



Limitations of Explainability for Established Prognostic Biomarkers of Prostate Cancer

Kalifa Manjang¹, Olli Yli-Harja^{2,3,4}, Matthias Dehmer^{5,6,7} and Frank Emmert-Streib^{1,4*}

¹ Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland, ² Computational Systems Biology, Tampere University, Tampere, Finland, ³ Institute for Systems Biology, Seattle, WA, United States, ⁴ Faculty of Medicine and Health Technology, Institute of Biosciences and Medical Technology, Tampere University, Tampere, Finland, ⁵ Department of Computer Science, Swiss Distance University of Applied Sciences, Brig, Switzerland, ⁶ Department of Mechatronics and Biomedical Computer Science, University for Health Sciences, Medical Informatics and Technology (UMIT), Hall, Austria, ⁷ College of Artificial Intelligence, Nankai University, Tianjin, China

OPEN ACCESS

Edited by:

Natalia Polouliakh,
Sony Computer Science Laboratories,
Japan

Reviewed by:

Tianshou Zhou,
Sun Yat-sen University, China
Padhmanand Sudhakar,
KU Leuven, Belgium

*Correspondence:

Frank Emmert-Streib
v@bio-complexity.com

Specialty section:

This article was submitted to
Systems Biology Archive,
a section of the journal
Frontiers in Genetics

Received: 04 January 2021

Accepted: 01 June 2021

Published: 22 July 2021

Citation:

Manjang K, Yli-Harja O, Dehmer M
and Emmert-Streib F (2021)
Limitations of Explainability for
Established Prognostic Biomarkers of
Prostate Cancer.
Front. Genet. 12:649429.
doi: 10.3389/fgene.2021.649429

High-throughput technologies do not only provide novel means for basic biological research but also for clinical applications in hospitals. For instance, the usage of gene expression profiles as prognostic biomarkers for predicting cancer progression has found widespread interest. Aside from predicting the progression of patients, it is generally believed that such prognostic biomarkers also provide valuable information about disease mechanisms and the underlying molecular processes that are causal for a disorder. However, the latter assumption has been challenged. In this paper, we study this problem for prostate cancer. Specifically, we investigate a large number of previously published prognostic signatures of prostate cancer based on gene expression profiles and show that none of these can provide unique information about the underlying disease etiology of prostate cancer. Hence, our analysis reveals that none of the studied signatures has a sensible biological meaning. Overall, this shows that all studied prognostic signatures are merely black-box models allowing sensible predictions of prostate cancer outcome but are not capable of providing causal explanations to enhance the understanding of prostate cancer.

Keywords: prostate cancer, biomarkers, prognostic biomarkers, survival analysis, data science, computational biology, biostatistics

1. INTRODUCTION

Prostate cancer (PCa) is the second most prevalent cancer among men, the average age of diagnosis is 66 years, and about 60% of diagnosed cases occur in men over 65 years old. In the United States, for example, 191,930 newly diagnosis cases of PCa are estimated in 2020, resulting in about 33,330 mortalities (Siegel et al., 2020). A substantial proportion of PCa is characterized as slow-growing and indolent requiring no immediate therapeutic intervention. However, tumor stages T1 and T2, and tumor stages higher than T2 are more aggressive and invade the surrounding organs and the patient is more likely to die from the disease (Chen et al., 2020). Specifically, for men with local or regional PCa, the 5-year survival rate is almost 100%, whereas the 5-year survival rate for men with metastatic PCa is 31%.

Since the inception of high-throughput technologies, a large number of molecular markers have been described in the literature capable of distinguishing cancer patients with good and bad prognosis. Nonetheless, few found their way into clinical decision making. Many biomarker studies have used genome-wide gene expression analysis to define unique gene expression signatures related to the prognosis of PCa. For example, Chen et al. (2012) developed a 7-gene prognostic signature through a cluster-correlation analysis to identify differentially expressed genes in various cell types associated with PCa progression. Likewise, in Liu et al. (2007), the gene expression of CD44+CD24 of low tumorigenic breast and normal breast epithelium cells were compared. They used the differentially expressed genes to construct a 186-gene “invasiveness” gene signature. The signatures were tested for their association with two clinical endpoints, overall survival and metastasis-free survival, in breast and other cancer patients. Interestingly, the signature was substantially correlated with the two survival endpoints in patients with breast cancer and other types of cancer. Another study by Ramaswamy et al. (2003) examined the molecular variations between human primary tumors and metastases. The gene expression profiles of different types of adenocarcinoma metastases and unmatched primary adenocarcinomas were compared, and the analysis identified a gene expression signature capable of separating primary from metastatic adenocarcinomas (Ramaswamy et al., 2003).

There are also studies that use more advanced approaches to derive the gene signatures. In a study by Irshad et al. (2013), a 19-gene signature enriched in indolent prostate tumors was identified. Their final signature includes three genes that, through a further classification of the 19-gene signature, was established by a decision tree (DT) model. Similarly, a combination of artificial neural network analysis and data from literature search and other studies resulted in a panel of PCa progression markers, which were used in a transcriptomic analysis of 29 radical prostatectomy samples correlated with clinical outcome (Larkin et al., 2012).

Aside from such potential success stories, there are several well-known problems with prognostic signatures. One such problem relates to the stability of the selection of prognostic genes. In Michiels et al. (2005), this has been studied for various cancer types and the authors found that the size of the training data as well as the patient data both crucially effect the selection of such genes. For breast cancer, this effect has been quantified by Ein-Dor et al. (2006). Specifically, the authors showed that thousands of patient samples are needed for achieving an overlap of 50% between two predictive sets of prognostic genes. Further examples of such studies reporting similar results can be found in Kim (2009), Haury et al. (2011), and Gilhodes et al. (2017). A well-recognized study by Venet et al. (2011) addressed yet another problem by showing that many random breast cancer gene sets have similar prognostic prediction capabilities as biomarker (BM) signatures. The study by Goh and Wong (2018) extended this by removing proliferation genes. A conceptual problem of both studies is that random gene sets could still share biological similarity on the level of biological processes (BPs). The reason for this is that no systematic mechanism has been

implemented that would eliminate such a similarity. In contrast, the study by Manjang et al. (2021) introduced a gene removal procedure (GRP) that accomplished this.

The purpose of this paper is to test a hypothesis about the systems behavior of PCa. Specifically, despite well-documented differences between breast cancer and PCa, e.g., PCa affects men exclusively, whereas breast cancer commonly affects women, likewise both tumors arise in different organs involving different physiological functions, we hypothesize that their functional similarity, e.g., via the hallmarks of cancer (Hanahan and Weinberg, 2000, 2011), induces similar results for prognostic signatures. In order to investigate this, we study 32 published prognostic PCa signatures from the literature and demonstrate that random gene sets can be found with similar prediction capabilities as these signatures but opposite biological meaning.

The paper is organized as follows. In the next section, we describe our methods and data used for our analysis. Then we present and discuss our results. The paper concludes with concluding remarks.

2. MATERIALS AND METHODS

In this section, we provide information about the data and methods used for our analysis.

2.1. Biomarker Signatures

We identified reported PCa gene signatures from a literature search. From this search, we found 32 signatures from 31 studies that have been published between 2002 and 2020. For all signatures, the Entrez gene IDs corresponding to the HGNC gene symbols are determined. All genes without an associated Entrez gene ID are discarded. **Table 1** shows an overview of the published gene signatures we use for our study.

2.2. Gene Expression Data

We collected RNA-seq data (HTSeq-FPKM and HTSeq-FPKM-UQ) of patients with PCa from the TCGA-PRAD project. We obtained the data from the UCSC Xena GDC data hub (<https://xenabrowser.net/datapages/>) on September 7, 2020. FPKM stands for *Fragments Per Kilobase of transcript per Million* mapped reads (Trapnell et al., 2010). It accounts for a situation in which only 1 end of a pair-end read is mapped. The FPKM of a gene is estimated as follows:

$$FPKM = \frac{10^9 \times \text{number of reads mapped to the gene}}{(\text{number of reads mapped to all protein-coding genes} \times \text{length of the gene in base pairs})} \quad (1)$$

Similarly, FPKM-UQ means Fragments Per Kilobase of transcript per Million mapped reads upper quartile. It is a modified estimate of FPKM where the total protein-coding read count is replaced by the 75th percentile read count for a sample. A notable difference between the two is the values of FPKM-UQ tends to be much higher due to the significant disparity between the total mapped number of reads in an alignment and the mapped number of reads to one gene.

The gene expression data set used in our study contains 551 samples, of which 498 are primary solid tumors, 52 are solid tissue normal, and one is metastatic. We exclude the metastatic and solid tissue normal samples from the data set. From these data, we used only protein-coding genes without missing information for the HTSeq-FPKM data cohort. Likewise, from the HTSeq-FPKM-UQ data we used only genes with < 2% missing information across all samples. The final HTSeq-FPKM data set contains 498 samples and 16,428 genes, whereas the HTSeq-FPKM-UQ data set contains 498 samples and 15,165 genes. Lastly, patient survival information for each sample was derived from Liu et al. (2018). Specifically, we used the progression-free interval end-points. In this paper, we refer to the HTSeq-FPKM and HTSeq-FPKM-UQ in our analysis as GDC cohort A and GDC cohort B, respectively.

2.3. Outcome Association

In order to determine the prognostic importance of a random gene set, we perform a survival analysis. We estimate Kaplan-Meier survival curves and compare these with a Mantel-Haenszel test (Emmert-Streib and Dehmer, 2019). That means each comparison provides a *p*-value from such a hypothesis test.

The patients are stratified into two classes (low and high risk) by using the PC1 method. This method categorizes patients according to a particular gene set. Hence, the resulting survival analysis is a function of the gene set used to categorize the patients. Overall, our study consists of three main steps: first, the selection/construction of random gene set; second, the classification of patients samples; and third, the survival analysis.

In the next section, we explain our method we use as GRP for constructing random gene sets.

2.4. Gene Removal Procedure

Our GRP entails the removal of both the BM signatures and genes that belong to the same BPs as the genes in the BM signatures. The gene ontology (GO) is hierarchical (Ashburner et al., 2000). Hence, we approach this analysis iteratively by removing genes of BPs successively from the same hierarchy level. The GRP we use is defined as follows:

1. *G*: the genes in the PCa data set (16,425 and 15,165 for GDC cohort A and B, respectively).
2. $BM_i : g_i, \dots, g_m$. BM_i is the gene signature i (i range from 1 to 32) and g_i, \dots, g_m are the genes in the respective signatures.
3. Removing biomarker genes in signature BM_i from *G*. This produces a new set of genes G'_i with $G'_i = G \setminus BM_i$.
- 3* Optional step: Remove the proliferation genes, PG from *G*. This gives a new set of genes G_i^* with $G_i^* = G' \setminus PG$.
4. Map the genes in BM_i to GO-terms and the corresponding hierarchy levels. This gives: $BM_i = \{g_i, \dots, g_m\} \rightarrow R = \{(GO_1, L_1), \dots, (GO_t, L_t)\}$ (Manjang et al., 2020).
 Note, each gene can be connected to more than one GO-term. For this reason, $m \leq t$.
5. Map each GO-term in *R*, i.e., GO_i with $i \in \{1, \dots, t\}$, to its gene set GS_i .
6. For each biomarker set i : Loop-over the elements in set *R*.

TABLE 1 | Overview of published and evaluated prognostic signatures for prostate cancer used for our study.

Acronym of a study	Number of genes*	Cancer type	Reference
AGELL	12	Prostate cancer	Agell et al., 2012
BIBIKOVA	16	Prostate cancer	Bibikova et al., 2007
BISMAR	12	Prostate cancer	Bismar et al., 2006
CHEN	4	Prostate cancer	Chen et al., 2020
CHEN_CC	7	Prostate cancer	Chen et al., 2012
CHEVILLE	2	Prostate cancer	Cheville et al., 2008
CHU	8	Prostate cancer	Chu et al., 2018
CUZICK	31	Prostate cancer	Cuzick et al., 2011
GLINSKY	11	Multiple cancers	Glinsky et al., 2005
IRSHAD	19	Prostate cancer	Irshad et al., 2013
IRSHAD_1	3	Prostate cancer	Irshad et al., 2013
LARKIN	7	Prostate cancer	Larkin et al., 2012
LI	6	Prostate cancer	Li et al., 2019
LIU	167	Multiple cancers	Liu et al., 2007
LONG	12	Prostate cancer	Long et al., 2011
NAKAGAWA	17	Prostate cancer	Nakagawa et al., 2008
PENNEY	157	Prostate cancer	Penney et al., 2011
RAMASWAMY	16	Multiple cancers	Ramaswamy et al., 2003
REDDY	16	Prostate cancer	Reddy and Balk, 2006
ROSS-ADAMS	100	Prostate cancer	Ross-Adams et al., 2015
ROSS	6	Prostate cancer	Ross et al., 2012
SAAL	162	Multiple cancers	Saal et al., 2007
SHARMA	15	Prostate cancer	Sharma et al., 2013
SINGH	5	Prostate cancer	Singh et al., 2002
SONG	15	Prostate cancer	Song et al., 2019
STEPHENSON	10	Prostate cancer	Stephenson et al., 2005
TALANTOV	3	Prostate cancer	Talantov et al., 2010
TANDEFELT	36	Prostate cancer	Tandefelt et al., 2013
TRUE	86	Prostate cancer	True et al., 2006
WANG	43	Prostate cancer	Wang et al., 2017
WU	29	Prostate cancer	Wu et al., 2013
YU	14	Multiple cancers	Yu et al., 2007

Number of genes corresponds to the number of genes in a signature after conversion from HGNC gene symbols to Entrez gene IDs.*

- a. Delete all the genes associated with the GO-terms in set *R*. This results in a new set given by $G'' = G' \setminus D$, where *D* is the set of genes having GO-terms in *R*, i.e., $D = \cup_{i=1}^t GS_i$.
7. From G'' , we loop from 1 to 1,000:
 - a. We sample new sets of biomarker genes of size $|BM_i \in G|$ and perform the prognostic task. We repeat this for 1,000 times.
 - b. Application of a Bonferroni correction to the *p*-values.
 - c. Set $G' = G''$. Stop if $l = L_{min}(i)$ or $|G''| < 2 \times |BM_i \in G|$.

In the above procedure, the optional step called 3* involves the removal of the 664 genes that are related to proliferation (this gene set is called PG). We extracted the genes in PG from Goh and Wong (2018).

The prediction results are assessed using the p -values obtained from the survival analysis. We call a random gene set with a significant p -value, a *surrogate gene set*.

2.5. Unsupervised Classification

The patient samples are categorized using the PC1 stratification method, which is based on a principal component analysis (PCA). Briefly, PCA is a dimension reduction technique (this involves reducing the size of the data set). The goal is to transform a large data set into a smaller ones having a lower dimensional representation. This method trades a little accuracy for simplicity, thus achieving interpretability as well as minimal loss of information (Lever et al., 2017). For performing the PC1 method, we use the R function "prcomp" to obtain the first principal component (PC1) of a signature. The patients are then divided into two groups according to the median of the PC1, i.e., a sample is either categorized as group -1 if the PC1 is below the median or as group $+1$ if the PC1 is above the median value. Hence, the PC1 method is used to classify (or group) the patients into two classes, whereas this separation depends on a signature gene set. This approach has been previously used (see, e.g., Venet et al., 2011).

Formally, our analysis is based on a gene expression matrix of the form $X \in \mathbb{R}^m \times \mathbb{R}^n$, where m is the number of genes and n is the number of samples. Importantly, here m corresponds to the number of genes in a particular signature gene signature and not to all genes that are available in a data set.

2.6. Survival Analysis

For evaluating the prognostic value of gene sets, we conduct a survival analysis. Specifically, we estimate a Kaplan–Meier survival curve for each patient group and compare different groups with the Mantel–Haenszel test (Emmert-Streib and Dehmer, 2019). Hence, each comparison is characterized by a p -value resulting from a statistical hypothesis test. For the survival analysis, we use the progression-free interval as endpoint.

We would like to remark that due to the fact that the PC1 method provides a categorization of the patients, the resulting survival analysis depends on the gene set used for obtaining the first principal component of the signature.

2.7. Measuring of Biological Meaning

In order to have a well-defined meaning of the term “biological meaning,” we use information from the GO (Ashburner et al., 2000). Specifically, GO defines the biological meaning of a gene by a list of GO-terms associated with this gene. For a list of genes, the biological meaning of this set can be defined by the union of the sets of GO-terms of the individual genes. For instance, given three genes, g_1, g_2, g_3 , with associate GO-terms the biological meaning (M) of these genes is given by

$$M(g_1) = \{GO_1(1), GO_1(2), \dots GO_1(m)\} \tag{2}$$

$$M(g_2) = \{GO_2(1), GO_2(2), \dots GO_2(n)\} \tag{3}$$

$$M(g_3) = \{GO_3(1), GO_3(2), \dots GO_3(o)\} \tag{4}$$

with $m, n, o \in \mathbb{N}$. Here, the GO-terms are from a category, e.g., BP. Similarly, the biological meaning of the set of genes $\{g_1, g_2, g_3\}$

is given by

$$M(\{g_1, g_2, g_3\}) = M(g_1) \cup M(g_2) \cup M(g_3) \tag{5}$$

whereas \cup is the union of the individual sets. Hence, the biological meaning of $\{g_1, g_2, g_3\}$ is given by the set of GO-terms $M(\{g_1, g_2, g_3\})$.

From this follows that, e.g., the similarity of two sets of genes, $\{g_1, g_2, g_3\}$ and $\{g_4, g_5, g_6\}$, is zero if

$$M(\{g_1, g_2, g_3\}) \cap M(\{g_4, g_5, g_6\}) = \emptyset. \tag{6}$$

Importantly, our GRP defined above constructs random gene sets (RGS) with this property, i.e.,

$$M(RGS) \cap M(BM) = \emptyset \tag{7}$$

with RGS a set of random genes and BM a set of biomarker genes.

3. RESULTS

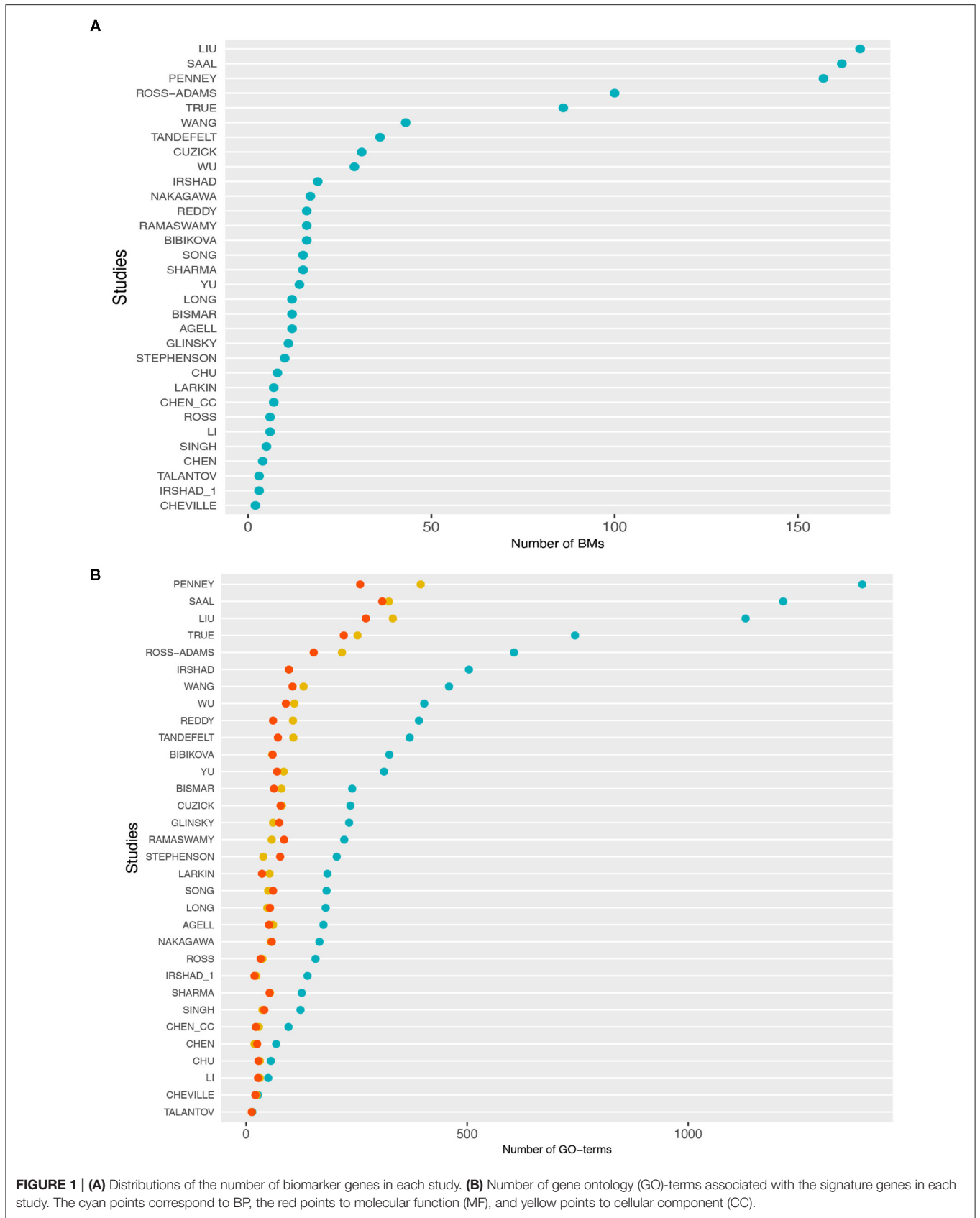
In this section, we present the results of our analysis. First, we study published prognostic biomarkers of PCa individually and comparatively. Then we study random gene set and show results for prognostic outcome.

3.1. Prognostic Biomarkers of Prostate Cancer

3.1.1. Size of Biomarker Sets and GO-Terms in Signatures

In **Table 1**, we show an alphabetically ordered overview of all 32 prognostic BM signatures included in our analysis. The smallest signature is from Cheville consisting of 2 genes only, whereas the signature from Liu is the largest containing 167 genes. Interestingly, there are some signatures that have the same number of genes. Specifically, the signatures of Irshad_1 and Talantov have 3 genes, the signatures of Li and Ross have 6 genes, the signatures of Chen_cc and Larkin have 7 genes, the signatures of Agell, Bismar, and Long have 12 genes, the signatures of Sharma and Song have 15 genes, and the signatures of Bibikova, Ramaswamy, and Reddy have 16 genes in their BM sets. An overall summary of the size distributions of all BM signatures is shown in **Figure 1A**.

In **Figure 1B**, we show information about the GO-terms associated with the genes in the signatures. The three colors correspond to the three GO categories: BP shown in cyan, molecular function (MF) shown in red, and cellular component (CC) shown in yellow. For each of these three categories, we show the absolute number of GO-terms in each study. Overall, from **Figure 1B** one can see that the present GO-terms in the signatures differ significantly from each other. That means some signatures are very specific because they contain only a very small number of different GO-terms, e.g., the signatures from Talantov, Cheville and Li, while others are rather generic containing many GO-terms, e.g., Penney, Liu and Saal. For GO-terms of BP (cyan), this variation is particularly large.



3.1.2. Pairwise Similarity of Signatures

In order to study differences between the 32 signatures, we conduct a pairwise comparison of these BM sets. Specifically, we study two different types of overlap. We study (i) the number of common genes and (ii) the number of common GO-terms among pairs of signatures. Formally, we define these two overlap measures as follows. Let S_i and S_j be two signature sets consisting either of genes or GO-terms. Then we find the percentage $z_i \in [0, 1]$ of common elements in S_i that are also present in S_j by

$$x_i = S_i \cap S_j \quad (8)$$

$$z_i = \frac{|x_i|}{|S_i|} \quad (9)$$

Here, z_i can assume values between zero and one and $|z|$ corresponds to the number of elements in set z . We would like to remark that the way we define the overlap is asymmetric, i.e., $z_i \neq z_j$ if $|S_i| \neq |S_j|$. That means the percentage of the overlap is taken with respect to the first signature set S_i on the right-hand side of Equation (8).

The two heatmaps in **Figures 2A,B** show the results of this analysis. From this analysis of the gene overlap, we find that the signatures of Chen_cc and Chu do not overlap with other signatures at all, i.e., both have a zero overlap with any other signature. This implies that the genes in the signatures of Chen_cc and Chu are unique concerning the genes in their corresponding BM sets. Every other BM signature has at least some overlap with another signature; see the last column in **Figure 2A** (red numbers) providing information about the number of signatures with a non-zero overlap.

The signature of Chevillat, which has the smallest number of genes, has a gene overlap with the three signatures of Cuzick, Li, and Penney. Surprisingly, the signature of Liu, which contains the highest number of genes, has only genes in common with 9 other signatures. Irshad_1 is the only signature that completely overlaps with another signature (Irshad); however, we would like to note that both signatures are from the same study (Irshad et al., 2013). Finally, we find that the signature of Penney has the highest gene overlap with other signatures (it has genes in common with 20 signatures). From this analysis, we see that there is a wide range of behaviors for the gene overlap reaching from zero overlap (for Chen_cc and Chu) to an overlap with 20 signatures (for Penny) corresponding to an overlap with 64.5% (= 20/31) of all signatures. This implies that all signatures are unique to a certain extent because this percentage would be much higher.

In contrast to these findings, **Figure 2B** shows the overlap of GO-terms among the signatures. Again, the overlap between the signatures varies considerably. For instance, the signatures of Saal and Penney share the highest overlap with 490 GO-terms. Interestingly, all the signatures have a non-zero overlap in their biological meaning.

Importantly, a difference to the gene overlap (see **Figure 2A**) is that for a GO-term overlap, all signatures share at least one GO-term with 26 other signatures (see last column in **Figure 2B**) and most signatures (25) have at least one common GO-term with all other signatures. This implies that on a GO-term level, the

signatures are much more similar to each other than on a gene level. This underlines the importance of a systems-view on PCa.

3.2. Prediction Abilities of Random Gene Signatures

Next, we systematically investigated the prognostic prediction capabilities of the 32 BM signatures and random gene sets. We begin by systematically removing BM signature genes from the available gene expression gene pool. Subsequently, we also omit hierarchically genes that share a biological meaning with the respective published signatures. We randomly sample 1,000 set of the same size as the BM signature from the gene set left to create random gene signature. The results are as follows:

The outcome of the study is given in three parts. First, from the gene pool, we systematically remove the published signatures and the genes that share a similar biological meaning with them and compute the outcome association. Next, we correct the obtained p -values by conservative Bonferroni correction and report the results. And finally, the analysis is repeated by omitting the proliferation genes from the gene pool, we correct the p -values by conservative Bonferroni correction, and present the results.

3.2.1. GDC Cohort a Data

The results for the GDC cohort A data are shown in **Figure 3**. The light/dark red points represent the outcome of the published signatures (without any gene removal), whereas light red indicates significant results and dark red non-significant one. The blue colored distributions are the result of random gene sets, whereas the shaded cyan bars correspond to the lower third percentile of the distributions and the bold black points are the median values of these distributions. The blue vertical line corresponds to a significant level of $\alpha = 0.05$. We would like to note that the p -values are on a logarithmic scale (i.e., \log_{10}).

First, from **Figure 3** we observed that not all published signatures (red points) lead to significant results. In order to highlight this, we show significant results by points in light red, whereas non-significant results are shown in dark red. A possible reason for this observation is that our analysis uses a different data set than the original studies and, hence, the observed results indicate to the well-known instability of biomarkers (lack of robustness) (Drier and Domany, 2011). Specifically, for our analysis 24 of the 32 biomarker signatures are significant and the remaining published signatures lack robustness for the independent validation data set.

Figure 3A shows results without a Bonferroni correction. This analysis is similar to the study by Venet et al. (2011), which also did not use a multiple testing correction even though many comparisons were conducted. Interestingly, in **Figure 3A** all lower third percentiles (cyan shaded bars) are significant. That means for all random gene sets we find at least 3% of these to be significant. When compared to the published signatures (red points), the lower third percentile of random gene sets outperform even 26 signatures. Five published signatures performed as well as the lower third percentile of random sets, or the random sets slightly outperformed them. Only one signature (Song) achieves a more significant outcome than the lower third percentile of the random gene sets. Two signatures Ross and

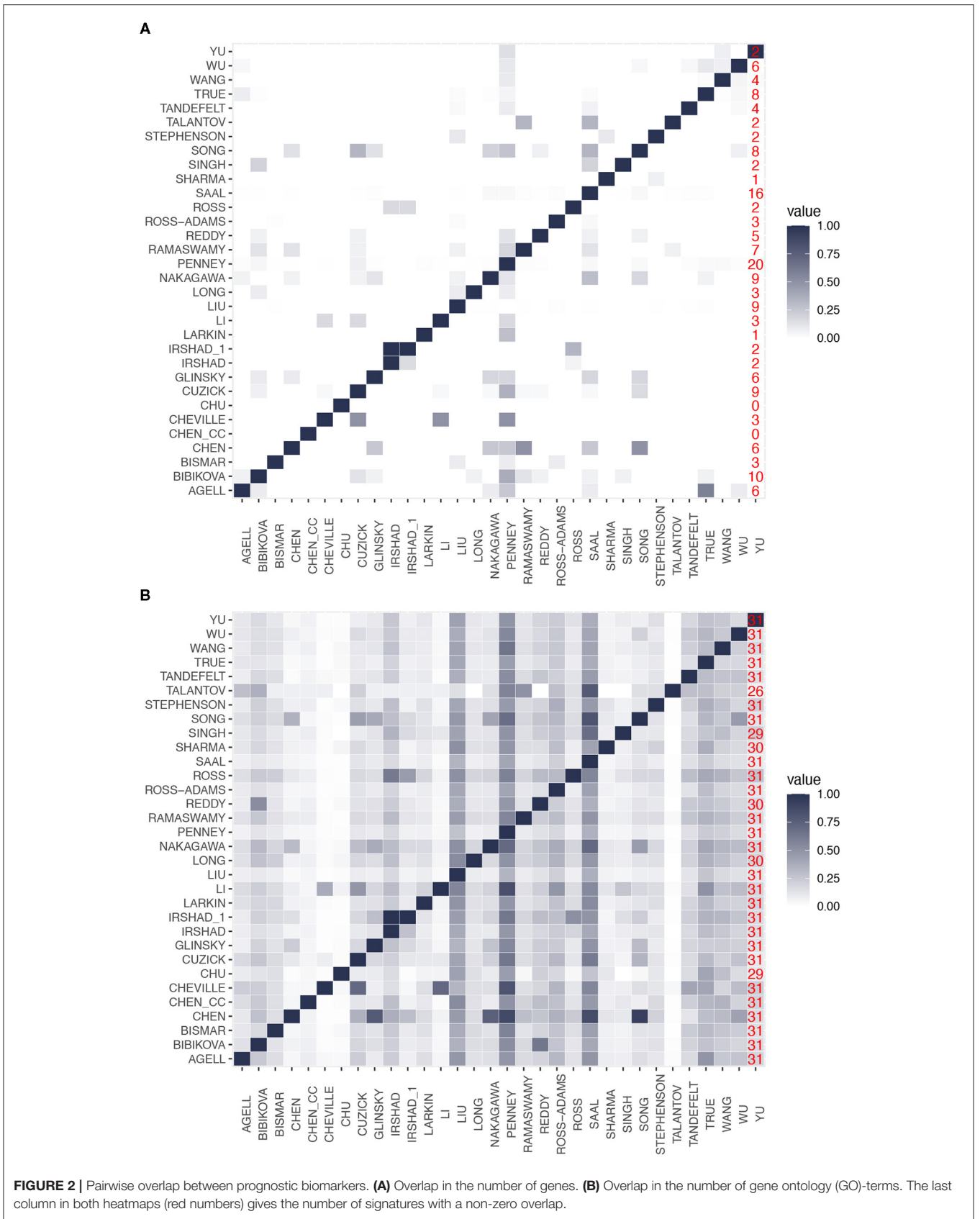
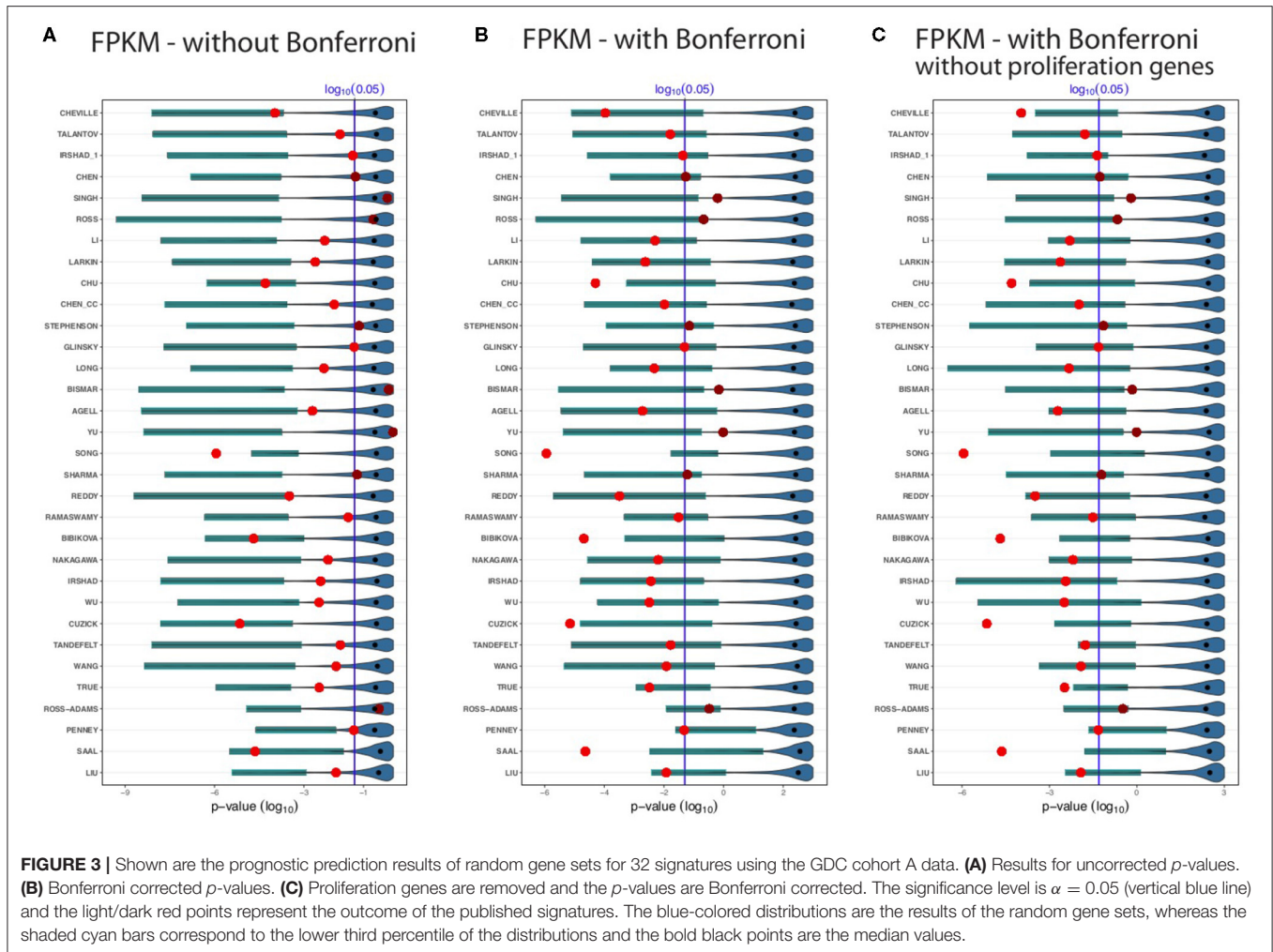


FIGURE 2 | Pairwise overlap between prognostic biomarkers. **(A)** Overlap in the number of genes. **(B)** Overlap in the number of gene ontology (GO)-terms. The last column in both heatmaps (red numbers) gives the number of signatures with a non-zero overlap.



Ross-Adams perform as worse as the median of the random sets and three signatures (Singh, Bismar, and Yu) perform even worse than the median of the random gene sets. The median of the random sets (bold black points) are all non-significant.

In **Figure 3B**, we repeated the analysis applying a conservative Bonferroni correction. With a Bonferroni correction, four signatures (Singh, Ross, Bismar, and Yu) performed worse than the lower third percentile of random signatures. Likewise, five published signatures, Chu, Song, Bibikova, Cuzick, and Saal, outperformed the random signatures. As one can see from **Figure 3B**, not all the lower third percentile are significant. However, for all random signatures (such as Penney and Liu), we find at least some significant random signatures. Interestingly, many smaller random signatures perform better in comparison to larger ones. For instance, Cheville, Talantov, Irshad_1, Chen, Singh, etc., all performed better than the top 5 largest signatures (True, Ross-Adams, Penney, Saal, and Liu).

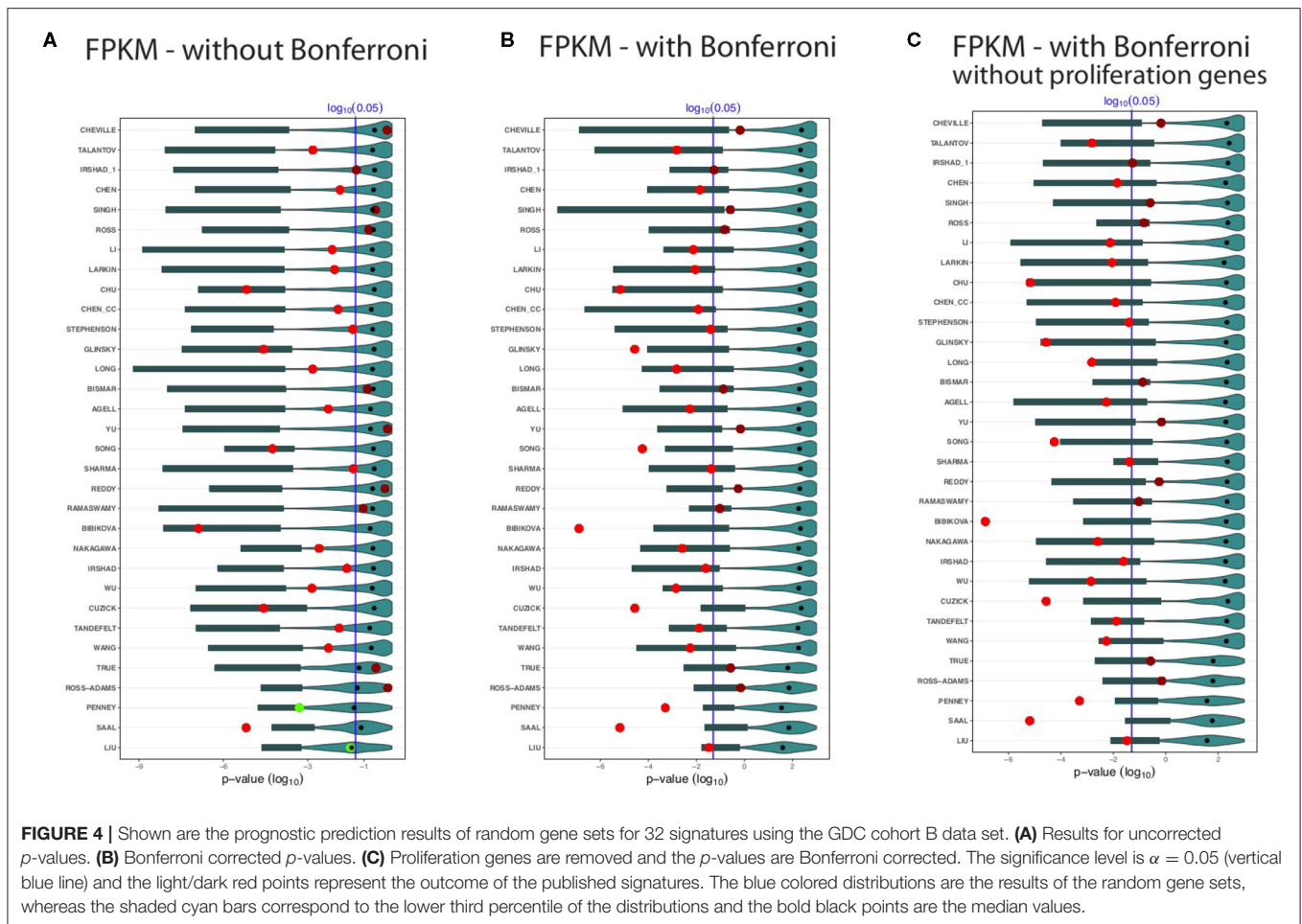
In a previous breast cancer study (Goh and Wong, 2018), it has been found that the removal of proliferation genes from random signatures leads to diminishing results of the prognostic performance of random signatures. In order to study this effect,

we removed additionally all proliferation genes from the gene pool and repeated our analysis with a Bonferroni correction. The results of this are shown in **Figure 3C**. Qualitatively, the results in **Figures 3B,C** are similar. Overall, for all results in **Figures 3A–C**, one can see that for all random signatures there are at least some that are statistically significant. We would like to emphasize that all random gene sets share *per construction* no biological meaning with the published signatures yet can perform prognosis as well as the BM signatures or better.

3.2.2. GDC Cohort B Data

In order to study the influence of the data processing, we repeat our analysis for the GDC cohort B data. The results of this analysis are shown in **Figures 4A–C**. In these figures, there are in addition to the dark and light red points, light green points indicate the BM signatures. These correspond to significant BM signatures, whereas the median values of the random gene sets (black points) are non-significant.

Again, we observe that not all BM signatures lead to a significant outcome. Specifically, we find 22 of the 32 signatures to be significant (**Figure 4**). Interestingly, we find also non-robust



results. For instance, Cheville, Irshad_1, Reddy, Ramaswamy, and True failed to predict the outcome in the GDC cohort B data set, but these signatures were significant for the GDC cohort A data (see **Figure 3**). Similarly, Chen, Stephenson, and Sharma are significant for GDC cohort B (see **Figure 4**) but not GDC cohort A (see **Figure 3**).

Also for the distributions of the results for the random gene sets, we observe very similar results as for the GDC cohort A data in **Figure 3**. Hence, overall, the results in **Figures 4A–C** confirm our analysis, which means there are always random gene sets leading to significant results.

4. DISCUSSION

Our hypothesis for the present study was that prognostic signatures of prostate cancer are lacking a sensible biological meaning. In order to investigate this, we used a GRP introduced in Manjang et al. (2021). This GRP allows to systematically construct random gene sets by omitting all biological similarities between published signatures and the genes in a gene pool from which random gene sets are drawn. These random gene sets are not assigned any

particular (biological) meaning or role. Importantly, such random gene sets do not necessarily have predictive capabilities as assessed by predicting progression-free survival as outcome variable. For this reason, we distinguish between random gene sets that are predictive (indicated by a significant p -value from a survival analysis) and non-predictive by calling the former ones surrogate gene sets. A published BM signature (see **Table 1**), on the other hand, is a gene set that is obtained in a targeted and non-random manner indicative of disease progression.

For testing our hypothesis, we studied 32 published BM signatures of prostate cancer from the literature (see **Table 1**). As a result, for all studied 32 signatures we found random gene sets with better or similar prognostic capabilities but no overlap in the biological meaning. In order to see if the preprocessing of the data has any effect on this, we extended our analysis by examining the effect of different data processing techniques. Regarding this, we conducted further analysis by using a data set with different processing methods applied to the raw data leading to a data set we call GDC cohort B. As a result from this analysis, we found no systematic influence of a particular data processing technique on the surrogate gene sets or the overall results (see **Figures 3, 4**). Finally, we also removed proliferation

genes (for both data sets, i.e., GDC cohort A and GDC cohort B) and found also for this setting no difference in our results (see **Figures 3C, 4C**).

As a conclusion from all these analyses, we can infer that any biological rationale provided for selecting the genes in the published gene signatures, as shown in **Table 1**, is anecdotal. This is taking into account the meaning of random gene sets arising from the GRP because the used GRP eliminates the risk of accidentally selecting genes for a random gene set that have the same biological meaning as the published gene signatures. Consequently, due to the discovery of surrogate gene sets with the same predictive capability but a completely distinct biological interpretation, as a result of the zero overlap in the GO-terms of the genes involved, any biological significance attributed to such BM signatures is required.

Interestingly, a similar interpretation has been found in a breast cancer study by Manjang et al. (2021). They showed that when all signs of the biological meaning of the BM signature genes are removed, surrogate gene sets can be determined among the remaining random gene sets with similar prognostic predictive capabilities but with contrasting biological meaning. Therefore, the research findings indicated that with regard to disease etiology, none of the studied signatures have a plausible biological interpretation or significance. The study concluded that prognostic signatures are black-box models that can yield accurate predictions of breast cancer outcome but with no benefit for disclosing causal, biological relations. Furthermore, this study also noted a relationship between the predictive accuracy and the size of the random gene sets by showing that the accuracy is higher for larger gene sets. It is interesting to note that in the current study, we could not establish this relationship. A possible explanation for this may be the relatively small size of published BM signatures of prostate cancer, which are all smaller than 200 genes (see **Figure 1**). In contrast, the breast cancer signatures studied in Manjang et al. (2021) are much larger in average reaching up to 1345 genes.

It is important to note that a similar study for breast cancer by Venet et al. (2011) has been unable to arrive at this conclusion since no GRP was used. As a consequence, BM signatures as well as genes from associated BP were not removed leaving the possibility to inadvertently select random genes with a common biological meaning as the original BM signatures because these genes belong to the same BPs as indicated by common GO-terms in the domain BP. Another statement by Venet et al. (2011) is that *most random signatures are significantly associated with prognostic outcome*. With respect to prostate cancer, this holds only for the random gene sets of Penney and Liu (see **Figure 4A**) because 50% of the surrogate gene sets are significant as indicated by the median values of the distributions (black points in **Figure 4**). However, generally, this assertion is not valid and only applies to some signatures.

To date, many studies investigated prognostic signatures of prostate cancer. For example, Bibikova et al. (2007) used a 16-gene expression signature to predict the prognosis of prostate cancer. They complemented their results by a discussion of the functional annotation of these genes, which were involved in proliferation, cell cycle, differentiation, signal transduction and

basic metabolism. Similarly, the studies by Saal et al. (2007), Sharma et al. (2013), and Song et al. (2019) argued that the biological importance of their prognostic signatures is based on the role of PI3K pathway signaling, altered signaling, P53 signaling and cell cycle process pathway respectively. In this paper, we studied those and other prognostic signatures of prostate cancer. Our results, however, demonstrate that such biological interpretations do not offer a causal explanation for the fundamental biology of prostate cancer since we can always find surrogate gene sets with no biological relationship to those signatures but similar or better prognostic prediction capabilities.

Considering that prostate cancer and breast cancer are two considerably different diseases yet our results demonstrate a similarity in the lack of biological meaning of both cancers one may wonder if there is a common factor giving raise to these findings. This is very difficult to answer, however, one common factor that comes to mind are the hallmarks of cancer (Hanahan and Weinberg, 2000). Specifically, the study by Hanahan and Weinberg (2000) highlighted six hallmarks of cancer (self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis), which are shared by all types of human cancers. Later this has been extended by four further hallmarks (deregulating cellular energetics, avoiding immune destruction, genome instability and mutation, tumor-promoting inflammation) (Hanahan and Weinberg, 2011). If our findings are actually related to the ten hallmarks of cancer is currently unclear. However, it seems not implausible to assume that there might be a connection because the hallmarks state that cancer is a system disease involving a multitude of pathways. We want to add that these pathways do not work in isolation but are connected among each other by intricate regulatory networks (Emmert-Streib et al., 2014).

On a technical note, we would like to remark that there could be other metrics for evaluating the prediction capabilities of random gene sets other than p -values. For instance, one could use information from pathology about disease states, which allow to use error measures for binary classifications. While this establishes sensible metrics, e.g., F-score or AUROC, such measures do not directly utilize survival information about the progression of patients. Instead, this is the strength of survival analysis comparing survival curves quantitatively. Hence, a regression framework, as provided by survival analysis (Kleinbaum and Klein, 2005), seems to be favorable over a classification framework allowing a more nuanced evaluation.

Finally, we would like to note that our study has similarities to recent investigations in Explainable Artificial Intelligence (XAI) (Xu et al., 2019; Emmert-Streib et al., 2020). Specifically, XAI explores the dichotomy of predictive and descriptive models (Emmert-Streib and Dehmer, 2021) in AI and aims to establish mechanisms for making predictive models also explainable in a sense that this can enhance our understanding of a system under investigation. On a wider scope, this discussion has a long history in the statistics community and refers to the distinction of black-box models and causal models (Holland, 1986; Breiman, 2001). Our study shows that prognostic biomarkers of prostate

cancer allow sensible predictions for cancer progression but do not establish a causal understanding with respect to the biological meaning of such prognostic signatures. Here, it is important to extend the considerations to the proposed gene selection mechanisms used by studies identifying prognostic signatures (see **Table 1**). Overall, such models have a predictive utility, e.g., for applications in the clinical practice but no biological utility for enhancing our understanding of cancer biology.

5. CONCLUSION

In this paper, we scrutinized the biological meaning of prognostic signatures of prostate cancer. Our study utilized a GRP that results in random gene sets without any overlap in the biological meaning with biomarker signatures yet a non-vanishing proportion of these random gene sets, called surrogate gene sets, achieve similar prediction results. Hence, our results demonstrate that none of the studied signatures of prostate cancer has a sensible biological interpretation with respect to disease etiology. To our knowledge, this is the

first study providing such results for prognostic biomarkers of prostate cancer.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://xenabrowser.net/datapages/>.

AUTHOR CONTRIBUTIONS

FE-S conceived the study. KM performed the analysis. KM and FE-S analyzed the data and interpreted the results. All authors wrote the manuscript.

FUNDING

KM has been supported by a fellowship from the Center for Prostate Cancer, Tampere University. MD thanks the Austrian Science Funds for supporting this work (project P30031).

REFERENCES

- Agell, L., Hernández, S., Nonell, L., Lorenzo, M., Puigdecamet, E., de Muga, S., et al. (2012). A 12-gene expression signature is associated with aggressive histological in prostate cancer: Sec14l1 and tceb1 genes are potential markers of progression. *Am. J. Pathol.* 181, 1585–1594. doi: 10.1016/j.ajpath.2012.08.005
- Ashburner, M., Ball, C., Blake, J., Botstein, D., and Butler H., et al. (2000). Gene Ontology: tool for the unification of biology. The Gene ontology consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556
- Bibikova, M., Chudin, E., Arsanjani, A., Zhou, L., Garcia, E. W., Modder, J., et al. (2007). Expression signatures that correlated with gleason score and relapse in prostate cancer. *Genomics* 89, 666–672. doi: 10.1016/j.ygeno.2007.02.005
- Bismar, T. A., Demichelis, F., Riva, A., Kim, R., Varambally, S., He, L., et al. (2006). Defining aggressive prostate cancer using a 12-gene model. *Neoplasia* 8:59. doi: 10.1593/neo.05664
- Breiman, L. (2001). Statistical modeling: the two cultures. *Stat. Sci.* 16, 199–231. doi: 10.1214/ss/1009213726
- Chen, X., Wang, J., Peng, X., Liu, K., Zhang, C., Zeng, X., et al. (2020). Comprehensive analysis of biomarkers for prostate cancer based on weighted gene co-expression network analysis. *Medicine* 99:e19628. doi: 10.1097/MD.00000000000019628
- Chen, X., Xu, S., McClelland, M., Rahmatpanah, F., Sawyers, A., Jia, Z., et al. (2012). An accurate prostate cancer prognosticator using a seven-gene signature plus gleason score and taking cell type heterogeneity into account. *PLoS ONE* 7:e45178. doi: 10.1371/journal.pone.0045178
- Chevillat, J. C., Karnes, R. J., Therneau, T. M., Kosari, F., Munz, J.-M., Tillmans, L., et al. (2008). Gene panel model predictive of outcome in men at high-risk of systemic progression and death from prostate cancer after radical retropubic prostatectomy. *J. Clin. Oncol.* 26:3930. doi: 10.1200/JCO.2007.15.6752
- Chu, J., Li, N., and Gai, W. (2018). Identification of genes that predict the biochemical recurrence of prostate cancer. *Oncol. Lett.* 16, 3447–3452. doi: 10.3892/ol.2018.9106
- Cuzick, J., Swanson, G. P., Fisher, G., Brothman, A. R., Berney, D. M., Reid, J. E., et al. (2011). Prognostic value of an rna expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. *Lancet Oncol.* 12, 245–255. doi: 10.1016/S1470-2045(10)70295-3
- Drier, Y., and Domany, E. (2011). Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? *PLoS ONE* 6:e17795. doi: 10.1371/journal.pone.0017795
- Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proce. Natl. Acad. Sci. U.S.A.* 103, 5923–5928. doi: 10.1073/pnas.0601231103
- Emmert-Streib, F., de Matos Simoes, R., Mullan, P., Haibe-Kains, B., and Dehmer, M. (2014). The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Front. Genet.* 5:15. doi: 10.3389/fgene.2014.00015
- Emmert-Streib, F., and Dehmer, M. (2019). Introduction to survival analysis in practice. *Mach. Learn. Knowl. Extract.* 1, 1013–1038. doi: 10.3390/make1030058
- Emmert-Streib, F., and Dehmer, M. (2021). Data-driven computational social network science: predictive and inferential models for web-enabled scientific discoveries. *Front. Big Data* 4:591749. doi: 10.3389/fdata.2021.591749
- Emmert-Streib, F., Yli-Harja, O., and Dehmer, M. (2020). Explainable artificial intelligence and machine learning: a reality rooted perspective. *WIREs Data Min. Knowl. Discov.* 10:e1368. doi: 10.1002/widm.1368
- Gilhodes, J., Zemmour, C., Ajana, S., Martinez, A., Delord, J.-P., Leconte, E., et al. (2017). Comparison of variable selection methods for high-dimensional survival data with competing events. *Comput. Biol. Med.* 91, 159–167. doi: 10.1016/j.compbiomed.2017.10.021
- Glinsky, G. V., Berezovska, O., Glinskii, A. B., et al. (2005). Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *J. Clin. Invest.* 115, 1503–1521. doi: 10.1172/JCI23412
- Goh, W. W. B., and Wong, L. (2018). Why breast cancer signatures are no better than random signatures explained. *Drug Discov. Today* 23, 1818–1823. doi: 10.1016/j.drudis.2018.05.036
- Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi: 10.1016/S0092-8674(00)81683-9
- Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013
- Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE* 6:e28210. doi: 10.1371/journal.pone.0028210
- Holland, P. (1986). Statistics and causal inference. *J. Am. Stat. Assoc.* 81, 945–960. doi: 10.1080/01621459.1986.10478354
- Irshad, S., Bansal, M., Castillo-Martin, M., Zheng, T., Aytes, A., Wenske, S., et al. (2013). A molecular signature predictive of indolent prostate cancer. *Sci. Transl. Med.* 5:202ra122. doi: 10.1126/scitranslmed.3006408

- Kim, S.-Y. (2009). Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinformatics* 10:147. doi: 10.1186/1471-2105-10-147
- Kleinbaum, D. and Klein, M. (2005). *Survival Analysis: A Self-Learning Text. Statistics for Biology and Health*. New York, NY: Springer.
- Larkin, S., Holmes, S., Cree, I., Walker, T., Basketter, V., Bickers, B., et al. (2012). Identification of markers of prostate cancer progression using candidate gene expression. *Br. J. Cancer* 106, 157–165. doi: 10.1038/bjc.2011.490
- Lever, J., Krzywinski, M., and Altman, N. (2017). Points of significance: principal component analysis. *Nat. Methods* 14, 641–642. doi: 10.1038/nmeth.4346
- Li, F., Ji, J.-P., Xu, Y., and Liu, R.-L. (2019). Identification a novel set of 6 differential expressed genes in prostate cancer that can potentially predict biochemical recurrence after curative surgery. *Clin. Transl. Oncol.* 21, 1067–1075. doi: 10.1007/s12094-018-02029-z
- Liu, R., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400–416. doi: 10.1016/j.cell.2018.02.052
- Liu, R., Wang, X., Chen, G. Y., Dalerba, P., Gurney, A., Hoey, T., et al. (2007). The prognostic role of a gene signature from tumorigenic breast-cancer cells. *New Engl. J. Med.* 356, 217–226. doi: 10.1056/NEJMoa063994
- Long, Q., Johnson, B. A., Osunkoya, A. O., Lai, Y.-H., Zhou, W., Abramovitz, M., et al. (2011). Protein-coding and microRNA biomarkers of recurrence of prostate cancer following radical prostatectomy. *Am. J. Pathol.* 179, 46–54. doi: 10.1016/j.ajpath.2011.03.008
- Manjang, K., Tripathi, S., Yli-Harja, O., Dehmer, M., and Emmert-Streib, F. (2020). Graph-based exploitation of gene ontology using goxplorer for scrutinizing biological significance. *Sci. Rep.* 10, 1–16. doi: 10.1038/s41598-020-73326-3
- Manjang, K., Tripathi, S., Yli-Harja, O., Dehmer, M., Glazko, G., and Emmert-Streib, F. (2021). Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning. *Sci. Rep.* 11, 1–18. doi: 10.1038/s41598-020-79375-y
- Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 365, 488–492. doi: 10.1016/S0140-6736(05)17866-0
- Nakagawa, T., Kollmeyer, T. M., Morlan, B. W., Anderson, S. K., Bergstralh, E. J., Davis, B. J., et al. (2008). A tissue biomarker panel predicting systemic progression after psa recurrence post-definitive prostate cancer therapy. *PLoS ONE* 3:e2318. doi: 10.1371/journal.pone.0002318
- Penney, K. L., Sinnott, J. A., Fall, K., Pawitan, Y., Hoshida, Y., Kraft, P., et al. (2011). mrna expression signature of gleason grade predicts lethal prostate cancer. *J. Clin. Oncol.* 29:2391. doi: 10.1200/JCO.2010.32.6421
- Ramaswamy, S., Ross, K. N., Lander, E. S., and Golub, T. R. (2003). A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* 33, 49–54. doi: 10.1038/ng1060
- Reddy, G. K., and Balk, S. P. (2006). Clinical utility of microarray-derived genetic signatures in predicting outcomes in prostate cancer. *Clin. Genitourin. cancer* 5, 187–189. doi: 10.3816/CGC.2006.n.035
- Ross, R. W., Galsky, M. D., Scher, H. I., Magidson, J., Wassmann, K., Lee, G.-S. M., et al. (2012). A whole-blood rna transcript-based prognostic model in men with castration-resistant prostate cancer: a prospective study. *Lancet Oncol.* 13, 1105–1113. doi: 10.1016/S1470-2045(12)70263-2
- Ross-Adams, H., Lamb, A., Dunning, M., Halim, S., Lindberg, J., Massie, C., et al. (2015). Integration of copy number and transcriptomics provides risk stratification in prostate cancer: a discovery and validation cohort study. *EBioMedicine* 2, 1133–1144. doi: 10.1016/j.ebiom.2015.07.017
- Saal, L. H., Johansson, P., Holm, K., Gruvberger-Saal, S. K., She, Q.-B., Maurer, M., et al. (2007). Poor prognosis in carcinoma is associated with a gene expression signature of aberrant pten tumor suppressor pathway activity. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7564–7569. doi: 10.1073/pnas.0702507104
- Sharma, N. L., Massie, C. E., Ramos-Montoya, A., Zecchini, V., Scott, H. E., Lamb, A. D., et al. (2013). The androgen receptor induces a distinct transcriptional program in castration-resistant prostate cancer in man. *Cancer Cell* 23, 35–47. doi: 10.1016/j.ccr.2012.11.010
- Siegel, R., Miller, K., and Jemal, A. (2020). Cancer statistics, 2020. *CA Cancer J. Clin.* 70, 7–30. doi: 10.3322/caac.21590
- Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1, 203–209. doi: 10.1016/S1535-6108(02)00030-2
- Song, Z., Huang, Y., Zhao, Y., Ruan, H., Yang, H., Cao, Q., et al. (2019). The identification of potential biomarkers and biological pathways in prostate cancer. *J. Cancer* 10:1398. doi: 10.7150/jca.29571
- Stephenson, A. J., Smith, A., Kattan, M. W., Satagopan, J., Reuter, V. E., Scardino, P. T., et al. (2005). Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer* 104, 290–298. doi: 10.1002/cncr.21157
- Talantov, D., Jatkoa, T. A., Böhm, M., Zhang, Y., Ferguson, A. M., Stricker, P. D., et al. (2010). Gene based prediction of clinically localized prostate cancer progression after radical prostatectomy. *J. Urol.* 184, 1521–1528. doi: 10.1016/j.juro.2010.05.084
- Tandefelt, D. G., Boormans, J. L., van der Korput, H. A., Jenster, G. W., and Trapman, J. (2013). A 36-gene signature predicts clinical progression in a subgroup of erg-positive prostate cancers. *Eur. Urol.* 64, 941–950. doi: 10.1016/j.eururo.2013.02.039
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621
- True, L., Coleman, I., Hawley, S., Huang, C.-Y., Gifford, D., Coleman, R., et al. (2006). A molecular correlate to the gleason grading system for prostate adenocarcinoma. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10991–10996. doi: 10.1073/pnas.0603678103
- Venet, D., Dumont, J. E., and Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* 7:e1002240. doi: 10.1371/journal.pcbi.1002240
- Wang, L.-Y., Cui, J.-J., Zhu, T., Shao, W.-H., Zhao, Y., Wang, S., et al. (2017). Biomarkers identified for prostate cancer patients through genome-scale screening. *Oncotarget* 8:92055. doi: 10.18632/oncotarget.20739
- Wu, C.-L., Schroeder, B. E., Ma, X.-J., Cutie, C. J., Wu, S., Salunga, R., et al. (2013). Development and validation of a 32-gene prognostic index for prostate cancer progression. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6121–6126. doi: 10.1073/pnas.1215870110
- Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. (2019). “Explainable AI: a brief survey on history, research areas, approaches and challenges,” in *CCF International Conference on Natural Language Processing and Chinese Computing* (Dunhuang: Springer), 563–574.
- Yu, J., Yu, J., Rhodes, D. R., Tomlins, S. A., Cao, X., Chen, G., et al. (2007). A polycomb repression signature in metastatic prostate cancer predicts cancer outcome. *Cancer Res.* 67, 10657–10663. doi: 10.1158/0008-5472.CAN-07-2498

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

Copyright © 2021 Manjang, Yli-Harja, Dehmer and Emmert-Streib. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.