Comprehensive evaluation of machine learning models and gene expression signatures for prostate cancer prognosis using large population cohorts

Ruidong Li[1,2*], Jianguo Zhu[3], Wei-De Zhong[4,5,6], Zhenyu Jia[1,2*]

Affiliations:

[1] Department of Botany and Plant Sciences, University of California, Riverside, CA, USA

[2] Graduate Program in Genetics, Genomics, and Bioinformatics, University of California, Riverside, CA, USA

[3] Department of Urology, Guizhou Provincial People's Hospital, Guizhou, China

[4] Department of Urology, Guangdong Key Laboratory of Clinical Molecular Medicine and Diagnostics, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China

[5] Urology Key Laboratory of Guangdong Province, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou Medical University, Guangzhou, China

[6] Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macau, China

Running title: Evaluation of prognostic signatures for prostate cancer

Key words: prostate cancer, prognostic signature, machine learning, gene expression, clinical decision-making

* To whom correspondence should be addressed: Ruidong Li at rli012@ucr.edu and Zhenyu Jia at arthur.jia@ucr.edu.

Financial Support

This work was supported by Z.J.'s UC Riverside Faculty Start-up Fund and UC Cancer Research Coordinating Committee Competition Award. J.Z. was supported by the Science and Technology Project of Guizhou Province in 2017 ([2017]5803), the High-level innovative talent project of Guizhou Province in 2018 ([2018]5639), and the Science and Technology Plan Project of Guiyang in 2019 ([2019]2-15). W.Z. was supported by the grants from National Natural Science Foundation of China (82072813, 8157142) and Guangzhou Municipal Science and Technology Project (201803040001).

Competing interests

The authors declare no potential conflicts of interest

Abstract

Overtreatment remains the pervasive problem in prostate cancer (PCa) management due to the highly variable and often indolent course. Molecular signatures derived from gene expression profiling have played critical roles in PCa treatment decision-making. Many gene expression signatures have been developed to improve the risk stratification of PCa and some of them have already been translationally applied to clinical practice, however, no comprehensive evaluation was performed to compare the performances of the signatures. In this study, we conducted a systematic and unbiased evaluation of 15 machine learning (ML) algorithms and 30 published PCa gene expression-based prognostic signatures leveraging 10 transcriptomics datasets with 1,558 primary PCa patients from public data repositories. The results revealed that survival analysis models outperformed binary classification models for risk assessment, and the performances of the survival analysis methods - Cox model regularized with ridge penalty (Cox-Ridge) and partial least squares regression for Cox model (Cox-PLS) – were generally more robust than the other methods. Based on the Cox-Ridge algorithm, a few top prognostic signatures that have comparable or even better performances than the commercial panels have been identified. The findings from the study will greatly facilitate the identification of existing prognostic signatures that are promising for further validations in prospective studies and promote the development of robust prognostic models to guide clinical decision-making. Moreover, the study provided a valuable data resource from large primary PCa cohorts, which can be used to develop, validate, and evaluate novel statistical methodologies and molecular signatures to improve PCa management.

Introduction

Prostate cancer (PCa) is the second most frequently diagnosed cancer in men worldwide, with an estimated 1.4 million new cases and 375,000 deaths in 2020 (1). Localized PCa is a highly heterogeneous disease which may lead to variable clinical outcomes. The majority of prostate tumors grow slowly and patients with low-risk PCa may only need active surveillance, whereas patients with aggressive PCa will require immediate treatment. Radical prostatectomy (RP) is the primary treatment for localized PCa with good oncologic outcomes (2). However, approximately 20-40% of patients will experience biochemical recurrence (BCR), i.e., escalated prostate-specific antigen (PSA) levels within 10 years after RP (3–5). Despite of considerable efforts, it remains challenging to accurately predict the aggressiveness of disease at the time of diagnosis to identify patients who only need active surveillance, and to select patients who may benefit from adjuvant treatments, such as chemotherapy, radiation, or immunotherapy following definitive therapy. Thus, overtreatment, which may cause various side effects and impact the quality of patients' lives, continues to be a pervasive problem in PCa management.

In the past decades, many gene expression-based signatures have been developed for risk assessment in PCa using various computational approaches or statistical methodologies. The aim was to identify a set of cancer-associated genes for constructing a statistical model to predict clinical outcomes. A variety of ML algorithms, which can efficiently handle the high-dimensional transcriptomics data where the number of variables (i.e., genes) is substantially greater than the number of observations (i.e., patients), have been widely used for the development of prognostic models in cancer research. The ML algorithms can be classified into two categories: (i) supervised learning such as least absolute shrinkage and selection operator (LASSO), support vector machines (SVM), and random forest (RF), etc. and (ii) unsupervised learning such as principal component analysis (PCA) and variational autoencoders (VAE), etc. Various strategies have been adapted to model the clinical outcome data. For instance, some of the prognostic

111 models were developed by dividing patients into two risk groups, i.e., aggressive
112 and indolent, to identify signature genes (e.g., differentially expressed genes) to
113 develop a binary classifier (6,7), whereas many other models were built by
114 selecting prognostic genes based on their associations with the time to BCR and
115 modeling the censored time-to-event data for the prediction of relapse-free survival
116 (RFS) (8–10). In addition, unsupervised ML techniques, such as hierarchical
117 clustering, non-negative matrix factorization (NMF) clustering, and latent process
118 decomposition (LPD) were also commonly used to identify cancer subtypes with
119 clinical relevance, and a binary or multi-class classifier will eventually be developed
120 for patient stratification (11–13).

121 Molecular signatures derived from gene expression profiling have become
122 important risk assessment tools to assist on therapeutic decision-making for PCa,
123 including the identification of candidates for active surveillance, selection of the
124 type and intensity of treatment, and determination of the benefit from adjuvant
125 therapy or multimodal therapy, etc. Some of the published signatures have already
126 been applied to clinical practice, including Decipher (6), Oncotype DX Genomic
127 Prostate Score (GPS) (14), and Prolaris (8). Evaluation of prognostic signatures
128 has been performed in several cancer types such as breast cancer and lung
129 cancer (15–17), however, no systematic and unbiased comparison has been
130 performed to evaluate the performances of multiple ML algorithms and prognostic
131 signatures for PCa. To fill this void, we leveraged ten public transcriptomics
132 datasets consisting of 1,558 primary PCa cases to comprehensively evaluate the
133 performances of 15 ML algorithms and 30 published gene expression signatures
134 for PCa prognosis. Generally, the survival models using the time-to-event (i.e.,
135 time to BCR) data outperformed the binary classification models which categorizes
136 patients into “aggressive” (BCR occurred) group and “indolent” (BCR-free) group
137 given 5-year follow-up time. Two survival analysis methods, i.e., Cox-Ridge and
138 Cox-PLS, were more robust than the other methods or algorithms. Most of the 30
139 prognostic signatures showed certain prognostic powers, while a few signatures,
140 including Penney (18), Wu (10), Li (19), and Sinnott (20), had comparable as or
141 even superior performance than the commercial panels. These promising

142 prognostic signatures, once validated with prospective trials, may be included in
143 clinical practice to improve PCa management. In addition, our study demonstrated
144 that the performances of random gene expression sets (RGESs), the whole
145 transcriptome, and individual genes in the published signatures all ranked lower
146 than the top published signatures, indicating that it's critical to identify a set of
147 signature genes that are significantly associated with clinical outcomes from the
148 transcriptome to achieve a high level of accuracy for cancer prognosis.

149 This is the first study that comprehensively evaluated the performances of
150 multiple ML algorithms and published prognostic signatures using PCa population
151 cohorts of large sizes. The findings from the study have demonstrated the
152 necessity and clinical impact of deriving molecular signatures from the
153 transcriptome profiling for PCa prognosis, greatly facilitated the selection of
154 promising prognostic signatures for further validations in prospective clinical
155 studies, and helped to refine the strategy for novel and robust prognostic model
156 development to assist on PCa treatment decision-making. Moreover, we have
157 established a valuable data resource, which consists of 10 transcriptomics
158 datasets with a total of 1,558 PCa cases and a collection of 30 gene expression
159 prognostic signatures, for the development, validation, and evaluation of novel
160 statistical methodologies and molecular signatures to improve PCa management.

161 Materials and Methods

162 Collection of PCa transcriptomics data from public data repositories

163 A comprehensive search for transcriptomics data of patients with primary PCa was
164 performed in the public data repositories, including the National Cancer Institute
165 (NCI) Genomic Data Commons (GDC) (21), cBioportal (22), National Center for
166 Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (23), and
167 ArrayExpress (19). The following criteria were used for dataset selection: (i) the
168 dataset must have a complete record of the time to BCR or the time to the last
169 follow-up if no BCR is incurred after treatment; (ii) the sample size should be
170 greater than 80; and (iii) the dataset must be generated using a genome-wide gene

171 expression profiling platform. A total of 10 datasets were selected in the study and
172 the detailed information were provided in Supplementary Table S1.

173 The HTSeq-Counts data from The Cancer Genome Atlas Prostate
174 Adenocarcinoma (TCGA-PRAD) project were downloaded and preprocessed
175 using the R package *GDCRNATools* (25). The raw .CEL files of the four Affymetrix
176 microarray datasets including CPC-Gene (GSE107299), Taylor (GSE21034),
177 CancerMap (GSE94764), and CIT (E-MTAB-6128) were downloaded from
178 GEO/ArrayExpress and normalized with the Robust Multichip Average (RMA)
179 method implemented in the R package *oligo* (26). The raw sequencing data for the
180 GSE54460 dataset was downloaded from SRA (https://www.ncbi.nlm.nih.gov/sra)
181 under the accession number SRP036848 using *fasterq-dump* in the SRA Toolkit
182 (version 2.10.8). *STAR* (version 2.7.2a) (27) was used for sequence alignment and
183 *featureCounts* (version 2.0.0) (28) was used for gene expression quantification.
184 The count data was then normalized using the Trimmed Mean of M values (TMM)
185 method implemented in the R package *edgeR* (29). The reads per kilobase per
186 million mapped reads (RPKM) values for the DFKZ dataset was downloaded from
187 cBioPortal and log2 transformation was performed. The processed intensity data
188 of the other datasets including Cambridge (GSE70768), Stockholm (GSE70769),
189 and Belfast (GSE116918) were downloaded directly from GEO using the R
190 package *GEOquery* (30). Gene/probe IDs in all the datasets were converted to the
191 corresponding Ensembl IDs. If multiple probes/genes matched to the same
192 Ensembl ID, only the most informative one with the maximum interquartile range
193 (IQR) for the gene expression is used for this Ensembl ID. The *ExpressionSet*
194 objects of the processed data, including the normalized gene expression data and
195 harmonized metadata, have been deposited in the PCaDB database (31). The
196 gene expression values in each dataset were rescaled by z-score transformation
197 for model evaluation.

198 **Collection of published gene expression signatures for PCa prognosis**

199 Gene expression-based signatures for PCa prognosis were collected by an
200 inclusive literature screening. The keywords ‘prostate cancer’, ‘prognosis, and

201 ‘gene expression signature’ were used to search the PubMed database. Some
202 signatures were identified in recent review papers on prostate cancer prognostic
203 signatures or in research papers with signature comparisons (32–34). Different
204 types of gene identifiers may be reported in the original papers, so the gene IDs
205 were harmonized by searching against the Ensembl (35), the HUGO Gene
206 Nomenclature Committee (HGNC) (36), and the NCBI Entrez Gene (37) databases.
207 A total of 30 gene expression signatures for PCa prognosis were collected and
208 evaluated in the study (Supplementary Table S2). The gene list of the 30 published
209 prognostic signatures were provided in Supplementary Table S3.

210 **Binary classification algorithms**

211 Nine binary classification algorithms were evaluated in the study, including Elastic
212 Net, support vector machines (SVM) using the linear function (SVM-Linear), the
213 polynomial kernel function (SVM-Poly), and the radial basis function (SVM-RBF),
214 random forest (RF), partial least square (PLS), linear discriminant analysis (LDA),
215 and XGBoost using the linear booster (XGBoost-Linear) and the tree booster
216 (XGBoost-Tree).

217 The R package *caret* (38), which has a set of functions to streamline the
218 process for creating predictive models, was used for model training and parameter
219 tuning. In each model, the predictor variables were the genes in a given signature
220 and the response variable was the 5-year BCR status, where 0 represents non-
221 BCR and 1 represents BCR incurred within 5 years after treatment. The grid search
222 with a tuning length of 10 was used for tuning parameters. The 10-fold cross
223 validation resampling scheme and ROC metric were used to select the optimal
224 model. A probability score for the BCR class was calculated for each patient, where
225 a higher score indicates a higher probability to experience BCR.

226 **Survival analysis methods**

227 Six survival analysis methods were evaluated in the study, including Cox
228 proportional hazards (CoxPH), Cox model regularized with ridge penalty (Cox-
229 Ridge) and lasso penalty (Cox-Lasso), supervised principal components

230 (SuperPC), partial least squares regression for Cox models (Cox-PLS), and
231 random survival forest (RSF).

232 The CoxPH models were built using the R package *survival (*https://CRAN.R-
233 project.org/package=survival*)*. A linear combination of the expression values and
234 coefficients of the signature genes were computed as the risk scores for the
235 patients. Cox-Ridge and Cox-Lasso models were built using the R package *glmnet*
236 (39) with the penalty α=0 and α=1, respectively. The R package *superpc* (40) was
237 used to build the SuperPC models. A subset of genes in a signature with the
238 univariate regression coefficients exceeding the threshold of 0.3 was used to
239 calculate the principal components to build the model. Cox-PLS models were built
240 using the R package *plsRcox* (41) and two components were included in the
241 models. The R package *randomForestSRC* (42) was used to build the RSF models
242 with 100 trees.

243 **Comparison of machine learning models and prognostic signatures**

244 *Intra-dataset (within dataset) comparison*

245 The 10-fold cross-validation (CV) was performed for any dataset to evaluate the
246 performance of a signature using a predictive model within the dataset. In a 10-
247 fold CV, the patient cases were randomly partitioned into ten portions with
248 approximately equal size. In each iteration, nine portions were used as the training
249 set to develop the prediction model, and the risk scores (or class probabilities for
250 binary classification models) were calculated for patients in the remaining one
251 portion (test set). This process was repeated ten times such that each portion was
252 used exactly once as the test data. Identical data partitioning was used to generate
253 10 folds for evaluating all the prediction models.

254 *Inter-dataset (across datasets) comparison*

255 For the inter-dataset comparisons, the prediction models were trained with one
256 dataset and then tested by the other nine independent datasets. Note we have
257 selected ten datasets that met the three selection criteria for this study.

258 *Evaluation Metrics*

259 We used three metrics, i.e., concordance index (C-index), time-dependent receiver
260 operating characteristics (ROC) curve, and hazard ratio (HR) estimated by the
261 Kaplan Meier (KM) analysis, to evaluate the performances of the machine learning
262 models and prognostic signatures. The C-index was calculated using the R
263 package *survcomp* (43), the area under the ROC curves (AUC) was estimated
264 using the R package *survivalROC* (44), and KM analyses were performed using
265 the R package *survival* to estimate the HR and 95% confidence intervals (CIs). For
266 the KM analyses, the median values of the risk scores were used to dichotomize
267 the patients into low- and high-risk groups.

268 **Differential expression analysis, CoxPH survival analysis, and functional**
269 **enrichment analysis for the signature genes**

270 The R packages *limma* (45) was used to perform differential expression (DE)
271 analysis and *survival* was used for CoxPH survival analysis to investigate whether
272 the genes in the prognostic signatures were differentially expressed between
273 tumor and normal samples, and if they were significantly associated with RFS
274 outcomes based on the TCGA-PRAD data. Kyoto Encyclopedia of Genes and
275 Genomes (KEGG) and Disease ontology (DO) enrichment analysis were
276 performed using the R package *clusterProfiler* (46).

277 **Results**

278 **Overview of the study design**

279 A total of 50 public transcriptomics datasets of patients with primary PCa were
280 identified from the public data repositories, including GDC, cBioportal, GEO, and
281 ArrayExpress, while only 14 of them have the BCR status data for the patients.
282 Eventually, 10 transcriptomics datasets (a total of 1,558 cases) which met three
283 data selection criteria (*see Methods*) were included in the study. The clinical
284 characteristics of the 10 cohorts were summarized in Supplementary Table S4.
285 We gleaned 30 published gene expression prognostic signatures for PCa through

286 a comprehensive literature search. Nine binary classification algorithms and six
287 survival analysis methods were used to build the prediction models based on the
288 expression data of genes in each prognostic signature. The comprehensive
289 evaluation included comparisons of: (i) the predictive abilities of binary
290 classification models versus survival analysis models as well as between different
291 survival algorithms; (ii) the prognostic performances of the 30 gene expression
292 signatures based on the robust survival algorithm, and (iii) the performances of
293 RGESs, individual signature genes, union of the signature genes, and the whole
294 transcriptome with the published signatures. Both intra-dataset comparison and
295 inter-dataset comparisons were conducted, and three metrics including C-index,
296 time-dependent AUC, and HR based on KM survival analysis were used to
297 evaluate the models and signatures. The overall design of the study was presented
298 in Figure 1.

299 **Functional characterization of the signature genes**

300 In total, there were 1,032 unique genes in the 30 published prognostic signatures
301 for PCa (union of the signature genes). Gene expression profiles were compared
302 between the primary tumor samples versus the tumor-adjacent normal samples by
303 leveraging the TCGA-PRAD data. The result showed that 223 among the 1,032
304 genes (21%) were differentially expressed with the absolute fold change (FC) > 2
305 and the false discovery rate (FDR) < 0.01 (Figure 2A). The CoxPH survival analysis
306 based on the TCGA-PRAD data showed that 558 out of the 1,032 genes (54%)
307 were significantly associated with RFS with P values < 0.05 (Figure 2B). The 1,032
308 genes were significantly enriched in many important cancer-associated pathways,
309 including the cell cycle pathway, apoptosis pathway, p53 signaling pathway, PI3K-
310 Akt signaling pathway, and prostate cancer pathway, etc. (Figure 2C), indicating
311 the biological importance of these signature genes. The genes involved in those
312 pathways were usually included in many signatures (Supplementary Figure S1).

313 Analysis of the gene overlap among different signatures indicated that the
314 prognostic signatures were largely distinct in terms of gene identity (Figure 2D).
315 For example, a maximum of three genes were shared between the Erho signature

316 (Decipher) and the Penney signature, while only one or two common genes were
317 identified between the Erho signature and eight other signatures. Among the 1,032
318 genes, 142 genes (13.8%) were shared by more than two signatures and 40 genes
319 (3.9%) were common in three or more signatures (Figure 2E). Based on the TCGA-
320 PRAD data, 28 among the 40 common genes (70%) were significantly differentially
321 expressed between the primary tumor samples versus the tumor-adjacent normal
322 samples, and 36 out of the 40 common genes (90%) were significantly associated
323 with RFS. The disease ontology analysis showed that these 40 genes were
324 primarily associated with cancers, including prostate cancer (Figure 2F). Taken
325 together, the results revealed that many the signature genes were indeed relevant
326 to PCa and disease outcomes.

327 **Comparison of survival analysis models versus binary classification models**

328 We first compared the performances of nine binary classification algorithms and
329 six survival analysis methods in risk assessment. The sample size of each dataset
330 for binary classification and survival analysis were summarized in Supplementary
331 Table S5. For each algorithm, the 30 published signatures will be used individually
332 to build the prediction model in the training dataset and then be validated in the
333 validation datasets. For instance, in the intra-dataset comparison, we used 10-fold
334 cross-validation to test a ML algorithm based on each of the 30 signatures across
335 10 datasets, whereas in the inter-dataset comparison, one dataset was used to
336 train the 30 prognostic models and the remaining nine datasets were used for
337 validations. In each strategy, the median values of the three metrics, i.e., C-index,
338 AUC, and HR, were calculated across all the signatures and all the datasets to
339 rank these ML algorithms. The results indicated that almost all the survival models
340 were superior to binary classification models in the intra-dataset comparison
341 (Figure 3), i.e., the median values of C-index, AUC, and HR across all the
342 signatures and all the datasets for the survival models were greater than those for
343 the binary classification models. The results in the inter-dataset comparison were
344 generally consistent with those in the intra-dataset comparison. Some of the top
345 binary classification models such as PLS, Elastic Net, and RF performed equally

346 well or slightly better than the CoxPH or SuperPC survival models, however, the
347 top survival models including Cox-Ridge, Cox-PLS, and RFS outperformed all the
348 binary classification models in the inter-dataset comparison.

349 There are two possible reasons why the survival analysis models were
350 superior to the binary classification models: (i) the sample size became smaller in
351 binary classification models if the censored data were excluded (Supplementary
352 Table S5), and (ii) the variability in survival time provides more information than
353 binary outcome, leading to increased data resolution and therefore statistical
354 power. To test these hypotheses, we compare the survival models with the binary
355 classification models using the same data for which the patients with censored
356 data were removed. The results indicated that the prediction performances for all
357 the survival models increased as sample size (Supplementary Figure S2), and the
358 survival models also outperformed the binary classification models based on the
359 same sample sizes (Supplementary Figure S3), especially in the intra-dataset
360 comparison, which supported our hypotheses.

361 **Performance evaluation of the survival analysis methods**

362 A comprehensive evaluation of the performances of the six algorithms for survival
363 analysis were carried out (Figure 4). For the intra-dataset comparison based on
364 10-fold CV, Cox-Ridge and Cox-PLS models generally performed better than the
365 other models, whereas the most widely used CoxPH method usually ranked
366 bottom among six survival analysis algorithms. The performances of the SuperPC
367 and RSF algorithms substantially varied and largely depended on the datasets that
368 were used for analysis. For example, based on the C-index, the performances of
369 RSF were very similar to the top 2 algorithms Cox-Ridge and Cox-PLS in the DKFZ,
370 GSE54460, and Cambridge datasets, whereas the performances of RSF ranked
371 the last or the second last in the CPC-Gene, Stockholm, CancerMap, CIT, and
372 Belfast datasets. The performances of Cox-Lasso usually ranked higher than RSF,
373 SuperPC, and CoxPH, but lower than Cox-Ridge and Cox-PLS. The result was
374 slightly different for the inter-dataset comparison, where the performances of Cox-
375 Ridge, Cox-PLS, and RSF were comparable and generally better than the other

algorithms. The rankings of Cox-Lasso, CoxPH, and SuperPC were consistent with that in the intra-dataset comparison, where Cox-Lasso > SuperPC = CoxPH.

It was as expected that the survival analysis models usually performed better in the intra-dataset comparison than that in the inter-dataset comparison. This was mainly because that in the intra-dataset comparison, the training and test data were generated in the same study using the same technology, whereas in the inter-dataset comparison, different technologies, different platforms, and even different bioinformatics pipelines may be used to generate the training and test sets. The cohorts may also be quite different between studies, yielding increased level of heterogeneity in inter-dataset analysis. These results in the study emphasized the importance of validation with multiple independent cohorts when building and evaluating prognostic models.

Evaluation of the published gene expression signatures for PCa prognosis

Based on the comprehensive comparisons for different survival analysis algorithms, Cox-Ridge, which performed consistently well across all the datasets was selected to further evaluate the performances of various prognostic signatures. The median values of the three metrics were calculated across all the datasets and were used to rank the signatures. The results indicated that almost all the signatures had some prognostic powers with median C-indexes and median AUCs greater than 0.5, and median HRs greater than 1 both in the intra-dataset and in the inter-dataset comparisons (Figure 5). Some prognostic signatures, including Penny, Li, Klein, Sinnott, Wu, Erho, Kamoun, Planche, and Long, almost always ranked among the top 10 signatures based on the three metrics. Two of the three commercially applied prognostic signatures, i.e., Klein (Oncotype DX GPS) and Erho (Decipher), performed consistently well across the datasets, especially that Klein always ranked among the top 5 signatures in the intra-dataset comparison and ranked the second in the inter-dataset comparison. The median values of C-index, AUC, and HR for the signature Penney usually ranked the first, which was the only signature performing better than the Klein signature in most comparisons. Some of the other top signatures including Li, Sinnott, and Wu always had higher

406 rankings than the Erho signature in the intra-dataset comparison, and Wu also
407 outperformed the Erho signature in the inter-dataset comparison.

408 We also did similar comparisons using the other well-performed algorithm Cox-
409 PLS to investigate whether the performances of the signatures relied on the
410 methods for model development. The results indicated that although the rankings
411 may be slightly different based on different algorithms, the top signatures using the
412 Cox-Ridge algorithm also ranked in the top list when Cox-PLS was used
413 (Supplementary Figure S4). The signature Ramaswamy performed slightly better
414 using the Cox-PLS algorithm, which made it always ranked among the top 10
415 signatures.

416 Since the test characteristics, i.e., C-index, AUC, and HR, may vary when
417 different datasets were used as training and test sets, we utilized the rankings of
418 the median values of these three metrics across all possible combinations of
419 training and test sets to reflect the overall performances of the signatures being
420 evaluated. Figure 6 showed details about the HR-based performances of the
421 signatures for each dataset in the intra-dataset comparison, while Figure 7
422 revealed such performances when each dataset was used as training set in the
423 inter-dataset comparison. The rankings of the top signatures such as Penny, Li,
424 Klein, Sinnott, Wu, Erho, etc. were generally high across multiple comparisons.
425 For instance, in the intra-dataset comparison, the signature Sinnott always ranked
426 among the top five signatures in six datasets, including CPC-Gene, Taylor, DKFZ,
427 GSE54460, CancerMap, and Belfast, whereas the signature Penney ranked
428 among the top five signatures in five datasets, including GSE54460, Cambridge,
429 Stockholm, CancerMap, and Belfast (Figure 6). In the inter-dataset comparison,
430 the signature Penney and Klein can be independently validated in more than half
431 of the nine test datasets no matter which training dataset was used when building
432 the model (Figure 7). The signature Wu can also be validated by multiple
433 independent cohorts when different training sets were used. Supplementary Figure
434 S5 showed the examples of survival curves for the top signatures - Penney and
435 Wu, and the two commercial panels – Klein and Erho when the TCGA-PRAD

436 dataset was used as the training set and the other datasets as validation sets. We
437 also observed that some signatures may perform well when certain datasets were
438 used for model development and validation, but they could show very low or even
439 no prognostic power when other datasets were tried. The results suggested the
440 importance of leveraging data from multiple cohorts to develop and validate new
441 prognostic models or to evaluate existing prognostic models.

442 **The necessity and importance of deriving molecular signatures from gene**
443 **expression profiling**

444 We compared the performances of the published signatures with RGESs, the
445 whole transcriptome, individual signature genes, and union of genes in the
446 published signatures to investigate the necessity of deriving gene expression
447 signatures for PCa prognosis. We generated 10 RGESs with 50 genes each and
448 performed both the intra-dataset and inter-dataset analyses as what have been
449 done for the published signatures using the Cox-Ridge algorithm. Consistent with
450 some previous studies (16,17), RGESs could be significantly associated with
451 clinical outcomes of cancer, however, it was observed that the top published
452 signatures always outperformed all the 10 RGESs in our study (Supplementary
453 Figure S6). We also investigated the performances of prediction models leveraging
454 the whole transcriptome without any feature selection for PCa prognosis. The
455 results showed that the prognostic models using the whole transcriptome generally
456 performed reasonably well, which were better than two third of the published
457 signatures, but they were not as good as the top signatures (Supplementary Figure
458 S7). To investigate whether the individual genes in a published signature have
459 comparable performances with the entire gene expression signature in risk
460 assessment, the expression values of each individual gene were used as the risk
461 scores and the C index, AUC, and HR of KM analysis were calculated in each
462 dataset. These metrics for individual genes were compared with those for the
463 published signatures calculated in the intra-dataset comparison using the robust
464 Cox-Ridge algorithm. As expected, the top 10 individual signature genes
465 performed better than some of the published signatures, however, none of them

466 outperformed the top signatures (Supplementary Figure S8). The union of the
467 signature genes was also modeled using the same strategy and compared with
468 the published signatures to test if integrated gene sets lead to a better performance.
469 The results showed that the performances of the signature with union genes were
470 comparable with the top signatures, which always ranked in the top 5 across all
471 the comparisons (Supplementary Figure S9). On the other hand, it also indicated
472 that a relatively small set of the most significant genes is able to achieve as high
473 prognostic accuracy as the large signatures do, and will be much easier to be
474 translated into clinical practice when the analytical validity and cost-effectiveness
475 are taken into consideration. The sum of the evidences indicated that it is desirable
476 to develop PCa prognostic models using the most significant outcome-associated
477 genes selected from transcriptome.

478 **Translational potential of the prognostic signatures in multiple race groups**

479 Although most published prognostic signatures for PCa were developed and
480 validated in cohorts of predominantly white men, the signatures could also be
481 significantly associated with clinical outcomes of PCa patients in other
482 race/ethnicity populations, thus may potentially be validated in well-designed
483 clinical trials to assist on clinical decision-making for non-white men with PCa. To
484 test this hypothesis, we carried out an analysis by leveraging the TCGA-PRAD
485 dataset that consists of primary PCa samples from 386 white men, 52 African
486 American men and 12 Asian men. The robust Cox-Ridge algorithm was used to
487 develop prediction models based on each of the 30 signatures using the 386
488 training samples of white men. The models were then validated by the cohort of
489 African American patients only (N=52) and by the combined cohort of the African
490 American patients and the Asian patients (N=64), respectively. KM survival
491 analysis was performed to evaluate the associations of the 30 published signatures
492 with the time to BCR. The results indicated that 12 signatures can be validated in
493 the African American cohort and 19 signatures can be validated in the combined
494 cohort (African Americans + Asians) with p values less than 0.05, including many
495 of the top signatures in the aforementioned study such as Li, Penney, Wu, Erho,

496 and Klein, as well as some other signatures such as Ross-Adams, Romos-
497 Montoya, and Varambally, etc. (Supplementary Figure S10). In spite of small
498 sample size, the promising results warrant further assessment of these signatures
499 with multiple large validation cohorts of different races and ethnicities.

500 **Discussion**

501 Unlike many other cancer types, a major challenge of localized PCa management
502 is to accurately predict the risk of the disease aggressiveness at the time of
503 diagnosis to make informed clinical decisions to avoid over-treatment because the
504 majority of localized PCa tumors grow slowly and will likely never cause health
505 problems but many patients will still undergo RP, a surgical procedure to remove
506 the entire prostate gland, which may adversely impact health-related quality of life.
507 Accurate risk assessment tools are desperately needed to assist on PCa treatment
508 decision-making. Molecular signatures derived from gene expression profiling are
509 becoming critical tools and have impacted the clinical practice in PCa management.
510 Once validated in various clinical settings, the molecular signatures will be of great
511 utility, including (i) identification of candidates for active surveillance to avoid
512 unnecessary surgery and selection of patients who will need immediate treatment
513 to avoid missing the opportunity for early intervention. (ii) determination of the type
514 and intensity of treatment, (iii) evaluation of the need of multimodal therapy or the
515 possible benefits of adjuvant therapies, and (iv) periodic monitoring of local
516 recurrence and distant metastasis after definitive therapy.

517 A systematic and unbiased evaluation of ML algorithms and gene expression
518 signatures for cancer prognosis using large cohorts is very critical to select the
519 optimal approaches for building the predictive models and identify the most
520 promising signatures for further validations in prospective clinical studies before
521 they can be applied to clinical practice. The results from our study revealed that
522 the survival analysis algorithms modeling the censored time-to-event data
523 generally outperformed the binary classification algorithms. This is mainly because
524 that the censored data were removed when building the binary classifiers which
525 resulted in smaller sample size, and the binary classification algorithms

526 themselves were not as good as the survival analysis algorithms for risk
527 assessment. Among the six survival analysis methods, Cox-Ridge and Cox-PLS
528 generally performed better than the other algorithms, whereas the most widely
529 used method CoxPH usually performed the worst or the second worst. The
530 performances of RSF were not good in the intra-dataset comparison, but were very
531 robust in the inter-dataset comparison. It was as expected that the prediction
532 models in the inter-dataset comparison usually do not perform as well as that in
533 the intra-dataset comparison because of the high heterogeneities between
534 different cohorts. This emphasized the importance of clinical validations in
535 independent cohorts for the development and evaluation of prognostic models.

536 In this study, the two commercial tests, Oncotype DX GPS (Klein signature)
537 and Decipher (Erho signature) almost always ranked in the top signatures in all the
538 comparisons. We were also able to identify a few other published signatures such
539 as Penney, Wu, Li, and Sinnott that generally had comparable or even better
540 performances than the commercial panels. These four signatures and some other
541 top signatures such as Kamoun, Planche, and Long could be promising for
542 developing robust prognostic models to be incorporated into clinical practice with
543 further prospective validations. The promising signatures were identified mainly
544 based on the overall performances. It's possible that a signature could perform
545 well using a given machine learning algorithm and a given training data to develop
546 the predictive model, while could have no prognostic power when other algorithms
547 or training datasets were used. Therefore, multiple robust survival analysis
548 algorithms and large PCa population cohorts were recommended when
549 developing new prognostic signatures. A few previous studies have demonstrated
550 that RGESs may be significantly associated with clinical outcomes of cancer,
551 which has also been observed in our study, however, the top published signatures
552 always outperformed all RGESs. We also evaluated the performances of
553 prognostic models using the whole transcriptome data and found that the
554 transcriptome models could outperform some of the signatures but were not
555 comparable with the top signatures. Although the performance of the union of the
556 signature genes were comparable with the top signatures and in certain cases

557 were slightly better than them, it would be challenging to translate the signatures
558 with a large number genes into clinical practice. These findings from our study
559 indicated that it's critical to derive molecular signatures with a relatively small
560 number of genes that are relevant to PCa outcomes from the transcriptome to
561 further improve the PCa prognosis to assist on clinical decision-making.

562 From the comprehensive review and evaluation of the published prognostic
563 signatures for PCa, we proposed a few practice principles should be followed when
564 developing and evaluating a molecular signature for PCa prognosis, such as (i)
565 demonstration of clinical utility, (ii) selection of the most relevant cohorts and
566 endpoints, (iii) selection of the robust machine learning algorithm and modeling
567 time-to-event data if available rather than categorical data, (iv) conducting multiple
568 external validations to demonstrate the clinical validity, and (v) comparison with
569 the top published signatures and the established clinico-pathological prognostic
570 markers. Some of these principles have also been extensively reviewed and
571 discussed elsewhere (47–49). Clinical utility is the most important criterion for
572 molecular signature development. The majority of the published gene expression
573 signatures for PCa prognosis have been developed based on RP tissue in surgical
574 cohorts. Thus, these signatures may be more relevant to patients with primary
575 treatment for assessing the risk of recurrence, metastasis, or prostate cancer–
576 specific mortality (PCSM), and determining if the patients will benefit from adjuvant
577 therapy. Although some of these signatures have also been proven to be useful to
578 identify candidates for active surveillance with prostate needle biopsy samples, the
579 signatures directly derived in the prostate needle biopsy setting are needed to
580 boost sensitivity and specificity of the test. It's also critical to select the most
581 relevant endpoints for future molecular signature development and validation.
582 Some published signatures were developed to predict the Gleason score, which
583 may have limited clinical impact. The surrogate endpoint BCR is most commonly
584 used to define disease progression or aggressiveness, however, it should be noted
585 that many patients with an increase of PSA level following the definitive therapy
586 may not experience local recurrence or develop distant metastasis for a few years.
587 The clinical recurrence, metastasis, and PCSM may be better endpoints, and if

588 possible, the time-to-event data should be modeled using the most robust survival
589 analysis algorithms to improve the accuracy of risk stratification. The endpoint
590 adverse pathology could also be used to define the disease aggressiveness when
591 developing molecular signatures for risk assessment at the time of biopsy. Clinical
592 validity is another critical criterion for molecular signature development. Prognostic
593 signatures are of no utility without multiple well-designed external validations. The
594 signature scores computed using any models will be significantly associated with
595 the endpoint of interest in the training dataset. The findings from our study
596 emphasized the importance of independent validation when building and
597 evaluating prognostic models. In addition, any new prognostic signatures should
598 be compared with the top published signatures and the established clinico-
599 pathological prognostic markers to assess the incremental values of the signatures.

600 In summary, this is the first study that performed a comprehensive evaluation
601 of the performances of multiple machine learning models and published gene
602 expression signatures using large PCa cohorts. The findings from the study have
603 greatly facilitated the identification of existing prognostic signatures for further
604 validations and the selection of the best strategies for the development of more
605 robust signatures to assist on PCa treatment decision making. In addition, our
606 study also provided a valuable resource with 10 transcriptomics datasets from
607 large primary PCa cohorts and a comprehensive collection of 30 gene expression
608 prognostic signatures that can be used to develop, validate, and evaluate novel
609 statistical methodologies and molecular signatures to improve PCa management.

610 **Availability of data and materials**

611 The processed data including the normalized expression data and the harmonized
612 metadata of all the 10 transcriptomics datasets were deposited in the PCaDB
613 database (http://bioinfo.jialab-ucr.org/PCaDB/) and the *ExpressionSet* object of
614 each dataset can be easily downloaded from this database. All the scripts used in
615 this study, including data preprocessing, model comparison, and data visualization,
616 are publicly available at https://github.com/rli012/PCaSignatures.

617 **References**

618 1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al.
619 Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and
620 Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal
621 for Clinicians. 2021;71:209–49.

622 2. Hruza M, Bermejo JL, Flinspach B, Schulze M, Teber D, Rumpelt HJ, et al.
623 Long-term oncological outcomes after laparoscopic radical prostatectomy.
624 BJU International. 2013;111:271–80.

625 3. Freedland SJ, Humphreys EB, Mangold LA, Eisenberger M, Dorey FJ, Walsh
626 PC, et al. Risk of Prostate Cancer–Specific Mortality Following Biochemical
627 Recurrence After Radical Prostatectomy. JAMA. 2005;294:433–9.

628 4. Stephenson AJ, Kattan MW, Eastham JA, Dotan ZA, Bianco FJ, Lilja H, et al.
629 Defining biochemical recurrence of prostate cancer after radical
630 prostatectomy: a proposal for a standardized definition. J Clin Oncol.
631 2006;24:3973–8.

632 5. Bhargava P, Ravizzini G, Chapin BF, Kundra V. Imaging Biochemical
633 Recurrence After Prostatectomy: Where Are We Headed? American Journal
634 of Roentgenology. American Roentgen Ray Society; 2020;214:1248–58.

635 6. Erho N, Crisan A, Vergara IA, Mitra AP, Ghadessi M, Buerki C, et al.
636 Discovery and validation of a prostate cancer genomic classifier that predicts
637 early metastasis following radical prostatectomy. PLoS One. 2013;8:e66855.

638 7. Irshad S, Bansal M, Castillo-Martin M, Zheng T, Aytes A, Wenske S, et al. A
639 Molecular Signature Predictive of Indolent Prostate Cancer. Science
640 Translational Medicine. American Association for the Advancement of
641 Science; 2013;5:202ra122-202ra122.

642 8. Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, et al.
643 Prognostic value of an RNA expression signature derived from cell cycle

644 proliferation genes in patients with prostate cancer: a retrospective study.
645 Lancet Oncol. 2011;12:245–55.

646 9. Long Q, Xu J, Osunkoya AO, Sannigrahi S, Johnson BA, Zhou W, et al. Global
647 transcriptome analysis of formalin-fixed prostate cancer specimens identifies
648 biomarkers of disease recurrence. Cancer Res. 2014;74:3228–37.

649 10. Wu C-L, Schroeder BE, Ma X-J, Cutie CJ, Wu S, Salunga R, et al.
650 Development and validation of a 32-gene prognostic index for prostate cancer
651 progression. Proc Natl Acad Sci U S A. 2013;110:6121–6.

652 11. Luca B-A, Brewer DS, Edwards DR, Edwards S, Whitaker HC, Merson S, et
653 al. DESNT: A Poor Prognosis Category of Human Prostate Cancer. Eur Urol
654 Focus. 2018;4:842–50.

655 12. Ross-Adams H, Lamb AD, Dunning MJ, Halim S, Lindberg J, Massie CM, et
656 al. Integration of copy number and transcriptomics provides risk stratification
657 in prostate cancer: A discovery and validation cohort study. EBioMedicine.
658 2015;2:1133–44.

659 13. You S, Knudsen BS, Erho N, Alshalalfa M, Takhar M, Al-Deen Ashab H, et al.
660 Integrated Classification of Prostate Cancer Reveals a Novel Luminal
661 Subtype with Poor Outcome. Cancer Res. 2016;76:4948–58.

662 14. Klein EA, Cooperberg MR, Magi-Galluzzi C, Simko JP, Falzarano SM,
663 Maddala T, et al. A 17-gene assay to predict prostate cancer aggressiveness
664 in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy
665 undersampling. Eur Urol. 2014;66:550–60.

666 15. Zhao X, Rødland EA, Sørlie T, Vollan HKM, Russnes HG, Kristensen VN, et
667 al. Systematic assessment of prognostic gene signatures for breast cancer
668 shows distinct influence of time and ER status. BMC Cancer. 2014;14:211.

669 16. Venet D, Dumont JE, Detours V. Most Random Gene Expression Signatures
670 Are Significantly Associated with Breast Cancer Outcome. PLOS
671 Computational Biology. Public Library of Science; 2011;7:e1002240.

672 17. Tang H, Wang S, Xiao G, Schiller J, Papadimitrakopoulou V, Minna J, et al.
673 Comprehensive evaluation of published gene expression prognostic
674 signatures for biomarker-based lung cancer clinical studies. Annals of
675 Oncology. Elsevier; 2017;28:733–40.

676 18. Penney KL, Sinnott JA, Fall K, Pawitan Y, Hoshida Y, Kraft P, et al. mRNA
677 Expression Signature of Gleason Grade Predicts Lethal Prostate Cancer.
678 JCO. Wolters Kluwer; 2011;29:2391–6.

679 19. Li R, Wang S, Cui Y, Qu H, Chater JM, Zhang L, et al. Extended application
680 of genomic selection to screen multiomics data for prognostic signatures of
681 prostate cancer. Brief Bioinform. 2021;22:bbaa197.

682 20. Sinnott JA, Peisch SF, Tyekucheva S, Gerke T, Lis R, Rider JR, et al.
683 Prognostic Utility of a New mRNA Expression Signature of Gleason Score.
684 Clin Cancer Res. 2017;23:81–7.

685 21. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data
686 Commons as an engine for precision medicine. Blood. 2017;130:453–9.

687 22. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al.
688 Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using
689 the cBioPortal. Sci Signal. 2013;6:pl1.

690 23. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et
691 al. NCBI GEO: archive for functional genomics data sets—update. Nucleic
692 Acids Research. 2013;41:D991–5.

693 24. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena
694 N, et al. ArrayExpress—a public repository for microarray gene expression
695 data at the EBI. Nucleic Acids Research. 2003;31:68–71.

696 25. Li R, Qu H, Wang S, Wei J, Zhang L, Ma R, et al. GDCRNATools: an
697 R/Bioconductor package for integrative analysis of lncRNA, miRNA and
698 mRNA data in GDC. Bioinformatics. 2018;34:2515–7.

699 26. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray
700 preprocessing. Bioinformatics. 2010;26:2363–7.

701 27. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR:
702 ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

703 28. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose
704 program for assigning sequence reads to genomic features. Bioinformatics.
705 2014;30:923–30.

706 29. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for
707 differential expression analysis of digital gene expression data. Bioinformatics.
708 2010;26:139–40.

709 30. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression
710 Omnibus (GEO) and BioConductor. Bioinformatics. 2007;23:1846–7.

711 31. Li R, Zhu J, Zhong W-D, Jia Z. PCaDB - a comprehensive and interactive
712 database for transcriptomes from prostate cancer population cohorts. bioRxiv.
713 Cold Spring Harbor Laboratory; 2021;2021.06.29.449134.

714 32. Luca B-A, Moulton V, Ellis C, Connell SP, Brewer DS, Cooper CS.
715 Convergence of Prognostic Gene Signatures Suggests Underlying
716 Mechanisms of Human Prostate Cancer Progression. Genes (Basel).
717 2020;11:E802.

718 33. Mo F, Lin D, Takhar M, Ramnarine VR, Dong X, Bell RH, et al. Stromal Gene
719 Expression is Predictive for Metastatic Primary Prostate Cancer. Eur Urol.
720 2018;73:524–32.

721 34. Yang L, Roberts D, Takhar M, Erho N, Bibby BAS, Thiruthaneeswaran N, et
722 al. Development and Validation of a 28-gene Hypoxia-related Prognostic
723 Signature for Localized Prostate Cancer. EBioMedicine. 2018;31:182–9.

724 35. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al.
725 Ensembl 2020. Nucleic Acids Research. 2020;48:D682–8.

726 36. Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al.
727 Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids
728 Research. 2021;49:D939–46.

729 37. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered
730 information at NCBI. Nucleic Acids Research. 2005;33:D54–8.

731 38. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of
732 Statistical Software. 2008;28:1–26.

733 39. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's
734 Proportional Hazards Model via Coordinate Descent. J Stat Softw. 2011;39:1–
735 13.

736 40. Bair E, Tibshirani R. Semi-Supervised Methods to Predict Patient Survival
737 from Gene Expression Data. PLoS Biol. 2004;2:e108.

738 41. Bastien P, Bertrand F, Meyer N, Maumy-Bertrand M. Deviance residuals-
739 based sparse PLS and sparse kernel PLS regression for censored data.
740 Bioinformatics. 2015;31:397–404.

741 42. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests.
742 The Annals of Applied Statistics. Institute of Mathematical Statistics;
743 2008;2:841–60.

744 43. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an
745 R/Bioconductor package for performance assessment and comparison of
746 survival models. Bioinformatics. 2011;27:3206–8.

747 44. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored
748 survival data and a diagnostic marker. Biometrics. 2000;56:337–44.

749 45. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers
750 differential expression analyses for RNA-sequencing and microarray studies.
751 Nucleic Acids Research. 2015;43:e47–e47.

46. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021;100141.

47. Michiels S, Ternès N, Rotolo F. Statistical controversies in clinical research: prognostic gene signatures are not (yet) useful in clinical practice. Annals of Oncology. 2016;27:2160–7.

48. Teutsch SM, Bradley LA, Palomaki GE, Haddow JE, Piper M, Calonge N, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative: methods of the EGAPP Working Group. Genet Med. 2009;11:3–14.

49. Simon RM, Paik S, Hayes DF. Use of Archived Specimens in Evaluation of Prognostic and Predictive Biomarkers. JNCI: Journal of the National Cancer Institute. 2009;101:1446–52.

Figure Legends

Figure 1. Overview of the study design. In general, the transcriptomics data and sample metadata for a total of 10 independent prostate cancer (PCa) cohorts (1,588 primary tumor samples) with the time to biochemical recurrence (BCR) data were collected and included in the study. The performances of 15 machine learning (ML) algorithms including 6 survival analysis methods and 9 binary classification algorithms were compared to identify the most robust one, and the 30 published gene expression signatures were analyzed and evaluated based on the most robust ML algorithm across all the independent cohorts. Both the intra-dataset (10-fold cross-validation within a dataset) and inter-dataset (train the model in one dataset and test the model in the remaining independent datasets) comparisons were conducted to evaluate the models and signatures.

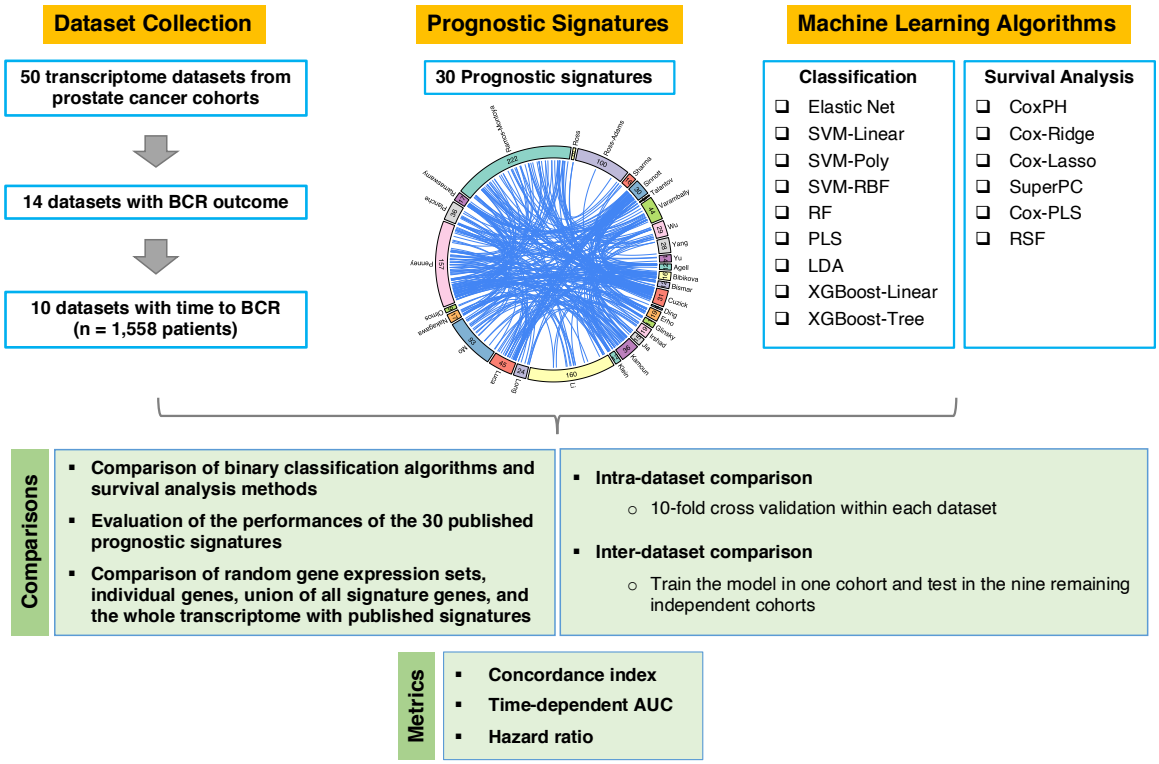
Figure 2. Functional characterization of the signature genes. (**A**) Differential expression analysis of the 1,042 unique signature genes (union of the 30 prognostic signatures) between tumor vs. normal samples based on the TCGA-PRAD data. Genes in more than 3 signatures are labeled. (**B**) Cox Proportional-Hazards (CoxPH) survival analysis of relapse-free survival (RFS) for the 1,042 genes based on the TCGA-PRAD data. Genes in more than 3 signatures are labeled. (**C**) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways that the 1,042 genes enriched in. **(D)** Overlap of genes between different prognostic signatures. (**E**) Heatmap showing the 40 common genes that were identified in three or more signatures. (**F**) Disease Ontology (DO) enrichment analysis of the 40 common genes identified in three or more signatures.

Figure 3. Comparison of survival analysis models versus binary classification models. Prognostic performances of the machine learning (ML) algorithms were evaluated based on all the 30 gene expression signatures across all the transcriptomics datasets. Both intra-dataset (**A**) and inter-dataset (**B**)

807 comparisons were performed. Median values of the three metrics, including
808 concordance index (C-index, *top*), area under the time-dependent receiver
809 operating characteristic curve (AUC, *middle*), and hazard ratio (HR, *bottom*)
810 estimated by the Kaplan Meier (KM) analysis of relapse-free survival (RFS) were
811 used to rank the algorithms, respectively. The diamond in each box indicates the
812 median value. Different colors of the box plots denote different categories of ML
813 algorithms, *i.e.*, blue for the six survival analysis methods and red for the nine
814 binary classification algorithms.
815
816 **Figure 4. Comprehensive evaluation of the six survival analysis methods.**
817 The performances of the methods were evaluated based on all the 30 gene
818 expression signatures within each individual dataset via 10-fold cross-validation
819 for intra-dataset comparison (**A**), or using an individual dataset to train the
820 prognostic models and the remaining independent datasets to test the models for
821 inter-dataset comparison (**B**). The title of each small panel indicates the dataset
822 (or the training dataset for inter-dataset comparison) that was being used for model
823 comparison. Median values of the three metrics, including concordance index (C-
824 index, *top*), area under the time-dependent receiver operating characteristic curve
825 (AUC, *middle*), and hazard ratio (HR, *bottom*) estimated by the Kaplan Meier (KM)
826 analysis of relapse-free survival (RFS) were used to rank the survival analysis
827 methods, respectively. The diamond in each box indicates the median value.
828
829 **Figure 5. Evaluation of the 30 gene expression prognostic signatures based**
830 **on the robust Cox-Ridge algorithm.** (**A**) Intra-dataset comparison via 10-fold
831 cross-validation. (**B**) Inter-dataset comparison using one dataset as the training set
832 and the remaining datasets as test sets. The prognostic signatures are ranked
833 based on the median values of the three metrics, including concordance index (C-
834 index, *top*), area under the time-dependent receiver operating characteristic curve
835 (AUC, *middle*), and hazard ratio (HR, *bottom*) estimated by the Kaplan Meier (KM)
836 analysis of relapse-free survival (RFS), respectively, in the plots. Each dot
837 represents the C-index, AUC, or HR value calculated within the dataset via 10-fold

838 cross-validation for the intra-dataset comparison or in a test dataset for the inter-
839 dataset comparison.
840
841 **Figure 6. Intra-dataset comparison of the 30 gene expression prognostic**
842 **signatures based on hazard ratio (HR) estimated by the Kaplan Meier (KM)**
843 **analysis of relapse-free survival (RFS).** The robust Cox-Ridge algorithm was
844 used to train the prognostic models. The signatures are ranked by HRs in each
845 individual dataset in the plots. The title of each small panel indicates the name of
846 the dataset.
847
848 **Figure 7. Inter-dataset comparison of the 30 gene expression prognostic**
849 **signatures based on hazard ratio (HR) estimated by the Kaplan Meier (KM)**
850 **analysis of relapse-free survival (RFS).** The robust Cox-Ridge algorithm was
851 used to train the prognostic models. The title of each small panel refers to the
852 training dataset. The x-axis indicates the test datasets and the y-axis indicates the
853 signatures. The prognostic signatures are ranked based on the numbers of test
854 datasets in which the prognostic models have been successfully validated (p value
855 < 0.05) based on KM survival analysis. Only the validations with p value < 0.05 are
856 shown in the figure.
857
858
859
860
861
862
863
864
865

866 **Figures**



867

868 **Figure 1. Overview of the study design.** In general, the transcriptomics data and
869 sample metadata for a total of 10 independent prostate cancer (PCa) cohorts
870 (1,588 primary tumor samples) with the time to biochemical recurrence (BCR) data
871 were collected and included in the study. The performances of 15 machine learning
872 (ML) algorithms including 6 survival analysis methods and 9 binary classification
873 algorithms were compared to identify the most robust one, and the 30 published
874 gene expression signatures were analyzed and evaluated based on the most
875 robust ML algorithm across all the independent cohorts. Both the intra-dataset (10-
876 fold cross-validation within a dataset) and inter-dataset (train the model in one
877 dataset and test the model in the remaining independent datasets) comparisons
878 were conducted to evaluate the models and signatures.

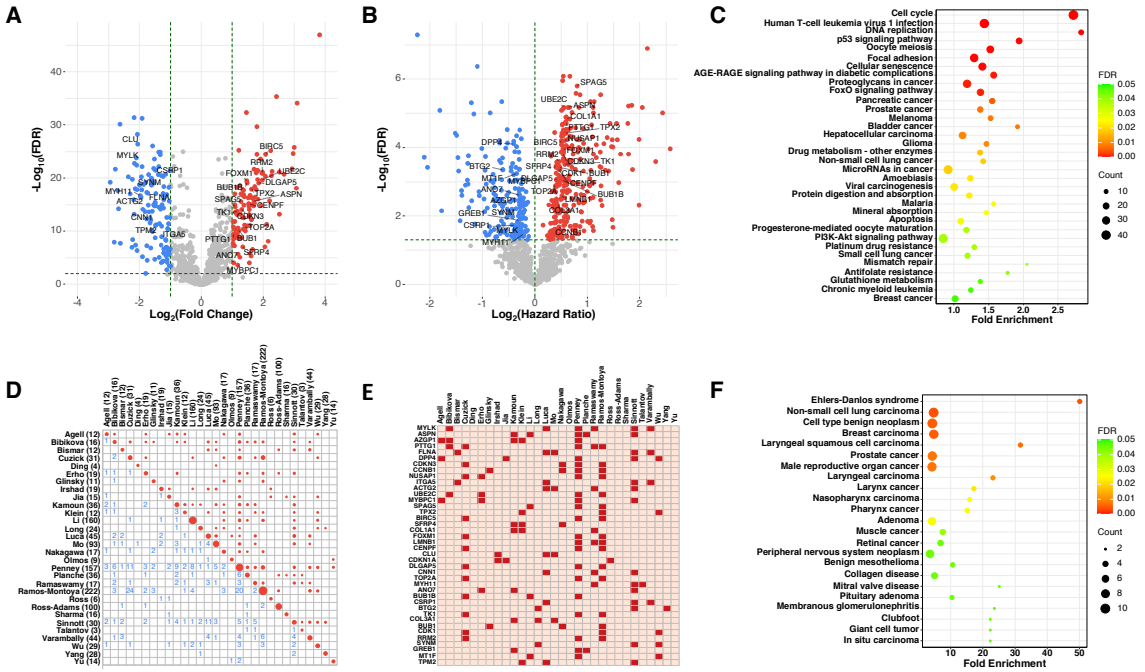


Figure 2. Functional characterization of the signature genes. (**A**) Differential expression analysis of the 1,042 unique signature genes (union of the 30 prognostic signatures) between tumor vs. normal samples based on the TCGA-PRAD data. Genes in more than 3 signatures are labeled. (**B**) Cox Proportional-Hazards (CoxPH) survival analysis of relapse-free survival (RFS) for the 1,042 genes based on the TCGA-PRAD data. Genes in more than 3 signatures are labeled. (**C**) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways that the 1,042 genes enriched in. **(D)** Overlap of genes between different prognostic signatures. (**E**) Heatmap showing the 40 common genes that were identified in three or more signatures. (**F**) Disease Ontology (DO) enrichment analysis of the 40 common genes identified in three or more signatures.

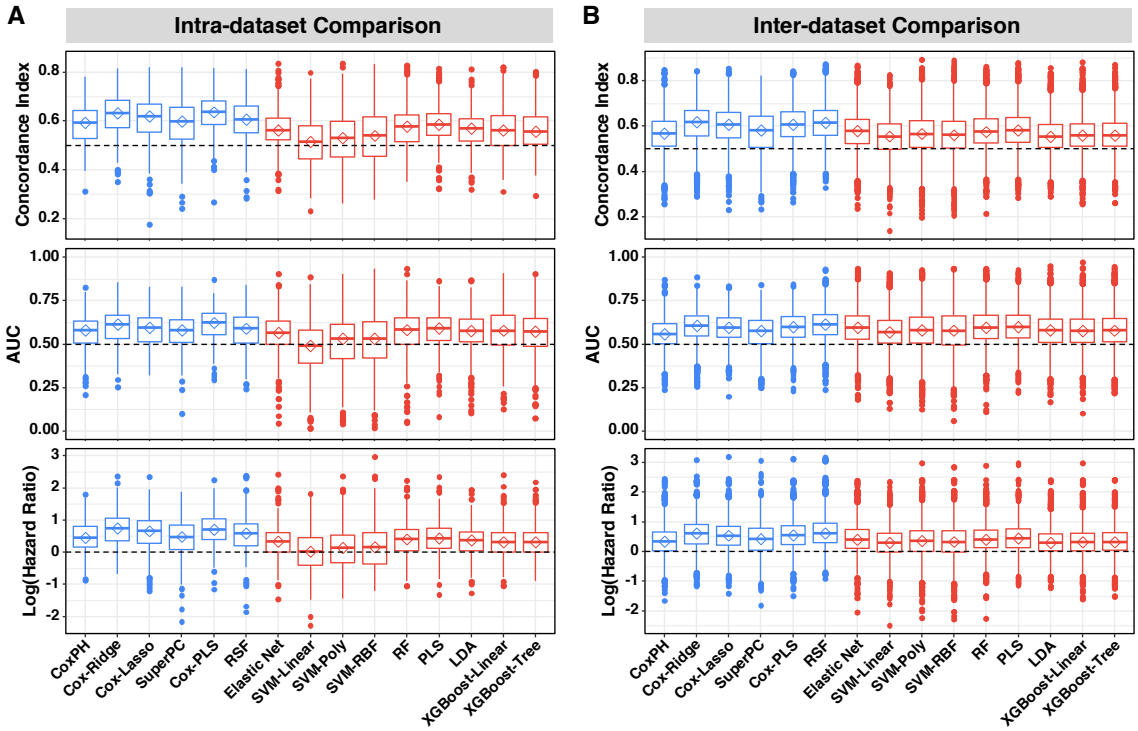


Figure 3. Comparison of survival analysis models versus binary classification models. Prognostic performances of the machine learning (ML) algorithms were evaluated based on all the 30 gene expression signatures across all the transcriptomics datasets. Both intra-dataset (**A**) and inter-dataset (**B**) comparisons were performed. Median values of the three metrics, including concordance index (C-index, *top*), area under the time-dependent receiver operating characteristic curve (AUC, *middle*), and hazard ratio (HR, *bottom*) estimated by the Kaplan Meier (KM) analysis of relapse-free survival (RFS) were used to rank the algorithms, respectively. The diamond in each box indicates the median value. Different colors of the box plots denote different categories of ML algorithms, *i.e.*, blue for the six survival analysis methods and red for the nine binary classification algorithms.

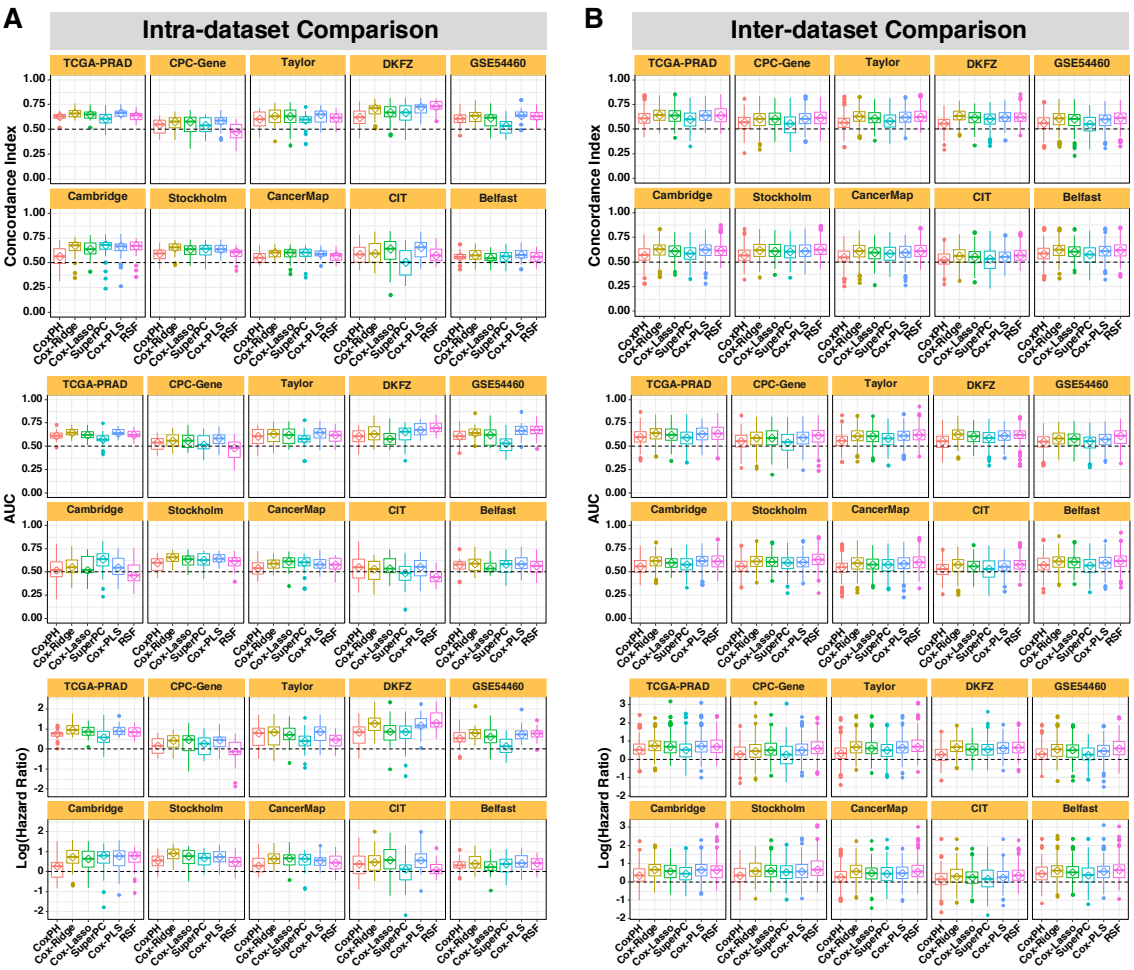


Figure 4. Comprehensive evaluation of the six survival analysis methods. The performances of the methods were evaluated based on all the 30 gene expression signatures within each individual dataset via 10-fold cross-validation for intra-dataset comparison (**A**), or using an individual dataset to train the prognostic models and the remaining independent datasets to test the models for inter-dataset comparison (**B**). The title of each small panel indicates the dataset (or the training dataset for inter-dataset comparison) that was being used for model comparison. Median values of the three metrics, including concordance index (C-index, *top*), area under the time-dependent receiver operating characteristic curve (AUC, *middle*), and hazard ratio (HR, *bottom*) estimated by the Kaplan Meier (KM) analysis of relapse-free survival (RFS) were used to rank the survival analysis methods, respectively. The diamond in each box indicates the median value.

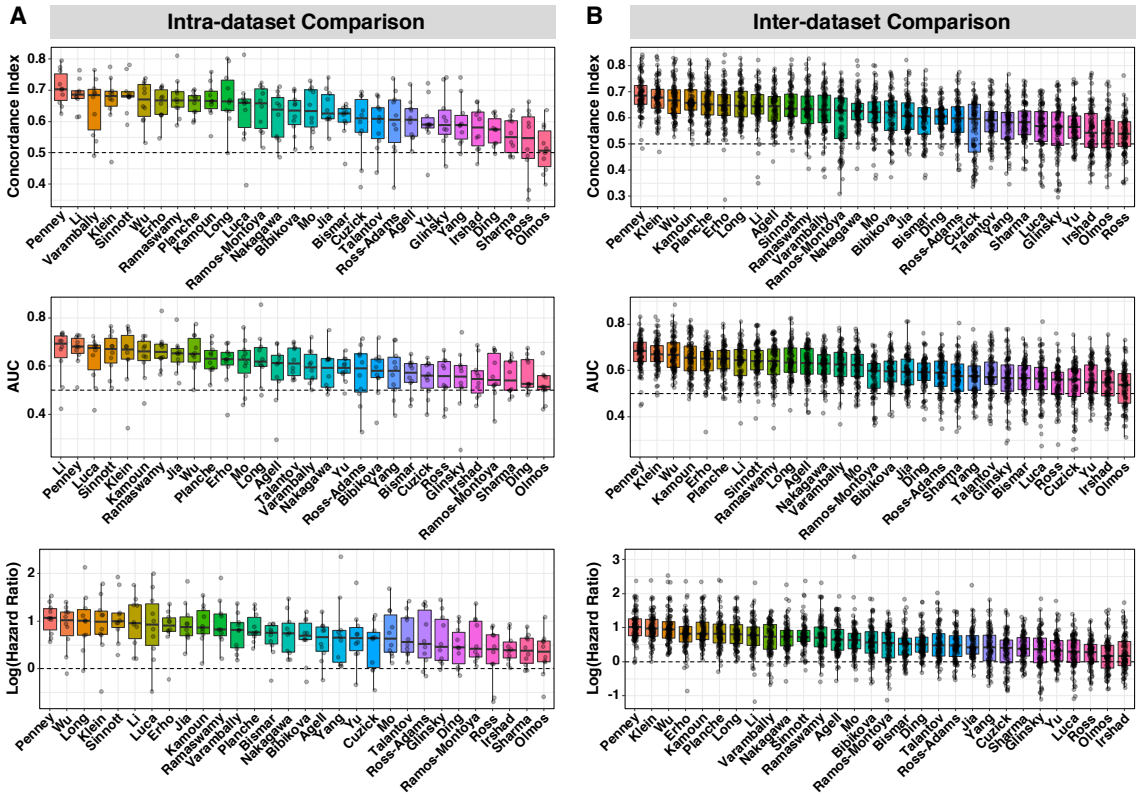


Figure 5. Evaluation of the 30 gene expression prognostic signatures based on the robust Cox-Ridge algorithm. (**A**) Intra-dataset comparison via 10-fold cross-validation. (**B**) Inter-dataset comparison using one dataset as the training set and the remaining datasets as test sets. The prognostic signatures are ranked based on the median values of the three metrics, including concordance index (C-index, *top*), area under the time-dependent receiver operating characteristic curve (AUC, *middle*), and hazard ratio (HR, *bottom*) estimated by the Kaplan Meier (KM) analysis of relapse-free survival (RFS), respectively, in the plots. Each dot represents the C-index, AUC, or HR value calculated within the dataset via 10-fold cross-validation for the intra-dataset comparison or in a test dataset for the inter-dataset comparison.

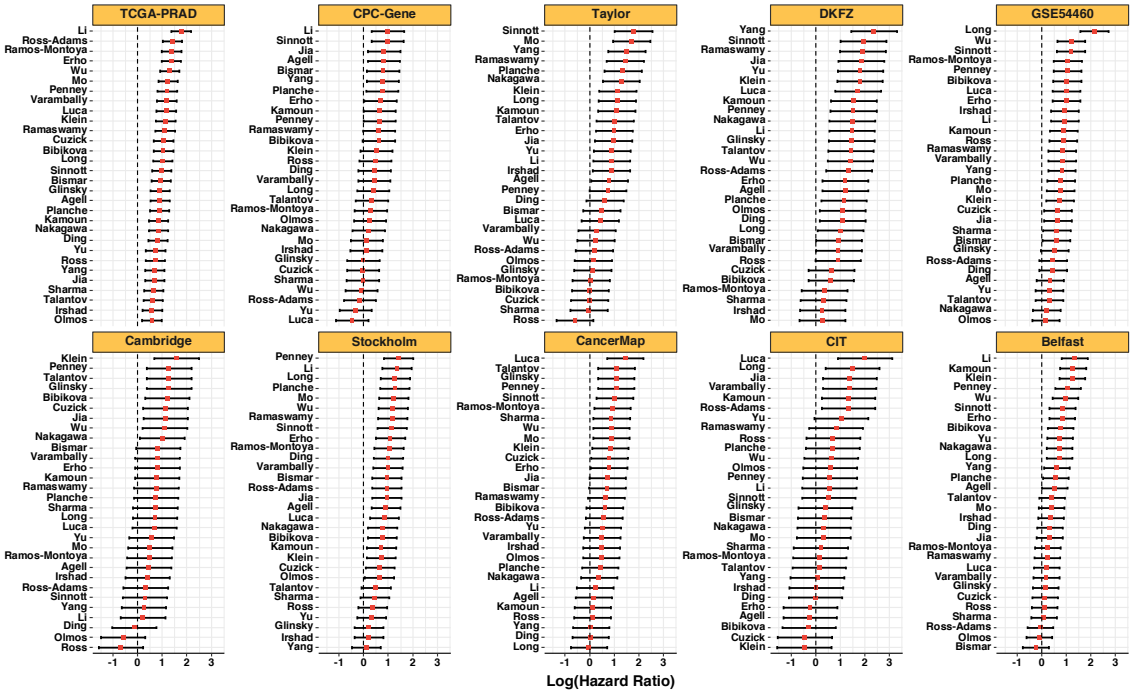


Figure 6. Intra-dataset comparison of the 30 gene expression prognostic signatures based on hazard ratio (HR) estimated by the Kaplan Meier (KM) analysis of relapse-free survival (RFS). The robust Cox-Ridge algorithm was used to train the prognostic models. The signatures are ranked by HRs in each individual dataset in the plots. The title of each small panel indicates the name of the dataset.

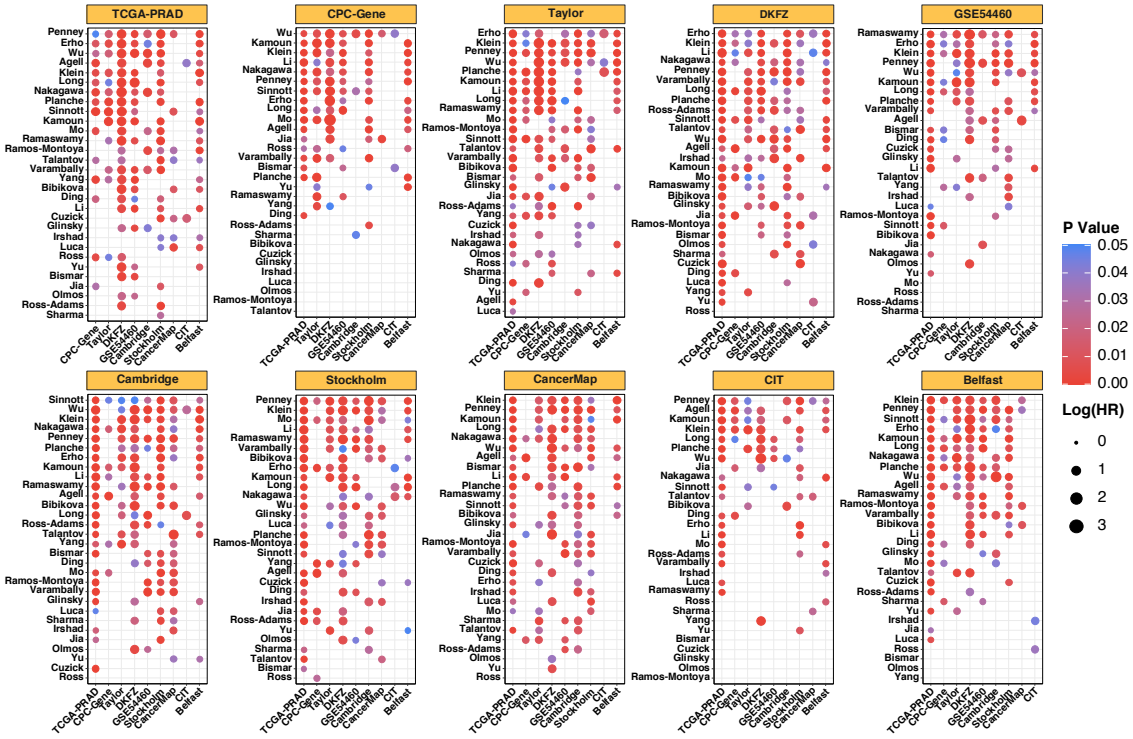


Figure 7. Inter-dataset comparison of the 30 gene expression prognostic signatures based on hazard ratio (HR) estimated by the Kaplan Meier (KM) analysis of relapse-free survival (RFS). The robust Cox-Ridge algorithm was used to train the prognostic models. The title of each small panel refers to the training dataset. The x-axis indicates the test datasets and the y-axis indicates the signatures. The prognostic signatures are ranked based on the numbers of test datasets in which the prognostic models have been successfully validated (p value < 0.05) based on KM survival analysis. Only the validations with p value < 0.05 are shown in the figure.