# Transcriptional landscape of DNA repair genes underpins a pan-cancer prognostic signature associated with cell cycle dysregulation and tumor hypoxia

Running title: **A signature of DNA repair genes in six cancers**

Wai Hoong Chang and Alvina G. Lai

Nuffield Department of Medicine, University of Oxford,

Old Road Campus, OX3 7FZ, United Kingdom

For correspondence: alvina.lai@ndm.ox.ac.uk

## Abstract

Overactive DNA repair contributes to therapeutic resistance in cancer. However, pan-cancer comparative studies investigating the contribution of *all* DNA repair genes in cancer progression employing an integrated approach have remained limited. We performed a multi-cohort retrospective analysis to determine the prognostic significance of 138 DNA repair genes in 16 cancer types (n=16,225). Cox proportional hazards analyses revealed a significant variation in the number of prognostic genes between cancers; 81 genes were prognostic in clear cell renal cell carcinoma while only two genes were prognostic in glioblastoma. We reasoned that genes that were commonly prognostic in highly correlated cancers revealed by Spearman's correlation analysis could be harnessed as a molecular signature for risk assessment. A 10-gene signature, uniting prognostic genes that were common in highly correlated cancers, was significantly associated with overall survival in patients with clear cell renal cell (P<0.0001), papillary renal cell (P=0.0007), liver (P=0.002), lung (P=0.028), pancreas (P=0.00013) or endometrial (P=0.00063) cancers. Receiver operating characteristic analyses revealed that a combined model of the 10-gene signature and tumor staging outperformed either classifier when considered alone. Multivariate Cox regression models incorporating additional clinicopathological features showed that the signature was an independent predictor of overall survival. Tumor hypoxia is associated with adverse outcomes. Consistent across all six cancers, patients with high 10-gene and high hypoxia scores had significantly higher mortality rates compared to those with low 10-gene and low hypoxia scores. Functional enrichment analyses revealed that high mortality rates in patients with high 10-gene scores were attributable to an overproliferation phenotype. Death risk in these patients was further exacerbated by concurrent mutations of a cell cycle checkpoint protein, *TP53*. The 10-gene

signature identified tumors with heightened DNA repair ability. This information has the potential to radically change prognosis through the use of adjuvant DNA repair inhibitors with chemotherapeutic drugs.

[298 words]

Keywords: DNA repair; pan-cancer; cell cycle; hypoxia; tumor microenvironment

List of abbreviations:

| | |
|---|---|
| DDR | DNA damage response |
| BER | Base excision repair |
| NER | Nucleotide excision repair |
| MR | Mismatch repair |
| HDR | Homology-directed repair |
| NHEJ | Non-homologous end joining |
| FA | Fanconi anemia |
| TCGA | The Cancer Genome Atlas |
| GO | Gene Ontology |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| HR | Hazard ratio |
| ROC | Receiver operating characteristic |
| AUC | Area under the curve |
| TNM | Tumor, node and metastasis |
| CDK | Cyclin-dependent kinase |
| DEG | Differentially expressed genes |

## Introduction

Genetic material must be transmitted in its original, unaltered form during cell division. However, DNA faces continuous assaults from both endogenous and environmental agents contributing to the formation of permanent lesions and cell death. To overcome DNA damage threats, living systems have evolved highly coordinated cellular machineries to detect and repair damages as they occur. However, DNA repair mechanisms and consequently DNA damage responses (DDR) are often deregulated in cancer cells and such aberrations may contribute to cancer progression and influence prognosis. Overexpression of DNA repair genes allows tumor cells to overcome the cytotoxic effects of radiotherapy and chemotherapy. As such, inhibitors of DNA repair can increase the vulnerability of tumor cells to chemotherapeutic drugs by preventing the repair of deleterious lesions[1].

There are six main DNA repair pathways in mammalian cells. Single-strand DNA damage is repaired by the base excision repair (BER), nucleotide excision repair (NER) and mismatch repair (MR) pathways. The poly(ADP-ribose) polymerase (PARP) gene family encodes critical players of the BER pathway involved in repairing damages induced by ionizing radiation and alkylating agents[2,3]. Replication errors are corrected by the MR pathway while the NER pathway is responsible for removing bulky intercalating agents[4,5]. Tumor cells with deficiencies in the NER pathway have increased sensitivity to platinum-based chemotherapeutic drugs (cisplatin, oxaliplatin, etc.)[6,7]. Double-strand breaks induced by ionizing radiation are more difficult to repair and thus are highly cytotoxic. Dysregulation of genes involved in the homology-directed repair (HDR), non-homologous end joining (NHEJ) and Fanconi anemia (FA) pathways are associated with altered repair of double-strand breaks.

Aberrations in DNA repair genes are widespread in most cancers; hence they represent attractive candidates for pharmacological targeting to improve radiosensitivity and chemosensitivity[8]. In a process known as 'synthetic lethality', faults in two or more DNA repair genes or pathways together would promote cell death, while defects in a single pathway may be tolerated[1]. Functional redundancies in repair pathways allow tumor cells to rely on a second pathway for repair if the first pathway is defective. Based on the principles of synthetic lethality, inhibition of the second pathway will confer hypersensitivity to cytotoxic drugs in cells with another malfunctioning pathway. This promotes cell death because DNA lesions can no longer be repaired by either pathway. For instance, PARP inhibitors (targeting the BER pathway) could selectively kill tumor cells that have *BRCA1* or *BRCA2* mutations (defective HDR pathway) while not having any toxic effects on normal cells[9,10].

Since one DDR pathway could compensate for another, there is a need for a pan-cancer, large-scale, systematic study on *all* DNA repair genes to reveal similarities and differences in DDR signaling between cancer types, which is limited at present. In this study, we explored pan-genomic expression patterns of 138 DNA repair genes in 16 cancer types. We developed and validated the prognostic significance of a 10-gene signature that can be used for rapid risk assessment and patient stratification. There are considerable variations in the success of chemotherapy and radiotherapy regimes between cancer types. Such differences may be explained by the complex cancer-specific nature of DDR defects. Prognostic biomarkers of DNA repair genes are needed to allow the use of repair inhibitors in a stratified, non-universal approach to expose the particular vulnerabilities of tumors to therapeutic agents.

## Materials and methods

A list of 138 DNA repair genes is available in Table S1.

### Study cohorts

We obtained RNA-sequencing datasets for the 16 cancers from The Cancer Genome Atlas (TCGA)[11] (n=16,225) (Table S2). TCGA Illumina HiSeq rnaseqv2 Level 3 RSEM normalized data were retrieved from the Broad Institute GDAC Firehose website. Gene expression profiles for each cancer types were separated into tumor and non-tumor categories based on TCGA barcodes and converted to $\log_2(x + 1)$ scale. To compare the gene-by-gene expression distribution in tumor and non-tumor samples, violin plots were generated using R. The nonparametric Mann-Whitney-Wilcoxon test was used for statistical analysis.

### Calculation of 10-gene scores and hypoxia scores

The 10-gene scores for each patient were determined from the mean $\log_2$ expression values of 10 genes: *PRKDC, NEIL3, FANCD2, BRCA2, EXO1, XRCC2, RFC4, USP1, UBE2T* and *FAAP24*). Hypoxia scores were calculated from the mean $\log_2$ expression values of 52 hypoxia signature genes[12]. For analyses in Figure 5, patients were delineated into four categories using median 10-gene scores and hypoxia scores as thresholds. The nonparametric Spearman's rank-order correlation test was used to determine the relationship between 10-gene scores and hypoxia scores.

### Differential expression analyses comparing expression profiles of high-score and low-score patients

Patients were median dichotomized into low- and high-score groups based on their 10-gene scores in each cancer type. Differential expression analyses were performed using the linear model and Bayes method executed by the limma package in R. P values were adjusted using the Benjamini-Hochberg false discovery rate procedure. We considered genes with $\log_2$ fold change of > 1 or < -1 and adjusted P-values < 0.05 as significantly differentially expressed between the two patient groups.

## Functional enrichment and pathway analyses

To determine which biological pathways were significantly enriched, differentially expressed genes were mapped against the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases using GeneCodis[13]. The Enrichr tool was used to investigate transcription factor protein-protein interactions that were associated with the differentially expressed genes[14,15].

## Survival analysis

Univariate Cox proportional hazards regression analyses were performed using the R survival and survminer packages to determine if expression levels of individual DNA repair genes as well as those of the 10-gene scores were significantly associated with overall survival. Multivariate Cox regression was employed to determine the influence of additional clinical variables on the 10-gene signature. Hazard ratios (HR) and confidence intervals were determined from the Cox models. HR greater than one indicated that a covariate was positively associated with even probability or increased hazard and negatively associated with survival

duration. Non-significant relationship between scaled Schoenfeld residuals supported the proportional hazards assumption in the Cox model. Both survival and survminer packages were also used for Kaplan-Meier analyses and log-rank tests. For Kaplan-Meier analyses, patients were median dichotomized into high- and low-score groups using the 10-gene signature. To determine the predictive performance (specificity and sensitivity) of the signature in relation to tumor staging parameters, we employed the receiver operating characteristic (ROC) analysis implemented by the R survcomp package, which also calculates area under the curve (AUC) values. AUC values can fall between 1 (perfect marker) and 0.5 (uninformative marker).

*TP53* mutation analysis

TCGA mutation datasets (Level 3) were retrieved from GDAC Firehose to annotate patients with mutant *TP53*. To ascertain the association of *TP53* mutation with the 10-gene signature on overall survival, we employed the Kaplan-Meier analysis and log-rank tests implemented in R.

All plots were generated using R pheatmap and ggplot2 packages[16]. Venn diagram was generated using the InteractiVenn tool[17].

## Results

### Prognosis of DNA repair genes in 16 cancer types and the development of a 10-gene signature

A total of 187 genes associated with six DDR pathways found in mammalian cells were curated: BER (33 genes), MR (23 genes), NER (39 genes), HDR (26 genes), NHEJ (13 genes) and FA (53 genes)[18] (Fig. 1, Table S1). Of the 187 genes, 49 were represented in two or more pathways, yielding 138 non-redundant candidates. To determine which of the 138 DNA repair genes conferred prognostic information, we employed Cox proportional hazards regression on all genes individually on 16 cancer types to collectively include 16,225 patients[11] (Table S2). In clear cell renal cell carcinoma, 81 genes were found to be significantly associated with overall survival; this cancer had the highest number of prognostic DNA repair genes (Table S3). This is followed by 54, 53, 46, 44 and 33 prognostic genes in cancers of the pancreas, papillary renal cell, liver, lung and endometrium respectively (Table S3). In contrast, cancers of the brain (glioblastoma: 2 genes), breast (5 genes), cervix (6 genes) and esophagus (7 genes) had some of the lowest number of prognostic DNA repair genes (Table S3), suggesting that there is a significant degree of variation in the contribution of DNA repair genes in predicting survival outcomes. Spearman's rank-order correlation analysis revealed a hub of five highly correlated cancers (lung, papillary renal cell, pancreas, liver and endometrium), indicating that a good number of prognostic DNA repair genes were shared between these cancers (Spearman's rho=0.21 to 0.44) (Fig. S1). We rationalized that prognostic genes that are common in these highly correlated cancers could form a new multigenic risk assessment classifier. Ten genes were prognostic in the five highly correlated cancers: *PRKDC* (NHEJ), *NEIL3* (BER), *FANCD2* (FA), *BRCA2* (HDR and FA), *EXO1* (MR), *XRCC2* (HDR), *RFC4* (MR and NER), *USP1* (FA), *UBE2T* (FA) and *FAAP24* (FA), which, interestingly, represent members from all six DDR pathways.

181

182 A 10-gene signature predictive of DDR signaling is an independent prognostic classifier in 6
183 cancer types

184 The ten genes above were employed as a new prognostic model to evaluate whether they
185 were significantly associated with overall survival in all 16 cancer types. A 10-gene score for
186 each patient was calculated by taking the mean expression of all ten genes. Patients were
187 median dichotomized based on their 10-gene scores into a low- and high-score groups. The
188 10-gene signature could predict patients at significantly higher risk of death in the five cancers
189 that were initially highly correlated (Fig. S1), and in one additional cancer (clear cell renal cell
190 carcinoma) (Fig. 2). Kaplan-Meier analyses demonstrated that patients categorized within
191 high-score groups had significantly poorer survival rates: clear cell renal cell (log-rank
192 $P<0.0001$), papillary renal cell ($P=0.0007$), liver ($P=0.002$), lung ($P=0.028$), pancreas
193 ($P=0.00013$) and endometrium ($P=0.00063$) (Fig. 2). Expression profiles of the 10 genes in
194 tumor and non-tumor samples showed a general distribution that were comparable among
195 the six cancer types. Mann-Whitney-Wilcoxon tests revealed that a vast majority of genes were
196 significantly upregulated in tumor samples with a few minor exceptions (Fig. S2). *USP1* was
197 significantly downregulated in tumors of papillary renal cell and endometrium (Fig. S2). Only
198 four non-tumor samples were available in the pancreatic cancer cohort, precluding robust
199 statistical analyses. Due to limitations in sample size, only *UBE2T* was observed to be
200 significantly upregulated in pancreatic tumors (Fig. S2).

201

202 To evaluate the independent predictive value of the signature over the current tumor, node
203 and metastasis (TNM) staging system, we applied the signature on patients separated by TNM
204 stage: early (stages 1 and/or 2), intermediate (stages 2 and/or 3) and late (stages 3 and/or 4)

disease stages. Remarkably, the signature successfully identified high-risk patients in early (liver, lung, pancreas, endometrium), intermediate (papillary renal cell, liver, pancreas, endometrium) and late (clear cell renal cell, papillary renal cell, liver, endometrium) TNM stages (Fig. 3). Collectively, this implied that the signature offered an additional resolution of prognosis within similarly staged tumors and that the signature retained excellent prognostic ability in individual tumor groups when considered separately.

To evaluate the predictive performance of the 10-gene signature on 5-year overall survival, we employed receiver operating characteristic (ROC) analyses on all six cancers. Comparing the sensitivity and specificity of the signature in relation to TNM staging revealed that the signature outperformed TNM staging in cancers of the papillary renal cell (AUC=0.832 vs. AUC=0.640), pancreas (AUC=0.697 vs. AUC=0.593) and endometrium (AUC=0.700 vs. AUC=0.674) (Fig. 4). Importantly, when the signature was used in conjunction with TNM staging as a combined model, its performance was superior to either classifier when they were considered individually: clear cell renal cell (AUC=0.792), papillary renal cell (AUC=0.868), liver (AUC=0.751), lung (AUC=0.693), pancreas (AUC=0.698) and endometrium (AUC=0.764) (Fig. 4).

We next employed multivariate Cox regression models to examine whether the association between high 10-gene scores and increased mortality was not due to underlying clinical characteristics of the tumors. Univariate analysis revealed that TNM staging is not prognostic in pancreatic cancer (hazard ratio [HR]=1.339, P=0.153); hence this cancer was excluded from the multivariate model involving TNM (Table 1). For the five remaining cancer types, even when TNM staging was considered, the signature significantly distinguished survival outcomes

in high- versus low-score patients, confirming that it is an independent prognostic classifier: clear cell renal cell (HR=1.555, P=0.0058), papillary renal cell (HR=1.677, P=0.032), liver (HR=1.650, P=0.029), lung (HR=1.301, P=0.032) and endometrium (HR=2.113, P=0.013) (Table 1).

## Crosstalk between DDR signaling and tumor hypoxia

Tumor hypoxia is a well-known barrier to curative treatment. It is often associated with poor prognosis[19,20], which may be a result of tumor resistance to chemotherapy and radiotherapy[21,22]. Since both the upregulation of DNA repair genes and hypoxia are linked to therapeutic resistance, we rationalized that incorporating hypoxia information in the 10-gene signature would allow further delineation of patient risk groups. Patients with high 10-gene scores had significantly poorer survival outcomes and we predict that these patients have tumors that are more hypoxic, and that oxygen deprivation could influence DDR signaling to enhance tumor resistance to apoptotic stimuli leading to more aggressive disease states. We calculated hypoxia scores for each patient using a mathematically derived hypoxia gene signature consisting of 52 genes[12]. Hypoxia scores were defined as the mean expression of the 52 genes. Patients for each of the six cancer types were divided into four categories using the median 10-gene and hypoxia scores: 1) high scores for both 10-gene and hypoxia, 2) high 10-gene and low hypoxia scores, 3) low 10-gene and high hypoxia scores and 4) low scores for both 10-gene and hypoxia (Fig. 5A). Remarkably, significant positive correlations were observed between 10-gene scores and hypoxia scores consistent across all six cancer types: clear cell renal cell (rho=0.363, P<0.0001), papillary renal cell (rho=0.518, P<0.0001), liver (rho=0.615, P<0.0001), lung (rho=0.753, P<0.0001), pancreas (rho=0.582, P<0.0001) and

endometrium (rho=0.527, P<0.0001) (Fig. 5A). This suggests that tumor hypoxia may influence DDR signaling and potentially, patient outcomes.

We generated Kaplan-Meier curves and employed the log-rank test to determine whether there were differences in overall survival outcomes among the four patient groups. Combined relation of hypoxia and 10-gene scores revealed significant associations with overall survival in all six cancers (Fig. 5B). Patients classified within the 'high 10-gene and high hypoxia' category had significantly poorer survival rates compared to those with low 10-gene and low hypoxia scores: clear cell renal cell (HR=2.316, P<0.0001), papillary renal cell (HR=7.635, P=0.0011), liver (HR=2.615, P=0.00013), lung (HR=1.832, P=0.0021), pancreas (HR=2.680, P=0.00079) and endometrium (HR=2.707, P=0.0075) (Table 2; Fig. 5B). Our results suggest that the combined effects of hypoxia and heightened expression of DNA damage repair genes may be linked to tumor progression and increased mortality risks. It remains unknown in this context whether the basis for differential sensitivity to chemotherapy would be explained, in part, by DNA repair ability of tumor cells exposed to chronic hypoxia environments.

## Patients with high 10-gene scores had an overproliferation phenotype due to cell cycle dysregulation

The cell cycle represents a cellular gatekeeper that controls how cells grow and proliferate. Cyclins and cyclin-dependent kinases (CDKs) allow cells to progress from one cell cycle stage to the next; a process that is antagonized by CDK inhibitors. Many tumors overexpress cyclins or inactivate CDK inhibitors, hence resulting in uncontrolled cell cycle entry, loss of checkpoint and uninhibited proliferation[23–25]. Targeting proteins responsible for cell cycle progression

would thus be an attractive measure to limit tumor cell proliferation. This has led to the development of various CDK inhibitors as anticancer agents[26,27]. DNA repair is tightly coordinated with cell cycle progression. Certain DNA repair mechanisms are dampened in non-proliferating cells, while repair pathways are often perturbed during tumor development. Perturbation can take the form of defective DNA repair or over-compensation of a pathway arising from defects in another pathway[28]. As a result, DNA repair inhibitors could prevent the repair of lesions induced by chemotherapeutic drugs to trigger apoptosis and to enhance the elimination of tumor cells.

We rationalize that patients with high 10-gene scores would have heightened ability for DNA repair thus allowing tumor cells to progress through the cell cycle and continue to proliferate. Using Spearman's rank-order correlation, we observed that the expression of each of the 10 signature genes was positively correlated with the expression of genes involved in cell cycle progression (cyclins and CDKs) and negatively correlated with genes involved in cell cycle arrest (CDK inhibitors) (Fig. 6A). Interestingly, the patterns of correlation were remarkably similar across all six cancer types, implying that elevated expression of DNA repair genes is associated with a hyperproliferative phenotype. We next asked whether patients within the high 10-gene score category had an overrepresentation of processes associated with cell cycle dysregulation as this could explain the elevated mortality risks in these patients. To answer this, we divided patients from each of the six cancer types into two groups (high score and low score) based on the mean expression of the 10 signature genes using the $50^{th}$ percentile cut-off. Differential expression analyses between the high- and low-score groups revealed that 394, 425, 1259, 1279, 714 and 977 genes were differentially expressed ($-1 > \log_2$ fold-change $> 1$, $P<0.05$) in

clear cell renal cell, papillary renal cell, liver, lung, pancreas and endometrial cancers respectively (Table S4).

Analyses of biological functions of these genes revealed functional enrichment of ontologies associated with cell division, mitosis, cell cycle, cell proliferation, DNA replication and homologous recombination consistent in all six cancer types (Fig. 6B). This suggests that the significantly higher mortality rates in patients with high 10-gene scores were due to enhanced tumor cell proliferation exacerbated by the ability of these cells to repair DNA lesions as they arise. Additional ontologies related to tumorigeneses such as *PPAR* and *TP53* signaling were also associated with poor prognosis (Fig. 6B). A total of 87 differentially expressed genes (DEGs) were found to be in common in all six cancer types (Fig. S3) (Table S5). To dissect the underlying biological roles of the 87 DEGs at the protein level, we evaluated the enrichment of transcription factor protein-protein interactions using the Enrichr platform[14].*TP53* represents the most enriched transcription factor involved in the regulation of the DEGs as evidenced by the highest combined score, which takes into account both Z score and P value (Table S6). This indirectly corroborated our results on enriched *TP53* signaling obtained from the KEGG pathway analysis (Fig. 6B). Taken together, these results highlight the interplay between DDR signaling, cell cycle regulation and *TP53* function in determining prognosis.

## Prognostic relevance of a combined model involving the 10-gene signature and *TP53* mutation status

An important role of *TP53* is its tumor suppressive function through *TP53*-mediated cell cycle arrest and apoptosis[29]. Hence, somatic mutations in *TP53* can confer tumor cells with a

growth advantage and indeed, this is a well-known phenomenon in many cancers[30–32]. We rationalized that *TP53* deficiency resulting in defective checkpoint may synergize with the overexpression of DNA repair genes to prevent growth arrest and promote tumor proliferation. To test this hypothesis, we examined *TP53* mutation status in all six cancer types and observed that *TP53* mutation frequency was the highest in pancreatic cancer patients (58%) followed by lung cancer (57%), endometrial cancer (21%), liver cancer (16%), papillary renal cell (1.8%) and clear cell renal cell (1.2%) (Table S7). Cancers with *TP53* mutation frequency of at least 10% were selected for survival analyses. Univariate Cox regression analyses revealed that *TP53* mutation status only conferred prognostic information in pancreatic (HR=1.657, P=0.044), endometrial (HR=1.780, P=0.041) and liver (HR=2.603, P<0.0001) cancers but not in lung cancer (HR=1.428, P=0.056) (Table 1). Cancers, where *TP53* mutation offered predictive value, were taken forward for analyses in relation to the 10-gene signature. Cox regression analyses revealed that a combination of *TP53* mutation and high 10-gene score resulted in a significantly higher risk of death (Table 3; Fig. 6C). Survival rates were significantly diminished in patients harboring high 10-gene scores and the mutant variant of *TP53* compared to those with low 10-gene scores and wild-type *TP53*: liver (HR=3.876, P<0.0001), pancreas (HR=4.881, P=0.0002) and endometrium (HR=3.719, P=0.00028) (Table 3; Fig. 6C). Moreover, in multivariate Cox models involving TNM staging and *TP53* mutation status, the 10-gene signature remained a significant prognostic factor (Table 1). This suggests that although the 10-gene signature provided additional resolution in risk assessment when used in combination with *TP53* mutation status, its function is independent. However, in the multivariate model, *TP53* was significant only in liver cancer (HR=2.085, P=0.0044), suggesting that *TP53* mutation was not independent of the signature or TNM staging in pancreatic and endometrial cancers (Table 1). Overall, the results suggest that defects in cell cycle checkpoint combined with augmented

DNA repair ability were adverse risk factors contributing to poor prognosis. Both *TP53* mutation status and 10-gene scores could offer additional predictive value in risk assessment by further delineation of patients into additional risk groups.

## Discussion and Conclusion

We systematically examined the associations between the expression patterns of 138 DNA repair genes in 16 cancer types and prognosis. Our pan-cancer multigenic approach revealed genes that work synergistically across cancers to inform patient prognosis that would otherwise remain undetected in analysis involving a single gene or a single cancer type. We developed a 10-gene signature that incorporates the expression profiles of 10 highly correlated DNA repair genes for use as risk predictors in six cancer types (n=2,257). This signature offers more precise discrimination of patient risk groups in these six cancers where high expression of signature genes is associated with poor survival outcomes. Importantly, we demonstrated that the signature could improve the prognostic discrimination of TNM when used as a combined model, which is particularly useful to allow further stratification of patients within similar TNM stage groups (Fig. 4).

Intrinsic differences in DNA repair machineries in cancer cells may pose a significant challenge to successful therapy. Mutations in DNA repair genes allow the generation of persistent DNA lesions that would otherwise be repaired. Germline mutations of DNA repair genes are linked to increased genome instability and cancer risks[33] and abrogation of genes in one DNA repair pathway can be compensated by another pathway[1]. *BRCA1* and *BRCA2* mutations sensitize cells to PARP1 inhibition, a protein involved in the BER pathway[10]. Since *BRCA1* and *BRCA2* are important for homology-directed repair, PARP1 inhibition in *BRCA1/2*-defective cells would result in dysfunctional HDR and BER pathways preventing lesion repair and thus leading to apoptosis[10].

In addition to genetic polymorphism, upregulation of DNA repair genes in tumors promotes resistance to radiotherapy and chemotherapy as the cells would have enhanced ability to repair cytotoxic lesions induced by these therapies. Overexpression of *ERCC1* involved in the NER pathway in non-small-cell lung cancer is linked to poor survival in cisplatin-treated patients[7]. The 1,2-d(GpG) cross-link lesion generated by cisplatin treatment is readily repaired by the NER pathway; hence *ERCC1* overexpression would promote cisplatin resistance. Low *MGMT* expression in astrocytoma is associated with longer survival outcomes in patients treated with temozolomide[34]; an observation that is consistent with the role of *MGMT* in repairing lesions caused by temozolomide thus allowing *MGMT* deficient tumor cells to accumulate enough unrepairable damage. The ribonucleotide reductase (RNR) enzyme plays an important role in DNA repair and RNR activity is tightly coordinated with cell cycle progression to maintain a balance between DNA replication and dNTP production[35]. Overexpression of the RNR subunit, RRM2, is associated with poor outcomes in breast[36,37] and colorectal cancers[38]. In prostate cancer, overexpression of RRM2 promotes 3D colony formation and invasive phenotypes where RRM2 activates epithelial-to-mesenchymal transition through the upregulation of E-cadherin and P-cadherin[39]; an observation that is consistent with our results showing significant positive correlation between expression of DNA repair genes with genes involved in cell cycle progression (Fig. 6). EXO1 is involved in DNA MR and HR[40]. EXO1 expression promotes survival of ovarian cancer cells post cisplatin treatment[41]. Overexpression of EXO1 is positively correlated with tumor aggression and unfavorable prognosis in astrocytoma[42] and in liver cancer[43], where in the latter, EXO1 knockdown suppresses clonogenic cell survival and increases radiosensitivity[43]. XRCC5 is a subunit of the Ku heterodimer protein involved in NHEJ and is overexpressed in multiple cancer types including head and neck[44], colorectal[45] and lung[46] cancers. Overexpression of

XRCC5 is also a poor prognostic marker in gastric cancer[47]. The MRN complex, consisting of MRE11, RAD50 and NBS1 proteins, is essential for repairing double-stranded breaks. Tumors deficient in the MRN complex are more sensitive to the DNA-damaging effects of radiotherapy and likewise, high MRN expression is associated with poor disease-free and overall survival in colorectal cancer patients receiving neoadjuvant radiotherapy[48]. FEN1 is an endonuclease involved in BER and NHEJ. FEN1 is overexpressed in breast, brain, lung, testis, prostate and gastric cancers[49–52]. Moreover, FEN1 overexpression is linked to high tumor grade and poor survival outcomes in ovarian and breast cancers[53]. Non-small-cell lung cancer patients with FEN1 overexpression have poor differentiation and poor prognosis and knock-down of FEN1 attenuates homologous DNA repair, which promotes the cytotoxic effects of cisplatin[54]. Similarly, FEN1 downregulation in glioma cells causes increased sensitivity to temozolomide damage [50]. *TP53* plays essential roles in cell-cycle arrest and apoptosis through the activation of checkpoint genes[29]. We show that patients with high 10-gene scores that concurrently have mutant *TP53* exhibited significantly higher mortality rates (Fig. 6C), suggesting that defects in cell cycle checkpoint coupled with an increased propensity for DNA repair may lead to dramatically poorer outcomes. Taken together, our study along with reports from others confirmed that hyperactive DNA repair is linked to tumor aggression and adverse patient outcomes.

Multiple studies have reported the associations between dysfunctional DNA repair pathways and cancer, but most of these studies are restricted to investigations on a limited number of genes and in one cancer at a time. One of the key advantages of our study is that it is an unbiased exploration transcending the candidate-gene approach that takes into account the

multifaceted interplay of DNA repair genes in diverse cancer types. We rationalize that since ionizing radiation and chemotherapy are the main treatment options currently available for cancer patients, a molecular signature capable of discriminating patients with increased expression of DNA repair genes that would benefit from adjuvant therapy through pharmacological inhibition of DNA repair to overall improve therapeutic outcomes.

Tumor hypoxia is also a well-known cause of therapy resistance. A notable finding of our study is that patients having both high 10-gene and hypoxia scores had significantly poorer survival rates compared to those with low 10-gene and hypoxia scores (Fig. 5). Previous reports suggest that low oxygen conditions may interfere with DNA damage repair. For example, hypoxia could compromise HR function through decreased *RAD51* expression[55]. However, results concerning the effects of hypoxia on DDR signaling have remained inconclusive. Genes associated with NHEJ were reported to be downregulated under hypoxia in prostate cancer cell lines[56], while hypoxia drove the upregulation of NHEJ-associated genes, *PRKDC* and *XRCC6,* in hepatoma cell lines[57]. The authors proposed an interaction between *PRKDC* and the hypoxia-responsive transcriptional activator, HIF-1α, hence suggesting that tumor hypoxia may lead to increase in NHEJ. Tumor cells within their 3D space are subjected to differential levels of oxygen over time and chronic exposures to these fluctuating conditions could result in very different biological outcomes. *In vitro* studies retain a significant caveat as many hypoxia assays are carried out short term using constant, predefined oxygen tensions. Although further work is needed to ascertain the clinical relevance of these findings, our results demonstrate that the integration of hypoxia assessment in molecular stratification using the 10-gene signature revealed a subset of high-risk individuals accounting for approximately 31% to 38%

in each cohort (Fig. 5B). Whether hypoxia could directly promote DNA damage repair *in vivo* remains an open question.

We reasoned that the expression patterns of DNA repair genes would positively correlate with genes involved in cell cycle progression since lesions could be repaired more effectively to prevent cell cycle arrest (Fig. 6A). Enhanced DNA repair ability may also confer tumor cells with a growth advantage. Consistent with this hypothesis, differential expression analyses between patients with high versus low 10-gene scores revealed an enrichment of ontologies involved in growth stimulation as a consequence of increased DNA repair gene expression (Fig. 6B). Enrichment of biological pathways involved in cell cycle, mitosis, cell division and DNA replication implied that the shorter life expectancy in patients with high 10-gene scores could in part be explained by an overproliferation phenotype commonly present in more aggressive tumors.

In summary, we developed a prognostic signature involving DNA repair genes and confirmed its utility as a powerful predictive marker for six cancer types. Although not currently afforded by this work due to its retrospective nature, it will be useful to determine if the signature can predict response to radiotherapy and chemotherapy in future research. While prospective validation is warranted, we would expect, based on our encouraging retrospective data, that the signature can guide decision making and treatment pathways. The confirmation of this hypothesis by a clinical trial using the 10-gene signature to select patients that would benefit from treatment with adjuvant DNA repair inhibitors could have a substantial impact on treatment outcomes.

Conflict of Interest: None declared.

Funding: None.

Authors contribution. WHC and AGL designed the study, analyzed the data and interpreted the data. AGL supervised the research. WHC and AGL wrote the initial manuscript draft. AGL revised the manuscript draft and approved the final version.

## References

1. Curtin NJ. DNA repair dysregulation from cancer driver to therapeutic target. Nat Rev Cancer. Nature Publishing Group; 2012;12:801.

2. Krishnakumar R, Kraus WL. The PARP side of the nucleus: molecular actions, physiological outcomes, and clinical targets. Mol Cell. Elsevier; 2010;39:8–24.

3. Sweasy JB, Lang T, DiMaio D. Is base excision repair a tumor suppressor mechanism? Cell cycle. Taylor & Francis; 2006;5:250–9.

4. Martin SA, Lord CJ, Ashworth A. Therapeutic targeting of the DNA mismatch repair pathway. Clin cancer Res. AACR; 2010;432–1078.

5. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JHJ. Understanding nucleotide excision repair and its roles in cancer and ageing. Nat Rev Mol cell Biol. Nature Publishing Group; 2014;15:465.

6. Usanova S, Piée-Staffa A, Sied U, Thomale J, Schneider A, Kaina B, et al. Cisplatin sensitivity of testis tumour cells is due to deficiency in interstrand-crosslink repair and low ERCC1-XPF expression. Mol Cancer. BioMed Central; 2010;9:248.

494 7. Rosell R, Taron M, Barnadas A, Scagliotti G, Sarries C, Roig B. Nucleotide excision repair
495 pathways involved in Cisplatin resistance in non-small-cell lung cancer. Cancer Control. SAGE
496 Publications Sage CA: Los Angeles, CA; 2003;10:297–305.
497 8. Martin LP, Hamilton TC, Schilder RJ. Platinum resistance: The role of DNA repair pathways.
498 Clin Cancer Res. 2008;14:1291–5.
499 9. Bryant HE, Schultz N, Thomas HD, Parker KM, Flower D, Lopez E, et al. Specific killing of
500 BRCA2-deficient tumours with inhibitors of poly (ADP-ribose) polymerase. Nature. Nature
501 Publishing Group; 2005;434:913.
502 10. Farmer H, McCabe N, Lord CJ, Tutt ANJ, Johnson DA, Richardson TB, et al. Targeting the
503 DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature. Nature Publishing
504 Group; 2005;434:917.
505 11. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The
506 cancer genome atlas pan-cancer analysis project. Nat Genet. Nature Publishing Group;
507 2013;45:1113.
508 12. Buffa FM, Harris AL, West CM, Miller CJ. Large meta-analysis of multiple cancers reveals a
509 common, compact and highly prognostic hypoxia metagene. Br J Cancer [Internet].
510 2010;102:428–35. Available from: http://dx.doi.org/10.1038/sj.bjc.6605450
511 13. Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A. GeneCodis3: a non-redundant
512 and modular enrichment analysis tool for functional genomics. Nucleic Acids Res. Oxford
513 University Press; 2012;40:W478--W483.
514 14. Kuleshov M V, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a
515 comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res.
516 Oxford University Press; 2016;44:W90--W97.
517 15. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and

518 collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. BioMed Central;
519 2013;14:128.
520 16. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New
521 York; 2016. Available from: http://ggplot2.org
522 17. Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R. InteractiVenn: a web-based
523 tool for the analysis of sets through Venn diagrams. BMC Bioinformatics. BioMed Central;
524 2015;16:169.
525 18. Wood RD, Mitchell M, Sgouros J, Lindahl T. Human DNA repair genes. Science (80- ).
526 American Association for the Advancement of Science; 2001;291:1284–9.
527 19. Chang WH, Forde D, Lai AG. A novel signature derived from immunoregulatory and
528 hypoxia genes predicts prognosis in liver and five other cancers. J Transl Med [Internet].
529 2019;17:14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30626396
530 20. Chang WH, Forde D, Lai AG. Dual prognostic role for 2-oxoglutarate oxygenases in ten
531 diverse cancer types: Implications for cell cycle regulation and cell adhesion maintenance.
532 bioRxiv [Internet]. Cold Spring Harbor Laboratory; 2018; Available from:
533 https://www.biorxiv.org/content/early/2018/10/14/442947
534 21. Samanta D, Gilkes DM, Chaturvedi P, Xiang L, Semenza GL. Hypoxia-inducible factors are
535 required for chemotherapy resistance of breast cancer stem cells. Proc Natl Acad Sci.
536 National Acad Sciences; 2014;111:E5429--E5438.
537 22. Semenza GL. Hypoxia-inducible factors: mediators of cancer progression and targets for
538 cancer therapy. Trends Pharmacol Sci. Elsevier; 2012;33:207–14.
539 23. Sutherland RL, Musgrove EA. Cyclin D1 and mammary carcinoma: new insights from
540 transgenic mouse models. Breast cancer Res. BioMed Central; 2001;4:14.
541 24. Buckley MF, Sweeney KJ, Hamilton JA, Sini RL, Manning DL, Nicholson RI, et al. Expression

542 and amplification of cyclin genes in human breast cancer. Oncogene. 1993;8:2127–33.
543 25. Musgrove EA, Caldon CE, Barraclough J, Stone A, Sutherland RL. Cyclin D as a therapeutic
544 target in cancer. Nat Rev Cancer. Nature Publishing Group; 2011;11:558.
545 26. Schwartz GK, Shah MA. Targeting the cell cycle: a new approach to cancer therapy. J Clin
546 Oncol. American Society of Clinical Oncology; 2005;23:9408–21.
547 27. Finn RS, Crown JP, Lang I, Boer K, Bondarenko IM, Kulyk SO, et al. The cyclin-dependent
548 kinase 4/6 inhibitor palbociclib in combination with letrozole versus letrozole alone as first-
549 line treatment of oestrogen receptor-positive, HER2-negative, advanced breast cancer
550 (PALOMA-1/TRIO-18): a randomised phase 2 study. Lancet Oncol. Elsevier; 2015;16:25–35.
551 28. Helleday T, Petermann E, Lundin C, Hodgson B, Sharma RA. DNA repair pathways as
552 targets for cancer therapy. Nat Rev Cancer. Nature Publishing Group; 2008;8:193.
553 29. Sengupta S, Harris CC. p53: traffic cop at the crossroads of DNA repair and
554 recombination. Nat Rev Mol cell Biol. Nature Publishing Group; 2005;6:44.
555 30. Petitjean A, Achatz MIW, Borresen-Dale AL, Hainaut P, Olivier M. TP53 mutations in
556 human cancers: functional selection and impact on cancer prognosis and outcomes.
557 Oncogene. Nature Publishing Group; 2007;26:2157.
558 31. Olivier M, Langer A, Carrieri P, Bergh J, Klaar S, Eyfjord J, et al. The clinical value of
559 somatic TP53 gene mutations in 1,794 patients with breast cancer. Clin cancer Res. AACR;
560 2006;12:1157–67.
561 32. Skinner HD, Sandulache VC, Ow TJ, Meyn RE, Yordy JS, Beadle BM, et al. TP53 disruptive
562 mutations lead to head and neck cancer treatment failure through inhibition of radiation-
563 induced senescence. Clin cancer Res. AACR; 2012;18:290–300.
564 33. Hoeijmakers JHJ. Genome maintenance mechanisms for preventing cancer. Nature.
565 Nature Publishing Group; 2001;411:366.

566 34. Hegi ME, Diserens A-C, Godard S, Dietrich P-Y, Regli L, Ostermann S, et al. Clinical trial
567 substantiates the predictive value of O-6-methylguanine-DNA methyltransferase promoter
568 methylation in glioblastoma patients treated with temozolomide. Clin cancer Res. AACR;
569 2004;10:1871–4.
570 35. Aye Y, Li M, Long MJC, Weiss RS. Ribonucleotide reductase and cancer: biological
571 mechanisms and targeted therapies. Oncogene. Nature Publishing Group; 2015;34:2011.
572 36. Shah KN, Wilson EA, Malla R, Elford HL, Faridi JS. Targeting ribonucleotide reductase M2
573 and NF-$κ$B activation with Didox to circumvent tamoxifen resistance in breast cancer. Mol
574 Cancer Ther. AACR; 2015;14:2411–21.
575 37. Kretschmer C, Sterner-Kock A, Siedentopf F, Schoenegg W, Schlag PM, Kemmner W.
576 Identification of early molecular markers for breast cancer. Mol Cancer. BioMed Central;
577 2011;10:15.
578 38. Grade M, Hummon AB, Camps J, Emons G, Spitzner M, Gaedcke J, et al. A genomic
579 strategy for the functional validation of colorectal cancer genes identifies potential
580 therapeutic targets. Int J cancer. Wiley Online Library; 2011;128:1069–79.
581 39. Mazzu YZ, Armenia J, Chakraborty G, Yoshikawa Y, Coggins S, Nandakumar S, et al. A novel
582 mechanism driving poor-prognosis prostate cancer: overexpression of the DNA repair gene,
583 ribonucleotide reductase small subunit M2 (RRM2). Clin Cancer Res. United States; 2019;
584 40. Schmutte C, Marinescu RC, Sadoff MM, Guerrette S, Overhauser J, Fishel R. Human
585 exonuclease I interacts with the mismatch repair protein hMSH2. Cancer Res. AACR;
586 1998;58:4537–42.
587 41. Zhou J, Wang Y, Wang Y, Yin X, He Y, Chen L, et al. FOXM1 modulates cisplatin sensitivity
588 by regulating EXO1 in ovarian cancer. PLoS One. Public Library of Science; 2014;9:e96989.
589 42. de Sousa JF, Torrieri R, Serafim RB, Di Cristofaro LFM, Escanfella FD, Ribeiro R, et al.

590 Expression signatures of DNA repair genes correlate with survival prognosis of astrocytoma
591 patients. Tumor Biol. SAGE Publications Sage UK: London, England;
592 2017;39:1010428317694552.
593 43. Dai Y, Tang Z, Yang Z, Zhang L, Deng Q, Zhang X, et al. EXO1 overexpression is associated
594 with poor prognosis of hepatocellular carcinoma patients. Cell Cycle. Taylor & Francis;
595 2018;17:2386–97.
596 44. Moeller BJ, Yordy JS, Williams MD, Giri U, Raju U, Molkentine DP, et al. DNA repair
597 biomarker profiling of head and neck cancer: Ku80 expression predicts locoregional failure
598 and death following radiotherapy. Clin Cancer Res. AACR; 2011;17:2035–43.
599 45. Grabsch H, Dattani M, Barker L, Maughan N, Maude K, Hansen O, et al. Expression of DNA
600 double-strand break repair proteins ATM and BRCA1 predicts survival in colorectal cancer.
601 Clin Cancer Res. AACR; 2006;12:1494–500.
602 46. Ma Q, Li P, Xu M, Yin J, Su Z, Li W, et al. Ku80 is highly expressed in lung adenocarcinoma
603 and promotes cisplatin resistance. J Exp Clin Cancer Res. BioMed Central; 2012;31:99.
604 47. Gu Z, Li Y, Yang X, Yu M, Chen Z, Zhao C, et al. Overexpression of CLC-3 is regulated by
605 XRCC5 and is a poor prognostic biomarker for gastric cancer. J Hematol Oncol. BioMed
606 Central; 2018;11:115.
607 48. Ho V, Chung L, Singh A, Lea V, Abubakar A, Lim SH, et al. Overexpression of the MRE11-
608 RAD50-NBS1 (MRN) complex in rectal cancer correlates with poor response to neoadjuvant
609 radiotherapy and prognosis. BMC Cancer. BioMed Central; 2018;18:869.
610 49. Singh P, Yang M, Dai H, Yu D, Huang Q, Tan W, et al. Overexpression and hypomethylation
611 of flap endonuclease 1 gene in breast and other cancers. Mol Cancer Res. AACR;
612 2008;6:1710–7.
613 50. Nikolova T, Christmann M, Kaina B. FEN1 is overexpressed in testis, lung and brain

614 tumors. Anticancer Res. International Institute of Anticancer Research; 2009;29:2453–9.
615 51. LaTulippe E, Satagopan J, Smith A, Scher H, Scardino P, Reuter V, et al. Comprehensive
616 gene expression analysis of prostate cancer reveals distinct transcriptional programs
617 associated with metastatic disease. Cancer Res. AACR; 2002;62:4499–506.
618 52. Wang K, Xie C, Chen D. Flap endonuclease 1 is a promising candidate biomarker in gastric
619 cancer and is involved in cell proliferation and apoptosis. Int J Mol Med. Spandidos
620 Publications; 2014;33:1268–74.
621 53. Abdel-Fatah TMA, Russell R, Albarakati N, Maloney DJ, Dorjsuren D, Rueda OM, et al.
622 Genomic and protein expression analysis reveals flap endonuclease 1 (FEN1) as a key
623 biomarker in breast and ovarian cancer. Mol Oncol. Wiley Online Library; 2014;8:1326–38.
624 54. Zhang K, Keymeulen S, Nelson R, Tong TR, Yuan Y-C, Yun X, et al. Overexpression of flap
625 endonuclease 1 correlates with enhanced proliferation and poor prognosis of Non--Small-Cell
626 lung cancer. Am J Pathol. Elsevier; 2018;188:242–51.
627 55. Bindra RS, Schaffer PJ, Meng A, Woo J, Måseide K, Roth ME, et al. Down-regulation of
628 Rad51 and decreased homologous recombination in hypoxic cancer cells. Mol Cell Biol. Am
629 Soc Microbiol; 2004;24:8504–18.
630 56. Meng AX, Jalali F, Cuddihy A, Chan N, Bindra RS, Glazer PM, et al. Hypoxia down-regulates
631 DNA double strand break repair gene expression in prostate cancer cells. Radiother Oncol.
632 Elsevier; 2005;76:168–76.
633 57. Um JH, Kang CD, Bae JH, Shin GG, Kim DW, Chung BS, et al. Association of DNA-
634 dependent protein kinase with hypoxia inducible factor-1 and its implication in resistance to
635 anticancer drugs in hypoxic tumor cells. Exp Mol Med. Nature Publishing Group;
636 2004;36:233.
637 58. Chang WH, Lai AG. Aberrations in Notch-Hedgehog signalling reveal cancer stem cells

638 harbouring conserved oncogenic properties associated with hypoxia and immunoevasion.
639 bioRxiv [Internet]. Cold Spring Harbor Laboratory; 2019; Available from:
640 https://www.biorxiv.org/content/early/2019/01/21/526202
641 59. Chang WH, Lai AG. Pan-cancer genomic amplifications underlie a Wnt hyperactivation
642 phenotype associated with stem cell-like features leading to poor prognosis. bioRxiv
643 [Internet]. Cold Spring Harbor Laboratory; 2019; Available from:
644 https://www.biorxiv.org/content/early/2019/01/13/519611
645 60. Chang WH, Lai AG. Timing gone awry: distinct tumour suppressive and oncogenic roles of
646 the circadian clock and crosstalk with hypoxia signalling in diverse malignancies. bioRxiv
647 [Internet]. Cold Spring Harbor Laboratory; 2019; Available from:
648 https://www.biorxiv.org/content/early/2019/02/21/556878
649 61. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer
650 genomics portal: an open platform for exploring multidimensional cancer genomics data.
651 AACR; 2012.
652 62. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of
653 complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. American
654 Association for the Advancement of Science; 2013;6:pl1--pl1.
655

## Figure legends

**Figure 1.** Schematic representation of the study design and development of the 10-gene signature. DNA repair genes from six major pathways were manually curated to generate a non-redundant list containing 138 genes. Cox proportional hazards regression was employed to determine the significance of each gene in predicting overall survival in 16 cancer types. Spearman's correlation analyses revealed that five cancer types exhibited a high degree of correlation in terms of their prognostic genes. Ten genes were found to be prognostic in all five cancers; these genes subsequently formed the 10-gene signature. The ability of the signature in predicting survival outcomes was tested using Kaplan-Meier, Cox regression and receiver operating characteristic methods. The signature could predict high-risk patients in six cancer types (n=2,257). Associations of the signature with tumor hypoxia, cell cycle deregulation and *TP53* mutation were investigated. Potential clinical applications of the signature were proposed.

**Figure 2.** Patient stratification using the 10-gene signature in six cancer types. Kaplan-Meier analyses of overall survival on patients stratified into high- and low-score groups using the 10-gene signature. P values were determined from the log-rank test.

**Figure 3.** Independence of the 10-gene signature over TNM staging. Kaplan-Meier analyses were performed on patients categorized according to tumor TNM stages that were further stratified using the 10-gene signature. The signature successfully identified patients at higher risk of death in all TNM stages. P values were determined from the log-rank test. TNM: tumor, node, metastasis.

680

681 **Figure 4.** Predictive performance of the 10-gene signature. Receiver operating characteristic
682 (ROC) was employed to determine the specificity and sensitivity of the signature in predicting
683 5-year overall survival in all six cancer types. ROC curves generated based on the 10-gene
684 signature, TNM staging and a combination of 10-gene signature and TNM staging were
685 depicted. AUC: area under the curve. AUC values for TNM staging employing TCGA datasets
686 were in accordance with our previously published work[19,58–60]. TNM: tumor, node,
687 metastasis.

688

689 **Figure 5.** Association between the 10-gene signature and tumor hypoxia. **(A)** Scatter plots
690 depict significant positive correlations between 10-gene scores and hypoxia scores in all six
691 cancers. Patients were color-coded and separated into four categories based on their 10-gene
692 and hypoxia scores. **(B)** Kaplan-Meier analyses were performed on the four patient categories
693 to assess the effects of the combined relationship of hypoxia and the signature on overall
694 survival.

695

696 **Figure 6.** Elevated DNA repair gene expression is associated with an overproliferation
697 phenotype. **(A)** Significant positive correlations between individual signature gene expression
698 and genes involved in cell cycle progression, while negative correlations were observed with
699 genes involved in cell cycle arrest. Heatmaps were generated using the R pheatmap package.
700 Cell cycle genes were depicted on the y-axis and the 10 signature genes on the x-axis. **(B)**
701 Patients were median-stratified into low- and high-score groups using the 10-gene signature
702 for differential expression analyses. Enrichment of GO and KEGG pathways associated with
703 differentially expressed genes were depicted for all six cancers. **(C)** Investigation of the

relationship between a gene involved in cell cycle checkpoint regulation, *TP53*, and the signature. Patients were categorized into four groups based on their *TP53* mutation status and 10-gene scores for Kaplan-Meier analyses. P values were determined from the log-rank test. Positions of individual mutation types were indicated and color-coded in the mutation diagram generated using cBioPortal[61,62].

**Table 1.** Univariate and multivariate Cox proportional hazards analyses of the 10-gene signature and additional clinical risk factors associated with overall survival in six cancers. Univariate values for TNM staging employing TCGA datasets were in accordance with our previously published work[19,58–60].

**Table 2.** Univariate Cox proportional hazards analysis of the relation between the 10-gene signature and hypoxia score.

**Table 3.** Univariate Cox proportional hazards analysis of the relation between the 10-gene signature and *TP53* mutation status.

720 Supplementary information

721

722 **Figure S1.** Correlation analyses of 138 prognostic DNA repair genes. Spearman's correlation
723 coefficients were determined from pairwise comparisons prognostic genes from 16 cancer
724 types. Five cancers were highly correlated as shown in the blue area of the heatmap. Numbers
725 represent correlation coefficient values. Refer to Table S2 for cancer abbreviations.

726

727 **Figure S2.** Expression distribution of the ten signature genes in tumor and non-tumor samples.
728 Boxplots overlaying violin plots were used to illustrate tumor and non-tumor distribution in six
729 cancers: **(A)** clear cell renal cell, **(B)** papillary renal cell, **(C)** liver, **(D)** lung, **(E)** pancreas and **(F)**
730 endometrium. Nonparametric Mann-Whitney-Wilcoxon tests were employed to determine
731 whether there were significant differences in expression distributions. Asterisks represent
732 significant P values: * < 0.05, *** < 0.0001.

733

734 **Figure S3**. Venn diagram depicts a six-way comparison of the differentially expressed genes (-
735 $1 > \log_2$ fold-change $> 1$, $P<0.05$) identified from high-score versus low-score patients in all six
736 cancers. Numbers in parentheses represent the number of differentially expressed genes in
737 each cancer. The Venn intersection of all cancers indicated that 87 genes were common.

738

739 **Table S1.** List of 138 DNA repair genes and associated pathways.

740

741 **Table S2.** Description of TCGA cancer cohorts.

742

743 **Table S3.** Univariate Cox proportional hazards analysis of the 138 genes in 16 cancers.

744

745 **Table S4.** Differentially expressed genes between high- and low-score patient groups in six
746 cancers.

747

748 **Table S5.** List of 87 differentially expressed genes that are common in all six cancers.

749

750 **Table S6.** Enrichr transcription factor protein-protein interaction analysis of the 87
751 differentially expressed genes.

752

753 **Table S7.** *TP53* mutation analysis in liver, pancreatic, endometrial and lung cancers.

# Figure 1

## DNA repair genes in humans

| DNA repair pathways | No. of genes |
|---|---|
| Base excision repair | 33 |
| Mismatch repair | 23 |
| Nucleotide excision repair | 39 |
| Homologous recombination | 26 |
| Non-homologous end joining | 13 |
| Fanconi anemia | 53 |

Total = 187 genes
49 genes are present in 2 or more pathways.
**138 genes are non-redundant.**

## Gene signature development

Cox regression on 138 genes in 16 cancers

↓

Spearman's correlation of prognostic genes between cancers.
Five cancer are highly correlated.

↓

Ten prognostic genes are present in all five cancers. Develop a signature based on 10 genes.

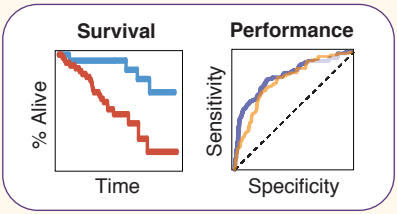
↓

Test signature in 16 cancers. Signature is prognostic in 6 cancers.

## Prognosis in 6 cancers

| Prognostic cohorts | Patient no. |
|---|---|
| Renal clear cell | 533 |
| Renal papillary cell | 290 |
| Liver | 371 |
| Lung | 515 |
| Pancreas | 178 |
| Endometrium | 370 |
| **Total** | **2,257** |

## Performance of signature




## Association with tumor biology

Tumor hypoxia
Cell cycle deregulation
*TP53* mutation

Diagnosis

↓

Risk stratification with 10-gene signature


## Clinical application

Alter clinical care and treatment plan

↓

DNA repair inhibitors (eg. PARP inhibitors).

↓

Allows the accumulation of fatal replication lesions contributing to the death of cancer cells.

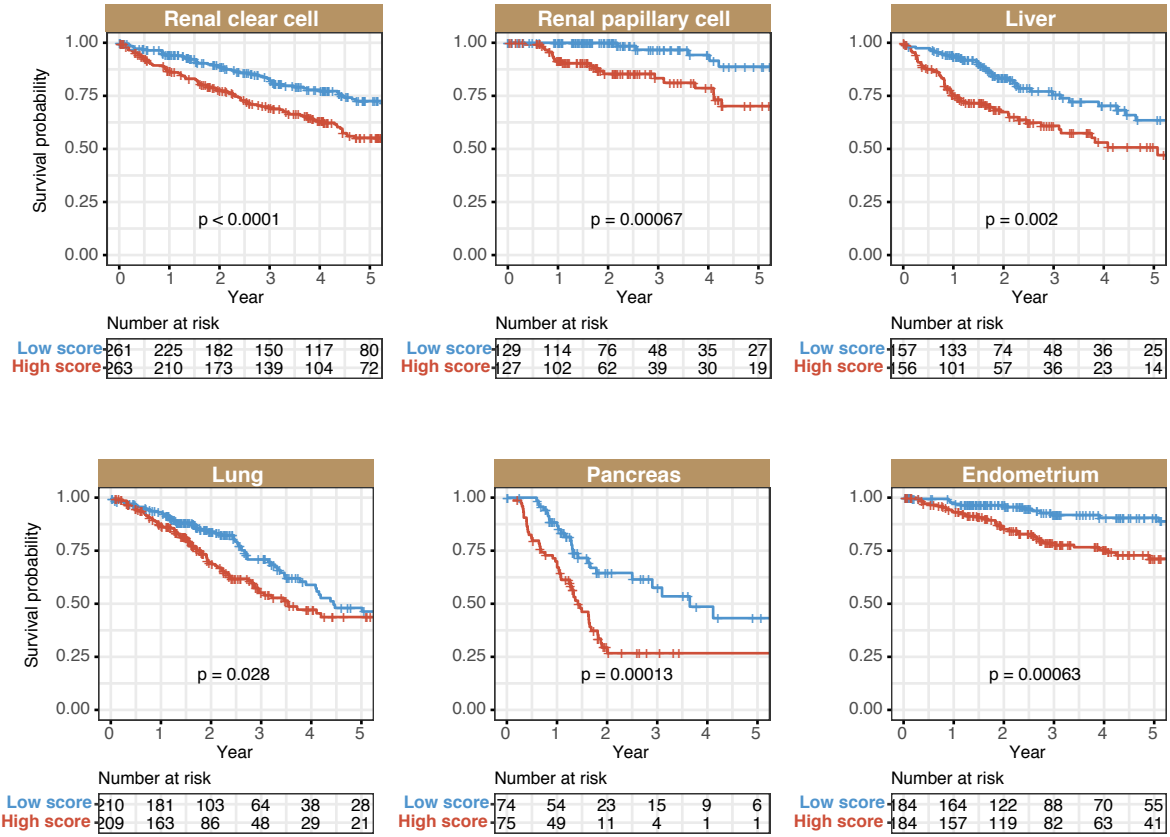
↓

Alter tumor chemo- and radiosensitivity

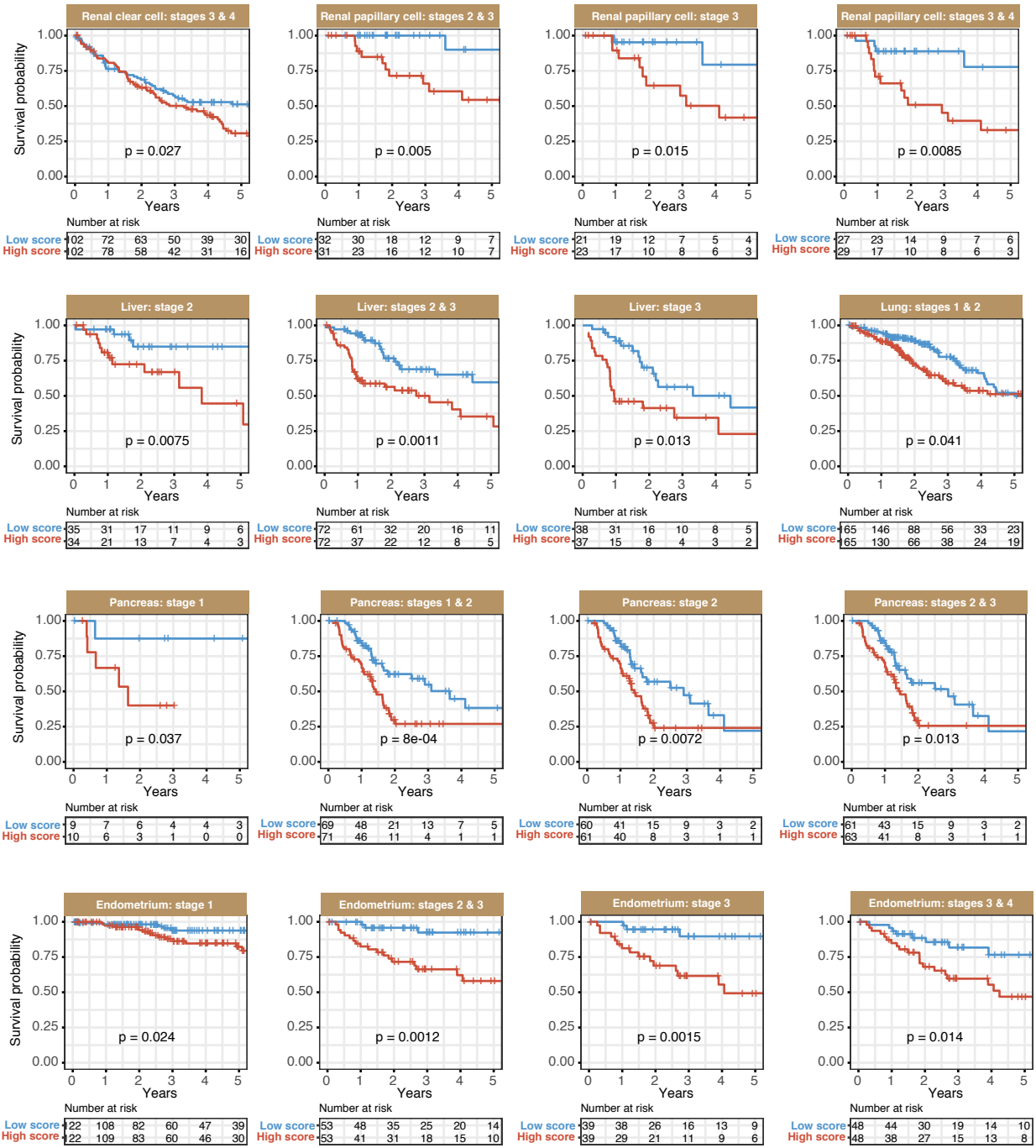
↓

Improved prognosis

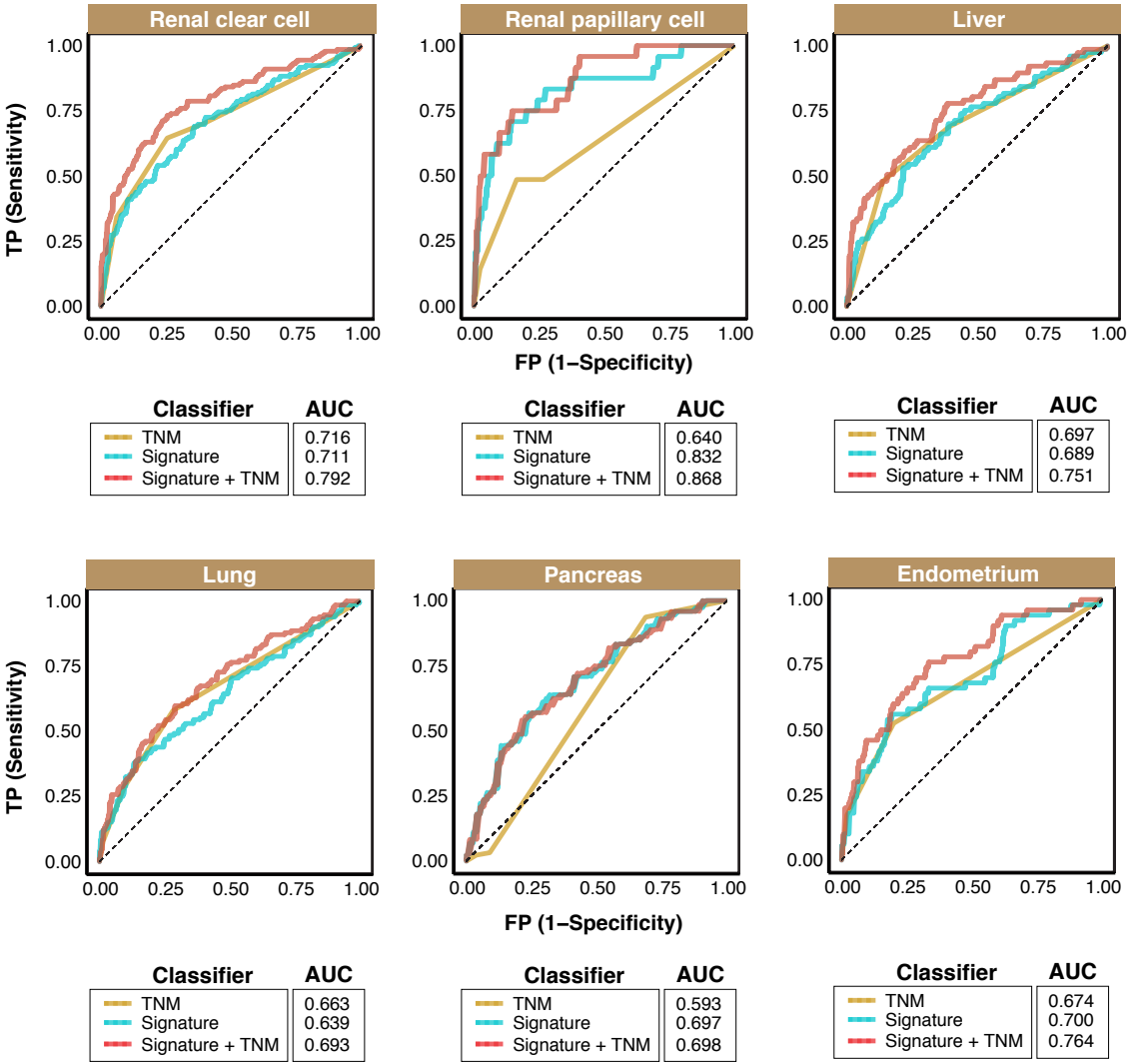
# Figure 2


Renal clear cell
Survival probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0 1 2 3 4 5
Year
Number at risk
Low score 261 225 182 150 117 80
High score 263 210 173 139 104 72
Renal papillary cell
p = 0.00067
Number at risk
Low score 129 114 76 48 35 27
High score 127 102 62 39 30 19
Liver
p = 0.002
Number at risk
Low score 157 133 74 48 36 25
High score 156 101 57 36 23 14
Lung
p = 0.028
Number at risk
Low score 210 181 103 64 38 28
High score 209 163 86 48 29 21
Pancreas
p = 0.00013
Number at risk
Low score 74 54 23 15 9 6
High score 75 49 11 4 1 1
Endometrium
p = 0.00063
Number at risk
Low score 184 164 122 88 70 55
High score 184 157 119 82 63 41

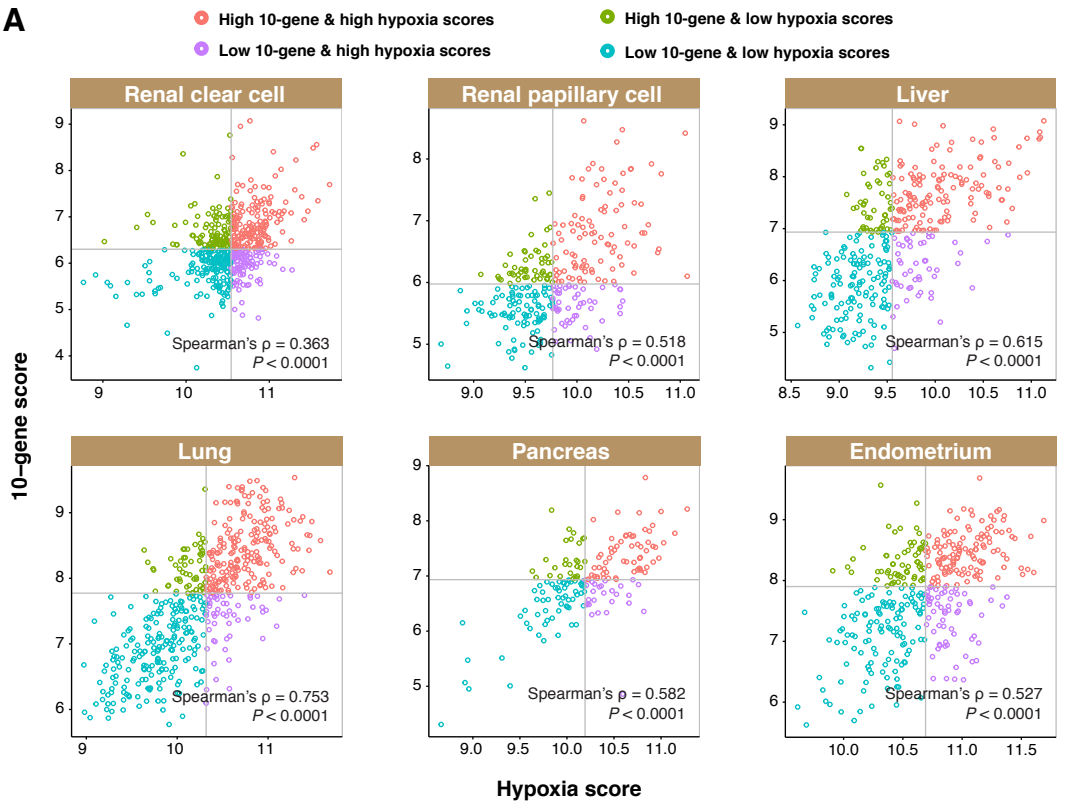
# Figure 3

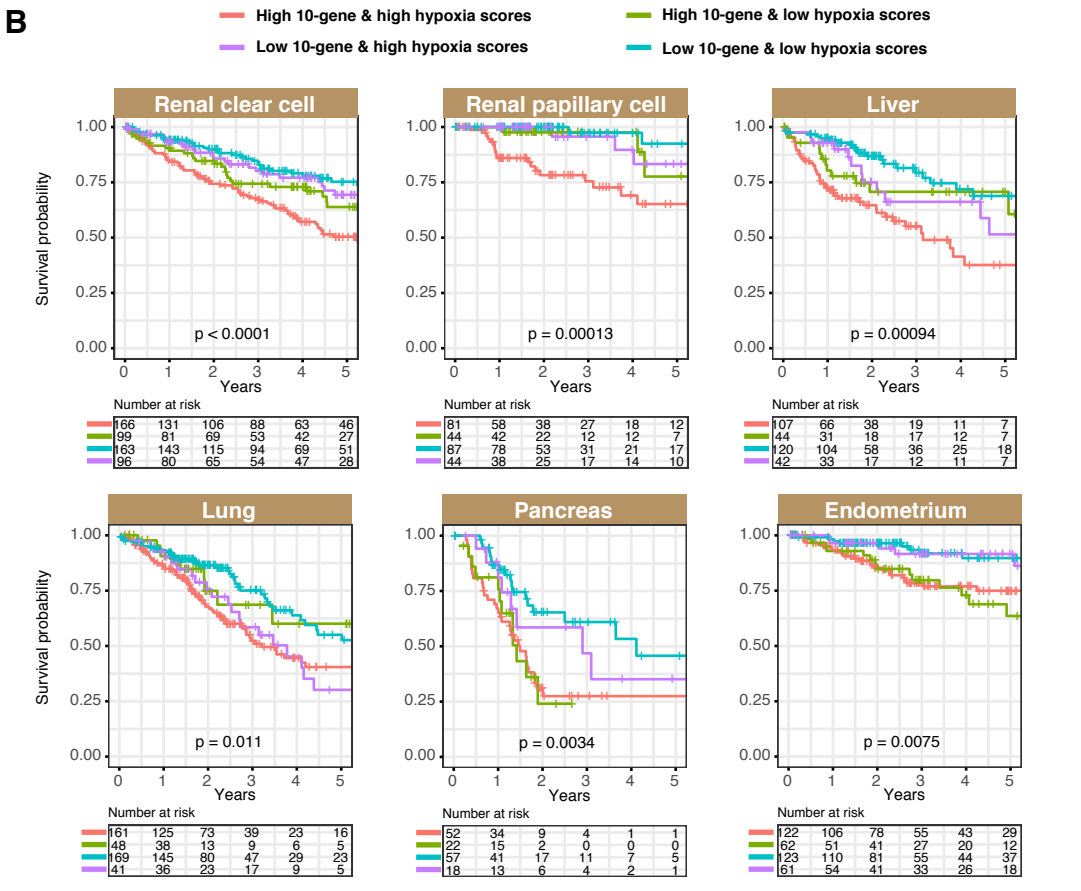

Renal clear cell: stages 3 & 4
p = 0.027
Number at risk
Low score 102 72 63 50 39 30
High score 102 78 58 42 31 16
Renal papillary cell: stages 2 & 3
p = 0.005
Number at risk
Low score 32 30 18 12 9 7
High score 31 23 16 12 10 7
Renal papillary cell: stage 3
p = 0.015
Number at risk
Low score 21 19 12 7 5 4
High score 23 17 10 8 6 3
Renal papillary cell: stages 3 & 4
p = 0.0085
Number at risk
Low score 27 23 14 9 7 6
High score 29 17 10 8 6 3
Liver: stage 2
p = 0.0075
Number at risk
Low score 35 31 17 11 9 6
High score 34 21 13 7 4 3
Liver: stages 2 & 3
p = 0.0011
Number at risk
Low score 72 61 32 20 16 11
High score 72 37 22 12 8 5
Liver: stage 3
p = 0.013
Number at risk
Low score 38 31 16 10 8 5
High score 37 15 8 4 3 2
Lung: stages 1 & 2
p = 0.041
Number at risk
Low score 65 146 88 56 33 23
High score 65 130 66 38 24 19
Pancreas: stage 1
p = 0.037
Number at risk
Low score 9 7 6 4 4 3
High score 10 6 3 1 0 0
Pancreas: stages 1 & 2
p = 8e-04
Number at risk
Low score 69 48 21 13 7 5
High score 71 46 11 4 1 1
Pancreas: stage 2
p = 0.0072
Number at risk
Low score 60 41 15 9 3 2
High score 61 40 8 3 1 1
Pancreas: stages 2 & 3
p = 0.013
Number at risk
Low score 61 43 15 9 3 2
High score 63 41 8 3 1 1
Endometrium: stage 1
p = 0.024
Number at risk
Low score 22 108 82 60 47 39
High score 22 109 83 60 46 30
Endometrium: stages 2 & 3
p = 0.0012
Number at risk
Low score 53 48 35 25 20 14
High score 53 41 31 18 15 10
Endometrium: stage 3
p = 0.0015
Number at risk
Low score 39 38 26 16 13 9
High score 39 29 21 11 9 6
Endometrium: stages 3 & 4
p = 0.014
Number at risk
Low score 48 44 30 19 14 10
High score 48 38 27 15 13 8
Survival probability
Years

# Figure 4

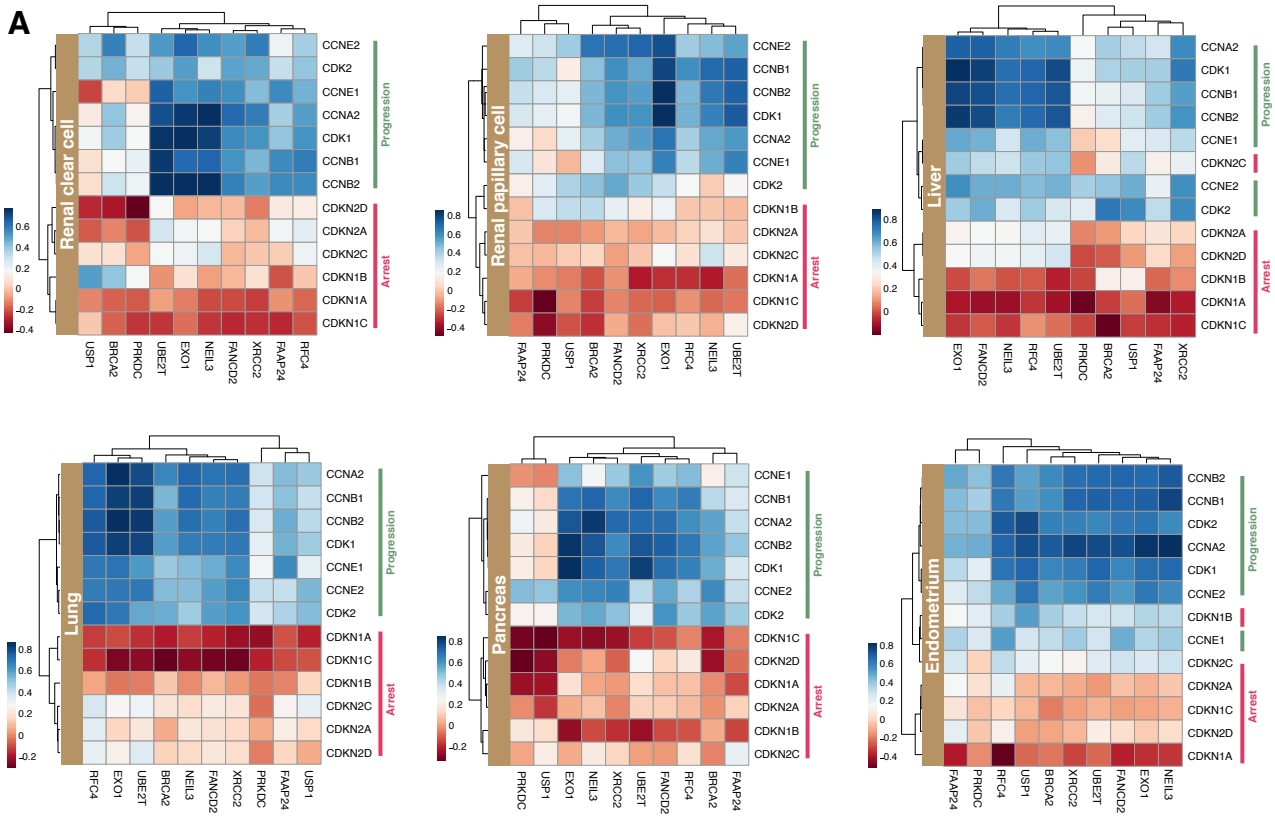

Renal clear cell
Renal papillary cell
Liver
Lung
Pancreas
Endometrium
TP (Sensitivity)
FP (1−Specificity)
0.00
0.25
0.50
0.75
1.00
Classifier
AUC
TNM
Signature
Signature + TNM
0.716
0.711
0.792
0.640
0.832
0.868
0.697
0.689
0.751
0.663
0.639
0.693
0.593
0.697
0.698
0.674
0.700
0.764

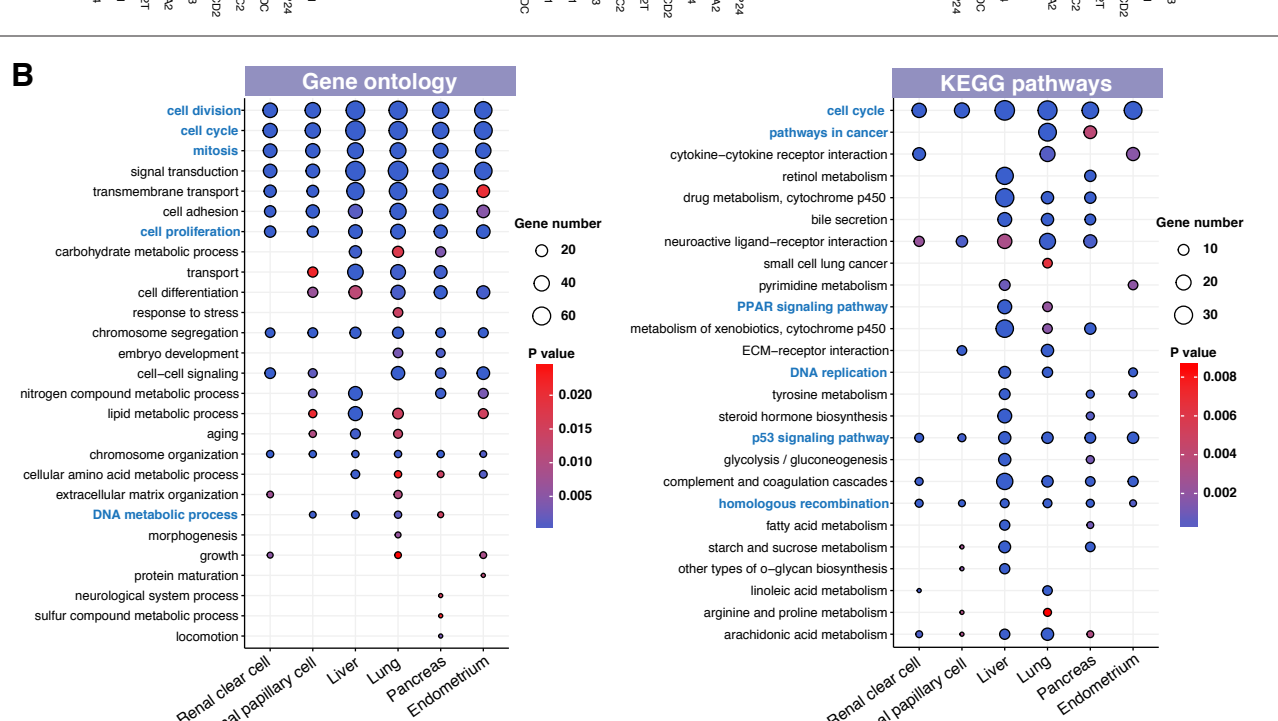
# Figure 5

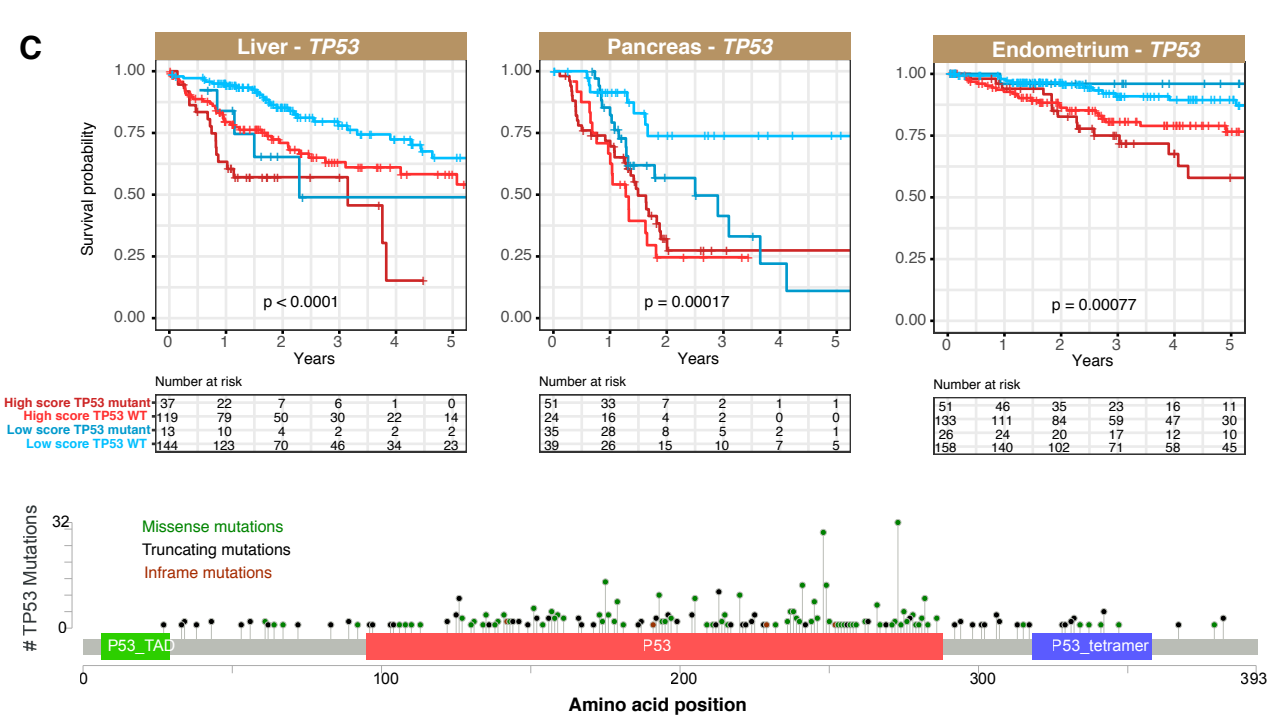

A
High 10-gene & high hypoxia scores
High 10-gene & low hypoxia scores
Low 10-gene & high hypoxia scores
Low 10-gene & low hypoxia scores
Renal clear cell
Spearman's ρ = 0.363
P < 0.0001
Renal papillary cell
Spearman's ρ = 0.518
P < 0.0001
Liver
Spearman's ρ = 0.615
P < 0.0001
Lung
Spearman's ρ = 0.753
P < 0.0001
Pancreas
Spearman's ρ = 0.582
P < 0.0001
Endometrium
Spearman's ρ = 0.527
P < 0.0001
10−gene score
Hypoxia score


B
High 10-gene & high hypoxia scores
High 10-gene & low hypoxia scores
Low 10-gene & high hypoxia scores
Low 10-gene & low hypoxia scores
Renal clear cell
p < 0.0001
Renal papillary cell
p = 0.00013
Liver
p = 0.00094
Lung
p = 0.011
Pancreas
p = 0.0034
Endometrium
p = 0.0075
Survival probability
Years
Number at risk

# Figure 6


A
Renal clear cell
Renal papillary cell
Liver
Lung
Pancreas
Endometrium
Progression
Arrest
CCNE2
CDK2
CCNE1
CCNA2
CDK1
CCNB1
CCNB2
CDKN2D
CDKN2A
CDKN2C
CDKN1B
CDKN1A
CDKN1C
USP1
BRCA2
PRKDC
UBE2T
EXO1
NEIL3
FANCD2
XRCC2
FAAP24
RFC4


B
Gene ontology
cell division
cell cycle
mitosis
signal transduction
transmembrane transport
cell adhesion
cell proliferation
carbohydrate metabolic process
transport
cell differentiation
response to stress
chromosome segregation
embryo development
cell−cell signaling
nitrogen compound metabolic process
lipid metabolic process
aging
chromosome organization
cellular amino acid metabolic process
extracellular matrix organization
DNA metabolic process
morphogenesis
growth
protein maturation
neurological system process
sulfur compound metabolic process
locomotion
Gene number
20
40
60
P value
0.020
0.015
0.010
0.005
KEGG pathways
cell cycle
pathways in cancer
cytokine−cytokine receptor interaction
retinol metabolism
drug metabolism, cytochrome p450
bile secretion
neuroactive ligand−receptor interaction
small cell lung cancer
pyrimidine metabolism
PPAR signaling pathway
metabolism of xenobiotics, cytochrome p450
ECM−receptor interaction
DNA replication
tyrosine metabolism
steroid hormone biosynthesis
p53 signaling pathway
glycolysis / gluconeogenesis
complement and coagulation cascades
homologous recombination
fatty acid metabolism
starch and sucrose metabolism
other types of o−glycan biosynthesis
linoleic acid metabolism
arginine and proline metabolism
arachidonic acid metabolism
10
20
30
0.008
0.006
0.004
0.002
Renal clear cell
Renal papillary cell
Liver
Lung
Pancreas
Endometrium


C
Liver - TP53
Pancreas - TP53
Endometrium - TP53
Survival probability
Years
p < 0.0001
p = 0.00017
p = 0.00077
Number at risk
High score TP53 mutant
High score TP53 WT
Low score TP53 mutant
Low score TP53 WT
37 22 7 6 1 0
119 79 50 30 22 14
13 10 4 2 2 2
144 123 70 46 34 23
51 33 7 2 1 1
24 16 4 2 0 0
35 28 8 5 2 1
39 26 15 10 7 5
51 46 35 23 16 11
133 111 84 59 47 30
26 24 20 17 12 10
158 140 102 71 58 45
# TP53 Mutations
Missense mutations
Truncating mutations
Inframe mutations
P53_TAD
P53
P53_tetramer
Amino acid position