frontiers in Genetics

ORIGINAL RESEARCH
published: 25 November 2020
doi: 10.3389/fgene.2020.550894

Identification of a Transcriptomic Prognostic Signature by Machine Learning Using a Combination of Small Cohorts of Prostate Cancer

Benjamin Vittrant[1,2], *Mickael Leclercq*[1,2]*, *Marie-Laure Martin-Magniette*[3,4], *Colin Collins*[5,6], *Alain Bergeron*[1,7], *Yves Fradet*[1,7]* and *Arnaud Droit*[1,2]*

[1] Centre de Recherche du CHU de Québec – Université Laval, Québec, QC, Canada, [2] Département de Médecine Moléculaire, Université Laval, QC, Canada, [3] Universities of Paris Saclay, Paris, Evry, CNRS, INRAE, Institute of Plant Sciences Paris Saclay (IPS2) 91192, Gif-sur-Yvette, France, [4] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, Paris, France, [5] Vancouver Prostate Cancer Centre, Vancouver, BC, Canada, [6] Department of Urologic Sciences, The University of British Columbia, Vancouver, BC, Canada, [7] Département de Chirurgie, Oncology Axis, Université Laval, Québec, QC, Canada

Determining which treatment to provide to men with prostate cancer (PCa) is a major challenge for clinicians. Currently, the clinical risk-stratification for PCa is based on clinico-pathological variables such as Gleason grade, stage and prostate specific antigen (PSA) levels. But transcriptomic data have the potential to enable the development of more precise approaches to predict evolution of the disease. However, high quality RNA sequencing (RNA-seq) datasets along with clinical data with long follow-up allowing discovery of biochemical recurrence (BCR) biomarkers are small and rare. In this study, we propose a machine learning approach that is robust to batch effect and enables the discovery of highly predictive signatures despite using small datasets. Gene expression data were extracted from three RNA-Seq datasets cumulating a total of 171 PCa patients. Data were re-analyzed using a unique pipeline to ensure uniformity. Using a machine learning approach, a total of 14 classifiers were tested with various parameters to identify the best model and gene signature to predict BCR. Using a random forest model, we have identified a signature composed of only three genes (JUN, HES4, PPDPF) predicting BCR with better accuracy [74.2%, balanced error rate (BER) = 27%] than the clinico-pathological variables (69.2%, BER = 32%) currently in use to predict PCa evolution. This score is in the range of the studies that predicted BCR in single-cohort with a higher number of patients. We showed that it is possible to merge and analyze different small and heterogeneous datasets altogether to obtain a better signature than if they were analyzed individually, thus reducing the need for very large cohorts. This study demonstrates the feasibility to regroup different small datasets in one larger to identify a predictive genomic signature that would benefit PCa patients.

Keywords: machine learning, prostate cancer, RNA-seq, biochemical recurrence, random forest, predictive signature

Abbreviations: ACC, accuracy; BER, balanced error rate; BCR, biochemical recurrence; AUC, area under the curve; MCC, matthews correlation coefficient; MMCE, mean misclassification error rate; PCa, prostate cancer; PSA, prostate specific antigen; TNM, tumor node metastasis.

OPEN ACCESS

Edited by:
Changning Liu,
Xishuangbanna Tropical Botanical Garden (CAS), China

Reviewed by:
Guosen Zhang,
Henan University, China
Zhitong Bing,
Lanzhou University, China
Xuan Zhang,
Xishuangbanna Tropical Botanical Garden (CAS), China

****Correspondence:***
Arnaud Droit
arnaud.Droit@crchudequebec.ulaval.ca
Yves Fradet
yves.fradet@chudequebec.ca
Mickael Leclercq
Mickael.Leclercq@gmail.com

Specialty section:
This article was submitted to Genomic Medicine, a section of the journal Frontiers in Genetics

Received: *05 June 2020*
Accepted: *29 October 2020*
Published: *25 November 2020*

Citation:
Vittrant B, Leclercq M, Martin-Magniette M-L, Collins C, Bergeron A, Fradet Y and Droit A (2020) Identification of a Transcriptomic Prognostic Signature by Machine Learning Using a Combination of Small Cohorts of Prostate Cancer. Front. Genet. 11:550894. doi: 10.3389/fgene.2020.550894

INTRODUCTION

Prostate Cancer (PCa) is the most common non-cutaneous cancer in American men. Around 160 000 men were diagnosed with PCa in 2017 (Siegel et al., 2017) and around 27 000 died of it. The burden of this disease on public health is important and expected to grow as a recent study revealed that the incidence of advanced PCa increased in the last few years (Weiner et al., 2016). PCa is a complex and heterogeneous disease (D'Amico et al., 2003; Buyyounouski et al., 2012) since the risk of relapse and death after treatment differs among cancers with the same clinico-pathological features, namely the grade (Gleason score), stage [Tumor, Node, Metastasis (TNM)] (Edge and Compton, 2010; Amin et al., 2018) and the level of prostatic specific antigen (PSA) (Papsidero et al., 1980).

Current treatments for localized PCa mainly include surgical removal or external beam radiation therapy of the prostate. If the initial treatments did not succeed to cure the patient then a recurrence will occur, revealed by an increase in seric PSA level, an event called biochemical recurrence (BCR). After surgery, about 70% of the patients will be cured and about 30% will relapse to a BCR. Since prostate tumor cells depend on androgens to grow, recurrences are treated with androgen deprivation therapy consisting in chemical or surgical castration either alone or in association with administration of anti-androgens. However, the cancer will inevitably recur and will then be called castration-resistant prostate cancer (CRPC). To treat CRPC, docetaxel (Tannock et al., 2004) was introduced in 2004, but more recently, second generation of androgen-deprivation therapies resulted in better survival (Tannock et al., 2004; Nevedomskaya et al., 2018). Ultimately all these tumors will relapse and patients will be offered palliative therapy. Consequently, in order to offer better treatments to these patients, there is a pressing need to identify earlier those tumors that will recur after surgery and evolve to become lethal.

One problem generally inherent to cancer care is to orient people to the adequate treatment corresponding to the stage of the disease and the individual characteristics of the patient (Terada et al., 2017). In PCa, the stage, grade and PSA level are currently the best standards to drive patients in the different treatment options. Currently, after radical prostatectomy the PSA level is actively monitored to assess the BCR, but there is no biomarker that is used clinically to predict a future BCR.

To reduce costs and continue to improve prognostic, omics data are promising. With the decreasing price of RNA sequencing (RNA-seq), the accessibility of affordable technologies [e.g., MinION from Oxford Nanopore Technologies (Menegon et al., 2017)], the available PCa cohorts and the efficient computational approaches, transcriptomics is becoming a valuable resource to identify biomarkers (Nikitina et al., 2017). The rapid development of omics technology has led to the availability of many omics databases (Marx, 2013; Almeida et al., 2014; Stephens et al., 2015), including The Cancer Genome Atlas Program (TCGA) (Tomczak et al., 2015) and those of the International Cancer Genome Consortium (ICGC) (International Cancer Genome Consortium Hudson et al., 2010), thus opening an opportunity to apply and test machine learning algorithms (Li et al., 2016). These algorithms have been utilized as an aim to model the progression and treatment of cancerous conditions, and resulted in effective and accurate decision-making (Kourou et al., 2015). However, many of the datasets results from patients cohorts that were either rather small and/or had insufficient follow-up of clinical history which limit their use for clinical outcome prediction.

Hence, there is a challenge to set up predictive models that could anticipate the event of BCR, thus predicting the evolution of cancer, immediately after surgery. Consequently, we propose here a method to discover a transcriptomic signature that could be used to predict BCR events using a combination of datasets to increase the discovery potential. To this purpose, we applied specific preprocessing and cleaning steps on three RNA-seq datasets and established a machine learning protocol.

MATERIALS AND METHODS

Research Pipeline

After recovering the raw data from the different studies, we processed them in a pipeline composed of three main steps: Samples quality control and selection, sequencing data processing, machine learning analysis (**Figure 1**). All developed scripts are available in the github repository (See section "Data Availability Statement").

Datasets

We retrieved three different RNA-Seq datasets of radical prostatectomy specimens with the associated clinical features. The first dataset is from TCGA cohort in the Prostate Adenocarcinoma (PRAD) project. The second dataset (GSE54460) is from a cohort constituted by Long et al. (2014) and the third dataset was provided by Prof C. Collins from the Vancouver Prostate Cancer Center (VPCC) (Wyatt et al., 2014).

Quality of the BCR event data is dependent on patient clinical follow-up. A patient followed only a few weeks or months after surgery without showing BCR would be considered as a non-BCR case. These cases are a bias since the patient could have experienced a BCR event after the period of follow-up. Consequently, we discarded from our analysis the patients with no BCR whose follow-up was inferior to 60 months.

TCGA-PRAD Dataset

Data from 498 samples were initially recovered from the PRAD project on the TCGA data portal[1]. According to the *TCGA Research Network* (Cancer Genome Atlas Research Network, 2015) 131 samples must be discarded because of the presence of RNA degradation, as we did. We also ignored samples with less than 40% of tumor cells (column *percent_tumor_cells* in clinical file) and follow-up inferior to 60 months. We ended up with 52 samples after these filters.

GSE54460 Dataset

The data were downloaded from NCBI website (GEO accession GSE54460) where sequencing and clinical data from 106 patients

[1] https://portal.gdc.cancer.gov/

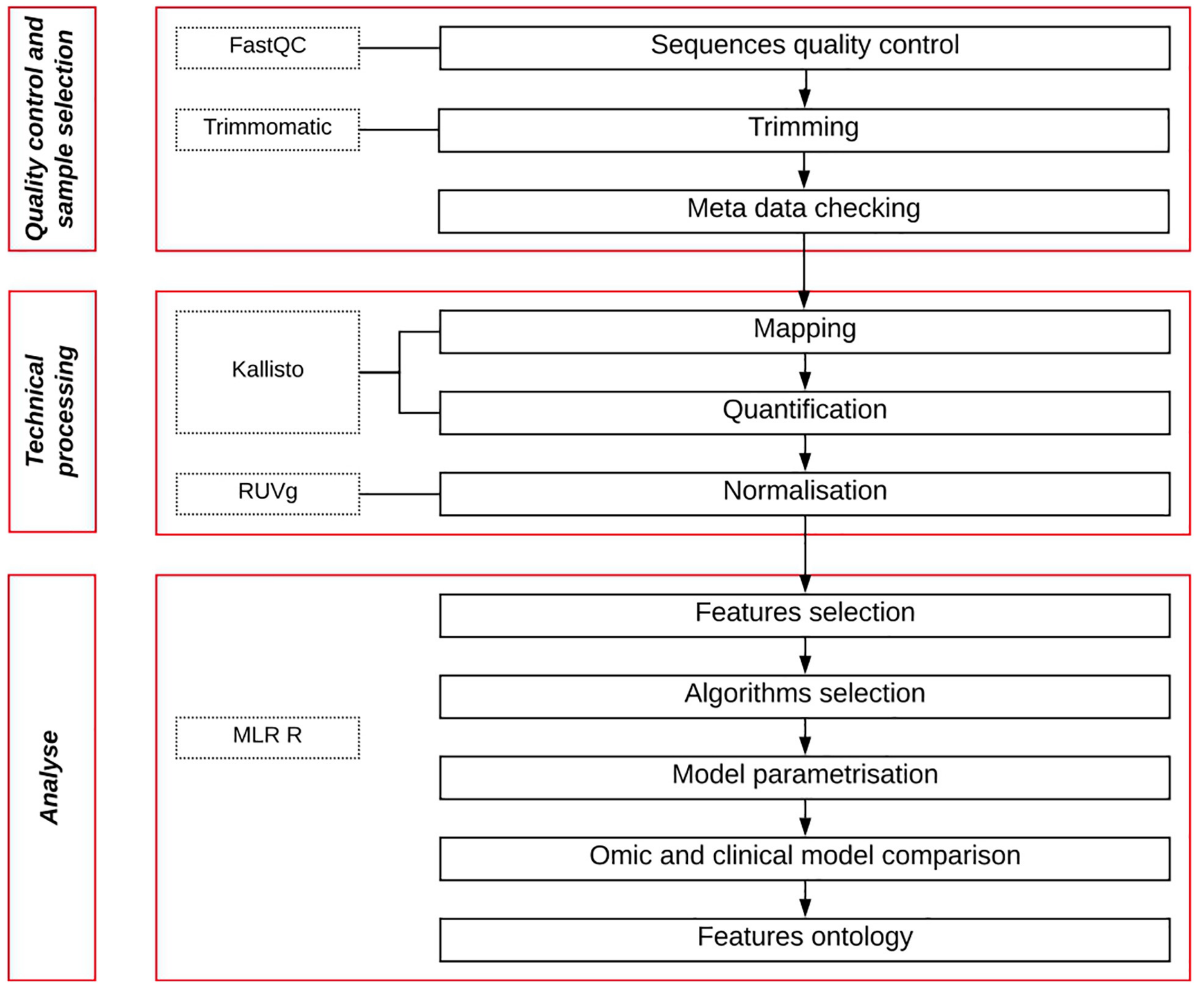

FIGURE 1 | Pipeline workflow. Quality control of raw data sequencing files is measured, then trimmed to remove their adaptors. Patient metadata are then filtered to keep only BCR patients with long follow-up. Retained sequences are then mapped, quantified and normalized. Finally, a machine learning approach is used to analyze the data to obtain a gene expression predictive signature and a model.

were recovered. After selecting cases with a minimum of 60 months of follow-up, we ended-up with 96 patients of whom 54 had a BCR.

VPCC Dataset

We obtained the raw fastq files and clinical data from 85 patients, available at European Nucleotide Archive of the EMBL-EBI under accession PRJEB6530. Patients treated with hormonal therapy before radical prostatectomy were removed because this treatment strongly alters RNA expression. After selecting patients for minimal follow-up we ended up with 23 patients of whom five experienced a BCR.

The baseline characteristics of the resulting individual and combined cohorts after selection of eligible cases are summarized in **Table 1**.

Quality Control, Alignment and Gene Expression

The quality of the raw fastq files from the TCGA cohort was measured using *FastQC* (Andrews et al., 2010) (v0.11.5) and *Trimmomatic* (Bolger et al., 2014) (v0.32). A threshold quality per base of 30 (based on Phred 33) and a minimal length of 40 bases were applied. The transcriptomes were then mapped on GrCH38.p7 using *Kallisto* (Bray et al., 2016) (v0.43.0). The software Kallisto was used to estimate isoform counts, adjusted for the amount of bias in the experiment to ensure a coherent no-naive mapping. Default paired end parameters indicated in kallisto's manual were used. The index needed to run Kallisto is provided on the official github repository[2],

[2] https://github.com/pachterlab/kallisto-transcriptome-indices/releases

TABLE 1 | Baseline characteristics of the cohorts.

	TCGA	GSE54460	VPCC	Total
Patients				
	52	96	23	171
Grade				
Low grade				
5	0	1	3	4
6	2	9	12	23
7	14	72	4	90
Total	16	82	19	117
High grade				
8	9	9	1	19
9	27	5	2	34
10	0	0	1	1
NA	0	0	0	0
Total	36	14	4	54
Total	52	96	23	171
Stage				
T1C	0	14	0	14
T2	0	7	0	7
T2A	1	21	3	25
T2B	2	10	0	12
T2C	9	26	17	52
T3	0	2	0	2
T3A	16	5	2	23
T3B	24	9	1	34
T4	0	1	0	1
NA	0	1	0	1
Total	52	96	23	171
BCR				
NO	14	54	5	73
YES	38	42	18	98
Total	52	96	23	171
PSA at dx/preop				
< = 10	31	64	21	116
10–20	16	17	1	34
> = 20	5	12	1	18
NA	0	3	0	3
Total	52	96	23	171

but can be manually created. Consequently, we computed gene counts with *tximport* (Soneson et al., 2015) (**Figure 2**). The Ensembl gene identifiers were converted with *Biomart tools* (Kinsella et al., 2011; Smedley et al., 2015) from transcript ID to gene ID. For both GSE54460 and VPCC datasets, we processed the raw fastq files using the same method as for the TCGA dataset. However, in GSE54460 the ribosomal sequences were still present within the reads, so we separated these sequences from the mapped reads and removed them. After mapping procedure, 29820 Ensembl genes were found in TCGA-PRAD dataset, 28704 in GSE54460 dataset and 32334 in VPCC dataset. The difference of number of Ensembl genes detected is explained by the sequencing depth of the datasets. A total of 25504 Ensembl genes were common to all sets and were retained for the analysis.

Normalization

The gene expression data were normalized with the *RUV method* (Gagnon-Bartsch and Speed, 2012; Risso et al., 2014) in each dataset separately following the default protocol indicated in the RUVseq package vignette. RUVg uses negative control genes [housekeeping genes (HKG)], assumed not to be differentially expressed. In order to normalize properly we selected in the literature a set of specific HKG candidates for PCa (de Kok et al., 2005; Ohl et al., 2005; Chua et al., 2011; Vajda et al., 2013): ACTB, PPIA, GAPDH, PGK1, GUSB, RRN18S, and RPL13A. The expression of these genes was tested by RT-qPCR in a series of 50 prostate tumors and the genes were shown to be stably expressed between tumor samples. We excluded from the final list the ribosomal genes RRN18S and RPL13A because ribosomal RNAs were removed from our RNA-seq datasets. PGK1 was also excluded according to recent results (Vajda et al., 2013). Finally, four genes were chosen: GUSB, PPIA, GAPDH, and ACTB.

Machine Learning

There are multiple approaches to treat biological data in a machine learning workflow (Al-Jarrah et al., 2015; Makridakis et al., 2018). Many machine learning libraries exist, in various programming languages, such as MLR in R (Lesmeister, 2015), Scikit-Learn (Garreta and Moncecchi, 2013) in python and WEKA (Hall et al., 2009) in Java. We chose the MLR (v2.8) package in R to set up our work. Our general workflow is described in **Figure 3**.

Validation Strategy

We performed a resampling to assess the performance of the learning algorithm, avoid over-optimistic results and get a more robust measure of the performance of our model. The entire dataset was split into a random stratified (i.e., class balance preserved) training and testing sets, 1000 times, hence the classification algorithm is trained and tested on different sets. The measure of performance is an aggregated value (e.g., average) of the individual performance on the test set. Because we have no repeated measures and independent variables (i.e., the patients) we chose the subsampling method which is also the best in general in different benchmarks but is less effective computationally (Bischl et al., 2012). The resampling strategy was run 200 times with a split of 2/3 for training and 1/3 for test sets. In the resampling methods the split is usually 4/5 or 9/10. In our case we wanted to avoid over-optimistic results then we chose a smaller train set closer to a classical cross validation (CV) approach.

Performance Metric

To evaluate the performance we used the balanced error rate (BER), the matthews correlation coefficient (MCC) and the mean misclassification error (MMCE). The BER is calculated as the average proportion of wrongly classified samples in each class and weights up small sample size classes (**Table 2**). The area under the curve (AUC) was also reported.

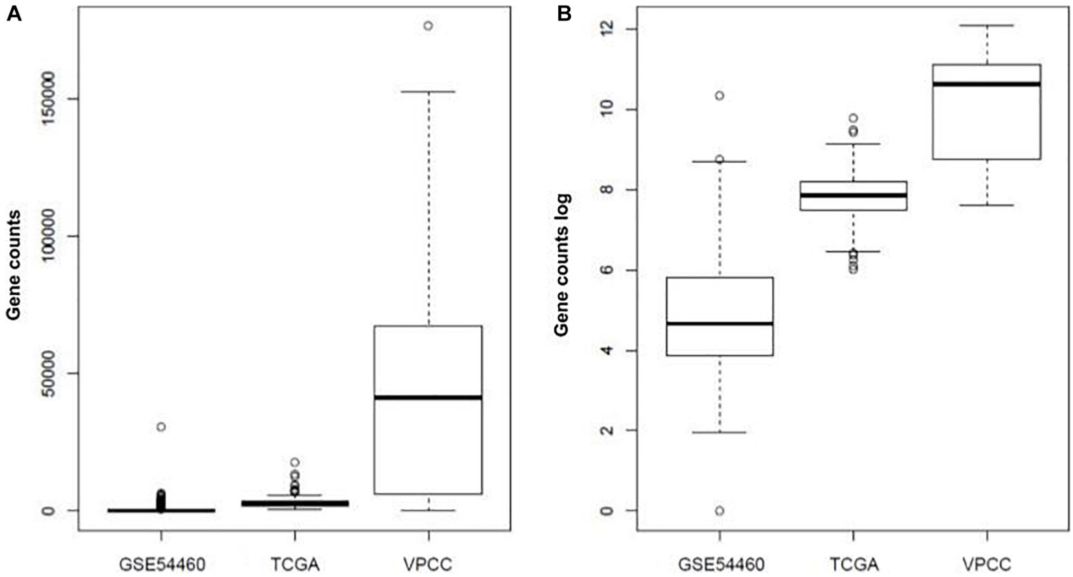

FIGURE 2 | Summary of gene expression value in each dataset **(A)** or log of the expression value **(B)**.

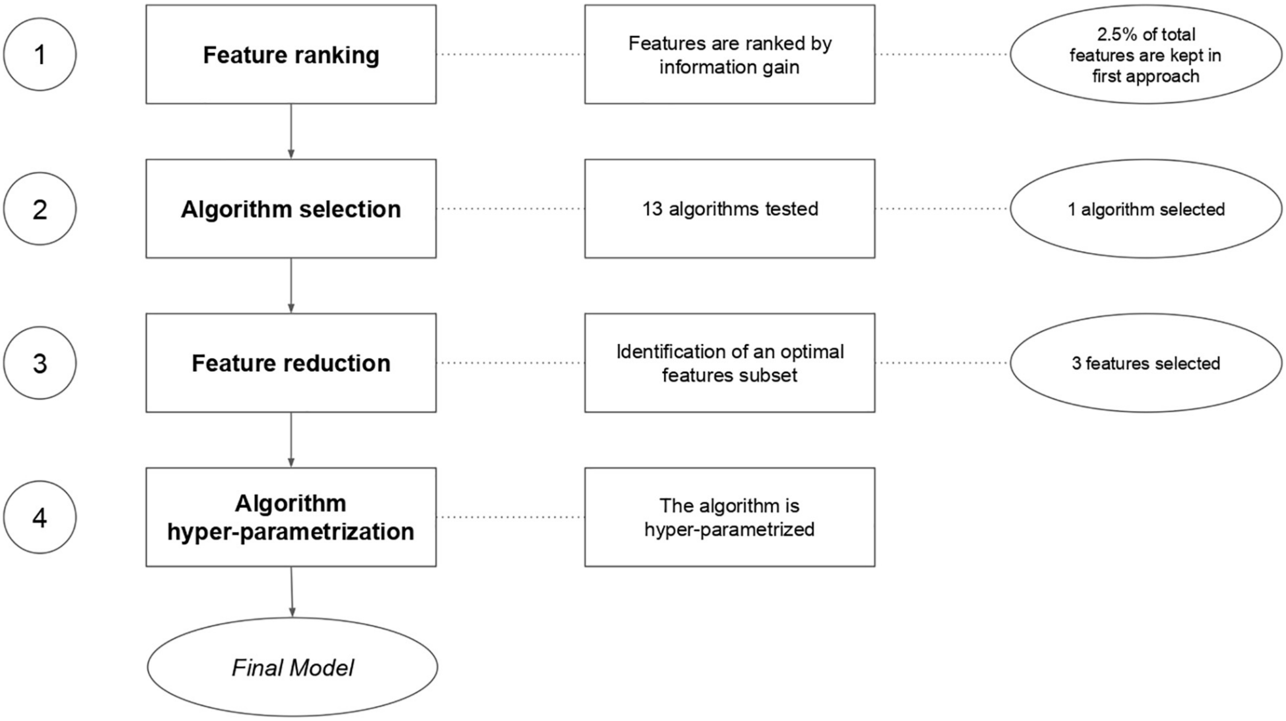
1
Feature ranking
Features are ranked by information gain
2.5% of total features are kept in first approach
2
Algorithm selection
13 algorithms tested
1 algorithm selected
3
Feature reduction
Identification of an optimal features subset
3 features selected
4
Algorithm hyper-parametrization
The algorithm is hyper-parametrized
Final Model

FIGURE 3 | Machine learning feature selection and model evaluation workflow.

Feature Selection

Feature selection was performed to reduce dimensionality to improve prediction performances by removing uninformative features, which has been proven successful in other studies (Novakovic et al., 2011). There are different approaches to identify relevant features (Hira and Gillies, 2015; Singh and Sivabalakrishnan, 2015; Raza and Qamar, 2019). We chose information gain ranking, an entropy based method, that can handle both numerical (e.g., gene expression) and categorical data (e.g., clinical data). In MLR this method relies on the package FSelector which is an entropy based selection method (Lin, 1991; Coifman and Wickerhauser, 1992).

TABLE 2 | Performance measures.

Performance metric	Formula
Sensitivity	TP/(TP + FN)
Specificity	TN/(TN + FP)
Accuracy	(TP + TN)/(TP + TN + FP + FN)*100
MCC	$\sqrt{(TP + TN)/(TP + TN + FP + FN)*100}$
BER	1–0.5 (Sensitivity + Specificity)
MMCE	Mean (response! = truth)

The detailed formula of our metrics.

Classifier Hyper-Parametrization

Algorithms typically require to change the settings of parameters to optimize their performance. The optimization method was the Irace method (López-Ibáñez et al., 2016) which is automated and implemented in an R package. We also work with a grid search algorithm for some specific parameters, which span the space in a number of chosen steps. These methods are also available within the MLR package to be used directly with the created tasks. The hyperparameters search depends on the algorithm iterated, defined in the MLR related man page.

RESULTS

Model and Features Selection

Following our machine learning pipeline (**Figure 3**), we first reduced the dimension of the dataset and removed non-informative features to obtain 400 top ranked features to train and benchmark 13 models (**Figure 4**). We observed that the random forest (RF) algorithm (Ho, 1995) performed best on our data. The classical RF was chosen as the main model for our further analysis.

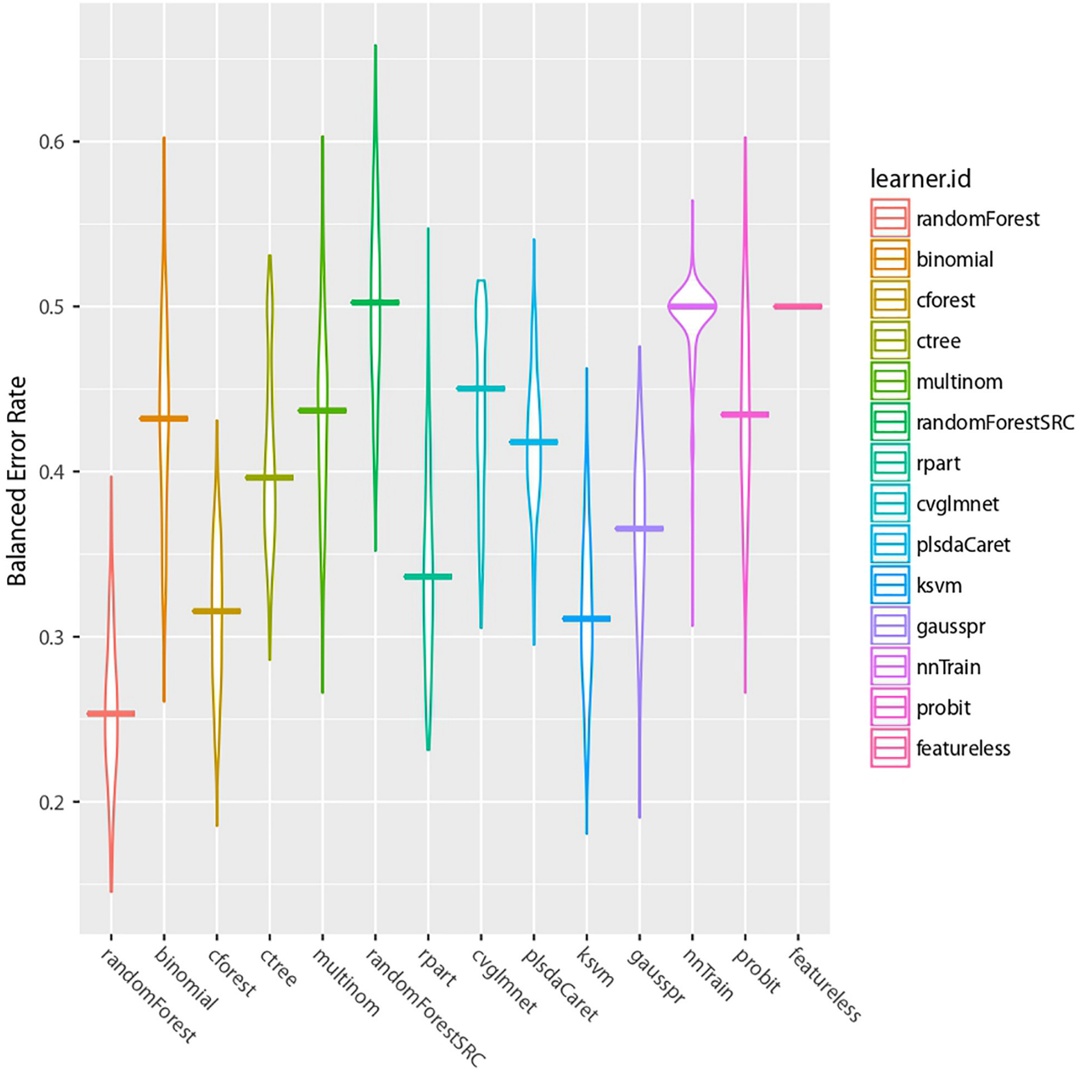

FIGURE 4 | Machine learning algorithms comparison. The BER results of our 13 benchmarked algorithms are presented. The last model is a featureless control case.

Since our goal was to identify a very short genomic signature we looked up the BER rate and other metrics while varying the number of selected features, from 1 to 400, used in the model. We observed that the BER and MMCE dropped rapidly with a few features selected (<3) then oscillated around 0.27 (**Figure 5**).

The MCC and the accuracy (ACC) went up rapidly and stabilized in the same way. After these observations, we focused the analysis on the first eight genes. The results are shown in **Table 3**. We observed a shift in BER value after adding the third most predictive gene to the signature. Afterward, BER begins to stabilize around 0.25–0.28 despite adding more informative genes. Consequently, we decided to keep the first three genes for the rest of the analysis. These genes are ENSG00000125534 (PPDPF), ENSG00000177606 (JUN), and ENSG00000188290 (HES4).

Hyper-Parameters Optimization and Final Model

Four hyper-parameters of the RF classifier were optimized: ntree, mtry, maxnode, and nodesize. Ntree refers to the number of decision trees in the model, mtry the number of variables selected from a decision split for the next split, maxnodes the maximal number of nodes in the forest and nodesize the minimal number of samples allowed in a node. Because we selected only three features, the parametrization step was not expected to drastically

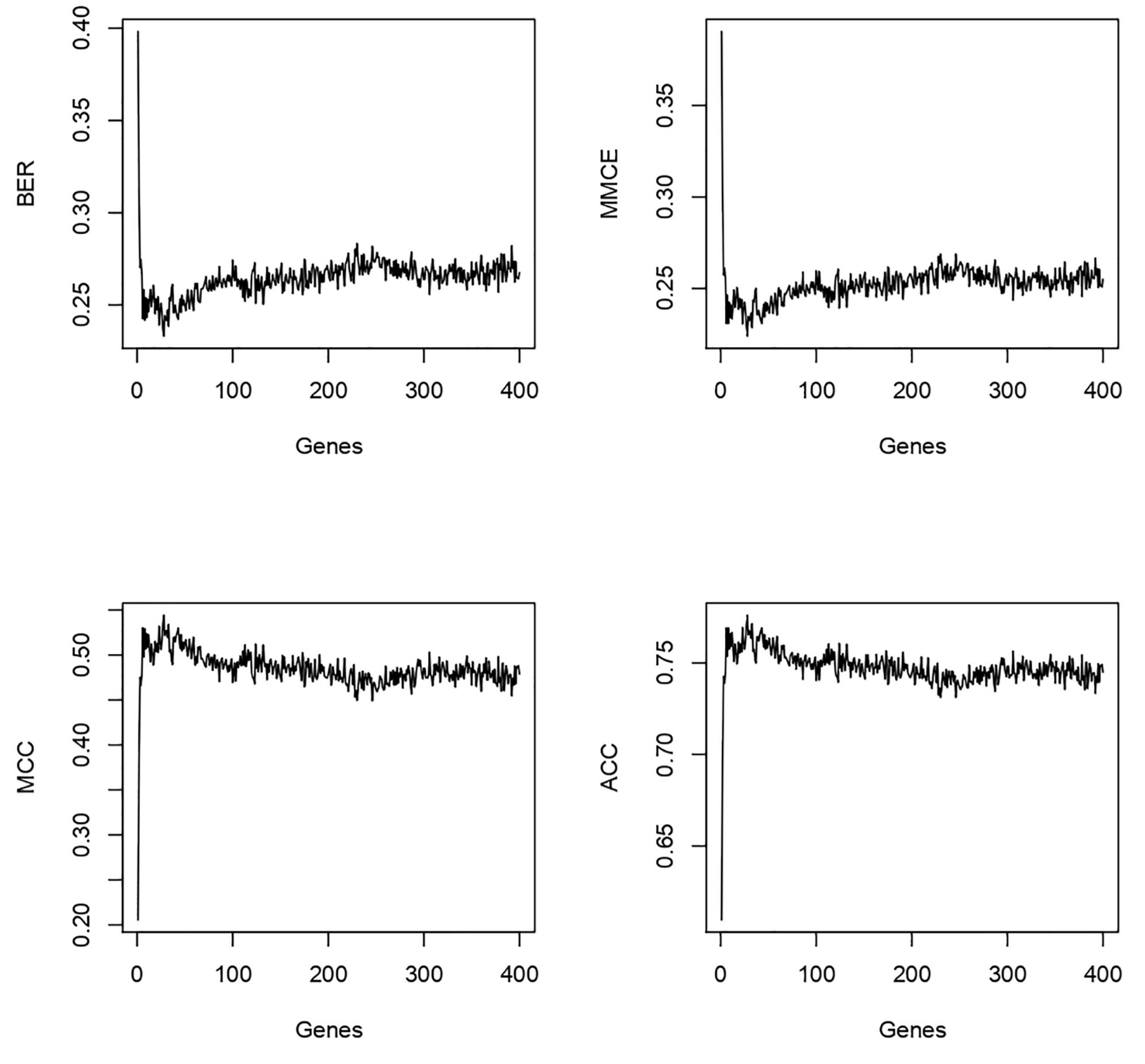

FIGURE 5 | For the 400 genes tested the best genes/performance ratio is obtained with less than 20 genes in our model.

TABLE 3 | Feature selection benchmark.

Nb of features	BER	MMCE	MCC	ACC	Gene name	ENSG
1	0.40	0.39	0.20	0.60	PPDPF	ENSG00000125534
2	0.32	0.30	0.38	0.69	HES4	ENSG00000188290
3	0.28	0.28	0.48	0.74	JUN	ENSG00000177606
4	0.28	0.26	0.47	0.73	GNB2	ENSG00000172354
5	0.28	0.26	0.48	0.74	PYROXD2	ENSG00000119943
6	0.25	0.23	0.53	0.77	MAP3K2	ENSG00000169967
7	0.27	0.25	0.50	0.75	RPL28	ENSG00000108107
8	0.25	0.23	0.53	0.77	DHCR24	ENSG00000116133

Benchmark on specific number of features has been performed and results of the performance metrics are presented.

change the performance of our optimization task. First we used a grid search method to define the best setting for each parameter taken individually, letting the others at default. The grid search provided us 500 (ntree), 1 (mtry), 24 (maxnodes), and 5 (nodesize) (**Figure 6**).

From these hyper-parameters an Irace search was performed around the space of those values. The best value was obtained with ntree, mtry, maxnodes and nodesize at 187, 1, 881 and 1 resp. for a BER of 0.27. We observed relative stability despite the modification of the hyperparameters.

To ensure the stability of our three-gene model, a subsampling test was done 100000 times for the last part of our work. From this subsampling, the results obtained are ber = 0.274, mmce = 0.26, mcc = 0.468, fpr = 0.368, tpr = 0.82, acc = 0.739. Then we calculated the associated AUC (0.761) and plotted the ROC curve **Figure 7**.

The proposed three genes signature (see gene distribution for each cohort in **Figure 8**) model can be retrained using the training data provided in the github repository (see "Data Availability Statement" section), and new data must be processed following the indications in Materials and Methods before being submitted to the model.

Comparison of Omics and Clinic Models

We compared the potential of omic data versus clinical data to assess the ACC of our omics model. A RF model for the clinical data (Grade, stage, and PSA) and a merged model combining clinic and omics data were set up following the same protocol used for the omics data. For the clinical model the best BER obtained was 0.311 and for the mixed model the best BER obtained was 0.276 (**Table 4**).

Single Cohort Performance

To further assess the performance of the three-gene model obtained with the combined dataset, we also performed the analysis with the individual cohorts. We used the RF algorithm iterated on the 50 best features from Information Gain on the three datasets evaluated by leave one out group validation (i.e., two datasets for training, one for testing), and the combined dataset evaluated by resampling (see section "Validation Strategy"). The results are displayed in **Figure 9** and show that the combined dataset offers better and more stable performances.

DISCUSSION

Machine Learning is one of the fastest growing fields in bioinformatics (Inza et al., 2010) and its application to healthcare is a challenge. In the past decade, various mathematical methods using combination of omics biomarkers (Halabi et al., 2003; Gaudreau et al., 2016), including non-coding RNAs, PCA3, TMPRSS2:ERG (Nilsson et al., 2009) were developed to improve PCa diagnosis (Wang et al., 2017; Guo et al., 2018), define the grade (Arvaniti et al., 2018), define the risk (Paulo et al., 2018) and predict survival time (Zupan et al., 2000). Machine learning approaches to predict BCR or other characteristics demonstrated good performances in various situations. Lalonde et al. (2014, 2017) built a 100 loci-DNA (CNV) signature for low to high risk cohort with 563 patients and a 60-month follow-up for BCR. The obtained AUC was 0.74, which is similar to our performance but with another technology (CNV assay) and for much fewer biomarkers. Moreover, a model containing so many features can be suspected of overfitting. Regnier-Coudert et al. (2012) built a model on Partin table from a large cohort of 1700 patients to improve cancer grading and staging, and obtained an AUC of 0.68. Mangiola et al. (2018) focused on gene expression but chose to predict dichotomous cohorts with low versus high risk patients. With a cohort of 80 patients and an average follow-up of 27–29 months they achieved an AUC of 0.72. Finally, Abou-Ouf et al. (2018) used a large cohort of 545 patients to define a ten-gene signature from microarray exon chips to predict BCR, but couldn't exceed an AUC of 0.65. Thus, there was a large room for improvement in terms of predictive performance, and a lack of focus on small gene signature, much easier to reproduce, to predict BCR with recent technology (RNA-Seq).

In this study, we took advantage of the power of machine learning to identify a biomarker signature composed of three genes. We showed that such short signature from omics data performs better to predict BCR than clinico-pathological features or a combination of these data (i.e., clinico-pathological + omics data). We have explored many machine learning algorithms, since each has its advantages and drawbacks in terms of computational time, hyper-parameters and range of application (class, type and dimension) and also because their performance depends on the type of data and their composition (Heung et al., 2016). Using this approach, we ended with a Random Forest model with a 27% BER with a three genes signature.

The identified signature contains three genes: JUN, HES4, and PPDPF. Gene JUN is well known for being a transcription factor acting as an oncogene (Maki et al., 1987; Vogt and Bos, 1990; Wasylyk et al., 1990; Mariani et al., 2007). Proteins of the JUN family combined with the Fos protein to form the heterodimeric AP-1 transcription factor. This complex can enter into the nucleus and bind specific DNA sequences to module targeted genes. AP-1 activity is induced by stimuli such as growth factors and cytokines that bind to

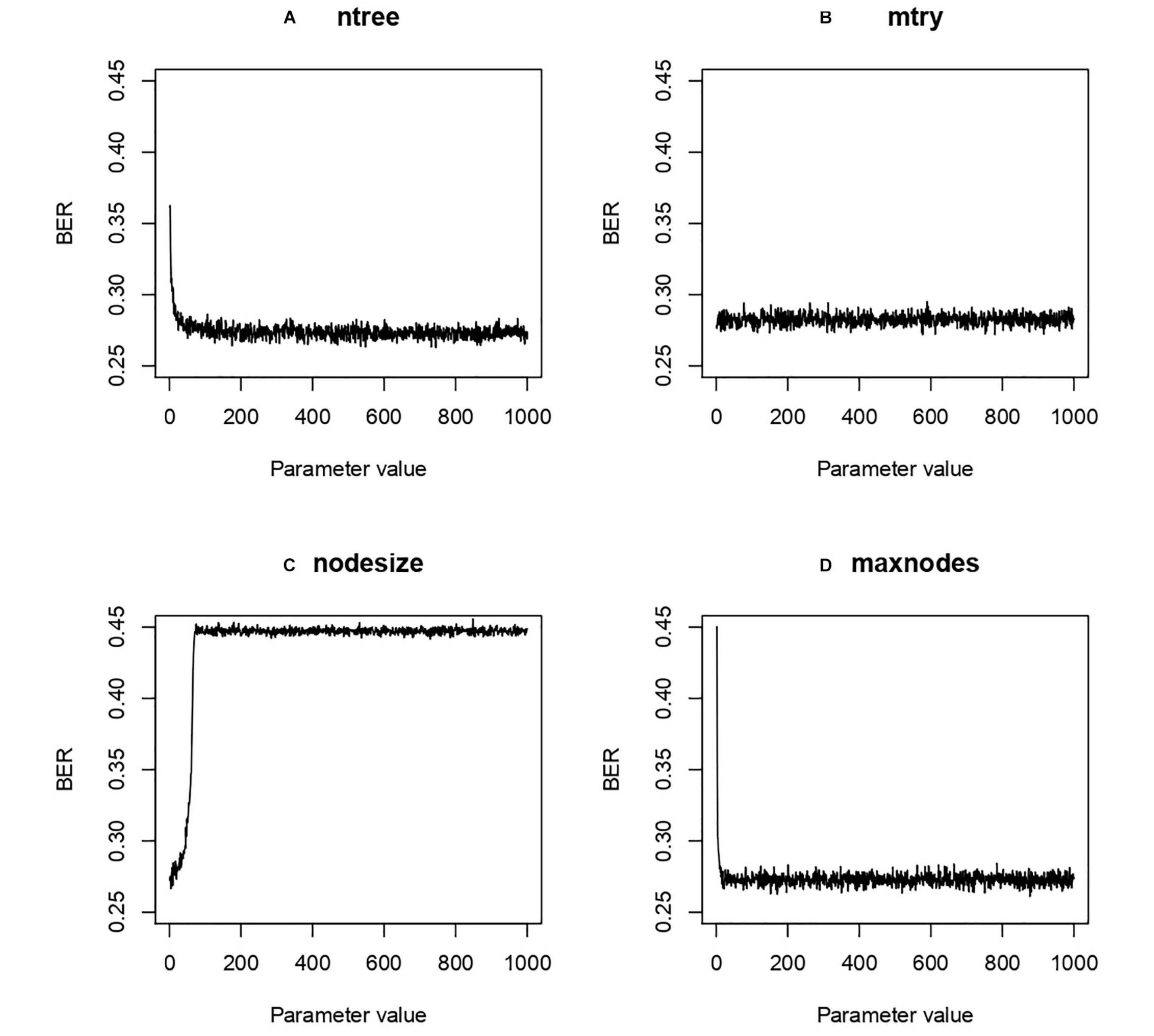

FIGURE 6 | Balanced Error Rate (BER) evolution according to modulation of Random Forest (RF) parameters. Four different RF hyper-parameters were tested while keeping the others at default value in a grid search approach. The results were then used in an Irace search to find optimal parameters. **(A)** ntree, number of decision trees; **(B)** mtry, number of variables selected from a decision split for the next split; **(C)** maxnodes, maximal number of nodes; **(D)** nodesize, minimal number of samples allowed in a node.

specific cell surface receptors (Yang et al., 1999). Recently a miRNA targeting JUN has been identified as tumor suppressor (Liu et al., 2015).

Hes Family BHLH Transcription Factor 4 (HES4) is a gene related to the PI3K-Akt signaling pathway. This gene is a transcription factor binding DNA. It is related to the NOTCH3 receptor and is a biomarker of PCa aggressiveness (Carvalho et al., 2012) and is also related to colorectal cancer in the same pathway (Sikandar et al., 2010). It was demonstrated as a high grade biomarker of osteosarcoma (McManus et al., 2017).

Finally, PPDPF is known to be expressed during pancreas development [Pancreatic Progenitor Cell Differentiation And Proliferation Factor (Breunig et al., 2017)] and differentially expressed in several types of cancer (Voena et al., 2013; Xue et al., 2015). But it was not previously associated with PCa.

We have attempted to understand the biological links between these three genes and the eventual relation with the BCR. This is not straightforward considering that Random Forest models tend to reflect a nonlinear approximation of statistical relationships, hence providing little insight of how

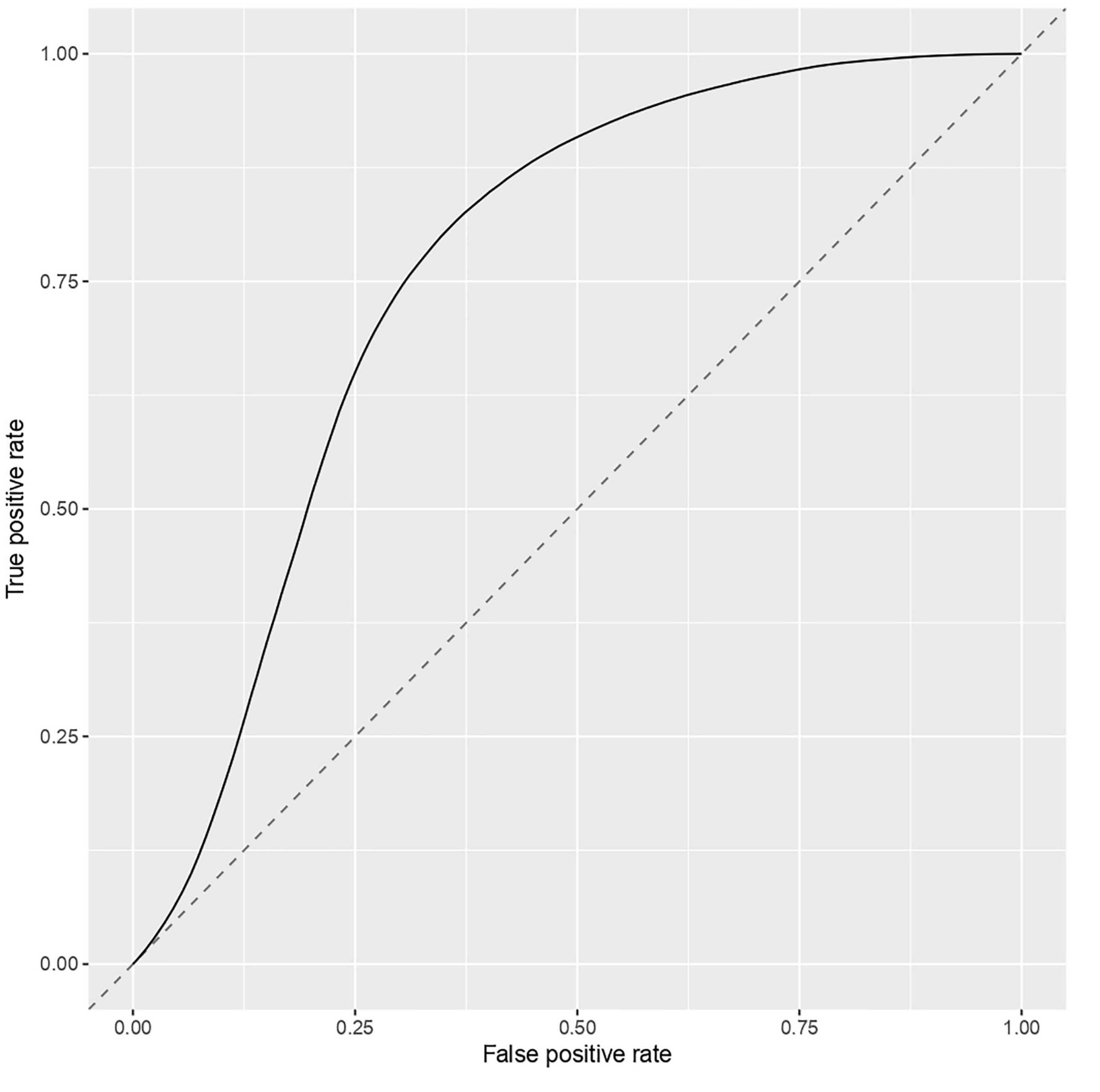

FIGURE 7 | ROC curve for the three-gene model.

elements of the signature are related. Thus, we have performed a protein-protein interaction networks functional enrichment analysis using String-DB (Szklarczyk et al., 2019) on the three identified genes, but no evident relations could be found, even after addition of intermediate protein nodes. We have also performed a gene list enrichment analysis and candidate gene prioritization based on functional annotations using ToppGene Suite (Chen et al., 2009) using the three identified genes. The only biologically relevant (i.e. cancer hormono-dependant as the PCa) and significant (q-value 2.1E-2 after FDR Benjamini-Yekutieli procedure correction) hit is that the three genes exist in the Human Breast Nam08 30 genes UpregulatedGeneList signature (Nam et al., 2008), provided by GeneSigDB (Culhane et al., 2012), but no evident and/or significant biological functions by ontology seem to link these three genes together. We have eventually expanded the list of three genes to 320 genes by retrieving correlated genes (>90% Pearson correlation) and observed that many genes were involved in mitochondrial functions, including mitochondrial translation, mitochondrial gene expression, mitochondrial translational termination and mitochondrial translational elongation, all having a q-value <5.9E-5 after FDR Benjamini-Yekutieli

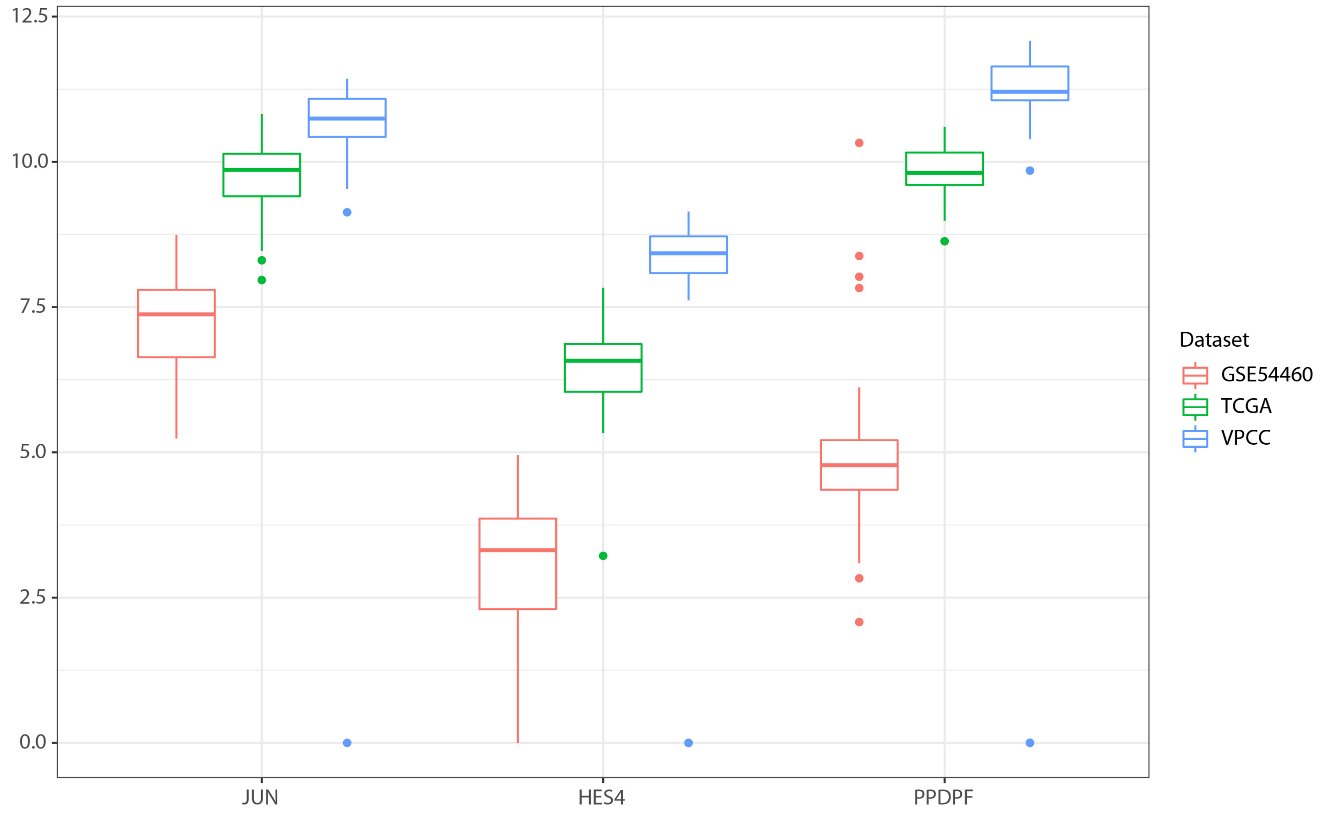

FIGURE 8 | Log2 transformed distribution of normalized read counts for the three genes signature in each cohort.

procedure correction. This observation is supported by other studies who have found a clear relation between mitochondrial genomic alterations and BCR (Ellinger et al., 2008; Kalsbeek et al., 2016; Xu et al., 2020).

This is not the first time that predictive three-genes signatures have been identified in various diseases (Sun et al., 2015; Thakkar et al., 2015; De Palma et al., 2016; Ibrahim et al., 2016; Wang et al., 2016; Li et al., 2017; Chen et al., 2018; Yang et al., 2018; Bao et al., 2019; Ding et al., 2019; Saidak et al., 2019; Xiao et al., 2020), hence showing that extensive research is ongoing to identify multigenic signatures containing a reasonable number of potential targets. The identified genes could be eventually verified in other cohorts or by experimental validations. One key point should be to add gradually smaller datasets to control the signature stability with various experiments and technologies. Integrate too large cohorts in this approach will imbalance model parameters in favor of that cohort, then all the advantages of using several small dataset will be lost. This approach has the advantage of offering a small research team the opportunity to integrate their own work in a larger view. After integrating more dataset, a set up in a specific technology such as TaqMan probe to evaluate gene expression could be proposed as diagnosis and maybe to develop drugs (Laetsch et al., 2018; Havel et al., 2019).

TABLE 4 | Comparison of model performance using clinic or omics data or both.

Metric	Omics	Clinic	Omics + Clinic
BER	0.27	0.32	0.28
MMCE	0.257	0.306	0.265
MCC	0.474	0.373	0.457
ACC	0.742	0.692	0.734
Parameters			
ntree	187	1402	667
mtry	1	3	1
maxnodes	881	30	25
nodesize	1	4	6

The omics model is based on three genes and the clinic model is a model based on the grade, stage and PSA. The omics + clinic model integrates all the selected features together.

CONCLUSION

By using an appropriate data transformation strategy and machine learning pipeline, we have identified a three-gene signature. With the decreasing price of RNA sequencing and its growing accuracy there are opportunities for less invasive and faster exams if the right biological variables are chosen. Other investigations on other omics data using the same machine learning approach could be undertaken, such as using miRNAs (Kristensen et al., 2016; Matin et al., 2018).

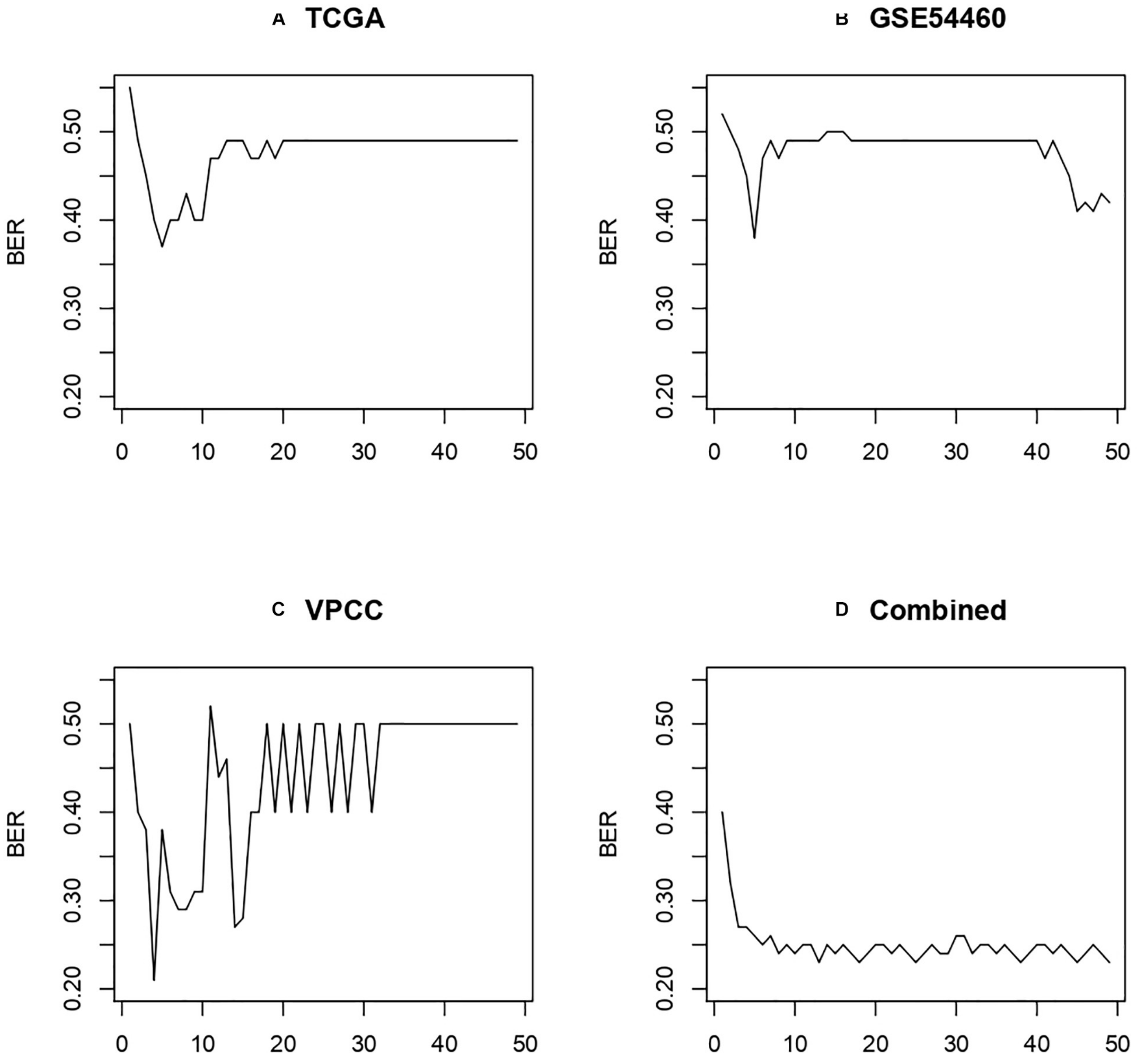

FIGURE 9 | Performance obtained using leave one out group validation. **(A)** Model trained on GSE54460 and VPCC then tested on TCGA. **(B)** Model trained on TCGA and VPCC then tested on GSE54460. **(C)** Model trained on GSE54460 and TCGA then tested on VPCC. **(D)** Combined dataset evaluated by subsampling method described in "Validation Strategy."

We also showed that it is possible to concatenate several cohorts to get stable and performing models from heterogeneous RNA-Seq PCa datasets, hence showing a robustness against batch effect. This study demonstrates the potential of taking advantage of many independent datasets produced on the same disease. Machine learning algorithms can handle the batch effect if there is the right preprocessing pipeline applied on the data.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: TCGA at GDC data portal; GEO accession GSE54460; The European Nucleotide Archive (ENA), accession number PRJEB6530 from Wyatt et al. (2014). Moreover, the scripts developed for this study and the processed read counts are available at github.com/ArnaudDroitLab/prostate_BCR_prediction.

ETHICS STATEMENT

This study was approved by the Research Ethics Committee of the CHU de Québec-Université Laval (Project 2018-3670). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

BV conducted literature searches, gathered the data, wrote the code to perform the research, and wrote the manuscript. He also created figures and tables, and wrote, formatted the manuscript for submission. ML participated to design the approach. ML, M-LM-M, and AB helped to improve the manuscript. YF and AD supervised and reviewed the design of the study. CC provided the VPCC data. All authors contributed to reviewing the manuscript.

FUNDING

This research was realized with internal funds from the Laboratoire d'Uro-Oncologie Expérimentale (Ulaval, Dr. Fradet). The production of RNA-seq data at VPCC was realized with funds from the Terry Fox Research Institute New Frontier Program Project Grant #1062 (TFRI NF PPG, UBC - Dr. Collins).

REFERENCES

Abou-Ouf, H., Alshalalfa, M., Takhar, M., Erho, N., Donnelly, B., Davicioni, E., et al. (2018). Validation of a 10-gene molecular signature for predicting biochemical recurrence and clinical metastasis in localized prostate cancer. *J. Cancer Res. Clin. Oncol.* 144, 883–891. doi: 10.1007/s00432-018-2615-7

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., and Taha, K. (2015). Efficient machine learning for big data: a review. *Big Data Res.* 2, 87–93. doi: 10.1016/j.bdr.2015.04.001

Almeida, H., Meurs, M.-J., Kosseim, L., Butler, G., and Tsang, A. (2014). Machine learning for biomedical literature triage. *PLoS One* 9:e115892. doi: 10.1371/journal.pone.0115892

Amin, M. B., Edge, S. B., Greene, F. L., Byrd, D. R., Brookland, R. K., Washington, M. K., et al. (2018). *AJCC Cancer Staging Manual*. Berlin: Springer.

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., and Wingett, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Babraham: Babraham Institute.

Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., et al. (2018). Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci. Rep.* 8:12054.

Bao, B., Zheng, C., Yang, B., Jin, Y., Hou, K., Li, Z., et al. (2019). Identification of subtype-specific three-gene signature for prognostic prediction in diffuse type gastric cancer. *Front. Oncol.* 9:1243. doi: 10.3389/fonc.2019.01243

Bischl, B., Mersmann, O., Trautmann, H., and Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evol. Comput.* 20, 249–275. doi: 10.1162/evco_a_00069

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi: 10.1038/nbt.3519

Breunig, M., Hohwieler, M., Seufferlein, T., Liebau, S., and Kleger, A. (2017). PPDPF impacts pancreatic differentiation of human pluripotent stem cell derived pancreatic organoids. *Z. Gastroenterol.* 55, e57–e299. doi: 10.1055/s-0037-1604922

Buyyounouski, M. K., Pickles, T., Kestin, L. L., Allison, R., and Williams, S. G. (2012). Validating the interval to biochemical failure for the identification of potentially lethal prostate cancer. *J. Clin. Oncol.* 30, 1857–1863. doi: 10.1200/jco.2011.35.1924

Cancer Genome Atlas Research Network (2015). The molecular taxonomy of primary prostate cancer. *Cell* 163, 1011–1025.

Carvalho, F. L. F., Simons, B., and Berman, D. M. (2012). Abstract B56: notch signaling in prostate cancer progression. *Cancer Res.* 72, B56–B56. doi: 10.1158/1538-7445.prca2012-b56

Chen, H., Liu, X., Jin, Z., Gou, C., Liang, M., Cui, L., et al. (2018). A three miRNAs signature for predicting the transformation of oral leukoplakia to oral squamous cell carcinoma. *Am. J. Cancer Res.* 8, 1403–1413.

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311.

Chua, S. L., See Too, W. C., Khoo, B. Y., and Few, L. L. (2011). UBC and YWHAZ as suitable reference genes for accurate normalisation of gene expression using MCF7, HCT116 and HepG2 cell lines. *Cytotechnology* 63, 645–654. doi: 10.1007/s10616-011-9383-4

Coifman, R. R., and Wickerhauser, M. V. (1992). Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory* 38, 713–718. doi: 10.1109/18.119732

Culhane, A. C., Schröder, M. S., Sultana, R., Picard, S. C., Martinelli, E. N., Kelly, C., et al. (2012). GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Res.* 40, D1060–D1066.

D'Amico, A. V., Moul, J., Carroll, P. R., Sun, L., Lubeck, D., and Chen, M.-H. (2003). Cancer-specific mortality after surgery or radiation for patients with clinically localized prostate cancer managed during the prostate-specific antigen era. *J. Clin. Oncol.* 21, 2163–2172. doi: 10.1200/jco.2003.01.075

de Kok, J. B., Roelofs, R. W., Giesendorf, B. A., Pennings, J. L., Waas, E. T., Feuth, T., et al. (2005). Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. *Lab. Invest.* 85, 154–159. doi: 10.1038/labinvest.3700208

De Palma, G., Sallustio, F., Curci, C., Galleggiante, V., Rutigliano, M., Serino, G., et al. (2016). The three-gene signature in urinary extracellular vesicles from patients with clear cell renal cell carcinoma. *J. Cancer* 7, 1960–1967. doi: 10.7150/jca.16123

Ding, T.-T., Ma, H., and Feng, J.-H. (2019). A three-gene novel predictor for improving the prognosis of cervical cancer. *Oncol. Lett.* 18, 4907–4915.

Edge, S. B., and Compton, C. C. (2010). The American joint committee on cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann. Surg. Oncol.* 17, 1471–1474. doi: 10.1245/s10434-010-0985-4

Ellinger, J., Müller, S. C., Wernert, N., von Ruecker, A., and Bastian, P. J. (2008). Mitochondrial DNA in serum of patients with prostate cancer: a predictor of biochemical recurrence after prostatectomy. *BJU Int.* 102, 628–632. doi: 10.1111/j.1464-410x.2008.07613.x

Gagnon-Bartsch, J. A., and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 13, 539–552. doi: 10.1093/biostatistics/kxr034

Garreta, R., and Moncecchi, G. (2013). *Learning Scikit-Learn: Machine Learning in Python*. Birmingham: Packt Publishing Ltd.

Gaudreau, P.-O., Stagg, J., Soulières, D., and Saad, F. (2016). The present and future of biomarkers in prostate cancer: proteomics, genomics, and immunology advancements. *Biomark. Cancer* 8, 15–33.

Guo, J., Yang, J., Zhang, X., Feng, X., Zhang, H., Chen, L., et al. (2018). A panel of biomarkers for diagnosis of prostate cancer using urine samples. *Anticancer Res.* 38, 1471–1477.

Halabi, S., Small, E. J., Kantoff, P. W., Kattan, M. W., Kaplan, E. B., Dawson, N. A., et al. (2003). Prognostic model for predicting survival in men with hormone-refractory metastatic prostate cancer. *J. Clin. Oncol.* 21, 1232–1237. doi: 10.1200/jco.2003.06.100

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explor. Newslett.* 11:10. doi: 10.1145/1656274.1656278

Havel, J. J., Chowell, D., and Chan, T. A. (2019). The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat. Rev. Cancer* 19, 133–150. doi: 10.1038/s41568-019-0116-x

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., and Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma* 265, 62–77. doi: 10.1016/j.geoderma.2015.11.014

Hira, Z. M., and Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Adv. Bioinform.* 2015:198363.

Ho, T. K. (1995). "International conference on document analysis and recognition," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, QC.

Ibrahim, M. K., Salama, H., Abd El Rahman, M., Dawood, R. M., Bader, El Din, N. G., et al. (2016). Three gene signature for predicting the development of hepatocellular carcinoma in chronically infected Hepatitis C virus patients. *J. Interf. Cytokine Res.* 36, 698–705. doi: 10.1089/jir.2016.0042

International Cancer Genome Consortium Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi: 10.1038/nature08987

Inza, I., Calvo, B., Armañanzas, R., Bengoetxea, E., Larrañaga, P., and Lozano, J. A. (2010). Machine learning: an indispensable tool in bioinformatics. *Methods Mol. Biol.* 593, 25–48. doi: 10.1007/978-1-60327-194-3_2

Kalsbeek, A. M. F., Chan, E. F. K., Grogan, J., Petersen, D. C., Jaratlerdsiri, W., Gupta, R., et al. (2016). Mutational load of the mitochondrial genome predicts pathological features and biochemical recurrence in prostate cancer. *Aging* 8, 2702–2712. doi: 10.18632/aging.101044

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011:bar030. doi: 10.1093/database/bar030

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Kristensen, H., Thomsen, A. R., Haldrup, C., Dyrskjøt, L., Høyer, S., Borre, M., et al. (2016). Novel diagnostic and prognostic classifiers for prostate cancer identified by genome-wide microRNA profiling. *Oncotarget* 7, 30760–30771. doi: 10.18632/oncotarget.8953

Laetsch, T. W., DuBois, S. G., Mascarenhas, L., Turpin, B., Federman, N., Albert, C. M., et al. (2018). Larotrectinib for paediatric solid tumours harbouring NTRK gene fusions: phase 1 results from a multicentre, open-label, phase 1/2 study. *Lancet Oncol.* 19, 705–714. doi: 10.1016/s1470-2045(18)30119-0

Lalonde, E., Alkallas, R., Chua, M. L. K., Fraser, M., Haider, S., Meng, A., et al. (2017). Translating a prognostic DNA genomic classifier into the clinic: retrospective validation in 563 localized prostate tumors. *Eur. Urol.* 72, 22–31. doi: 10.1016/j.eururo.2016.10.013

Lalonde, E., Ishkanian, A. S., Sykes, J., Fraser, M., Ross-Adams, H., Erho, N., et al. (2014). Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. *Lancet Oncol.* 15, 1521–1532. doi: 10.1016/s1470-2045(14)71021-6

Lesmeister, C. (2015). *Mastering Machine Learning with R*. Birmingham: Packt Publishing Ltd.

Li, B., Feng, W., Luo, O., Xu, T., Cao, Y., Wu, H., et al. (2017). Development and validation of a three-gene prognostic signature for patients with hepatocellular carcinoma. *Sci. Rep.* 7:5517.

Li, Y., Wu, F.-X., and Ngom, A. (2016). A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.* 19, 325–340. doi: 10.1093/bib/bbw113

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory* 37, 145–151. doi: 10.1109/18.61115

Liu, J., Yan, J., Zhou, C., Ma, Q., Jin, Q., and Yang, Z. (2015). miR-1285-3p acts as a potential tumor suppressor miRNA via downregulating JUN expression in hepatocellular carcinoma. *Tumour Biol.* 36, 219–225. doi: 10.1007/s13277-014-2622-5

Long, Q., Xu, J., Osunkoya, A. O., Sannigrahi, S., Johnson, B. A., Zhou, W., et al. (2014). Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence. *Cancer Res.* 74, 3228–3237. doi: 10.1158/0008-5472.can-13-2699

López-Ibáñez, M., Dubois-Lacoste, J., Cáceres, L. P., Birattari, M., and Stützle, T. (2016). The irace package: iterated racing for automatic algorithm configuration. *Operat. Res. Perspect.* 3, 43–58. doi: 10.1016/j.orp.2016.09.002

Maki, Y., Bos, T. J., Davis, C., Starbuck, M., and Vogt, P. K. (1987). Avian sarcoma virus 17 carries the jun oncogene. *Proc. Natl. Acad. Sci. U.S.A.* 84, 2848–2852. doi: 10.1073/pnas.84.9.2848

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: concerns and ways forward. *PLoS One* 13:e0194889. doi: 10.1371/journal.pone.0194889

Mangiola, S., Stuchbery, R., Macintyre, G., Clarkson, M. J., Peters, J. S., Costello, A. J., et al. (2018). Periprostatic fat tissue transcriptome reveals a signature diagnostic for high-risk prostate cancer. *Endocrine Relat. Cancer* 25, 569–581. doi: 10.1530/erc-18-0058

Mariani, O., Brennetot, C., Coindre, J.-M., Gruel, N., Ganem, C., Delattre, O., et al. (2007). JUN oncogene amplification and overexpression block adipocytic differentiation in highly aggressive sarcomas. *Cancer Cell* 11, 361–374. doi: 10.1016/j.ccr.2007.02.007

Marx, V. (2013). The big challenges of big data. *Nature* 498, 255–260. doi: 10.1038/498255a

Matin, F., Australian Prostate Cancer BioResource, Jeet, V., Moya, L., Selth, L. A., Chambers, S., et al. (2018). A plasma biomarker panel of four MicroRNAs for the diagnosis of prostate cancer. *Sci. Rep.* 8:6653. doi: 10.1038/s41598-018-24424-w

McManus, M., Kleinerman, E., Yang, Y., Livingston, J. A., Mortus, J., Rivera, R., et al. (2017). Hes4: a potential prognostic biomarker for newly diagnosed patients with high-grade osteosarcoma. *Pediatr. Blood Cancer* 64:10.1002/bc.26318. doi: 10.1002/pbc.26318

Menegon, M., Cantaloni, C., Rodriguez-Prieto, A., Centomo, C., Abdelfattah, A., Rossato, M., et al. (2017). On site DNA barcoding by nanopore sequencing. *PLoS One* 12:e0184741. doi: 10.1371/journal.pone.0184741

Nam, D. H., Jeon, H. M., Kim, S., Kim, M. H., Lee, Y. J., Lee, M. S., et al. (2008). Activation of notch signaling in a xenograft model of brain metastasis. *Clin. Cancer Res.* 14, 4059–4066. doi: 10.1158/1078-0432.CCR-07-4039

Nevedomskaya, E., Baumgart, S. J., and Haendler, B. (2018). Recent advances in prostate cancer treatment and drug discovery. *Int. J. Mol. Sci.* 19:1359. doi: 10.3390/ijms19051359

Nikitina, A. S., Sharova, E. I., Danilenko, S. A., Butusova, T. B., Vasiliev, A. O., Govorov, A. V., et al. (2017). Novel RNA biomarkers of prostate cancer revealed by RNA-seq analysis of formalin-fixed samples obtained from Russian patients. *Oncotarget* 8, 32990–33001. doi: 10.18632/oncotarget.16518

Nilsson, J., Skog, J., Nordstrand, A., Baranov, V., Mincheva-Nilsson, L., Breakefield, X. O., et al. (2009). Prostate cancer-derived urine exosomes: a novel approach to biomarkers for prostate cancer. *Br. J. Cancer* 100, 1603–1607. doi: 10.1038/sj.bjc.6605058

Novakovic, J., Strbac, P., and Bulatovic, D. (2011). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav J. Operat. Res.* 21, 119–135. doi: 10.2298/yjor1101119n

Ohl, F., Jung, M., Xu, C., Stephan, C., Rabien, A., Burkhardt, M., et al. (2005). Gene expression studies in prostate cancer tissue: which reference gene should be selected for normalization? *J. Mol. Med.* 83, 1014–1024. doi: 10.1007/s00109-005-0703-z

Papsidero, L. D., Wang, M. C., Valenzuela, L. A., Murphy, G. P., and Chu, T. M. (1980). A prostate antigen in sera of prostatic cancer patients. *Cancer Res.* 40, 2428–2432.

Paulo, P., Maia, S., Pinto, C., Pinto, P., Monteiro, A., Peixoto, A., et al. (2018). Targeted next generation sequencing identifies functionally deleterious germline mutations in novel genes in early-onset/familial prostate cancer. *PLoS Genet.* 14:e1007355. doi: 10.1371/journal.pone.1007355

Raza, M. S., and Qamar, U. (2019). "Introduction to feature selection," in *Understanding and Using Rough Set Based Feature Selection: Concepts, Techniques and Applications*, eds U. Qamar, and M. S. Raza (Singapore: Springer), 1–25. doi: 10.1007/978-981-32-9166-9_1

Regnier-Coudert, O., McCall, J., Lothian, R., Lam, T., McClinton, S., and N'dow, J. (2012). Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers. *Artif. Intell. Med.* 55, 25–35. doi: 10.1016/j.artmed.2011.11.003

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902. doi: 10.1038/nbt.2931

Saidak, Z., Pascual, C., Bouaoud, J., Galmiche, L., Clatot, F., and Dakpé, S. (2019). A three-gene expression signature associated with positive surgical margins in tongue squamous cell carcinomas: predicting surgical resectability from tumour biology? *Oral Oncol.* 94, 115–120. doi: 10.1016/j.oraloncology.2019.05.020

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer statistics, 2017. *CA Cancer J. Clin.* 67, 7–30. doi: 10.3322/caac.21387

Sikandar, S. S., Pate, K. T., Anderson, S., Dizon, D., Edwards, R. A., Waterman, M. L., et al. (2010). NOTCH signaling is required for formation and self-renewal of tumor-initiating cells and for repression of secretory cell differentiation in colon cancer. *Cancer Res.* 70, 1469–1478. doi: 10.1158/0008-5472.can-09-2557

Singh, R. K., and Sivabalakrishnan, M. (2015). Feature selection of gene expression data for cancer classification: a review. *Proc. Comput. Sci.* 50, 52–57. doi: 10.1016/j.procs.2015.04.060

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* 43, W589–W598.

Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* 4:1521. doi: 10.12688/f1000research.7563.2

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big data: astronomical or genomical? *PLoS Biol.* 13:e1002195. doi: 10.1371/journal.pone.1002195

Sun, L.-L., Wu, J.-Y., Wu, Z.-Y., Shen, J.-H., Xu, X.-E., Chen, B., et al. (2015). A three-gene signature and clinical outcome in esophageal squamous cell carcinoma. *Int. J. Cancer* 136, E569–E577.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., 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.

Tannock, I. F., de Wit, R., Berry, W. R., Horti, J., Pluzanska, A., Chi, K. N., et al. (2004). Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. *New Engl. J. Med.* 351, 1502–1512. doi: 10.1056/nejmoa040720

Terada, N., Akamatsu, S., Kobayashi, T., Inoue, T., Ogawa, O., and Antonarakis, E. S. (2017). Prognostic and predictive biomarkers in prostate cancer: latest evidence and clinical implications. *Therap. Adv. Med. Oncol.* 9, 565–573. doi: 10.1177/1758834017719215

Thakkar, A., Raj, H., Ravishankar, L., Muthuvelan, B., Balakrishnan, A., and Padigaru, M. (2015). High expression of three-gene signature improves prediction of relapse-free survival in estrogen receptor-positive and node-positive breast tumors. *Biomark. Insights* 10, 103–112.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77.

Vajda, A., Marignol, L., Barrett, C., Madden, S. F., Lynch, T. H., Hollywood, D., et al. (2013). Gene expression analysis in prostate cancer: the importance of the endogenous control. *Prostate* 73, 382–390. doi: 10.1002/pros.22578

Voena, C., Di Giacomo, F., Panizza, E., D'Amico, L., Boccalatte, F. E., Pellegrino, E., et al. (2013). The EGFR family members sustain the neoplastic phenotype of ALK+ lung adenocarcinoma via EGR1. *Oncogenesis* 2:e43. doi: 10.1038/oncsis.2013.7

Vogt, P. K., and Bos, T. J. (1990). jun:Oncogene and transcription factor. *Adv. Cancer Res.* 55, 1–35. doi: 10.1016/s0065-230x(08)60466-2

Wang, W., Zhang, L., Wang, Z., Yang, F., Wang, H., Liang, T., et al. (2016). A three-gene signature for prognosis in patients with MGMT promoter-methylated glioblastoma. *Oncotarget* 7, 69991–69999. doi: 10.18632/oncotarget.11726

Wang, X., An, P., Zeng, J., Liu, X., Wang, B., Fang, X., et al. (2017). Serum ferritin in combination with prostate-specific antigen improves predictive accuracy for prostate cancer. *Oncotarget* 8, 17862–17872. doi: 10.18632/oncotarget.14977

Wasylyk, C., Schneikert, J., and Wasylyk, B. (1990). Oncogene v-jun modulates DNA replication. *Oncogene* 5, 1055–1058.

Weiner, A. B., Matulewicz, R. S., Eggener, S. E., and Schaeffer, E. M. (2016). Increasing incidence of metastatic prostate cancer in the United States (2004-2013). *Prostate Cancer Prostat. Dis.* 19, 395–397. doi: 10.1038/pcan.2016.30

Wyatt, A. W., Mo, F., Wang, K., McConeghy, B., Brahmbhatt, S., Jong, L., et al. (2014). Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer. *Genome Biol.* 15:426.

Xiao, K., Liu, Q., Peng, G., Su, J., Qin, C.-Y., and Wang, X.-Y. (2020). Identification and validation of a three-gene signature as a candidate prognostic biomarker for lower grade glioma. *PeerJ* 8:e8312. doi: 10.7717/peerj.8312

Xu, J., Chang, W.-S., Tsai, C.-W., Bau, D.-T., Davis, J. W., Thompson, T. C., et al. (2020). Mitochondrial DNA copy number in peripheral blood leukocytes is associated with biochemical recurrence in prostate cancer patients in African Americans. *Carcinogenesis* 41, 267–273. doi: 10.1093/carcin/bgz139

Xue, T.-C., Zhang, B.-H., Ye, S.-L., and Ren, Z.-G. (2015). Differentially expressed gene profiles of intrahepatic cholangiocarcinoma, hepatocellular carcinoma, and combined hepatocellular-cholangiocarcinoma by integrated microarray analysis. *Tumour Biol.* 36, 5891–5899. doi: 10.1007/s13277-015-3261-1

Yang, J. T., Bader, B. L., Kreidberg, J. A., Ullman-Culleré, M., Trevithick, J. E., and Hynes, R. O. (1999). Overlapping and independent functions of fibronectin receptor integrins in early mesodermal development. *Dev. Biol.* 215, 264–277. doi: 10.1006/dbio.1999.9451

Yang, Y., Lu, Q., Shao, X., Mo, B., Nie, X., Liu, W., et al. (2018). Development of A three-gene prognostic signature for Hepatitis B virus associated hepatocellular carcinoma based on integrated transcriptomic analysis. *J. Cancer* 9, 1989–2002. doi: 10.7150/jca.23762

Zupan, B., Demsar, J., Kattan, M. W., Beck, J. R., and Bratko, I. (2000). Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artif. Intell. Med.* 20, 59–75. doi: 10.1016/s0933-3657(00)00053-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.