

Reboot: a straightforward approach to identify genes and splicing isoforms associated with cancer patient prognosis

Felipe R. C. dos Santos^{1,2,*}, Gabriela D. A. Guardia^{1,*}, Filipe F. dos Santos^{1,3,*} and Pedro A. F. Galante^{1,#}

1 - Centro de Oncologia Molecular, Hospital Sirio-Libanes, Sao Paulo, SP.

2 - Programa Interunidades em Bioinformatica, Universidade de São Paulo, Sao Paulo, SP.

3 - Departamento de Bioquimica, Universidade de Sao Paulo, SP, Brazil.

* These authors contributed equally to this work

Corresponding author: pgalante@mochsl.org.br

Abstract

Nowadays, the massive amount of data generated by modern sequencing technologies provides an unprecedented opportunity to find genes associated with cancer patient prognosis, connecting basic and translational research. However, treating high dimensionality of gene expression data and integrating it with clinical variables are major challenges to carry out these analyses. Here, we present Reboot, an original and efficient algorithm to find genes and splicing isoforms associated with cancer patient survival, disease progression, or other clinical endpoints. Reboot innovates by using a multivariate strategy with penalized Cox regression (LASSO method) combined with a bootstrap approach, in addition to statistical tests for supporting the findings, which are automatically plotted. Applying Reboot on data from 154 glioblastoma patients, we identified a three-gene signature (IKBIP, OSMR, PODNL1) whose increased derived risk score was significantly associated with worse patients' prognosis, even in conjunction with other well-established clinical parameters. Similarly, Reboot was able to find a seven-splicing isoforms signature (CENPF-201; MLKL-202; NUP54-201; MCF2L-201; TFDP1-207; BBS1-206; HTT-202) related to worse overall survival in 177 pancreatic adenocarcinoma patients with elevated risk scores after uni- and multivariate analyses. In summary, Reboot is an efficient, intuitive, and straightforward way for finding genes or splicing isoforms (transcripts) signatures relevant to patient prognosis, which can democratize this kind of analysis and shed light on still under-investigated sets of cancer-related genes. Reboot effectively runs on either servers or personal computers and it is freely available at github.com/galantelab/reboot.

Keywords: prognosis, cancer genomics; transcriptomics; alternative splicing; polygenic score.

INTRODUCTION

The improvement of prognostic prediction and the identification of potential therapeutic targets are major interests in oncology (Hussein et al. 2018; Schirmacher 2019). To achieve these goals, large consortiums (International Cancer Genome Consortium et al. 2010; Grossman et al. 2016) have been created, generated and made available an unprecedented amount of data, which includes clinical (*e.g.*, survival time, tumor recurrence, treatment) and molecular information (*e.g.*, mutation and gene expression profiles) from cancer patients and their samples. In particular, a number of studies have shown that alterations in gene expression (Huang et al. 2013; Sturtz et al. 2014; Gutierrez et al. 2019) and in splicing profiles (Yu et al. 2020; Meng et al. 2020; Liu et al. 2020) are pivotal to tumorigenesis. Once these alterations are established, researchers are often interested in pinpointing genes or splicing isoforms impacting in the prognosis of patients, which are naturally suitable therapeutic targets.

In this scenario, Cox regression models are the gold standard methodology to find genes or splicing isoforms associated with cancer patient survival. Most commonly, analyses performed on datasets with a large number of covariates are either based on simple univariate regression models or their derived forms for variable selection (Zhang 2016). However, multivariate regression models are more suitable for multifactorial phenomena due to their ability to provide synergistic and antagonistic interrelation for explanatory variables (Xing et al. 2017; Glover et al. 2015; Martinez and Kolodner 2010), a typical condition when dealing with complex diseases like cancer.

Nevertheless, such traditional models are susceptible to data idiosyncrasy. For instance, considering the high number of covariates usually present in gene expression data, it may be a challenging task to build Cox models accounting for all of them with high accuracy (Xu 2012). In a first attempt to overcome this limitation, some methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) have been implemented to simultaneously estimate coefficients and treat data high dimensionality using variable selection techniques (Tibshirani 1997). Nonetheless, these implementations ordinarily exhibit poor performance for large datasets (*e.g.*, gene expression data generated by RNA sequencing methodologies), leading to struggling in the algorithms' convergence steps. Additionally, high collinearity and low variance of gene

expression may result in incorrect estimations of the individual contributions of genes and even the identification of redundant variables in a derived model (Lesaffre and Marx 1993). Moreover, finding and testing the prognostic value or biomarker potential of a gene set is a demanding task for researchers and clinicians without extensive bioinformatics training (Zhang et al. 2019). To aid, several computational tools have been created, but still with flaws inherent to them, namely: (i) finding genes that are suitable for accomplishing the user's goals; (ii) difficulties to determine the exact data type and even the appropriate method for user's experiments; (iii) impossibility to customize analyses and inputs, among others (Gill et al. 2016). An easy-to-use command-line tool is routinely a worthy and more powerful option.

Here, we present Reboot, an algorithm to identify and validate gene or transcript signatures highly associated with patient prognosis. Reboot innovates by using a multivariate strategy with penalized Cox regression (LASSO method (Tibshirani 1997)) combined with a bootstrap approach (Efron 1992). Our algorithm deals with collinear variables inherent in gene expression data preventing redundancies and incorrect estimates, also/thereby removing genes with low impact on survival (*i.e.*, low expression variance among individuals). Reboot provides complementary statistical tests to bolster gene signatures associated with patient survival or any other endpoint chosen. Furthermore, Reboot generates supporting figures, such as Kaplan-Meier curves (Jager et al. 2008) and Forest Plots (Lewis and Clarke 2001) to facilitate the interpretation of survival outcomes. Finally, Reboot seeks not only genes but also splicing isoforms (transcripts) associated with patient prognosis, successfully managing to cope with the escalation of the analysis and incorporating a deeper level of transcriptomic data interpretation to survival analyses in a practical way.

RESULTS

Implementation and availability

Reboot is written in R (v3.6) and comprises two major modules: regression and survival (**Figure 1**). These two modules were designed to work independently, allowing users to identify genetic signatures using the "regression" module, and to test the significance of these signatures in prognosis using the "survival" module, possibly with additional validation datasets. Moreover, we also provide a "complete" mode option which enables the integrated execution of the two modules in case the same dataset is intended to be used in both analyses.

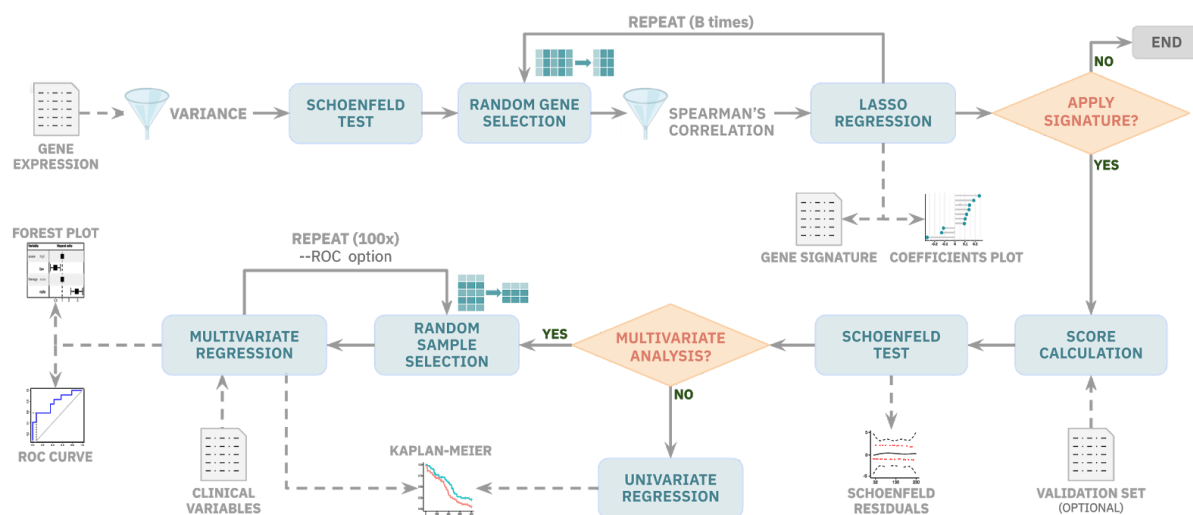


Figure 1. Reboot pipeline automatically integrates robust statistical tests, provides plots, and allows users to control parameters straightforwardly. In module I, gene or transcript expression data are filtered for variance and cox proportional hazard assumptions. Then, genes go through a bootstrap resampling random selection for LASSO regression and signature generation if they are not significantly correlated. In module II, a signature-based score is created and applied in survival analysis. Users are able to perform multivariate analyses, with or without bootstrap resampling and ROC curves, if clinical data are available. Plots are automatically yielded to the users.

The Reboot "regression" module is an easy-to-run step, which encapsulates statistical models to identify genes or splicing isoforms (transcripts) signatures. In brief, this module starts by checking and removing data attributes with variance lower than a user-defined or default cutoff. Next, a Schoenfeld test (Abeysekera and Sooriyarachchi 2009) is applied in a univariate way for each remaining attribute in the dataset using the packages "survival" (Therneau 2015)

and "survminer" (Kassambara et al. 2019). Every attribute that fails this proportional hazard assumption test is automatically removed from the analysis. After that, a Spearman's correlation (Akoglu 2018) filter is applied to every iteration of the bootstrap process (Efron 1992) based on the settable fraction of pairs with a correlation coefficient higher than 0.8 and a p-value < 0.05. Lastly, a bootstrap process is executed, which consists of a random sampling of attributes to be evaluated in a multivariate analysis. Regression itself is performed using a Least Absolute Shrinkage and Selection Operator (LASSO) algorithm (Tibshirani 1997) from the R packages "penalized" (Goeman 2010) and "survival" (Therneau 2015).

The next step in Reboot is the "survival" module, which is also easily executable. It receives and tests a gene/transcript signature produced in the previous (regression) module. In this step, the Reboot algorithm first produces and applies a score for all samples based on the gene/transcript signature coefficients obtained from the "regression" module and their corresponding expression values using the following formula: $\sum_{n=1}^n C \times E$, where "C" is the coefficient and "E" is the expression value. Next, the Schoenfeld test is applied to verify whether the score addresses the Cox model assumptions. Based on the median value (default) of the obtained scores, all individuals being tested are stratified into two groups, "low" or "high" score. The log-rank test (Bland and Altman 2004) is then applied in order to assess the relevance of the observed differences and to evaluate the relevance of the signature score as a prognostic factor for a given event, such as overall survival, progression- or recurrence-free survival. Next, a Kaplan-Meier survival curve is generated (Jager et al. 2008)(Schröder et al. 2011)(Jager et al. 2008) Of note, Reboot offers a multivariate option that allows extension of the survival model by additional clinical variables (e.g., therapy, age, and gender) (van Dijk et al. 2008). If this option is chosen, after applying the Schoenfeld test to all variables, multiple univariate analyses are performed and only variables under a minimal threshold (see Methods) are selected for the final multivariate model (Bradburn et al. 2003b) and illustrated in a forest plot using the R package "forestmodel" (Kennedy 2018).

Moreover, Reboot has an alternative to the use of the median value as a cutoff to stratify patients (e.g., into "low" and "high" groups, based on gene expression); this cutoff value may be defined using a Receiver Operator Characteristic (ROC) curve (López-Ratón et al.

2014)(Obuchowski and Bullen 2018)(López-Ratón et al. 2014) with the Nearest Neighbour Estimate (NNE) method and the Youden statistics (Hajian-Tilaki 2013). In this case, a patient-oriented bootstrap resampling strategy is performed using the R package "sjstats" (Lüdecke 2020). In order to derive highly confident and robust results, additional filters are applied such as null data removal, the minimum number of co-variables available and proportionality requirements (Clark et al. 2003). As a consequence, these filters ensure that the final dataset is composed of at least 70% of patients' data present in the original one. Additionally, the final dataset also has a minimum of two co-variables to be tested with the score, whose less abundant categories' frequencies are not smaller than 20%. After 100 cycles, the relevance frequency of each co-variable with the event is calculated and only the ones with at least 25% are plotted.

Usage and performance

Reboot was designed to be easy-to-install and of straightforward use. To generate a genetic signature, Reboot only requires a matrix of survival data along with gene/transcript expression values as input in the form of a ".tsv" file. This file should contain as its first three columns the sample identifiers, the survival status of individuals (*e.g.*, 0 = alive or 1 = dead) and the follow-up time, as well as gene expression values for multiple genes or transcripts across all individuals (file rows). In order to test a genetic signature, Reboot requires in addition to survival and expression data, a signature matrix with the previously produced regression coefficients ("tsv" file), which is not necessary when "Complete" mode is run. In case a multivariate survival analysis is requested by the user, an additional file containing clinical variables to be tested should also be provided ("tsv" file; (Bradburn et al. 2003a)). Figure S6 shows summarized examples of inputs to Reboot.

As output, Reboot generates two main textual results ("tsv" files): (i) a list of genes or transcripts that comprise the genetic signature and their corresponding regression coefficients, which explain the contribution of each gene or transcript to the signature; and (ii) the survival impact of the signature score, including hazard ratio estimates, log-rank p-values, number of samples and median survival per group, among others. In addition, multiple plots are produced: (i) a lollipop plot, displaying the ten most significant gene coefficients comprising the signature;

(ii) a coefficient histogram, displaying the distribution of all gene coefficients in the signature; (iii) a proportional hazard assumptions plot (Schoenfeld test); and (iv) a Kaplan-Meier plot (**Figure 1**). Should the multivariate option be chosen, Reboot returns all files and plots generated in the univariate analysis plus an additional file (in “.tsv” format as well) containing the survival results of the signature score along with all other clinical variables, also visible as a forest plot. Furthermore, if the score stratification is performed with the ROC method, the curve is also plotted. Finally, a histogram of co-variable frequencies is also provided in case the multivariate option was done with bootstrap resampling.

In order to analyze the performance and features of Reboot, we built a toy dataset containing clinical and random gene expression data from The Cancer Genome Atlas (TCGA; (Grossman et al. 2016)) **Tables S1 and S2**. Correlation between the number of iterations and execution time was assessed by varying the number of iterations and keeping group size and number of instances (patients) constant in two independent tests using either a server or a laptop (see Methods for details). As expected, a linear behavior for running time was observed and server performance was slightly better than laptop. Considering Reboot modules separately, regression massively accounts for the total running time (**Figure 2A and Figure 2B**). In another test, group size variation was also evaluated, while the number of iterations and patients was kept constant. In contrast to the previous experiment, group size variation presented a weak linear correlation. Nevertheless, the server again performed slightly better both in regression and survival modules, and time proportion between them showed a preponderance of the regression module, as expected (**Figure S2A, Figure S2B**). Furthermore, the number of patients also presented a linear behavior when tested in an isolated way (**Figure S2C**).

The frequency of sampling for the analysis follows a distribution in which the expected average is given by equation 1, where “G” is the group size, “B” the number of iterations, and “N” the total number of attributes.

$$\text{Equation (1)} \quad f = \frac{BG}{N}$$

Since “B” and “G” are both directly proportional to attribute frequency, we also sought to compute time correlation of different increasing rates of group size and number of iterations with time. For this analysis, a larger group of 500 genes was randomly selected, similar to data retrieved previously (**Table S3**). Both variables were increased by powers of 2 and 3 and multiplied, resulting in two curves containing points with the same frequency (**Figure 2C**). Group size increase was more efficient (lower time consumption) for small frequency values when compared to the number of iterations. However, group size exhibits exponential behavior, whereas number of iterations remains linear, even for high values, indicating that increasing the number of iterations is more efficient for high attribute coverage (**Figure 2C**). Besides time performance, large group sizes may introduce convergence issues as dimensionality increases. Reboot deals with convergence failure by reshuffling and recomputing coefficients, which may also introduce bias for large groups.

Additionally, we performed assays to compare the impact of algorithm differences and to justify why we picked LASSO instead of others. Firstly, LASSO and Ridge regressions (Hoerl and Kennard 1970) were run with a group size of 10 and 1000 iterations (**Figure 2D**). Non-zero coefficients were extracted and a distribution was built for each analysis. As expected, the LASSO algorithm used in Reboot compresses coefficients more efficiently, denoted by the highly populated regions around 0 in relation to Ridge (**Figure 2D**). Remarkably, LASSO also tends to keep significant coefficients closer to 0, while Ridge allows drastic distancing from coefficients (**Figure 2D**). LASSO also exhibits its variable selection feature, compressing coefficients to zero, and leaving a lower number of significant coefficients (**Figure S1D**).

Finally, we scrutinized the total sampling frequency of the attributes. Data obtained for **Figure S1A** was used to compute gene frequency, according to equation 1, by varying “G” (**Figure 2E**). Mean standard deviation for all eight distributions was 4.93, contributing to a reliable uniformity of variable assessment despite the stochastic process associated with the iterative process. Obtaining a satisfactory frequency depends on the total number of attributes “N”, the group size chosen “G” and the number of iterations “B”. It is recommended that the frequency of each attribute be N/G . In accordance with equation 1, “B” may be chosen in terms of equation 2.

$$B = \left(\frac{N}{G} \right)^2$$

Given that a free variation of “B” performs better in terms of computational time and prevents bias, “G” may be chosen for restricted lower values and “B” estimated, with no restrictions.

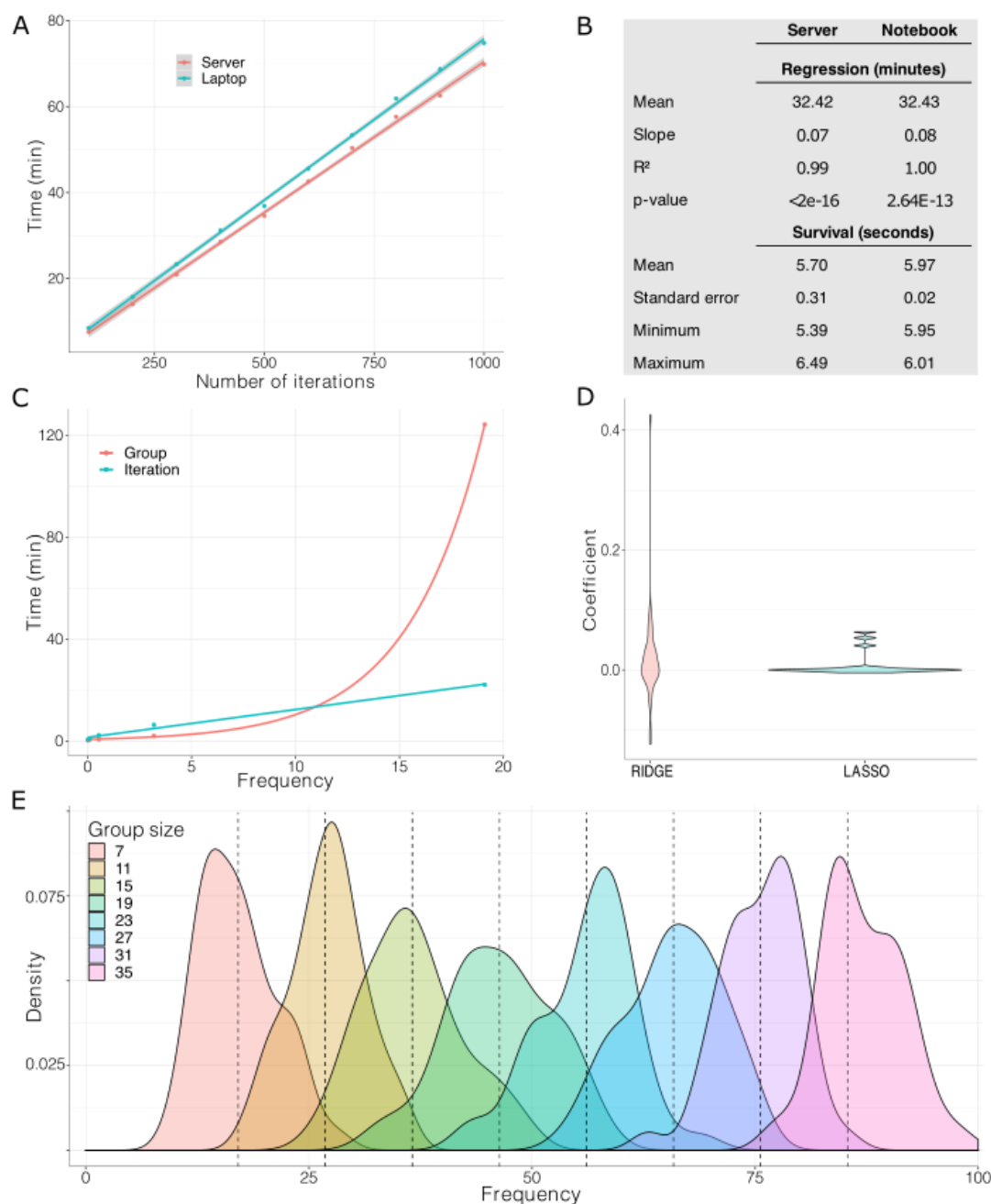


Figure 2: Reboot is computationally efficient. (A) Evaluation of total execution time for a Reboot complete run in a server and a laptop according to the number of iterations. Number of iterations varied from 100 to 1000 in steps of 100, keeping group size 20, 145 patients, and 50 randomly chosen genes. Survival was performed in multivariate

mode. (B) Table with extracted parameters obtained in A. (C) Time assay comparing the impact of group size and the number of iterations on time of execution. Group size and the number of iterations varied from 3 to 243 in powers of 3 and from 2 to 32 in powers of 2, respectively, in both curves. Legend attribution corresponds to the variable that varied in powers of 3. (D) Coefficient distribution profile obtained from LASSO and Ridge algorithms. (E) Frequency distribution of attribute selection performed with group size variation. Theoretical average is shown in dashed lines.

Using Reboot to identify genes related to prognosis in glioblastoma

To show how straightforward, useful and fast Reboot can be, we have applied it to a set of 1,013 protein-coding genes up-regulated in glioblastoma (GBM) in comparison to low-grade glioma (LGG) patients ($\log_2\text{FoldChange} \geq 2$ and FDR adjusted p-value < 0.05 ; **Table S4**). Reboot was executed using the “regression” module parameters “-G 10 -P 0.3 -V 0.01 -B 1000” and its execution took 1.15 hours in a standard server (see Methods). As a result, we identified 255 genes associated with patients' overall survival (**Table S5**).

To determine whether these 255 genes could be important in GBM patient prognosis, we further investigated them. First, we performed functional enrichment analysis that revealed 131 genes (51.37%) associated with several hallmarks of cancer according to the Molecular Signatures Database (FDR < 0.01 , hypergeometric test; **Table S6, Figure 3A**). Among the top 10 enriched hallmarks, we found 49 genes linked to at least 2 hallmarks relevant for glioblastoma progression and invasion, including those defining epithelial-mesenchymal transition (Iwadata 2016), encoding components of blood coagulation (Navone et al. 2019), as well as genes up-regulated in response to hypoxia (Monteiro et al. 2017) and/or by KRAS activation (Holmen and Williams 2005), among others. Genes associated with GBM patients' survival were also enriched in a number of GO biological processes and glioblastoma-related KEGG pathways (FDR < 0.01 , hypergeometric test; **Tables S7-S8, Figure 3B**). GO groups include, but are not limited to, processes related to inflammatory response, cell adhesion, proliferation, and motility, while the glioblastoma-related KEGG pathways with the greatest number of genes were proteoglycans/pathways in cancer, PI3K-Akt signaling pathway, and focal adhesion.

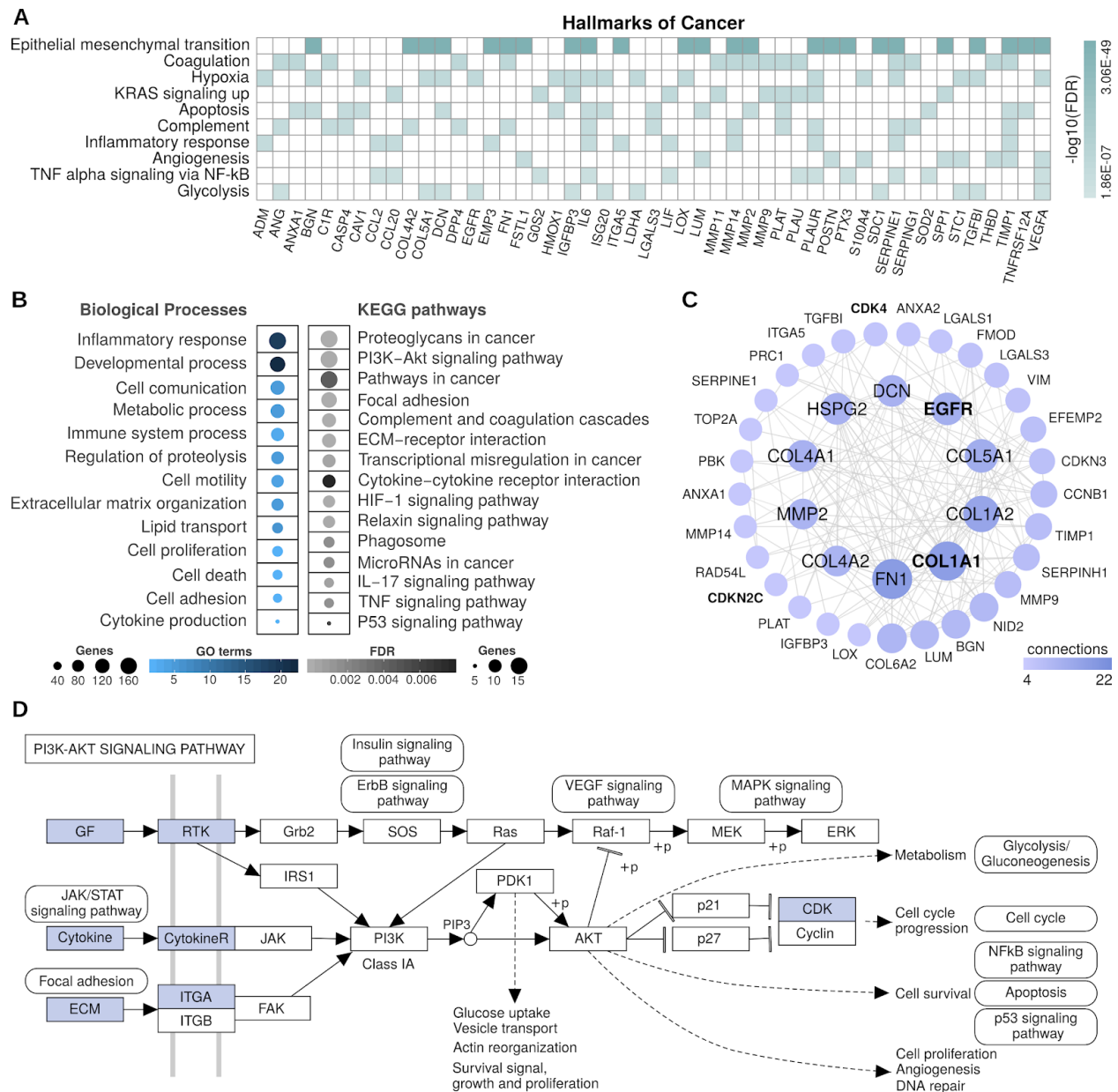


Figure 3. Functional enrichment analysis of genes associated with glioblastoma patients' overall survival obtained using Reboot regression. (A) Top 10 enriched hallmarks of cancer and genes associated with at least 2 of them; (B) Groups of enriched GO biological processes and glioblastoma-related KEGG pathways; (C) Network based on protein-protein interactions from STRING database with cancer driver genes highlighted in bold; (D) Schematic diagram of how up-regulation of 15 genes in glioblastoma may lead to activation of the PI3K-Akt signaling pathway (simplified KEGG representation with gene products highlighted in blue).

We also conducted a protein-protein interaction analysis using these 255 genes, which displayed a highly connected gene network comprising four cancer driver genes according to the Cancer Gene Census (CGC) database (<https://cancer.sanger.ac.uk/cosmic/census>; **Figure 3C**):

COL1A1, EGFR, CDK4, and CDKN2C. Moreover, according to CGC, six other driver genes were observed in the produced signature, most of them having an oncogenic role (**Table S9**). Several genes associated with GBM initiation and progression were also observed in the network, including EGFR, MMP2, HSPG2, and various members of the collagen gene family (*e.g.*, COL1A1, COL1A2, and COL5A1), which encode components of the extracellular matrix. Of note, fibronectin (FN1) was the top enriched gene in our network. An intracranial GBM xenograft model (Serres et al. 2014) showed that expression of FN1 promotes cell proliferation and resistance to ionizing irradiation, facilitates cell invasion and enhances angiogenic tumor growth. More recently, (Liao et al. 2018) provided evidence that fibronectin silencing in gliomas is associated with disruption of the PI3K-Akt signaling pathway and subsequent inhibition of cell proliferation, as well as promotion of cell apoptosis and senescence. Accordingly, we observed 15 genes highly expressed in GBM mostly encoding activators of the PI3K-Akt signaling pathway (**Figure 3D**), which is frequently activated in GBM (approximately 90%; (Langhans et al. 2017; Li et al. 2016). Altogether, these 255 candidates contain many genes already reported as relevant to GBM origin, maintenance and progression, suggesting that Reboot consistently selected a gene list potentially related to prognosis in glioblastoma.

Using Reboot to identify a minimal gene signature relevant to GBM survival

Next, we sought to determine the minimum gene set with the highest regression coefficients that are capable of explaining differences in overall survival (OS) of GBM patients using Reboot “survival” module in multivariate mode (run in docker with parameters “-M -C”; execution time ~10 seconds in a standard laptop). Out of the total 255 genes associated with patients' overall survival using Reboot (**Figure 4A; Table S4**), we identified three candidates: IKBIP, OSMR, PODNL1. They are among the top 10 genes identified as the most relevant for the prognosis of GBM patients (**Figure 4B**) and are over-expressed in GBM samples in comparison to low-grade glioma (LGG) (**Figure 4C**). Moreover, IKBIP, OSMR, PODNL1 combined score has a significant impact on survival of GBM patients (HR=0.48 95% CI: [0.32-0.71], p-value < 0.001; **Figure 4D**). The median OS for patients with a high score (> 0.34) was 335 days, yet for the low

score group was 468 days. More importantly, the obtained risk score remained significant (HR = 0.53 95% CI: [0.33-0.86], p-value = 0.01, **Figure 4E**) even when considered together with relevant clinical parameters for GBM patients, including age at diagnosis, chromosome 19/20 co-gain, G-CIMP, *IDH1* mutation, and *MGMT* methylation status.

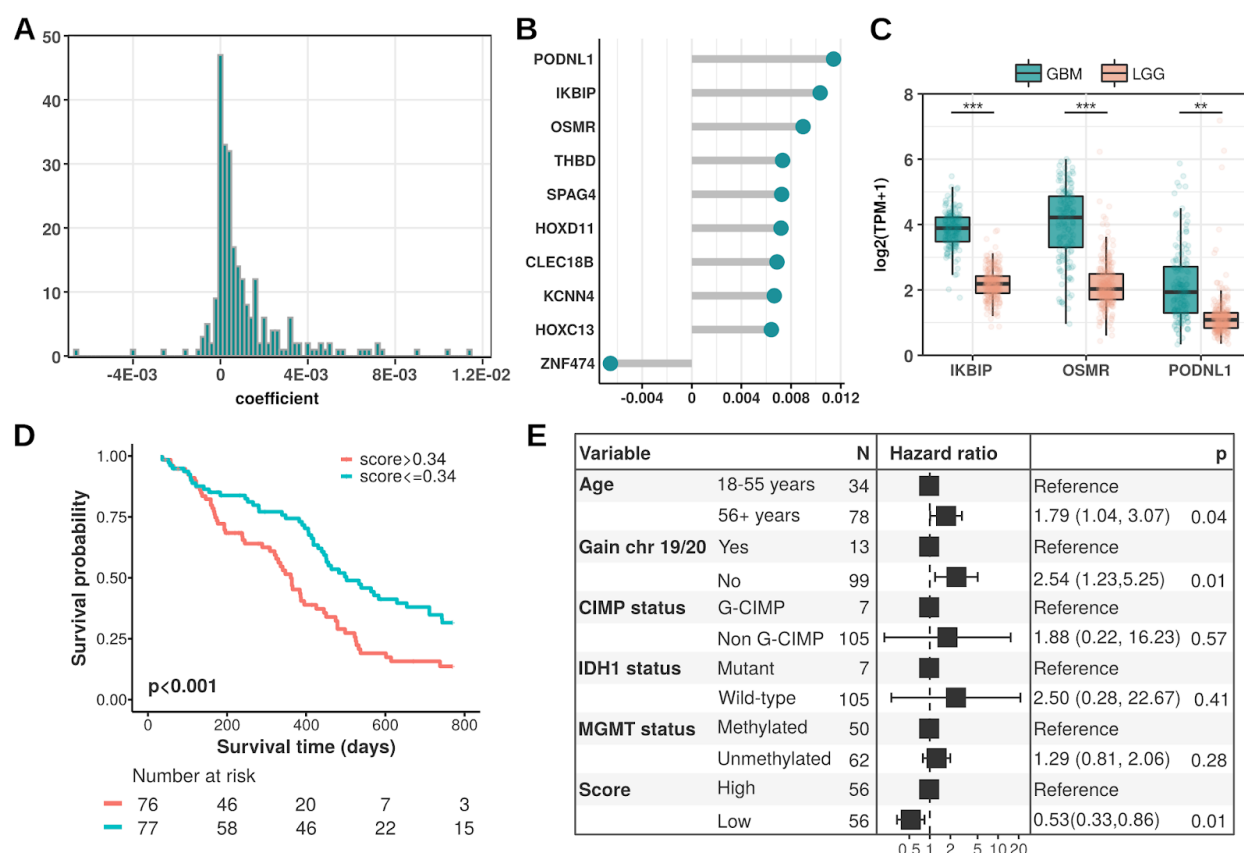


Figure 4. Reboot's application on the glioblastoma dataset. (A) Histogram displaying the distribution of all gene coefficients obtained using Reboot regression; (B) Top 10 genes identified as relevant for the prognosis of GBM patients; (C) Boxplots displaying the expression values of a 3-gene signature identified in GBM patients with Reboot (Wilcoxon test; **p<0.01, ***p<0.001); (D) Kaplan-Meier curve based on the 3-gene signature score identified in GBM patients with Reboot; and (E) Forest plot of a multivariate model including the 3-gene signature score adjusted for clinical parameters relevant to prognosis in glioblastoma.

In addition, we attempted to validate this three-gene signature in an independent cohort of 71 primary glioblastoma patients from the Chinese Glioma Genome Atlas (CGGA). Similarly, higher combined scores tended to be associated with worse prognosis of GBM patients (HR=0.66 95% CI: [0.38-1.15], p-value = 0.14; **Figure S1**). The median OS for patients with

higher scores (> 0.44) was 381 days versus 550 days for the low score group. Although we observed a clear separation between the higher and lower score groups in the CGGA cohort, the lack of statistical support might be explained by the smaller CGGA cohort size and sequencing depth (71 samples, ~22.5 million *reads* on average) compared to TCGA (154 samples, ~64.8 million *reads* on average). Thus, this result indicated that Reboot was efficient in selecting a minimal gene signature, in which a high expression is associated with worse prognosis to GBM patients.

Finding alternative splicing isoforms signature relevant to pancreatic adenocarcinoma patients' prognosis with Reboot

Next, we used Reboot to find splicing isoforms related to pancreatic adenocarcinoma patients' prognosis. First, we retrieved expression data from 167 healthy pancreatic tissues from The Genotype-Tissue Expression (GTEx; (Carithers et al. 2015) and from 177 pancreatic adenocarcinomas (PAAD) samples from TCGA (Grossman et al. 2016). For cancer data, we also obtained their respective clinical information (Liu et al. 2018). Next, by using Kallisto (Bray et al. 2016) and SUPPA2 (Trincado et al. 2018) tools, we found a complete set of alternative splicing isoforms (ASI) in pancreatic adenocarcinoma, which fed the Reboot's algorithm to perform the signature (part I) and the survival (part II) analyses (**Figure 5A**).

We found 386 significantly ASI, of which 224 and 162 were up-regulated and down-regulated, respectively, in PAAD versus healthy pancreas tissue (**Figure 5B**). To prove the Reboot robustness in candidate selection, we randomly split the initial ASI data into training (70%) and validation (30%) sets (**Figure 5C**). When applying the “regression” module on the training dataset (parameters: -B 100 -G 10 -P 0.3 -V 0.036 -F FALSE; its execution took 4.71 minutes in a standard laptop), a signature with 62 transcripts emerged (**Table S10 and Figure S3**). After setting a cutoff of 0.035 to coefficients, Reboot found a minimal signature of seven transcripts (CENPF-201; MLKL-202; NUP54-201; MCF2L-201; TFDP1-207; BBS1-206; HTT-202) presenting significant survival results (**Figure 5C and S4, Table S10**).

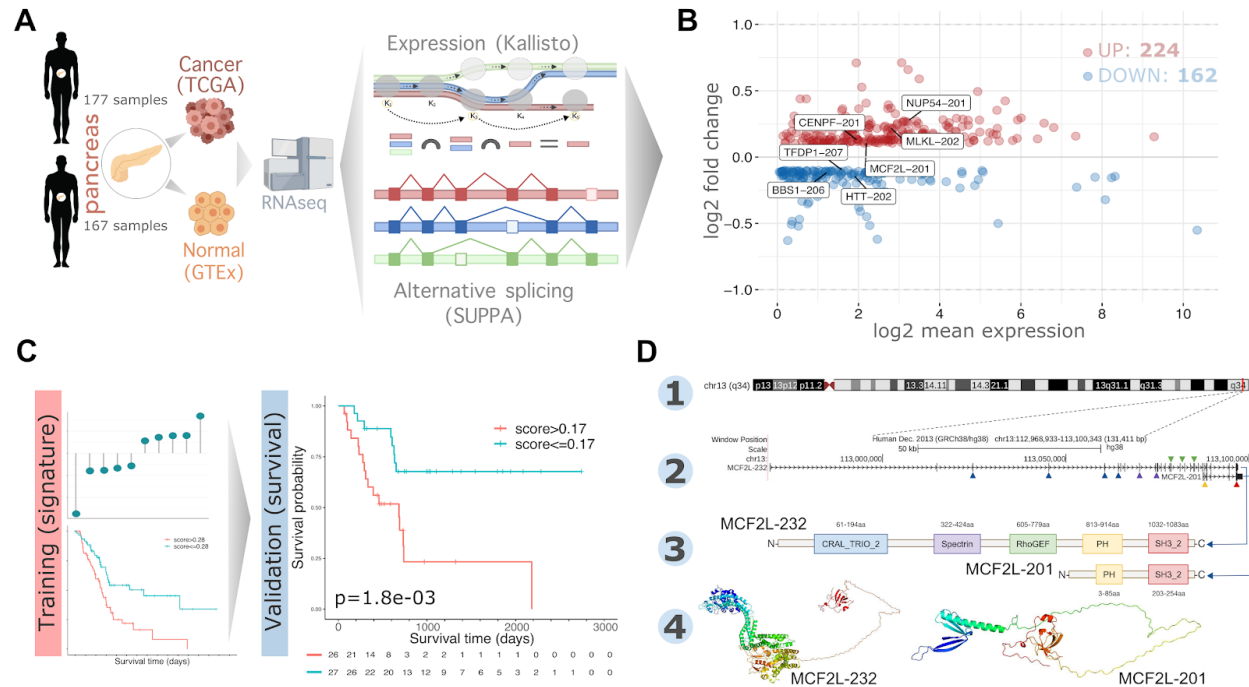


Figure 5. Reboot selects alternative splicing isoforms associated with pancreatic adenocarcinoma tumorigenesis and its patient's prognosis. (A) Selection of alternative splicing isoform (ASI) based on transcript expression data from healthy (GTEx) and tumoral (TCGA) pancreas. (B) MA plot showing the mean expression (in TPM) and Δ PSI (percent spliced in) values of all ASI. Highlighted ASIs compose the seven-transcripts signature generated with Reboot. (C) ASI data was split into training (70%) and validation (30%) set to find a transcript signature in survival analysis. Kaplan-Meiers made by Reboot when using both the training (HR: 0.4428 [0.2719 - 0.7211]; $p = 8e-04$) and the validation dataset (HR: 0.2791 [0.1191 - 0.6541]; $p = 0.0018$) showed a worse survival outcome for patients with higher scores (above median value). Follow-up time (days) is shown in the bottom for each group. (D) MCF2L mapping on the reference genome (1). Canonical (longer: MCF2L-232) and ASI (shorter: MCF2L-201) isoforms, respectively(2). Protein domains encoded from canonical (MCF2L-232) and ASI (MCF2L-201) isoforms, respectively. (3). Predicted 3D protein structure for canonical (MCF2L-232) and ASI (MCF2L-201) transcripts (4).

When testing the signature with module II (survival) of Reboot on the validation dataset (53 patients; parameters: -M TRUE -R FALSE -F FALSE; execution time ~5 seconds in a standard laptop), we found that patients with higher scores (values above the median) had worse overall survival (HR: 0.2791 [0.1191 - 0.6541]; p -value = 0.0018; **Figure 5C** and **Figure S4**). The median OS for patients in the high score group (> 0.17) was 684 days, whereas this value could not be calculated in the low score group since less than half of the patients died.

Furthermore, this result remained statistically significant after the multivariate analysis, accounting for relevant clinical variables such as age, gender, race, Tumor Node Metastasis (TNM) classification (Amin et al. 2017), histology, and grade (HR: 0.3806 [0.1569 - 0.923]; p-value = 0.0326; **Figure S4**). Additionally, the same results were observed when applying the score on the training dataset, as expected (**Figure S4**), where the median OS for patients in the high score group (> 0.28) was 517 days versus 1332 days for the low score group. Of note, other endpoints such as Disease-Specific (DSS), Progression-Free (PFI) or Disease-Free Interval (DFI) may be used instead in order to better fit the data and meet survival requirements (Liu et al. 2018).

Further investigation was performed for transcripts with major contributions to the genetic score (**Figure S4**). MCF2L-201, which had a significant positive score, lacks three protein domains: RhoGEF, Spectrin and CRAL_TRIO_2, which are all present in the canonical isoform MCF2L-232 (**Figure 5D**). However, MCF2L-201 maintains PH and SH3_2, which are also present in the canonical isoform. Another isoform, HTT-202, which scored negatively in our signature, lacks the huntingtin protein region DUF3652, present in the canonical isoform HTT-201 (**Figure S5**). Taken together, these results demonstrate that Reboot is effective not only to identify relevant genes but also splicing isoforms related to cancer.

DISCUSSION

In the past few years, advances in RNA sequencing technology have provided us an unprecedented opportunity to find novel gene signatures acting as prognostic or diagnostic biomarkers in cancer (Byron et al. 2016). Notwithstanding, treating high dimensionality of gene expression integrated with clinical variables are major challenges when performing this analysis, notably by researchers without extensive training in computational biology. It is therefore an urgent task to establish robust and straightforward methods capable of handling ultra-large datasets and finding these genes. Here we describe Reboot, a user-friendly algorithm to seek, evaluate, and validate genes and splicing isoforms signatures acting as prognostic or diagnostic biomarkers in cancer. Reboot is original and efficient: i) it combines a multivariate strategy with penalized Cox regression (LASSO method) and a bootstrap approach, plus a variety of statistical tests to find genes or transcripts candidates; ii) it is easy-to-use, well documented, and of simple installation; iii) it includes effortless steps to plot its results and to facilitate data interpretation and further analyses.

Availability of tools that manage to escalate genetic score analysis with high dimensional datasets, such as those found in gene expression data, are scarce. Beyond that, to the best of our knowledge, there is no state-of-the-art pipeline that integrates posterior validation, including clinical data, gene, and splicing isoforms expression. As genetic analyses get wider in order to capture the complexity of intricate diseases such as cancer (Frampton et al. 2013; Cheng et al. 2015; Campesato et al. 2015)), coverage of the transcriptome with more genes or transcripts (splicing isoforms) becomes crucial, which significantly raises the dimension of input datasets. Moreover, beyond all implemented filters that automate pre-processing, bootstrapping strategy allows a wide range of dataset dimensions. For instance, the number of iterations has a strong linear correlation with time. The association with the LASSO algorithm, which presents a strong variable selection feature, balances score attributes in the analysis, yielding less false positives but still exhibiting gains of a multivariate analysis. All of these approaches were included in Reboot, which introduces a robust, complete, and computationally efficient algorithm.

We selected and tested Reboot on data from glioblastoma (GBM) and pancreatic adenocarcinoma (PAAD), two cancer types presenting a poor survival rate and limited therapeutic options for their patients (Kleeff et al. 2016)(Di Carlo et al. 2019)(Kleeff et al. 2016). First, we identified prognostic genes in GBM, which were associated with a number of processes relevant for glioblastoma tumorigenesis, progression, and invasion (*e.g.*, epithelial-mesenchymal transition, inflammatory response, and cell proliferation). This list includes genes already described as related to GBM or other gliomas. In particular, the epidermal growth factor receptor (EGFR) is a primary driver of glioblastoma tumorigenesis, contributing mainly to cell proliferation and invasion (Huang et al. 2009). MMP2 is also highly expressed in gliomas and it was recently associated with stimulation of vasculogenic mimicry in glioma cells (Liu et al. 2019). HSPG2, in glioma tissues, is associated with the transformation of brain extracellular matrix into tumour microenvironment and represents a negative prognostic factor in overall and relapse-free survival (Ma et al. 2018; Kazanskaya et al. 2018).

Next, by using the “survival” module in multivariate mode, Reboot found a signature containing three genes (IKBIP, OSMR, and PODNL1) associated with GBM patients’ overall survival. Interestingly, they have emerged as prominent genes in glioblastoma: i) IKBIP has been described as a novel p53 target with pro-apoptotic activity, whose high expression is associated with poor prognosis in GBM (Cao et al. 2019; Long and Li 2019). Although in our results *MGMT* methylation was not considered a significant co-variable, another study has identified gene IKBIP as part of a signature that predicts prognosis only in GBM patients with methylated *MGMT* promoter (Wang et al. 2016). OSMR, characterized as a novel key regulator of glioblastoma tumorigenesis through EGFRvIII-STAT3 signaling, also correlates with poor prognosis in GBM patients both independently and also as part of a 4-gene signature (Mohan et al. 2017; Cao et al. 2019). Interestingly, PODNL1 encodes a protein involved in extracellular matrix formation with an unclear role in GBM tumorigenesis. The latter gene up-regulation has also been correlated with the poorest survival rates in GBM patients in distinct studies (Shergalis et al. 2018; Yan et al. 2013). Altogether, Reboot identified a valuable set of genes to be further and deeper investigated in GBM.

Second, we used Reboot to seek for alternative splicing isoforms associated with pancreatic adenocarcinoma (PAAD) patients' prognosis. As illustrated in our analyses, a genetic score obtained from differentially expressed transcripts stratifies patients with bad and good prognosis as efficiently as from gene analyses. Interestingly, a signature score with only seven transcripts was enough to yield statistical significance in the survival analysis of PAAD patients. Among them, only three isoforms are canonical (CENPF, MLKL, NUP54). Some of these genes (*e.g.*, CENPF, MLKL, TFDPI, MCF2L) have a known influence on cancer, while others (*e.g.*, NUP54, BBS1, and HTT) have been superficially studied in a tumoral context. CENPF, for instance, was already related to worse outcomes and survival in several cancer types (R. Li et al. 2020; H.-B. Liu, Huang, and Luo 2020; Garcés et al. 2020; Ying Liu et al. 2020). Another great example is the MLKL gene. This gene was shown to be up-regulated in pancreatic cancer, as we observed with Reboot, especially under the context of invasion (Ando et al. 2020). The transcription factor TFDPI is another gene whose functions remain uncovered in cancer, but significant somatic copy number alterations and corresponding somatic gene expression changes were observed in papillary thyroid carcinomas (Yang et al. 2020). Additionally, it is considered a prognostic marker in liver cancer (unfavorable), stomach cancer (favorable) and renal cancer (favorable) according to The Human Protein Atlas (Uhlen et al. 2017). Inconsistencies in these results may have arisen from a possible divergence of the role of different isoforms from this gene. Our results indicate that an isoform (TFDPI-207, down-regulated in our analysis) other than the canonical (TFDPI-201, up-regulated in our data) is of great significance for PAAD patient prognosis, an evidence that more detailed scrutiny is required for this gene (<https://www.proteinatlas.org/ENSG00000198176-TFDPI/pathology>). Taken together, it is clear that transcript-centered analysis may shed light on more detailed molecular mechanisms that would not be possible in a gene-based analysis. Therefore, the results provided by Reboot seem trustworthy and robust.

Nucleoporin 54 (NUP54), on the other hand, has not been extensively studied in cancer. Even though an association between larger tumor size and loss of heterozygosity (LOH) in NUP54 was found in patients with hepatocellular carcinoma, potentially playing an important role in its development and progression (G.-L. Huang et al. 2010). Moreover, it was recently

reported a novel function for NUP54 in the response to ionizing radiation and the maintenance of homologous recombination-mediated genome integrity in cell lines (Rodriguez-Berriguete et al. 2018).

Among the best-scored transcripts, MCF2L-201, which was found to be up-regulated in PAAD, is a compelling example. The canonical isoform of the MCF2L gene (MCF2L-232) encodes DBL from the Guanine Exchange Factor protein family, known to directly interact and regulate important factors for cell cycle such as Cdc42 and RhoA complexes (Jaiswal et al. 2013). It has been shown that the minimal and sufficient catalytic activity of DBL is composed of a DH and a PH domain linked in tandem (Cerione and Zheng 1996). Although MCF2L-201 does not present a DH domain, it keeps a PH and a SH3 domain. PH domains perform essential contact with Cdc42 and RhoA in the DBL structure (Snyder et al. 2002). They are also known to be responsible for protein subcellular localization and phosphoinositide interaction (Zheng et al. 1996; Han et al. 1998). Moreover, SH3 (Src homology 3) domains are abundant in oncogenic pathways such as cell migration and proliferation, distributed along with many different protein structures (Birge et al. 1996). SH3 domains have also been implicated in pancreatic cancer specifically due to its relevance for oncogenic pathways (Thalappilly et al. 2008). Although a few isoforms of MCF2L have been initially explored, such as MCF2L-203 - which does not catalyze guanine nucleotide exchange on CDC42 - and MCF2L-205 - which, on the other hand, activates CDC42 (Ueda et al. 2004) - MCF2L-201 requires further investigation. Details about how the PH-SH3 protein may act and its role in pancreatic cancer deserve deeper analyses, even though our study provides some guidance on that.

Despite the fact that Huntingtin is mostly known to cause Huntington disease by the expansion of the trinucleotide CAG in its first exon, it has a wide tissue expression and its trinucleotide expansion has been correlated to tumor progression, including metastasis, and inversely correlated to carcinogenesis (Thion and Humbert 2018). Huntingtin transcript HTT-202 is non-canonical and we found it down-regulated in pancreatic tumors. Its protein structure presents neither the characteristic polymorphic trinucleotide repetitive region nor the main huntingtin annotated domain: DUF3652; thus, its function is an enigma. A similar case involves the BBS1 gene since it is most known for its association with the Bardet-Biedl Syndrome (BBS)

(Forsythe and Beales 2003)(Beales et al. 2000)(Forsythe and Beales 2003). More interesting is the fact that higher expression of BBS1 was related to better survival in patients with malignant pleural mesothelioma (Vavougios et al. 2015), although in our PAAD signature this gene was down-regulated. Furthermore, BBS1 was part of a 15-gene signature associated with bone metastasis in breast carcinomas. Specifically, its up-regulation was correlated to the epithelial to mesenchymal transition status of the tumor (Savci-Heijink et al. 2016). Thus, Reboot's algorithm makes splicing isoform expression analysis feasible in cancer prognosis, also providing clues to decipher unclear molecular mechanisms of tumorigenesis and cancer progression.

CONCLUSIONS

In conclusion, here we present Reboot, an algorithm to seek, evaluate, and validate genes and transcripts signatures acting as prognostic or diagnostic biomarkers in cancer. Reboot brings novelties by combining a multivariate strategy with penalized Cox regression (LASSO method) and bootstrap approach, plus a variety of statistical tests to find genes or transcripts candidates. Reboot shows its usefulness by identifying prognostic genes and a minimal set of genes associated with glioblastoma patients' survival and a splicing isoforms signature associated with pancreatic adenocarcinoma. We believe that Reboot will be of immediate interest to the cancer research community because it will accelerate and democratize the search for genes and transcripts biomarkers, even by researchers and clinicians without extensive bioinformatics training.

METHODS

Usage and performance

Expression and clinical data from TCGA were obtained from individuals that presented only a single primary glioblastoma tumor by an *in-house* R script (toyfordocker.R found in <https://galantelab.github.io/reboot>). Gene expression was obtained from processed datasets in FPKM. The same 50 randomly picked genes were used in all essays with exception of concomitant group size and number of iterations variation, in which 500 genes were randomly picked. For time comparisons, laptop and server specifications are: CPU: Intel(R) Xeon(R) Silver 4114, 2.20 GHz, 128 GB of ram; and CPU: Intel(R) Core (TM) i7-8550U 1.80 GHz, 16GB of RAM, respectively. All-time essays were computed with the parameter “M” and all others were set default unless otherwise stated. All linear regressions (Pearson) and plots were generated in R.

Gene expression profiles

Mapped RNA-seq and clinical data of 154 and 248 samples from patients with primary glioblastoma (GBM) and low grade glioma (LGG grade II), respectively, were downloaded from TCGA (Grossman et al. 2016). We also obtained RNA sequencing and clinical data of 71 samples from primary GBM patients from the Chinese Glioma Genome Atlas (CGGA) for validation purposes. All datasets were processed using Kallisto (Bray et al. 2016) with GENCODE (v29) as reference to the human transcriptome. Gene-level counts (TPM normalized) were obtained using tximport (Soneson et al. 2015).

Differential gene expression

Differential gene expression of GBM versus LGG-II samples from TCGA was performed using DESeq2 (Love et al. 2014), and we considered as up-regulated only genes presenting a $\log_2\text{FoldChange} \geq 2$ and false discovery rate (FDR) adjusted p-value < 0.05 .

Functional analyses

For Gene Ontology enrichment analysis, we used ShinyGO (Ge et al. 2019) and REVIGO (Supek et al. 2011) web tools. ShinyGO was also used to evaluate cancer hallmarks from the Molecular Signatures Database - MSigDB (www.gsea-msigdb.org/gsea/msigdb/) and KEGG pathways (www.genome.jp/kegg/). Only functional terms with an FDR < 0.01 (hypergeometric test) were considered relevant. Protein-protein interaction analysis was performed in Cytoscape (Shannon et al. 2003) using the STRING database (Szklarczyk et al. 2019).

Transcript expression profiles

Mapped RNA-seq data from 177 tumor patients with pancreatic adenocarcinoma from TCGA (Grossman et al. 2016), along with their respective clinical data (J. Liu et al. 2018), were downloaded. The TCGA dataset was reprocessed using Kallisto (Bray et al. 2016) with GENCODE (v29) as reference to the human transcriptome and transcript-level counts (TPM normalized) were obtained using tximport (Soneson et al. 2015). We directly downloaded transcript-level TPM normalized data of pancreatic samples from 167 healthy individuals from The Genotype-Tissue Expression (GTEx; Carithers et al. 2015) through the UCSC Xena portal (toil.xenahubs.net/). Only transcripts available in both datasets were further considered in the analyses.

Differential transcript expression

Analysis of differential transcripts usage between tumoral (pancreatic adenocarcinoma) and healthy pancreatic samples from TCGA and GTEx was performed using SUPPA2 (Trincado et al. 2018; version 2.3). We considered as significant only transcripts presenting a $|\Delta\text{PSI}| \geq 0.1$ and FDR adjusted p-value ≤ 0.05 .

3D structure prediction

MCFL2-201, MCF2L-232, HTT-201 and HTT-202 transcript sequences were submitted to ORFfinder (Tatusov and Tatusov 2007) with default values. The longest positive open reading frames (ORFs) were then submitted to Pfam (Punta et al. 2012). Finally, the amino acid

sequences of both transcripts were submitted to Phyre2 ((Kelley et al. 2015)version 2.0(Kelley et al. 2015) for 3D structure prediction in “intensive” mode.

Availability and updates

Reboot is implemented as an R package that is freely available under the GNU General Public Licence version 3 (GPL3) at <https://galantelab.github.io/reboot/>. Reboot updates will be announced at its webpage and each respective Docker image will be released along with new versions.

Funding

This work was partially supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (2018/15579-8), Instituto Serrapilheira, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; to PAFG). FRCS (2017/18246-7), GDAG (2017/19541-2), and FFS (2017/17974-9) are supported by fellowships from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Conflict of Interest: none declared.

REFERENCES

- Abeysekera WWM, Sooriyarachchi MR. 2009. Use of Schoenfeld’s global test to test the proportional hazards assumption in the Cox proportional hazards model: an application to a clinical study. *Journal of the National Science Foundation of Sri Lanka* **37**: 41. <http://dx.doi.org/10.4038/jnsfsr.v37i1.456>.
- Akoglu H. 2018. User’s guide to correlation coefficients. *Turk J Emerg Med* **18**: 91–93.
- Amin MB, Greene FL, Edge SB, Compton CC, Gershewald JE, Brookland RK, Meyer L, Gress DM, Byrd DR, Winchester DP. 2017. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging. *CA Cancer J Clin* **67**: 93–99.
- Beales PL, Reid HA, Griffiths MH, Maher ER, Flinter FA, Woolf AS. 2000. Renal cancer and malformations in relatives of patients with Bardet-Biedl syndrome. *Nephrol Dial Transplant* **15**: 1977–1985.
- Bland JM, Altman DG. 2004. The logrank test. *BMJ* **328**: 1073.
- Bradburn MJ, Clark TG, Love SB, Altman DG. 2003a. Survival analysis Part III: multivariate data analysis -- choosing a model and assessing its adequacy and fit. *Br J Cancer* **89**: 605–611.
- Bradburn MJ, Clark TG, Love SB, Altman DG. 2003b. Survival analysis part II: multivariate data analysis--an introduction to concepts and methods. *Br J Cancer* **89**: 431–436.
- Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat*

- Biotechnol* **34**: 525–527.
- Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. 2016. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* **17**: 257–271.
- Campeato LF, Barroso-Sousa R, Jimenez L, Correa BR, Sabbaga J, Hoff PM, Reis LFL, Galante PAF, Camargo AA. 2015. Comprehensive cancer-gene panels can be used to estimate mutational load and predict clinical benefit to PD-1 blockade in clinical practice. *Oncotarget* **6**: 34221–34227.
- Cao M, Cai J, Yuan Y, Shi Y, Wu H, Liu Q, Yao Y, Chen L, Dang W, Zhang X, et al. 2019. A four-gene signature-derived risk score for glioblastoma: prospects for prognostic and response predictive analyses. *Cancer Biol Med* **16**: 595–605.
- Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS, Peter-Demchok J, Gelfand ET, et al. 2015. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv Biobank* **13**: 311–319.
- Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN, et al. 2015. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J Mol Diagn* **17**: 251–264.
- Clark TG, Bradburn MJ, Love SB, Altman DG. 2003. Survival analysis part I: basic concepts and first analyses. *Br J Cancer* **89**: 232–238.
- Di Carlo DT, Cagnazzo F, Benedetto N, Morganti R, Perrini P. 2019. Multiple high-grade gliomas: epidemiology, management, and outcome. A systematic review and meta-analysis. *Neurosurg Rev* **42**: 263–275.
- Efron B. 1992. Bootstrap Methods: Another Look at the Jackknife. In *Breakthroughs in Statistics: Methodology and Distribution* (eds. S. Kotz and N.L. Johnson), pp. 569–593, Springer New York, New York, NY.
- Forsythe E, Beales PL. 2003. Bardet-Biedl Syndrome. In *GeneReviews* (eds. M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, K. Stephens, and A. Amemiya), University of Washington, Seattle, Seattle (WA).
- Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, Schnall-Levin M, White J, Sanford EM, An P, et al. 2013. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol* **31**: 1023–1031.
- Ge SX, Jung D, Yao R. 2019. ShinyGO: a graphical enrichment tool for animals and plants. *Bioinformatics*. <http://dx.doi.org/10.1093/bioinformatics/btz931>.
- Gill SK, Christopher AF, Gupta V, Bansal P. 2016. Emerging role of bioinformatics tools and software in evolution of clinical research. *Perspect Clin Res* **7**: 115–122.
- Glover KP, Chen Z, Markell LK, Han X. 2015. Synergistic Gene Expression Signature Observed in TK6 Cells upon Co-Exposure to UVC-Irradiation and Protein Kinase C-Activating Tumor Promoters. *PLoS One* **10**: e0139850.
- Goeman JJ. 2010. L1 penalized estimation in the Cox proportional hazards model. *Biom J* **52**: 70–84.
- Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. 2016. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* **375**: 1109–1112.
- Gutierrez E, Cahatol I, Bailey CAR, Lafargue A, Zhang N, Song Y, Tian H, Zhang Y, Chan R, Gu K, et al. 2019. Regulation of RhoB Gene Expression during Tumorigenesis and Aging Process and Its Potential Applications in These Processes. *Cancers* **11**. <http://dx.doi.org/10.3390/cancers11060818>.

- Hajian-Tilaki K. 2013. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med* **4**: 627–635.
- Hoerl AE, Kennard RW. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**: 55–67.
- Holmen SL, Williams BO. 2005. Essential role for Ras signaling in glioblastoma maintenance. *Cancer Res* **65**: 8250–8255.
- Huang M-Y, Chen H-C, Yang I-P, Tsai H-L, Wang T-N, Juo S-HH, Wang J-Y. 2013. Tumorigenesis and tumor progression related gene expression profiles in colorectal cancer. *Cancer Biomark* **13**: 269–279.
- Huang PH, Xu AM, White FM. 2009. Oncogenic EGFR signaling networks in glioma. *Sci Signal* **2**: re6.
- Hussein AA, Forouzanfar T, Bloemena E, de Visscher J, Brakenhoff RH, Leemans CR, Helder MN. 2018. A review of the most promising biomarkers for early diagnosis and prognosis prediction of tongue squamous cell carcinoma. *Br J Cancer* **119**: 724–736.
- International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, et al. 2010. International network of cancer genome projects. *Nature* **464**: 993–998.
- Iwade Y. 2016. Epithelial-mesenchymal transition in glioblastoma progression. *Oncol Lett* **11**: 1615–1620.
- Jager KJ, van Dijk PC, Zoccali C, Dekker FW. 2008. The analysis of survival data: the Kaplan-Meier method. *Kidney Int* **74**: 560–565.
- Kazanskaya GM, Tsidulko AY, Volkov AM, Kiselev RS, Suhovskih AV, Kobozev VV, Gaytan AS, Aidagulova SV, Krivoschapkin AL, Grigorieva EV. 2018. Heparan sulfate accumulation and perlecan/HSPG2 up-regulation in tumour tissue predict low relapse-free survival for patients with glioblastoma. *Histochem Cell Biol* **149**: 235–244.
- Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**: 845–858.
- Kleeff J, Korc M, Apte M, La Vecchia C, Johnson CD, Biankin AV, Neale RE, Tempero M, Tuveson DA, Hruban RH, et al. 2016. Pancreatic cancer. *Nat Rev Dis Primers* **2**: 16022.
- Langhans J, Schnee L, Trenkler N, von Bandemer H, Nonnenmacher L, Karpel-Massler G, Siegelin MD, Zhou S, Halatsch M-E, Debatin K-M, et al. 2017. The effects of PI3K-mediated signalling on glioblastoma cell behaviour. *Oncogenesis* **6**: 398.
- Lesaffre E, Marx BD. 1993. Collinearity in generalized linear regression. *Communications in Statistics - Theory and Methods* **22**: 1933–1952.
- Lewis S, Clarke M. 2001. Forest plots: trying to see the wood and the trees. *BMJ* **322**: 1479–1480.
- Liao Y-X, Zhang Z-P, Zhao J, Liu J-P. 2018. Effects of Fibronectin 1 on Cell Proliferation, Senescence and Apoptosis of Human Glioma Cells Through the PI3K/AKT Signaling Pathway. *Cell Physiol Biochem* **48**: 1382–1396.
- Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al. 2018. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**: 400–416.e11.
- Liu Y, Li F, Yang YT, Xu XD, Chen JS, Chen TL, Chen HJ, Zhu YB, Lin JY, Li Y, et al. 2019. IGFBP2 promotes vasculogenic mimicry formation via regulating CD144 and MMP2 expression in glioma. *Oncogene* **38**:

1815–1831.

- Liu Z, Yoshimi A, Wang J, Cho H, Chun-Wei Lee S, Ki M, Bitner L, Chu T, Shah H, Liu B, et al. 2020. Mutations in the RNA Splicing Factor SF3B1 Promote Tumorigenesis through MYC Stabilization. *Cancer Discov* **10**: 806–821.
- Li X, Wu C, Chen N, Gu H, Yen A, Cao L, Wang E, Wang L. 2016. PI3K/Akt/mTOR signaling pathway and targeted therapy for glioblastoma. *Oncotarget* **7**: 33440–33450.
- Long S, Li G. 2019. Comprehensive analysis of a long non-coding RNA-mediated competitive endogenous RNA network in glioblastoma multiforme. *Exp Ther Med* **18**: 1081–1090.
- López-Ratón M, Rodríguez-Álvarez MX, Suárez CC, Sampedro FG. 2014. OptimalCutpoints : An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. *J Stat Softw* **61**. <http://www.jstatsoft.org/v61/i08/>.
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.
- Martinez SL, Kolodner RD. 2010. Functional analysis of human mismatch repair gene mutations identifies weak alleles and polymorphisms capable of polygenic interactions. *Proc Natl Acad Sci U S A* **107**: 5070–5075.
- Ma X-L, Shang F, Ni W, Zhu J, Luo B, Zhang Y-Q. 2018. Increased HSPG2 expression independently predicts poor survival in patients with oligoastrocytoma and oligodendroglioma. *Eur Rev Med Pharmacol Sci* **22**: 6853–6863.
- Meng N, Chen M, Chen D, Chen X-H, Wang J-Z, Zhu S, He Y-T, Zhang X-L, Lu R-X, Yan G-R. 2020. Small Protein Hidden in lncRNA Promotes “Cancerous” RNA Splicing and Tumorigenesis. *Adv Sci* **7**: 1903233.
- Mohan S, Bonni A, Jahani-Asl A. 2017. Targeting OSMR in glioma stem cells. *Oncotarget* **8**: 16103–16104.
- Monteiro AR, Hill R, Pilkington GJ, Madureira PA. 2017. The Role of Hypoxia in Glioblastoma Invasion. *Cells* **6**. <http://dx.doi.org/10.3390/cells6040045>.
- Navone SE, Guarnaccia L, Locatelli M, Rampini P, Caroli M, La Verde N, Gaudino C, Bettinardi N, Riboni L, Marfia G, et al. 2019. Significance and Prognostic Value of The Coagulation Profile in Patients with Glioblastoma: Implications for Personalized Therapy. *World Neurosurg* **121**: e621–e629.
- Obuchowski NA, Bullen JA. 2018. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys Med Biol* **63**: 07TR01.
- Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. 2012. The Pfam protein families database. *Nucleic Acids Res* **40**: D290–301.
- Savci-Heijink CD, Halfwerk H, Koster J, van de Vijver MJ. 2016. A novel gene expression signature for bone metastasis in breast carcinomas. *Breast Cancer Res Treat* **156**: 249–259.
- Schirmacher V. 2019. From chemotherapy to biological therapy: A review of novel concepts to reduce the side effects of systemic cancer treatment (Review). *Int J Oncol* **54**: 407–419.
- Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. 2011. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics* **27**: 3206–3208.
- Serres E, Debarbieux F, Stanchi F, Maggiorella L, Grall D, Turchi L, Burel-Vandenbos F, Figarella-Branger D, Virolle T, Rougon G, et al. 2014. Fibronectin expression in glioblastomas promotes cell cohesion, collective invasion of basement membrane in vitro and orthotopic tumor growth in mice. *Oncogene* **33**: 3451–3462.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003.

- Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.
- Shergalis A, Bankhead A 3rd, Luesakul U, Muangsins N, Neamati N. 2018. Current Challenges and Opportunities in Treating Glioblastoma. *Pharmacol Rev* **70**: 412–445.
- Soneson C, Love MI, Robinson MD. 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**: 1521.
- Sturtz LA, Deyarmin B, van Laar R, Yarina W, Shriver CD, Ellsworth RE. 2014. Gene expression differences in adipose tissue associated with breast tumorigenesis. *Adipocyte* **3**: 107–114.
- Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**: e21800.
- Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. 2019. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**: D607–D613.
- Tatusov T, Tatusov R. 2007. ORF Finder (Open Reading Frame Finder). *National Center for Biotechnology Information National Institute of Health* **9**.
- Therneau T. 2015. A Package for Survival Analysis in S. version 2.38.
- Tibshirani R. 1997. The lasso method for variable selection in the Cox model. *Stat Med* **16**: 385–395.
- Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyraas E. 2018. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* **19**: 40.
- Ueda S, Kataoka T, Satoh T. 2004. Role of the Sec14-like domain of Dbl family exchange factors in the regulation of Rho family GTPases in different subcellular sites. *Cell Signal* **16**: 899–906.
- van Dijk PC, Jager KJ, Zwinderman AH, Zoccali C, Dekker FW. 2008. The analysis of survival data in nephrology: basic concepts and methods of Cox regression. *Kidney Int* **74**: 705–709.
- Vavougios GD, Solenov EI, Hatzoglou C, Baturina GS, Katkova LE, Molyvdas PA, Gourgoulis KI, Zarogiannis SG. 2015. Computational genomic analysis of PARK7 interactome reveals high BBS1 gene expression as a prognostic factor favoring survival in malignant pleural mesothelioma. *Am J Physiol Lung Cell Mol Physiol* **309**: L677–86.
- Wang W, Zhang L, Wang Z, Yang F, Wang H, Liang T, Wu F, Lan Q, Wang J, Zhao J. 2016. A three-gene signature for prognosis in patients with MGMT promoter-methylated glioblastoma. *Oncotarget* **7**: 69991–69999.
- Xing P, Chen Y, Gao J, Bai L, Yuan Z. 2017. A fast approach to detect gene–gene synergy. *Sci Rep* **7**: 1–8.
- Xu J. 2012. High-Dimensional Cox Regression Analysis in Genetic Studies with Censored Survival Outcomes. *J Probab Stat* **2012**. <https://www.hindawi.com/journals/jps/2012/478680/abs/> (Accessed March 19, 2020).
- Yan Y, Zhang L, Xu T, Zhou J, Qin R, Chen C, Zou Y, Fu D, Hu G, Chen J, et al. 2013. SAMS1 is highly expressed and associated with a poor survival in glioblastoma multiforme. *PLoS One* **8**: e81905.
- Yu L, Kim J, Jiang L, Feng B, Ying Y, Ji K-Y, Tang Q, Chen W, Mai T, Dou W, et al. 2020. MTR4 drives liver tumorigenesis by promoting cancer metabolic switch through alternative splicing. *Nat Commun* **11**: 708.
- Zhang H, Klareskog L, Matussek A, Pfister SM, Benson M. 2019. Translating genomic medicine to the clinic: challenges and opportunities. *Genome Med* **11**: 9.

Zhang Z. 2016. Variable selection with stepwise and best subset approaches. *Ann Transl Med* 4: 136.

Campesato, Luís Felipe, Romualdo Barroso-Sousa, Leandro Jimenez, Bruna R. Correa, Jorge Sabbaga, Paulo M. Hoff, Luiz F. L. Reis, Pedro Alexandre F. Galante, and Anamaria A. Camargo. 2015. “Comprehensive Cancer-Gene Panels Can Be Used to Estimate Mutational Load and Predict Clinical Benefit to PD-1 Blockade in Clinical Practice.” *Oncotarget* 6 (33): 34221–27.

Cheng, Donavan T., Talia N. Mitchell, Ahmet Zehir, Ronak H. Shah, Ryma Benayed, Aijazuddin Syed, Raghu Chandramohan, et al. 2015. “Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology.” *The Journal of Molecular Diagnostics: JMD* 17 (3): 251–64.

Frampton, Garrett M., Alex Fichtenholtz, Geoff A. Otto, Kai Wang, Sean R. Downing, Jie He, Michael Schnall-Levin, et al. 2013. “Development and Validation of a Clinical Cancer Genomic Profiling Test Based on Massively Parallel DNA Sequencing.” *Nature Biotechnology* 31 (11): 1023–31.