP1* and *LECT2*) found common in two of the datasets.

We believe this discrepancy stems from the use of Least Absolute Shrinkage Selector Operator (LASSO) regression, which selects good predictors without offering a holistic view of the disease landscape. The variability among different publications indicates a need for a better approach to understand the full picture of hepatocellular carcinoma predictors. To help better understand the full picture of hepatocellular carcinoma predictors and their importance, we identified these 34 genes by their ranking by LOCC score and ROC AUC to investigate any trends.

We summarize the categorization of scores of all ranked genes and the 34 different genes in Table 1 and the full data regarding individual predictor genes in Supplemental Table 4. The trend is that 34 predictor genes are clustered at the top of LOCC scores and ROC AUC. However, it appears that the LOCC score better explains the selection of gene predictors. Using a cutoff of absolute LOCC score greater than 1, there were 2455 genes (13.1% of all genes) fitting that condition and contain 28 of the 34 predictor genes (Table 1). Meanwhile, using a cutoff of absolute ROC AUC greater than 0.56, there were 2731 genes (14.5% of all genes) fitting that condition and also contain 28 of the 34 predictor genes (Table 1). Using a cutoff of absolute LOCC score greater than 8 or absolute ROC AUC over 0.60 yields 214 genes (1.1% of total). Of the top 214 LOCC score genes, 14 of them were from the 34 gene predictors (Table 1). On the other hand, of the top 203 genes by ROC AUC ($AUC > 0.6$) there were only 9 out of the 34 predictors (Table 1). Finally, the average percentile of gene predictors in each gene signature was lower using LOCC score ranking compared to ROC AUC ranking (Supplemental Table 4).

Thus, while it appears that the gene predictors were clustered near the top of both scores, the LOCC score consistently better explained the selection of the gene predictors. In particular, we compare the ROC AUC and LOCC visualization and score for the *KIF20A*, a predictor in one of the published gene signatures [24] (Fig. 4C–E). The ROC AUC of *KIF20A* is 0.60, which is in the top 200 but its LOCC score is 46.8, the highest of all ranked genes. From the LOCC visualization, there is one very distinct peak and there is a very significant survival difference between the groups using that cutoff (Fig. 4D, E).

Next, we demonstrate even among the predictors in the gene signatures, LOCC score can help select which genes are contributing significantly or not as significantly. We used LOCC scoring to rank the 12 individual genes from a previously published prognostic gene signature (RISK) in hepatocellular carcinoma [24]. We found the top eight gene expression by LOCC score (*KIF20A*, *TTK*, *TPX2*, *LCAT*, *SPP1*, *HMMR*, *CYP2C9*, *ANXA10*) had a significantly higher LOCC score than the other 4 genes (LOCC Score < 0.05 and \log_{10} LOCC p value < 0.05 , Table 2). ROC AUC of the 12 genes were all between 0.57 and 0.62 with 10 genes between 0.59 and 0.62 though the top 8 were significantly higher in ROC AUC than the bottom 4 genes ($p < 0.05$, Table 2)[24]. Meanwhile,

Table 2 LOCC score of RISK gene signature

Gene	-log (lowest p value)	Highly significantly range (p < 0.01)	Highest HR	Lowest HR	HR at significant cutoff	LOCC score	Log ₂ LOCC score	ROC AUC	Good or poor prognosis	Cox regression exp(β)	Absolute value (β)	Cox regression p value	-log(cox p value)
KIF20A	16.097	0.709	4.094	1.470	4.094	46.8	5.548	0.604	Poor	1.320	0.278	0.000006	5.22
TTK	9.874	0.694	3.291	1.232	3.291	22.2	4.472	0.593	Poor	1.260	0.231	0.00002	4.70
TPX2	5.650	0.818	2.761	1.661	2.761	12.8	3.678	0.617	Poor	1.375	0.318	0.000004	5.40
LCA1	8.077	0.598	0.714	0.381	0.381	12.7	3.667	0.614	Good	0.795	0.229	0.000001	6.00
SPP1	5.692	0.799	2.415	1.740	2.312	10.5	3.392	0.617	Poor	1.125	0.118	0.000004	5.40
HMMR	5.363	0.780	2.480	1.417	2.199	9.2	3.202	0.607	Poor	1.299	0.262	0.00002	4.70
CYP2C9	5.599	0.658	0.677	0.410	0.449	8.2	3.036	0.611	Good	0.900	0.105	0.0001	4.00
ANXA10	4.993	0.738	0.781	0.420	0.469	7.9	2.982	0.607	Good	0.887	0.120	0.0001	4.00
RDH16	5.332	0.562	0.689	0.452	0.452	6.6	2.722	0.593	Good	0.897	0.109	0.0002	3.70
LECT2	5.041	0.501	0.847	0.430	0.430	5.9	2.555	0.596	Good	0.920	0.083	0.0002	3.70
LAPTM4B	4.836	0.543	2.532	1.444	2.094	5.5	2.459	0.583	Poor	1.270	0.239	0.0003	3.52
MAGEA6	3.540	0.256	2.030	1.000	2.335	1.8	0.880	0.570	Poor	1.060	0.058	0.001	3.00
Top 8 versus Bot 4 (t test p value)	0.065	0.028	0.356	0.442	0.380	0.046	0.022	0.019		0.499	0.135	0.134	0.00070

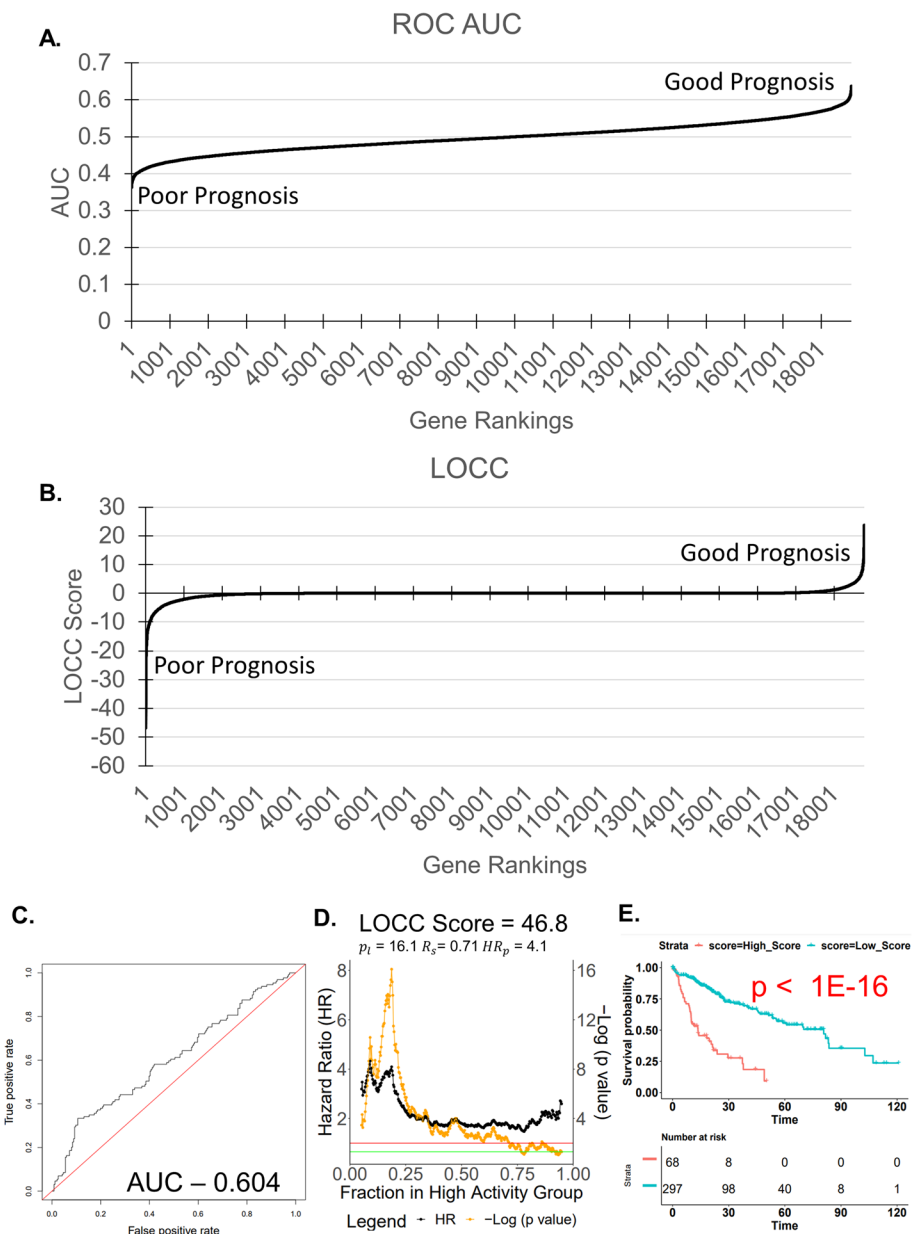


Fig. 4 LOCC score helps rank prognostic importance of predictors **A** Genes were evaluated and ranked by ROC AUC. **B** Genes were evaluated and ranked by LOCC score. **C** The ROC curve is shown for TCGA hepatocellular carcinoma *KIF20A* expression and the AUC is calculated. **D** The LOCC cutoff selection is graphed for TCGA hepatocellular carcinoma *KIF20A* expression. **E** A Kaplan–Meier curve is plotted for *KIF20A* at the most significant cutoff. Patients’ survival times are expressed in days. *P* values are calculated using log-rank test. HR are calculated using Cox proportional hazard regression

cox regression analysis also found a significant difference in the $-\log(p \text{ value})$ between the 8 genes from LOCC analysis and the other 4 genes ($p = 0.0007$, Table 2).

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, ROC, and cox regression. To do this, we used the same weight coefficient as the original article and used it to calculate the 8-gene RISK score [24]. 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 (Supplemental Fig. S4A–E). 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. (Supplemental Fig. S4D–G). Finally, the Akaike information criterion (AIC), a number to measure relative model quality, shows that the 8-gene RISK score is a similar if not better predictor model (Supplemental Fig. S4E). 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 signature demonstrating LOCC score ability to decipher the key predictors.

While the LOCC score is correlated with the cox regression p values and ROC AUC (Supplemental Fig. S4A, B), 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*. Although the three gene expression profiles are strongly correlated (Supplemental Fig. S4D, E), LOCC scoring values very significant cutoffs while cox regression and ROC AUC favor having a more consistent HR (Table 2). 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 [11, 13]. 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 more gene expression prognosis analysis. However, we argue that LOCC score better represents the overall gene signature. This can be seen as *KIF20A* and *TTK* are ranked higher by LOCC score than *TPX2* while *TPX2* has higher ROC AUC and cox regression p value than both *KIF20A* and *TTK* (Table 2). When we compare the LOCC visualization of the original 12-gene or 8-gene RISK signature (Supplemental Fig. S3A, B), both the HR and significance lines more closely resemble the LOCC visualization of *KIF20A* and *TTK* compared to *TPX2* (Fig. 4D, Supplemental Fig. S4F, G). This suggests that the contribution of *KIF20A* and *TTK* are likely more significant than *TPX2* in shaping the overall prognostic gene signature.

Variability and reproducibility of LOCC scores and ROC AUC

To scrutinize the consistency and reproducibility of LOCC scores and ROC AUC, we explored the gene expression of three distinct prognostic markers within a diverse sampling of the TCGA dataset. The markers selected include *KIF20A*, which has the highest LOCC score and is a critical predictor in the RISK score [24] (Fig. 4C–E), *E2F1*, a verified prognostic indicator presented in Fig. 1 and 2; and *MTHFR*, a marker previously identified in a hepatocellular carcinoma prognostic gene set [38] (Supplemental Fig. S6A, B). The LOCC scores and ROC AUC of *KIF20A*, *E2F1*, and *MTHFR* respectively are 46.7, 2.58, 0.01 and 0.604, 0.578, 0.528 (Supplemental Table 5). The trends revealed by these scores suggest a coherent pattern among the three genes, a pattern that we aim to sustain and clarify in random samples from the dataset.

twofold cross-validations were utilized to examine the generalizability and variability of the data set. Upon conducting a 100 twofold cross-validation, we discerned that the general trend was maintained across all features investigated, yet the degree of variability differed by feature (Fig. 5A–F). Significant HR, highest $-\log(p)$ value, and percent highly significant each demonstrated a preserved trend with non-overlapping

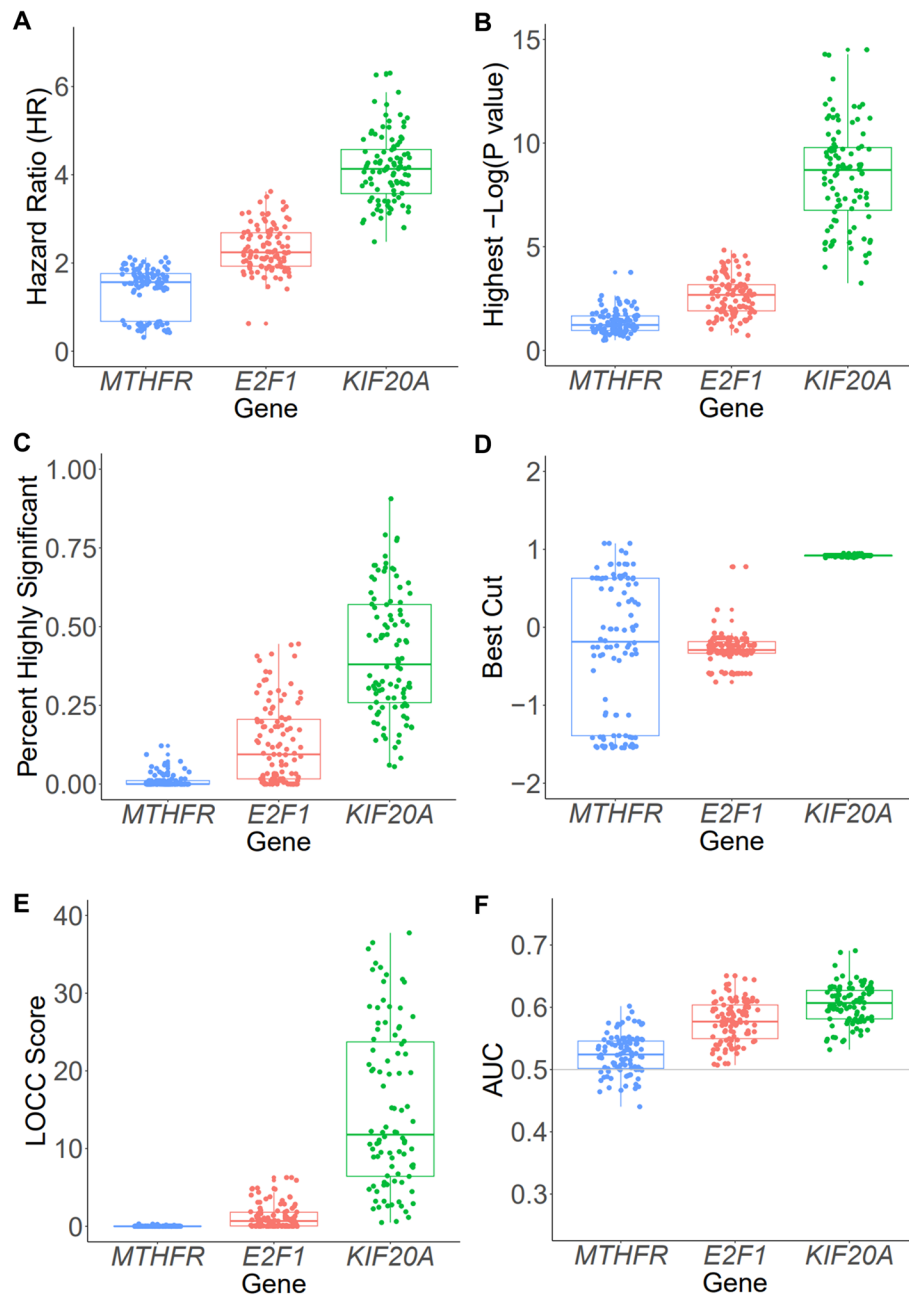


Fig. 5 Variability and Reproducibility of LOCC Scores and ROC AUC **A** The 8-gene modified RISK score was ordered in descending order and plotted against the ranking of the samples. A horizontal line is graphed at 2.00 to separate patients into high and low risk groups using ideal LOCC cutoff. **B** 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

interquartile ranges (Fig. 5A–C and Table 2). The best cut, determined by the cutoff with the most significant p value, exhibited considerable variability for *MTHFR*, minimal for *E2F1*, and virtually none for *KIF20A* (Fig. 5D and Supplemental Table 5). The LOCC scores effectively differentiated between the three genes though with significant variability for *KIF20A* (Fig. 5E and Table 2). Finally, ROC AUC exhibited considerable variability but also had considerable overlap between *E2F1* and *KIF20A* AUC (Fig. 5F and Table 2).

Comparison of features between the full dataset and individual samples reveals several patterns (Table 2). The significant HR is relatively consistent between the full dataset and the samples. However, the highest $-\log(P)$ value and percent highly significant are considerably lower in the samples, a foreseeable outcome given the dependency of p values on the number of samples. Consequently, the LOCC scores are on average lower in the samples as compared to the full dataset. The ROC AUC was similar on average between the samples and the full dataset but demonstrated substantial variability among individual samples. The optimal cutoffs were similar for the full dataset and individual samples, with the exception of *MTHFR*, which exhibited three distinct clusters of points, leading to the discrepancy (Fig. 5D). Further, leveraging the twofold cross validation, we were able to investigate the “validation success rate” of our data set. The validation success rate represented the percentage of the twofold cross-validation that exhibited $p < 0.05$ when using the full data best cutoff. This percentage represents the probability that the calculated LOCC cutoff can be replicated or validated in miniature cohort; if the twofold cross-validation success rate is high, then the likelihood of validation using similar methods is high as the variance within the data set is low, and if the twofold cross-validation success rate is low, then the likelihood of validation using methods is low due to high variance within the data set. With this approach, we examined the validation success rate of *KIF20A*, *E2F1*, and *MTHFR*. Their validation success rates were 100%, 62%, and 1%, respectively. Notably, *KIF20A* successfully passed every twofold cross-validation while *E2F1* passed the majority, but *MTHFR* only managed 1% of the twofold cross-validations, suggesting its limited utility as a prognostic predictor (Supplemental Table 5).

To better illustrate certain samples, we chose three from *MTHFR* and *KIF20A* for further investigation using LOCC (Supplemental Fig. S6C, D). By comparing the visualization of the full dataset *MTHFR* and the three samples of *MTHFR*, we observed both similarities and distinctions (Supplemental Fig. S6A, C). The most dramatic changes occur at the edges, yet each sample bears resemblance to the full dataset. *KIF20A*, on the other hand, showcases a distinct peak at the high score group (Fig. 5D), a feature retained in all its samples (Supplemental Fig. S6D). High peaks, marked by the yellow significant line, imply that the cutoff is more resilient to sampling bias, offering another interpretation of p values. However, we did notice edge noise, particularly in sample 3 of *KIF20A*. Despite the high HR at the right edge, the yellow line does not surpass the red line. As such, the optimal cutoff remains the same as the original where the yellow line peaks around 0.20. In conclusion, we believe LOCC visualization and scoring effectively upheld their integrity during the variability and reproducibility sampling, underscoring their utility as computational tools for the analysis and comprehension of cutoffs and prognostic predictors.

Individual gene expression and gene signatures associated with prognosis in hepatocellular carcinoma

While individual gene expression profiles are interesting biomarkers for diseases, it is of greater biological relevance to understand which pathways are of prognostic significance, which can be achieved through LOCC analysis. Leveraging LOCC scores, we can rank each gene based on its prognostic potential. Then, by utilizing GSEA, we can determine the key physiological pathways and gene sets associated with the individual genes identified by LOCC. While LOCC classifies the prognostic value of individual genes, further examining LOCC scores with GSEA allows us to discern which gene sets and pathways are crucial for prognosis. We applied this approach to hepatocellular carcinoma, scoring 18,789 genes with LOCC, out of which 10,145 genes exhibited non-zero LOCC scores. Employing these LOCC scores (Supplemental Table 1), we utilized GSEA pre-rank enrichment to identify their associated prognostic pathways and gene clusters. A handful of hallmark pathways emerged as significantly relevant (Fig. 6A, B, Supplemental Table 6). In particular, G2M checkpoint and E2F targets correlated with poor prognosis, while bile acid metabolism and oxidative phosphorylation were linked to good prognosis (Fig. 6C–F). These findings align with previous research [39], further confirming the validity of the LOCC scores. However, our method also flagged oxidative phosphorylation, which was likely overlooked in the prior study as they focused on contrasting gene expression patterns between two groups, rather than examining all genes in a continuous manner.

To ensure the robustness of our GSEA model, we also examined possible redundancy and overlap between gene signatures. Between bile acid metabolism and oxidative phosphorylation, we identified only 4 overlapping genes and over 100 distinct genes within each gene set. As such, its redundancy is negligible and will have minimal to no impact on GSEA. On the contrary, hallmark E2F targets and G2M checkpoints have significant overlap with 73 overlapping genes and 127 distinct genes in each gene set (Supplemental Fig. S7A). We conducted a secondary GSEA to determine whether overlapping gene signatures would meaningfully impact our GSEA results. When overlapping genes are removed, we observe very similar NES, p values, and FDR values for each gene set (Supplemental Fig. S7B, C), illustrating that the overlapping portion of genes does not significantly alter the overall analysis and confirming that G2M checkpoint and E2F targets are associated with poor prognosis.

We also performed validation that the p values from preranked GSEA of randomly created gene sets followed an expected uniform distribution of p values. We generated 100 gene sets of up to 200 genes per set and performed preranked GSEA using LOCC scores which did demonstrate an uniform distribution of p values (Supplemental Fig. S7D, Supplemental Table 6). As such, the very low p values for the top Hallmark pathways associated with prognosis are unlikely to be due to pure chance.

Discussion

In this study, we introduce LOCC, an innovative visualization tool that depicts the distribution, significance, and hazard ratios associated with a continuous variable. This enables the selection of an optimized cutoff or multiple cutoffs while surveying the entire

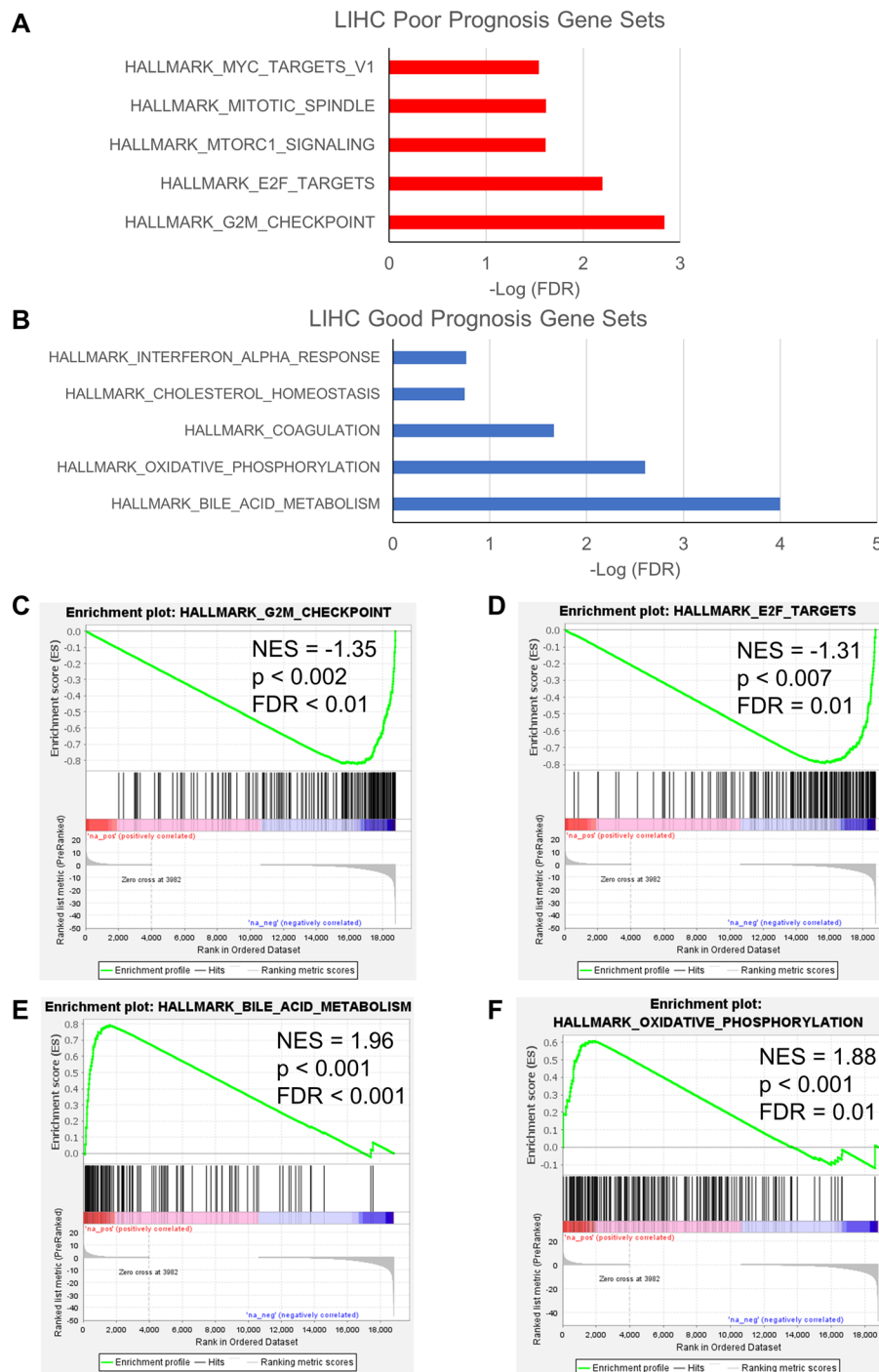


Fig. 6 Individual gene expression and gene signatures associated with prognosis in hepatocellular carcinoma **A** The most significant poor prognosis gene sets from GSEA are shown. Gene sets are sorted by FDR (False Discovery Rate). **B** The most significant good prognosis gene sets from GSEA are shown. Gene sets are sorted by FDR (False Discovery Rate). **C** The enrichment plot of the hallmark G2M checkpoint is shown. **D** The enrichment plot of the hallmark E2F targets is shown. **E** The enrichment plot of the hallmark bile acid metabolism is shown. **F** The enrichment plot of the hallmark oxidative phosphorylation is shown

range of potential cutoffs. While this study investigates hepatocellular carcinoma, LOCC is not limited to analysis within cancer. LOCC can analyze any data set containing a continuous variable and its corresponding prognostic outcomes. In this study, we've demonstrated the applicability of LOCC towards elucidating the relationship between gene expression and hepatocellular carcinoma prognosis. However, we anticipate that LOCC may be applied towards a broad spectrum of carcinomas and even other diseases as a tool to examine the relationship of a continuous variable and its prognostic outcomes.

LOCC offers several advantages over prior methods, especially ROC curve and Cox PH model analysis for prognosis. By fusing the hazard ratio and p value into a single graph, the significance and impact of a variable can be comprehended simultaneously. We established that LOCC holds more informative content compared to ROC curves through direct comparison. In the context of prognosis, sensitivity, specificity, and the c -statistic are challenging to interpret, while HR and p values are straightforward. LOCC analysis is not also limited by the assumptions that are required by Cox PH to have accurate results. Thus, we propose the LOCC visualization and scoring tool as a valuable resource for the analysis of many continuous variables to outcome analyses, potentially outperforming ROC curves and Cox PH analysis [17].

We suggested the LOCC score as a representative metric for ranking variables by significance, range, and impact. Using a combination of three variables, we can generate a score to help pinpoint the most significant variables related to the outcome. This score represents our attempt to numerically encapsulate the comprehensive LOCC visualization for ease of understanding and sorting. The LOCC score outperforms ROC for prognostic studies, as the AUC has a relatively small range and large variability in sampling tests. Even though both LOCC score and ROC AUC assist in enriching prognostic gene predictors, the LOCC score demonstrated superior enrichment performance for gene predictors. A notable difference is LOCC's incorporation of significance into its scoring, which aids in determining valuable predictors of outcome, while ROC AUC cutoffs appear arbitrary. The inclusion of significance, range, and impact into the LOCC score allows LOCC to more effectively portray the prognostic value of a biomarker; a biomarker with no prognostic value will always score a low LOCC score and vice versa. Additionally, biomarkers with an LOCC score of zero can easily be categorized as having little to no prognostic value; the ROC AUC approach does not clearly delineate what is considered a significant vs insignificant prognostic number. As noted previously ROC AUC cutoffs are arbitrary at best, and ROC AUC cutoffs are not always correlated with prognostic potential. As a result, the LOCC score better characterizes the relationship between a gene set and prognostic outcomes. All in all, LOCC is a more practical, robust approach for prognostic evaluation.

As discussed previously, the Cox PH model has also been utilized to characterize prognostic studies, but its reliance on the assumption of proportionality and linearity across an entire distribution within a data set limits its practicality and feasibility. The Cox PH model often overlooks cutoffs identified by LOCC as the Cox PH model is unable to account for non-proportional associations within datasets [41]. In many cases, particularly in complex biological prognostic studies, continuous variables hold predictive value at certain ranges but do not hold linearity and proportionality across an entire data distribution, which can cause biased estimates and misleading information when relying on

the Cox PH model [19, 37]. This can be seen in our specific examples of when assumptions of Cox PH model was violated leading to misleading p values. The LOCC model resolves this limitation by employing an alternative approach; by identifying specific cut-offs within the data set that hold significant prognostic value, we can measure the predictive power of a data set with or without an assumption of proportionality. Further, the LOCC is able to pinpoint specific cut-offs with the greatest prognostic relevance. As a result, LOCC provides a more practical realistic methodology to accurately examine the prognostic power of continuous variables and their associated outcomes.

There are still challenges in the bioinformatics field regarding the categorization of continuous variables. Within the context of cancer and gene expression, LOCC can assist in ranking variables and visualizing the relevance of the variable and outcome, but its multi-variable integration needs refinement. In our example, we utilized weighted coefficients from the original study that applied LASSO (Least Absolute Shrinkage Selector Operator) regression. Given that many models like LASSO employ single weight approaches, novel methods might be necessary for selecting and weight variables for optimal LOCC score and prognosis prediction [42]. This notion stems from the fact that LASSO regression may omit potentially active predictors, especially if predictors are correlated or if the number of predictors isn't significantly larger than the number of samples [42, 43]. We saw this kind of issue where several studies used LASSO regression to select prognostic gene signatures for hepatocellular carcinoma with minimal overlapping genes. Furthermore, these prognostic gene signatures often lack biological significance or understanding, leading others to ignore them. Ideally, a gene signature should be prognostic and contain elements that can be interpreted individually or as a group. In addition, we plan to integrate other aspects of cancer data such as cancer stage and age of the patient to better characterize prognostic markers. Therefore, the challenge of finding a superior method for integrating these variables remains [44].

A notable advantage of LOCC is its ability to visualize the selection process and hazard risks, comparable to showing one's work in mathematics. This allows the audience to understand which cutoffs are selected, how they were chosen, and how the validation process succeeded or faltered. Despite many calculations being computer-based, it becomes easier to validate others' work by checking if LOCC graphs align. Moreover, when validation doesn't proceed as expected, it is possible to visualize what went wrong, such as the absence of a peak or a weaker than expected signal. Consequently, LOCC can supplement existing and future prognosis studies by presenting a complete overview of the predictor or biomarker, increasing the audience's confidence.

Limitations

One main limitation of LOCC is that there is while cut-offs are ranked in order, this could be of very minor numerical difference in the continuous variable. A small change in the continuous variable could lead to large relative ranking changes and some effects on the overall LOCC graph and score. However, with large samples and internal sampling, this limitation will likely be minor. Additionally, future methods could include a noise validation test to introduce a noise multiplier to simulate noisy measurements in the real-world though additional experiments should be performed to determine how much noise to signal is present in the model.

LOCC does heavily rely on the best specific cutoff for its rankings which can be problematic. However, the LOCC score includes a range component, helping to counteract overfitting at a specific cutoff. One aspect of LOCC that is both a limitation and benefit is the dependency on the p value, meaning that large datasets can perform better with significance. However, it is also an advantage to incorporate p values to assist in our understanding and selection of prognostic predictors. Ultimately, the best way to verify a predictor is through external validation with another separate dataset.

Finally, this work is primarily focused on the univariate analysis of LOCC. We know that univariate analysis using gene expression has limitations in understanding and predicting prognosis in the big picture of cancer; however, we do demonstrate that LOCC has advantages over ROC and Cox PH in our univariate analysis, a good sign for future development of LOCC. Multivariable visualization and integration will be helpful in better understanding and prognostic modeling of cancer biomarkers but extensive work and comparisons will be required to incorporate these aspects.

Conclusion

LOCC serves as a powerful visualization and prognostic tool, adept at representing the extensive data of continuous variables in relation to survival or prognosis. It facilitates the selection of optimal cutoffs and provides a comprehensive analysis of the relationship of the variable with a particular outcome. The LOCC score excels in pinpointing variables that correlate with survival, and its capacity to rank predictors has proven instrumental in elucidating prognostic genes and pathways. Further, when paired with GSEA, it can be used to identify key physiological pathways which may substantiate our understanding of the biological underpinnings of prognosis and survival. Serving as a viable alternative to ROC curves and Cox PH for prognosis, LOCC brings forth superior visualization and scoring capabilities, consequently providing a deeper insight into predictors and outcomes with better reliability and clarity.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-024-05932-1>.

Supplementary material 1.
Supplementary material 2.
Supplementary material 3.
Supplementary material 4.
Supplementary material 5.
Supplementary material 6.
Supplementary material 7.

Acknowledgements

The graphical abstract was created with Biorender.

Author contributions

George Luo: Conceptualization, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Toby Chen: Conceptualization, Validation, Writing—review & editing. John J Letterio: Supervision, Validation, Writing—review & editing.

Funding

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.

Availability of data and materials

Gene expression and mutation data was accessed from cBioportal.com for TCGA LIHC and <https://dcc.icgc.org/projects/LIRI-JP> for LIRI-JP. Code and instructions for LOCC on R is available at <https://github.com/gluo30/locc>.

Declarations**Ethics approval and consent to participate**

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 22 July 2023 Accepted: 16 September 2024

Published online: 27 September 2024

References

- Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Med Res Methodol*. 2012;12:21.
- 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.
- Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, Geyer CE Jr, Dees EC, Goetz MP, Olson JA Jr, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N Engl J Med*. 2018;379(2):111–21.
- Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, Pierga JY, Brain E, Causeret S, DeLorenzi M, et al. 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. *N Engl J Med*. 2016;375(8):717–29.
- Iizuka N, Oka M, Yamada-Okabe H, Nishida M, Maeda Y, Mori N, Takao T, Tamesa T, Tangoku A, Tabuchi H, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet*. 2003;361(9361):923–9.
- Marquardt JU, Galle PR, Teufel A. Molecular diagnosis and therapy of hepatocellular carcinoma (HCC): an emerging field for advanced technologies. *J Hepatol*. 2012;56(1):267–75.
- Lee JS, Chu IS, Heo J, Calvisi DF, Sun Z, Roskams T, Durnez A, Demetris AJ, Thorgeirsson SS. Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology*. 2004;40(3):667–76.
- Kim K, Zakharkin SO, Allison DB. Expectations, validity, and reality in gene expression profiling. *J Clin Epidemiol*. 2010;63(9):950–9.
- Pinyol R, Montal R, Bassaganyas L, Sia D, Takayama T, Chau GY, Mazzaferro V, Roayaie S, Lee HC, Kokudo N, et al. Molecular predictors of prevention of recurrence in HCC with sorafenib as adjuvant treatment and prognostic factors in the phase 3 STORM trial. *Gut*. 2019;68(6):1065–75.
- Qian Y, Daza J, Itzel T, Betge J, Zhan T, Marmé F, Teufel A. Prognostic cancer gene expression signatures: current status and challenges. *Cells*. 2021;10(3):648.
- Budczies J, Klauschen F, Sinn BV, Györfy B, Schmitt WD, Darb-Esfahani S, Denkert C. Cutoff finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. *PLoS ONE*. 2012;7(12): e51862.
- Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006;332(7549):1080.
- Nagy Á, Munkácsy G, Györfy B. Pancancer survival analysis of cancer hallmark genes. *Sci Rep*. 2021;11(1):6047.
- Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem*. 2008;54(1):17–23.
- Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007;115(5):654–7.
- 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.
- 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.
- Abd ElHafeez S, D'Arrigo G, Leonardis D, Fusaro M, Tripepi G, Roumeliotis S. Methods to analyze time-to-event data: the cox regression analysis. *Oxid Med Cell Longev*. 2021;2021:1302811.
- Stensrud MJ, Hernán MA. Why test for proportional hazards? *JAMA*. 2020;323(14):1401–2.
- Rulli E, Ghilotti F, Biagioli E, Porcu L, Marabese M, D'Incalci M, Bellocco R, Torri V. Assessment of proportional hazard assumption in aggregate data: a systematic review on statistical methodology in clinical trials using time-to-event endpoint. *Br J Cancer*. 2018;119(12):1456–63.
- TCGA. GA Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*. 2017;169(7):1327–41.
- Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2(5):401–4.
- North BV, Curtis D, Sham PC. A note on the calculation of empirical P values from Monte Carlo procedures. *Am J Hum Genet*. 2002;71(2):439–41.
- 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.

25. Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J Clin Epidemiol*. 2014;67(8):850–7.
26. Kent LN, Leone G. The broken cycle: E2F dysfunction in cancer. *Nat Rev Cancer*. 2019;19(6):326–38.
27. 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.
28. Huang YL, Ning G, Chen LB, Lian YF, Gu YR, Wang JL, Chen DM, Wei H, Huang YH. Promising diagnostic and prognostic value of E2Fs in human hepatocellular carcinoma. *Cancer Manag Res*. 2019;11:1725–40.
29. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol*. 2012;12:82.
30. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med Res Methodol*. 2017;17(1):53.
31. TCGA. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell*. 2017;171(4):950–65.
32. TCGA. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell*. 2017;32(2):185–203.
33. Linehan WM, Spellman PT, Ricketts CJ, Creighton CJ, Fei SS, Davis C, Wheeler DA, Murray BA, Schmidt L, Vocke CD, et al. Comprehensive molecular characterization of papillary renal-cell carcinoma. *N Engl J Med*. 2016;374(2):135–45.
34. Brat DJ, Verhaak RG, Aldape KD, Yung WK, Salama SR, Cooper LA, Rheinbay E, Miller CR, Vitucci M, Morozova O, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med*. 2015;372(26):2481–98.
35. Wang X, Wang H, Xu J, Hou X, Zhan H, Zhen Y. Double-targeting CDCA8 and E2F1 inhibits the growth and migration of malignant glioma. *Cell Death Dis*. 2021;12(2):146.
36. Gjerstorff MF, Ditzel HJ. An overview of the GAGE cancer/testis antigen family with the inclusion of newly identified members. *Tissue Antigens*. 2008;71(3):187–92.
37. Matsuo K, Purushotham S, Jiang B, Mandelbaum RS, Takiuchi T, Liu Y, Roman LD. Survival outcome prediction in cervical cancer: cox models vs deep-learning model. *Am J Obstet Gynecol*. 2019;220(4):381.e381–381.e314.
38. Liu GM, Zeng HD, Zhang CY, Xu JW. Identification of a six-gene signature predicting overall survival for hepatocellular carcinoma. *Cancer Cell Int*. 2019;19:138.
39. Zhou T, Cai Z, Ma N, Xie W, Gao C, Huang M, Bai Y, Ni Y, Tang Y. A novel ten-gene signature predicting prognosis in hepatocellular carcinoma. *Front Cell Dev Biol*. 2020;8:629.
40. Chen R, Zhao M, An Y, Liu D, Tang Q, Teng G. A prognostic gene signature for hepatocellular carcinoma. *Front Oncol*. 2022;12: 841530.
41. Bewick V, Cheek L, Ball J. Statistics review 12: survival analysis. *Crit Care*. 2004;8(5):389–94.
42. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.
43. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, Tibshirani RJ. Strong rules for discarding predictors in lasso-type problems. *J R Stat Soc Ser B Stat Methodol*. 2012;74(2):245–66.
44. McDermott JE, Wang J, Mitchell H, Webb-Robertson BJ, Hafen R, Ramey J, Rodland KD. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. *Expert Opin Med Diagn*. 2013;7(1):37–51.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.