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

Yasser Perera (1, 2), Augusto Gonzalez (3, 4), Rolando Perez (3,5)

Affiliations:

(1) China-Cuba Biotechnology Joint Innovation Center, Zhong Gu Biotechnology Co., Ltd,Yongzhou City, Hunan Province, People Republic of China
(2) Center of Genetic Engineering and Biotechnology, Havana, 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: China-Cuba Biotechnology Joint Innovation Center (CCBJIC), Yongzhou Zhong Gu Biotechnology Co., Ltd, Yangjiaqiao Street, Lengshuitan District, Yongzhou City 425000, Hunan Province, China; email: yasserperera@outlook.com

Abstract

Prostate cancer is a highly heterogeneous disease and the second more common tumor in males. Molecular and genetic profiles are currently used to identify subtypes and to guide therapeutic intervention. However, roughly 26% of primary prostate cancers of both good and poor prognosis is driven by unknown molecular lesions. Thus, ongoing research aims at providing better prognostic biomarkers and causal molecular alterations to intervene. Here, 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 what we call 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, others with validated/predicted roles as biomarkers (HOXC6, TDRD1, DLX1), and/or cancer drivers (PCA3, ARLN1, PCAT-14). Of note, the PRAD-CES also comprises six over-expressed LncRNAs with no previous association with Pca biology. Overall, the PCA allows us to capture roughly 3/4 of data complexity in the first eight PCs, whereas PC1, PC2 and PC3 altogether comprise 57% of such variability. Finally, GO enrichment analysis and PCs correlation with major PRAD clinical features (i.e. Gleason Score, AR Score, TMPRSS2-ERG fusions and Tumor Cellularity) suggest that PC2 and PC3 gene signatures might describe more aggressive and inflammation-prone transitional forms of PRAD with infiltrating immunosystem. The list of surfaced genes may entails putative novel prognostic biomarkers and/or molecular alterations to intervene.

Introduction

Prostate cancer (Pca) is the second most common cancer in men (Bray et al., 2018). Multiple genetic and demographic factors contribute to the incidence of Pca (Al Olama et al., 2014). Prostate-specific antigen (PSA) screening allows detection of nearly 90% of prostate cancers when still are clinically localized and then surgical removal of malignant tissues is the standard therapeutic intervention (Penney et al., 2013). However, for those tumors spreading beyond the prostatic gland including local or distant metastasis, the prognosis is more dismal (Wang et al., 2020). Major risk stratification systems are based on clinical and pathological parameters such as Gleason score, PSA levels, TNM system and surgical margins status (Sathianathen et al., 2018). Of note, the above risk stratification systems do not adequately predict outcome in all cases (D'Amico et al., 1998; Cooperberg et al., 2009). Therefore, there are renewed expectations on recently emerging molecular features which may improve early prognosis of aggressive forms of Pca and at the same time provide novel therapeutic targets to intervene.

Comprehensible genome-wide analysis of primary Prostate Adenocarcinoma (PRAD) revealed both known and novel molecular lesions for 74% of all tumors (**T**he **C**ancer **G**enome **A**tlas, **TCGA** Research Network, 2015). 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; however, somatic point mutations are relatively less common (Alexandrov et al., 2013; Lawrence et al., 2013). Despite this detailed molecular taxonomy of PRAD, roughly 26% of primary Pca of both, good and poor prognosis, are driven by unknown molecular lesions (**TCGA** Research Network, 2015).

Recently, we used **P**rincipal **C**omponent **A**nalysis (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 (https://arxiv.org/abs/2003.07828). Here by combining such PCA along with resources allocated at Cbioportal (Cerami et al., 2012; Gao et al., 2013), we analyze primary PRAD RNAseq data to uncover a novel **PRAD-C**ore **E**xpression **S**ignature (**PRAD-CES**) which may "describe" Pca. 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 entails putative novel prognostic biomarkers and/or molecular alterations to intervene.

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 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, $\bar{e} = e/e_{ref}$. Finally, we take the base 2 logarithm, $\hat{e} = \mathrm{Log}_2(\bar{e})$, 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: $\sigma^2_{ij} = \Sigma\ \hat{e}_i(s)\ \hat{e}_j(s)\ /$ (Nsamples-1), where the sum runs over the samples, s, and $N_{samples}$ is the total number of samples in the study. $\hat{e}_i(s)$ is the fold variation of gene i in sample s.

The dimension of matrix σ^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 σ^2. To this end, we diagonalize σ^2 by means of a Lanczos routine in Python language, from which we get the first 100 eigenvalues and their corresponding eigenvectors.

Gene information and genome visualization

General gene information was collected from Genecards integrated data sources (www.genecards.org) (Rebhan et al., 1997) including but not limited to expression, tissues-specificity, sub-cellular localization and diseases association data. Genome visualizations were done with Ensembl release 100 - April 2020

(https://www.ensembl.org), Genome assembly: GRCh38.p13 (GCA_000001405.28) (Yates et al., 2020).

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 (Gao et al., 2019). LncRNADisease 2.0: contains experimentally and/or computationally supported data (Bao et al., 2019). Cancer LncRNA Census (CLC): a compilation of 122 GENCODE lncRNAs with causal roles in cancer phenotypes (Carlevaro-Fita et al., 2017). The miRTarBase http://miRTarBase.mbc.nctu.edu.tw/ was used to uncover ceRNAs among selected LncRNAs (Chou et al., 2016).

Enrichment analysis

The enrichment analysis was performed using the Enrich platform (Kuleshov et al., 2016) and the following categories: Ontologies (GO_Biological_Process_2018) and Pathways (Reactome_2016).

Cbioportal

Oncoprint visualizations for selected Genomic Profiles, Alteration Frequency, and mutations representation were obtained from Cbioportal, https://www.cbioportal.org/ (Cerami et al., 2012; Gao et al., 2013).

Cancer Driver repositories and driver prediction platforms

To search for any previous cancer association of identified genes the Cancer Gene Census (Sondka et al., 2018) and OncoKB (Chakravarty et al., 2017, http://oncokb.org) databases were reviewed. The driver prediction platforms IntoGene (Martínez-Jiménez et al., 2020, https://www.intogen.org/search) and ExInAtor (Lanzós et al., 2017, https://www.gold-lab.org/cancer-driver-lncrna-prediction-sof) were used to predict a potential driver role for protein-coding and non-coding genes.

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 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 (**Fig 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)(**Supplementary Fig 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 in 4 new PRAD diagnostic cases occurs below 60 years, whereas the mean diagnosis age is 66 years (Li *et al.*, 2012).

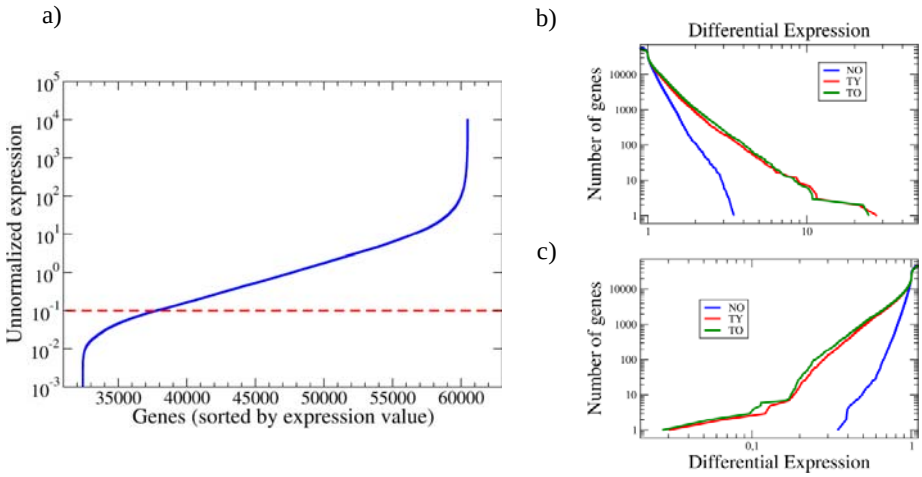

Fig. 1 Un-normalized expression data and differential expression profiles in PRAD. a) Typical range of (un-normalized) expression values from one representative patient (log scale). The red dashed line denote the expression threshold for statistical significance. Genes with expression below the threshold 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) (**Fig 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.

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 **C**ore **E**xpression **S**ignature 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 (**Fig. 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.

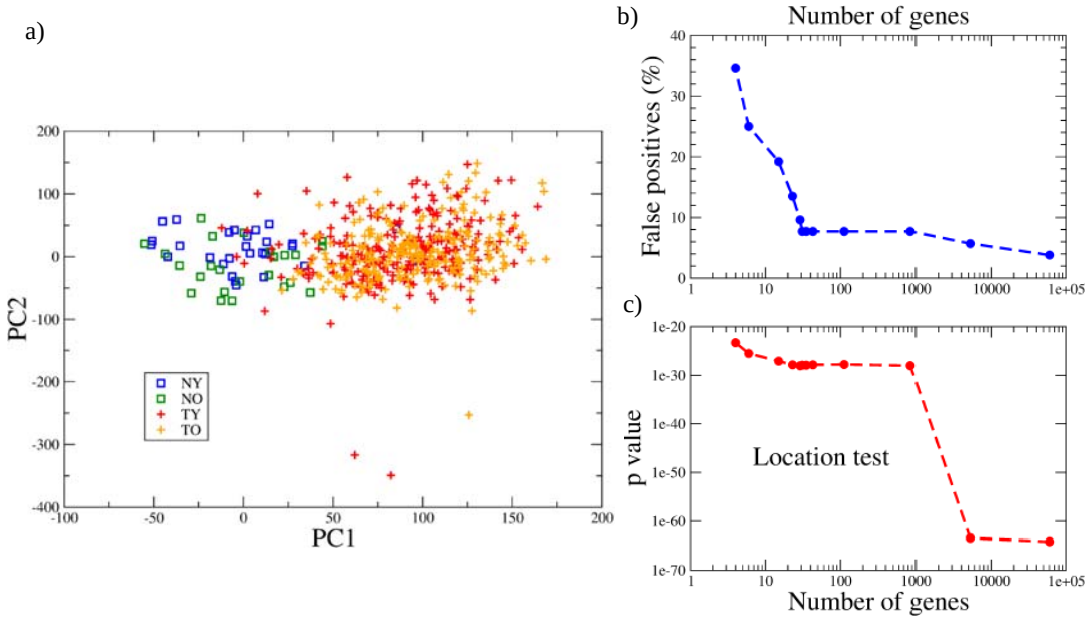

Fig 2. Principal Component Analysis (PCA) of RNAseq-based expression data from PRAD patients. The tumor samples (cloud mean=+91.3) fall apart the distribution of "normal" ones (cloud mean= 0.0) along the PC1 axis defined here as the "cancer axis" (p-value = 10^-65, Mann-Whitney test). Selecting a PC1 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. 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 $x_1 = \Sigma \hat{e}_i v_{1i}$, where v_{1i} 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 Fig 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

(Bussemakers et al., 1999; Salameh et al., 2015). Otherwise, the most under-expressed genes within this PRAD-CES is the protein coding gene SEMG1 (Canacci et al., 2011). Further bardcode-like analysis of top-100 genes contributing to PC1 axis shown a similar profile (**Supplementary Fig S2**). Detailed information about the 33 genes included in the core signature are described in **Supplementary Table 1**.

Notice that a picture like **Fig 2, right panel** 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 $\mathbf{v}_1$.

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). This results imply that not only the global normalized gene expression profile are similar among TY and TO in PRAD cases, rather that an 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, LncRNA, and ncRNA (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 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**).

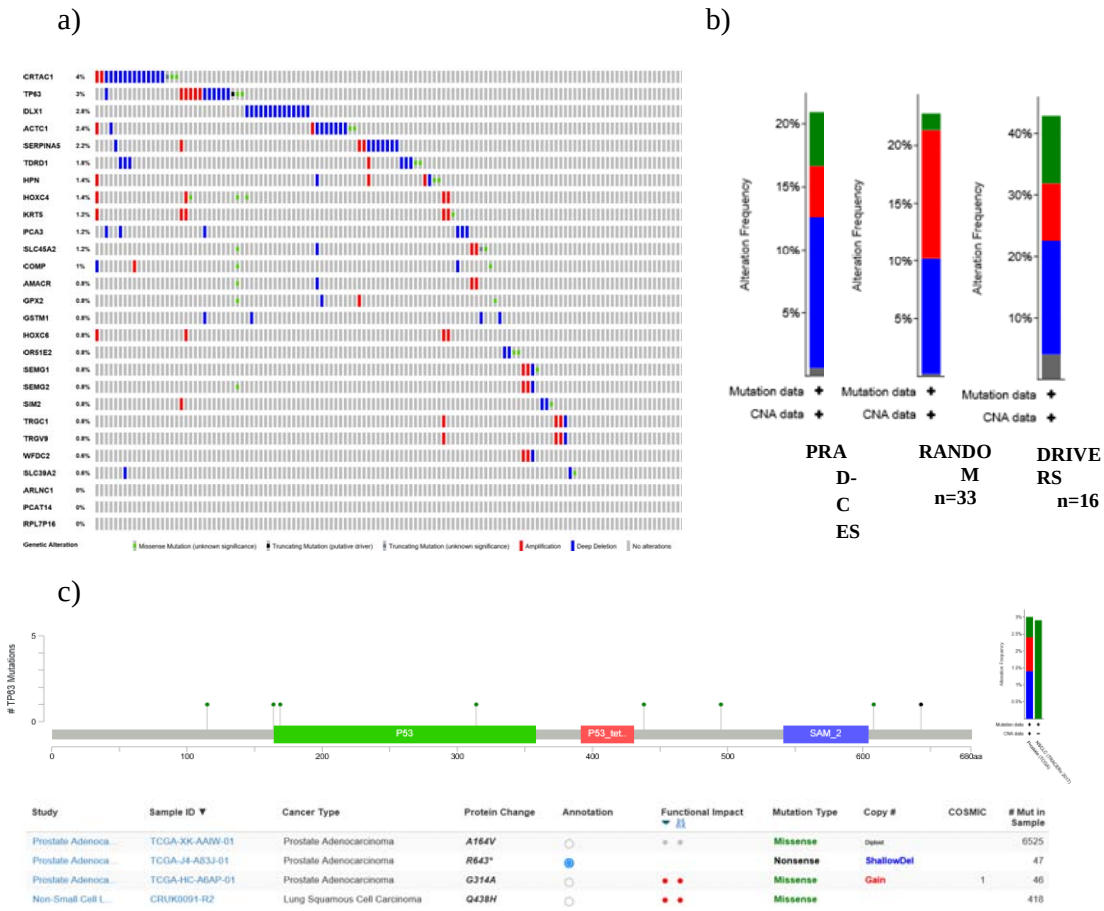

Figure 3. Representation of the pattern and type of molecular alteration in PRAD samples according to Cbioportal (TCGA, Firehose Legacy) data cohort. a) Oncoprint representing genomic alterations for PRAD-CES33 composite genes; b) Frequency of genomic alterations observed in PRAD for PRAD-CES33 genes, 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).

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 predicted by ExInAtor as a driver, despite of 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 LncRNAs (i.e. ARLNC1, PCA3, PCAT-14), six LncRNAs (i.e. AC092535.4, AP001610.2, AP002498.1, AP006748.1, PCA3 and PCAT14), and one LncRNAs (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.

Table1. LncRNA included in PRAD-CES and their association with cancer according to indicated databases.

LncRNA	Databases	Method	Tumor type*	Role	mRNA targets	Refererence
ARLNC1	Lnc2Cancer 2.0[a]	RNA-seq, qPCR, Northern blot	Prostate	Driver[α]	CDYL2[β]	29808028
PCA3		qPCR, Western blot	Prostate & Others	Driver; Biomarker	PRUNE2	30569456
PCAT-14		RNA-seq, qPCR, RNAi, ISH	Prostate & Others	Driver; Biomarker	IGLL1, DRICH1	27566105
AC092535.4	LncRNADisease[b]	Predicted lncRNA-disease	Cervical cancer & Others (not Prostate)	?	CTBP1, SPON2, RNF212	not found
AP001610.2		Predicted lncRNA-disease	Cervical cancer & Others (not Prostate)	?	TMPRSS2, MX1	not found
AP002498.1		Predicted lncRNA-disease	Cervical cancer & Others (not Prostate)	?	CAPN5, B3GNT6, ACER3	not found
AP006748.1		Predicted lncRNA-disease	Cervical cancer & Others (not Prostate)	?	TMPRSS2	not found
PCA3		ncRNA-disease causality	Prostate & Others	Driver; Biomarker	PRUNE2	27743381; 26594800
PCAT-14		ncRNA-disease causality	Prostate & Others	Driver; Biomarker	IGLL1, DRICH1	27460352; 27566105
PCA3	Cancer LncRNA Census[c]	ncRNA-disease causality	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 other 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 (Salameh et al., 2015); whereas AC139783.1 is transcribed within the AMACR protein coding gene (**Figure 4a, b**).

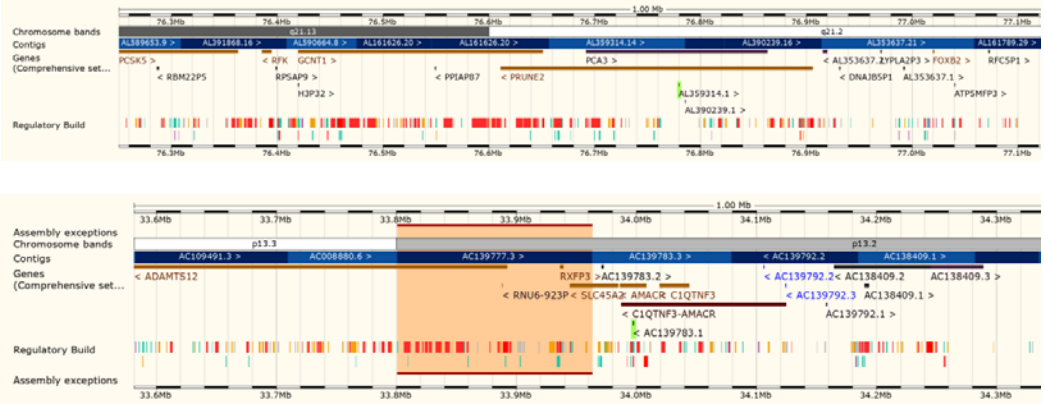

Figure 4. Genome location of the surfaced LncRNAs AL359314.1 (a) and AC139783.1 (b) indicating sequence overlap with protein coding genes PRUNE2 (anti-sense direction) and AMACR (sense direction). Representation of the regions of interest using Ensembl release 100.

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 Fig 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 Fig 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. 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 (**Fig 5, Supplementary Fig 4**).

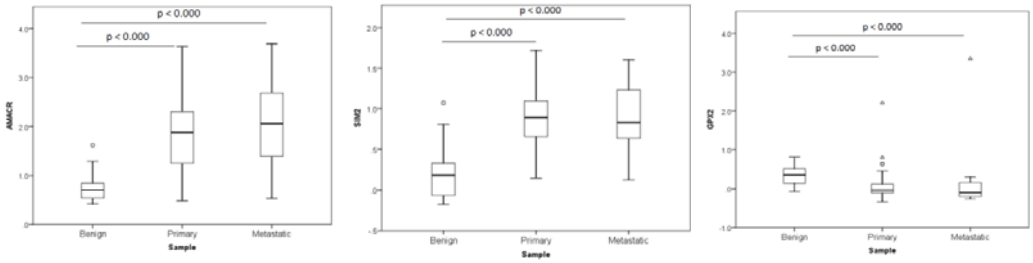

Fig 5. The figures shown Box plots of z-scores of Benign vs malignant tissues (Taylor et al., 2010). For statistics analysis a Kruskal-Wallis test with Bonferroni correction for multiple tests was conducted.

Overall, 14 of 33 PRAD-CES genes were included in the Lapointe dataset (**Supplementary Fig 4**). 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. 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 Fig 4**).

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 (**Fig. 6a-c, Supplementary Table 3**).

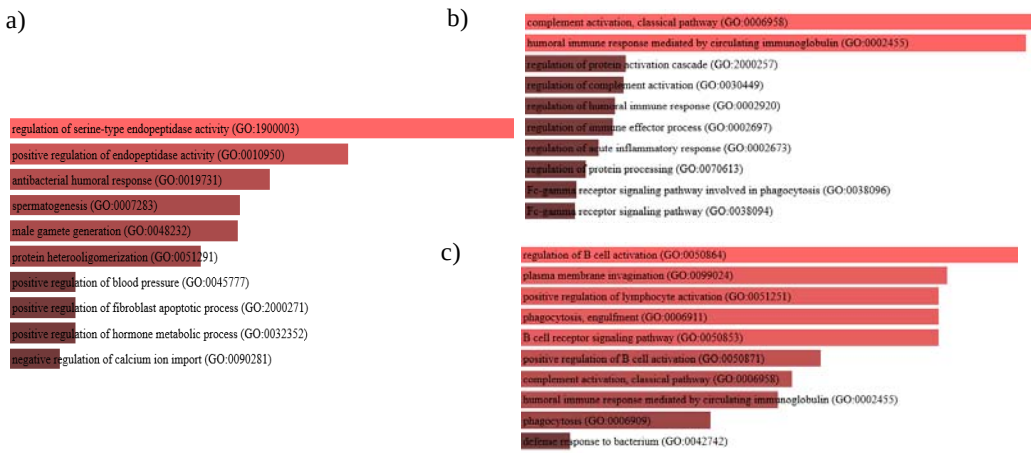

Fig 6. Enrichment analysis for Biological Process using the tool Enrich. PRAD_CES composite genes from PC1 (a) and top 33 genes from PC2 (b) and PC3 (c) were included in the analysis. Statistical significance are in accordance with color from light (highly significant) to dark tones (less significant).

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.

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

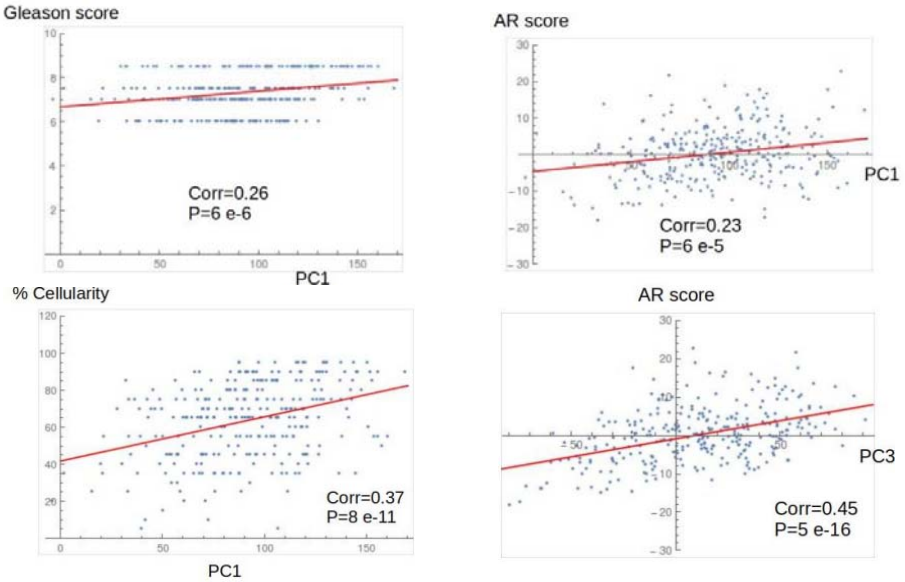

Fig 7. Correlation between PCs 1-3 and major clinical features of PRAD using the data cohort: Prostate Adenocarcinoma (TCGA, Cell 2015) comprising 333 primary tumors. Major features include R-Score, EGR-fusion, Gleason-Score and tumor cellularity.

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 simply reflects the underlying PRAD biology which is in line with the fact that PC1 explain up to 39% of data dispersion, 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)**.

Discussion

Here, we use Principal Component Analysis (PCA) and custom-made normalization to surface gene expression patterns which may characterize primary PRAD. Such dimensionality reduction algorithm segregates tumor from normal samples, with eight PCs capturing roughly 3/4 of data variance. The RNA-seq input data was obtained from the Prostate Adenocarcinoma (TCGA, Firehose Legacy) cohort, which comprised a significant number of both Pca tumor and normal tissue data. Furthermore, we used two other major Prostate Adenocarcinoma genomic data cohorts (i.e. TCGA, Cell 2015 and MSKCC, Cancer Cell 2010), since they provide a required clinical data (TCGA, Cell 2015) and a similar data normalization for RNA-seq based gene expression analysis (MSKCC, Cancer Cell 2010). Whereas other excellent repositories of Pca are available at Cbioportal (Cerami et al., 2012), the above mentioned data cohorts fullfit our attempts to describe primary PRAD through an RNAseq-based gene expression analysis.

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) (Sen et al., 2016). Since we used "normal Young" data as reference, the obtained pattern may suggest that neoplastic transformation over-impose 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 by using the same normalization procedure (data not shown). Of note, this normalization was used for subsequent PCA analysis.

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) (Dhillon et al., 2009), three RNA genes with causative roles were surfaced: ARLNC1, PCA3, and PCAT-14 (Zhang et al., 2018; Salameh et al., 2015; Wang et al., 2017). Otherwise, six protein coding genes awaits further validation concerning PRAD driver roles: OR51E2, HPN, AMACR, DLX1, HOXC6 and WFDC2 (Rodriguez et al., 2016; Tang et al., 2014; Takahara et al., 2009; Liang et al., 2018; Vinarskaja et al., 2011; Gao et al., 2011).

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 (Leyten *et al.*, 2015). TDRD1 might also play an important role in prostate cancer

development, and as a cancer/testis antigen, a potential therapeutic target for cancer immunotherapy (Xiao et al., 2016).
Of note, cross-validation of genes included in PRAD-CES against independent data cohorts and/or previous studies, indicated that most of these genes were consistently and significantly 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). 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 LncRNAs within the PRAD-CES. LncRNA over-expression have been related with progression of PCa (Ahadi et al., 2016). Furthermore, LncRNA over-expression in Pca could be used as prognostic factors and potential therapeutic targets (Mehra et al., 2016; Ren et al., 2013). Of note, at least half of surfaced LncRNAs may exerts driver roles in PCa according to annotated databases and genome localization. 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 (Yu et al., 2014). On the other hand, among the several miRNAs which may down-regulate AMACR expression in Pca, the potential sponging of hsa-miR-26a-5p by the surfaced LncRNA 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 (Takahara et al., 2009).

The most frequent molecular abnormalities in PRAD involved gene-fusions, copy-number alterations and epigenetic deregulation (TCGA Research Network, 2015). Such findings imply that a significant proportion of the neoplastic events might be captured by overall transcript expression analysis. As a matter of facts, the mutational burden observed in surfaced PRAD-CES genes was scarce, suggesting that expression levels and not mutations may play a significant role within the surfaced pattern. Otherwise, less than 3% of the PRAD samples included in our study displayed CNV, thus suggesting that most of gene expression deregulation arose from transcription-based regulation 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 (Lin et al., 2009; Mani et al., 2009; Tomlins et al., 2005). Among such ETS fusion genes, TMPRSS2-ERG fusion accounts for 46% of cases (TCGA Research Network, 2015). On the other hand, Gleason score remains as a cornerstone pathological criteria for risk-stratification and prognosis (Delahunt et al.,

2012). Finally, we include tumor cellularity as a proxy for non-prostatic yet-relevant infiltrating populations (Krušlin et al., 2015).
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 last variable and AR Score. On the other hand, 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 (Strasner and Karin, 2015). 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 endowed to metastasize.

Altogether, our study is limited by data availability/structure and biopsy bias as any global transcriptome inquire (Joung et al., 2016). Primary prostate tumors are multifocal and molecularly heterogeneous, thus the surfaced signature may reflect only gene expression features from the inquired site (Cooper et al., 2015; Boutros et al., 2015). However, our PCA simultaneously uncover relevant PRAD genes found dispersed across previous studies; whereas at the same time emerged new putative drivers and/or biomarkers that needs to be validated. Considering 26% of primary prostate cancers is driven by as-yet-unexplained molecular alterations, the evaluation of the proposed PRAD-CES seems appealing.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/...

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. AG acknowledges the Cuban Program for Basic Sciences and the Office of External Activities of the Abdus Salam Centre for Theoretical Physics for support. The research is carried on under a project of the Platform for Bio-informatics of BioCubaFarma, Cuba. The data for the present analysis come from the TCGA Research Network: https://www.cancer.gov/tcga.

AUTHOR CONTRIBUTIONS

Conceptualization & Investigation, all authors; Software and formal analysis, A.G.; Writing –Original Draft, Y.P.; Writing – Review & Editing, all authors.

DECLARATION OF INTERESTS

The authors have no competing interests to declare.

References

Ahadi A, Brennan S, Kennedy PJ, Hutvagner G and Tran N: Long non-coding RNAs harboring miRNA seed regions are enriched in prostate cancer exosomes. Sci Rep 6: 24933, 2016.

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.; Breast and Prostate Cancer Cohort Consortium (BPC3); PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium; COGS (Collaborative Oncological Gene-environment Study) Consortium; GAME-ON/ELLIPSE Consortium (2014). A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat. Genet. 46, 1103–1109.

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.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013). Signatures of mutational processes in human cancer. Nature 500, 415–421.

Bao Z, Yang Z, Huang Z, et al. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res 2019 47(D1):D1034-D1037. doi: 10.1093/nar/gky905. PMID: 30285109

Boutros PC, Fraser M, Harding NJ, et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. Nat. Genet. 2015 47:736–745.

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018, 68(6):394-424. doi: 10.3322/caac.21492.

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

Canacci AM, Izumi K, Zheng Y, et al. Expression of semenogelins I and II and its prognostic significance in human prostate cancer. Prostate . 2011 71(10):1108-14. doi: 10.1002/pros.21323. PMID: 21557275

Carlevaro-Fita J, Lanzós A, Feuerbach L, et al. Unique genomic features and deeply-conserved functions of long non-coding RNAs in the Cancer LncRNA Census (CLC) https://www.biorxiv.org/content/early/2017/08/25/152769). doi: 10.1101/152769.

Cerami E, Gao J, Dogrusoz U,et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery. 2012 2; 401. doi: 10.1158/2159-8290.CD-12-0095. PMID: 22588877

Chakravarty D, Gao J, Phillips S, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precision Oncology 2017 :1, 1-16 doi:10.1200/PO.17.00011

Chen et al., LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 2013 Jan 1;41(D1):D983-6.

Chou CH, Chang NW, Shrestha S, et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Research 2016 44(D1):D239–D247. DOI: 10.1093/nar/gkv1258

Cooper CS, Eeles R, Wedge DC, et al. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nat. Genet. 2015 47:367–372.

Cooperberg, M.R., Broering, J.M., and Carroll, P.R. (2009). Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. J. Natl. Cancer Inst. 101, 878–887.

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., and Wein, A. (1998). Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. JAMA 280, 969–974.

Delahunt B, Miller RJ, Srigley JR, Evans AJ, Samaratunga H. Gleason grading: past, present and future. Histopathology. 2012;60(1):75–86. doi: 10.1111/j.1365-2559.2011.04003.x.

Dhillon PK, Barry M, Stampfer MJ,et al. ABERRANT CYTOPLASMIC EXPRESSION OF P63 AND PROSTATE CANCER MORTALITY. Cancer Epidemiol Biomarkers Prev. 2009 Feb; 18(2): 595–600. doi: 10.1158/1055-9965.EPI-08-0785. PMID: 19155438

Gao J, Aksoy BA, Dogrusoz U et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 6, pl1 (2013). doi: 10.1126/scisignal.2004088. PMID: 23550210

Gao L, Cheng HY, Dong L, et al. The role of HE4 in ovarian cancer: inhibiting tumour cell proliferation and metastasis. J Int Med Res 2011 39(5):1645-60. doi:10.1177/147323001103900507. PMID: 22117965

Gao Y, Wang P, Wang Y, et al. Lnc2Cancer v2.0: updated database of experimentally supported long non-coding RNAs in human cancers. Nucleic Acids Res 2019 47(D1):D1028-D1033. doi: 10.1093/nar/gky1096. PMID: 30407549

Joung JG, Bae JS, Kim SC, et al. Genomic Characterization and Comparison of Multi-Regional and Pooled Tumor Biopsy Specimens. PLoS One 2016 11(3):e0152574. doi: 10.1371/journal.pone.0152574
Kuleshov MV, JonesMR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, JenkinsSL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, GundersenGW, Ma'ayan A. Enrichr: a comprehensive gene set enrichmentanalysis web server 2016 update. Nucleic Acids Research.2016; gkw377.

Krušlin B, Ulamec M, Tomas D. Prostate cancer stroma: an important factor in cancer growth and progression. Bosn J Basic Med Sci. 2015;15(2):1–8. doi: 10.17305/bjbms.2015.449.
Lanzós A, Carlevaro-Fita J,Mularoni L, et al. Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. Sci Rep. 2017; 7: 41544. doi: 10.1038/srep41544 PMID: 28128360

Lapointe J, Li C, Higgins JP, van de Rijn M et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci USA 2004 Jan 20;101(3):811-6.

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. (2013). Mutational heterogeneity in cancer and the search for new cancer associated genes. Nature 499, 214–218.

Leyten GHJM, Hessels D, Smit FP, et al. Identification of a Candidate Gene Panel for the Early Diagnosis of Prostate Cancer. Clin Cancer Res; 21(13) July 1, 2015

Li J, Djenaba JA, Soman A, et al., 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

Liang M, Sun Y, Yang HL, et al. DLX1, a binding protein of beta-catenin, promoted the growth and migration of prostate cancer cells. Exp Cell Res 2018 Feb 1;363(1):26-32. doi: 10.1016/j.yexcr.2018.01.007.PMID: 29317218

Martínez-Jiménez F, Muiños F, Sentís I, et al. A compendium of mutational cancer driver genes. Nature Reviews Cancer 2020; doi:10.1038/s41568-020-0290-x

Mehra R, Udager AM, Ahearn TU, Cao X, Feng FY, Loda M, Petimar JS, Kantoff P, Mucci LA and Chinnaiyan AM: Overexpression of the long non-coding RNA SChLAP1 independently predicts lethal prostate cancer. Eur Urol 70: 549-552, 2016.

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. (2013). Gleason grade progression is uncommon. Cancer Res. 73, 5163–5168.

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

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. J Urol 190: 2278-2287, 2013

Rodriguez M, Siwko S, Liu M. Prostate-Specific G-Protein Coupled Receptor, an Emerging Biomarker Regulating Inflammation and Prostate Cancer Invasion. Curr Mol Med 2016;16(6):526-32. doi: 10.2174/1566524016666160607091333. PMID: 27280498

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

Salameh A, Lee AK, Cardó-Vila M. PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3 . Proc Natl Acad Sci U S A . 2015 Jul 7;112(27):8403-8. doi: 10.1073/pnas.1507882112.

Sathianathen NJ, Konety BR, Crook J, Saad F, Lawrentschuk N. Landmarks in prostate cancer. Nat Rev Urol. 2018;15(10):627–642. doi: 10.1038/s41585-018-0060-7

Sen P, Shah PP, Nativio R, et al. Epigenetic mechanisms regulating longevity and aging. Cell. 2016 Aug 11; 166(4): 822–839. doi: 10.1016/j.cell.2016.07.050. PMID: 27518561

Siegel, R.L., Miller, K.D., and Jemal, A. (2015). Cancer statistics, 2015. CA Cancer J. Clin. 65, 5–29.

Taylor BS, Schultz N, Hieronymus H, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010;18(1):11–22.

Sondka, Z., Bamford, S., Cole, C.G. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer 18, 696–705 (2018). https://doi.org/10.1038/s41568-018-0060-1

Strasner A, Karin M. Immune infiltration and prostate cancer. Front. Oncol. 2015 doi.org/10.3389/fonc.2015.00128

Takahara K, Azuma H, Sakamoto T, 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 Res 2009 29(7):2497-505. PMID: 19596919

Tang X, Mahajan SS, Nguyen LT, et al. Targeted inhibition of cell-surface serine protease Hepsin blocks prostate cancer bone metastasis. Oncotarget. 2014; 5:1352-1362. https://doi.org/10.18632/oncotarget.1817

The Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015 Nov 5; 163(4):1011-25. doi: 10.1016/j.cell.2015.10.025.

Tomlins, S.A., Alshalalfa, M., Davicioni, E., Erho, N., Yousefi, K., Zhao, S., Haddad, Z., Den, R.B., Dicker, A.P., Trock, B.J., et al. (2015). Characterization of 1577 primary prostate cancers reveals novel biological and clinicopathologic insights into molecular subtypes. Eur. Urol. 68, 555–567.

Tomlins, S.A., Bjartell, A., Chinnaiyan, A.M., Jenster, G., Nam, R.K., Rubin, M.A., and Schalken, J.A. (2009). ETS gene fusions in prostate cancer: from discovery to daily clinical practice. Eur. Urol. 56, 275–286.

Tomlins, S.A., Laxman, B., Dhanasekaran, S.M., Helgeson, B.E., Cao, X., Morris, D.S., Menon, A., Jing, X., Cao, Q., Han, B., et al. (2007). Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 448, 595–599.

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. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310, 644–648.

Vinarskaja A, Yamanaka M, Ingenwerth M, Schulz WA. DNA Methylation and the HOXC6 Paradox in Prostate Cancer. Cancers 2011, 3(4), 3714-3725; doi:10.3390/cancers3043714

Wang Y, Hu Y, Wu G, et al. Long noncoding RNA PCAT-14 induces proliferation and invasion by hepatocellular carcinoma cells by inducing methylation of miR-372. Oncotarget. 2017 8(21): 34429–34441. doi: 10.18632/oncotarget.16260. PMID: 28415780.

Wang Z, Ni Y, Chen J, et al. The efficacy and safety of radical prostatectomy and radiotherapy in high-risk prostate cancer: a systematic review and meta-analysis. World J Surg Oncol. 2020; 18: 42.doi: 10.1186/s12957-020-01824-9

Xiao L, Lanz RB, Frolov A, et al. The Germ Cell Gene TDRD1 as an ERG Target Gene and a Novel Prostate Cancer Biomarker. Prostate 2016, 76:1271-1284.

Yates AD, Achuthan P, Akanni W, et al. Ensembl 2020. Nucleic Acids Research 2020 48(D1):D682–D688. doi: 10.1093/nar/gkz966

Yu YP, Ding Y, Chen Z, et al. Novel Fusion Transcripts Associate with Progressive Prostate Cancer. The American Journal of Pathology 2104 184(10). doi:10.1016/j.ajpath.2014.06.025

Zhang Y, Pitchiaya S, Cieślik M, et al. Analysis of the androgen receptor-regulated lncRNA landscape identifies a role for ARLNC1 in prostate cancer progression. Nat Genet 2018 50(6):814-824. doi: 10.1038/s41588-018-0120-1. PMID: 29808028