Reference-free transcriptome signatures for prostate cancer prognosis

Ha Tn Nguyen, Haoliang Xue, Virginie Firlej, Yann Ponty, Mélina Gallopin, Daniel Gautheret

A Comparative Analysis of Reference-Free and Conventional Transcriptome Signatures for Prostate Cancer Prognosis

Ha TN Nguyen[1], Haoliang Xue[1], Virginie Firlej[2], Yann Ponty[3], Mélina Gallopin[1], Daniel Gautheret[1*]

September 20, 2020

* Correspondence: daniel.gautheret@universite-paris-saclay.fr

[1] Institute for Integrative Biology of the Cell, UMR 9198, CEA, CNRS, Université Paris-Saclay, Gif-Sur-Yvette, France.

[2] Université Paris Est Creteil, TRePCa, Creteil, France.

[3] LIX UMR 7161, Ecole Polytechnique, Institut Polytechnique de Paris, France.

Abstract

Background RNA-seq data are increasingly used to derive prognostic signatures for cancer outcome prediction. A limitation of current predictors is their reliance on reference gene annotations, which amounts to ignoring large numbers of non-canonical RNAs produced in disease tissues. A recently introduced kind of transcriptome classifier operates entirely in a reference-free manner, relying on k-mers extracted from patient RNA-seq data.

Methods In this paper, we set out to compare conventional and reference-free signatures in risk and relapse prediction of prostate cancer. To compare the two approaches as fairly as possible, we set up a common procedure that takes as input either a k-mer count matrix or a gene expression matrix, extracts a signature and evaluates this signature in an independent dataset.

Results We find that both gene-based and k-mer based classifiers had similarly high performances for risk prediction and a markedly lower performance for relapse prediction. Interestingly, the reference-free signatures included a set of sequences mapping to novel lncRNAs or variable regions of cancer driver genes that were not part of gene-based signatures.

Conclusions Reference-free classifiers are thus a promising strategy for the identification of novel prognostic RNA biomarkers.

Keywords

Reference-free transcriptomic, supervised learning, prostate cancer signature

1 Introduction

The outcome of human cancer can be predicted in part through gene expression profiling[1,2,3]. Outcome prediction is particularly important in prostate cancer (PCa), where distinguishing indolent from aggressive tumors would prevent unnecessary treatment and improve patients' quality of life. However, currently there is no reliable signature of aggressive prostate cancer. Pathologists classify prostate tumor biopsies using scoring systems such as the Gleason score that evaluates tumor differentiation and the Tumour, Node, Metastasis (TNM) grade that evaluates tumor extent and propagation. Gleason, TNM and Prostate-specific antigen (PSA) levels can be combined into a low, medium or high risk status[4]. Several studies used gene expression profiles to derive predictors of Gleason score or risk[5,6,7,8]. Other studies predicted actual clinical progression (tumor recurrence or metastasis) after several years of patient followup. Clinical progression can be evaluated either indirectly through monitoring of PSA levels (BCR=biochemical relapse)[9,10,11,12] or upon direct clinical observation[13,14,15,16]. Gene expression predictors usually take the form a of signature, that is a set of genes or transcripts and associated coefficients of a model that can be used to predict risk or outcome from a patient sample.

Gene expression profiling of prostate biopsies is performed either using DNA microarrays[13,14,15,16] or high throughput RNA sequencing (RNA-seq)[5,6,7,8]. An important advantage of RNA-seq is its ability to identify novel genes or transcripts, which can in principle be incorporated into predictive signatures. However, RNA-seq analysis is usually performed in a "reference-based" fashion, ie. by using RNA-seq reads to quantify a predetermined set of transcripts. This amounts to using RNA-seq in the same way as a microarray that only quantifies a predetermined set of probes. Yet, there is abundant evidence that non-reference RNAs are frequent in disease tissues and may constitute clinically useful biomarkers[17]. Therefore one may expect that prognostic models incorporating non-reference RNAs may carry substantial benefits.

Our group[18,19] and others[20] introduced new k-mer based strategies to analyse RNA-seq data in a "reference-free" manner, that is without mapping sequence reads to a predefined set of genes or transcripts. K-mers are sub-sequences of fixed length which are extracted and quantified from sequence files. When applied to medical RNA-seq datasets using appropriate statistical methods, this strategy identifies any sub-sequence whose increased abundance is associated to a given clinical label. This may include novel splice variants, long non-coding RNAs (lncRNAs) or RNAs from repeated retroelements[18,19] which are ignored by conventional protocols based on reference gene annotations.

Although attractive in principle, k-mer derived prognostic signatures pose two major challenges. First, a single RNA-seq dataset commonly contains tens to hundreds of millions distinct k-mers. Therefore false positive and replicability issues encountered with gene expression profiles[21,22,23,24] are expected to worsen with k-mer count matrices. The second challenge is related to the transfer of a k-mer signatures across independent datasets. Signatures inferred from an initial discovery set are expected to generalize to any independent dataset. In the absence of a unifying gene concept, independent validation requires matching signature k-mers to read sequences from the new dataset. This may cause significant signal loss if sequencing or library preparation technologies differ.

Our main objective here was to compare the characteristics and performances of reference-based and reference-free classifiers for PCa risk and relapse prediction. We built both types of classifiers using the same discovery dataset and assessed their performances in independent datasets using equivalent pipelines and parameters. For the reference-free approach, this required special developments to reduce the number of variables and to transfer expression measures between datasets. We present below a detailed analysis of the relative performances

and sequence contents of the different classifiers and discuss possible future developments to improve performances of models.

2 Materials and Methods

2.1 Data acquisition and outcome labelling

We used tumor samples from the TCGA-PRAD data collection[25] (N=505) for signature discovery and from the ICGC-PRAD data collection[26] (N=284) and from Stelloo et al.[27] (N=91) for independent validation. All three datasets used similar technologies for library preparation (frozen samples, poly(A)+ RNA selection) and Illumina sequencing, however they differed by read-size, read depth, strandedness and use of single or paired ends sequencing (Table 1).

TCGA-PRAD RNA-seq data were retrieved from dbGAP accession phs000178.v9.p8 with permission. ICGC-PRAD-CA RNA-seq data (EGAD00001004424) were downloaded from the European Genome-Phenome Archive (EGA) with permission. The RNA-seq files from the "Porto" cohort[27] were retrieved from GEO, under accession GSE120741. Clinical information was retrieved from Liu et al.[28] for TCGA-PRAD, from Fraser et al.[26] for ICGC-PRAD and from sample metadata of GEO accession GSE120741 for Stello et al.[27].

Table 1: Characteristics of prostate tumor RNA-seq datasets

Study	**RNA-seq library type**	**Reads/sample**	**#Tumor samples**	**Risk**		**Relapse**	
				LR	**HR**	**NO**	**YES**
TCGA-PRAD	Poly(A)+ unstranded 2x50nt	130M	505	134	240	56	58
ICGC-PRAD	Poly(A)+ stranded 2x100nt	313M	284	40	23	7	49
STELLOO	Poly(A)+ stranded 1x65nt	20M	91			43	48

We built predictors for risk and relapse using two-class prediction models. To achieve a clear separation between the two classes, we only focused on high risk (HR) samples versus low risk (LR) samples, ignoring the medium risk, and we focused on relapse prior to a given year and non-relapse after a given year. For this reason, only a fraction of samples could be labelled for a given class in each set. Risk information was not available in the Stelloo dataset and relapse labelling on the ICGC dataset led to a small validation set (only 7 non-relapse samples).

We classified tumor specimens into low-risk and high-risk groups using an adaptation of d'Amico's classification which does not take into account the PSA rate but only the anatomo-pathological data on the basis of Gleason and TNM features as performed previously[19]. Tumors with Gleason score 6/7 (3+4) and TNM grade pT1/2 were classified as low risk. Tumors with Gleason score 8/9 and/or TNM grade pT3b/4 were defined as high-risk. 374 TCGA-PRAD tumors and 63 ICGC-PRAD-CA tumors could be labelled for LR or HR. We could not obtain Gleason/TMN scores for Stelloo et al, hence we did not annotate risk for this cohort.

For relapse analysis, we distinguished patients with biochemical relapse (BCR) and time to BCR $<$ 2yr and patients with no BCR after 5 years or longer, except for Stelloo et al. where only precomputed relapse data was available with cutoffs at 5yr and 10yr, respectively (Table 2). BCR information was obtained from table S1 of Liu et al.[28] for TCGA-PRAD and from table S1 (PFS field) of Fraser et al.[26] for ICGC-PRAD. Precomputed relapse data for Stelloo et al. was taken from SRA accession PRJNA494345.

Table 2: Relapse group definitions

Relapse group	TCGA-PRAD	ICGC-PRAD	STELLOO
Relapse (YES)	PFS = 1 and PFS.time < 2yr	BCR = "Yes" and BCR.time < 2yr	BCR = "Yes" and BCR.time < 5yr
Non relapse (NO)	PFS = 0 and PFS.time > 5yr	BCR = "No" and BCR.time > 5yr	BCR = "No" and BCR.time > 10yr

2.2 A generic framework to infer reference-based and reference-free signatures

Risk and relapse predictors were derived using a combination of feature selection and supervised learning (Figure 1). The predictive model was tuned over a discovery (or training) dataset and its performance was then evaluated on an independent validation (or testing) dataset, to avoid selection bias[29]. The same procedure was used for reference-based and reference-free models, however two extra steps were included to obtain and validate reference-free signatures. First a procedure was implemented to reduce the k-mer matrix using a sequence assembly-like algorithm to merge k-mers into contigs based on their sequence overlap and on the similarity of their count vectors. This step led to a contig count table an order of magnitude smaller than the initial k-mer count table (see results below). Feature selection and model fitting were performed over this contig table. A second adaptation was necessary to validate the reference-free signature in an independent dataset. This required extracting k-mers from both the signature and the sequence files of the independent set, and compute the signature expression in the independent set based on counts of matching k-mers. The pipeline is detailed in Methods. Note that we select features and train a predictive model only on the discovery dataset. The model is then applied to the validation set with no retraining (*i.e.* with the same coefficients) for an unbiased evaluation of the signature.

2.3 Gene and k-mer count matrices

DEkupl-run[18] was used to produce gene and k-mer count matrices for each dataset. DEkupl-run converts FASTQ files to k-mer counts using Jellyfish[30], joins individual sample counts into a single count table and filters out low count k-mers. K-mer size was set to 31, lib_type to unstranded, and parameters min_recurrence and min_recurrence_abundance were set for each dataset as in Supplementary Table S1. K-mer size was set to 31 as commonly adopted for human transcriptome applications[31,18]. Note that contrary to TCGA-PRAD, ICGC-PRAD uses stranded RNA-seq libraries. However we could not use this information as signatures were produced from unstranded libraries. We thus built all k-mer tables in canonical mode, which amounts to consider all libraries as unstranded. Gene expression was computed using Kallisto v0.43.0[31] with Gencode V24 as a reference transcriptome. Gene-level counts were obtained by summing counts for all transcripts of each gene. Gene expression matrices were submitted to the same recurrence filters as k-mer tables to remove low expression genes. After count tables were generated and filtered, the k-mer merging and differential expression analysis module of DEkupl-run were not used. Instead, tables were further processed as explained below.

2.4 Reduction of k-mer matrix via contig extension

k-mer occurence tables were converted into contig occurence tables using an extension procedure similar to that described in Audoux et al.[18]. We define here as contig any sequence produced by merging 1 or more k-mers. Briefly, contigs overlapping by (k-1) to (k-15) nucleotide were iteratively merged into longer contigs till any of the following condition was encountered. In a straightforward case, extension stops when no more overlapping contig is available. Alternatively, extension stops when ambiguity is introduced i.e. when competing extension paths

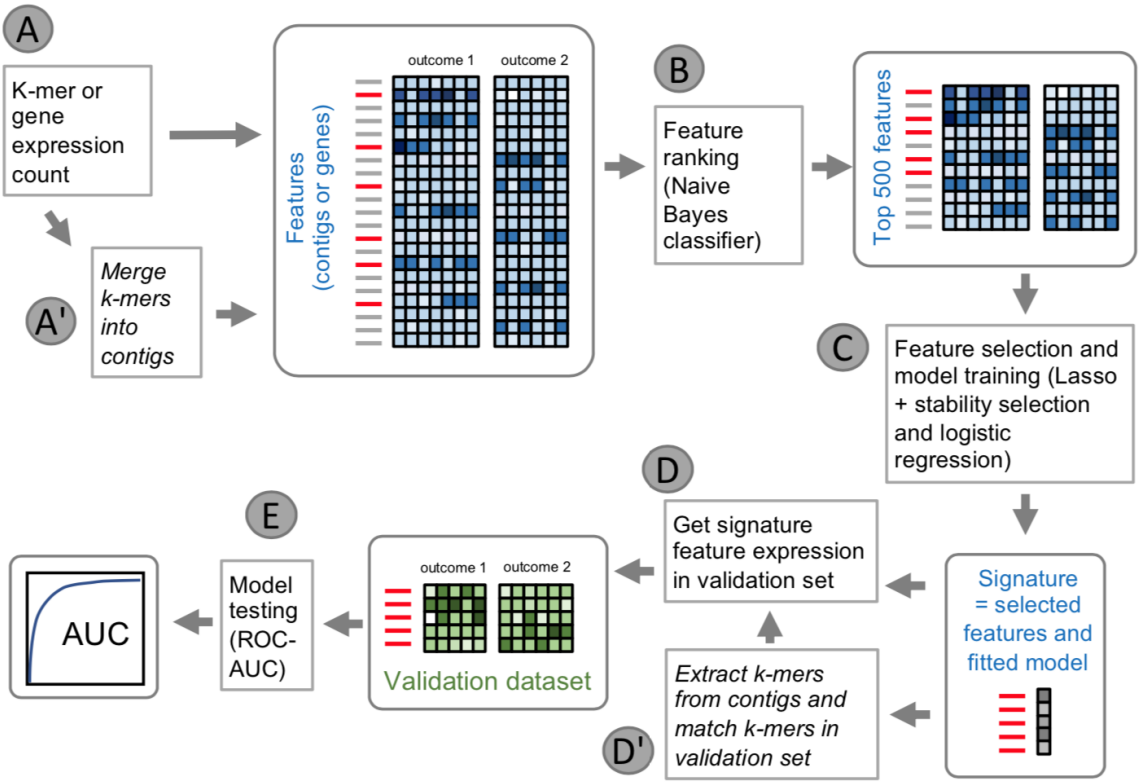

Figure 1: Uniform procedure for signature inference based on k-mer or gene expression. **A**: The discovery matrix is built from normalized k-mer counts or gene expression counts. Samples are labelled by their outcome (risk or relapse) status. Normalization is performed as count per billion for k-mers or count per million for genes. **B**. Features are ranked according to their F1-score computed by cross validation using a Naive Bayes classifier (NB). The top 500 features are retained. **C**. Among the top 500, features are selected using lasso logistic regression combined with stability selection. A logistic regression is tuned on the selected features. **D**. Features from the signature are measured in the count matrix from an independent dataset. **E**. Performance of the signature (selected features + tuned logistic regression) is evaluated using Area Under ROC Curve (AUC) on the validation dataset. To deal with the specificity of k-mer matrices, extra steps A' and D' are introduced: **A'**. the k-mer matrix in converted into a much smaller contig matrix by merging overlapping k-mers with compatible counts. **D'**. k-mers are extracted from the signature contigs and their counts in the validation matrix are aggregated.

occur. Lastly, we applied here an intervention not included in Audoux et al.[18] by considering sample count compatibility between contigs, as shown in Figure 2. Sample count compatibility is measured by the mean value of absolute contrast (MAC) between the counts of the two contigs across all samples, i.e.

$$\mathrm{MAC}(\mathbf{c_1}, \mathbf{c_2}) = mean_{s \in \{samples\}} \left(\left| \frac{c_{1,s} - c_{2,s}}{c_{1,s} + c_{2,s}} \right| \right)$$

where $\mathbf{c_1}$ and $\mathbf{c_2}$ are count vectors of two contigs to be merged, and $c_{1,s}$ and $c_{2,s}$ are counts in sample s from the corresponding count vectors. The extension is rejected if MAC > 0.25. In this way, all contigs are guaranteed to have member k-mers with consistent sample count vectors. After the merging procedure, the new contig's sample count vector is set to the mean of composite k-mer's sample count vectors. The algorithm is implemented in C++ to be published

(https://github.com/i2bc/PCa-gene-based_vs_gene-free/tree/master/KaMRaT)

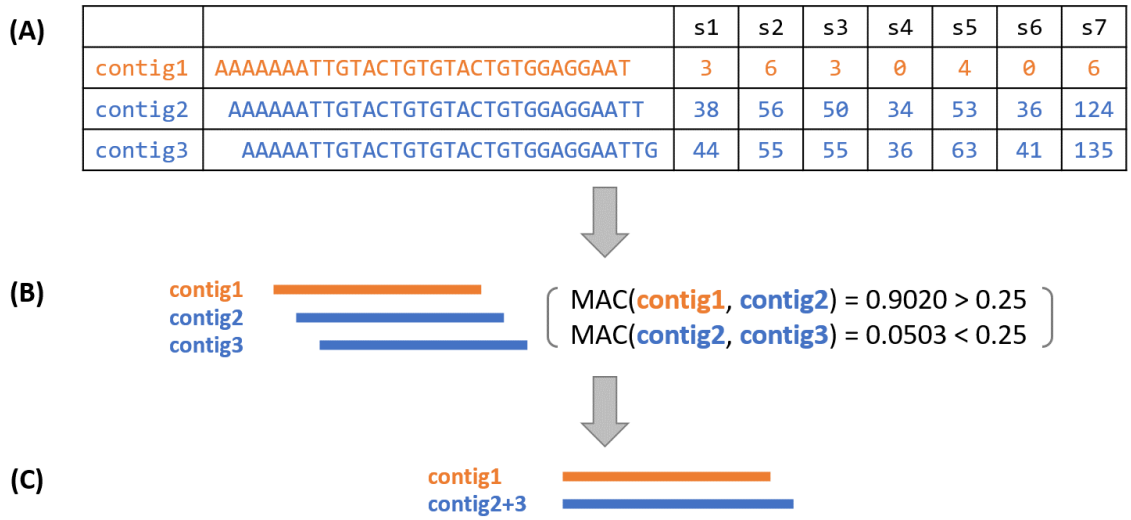

Figure 2: Merging procedure of 3 example contigs: **A**. Count table of contigs in samples. Both pairs $(contig1, contig2)$ and $(contig2, contig3)$ have good overlaps shifting by only one nucleotide, but the sample count vectors of $contig1$ and $contig2$ are not compatible. **B**. Merging intervention considering sample count compatibility between contigs. The mean absolute contrast (MAC) is calculated for each pair, and merging of $(contig1, contig2)$ is rejected due to a MAC value exceeding threshold. **C**. The resulting contigs are the initial $contig1$ and the merged contig from the initial $(contig2, contig3)$ pair.

2.5 Count normalization

To account for differences in sequencing depth among samples, we applied a normalization step on feature counts (genes or contigs) in discovery and validation datasets. Each feature count in a sample is divided by the sum of all feature counts in this sample, then multiplied by a constant base number:

$$e_{f,s} \leftarrow \frac{e_{f,s}}{\sum_{f \in \{features\}} e_{f,s}} \cdot C_b,$$

where $e_{f,s}$ refers to count of feature f in sample s, and C_b is the base constant. For genes, $C_b = 10^6$ resulting in a conventional count per million (CPM) normalization, while for contigs, we used $C_b = 10^9$, or count per billion (CPB). For contigs, normalization is applied on the contig count table produced after contig extension and for genes it is applied on the recurrence filtered gene expression matrix.

2.6 Univariate features ranking

Given the limited number of samples, it was necessary to reduce the number of features (genes or contigs) in the dataset. We discarded irrelevant features to focus on a subset of 500 top candidates for subsequent feature selection. To rank features, we performed prediction of status (risk/relapse) using a Naïve Bayes classifier on each independent feature, after log transformation of the normalized counts (after adding an offset 1 to avoid numerical problem). To assess the quality of the prediction, we computed the average f_1 score by 5-fold cross validation ($f_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$, where precision $= TP/(TP + FP)$ and recall $= TP/(TP + FN)$ and FP, TP, FN are respectively the False Positive, True Positive and False Negative). In cases

where 5-fold cross-validation returned an undefined value, f_1 score was set to 0 (the worst). The average f_1 score was used to rank features. The Naïve Bayes classifier implementation was taken from the `MLPack` library[32]. The C++ code to perform feature ranking is available at `https://github.com/i2bc/PCa-gene-based_vs_gene-free/tree/master/KaMRaT`.

2.7 Feature selection, model fitting and predictor evaluation

To select a subset of non-correlated features (genes or contigs) among the top 500 candidates, we performed penalized logistic regression using the implementation from the `glmnet` R package[33]. We implemented stability selection[34]: only features selected with a frequency of being selected above 0.5 upon 2000 resamples of the input dataset were retained. To evaluate the performance of the selected features on the discovery (training dataset), we fitted a logistic regression and computed the area under the ROC curve (AUC) using a 10-fold cross validation scheme, repeated 20 times, as implemented in the `caret` package[35]. To assess the performance of the signature on the external validation datasets, we fitted a logistic regression on the whole discovery dataset and applied the predictor to the validation datasets. In the reference-free approach, some features present in the signature were not found in the validation (see below). In this case, the coefficient of the logistic regression corresponding to missing features were set to zero. Signature contigs were annotated through BLAST alignment *vs.* Gencode V34 transcripts. HGNC symbols for signature genes were obtained from the Ensembl EnsDb.Hsapiens.v79 R package[36]. R scripts to perform the feature selection, model fitting and evaluation on the discovery and validation sets are available at: `https://github.com/i2bc/PCa-gene-based_vs_gene-free`.

2.8 Matching signature contigs in the validation cohort

To measure contig expression in the validation cohort we implemented the procedure schematized in Figure 3. The procedure comprises two main steps: (1) all k-mers from signature contigs were extracted and identified in the k-mer count matrix generated from the validation cohort and (2) the resulting sub-matrix was used to estimate each contig's expression in the validation cohort, measured for each sample as the median of extracted k-mer counts. Step 1 is implemented in C++ at: `https://github.com/i2bc/PCa-gene-based_vs_gene-free/tree/master/kmerFilter`, step 2 is implemented in R at: `https://github.com/i2bc/PCa-gene-based_vs_gene-free/blob/master/infer_gene-free_risk_signature.R`.

2.9 Data sharing

Data sharing not applicable – no new data generated.

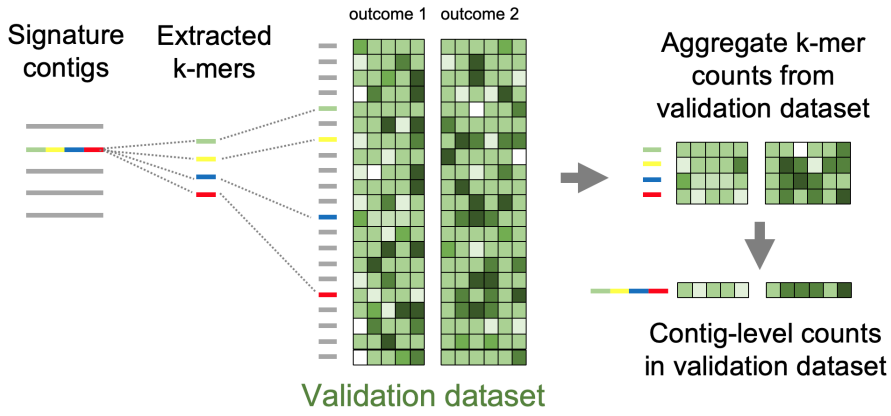

Figure 3: Procedure for inferring signature contig expression in an independent validation dataset. The colored contig from the signature is quantified in the validation cohort by extracting all its constituent k-mers and retrieving the corresponding k-mer counts from validation k-mer count matrix. The count vector of the contig in each sample of the validation dataset is taken as the median of counts for k-mers in this sample.

3 Results

3.1 A reference-free risk signature for prostate cancer

We first applied the gene-free and gene-based signature discovery procedures detailed in section 2.2 to infer PCa risk signatures. The k-mer table for 374 TCGA-PRAD risk-labelled samples had 94M k-mers after low count filtering. The merging step reduced it to 5.2M contigs, i.e. achieving a considerable 18-fold reduction in size (Table 3). Contig sizes (mean=49nt, median=34nt, Table 4) were small relatively to a typical human RNA, which is characteristic of the adopted contig extension procedure[18] (see section 2.4).

Table 3: Result of filtering procedure on the k-mer and gene matrices for risk analysis

	Initial matrix	Low expression filter	k-mer merging	Naive Bayes ranking	Feature Selection by Lasso LR	Validation
k-mers or contigs	(not generated)	94,539,338 k-mers	5,234,940 contigs	500 contigs	26 contigs (1,444 k-mers)	21 contigs (1,404 k-mers)
genes	60,554	38,382	NA	500	14	14

Table 4: Contig sizes (Risk model)

	After k-mer merging	After Naive Bayes ranking
mean contig size (nt)	49.1	189
median contig size (nt)	34	61

The 5.2M contig matrix and the 38k gene expression matrix were submitted to screening using univariate Naive Bayes classification and the top scoring 500 features were retained for feature selection (section 2.6). Interestingly, the 500 top scoring contigs were significantly longer than prior to selection (median 61nt vs. 34nt, Table 4), suggesting the procedure tended to eliminate spurious short contigs.

Finally, Lasso logistic regression produced a reference-free signature of 26 contigs and a reference-based signature of 14 genes (Table 3, Figure 4, Suppl. Figure S5). Ten-fold cross validation performances of both signatures were very high on the discovery dataset (0.90 and 0.93 for genes and k-mers, respectively) (Table 5), which is an over-estimated performance since features here were tested on the same dataset used to select features[29].

Figure 4.A shows the 26 contigs in the reference-free risk signature and their abundance distribution in LR and HR samples. 24/26 contigs mapped Gencode transcripts from 21 unique genes (Supplementary file 1). Eleven of the 21 genes were also found in a list 180 genes compiled from published PCa outcome signatures (Supplementary file 2), which is a highly significant enrichment (P-value = 7.9e-9, Fisher's exact test), especially when considering that no gene information was used to infer our signature. The gene and contig signatures involved five shared genes: MYBPC1, ASPN, SLC22A3, SRD5A2 and CD38 (Supplementary file 2, Figure S6.A, Figure 4.A). The first four genes are part of published prostate risk signatures. CD38 is particular in that it is the most downregulated in both signatures and it is not part of previous signatures. However, downregulation of this gene has been associated with poor outcome in prostate cancer[37], supporting its status as a high risk biomarker. Risk signature contigs mapped at least five other genes with established driver roles in PCa or other cancers: CAMK2N1[38], COL1A1[39], GTSE1[40] and PTPRN2[41], supporting the relevance of these sequence contigs in PCa etiology.

Of the two contigs that did not map any Gencode transcript, one aligned to an intron of GMNN (ctg_20), a gene also mapped by an exonic contig, the other an intron of LDLRAD4

(ctg_23). Contig ctg_23 corresponds to a 1.29 kb spliced transcript located between exons 4 and 5 of LDLRAD4 and is strongly upregulated in HR samples, as displayed in the Integrative Genomics Viewer (IGV)[42] in Supplementary Figure S1. Although ctg_23 partly maps short annotated LDLRAD4 isoforms, its expression seems unrelated to that of the longer LDLRAD4 transcripts whose coverage in flanking exons is 4-6 times lower than ctg_23 (Supplementary Figure S2.) Therefore ctg_23 likely comes from an independent lncRNA. The host gene LDLRAD4 is a negative regulator of TGF-beta signaling with roles in proliferation and apoptosis and was recently associated to negative outcome in other tumor types[43][44]. Lastly, one contig (ctg_11, EFNA2) was probably misassigned to the EFNA2 gene since it maps to a highly expressed discrete area just 3' of EFNA2 while EFNA2 seems silent. Thus ctg_11 probably comes from an independent lncRNA as well (Supplementary Figure S3.).

To assess the replicability of risk signatures, we evaluated their performance in the ICGC-PRAD independent dataset. To this aim, we developed a specific procedure to estimate the expression of an arbitrary sequence contig across datasets using matched k-mers (see Methods). The 26 contigs represented 1444 k-mers, of which 97% were present in the ICGC-PRAD validation dataset. Overall 5 contigs (SFRP4, GTSE1, COL3A1, COL1A1.a, COL1A1.c) could not be quantified in the validation set due to lack of supporting k-mers (see Table 3 and Figure 4B). In spite of this, the reference-free signature had similar performance in the validation set as the reference-based signature (0.85 and 0.86 respectively, Table 5), although the later did not sustain any loss when transferred to the independent cohort (Table 3). High validation AUCs indicate a strong replicability of both the reference-free and reference-based risk signatures.

Table 5: Signature performances for risk prediction

	AUC - risk prediction	
	TCGA Cross-validation	**ICGC Independent dataset**
Reference-free	0.93 +/- 0.04	0.85
Reference-based	0.90 +/- 0.05	0.86

3.2 Relapse signatures contain key PCa drivers

Application of the gene-free and gene-based signature discovery procedures (section 2.2) to relapse analysis produced a 14-contig reference-free signature and a 10-gene reference-based signature (Supplementary File 2, Figure 5A, Supplementary Figure S6 A). The reference-free signature was populated by obvious PCa drivers. Strikingly, 3 contigs matched KLK2, AR and KLK3, which are among the most important genes in PCa onset and progression[45], the androgen receptor (AR) and two of its main targets, KLK2 and KLK3, the later encoding the PSA protein (Figure 5A). Another contig matched SPDEF, a gene whose loss is associated to PCa metastasis[46].

Contigs matching KLK2 and AR were overexpressed 23-fold and 7-fold, respectively in relapsed patients while the contig matching KLK3 was depleted 1.8 fold. The AR contig matches exon 1 of AR and contains an non-templated poly-A end but no visible polyadenylation signal. The KLK2 contig is intronic and harbours a common SNP (rs62113074). The KLK3 contig is located in a distal part of the 3' UTR region present only in longer isoforms of KLK3. Its lower expression in relapsed patients was unexpected as low expression of PSA is usually associated to a lower risk. It is possible though that only this longer isoform is depleted in relapsing samples. The expression boxplot shows the KLK2 contig occurs only in a few outlier patients while the AR and KLK3 contigs are common (Figure 5A). The contig matching SPDEF is a special variant of the 3' exon including two nonsynonymous SNPs. The SPDEF gene as a whole was highly expressed in both relapse and non-relapse samples but the contig expression

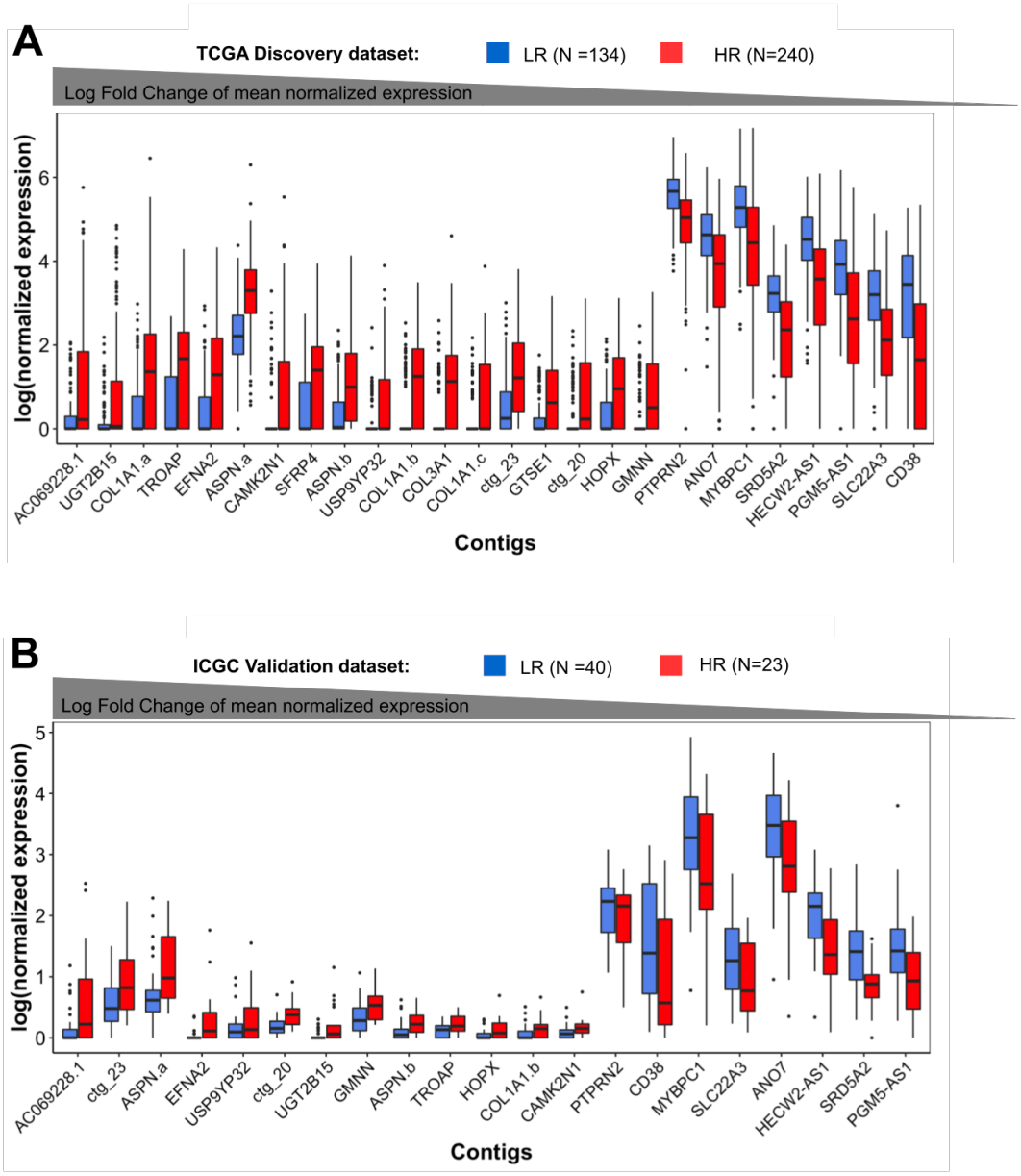

Figure 4: Expression of risk signature contigs in LR and HR samples. A: TCGA-PRAD discovery cohort. B: ICGC-PRAD validation cohort

was twice lower in average in relapse samples. Two contigs matched no known transcript: ctg_7 is a low complexity sequence of unknown origin and ctg_1 matches an intron of RPL9.

The contig matching lncRNA AC069228.1 also raised our attention since AC069228.1 is the only gene mapped by contigs in both relapse and risk signatures. The AC069228.1 lncRNA is antisense of PPFIA2, a protein tyrosine phosphatase that is itself an alleged urine biomarker of PCa[47]. The contigs from risk and relapse models match different regions of AC069228.1 (Figure S4). One is spliced, the other is a continuous 864 bp segment of a long exon. In both cases, a negative outcome (HR or relapse) is associated to a clearly higher expression of the contig, while the antisense gene PPFIA2 does not appear to follow the same trend (Figure S4).

Of note, the 10 genes in the reference-based signature were also clearly PCa-related: one was the major PCa biomarker PCA3[48] and 5 others (DDC, RRM2, FEV, TSPAN1, HMGCS2) are involved in PCa etiology[49,50,51,52,53]. Therefore both gene-based and gene-free relapse signatures were significant in terms of PCa related functions of their component genes or contigs.

3.3 Relapse signatures do not accurately classify independent cohorts

Table 6: Signatures performances for relapse prediction

Method	**AUC - relapse prediction**		
	TCGA Cross-validation	**ICGC Independent dataset**	**STELLOO Independent dataset**
Reference-free	0.93 +/- 0.1	0.51	0.62
Reference-based	0.84 +/- 0.11	0.66	0.59

Contrary to the risk signatures, relapse signatures showed little overlap with each other and with published PCa signatures (Supplementary File 2). Only PCA3 and KLK2 were found in prior signatures[16,54] and the only gene found shared between relapse and risk signatures in this study was AC069228.1. The poor overlap in this study was not unexpected as the discovery samples for risk and relapse information were quite disjointed and not always consistent: for instance only 25% of the high risk samples were labelled for relapse and 28% of these did not relapse. Conversely, 51% of non-relapse patients were labelled as HR. Therefore risk and relapse classifiers were trained to recognize quite different phenotypes.

As in the risk model, both reference-based and reference-free signatures had excellent cross-validation performance on the discovery set (AUC of 0.84 and 0.93 respectively, Table 6). However this should again be considered as an overly optimistic estimation due to the experimental design. Indeed, performances of both relapse signatures on the ICGC-PRAD and Stelloo validation sets were much lower (AUC 0.51 to 0.66), bordering randomness and confirming overfitting of the trained signatures. The reference-based model performed slightly better over ICGC-PRAD, and the reference-free model was slightly better over the Stelloo dataset (Table 6). Furthermore, several genes and contigs in the discovery signatures had inconsistent expression variations in the validation datasets (Fig 5B,C, Supplementary Figure S6B,C, Supplementary File 3). Overall two genes from the reference-based signature (ALB and CTD-2228K2.7) and 5 contigs from the reference-free signature (KLK2, AC069228.1, PDLIM5, RTN4, ctg_1) changed logFC sign between the discovery and either validation cohort. This problem, which was not observed in risk models, underlines the poor replicability of the relapse signatures, whether or not reference-free.

Low replicability of the relapse model may be caused in part by weaknesses in validation datasets: the ICGC dataset had only 7 samples labelled for non-relapse (Table 1) and the Stello dataset had very low coverage (Table 1) which caused considerable loss when computing contig expression. Only three of the 14 signature contigs (AC069228.1, KLK2 and KLK3) could be quantified in the Stelloo dataset (Table 7, Figure 5.C). Yet, we note that in spite of this loss

the reference-free model still outperformed the reference-based model on this set (AUC of 0.62 vs. 0.59, Table 6). Other limitations of the relapse model are addressed in the discussion.

Table 7: Result of filtering procedure on the k-mer and gene matrices for relapse analysis

	Initial matrix	**Low expression filter**	**k-mer merging**	**Naive Bayes ranking**	**Feature Selection by Lasso LR**	**Validation in ICGC**	**Validation in Stelloo**
k-mers or contigs	(not generated)	97,731,857	6,184,108 contigs	500 contigs	14 contigs (219 k-mers)	12 contigs (215 k-mers)	3 contigs (71 k-mers)
genes	60,554	36,006	NA	500	10	10	10

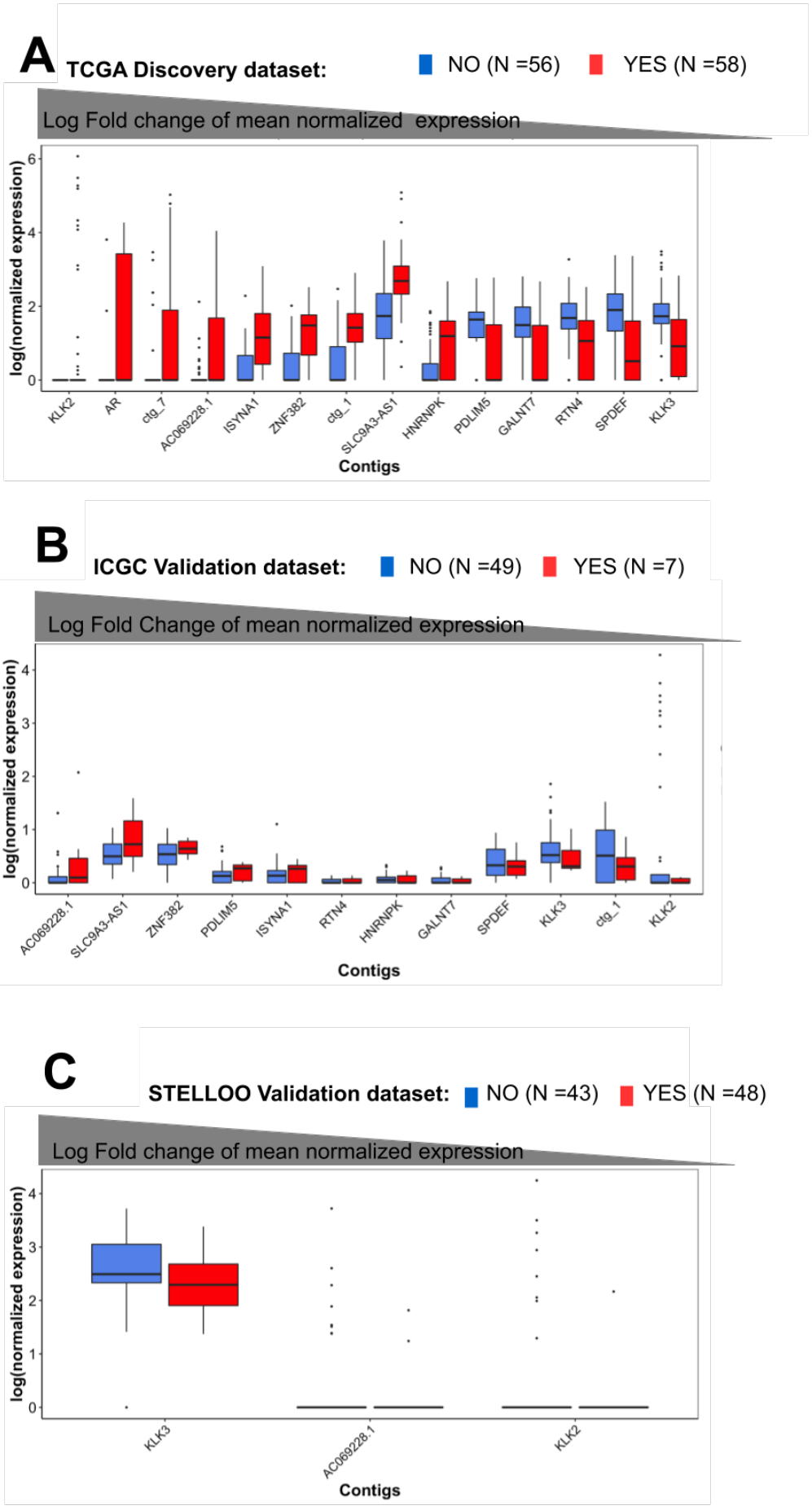

Figure 5: Expression of relapse signature contigs in relapse/non relapse samples. A: TCGA-PRAD discovery cohort. B: ICGC-PRAD validation cohort. C: Stelloo validation cohort.

4 Discussion

4.1 Properties of reference-free signatures

We evaluated here a method for building transcriptome classifiers that are totally reference-free, *i.e.* that do not require prior knowledge of genes or genome. The major interest of this approach lies in its ability to discover and incorporate in models previously unknown RNA biomarkers. Multiple examples exist of such disease-specific RNAs produced by genome alterations or deficient RNA processing and we hypothetized their inclusion in predictive models would be beneficial[17]. Applying a reference-free strategy to PCa outcome prediction, we obtained signatures made of short RNA contigs (median size 33 to 45 nt). These contigs are not full transcript models as can be produced by usual *denovo* assembly procedures. Instead, they often match SNPs or splice variants thus describing specific genetic or transcriptional events enriched in a patient group. Our strategy thus identifies RNA variations independently instead of lumping them into a full transcript model. Yet, the mapped genes were highly relevant to PCa etiology and included known cancer drivers LDLRAD4, GMNN, COL1A1, CD38, PTPRN2, GTSE1 and CAMK2N1 in the risk signature and KLK2, AR, KLK3, SPDEF in the relapse signature. Furthermore the risk signature comprised contigs matching two potential novel lncRNAs, located within LDLRAD4 and immediately downstream of EFNA2.

To our knowledge the only other software using a reference-free approach for inferring predictive signatures is Gecko[20]. Gecko uses machine learning (genetic algorithm) directly on the k-mer count matrix while we first reduce the matrix by grouping k-mers into contigs, before classification and machine learning. This enabled us to produce a signature composed of sequences larger than k, hence easier to interpret and quantify in an independent dataset.

Transferring a reference-free model to a new dataset is challenging. This requires that important features, such as SNPs, are precisely evaluated in the independent dataset. To this aim, we transferred signatures between datasets based on exact k-mer matches. As k-mer contents vary a lot between library preparation protocols, we expected this strategy to show poor sensitivity when discovery and validation datasets differed substantially. Indeed, transfer of signatures trained on the TCGA-PRAD dataset to the low coverage Stelloo dataset caused the loss of a majority of contigs. However, in this particular case, the remaining contigs were sufficient to maintain a prediction performance at the same level as that of the gene-based signature.

4.2 Performances and generalization issues

To compare the reference-free and reference-based strategies, a common evaluation framework was adopted. For both risk and relapse predictions, performances of the reference-free classifiers were on a par with that of reference-based classifiers. However while risk signatures showed satisfying reproducibility, relapse signatures performed poorly in independent datasets.

On possible reason for the low performance of relapse models is our grouping of patients in discrete relapse and non relapse categories as done in other studies[9,13,15,16]. This allowed us to address relapse prediction using the same logistic regression method as for risk, however this meant valuable patient information was left unused. A more accurate prediction of relapse may be achieved using survival models[55,54,10,14,12]. Adaptation of survival analysis tools to large k-mer matrices require additional developments that are certainly worth considering in the future.

A more general concern with relapse analysis is related to difficulty of predicting an outcome occurring several years after a sample is biopsied and analyzed. There might just be too little information available in the training data to infer a reliable classifier, a problem that would be

independent of the use of contigs or genes. However, both gene-level and contig-level signatures were highly enriched in PCa driver genes, which suggests information about tumor progression was indeed present in the primary tumor biopsy. The key problem with relapse analysis was more likely related to sample heterogeneity. The diversity of relapse mechanisms was not properly represented in a training set of 100 patients as we used here. Patient stratification have been proposed to deal with sample heterogeneity in omics data[56,57]. Adaptations of these solutions to large k-mers matrices will also be considered in the future.

5 Conclusion

For prediction of PCa risk and relapse, reference-free classifiers did not significantly outperform reference-based classifiers, however they incorporated a distinct set of RNA sequences including unannotated RNAs and novel variants of annotated RNAs. It is likely than with other diseases and datasets, novel biomarkers will be identified with an even greater impact on prediction performance. The reference-free approach will be of particular interest in problems where unknown RNAs are expected to play an important role, such as when studying rare diseases, poorly studied tissue types or when analysing dual human-pathogen RNA-seq samples. Our strategy also permits to infer efficient transcriptome classifiers in species lacking an accurate genome or transcriptome reference.

6 Acknowledgements

This work was funded in part by Agence Nationale de la Recherche grant ANR-18-CE45-0020. Conflicts of interests: none declared.

References

[1] Perou C. M., Sorlie T., Eisen M. B., et al. Molecular portraits of human breast tumours *Nature.* 2000;406:747–752.

[2] Singh Dinesh, Febbo Phillip G., Ross Kenneth, et al. Gene expression correlates of clinical prostate cancer behavior *Cancer Cell.* 2002;1:203–209.

[3] van 't Veer Laura J., Dai Hongyue, Vijver Marc J., et al. Gene expression profiling predicts clinical outcome of breast cancer *Nature.* 2002;415:530–536.

[4] D'Amico Anthony V, Whittington Richard, Malkowicz S Bruce, et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer *Jama.* 1998;280:969–974.

[5] Bibikova Marina, Chudin Eugene, Arsanjani Amir, et al. Expression signatures that correlated with Gleason score and relapse in prostate cancer *Genomics.* 2007;89:666–672.

[6] Penney Kathryn L, Sinnott Jennifer A, Fall Katja, et al. mRNA expression signature of Gleason grade predicts lethal prostate cancer *Journal of Clinical Oncology.* 2011;29:2391.

[7] Sinnott Jennifer A, Peisch Sam F, Tyekucheva Svitlana, et al. Prognostic utility of a new mRNA expression signature of Gleason score *Clinical Cancer Research.* 2017;23:81–87.

[8] Jhun Min A, Geybels Milan S, Wright Jonathan L, et al. Gene expression signature of Gleason score is associated with prostate cancer outcomes in a radical prostatectomy cohort *Oncotarget.* 2017;8:43035.

[9] Latil Alain, Bièche Ivan, Chêne Laurent, et al. Gene Expression Profiling in Clinically Localized Prostate Cancer: A Four-Gene Expression Model Predicts Clinical Behavior *Clinical Cancer Research.* 2003;9:5477–5485.

[10] Long Q., Xu J., Osunkoya A. O., et al. Global Transcriptome Analysis of Formalin-Fixed Prostate Cancer Specimens Identifies Biomarkers of Disease Recurrence *Cancer Research.* 2014;74:3228–3237.

[11] Ren Shancheng, Wei Gong-Hong, Liu Dongbing, et al. Whole-genome and transcriptome sequencing of prostate cancer identify new genetic alterations driving disease progression *European urology.* 2018;73:322–339.

[12] Sinha Ankit, Huang Vincent, Livingstone Julie, et al. The proteogenomic landscape of curable prostate cancer *Cancer Cell.* 2019;35:414–427.

[13] Erho Nicholas, Crisan Anamaria, Vergara Ismael A, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy *PloS one.* 2013;8:e66855.

[14] Karnes R. Jeffrey, Bergstralh Eric J., Davicioni Elai, et al. Validation of a Genomic Classifier that Predicts Metastasis Following Radical Prostatectomy in an At Risk Patient Population *Journal of Urology.* 2013;190:2047–2053.

[15] Klein Eric A., Yousefi Kasra, Haddad Zaid, et al. A genomic classifier improves prediction of metastatic disease within 5 years after surgery in node-negative high-risk prostate cancer patients managed by radical prostatectomy without adjuvant therapy *European Urology.* 2015;67:778–786.

[16] Shahabi Ahva, Lewinger Juan Pablo, Ren Jie, et al. Novel Gene Expression Signature Predictive of Clinical Recurrence After Radical Prostatectomy in Early Stage Prostate Cancer Patients *The Prostate.* 2016;76:1239–1256.

[17] Morillon Antonin, Gautheret Daniel. Bridging the gap between reference and real transcriptomes *Genome biology.* 2019;20:1–7.

[18] Audoux Jérôme, Philippe Nicolas, Chikhi Rayan, et al. DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition *Genome Biology.* 2017;18:243.

[19] Pinskaya Marina, Saci Zohra, Gallopin Mélina, et al. Reference-free transcriptome exploration reveals novel RNAs for prostate cancer diagnosis *Life Science Alliance.* 2019;2:1–12.

[20] Thomas Aubin, Barriere Sylvain, Broseus Lucile, et al. GECKO is a genetic algorithm to classify and explore high throughput sequencing data *Communications Biology.* 2019;2:222.

[21] Michiels Stefan, Koscielny Serge, Hill Catherine. Prediction of cancer outcome with microarrays: a multiple random validation strategy *The Lancet.* 2005;365:488–492.

[22] Ein-Dor L., Zuk Or, Domany Eytan. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer *Proceedings of the National Academy of Sciences.* 2006;103:5923–5928.

[23] Michiels S, Koscielny S, Hill C. Interpretation of microarray data in cancer *British Journal of Cancer.* 2007;96:1155–1158.

[24] Venet David, Dumont Jacques E., Detours Vincent. Most random gene expression signatures are significantly associated with breast cancer outcome *PLoS computational biology.* 2011;7:e1002240.

[25] Abeshouse Adam, Ahn Jaeil, Akbani Rehan, et al. The molecular taxonomy of primary prostate cancer *Cell.* 2015;163:1011–1025.

[26] Fraser Michael, Sabelnykova Veronica Y, Yamaguchi Takafumi N, et al. Genomic hallmarks of localized, non-indolent prostate cancer *Nature.* 2017;541:359–364.

[27] Stelloo Suzan, Nevedomskaya Ekaterina, Kim Yongsoo, et al. Integrative epigenetic taxonomy of primary prostate cancer *Nature communications.* 2018;9:1–12.

[28] Liu Jianfang, Lichtenberg Tara, Hoadley Katherine A, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics *Cell.* 2018;173:400–416.

[29] Ambroise Christophe, McLachlan Geoffrey J.. Selection bias in gene extraction on the basis of microarray gene-expression data *Proceedings of the National Academy of Sciences.* 2002;99:6562–6566.

[30] Marçais Guillaume, Kingsford Carl. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers *Bioinformatics.* 2011;27:764–770.

[31] Bray Nicolas L, Pimentel Harold, Melsted Páll, Pachter Lior. Near-optimal probabilistic RNA-seq quantification *Nature biotechnology.* 2016;34:525–527.

[32] Curtin Ryan R., Edel Marcus, Lozhnikov Mikhail, Mentekidis Yannis, Ghaisas Sumedh, Zhang Shangtong. mlpack 3: a fast, flexible machine learning library *Journal of Open Source Software.* 2018;3:726.

[33] Friedman Jerome, Hastie Trevor, Tibshirani Robert. Regularization Paths for Generalized Linear Models via Coordinate Descent *Journal of Statistical Software.* 2010;33:1–22.

[34] Meinshausen Nicolai, Bühlmann Peter. Stability selection *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2010;72:417–473.

[35] Kuhn Max. Building Predictive Models in R Using the caret Package *Journal of Statistical Software, Articles.* 2008;28:1–26.

[36] Rainer Johannes. *EnsDb.Hsapiens.v79: Ensembl based annotation package* 2017. R package version 2.99.0.

[37] Liu Xian, Grogan Tristan R, Hieronymus Haley, et al. Low CD38 identifies progenitor-like inflammation-associated luminal cells that can initiate human prostate cancer and predict poor outcome *Cell reports.* 2016;17:2596–2606.

[38] Wang Tao, Liu Zhuo, Guo Shuiming, et al. The tumor suppressive role of CAMK2N1 in castration-resistant prostate cancer *Oncotarget.* 2014;5:3611.

[39] Liu Jing, Shen Jia-Xin, Wu Hua-Tao, et al. Collagen 1A1 (COL1A1) promotes metastasis of breast cancer and is a potential therapeutic target *Discovery medicine.* 2018;25:211–223.

[40] Wu Xiaojuan, Wang Hongbo, Lian Yifan, et al. GTSE1 promotes cell migration and invasion by regulating EMT in hepatocellular carcinoma and is associated with poor prognosis *Scientific reports.* 2017;7:1–12.

[41] Chen Chun-Liang, Mahalingam Devalingam, Osmulski Pawel, et al. Single-cell analysis of circulating tumor cells identifies cumulative expression patterns of EMT-related genes in metastatic prostate cancer *The Prostate.* 2013;73:813–826.

[42] Robinson James T, Thorvaldsdóttir Helga, Winckler Wendy, et al. Integrative genomics viewer. *Nature biotechnology.* 2011;29:24–6.

[43] Xie Wei, Xiao He, Luo Jia, et al. Identification of low-density lipoprotein receptor class A domain containing 4 (LDLRAD4) as a prognostic indicator in primary gastrointestinal stromal tumors *Current Problems in Cancer.* 2020:100593.

[44] Mo Shaobo, Zhang Long, Dai Weixing, et al. Antisense lncRNA LDLRAD4-AS1 promotes metastasis by decreasing the expression of LDLRAD4 and predicts a poor prognosis in colorectal cancer *Cell death & disease.* 2020;11:1–16.

[45] Chen Charlie D, Welsbie Derek S, Tran Chris, et al. Molecular determinants of resistance to antiandrogen therapy *Nature medicine.* 2004;10:33–39.

[46] Chen Wei-Yu, Tsai Yuan-Chin, Yeh Hsiu-Lien, et al. Loss of SPDEF and gain of TGFBI activity after androgen deprivation therapy promote EMT and bone metastasis of prostate cancer *Science Signaling.* 2017;10:eaam6826.

[47] Leyten Gisele HJM, Hessels Daphne, Smit Frank P, Jannink Sander A, Jong Hans, Melchers Willem JG. Identification of a candidate gene panel for the early diagnosis of prostate cancer *Clinical Cancer Research.* 2015;21:3061–3070.

[48] Bussemakers Marion J.G., Bokhoven A, Verhaegh Gerald W., et al. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer research.* 1999;59:5975–9.

[49] Koutalellis Georgios, Stravodimos Konstantinos, Avgeris Margaritis, et al. L-dopa decarboxylase (DDC) gene expression is related to outcome in patients with prostate cancer *BJU international.* 2012;110:E267–E273.

[50] Mazzu Ying Z, Armenia Joshua, Chakraborty Goutam, et al. A novel mechanism driving poor-prognosis prostate cancer: overexpression of the DNA repair gene, ribonucleotide reductase small subunit M2 (RRM2) *Clinical Cancer Research.* 2019;25:4480–4492.

[51] Zhong Wei-De, Liang Yu-Xiang, Liang Ying-Ke, et al. Tumor Suppressor Role and Clinical Implication of the Fifth Ewing Variant (FEV) Gene, an ETS Family Gene, in Prostate Cancer *Prostate Cancer (April 15, 2019).* 2019.

[52] Munkley Jennifer, McClurg Urszula L, Livermore Karen E, et al. The cancer-associated cell migration protein TSPAN1 is under control of androgens and its upregulation increases prostate cancer cell migration *Scientific reports.* 2017;7:1–11.

[53] Wan Song, Xi Ming, Zhao Hai-Bo, et al. HMGCS2 functions as a tumor suppressor and has a prognostic impact in prostate cancer *Pathology-Research and Practice.* 2019;215:152464.

[54] Klein Eric A., Cooperberg Matthew R., Magi-Galluzzi Cristina, et al. A 17-gene assay to predict prostate cancer aggressiveness in the context of gleason grade heterogeneity, tumor multifocality, and biopsy undersampling *European Urology.* 2014;66:550–560.

[55] Witten Daniela M, Tibshirani Robert. Survival analysis with high-dimensional covariates *Statistical Methods in Medical Research.* 2010;19:29–51.

[56] Ronde Jorma J., Rigaill Guillem, Rottenberg Sven, Rodenhuis Sjoerd, Wessels Lodewyk F. A.. Identifying subgroup markers in heterogeneous populations *Nucleic Acids Research.* 2013;41:e200–e200.

[57] Campos-Laborie F J, Risueño A, Ortiz-Estévez M, et al. DECO: decompose heterogeneous population cohorts for patient stratification and discovery of sample biomarkers using omic data profiling *Bioinformatics.* 2019;35:3651-3662.

Supplementary figures and tables

Supplementary file 1: Contig sequences and mapping locations in the risk and relapse signatures.

Supplementary file 2: Published PCa risk and relapse signatures. Genes in common between published and this publication's signatures.

Supplementary file 3: Contents and expression characteristics of all signatures in the discovery and validation datasets.

Table S1. Filtering parameters for count tables

	Analysis	min_recurrence	min_recurrence_abundance
TCGA-PRAD	Risk	3	10
ICGC-PRAD		5	5
TCGA-PRAD	Relapse	3	5
ICGC-PRAD		4	2
STELLOO		3	5

Figure S1. IGV view of RNA-seq reads from the TCGA-PRAD discovery set aligned at the genomic location of risk signature contig ctg_23 (red box). This contig is located in an intron of LDLRAD4. Frames LR and HR show reads sampled from all samples in the LR and HR subsets, respectively, at identical depth for each. Blue boxes and lines in the bottom frame correspond to Gencode annotations of LDLRAD4 transcript isoforms (thick lines: exons, thin lines with arrows: introns).

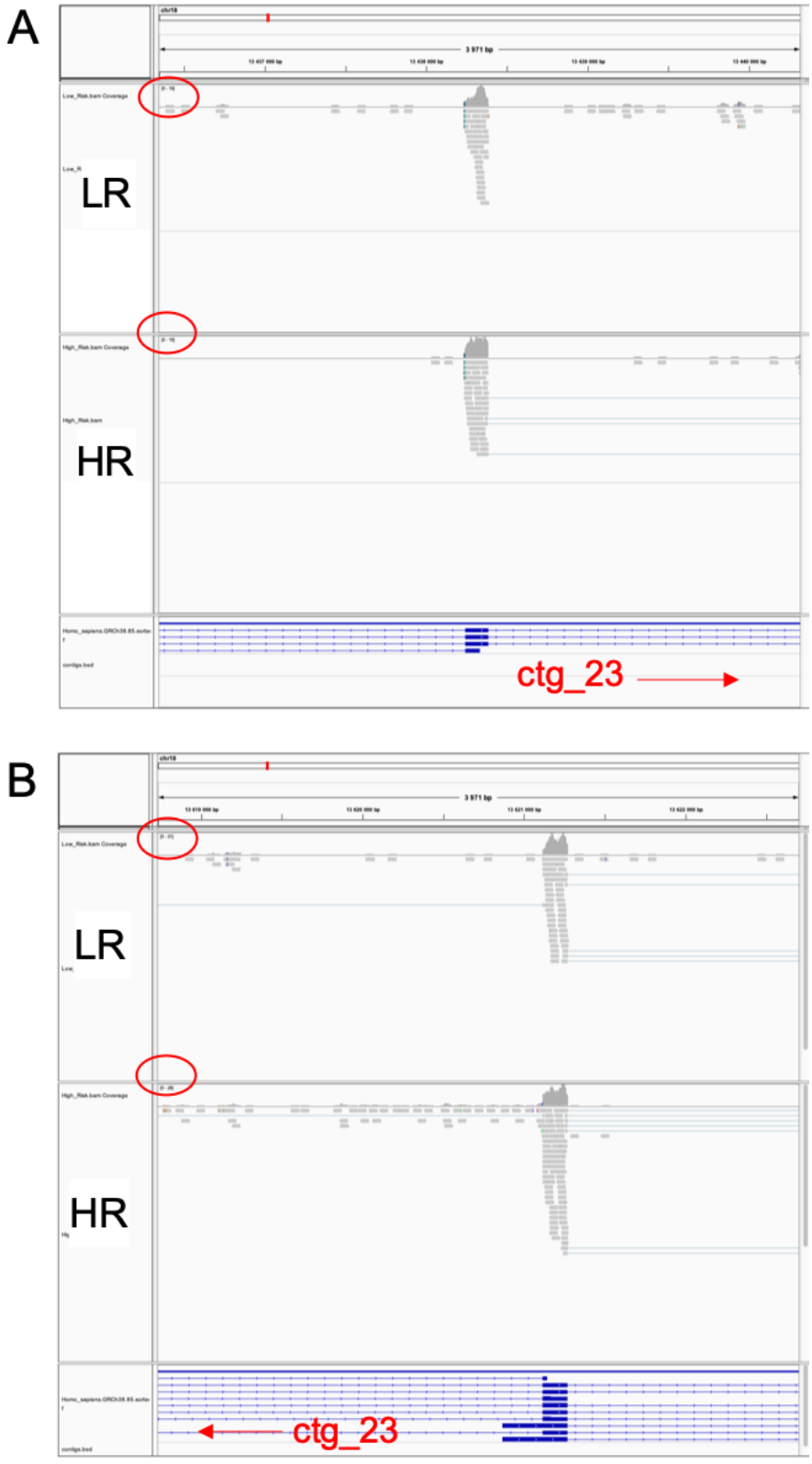

Figure S2. IGV view of RNA-seq reads aligned on the LDLRAD4 exons flanking signature contig ctg_23 on the left (A) and right (B) side of the genomic location of the contig. HR and LR frames are as described above. Note the coverage depth about 6 times lower than ctg_23 coverage in HR condition (red circles) and its lack of variation between LR and HR conditions.

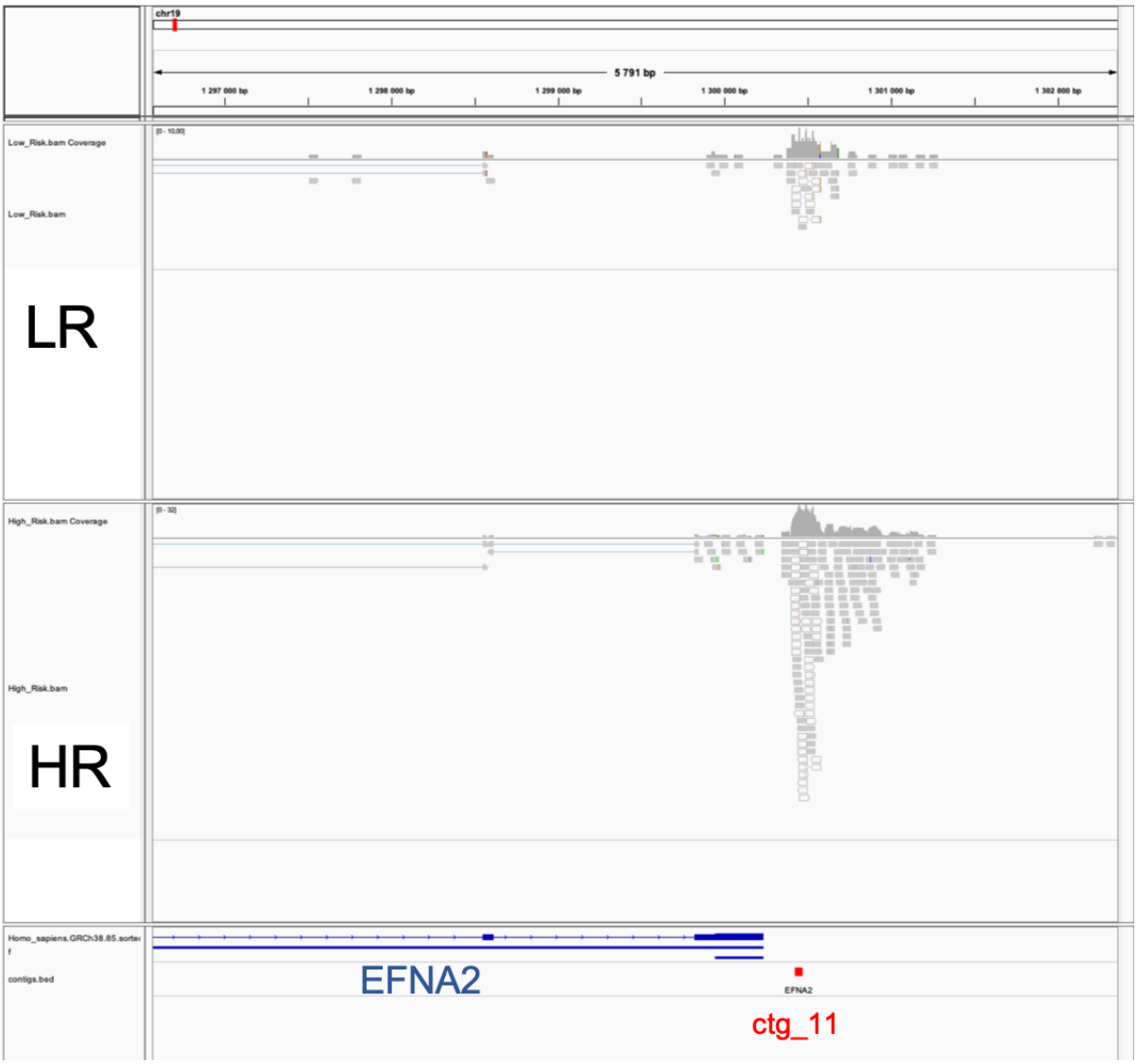

Figure S3. IGV view of RNA-seq reads aligned at risk signature contig ctg_11. Figure legend is as above. ctg_11 was assigned to EFNA2 based on an 3' extended isoform (not shown), but it appears it is more likely an independent transcript.

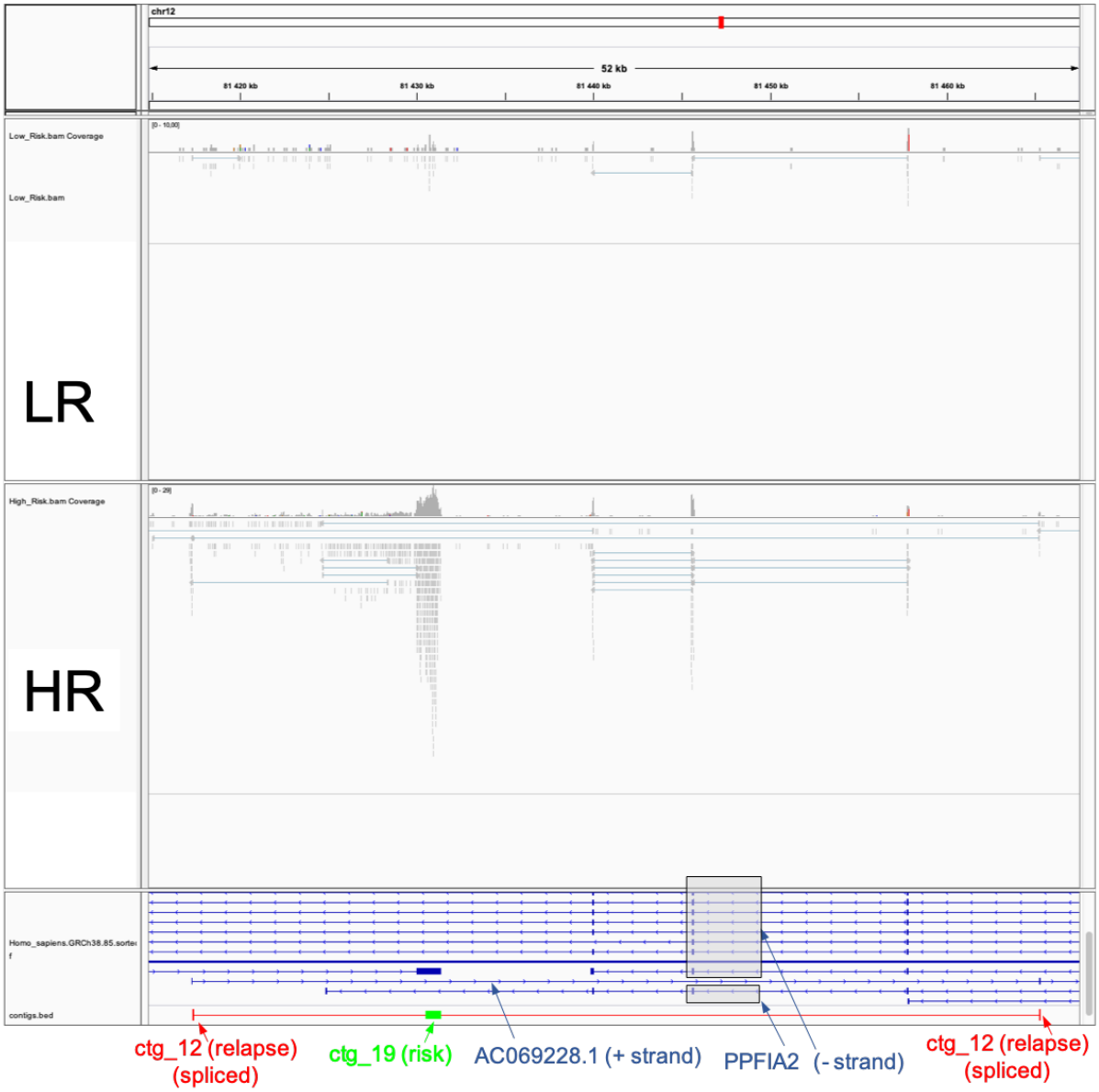

Figure. S4. IGV view of RNA-seq reads aligned at locus AC069228.1, where two signature contigs (ctg_19 from the risk model and ctg_12 from the relapse model) are aligned. Figure legend is as above. Contigs match two different transcripts of the AC069228.1 lncRNA gene, located antisense of gene PPFIA2 (boxed transcripts). In spite of the unstranded nature of aligned reads, mapping to AC069228.1 is unambiguous as only this gene has annotated exons at the corresponding locations.

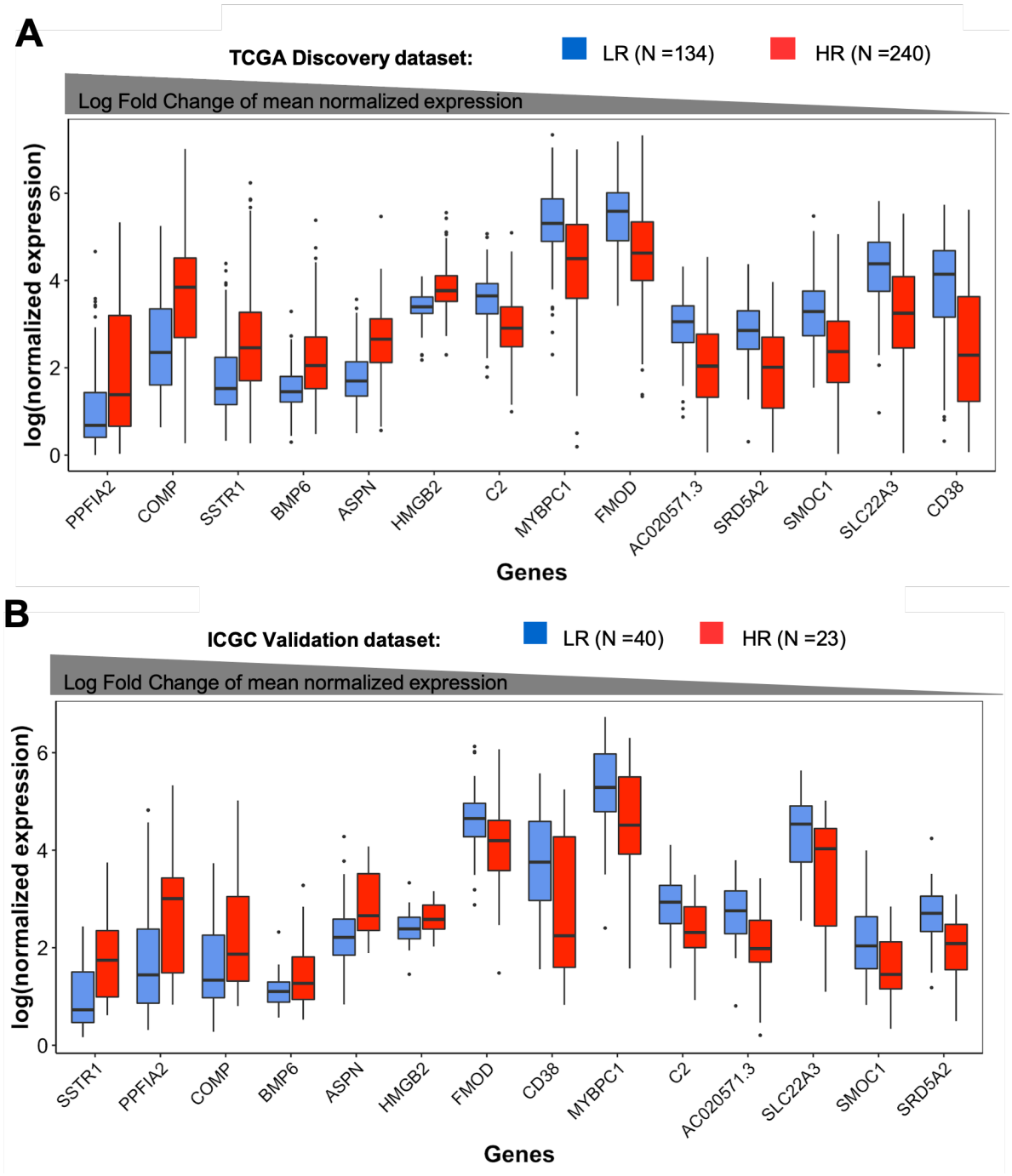

Figure S5. Expression of risk signature genes in relapse/non relapse samples. A: TCGA-PRAD discovery cohort. B: ICGC-PRAD validation cohort.

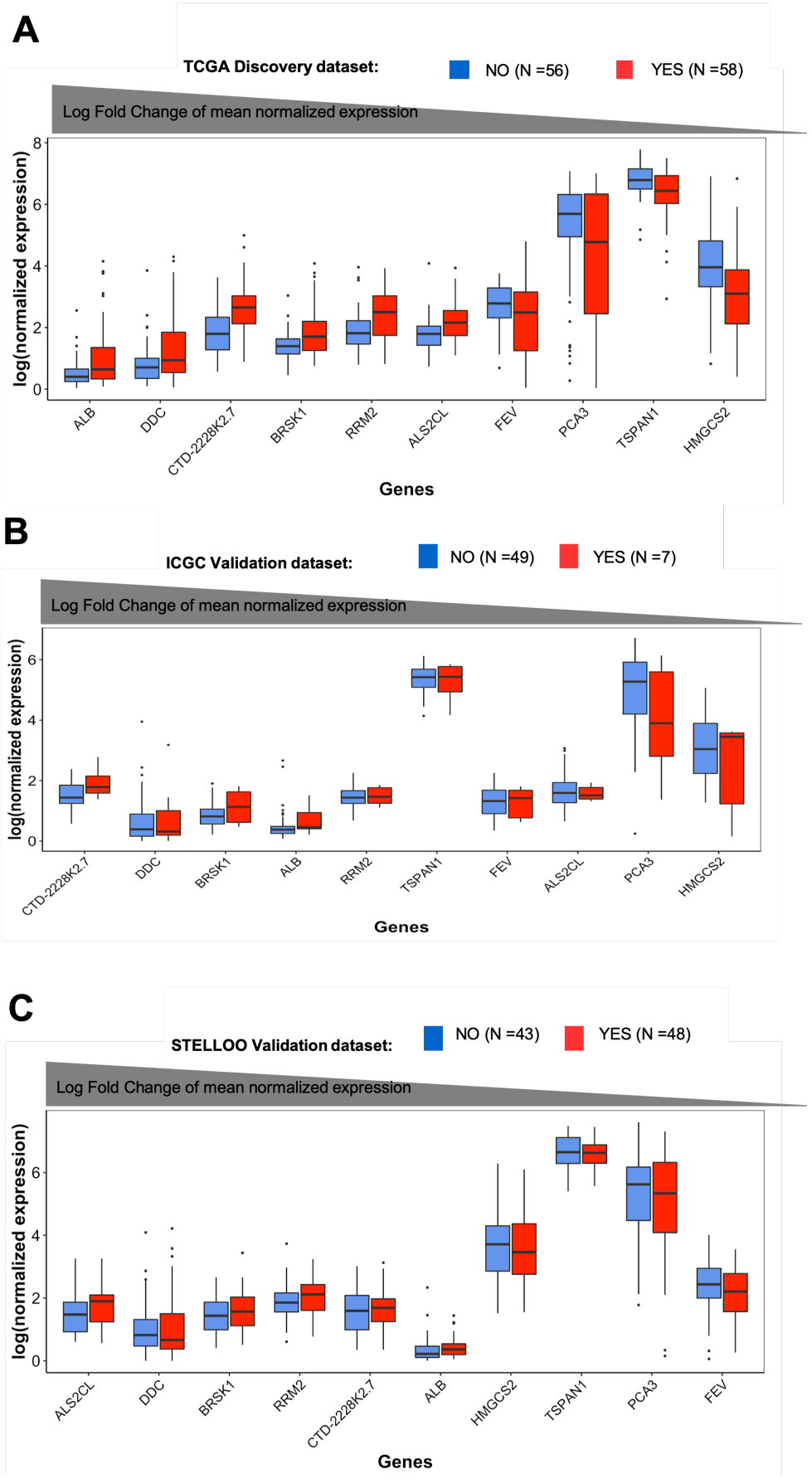

Figure S6. Expression of relapse signature genes in relapse/non relapse samples. A: TCGA-PRAD discovery cohort. B: ICGC-PRAD validation cohort. C: Stelloo validation cohort.