Article

Principal component analysis of RNA-seq data unveils a novel prostate cancer-associated gene expression signature

Yasser Perera-Negrin[1,2,*], Augusto Gonzalez[3,4] and Rolando Perez [3,5]

[1] China-Cuba Biotechnology Joint Innovation Center, Yongzhou Zhong Gu Biotechnology Co., Ltd,Yongzhou City, 425000, Hunan Province, People Republic of China; ypereranegrin@ccbjic.com
[2] Laboratory of Molecular Oncology, Center of Genetic Engineering and Biotechnology, Havana, 10600, Cuba; yasser.perera@cigb.edu.cu
[3] Joint China-Cuba neuroinformatics laboratory and Academic Unit, University of Electronic Science and Technology of China, Chengdu, People Republic of China
[4] Institute of Cybernetics, Mathematics and Physics, Havana, Cuba; agonzale@icimaf.cu
[5] Center of Molecular Immunology, Havana, Cuba; rolando@oc.biocubafarma.cu

* Correspondence: ypereranegrin@ccbjic.com

Simple Summary: Prostate cancer (Pca) is a highly heterogeneous disease and the second more common tumor in males. Molecular and genetic profiles have been used to identify subtypes and guide therapeutic intervention. However current risk stratification systems still fail to adequately predict outcome, resulting in frequent patient over-treatment. In addition, therapeutic options for poorly tractable Pca are limited, thus the discovery of novel molecular targets to intervene is also needed. Our Principal Component Analysis (PCA) of RNAseq-data uncovered a Core-Expression Signature (CES) which segregates primary PRAD from normal prostate tissues. The 33 surfaced genes (PRAD-CES) include validated/predicted biomarkers and emerging/putative Pca drivers, as well as six novel RNA genes not previously associated to Pca. GO enrichment and correlation analysis involving major clinical features (i.e., Gleason Score, AR Score, TMPRSS2-ERG fusion and Tumor Cellularity) suggest that PC2 and PC3 gene signatures might describe more aggressive and inflammation-prone transitional forms of PRAD.

Abstract: Prostate cancer (Pca) is a highly heterogeneous disease and the second more common tumor in males. Molecular and genetic profiles have been used to identify subtypes and guide therapeutic intervention. However, roughly 26% of primary Pca are driven by unknown molecular lesions. We use Principal Component Analysis (PCA) and custom RNAseq-data normalization to identify a gene expression signature which segregates primary PRAD from normal tissues. This Core-Expression Signature (PRAD-CES) includes 33 genes and accounts for 39% of data complexity along the PC1-cancer axis. The PRAD-CES is populated by protein-coding (AMACR, TP63, HPN) and RNA-genes (PCA3, ARLN1) sparsely found in previous studies, validated/predicted biomarkers (HOXC6, TDRD1, DLX1), and/or cancer drivers (PCA3, ARLN1, PCAT-14). Of note, the PRAD-CES also comprises six over-expressed LncRNAs without previous Pca association, four of them potentially modulating driver's genes TMPRSS2, PRUNE2 and AMACR. Overall, our PCA capture 57% of data complexity within PC1-3. GO enrichment and correlation analysis involving major clinical features (i.e., Gleason Score, AR Score, TMPRSS2-ERG fusion and Tumor Cellularity) suggest that PC2 and PC3 gene signatures might describe more aggressive and inflammation-prone transitional forms of PRAD. Of note, surfaced genes may entail novel prognostic biomarkers and molecular alterations to intervene. Particularly, our work uncovered RNA genes with appealing implications on Pca biology and progression.

Keywords: Principal Component Analysis, RNA-seq, prostate cancer, biomarkers, RNA genes

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Cancers* **2021**, *13*, x. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname
Received: date
Accepted: date
Published: date

 doi:10.20944/preprints202102.0234.v1

1. Introduction

Prostate cancer (Pca) is the second most common cancer in men [1]. Multiple genetic and demographic factors contribute to the incidence of Pca [2]. Prostate-specific antigen (PSA) screening allows detection of nearly 90% of prostate cancers at initial stages when their surgical removal is the selected medical intervention [3]. However, most of these patients never experience the disease during their life-time, therefore Pca is considered over-diagnosed and over-treated, impacting on patient's quality of life, medical work-loads/costs, and overall society [4].

On the other hand, the clinical outcome of Pca is highly variable, and precise prediction of disease's course, once diagnosed, is not possible [5]. Major risk stratification systems are based on clinical and pathological parameters such as Gleason score, PSA levels, TNM system and surgical margins [6]. However, the above risk stratification systems fail to adequately predict outcome in many cases [7,8]; thus, novel serum-, urinary-, and tissue-based biomarkers are constantly tested and implemented [9]. Of note, for those tumors spreading beyond the prostatic gland (i.e., local and/or distant metastasis) the prognosis is more dismal, and effective therapies are needed [10,11]. Renewed expectations are still rooted into emerging and hopefully tractable Pca molecular alterations [12,13].

Comprehensible genome-wide analysis of primary Prostate Adenocarcinoma (PRAD) revealed both already known and novel molecular lesions for 74% of all tumors [14]. The most common alterations were fusions of androgen-regulated promoters with ERG and other members of the E26 transformation-specific (ETS) family of transcription factors. Particularly, the TMPRSS2-ERG fusion is the most representative molecular lesion, accounting for 46% of study cases. Pca also show varying degrees of DNA copy-number alteration, whereas somatic point mutations are relatively less common [15,16]. Despite this detailed molecular taxonomy of PRAD, roughly 26% of primary Pca of both, good and poor prognosis, are driven by unknown molecular lesions [14].

Principal Component Analysis (PCA) is an unsupervised analysis method providing information about major directions of data variability and structure, thus reducing the overall dimensionality of complex datasets to a few dominant components [17]. Based on global gene expression data, PCA usually reveals underlying population heterogenicity, including cell differentiation stages, malignant phenotypes and treatment-induced changes, which can be linked to phenotypes and further characterized [18]. Biological meanings are usually capture by the first 3-4 PCs, although further improvements on PCA revealed that higher dimensions may also entail biology information [19].

Recently, we used Principal Component Analysis (PCA) analysis of RNA-seq expression data to demonstrated that a relatively small number of "core genes" can segregate normal from neoplastic tissues for different tumor localizations [20]. Here by combining such PCA along with resources allocated at Cbioportal, we analyze primary PRAD RNAseq data to uncover a novel PRAD-Core Expression Signature (PRAD-CES) which may "describe" Pca [21,22]. Furthermore, whereas this PRAD-CES segregates tumor from normal samples along what we call the cancer axis (i.e., PC1), top genes populating PC2 and PC3 might reflects a more aggressive and inflammation-prone transitional forms of PRAD. Overall, the list of surfaced genes may entail novel prognostic biomarkers and/or molecular alterations to intervene. Particularly appealing, was the identification of several RNA genes with potential implications on Pca biology and progression.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2021 doi:10.20944/preprints202102.0234.v1

2. Materials and Methods

RNA-seq data

For PCA we take RNA-seq tissue expression data from the TCGA Prostate Adenocarcinoma project (TCGA-PRAD, https://portal.gdc.cancer.gov/repository, Accessed in March 2019). The data is in the number of fragments per kilo base of gene length per mega-base of reads format (FPKM). The studied cases include 499 tumor samples and 52 normal samples. At Cbioportal (https://www.cbioportal.org/) such data belong to Prostate Adenocarcinoma (TCGA, Firehose Legacy) cohort. Two other data cohorts were used in particular analysis: Prostate Adenocarcinoma (TCGA, Cell 2015) and Prostate Adenocarcinoma (MSKCC, Cancer Cell 2010).

PCA analysis

Fig. 1a shows in a typical PRAD sample that the expression of more than 35000 genes is below 0.1. We shift the ex-pression by 0.1 in such a way that, when computed the differential expressions, genes with not statistically significant expressions are ruled out of the analysis. Then, we take the mean geometric average over normal samples in order to define the reference expression for each gene, and normalize accordingly to obtain the differential expressions, $\bar{e} = e/e_{ref}$. Finally, we take the base 2 logarithm, $\hat{e} = \text{Log2}(\bar{e})$, to define the fold variation. Besides reducing the variance, the logarithm allows treating over- and sub-expression in a symmetrical way. The co-variance matrix is defined in terms of $\hat{e}$. We forced the reference for the PC analysis to be at the center of the cloud of normal samples, $\hat{e} = 0$. This is what actually happens in a population, where most individuals are healthy and cancer situations are rare.

With these assumptions, the covariance matrix is written: $\sigma 2ij = \Sigma\, \hat{e}i(s)\, \hat{e}j(s) / (Nsamples-1)$, where the sum runs over the samples, s, and Nsamples is the total number of samples in the study. $\hat{e}i(s)$ is the fold variation of gene i in samples. The dimension of matrix $\sigma 2$ is 60483, that is equals the number of genes in the data. By diagonalizing this matrix, we get the axes of maximal variance: The Principal Components (PCs). They are sorted in descending order of their contribution to the variance. As mentioned, PC1 captures 39% of the total data variance, PC2 11%, PC3 7%, etc. These results suggest that we may achieve a reasonable description of the main biological characteristics of PRAD using only a small number of the eigenvalues and eigenvectors of $\sigma 2$. To this end, we diagonalize $\sigma 2$ by means of a Lanczos routine in Python language, from which we get the first 100 eigenvalues and their corresponding eigen-vectors.

Gene information and genome visualization

General gene information was collected from Genecards integrated data sources (www.genecards.org) including but not limited to expression, tissues-specificity, sub-cellular localization and diseases association data [23]. Genome visualizations were done with Ensembl release 100 - April 2020 (https://www.ensembl.org), Genome assembly: GRCh38.p13 (GCA_000001405.28) [24].

LncRNA databases

To identify any previous association among identified LncRNAs and cancer, the following non-redundant databases were reviewed: Lnc2Cancer 2.0: An updated database that provides comprehensive experimentally supported associations between lncRNAs and human cancers [25]. LncRNADisease 2.0: contains experimentally and/or computationally supported data [26]. Cancer LncRNA Census (CLC): a compilation of 122 GENCODE lncRNAs with causal roles in cancer phenotypes [27]. The miRTarBase http://miRTarBase.mbc.nctu.edu.tw/ was used to uncover ceRNAs among selected LncRNAs [28].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2021 doi:10.20944/preprints202102.0234.v1

Enrichment analysis
The enrichment analysis was performed using the Enrich platform and the following categories: Ontologies (GO_Biological_Process_2018) and Pathways (Reactome_2016) [29].

Cbioportal
Oncoprint visualizations for selected Genomic Profiles, Alteration Frequency, and mutations representation were obtained from Cbioportal, https://www.cbioportal.org/ [21,22].

Cancer Driver repositories and driver prediction platforms
To search for any previous cancer association of identified genes the Cancer Gene Census and OncoKB (http://oncokb.org) databases were reviewed [30,31]. The driver prediction platforms IntoGene (https://www.intogen.org/search) and ExInAtor (https://www.goldlab.org/cancer-driver-lncrna-prediction-sof) were used to predict a potential driver role for protein-coding and non-coding genes [32,33].

Pearson Correlation
Correlations among selected Pca clinical features and the PCs variables were performed using a Mathematica function (Pearson Correlation Test). A normal distribution of the variables is required.

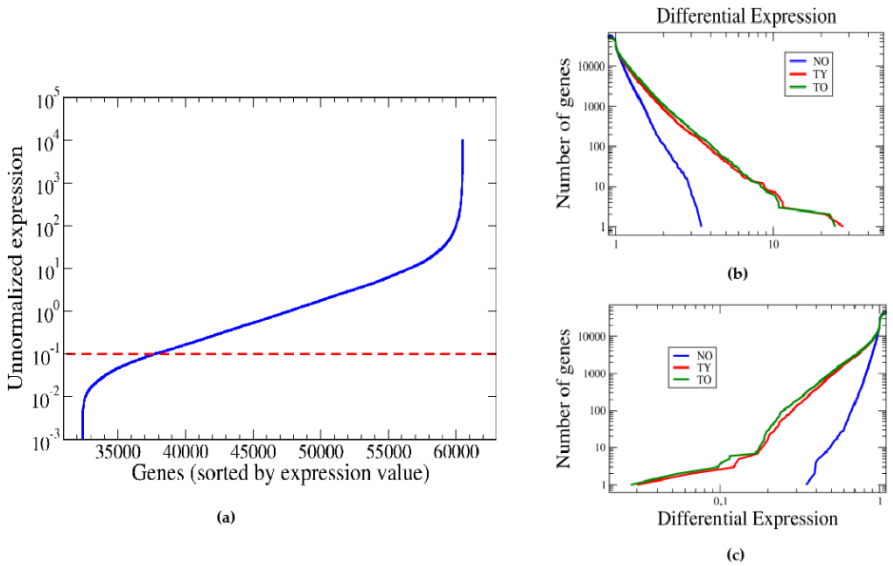
.

3. Results

3.1. Data normalization surfaced an age-independent global gene expression profile

In our analysis there are 52 samples of "normal" prostate tissues, 498 primary tumors samples, and one metastatic sample. RNA-seq data comprise expression values for 60483 independent genes, roughly 35000 of them are not transcribed at significant levels in prostate samples (Figure 1a).

Considering sample availability, we dichotomized the RNAseq data from "normal" and "neoplastic" tissues into two arbitrary age cohorts, with the "old" threshold set at ≥ 62 years (age range: 42-78, median=62) (Supplementary Figure 1). Thus, "normal" patient samples were divided in "young" samples (n=28, NY) and "old" patient samples (n=24, NO); whereas primary tumors samples were divided in "young" tumor samples (n=249, TY) and "old" ones (n=250, TO). While such distribution seems arbitrary and dictated by data availability, only 1 out 4 new PRAD diagnostic cases occurs below 60 years, whereas the mean diagnosis age is 66 years [34].



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2021 doi:10.20944/preprints202102.0234.v1

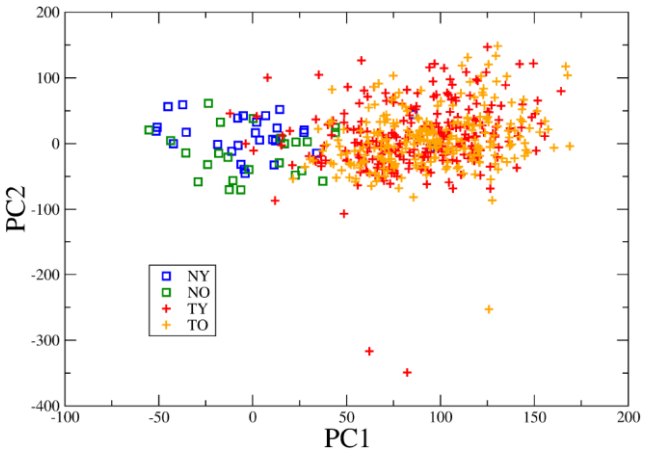
Figure 1. Un-normalized expression data and differential expression profiles in PRAD. (a) Typical range of (un-normalized) expression values from one representative patient (log scale). The red dashed line denotes the expression threshold for statistical significance. Genes with expression below the threshold both in normal tissue and in tumor are mapped to differential expressions very near one. (b,c) Differential-expression profiles for each of the data cohorts NO, TY and TO. The geometric average over the NY group is taken as reference. Notice that there are around 1000 genes with differential expression values below 1/2 (down-regulated) (b) and around 1000 genes with differential expression above 2 (up-regulated) (c). Notice also that the expression profiles practically coincide for the TY and TO groups, and apparently differ from the NO profile.

The normalization of expression values for each of the data cohorts TY and TO against NY group data indicates that the neoplastic transformation entails a similar and genome-wide over- and under-expression of genes, irrespective of the age of the patients (i.e., TY vs TO) (Figure 1b, c). Overall, we found roughly 1000 genes with normalized expression values above 2 and about the same number of genes with normalized expression values below 0.5.

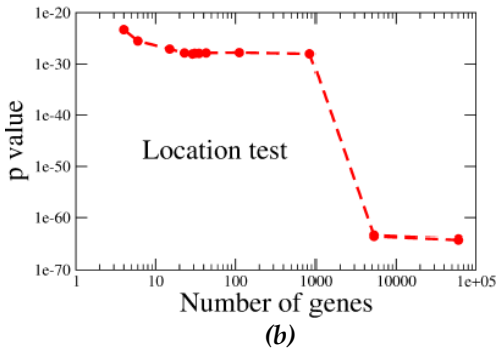
3.2 Principal Component Analysis unveils a Core Expression Signature

The eigenvectors of the covariance matrix defined the PCs axes: PC1, PC2, etc., and projection over them define the new state variables. By definition, PC1 captures the highest fraction of the total variance in the sample set (i.e., PC1=39%), whereas the rest of components are sorted in descending order of their contribution to the variance 11% (PC2), 7% (PC3), 5% (PC4) and so on. Overall, the 8 first PCs comprised 74% of the data variance. Of note, 50% of data variance can be captured by the two major Principal Components (i.e., PC1 and PC2).

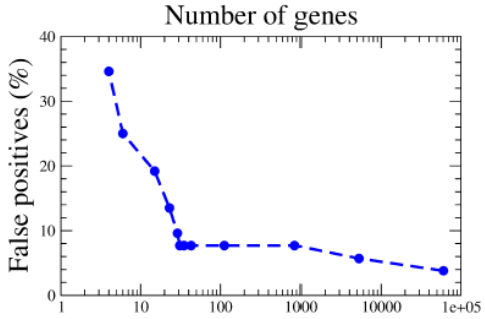
The PCA reveals that a Core Expression Signature composed of 33 genes from PC1 (hereafter, PRAD-CES33) can segregate primary neoplastic samples from normal prostatic tissues with roughly 4% and 8% of false positives and false negatives, respectively (Figure 2). Beyond such 33 genes, the addition of subsequent genes only slightly improves the ratio of false positives and the segregation of neoplastic from normal samples along the PC1 axis.



(a)



(b)



(c)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2021 doi:10.20944/preprints202102.0234.v1

Figure 2. Principal Component Analysis (PCA) of RNAseq-based expression data from PRAD patients. (a) The tumor samples (cloud mean=+91.3) fall apart the distribution of "normal" ones (cloud mean= 0.0) along the PC1 axis defined here as the "cancer axis" (p-value = 10^-65, Mann-Whitney test). (b) Selecting a PC1 value of 45 as a frontier, 2/52 (3.8%) of normal samples are false positives, whereas 38/499 (7.7%) deemed as false negatives. (b, c) The optimal number of "core" genes within the PC1 gene subset is selected according to the ratio of False Positives and the Location Test.

The position along the PC1 axis of a sample is computed as x1 = Σ êi v1i, where v1i are the components of the unitary vector along this axis. A bardcode-like representation of the amplitudes for such 33 genes is represented in Supplementary Figure S2. The greatest value (i.e., over-expression) corresponds to a well know driver and biomarker gene in PRAD, the Prostate Cancer Associated 3 (PCA3) antisense [35,36]. Otherwise, the most under-expressed genes within this PRAD-CES are the protein coding gene SEMG1 [37]. Further bardcode-like analysis of top-100 genes contributing to PC1 axis shown a similar profile (Supplementary Figure S2). Detailed information about the 33 genes included in the core signature are described in Supplementary Table 1.

Notice that a picture like Figure 2b is drawn by recomputing the positions of samples along PC1, the ratio of false positives, etc. by using only the first n genes, ordered according to the module of their amplitudes in vector v1.

Finally, the distribution of tumor samples according to PRAD-CES on the PC1-PC2 plane was similar, irrespective of the age range (i.e., TY cloud median=87, TO cloud median=64). These results imply that not only the global normalized gene expression profile is similar among TY and TO in PRAD cases, rather than a small number of core genes could become a molecular signature of the neoplastic state, irrespective of the age of the patient (i.e., PRAD-CES33).

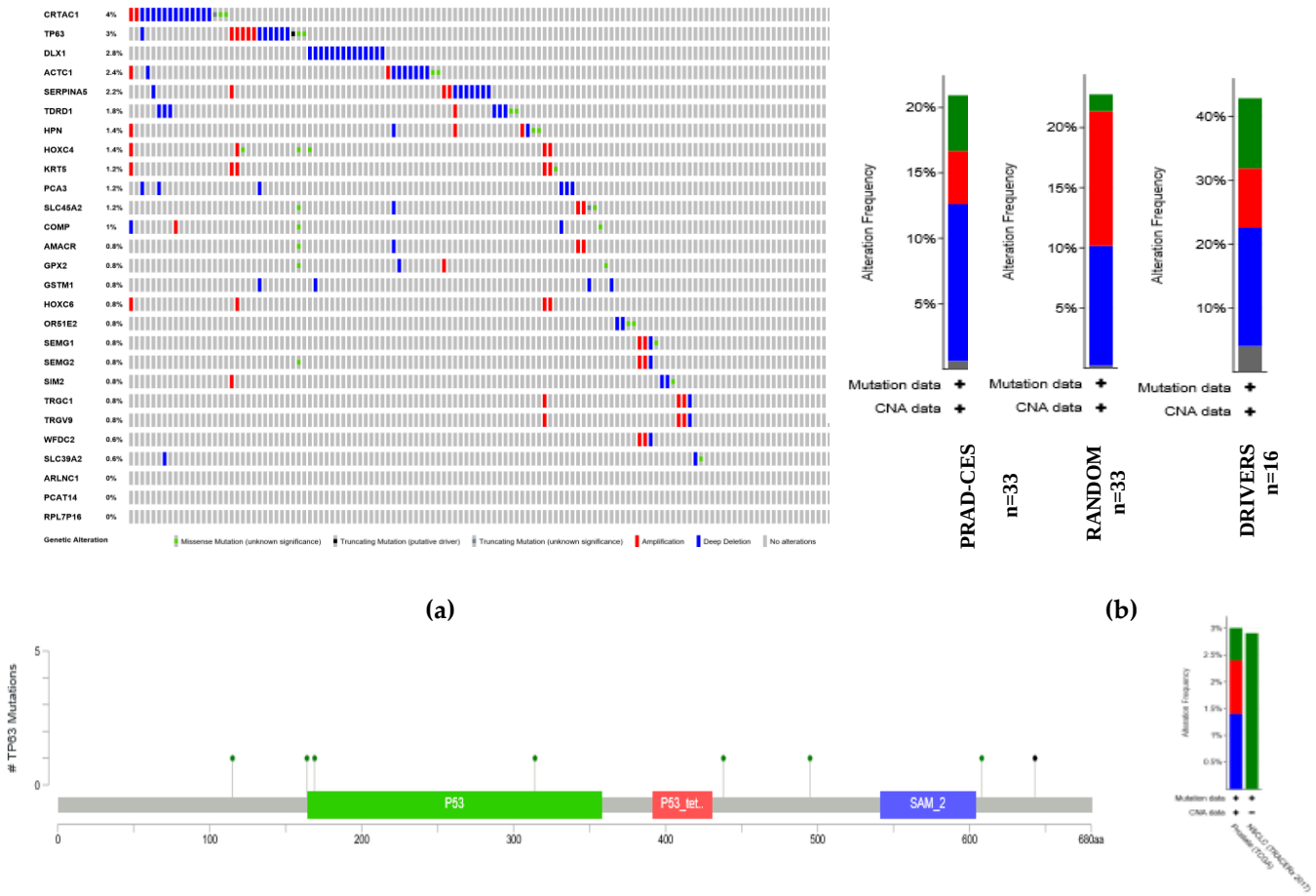
3.3 Protein coding and RNA-genes compose the PRAD-CES33

The surfaced PRAD molecular signature its composed by protein coding (70%), as well as RNA-genes, including antiSense, pseudogene, and LncRNA (30%). Of note, 20/23 (87%) of protein coding genes have been previously associated to cancer, 18 of them (78%) particularly to Pca (Supplementary Table 2). On the other hand, 3/10 RNA genes have been connected to Pca (33%).

We further verified the expression of PRAD-CES33 gene products on available databases. The expression of the corresponding proteins in malignant prostate tissues was observed for 9/23 coding genes, whereas 6/10 RNA genes were detected in such malignant tissues (Supplementary Table 2). Finally, 14/23 protein products were mainly located to the plasma membrane and/or the extra-cellular space.

PRAD-CES genes displayed very low mutational burden with less than 5% of all samples displaying any mutation (Figure 3a, b). Otherwise, roughly 15% of primary PRAD samples harbor CNV on PRAD-CES genes, being predominant deep deletions. The overall alteration frequency of PRAD-CES genes is half of Pca driver genes annotated in the CGC (i.e., 21% vs 42% of cumulative alteration frequency, respectively). Top altered genes include CRTAC1 (4%), TP63 (3%) and DLX1 (2.8%) (Figure 3a).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2021 doi:10.20944/preprints202102.0234.v1



Study	Sample ID	Cancer Type	Protein Change	Annotation	Functional Impact	Mutation Type	Copy #	COSMIC	# Mut in Sample
Prostate Adenoca...	TCGA-XK-AAIW-01	Prostate Adenocarcinoma	*A164V*			Missense	Diploid		6525
Prostate Adenoca...	TCGA-J4-A83J-01	Prostate Adenocarcinoma	*R643**			Nonsense	ShallowDel		47
Prostate Adenoca...	TCGA-HC-A6AP-01	Prostate Adenocarcinoma	*G314A*			Missense	Gain	1	46
Non-Small Cell L...	CRUK0091-R2	Lung Squamous Cell Carcinoma	*Q438H*			Missense			418

(c)

Figure 3. Representation of the pattern and type of molecular alteration in PRAD samples according to Cbioportal (TCGA, Firehose Legacy) data cohort. (a) Oncoprint representing genomic alterations for PRAD-CES33 composite genes. (b) Frequency of genomic alterations observed in PRAD for PRAD-CES33 genes, a random selected gene list and Pca driver genes. (c) Recurrent somatic and germline mutations for the protein-coding Tumor Protein P63 (TP63) in NSCLC and PRAD. Non-Small Cell Lung Cancer mutations from TRACERx, NEJM & Nature 2017 dataset (n= 447).

3.4 Core expression signature includes emerging drivers and biomarkers

A simply text-mining indicated roughly 18 surfaced genes may play driver roles in PRAD (Supplementary Table 2). To verify if PRAD-CES genes include validated cancer drivers we searched the Cancer Gene Census (CGC) and OncoKB databases. According to recurrent somatic and germline mutations, the CGC enlisted TP63 as Tier 1 driver for NSCLC, HNSCC and DLBCL cancers, but not Pca. Both the frequency and affected residues differed among NSCLC and Pca (Figure 3c). None of the remaining 23 protein coding genes within PRAD-CES genes populate such CGC-Pca driver list.

Furthermore, we seek for potential driver roles for genes included in PRAD-CES using the mutational pattern-based prediction platforms IntoGene and ExInAtor (see methods). None of the surfaced PRAD-CES coding genes were predicted as drivers by such orthogonal tools. Otherwise, we search for non-coding genes that could be predicted by ExInAtor as potential cancer drivers. PCA3 was the only significantly mutated LncRNA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2021 doi:10.20944/preprints202102.0234.v1

predicted by ExInAtor as a driver, despite four of the six LncRNAs were analyzed (i.e., PCA3, AP006748.1, AP001610.2, ARLNC1).

To further investigate RNA genes previously associated with cancer, we search three LncRNA databases Lnc2Cancer, LncRNADisease and Cancer LncRNA Census. Three (i.e., ARLNC1, PCA3, PCAT-14), six (i.e., AC092535.4, AP001610.2, AP002498.1, AP006748.1, PCA3 and PCAT14), and one RNA gene (i.e., PCA3), respectively; were previously associated with cancer (Table 1). Of note, the PRAD driver TMPRSS2 was predicted as mRNA target for AP001610.2 and AP006748.1 LncRNAs according to LncRNADisease database.

Table 1. LncRNA included in PRAD-CES and their association with cancer according to indicated databases.

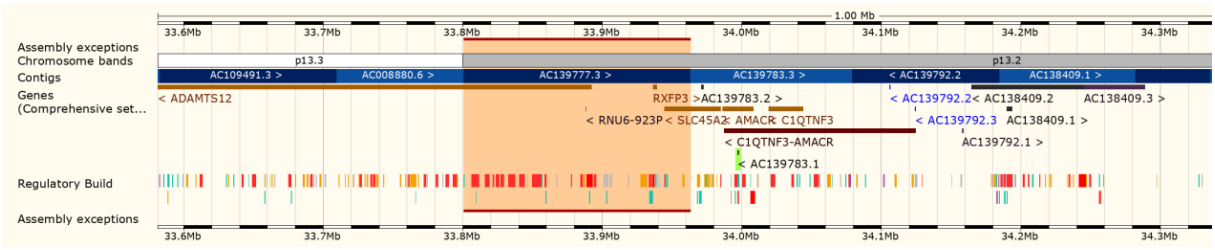
LncRNA	Database	Method	Tumor Type*	Role	mRNA target(s)	Reference
ARLNC1	*Lnc2Cancer 2.0*[a]	RNA-seq, qPCR, Northern blot	Prostate	*Driver*[α]	*CDYL2*[β]	29808028
PCA3	*Lnc2Cancer 2.0*[a]	qPCR, Western blot	Prostate & Others	Driver; Biomarker	PRUNE2	30569456
PCAT-14	*Lnc2Cancer 2.0*[a]	RNA-seq, qPCR, RNAi, ISH	Prostate & Others	Driver; Biomarker	IGLL1, DRICH1	27566105
AC092535.4	*LncRNADisease*[b]	Predicted lncRNA-disease	Cervical & Others	?	CTBP1, SPON2, RNF212	not found
AP001610.2	*LncRNADisease*[b]	Predicted lncRNA-disease	Cervical & Others	?	TMPRSS2, MX1	not found
AP002498.1	*LncRNADisease*[b]	Predicted lncRNA-disease	Cervical & Others	?	CAPN5, B3GNT6, ACER3	not found
AP006748.1	*LncRNADisease*[b]	Predicted lncRNA-disease	Cervical & Others	?	TMPRSS2	not found
PCA3	*LncRNADisease*[b]	ncRNA-disease causality	Prostate & Others	Driver; Biomarker	PRUNE2	27743381; 26594800
PCAT-14	*LncRNADisease*[b]	ncRNA-disease causality	Prostate & Others	Driver; Biomarker	IGLL1, DRICH1	27460352; 27566105
PCA3	*Cancer LncRNA Census*[c]	qPCR, Western blot	Prostate & Others	Driver; Biomarker	PRUNE2	27743381; 26594800

[a] Experimentally supported; [b] Experimentally and/or computationally supported; [c] GENCODE lncRNAs with causal roles; *Top associated tumor; [α] from text-mining; [β] Predicted using LncRNADiseaseb tool.

Finally, two others surfaced LncRNAs may impinge on Pca relevant genes according to a genomic inspection. The LncRNA AL359314.1 overlap with PCA3 and may reinforce the negative regulation of PCA3 over PRUNE2 [36]; whereas AC139783.1 is transcribed within the AMACR protein coding gene (Figure 4a, b).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2021 doi:10.20944/preprints202102.0234.v1

(a)



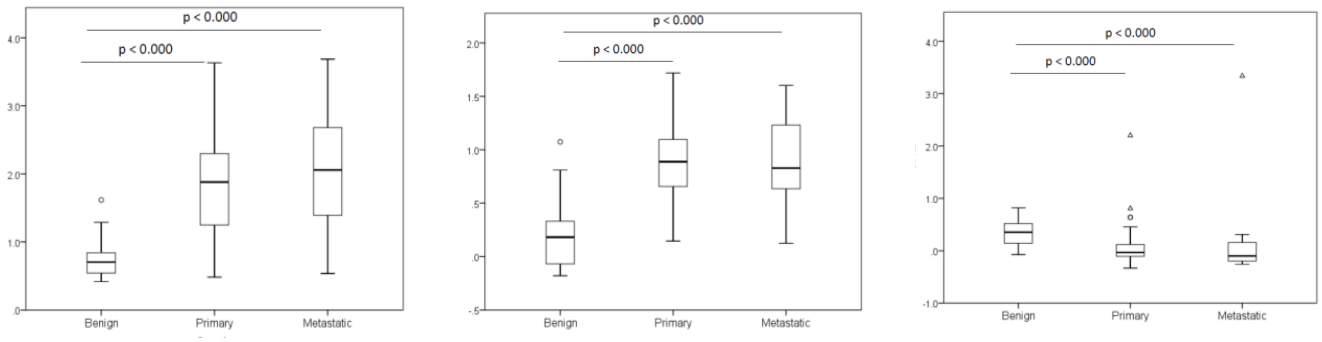
(b)

Figure 4. Genome location of the surfaced LncRNAs AL359314.1 (a) and AC139783.1 (b) indicating sequence overlap with protein coding genes PRUNE2 (anti-sense direction) and AMACR (sense direction). Representation of the regions of interest using Ensembl release 100.

3.5 Aberrant expression of PRAD-CES genes on independent datasets

To verify the deregulated expression of PRAD-CES genes in Pca, we used Prostate Adenocarcinoma (MSKCC, Cancer Cell 2010) data cohorts since such study also used normal samples/tissues to normalize. Most of PRAD-CES genes (i.e., 15/27 detected) showed a consistent expression (i.e., over-expression or under-expression) in a significant proportion of such a patient cohort (>20%) (Supplementary Figure 3). For 10 genes: TRGC1, PCAT14, ARLNC1, SERPINA5, COMP, CRTAC1, SLC39A2, SEMG1, SEMG2 and TRGV9, the aberrant expression was seen in a minor proportion of patients (i.e., <10%) or not patients at all (i.e., TRGC1, PCAT14, ARLNC1, SLC39A2, SEMG2 and TRGV9) (Supplementary Figure 3).

The expression of PRAD-CES genes were further analyzed on three independent prostate cancers studies from Lapointe et al., 2004, Taylor et al., 2010 and Ross-Adams et al., 2015 [38-40]. Three putative emerging drivers in PRAD were consistently deregulated across the analyzed datasets. AMACR, SIM2 and GPX2 protein-coding genes were significantly up-regulated (AMACR, SIM2) or down-regulated (GPX2) in both primary and metastatic samples from lymph node or multiple sites (Figure 5, Supplementary Figure 4).



(a) **(b)** **(c)**

Figure 5. The figures shown Box plots of z-scores of Benign vs malignant tissues for AMACR (a), SIM2 (b) and GPX2 (c) genes. For statistics analysis a Kruskal-Wallis test with Bonferroni correction for multiple tests was conducted. Data taken from (Taylor et al., 2010)[40].

Overall, 14 of 33 PRAD-CES genes were included in the Lapointe dataset (Supplementary Figure 4). Whereas, the expression of 11 of them were consistently up- or down-regulated

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2021 doi:10.20944/preprints202102.0234.v1

in this dataset, three showed no statistical differences (i.e., COMP, SEMG1 and SEMG2). On the other hand, in the Taylor dataset 17 of 33 PRAD-CES genes were detected. The expression of 11 genes were found consistently up- or down-regulated in primary tumors vs. benign tissues in agreement with our RNAseq-data, whereas no significant differences were found for 6 genes (i.e., GSTM1, SERPINA5, COMP, SLC39A2, SEMG1 and SEMG2). Finally, in the Ross-Adams dataset 19 of the 33 PRAD-CES genes were detected. The expression of 17 genes were found consistently up- or down-regulated in primary tumors vs. benign tissues, whereas no significant differences were found for 2 genes (i.e., SEMG1 and SEMG2) (Supplementary Figure 4).

3.6 PCs: Enriched Biological Processes and correlation with major clinical features

To seek for biological meanings beyond that of the individual genes populating the PCs, the top 33 genes from PC1, PC2 and PC3 were submitted to enrichment analysis to identify associated Biological Process. Of note, the top 33 genes populated PC1 (i.e., PRAD-CES) were mainly associated with tumor-intrinsic processes (GO:1900003, GO:0010950, GO:0007283, GO:0048232; $p<0.01$); whereas the Biological Process related to PC2 (GO:0006958, GO:0002455, GO:2000257, GO:0030449; $p<0.001$) and PC3 (GO:0050864, GO:0099024, GO:0051251, GO:0006911; $p<0.001$) suggested involvement of the Innate and adaptive Immune System (Figure 6a-c, Supplementary Table 3).

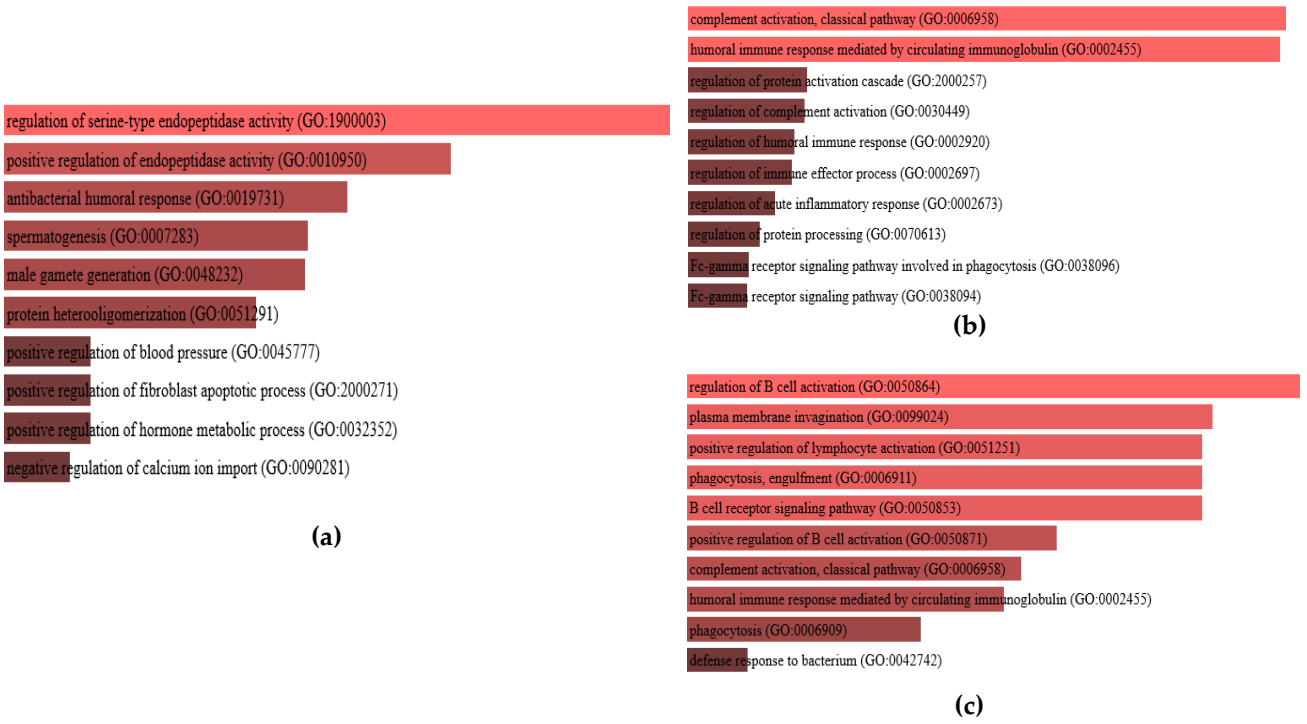


Figure 6. Enrichment analysis for Biological Process using the tool Enrich. PRAD_CES composite genes from PC1(a) and top 33 genes from PC2 (b) and PC3 (c) were included in the analysis. Statistical significance is in accordance with color from light (highly significant) to dark tones (less significant).

Overall, the PRAD-CES genes (PC1) participate in more diverse BP and pathways compared to genes populated PC2 and PC3 which its reflected by the lower combined scores for top identified BP within PC1 (Supplementary Table 3 and 4). Otherwise, PC2 and PC3 populated genes seemed mainly involved in the complement activation, humoral immune response, regulation of B cell activation, phagocytosis, engulfment and regulation of acute inflammatory response.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2021 doi:10.20944/preprints202102.0234.v1

To analyze the underlying distribution of major PRAD clinical features across PCs 1-3, a correlation analysis between each PC and the Gleason-Score, AR-Score, TMPRSS2-ERG and tumor cellularity were performed (Figure 7, Table 2).

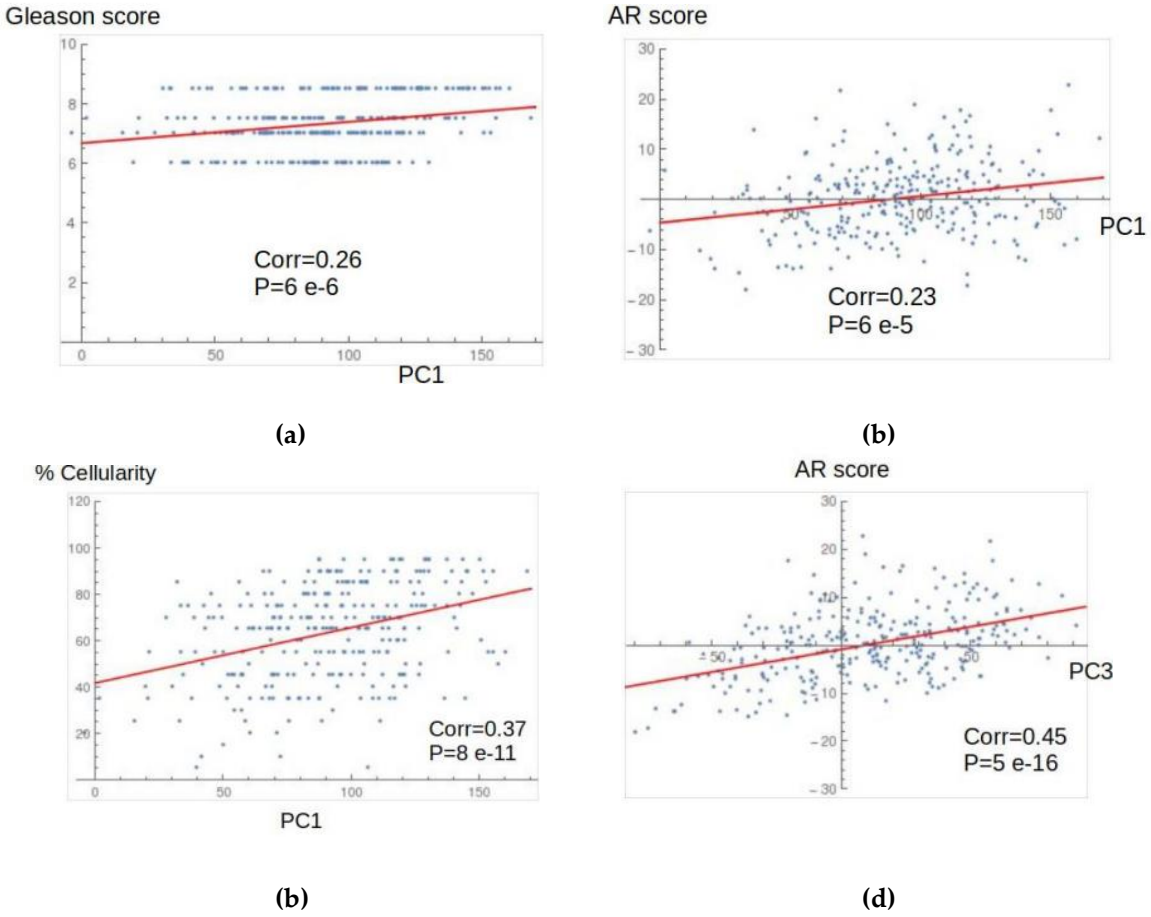


Figure 7. Correlation between PC1 and PC3 and major clinical features of PRAD using the data cohort Prostate Adenocarcinoma (TCGA, Cell 2015), comprising 333 primary tumors. Major features include Gleason-Score (a), AR-Score (b,d), TMPRSS2-ERG and tumor cellularity (c) (see Table 2 for details).

Table 2. Correlations among PCs and selected clinical features of PRAD (TCGA, Cell 2015). A Pearson Correlation Test was performed. **Bold** numbers indicate significant correlation p<0.05.
See also Supplementary Table 5.

Clinical Features	PC1	PC2	PC3
Gleason Score	**0.26**	**-0.16**	0.04
TMPRSS2-ERG	0.02	**-0.18**	**0.24**
AR Score	**0.23**	**0.32**	**0.45**
Cellularity	**0.37**	**0.14**	**0.19**

Our analysis revealed that PC1 values shown a weak-yet positive correlation with Gleason (R<0.30, p=6.0E-06), and AR Score (R<0.30, p=5.0E-5); whereas a medium-strength positive association with Tumor cellularity (R=0.37, p=8.0E-11) was seen. Of note, independent correlations among clinical features in this dataset indicated that the Gleason score weakly correlates with Cellularity (R=0.26, p=8.0E-6) and TMPRSS2-ERG fusion anti-correlates with AR Score (R=-0.24, p=4.0E-5) (Supplementary Table 5). Therefore, the observed correlation between PC1 values and the above-mentioned clinical features may reflect the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2021 doi:10.20944/preprints202102.0234.v1

underlying PRAD biology which is in line with the fact that PC1 explain up to 39% of data complexity, being a more "general" expression signature.

Concerning PC2, we observed an anti-correlation among TMPRSS2-ERG and AR Score which goes along the underlying PRAD biology; however, in this PC the Gleason Score anti-correlated with Tumor Cellularity. Finally, the genes included in PC3 showed positive correlations with TMPRSS2-ERG, AR Score and Tumor Cellularity (Table 2, Supplementary Table 5).

4. Discussion

Here, we use Principal Component Analysis (PCA) to surface gene expression patterns which may "describe" primary PRAD, providing new putative biomarkers and/or molecular targets to intervene. Such dimensionality reduction algorithm clearly segregates tumor from normal samples, with eight PCs capturing roughly 3/4 of data complexity. The RNA-seq input data was obtained from the Prostate Adenocarcinoma cohort TCGA_Firehose Legacy, which comprised a significant number of tumor and normal samples, the ultimate required to perform a custom-made normalization. Furthermore, the use of only one Pca data cohort, among severals available, is in line with findings indicating PCA lose resolution on highly heterogeneous and pooled data [13]. In addition, we used two other major PRAD data cohorts (i.e., TCGA, Cell 2015 and MSKCC, Cancer Cell 2010) as they provide clinical data (TCGA, Cell 2015) or used similar normalization for RNA-seq based gene expression analysis (MSKCC, Cancer Cell 2010).

Our custom-made normalization revealed a long-tail distribution of expression values which might reflect global deregulation events associated with aging and/or malignant transformation (i.e., genetic and epigenetics events) [41]. Since we used "normal Young" data as reference, the obtained pattern may suggest that neoplastic transformation overimpose on an already age-adjusted global expression profile (i.e., TY and TO profiles seem alike). A similar global gene-expression pattern emerged when a micro-array expression data from a previous study was analyzed (data not shown).

Our PCA allow us to identify a Core-Expression Signature (PRAD-CES) composed of 33 genes which accounts for 39% of data variance along what we call the cancer axis (PC1). The biological meaning of PC2 and PC3 seems more elusive, accounting for an additional 18% of variability. The PRAD-CES includes validated, emerging and putative PRAD drivers and/or biomarkers. Although only one validated protein-coding driver was found (i.e., TP63), three RNA genes with causative roles were surfaced: ARLNC1, PCA3, and PCAT-14 [36,42-44]. Otherwise, six protein coding genes awaits further validation concerning PRAD driver roles: OR51E2, HPN, AMACR, DLX1, HOXC6 and WFDC2 [45-50].

Concerning potential or validated biomarkers, the PRAD-CES list contains 15 RNA- or protein-coding genes with such a role. Among them HOXC6, TDRD1, and DLX1 have been already proposed to identify patients with aggressive prostate cancer [51]. TDRD1 might also play an important role in prostate cancer development, and as a cancer/testis antigen, a potential therapeutic target for cancer immunotherapy [52].

Of note, cross-validation of genes included in PRAD-CES against independent data cohorts, indicated that most of these genes were consistently deregulated in primary PRAD, with notable exceptions on comp, semg1 and semg2 genes. Otherwise, the expression of 14 genes could not be verified in some of the above-mentioned datasets (i.e., Lapointe et al., 2004, Taylor et al., 2010 and Ross-Adams et al., 2015)[38-40]. Overall, the most

 doi:10.20944/preprints202102.0234.v1

consistent genes among those detected across all analyzed data were OR51E2, SIM2, HPN, SLC45A2, TDRD1, PCA3, DLX1, AMACR, WFDC2, and HOXC6.

On the other hand, our PCA surfaced nine over-expressed RNA genes, six of them lacking previous association with Pca. Particularly, four LncRNAs could target PRAD driver's genes TMPRSS2, PRUNE2 and AMACR. One interesting finding was the genome proximity/overlap among PRAD-CES over-expressed genes AC139783.1, AMACR and SLC45A2 on Chromosome 5. SLC45A2-AMACR was reported as a novel fusion protein which is associated with progressive Pca disease [53]. Otherwise, among several miRNAs which may down-regulate AMACR expression in Pca, the potential sponging of hsa-miR-26a-5p by the surfaced AC139783.1, needs to be addressed. AMACR over-expression have been associated with Pca evolution towards hormone-independency, whereas AMACR inhibition seems a feasible strategy to treat hormone-refractory prostate cancer patients [47]. Of note, LncRNA over-expression in Pca has been related with disease progression, used as prognostic factor, or proposed as therapeutic targets [54-56].

The most frequent molecular abnormalities in PRAD involved gene-fusions, copy-number alterations and epigenetic deregulation [14]. As a matter of facts, the mutational burden observed in surfaced PRAD-CES genes was low, suggesting that expression levels and not co-existing mutations determine the PCA-based segregation of tumor from normal samples. Furthermore, less than 3% of PRAD samples included in our study displayed CNV, thus suggesting that most of the observed gene expression deregulation arose from epigenetic and/or transcription-based regulatory mechanism.

In this work we selected four molecular/clinical features to correlate with PCs. Primary prostate cancer is androgen dependent, and androgen-mediated signaling is crucial in prostate cancer pathogenesis, driving the creation and over-expression of most ETS fusion genes [57,58]. Among such ETS fusion genes, TMPRSS2-ERG fusion accounts for 46% of cases [14]. On the other hand, Gleason score remains as a cornerstone pathological criterion for risk-stratification and prognosis [59]. Finally, we included tumor cellularity as a proxy for non-prostatic yet-relevant infiltrating populations [60].

The observed correlations indicated PC1 might reflects the underlying primary PRAD biology with positive correlation among Gleason Score and Tumor Cellularity, as well as among this variable and AR Score. Otherwise, genes comprising PC2 and PC3 may reveal a transition towards a more aggressive and inflammation-prone phenotype, with a mixture of tumor epithelial cells and infiltrating immune cells [61]. This notion seems also supported by a weaker correlation of PC2 and PC3 genes with tumor cellularity, but also by the increasingly positive correlation among genes populating PCs 1-3 and the AR Score (i.e., from 0.23 to 0.45). Of note, only PC3 genes positively correlated with TMPRSS2-ERG fusion. Altogether, an intriguing possibility is whether PC3-populating genes may describe an inherent fraction of primary tumors cells endowed to metastasize.

5. Conclusions

Our study is limited by data availability/structure and biopsy bias as any global transcriptome inquire [62]. Primary prostate tumors are multifocal and molecularly heterogeneous; thus, the surfaced signature may reflect only gene expression features representative from the sampled site [63,64]. However, our PCA indeed uncover relevant PRAD genes found dispersed across several studies, providing new putative biomarkers and/or drivers. In this sense, the inclusion of PCA3 within our PRAD-CES seems encouraging since this LncRNA is well recognized as causative, prostate-specific and feasible biomarker which is secreted to an easy-to-inquire biological fluids like urine [65]. Furthermore, as

therapeutic options for poorly tractable Pca are limited, the evaluation of the putative novel molecular targets populating PRAD-CES seems appealing.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Histogram showing the age range for patients included in the PRAD-TCGA Firehose Legacy cohort used for PCA of RNA-seq expression data, Figure S2: A Barcode-like representation of PRAD genes comprising the unitary vectors along PC1. Top panel, contains 33 genes identified for a delta=0.027; Lower panel representing 100 genes for a delta=0.0235. Within the barcode the major value (i.e., over-expression) corresponds to PCA3, whereas the lower value (i.e., under-expression) belongs to SEMG1, Figure S3: Differential expression of PRAD-CES genes in the MSKCC, Cancer Cell 2010 cohort. The oncoprint representation tool from Cbioportal is used. Z scores>2, normal vs tumor expression values, Figure S4: Expression analysis of genes from the PRAD-CES using three independent data cohorts of primary prostate tumors from radical prostatectomy, Table S1: Detailed information about the 33 genes included in the PRAD core signature (PRAD-CES33), Table S2: Gene Classification, Disease Association and Gene Product Expression, Table S3: Enrichment analysis for Biological Process using the tool Enrich, Table S4: Enrichment analysis for Biological Pathways using the tool Enrich, Table S5: Correlations among PCs and selected clinical features of PRAD (TCGA, Cell 2015).

Author Contributions: Conceptualization, all authors; methodology, A.G. and Y.PN.; software, A.G.; formal analysis, A.G. and Y.PN.; writing—original draft preparation, Y.PN.; writing—review and editing, all authors; funding acquisition, R. P and Y.PN. "All authors have read and agreed to the published version of the manuscript."

Funding: This research was funded by Platform for Bio-informatics of BioCubaFarma, Cuba, grant number 01Y19 and the "Hunan Provincial Base for Scientific and Technological Innovation Cooperation", China, grant number 2019CB1012. The APC was funded by grant number 2019CB1012.

Institutional Review Board Statement: "Not applicable."

Informed Consent Statement: "Not applicable."

Data Availability Statement: The data come from the TCGA Research Network: https://www.cancer.gov/tcga and Cbioportal: https://www.cbioportal.org/ as for March, 2019.

Acknowledgments: We would like to thanks to Dr. Simone Chevalier from the Centre for Translational Medicine, Research Institute, McGill University for expression data cross-validation. A.G acknowledges the Cuban Program for Basic Sciences and the Office of External Activities of the Abdus Salam Centre for Theoretical Physics for support.

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **2018**, *68*, 394-424, doi:10.3322/caac.21492.
2. Al Olama, A.A.; Kote-Jarai, Z.; Berndt, S.I.; Conti, D.V.; Schumacher, F.; Han, Y.; Benlloch, S.; Hazelett, D.J.; Wang, Z.; Saunders, E., et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nature genetics* **2014**, *46*, 1103-1109, doi:10.1038/ng.3094.
3. Penney, K.L.; Stampfer, M.J.; Jahn, J.L.; Sinnott, J.A.; Flavin, R.; Rider, J.R.; Finn, S.; Giovannucci, E.; Sesso, H.D.; Loda, M., et al. Gleason grade progression is uncommon. *Cancer research* **2013**, *73*, 5163-5168, doi:10.1158/0008-5472.can-13-0427.
4. Vickers, A.J. Redesigning Prostate Cancer Screening Strategies to Reduce Overdiagnosis. *Clinical chemistry* **2019**, *65*, 39-41, doi:10.1373/clinchem.2018.287094.

5. Buyyounouski, M.K.; Pickles, T.; Kestin, L.L.; Allison, R.; Williams, S.G. Validating the interval to biochemical failure for the identification of potentially lethal prostate cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **2012**, *30*, 1857-1863, doi:10.1200/jco.2011.35.1924.
6. Sathianathen, N.J.; Konety, B.R.; Crook, J.; Saad, F.; Lawrentschuk, N. Landmarks in prostate cancer. *Nature reviews. Urology* **2018**, *15*, 627-642, doi:10.1038/s41585-018-0060-7.
7. D'Amico, A.V.; Whittington, R.; Malkowicz, S.B.; Schultz, D.; Blank, K.; Broderick, G.A.; Tomaszewski, J.E.; Renshaw, A.A.; Kaplan, I.; Beard, C.J., et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *Jama* **1998**, *280*, 969-974, doi:10.1001/jama.280.11.969.
8. Cooperberg, M.R.; Broering, J.M.; Carroll, P.R. Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. *J Natl Cancer Inst* **2009**, *101*, 878-887, doi:10.1093/jnci/djp122.
9. Duffy, M.J. Biomarkers for prostate cancer: prostate-specific antigen and beyond. *Clinical chemistry and laboratory medicine* **2020**, *58*, 326-339, doi:10.1515/cclm-2019-0693.
10. Powers, E.; Karachaliou, G.S.; Kao, C.; Harrison, M.R.; Hoimes, C.J.; George, D.J.; Armstrong, A.J.; Zhang, T. Novel therapies are changing treatment paradigms in metastatic prostate cancer. *Journal of hematology & oncology* **2020**, *13*, 144, doi:10.1186/s13045-020-00978-z.
11. Wang, Z.; Ni, Y.; Chen, J.; Sun, G.; Zhang, X.; Zhao, J.; Zhu, X.; Zhang, H.; Zhu, S.; Dai, J., et al. The efficacy and safety of radical prostatectomy and radiotherapy in high-risk prostate cancer: a systematic review and meta-analysis. *World journal of surgical oncology* **2020**, *18*, 42, doi:10.1186/s12957-020-01824-9.
12. Mateo, J.; Seed, G.; Bertan, C.; Rescigno, P.; Dolling, D.; Figueiredo, I.; Miranda, S.; Nava Rodrigues, D.; Gurel, B.; Clarke, M., et al. Genomics of lethal prostate cancer at diagnosis and castration resistance. *The Journal of clinical investigation* **2020**, *130*, 1743-1751, doi:10.1172/jci132031.
13. Luca, B.A.; Moulton, V.; Ellis, C.; Edwards, D.R.; Campbell, C.; Cooper, R.A.; Clark, J.; Brewer, D.S.; Cooper, C.S. A novel stratification framework for predicting outcome in patients with prostate cancer. *British journal of cancer* **2020**, *122*, 1467-1476, doi:10.1038/s41416-020-0799-5.
14. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **2015**, *163*, 1011-1025, doi:10.1016/j.cell.2015.10.025.
15. Alexandrov, L.B.; Nik-Zainal, S.; Wedge, D.C.; Aparicio, S.A.; Behjati, S.; Biankin, A.V.; Bignell, G.R.; Bolli, N.; Borg, A.; Børresen-Dale, A.L., et al. Signatures of mutational processes in human cancer. *Nature* **2013**, *500*, 415-421, doi:10.1038/nature12477.
16. Lawrence, M.S.; Stojanov, P.; Polak, P.; Kryukov, G.V.; Cibulskis, K.; Sivachenko, A.; Carter, S.L.; Stewart, C.; Mermel, C.H.; Roberts, S.A., et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **2013**, *499*, 214-218, doi:10.1038/nature12213.
17. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2*, 37-52, doi:https://doi.org/10.1016/0169-7439(87)80084-9.
18. Lukk, M.; Kapushesky, M.; Nikkilä, J.; Parkinson, H.; Goncalves, A.; Huber, W.; Ukkonen, E.; Brazma, A. A global map of human gene expression. *Nature biotechnology* **2010**, *28*, 322-324, doi:10.1038/nbt0410-322.
19. Lenz, M.; Müller, F.-J.; Zenke, M.; Schuppert, A. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. **2016**, *6*, 25696, doi:10.1038/srep25696.
20. Gonzalez, A.; Perera, Y.; Perez, R.J.a.p.a. On the gene expression landscape of cancer. **2020**.
21. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E., et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2012**, *2*, 401-404, doi:10.1158/2159-8290.Cd-12-0095.

22. Gao, J.; Aksoy, B.A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S.O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E., et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **2013**, *6*, pl1-pl1, doi:10.1126/scisignal.2004088.
23. Rebhan, M.; Chalifa-Caspi, V.; Prilusky, J.; Lancet, D. GeneCards: integrating information about genes, proteins and diseases. *Trends in Genetics* **1997**, *13*, 163, doi:10.1016/S0168-9525(97)01103-7.
24. Yates, A.D.; Achuthan, P.; Akanni, W.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Azov, A.G.; Bennett, R., et al. Ensembl 2020. *Nucleic acids research* **2020**, *48*, D682-d688, doi:10.1093/nar/gkz966.
25. Gao, Y.; Wang, P.; Wang, Y.; Ma, X.; Zhi, H.; Zhou, D.; Li, X.; Fang, Y.; Shen, W.; Xu, Y., et al. Lnc2Cancer v2.0: updated database of experimentally supported long non-coding RNAs in human cancers. *Nucleic acids research* **2019**, *47*, D1028-D1033, doi:10.1093/nar/gky1096.
26. Bao, Z.; Yang, Z.; Huang, Z.; Zhou, Y.; Cui, Q.; Dong, D. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic acids research* **2019**, *47*, D1034-d1037, doi:10.1093/nar/gky905.
27. Carlevaro-Fita, J.; Lanzós, A.; Feuerbach, L.; Hong, C.; Mas-Ponte, D.; Pedersen, J.S.; Johnson, R. Unique genomic features and deeply-conserved functions of long non-coding RNAs in the Cancer LncRNA Census (CLC). **2017**, 10.1101/152769 %J bioRxiv, 152769, doi:10.1101/152769 %J bioRxiv.
28. Chou, C.-H.; Chang, N.-W.; Shrestha, S.; Hsu, S.-D.; Lin, Y.-L.; Lee, W.-H.; Yang, C.-D.; Hong, H.-C.; Wei, T.-Y.; Tu, S.-J., et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic acids research* **2016**, *44*, D239-D247, doi:10.1093/nar/gkv1258.
29. Kuleshov, M.V.; Jones, M.R.; Rouillard, A.D.; Fernandez, N.F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S.L.; Jagodnik, K.M.; Lachmann, A., et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **2016**, *44*, W90-97, doi:10.1093/nar/gkw377.
30. Sondka, Z.; Bamford, S.; Cole, C.G.; Ward, S.A.; Dunham, I.; Forbes, S.A. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature reviews. Cancer* **2018**, *18*, 696-705, doi:10.1038/s41568-018-0060-1.
31. Chakravarty, D.; Gao, J.; Phillips, S.M.; Kundra, R.; Zhang, H.; Wang, J.; Rudolph, J.E.; Yaeger, R.; Soumerai, T.; Nissan, M.H., et al. OncoKB: A Precision Oncology Knowledge Base. *JCO precision oncology* **2017**, *2017*, doi:10.1200/po.17.00011.
32. Martínez-Jiménez, F.; Muiños, F.; Sentís, I.; Deu-Pons, J.; Reyes-Salazar, I.; Arnedo-Pac, C.; Mularoni, L.; Pich, O.; Bonet, J.; Kranas, H., et al. A compendium of mutational cancer driver genes. *Nature reviews. Cancer* **2020**, *20*, 555-572, doi:10.1038/s41568-020-0290-x.
33. Lanzós, A.; Carlevaro-Fita, J.; Mularoni, L.; Reverter, F.; Palumbo, E.; Guigó, R.; Johnson, R. Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. In *Sci Rep*, 2017; Vol. 7, p 41544.
34. Li, J.; Djenaba, J.A.; Soman, A.; Rim, S.H.; Master, V.A. Recent trends in prostate cancer incidence by age, cancer stage, and grade, the United States, 2001-2007. *Prostate cancer* **2012**, *2012*, 691380, doi:10.1155/2012/691380.
35. Bussemakers, M.J.; van Bokhoven, A.; Verhaegh, G.W.; Smit, F.P.; Karthaus, H.F.; Schalken, J.A.; Debruyne, F.M.; Ru, N.; Isaacs, W.B. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer research* **1999**, *59*, 5975-5979.
36. Salameh, A.; Lee, A.K.; Cardó-Vila, M.; Nunes, D.N.; Efstathiou, E.; Staquicini, F.I.; Dobroff, A.S.; Marchiò, S.; Navone, N.M.; Hosoya, H., et al. PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3. *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112*, 8403-8408, doi:10.1073/pnas.1507882112.
37. Canacci, A.M.; Izumi, K.; Zheng, Y.; Gordetsky, J.; Yao, J.L.; Miyamoto, H. Expression of semenogelins I and II and its prognostic significance in human prostate cancer. *The Prostate* **2011**, *71*, 1108-1114, doi:10.1002/pros.21323.
38. Ross-Adams, H.; Lamb, A.D.; Dunning, M.J.; Halim, S.; Lindberg, J.; Massie, C.M.; Egevad, L.A.; Russell, R.; Ramos-Montoya, A.; Vowler, S.L., et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *EBioMedicine* **2015**, *2*, 1133-1144, doi:10.1016/j.ebiom.2015.07.017.

39. Lapointe, J.; Li, C.; Higgins, J.P.; van de Rijn, M.; Bair, E.; Montgomery, K.; Ferrari, M.; Egevad, L.; Rayford, W.; Bergerheim, U., et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, 811-816, doi:10.1073/pnas.0304146101.
40. Taylor, B.S.; Schultz, N.; Hieronymus, H.; Gopalan, A.; Xiao, Y.; Carver, B.S.; Arora, V.K.; Kaushik, P.; Cerami, E.; Reva, B., et al. Integrative genomic profiling of human prostate cancer. *Cancer cell* **2010**, *18*, 11-22, doi:10.1016/j.ccr.2010.05.026.
41. Sen, P.; Shah, P.P.; Nativio, R.; Berger, S.L. Epigenetic Mechanisms of Longevity and Aging. *Cell* **2016**, *166*, 822-839, doi:10.1016/j.cell.2016.07.050.
42. Zhang, Y.; Pitchiaya, S.; Cieślik, M.; Niknafs, Y.S.; Tien, J.C.; Hosono, Y.; Iyer, M.K.; Yazdani, S.; Subramaniam, S.; Shukla, S.K., et al. Analysis of the androgen receptor-regulated lncRNA landscape identifies a role for ARLNC1 in prostate cancer progression. *Nature genetics* **2018**, *50*, 814-824, doi:10.1038/s41588-018-0120-1.
43. Wang, Y.; Hu, Y.; Wu, G.; Yang, Y.; Tang, Y.; Zhang, W.; Wang, K.; Liu, Y.; Wang, X.; Li, T. Long noncoding RNA PCAT-14 induces proliferation and invasion by hepatocellular carcinoma cells by inducing methylation of miR-372. *Oncotarget* **2017**, *8*, 34429-34441, doi:10.18632/oncotarget.16260.
44. Dhillon, P.K.; Barry, M.; Stampfer, M.J.; Perner, S.; Fiorentino, M.; Fornari, A.; Ma, J.; Fleet, J.; Kurth, T.; Rubin, M.A., et al. Aberrant cytoplasmic expression of p63 and prostate cancer mortality. *Cancer Epidemiol Biomarkers Prev* **2009**, *18*, 595-600, doi:10.1158/1055-9965.epi-08-0785.
45. Rodriguez, M.; Siwko, S.; Liu, M. Prostate-Specific G-Protein Coupled Receptor, an Emerging Biomarker Regulating Inflammation and Prostate Cancer Invasion. *Current Molecular Medicine* **2016**, *16*, 526-532, doi:http://dx.doi.org/10.2174/1566524016666160607091333.
46. Tang, X.; Mahajan, S.S.; Nguyen, L.T.; Béliveau, F.; Leduc, R.; Simon, J.A.; Vasioukhin, V. Targeted inhibition of cell-surface serine protease Hepsin blocks prostate cancer bone metastasis. *Oncotarget* **2014**, *5*, 1352-1362, doi:10.18632/oncotarget.1817.
47. Takahara, K.; Azuma, H.; Sakamoto, T.; Kiyama, S.; Inamoto, T.; Ibuki, N.; Nishida, T.; Nomi, H.; Ubai, T.; Segawa, N., et al. Conversion of prostate cancer from hormone independency to dependency due to AMACR inhibition: involvement of increased AR expression and decreased IGF1 expression. *Anticancer research* **2009**, *29*, 2497-2505.
48. Liang, M.; Sun, Y.; Yang, H.-L.; Zhang, B.; Wen, J.; Shi, B.-K. DLX1, a binding protein of beta-catenin, promoted the growth and migration of prostate cancer cells. *Experimental Cell Research* **2018**, *363*, 26-32, doi:https://doi.org/10.1016/j.yexcr.2018.01.007.
49. Vinarskaja, A.; Yamanaka, M.; Ingenwerth, M.; Schulz, W.A. DNA Methylation and the HOXC6 Paradox in Prostate Cancer. *Cancers (Basel)* **2011**, *3*, 3714-3725, doi:10.3390/cancers3043714.
50. Gao, L.; Cheng, H.Y.; Dong, L.; Ye, X.; Liu, Y.N.; Chang, X.H.; Cheng, Y.X.; Chen, J.; Ma, R.Q.; Cui, H. The role of HE4 in ovarian cancer: inhibiting tumour cell proliferation and metastasis. *The Journal of international medical research* **2011**, *39*, 1645-1660, doi:10.1177/147323001103900507.
51. Leyten, G.H.; Hessels, D.; Smit, F.P.; Jannink, S.A.; de Jong, H.; Melchers, W.J.; Cornel, E.B.; de Reijke, T.M.; Vergunst, H.; Kil, P., et al. Identification of a Candidate Gene Panel for the Early Diagnosis of Prostate Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **2015**, *21*, 3061-3070, doi:10.1158/1078-0432.Ccr-14-3334.
52. Xiao, L.; Lanz, R.B.; Frolov, A.; Castro, P.D.; Zhang, Z.; Dong, B.; Xue, W.; Jung, S.Y.; Lydon, J.P.; Edwards, D.P., et al. The Germ Cell Gene TDRD1 as an ERG Target Gene and a Novel Prostate Cancer Biomarker. *The Prostate* **2016**, *76*, 1271-1284, doi:10.1002/pros.23213.
53. Yu, Y.P.; Ding, Y.; Chen, Z.; Liu, S.; Michalopoulos, A.; Chen, R.; Gulzar, Z.G.; Yang, B.; Cieply, K.M.; Luvison, A., et al. Novel fusion transcripts associate with progressive prostate cancer. *The American journal of pathology* **2014**, *184*, 2840-2849, doi:10.1016/j.ajpath.2014.06.025.
54. Ahadi, A.; Brennan, S.; Kennedy, P.J.; Hutvagner, G.; Tran, N. Long non-coding RNAs harboring miRNA seed regions are enriched in prostate cancer exosomes. *Sci Rep* **2016**, *6*, 24922, doi:10.1038/srep24922.

55. Mehra, R.; Udager, A.M.; Ahearn, T.U.; Cao, X.; Feng, F.Y.; Loda, M.; Petimar, J.S.; Kantoff, P.; Mucci, L.A.; Chinnaiyan, A.M. Overexpression of the Long Non-coding RNA SChLAP1 Independently Predicts Lethal Prostate Cancer. *European urology* **2016**, *70*, 549-552, doi:10.1016/j.eururo.2015.12.003.
56. Ren, S.; Liu, Y.; Xu, W.; Sun, Y.; Lu, J.; Wang, F.; Wei, M.; Shen, J.; Hou, J.; Gao, X., et al. Long noncoding RNA MALAT-1 is a new potential therapeutic target for castration resistant prostate cancer. *The Journal of urology* **2013**, *190*, 2278-2287, doi:10.1016/j.juro.2013.07.001.
57. Tomlins, S.A.; Rhodes, D.R.; Perner, S.; Dhanasekaran, S.M.; Mehra, R.; Sun, X.W.; Varambally, S.; Cao, X.; Tchinda, J.; Kuefer, R., et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science (New York, N.Y.)* **2005**, *310*, 644-648, doi:10.1126/science.1117679.
58. Mani, R.-S.; Tomlins, S.A.; Callahan, K.; Ghosh, A.; Nyati, M.K.; Varambally, S.; Palanisamy, N.; Chinnaiyan, A.M. Induced chromosomal proximity and gene fusions in prostate cancer. *Science (New York, N.Y.)* **2009**, *326*, 1230, doi:10.1126/science.1178124.
59. Delahunt, B.; Miller, R.J.; Srigley, J.R.; Evans, A.J.; Samaratunga, H. Gleason grading: past, present and future. **2012**, *60*, 75-86, doi:https://doi.org/10.1111/j.1365-2559.2011.04003.x.
60. Krušlin, B.; Ulamec, M.; Tomas, D. Prostate cancer stroma: an important factor in cancer growth and progression. *Bosnian journal of basic medical sciences* **2015**, *15*, 1-8, doi:10.17305/bjbms.2015.449.
61. Strasner, A.; Karin, M. Immune Infiltration and Prostate Cancer. *Frontiers in oncology* **2015**, *5*, 128, doi:10.3389/fonc.2015.00128.
62. Joung, J.G.; Bae, J.S.; Kim, S.C.; Jung, H.; Park, W.Y.; Song, S.Y. Genomic Characterization and Comparison of Multi-Regional and Pooled Tumor Biopsy Specimens. *PloS one* **2016**, *11*, e0152574, doi:10.1371/journal.pone.0152574.
63. Cooper, C.S.; Eeles, R.; Wedge, D.C.; Van Loo, P.; Gundem, G.; Alexandrov, L.B.; Kremeyer, B.; Butler, A.; Lynch, A.G.; Camacho, N., et al. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nature genetics* **2015**, *47*, 367-372, doi:10.1038/ng.3221.
64. Boutros, P.C.; Fraser, M.; Harding, N.J.; de Borja, R.; Trudel, D.; Lalonde, E.; Meng, A.; Hennings-Yeomans, P.H.; McPherson, A.; Sabelnykova, V.Y., et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nature genetics* **2015**, *47*, 736-745, doi:10.1038/ng.3315.
65. Visser, W.C.H.; de Jong, H.; Melchers, W.J.G.; Mulders, P.F.A.; Schalken, J.A. Commercialized Blood-, Urinary- and Tissue-Based Biomarker Tests for Prostate Cancer Diagnosis and Prognosis. **2020**, *12*, 3790.