Article

Optimal microRNA sequencing depth to predict cancer patient survival with random forest and Cox models

Rémy Jardillier [1,2], Dzenis Koca [1], Florent Chatelain [2,†] and Laurent Guyon [1,†,*]

[1] Univ. Grenoble Alpes, Inserm, CEA, IRIG, BioSanté U1292, BCI, 38000 Grenoble, France
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Institute of Engineering University Grenoble Alpes, 38000 Grenoble, France
* Correspondence: laurent.guyon@cea.fr; Tel.: +33.438.780453 (L.G.) - florent.chatelain@gipsa-lab.grenoble-inp.fr; Tel.: +33.476.574371 (F.C.)
† These authors contributed equally to this work.

Abstract: (1) Background: tumor profiling enables patient survival prediction. The two essential parameters to be calibrated when designing a study based on tumor profiles from a cohort are the sequencing depth of RNA-seq technology and the number of patients. This calibration is carried out under cost constraints, and a compromise has to be found. In the context of survival data, the goal of this work is to benchmark the impact of the number of patients and of the sequencing depth of miRNA-seq and mRNA-seq on the predictive capabilities for both the Cox model with elastic net penalty and random survival forest. (2) Results: we first show that the Cox model and random survival forest provide comparable prediction capabilities, with significant differences for some cancers. Second, we demonstrate that miRNA and/or mRNA data improve prediction over clinical data alone. mRNA-seq data leads to slightly better prediction than miRNA-seq, with the notable exception of lung adenocarcinoma for which the tumor miRNA profile shows higher predictive power. Third, we demonstrate that the sequencing depth of RNA-seq data can be reduced for most of the investigated cancers without degrading the prediction abilities, allowing the creation of independent validation sets at lower cost. Finally, we show that the number of patients in the training dataset can be reduced for the Cox model and random survival forest, allowing the use of different models on different patient subgroups. (3) Availability: R script is available at https://github.com/remyJardillier/Survival_seq_depth

Keywords: Sequencing depth; cancer; microRNA; survival; Cox model; random survival forest model

Citation: Jardillier, R.; Koca, D.; Chatelain, F.; Guyon, L. miRNA sequencing depth required for prognosis. *Preprints* **2022**, *1*, 0. https://doi.org/

1. Introduction

microRNAs (miRNAs) are near 22-nucleotide long RNAs repressing protein coding gene expression at post-transcriptional level [1]. miRNAs have been shown to be implicated in various steps of carcinogenesis: initiation, propagation and metastasis [2]. The cancer genome atlas (TCGA) project has provided microRNA sequencing on thousands of tumor samples over 33 cancer types, together with patient follow-up [3]. miRNAs represent promising biomarkers to predict patient survival in cancer [4]. TCGA datasets are extremely valuable to build survival models with tumor miRNA expression, and contain large enough cohorts for many tumor types to evaluate the predictions. A semi-parametric and popular model to link patient survival with genomics variables, dealing with censored data and assuming proportional hazards, has been proposed by D.R. Cox [5]. Classically, in the case of high dimensional datasets, a penalty term is used to constrain the coefficients of the model, and to select only a subset of genes. Different forms of penalties exist [6], but we will focus on the elastic net penalty [7] in this paper, as we have recently shown that they provide similar performances [8]. More recently, non-parametric machine learning algorithms have been proposed and adapted to deal with survival data, including random survival forest [9]. Random survival forest potentially offers more flexibility, as it does not assume any proportionality between hazards, and takes into account non-linear effects and interactions between variables [10,11].

Building and evaluating efficient models from miRNA expression, applicable in clinics, requires to build large datasets from patient cohorts. It implies the recruitment of many patients, and the tumor miRNA profiling at a high enough sequencing depth. Increasing the number of patients and/or the sequencing depth means increasing the cost, but may not lead to direct improvement of the prediction performance of the models. Moreover, validation datasets remain scarce and expensive to build [12]. Thus, a compromise has to be found. For tumor mRNA profiling, P. Milanez-Almeida *et al.* [13] showed that the sequencing depth could be decreased by typically two orders of magnitude for TCGA datasets for most cancers when using the Cox model with elastic net penalty. They use C-index and p-value from single variable Cox model as prediction performance metrics. They argue that the saved cost could be used to increase the number of patients and/or to perform longitudinal studies.

The goal of the present work is to investigate the required miRNA and mRNA sequencing depth together with the number of patients in the training dataset to build optimal performance models according to well-established metrics (*i.e.* C-index and integrated Brier score), both with the classical Cox model with elastic net penalty and random survival forest. Additionally, we have validated our results with an independent cohort.

2. Materials and Methods

2.1. Cox model with elastic net penalty and random survival forest: the link between genetic and survival data

2.1.1. Cox proportional hazards model with elastic net penalty

Let T denote the survival time (also called the 'time-to-event'). The Cox model [5] is widely used in medicine to link covariates to survival data through the hazard function, defined for all time instants $t > 0$, $h(t) = \lim_{h \to 0} \frac{P(t \leq T < t+h | T \geq t)}{h}$, which represents the instantaneous death probability per unit of time. In the Cox model, the hazard function for patient i is modeled as follows:

$$h(t; \boldsymbol{X}^i) = h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{X}^i),$$

where $h_0(t)$ is the baseline hazard function, $\boldsymbol{X}^i = (X_1^i, \ldots, X_p^i)^T$ the vector of covariates for patient i (here as mRNA or miRNA expression, with p the number of coding or miRNA genes), and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ the vector of associated coefficients. We define Δ_i to be the associated status, as 1 for death and 0 for censoring. The vector of coefficients $\boldsymbol{\beta}$ can be estimated by maximizing the Cox pseudo-likelihood, as proposed by Breslow [14].

The elastic net methodology [7] consists of the addition of a penalty term to the log-pseudo-likelihood $l(\boldsymbol{\beta})$ before the maximization:

$$\hat{\boldsymbol{\beta}}(EN) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, l(\boldsymbol{\beta}) - \lambda \left(\alpha ||\boldsymbol{\beta}||_1 + \frac{1-\alpha}{2} ||\boldsymbol{\beta}||_2^2 \right)$$

We used the R package *glmnet* [15] to estimate Cox model with elastic net penalty. In the followig, 'Cox model' refers to 'Cox model with elastic net penalty'. For more details on the Cox model and the choice of the hyperparameters used for penalties (*i.e.* $\alpha = 0.3$), we refer the reader to Supplementary Figure S1 and Materials.

2.1.2. Random survival forest

Random forest, introduced by Breiman, is a classical ensemble algorithm for regression and classification whose principle is to build multiple decision trees and create a forest [16]. Results are averaged over all the trees. H. Ishwaran *et al.* then extended classical random forest algorithm to survival analysis with censored data [10]. At each node of each tree, m explanatory variables are randomly chosen, and the variable that best separates patients into two groups according to their survival curve is retained. Tree depth is controlled by a threshold on the minimum number of patients in the node. Random survival forest has

the advantage of possibly taking into account non-linear effects and interactions between variables [10,11].

We used the R package *tuneRanger* [9] to learn random survival forest for survival data, which is based on the *ranger* package [17] but with a fast implementation for tuning the number m of variables randomly drawn at each node. We used default hyperparameters suggested by the authors (*i.e.* patients used to build a tree chosen with boostrapping, at least 3 patients in a terminal node, 50 trees in a forest, log-rank test as splitting rule, $\sqrt{p}$ as starting value for tuning m, with p the total number of genes), and the function *tuneMtryFast*. To decrease computation time for mRNA datasets, we only retain the 2,500 genes with highest association with survival according to likelihood tests in single variable Cox models (*i.e.* we learn one single variable Cox model for each gene to compute the p-values).

2.2. Prediction performance metrics

As schemed Supplementary Fig. S2, we estimate the prediction performance of the models by 10 repetitions of a K-fold cross-validation ($K = 5$). We learn a model (*i.e.* Cox model or random survival forest) on a training dataset ($\frac{4}{5}$ of the patients), and we define a risk score from this estimation for each patient of the testing dataset ($\frac{1}{5}$ of the patients). The risk score (RS) is defined for a given patient i as the sum of $\beta_j X_j^i$ for the Cox model (X_j corresponds to the expression of gene j), and the mean of the estimated cumulative hazard function for the random survival forest:

- $\hat{\text{RS}}_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^i$ for the Cox model, with $\hat{\boldsymbol{\beta}}$ the estimator of the coefficients, and $\boldsymbol{X}^i$ the gene expression vector for patient i.
- $\hat{\text{RS}}_i = \frac{1}{\text{Card(T)}} \sum_{j \in \text{T}} \hat{H}(t_j | X_i)$ for random survival forest, with 'Card' the cardinal function, $\hat{H}(t|X_i)$ the estimated cumulative hazard function at time t for patient i, and T the times at which the hazard function is estimated.

This procedure allows to assess prediction performance by computing the C-index and the Integrated Brier Score (IBS), as defined below. Then, at the end of the procedure, 50 C-indices and 50 IBS are computed for each method.

The C-index allows the discrimination ability of a model to be assessed by quantifying the proportion of patient pairs for whom risk scores are in good agreement with their survival data. For two patients i and k with risk scores RS_i and RS_k, and with survival times T_i and T_k, the C-index is defined as $C = P(T_i < T_k \,|\, \text{RS}_i > \text{RS}_k)$. A C-index of 1 indicates perfect agreement, and a C-index of $\frac{1}{2}$ corresponds to random chance agreement. We took the estimator of the C-index given by [18] and theorized by [19].

The Brier Score [20] measures the average squared distance between the observed survival status and the predicted survival probability at a particular time t. It is always a number between 0 and 1, with 0 being the best possible value. We used the IBS that integrates the Brier Score between 0 and the maximum event time of the test set, and divides this quantity by the maximum integration time. Then, while the C-index measures the ability of a model to rank patients according to their risks, the IBS estimates the ability of a model to predict survival probabilities along time. The IBS is a global performance metric that assesses both discrimination and calibration. These two metrics are widely used to estimate prediction performance in practice and are complementary.

We used the R packages *survcomp* [21] to compute the C-index, and *pec* for the IBS [22].

2.3. The Cancer Genome Atlas and E-MTAB-1980 datasets

Cancer acronyms, as provided by the TCGA consortium, are available in Supplementary Table S1. First, we included cancers available in TCGA for which there were more than 75 patients with miRNA-seq and survival data. Then, we followed recent formal recommendations [23] to exclude the PCPG cancer that has too few death events and the SKCM cancer that has a high ratio of metastatic samples sequenced. We used overall survival as the disease-outcome, except when the authors recommend the use of progression-free

Table 1. Characteristics of the 11 cancers investigated. We computed the C-indices with 10 repetitions of 5-fold cross-validation for both the Cox-elastic net model (EN) and random survival forest (RF). Datasets are ordered according to their median C-index computed with Cox-elastic net model (decreasing order).

Cancer	n patients	p miRNA	Censoring rate	Survival - 3 years	C-index - EN	C-index - RF
UVM	77	536	0.73	0.74	0.81	0.83
ACC	77	518	0.65	0.75	0.8	0.84
KIRP	269	486	0.84	0.87	0.79	0.82
MESO	85	519	0.14	0.19	0.7	0.69
KIRC	508	462	0.66	0.75	0.7	0.66
LGG	506	548	0.62	0.56	0.7	0.69
CESC	288	542	0.76	0.72	0.68	0.59
LIHC	355	540	0.65	0.62	0.67	0.66
PRAD	486	470	0.81	0.8	0.66	0.59
LUAD	483	529	0.63	0.61	0.66	0.6
UCEC	532	554	0.83	0.83	0.61	0.64

interval (BRCA, LGG, PRAD, READ, TGCT, THCA and THYM). After these two steps, we retained 25 cancers. Finally, we computed the C-index and IBS estimates after running the Cox model or random survival forest applied on the miRNA profiles for these 25 cancers as schemed in Supplementary Fig. S2. To focus on cancers for which the sequencing data convey prognostic values, we decided to retain only the datasets for which the median C-index is significantly higher than 0.6 for at least one of the algorithms (*i.e.* Cox model or random survival forest) according to a one-sided Wilcoxon test at level 0.05. At the end of this procedure, we retained 11 cancers (Table 1, Supplementary Fig. S3).

We used the Broad GDAC FIREHOSE utility[1] to obtain clinical, miRNA-seq, and mRNA-seq datasets. We applied a Trimmed Mean of M-values (TMM) procedure to correct for between sample variance [24]. We first used the *calcNormFactors* function of package *EdgeR* [25] to compute a normalization factor for each patient, and we then applied the *voom* function of the *limma* package to compute log2-CPM data corrected with the normalization factors computed earlier [26]. We then standardized the expression of each gene both in the training dataset and the testing datasets using the mean and standard deviation values among patients of the training data.

To confirm the impact of sub-sampling on survival metrics in an external validation cohort, we acquired the processed E-MTAB-1980 dataset [27] from ArrayExpress[2]. Survival metrics were calculated as previously described, while using E-MTAB-1980 as a testing dataset. We standardized the testing dataset independently from the training dataset, by using the mean and standard deviation of the testing dataset.

2.4. Integration of miRNA-seq data together with clinical data

We verified whether the tumor miRNA profiles added predictive value to the clinical data [28]. Different strategies exist for integration of miRNA-seq and clinical data [29, 30]. In order to avoid the dilution of the few clinical parameters among all the miRNA covariates, we added the risk scores computed with miRNA-seq data alone (RS_{miRNA}) to ones computed using classical clinical features (age, gender, grade, T, N, M), when available ($RS = \beta.RS_{miRNA} + \sum_l \beta_l.Clin_l$, where $Clin_l$ is the l^{th} clinical variable). T is a score which stands for the extent of the tumor, N for the extent of spread to the lymph nodes, and M for the presence of metastasis. We did not include gender for sex-specific cancers (CESC,

[1] https://gdac.broadinstitute.org/
[2] https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1980/

UCEC, PRAD). Age is available for all cancers, and we specify whether the other variables are available in Supplementary Figures S4 and S5.

To emphasize if the miRNA-seq data added prediction value over clinical data for both Cox model and random survival forest, we performed a one-sided Wilcoxon signed rank test for each of the 11 cancers studied. We considered a difference significant when the p-value corrected with Benjamini-Hochberg method is below 0.05, even though this is purely indicative as discussed below.

2.5. Degradation of miRNA-seq data

2.5.1. Subsampling of miRNA-seq data

"Sequencing depth" is defined here as the sum of the number of aligned reads per patient, and can vary according to the patients, and is equivalent to the notion of "library size". These two nomenclatures will be used interchangeably in the following text.

To reduce the sequencing depth, we used a subsampling method [31]. The key parameter to calibrate fold reduction is the proportion of subsampling, $\varepsilon \in (0, 1]$. For each count data (*i.e.* number of reads) R_{ij} obtained for a patient i and a gene j, a subsampled count data of a proportion ε, noted $\tilde{R}_{ij}$, is drawn according to a binomial distribution of parameters R_{ij} and ε:

$$\tilde{R}_{ij} \sim \text{Binom}(R_{ij}, \varepsilon),$$

for each patient $i = 1, \ldots, n$ and each gene $j = 1, \ldots, p$.

Thus, the closer the parameter ε is to 0, the smaller the read depth: a proportion of ε (*e.g.* 0.01) corresponds to a subsampling of the sequencing data by a factor $\delta = \frac{1}{\varepsilon}$ (*e.g.* 100). In this study, we examine the effect of 1 (no subsampling), 10, 100, 1,000 and 10,000 subsampling factors δ. We then draw saturation curves, which are the evaluation metrics (*i.e.* C-index, IBS), performed on the test set which is not subsampled, as a function of the subsampling factor δ [32,33].

2.5.2. Reduction of the number of patients in the training dataset

To study the impact of the number of patients on prediction capabilities, we artificially decreased the percentage x of patients in the learning dataset. In this study, we chose $x = 10, 20, \ldots, 80\%$. However, to ensure that the C-index and IBS are not biased, the testing dataset is always composed of 20% of patients.

3. Results

3.1. Library sizes of mRNA-seq data are ten times larger than the ones of miRNA-seq data

The library sizes are equivalent between the 25 cancers, with a few exceptions, and are distributed around 5.10^6 reads for miRNAs, and 5.10^7 for mRNAs (Supplementary Fig. S6). Recall here that only aligned reads, and not raw reads, are taken into account. The sequencing depth for mRNA datasets is therefore higher than that of the miRNAs by a factor of 10 on average, which is not surprising as it spans on 40 times more genes. There are thus, on average, 4 times more aligned reads per gene for miRNAs than for mRNAs. The lengths of genes are also very different between mRNAs and miRNAs. Also, there is no particular relationship between the sequencing depth chosen for the mRNAs and for the miRNAs between the different cancers. Note that for the LAML cancer, we observe a lower sequencing depth than for the other cancers: the median sequencing depth is 720,000 reads for LAML, 2.5 million for KIRC and 7.5 million for LGG (Supplementary Fig. S6).

3.2. C-index highlighted noticeable prediction differences between Cox and random survival forest models for eight out of twenty-five cancers

Using miRNA-seq datasets and according to the C-index metric, the Cox model shows better prediction than the random survival forest for KIRC, CESC, PRAD, LUAD, HNSC, and to a lesser extent LIHC (Supplementary Fig. S3A). Conversely, random survival forest shows better predictions for KIRP, THYM, THCA, and to a lesser extent UCEC. For the

other cancers, we did not observe clear differences. Noticeably, while the Cox model is not able to capture any prediction abilities for THCA (*i.e.* median C-index of 0.46), random survival forest exhibits a median C-index of 0.62. However, if we choose the IBS as the prediction metric, random survival forest shows better prediction than Cox except for LGG (Supplementary Fig. S3B). We discuss this difference observed between the metrics below.

3.3. mRNA-seq data provides slightly better prediction performance than miRNA-seq data for most of the 11 investigated cancers

In this section, we use a Wilcoxon signed-rank test to highlight the situation in which there exist differences, and we corrected the 11 p-values computed with the Benjamni-Hochberg procedure.

The median C-indices reached with the Cox model is higher with mRNA-seq data than miRNA-seq data for all selected cancers except LUAD. More precisely, the C-indices are higher with mRNA-seq data for 8 cancers (ACC, KIRP, MESO, KIRC, LGG, CESC, PRAD, UCEC), and higher with miRNA-seq data only for LUAD (Supplementary Fig. S7A). When using IBS as metric, median IBS obtained from 5 cancers (UVM, KIRC,LIHC,LUAD,UCEC) vas lower compared to median IBS obtained with mRNA-seq data. Additionally, overall IBS obtained by using mRNA-seq data was lower in cases of KIRP, MESO, LGG, CESC, and higher while using miRNA-seq data of UVM, KIRC and LIHC. Comparable results were obtained using random survival forest as prediction model. Overall, mRNA-seq data provides better predictions, but the absolute differences remain small.

3.4. miRNA-seq data improves predictions over clinical data alone for most of the investigated cancers

For 8 cancers (UVM, ACC, MESO, LGG, CESC, LIHC, PRAD, LUAD) out of the 11 studied, the addition of miRNA-seq data to generic clinical data significantly improved the C-index compared to clinical data alone for the Cox model (Supplementary Fig. S4A, integration of clinical and miRNA data is described section 2.4). Similarly, for random survival forest, the median C-index is improved for 6 cancers (UVM, ACC, MESO, LGG, LIHC and PRAD, Supplementary Fig. S5A). When using the IBS as the performance metric, the difference are often not as clear: predictions appear better for 5 cancers when taking tumor miRNA profiles into account in the Cox model (KIRP, MESO, KIRC, LGG, and CESC, Supplementary Fig. S4B), and for 5 cancers in the random survival forest model (ACC, KIRP, MESO, LGG, and LIHC, Supplementary Fig. S5B).

Overall, the addition of miRNA-seq data to classical clinical data improves prediction performance as assessed by C-index and/or IBS for all the 11 cancers investigated but UCEC with the Cox model. This performance drops to 7 cancers with random survival forest; KIRC, CESC, LUAD, UCEC do no show improvement. The use of miRNA-seq data seems not as interesting as clinical data alone to build predictive risk scores for UCEC. However, we have included this cancer because RNA-seq data can be used in other contexts as part of a survival model: stratifying patients according to transcriptomic profiles [34], identifying predictive markers of response to treatments [35], identifying potential therapeutic targets [36].

3.5. Shallow tumor miRNA or mRNA sequencing keeps survival prediction performance for many cancers

We consider that the sequencing depth can be reduced if and only if none of the two prediction metrics (*i.e.* C-index or IBS) is degraded at level 0.05 according to a one sided Wilcoxon test. Fig. 1 shows the C-index as a function of miRNA (or mRNA) library size reduction and/or number of patients subsampling for the kidney cancer subtype KIRC (corresponding to clear cell renal cell carcinoma, ccRCC). We highlight this cancer subtype for the following reasons: the dataset contains many patient data (table 1), the prediction performance is quite high (C = 0.7), the availability of an independent dataset (section 3.8), and as we are more specialized on this subtype [37]. For this cancer, it appears that both the number of patients and the sequencing depth could have been decreased while

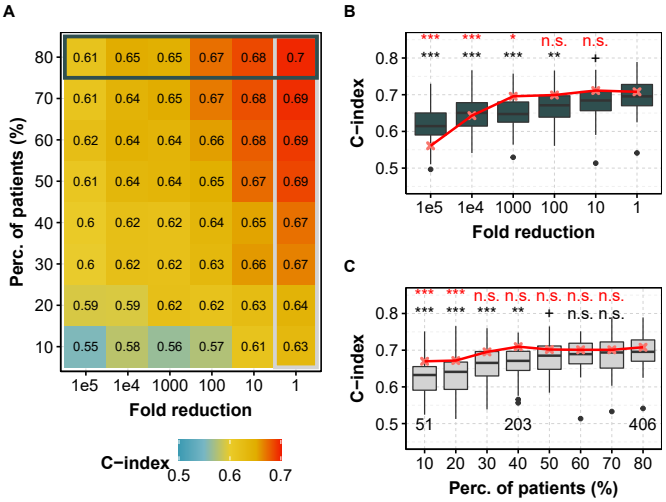

Figure 1. C-index obtained for different fold reduction factors and percentage of patients in the training dataset for KIRC (ccRCC, TCGA) with the Cox model. (A) Median C-index for different degradation of both sequencing depth (x axis) and percentage of patients (y axis) in the training dataset for miRNA-seq data. Horizontal box highlights the case where all of the 80% of patients are used and corresponds to (B), whereas vertical box focuses on the full available library size and corresponds to (C). **(B)** C-index for different fold reduction factors for miRNA-seq (gray boxplots) and mRNA-seq data (median values, in red) with 80% of the patients in the training dataset. Above is the p-value of a one-sided Wilcoxon test compared to no subsampling (*i.e.* $\delta = 1$). **(C)** C-index for different percentage of patients in the training dataset for miRNA-seq (light gray boxplots) and mRNA-seq data (median values, in red) with original TCGA sequencing depth. Above is the p-value compared to full dataset (*i.e.* 80%). red, mRNA-seq; gray boxplots, miRNA-seq. In each case, we computed the C-indices by 10 repetitions of a 5-fold cross validation. ***: $p \leq 0.001$, **: $p \leq 0.01$, *: $p \leq 0.05$, +: $p < 0.1$, n.s. : $p \geq 0.1$.

keeping similar prediction capacity. More precisely, using 60% of the patients ($n = 304$) in the training set leads to similar model performance, even though a small but noticeable performance decrease is noticed with IBS in the Cox model (Supplementary Fig. S9). Also, decreasing the miRNA and mRNA sequencing depth by one order of magnitude has no measurable consequences. However, both the number of patients and the sequencing depth should not be decreased to their maximum extents altogether, as shown by the shape of the color map (Fig. 1A).

Tables 2 and 3 summarize the maximum lowering of sequencing depth without affecting prediction performance. For most cancers, reducing the sequencing depth for miRNAs and mRNAs leads to similar performance, and the possible library size reduction is correlated between miRNAs and mRNAs when chosen as covariates in the Cox model (Supplementary Fig. S10). For mRNAs, one order of magnitude reduction or more is permitted for 11 cancers but CESC, which only tolerates a 50% reduction ($\sim 20,000,000$ aligned reads). For miRNAs, there are 3 exceptions: CESC again, which also tolerates a 50% reduction ($\sim 2,000,000$ aligned reads), PRAD with an 80% reduction ($\sim 1,000,000$ aligned reads), and LUAD which do not tolerate any sequencing depth reduction, and may even show improved performance with an increase in library size ($\geq 5,000,000$ aligned reads). For cancers tolerating $500,000$ aligned reads in mRNAs or less, the sequencing depth could be reduced at least for one order of magnitude in miRNAs. Thus, mRNA sequencing data might inform of the required sequencing depth for miRNAs.

For random survival forest and mRNA-seq data, the sequencing depth of all cancers can be reduced by a factor 100 without degrading the C-index and the IBS (Supplementary Tab. S2, and Fig. S11). This fold reduction corresponds to median sequencing depth of about $500,000$ aligned reads. For KIRC and LGG the sequencing depth can be even more

Table 2. Maximum miRNA-seq library size reduction before the decreasing of prediction performance, corresponding median sequencing depth (in thousands of aligned reads), and prediction metric degraded first, for the Cox model, and the 11 investigated cancers.

Cancer	UVM	ACC	KIRP	MESO	KIRC	LGG	CESC	LIHC	PRAD	LUAD	UCEC
Fold reduction	1000	1000	100	100	10	10	2	10	5	< 1	10000
Median library size (in 1000 reads)	5	6	60	50	200	700	2000	500	900	> 5000	1
Metric degraded first	C-index	both	both	C-index	both	IBS	C-index	both	C-index	both	C-index

Table 3. Maximum mRNA-seq library size reduction before the decreasing of prediction performance, corresponding median sequencing depth (in thousands of aligned reads), and prediction metric degraded first, for the Cox model, and the 11 investigated cancers.

Cancer	UVM	ACC	KIRP	MESO	KIRC	LGG	CESC	LIHC	PRAD	LUAD	UCEC
Fold reduction	100	1000	100	100	10	100	2	10	10	10	10
Median library size (in 1000 reads)	400	40	400	500	5000	500	20000	5000	5000	4000	2000
Metric degraded first	C-index	both	IBS	both	IBS	both	C-index	IBS	both	IBS	both

reduced, by a factor $1,000$ ($\sim 50,000$ aligned reads). The results are more heterogeneous for miRNA-seq data as the sequencing depth can be reduced by a factor $1,000$ for CESC ($\sim 5,000$ aligned reads) down to 5 for KIRP and MESO ($\sim 1,000,000$ aligned reads). Noticeably, for CESC, the possible fold reduction is much larger for random survival forest than for the Cox model for both miRNAs and mRNAs. We hypothesized that these differences are the consequence of a better C-index obtained with the Cox model than with random survival forest (supplementary Fig. S3A, S10, S11).

3.6. Models trained with fewer patients do not degrade prognosis for most of the investigated cancers

For the Cox model and miRNA-seq data, the number of patients in the training dataset can be reduced for 9 of the 11 cancers, at least for a small proportion (Supplementary Tab. S3). The two exceptions concern PRAD and LUAD, for which diminishing the number of patients in the training set decreases the prediction performance. We obtained comparable results for mRNA-seq data, except for UVM and UCEC which require more patients to achieve maximum performance - for UVM with similar C-index between miRNAs and mRNAs, but for UCEC with better performance with miRNAs (Supplementary Fig. S7). Surprisingly, random survival forest need less patients in the training dataset to achieve optimal prediction performance. This result makes it possible to consider stratifying patients into subgroups and to learn models separately for each subgroup.

3.7. Very small sequencing depth is responsible for the performance loss

When reducing sequencing depth, we automatically reduce the number of detected genes. We define a non-coding / coding gene as 'detected' if its CPM-normalized expression level is greater than 1 for at least 1% of the patients in the training dataset. Two hypotheses can be put forward to explain the decrease in predictive abilities induced by the subsampling of sequencing depth:

1. the number of detected genes decreases as the subsampling rate increases (Supplementary Fig. S12A, [38]), and only the level of expression of the most highly expressed genes can be measured (Supplementary Fig. S12B). However, genes with a low level of expression may have significant predictive power and go undetected, which would diminish overall predictive capabilities.
2. more generally, the signal-to-noise ratio decreases for all genes (the standard deviation of the measurements varies in $\sqrt{N}$, with N the number of aligned reads per gene).

To test the first hypothesis, we compared the C-indices obtained with all miRNAs, and those obtained only with the most expressed genes, using the Cox model. This corresponds on average to a 2-fold decrease in the number of predictors (*i.e.* 210 miRNAs on average detected after subsampling of the miRNA-seq data by a factor of 10,000 for all cancers studied). In these two scenarios, the sequencing data keep the same read depth per gene, but only the number of genes taken into account differs. For miRNAs, reducing the dimension by keeping only the 210 most expressed genes does not significantly impact prediction capabilities (*i.e.* C-index and IBS) for the 11 cancers studied, except for CESC and LUAD with the C-index (p-value < 0.001, one-sided Wilcoxon signed-rank test with Benjamini-Hochberg correction), and to a lesser extent for LIHC with the C-index (p-value < 0.05, Supplementary Fig. S13). The first hypothesis is therefore not verified.

To test the second hypothesis, we calculated the C-indices obtained after subsampling by a factor of 10,000 (*i.e.* 210 genes on average are detected, and the count data are subsampled), and those obtained with the same 210 genes (on average) but without subsampling, also with the Cox model. For all 11 cancers, the median C-index obtained after subsampling is lower than that obtained without subsampling but with the same predictors (Supplementary Fig. S13). For the IBS, we observed the same results, except for UVM, LUAD and UCEC. The second hypothesis is verified: the subsampling induces a decrease in the signal-to-noise ratio of the RNA-seq count data explaining the decrease of the prediction if a strong subsampling is applied. It is also interesting to notice that for most cancers, selecting half of the most expressed predictors does not affect prediction performance.

We obtained comparable results for random survival forest (Supplementary Fig. S14) and for mRNA-seq data (data not shown).

3.8. Prognostic performances follow similar trend after subsampling when tested on an independent dataset

To evaluate to which extent these results can be used in practice, we checked whether the models learned on TCGA, with an incremental decrease in sequencing depth, also do not degrade the predictions on an independent dataset. However, we were unable to find a dataset comparable to TCGA with both microRNA tumor profiling and patient survival, so we chose to demonstrate the reproducibility on mRNA profiling. We chose an mRNA profiling dataset measured with microarray in order to improve generalizability, and keeping ccRCC for the reasons detailed section 3.5. Figure 2 shows that first the C-index is higher when the test set is the independent one, and second that the trend is comparable for both test sets (E-MTAB-1980 and TCGA datasets). More precisely, reducing the mRNA sequencing depth by a factor of 10 or even 100 on the training set does not affect the C-index performance. The E-MTAB-1980 dataset also provide improved IBS performance and smaller variability. We hypothesize that the smaller variability is due to the fact that the E-MTAB-1980 test set remains the same, as compared to the TCGA test set gathering the 20% of remaining patients, not used in the training set, and so the difference in performance is more sensitive (Supplementary Fig. S15 and S16). Halving the number n of patients in the training set (from 80% to 40%, *i.e.* from $n = 406$ to 203) increases the IBS measured on the E-MTAB-1980 dataset in a modest manner (from a median of 0.112, 95% confidence interval of the median [0.11;0.114] to 0.133 [0.125;0.14]), whereas only 10% ($n = 51$) of the patients lead to low IBS performance (0.205 [0.195;0.214]). This information is useful when trying to improve the model by clustering the patients into refined cancer subtypes, and further applying different Cox model in each of these subtypes: while 2 subtypes are reasonable for ccRCC on this large TCGA cohort, more than 5 may lead to too large IBS values only because of the small number of patients in each learning subset.

4. Discussion

Our work shows the benefit of using tumor mRNA or miRNA profiling to predict patient survival. We thus estimated the optimal sequencing depth and number of patients

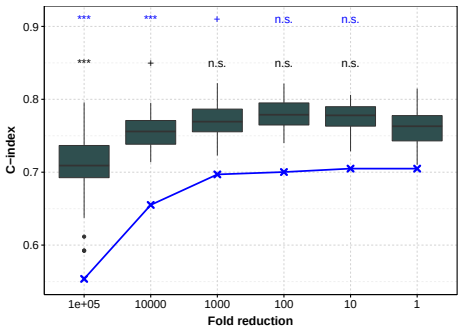

Figure 2. C-index as a function of sequencing depth reduction tested on the E-MTAB-1980 dataset and TCGA subset for mRNA profiling in ccRCC. In dark gray, performance measured with the C-index calculated on the E-MTAB-1980 dataset, after training on an 80% sub-sample of the TCGA dataset (the procedure is repeated to obtain 50 C-indices). In blue, the test is performed on the remaining 20% of TCGA data (median C-index). ***: $p \leq 0.001$, **: $p \leq 0.01$, *: $p \leq 0.05$, n.s. : $p \geq 0.1$.

to achieve this prediction, using Cox and random forest models. We considered that the sequencing depth can be reduced if both C-index and IBS are not significantly degraded. This choice is subjective and can easily be adapted with the R script provided. For example, if discrimination among patients at risk is the only important aspect for a particular study, the C-index should be considered as the only prediction metrics. Second, as the same initial set of patients is used in multiple Monte-Carlo runs (10 repetitions of 5-fold cross-validation) to estimate prediction performance, the 50 metrics (*i.e.* C-index or IBS) are not independent. Besides, the p-values can be reduced toward 0 when small differences are observed by increasing the number of repetitions. The computed p-values are thus only indicative, but helps the readability of the graphics.

Thus, because of the methodology used, we obtain comparable results with slight differences with [13] about shallow tumor mRNA sequencing to predict patient survival with Cox model. We have extended it to miRNA-sequencing, and to random forest survival.

Then, estimation of the IBS with random survival forest is direct in the sense that the survival function S is an output of the algorithm for patients of the testing dataset. However, it is not the case with the Cox model as the baseline hazard function h_0 is not estimated in the pseudo-likelihood. This function h_0 is estimated with the Breslow estimator [14]. This dissimilarity in the way individual survival functions are estimated may explain the large advantage on prediction performance for random survival forest as compared to the Cox model (Supplementary Fig. S3B). More investigations are still needed to better understand this observation.

This work is the first one to focus on the sequencing depth of miRNA profiling for survival prediction, and could serve as a proxy to calibrate future experiments. Lung adenocarcinoma (LUAD) showed a noticeable difference with other cancers as tumor miRNAs better predict patient survival than mRNAs. This may indicate the particular role of miRNAs in this tumor type, and would be worth to further investigate.

Finally, the number of patients required in the training dataset is lower for random survival forest than for the Cox model (section 3.6). This result is surprising as random forest has more degrees of freedom, and further work is needed to investigate this point.

5. Conclusion

In this work, we present a methodology and results on the possibility of (i) reducing sequencing costs to, for example, create validation datasets, and (ii) reducing the number of samples in training datasets to stratify patients into subgroups in the context of prediction of survival in cancer. Cox model and random survival forest provide comparable C-indices for some cancers but not all (*e.g.* the Cox model outperforms random survival forest for CESC and miRNA-seq data; the contrary being true for THCA). However, with IBS as

the performance metric, the performance are better with random survival forest. We also pointed out that mRNA-seq data provide slightly better performance than miRNA-seq data on average, with the noticeable exception of lung adenocarcinoma (LUAD). Integration of miRNA-seq data with clinical data allows to improve predictions over clinical data alone for most of the 11 investigated cancers. Importantly, we demonstrated that sequencing depth of miRNA-seq and mRNA-seq can be reduced without degrading prediction performance for most of the 11 cancers retained in a cancer, data (*i.e.* miRNAs or mRNAs) and metric (*i.e.* C-index or IBS) dependent manner, thus allowing the reduction of sequencing cost to create independent validation datasets. Finally, we demonstrated that the number of patients in the training dataset can be reduced for both miRNA and mRNA data without degrading the prediction performance for the Cox model, and in a larger extent for random survival forest. Finally, our results were confirmed on an independent dataset.

Funding

This article was developed in the framework of the Grenoble Alpes Data Institute, supported by the French National Research Agency under the *Investissements d'avenir* programme (ANR-15-IDEX-02).

1. Bartel, D.P. Metazoan micrornas. *Cell* **2018**, *173*, 20–51.
2. Peng, Y.; Croce, C.M. The role of MicroRNAs in human cancer. *Signal transduction and targeted therapy* **2016**, *1*, 1–9.
3. Chu, A.; et al. Large-scale profiling of microRNAs for the cancer genome atlas **2016**. *44*, e3–e3.
4. Capula, M.; et al. New avenues in pancreatic cancer: exploiting microRNAs as predictive biomarkers and new approaches to target aberrant metabolism. *Expert Review of Clinical Pharmacology* **2019**, *12*, 1081–1090.
5. Cox, D.R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **1972**, *34*, 187–202.
6. Jardillier, R.; et al. Bioinformatics Methods to Select Prognostic Biomarker Genes from Large Scale Datasets: A Review. *Biotechnology Journal* **2018**, *13*, 1800103. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/biot.201800103, https://doi.org/https://doi.org/10.1002/biot.201800103.
7. Zou, H.; Hastie, T. Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society* **2005**, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
8. Jardillier, R.; Koca, D.; Chatelain, F.; Guyon, L. Prognosis of lasso-like penalized Cox models with tumor profiling improves prediction over clinical data alone and benefits from bi-dimensional pre-screening. *BMC cancer* **2022**, *22*, 1–16.
9. Probst, P.; et al. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2019**, *9*, e1301.
10. Ishwaran, H.; et al. Random survival forests. *Annals of Applied Statistics* **2008**, *2*, 841–860. Publisher: Institute of Mathematical Statistics, https://doi.org/10.1214/08-AOAS169.
11. Wright, M.N.; et al. Do little interactions get lost in dark random forests? *BMC Bioinformatics* **2016**, *17*, 145. https://doi.org/10.1186/s12859-016-0995-8.
12. Kourou, K.; et al. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* **2015**, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005.
13. Milanez-Almeida, P.; et al. Cancer prognosis with shallow tumor RNA sequencing. *Nature Medicine* **2020**, *26*, 188–192.
14. Breslow, N. Contribution to the Discussion of the Paper by D.R. Cox. *Journal of the Royal Statistical Society B* **1972**, *34*, 2016–2017.
15. Friedman, J.; et al. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **2010**, *33*, 1–22.
16. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
17. Wright, M.N.; Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **2017**, *77*, 1–17. Number: 1, https://doi.org/10.18637/jss.v077.i01.

18. Harrell Jr, F.E.; Lee, K.L.; Mark, D.B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* **1996**, *15*, 361–387.
19. Pencina, M.J.; D'Agostino, R.B. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine* **2004**, *23*, 2109–2123, [https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1802]. https://doi.org/10.1002/sim.1802.
20. Gerds, T.A.; Schumacher, M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom J* **2006**, *48*, 1029–1040.
21. Schroeder, M.; et al. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics* **2011**, *27(22)*, 3206–3208.
22. Mogensen, U.B.; et al. Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *Journal of Statistical Software* **2012**, *50*, 1–23.
23. Liu, J.; et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **2018**, *173*, 400–416.e11. https://doi.org/10.1016/j.cell.2018.02.052.
24. Robinson, M.D.; Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **2010**, *11*, 1–9.
25. Robinson, M.D.; et al. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **2010**, *26*, 139–140. https://doi.org/10.1093/bioinformatics/btp616.
26. Ritchie, M.E.; et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **2015**, *43*, e47. https://doi.org/10.1093/nar/gkv007.
27. Sato, Y.; Yoshizato, T.; Shiraishi, Y.; Maekawa, S.; Okuno, Y.; Kamura, T.; Shimamura, T.; Sato-Otsubo, A.; Nagae, G.; Suzuki, H.; et al. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nature genetics* **2013**, *45*, 860–867.
28. Volkmann, A.; et al. A plea for taking all available clinical information into account when assessing the predictive value of omics data. *BMC Medical Research Methodology* **2019**, *19*, 162. https://doi.org/10.1186/s12874-019-0802-0.
29. López de Maturana, E.; et al. Challenges in the Integration of Omics and Non-Omics Data. *Genes* **2019**, *10*. https://doi.org/10.3390/genes10030238.
30. De Bin, R.; et al. Combining clinical and molecular data in regression prediction models: insights from a simulation study. *Briefings in Bioinformatics* **2019**, [https://academic.oup.com/bib/advance-article-pdf/doi/10.1093/bib/bbz136/31080858/bbz136.pdf]. bbz136, https://doi.org/10.1093/bib/bbz136.
31. Robinson, D.G.; Storey, J.D. subSeq: Determining Appropriate Sequencing Depth Through Efficient Read Subsampling. *Bioinformatics* **2014**, *30*, 3424–3426. https://doi.org/10.1093/bioinformatics/btu552.
32. Tarazona, S.; et al. Differential expression in RNA-seq: A matter of depth. *Genome Research* **2011**, *21*, 2213–2223. https://doi.org/10.1101/gr.124321.111.
33. Bass, A.J.; et al. Determining sufficient sequencing depth in RNA-Seq differential expression studies. *bioRxiv* **2019**. https://doi.org/10.1101/635623.
34. Ricketts, C.J.; et al. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Reports* **2018**, *23*, 313–326.e5. Publisher: Elsevier, https://doi.org/10.1016/j.celrep.2018.03.075.
35. Ternès, N.; et al. Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. *Biometrical Journal. Biometrische Zeitschrift* **2017**, *59*, 685–701. https://doi.org/10.1002/bimj.201500234.
36. Wei, H.; et al. MiR-638 inhibits cervical cancer metastasis through Wnt/beta-catenin signaling pathway and correlates with prognosis of cervical cancer patients. *European Review for Medical and Pharmacological Sciences* **2017**, *21*, 5587–5593. https://doi.org/10.26355/eurrev_201712_13999.
37. Roelants, C.; Pillet, C.; Franquet, Q.; Sarrazin, C.; Peilleron, N.; Giacosa, S.; Guyon, L.; Fontanell, A.; Fiard, G.; Long, J.A.; et al. Ex-vivo treatment of tumor tissue slices as a predictive preclinical method to evaluate targeted therapies for patients with renal carcinoma. *Cancers* **2020**, *12*, 232.
38. Sims, D.; et al. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* **2014**, *15*, 121–132. https://doi.org/10.1038/nrg3642.