RESEARCH ARTICLE

Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes

Shixiang Wang[1,2,3], **Huimin Li**[1,2,3], **Minfang Song**[1,2,3], **Ziyu Tao**[1,2,3], **Tao Wu**[1,2,3], **Zaoke He**[1,2,3], **Xiangyu Zhao**[1,2,3], **Kai Wu**[4], **Xue-Song Liu**[1]*

1 School of Life Science and Technology, ShanghaiTech University, Shanghai, China, **2** Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China, **3** University of Chinese Academy of Sciences, Beijing, China, **4** Department of Thoracic Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

* liuxs@shanghaitech.edu.cn

Abstract

Genome alteration signatures reflect recurring patterns caused by distinct endogenous or exogenous mutational events during the evolution of cancer. Signatures of single base substitution (SBS) have been extensively studied in different types of cancer. Copy number alterations are important drivers for the progression of multiple cancer. However, practical tools for studying the signatures of copy number alterations are still lacking. Here, a user-friendly open source bioinformatics tool “sigminer” has been constructed for copy number signature extraction, analysis and visualization. This tool has been applied in prostate cancer (PC), which is particularly driven by complex genome alterations. Five copy number signatures are identified from human PC genome with this tool. The underlying mutational processes for each copy number signature have been illustrated. Sample clustering based on copy number signature exposure reveals considerable heterogeneity of PC, and copy number signatures show improved PC clinical outcome association when compared with SBS signatures. This copy number signature analysis in PC provides distinct insight into the etiology of PC, and potential biomarkers for PC stratification and prognosis.

Author summary

Genomic DNA alteration signatures are recurring genomic patterns that are the imprints of mutagenic processes accumulated over the lifetime of cancer cell. Copy number alteration is a key driver for the progression of multiple cancer, including prostate cancer, which is particularly driven by complex genome alterations. However, practical tools for studying the signatures of copy number alterations are still lacking. Here a novel bioinformatics tool for copy number signature analysis has been constructed. With this newly developed bioinformatics tool “sigminer”, we performed the first copy number signature analysis in prostate cancer, and an unprecedentedly clear map connecting genome alteration driving factors and prostate cancer clinical outcomes have been illustrated. These

OPEN ACCESS

Citation: Wang S, Li H, Song M, Tao Z, Wu T, He Z, et al. (2021) Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. PLoS Genet 17(5): e1009557. https://doi.org/10.1371/journal.pgen.1009557

Editor: Dmitry A. Gordenin, National Institute of Environmental Health Sciences, UNITED STATES

Received: December 22, 2020

Accepted: April 19, 2021

Published: May 4, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1009557

Data Availability Statement: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files. All code required to reproduce the analysis outlined in this manuscript are freely available at https://github.com/XSLiuLab/PC_CNA_signature. Analyses can be read online at https://xsliulab.github.io/PC_CNA_signature.

Funding: This work was supported by The National Natural Science Foundation of China (http://www.nsfc.gov.cn): 31771373 [X.-S.L.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

analyses provide novel insight into the mutational processes and clinical outcomes of prostate cancer.

Introduction

Cancer are primarily caused by somatic alterations in the genomic DNA. Based on the size and feature of genome alterations, these cancer associated DNA alterations can be classified into the following four types: single base substitution (SBS), small insertion and deletion (INDEL), structural alteration including translocation/inversion, and copy number alteration (CNA). Somatic copy number alterations are extremely common in cancer, and have been reported as important drivers for the progression of multiple types of cancer [1,2].

Genomic DNA alteration signatures are recurring genomic patterns that are the imprints of mutagenic processes accumulated over the lifetime of cancer cell [3,4]. Genome alteration signature analysis can not only provide the mutational process information, but also biomarker for cancer precision medicine [5,6]. SBS signature analysis has been extensively studied, and represents a prototype for other types of signature study [3]. Despite the importance of copy number alteration in cancer progression, practical tools for copy number signature study are still lacking.

Prostate cancer (PC) is a common cancer type among men [7]. PC is particularly driven by copy number alterations, indolent and low-Gleason tumors have few alterations, whereas more aggressive primary and metastatic tumors have extensive copy number alterations [8,9]. In contrast, somatic point mutations are less common in PC than in most other solid tumors [10]. The most frequently mutated genes in primary PC are *SPOP*, *TP53*, *FOXA1*, and *PTEN* [11]. TMPRSS2-ETS fusion genes are frequently observed in prostate cancer patients [12]. Several studies reported that percent of copy number altered genome, termed "CNA burden" is associated with the recurrence and death of prostate tumors [9] and other tumor types [13]. Aneuploidy, defined as chromosome gains and losses, has been reported to drive lethal progression in prostate cancer [14]. All these results suggest a key role of CNA in PC progression and clinical outcome association. However, both CNA burden and aneuploidy only capture particular aspect of CNA, comprehensive understanding of CNA in PC remains obscure, and the underlying mechanisms and the specific mutational processes involved also remain unknown.

To address these challenges, we developed a novel method to investigate the signatures of copy number alteration, and built an open source R/CRAN package for the scientific community. With this tool, we performed the first copy number signature analysis in PC. The driving forces and mechanisms underlying each distinct copy number signature have been proposed, and the clinical outcome for each signature has been further investigated. This copy number signature analysis provides a new insight into the mutational processes of prostate tumors, and also novel biomarkers for PC stratification and prognosis.

Results

Copy number signature analysis framework

SBS mutations are primarily the consequences of single-strand DNA lesions and single-strand break repair mistakes. Copy number alterations reflect double-strand DNA lesions and double-strand break repair problems. SBS signatures have been well studied in many cancers, and summarized in the COSMIC (Catalogue of Somatic Mutations in Cancer) database [4].

However, copy number signatures are less studied. Macintyre et al. performed copy number signature analysis in high-grade serous ovarian cancer [15], and represented the currently only available study related to copy number signature. Macintyre et al. employed a mixture modeling based method for copy number component extraction, the biological meaning of the copy number component is unclear and not consistent among different cancer types or datasets, this limits the application of their method in cancer genome study [15]. A unified and extensible copy number component classification method and practical bioinformatical toolkit are required.

Here a unified copy number signature extraction method has been constructed (Fig 1). It incorporated the following 8 copy number features: the breakpoint count per 10 Mb (named "BP10MB"); the breakpoint count per chromosome arm (named "BPArm"); the absolute copy number of the segments (named "CN"); the difference in copy number values between adjacent segments (named copy number change point, or "CNCP"); the lengths of oscillating copy number segment chains (named "OsCN"); the log10 based copy number segment size (named "SS"); the minimal number of chromosomes with 50% copy number alterations (named "NC50"); the distribution of copy number alterations in each chromosome (burden of chromosome, named "BoChr") (Fig 1). These features were selected as hallmarks of previously reported genomic aberrations like chromothripsis, tandem duplication or to denote the genome distribution pattern of copy number alteration events [15–17].

We classified the distributions of the above mentioned 8 copy number features into 80 components (Fig 1 and S1 Table). Each copy number component has a clear biological meaning, for example: "CN[2]" indicate the absolute copy number of the DNA segment is 2. For each tumor, the value for each component of copy number feature was counted based on the segmented absolute copy number profile of that tumor. The absolute copy number profiles can be extracted from whole exome sequencing (WES), whole genome sequencing (WGS), or SNP array data [18,19]. Then a tumor by copy number component value matrix was generated by combining component values in all tumors. This matrix was subjected to non-negative matrix factorization (NMF), a method previously used for deriving SBS signatures [3].

Compared to Macintyre et al. method [15], our method is much more computationally efficient, it took about 1.5 minutes with our method rather than about 1.5 hours using Macintyre et al. method to generate the matrix for copy number signature extraction in this study. Besides, each of the predefined 80 copy number components has clear and fixed biological meaning, and this facilitates not only the intelligibility of signature result, but also the scalability of this copy number signature analysis method with known SBS signature analysis. The biological meaning of the CNA component in Macintyre et al. method is not consistent for different datasets. This is problematic especially if the investigator want to compare the signature generated with different dataset using cosine similarity analysis. To illustrate this difference, we compared our method with Macintyre et al. method in ovarian cancer with different sample size. The CNA component is consistent for three datasets with our method but not Macintyre et al. method, consequently the signatures in different dataset cannot be compared when using Macintyre et al method (S1 Fig). Due to this problem Macintyre et al. method cannot be applied in single sample CNA signature analysis. The major differences between our method and Macintyre et al. method are listed in S2 Fig.

A R/CRAN package (Sigminer, https://cran.r-project.org/package=sigminer) has been developed based on this novel copy number analysis method for bioinformatics community and cancer researchers to explore and analyze copy number signatures (Fig 1). To our knowledge, sigminer is the first practical bioinformatics tool for extracting the signatures of copy number alterations, supporting both our new method and the method from Macintyre et al. study [15].

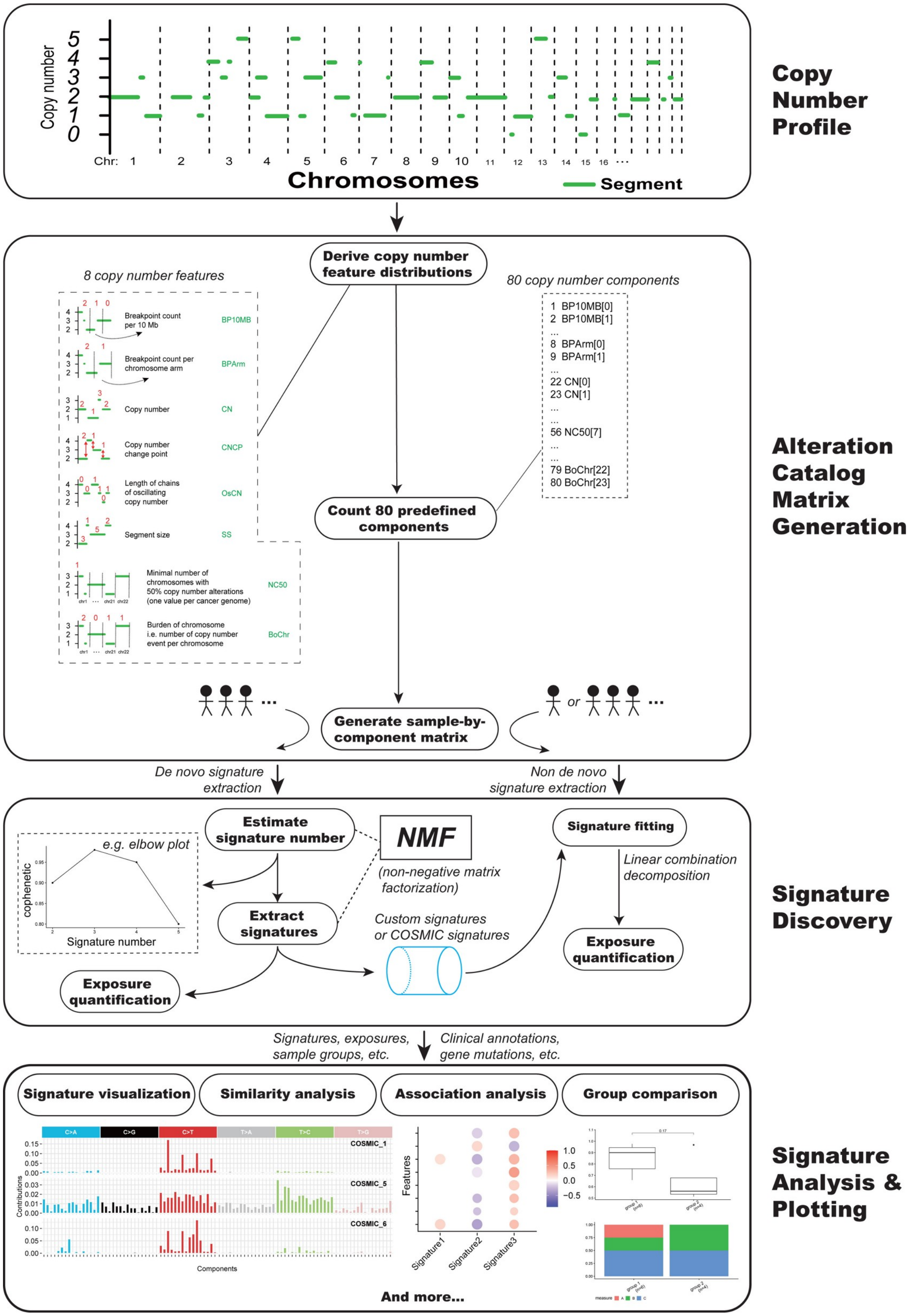
Copy Number Profile
Copy number
5
4
3
2
1
0
Chr: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 ···
Chromosomes
Segment
Alteration Catalog Matrix Generation
8 copy number features
Breakpoint count per 10 Mb
BP10MB
Breakpoint count per chromosome arm
BPArm
Copy number
CN
Copy number change point
CNCP
Length of chains of oscillating copy number
OsCN
Segment size
SS
Minimal number of chromosomes with 50% copy number alterations (one value per cancer genome)
NC50
Burden of chromosome i.e. number of copy number event per chromosome
BoChr
chr1 ··· chr21 chr22
Derive copy number feature distributions
80 copy number components
1 BP10MB[0]
2 BP10MB[1]
...
8 BPArm[0]
9 BPArm[1]
...
22 CN[0]
23 CN[1]
...
...
56 NC50[7]
...
...
79 BoChr[22]
80 BoChr[23]
Count 80 predefined components
Generate sample-by-component matrix
or
De novo signature extraction
Non de novo signature extraction
Signature Discovery
Estimate signature number
e.g. elbow plot
cophenetic
Signature number
NMF
(non-negative matrix factorization)
Extract signatures
Exposure quantification
Custom signatures or COSMIC signatures
Signature fitting
Linear combination decomposition
Exposure quantification
Signatures, exposures, sample groups, etc.
Clinical annotations, gene mutations, etc.
Signature Analysis & Plotting
Signature visualization
Similarity analysis
Association analysis
Group comparison
C>A
C>G
C>T
T>A
T>C
T>G
COSMIC_1
COSMIC_5
COSMIC_6
Contributions
Components
Features
Signature1
Signature2
Signature3
1.0
0.5
0.0
−0.5
measure A B C
And more...

Fig 1. Copy number signature analysis framework.

https://doi.org/10.1371/journal.pgen.1009557.g001

Genome alteration landscape of PC

We assembled and uniformly analyzed WES data from 1,003 pairs of prostate cancers and matched germline control (649 primary and 354 metastatic tumors) that passed quality control parameters from six independent studies (S3 Fig) [10,20–24]. Patient characteristics, including age at diagnosis, Gleason score, and metastatic site, are shown in S2 Table.

Small scale genome alterations include single base substitution (SBS) and small insertion and deletion (INDEL). Mutation significance analysis with these small scale genome alterations revealed 47 cancer driver genes (MutSig $q < 0.05$) (S4 Fig). Similar to previous study [11], the top frequently mutated genes in this PC genome dataset are *TP53* (19%), *SPOP* (9%), *FOXA1* (8%), *PTEN* (4%) (S4 Fig). About half of PC samples (56.44%, 552 of 978) have small scale cancer driving genome alterations. These small scale genome alterations were summarized based on different variant classifications (S5A–S5C Fig). The median number of small scale variants in PC is 28, and most of them are SBS, suggesting a relatively low SBS number in PC compared with other cancer types (S5B Fig).

Distribution pattern of CNA segment length and chromosome distribution of CNA counts in available PC samples are summarized and shown (S5D Fig). Majority of copy number alterations are focal, and 14.9% are chromosome arm level or whole chromosome level alterations (S5D Fig). Median values of CNA segment count, CNA amplification count, CNA deletion count and CNA burden are shown in S3 Table. The distribution of 8 CNA features selected in this study are shown (S5E Fig).

Genome alteration signatures identified in PC

With sigminer developed in this study, we identified five copy number signatures and three SBS signatures from 1003 tumor-normal pairs of PC WES data. The number of signature was determined after comprehensive consideration of result stability shown by cophenetic plot and biological interpretability (S6 Fig).

Three SBS mutational signatures named SBS-sig 1, SBS-sig 2, SBS-sig 3 are identified (S7A Fig). Etiology for the SBS mutational signatures has been well explored and stored in COSMIC database. Cosine similarity analysis has been performed between the three SBS signatures identified in this study and COSMIC signatures (S8 Fig). The SBS signatures identified in this study are similar to those identified by COSMIC based on TCGA PC datasets.

Five copy number alteration signatures are identified, namely CN-sig 1 to CN-sig 5 (Fig 2A). These copy number signatures are ranked based on the median length of CNA segment. CN-sig 1 has the smallest median size of CNA segment, and CB-sig 5 has the largest median size of CNA segment. Copy number components of the same feature are row normalized, and this facilitates copy number component value comparisons within the signature. Macintyre et al. performed column normalization for each copy number component, and the component values of same CNA feature cannot be compared within each signature [15]. Representative absolute copy number profile for each signature enriched PC patient is shown (Fig 2B).

Signature stability and single sample copy number signature analysis

In cancer precision medicine, it is important to confidently associate known signatures and their activities to patients. One of the key issues in signature analysis is the stability of signature extracted. Bootstrap analysis as previous reported [25] was performed to evaluate the stability of signatures extracted through *do novo* NMF (S7B Fig). Comparison analysis of signature exposure instability between all signatures has been applied. Every signature instability measured as the root mean squared error (RMSE) between its exposures in 1000 bootstrap mutation catalogs and its exposures in the original mutation catalog for each tumor are shown (Fig

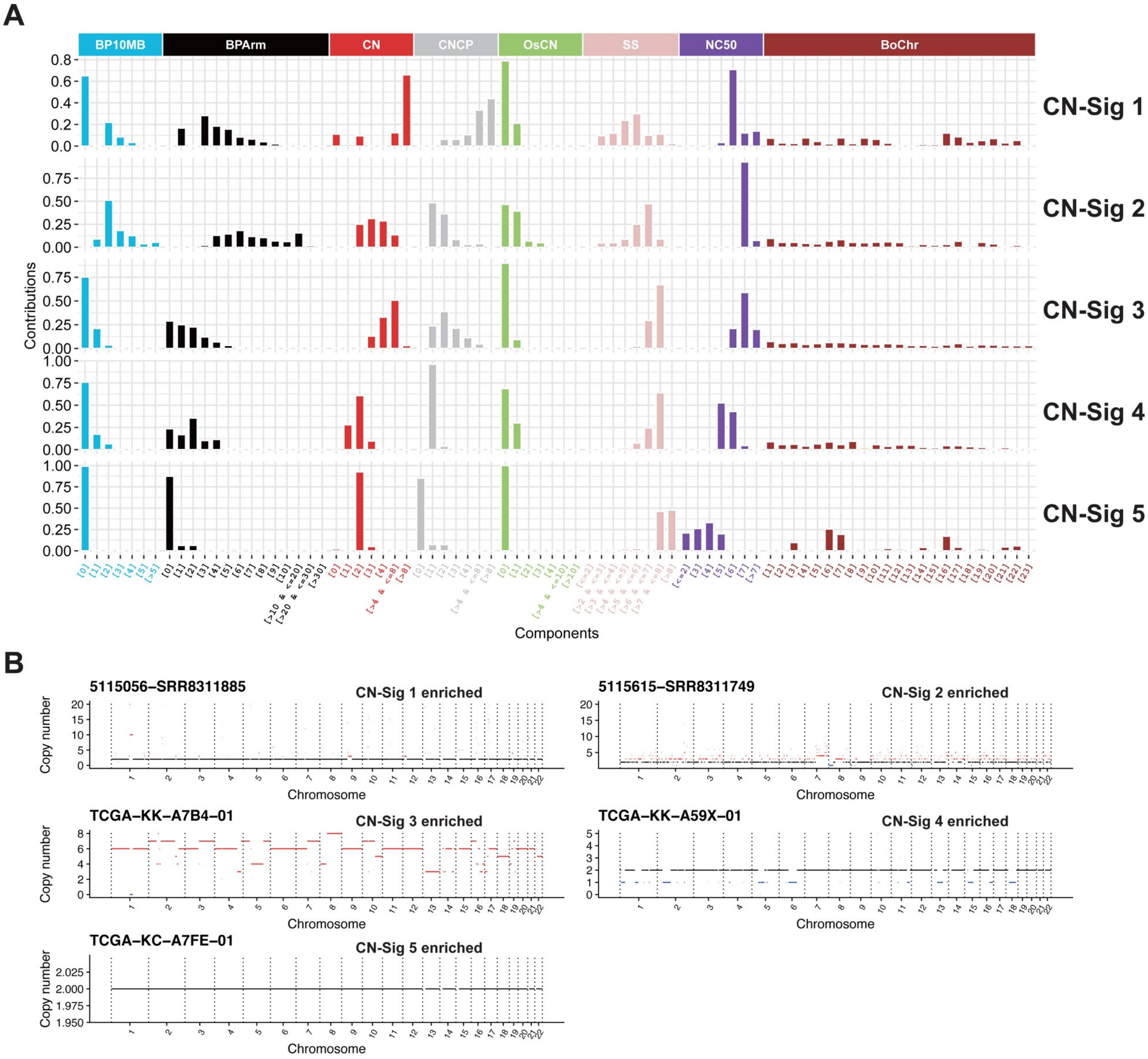

Fig 2. Copy number signatures identified in PC. (A) Five copy number signatures are identified from PC WES data. For copy number signatures, the signature profile was row normalized within each feature. Copy number signature has 8 different features with 80 components in together. (B) Representative absolute copy number profile for each copy number signature. Selected samples with enriched copy number signature CN-Sig 1, CN-Sig 2, CN-Sig 3, CN-Sig 4 and CN-Sig 5 are shown. The segments with copy number gain and loss are labeled by red and blue color, respectively.

https://doi.org/10.1371/journal.pgen.1009557.g002

3A). For the combined PC WES dataset, at least 57 tumors are required for extraction of all five copy number signatures (Fig 3B).

In addition to *de novo* copy number signature extraction with NMF, method for single sample copy number signature analysis has also been developed based on signature fitting and optimization with simulated annealing [25]. As an example, the five PC samples in Fig 2B are reanalyzed with single sample signature fitting method. Each tumor is clearly dominated by the same type of copy number signature as the signature derived from *de novo* NMF (Fig 3C). Bootstrap procedure was also adopted to evaluate the stability of signatures extracted through single sample copy number signature fitting in two different samples, and consistent results

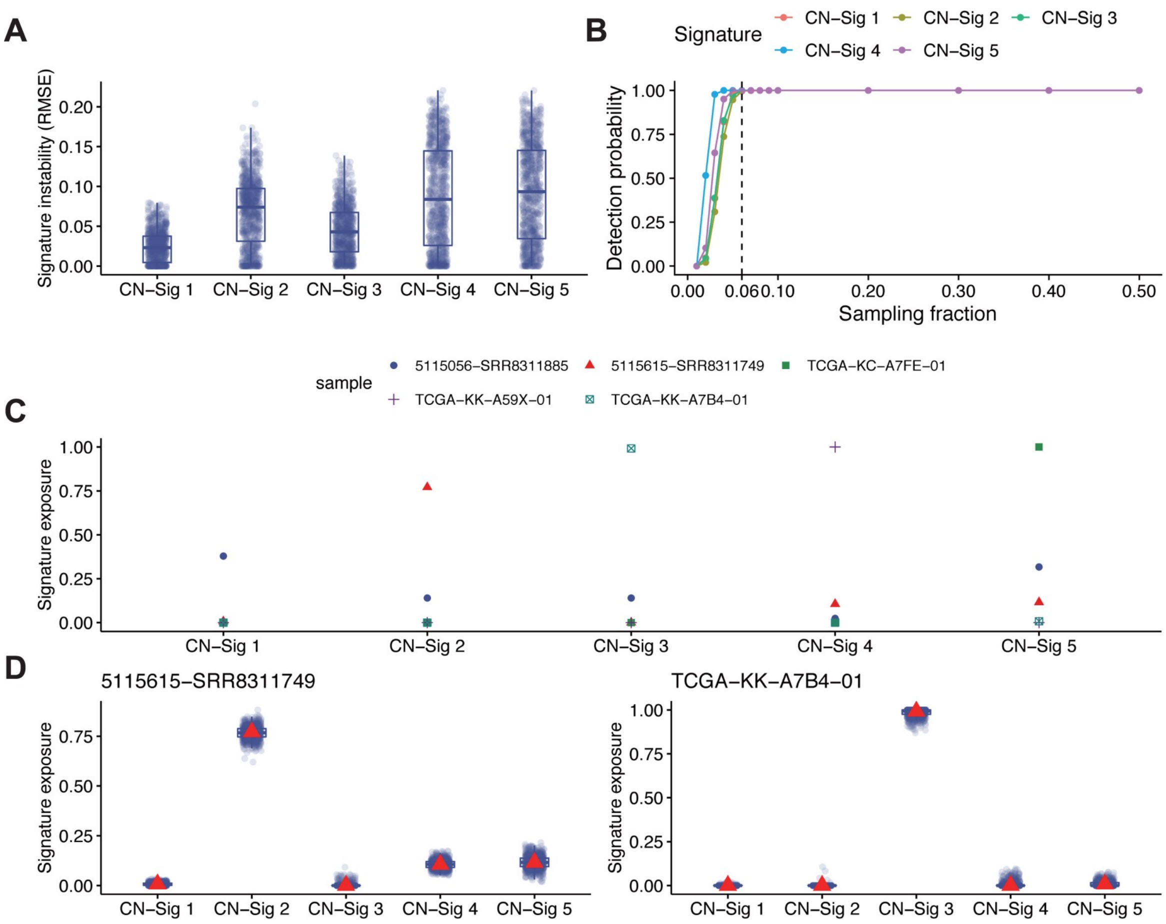

Fig 3. Signature instability analysis and single sample copy number signature analysis. (A) Comparison of signature exposure instability measured as RMSE (root mean squared error) between exposures in 1000 bootstrap mutation catalogs and exposures in the original mutation catalog for each tumor. (B) Signature detection probability versus different sampling fraction. The dotted line indicates we could steadily detect all five copy number signatures with at least 6% (~57) tumors. A signature is detected if more than 10 tumors have at least 1% relative exposure under p-value cutoff 0.05. The sampling process was repeated 1000 times for each sampling fraction to estimate the detection probability. (C) Relative signature exposures measured through single sample signature fitting for the five tumors in Fig 2B. (D) Single sample signature instability analysis for two of the five tumors in Fig 2B based on 1000 bootstrap signature exposures (indicated as blue dots and boxplot) and the original signature exposure (indicated as red triangle). The boxplot is bounded by the first and third quartile with a horizontal line at the median.

https://doi.org/10.1371/journal.pgen.1009557.g003

are obtained (Fig 3D). All these analyses suggest that the copy number signature extraction method developed in this study can provide reliable and consistent results.

Mutational processes underlying copy number signatures

Five copy number signatures and three SBS signatures have been identified in 1003 tumor-normal paired PC WES dataset. These signatures reflect distinct genomic DNA alteration patterns driving by distinct molecular events. To investigate the interconnection among genome alteration signatures, we computed the associations between the exposures of signatures and various types of genomic alteration features including tumor purity, MATH (A simple quantitative measure of intra-tumor heterogeneity [26]), tumor ploidy, CNA burden, INDEL number, tandem duplication phenotype score (TDP score, see methods) [17], chromothripsis state score (see methods) [16] (Fig 4A), genetic alteration in known PC driving genes (Fig 4B), and PC associated signaling pathways (Fig 4C and S4 Table).

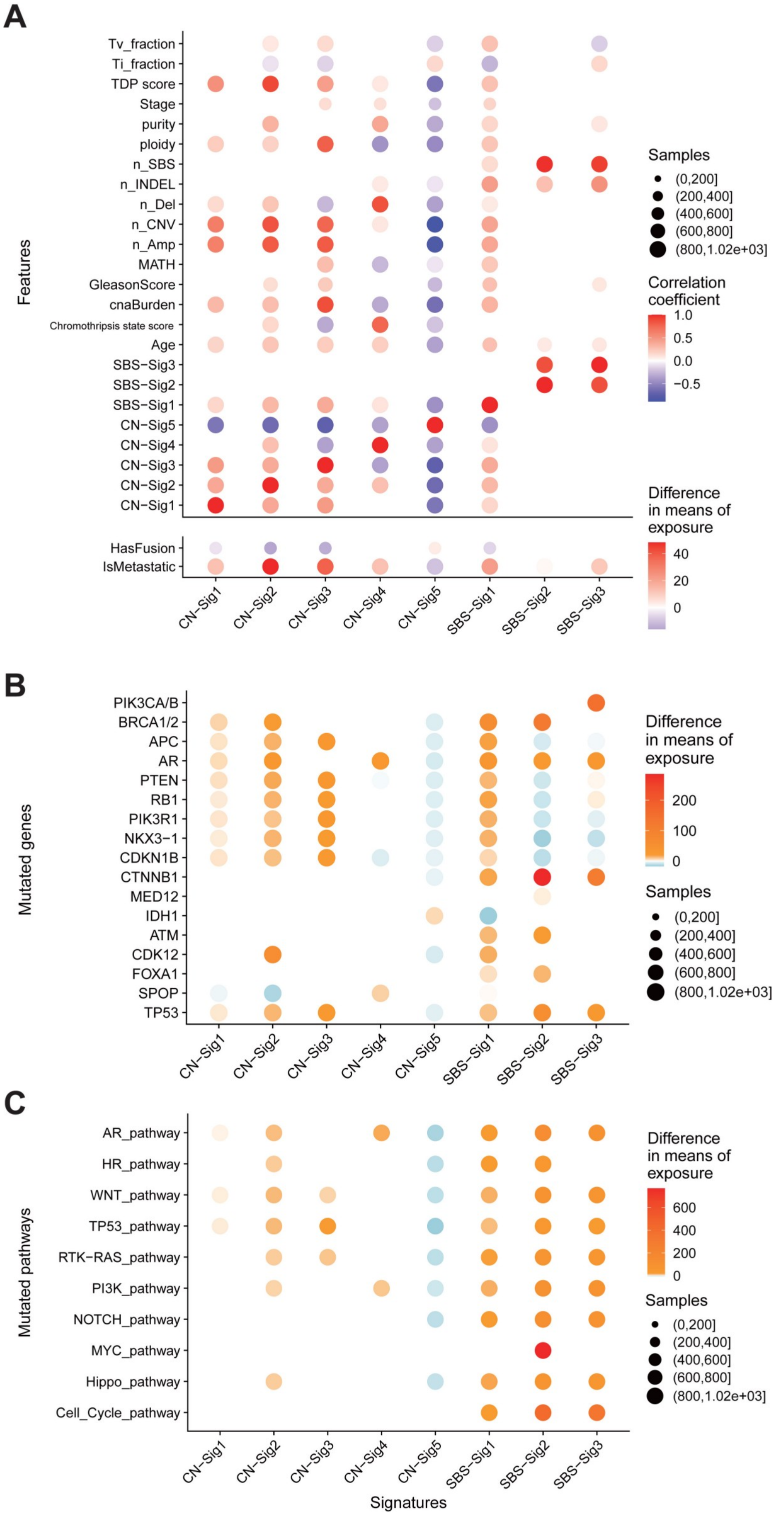

Fig 4. Genome alteration signatures and mutational processes in PC. (A) Associations between the exposures of genome alteration signatures and features including clinical parameters, somatic DNA alteration quantifications. (B,C) Associations between the exposures of genome alteration signatures and mutations in selected genes (B) or pathways (C). Only differences with false discovery rate P < 0.05 (Mann–Whitney U-test) are shown. Red, positive correlation; blue, negative correlation. The depths of colors in filled circles indicate the extent of difference. Sizes of circle in all plots indicate the number of cases included in each analysis.

https://doi.org/10.1371/journal.pgen.1009557.g004

Table 1. Genome alteration signatures in PC. Associations for each genome alteration signature and proposed mechanisms are shown.

Signature	Proposed mechanism or etiology	Evidences	Notable associations
CN-Sig 1	Focal amplification	Many short, focally localized DNA segments with very high absolute copy number	-
CN-Sig 2	Tandem duplication	Many evenly distributed medium-length DNA segment amplification; high number of breakpoints	Positive association: *CDK12* mutation; Homologous recombination pathway mutation; TDP score; metastasis. Negative association: SPOP mutation; *TMPRSS2–ETS* fusion.
CN-Sig 3	Whole-genome duplication	High ploidy, High absolute copy number; Few OsCN; global CNA distribution	Positive association: *TP53* mutation; CNA burden; ploidy; Negative association: *TMPRSS2–ETS* fusion; CN-sig 4; CN-sig 5.
CN-Sig 4	Chromothripsis	Copy number change point is one; considerable OsCN	Positive association: *AR* mutation; *SPOP* mutation; copy number deletion number; chromothripsis state score. Negative association: CN-Sig3; MATH.
CN-Sig 5	Copy number neutral like events including oncogenic *TMPRSS2–ETS* fusion	Few CNA; Focal distribution of CNA	Positive association: *TMPRSS2–ETS* fusion; *IDH1* mutation; Negative association: mutations (except *IDH1* mutation) and other genome alterations.
SBS-Sig 1	Homologous recombination defect or unknown	Highly similar to COSMIC SBS signature 3/5	Positive association: copy number signatures.
SBS-Sig 2	Mismatch repair defect	Highly similar to COSMIC SBS signature15/6	-
SBS-Sig 3	Aging	Highly similar to COSMIC SBS signature 1	-

https://doi.org/10.1371/journal.pgen.1009557.t001

CN-sig 5 shows negative association with nearly all cancer genome alteration features excluding *TMPRSS2-ETS* oncogenic fusions. And CN-sig 5 is the only genome alteration signature that shows positive correlation with the presence of *TMPRSS2-ETS* oncogenic fusions. *TMPRSS2-ETS* fusion happens in early stage of PC [27]. All these observations suggest that CN-sig 5 reflects a stable genome or early PC evolving state.

Multidimensional clustering with genome alteration signatures and PC clinical parameters results in several clusters (S9 Fig). CNA burden feature is clustered with CN-sig 1, 2, 3, 5. And in this cluster, CN-sig 5 show negative correlation with other features or signatures. CN-sig 4 is associated with number of deletions, and forms a distinct cluster when compared to other copy number signatures. SBS signatures form separate clusters from copy number signatures. SBS-sig 2 and SBS-sig 3 form cluster with number of INDEL and total mutation. SBS-sig 1 form a different cluster. These analyses revealed underlying connections between different genome alteration signatures.

Based on the distribution of the value of each copy number component (Fig 2), copy number profiles of representative tumors (Fig 2B), and signature correlation analyses (Figs 4 and S9), the etiologies and mechanisms for copy number signatures has been proposed as following (Table 1):

CN-sig 1 is represented by many short (0.01–0.1Mb), focally localized DNA segments with very high absolute copy number (>8). This signature is caused by focal amplification of DNA segments.

CN-sig 2 is represented by evenly or globally distributed, large amount of medium-length (0.1–10 Mb) DNA segment amplifications. This signature is associated with *CDK12*, homologous recombination (HR) pathway gene mutations and high TDP score, and shows highest association with PC metastasis. This signature can be a result of defective HR DNA repair and consequently tandem duplication phenotypes.

CN-sig 3 is represented by high tumor ploidy, and is associated with high CNA burden and *TP53* mutation. This signature reflects the occurrence of whole genome doubling events.

CN-sig 4 is featured by DNA segment one copy deletion, and considerable oscillating copy number. This signature is associated with *AR*, *SPOP* mutation and high chromothripsis state score, suggesting a state of chromothripsis [28].

CN-sig 5 is represented by large copy number segment size with few background copy number alterations, and is the only signature positively associated with PC oncogenic fusions. This signature is also the only copy number signature enriched specifically in non-metastatic PC, and is associated with good survival. This signature is a result of relatively stable PC genome, reflecting copy number neutral mutation processes (e.g. *TMPRSS2-ETS* oncogenic fusions).

Copy number signature and PC patients' stratification

PC samples are clustered into five groups based on the consensus matrix from multiple NMF runs, and each group is specified by one enriched copy number signature (Fig 5A). Clinical and genomic parameters including patients' age, SBS count (n_SBS), number of insertion/deletions (n_INDEL), transition (Ti) fraction, transversion (Tv) fraction, CNA number, copy number amplification number (n_Amp), copy number deletion number (n_Del), CNA burden, tumor purity, and tumor ploidy are compared in each copy number signature enriched PC patient group (Fig 5B). Significant differences are observed in all above mentioned clinical parameters, except total mutation and number of INDEL among different copy number signature enriched PC groups (Fig 5B). Similar analysis was performed using SBS signature to cluster PC patients. Most clinical and genomic parameters do not show significant difference among three SBS signature enriched PC groups (S10 Fig). This analysis suggests that copy number signatures could have more PC stratification power when compared with SBS signatures.

Metastasis is the leading cause of PC associated death, the associations between each genome alteration signatures and PC metastasis are shown as Sankey diagram (S11A Fig). CN-Sig 2 is nearly exclusively found to be associated with PC metastasis, and CN-Sig 5 is nearly exclusively found not to be associated with PC metastasis. Compared to copy number signatures, SBS signatures do not show strong associations with PC metastasis. This metastasis association analysis suggests that copy number signature can be PC metastasis specific biomarker, and this analysis further demonstrates that copy number signatures provide more PC stratification information than SBS signatures.

The associations between copy number signatures and metastasis, Gleason score and clinical stage have been statistically analyzed (S11B Fig). Similar analysis has been performed with SBS signatures (S11C Fig). Significant differences in metastasis, Gleason score and clinical stage status are observed among different copy number signature enriched PC patients. However, among SBS signature enriched PC patients, no significant differences are observed. These analyses again demonstrate that copy number signatures carry more PC stratification information than SBS signatures.

Copy number signatures predict PC patients' survival

Univariate Cox regression analyses were performed to evaluate the associations between each genome alteration signature and PC patients' survival time (Fig 6). In overall survival (OS) analysis, only copy number signatures show significant correlations, none of SBS signatures show significant correlations (Fig 6A). CN-sig 2 is significantly associated with poor OS, and CN-sig 5 is significantly associated with improved OS. In regards to progression-free survival (PFS), CN-sig 3 is significantly associated with poor PFS, and CN-sig 5 is significantly associated with improved PFS (Fig 6B).

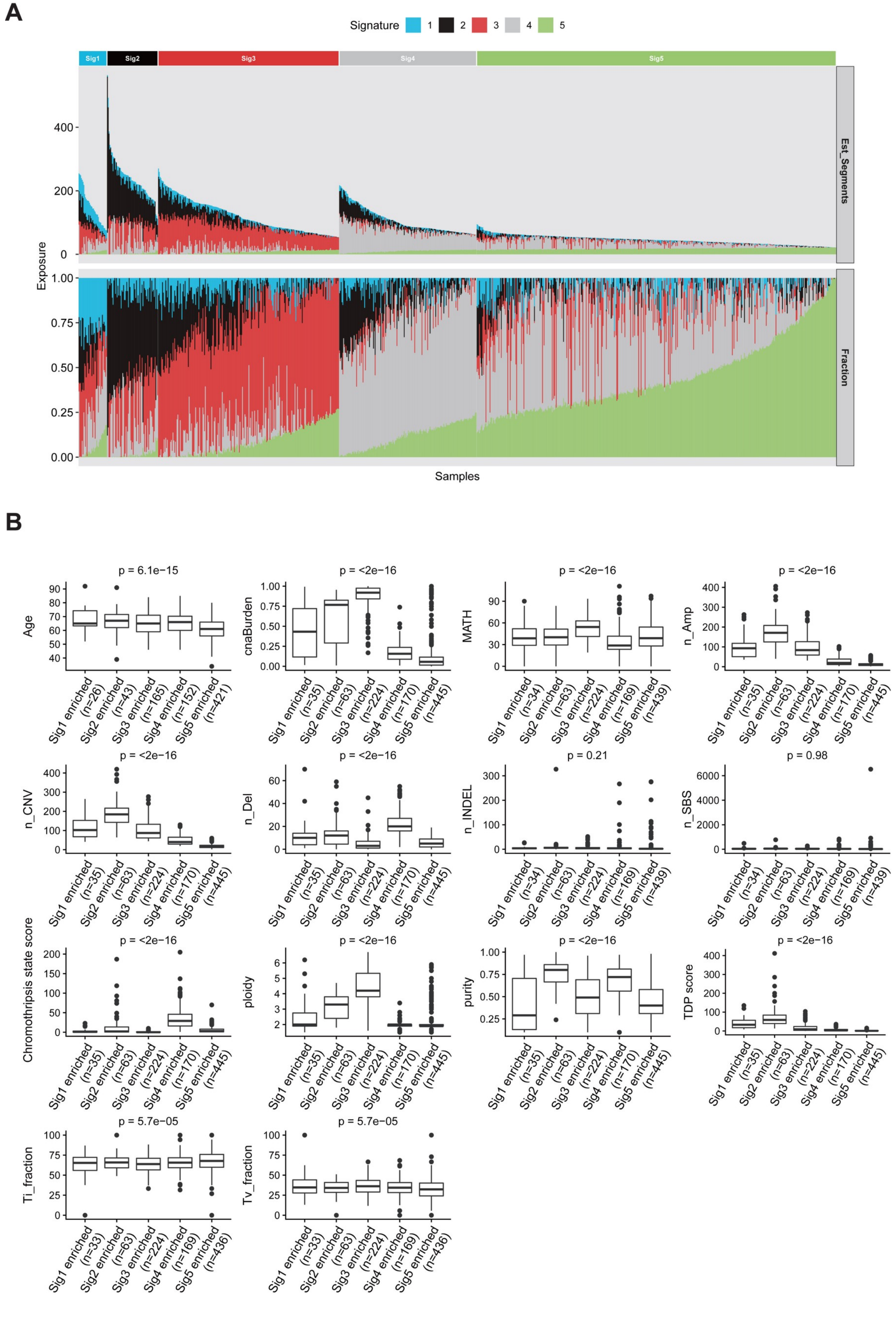
A
Signature 1 2 3 4 5
Sig1
Sig2
Sig3
Sig4
Sig5
Exposure
400
200
0
1.00
0.75
0.50
0.25
0.00
Est_Segments
Fraction
Samples
B
p = 6.1e−15
Age
Sig1 enriched (n=26)
Sig2 enriched (n=43)
Sig3 enriched (n=165)
Sig4 enriched (n=152)
Sig5 enriched (n=421)
p = <2e−16
cnaBurden
Sig1 enriched (n=35)
Sig2 enriched (n=63)
Sig3 enriched (n=224)
Sig4 enriched (n=170)
Sig5 enriched (n=445)
p = <2e−16
MATH
Sig1 enriched (n=34)
Sig2 enriched (n=63)
Sig3 enriched (n=224)
Sig4 enriched (n=169)
Sig5 enriched (n=439)
p = <2e−16
n_Amp
p = <2e−16
n_CNV
p = <2e−16
n_Del
p = 0.21
n_INDEL
p = 0.98
n_SBS
p = <2e−16
Chromothripsis state score
p = <2e−16
ploidy
p = <2e−16
purity
p = <2e−16
TDP score
p = 5.7e−05
Ti_fraction
Sig1 enriched (n=33)
Sig2 enriched (n=63)
Sig3 enriched (n=224)
Sig4 enriched (n=169)
Sig5 enriched (n=436)
p = 5.7e−05
Tv_fraction

Fig 5. Sample clustering and heterogeneity analysis of PC based on the exposures of copy number signatures. (A) For each PC patient, the relative contribution (bottom panel) and estimated copy number segment counts (top panel) of each signature are shown as a staked barplot. PC samples are clustered into five groups based on the consensus matrix from multiple NMF runs, and each group is specified by one enriched copy number signature. (B) Quantitative comparison for somatic and clinical parameters among each copy number signature enriched PC group by boxplot. The boxplot is bounded by the first and third quartile with a horizontal line at the median. ANOVA p values are shown. Abbr.: n_, number of, e.g. n_SNV, number of SNV; INDELs, small insertions and deletions; Ti_fraction, transition fraction; Tv_fraction, transversion fraction; Amp, copy number segments with amplification; Del, copy number segments with deletion; TDP score, tandem duplication phenotype score; cnaBurden, copy number alteration burden; MATH, MATH score is a quantitative measure of intra-tumor heterogeneity.

https://doi.org/10.1371/journal.pgen.1009557.g005

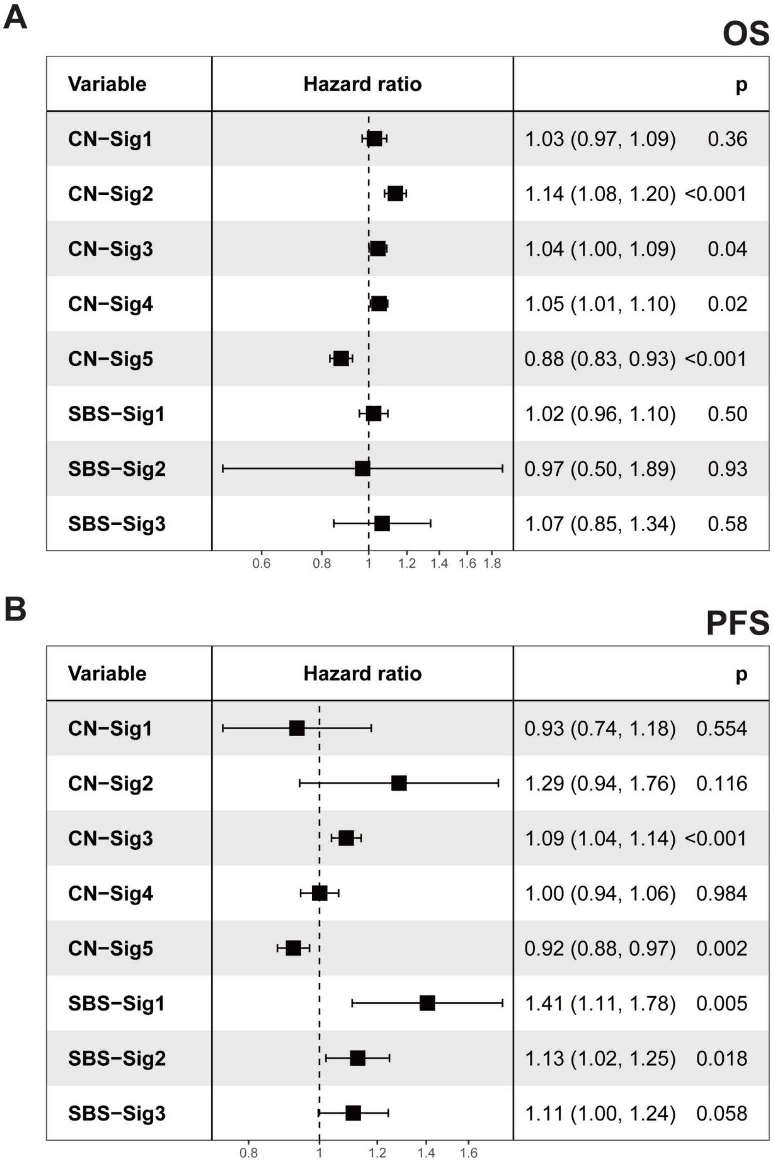

Fig 6. Genome alteration signature exposures and PC survival. (A,B) Forest plots showing the relative risk of signature exposures in overall survival (OS) (A) and progression-free survival (PFS). Due to the availability of clinical information, OS analyzes were performed with phs000178 dataset (TCGA, 498 primary prostate cancer) and phs000554 dataset (48 metastatic prostate cancer), PFS analyzes were performed with phs000178 dataset. (B). Exposures for all signatures were normalized to 1–20 for evaluating the hazard ratios per 5% exposure increase. Hazard ratios and p values were calculated by univariable Cox analysis. Squares represent hazard ratio, horizontal lines indicate the 95% confidence interval.

https://doi.org/10.1371/journal.pgen.1009557.g006

The associations between other clinical genomic parameters and OS, PFS are analyzed with univariate Cox regression method (S12 Fig). As expected, high Gleason scores are significantly associated with both poor OS and poor PFS. Late clinical stages are significantly associated with poor PFS. CNA burden is significantly associated with poor PFS.

This survival time analysis suggests that copy number signatures are associated with PC patients' survival. CN-sig 5, which has the lowest CNA burden, is associated with improved survival, while CN sig-2 is associated with poor survival. This is consistent with previous report that increased CNA burden was associated with poor PC survival [9]. In addition, this study further suggests that specific type of CNA represented by CN-sig 2 is associated with poor PC survival. CN-sig 2 has the highest hazard ratio among all five copy number signatures in both OS and PFS analysis. Of note, CN-sig 2 is not the copy number signature with the highest CNA burden (Fig 6).

Discussion

Here a novel copy number alteration signature analysis method has been constructed. This method is featured by enhanced signature extraction efficiency, enhanced component visualization and comprehension, and can be coherently integrated with the known SBS signature analysis method. With this newly developed bioinformatics tool "sigminer", we performed the first copy number signature analysis in PC, and identified five copy number signatures. The mutational processes underlying each copy number signature have been illustrated. Copy number signatures show improved performance in PC patients' stratification and survival prediction compared with SBS signatures. Specific copy number signatures associated with PC metastasis and survival have been identified. These analyses provide novel insight into the etiology of PC, and novel biomarkers for PC stratification and prognosis. We demonstrated the performance of "sigminer" in PC, and this tool can be applied in other cancer types for the deep understanding of copy number alterations.

In survival analysis, the prognostic performance of CNA signature is much improved compared with SBS signature. In addition, copy number signature also shows improved association with PC metastasis and clinical stages compared with SBS signature. This implicates that the factors underlying copy number alterations can be major driving forces for PC progression and metastasis. And this difference needs to be further investigated in other cancer types. In addition, different types of copy number signature can have different impacts on PC prognosis, this greatly extends previous understanding that CNA burden is associated with PC prognosis [9,13].

Our copy number signature analysis tool "sigminer" only needs absolute copy number profile derived from WES or SNP array data, and high coverage WGS data is not needed. This is different from the recently described rearrangement signatures [29], which require high coverage WGS data for analysis. Copy number signature analysis can ultimately improve cancer patients' stratification, clinical prognosis and therapeutic outcome prediction.

Methods

Cohort collection and preprocessing

PC samples were included in this study when tumor and matched germline whole exome sequencing (WES) raw data (BAM or SRA files) from fresh frozen samples were accessible. The preprocessing steps are abstracted in S2 Fig. In total, 6 cohorts are included in this study, their dbGaP accession numbers are phs000178 (TCGA) [20], phs000447 [21], phs000554 [10], phs000909 [22], phs000915 [23] and phs001141 [24]. In total 1003 tumor-normal paired WES data are available for this study. BAM files for TCGA cohort were downloaded from GDC portal (https://portal.gdc.cancer.gov/) with tool gdc-client. The average sequencing depth of each BAM file was summarized in S5 Table. SRA files for other cohorts were obtained from dbGaP database (https://dbgap.ncbi.nlm.nih.gov/) and converted to FASTQ files by SRA toolkit. Adapters were removed from the FASTQ files by trimgalore (https://github.com/FelixKrueger/TrimGalore). BWA MEM algorithm were then applied using hg38 as reference genome [30]. The result SAM files were converted to BAM files with samtools [31], followed by Picard toolkit (https://broadinstitute.github.io/picard/) to sort BAM files and mark duplications for variant and copy number calling. The corresponding codes are provided in section "Code availability".

Clinical data

Clinicopathological annotations for cohorts phs000447, phs000554, phs000909, phs000915 and phs001141 were obtained from the original papers and dbGap database. Clinicopathological annotations for TCGA prostate cohort were downloaded from UCSC Xena by R package UCSCXenaTools v1.2.10 [32]. The data were cleaned and organized by in-house R scripts available in "Code availability" section.

Absolute copy number calling from WES

Two bioinformatics tools including Sequenza [18] and FACETS [19] were performed to generate absolute copy number profiles from tumor-normal paired WES BAM files. For Sequenza, we followed its standard pipeline described in its vignette (https://cran.r-project.org/web/packages/sequenza/vignettes/sequenza.html) but with parameter 'female' set to 'FALSE' and a modified R package copynumber (https://github.com/ShixiangWang/copynumber) to work with hg38 genome build. For FACETS, we strictly followed its standard workflow described in its vignette (https://github.com/mskcc/facets/blob/master/vignettes/FACETS.pdf). Samples failed in calling were excluded from downstream analysis. In total Sequenza successfully called 937 pairs of WES data, and FACETS called 933 pairs of WES data. Results from FACETS and Sequenza are compared, and false positive signals have been carefully filtered. The somatic absolute copy number profiles detected in this study were summarized in S6 Table.

Variant calling from WES

Somatic mutations including SBS and INDEL were detected by following the best practices of Genome Analysis Toolkit (GATK v4.1.3) with Mutect2 [33]. For comparison and data validation, additional variant caller, including varscan2 [34], muse [35], somaticsniper [36], have been applied. The lists of genome variants that passed all quality control filters were converted into VCF files by VCFtools [37]. The resulting VCF files were annotated by VEP [38] and further converted to MAF file by vcf2maf.pl (https://github.com/mskcc/vcf2maf). The MAF file was loaded into R, analyzed and visualized by Maftools [39]. The somatic genome variants detected in this study are available: https://zenodo.org/record/4591005.

Genome alteration signature extraction procedures

To identify SBS and copy number signatures in a similar way, the analysis procedure has been properly abstracted into the following four common steps and implemented as an R/CRAN package sigminer (https://cran.r-project.org/web/packages/sigminer/).

1. Tally variation components. a) for SBS profile, same as previously reported [3, 4], for each tumor, we firstly classified mutation records into six substitution subtypes: C>A, C>G, C>T, T>A, T>C, and T>G (all substitutions are referred to by the pyrimidine of the

mutated Watson–Crick base pair). Further, each of the substitutions was examined by incorporating information on the bases immediately 5' and 3' to each mutated base generating 96 possible mutation types (6 types of substitution * 4 types of 5' base * 4 types of 3' base). b) For copy number profile, we firstly computed the genome-wide distributions of 8 fundamental copy number features for each tumor: the breakpoint count per 10 Mb (named BP10MB); the breakpoint count per chromosome arm (named BPArm); the copy number of the segments (named CN); the difference in copy number between adjacent segments (named CNCP); the lengths of oscillating copy number segment chains (named OsCN); the log10 based copy number segment size (named SS); the minimal number of chromosome with 50% copy number variation (named NC50); the burden of chromosome (named BoChr). These features were selected as hallmarks of previously reported genomic aberrations like chromothripsis or to denote the distribution pattern of copy number events [16,17,40,41]. The former 6 features have been used in previous study [15] to uncover the mutational processes in ovarian carcinoma. Next, unlike previous study[15], which applied mixture modeling to separate the 6 copy number features' distributions into mixtures of Poisson or Gaussian distributions, we directly classified 8 copy number features' distributions into 80 components according to the comprehensive consideration of value range, abundance and biological significance (S5E Fig). Most of the copy number components are discrete values, and the remaining are range values (see S1 Table). Based on the genome alteration component definition described above, two tumor-by-component matrices (one for SBS and the other for copy number) were generated and treated as the input of non-negative matrix factorization (NMF) algorithm for extracting signatures as previously reported, individually [3,4,42].

2. Estimate signature number. Signature number (or factorization rank) is a critical value which affects both the performance of NMF algorithm and biological interpretability. A common way for determining signature number is to try different values, compute some quality measures of the results, and choose the best value according to this quality criteria [42]. The most common criteria is the cophenetic correlation coefficient [43]. As suggested, performing 30–50 runs is considered sufficient to obtain a robust estimation of signature number value [42]. We performed 50 runs for both SBS (signature number range from 2 to 10) and copy number signatures (signature number range from 2 to 12). According to the cophenetic vs signature number plot (S5a and S5b Fig), the number of SBS signatures and copy number signatures were selected to be three and five respectively.

3. Extract signatures. After determining the signature number, we performed NMF with 50 runs to extract signatures for downstream analysis.

4. Quantify signature exposure (or signature activity). Sigminer package can provide both relative and absolute exposures of each signature for a tumor. The relative exposure of each signature represents its contribution proportion in a tumor among all signatures and can be directly obtained after NMF. The absolute exposure for each signature across tumors was determined after a scaling transformation as previously described [44]. For SBS signature, the absolute exposure represents the expected number of mutations associated with each SBS signature. For copy number signature, the absolute exposure represents the expected number of copy number segment records associated with each copy number signature.

Signature profile normalization

Signature profile is essentially a matrix with row representing signature components and column representing the contributions of each component. Same as previous reported [4], SBS signature profile was normalized within each signature (i.e. by row), and a component value can be compared to another component value of the same copy number feature within each signature.

Single sample copy number signature analysis and cancer subtype classification

After *de novo* signature discovery, relative and absolute signature exposures can be readily obtained with sigminer package. In clinical practice, it would be useful to quantify the signature exposures for a single tumor based on existing signatures (e.g. signatures from *de novo* signature discovery or public signature database). Sigminer package adopted the quadratic programming method to fit the existing signatures to one or more tumors. To classify a single tumor based on signature exposures, we built a 5-layer neural network model (input layer + hidden layer + 2 dropout layers + output layer) with Keras library (https://keras.rstudio.com/) and trained it with datasets used in our study. The evaluation and selection of hyper-parameters of models were determined by grid search. This model can predict the subtype for single PC sample with copy number data available in high accuracy (average >0.9, see the link below). All steps to build and train the model for practical use accompanying with model training and validation results are packaged and available at https://github.com/ShixiangWang/sigminer.prediction for research community.

Association analysis

Same as previously reported [15], associations between the exposures of signatures and other clinical or genomic features was performed using one of two procedures: 1) for a continuous association variable (including ordinal variable like clinical stage), Pearson correlation was performed; 2) for a binary variable, patients were divided into two groups and a Mann-Whiney U-test was performed to test for differences in average exposures of signatures between the two groups.

Correlation network analysis

To investigate the structure of signature associations, correlation network analysis was performed with R package corrr (https://github.com/tidymodels/corrr) with continuous association variables. Variables that are more correlated appear closer together and are joined by thicker curves. Red and blue curves indicate positive and negative correlation respectively. The proximity of the points was determined with multidimensional clustering.

Score definitions

To quantify the tandem duplication (TD) status for each tumor, we defined a tandem duplication phenotype (TDP) score, which capture the distribution and length information of TD across chromosomes [17,45]. This TDP score is calculated as:

$$\frac{TD_{total}}{\Sigma_{chr}|TD_{obs} - TD_{exp}| + 1} \times L$$

Where TD_{total} is the total number of TD in a sample, TD_{obs} and TD_{exp} are the observed and expected TD for a chromosome in the tumor, L is the total size of TD in Mb unit. TD is defined as copy number amplification segments with size range from 1Kb to 2Mb.

To quantify the chromothripsis state for each tumor, we defined a chromothripsis state score. This score is based on one key feature of chromothripsis, which forms tens to hundreds of locally clustered segmental losses being interspersed with regions displaying normal (disomic) copy-number [16]. This chromothripsis state score is calculated as:

$$\sum_{chr} N_{OsCN}^{2}$$

Where N_{OsCN}^{2} is the square of total number of copy number fragments with absolute copy number value following "2-1-2" pattern. This score is a simplified estimation of chromothripsis state for each tumor and was visually inspected.

Survival analysis

Univariate Cox analysis and visualization were performed by R package survival and ezcox (https://github.com/ShixiangWang/ezcox). The cox model returns hazard ratio value (including 95% confidence interval) per variable unit increase. To properly determine the relevance between variables and PC patients' OS (overall survival) and PFS (progression-free survival), exposures of all signatures were normalized to 1–20 for indicating the hazard ratio per 5% exposure increase. In a similar way, we multiplied some variables including CNA burden, Ti fraction (transition fraction) and tumor purity by 20 (the raw values of these variables range from 0 to 1). For other variables, we directly evaluated the hazard ratio per variable unit increase, e.g. hazard ratio for n_SBS indicates the hazard ratio per SBS count increase in a sample.

Statistical analysis

Correlation analysis was performed using the Pearson method. Mann-Whiney U-test was performed to test for differences in signature exposure medians between the two groups. ANOVA was performed to test for differences across more than 2 groups. Fisher test was performed to test the association between signatures and categorical variables. In larger than 2 by 2 tables, Fisher p-values were calculated by Monte Carlo simulation. For multiple hypothesis testing, p values were adjusted using the false discovery rate method. All reported p-values are two-tailed, and for all analyses, $p< = 0.05$ is considered statistically significant, unless otherwise specified. Statistical analyses were performed by R v3.6 (https://cran.r-project.org/).

Supporting information

S1 Table. Copy number component parameter setting.
(XLSX)

S2 Table. Cohort characteristics.
(XLSX)

S3 Table. Summary statistics for somatic genome variations.
(XLSX)

S4 Table. Selected pathways and genes for association analysis.
(XLSX)

S5 Table. Sequencing depth information for BAM files.
(XLSX)

S6 Table. Summary table for absolute copy number profiles detected in this study.
(XLSX)

S1 Fig. Application comparison of two copy number signature extraction methods (our method and Macintyre et al. method) to ovary cancer with different sample size. (A) The workflow of this analysis. Copy number data with different sample size are inputted into sigminer for matrix generation and signature extraction using two methods. (B, C) Signature profiles of different sample size generated through our method (B) and Macintyre et al. method (C) are shown. (D) Cosine similarity heatmaps for signatures extracted using our method. Cosine similarity analysis cannot be applied to signature profiles generated using Macintyre et al. method, because of unequal CNA components in ovarian cancer cohorts with different sample size.
(TIF)

S2 Fig. Comparison between Macintyre et al. method and the new method developed in this study. (A) Copy number signatures extracted with the Macintyre et al. method. (B) Representative copy number signature extracted with the new method in this study. The following differences are listed: (1) Here we used predefined copy number component, and calculate the weight of each component with a counting based method. Macintyre et al. applied mixture modeling to separate the copy number feature distributions into mixtures of Poisson or Gaussian distributions. The biological meaning of each component of Macintyre et al. method is not clear. While the biological meaning of our predefined component is clear. (2) Our method is much more computationally efficient than Macintyre et al. method. (3) Different normalization, we performed inter-feature row normalization, thus the values of the same signature can be compared, however, Macintyre et al. performed column normalization, and the values of the same signature cannot be compared. (4) Our method includes two additional copy number features “NC50” and “BoChr” to reflect the chromosome distribution pattern of copy number alterations. These features provided additional information for understanding copy number alterations. (C) List of the differences between Macintyre *et al* study and this study.
(TIF)

S3 Fig. Study design and flowchart depicting the processing steps from raw sequencing data to genome alteration signatures.
(TIF)

S4 Fig. Mutational landscape of PC WES datasets. Driver genes are identified by MutSig with q value < 0.05. The right panel indicates log10 based MutSig q values for driver genes. This plot was generated by Maftools with default setting.
(TIF)

S5 Fig. Summary for small scale variants and copy number alteration features in PC WES dataset. (A) Number of variants of different types in PC WES datasets. TNP: triple nucleotide polymorphism; SNP: single nucleotide polymorphism; INS: insertion; DNP: double nucleotide polymorphism; DEL: deletion. (B) Number of variants in each sample as a stacked barplot and variant classification as a boxplot. (C) Top 10 mutated genes as a stacked barplot by variant classification. (D) Length distribution of somatic copy number alterations (SCNA) segment and chromosome distribution of SCNAs. Location ‘pq’ represents a segment across both p arm and q arm. (E) Frequency distribution of 8 copy number features in combined PC WES

dataset. The x coordinate 23 in feature 'BoChr' represents chromosome X.
(TIF)

S6 Fig. Cophenetic plot for determining the number of signatures. (A,B) Cophenetic plot analysis for copy number signatures (A) and SBS signatures (B). (C,D) The consensus matrices for copy number signatures (C) and SBS signatures (D).
(TIF)

S7 Fig. SBS signatures identified in PC and the workflow for signature stability analysis. (A) Three SBS signatures are identified from PC WES data. Each signature was row normalized. (B) Bootstrap analysis workflow for signature stability applied to each tumor. For a tumor, firstly a mutation catalog was obtained based on component classification (96 components for SBS and 80 components for CNA), then a bootstrap procedure was adopted to get 1000 bootstrap catalogs based on observed mutation probability, this can generate a distribution of signature exposure for a p-value calculation in statistical test, and also quantify the signature stability by calculating RMSE between observed signature exposures (from observed catalogs) and bootstrap signature exposures (from bootstrap catalogs).
(TIF)

S8 Fig. Cosine similarity analysis between SBS signatures identified in this study and COSMIC signatures.
(TIF)

S9 Fig. Correlation network analysis for genome alteration signatures. Variables that are more highly correlated appear closer together and are joined by stronger curves. Red color indicates positive correlation and blue color indicate negative correlation. The proximity of the points are determined using multidimensional clustering. Associations with Pearson correlation coefficient r>0.2 are shown.
(TIF)

S10 Fig. Sample clustering and heterogeneity analysis of PC based on the exposures of SBS signatures. (A) For each PC patient, the relative contribution (bottom panel) and estimated SBS mutational counts (top panel) of each signature are shown as a staked barplot. PC samples are clustered into three groups based on the consensus matrix from multiple NMF runs, and each group is specified by one enriched SBS signature. (B) Quantification comparison for somatic and clinical parameters among each SBS signature enriched PC group by boxplot. The boxplot is bounded by the first and third quartile with a horizontal line at the median. ANOVA p values are shown. Abbr.: n_, number of, e.g. n_SNV, number of SNV; INDELs, small insertions and deletions; Ti_fraction, transition fraction; Tv_fraction, transversion fraction; Amp, copy number segments with amplification; Del, copy number segments with deletion; TDP score, tandem duplication phenotype score; cnaBurden, copy number alteration burden; MATH, MATH score is a quantitative measure of intra-tumor heterogeneity.
(TIF)

S11 Fig. Associations between genome alteration signatures and PC clinical variables including metastatic status, clinical stage and Gleason score. (A) Relationship between copy number signature exposure or SBS signature exposure and metastasis is shown by a Sankey plot. The sample size are indicated on the left of plot. (B,C) Fraction changes of clinical variables including metastasis, clinical stage and Gleason score in each copy number signature enriched PC groups (B) or each SBS signature enriched PC groups (C). p values were calculated by Fisher test with Monte Carlo simulation.
(TIF)

S12 Fig. Genomic and clinical features and PC survival. Forest plots showing the relative risk of selected genomic and clinical features in overall survival (OS) (A) and progression-free survival (PFS) (B). Due to the availability of clinical information, OS analyzes were performed with phs000178 dataset (TCGA, 498 primary prostate cancer) and phs000554 dataset (48 metastatic prostate cancer), PFS analyzes were performed with phs000178 dataset. Hazard ratios and p values were calculated by univariable Cox analysis. Squares represent hazard ratio, horizontal lines indicate the 95% confidence interval. Abbr.: n_, number of, e.g. n_SNV, number of SNV; INDELs, small insertions and deletions; Ti_fraction, transition fraction; Tv_fraction, transversion fraction; Amp, copy number segments with amplification; Del, copy number segments with deletion; TDP score, tandem duplication phenotype score; MATH, MATH score is a quantitative measure of intra-tumor heterogeneity.
(TIF)

Acknowledgments

We thank Raymond Shuter for editing the text. Thank ShanghaiTech University High Performance Computing Public Service Platform for computing services.

Author Contributions

Conceptualization: Xue-Song Liu.

Data curation: Shixiang Wang, Huimin Li, Minfang Song, Ziyu Tao.

Formal analysis: Shixiang Wang, Xue-Song Liu.

Funding acquisition: Xue-Song Liu.

Investigation: Shixiang Wang, Tao Wu, Zaoke He, Xiangyu Zhao, Xue-Song Liu.

Methodology: Shixiang Wang, Xue-Song Liu.

Project administration: Xue-Song Liu.

Resources: Kai Wu, Xue-Song Liu.

Software: Shixiang Wang.

Supervision: Xue-Song Liu.

Visualization: Shixiang Wang.

Writing – original draft: Xue-Song Liu.

Writing – review & editing: Xue-Song Liu.

References

1. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463(7283):899–905. Epub 2010/02/19. https://doi.org/10.1038/nature08822 PMID: 20164920; PubMed Central PMCID: PMC2826709.

2. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013; 45(10):1134–40. Epub 2013/09/28. https://doi.org/10.1038/ng.2760 PMID: 24071852; PubMed Central PMCID: PMC3966983.

3. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500(7463):415–21. Epub 2013/08/16. https://doi.org/10.1038/nature12477 PMID: 23945592; PubMed Central PMCID: PMC3776390.

4. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020; 578(7793):94–101. Epub 2020/02/07. https://doi.org/10.1038/s41586-020-1943-3 PMID: 32025018; PubMed Central PMCID: PMC7054213.

5. Gulhan DC, Lee JJ, Melloni GEM, Cortes-Ciriano I, Park PJ. Detecting the mutational signature of homologous recombination deficiency in clinical samples. Nat Genet. 2019; 51(5):912–9. Epub 2019/04/17. https://doi.org/10.1038/s41588-019-0390-2 PMID: 30988514.

6. Wang S, Jia M, He Z, Liu XS. APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. Oncogene. 2018; 37(29):3924–36. Epub 2018/04/27. https://doi.org/10.1038/s41388-018-0245-9 PMID: 29695832; PubMed Central PMCID: PMC6053356.

7. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin. 2015; 65(2):87–108. Epub 2015/02/06. https://doi.org/10.3322/caac.21262 PMID: 25651787.

8. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010; 18(1):11–22. Epub 2010/06/29. https://doi.org/10.1016/j.ccr.2010.05.026 PMID: 20579941; PubMed Central PMCID: PMC3198787.

9. Hieronymus H, Schultz N, Gopalan A, Carver BS, Chang MT, Xiao Y, et al. Copy number alteration burden predicts prostate cancer relapse. Proc Natl Acad Sci U S A. 2014; 111(30):11139–44. Epub 2014/07/16. https://doi.org/10.1073/pnas.1411446111 PMID: 25024180; PubMed Central PMCID: PMC4121784.

10. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature. 2012; 487(7406):239–43. Epub 2012/06/23. https://doi.org/10.1038/nature11125 PMID: 22722839; PubMed Central PMCID: PMC3396711.

11. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012; 44(6):685–9. Epub 2012/05/23. https://doi.org/10.1038/ng.2279 PMID: 22610119; PubMed Central PMCID: PMC3673022.

12. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005; 310(5748):644–8. https://doi.org/10.1126/science.1117679 WOS:000232997700032. PMID: 16254181

13. Hieronymus H, Murali R, Tin A, Yadav K, Abida W, Moller H, et al. Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. Elife. 2018;7. Epub 2018/09/05. https://doi.org/10.7554/eLife.37294 PMID: 30178746; PubMed Central PMCID: PMC6145837.

14. Stopsack KH, Whittaker CA, Gerke TA, Loda M, Kantoff PW, Mucci LA, et al. Aneuploidy drives lethal progression in prostate cancer. Proc Natl Acad Sci U S A. 2019; 116(23):11390–5. Epub 2019/05/16. https://doi.org/10.1073/pnas.1902645116 PMID: 31085648; PubMed Central PMCID: PMC6561291.

15. Macintyre G, Goranova TE, De Silva D, Ennis D, Piskorz AM, Eldridge M, et al. Copy number signatures and mutational processes in ovarian carcinoma. Nat Genet. 2018; 50(9):1262–70. Epub 2018/08/15. https://doi.org/10.1038/s41588-018-0179-8 PMID: 30104763; PubMed Central PMCID: PMC6130818.

16. Korbel JO, Campbell PJ. Criteria for inference of chromothripsis in cancer genomes. Cell. 2013; 152(6):1226–36. Epub 2013/03/19. https://doi.org/10.1016/j.cell.2013.02.023 PMID: 23498933.

17. Menghi F, Inaki K, Woo X, Kumar PA, Grzeda KR, Malhotra A, et al. The tandem duplicator phenotype as a distinct genomic configuration in cancer. Proc Natl Acad Sci U S A. 2016; 113(17):E2373–82. Epub 2016/04/14. https://doi.org/10.1073/pnas.1520010113 PMID: 27071093; PubMed Central PMCID: PMC4855596.

18. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol. 2015; 26(1):64–70. Epub 2014/10/17. https://doi.org/10.1093/annonc/mdu479 PMID: 25319062; PubMed Central PMCID: PMC4269342.

19. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016; 44(16):e131. Epub 2016/06/09. https://doi.org/10.1093/nar/gkw520 PMID: 27270079; PubMed Central PMCID: PMC5027494.

20. Cancer Genome Atlas Research N. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015; 163(4):1011–25. Epub 2015/11/07. https://doi.org/10.1016/j.cell.2015.10.025 PubMed Central PMCID: PMC4695400. PMID: 26544944

21. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. Nature. 2011; 470(7333):214–20. Epub 2011/02/11. https://doi.org/10.1038/nature09744 PMID: 21307934; PubMed Central PMCID: PMC3075885.

22. Beltran H, Rickman DS, Park K, Chae SS, Sboner A, MacDonald TY, et al. Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets. Cancer Discov. 2011; 1

(6):487–95. Epub 2012/03/06. https://doi.org/10.1158/2159-8290.CD-11-0130 PMID: 22389870; PubMed Central PMCID: PMC3290518.

23. Robinson D, Van Allen EM, Wu YM, Schultz N, Lonigro RJ, Mosquera JM, et al. Integrative clinical genomics of advanced prostate cancer. Cell. 2015; 161(5):1215–28. Epub 2015/05/23. https://doi.org/10.1016/j.cell.2015.05.001 PMID: 26000489; PubMed Central PMCID: PMC4484602.

24. Kohli M, Wang L, Xie F, Sicotte H, Yin P, Dehm SM, et al. Mutational Landscapes of Sequential Prostate Metastases and Matched Patient Derived Xenografts during Enzalutamide Therapy. PLoS One. 2015; 10(12):e0145176. Epub 2015/12/24. https://doi.org/10.1371/journal.pone.0145176 PMID: 26695660; PubMed Central PMCID: PMC4687867.

25. Huang X, Wojtowicz D, Przytycka TM. Detecting presence of mutational signatures in cancer with confidence. Bioinformatics. 2018; 34(2):330–7. Epub 2017/10/14. https://doi.org/10.1093/bioinformatics/btx604 PMID: 29028923; PubMed Central PMCID: PMC5860213.

26. Mroz EA, Rocco JW. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. Oral Oncol. 2013; 49(3):211–5. Epub 2012/10/20. https://doi.org/10.1016/j.oraloncology.2012.09.007 PMID: 23079694; PubMed Central PMCID: PMC3570658.

27. Gerhauser C, Favero F, Risch T, Simon R, Feuerbach L, Assenov Y, et al. Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. Cancer Cell. 2018; 34(6):996–1011 e8. Epub 2018/12/12. https://doi.org/10.1016/j.ccell.2018.10.016 PMID: 30537516

28. Yi K, Ju YS. Patterns and mechanisms of structural variations in human cancer. Exp Mol Med. 2018; 50 (8):98. Epub 2018/08/10. https://doi.org/10.1038/s12276-018-0112-3 PMID: 30089796; PubMed Central PMCID: PMC6082854.

29. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016; 534(7605):47–54. Epub 2016/05/03. https://doi.org/10.1038/nature17676 PMID: 27135926; PubMed Central PMCID: PMC4910866.

30. Li H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. ArXiv. 2013; 1303. 3997.

31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. Epub 2009/06/10. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943; PubMed Central PMCID: PMC2723002.

32. Wang S, Liu X. The UCSCXenaTools R Package: A Toolkit for Accessing Genomics Data from UCSC Xena Platform, from Cancer Multi-Omics to Single-Cell RNA-Seq. Journal of Open Source Software. 2019; 4(40). https://doi.org/10.21105/joss.01627

33. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. bioRxiv. 2019. https://doi.org/10.1101/861054.

34. Koboldt DC, Zhang QY, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. 2012; 22 (3):568–76. WOS:000300962600016. https://doi.org/10.1101/gr.129684.111 PMID: 22300766

35. Fan Y, Xi L, Hughes DST, Zhang JJ, Zhang JH, Futreal PA, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biology. 2016;17. WOS:000383427300001. https://doi.org/10.1186/s13059-016-0875-6 PMID: 26832153

36. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012; 28(3):311–7. WOS:000300043200003. https://doi.org/10.1093/bioinformatics/btr665 PMID: 22155872

37. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–8. Epub 2011/06/10. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522; PubMed Central PMCID: PMC3137218.

38. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17(1):122. Epub 2016/06/09. https://doi.org/10.1186/s13059-016-0974-4 PMID: 27268795; PubMed Central PMCID: PMC4893825.

39. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018; 28(11):1747–56. Epub 2018/10/21. https://doi.org/10.1101/gr.239244.118 PMID: 30341162; PubMed Central PMCID: PMC6211645.

40. Murnane JP. Telomere dysfunction and chromosome instability. Mutat Res. 2012; 730(1–2):28–36. Epub 2011/05/18. https://doi.org/10.1016/j.mrfmmm.2011.04.008 PMID: 21575645; PubMed Central PMCID: PMC3178001.

41. Ng CK, Cooke SL, Howe K, Newman S, Xian J, Temple J, et al. The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer. J Pathol. 2012; 226(5):703–12. Epub 2011/12/21. https://doi.org/10.1002/path.3980 PMID: 22183581.

42. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 2010; 11:367. Epub 2010/07/06. https://doi.org/10.1186/1471-2105-11-367 PMID: 20598126; PubMed Central PMCID: PMC2912887.

43. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004; 101(12):4164–9. Epub 2004/03/16. https://doi.org/10.1073/pnas.0308531101 PMID: 15016911; PubMed Central PMCID: PMC384712.

44. Kim J, Mouw KW, Polak P, Braunstein LZ, Kamburov A, Kwiatkowski DJ, et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat Genet. 2016; 48(6):600–6. Epub 2016/04/26. https://doi.org/10.1038/ng.3557 PMID: 27111033; PubMed Central PMCID: PMC4936490.

45. Menghi F, Barthel FP, Yadav V, Tang M, Ji B, Tang Z, et al. The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. Cancer Cell. 2018; 34(2):197–210 e5. Epub 2018/07/19. https://doi.org/10.1016/j.ccell.2018.06.008 PMID: 30017478; PubMed Central PMCID: PMC6481635.