Research Article

iMedPub Journals
www.imedpub.com

Archives in Cancer Research
ISSN 2254-6081

2021
Vol.9 No.S6:002

Principal Component Analysis of RNA-Seq Data Unveils a Novel Prostate Cancer-Associated Gene Expression Signature

Yasser Perera[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

[2]Laboratory of Molecular Oncology, Center of Genetic Engineering and Biotechnology, Havana, 10600, Cuba

[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

[5]Center of Molecular Immunology, Havana, Cuba

[*]**Corresponding author:** Yasser Perera, China-Cuba Biotechnology Joint Innovation Center, Yongzhou Zhong Gu Biotechnology Co., Ltd, Yongzhou City, 425000, Hunan Province, People Republic of China; E-mail: ypereranegrin@ccbjic.com

Received date: September 23, 2021; **Accepted date:** October 07, 2021; **Published date:** October 14, 2021

Citation: Perera Y, Gonzalez A, Perez R (2021) Principal Component Analysis of RNA-Seq Data Unveils a Novel Prostate Cancer-Associated Gene Expression Signature. Arch Can Res Vol.9 No.S6: 002.

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 RNA-seq data normalization to identify a gene expression signature which segregates primary Prostate Adenocarcinoma (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*), 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. Gene Ontology enrichment and correlation analysis comprising major clinical features (i.e., Gleason Score, AR Score, *TMPRSS2-ERG* fusion and Tumor Cellularity) suggest that PC2 and PC3 gene signatures may 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

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 preferred medical intervention [3]. Of note, during their life-time, most of these patients would never experience Pca, therefore the disease is considered over-diagnosed and over-treated [4].

The clinical outcome of Pca is highly variable, and precise prediction of disease's course 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 more tractable Pca molecular alterations [12,13].

Comprehensible genome-wide analysis of primary Prostate Adenocarcinoma (PRAD) revealed 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.

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 heterogeneity, including cell differentiation stages, malignant phenotypes and treatment induced changes, which can be linked to phenotypes and further characterized [18].

Biological meanings are usually captured 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 show that a relatively small number of "core genes" can segregate normal from neoplastic tissues in different tumor localizations [20].

Here, by using such PCA we analyze primary PRAD RNA-seq data to uncover and characterize a novel PRAD-Core Expression Signature (PRAD-CES) which may may "describe" at expression level Pca [21,22].

The PRAD-CES segregates tumor from normal samples along what we call the cancer axis (i.e., PC1), whereas 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.

Materials and Methods

RNA-seq data

For PCA we take RNA-seq tissue expression data from the TCGA Prostate Adenocarcinoma project (TCGA-PRAD, 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 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

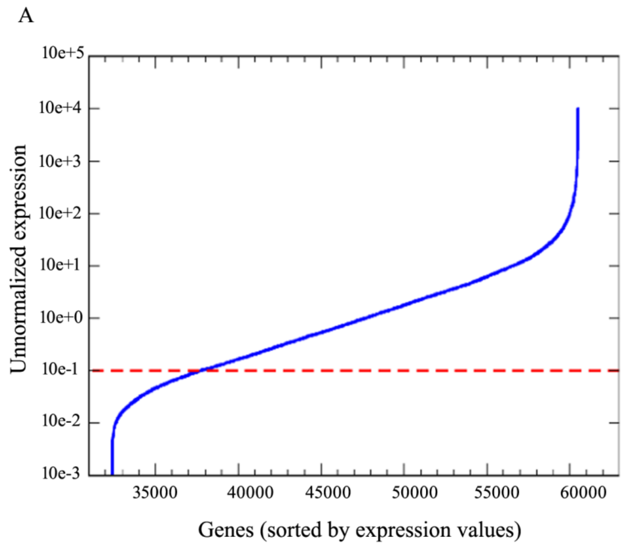

Figure 1: (A) Un-normalized expression data and differential expression profiles in PRAD. Typical range of (un-normalized) expression values from one representative patient (log scale). The red dashed line denotes the expression threshold for statistical significance. Genes with expression below the threshold both in normal tissue and in tumor are mapped to differential expressions very near one.

Figure 1A shows in a typical PRAD sample that the expression of more than 35000 genes is below 0.1. We shift the expression 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, ē=e/eref. Finally, we take the base 2 logarithm, ê=Log2 (ē), 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 ê. We forced the reference for the PC analysis to be at the center of the cloud of normal samples, ê=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: $\sigma^2_{ij}=\Sigma\ \hat{e}_i(s)\ \hat{e}_j(s)/(N_{samples}-1)$, where the sum runs over the samples, s, and Nsamples is the total number of samples in the study. êi(s) is the fold variation of gene i in samples. The dimension of matrix σ2 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 σ2. To this end, we diagonalize σ2 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 Gene cards integrated data sources including but not limited to expression, tissues-specificity, sub-cellular localization and diseases association data [23]. Genome visualizations were done with Ensemble release 100-April 2020, 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]. LncRNA disease 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 was used to uncover ceRNAs among selected LncRNAs [28].

Enrichment analysis

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

Cbioportal

Oncoprint visualizations for selected Genomic Profiles, Alteration Frequency, and mutations representation were obtained from Cbioportal.

Cancer driver repositories and driver prediction platforms

To search for any previous cancer association of identified genes the Cancer Gene Census and OncoKB databases were reviewed [30,31]. The driver prediction platforms IntoGene and ExInAtor 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.

Results

Data normalization surfaced a n age-independent aberrant 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 shown in **Figure 1A**.

Considering sample availability, we dicotomized 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) (Figure S1). 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].

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 included in the analyzed data cohort (i.e., TY *vs.* TO) (**Figures 1B and 1C**).

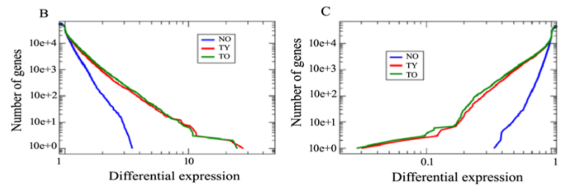

Figure 1: (B, C) Un-normalized expression data and differential expression profiles in PRAD. (B) and around 1000 genes with differential expression above 2 (up-regulated) (C). Notice also that the expression profiles overlap for the TY and TO groups, and apparently differ from the NO profile.

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.

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 defines 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-CES 33) can segregate primary neoplastic samples from normal prostatic tissues with roughly 4% and 8% of false positives and false negatives, respectively **(Figures 2A-2C)**.

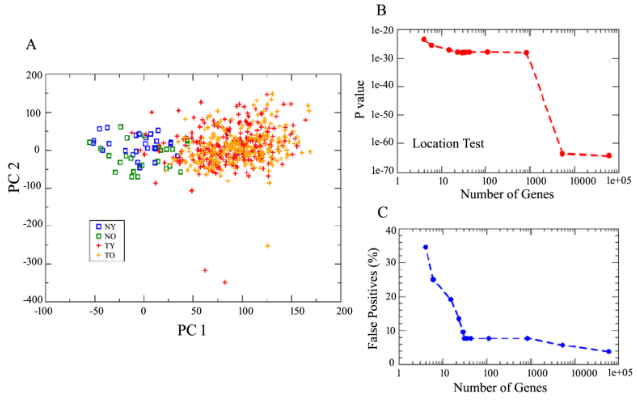

Figure 2: Principal Component Analysis (PCA) of RNAseq-based expression data from PRAD patients. (A) The tumor samples (cloud mean=+91.3) fall apart the distribution of "normal" ones (cloud mean=0.0) along the PC1 axis defined here as the "cancer axis" (p-value=10^-65, Mann-Whitney test). (B) Selecting 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.

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.

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 (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 underexpressed 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. Detailed information about the 33 genes included in the core signature are described (Table S1).

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).

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%). The expression of the corresponding proteins was observed for 9/23 coding genes, whereas 6/10 RNA genes were detected in malignant prostate tissues. Of note, 20/23 (87%) protein coding genes have been previously associated to cancer, 18 (78%) particularly to Pca. Otherwise, 3/10 RNA genes have been connected to Pca (33%)(Table S2).

PRAD-CES genes displayed low mutational burden with less than 5% of all samples having mutations (Figure S3). 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 roughly half of Pca driver genes annotated in the CGC (i.e., 21% *vs.* 42% of cumulative alteration frequency, respectively).

Core expression signature includes emerging drivers and biomarkers

A text-mining indicated at least 18 surfaced genes may play driver roles in PRAD. However, only TP63 is enlisted in the CGC database as Tier 1 driver for NSCLC, HNSCC and DLBCL cancers, but not Pca. None of the remaining 23 protein coding genes populate CGC or OncoKB databases, nor two orthogonal driver prediction tools (i.e. IntoGene and ExInAtor) found further drivers among PRAD-CES genes.

Otherwise, we search for non-coding genes that might be predicted as drivers by Ex-InAtor. *PCA3* was the only significantly mutated LncRNA predicted as a driver, de-spite four of the six LncRNAs were analyzed (i.e., *PCA3, AP006748.1, AP001610.2, ARLNC1*) (data not shown).

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 2.0a	RNA-seq, qPCR, Northern blot	Prostate	Driverα	*CDYL2β*	29808028
PCA3	Lnc2Cancer 2.0a	qPCR, Western blot	Prostate and others	Driver; Biomarker	*PRUNE2*	30569456
PCAT-14	Lnc2Cancer 2.0a	RNA-seq, qPCR, RNAi, ISH	Prostate and others	Driver; Biomarker	*IGLL1, DRICH1*	27566105

AC092535.4	LncRNADiseaseb	Predicted lncRNA-disease	Cervical and others	?	*CTBP1, SPON2, RNF212*	not found
AP001610.2	LncRNADiseaseb	Predicted lncRNA-disease	Cervical and others	?	*TMPRSS2, MX1*	not found
AP002498.1	LncRNADiseaseb	Predicted lncRNA-disease	Cervical and others	?	*CAPN5, B3GNT6, ACER3*	not found
AP006748.1	LncRNADiseaseb	Predicted lncRNA-disease	Cervical and others	?	*TMPRSS2*	not found
PCA3	LncRNADiseaseb	ncRNA-disease causality	Prostate and others	Driver; Biomarker	*PRUNE2*	27743381; 26594800
PCAT-14	LncRNADiseaseb	ncRNA-disease causality	Prostate and others	Driver; Biomarker	*IGLL1, DRICH1*	27460352; 27566105
PCA3	Cancer LncRNA Censusc	qPCR, Western blot	Prostate and others	Driver; Biomarker	*PRUNE2*	27743381; 26594800

Note: 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*; whereas AC139783.1 is transcribed within the *AMACR* protein coding gene **(Figures 3A and 3B)**.

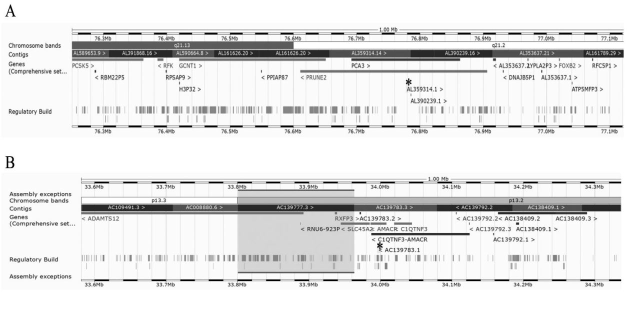

Figure 3: Genome location (*) of the surfaced LncRNAs AL359314.1 (A) and AC139783.1 (B) indicating sequence overlap with protein coding genes PRUNE2 (anti-sense direction) and AMACR (sense direction). Representation of the regions of interest using Ensembl release 100.

Aberrant expression of PRAD-CES genes on independent datasets

The expression of PRAD-CES genes were further analyzed on three independent prostate cancers studies [38-40]. Three putative emerging drivers in PRAD were consistently de-regulated 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 and multiple sites (Figures 4A-4C, Table S3).

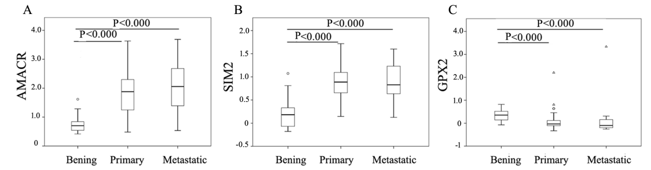

Figure 4: Box plots of z-scores of Benign vs. malignant tissues for AMACR (A), SIM2 (B) and GPX2 (C) genes. For statistics analysis a Kruskal-Wallis test with Bonferroni correction for multiple tests was conducted. Data taken from (Taylor et al., 2010) [40].

Overall, 14 of 33 PRAD-CES genes were included in the Lapointe dataset (Figure S4). Whereas, the expression of 11 of them were consistently up or down regulated 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 (Figure S5). 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*) (Figure S6).

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 **(Figures 5A-5C, Data S1)**.

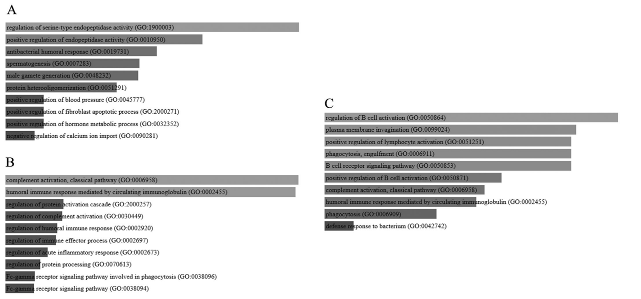

Figure 5: Enrichment analysis for Biological Process using the tool Enrich. PRAD_CES composite genes from PC1 (A) and top 33 genes from PC2 (B) and PC3 (C) were included in the analysis. Statistical significance is in accordance with color from light (highly significant) to dark tones (less significant).

Overall, the PRAD-CES genes (PC1) participate in more diverse BP and pathways compared to genes populated PC2 and PC3 (Data S2). 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.

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 **(Figures 6A-6D, Table 2)**.

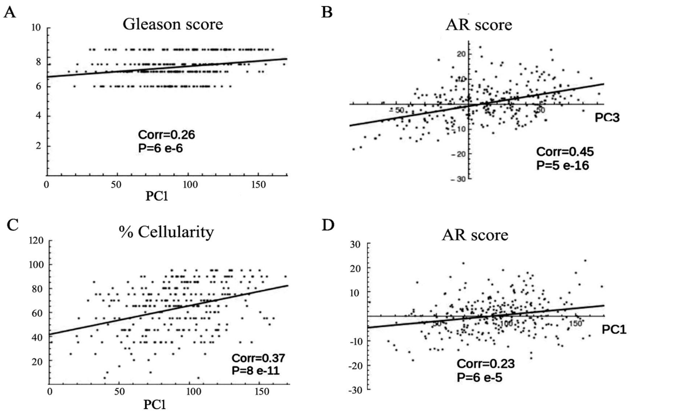

Figure 6: Correlation between PC1 and PC3 and major clinical features of PRAD using the data cohort Prostate Adenocarcinoma (TCGA, Cell 2015), comprising 333 primary tumors. Major features include Gleason-Score (A), AR-Score (B,D), TMPRSS2-ERG and tumor cellularity (C) (see Table 2 for details).

Table 2: Correlations among PCs and selected clinical features of PRAD (TCGA, Cell 2015). A Pearson correlation test was performed.

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}) (Data S3). Therefore, the observed correlation between PC1 values and the above-mentioned clinical features may reflect the underlying PRAD biology which is in line with the fact that PC1 may 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.

Discussion

Here, we use Principal Component Analysis (PCA) to surface a gene expression sig-nature 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 our custom-made normalization. Furthermore, considering that PCA lose resolution on highly heterogeneous and pooled data, we selected only this Pca data cohort to perform our PCA.

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 [41]. Since we used "Normal Young" data as reference, the obtained pattern may suggest that neoplastic transformation over-impose on already age-adjusted global expression profile (i.e., similar TY and TO distribution). However, this notion needs to be verified by using larger and better dichotomized age-based patient cohorts. Of note, the observed long-tail distribution is independent of the type of expression data (i.e. RNA-seq), since similar global gene-expression patterns emerged after analyzing micro-array data using our normalization procedure (data not shown).

The 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 pro-tein-coding driver was found (i.e., *TP63*), three RNA genes with causative roles were surfaced: *ARLNC1, PCA3*, and *PCAT-14* [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 PRAD-CES genes using 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 PRAD-CES genes could not be verified in all datasets. Overall, the most 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 ad-dressed. *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. 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. 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 other transcription-based mechanism.

Finally, we selected four Pca molecular/clinical features to correlate with PCs 1-3. The first, Gleason score, remains as a cornerstone pathological criteria for risk-stratification and disease prognosis [57]. Furthermore, 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 [58,59]. Among such ETS fusion genes, *TMPRSS2-ERG* fusion accounts for 46% of cases. The fourth clinical feature, i.e. tumor cellularity, was used here as a proxy for non-prostatic yet-relevant infiltrating populations [60]. The observed correlations indicated PC1 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 might reveal a transition towards a more aggressive and inflammation-prone phenotype, with a mixture of tumor epithelial cells and infiltrating immune cells [61,62]. 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 highly infiltrated tumor cells endowed to metastasize [63-65].

Conclusion

Overall, our study is limited by data availability/structure and biopsy bias as any global transcriptome inquire. Primary prostate tumors are multi-focal and molecularly heterogeneous; thus, the surfaced gene expression signature may “described” only the sampled site, which also contains different cells from the tumor micro-environment. 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. Finally, as therapeutic options for poorly tractable Pca are limited, the evaluation of putative novel molecular targets populating PRAD-CES seems appealing.

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.

Funding

This research was supported 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.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, et al. (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68(6): 394-424.
2. Al Olama AA, Kote-Jarai Z, Berndt SI, Conti DV, Schumacher F, et al. (2014) A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat Genet 46(10): 1103–1109.

3. Penney KL, Stampfer MJ, Jahn JL, Sinnott JA, Flavin R, et al. (2013) Gleason grade progression is uncommon. Cancer Res 73(16): 5163–5168.

4. Vickers AJ (2019) Redesigning Prostate Cancer Screening Strategies to Reduce Overdiagnosis. Clin Chem 65(1): 39-41.

5. Buyyounouski MK, Pickles T, Kestin LL, Allison R, Williams SG (2012) Validating the interval to biochemical failure for the identification of potentially lethal prostate cancer. J Clin Oncol 30(15): 1857-63.

6. Sathianathen NJ, Konety BR, Crook J, Saad F, Lawrentschuk N (2018) Landmarks in prostate cancer. Nat Rev Urol 15(10): 627-642.

7. D'Amico AV, Whittington R, Malkowicz SB, Schultz D, Blank K, et al. (1998) Biochemical outcome after radical pros-tatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically lo-calized prostate cancer. Jama 280(11): 969-974.

8. Cooperberg MR, Broering JM, Carroll PR (2009) Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. J Natl Cancer Inst 101(12): 878-87.

9. Duffy MJ (2020) Biomarkers for prostate cancer: prostate-specific antigen and beyond. Clin Chem Lab Med 58(3): 326-339.

10. Powers E, Karachaliou GS, Kao C, Harrison MR, Hoimes CJ et al. (2020) Novel therapies are changing treatment paradigms in metastatic prostate cancer. J Hematol Oncol 13: 144.

11. Wang Z, Ni Y, Chen J, Sun G, Zhang X, et al. (2020) The efficacy and safety of radical prostatectomy and radiotherapy in high-risk prostate cancer: a systematic review and meta-analysis. World J Surg Oncol 18(1): 42.

12. Mateo J, Seed G, Bertan C, Rescigno P, Dolling D, et al. (2020) Genomics of lethal prostate cancer at diagnosis and castration resistance. J Clin Invest 130(4): 1743-1751.

13. Luca BA, Moulton V, Ellis C, Edwards DR, Campbell C, et al. (2020) A novel stratification framework for predicting outcome in patients with prostate cancer. British Journal of Cancer 122(10): 1467-1476.

14. Cancer Genome Atlas Research Network (2015) The Molecular Taxonomy of Primary Prostate Cancer. Cell 163(4): 1011-1025.

15. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, et al. (2013) Signatures of mutational processes in human cancer. Nature 500(7463): 415-421.

16. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499(7457): 214-218.

17. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometrics and Intelligent Laboratory Systems 2: 37-52.

18. Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, et al. (2010) A global map of human gene expression. Nat biotechnol 28(4): 322-4.

19. Lenz M, Müller FJ, Zenke M, Schuppert A (2010) Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. Sci Rep 6: 25696.

20. Gonzalez A, Perera Y, Perez RJ (2020) On the gene expression landscape of cancer. Quantitative Biology eprint: 2003.07828.

21. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, et al. (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discov 2(5): 401-404.

22. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, et al. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 6: pl1.

23. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D (1997) GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13(4): 163.

24. Yates AD, Achuthan P, Akanni W, Allen J, Alvarez-Jarreta J, et al. (2020) Ensembl 2020. Nucleic Acids Res 48(D1): D682-D688.

25. Gao Y, Wang P, Wang Y, Ma X, Zhi H, et al. (2019) Lnc2Cancer v2.0: updated database of experimentally sup-ported long non-coding RNAs in human cancers. Nucleic Acids Res 47(D1): D1028-D1033.

26. Bao Z, Yang Z, Huang Z, Zhou Y, Cui Q, et al. (2019) LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases.: Nucleic acids Res 47(D1): D1034-d103.

27. Carlevaro-Fita J, Lanzós A, Feuerbach L, Hong C, Mas-Ponte D, et al. (2020) Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. Commun Biol 3: 56.

28. Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, et al. (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res 44(D1): D239-47.

29. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, et al. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44(W1): W90-7.

30. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, et al. (2018) The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer 18(11): 696-705.

31. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, et al. (2017) OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol 2017: PO.17.00011.

32. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, et al. (2020) A compendium of mutational cancer driver genes. Nat Rev Cancer 20(10): 555-572.

33. Lanzós A, Carlevaro-Fita J, Mularoni L, Reverter F, Palumbo E, et al. (2017) Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. Sci Rep 7: 41544.

34. Li J, Djenaba JA, Soman A, Rim SH, Master VA (2012) Recent trends in prostate cancer incidence by age, cancer stage, and grade, the United States, 2001-2007. Prostate Cancer 2012: 691380.

35. Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, et al. (1999) DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. Cancer Res 59(23): 5975-5979.

36. Salameh A, Lee AK, Cardó-Vila M, Nunes DN, Efstathiou E, et al. (2015) PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3. Proc Natl Acad Sci U S A 112(27): 8403-8408.

37. Canacci AM, Izumi K, Zheng Y, Gordetsky J, Yao JL,et al. (2011) Expression of se-menogelins I and II and its prognostic significance in human prostate cancer. Prostate 71(10): 1108-1114.

38. Ross-Adams H, Lamb AD, Dunning MJ, Halim S, Lindberg J, et al. (2015) Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. EBioMedicine 2(9): 1133-1144.

39. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, et al. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci U S A 101(3): 811-6.

40. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, et al.(2010) Integrative genomic profiling of human prostate cancer. Cancer cell 18(1): 11-22.

41. Sen P, Shah PP, Nativio R, Berger SL (2016) Epigenetic Mechanisms of Longevity and Aging. Cell 166(4): 822-839.

42. Zhang Y, Pitchiaya S, Cieślik M, Niknafs YS, Tien JC, et al. (2018) Analysis of the androgen receptor-regulated lncRNA landscape identifies a role for ARLNC1 in prostate cancer progression. Nat Genet 50(6): 814-824.

43. Wang Y, Hu Y, Wu G, Yang Y, Tang Y, et al. (2017) Long noncoding RNA PCAT-14 induces proliferation and invasion by hepatocellular carcinoma cells by inducing methylation of miR-372. Oncotarget 8(21): 34429-34441.

44. Dhillon PK, Barry M, Stampfer MJ, Perner S, Fiorentino M, et al. (2009) Aberrant cytoplasmic expression of p63 and prostate cancer mortality. Cancer Epidemiol Biomarkers Prev 18(2): 595-600.

45. Rodriguez M, Siwko S, Liu M (2016) Prostate-Specific G-Protein Coupled Receptor, an Emerging Biomarker Regulating Inflammation and Prostate Cancer Invasion. Curr Mol Med 16(6): 526-32.

46. Tang X, Mahajan SS, Nguyen LT, Béliveau F, Leduc R, et al. (2014) Targeted inhibition of cell-surface serine protease Hepsin blocks prostate cancer bone metastasis. Oncotarget 5(5): 1352-1362.

47. Takahara K, Azuma H, Sakamoto T, Kiyama S, Inamoto T, et al. (2009) Conversion of prostate cancer from hormone independency to dependency due to AMACR inhibition: involvement of increased AR expression and decreased IGF1 expression. Anticancer research 29(7): 2497-2505.

48. Liang M, Sun Y, Yang HL, Zhang B, Wen J, et al. (2018) DLX1, a binding protein of be-ta-catenin, promoted the growth and migration of prostate cancer cells. Exp Cell Res 363(1): 26-32.

49. Vinarskaja A, Yamanaka M, Ingenwerth M, Schulz WA (2011) DNA Methylation and the HOXC6 Paradox in Prostate Cancer. Cancers (Basel) 3(4): 3714-25.

50. Gao L, Cheng HY, Dong L, Ye X, Liu YN, et al. (2011) The role of HE4 in ovarian cancer: inhibiting tumour cell proliferation and metastasis. J Int Med Res 39(5): 1645-1660.

51. Leyten GH, Hessels D, Smit FP, Jannink SA, de Jong H, et al. (2015) Identification of a Candidate Gene Panel for the Early Diagnosis of Prostate Cancer. Clin Cancer Res 21(13): 3061-70.

52. Xiao L, Lanz RB, Frolov A, Castro PD, Zhang Z, et al. (2016) The Germ Cell Gene TDRD1 as an ERG Target Gene and a Novel Prostate Cancer Biomarker. Prostate 76(14): 1271-1284.

53. Yu YP, Ding Y, Chen Z, Liu S, Michalopoulos A, et al. (2014) Novel fusion transcripts associate with progressive prostate cancer. Am J Pathol 184(10): 2840-2849.

54. Ahadi A, Brennan S, Kennedy PJ, Hutvagner G, Tran N (2016) Long non-coding RNAs harboring miRNA seed regions are enriched in prostate cancer exosomes. Sci Rep 6: 24922.

55. Mehra R, Udager AM, Ahearn TU, Cao X, Feng FY, et al. (2016) Overexpression of the Long Non-coding RNA SChLAP1 Independently Predicts Lethal Prostate Cancer. Eur Urol 70(4): 549-552.

56. Ren S, Liu Y, Xu W, Sun Y, Lu J, et al. (2013) Long noncoding RNA MALAT-1 is a new potential therapeutic target for castration resistant prostate cancer. J Urol 190(6): 2278-2287.

57. Delahunt B, Miller RJ, Srigley JR, Evans AJ, Samaratunga H (2012) Gleason grading: past, present and future. Histopathology 60(1): 75-86.

58. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, et al. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310(5748): 644-8.

59. Mani RS, Tomlins SA, Callahan K, Ghosh A, Nyati MK, et al. (2009) Induced chromosomal proximity and gene fusions in prostate cancer. Science 326(5957): 1230.

60. Krušlin B, Ulamec M, Tomas D (2015) Prostate cancer stroma: an important factor in cancer growth and progression. Bosn J Basic Med Sci 15(2): 1-8.

61. Strasner A, Karin M (2015) Immune Infiltration and Prostate Cancer. Front Oncol 5: 128.

62. Joung JG, Bae JS, Kim SC, Jung H, Park WY, et al. (2016) Genomic Characterization and Comparison of Multi-Regional and Pooled Tumor Biopsy Specimens. PloS one 11(3): e0152574.

63. Cooper CS, Eeles R, Wedge DC, Van Loo P, Gundem G, et al. (2015) Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nat Genet 47(6): 367-372.

64. Boutros PC, Fraser M, Harding NJ, de Borja R, Trudel D, et al. (2015) Spatial genomic heterogeneity within localized, multifocal prostate cancer. Nat Genet 47(7): 736-745.

65. Visser WCH, de Jong H, Melchers WJG, Mulders PFA, Schalken JA (2020) Commercialized Blood-, Urinary-and Tissue-Based Biomarker Tests for Prostate Cancer Diagnosis and Prognosis. Cancers (Basel) 12(12): 3790.