EBioMedicine 74 (2021) 103716

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Original Research

Identification of copy number variation-driven molecular subtypes informative for prognosis and treatment in pancreatic adenocarcinoma of a Chinese cohort

Qian Zhan[a,b,§], Chenlei Wen[a,b,§], Yi Zhao[c,§], Lu Fang[c], Yangbing Jin[a,b], Zehui Zhang[a,b], Siyi Zou[a,b], Fanlu Li[a,b], Ying Yang[c], Lijia Wu[c], Jiabin Jin[a,b], Xiongxiong Lu[a,b], Junjie Xie[a,b], Dongfeng Cheng[a,b], Zhiwei Xu[a,b], Jun Zhang[a,b], Jiancheng Wang[a,b], XiaXing Deng[a,b], Hao Chen[a,b], Chenghong Peng[a,b], Hongwei Li[a,b], Henghui Zhang[c,****], Hai Fang[4,***], Chaofu Wang[e,**], Baiyong Shen[a,b,*]

[a] *Department of General Surgery, Pancreatic Disease Center, Research Institute of Pancreatic Diseases, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China.*

[b] *State Key Laboratory of Oncogenes and Related Genes, National Research Center for Translational Medicine (Shanghai), Shanghai, China.*

[c] *Genecast Biotechnology Co., Ltd, Wuxi, China.*

[d] *Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine (Shanghai), Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China.*

[e] *Department of Pathology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China*

ARTICLE INFO

Article History:
Received 16 May 2021
Revised 10 November 2021
Accepted 11 November 2021
Available online xxx

Keywords:
Pancreatic adenocarcinoma
Molecular subtyping
Copy number variations
DNA repair
Receptor tyrosine kinase signaling

ABSTRACT

Background: Pancreatic adenocarcinoma (PAAD) is one of the most lethal carcinomas, and the current histopathological classifications are of limited use in clinical decision-making. There is an unmet need to identify new biomarkers for prognosis-informative molecular subtyping and ultimately for precision medicine.

Methods: We profiled genomic alterations for 608 PAAD patients in a Chinese cohort, including somatic mutations, pathogenic germline variants and copy number variations (CNV). Using the CNV information, we performed unsupervised consensus clustering of these patients, differential CNV analysis and functional/pathway enrichment analysis. Cox regression was conducted for progression-free survival analysis, the elastic net algorithm used for prognostic model construction, and rank-based gene set enrichment analysis for exploring tumor microenvironments.

Findings: Our data did not support prognostic value of point mutations in either highly mutated genes (such as *KRAS, TP53, CDKN2A* and *SMAD4*) or homologous recombination repair genes. Instead, associated with worse prognosis were amplified genes involved in DNA repair and receptor tyrosine kinase (RTK) related signalings. Motivated by this observation, we categorized patients into four molecular subtypes (namely repair-deficient, proliferation-active, repair-proficient and repair-enhanced) that differed in prognosis, and also constructed a prognostic model that can stratify patients with low or high risk of relapse. Finally, we analyzed publicly available datasets, not only reinforcing the prognostic value of our identified genes in DNA repair and RTK related signalings, but also identifying tumor microenvironment correlates with prognostic risks.

Interpretation: Together with the evidence from genomic footprint analysis, we suggest that repair-deficient and proliferation-active subtypes are better suited for DNA damage therapies, while immunotherapy is highly recommended for repair-proficient and repair-enhanced subtypes. Our results represent a significant step in molecular subtyping, diagnosis and management for PAAD patients.

[*] Corresponding author: Dr. Baiyong Shen, Department of General Surgery, Pancreatic Disease Center, Research Institute of Pancreatic Diseases, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, 197 Ruijin Er Road, Shanghai, 200025, China

[**] Corresponding author: Chaofu Wang, Department of Pathology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, 197 Ruijin Er Road, Shanghai, 200025, China

[***] Corresponding author: Hai Fang, National Research Center for Translational Medicine (Shanghai), State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, 197 Ruijin Er Road, Shanghai, 200025, China

[****] Corresponding author: Henghui Zhang, Genecast Biotechnology Co., Ltd, 88 Danshan Road, XidongChuangrong Building, Suite D-401, Xishan District, Wuxi City, Jiangsu, 214104, China

E-mail addresses: zhang.henghui@genecast.com.cn (H. Zhang), fh12355@rjh.com.cn (H. Fang), wcf11956@rjh.com.cn (C. Wang), shenby@shsmu.edu.cn (B. Shen).

[§] These authors make the same contributions to this paper.

https://doi.org/10.1016/j.ebiom.2021.103716

Funding: This work was supported by the National Natural Science Foundation of China (grant numbers 81470894, 81502695, 81672325, 81871906, 82073326, 82103482 and 32170663), the Shanghai Sailing Program (grant number 20YF1426900), and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (awarded to H.F.).

Research in context

Evidence before this study

This study includes more than 600 patients with pancreatic adenocarcinoma (PAAD) from a single hospital in China. We searched PubMed for research articles published between database inception and July 31, 2021, using the terms "molecular subtypes" or "molecular stratification" or "biomarkers" or "precision medicine" or "personalized medicine", along with "pancreatic adenocarcinoma" and "Chinese cohort". Candidate papers were manually checked to generate the reference list, from which no nation-wide effort has been reported to identify prognosis-informative molecular subtypes in China.

Added value of this study

For the first time to our knowledge, we contribute to molecular subtyping for PAAD patients in a large and Chinese cohort. This is enabled by our efforts generating the most comprehensive resource on genetic alternations for 608 PAAD patients: genetic mutations, copy number variations (CNVs) and many others. Only using the CNV information on DNA repair and receptor tyrosine kinase related genes allows us to stratify patients into four prognosis-informative subtypes, each associated with distinct survival outcomes. We also make a contribution to diagnosis for PAAD patients by reporting a prognostic model that can stratify patients with low or high risk of relapse.

Implications of all the available evidence

Our data suggest that patients in repair-deficient and proliferation-active subtypes are better suitable for DNA damage therapies, and immunotherapy highly recommended for patients in repair-proficient and repair-enhanced subtypes. Our study highlights the importance of routinely performing molecular subtyping to help manage PAAD patients.

Introduction

Pancreatic adenocarcinoma (PAAD) is one of the most aggressive and deadly cancers in China, with an estimate of 90,100 new cases and 79,400 new deaths according to the cancer statistics [1]. The high rate of mortality indicates worse prognosis over time and lacking effective systematic therapies available for patients. The first and the most urgent is to stratify patients ideally driven by molecular subtyping coupled with effective prognostic models. With patients stratified into, for example, high or low risk of relapse, the right treatment can be applied to the right patient ('precision medicine') [2], ultimately reducing mortality.

The prevailing stratification and treatment for PAAD patients are based on the mutation profiles of genes, particularly those genes involved in homologous recombination repair (HRR) pathway [3,4]. It has been reported that patients with deficiency in HRR genes have better prognosis than patients with proficiency in HRR genes, when treated with platinum-based chemotherapy or poly (ADP-ribose) polymerase inhibitor (PARPi) [5,6]. However, for resected or advanced patients unable to receive platinum-based chemotherapy, there is no value in prognosis, irrespective of HRR genes mutated or not [6]. Moreover, the overall prevalence of point mutations in HRR genes for PAAD patients is as low as 15.4% [95% Confidence Interval (CI) = 13.0%-18.0%] [7], limiting wide application of this prognostic marker.

In addition to HRR genes, other genetic alterations have also been identified from a comprehensive study involving a cohort of 3,594 PAAD patients [8]. Newly identified genetic alterations are mostly somatic mutations in *KRAS, TP53, CDKN2A* and *SMAD4*. For example, somatic mutations in *KRAS* are found in as many as 88% of patients. A majority of mutated genes, however, are not druggable. Only 4% of patients have genetic alterations occurring in druggable genes, and most of these genes are involved in receptor tyrosine kinase (RTK), RAS or MAPK signalings. Only 0.5% of patients have high tumor mutational burden (TMB) or high microsatellite instability (MSI). Therefore, there is an unmet need to seek new genomic markers that can be used to guide prognosis and treatment for PAAD patients.

Exploring tumor microenvironments is an active area of research in PAAD, as highlighted by the effort stratifying patients by cytolytic activity which can be estimated from, for example, RNA-seq transcriptome data of PAAD patients in The Cancer Genome Atlas (TCGA) cohort[9]. Patients with low cytolytic activity tend to be more instable in the genome with increased copy number alterations, including recurrent amplification of *MYC* and *NOTCH2* as well as deletion of *CDKN2A/B*[9]. On the other hand, for patients with high cytolytic activity, immune checkpoint genes (except for *PD-L1*) are highly expressed *[9]*. However, prognostic values are far from clear when patients stratified in this way, awaiting further studies.

We recognize that the high mortality of PAAD in China can be traced back to the lack of comprehensive molecular subtyping. With this end, we profile the mutational landscape of 608 PAAD patients, the largest cohort ever reported in China, generating the most comprehensive resource on genetic alternations. Genetic alterations profiled include somatic mutations, pathogenic germline variants, copy number variations (CNV) and well-known genomic markers, such as TMB, copy number instability (CNI) and somatic mutational signatures. To the best of our knowledge, we are the first to report that the poor prognosis is associated with amplification of genes involved in DNA repair and RTK related signalings. Based on this finding, we are able to stratify patients into four molecular subtypes, each associated with distinct prognosis and treatment. Then, we construct a prognostic model incorporating the information on CNV of DNA repair and RTK related genes, and apply the constructed model to distinguish patients with high or low risk of relapse. Finally, we analyze PAAD patients from a Western cohort [10] to reinforce the informativeness of CNV of DNA repair and RTK related genes in identifying molecular subtypes informative for prognosis.

Methods

Patient cohort

This study enrolled a Chinese cohort consisting of 608 patients pathologically diagnosed with PAAD from a commercial genetic testing database (Genecast Connect) and Ruijin Hospital (Shanghai, China) between March 2018 and April 2020. Informed consent form was obtained from each participant. Amongst 608 patients, 233 underwent a follow-up of ≥6 months and received radical pancreatectomy (R0) in Ruijin Hospital (**Supplementary Table 1**), thus selected for progression-free survival (PFS) analysis. The associated clinical data, including histological grade, TNM stages and adjuvant treatment data, were collected from Ruijin Hospital. Clinical stage was classified according to the 8th edition of the American Joint Committee on Cancer staging criteria. The overall study protocol (NO. 2013-70) was approved by the Medical Ethical Committee of Ruijin Hospital and the research was conducted in accordance with relevant ethical guidelines. The copy of the study protocol was provided in **Supplementary File 1**.

DNA extraction

Genomic DNA of tumor was extracted from formalin-fixed paraffin-embedded (FFPE) samples using MagPure FFPE DNA Kit B (Magen, China, ID: D6323-02). Genomic DNA of peripheral blood lymphocyte (PBL) was extracted using TGuide S32 Magnetic Blood Genomic DNA Kit (Tiangen, China, ID: DP601). The concentration of DNA was measured by Qubit dsDNA HS (High Sensitivity) Assay Kit (Thermo Fisher, USA, ID: Q32851), while the quality of DNA was assessed by Agilent 2100 BioAnalyzer (Agilent, USA). All samples for DNA extraction were obtained before the treatments started.

Library preparation

30 to 300 ng genomic DNA extracted from samples of FFPE and PBL was sheared with Covaris LE220 to the length of 200 bp with recommended settings. Then, fragmented DNA was used to construct library using KAPA Hyper Preparation Kit (Kapa Biosystems, USA, ID: KK8504) according to the manufacturer's instructions. Quantity of libraries was measured using AccuGreen High Sensitivity dsDNA Quantitation Kit (Biotium, USA, ID: Q32854) and size was determined on Agilent Bioanalyzer 2100 (Agilent, USA).

Targeted-region capture and sequencing

Targeted-region capture was performed using xGen Hybridization and wash kit box (IDT, USA, ID:1080584). Two gene panels, specifically designed for cancer gene detection by our project partner (Genecast Biotechnology Co., Ltd), include 566 and 764 genes, respectively. Panels cover frequently mutated genes in solid tumors, and gene lists were provided in **Supplementary Table 2**. For 608 patients involved in this study, 562 patients were profiled using the panel with 566 genes, and 46 patients profiled using the panel with 764 genes. Hybridization and washing were implemented according to the manufacturer's protocol. Captured libraries were sequenced on the instrument of Illumina NovaSeq 6000 according to the manufacturer's protocol, producing reads with the length of 150 bp in pairs.

Somatic mutation calling

Raw reads were first processed by Trimmomatic (v0.36) [11] to remove adaptor sequence and low-quality base. Clean reads were mapped to human genome (version hg19) by BWA aligner (v0.7.17) [12]. Mapping results were then sorted and marked for duplications via Picard (v2.23.0) [13]. SNVs and InDels as well as complex mutations were called via VarDict (v1.5.1) [14] and FreeBayes (v 1.2.0) from processed mapping results in pair of tumor and control samples, respectively. Somatic mutations appeared in genomic regions overlapped with lowly mappable regions defined by ENCODE [15] as well as low complex repeats were removed. Segmental duplications and recurrent sequence specific errors (SSEs) were also removed to promise reliable calling results. Retained somatic mutations were annotated with ANNOVAR [16]and further filtered according to these criteria: i) VAF (variant allele frequency) >= 2%, supported reads >= 6 and without strand bias; ii) annotated as nonsynonymous mutations; iii) MAF (minor allele frequency) <= 0.2% in both databases of Exome Aggregation Consortium (ExAC) [17]and Genome Aggregation Database (gnomAD) [18]. The information on the identity of somatic mutations (together with VAF) was provided in **Supplementary Table 3**.

Germline variant calling for HRR genes

Germline variants were called for 18 HRR genes (*ATM, ATR, BARD1, BLM, BRCA1, BRCA2, BRIP1, CDK12, CHEK1, CHEK2, NBN, PALB2, RAD50, RAD51B, RAD51C, RAD51D, RAD54L* and *MRE11A*) from mapping results of control samples. Firstly, only germline variants with supported reads not smaller than 15 and with VAF not smaller than 1% were retained. Secondly, pathogenicity (pathogenic and likely pathogenic) of germline variants were evaluated by CharGer [19], ClinVar [20] and manual curation-ACMG, and those labeled as "pathogenic" and "likely pathogenic" were retained. Lastly, pathologic germline variants in HRR genes were further checked by manual curation. Notably, all of patients in our cohort carry heterozygous pathogenic germline variants in DNA repair genes, with the detailed information available in **Supplementary Table 4**.

Somatic CNV identification

Somatic CNVs were identified using CNVkit (v0.9.2) [21] in the reference mode. The baseline of normalized sequencing depth on targeted regions were first constructed based on a panel of normal samples. For each tumor sample, log2 copy number ratio of normalized sequencing depth on each targeted region between the tumor sample and the baseline was then calculated. If a gene contained at least 5 targeted regions, median log2 ratio was considered as the CNV value for this gene. Gene depletion was called if the CNV value was not larger than -0.74 ($\log_2 0.6$), and gene amplification called if the CNV value was not smaller than 0.68 ($\log_2 1.6$). When log2 copy number ratio > 0, the higher positive value indicates the higher level of gene amplification. Inversely when log2 copy number ratio < 0, the lower negative value indicates the higher level of gene depletion.

Patient clustering based on CNV value

The methodology used to cluster patients was based on the concept of consensus clustering, implemented by the Consensus Cluster Plus package [22]. This package was designed to allow the optimization of parameters, including clustering algorithms (such as hierarchical, K-means and PAM), distance measures (such as Euclidean distance *versus* Pearson correlation) and the optimal number of clusters/groups. The analysis was detailed as follows. We first normalized CNV values of genes in samples by subtracting the mean and dividing the standard deviation. Based on normalized CNV, we then performed sensitivity analysis to optimize parameters mentioned above: 1) the subsampling was done both on genes and samples; 2) these subsamples were tested against different clustering algorithms, distance measures and the number of groups (from 2 to 6); and 3) item-consensus and cluster-consensus plots were drawn to evaluate the stability of clusters (**Supplementary Figure1**), showing that the use of PAM in combination with the Euclidean distance identified two groups (the optimal) of patients. Genes with significantly differential CNV value between patients of two clusters were identified through Wilcoxon rank-sum test [23], a non-parametric test that relaxes distribution assumptions and thus is more widely applicable than parameter-based tests.

CNV score calculation

Genes with significantly differential CNV value between different patients were grouped into two clusters via the same method mentioned above. Then, the first principal component (PC1) was calculated for all patients based on normalized CNV values of genes in each cluster, respectively. Next, univariate Cox regression was conducted to calculate the hazard ratio (HR) for PC1 per cluster. Following the approach [24], we constructed the following fomula to calculate the CNV score:

$$\text{CNV score} = \sum^{\text{PC}} 1_{\text{i}} - -\sum^{\text{PC}} 1_{\text{j}}$$

where *i* represented the gene cluster with HR larger than 1, and *j* for the cluster with HR smaller than 1. Codes for CNV score calculation are made available at https://github.com/corefacilitygenecast/FACTORscore. Based on the CNV score, patients were stratified into two groups with high or low CNV score, using the optimal cutoff (4.5) determined by the maxstat package[25] to maximize the separation of these two groups. The maxstat package implements a significance test on standardized maximally selected rank statistic of CNV scores under the null hypothesis that any cutoff has no influence on the distribution of survival time. Also, genes with significantly differential CNV value between patients in different groups were identified through Wilcoxon rank-sum test [23].

Functional and pathway enrichment analysis

Enrichment analysis was conducted for genes with significantly higher CNV value in each cluster or group, through Fisher's exact test implemented by the cluster Profiler package [26] using functional annotations of Gene Ontology (GO) [27,28], and pathway resources obtained from Kyoto Encyclopedia of Genes and Genomes (KEGG) [29] as well as Reactome [30].

Risk score calculation

As mentioned above, in our cohorts there were 233 patients with available prognostic information. These patients were first randomly partitioned into training and validation sets according to the ratio of 2:1, resulting in 155 patients in the training set and 78 patients in the validation set. Such random allocation was not stratified on any clinical variables. Based on patients in the training set, a prognostic model was constructed through the algorithm of elastic net implemented by the glmnet package [31,32], taking as inputs normalized CNV values of 73 genes involved in DNA repair and RTK related signalings as well as in the HRR pathway (**Supplementary Table 5**). More specifically, these genes, as an initial gene set, were used to fit a regularized Cox model for survival times (PFS), where the elastic net penalty was adopted to maximize the partial likelihood of coefficients of selected genes. After the process of leave-one-out cross-validation, 5 genes closely associated with PFS were selected and tuned for their coefficients. The resulting prognostic model was shown below (Equation 1):

Risk score = -0.1136**RAD50* + 0.6140**AKT1* + 0.5643**CSF1R* - 0.3089**JAK2* -0.2088**ABL1* where each gene name represented its CNV value of patients. Then, risk score for each paitent was calculated according to the constructed prognostic Cox model with CNV values of 5 genes composing it. The optimal cutoff (-0.0958) of risk score was determined via the maxstat package [25] to stratify patients into two groups with high or low risk of relapse, while prognosis of patients in these two groups was compared. Lastly, the same prognostic and cutoff were applied in the validation set.

Risk score calculation for the TCGA cohort

Datasets of genomic alteration and clinical information of PAAD patients collected in TCGA were downloaded from the website of cBioPortal for cancer genomics (https://www.cbioportal.org/study/summary?id=paad_tcga_pan_can_atlas_2018). For PADD, 182 patients had both CNV values and prognostic information available. We then conducted the downstream analysis for these 182 PAAD patients in the TCGA cohort. We constructed a prognostic model for patients of the TCGA cohort with the same strategy and method as described above (Equation 2):

Risk score = 0.1707*PALB2 - 0.8363*RAD51C + 0.1818*FGF3 - 0.1096*NF1 + 1.2076e-05*FGF4 + 1.9914*PIK3CA - 0.1841*MAPKAP1 - 0.1357*RICTOR

where each gene name represented its CNV value of patients. Following that, risk score was calculated for each patient based on this prognostic model. The optimal cutoff (0.2833) was determined by the maxstat package [25] to stratify patients into two groups and the prognosis was compared between them.

Tumor microenvironment (TME) analysis for TCGA cohort

We obtained transcriptional signatures associated with infiltration of immune cells [33], that is, marker gene sets signifying immune cells of different types (**Supplementary Table 6**). The abundance of each immune cell type in TME for each patient in the TCGA cohort was estimated through the algorithm of single sample Gene Set Enrichment Analysis (ssGSEA) [34]. This algorithm calculates an enrichment score that quantifies the degree of absolute enrichment of a cell-type-specific gene set in each patient. Additionally, enrichment score of transcriptional signatures associated with anti-PD-1 resistance (IPRES) [35] was also calculated for each patient using Gene Set Variation Analysis (GSVA) [36]. Both analyses were conducted based on RNA-seq data of patients in the TCGA cohort downloaded from the website of cBioPortal for cancer genomics, and the analysis methods are the same as above. Significantly differential transcriptional signatures between patients in groups with high- or low-risk score were identified through Wilcoxon rank sum test.

Statistical analysis and visualization

Prognosis of patients in different clusters or groups was analyzed and compared via Kaplan-Meier estimate and log-rank test [37] implemented by the survival package [38,39], with results visualized by the survminer package. Effects on PFS of other independent variables were evaluated via multivariate Cox regression analysis implemented in the survival package [38,39]. ROC curves together with AUC values were calculated using the pROCpackage [40]. Comparison of values between two distributions was conducted via Wilcoxon rank sum test, while multiple testing correction was done using false discovery rate (FDR) method. Landscape of genomic alterations of PAAD patients in the Chinese cohort was plotted through the ComplexHeatmap package [41].

Statistics

Wilcoxon rank-sum test was used to compare the distributions of continuous values between two groups, which is a non-parametric test that relaxes distribution assumptions and widely applicable in statistical analysis. Fisher's exact test was used to examine the significance of the association between the two kinds of classification of patients, which is much more robust for the case of small sample size in any cell of the contingency table. Similarly, pathway and functional enrichment analysis was implemented via Fisher's exact test. Additionally, log-rank test was used to compare the survival distributions of two kinds of stratification of patients, and Wald test used to test whether the beta coefficient of a given variable is significantly different from zero in the multivariate Cox regression analysis; both are conventional in prognosis analyses. In total, 608 PAAD patients are involved in this study, and this cohort is enough to promise the power for statistical analysis.

Role of funders

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Results

Mutational landscape of PAAD patients

We carried out a cohort study involving 608 PAAD patients in China and, for each patient, profiled genetic alterations including somatic mutations, pathogenic germline variants and CNV along with several genomic markers, such as TMB, CNI and somatic mutational signatures. From this Chinese cohort we identified a list of frequently mutated genes, with the top 30 illustrated in **Figure 1A**. Consistent with the previous findings based on two larger cohorts of PAAD patients (collected out of China) [8,42], the highest mutated genes included *KRAS* (518 out of 608 patients, 85%), *TP53* (63%), *SMAD4* (20%) and *CDKN2A* (18%). For the gene *KRAS*, the somatic mutations tended to occur in G12D (233 out of 608 patients, 38%), G12V (29%) and G12R (11%); this mutation pattern was similar to the observation based on the TCGA cohort [42]. More interestingly, we found that 73 out of 608 patients (12%) in our cohort carried point mutations in one or more genes involved in the HRR pathway (**Table 1**), including *ATM, ATR, BARD1, BLM, BRCA1/2, BRIP1, CDK12, CHEK1/2, NBN, PALB2, RAD50/51B/51C/51D/54L* and *MRE11A*. When considering the HRR genes as a whole, aggregated carrier ratio (12.01%) was similar to that in the Caris cohort [7].

Next, we explored prognostic values of mutated genes identified above, focusing on a subset of PAAD patients (233 out of 608) treated with R0 resection and containing prognostic information (≥6-month follow-up; **Supplementary Table 1**). Firstly, we found no significant difference in prognosis when stratifying patients by mutated status of *KRAS, TP53, SMAD4* or *CDKN2A* (**Figure 1B-E**). These four genes were mutated prevalently in PAAD patients, and our finding did not support their use in prognosis. When examining specific mutations within *KRAS*, we also found no significant association with the prognosis; for example, the median survival time (310 days) of patients with G12D *versus* that (406 days) of patients with wild type [**Figure 1F**; hazard ratio (HR) = 1.6, 95% CI = 0.85-2.9, *P* = 0.15 on Wald test]. Secondly, we found no prognostic value of HRR genes (**Figure 1G** and **Supplementary Figure 2**), consistent with the previous findings from cohorts collected in Know Your Tumor program [6]. Thirdly, we also observed no difference in prognosis when patients were stratified into two groups according to mutated status of all genes (**Supplementary Figure 3**).

In addition to prognostic values, we also explored therapeutic values for genes with genetic alternations. Firstly, we found that 42 out of 608 patients (6.9%) had clinically actionable genetic alterations that might be targeted with existing drugs, including gene amplifications (n = 9), gene deletions (n = 4) and somatic mutations (n = 29). Secondly, we found that most of genetic alterations occurred within druggable candidate genes, such as genes involved in the RTK/RAS signaling (*EGFR, ERBB2, MET* and *BRAF*) and the PI3K pathway (*PTEN, PIK3CA* and *AKT1*); this is similar to a previous report based on real-time targeted genome profile analysis [8]. Third, we found that 7 patients (1.2%) had somatic mutations in genes involved in the mismatch repair (MMR) pathway (*MLH1* and *MSH2*), indicating potential benefits on immunotherapy [43].

Amplification of DNA repair genes confers worse prognosis in PAAD patients

The above analysis suggested that in terms of prognostic use, a limited number of mutated genes might be not informative for such complex cancers as PAAD. This motivated us to consider other genomic information, particularly CNV (quantified as logarithmically transformed ratio of normalized read depth on genes between tumor samples and normal samples; see **Methods**). We classified 608 PAAD patients into two groups according to CNV (**Figure 2A**): the first group (CNV-G1 with 321 patients) *versus* the second group (CNV-G2 with 287 patients). Patients in CNV-G2 had a worse prognosis than patients in CNV-G1 (HR = 2.0, 95% CI = 1.3-3.2, $P = 1.7 \times 10^{-3}$ on Log-rank test), with median survival time (410 days) for CNV-G1 *versus* that (239 days) for CNV-G2 (**Figure 2B**).

To uncover genomic factors influencing the survival of PAAD patients, we performed differential CNV analysis using Wilcoxon rank sum test, identifying 155 genes with higher CNV in CNV-G1 than in CNV-G2, and 215 genes with higher CNV in CNV-G2 under the cutoff of FDR < 0.05. Through enrichment analysis using Reactome pathways[30], we found that 26 genes with significantly higher CNV value in CNV-G2 patients were enriched in DNA repair (**Figure 2C**; odds ratio (OR) = 3.7, FDR = 5.4×10^{-4} on Fisher's exact test), including 11 genes involved in the HRR pathway (*ATM, ATR, BLM, BRCA1/2, BRIP1, CHEK2, PALB2* and *RAD51B/51C/51D*). This finding was visually confirmed by the CNV landscape (**Figure 2D**), in which CNV-G2 patients possessed more amplified DNA repair genes, especially those in the HRR pathway. Contrary to that, CNV-G1 patients possessed more depleted DNA repair genes, with 11.5% of these patients (37 out of 321) having depleted *TP53* that functions in modulating the DNA damage repair pathway[44]. In summary, as compared to CNV-G1, CNV-G2 patients seemed to obtain an enhanced function of DNA repair, likely HRR. For convenience, we renamed CNV-G2 as 'repair-proficient', and CNV-G1 as 'repair-deficient'. Such relabeling was also justified by our observation that repair-deficient patients (43 out of 321 patients, 13%) had more point mutations in genes involved in the HRR pathway than repair-proficient patients (30 out of 287 patients, 10%), though no significance reached (OR = 1.32, *P* = 0.32 on Fisher's exact test).

Next, we explored genomic signatures that could signify the difference between repair-deficient patients and repair-proficient patients. We found that CNI of repair-deficient patients was significantly higher than that of repair-proficient patients ($P = 9.1 \times 10^{-11}$ on Wilcoxon rank sum test; **Figure 2E**), whereas TMB of repair-proficient patients was significantly higher than that of repair-deficient patients ($P = 2.6 \times 10^{-2}$ on Wilcoxon rank sum test; **Figure 2E**), likely reflecting the involvement of different oncogenic mechanisms in these two groups of patients. Consistent with these observations, we also found distinct patterns of somatic mutational signatures that were associated with each of these two groups (**Figure 2F**). Mutational signatures for each patient were systematically identified based on the point somatic mutation profiles of patients in our cohort. Comparing the abundance distribution of mutational signatures between repair-proficient and repair-deficient patients (Wilcoxon rank sum test), we found that mutational signatures of defects in DNA-DSB repair by homologous recombination (FDR = 2.7×10^{-38}) as well as defective DNA mismatch repair (FDR = 1.5×10^{-33}) were significantly higher in repair-deficient patients (in other words, prevalent in repair-deficient patients). On the contrary, mutational signatures of defects in polymerase *POLE* (FDR = 1.9×10^{-108}) were prevalent in repair-proficient patients, consistent with higher TMB observed in these patients shown in **Figure 2E**. Among 287 repair-proficient patients, 4 had TMB larger than 10, which can be considered as the hypermutation phenotype associated with loss of *POLE*. More intriguingly, mutational signatures of exposure to tobacco (smoking) mutagens were observed only in

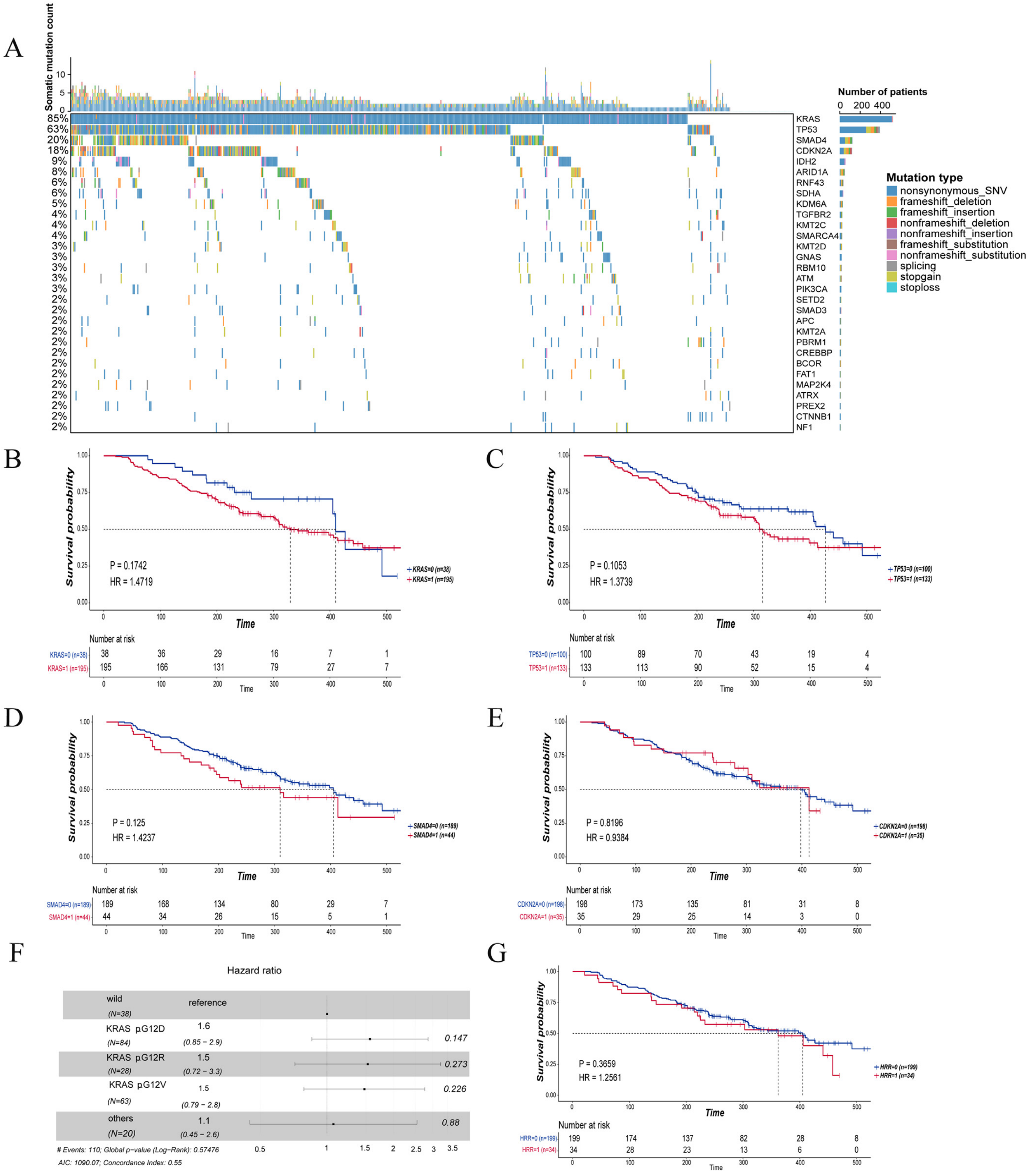

Figure 1. Genomic landscape of PAAD patients and prognosis analysis based on mutational status of selected genes. (A) The somatic mutation landscape of PAAD patients. Top 30 genes with the highest frequency were plotted, with types of somatic mutations color-coded. **(B-E)** Kaplan-Meier survival curves for patients with *KRAS*, *TP53*, *SMAD4* or *CDKN2A* mutated (denoted as 1) or not (0). **(F)** The forest plot showed the result of multivariate Cox regression. Multivariate Cox regression determined the correlation between different types of somatic mutations in *KRAS* (including G12 D, G12R, G12V and others) and prognosis. Hazard ratio and *P*value ranked in the second and fourth column, respectively. Horizontal lines represent the 95% confidence interval. **(G)** Kaplan-Meier survival curve for patients with any HRR genes mutated (denoted as 1) or not (0).

Table 1
Carrier ratio of pathogenic germline variants and somatic mutations in genes of HRR pathway in PAAD patients.

Gene	Germline variant (%)	Somatic mutation (%)	All (%)
ATM	1.15	3.45	4.44
ATR	0.33	1.48	1.81
BARD1	0.16	1.15	1.32
BLM	0.00	0.16	0.16
BRCA1	0.00	0.82	0.82
BRCA2	0.33	1.32	1.65
BRIP1	0.16	0.49	0.66
CDK12	0.00	0.99	0.99
CHEK1	0.00	0.33	0.33
CHEK2	0.16	0.49	0.66
MRE11A	0.00	0.49	0.49
NBN	0.00	0.33	0.33
PALB2	0.49	0.33	0.66
RAD50	0.16	1.15	1.32
RAD51B	0.00	0.16	0.16
RAD51C	0.00	0.66	0.66
RAD51D	0.00	0.16	0.16
RAD54L	0.33	0.99	1.32
ALL	3.29	9.21	12.01

repair-deficient patients (**Supplementary Figure 4A**), and mutational signatures of exposure to alkylating agents observed only in repair-proficient patients (**Supplementary Figure 4B**). Noteworthily, the latter has also been found in melanoma, the cancer with high TMB in patients from Western countries [45].

Amplification of RTK related genes is associated with worse prognosis in PAAD patients

To further identify sub-clusters from repair-deficient patients (and/or repair-proficient patients, though limited by the numbers available), we first calculated CNV scores for each of 608 patients, and then used such information for prognostic analysis in terms of PFS as the clinical outcome endpoint (see **Methods**). We partitioned patients into two subgroups based on an optimal cutoff[25]: one subgroup with high CNV score, and the other subgroup with low CNV score. Patients with high CNV score were associated with worse prognosis, including 50 repair-deficient patients and 6 repair-proficient patients. As expected, repair-deficient patients tended to have higher CNV score than repair-proficient patients (OR = 8.6, $P = 1.7 \times 10^{-9}$ on Fisher's exact test). Regarding the repair-deficient group, prognostic analysis showed that worse prognosis was significantly associated with the subgroup with high CNV score compared to the subgroup with low CNV score (HR = 2.2, 95% CI = 1.3-3.8, $P = 1.8 \times 10^{-3}$ on Log-rank test;**Figure 3A**). Regarding the repair-proficient group, though the subgroup with high CNV score contained only one patient with available prognostic information, we noted that this patient had PFS as short as 48 days.

Comparing repair-deficient patients with high or low CNV score, we identified 203 genes with differential CNV (FDR $<$ 0.05 on Wilcoxon rank sum test). Genes with higher CNV (in the patient subgroup with higher CNV score) were largely involved in the RTK related signalings (**Figure 3B**), including RTK signaling (OR = 3.32, FDR = 1.7×10^{-2} on Fisher's exact test), Ras signaling (OR = 4.83, FDR = 5.9×10^{-4} on Fisher's exact test), Rap1 signaling (OR = 3.25, FDR = 1.4×10^{-2} on Fisher's exact test) and MAPK signaling (OR = 2.96, FDR = 1.6×10^{-2} on Fisher's exact test). In addition, we found that patients with higher CNV score possessed significantly higher CNI than those with low CNV score ($P = 6 \times 10^{-3}$ on Wilcoxon rank sum test; **Supplementary Figure 5**). One of possible explanations why patients with high CNV score had worse prognosis was that the genome with high CNV tended to be instable, likely inducing extensive amplification of, for example, RTK related genes.

Comparing repair-proficient patients with high or low CNV score, we identified 95 genes with differential CNV (FDR $<$ 0.05 on Wilcoxon rank sum test). Genes with higher CNV (in the patient subgroup with higher CNV score) were largely involved in HRR (**Figure 3C**), including homology directed repair (OR = 17.23, FDR = 8.5×10^{-5} on Fisher's exact test), homologous DNA pairing and strand exchange (OR = 16.00, FDR = 4.2×10^{-4} on Fisher's exact test) and homology directed repair through homologous recombination (OR = 13.05, FDR = 7.8×10^{-4} on Fisher's exact test). This functional enrichment pattern implied that further amplification of HRR genes in patients with high CNV likely enhanced the self-repair ability of cancer genome, ultimately resulting in worse prognosis.

Identification of molecular subtypes improves prognosis in PAAD patients

Collectively considering the information obtained from unsupervised clustering and CNV-based stratification, we were able to categorize PAAD patients into four molecular subtypes (namely *repair-deficient, proliferation-active, repair-proficient* and *repair-enhanced*). More specifically, we subdivided patients from the repair-deficient group into two subtypes: the repair-deficient subtype with low CNV score, and the proliferation-active subtype with high CNV score. Similarly, we subdivided patients from the repair-proficient group into two subtypes: the repair-proficient subtype with low CNV score, and the repair-enhanced with high CNV score (notably lacking the survival information for prognostic analysis). Such categorization was largely driven by the CNV information of genes involved in DNA repair and RTK related signalings. We further showed that identified subtypes were informative in prognosis. Repair-deficient patients had the best prognosis with median survival time of 410 days, whereas proliferation-active and repair-proficient patients had worse prognosis with median survival times of 197 and 239 days, respectively (HR = 2.2, 95% CI = 1.4-3.6, $P = 2 \times 10^{-4}$ on Log-rank test; **Figure 3D**). We also confirmed our findings using multivariate Cox regression analysis (**Supplementary Figure 6**).

To evaluate the power of using the CNV information to predict relapse in a range of follow-up windows (six months, twelve months and median survival time), we compared receiver operating characteristic (ROC) curves for patients with repair-deficient and proliferation-active subtypes (**Supplementary Figure 7A**) as well as with repair-proficient and repair-enhanced subtypes (**Supplementary Figure 7B**). Patients with higher CNV score experienced earlier relapse, with predictive power achieved acceptably.

Validation of molecular subtypes using TCGA-PAAD datasets

Using TCGA-PAAD datasets, we performed three levels of validations on molecular subtypes identified from our Chinese-PAAD datasets. Firstly, we validated CNV-G1 (repair-deficient) and CNV-G2 (repair-proficient). Based on our Chinese PAAD cohort, standardized shrunken centroids of CNV value were calculated for two groups of CNV-G1 (repair-deficient) and CNV-G2 (repair-proficient). Using the method of nearest shrunken centroids [46] each patient from the TCGA-PAAD cohort was assigned to either of these two groups. For example, a patient will be assigned to the CNV-G1 (repair-deficient) group if this patient has the CNV profile closest squared distance to the centroid of the CNV-G1 group. As such, we stratified TCGA-PAAD patients into CNV-G1 (repair-deficient) and CNV-G2 (repair-proficient). When comparing CNV values of genes between these two groups, we found 7 genes (*ATM, BARD1, BRCA1, CDK12, RAD51B/51D* and *MRE11A*) involved in the HRR pathway possessed significantly lower CNV values in CNV-G1 than in CNV-G2, which is consistent with results obtained from our Chinese-PAAD datasets.

Secondly, we validated CNV score-driven subclusters of CNV-G1 (repair-deficient). According to the eigenvectors of two PC1s

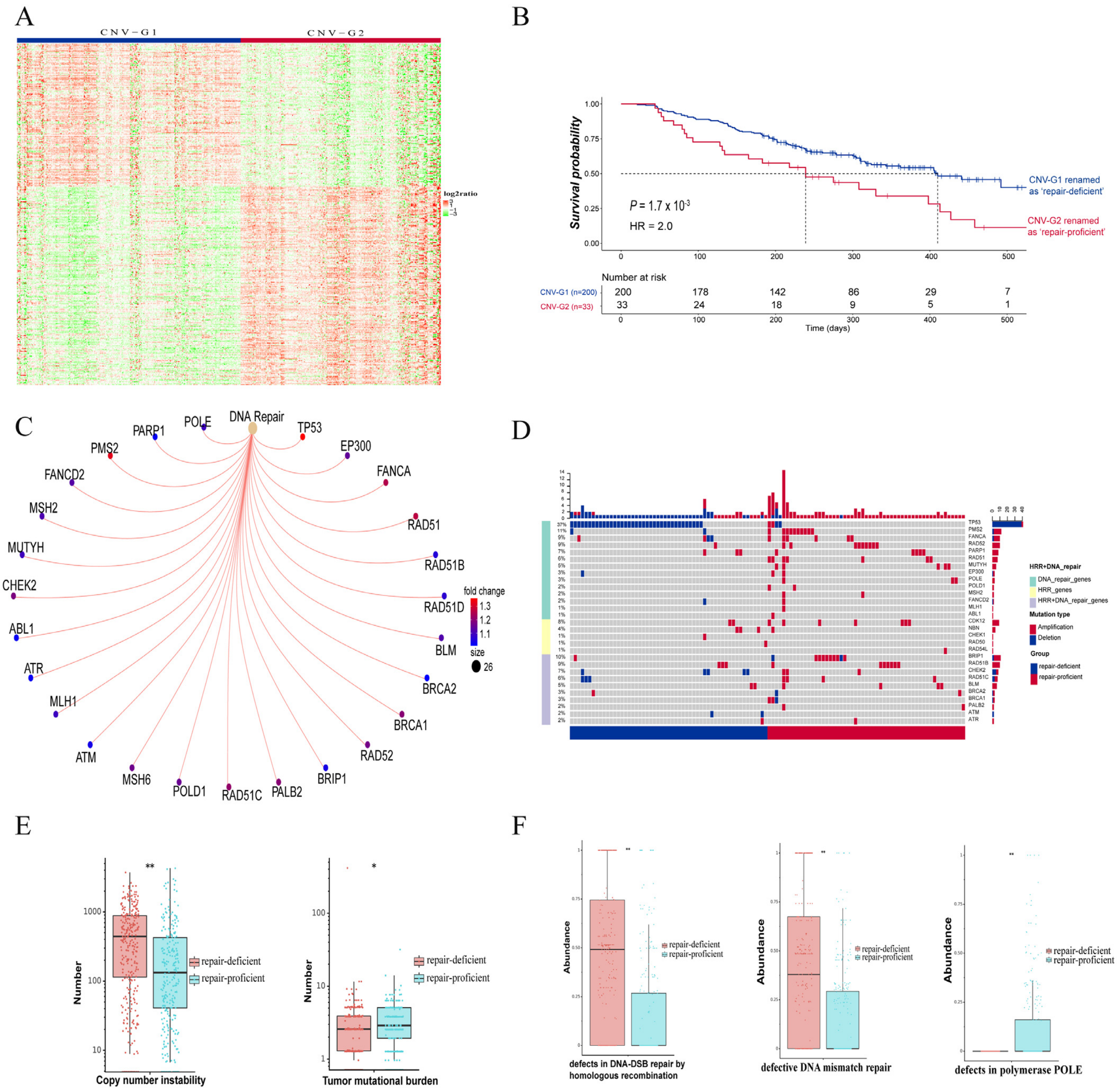

Figure 2. Amplification of HRR genes associated with worse prognosis. (A) Unsupervised clustering based on the CNV information identifying two groups of patients, visualized as a heatmap with columns for patients and rows for genes. **(B)** Kaplan-Meier survival curves for patients of two groups, with patients of the repair-proficient (CNV-G2) group having worse prognosis compared to patients of the repair-deficient (CNV-G1) group. **(C)** Genes with significantly higher CNV in patients of the repair-proficient (CNV-G2) group were enriched in the function of DNA repair, including 11 genes involved in the HRR pathway. **(D)** CNV landscape of patients of repair-deficient (CNV-G1) and repair-proficient (CNV-G2) groups. Red denotes gene amplification, and blue for gene deletion. **(E)** Boxplots for CNI and TMB distribution of patients in repair-deficient (CNV-G1) and repair-proficient (CNV-G2) groups, showing that CNIs of patients in the repair-deficient group was significantly higher than those of patients in the repair-proficient group, whereas TMBs of patients in the repair-proficient group was significantly higher than those of patients in the repair-deficient group. **(F)** Boxplots for contributions of somatic signatures of patients in repair-deficient (CNV-G1) and repair-proficient (CNV-G2) groups, showing that somatic signatures of defects in DNA-DSB repair by homologous recombination and defective DNA mismatch repair enriched in patients of the repair-deficient group, whereas somatic signature of defects in polymerase *POLE* enriched in patients of the repair-proficient group.

extracted from the Chinese-PAAD datasets, we calculated the value for each of them based on CNV values measured in the TCGA-PAAD datasets. After prognostic analysis, we observed that both of these two PC1s were associated with worse prognosis. This observation motivated us to calculate CNV score for each patient in the TCGA-PAAD cohort (similar to the 'CNV score calculation' of **Methods**), obtaining two subclusters of CNV-G1 (repair-deficient). When compared CNV values of genes between two subclusters of CNV-G1 (repair-deficient), we found 6 genes (*ALK*, *CBL*, *ERBB4*, *ERRFI1*, *FGFR1* and *ROS1*) involved in RTK related signalings possessed significantly higher CNV value in the subcluster with high CNV score than in the subcluster with low CNV score, which is in line with results obtained from the Chinese-PAAD datasets.

Thirdly, we validated molecular subtypes. Collectively considering the information obtained from unsupervised clustering and CNV score-driven subclusters, we categorized patients in the TCGA-PAAD

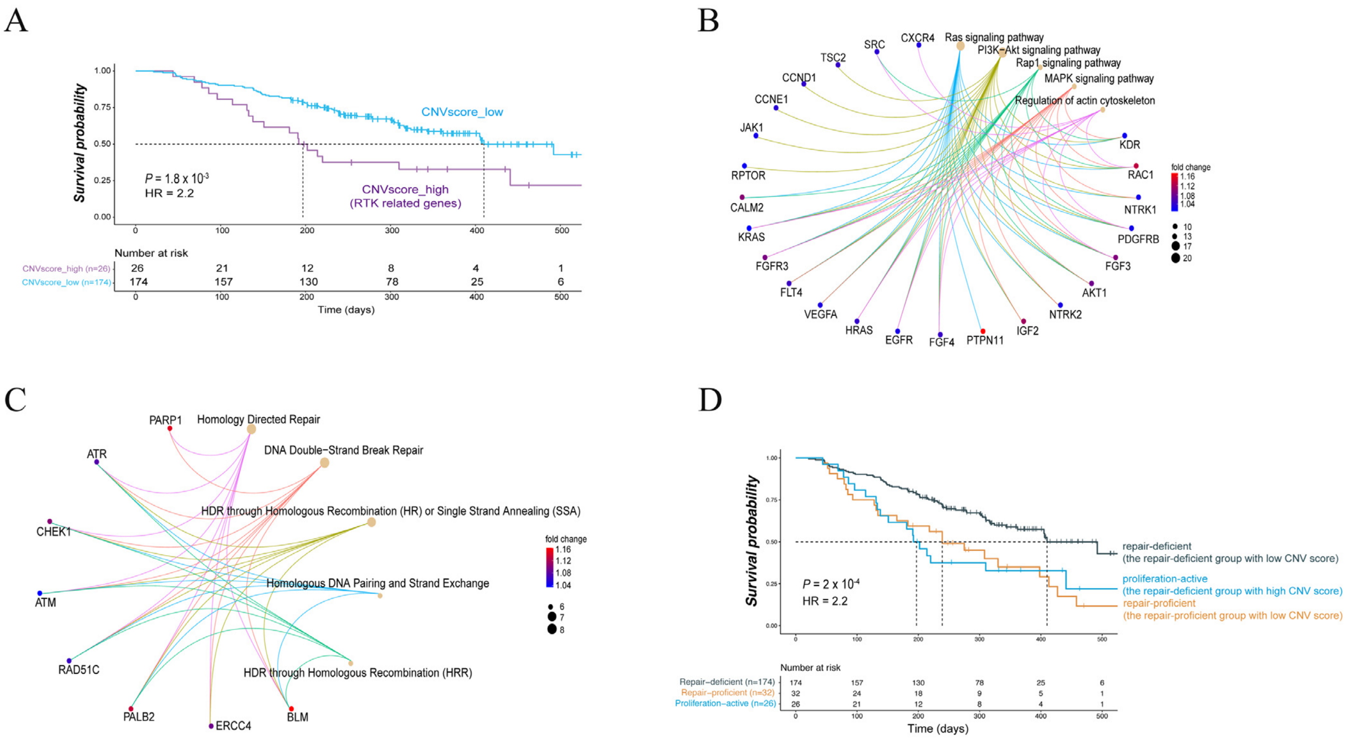

Figure 3. Amplification of RTK-related genes associated with worse prognosis. (A) Kaplan-Meier survival curves for patients in the repair-deficient group, showing that patients in the subgroup with high CNV score had worse prognosis compared to patients in the subgroup with low CNV score. **(B)** Genes with significantly higher CNV in patients of the repair-deficient group with high CNV score were mostly RTK signaling related. **(C)** Genes with significantly higher CNV in patients of the repair-proficient group with high CNV score were enriched in the function of homologous recombination repair. **(D)** Kaplan-Meier survival curves for patients of three molecular subtypes, including repair-deficient, proliferation-active and repair-proficient.

cohort into three molecular subtypes, including repair-deficient (n=18, derived from the repair-deficient group with low CNV score), (2) proliferation-active (n=121, the repair-deficient group with high CNV score), and (3) repair-enhanced (n=44, the repair-proficient group with high CNV score). Patients of the repair-deficient subtype showed the best prognosis (50% of patients above are still alive at last point; **Supplementary Figure 8**) compared to ones in proliferation-active (median survival time is 599 days) and repair-enhanced (median survival time is 538 days) subtypes, which is consistent with results obtained from the Chinese-PAAD datasets. Interestingly, much more patients were classified as repair-enhanced subtype in the TCGA-PAAD cohort than in the Chinese-PAAD cohort (OR = 31.83, $P< 2.2 \times 10^{-16}$ on Fisher's exact test).

Construction of an effective prognostic model for PAAD patients

We proceeded to explore how to utilize the CNV information for precision medicine in the field of PAAD. With this aim, we attempted to construct a prognostic model that may be clinically actionable. We selected a total of 73 genes that are mainly involved in DNA repair and RTK related signalings, and these genes were used as the initial gene set for model construction. Using the elastic net algorithm [31,32] and dividing patients into training and validation sets (see **Methods**), we first constructed a prognostic model from the training set. The constructed model consisted of *RAD50* (involved in the HRR pathway), *ABL1* (DNA repair), and 3 RTK related genes (*JAK2, AKT1* and *CSF1R*). The CNV distribution for these 5 genes was illustrated in **Figure 4A**. Afterwards, we calculated risk score for each patient based on the constructed model, and stratified patients into two groups with high- or low-risk score maximizing the PFS-based rank statistics [25]. For the training set, patients with high-risk score (119 out of 155 patients, 77%) had significantly worse prognosis than those with low-risk score (36 patients, 23%; HR = 5.9, 95% CI = 2.36-14.61, $P = 1.6\times 10^{-5}$ on Log-rank test; **Supplementary Figure 9**). This model also performed well for patients in the validation set that were not considered during the model construction (HR = 2.6, 95% CI = 1.02-6.69, $P = 3.8 \times 10^{-2}$ on Log-rank test;**Figure 4B**). Using multivariate Cox regression model, we observed similar results (**Supplementary Figure 10**). Thus, this prognostic model can be of potential use aiding in clinical decision-making to identify patients with high-risk relapse for receiving the right treatment and management.

As expected, most of patients with low-risk score were grouped into the repair-deficient subtype which had the best prognosis (50 out of 56, 89%; OR = 3.6, $P = 1.5 \times 10^{-13}$ on Fisher's exact test; **Figure 4C**). In addition, patients of the same subtype may be classified as low- or high-risk score patients, thereby revealing the complex relationship between molecular subtypes and prognosis. We also calculated ROC curves of risk score to evaluate performance of this prognostic model in predicting relapse of patients (**Figure 4D**). Area under curves (AUCs) were 0.64 (95% CI = 0.56-0.72), 0.65 (95% CI = 0.58-0.72) and 0.73 (95% CI = 0.66-0.79) for relapse within six months, twelve months and median survival time, respectively. Thus, our definition of higher risk score indeed can be used to signify earlier relapse, with better predictive power than shown in **Supplementary Figure 7**.

Exploring tumor microenvironments of PAAD patients that correlate with prognostic risk using publicly available datasets

The previous study [47] has demonstrated that CNV burden and homologous recombination deficiency, as well as somatic alterations of RTK related genes, have effects on immune microenvironments in various cancer types. Thus, we proceeded to provide the clue showing that tumor microenvironments differed between PAAD patients with high- or low-risk prognostic score; doing so exclusively based on CNV of genes involved in DNA repair and RTK related signalings.

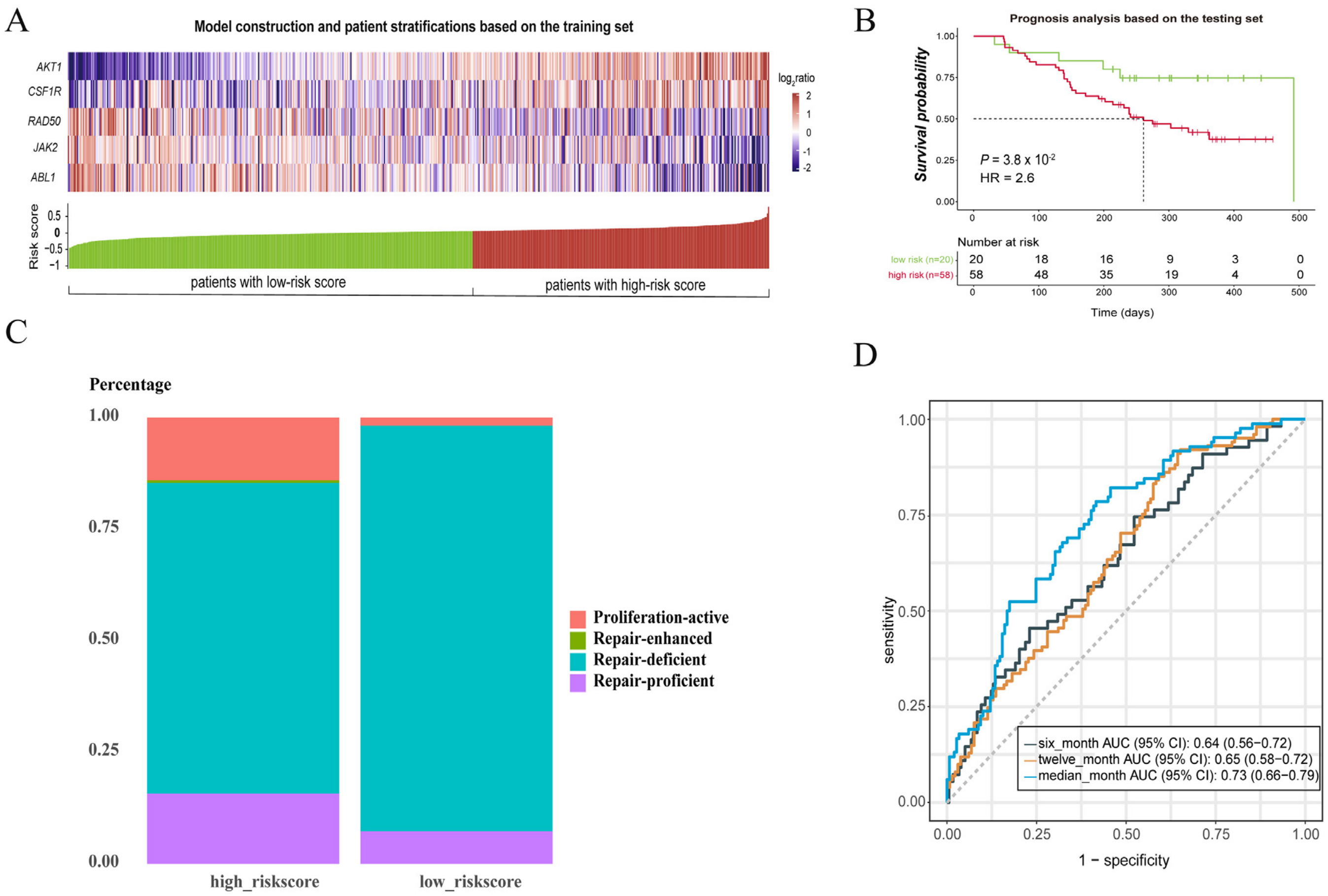

Figure 4. Prognostic model for PAAD patients. (A) CNV heatmap of 5 genes, which were selected by the algorithm of elastic net to construct the prognostic model from the training set. Genes in rows were sorted according to the direction of coefficients, while patients in columns were sorted according to the risk score, with green denoting patients with low risk of relapse and red for patients with high risk of relapse. (B) Kaplan-Meier survival curves for patients with low and high risk of relapse in the independent testing set. (C) Distribution of four molecular subtypes of patients over groups with low and high risk of relapse. (D) ROC curves of risk score of patients to evaluate its performance in predicting relapse status within six months, twelve months and median survival time.

We first constructed prognostic model using the same initial set of 73 genes (as justified and used in the previous subsections) and the same method applied to 182 PAAD patients collected in the TCGA cohort where both CNV values and prognostic information are available. Patients with high-risk score (132 out of 182 patients, 73%) had significantly worse prognosis than those with low-risk score (50 patients, 27%; HR = 3.8, 95% CI = 2.2-6.7, $P = 9.6 \times 10^{-7}$ on Log-rank test; **Supplementary Figure 11**).

Finally, we compared the immune microenvironments between two groups of patients with high- or low-risk score, using TCGA RNA-seq datasets and transcriptional signatures associated with immune cell infiltration[33] and innate anti-PD-1 resistance (IPRES) [35]. As shown in **Figure 5A**, we found that immune cells infiltrated into the tumor of low-risk patients were more likely to be immune cells with antitumor capability, such as effector memory CD4 T cell [48,49] (FDR = 1.4×10^{-2} on Wilcoxon rank sum test), monocyte [50] (FDR = 1.4×10^{-2} on Wilcoxon rank sum test) and eosinophil [51] (FDR = 4.4×10^{-2} on Wilcoxon rank sum test). In low-risk patients, we also observed the suggestively significant enrichment of immune cells responsible for killing tumor cells, including CD56 bright natural killer cell [52,53], activated CD8 T cell [54] and T follicular helper cell [55]. On the contrary, we found that transcriptional signatures associated with IPRES were enriched in high-risk patients (**Figure 5B**), including the up-regulation of carcinoma associated fibroblast (FDR = 1.8×10^{-2} on Wilcoxon rank sum test) and MAPK inhibitor induced epithelial mesenchymal transition (FDR = 2.1×10^{-2} on Wilcoxon rank sum test). Taken together, analysis of the TCGA datasets supported that the tumor microenvironment features of PAAD patients can be relevant to their prognostic risks.

Discussion

PAAD is one of the most lethal carcinomas in China with high incidence as well as mortality [1]. In this study, we attempt to define molecular subtypes and develop prognostic model for PAAD patients to assist in selection of the most appropriate treatment by comprehensively profiling the mutational landscape of PAAD patients in a Chinese cohort. Interestingly, we found that amplification of genes in DNA repair and RTK related signalings was associated with worse prognosis. Motivated by this finding, we further used CNV of DNA repair and RTK related genes to categorize our PAAD patients into four molecular subtypes (including repair-deficient, proliferation-active, repair-proficient and repair-enhanced), with the repair-deficient subtype having the best prognosis and the worst prognosis observed for the repair-enhanced subtype. These molecular subtypes identified from our Chinese-PAAD cohort were validated using the TCGA-PAAD cohort. Furthermore, we used DNA repair and RTK related genes as the initial gene set to construct a clinically usable prognostic model built on our Chinese cohort, which performed well in discriminating high-risk PAAD patients from low-risk ones. We have also used the TCGA-PAAD cohort to illustrate the informativeness of CNV of DNA repair and RTK related genes in prognosis,

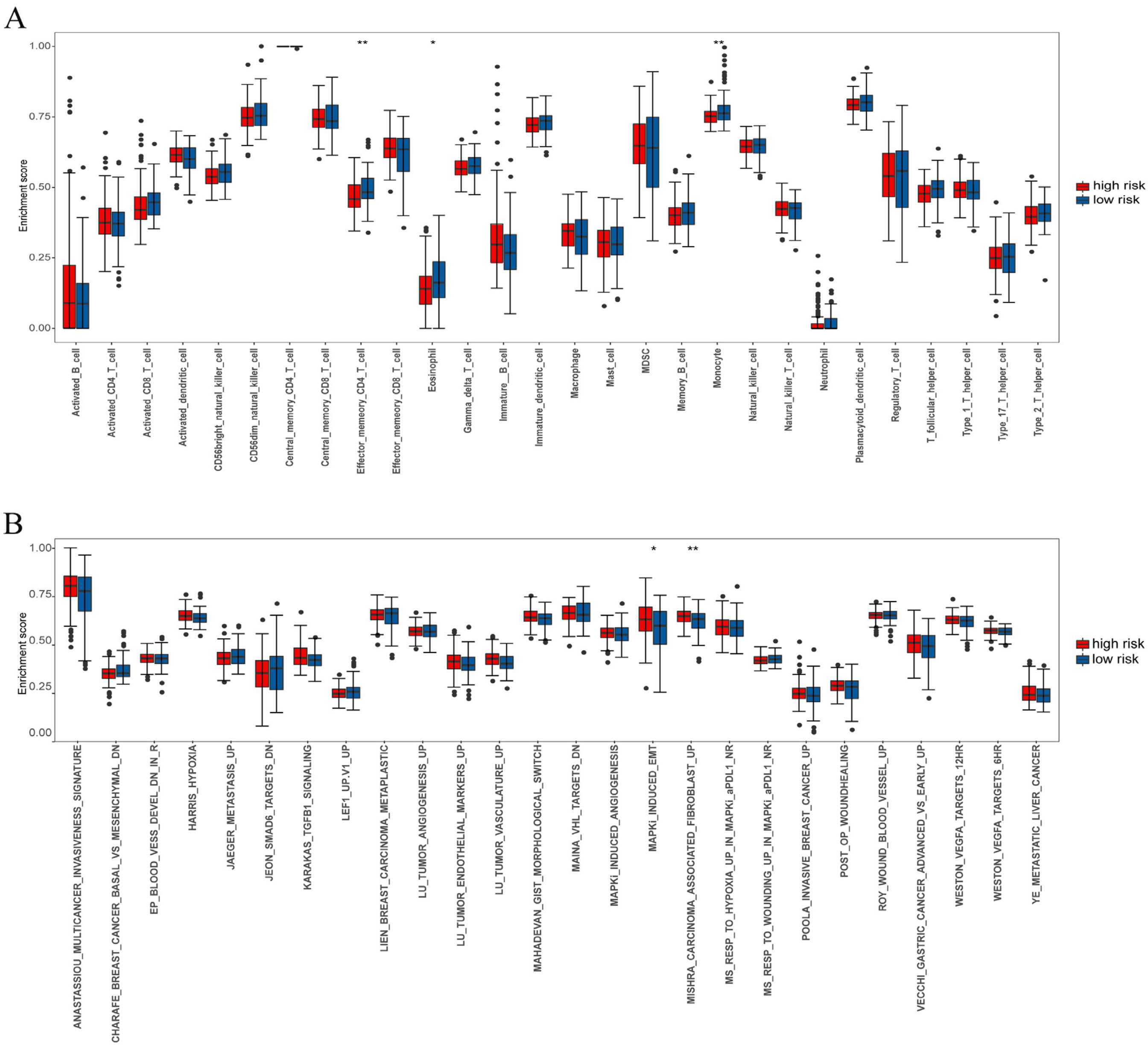

Figure 5. Comparisons of tumor microenvironments of PAAD patients with high- or low-risk prognostic score. (A) Comparison of enrichment scores of immune cells infiltration in PAAD patients between high-risk and low-risk group. (B) Comparison of enrichment scores of innate anti-PD-1 resistance in PAAD patients between high-risk and low-risk group.

showing that TCGA-PAAD patients have poor prognosis if having a high level of amplification/depletion of genes involved in DNA break repair. It should be noted that, we did not use the TCGA-PAAD cohort to validate 'the model' built on the Chinese cohort (notably, 5 genes included in the prognostic model for our Chinese-PAAD cohort, and 8 genes in the model built based on the TCGA-PAAD cohort). Instead, we used this external cohort to strengthen our findings, that is, the prognostic value of CNV of DNA repair and RTK related genes to distinguish PAAD patients with high or low risk of relapse. Indeed, we also attempted to calculate the risk score for each patient in the TCGA-PAAD cohort according to the prognostic model constructed based on our Chinese PAAD datasets. We used the same cutoff (-0.0958) as determined in our Chinese PAAD datasets to separate TCGA-PAAD patients into two groups with high and low risk. Prognostic analysis showed that TCGA-PAAD patients with high-risk score showed decreased median survival time compared to patients with low-risk score (599 days *versus* 728 days, though no significance reached).

As before, we analyzed somatic mutations and pathogenic germline variants of PAAD patients in detail. Similar to Western cohorts, the most patients had somatic mutations in *KRAS*, followed by *TP53, CDKN2A* and *SMAD4*. However, none of these four mutated genes can tell difference in prognosis, despite their high prevalence in our Chinese cohort. Although mutated *KRAS* is a recognized risk factor for PAAD prognosis in previous study [56], we only observed the trend that mutated *KRAS* was associated with worse prognosis (**Figure 1B**), which might be due to different ethnic backgrounds and operation methods (notably, all of PAAD patients in our Chinese cohort received R0 operation). More interestingly, there was no bias in patients with highly mutated genes (*KRAS, TP53, CDKN2A* and *SMAD4*) distributed over four molecular subtypes ($P > 0.05$ on Fisher's exact test), likely explaining why there was no difference in prognosis observed between patients stratified by mutated status of those genes. We detected somatic mutations and pathogenic germline variants in the HRR genes for 12.01% of patients, the prevalence similar to that in the Caris cohort [7]. Although mutated status of HRR genes can stratify platinum treated patients with different prognosis [5,6], we found this genomic marker was unrelated to the prognosis of general patients. Moreover, we clustered patients into two groups based on mutated status of all genes and no different prognosis observed between them. Taken together, we found point mutations are not informative as prognostic markers for PAAD based on our Chinese cohort.

In this study, we focused on analysis of CNVs for PAAD patients. Through combination of unsupervised clustering and stratification by CNV score, we categorized patients into four subtypes, named as repair-deficient, proliferation-active, repair-proficient and repair-enhanced subtypes. For repair-deficient and proliferation-active subtypes, deletion tended to appear in DNA repair genes, including those involved in HRR, mismatch repair and Fanconi anemia. These two subtypes were also signified by higher CNI and stronger signatures of defects in DNA-DSB repair by HRR as well as defective DNA mismatch repair. On the other hand, for repair-proficient and repair-enhanced subtypes, amplification tended to appear in DNA repair genes, accompanied by higher TMB and exclusive signature of defects in polymerase *POLE*. Prognosis of the repair-deficient subtype was better than that of other three subtypes, indicating that deletion of genes in the DNA repair pathway (especially the HRR pathway) induces higher genomic instability and is disadvantaged for survival of cancer cells. In addition to CNV of DNA repair genes, CNV of RTK related genes also impacts prognosis. Compared to the repair-deficient subtype, amplification tended to appear in genes of RTK related signalings in the proliferation-active subtype. Prognosis of the proliferation-active subtype was worse than that of the repair-deficient subtype, indicating that amplification of genes in RTK related signalings promotes proliferation of cancer cells and thus confers worse prognosis. Considering genomic footprints of patients, DNA damage therapies (such as platinum-based chemotherapy and PARPi) are suited for repair-deficient and proliferation-active subtypes (which have higher CNI and defects in DNA-DSB repair by HRR), while immunotherapy is suited for repair-proficient and repair-enhanced subtypes (which have higher TMB and defects in polymerase *POLE*). Intriguingly, Waddell et al observed that PAAD with high *BRCA* mutational signature burden had much better response to platinum-based chemotherapy [57].

Indeed, amplification and deletion events of oncogenes and tumor suppressor genes in PAAD patients have been observed in previous studies [10,58,59], but the association between them and prognosis was not clearly identified. Particularly, a study on 109 micro-dissected PAAD found that patients with high level of amplification or depletion of genes involved in DNA break repair had relatively poor prognosis compared to others [10]. Although the trend discovered by that study was similar with ours, it failed to categorize molecular subtypes for patients and uncover respective characteristics behind them.

It is important to assess whether the genes targeted by CNV gains in this study are amplified or merely affected by broader background gains. To partially address this, we calculated the CNV values for 44 arms of 22 chromosomes (with chromosomes X and Y excluded from this analysis) with the same method as described above, and evaluated the correlation between CNV value of each amplified gene involved in selected pathways (including HRR and RTK related signalings, and the chromosome arm in which these genes located). Our results (**Supplementary Table 7**) showed there was only relatively weak correlation between them (the mean correlation between all pairs is 0.369), suggesting that the genes targeted by CNV gains are likely amplified but not mainly affected by broader background gains. CNVs include gains and losses. We thus also incorporated such distinction into CNV-based analysis. Comparing repair-deficient patients with high (proliferation-active) and low (proliferation-active) CNV score (**Supplementary Figure 12**), we identified 7 genes involved in the RTK and RAS pathways having significantly higher CNV value in proliferation-active patients than in repair-deficient patients. For genes *PTPN11* and *RAC1*, 42% and 56% patients in the proliferation-active subtype were identified as amplification separately, whereas just 17.34% and 32.1% patients in the repair-deficient subtype appeared at lower percentage.

Transcriptomic data has been also utilized for molecular subtyping in PAAD. Collisson et al defined three subtypes (classical, quasi-mesenchymal and exocrine-like) [60], Moffitt et al defined two subtypes (normal stromal and activated stromal) [61], while Bailey et al defined four subtypes (squamous, pancreatic progenitor, immunogenic and aberrantly differentiated endocrine exocrine) [58]. In those studies, prognostic outcome varied among subtypes with moderate significance reported ($0.01 < P < 0.05$). Instead, subtypes defined in our study were highly evident at least from a statistical viewpoint ($P = 2 \times 10^{-4}$ on Log-rank test), indicating that subtyping based on CNV profile was much more powerful for prognostic purpose than doing so based on transcriptome data. This might be due to the fact that the detection for mutations (and subsequent CNV) is more robust (i.e. more specific to tumor cells), whereas the information of gene expression might be influenced by, for example, the presence of stroma (i.e. genes specifically expressed in stromal cells). **Supplementary Table 8**provides the information on tumor contents for the samples profiled in this study. More importantly, we found that genomic footprints behind CNV-driven subtypes were of great use to infer clinical treatments (for example, DNA damage therapies *versus* immunotherapy), providing much clearer guidance for clinical decision-making than transcriptomic subtypes (lacking such clues on how to guide the treatment management). Furthermore, from the viewpoint of clinical practice, transcriptomic measures require higher quality of tumor samples compared to CNV profiling, which is not so easy to achieve, especially for formalin-fixed paraffin-embedded samples.

Finally, we attempted to construct a clinically usable prognostic model to stratify PAAD patients with high and low risk of relapse, providing aids for clinical decision-making. Considering the importance of DNA repair and RTK related genes in molecular subtyping, we used these genes as the initial set to construct the prognostic model. The prognostic model constructed based on the training set also performed well based on the independent testing set. Intriguingly, using the same initial set of genes, a similar prognostic model was also constructed for PAAD patients in the TCGA cohort, suggesting the generalized values of DNA repair and RTK related genes in prognosis. Furthermore, we established the connection between prognostic risk and tumor microenvironments for PAAD patients in the TCGA cohort, in which low-risk patients had more beneficial tumor microenvironments than high-risk patients had.

We note limitations of our study as discussed below. First, we found that mutated genes are not informative in terms of prognostic use; this conclusion is restricted to genes assayed in our gene panels. Second, it is well-recognised that the robustness of information is in order as such: mutations > CNV > expression. Third, the model built on the Chinese-PAAD cohort can not be naively and directly applied to the TCGA-PAAD cohort; the inclusion of genes in the built model is very specific to the input genomic and clinical datasets. Though we can validate the prognostic value of CNV of DNA repair and RTK related genes, the specific genes included in the model are varied (it is also expected from the viewpoint of model building).

In conclusion, based on CNV landscape we categorized PAAD patients into four molecular subtypes (including repair-deficient, proliferation-active, repair-proficient, and repair-enhanced subtypes), each with distinct genomic characteristics, prognostic status and suited treatment (**Figure 6**). In addition, we constructed a clinically actionable prognostic model to stratify patients with high or low risk of relapse. Our molecular subtyping and prognostic model can be of translational use to improve diagnosis, treatment and management for PAAD patients. Considering the relevance of immune microenvironments to prognostic risks, we anticipate that our work

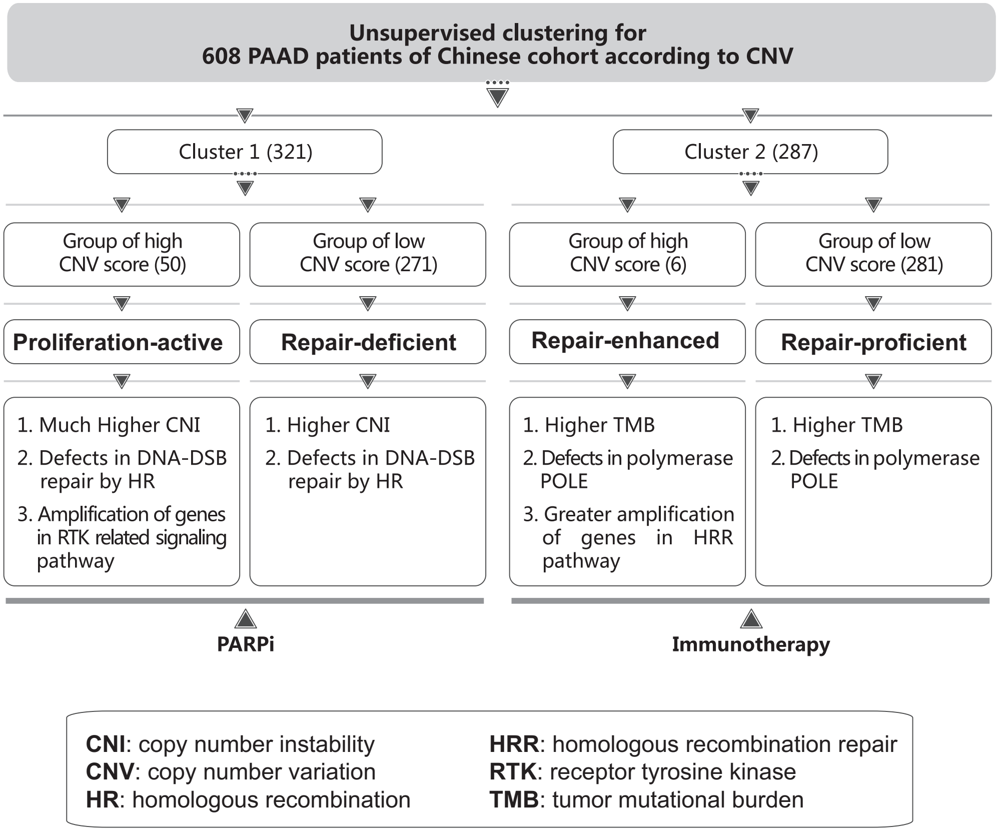
Unsupervised clustering for
608 PAAD patients of Chinese cohort according to CNV
Cluster 1 (321)
Cluster 2 (287)
Group of high CNV score (50)
Group of low CNV score (271)
Group of high CNV score (6)
Group of low CNV score (281)
Proliferation-active
Repair-deficient
Repair-enhanced
Repair-proficient
1. Much Higher CNI
2. Defects in DNA-DSB repair by HR
3. Amplification of genes in RTK related signaling pathway
1. Higher CNI
2. Defects in DNA-DSB repair by HR
1. Higher TMB
2. Defects in polymerase POLE
3. Greater amplification of genes in HRR pathway
1. Higher TMB
2. Defects in polymerase POLE
PARPi
Immunotherapy
CNI: copy number instability
CNV: copy number variation
HR: homologous recombination
HRR: homologous recombination repair
RTK: receptor tyrosine kinase
TMB: tumor mutational burden

Figure 6. Diagram of the treatment management recommended for molecular subtypes of PAAD in the Chinese cohort.

can be further extended, with the priority on either developing new immunomodulators or repurposing existing immunomodulatory therapies particularly for repair-proficient patients who have higher TMB and defects in polymerase *POLE*.

Declaration of Competing Interest

The authors declare no potential conflicts of interest.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant numbers 81470894, 81502695, 81672325, 81871906, 82073326, 82103482 and 32170663), the Shanghai Sailing Program (grant number 20YF1426900), and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (awarded to H.F.).

Data Sharing Statement

In addition to supplementary files, the data that support the findings of this study are also available from the project webpage (https://23verse.github.io/PAAD). The raw sequence data reported in this paper have been deposited into the Genome Sequence Archive in National Genomics Data Center under accession number HRA000456.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2021.103716.

References

[1] Chen W, et al. Cancer statistics in China. CA: a cancer journal for clinicians 2015;66:115–32 (2016). doi: 10.3322/caac.21338.
[2] Zeggini E, Gloyn AL, Barton AC, Wain LV. Translational genomics and precision medicine: Moving from the lab to the clinic. Science 2019;365:1409–13. doi: 10.1126/science.aax4588.
[3] Teo MY, O'Reilly EM. Is it time to split strategies to treat homologous recombinant deficiency in pancreas cancer? J Gastrointest Oncol 2016;7:738–49. doi: 10.21037/jgo.2016.05.04.
[4] Perkhofer L, et al. DNA damage repair as a target in pancreatic cancer: state-of-the-art and future perspectives. Gut 2020. doi: 10.1136/gutjnl-2019-319984.
[5] Park W, et al. Genomic Methods Identify Homologous Recombination Deficiency in Pancreas Adenocarcinoma and Optimize Treatment Selection. Clin Cancer Res 2020;26:3239–47. doi: 10.1158/1078-0432.CCR-20-0418.
[6] Pishvaian MJ, et al. Outcomes in patients with pancreatic adenocarcinoma with genetic mutations in DNA damage response pathways: Results from the Know Your Tumor Program. JCO Precision Oncology 2019;3:1–10.
[7] Heeke AL, et al. Prevalence of Homologous Recombination-Related Gene Mutations Across Multiple Cancer Types. JCO Precis Oncol 2018;2018. doi: 10.1200/PO.17.00286.
[8] Singhi AD, et al. Real-Time Targeted Genome Profile Analysis of Pancreatic Ductal Adenocarcinomas Identifies Genetic Alterations That Might Be Targeted With Existing Drugs or Used as Biomarkers. Gastroenterology 2019;156:2242–53 e2244. doi: 10.1053/j.gastro.2019.02.037.
[9] Balli D, Rech AJ, Stanger BZ, Vonderheide RH. Immune Cytolytic Activity Stratifies Molecular Subsets of Human Pancreatic Cancer. Clin Cancer Res 2017;23:3129–38. doi: 10.1158/1078-0432.CCR-16-2128.
[10] Witkiewicz AK, et al. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. Nat Commun 2015;6:6744. doi: 10.1038/ncomms7744.
[11] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20. doi: 10.1093/bioinformatics/btu170.
[12] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60. doi: 10.1093/bioinformatics/btp324.
[13] Picard toolkit. *Broad Institute, GitHub repository* (2019).
[14] Lai Z, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res 2016;44:e108. doi: 10.1093/nar/gkw227.

[15] Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep 2019;9:9354. doi: 10.1038/s41598-019-45839-z.

[16] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164. doi: 10.1093/nar/gkq603.

[17] Karczewski KJ, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res 2017;45:D840–5. doi: 10.1093/nar/gkw971.

[18] Wang Q, et al. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nature communications 2020;11:2539. doi: 10.1038/s41467-019-12438-5.

[19] Scott AD, et al. CharGer: clinical Characterization of Germline variants. Bioinformatics 2019;35:865–7. doi: 10.1093/bioinformatics/bty649.

[20] Landrum MJ, et al. ClinVar: improvements to accessing data. Nucleic Acids Res 2020;48:D835–44. doi: 10.1093/nar/gkz972.

[21] Talevich E, Shain AH, Botton T, Bastian BCCNVkit. Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol 2016;12:e1004873. doi: 10.1371/journal.pcbi.1004873.

[22] Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572–3. doi: 10.1093/bioinformatics/btq170.

[23] Hazra A, Gogtay N. Biostatistics Series Module 3: Comparing Groups: Numerical Variables. Indian J Dermatol 2016;61:251–60. doi: 10.4103/0019-5154.182416.

[24] Zeng D, et al. Tumor Microenvironment Characterization in Gastric Cancer Identifies Prognostic and Immunotherapeutically Relevant Gene Signatures. Cancer Immunol Res 2019;7:737–50. doi: 10.1158/2326-6066.CIR-18-0436.

[25] Hothorn T. maxstat: Maximally Selected Rank Statistics. R package 2017.

[26] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7. doi: 10.1089/omi.2011.0118.

[27] Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9. doi: 10.1038/75556.

[28] The Gene Ontology C. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 2019;47:D330–8. doi: 10.1093/nar/gky1055.

[29] Ogata H, et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 1999;27:29–34. doi: 10.1093/nar/27.1.29.

[30] Jassal B, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2020;48:D498–503. doi: 10.1093/nar/gkz1031.

[31] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1–22.

[32] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J Stat Softw 2011;39:1–13. doi: 10.18637/jss.v039.i05.

[33] Charoentong P, et al. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell Rep 2017;18:248–62. doi: 10.1016/j.celrep.2016.12.019.

[34] Subramanian A, 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. doi: 10.1073/pnas.0506580102.

[35] Hugo W, et al. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell 2016;165:35–44. doi: 10.1016/j.cell.2016.02.065.

[36] Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 2013;14(7). doi: 10.1186/1471-2105-14-7.

[37] Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. International journal of Ayurveda research 2010;1:274–8. doi: 10.4103/0974-7788.76794.

[38] Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. R package 2000.

[39] Therneau TM. A Package for Survival Analysis in R. R package 2020.

[40] Robin X, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77. doi: 10.1186/1471-2105-12-77.

[41] Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016;32:2847–9. doi: 10.1093/bioinformatics/btw313.

[42] Cancer Genome Atlas Research Network. Electronic address, a. a. d. h. e. & Cancer Genome Atlas Research, N. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell 2017;32:185–203 e113. doi: 10.1016/j.ccell.2017.07.007.

[43] Zhao P, Li L, Jiang X, Li Q. Mismatch repair deficiency/microsatellite instability-high as a predictor for anti-PD-1/PD-L1 immunotherapy efficacy. J Hematol Oncol 2019;12:54. doi: 10.1186/s13045-019-0738-1.

[44] Knijnenburg TA, et al. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell reports 2018;23:239–54 e236. doi: 10.1016/j.celrep.2018.03.076.

[45] Saini N, et al. Mutation signatures specific to DNA alkylating agents in yeast and cancers. Nucleic Acids Res 2020;48:3692–707. doi: 10.1093/nar/gkaa150.

[46] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99:6567–72. doi: 10.1073/pnas.082099299.

[47] Thorsson V, et al. The Immune Landscape of Cancer. Immunity 2018;48:812–30 e814. doi: 10.1016/j.immuni.2018.03.023.

[48] Gu-Trantien C, Willard-Gallo K. Tumor-infiltrating follicular helper T cells: The new kids on the block. Oncoimmunology 2013;2:e26066. doi: 10.4161/onci.26066.

[49] Harrington LE, Janowski KM, Oliver JR, Zajac AJ, Weaver CT. Memory CD4 T cells emerge from effector T-cell progenitors. Nature 2008;452:356–60. doi: 10.1038/nature06672.

[50] Olingy CE, Dinh HQ, Hedrick CC. Monocyte heterogeneity and functions in cancer. J Leukoc Biol 2019;106:309–22. doi: 10.1002/JLB.4RI0818-311R.

[51] Grisaru-Tal S, Itan M, Klion AD, Munitz A. A new dawn for eosinophils in the tumour microenvironment. Nature reviews. Cancer 2020. doi: 10.1038/s41568-020-0283-9.

[52] Poli A, et al. CD56bright natural killer (NK) cells: an important NK cell subset. Immunology 2009;126:458–65. doi: 10.1111/j.1365-2567.2008.03027.x.

[53] Wagner JA, et al. CD56bright NK cells exhibit potent antitumor responses following IL-15 priming. The Journal of clinical investigation 2017;127:4042–58. doi: 10.1172/JCI90387.

[54] van der Leun AM, Thommen DS, Schumacher TN. CD8(+) T cell states in human cancer: insights from single-cell analysis. Nature reviews. Cancer 2020;20:218–32. doi: 10.1038/s41568-019-0235-4.

[55] Crotty S. T follicular helper cell differentiation, function, and roles in disease. Immunity 2014;41:529–42. doi: 10.1016/j.immuni.2014.10.004.

[56] Qian ZR, et al. Association of Alterations in Main Driver Genes With Outcomes of Patients With Resected Pancreatic Ductal Adenocarcinoma. JAMA Oncol 2018;4:e173420. doi: 10.1001/jamaoncol.2017.3420.

[57] Waddell N, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. Nature 2015;518:495–501. doi: 10.1038/nature14169.

[58] Bailey P, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature 2016;531:47–52. doi: 10.1038/nature16965.

[59] Biankin AV, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 2012;491:399–405. doi: 10.1038/nature11547.

[60] Collisson EA, et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat Med 2011;17:500–3. doi: 10.1038/nm.2344.

[61] Moffitt RA, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat Genet 2015;47:1168–78. doi: 10.1038/ng.3398.