Original Article

Unsupervised clustering reveals new prostate cancer subtypes

Shaowei Gao[1*], Zeting Qiu [1*], Yiyan Song[2*], Chengqiang Mo[3], Wulin Tan[1], Qinchang Chen[2], Dong Liu[2], Mengyu Chen[2], Huaqiang Zhou[1,2]

[1]Department of Anesthesia, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, China; [2]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China; [3]Department of Urology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, China

Contributions: (I) Conception and design: H Zhou, S Gao; (II) Administrative support: C Mo, W Tan; (III) Provision of study materials or patients: Z Qiu, C Mo; (IV) Collection and assembly of data: Y Song, Q Chen; (V) Data analysis and interpretation: S Gao, Z Qiu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*These authors contributed equally to this work.

Correspondence to: Huaqiang Zhou. Department of Anesthesia, The First Affiliated Hospital of Sun Yat-sen University, 58 Zhongshan 2 Road, Guangzhou 510080, China. Email: liujiaosuan@gmail.com.

Background: Prostate cancer is the second most common cancer in men. It is urgent to develop a genetic classification for prostate cancer. We aimed to establish the basis of genetic typing.

Methods: We used four series of prostate cancer data. The Cancer Genome Atlas (TCGA) RNA-Seq data were used to train the classifier. Three subgroups based on the classifier were tested whether to have significant differences in the clinical data. The other three sets were classified by the classifier and validated with respective clinical data.

Results: The classifier had 183 genes. Prostate cancer subtype 1 (PCS1) was characterized by high expression of GSTP1, with lower Gleason scores (P<0.001). PCS2 had higher Gleason score, more lymph node invasion (P=0.005) and higher pathology T stage (pT stage) (P<0.001). Three GEO (Gene Expression Omnibus) validation datasets had similar results. We even observed significances in the recurrence time among different subgroups (P=0.005 in GSE70768).

Conclusions: We established a PCS classifier (183 genes) based on RNA-Seq data, and identified three PCSs. The classification was robustly relating to clinical data which may have potential for clinical use.

Keywords: Prostate cancer; the Cancer Genome Atlas (TCGA); Gleason score; prostate cancer subtype classifier (PCS classifier)

Submitted Dec 10, 2016. Accepted for publication Mar 29, 2017.
doi: 10.21037/tcr.2017.05.15
View this article at: http://dx.doi.org/10.21037/tcr.2017.05.15

Introduction

Despite the fact that the advent and innovation of the screening technology has made prostate cancer easier to be diagnosed, with an ever rising incidence, it has been the second cause of cancer-related death among American males (1-3). More than 95 percent of prostate cancer presents as adenocarcinoma. Thus, it is of great importance to explore the developing and prognostic stratification of prostate cancer, especially the molecular classification.

Based upon architectural features of prostate cancer cells, the Gleason score is an efficient way to evaluate the malignancy of prostate cancer. Traditionally, the Gleason score is calculated by adding together the numerical values of the two most prevalent differentiation patterns (a primary grade and a secondary grade). However, a revised grading system (*Table S1*) has been adopted in 2016 World Health Organization classification of genitourinary tumors (4,5). An increasing risk of biochemical recurrence (BCR) with elevating grade has been observed (6). Besides the clinical indexes, some protein-coding genes such as c-myc, Bcl-2 and p53 have also been reported to be related to the

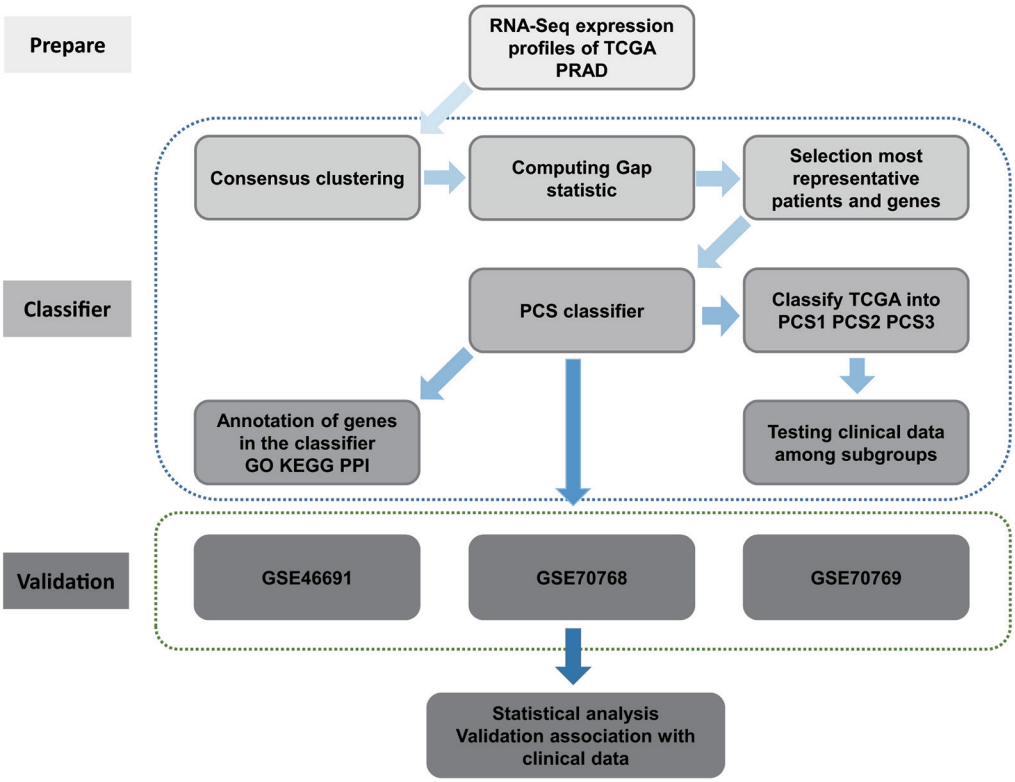

Figure 1 The workflow of our data analyses.

prognosis of prostate cancer over the last decades (7,8).

The Cancer Genome Atlas (TCGA) is a large project composed of multiple components and equipped with various custom applications to handle large volumes of research data (9). It contains genomic characterization data, high level sequencing data and corresponding clinical data of all common tumors and several rare tumors. The integrative analyses made by the TCGA research network in some tumors like glioma and ovarian cancer provided a new insight into their diagnoses and treatment strategies (10,11). A global analysis of prostate adenocarcinoma has been published by the TCGA research network, revealing molecular heterogeneity and potentially actionable molecular defects among primary prostate cancers (12).

As a subfield of computer science, machine learning plays an important role in bioinformatics for its incomparable advantage in handling big data. Up to now, it has made great contribution in the discovery of many valuable results (13,14). In the current study, we applied the previous methods to the TCGA database analysis, and found out some interesting results, which could be validated stably in three other datasets regardless of platforms.

Methods

We summarized our workflow in the *Figure 1*.

Data preparing

We used four different series of prostate cancer data with 1258 unique patients in this study (*Table S2*). TCGA RNAseqV2 dataset consisted of 497 prostate adenocarcinoma samples as a training dataset. The validation datasets comprised three prostate cancer series from GEO (Gene Expression Omnibus): Erho's series (GSE46691) with 545 patients (15), Lamb's set (GSE70768) with 125 patients (16) and Ross's set (GSE70769) with 91 patients (16) respectively.

The training dataset (TCGA RNAseqV2 and Clinical data)

Prostate adenocarcinoma level-3 data were obtained from TCGA consortium (https://tcga-data.nci.nih.gov/tcga/). This dataset included 497 individuals with prostate adenocarcinoma. The RNA sequences of each sample were

profiled based on Illumina HiSeq 2000 RNA Sequencing Version 2 analysis (https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2). Tumor samples from TCGA were not only in different institutions and distinct times, but also processed in batches rather than at the same time, which would lead to misleading analysis resulting from systematic noise such as batch effects and trend effects. TCGA BatchEffects website (http://bioinformatics.mdanderson.org/tcgambatch/) was used to assess and correct for batch effects in prostate adenocarcinoma data. We downloaded the batch effects corrected data processed by Empirical Bayes online. Besides, clinical data was downloaded through R package of TCGA-Assembler (17).

The validation dataset (GEO series)

The expression profiles as well as clinical data of GSE46691, GSE70768 and GSE70769 were downloaded through R package of GEOquery. For each dataset, the expression profiles were annotated from probesets to genes and median centred across all samples. We filled target gene expression with zero when missing value occurred. For clinical data, we concentrated on Gleason score, T staging and prognostic information.

Generation of the prostate cancer subtype (PCS) classifier and identification of subtypes

We built up a PCS classifier to identify three subtypes based on RNA-Seq expression profiles of TCGA with algorithm of Hierarchical clustering and nearest shrunken centroids.

Consensus clustering

Consensus clustering is one way of assessing the clustering stability. With the R package of "ConsensusClusterPlus" (18), we carried out hierarchical clustering with agglomerative average linkage and classified these patients into 3–12 clusters via the 10426 most variable genes. The most variable genes were defined according to the criterion of median absolute deviation >0.5. We had median centred all expression arrays before all the computations and set 1,000 iterations, 0.98 subsampling ratio for consensus clustering (19).

Computing gap statistic

Gap statistic is a standard method to determine the optimal number of clusters in a dataset via comparing the change of the observed and expected within-cluster dispersion (20). To identify the ideal number of clusters, gap statistic was computed from k=1 to 6 for selected top variable genes by the R package of "cluster" (21).

Selection of patients and genes for machine learning

For patient samples, we computed Silhouette width to recognize the most representative patients in each cluster (22). Patients with a positive silhouette width were selected for the following classifier. Similarly, two filtering steps were applied to select the most representative and predictive genes. Firstly, SAM (Significance Analysis of Microarrays) was used to identify significantly differentially expressed genes (FDR <0.01, False Discovery Rate) between one subtype and others with R packages of "siggenes" (23). Secondly, AUC (area under receiver operating characteristic curve curve) values were calculated to estimate one gene's predictive ability to divide one subtype from others with R packages of "ROCR" (24). Only patients with FDR <0.01 and AUC >0.9 were kept to build the PCS classifier.

Building the PAM classifier and identifying subtypes

With the filtered patients and selected genes, a robust classier was established by the R packages of "PAMR" (prediction analysis for microarrays R package) based on the algorithm of nearest shrunken centroids (25). We set up a 10-fold cross-validation for 1,000 iterations to choose the optimal threshold of centroid shrinkage. Finally, we selected the classifier providing error rate <2% with the least number of genes. As a result, with the built PAM classifier, we classified all the TCGA prostate adenocarcinoma patients into three subgroups for the subsequent analyses. Firstly, we fit the clinical data to the three subtypes of prostate adenocarcinoma to check if there would be a difference in Gleason score, T staging, or prognosis. Secondly, in order to explore the molecular heterogeneity, we selected some popular biomarkers or mutations to check their expression variation in the different subtypes.

Annotation of genes in the classifier

To understand the biological significance of PCS genes, we performed annotation for genes by the following methods.

Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis

Gene ontology annotation associated with biological pathway and KEGG pathway enrichment analysis of the classifiers was achieved using the Database for Annotation

Visualization and Integrated Discovery (DAVID, https://david.ncifcrf.gov/) online tool. A P value <0.001 and gene counts >2 were set as threshold for biological pathway analysis, and P value <0.05 and gene counts >2 for KEGG pathway analysis.

Protein-protein interaction (PPI) network construction

We mapped the classifiers into the Search Tool for the Retrieval of interaction Genes database (String, http://string.embl.de/) and constructed a PPI network, which provided the information to predict the protein interactions. The minimum required interaction score was set to 0.400. We analyzed the PPI network with Cytoscape software (http://www.cytoscape.org/).

Validation on three GEO dataset

GSE70769, GSE70768 and GSE46691 from GEO dataset were selected as validation sets. Based on the PCS classifier generated in section 2.2.4, we classified each validation dataset and analyzed the relation between the given subgroups and clinical data. Some clinical information of all series was not identical. For example, the GSE70768 gave the primary and secondary Gleason score, while the GSE46691 just gave total Gleason score. Concerning the issues above, we utilized the majority of clinical data and presented the details in the results part.

Expression of the luminal and basal markers in the PCS subtypes.

To determine whether the PCS subtypes correspond to luminal or basal tumors, we analyzed the mean expression of genes known to be markers of luminal (EZH2, AR, MKI67, NKX3-1, KLK2/3 and ERG) or basal (ACTA2, GSTP1, IL6, KRT5 and TP63) prostate cancers among PCS subtypes (26). We also performed the gene set enrichment analysis (GSEA) to provide some biological insights of the PCS clusters (27).

Statistical analysis

Sample clustering and classifying were performed with the corresponding R packages mentioned above under R software (version of 3.3). Clinical data were treated as discrete variable (time to event) and categorical variable (Gleason score, pathology T stage (pT stage), metastasis, and prognostic end-point). We utilized Chi-square test and fisher's exact test when detecting the relation between categorical variables. Kaplan-Meier curves were used to describe the time-event data and log-rank method was used to test the differences. It was regarded as significant statistically when P value was <0.05. We applied the statistical packages SPSS v20 (IBM) to manage the clinical data. R software and Microsoft PowerPoint (v2016) were used for visualizing the results.

Results

Generation of the PCS classifier and subtypes identification

A total of 10,426 genes with most variability across samples were retained and median centred. Next, we performed hierarchical clustering with agglomerative average linkage to cluster these samples. We employed consensus clustering to assess the clustering stability. A significant increase in clustering stability was observed from k=2 to 3, but not for k>3 (*Figure 2A*). To further confirm it, we computed Gap statistic for k=1 to 5. A peak was mainly at k=3 or 4, indicating that three or four subtypes were ideal to explain the inner construction of dataset. To simplify the explanation for the result, we chose three subtypes for our following analysis (*Figure 2A*). We computed Silhouette width to identify the most representative samples within each cluster. To build the PCS classifier, we retained samples with positive silhouette width (n=412) (*Figure 2A*). We also applied two filtering methods (SAM and AUC) to select the most representative and predictive genes. The retained 256 genes were trained by PAM to build a classifier. To select the optimal threshold for centroid shrinkage, we performed 10-fold cross-validation and selected the one yielding a good performance (error rate <2%) with the least number of genes. Using this strategy, we built a classifier of 183 unique genes (*Figure 2B*) (*Table S3*—The gene list of PCS classifier). These genes were annotated with biologic process, biologic pathway, protein-protein interaction databases. Then we used it to classify all the prostate adenocarcinoma samples. 250 samples were classified for the first subtype (PCS1, prostate cancer cluster 1), 153 were PCS2, and the rest were PCS3 (*Figure 2C*). Some candidate genes were found to be validated in this dataset. They were KLF5, BCL2 and etc., and *Figure 2B* gave the details.

Function annotation and PPI analysis of our classifier

As partially displayed in *Figure 2D*, the classifier was mainly

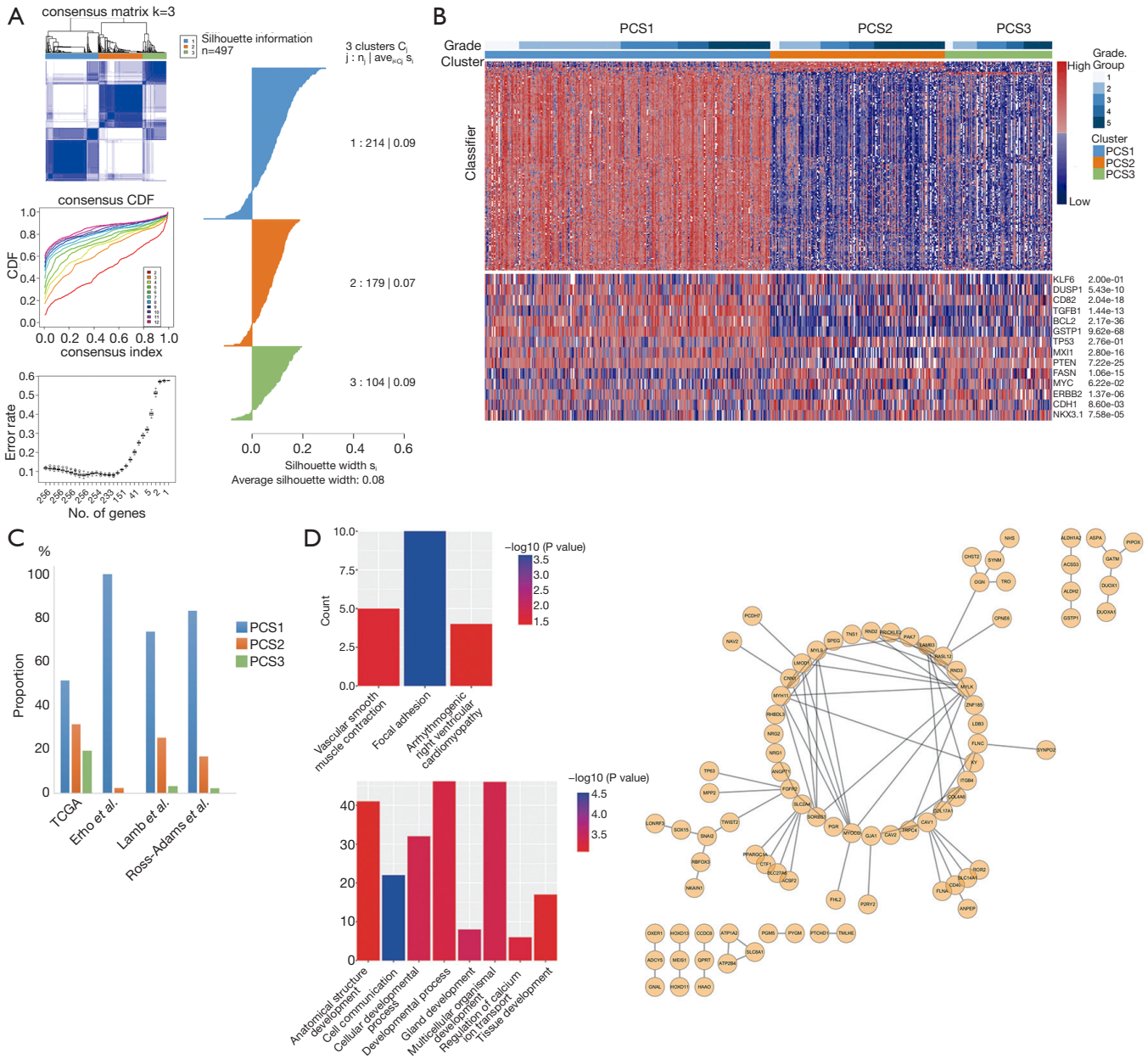

Figure 2 Unsupervised classification identified three genetic distinct subgroups with 183 genes. (A) The consensus clustering method displayed the optimal number of classification. The color scale represented the frequency that two samples belong to the same cluster (top left). Empirical CDF of consensus clustering improved sharply from two to three clusters (middle left). Samples with positive Silhouette values were selected as core samples to build the classifier (right). The PAM method indicated 183 was the optimal number for classification; (B) the TCGA dataset was classified in three subgroups according to the classifier. The top-right bar indicates the subgroups; light blue; PCS1, orange; PCS2, green; PCS3. In the heatmap, rows indicated genes from the classifier and columns indicated patients. The expression was color-coded and transformed with median centred log2 (red, high expression; blue, low expression). The heatmap below was selected with having been validated by previous studies; (C) all the four datasets were classified with the 183 genes. We displayed the proportion of each group in the respective dataset; (D) the 183 genes classifier was functionally annotated with GO and KEGG (the left two histogram). The height of bars indicated gene counts. Color represented P value. Protein-protein interaction graph was displayed on the right. Nodes standard for proteins and lines meant the possible relationships. CDF, cumulative distribution function; PAM, prediction analysis for microarrays; TCGA, the Cancer Genome Atlas; PCS1, prostate cancer subtype 1; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

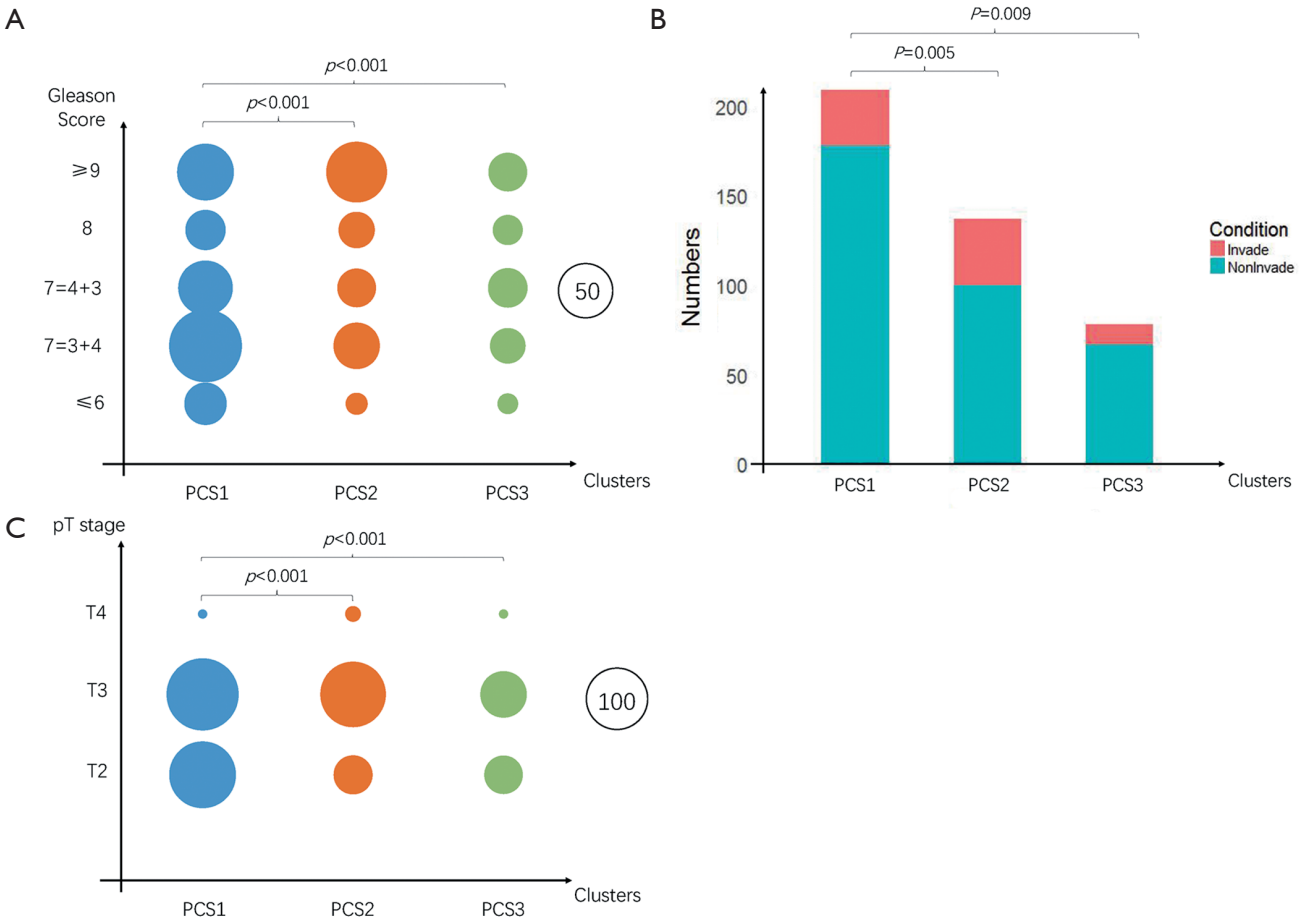

Figure 3 Three kinds of clinical data were compared among subgroups. (A) The numbers of patients in three subgroups were significantly different among five Gleason levels. The area of the bubble represented the number of patients. Three subgroups shared three colors. The vertical axis indicated five Gleason grade groups; (B) the proportions of lymph-node invasion patients in the three subgroups were still different. The height of split bars represented the population of each cluster or each group. Two colors marked two different conditions; (C) in terms of pathology T stage, the differences were still significant. Similarly, the area of the bubble represented the number of patients. Different T stages were on the vertical axis. PCS1, prostate cancer subtype 1.

annotated with 48 biologic process and biologic pathway terms, most of which were concerning cell communication, development process, morphogenesis and ion transport. KEGG pathway analysis revealed that these classifiers mostly enriched in focal adhesion.

The PPI network included 86 nodes and 112 edges. Some genes validated by other research were in the centre of the network, such as SNAI2, FGFR2, SPEG, CAV1 and so on, which might serve as intermediate regulation. Other genes validated by former research included FHL2, CD40, ANPEP, GSTP1, FLNA (28-33), which were not so highly interconnected with other genes.

Analysis with clinical data

After statistical test with the TCGA clinical data, we found PCS2 with an obviously higher Gleason score when comparing to PCS1 (*Figure 3A*). PCS3 had the least number of patients, and seemed to be the buffer zone for PCS1 and PCS2. About 1/3 PCS1 patients were in the WHO grade group 2 (Gleason score at 7=3+4). As for the PCS2, more than 1/3 patients were in the group 5 (Gleason score at 10).

The pT stage and lymph node invasion had the same tendency (*Figure 3B,C*). PCS2 had higher levels of pT stage and N stage comparing to PCS1. However, the recurrent

times had no significant difference among these subtypes, just like what the TCGA research network reported (12). The insufficient follow-up time (median follow-up time was 15.1 months) and low follow-up rate (less than 66%) might account for that.

Validation in three other datasets

The interesting results of above had been validated only in TCGA dataset. To push them to universal, we found three more datasets for further validation. Some information about these series was summarized in *Table S2*.

GSE46691

We applied the classifier to this dataset. Only two subtypes (PCS1 and PCS2) were found, but none for PCS3. Most of them were PCS1(532/545 *vs.* 13/545) (*Figure 2C*).

Different from TCGA dataset, this series provided too little information for us to use the WHO grading system. Thus, we separated all the patients to two levels with a cutoff of 8 (≤8 and >8). After comparison, Gleason scores and metastasis status had extremely similar patterns in these two different subtypes. The different results between PCS1 and PCS2 were shown in *Figure 4A*.

GSE70769

With the similar classification, there were only two patients in PCS3. Thus, we only analyzed PCS1 [74] and PCS2 [15] for the following items. The significant difference was observed between distinct levels of Gleason grading system (P=0.023) (*Figure 4B*). Unlike TCGA dataset, pT stages were not significantly different between PCS1 and PCS2 (*Figure 4B*). As for metastasis status, positive results could be observed (P=0.001), despite the paucity of metastatic cases (*Figure 4B*). Unfortunately, we observed no positive results concerning the recurrences between PCS1 and PCS2 in the survival analysis (P=0.059) (*Figure 4B*). However, the tendency of survival curve and critical P value implied a possibility of difference.

GSE70768

After classification, 90 cases were tagged as PCS1, while 31 were in PCS2 and four were in PCS3. Following GSE70769, we ignored PCS3 for further analysis. Besides the Gleason scores, the recurrence time was also significantly different between PCS1 and PCS2 in the survival analysis (P=0.005) (*Figure 4C*).

Expression of the luminal and basal markers in the PCS subtypes

There were an association between luminal genes and "PCS2 and PCS3", and basal genes and PCS1 (P<0.05). We also verified this observation in other 3 data sets. It provides evidences that PCS1 tumors correspond to basal subtypes, PCS2 and PCS3 reflect luminal subtypes. *Table S4* shows the gene sets enriched in each of PCS subtypes by GSEA. There are 17 gene sets significant enriched at FDR <25% and nominal P<5% in PCS1 subtypes, including epithelial to mesenchymal transition, myogenesis, inflammatory response and so on. PCS2 and PCS3 both enriched in HALLMARK_MYC_TARGETS_V2 (a subgroup of genes regulated by MYC).

Discussion

In order to explore the prostate adenocarcinoma, we first studied RNA-Seq data of prostate tumor samples from TCGA network (n=497). The most robust and optimal number of classification was three identified by an unsupervised consensus-based clustering algorithm. We built up the PCS classifier (183 genes) based on the expression of encoded genes from RNA-Seq level, rather than those specialized genes filtered or corrected with clinical data. Those samples with same gene-expression pattern would be more likely to be identified and classified into the same category. Then the classifier was validated in three independent datasets, which were completely different in data platform, even in data type (two were mRNA microarray data, one was extron microarray data). From the results, we can see that the variation of genes would indicate diverse clinical outcome.

According to previous studies, tumorigenesis is closely related to internal or genetic disorders. These disorders sometimes can be observed by different gene expression, such as tumor suppressor gene mutations with reduced expression and oncogene activation with increased expression. Tumors need a continuous process. The patterns of disorder vary among different tumors. In other words, there is only one way to be normal, but many ways to be abnormal. These compose the basis of tumor subgrouping and precision medicine (34). Our three distinct prostate adenocarcinoma subtypes could be identified each associating with unique molecular features. For example, PCS1 was featured with highly expressed GSTP1

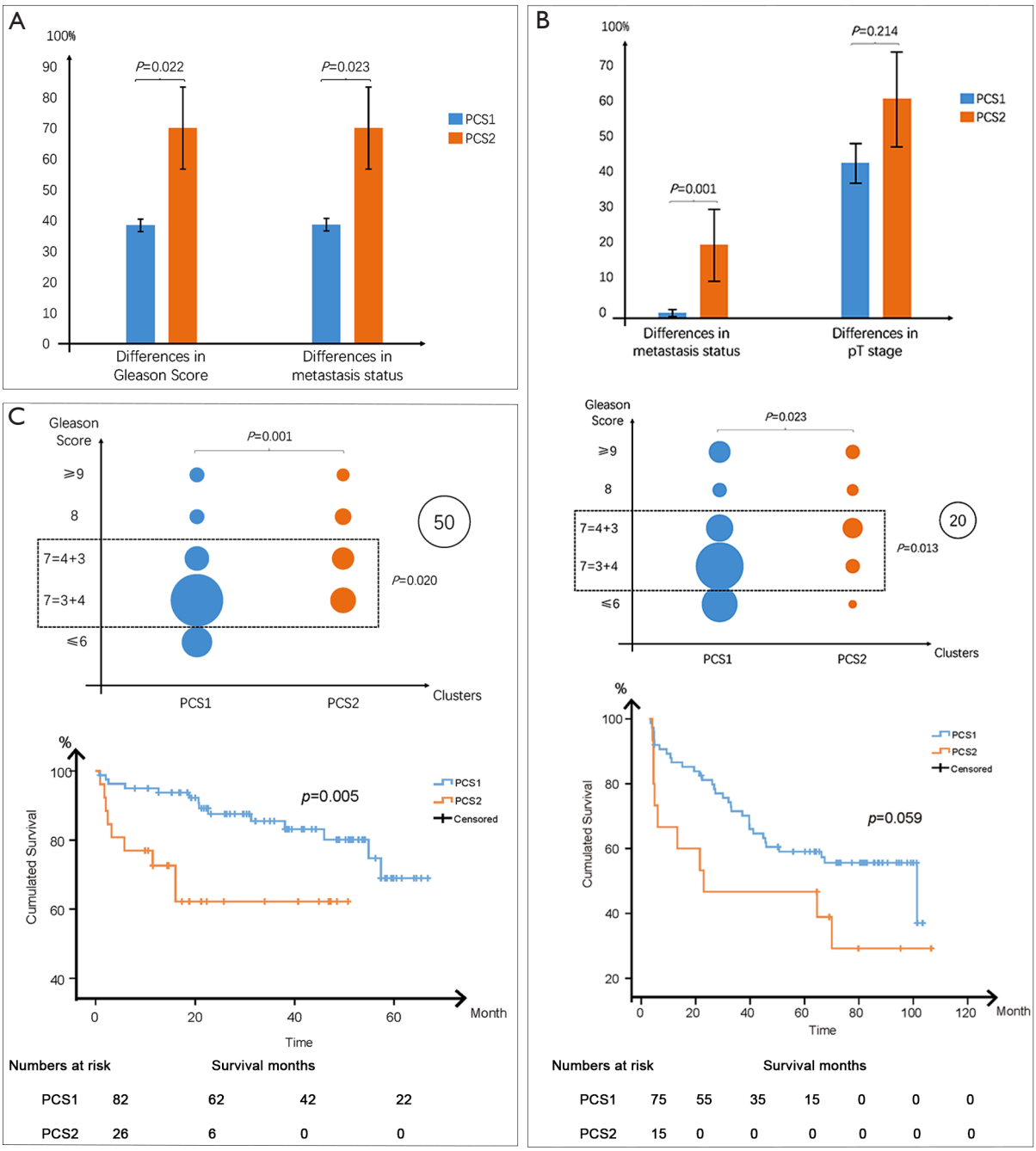

Figure 4 All three GEO datasets were validated positively. (A) GSE46991 was proved to have significances in Gleason Scores (P=0.022) and metastasis status (P=0.023). The columns indicated the proportions of patients who had >8 of Gleason score. The bars on top of columns were standard deviations; (B) for GSE70769, metastasis status had significances comparing PCS1 and PCS2 (P=0.001) (lower left), while the pT stages were not (P=0.214) (top). The number of patients in PCS1 and PCS2 was significant different among five Gleason levels (middle) (P=0.023). The area of the bubble represented the number of patients. The two levels of Gleason score in the dotted line box also indicated a significant difference (P=0.013). The recurrence time of two PCSs was displayed on the bottom; (C) the similar analyses were done in GSE70768 for Gleason grade (top) and recurrence time (bottom). GEO, gene expression omnibus; PCS1, prostate cancer subtype 1.

(P-class glutathione S-transferase gene 1) ($P=9.62\times10^{-68}$), which has been recognized as a tumor suppressor gene. Furthermore, there is a low hypermethylation of GSTP1 promoter in these PCS1 samples (P<0.01, *Figure S1*). Hypermethylation of the GSTP1 promoter is possibly the most common genetic event in prostate cancer and appears to be an early event in tumorigenesis (35). Usually, a low hypermethylation also means a high gene expression. Our observation corresponds to the fact that PCS1 patients have a relatively better prognosis. All the classifier genes (n=183) were functionally annotated with GO and KEGG databases. The most possible items were cell communication, development process and focal adhesion, most of which were closely related to tumor biology behavior.

We performed statistical test with matched TCGA clinical data among different groups, and found some interesting phenomena surprisingly. There were significant differences among these five Gleason grade groups (P<0.001). The same tendency was observed on pathology T stage and N stage, though we could not rule out the confounding effect here. We didn't find any significant difference in recurrent time among three subtypes. Just as what the prostate adenocarcinoma global analysis said (12), the follow-up time in the TCGA project is insufficient in terms of those slow progression tumors such as prostate cancer. Besides, the follow-up rate of prostate adenocarcinoma dataset (less than 66%) and median follow-up time (15.1 months) might be additional reasons that alter the accuracy and credibility of the survival analysis.

For further validation, the same workflow was applied to another three GEO series. It showed significant differences among clusters in Gleason scores (P<0.05). In addition, Gleason grade group 2 (7=3+4) and Gleason grade group 3 (7=4+3) had obvious differences in GSE70769. Previous studies indicated the different prognosis between these two groups (36,37). Therefore, the new WHO grading system regarded them as 2 groups. Our result could show this difference, which further illustrated the subtle relationship between tumor subtypes and prognosis.

GSE70768 and GSE70769, which include more complete follow-up, have follow-up time of 27.5 and 81.1 months respectively. For GSE70769, although the difference of recurrences between PCS1 and PCS2 was near a critical position of statistical (P=0.059), PCS2 still has a tendency of bad prognosis. GSE70768 consisted of 125 samples, and the recurrence times between PCS1 and PCS2 was significantly different (P=0.005). We all know that the Gleason score is based on microscopic appearance. The morphology change may be based on the inherent genetic change of tumor cells. Is the tumor subtypes we obtain through PCS classifier the basis of architectural patterns? The positive results above, must mean something, which we should run after for a long time.

There are several prostate cancer classifications over the past years (26,38,39). Tomlins described 4 subtypes based on microarray gene expression patterns that are related to several genomic aberrations; You and colleges used pathway activation signatures of known relevance to prostate cancer to developed a novel classification system. Different from above, our PCS classifier mainly built on the mining of the whole expression profile by unsupervised clustering. Whether our model corresponds to the former reported classification, especially the basal and luminal subtypes? We analyzed the mean expression of basal and luminal markers among PCS subtypes. The results provide evidences that PCS1 tumors correspond to basal subtypes, PCS2 and PCS3 reflect luminal subtypes (*Figure 5*). PCS1 (basal like) subtypes have a better prognosis, the same as what You *et al.* reported. Aberrant CpG methylation hypermethylation of GSTP1 promoter mentioned above may participated in these biological processes. GSEA may give us some biological insights about 3 PCS subtypes. As a result, epithelial-to-mesenchymal transition gene sets are upregulated in PCS1 and significant at FDR <25% (*Figure S2*). Epithelial to mesenchymal transition, has been recognized as a feature of aggressive tumors, and plays a crucial importance in cancer invasion and metastasis (40,41). PCS2 and PCS3 both enriched in a gene sets regulated by MYC (*Figures S3*,*S4*). The MYC oncogene encodes a nuclear protein that is involved in the control of normal cell growth, differentiation, and apoptosis. Overexpression correlate with advancing stage and grade of prostate cancer (42-44).

There were still some limitations in our study. Only RNA-Seq data were concerned in subgrouping the tumors. In addition, we have some doubts about the meaning of PCS3, for the limited number especially in the validation datasets. Genetically, it was nearer to PCS2 (*Figure 2B*), though it had ambiguous outcome. Those above may implied the complexity of genes.

Conclusions

In conclusion, we explored the TCGA prostate adenocarcinoma database by machine learning. We built up the PCS classifier (183 genes) based on gene expression profile, and got three prostate adenocarcinoma subtypes.

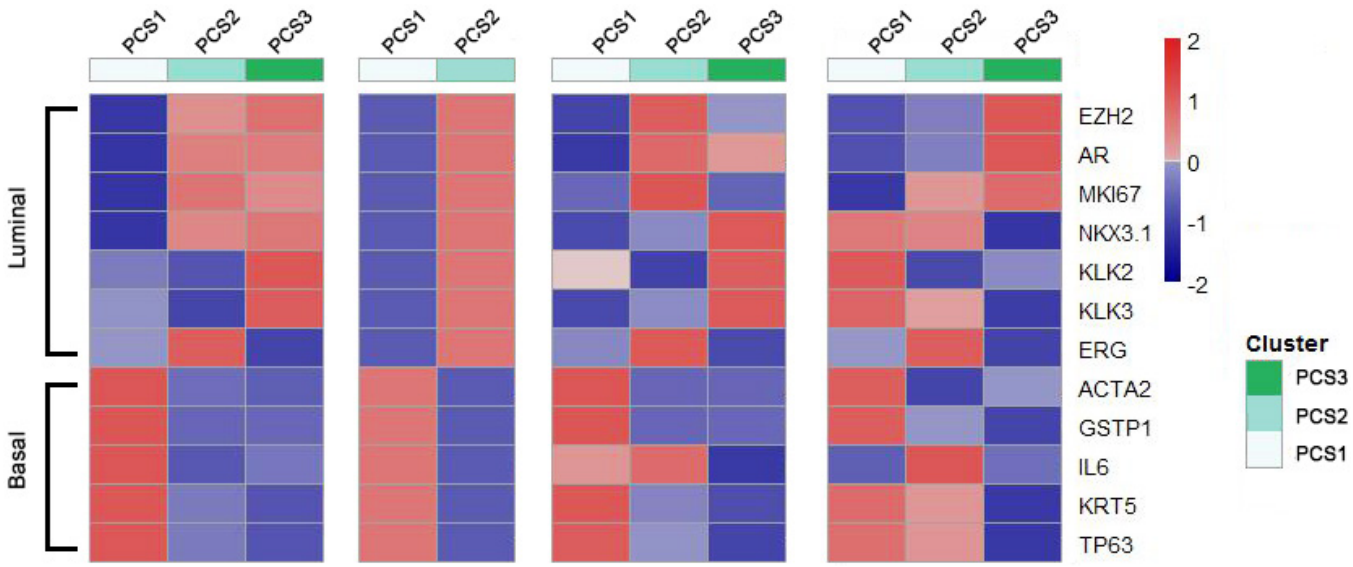

Figure 5 Expression of the basal and luminal markers in the three PCS subtypes. PCS, prostate cancer subtype. From left to right, it corresponds to the four datasets: TCGA, Erho *et al.*, Lamb *et al.*, Ross-Adams *et al.*

There were distinct differences among subgroups in Gleason grade group, pathological T, lymph node invasive and so on. These results were validated in other GEO datasets, and prognostic information also had significance in survival analysis, which might be helpful to further study and clinical use.

Acknowledgments

We would like to extend our gratitude to all staff working in TCGA network and the donors of tissues. We appreciate Desousa's methods, which provided us with so much guidance.
Funding: None.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/tcr.2017.05.15). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This article does not contain any studies with human participants or animals performed by any of the authors. TCGA is a deidentified database. The access to the TCGA database is approved by the National Cancer Institute.

References

1. Ketchandji M, Kuo YF, Shahinian VB, et al. Cause of death in older men after the diagnosis of prostate cancer. J Am Geriatr Soc 2009;57:24-30.
2. Garnick MB. Prostate cancer: screening, diagnosis, and management. Ann Intern Med 1993;118:804-18.
3. Walsh PC. Cancer surveillance series: interpreting trends in prostate cancer--part I: evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates. J Urol 2000;163:364-5.
4. Epstein JI, Egevad L, Amin MB, et al. The 2014

International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. Am J Surg Pathol 2016;40:244-52.
5. Humphrey PA, Moch H, Cubilla AL, et al. The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part B: Prostate and Bladder Tumours. Eur Urol 2016;70:106-19.
6. van den Bergh RC, van der Kwast TH, de Jong J, et al. Validation of the novel International Society of Urological Pathology 2014 five-tier Gleason grade grouping: biochemical recurrence rates for 3+5 disease may be overestimated. BJU Int 2016;118:502-5.
7. Qian J, Hirasawa K, Bostwick DG, et al. Loss of p53 and c-myc Overrepresentation in Stage T2-3N1-3M0 Prostate Cancer are Potential Markers for Cancer Progression. Mod Pathol 2002;15:35-44.
8. Catz SD, Johnson JL. BCL-2 in prostate cancer: a minireview. Apoptosis 2003;8:29-37.
9. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113-20.
10. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061-8.
11. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature 2011;474:609-15.
12. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell 2015;163:1011-25.
13. De Sousa E Melo F, Wang X, Jansen M, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat Med 2013;19:614-8.
14. Wong KC, Li Y, Peng C, et al. Computational learning on specificity-determining residue-nucleotide interactions. Nucleic Acids Res 2015;43:10180-9.
15. Erho N, Crisan A, Vergara IA, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. Plos One 2013;8:e66855.
16. 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;2:1133-44.
17. Zhu Y, Qiu P, Ji Y. TCGA-Assembler: open-source software for retrieving and processing TCGA data. Nat Methods 2014;11:599-600.
18. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572-3.
19. Monti S, Tamayo P, Mesirov J, et al. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Mach Learn 2003;52:91-118.
20. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Statist Soci 2001;63:411-23.
21. Kaufman L, Rousseeuw PJ. Finding groups in data. an introduction to cluster analysis. Biometrics, 1990.
22. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53-65.
23. Schwender H, Ickstadt K. Empirical Bayes analysis of single nucleotide polymorphisms. BMC Bioinformatics 2008;9:144.
24. Sing T, Sander O, Beerenwinkel N, et al. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21:3940-1.
25. Tibshirani R, Hastie T, Narasimhan B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99:6567-72.
26. You S, Knudsen BS, Erho N, et al. Integrated Classification of Prostate Cancer Reveals a Novel Luminal Subtype with Poor Outcome. Cancer Res 2016;76:4948-58.
27. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.
28. Lee WH, Isaacs WB, Bova GS, et al. CG island methylation changes near the GSTP1 gene in prostatic carcinoma cells detected using the polymerase chain reaction: a new prostate cancer biomarker. Cancer Epidemiol Biomarkers Prev 1997;6:443-50.
29. Brooks JD, Weinstein M, Lin X, et al. CG island methylation changes near the GSTP1 gene in prostatic intraepithelial neoplasia. Cancer Epidemiol Biomarkers Prev 1998;7:531-6.
30. Palmer DH, Hussain SA, Ganesan R, et al. CD40 expression in prostate cancer: a potential diagnostic and therapeutic molecule. Oncol Rep 2004;12:679-82.
31. McGrath MJ, Binge LC, Sriratana A, et al. Regulation of the transcriptional coactivator FHL2 licenses activation of the androgen receptor in castrate-resistant prostate cancer.

Cancer Res 2013;73:5066-79.
32. Sørensen KD, Abildgaard MO, Haldrup C, et al. Prognostic significance of aberrantly silenced ANPEP expression in prostate cancer. Br J Cancer 2013;108:420-8.
33. Sun GG, Lu YF, Zhang J, et al. Filamin A regulates MMP-9 expression and suppresses prostate cancer cell migration and invasion. Tumour Biol 2014;35:3819-26.
34. Fisher KW, Montironi R, Lopez Beltran A, et al. Molecular foundations for personalized therapy in prostate cancer. Curr Drug Targets 2015;16:103-14.
35. Lee WH, Morton RA, Epstein JI, et al. Cytidine methylation of regulatory sequences near the pi-class glutathione S-transferase gene accompanies human prostatic carcinogenesis. Proc Natl Acad Sci U S A 1994;91:11733-7.
36. Jo JK, Hong SK, Byun SS, et al. Comparison of clinical outcomes between upgraded pathologic Gleason score 3 + 4 and non-upgraded 3 + 4 prostate cancer among patients who are candidates for active surveillance. World J Urol 2015;33:1729-34.
37. Koontz BF, Tsivian M, Mouraviev V, et al. Impact of primary Gleason grade on risk stratification for Gleason score 7 prostate cancers. Int J Radiat Oncol Biol Phys 2012;82:200-3.
38. Smith BA, Sokolov A, Uzunangelov V, et al. A basal stem cell signature identifies aggressive prostate cancer phenotypes. Proc Natl Acad Sci U S A 2015;112:E6544-52.
39. Tomlins SA, Alshalalfa M, Davicioni E, et al. Characterization of 1577 primary prostate cancers reveals novel biological and clinicopathologic insights into molecular subtypes. Eur Urol 2015;68:555-67.
40. Thiery JP, Acloque H, Huang RYJ, et al. Epithelial-mesenchymal transitions in development and disease. Cell 2009;139:871-90.
41. Gravdal K, Halvorsen OJ, Haukaas SA, et al. A switch from E-cadherin to N-cadherin expression indicates epithelial to mesenchymal transition and is of strong and independent importance for the progress of prostate cancer. Clin Cancer Res 2007;13:7003-11.
42. Buttyan R, Sawczuk IS, Benson MC, et al. Enhanced expression of the c-myc protooncogene in high-grade human prostate cancers. Prostate 1987;11:327-37.
43. Jenkins RB, Qian J, Lieber MM, et al. Detection of c-myc oncogene amplification and chromosomal anomalies in metastatic prostatic carcinoma by fluorescence in situ hybridization. Cancer Res 1997;57:524-31.
44. Qian J, Hirasawa K, Bostwick DG, et al. Loss of p53 and c-myc overrepresentation in stage T(2-3)N(1-3) M(0) prostate cancer are potential markers for cancer progression. Mod Pathol 2002;15:35-44.

Cite this article as: Gao S, Qiu Z, Song Y, Mo C, Tan W, Chen Q, Liu D, Chen M, Zhou H. Unsupervised clustering reveals new prostate cancer subtypes. Transl Cancer Res 2017;6(3):561-572. doi: 10.21037/tcr.2017.05.15

Supplementary

Table S1 The new WHO grade group system of prostate cancer

Grade group	Gleason score and pattern
1	Grade 6 (3+3)
2	Grade 7 (3+4)
3	Grade 7 (4+3)
4	Grade 8 (4+4, 3+5, or 5+3)
5	Grade 9 or 10 (4+5, 5+4, or 5+5)

Table S2 Basic information about TCGA and three validation series

Series	No. of samples	Platform
TCGA (12)	497	Illumina HiSeq 2000 RNA Sequencing Version 2 analysis
GSE46691 (15)	545	Affymetrix Human Exon 1.0 ST Array [probe set (exon) version
GSE70768 (16)	125	Illumina HumanHT-12 V4.0 expression beadchip
GSE70769 (16)	91	Illumina HumanHT-12 V4.0 expression beadchip

Table S3 The gene list of PCS classifier

Gene symbol	Entrez ID
ADCY5	111
ALDH2	217
ALOX15	246
ANGPT1	284
ANPEP	290
AOX1	316
RND3	390
ASPA	443
ATP1A2	477
ATP2B4	493
CAV1	857
CAV2	858
CD40	958
CDH7	1005
CNN1	1264
COL4A6	1288
COL17A1	1308
CTF1	1489
CYP27A1	1593
EYA4	2070
FGFR2	2263
FHL2	2274
FLNA	2316
FLNC	2318
GAS1	2619
GATM	2628
GJA1	2697
GNAL	2774
GSTP1	2950
HFE	3077
NRG1	3084
HOXD10	3236
HOXD11	3237
HOXD13	3239
ITGB4	3691
KCNMB1	3779
LAMB3	3914
VWA5A	4013
MCC	4163
MEIS1	4211
MEIS2	4212
MPP2	4355
MYH11	4629
MYLK	4638
NBL1	4681
NHS	4810
ROR2	4920
OGN	4969
P2RY2	5029
PCDH7	5099
PGM5	5239
PGR	5241
PIK3C2G	5288
PYGM	5837
SCN2B	6327
SLC2A4	6517
SLC8A1	6546
SLC14A1	6563
SNAI2	6591
SNAP25	6616
SOX15	6665
SRD5A2	6716
TNS1	7145
TRO	7216
TRPC4	7223
ZNF154	7710
ZNF185	7739
RND2	8153
TP63	8626
B3GALT2	8707
KSR1	8844
ALDH1A2	8854
SLC16A5	9121
CPNE6	9362
NTN1	9423
CHST2	9435
NRG2	9542
ZNF516	9658
GPRASP1	9737
TOX	9760
HS3ST3A1	9955
SPEG	10290
MYL9	10398
WFDC2	10406
SORBS1	10580
PDPN	10630
KHDRBS3	10656
C5orf4	10826
PPARGC1A	10891
LDB3	11155
FAM107A	11170
CAND2	23066
NACAD	23148
SYNM	23336
QPRT	23475
HAAO	23498
TRIM29	23650
BHMT2	23743
LMOD1	25802
HSPB8	26353
MYOF	26509
APOBEC3C	27350
SLC27A6	28965
PIPOX	51268
RASL12	51285
SCARA3	51435
ASB2	51676
LRP1B	53353
DUOX1	53905
CLIC6	54102
A2BP1	54715
TMLHE	55217
SLC47A1	55244
DBNDD2	55861
KCNQ5	56479
SLC2A9	56606
PRDM8	56978
PAK7	57144
JPH2	57158
NYNRIN	57523
ARHGAP20	57569
ARHGAP23	57636
CACHD1	57685
TMEM35	59353
NECAB1	64168
KCTD14	65987
NKAIN1	79570
ACSS3	79611
ZNF750	79755
IQCA1	79781
LONRF3	79836
ANKRD53	79998
ACSF2	80221
MAGED4B	81557
TCF7L1	83439
CCDC8	83987
C2orf88	84281
JPH4	84502
ATCAY	85300
KIAA1644	85352
NAV2	89797
DUOXA1	90527
L3MBTL4	91133
CHRDL1	91851
DMKN	93099
MYOCD	93649
TMEM106A	113277
FBXO17	115290
FAM46B	115572
MRGPRF	116535
HRASLS5	117245
TWIST2	117581
FAT3	120114
FRMD6	122786
EVC2	132884
PTCHD1	139411
TCEAL2	140597
PRIMA1	145270
GCOM1	145781
HRNBP3	146713
ANKRD35	148741
GLIS1	148979
ITPRIPL1	150771
RHBDL3	162494
OXER1	165140
PRICKLE2	166336
SYNPO2	171024
FAM124A	220108
NKAPL	222698
MSRB3	253827
C20orf200	253868
LOC284276	284276
C3orf70	285382
RNF175	285533
KY	339855
ANKRD34B	340120
C18orf34	374864
B3GNT8	374907
C1orf175	374977
C1orf190	541468
LOC572558	572558
PGM5P2	595135
LOC644538	644538

Table S3 *(continued)*

Table S3 *(continued)*

Table S4 Gene set enrichment analysis (GSEA) results for the 'HALLMARKS' gene sets

Gene sets enriched in PCS subtype	SIZE	ES	NES	NOM p	FDR q
Gene sets enriched in PCS1 subtype					
Gene sets					
MYOGENESIS	187	0.59833	1.536722	0	0.016098
APICAL_SURFACE	41	0.6375206	1.5354674	0.00102145	0.008049
ALLOGRAFT_REJECTION	185	0.5654809	1.4592131	0	0.0161038
INFLAMMATORY_RESPONSE	191	0.556809	1.4327916	0	0.019519
IL6_JAK_STAT3_SIGNALING	78	0.5575616	1.4086231	0.00200803	0.0240011
EPITHELIAL_MESENCHYMAL_TRANSITION	198	0.5434884	1.4019129	0	0.0221452
TNFA_SIGNALING_VIA_NFKB	197	0.5392981	1.3915504	0	0.0219671
APICAL_JUNCTION	183	0.5268314	1.3486778	0	0.0387194
KRAS_SIGNALING_DN	181	0.5145873	1.3231847	0	0.0508606
INTERFERON_GAMMA_RESPONSE	182	0.5051111	1.3053231	0	0.0624133
COMPLEMENT	180	0.503827	1.2966136	0.004	0.0652992
KRAS_SIGNALING_UP	193	0.4935308	1.2788059	0.002	0.0759542
UV_RESPONSE_DN	136	0.4963331	1.2678461	0.009	0.0830991
HYPOXIA	190	0.4763444	1.22983	0.01	0.1335384
APOPTOSIS	142	0.4746551	1.2161257	0.023	0.1514477
ESTROGEN_RESPONSE_EARLY	193	0.4668625	1.207341	0.028	0.1594714
IL2_STAT5_SIGNALING	192	0.4640342	1.1963452	0.02	0.1754912
Gene sets enriched in PCS2 subtype					
Gene sets					
MITOTIC_SPINDLE	170	−0.380714			1
G2M_CHECKPOINT	152	−0.602883			1
MYC_TARGETS_V1	116	−0.628538	−3.292911	0	0
E2F_TARGETS	160	−0.649992	−3.196251	0	0
MYC_TARGETS_V2	45	−0.56	−2.113879	0	0
MTORC1_SIGNALING	161	−0.203905	−0.937359	1	0.8321512
UNFOLDED_PROTEIN_RESPONSE	85	−0.168492	−0.869125	1	0.931307
OXIDATIVE_PHOSPHORYLATION	138	−0.160268	−0.812112	1	0.8691489
DNA_REPAIR	104	−0.179691	−0.728583	1	0.8586289
PROTEIN_SECRETION	69	−0.110638	−0.529586	1	0.7965958
Gene sets enriched in PCS3 subtype					
Gene sets					
MITOTIC_SPINDLE	170	−0.347363			1
DNA_REPAIR	104	−0.290227			1
G2M_CHECKPOINT	152	−0.609389			1
ANDROGEN_RESPONSE	92	−0.347315			1
UNFOLDED_PROTEIN_RESPONSE	85	−0.237386			1
MTORC1_SIGNALING	161	−0.209816			1
E2F_TARGETS	160	−0.699309			1
MYC_TARGETS_V1	116	−0.613631			1
OXIDATIVE_PHOSPHORYLATION	138	−0.331058			1
MYC_TARGETS_V2	45	−0.596695	−2.444128	0	0

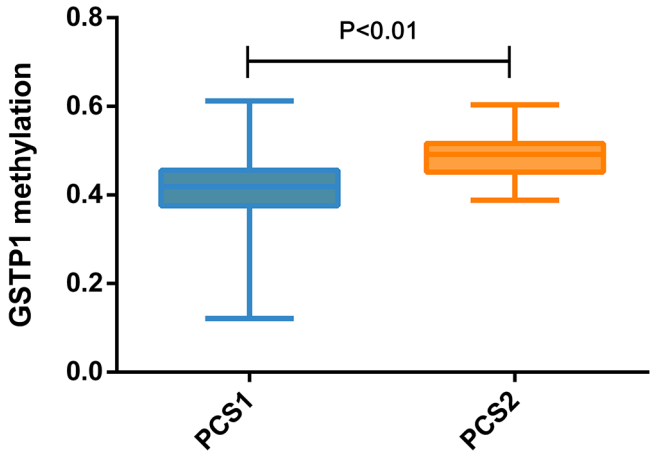

Figure S1 PCS1 patients have a lower hypermethylation of the GSTP1 promoter (P<0.01). PCS1, prostate cancer subtype 1.

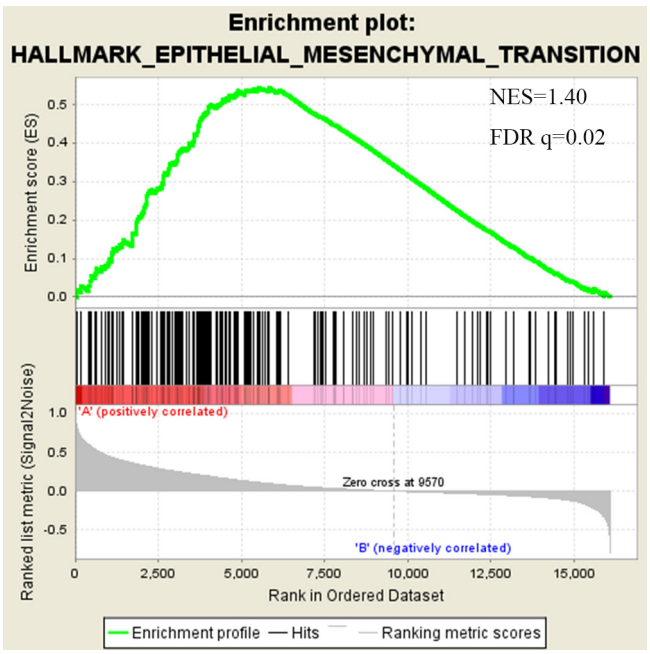

Figure S2 GSEA plots showing upregulation of Epithelial-Mesenchymal Transitions in PCS1. PCS1, prostate cancer subtype 1. GSEA, gene set enrichment analysis.

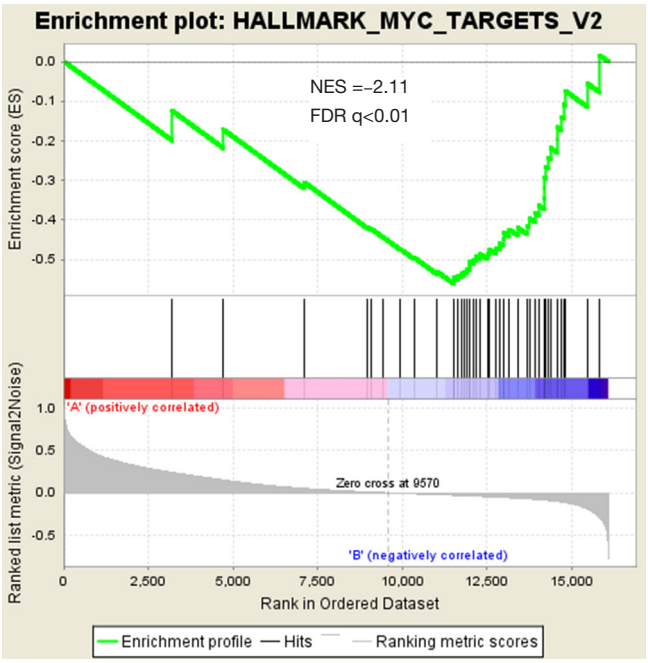

Figure S3 GSEA plots showing downregulation of MYC targets in PCS2. GSEA, gene set enrichment analysis.

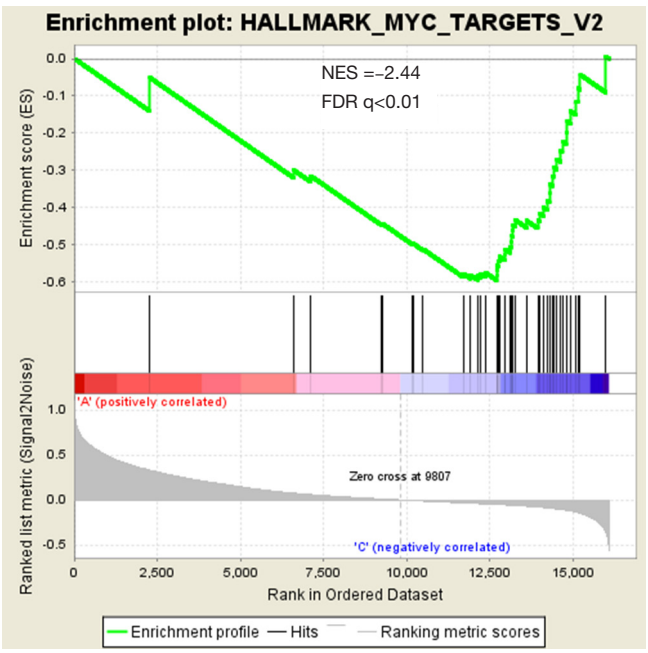

Figure S4 GSEA plots showing downregulation of MYC targets in PCS3. GSEA, gene set enrichment analysis.