Peer-reviewed version available at *Cancers* **2019**, *11*, 501; doi:10.3390/cancers11040501

Article

DNA Repair Gene Expression Adjusted by the PCNA Metagene Predicts Survival in Multiple Cancers

Leif E. Peterson [1,2], Tatiana Kovyrshina[2]

[1] Department of Healthcare Policy and Research, Weill Cornell Medical College, Cornell University, New York City, New York 10065 USA

[2] Center for Biostatistics, Institute for Academic Medicine, Houston Methodist Research Institute, 6565 Fannin Street, Houston, Texas 77030 USA

* Correspondence: peterson.leif.e@gmail.com; Tel.: +001 (281) 381-6218

Abstract: Removal of the proliferation component of gene expression by PCNA adjustment has been addressed in numerous survival prediction studies for breast cancer and all cancers in the TCGA. These studies indicate that widespread co-regulation of proliferation upwardly biases survival prediction when gene selection is performed on a genome-wide basis. In addition, removal of the correlative effects of proliferation does not reduce the random bias associated with survival prediction using random gene selection. Since most cancers become addicted to DNA repair as a result of forced cellular replication, increased oxidation, and repair deficiencies from oncogenic loss or genetic polymorphisms, we pursued an investigation to remove the proliferation component of expression in DNA repair genes to determine survival prediction. This translational hypothesis-driven focus on DNA repair genes is directly amenable to finding new sets of DNA repair genes that could potentially be studied for inhibition therapy. Overall survival (OS) prediction was evaluated in 18 cancers by using normalized RNA-Seq data for 126 DNA repair genes with expression available in TCGA. Transformations for normality and adjustments for age at diagnosis, stage, and PCNA metagene expression were performed for all DNA repair genes. We also analyzed genomic event rates (GER) for somatic mutations, deletions, and amplification in driver genes and DNA repair genes. After performing empirical p-value testing with use of randomly selected gene sets, it was observed that OS could be predicted significantly by sets of DNA repair genes for 61% (11/18) of the cancers. Interestingly, PARP1 was not a significant predictor of survival for any of the 11 cancers. Results from cluster analysis of GERs indicates that the most opportunistic cancers for inhibition therapy may be AML, colorectal, and renal papillary, because of potentially less confounding due to lower GERs for mutations, deletions, and amplifications in DNA repair genes. However, the most opportunistic cancer for inhibition therapy is likely to be AML, since it showed the lowest GERs for mutations, deletions, and amplifications in DNA repair genes. In conclusion, our hypothesis-driven focus to target DNA repair gene expression adjusted for the PCNA metagene as a means of predicting OS in various cancers resulted in statistically significant sets of genes.

Keywords: RNA-Seq; Oncology; DNA repair; Survival; PCNA metagene

1. Introduction

Genomic instability is an important hallmark of cancer which leads to mutations that dysregulate cellular growth [1,2]. Mutations play an important role in oncogenic transformation and can be catastrophic during mitosis [3,4]. The additional replication stress and increased oxidative damage that arises from continuous forced cell division in tumor cells requires several DNA repair components [5,6]. However, inherited genetic polymorphisms and oncogenic loss result in DNA repair deficiencies, and therefore alternative DNA repair pathways must be found if replication is to continue [7]. The addiction to alternative DNA repair pathways by cancer can therefore be targeted to prevent the repair and restart of stressed replication forks [8,9,10].

Peer-reviewed version available at Cancers 2019, 11, 501; doi:10.3390/cancers11040501

Genetic stability relies on DNA repair, which is a complex process that depends on several molecular pathways to correct damage to DNA. DNA damage ranges from minor mismatched bases and methylation events to oxidized bases, intra- and interstrand DNA crosslinks, protein-DNA adducts, double strand breaks (DSBs), and stalled forks. DNA repair pathways include mismatch repair, base excision repair, nucleotide excision repair, and the homology directed repair/Fanconi anemia pathway. Earlier forms of chemotherapy and radiation therapy focused on damaging DNA to promote excessive lethal mutations and cellular death; however, it has been demonstrated that cancer cells can repair therapy-induced DNA damage [11,12]. This has led to the concept of synthetic lethality and DNA repair inhibition, in which specific DNA repair pathways and their proteins are targeted for increasing sensitivity to traditional therapeutics [13,14].

Recently, it was reported that the proliferating cell nuclear antigen (*PCNA*) DNA repair protein is widely co-regulated in the genome [15,16]. *PCNA* is a ring-like protein which serves as a co-factor for polymerase δ, and surrounds DNA during strand synthesis to recruit proteins needed for DNA replication and repair [17]. *PCNA* by itself is not a tumor suppressor gene or oncogene, but rather is a proliferation promoting protein whose expression is upregulated during cell replication. Ge et al. [18] identified 131 mRNAs in 36 types of normal tissues whose expression correlated r>0.65 with expression of *PCNA*. Expression patterns of these 131 genes were collapsed into a median value (called *PCNA* "metagene") by Venet et al., who removed the proliferation effect by performing multivariate analysis of expression profiles published by 47 breast cancer (survival) studies [16]. Their results indicated that 91% of the genes predictive of survival were significantly correlated with the *PCNA* metagene. Shimoni also investigated random bias and *PCNA* adjustment of RNA-Seq expression during survival analysis for 34 cancers from TCGA and reported that *PCNA* adjustment did not remove random bias, and that tumor sub-classification significantly improved survival prediction [19].

While genome-wide approaches have revealed that PCNA adjustment to expression cannot remove random bias and reduces survival prediction, a hypothesis-driven approach specifically focusing on removal of the proliferation component of expression in DNA repair genes used for survival analysis has eluded systematic investigation. We therefore pursued a candidate gene approach to determine survival prediction with DNA repair gene expression adjusted for *PCNA* using TCGA RNA-Seq data. The hypothesis was that *PCNA* adjustment would result in a significant association with survival in common cancers. Not all of the cancers available in TCGA were investigated because the body of information on DNA repair inhibition therapy is more strongly hinged to common cancers [20,21,14]. Another hypothesis was that patterns of somatic mutations, deletions, and amplifications in cancer-specific driver genes and the DNA repair genes considered would provide new insight into the patterns of genomic alteration observed in tumor cells [22]. Results of the computational analyses were used for generating lists of DNA repair genes whose upregulation was associated with shortened survival, which would be potentially amenable for inhibition therapy to prolong survival.

2. Data

Cancer data. The data used in this investigation were derived from genomic sequencing of tumors in The Cancer Genome Atlas (TCGA) [23]. We investigated DNA repair gene expression in 18 cancers for which genomic sequencing and RNA-seq expression data were available from cBio-Portal (http://www.cbioportal.org) [24,25]. Specifically, the cancers for which only age at diagnosis was available, and not pathological stage, in the TCGA data were acute myelogenous leukemia (AML), bladder, low grade gliomas, glioblastoma multiforme (GBM), head and neck, and sarcoma. Cancers for which both age at diagnosis and pathological stage were available included breast, cervical, colorectal, liver, lung, lung squamous cell, melanoma, ovarian, renal clear cell, renal papillary, stomach, and uterine. Altogether, this resulted in a total of 18 cancers that were considered.

Expression, mutations, deletions, and amplifications. RNA-Seq based normalized expression values for somatic mutations in DNA repair genes were also obtained from cBio-Portal

[24,25]. We also acquired high-confidence deletions and amplifications from cBio-Portal, where a deletion was defined as full homozygous loss with a GISTIC score [26] of -2, and an amplification was defined as high-level gain with a GISTIC score of 2. Low-level deletions (heterozygous loss) and low-level gain (low-level amplifications) with GISTIC scores of -1 and 1, respectively, were not used.

DNA Repair gene lists. We used the comprehensive list of DNA repair genes provided at https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html, which are described and published in [27,28,29,30,31,32,33]. Table S1 of the Supplemental Information lists the DNA repair genes used which were available in TCGA expression data. The various DNA repair mechanisms and pathways for repair genes used are described below.

Direct Reversal Repair, DRR. Direct reversal DNA repair (DRR) is a single step reaction of removal of the methyl- or photoadducts. DRR is provided to methylphosphotriesters (direct removal of the alkyl damage by nucleophilic Cys residues), alkyltransferases (repair of O6-alkylguanine by *MGMT*), oxidative dealkylation (*ALKB* proteins), and photolyase (direct reversal of the thymine dimer created by UV light in CTD photolyase and 6-4TT photolyase). The DRR genes used in this study were: *ALKBH2, ALKBH3*, and *MGMT*.

Base Excision Repair, BER. Base excision DNA repair (BER) corrects base lesions generated by oxidative, alkylation, deamination, and depurinatiation/depyrimidination damage. BER is initiated by DNA glycosylases, which recognize and catalyze removal of damaged bases. Downstream enzymes carry out strand incision, gap-filling, and ligation. BER involves two general pathways: short-patch (SP-BER) and long-patch (LP-BER). BER genes employed in this study included: *APEX1, APEX2, APTX, FEN1, HUS1, LIG1, LIG3, MBD4, MPG, MUTYH, NEIL1, NEIL2, NEIL3, NTH, OGG1, PARP1, PARP2, PCNA, PNKP, POLB, POLD1, POLE, POLH, POLL, RECQL2, SMUG1, TDG, UNG*, and *XRCC1*.

Non-homologous End-joining, NHEJ. Non-homologous end-joining (NHEJ) repairs DSBs at all stages of the cell-cycle, bringing about the ligation of two double-strand breaks (DSBs) without the need for sequence homology, and therefore NHEJ is error-prone. NHEJ is referred to as "non-homologous" because the break ends are directly ligated without the need for a homologous template, in contrast to homologous recombination, which requires a homologous sequence to guide repair. NHEJ typically utilizes short homologous DNA sequences called microhomologies, which are often present in single-stranded overhangs on the ends of DSBs. Inappropriate NHEJ can lead to translocations and telomere fusion, which are common hallmarks of tumor cells. The NHEJ genes considered in this investigation were: *DCLRE1C, XRCC6, XRCC5, LIG4, NHEJ1, POLL, POLM, PRKDC, RECQL2, XPF*, and *XRCC4*.

Mismatch Repair, MMR. DNA mismatch repair (MMR) is responsible for correction of replication errors (mismatches, small insertions, deletions, and microsatellites) that escape the proofreading activity of a DNA polymerase [34]. Cells harboring deficiencies in MMR can acquire microsatellite instability (MSI) following several cell divisions, and fail to regulate microsatellite lengths during cell division. MMR is initiated by two proteins homologous to MutS and MutL: MutS*pha* and MutLα. Mutations in the genes coding MutS and MutL homologs have been linked with the Lynch syndrome, which is characterized by an increased risk of developing cancer. MMR genes included in this study were: *EXO1, LIG1, MLH1, MLH3, MSH2, MSH3, MSH6, PMS1, PMS2*, and *POLD1*.

Translesion Synthesis, TLS. Translesion synthesis (TLS) is a process involving specialized DNA polymerases which replicate across from DNA lesions. TLS aids in resistance to DNA damage, presumably by restarting stalled replication forks or filling in gaps that remain in the genome due to the presence of DNA lesions. TLS has the potential to produce mutations. The TLS genes considered in this study were: *POLH, POLI, POLK, POLM, POLN, POLQ, REV1*, and *REV3L*.

DNA Damage Signaling, DDS. DNA damage induces several cellular responses including DNA repair, cell-cycle checkpoint activity, and triggering of apoptotic pathways. DNA damage checkpoints are associated with biochemical pathways that end delay or arrest of cell-cycle progression. Such checkpoints engage damage sensor proteins, such as the *RAD9-RAD1-HUS1* (9-1-1) complex, and the *RAD17-RFC* complex, in the detection of DNA damage and transduction of

signals to *ATM, ATR, CHK1* and *CHK2* kinases. In addition, *CHK1* and *CHK2* kinases regulate *CDC25*, p21 and p53 that ultimately inactivate cyclin-dependent kinases (CDKs), which inhibit cell-cycle progression. The DDS genes considered included: *ATM, ATR, ATRIP, BLM, BRCA1, CCNH, CDK7, CDKN1A, CHEK1, CHEK2, COPS5, DCLRE1A, DCLRE1B, FANCA, FANCC, GPS1, HUS1, MDC1, MNAT1, MRE11A, NBN, RAD1, RAD17, RAD18, RAD23A, RAD50, RAD9A, RFC1, RFC2, RFC3, RFC4, RFC5, TOPBP1*, and *TP53*.

Homologous Recombination Repair, HRR. Homologous recombination repair (HRR) is a type of genetic recombination in which nucleotide sequences are exchanged between two similar or identical molecules of DNA. HRR is an "error free" mechanism which acts on DSBs occurring within replicated DNA (replication-independent DSBs) or on DSBs that are generated at broken replication forks (replication-dependent DSBs). HRR also involves processing of the ends of the DNA double-strand break, homologous DNA pairing and strand exchange, repair DNA synthesis, and resolution of the heteroduplex molecules. The HRR genes used in this project included: *BLM, BRCA1, BRCA2, C19ORF40, EME1, EME2, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, MRE11A, MSH4, MSH5, MUS81, RAD51*, and *RAD52*.

Nucleotide Excision Repair, NER. Nucleotide excision repair (NER) removes UV-induced damage (thymine dimers and 6-4-photoproducts) as well as other kinds of DNA damage, which produce bulky distortions in the shape of DNA double helix. NER enzymes recognize bulky distortions in the shape of the DNA double helix, and only repair damaged bases that can be removed by a specific glycosylase. Specifically, NER entails removal of a short single-stranded DNA segment that includes the lesion, creating a single-strand gap (20-30 nucleotides) in the DNA, which is subsequently filled in by DNA polymerase d or e by copying the undamaged strand. Polymorphisms in NER proteins include Xeroderma Pigmentosum and Cockayne's syndrome. The NER genes considered in this study included: *CSA, CSB, CUL4A, DDB1, DDB2, ERCC1, GTF2H1, GTF2H2, GTF2H3, GTF2H4, GTF2H5, LIG1, MMS19, PCNA, POLD1, POLE, RAD23B, RFC1, RPA1, XPA, XPB, XPC, XPD, XPF*, and *XPG*.

2. Results

Table 1 lists sample sizes and cancer sites for which Kaplan-Meier logrank tests of OS based on the best binarized PC from correlation of multiple DNA repair gene expression adjusted for age and the PCNA metagene. For the simultaneous adjustment of expression by age and PCNA metagene (first A,P column), all of the cancers resulted in a significant KM logrank test for the best binarized PC; however, the empirical p-values for random selection of genes resulted in a significant KM test (second A,P column) for 3 cancers, namely, AML, bladder, and sarcoma.

Table 2 lists sample sizes and cancer sites for KM tests performed on the best binarized PC for multiple DNA repair gene expression adjusted for age, stage, and the *PCNA* metagene. For the simultaneous adjustment of expression for age, stage, and *PCNA* metagene (first A,S,P column), 11 of the 12 cancer resulted in PCs whose KM test result were significant. However, when random gene sets of the same size were selected, (second A,S,P column), only 8 of the 12 cancers were significant: breast, colorectal, liver, lung, lung squamous cell, melanoma, renal papillary cancer, and stomach. The combined results in Tables 1 and 2 suggest that while age, stage, and *PCNA* metagene adjusted expression of DNA repair gene expression resulted in significant KM tests for 94% of the cancers (17/18 tests), the selection of random gene sets of the same size resulted in 61% (11/18) cancers showing significant prediction of OS by adjusted DNA repair gene expression.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2019 doi:10.20944/preprints201902.0042.v1

Peer-reviewed version available at *Cancers* **2019**, *11*, 501; doi:10.3390/cancers11040501

Table 1. Kaplan-Meier logrank test p-values for best principal component (best binarized PC) derived from correlation matrix of DNA repair genes with significant individual KM tests after adjustment of expression for age and *PCNA* metagene effects. Cancers listed had only age at diagnosis available in TCGA clinical data. Bold highlighting for p-values denotes cancers for which KM logrank or randomization tests were significant when both age at diagnosis and PCNA adjustments were made.

Cancer	n	Kaplan-Meier logrank[a]				Random genes (*B*=1000)[b]			
		A	P	A,P	N	A	P	A,P	N
AML	200	0.0062	0.0018	**0.0157**	0.0062	0.0120	0.0020	**0.0130**	0.0130
Bladder	413	0.0000	0.0000	**0.0000**	0.0001	0.0190	0.0180	**0.0160**	0.0360
Low Grade Gliomas	530	0.0000	0.0000	**0.0000**	0.0000	0.5740	0.2760	0.3880	0.3520
GBM	604	0.0090	0.0005	**0.0231**	0.0008	0.0270	0.0040	0.1600	0.0030
Head & Neck	530	0.0005	0.0009	**0.0006**	0.0008	0.0150	0.1370	0.1120	0.0110
Sarcoma	265	0.0011	0.0032	**0.0008**	0.0001	0.1200	0.0520	**0.0120**	0.0150

A-Expression adjusted for age at diagnosis.

P-Expression adjusted for PCNA metagene expression.

A,P-Expression adjusted for age at diagnosis and PCNA metagene expression.

N-No adjustment to expression.

a - KM logrank test of binarized principal component scores from correlation matrix of DNA repair genes with significant logrank tests.

b - P-values based on number of times KM logrank test $\chi 2$(b) based on randomly selected genes exceeded $\chi 2$ based on non-random genes.

Table 2. Kaplan-Meier logrank test p-values for best principal component (best binarized PC) derived from correlation matrix of DNA repair genes with significant individual KM tests after adjustment of expression for age, stage, and *PCNA* metagene effects. Cancer listed had both age at diagnosis and stage available in TCGA clinical data. Bold highlighting for p-values denotes cancers for which KM logrank or randomization tests were significant when age, stage, and PCNA adjustments were made.

Cancer	n	Kaplan-Meier logrank[a]						Random genes (B=1000)[b]					
		A	S	P	A,S	A,S,P	N	A	S	A,S	P	A,S,P	N
Breast	1105	0.0057	0.0049	0.0004	0.0069	**0.0015**	0.0060	0.0790	0.0450	0.0050	0.0860	**0.0350**	0.0650
Cervical	309	0.0071	0.0014	0.0015	0.0100	**0.0151**	0.0088	0.2150	0.0630	0.0120	0.2530	0.0730	0.2810
Colorectal	633	0.0550	0.0255	0.0048	0.0188	**0.0106**	0.0330	0.2050	0.0700	0.0120	0.0450	**0.0340**	0.1160
Liver	379	0.0000	0.0002	0.0004	0.0001	**0.0027**	0.0001	0.1150	0.2220	0.0020	0.2140	**0.0130**	0.1720
Lung	522	0.0120	0.0008	0.0003	0.0021	**0.0036**	0.0070	0.4850	0.0560	0.0070	0.1030	**0.0400**	0.3570
Lung SC	505	0.0057	0.0050	0.0040	0.0050	**0.0040**	0.0057	0.0400	0.0290	0.0110	0.0290	**0.0100**	0.0330
Ovarian	603	0.0259	0.0088	0.0184	0.0722	**0.0100**	0.0183	0.2310	0.1250	0.2010	0.5240	0.0700	0.2060
Melanoma	479	0.0000	0.0001	0.0002	0.0000	**0.0001**	0.0001	0.0170	0.0710	0.0190	0.0250	**0.0040**	0.0670
Renal CC	538	0.0000	0.0000	0.0000	0.0000	**0.0003**	0.0000	0.2570	0.4230	0.0100	0.1340	0.2780	0.4710
Renal Pap.	292	0.0000	0.0001	0.0011	0.0000	**0.0012**	0.0000	0.0770	0.0100	0.1060	0.0040	**0.0190**	0.0320
Stomach	478	0.6412	0.0103	0.0180	0.0103	**0.0107**	0.6412	0.6570	0.0070	0.0300	0.0060	**0.0160**	0.6780
Uterine	548	0.0013	0.0024	0.0150	0.0010	**0.0072**	0.0009	0.0250	0.0390	0.3620	0.0070	0.1070	0.0190

A-Expression adjusted for age at diagnosis.

S-Expression adjusted for stage.

P-Expression adjusted for *PCNA* metagene expression.

A,S-Expression adjusted for age at diagnosis and stage.

A,S,P-Expression adjusted for age at diagnosis, stage, and *PCNA* metagene expression.

N-No adjustment to expression.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2019 doi:10.20944/preprints201902.0042.v1

Peer-reviewed version available at Cancers 2019, 11, 501; doi:10.3390/cancers11040501

[a] - KM logrank test of binarized principal component scores from correlation matrix of DNA repair genes with significant logrank tests.

[b] - P-values based on number of times KM test $\chi^{2(b)}$ based on random genes exceeded χ^2 based on non-random genes.

Table 4. Qualitative patterns of genomic event rates (GER) per gene-tumor for each cluster identified during hierarchical cluster analysis (from Figure 1). Opportunistic cancers for further study in cluster 2 are AML, Colorectal, and Renal Papillary.

		Driver genes			DNA Repair genes		
Cluster	**Cancer**	**Mutations**	**Deletions**	**Amplifications**	**Mutations**	**Deletions**	**Amplifications**
1	Uterine*, Stomach, Bladder, Head & Neck*, Lung, Breast, Lung Sq. Cell, Liver*, Cervical*, Sarcoma	↑	↑	↑	↑	↓	↑
2	AML, Colorectal, GBM*, Low Grade Gliomas*, Renal Papillary, Renal Clear Cell*	↑	↓	↓	↓	↓	↓
3	Melanoma	↑	↓	↑	↑	↓	↑
4	Ovarian*	↑	↓	↑	↓	↑	↑

*Not significant during empirical p-value testing, i.e., DNA repair gene expression adjusted for age, stage, and *PCNA* metagene does not predict OS significantly.

Table 3. DNA Repair genes whose upregulation or downregulation prolongs overall survival (OS) for subjects with RNA-Seq data in TCGA. Cancers listed had significant empirical p-value test results (P<0.05).

Cancer	**Upregulation prolongs OS**	**Downregulation prolongs OS**
AML*	*POLN*	*RAD23A, APEX2, EME2*
Bladder*	*BLM, RAD9A, MGMT, LIG1, MUTYH, DDB1, ERCC5, XPC, FANCD2, MSH5, DCLRE1C, REV1*	*FANCC, ALKBH2, APEX2, LIG3, POLB, GTF2H5, PMS1, PRKDC, REV3L*
Sarcoma*	*MNAT1, APEX1, APTX, FEN1, NEIL3, DDB1, GTF2H3, FANCI, PRKDC*	*DCLRE1B, POLL, CUL4A, ERCC2, MSH2, FANCG*
Breast	*RAD50, PMS1*	*ATRIP, FANCC, RAD1, RFC3, NEIL3, EXO1, FANCB, FANCD2, FANCI, RAD51, XRCC4*
Colorectal	*DCLRE1C*	*RAD23A, RFC2, POLL, MLH3, FANCL*
Renal Papillary	*RAD17, OGG1, DDB2, ERCC2*	*BLM, RAD1, FEN1, LIG1, EXO1, MSH6, BRCA2, EME1, FANCB, LIG4*
Lung	*RAD17, ALKBH3, MGMT, MPG, NEIL1, XPC, LIG4, POLK, REV3L*	*BRCA1, NBN, RAD1, NEIL3, MMS19, FANCI, XRCC5*
Lung Sq. Cell	*CHEK2, MNAT1, APTX, TDG, FANCE, FANCL*	*XRCC1*

Melanoma	*ATM, MNAT1, MBD4, NEIL1, ERCC5, RAD23B, DCLRE1C*	*MDC1, NBN, MUTYH, POLE, UNG, FANCE, FANCI, POLI, POLK*
Stomach	*CUL4A, POLQ*	

*Stage not available, expression adjusted only by age and *PCNA* metagene.

Table 3 lists the DNA repair genes of the best binarized PC for the 11 cancers showing empirical p-values less than 0.05 when adjusting for age, stage, and PCNA metagene. When considering the composite of all the genes which were significant survival predictors, pathway activation results indicate upregulation of the *BRCA1* pathway, NHEJ pathway, BER pathway, MMR pathway, and downregulation of the NER pathway. *ATM* signaling (ATM, RAD17, RAD50) was downregulated due to strong downregulation of *ATM*, which has been observed to be due to promoter hypermethylation [38]. With downregulation of the tumor suppressor kinase *ATM* and the checkpoint kinase *CHEK2*, it follows that the *ATM* intraphase checkpoint would also likely be inactivated. Cells deficient in *ATM* have also been observed to not reduce transcription following DSBs, a phenotype which has been called "radiosensitized DNA synthesis." In addition, ataxia-telangiaectasia (A-T) cells deficient in *ATM* are known to repair DSBs following exposure to IR [39]. With regard to the *BRCA1* pathway, Complex B was highly upregulated with downstream upregulation of G1/S-Phase as well as the HRR pathway (*BLM*, *BRCA1*, *MSH2*, *MS62*, and *RFC*). In addition, *FANCD2* was upregulated, which activates S-Phase checkpoint control. However, Complex C was downregulated (mostly due to *RAD50*), possibly suggesting downregulation of downstream G2/M Phase. In the NHEJ pathway, *ATM*, the cross-linking enzyme Artemis (*DLCRE1C*), and *RAD50* were downregulated, however, *LIG3*, *LIG4*, *NBN*, *PRKDC*, *XRCC1*, *XRCC2*, and *XRCC5* were upregulated. Activated genes in the BER pathway included *POLB*, *POLE*, *XRCC1*, *LIG1*, *LIG3*, and *FEN1*. The MMR pathway activated genes were *EXO1*, *FEN1*, *MSH2*, *MSH6*, *RFC2*, and *RFC3*. Downregulated genes in the NER pathway were the sensitizer *APEX1* and the DNA glycosylate *OGG1*. The NER pathway was mostly downregulated due to downregulation of *HR23B*, *TFIH*, *XPC*, and *XPG*. We also noted that *MGMT* was downregulated, which would increase SSBs and their conversion to DSBs at replication forks [40]. Another observation was that *PARP1* was not in any of the lists of significant genes, which may indicate co-regulation with *PCNA*, whose effect on expression of other DNA repair genes was removed.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2019 doi:10.20944/preprints201902.0042.v1

Peer-reviewed version available at Cancers 2019, 11, 501; doi:10.3390/cancers11040501

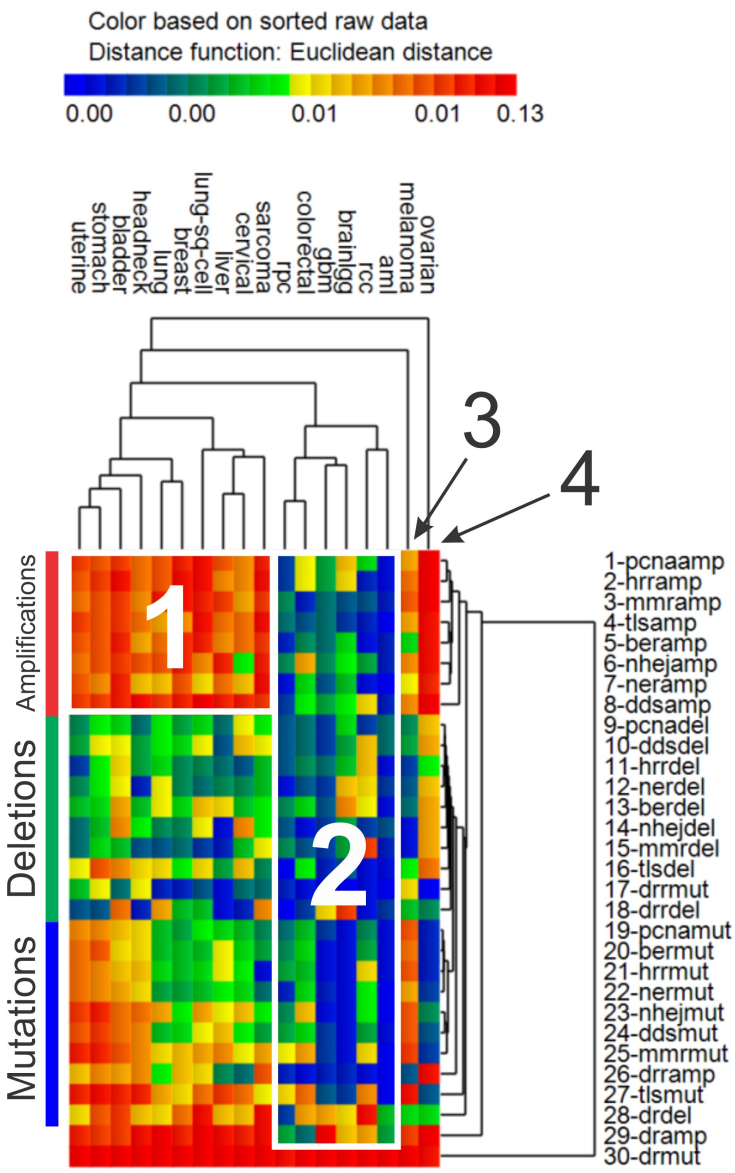

Figure 1. Genomic event rates. Hierarchical cluster analysis results of genomic event rates (GER, per tumor-gene) for somatic mutations ("mut"), deletions ("del"), and amplifications ("amp") in driver genes ("dr") and the 8 groups of DNArepair genes (DRR, BER, NHEJ, MMR, TLS, DDS, HRR, NER). Euclidean distance used as the distance function, with unweighted pair group method with arithmetic mean (UPGMA) as the agglomeration method.

Results of cluster analysis of the GER for various cancers is shown in Figure 1. A total of 4 clusters of cancers were discernible in the data. In spite of all the cancers exhibiting high GERs for driver mutations, cancers in cluster 1 portrayed strong upregulation of genomic amplification in DNA repair genes, while cancers in cluster 2 reveal downregulation of amplification, deletion, and mutation in DNA genes. Melanoma and ovarian cancer clustered furthest away from the previously described clusters, mostly because of the unique patterns among GERs which emerged. Regarding driver genes, both melanoma and ovarian cancer exhibited greater rates of amplifications but had lower rates of deletions. Additionally, while melanoma revealed increased rates of mutations and amplifications and decreased rates of deletions in DNA repairs genes, ovarian cancer showed the

opposite pattern, with lower rates of mutations in DNA repair genes and greater rates of amplifications and deletions. Table 4 lists qualitative patterns which emerged from the cluster analysis of GERs shown in Figure 1.

Figures S1-S36 illustrate for each cancer investigated a KM plot and density (pdf) plot of p-values during random selection of gene sets for the best binarized PCs derived from sets of genes, for which each gene had its own significant KM test after the various adjustments for age, stage, and PCNA metagene. P-values listed in Tables 1 and 2 were extracted from Figures S1-S36.

3. Discussion

Cancer is a multifactorial disease which depends on a constellation of factors involving genomic instability, selective genetic pressure from somatic mutations and polymorphisms, and gene-environment interactions. Two important hallmarks of cancer are the persistent high level of somatic mutations in driver genes and DNA repair addiction. Together, these mechanisms directly and indirectly support a growth advantage and prolonged cell survival. As the costs of genomic DNA sequencing and RNA-Seq analysis decrease, there will continue to be new information available regarding cancer's addiction for DNA repair.

Our approach employed two levels of statistical testing, one that merely involved straightforward ML estimation and another based on randomization tests, which resulted in empirical p-values. The ML-based survival prediction with adjusted DNA repair gene expression was significant for most of the cancers; however, survival prediction based on empirical p-values was significant for fewer cancers. It is now known that identification of sets of genes from genome-wide annotation lists will result in false positives that are associated with the *PCNA* metagene. Our focus was to specifically target DNA repair gene expression, remove the effect of the *PCNA* metagene, age at diagnosis, and stage, to determine if significant lists can be obtained. Not surprisingly, after the adjustments, many of the cancers revealed DNA repair genes which significantly predicted OS.

It is important to realize that our use of randomly selected same-size gene sets was performed in order to develop randomization tests for empirical p-value testing. Since the study was hypothesis-driven and used a candidate gene approach with a constrained list of DNA repair genes, it would be impossible to determine the effects of random bias based on randomly selecting genes from a genome-wide perspective as was done by Venet et al. and Shimon [16,19]. There may likely be genes in the genome which predict survival better than the DNA repair genes considered; however, they would not be DNA repair genes. There was also no freedom to use genome-wide selection of genes for survival prediction, due to our candidate gene approach.

Pathway analysis results indicate a pattern suggestive of downregulation of primary damage signaling kinase (*ATM*) and initial BER pathway components (*APEX1* and *OGG1*), and when combined with increased pathogenic somatic mutations in driver genes (e.g., TP53), our results may indicate that damage signaling in the initial portion of repair pathways is abrogated, while the remainder of the pathway is intact.

We also assessed GERs of the various cancers and confirmed that all of the cancers had high somatic mutation rates in driver genes. There were also two main clusters of cancers identified, which portrayed either high levels of amplification in DNA repair genes or low GERs for mutations, deletions, and amplification in DNA repair genes. The latter group of cancers including AML, colorectal, GBM, low grade gliomas, and renal papillary (cluster 2 in Table 4), may be more opportunistic for repair inhibition therapy because of less confounding associated with low levels of mutations, deletions, and amplification exhibited in DNA repair genes for these cancers. It warrants noting that the heat map in Figure 1 represents cancer-specific GERs in DNA repair pathways and not genome-wide mutations levels in cancers as reported by Lawrence et al. [41]. While melanoma was reported by Lawrence et al. to have the greatest levels of genome-wide mutation levels, our results indicate that melanoma had high levels of gene amplification and somatic mutations in DNA

Peer-reviewed version available at Cancers 2019, 11, 501; doi:10.3390/cancers11040501

repair genes, causing its separate clustering. In addition, ovarian cancer clustered by itself because of high levels of amplification in DNA repair genes.

Random somatic mutations considered in TCGA data are not genetic polymorphisms (e.g. SNPs) occurring in the same regions of DNA, for which allelic phenotypes, associated risks, prognosis, and recommended treatment options are known. Instead, they are rarely found in the same location in DNA and are rarely of the same type. Somatic mutations merely accumulate from genetic selective pressure in driver genes, which is one of the most important hallmarks of sporadic cancers. Along these lines, there is uncertainty related to experimentally verifying the effect and pathogenicity of a single somatic mutation. Numerous algorithms can be employed for computationally predicting pathogenicity of somatic mutations (our results are based on the FATHMM algorithm used by TCGA and COSMIC [42]), but the experimental laboratory costs required to fully understand how a single somatic mutation alters a protein and how any change in function impacts the disease phenotype are exorbitant.

The translational value of our results is established by the potential of novel patterns of DNA repair gene expression in cancer, which could prove useful in animal studies, transgenics, and xenograft models, etc., in order to understand if inhibition of the genes identified inhibit tumor growth and improve survival [43,44,45]. Adjustment of DNA repair gene expression by the PCNA metagene has enabled us to view cancer from a distant perspective based on high-granularity involvement of DNA repair pathways in cancer. This view will hopefully enforce an appreciation among biologists and oncologists for the translational value of pursuing experimental inhibition studies, as well as randomized control trials for establishing safety and evaluating efficacy.

We did not comparatively assess numerous techniques for their computational efficiency, scalability, or differences in OS survival prediction. We also did not evaluate differences between using progression free survival (PFS) vs. OS, or bootstrapping effects on results. TCGA was primarily undertaken for molecular studies, and therefore clinical data standardization and collection was a secondary effort [35]. TCGA was also not a clinical trial, and therefore outcomes were considered from both retrospective and prospective cases, without a standardized patient follow-up plan. Molecular data were also obtained from single sections of primary tumors, and therefore spatial and temporal variation in tumor heterogeneity cannot be addressed. Use of OS as a survival endpoint is supported for most of the cancers available in TCGA [35]. For breast cancer subtypes with varying aggressiveness, OS is likely appropriate for the basal-like subtype but not for luminal A. GBM is also considered an aggressive cancer, although OS is considered suitable for use with TCGA data. Prostate cancer OS in TCGA is not a suitable endpoint, and therefore prostate cancer was not evaluated. Regarding confounding, we removed the effects of age and stage at diagnosis from expression data using skew-zero inputs to multivariate regression. Confounding caused by competing risks and outcomes in TCGA is more relevant for disease-specific survival (DSS), disease-free interval (DFI), and PFS and less of an issue for OS, so we do not envision a sufficient level of competing risks bias which would trigger a concern. The Cox PH assumption is also supported for most of the TCGA cancers, with only a few exceptions. While KM analysis was employed as the primary survival hypothesis test, Cox PH regression was only used for determination of risk directionality as a function of increasing expression level.

There is also the problem of unknown upstream effects of germline polymorphisms and DNA repair deficiencies which may result in a variety of unknown influences. The difficulty presented by cellular niching and high levels of clonal heterogeneity in tumors also presents a challenge for fully unraveling the associations observed in this study. The TCGA data used are not based on single-cell RNA-Seq analysis, which would be helpful for elucidating heterogeneity effects; however, the large variation in genotypes would exacerbate the present uncertainties surrounding our attempt to portray the role of DNA repair genes in cancer survival. We also did not employ empirical data from TCGA for DNA microsatellite instability, methylation status, or chromosome aberrations, which would overlay more complexity on the models developed. Although we did include DNA repair genes in the MMR pathway, which verify repeat count of microsatellites during cell division [46].

4. Materials and Methods

PCNA Metagene. For each tumor, we obtained RNA-Seq derived normalized expression values of the 131 PCNA-related genes [16], and collapsed their expression values down to median expression, which is termed the "PCNA metagene." Next, the *PCNA* metagene (median) and expression values for all of the DNA repair genes listed above were transformed into van der Waerden (VDW) scores. This transformation simply involved first transforming expression values for each gene into percentiles, and then substituting the percentiles as probabilities in the inverse standard normal function, i.e., $Z=\Phi^{-1}(ptile)$, to obtain standard normal variates of expression, which are distributed with mean zero and variance unity.

Maximum Likelihood (ML) Survival Prediction. VDW scores for each DNA repair genes were regressed separately on the VDW scores for age, VDW scores for tumor stage, and VDW scores for *PCNA* metagene, and the residuals were taken as the new DNA repair gene expression values. Residuals for each DNA repair gene were then binarized into (0,1) by splitting on the median, to form a grouping variable which was employed in Kaplan-Meier (KM) survival analysis with overall survival (OS) status. OS was reported to be the most accurately derived outcome for TCGA data [35]. Each DNA repair gene which resulted in significant maximum likelihood (ML) estimates of the KM logrank tests, i.e., χ^2 p-value <0.05 (1 d.f. chi-squared), was appended to a list of p genes. Eigendecomposition of the correlation matrix of the p significant genes was then performed with Varimax orthogonal rotation, and the principal components (PCs) for all p dimensions were extracted. Each PC was then transformed into a binary grouping variable for KM input, by assigning negative PC values to 0 and positive to 1. The single group-transformed PC which resulted in the greatest χ^2 value during KM analysis was identified and called the "best binarized PC." For cancers without stage available in TCGA, the best binarized PC when age and *PCNA* were used for residual generation was input into Cox proportional hazards (PH) regression as a continuous variable (i.e., PC score) to determine whether positive values prolonged or shortened OS. Whereas, for cancers with stage available in TCGA, the best binarized PC when age, stage, and *PCNA* were used for residual generation was input into Cox proportional hazards (PH) regression. For the best binarized PC under evaluation, if the Cox PH hazard ratio (*HR*)<1, then genes having positive loading on the best binarized PC were beneficial to OS if upregulated, whereas genes whose loadings were negative were considered hazardous, and would need to be downregulated in order to be beneficial. Analogously, if the Cox PH *HR*>1, it meant that positive PC values were deleterious, and therefore genes that loaded negatively on this PC would need to be upregulated to be beneficial, and genes that loaded positively on this PC would need to be downregulated in order to be beneficial to OS. Figure 1 illustrates the workflow employed, outlining the various steps used for establishing the best binarized PC for each cancer, and whether positive loadings on the best binarized PC prolonged or shortened OS. Justification for using KM analyses was hinged to our observation that for the same genes, results from Cox PH regression of continuously-scaled expression were consistently more significant when compared with grouped analysis based on KM. Thus, we chose KM analysis for prediction significance due to conservativeness, and Cox PH to determine directionality of survival risk as a function of increasing expression values.

Empirical P-value Tests of Survival Using Randomly Selected Genes. For each cancer, the single best binarized PC that resulted in the greatest chi-squared statistic during maximum likelihood KM analysis was considered to the be the "observed" test statistic. Recall, this test statistic for the best binarized PC was initially based on individual DNA repair genes whose adjusted gene expression resulted in a significant KM test. Let the number of significant DNA repair genes for a best binarized PC be p. We used B=1,000 iterations for empirical p-value testing. During each bth iteration, a random set of p DNA repair genes with the same adjustment to expression were selected, followed by correlation analysis, and then PC extraction via eigendecomposition of the $p \times p$ correlation matrix. Each PC was then binarized and used in KM analysis to determine which PC resulted in the greatest chi-squared test statistic for the set of p random genes. After B iterations, the

empirical p-value was equal to $P = \#\{b: \chi^{2(b)} > \chi^2\} / B$, where χ^2 is the "observed" 1 d.f. chi-squared test statistic from ML-based KM analysis based on the best binarized PC, and $\chi^{2(b)}$ is the chi-squared statistic from the best binarized PC extracted from the correlation matrix of p randomly selected DNA repair genes used in KM analysis during the bth iteration. The bottom of Figure 2 illustrates how the correlation matrix of p genes with significant KM analysis were employed to obtain the best binarized PC for predicting OS.

Genomic Event Rates (GERs). For each cancer, we also summed the number of pathogenic somatic mutations, deletions, and amplifications in the set of 20 driver genes, and in each of the 8 groups of DNA repair genes (DRR, BER, NHEJ, MMR, TLS, DDS, HRR, NER). Table S2 of the Supplemental Information lists the cancer-specific driver genes used, which we previously reported [37]. Driver gene selection was based on the top 20 driver genes identified by at least 2 tools in the DRIVERDB database [36]. The genomic event rate (GER) of each type of event was determined by dividing the sum by the number of genes in the group and the number of tumors obtained for each of the cancers considered. This led to the GER in units of events per gene-tumor. Hierarchical cluster analysis was then used to cluster values of GER for each cancer. Euclidean distance was used as the distance function, while the unweighted pair group method with arithmetic mean (UPGMA) was used for the agglomeration method.

Removing Redundant Genes in Gene Lists. *PCNA* itself was listed as a BER gene and was removed from survival analysis because our primary goal was to remove the genome-wide association of *PCNA* with other genes from the expression of DNA repair genes. For most cancers, *TP53* was listed as a driver gene with high levels of somatic mutations, so it was not included as a DNA repair gene during survival analysis for those cancers. DNA repair genes listed multiple times in the repair pathways described above included *POLD1, POLE, PLOH, POLM, RECOL2, PCNA, LIG1, BLM, BRCA1, FANCA, FANCC,* and *XPF*. Only the first occurrence of these genes in their respective lists was used. Altogether, a final list of 126 unique (non-redundant) DNA repair genes was constructed and used for all cancers. Table S3 of the Supplemental Information lists the DNA repair genes which were excluded from survival analysis because of duplicate listing in the PCNA gene list or list of driver genes for the given cancer, and also lists driver genes excluded from GER analysis because of duplicate listing in the PCNA gene list.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2019 doi:10.20944/preprints201902.0042.v1

Peer-reviewed version available at Cancers 2019, 11, 501; doi:10.3390/cancers11040501

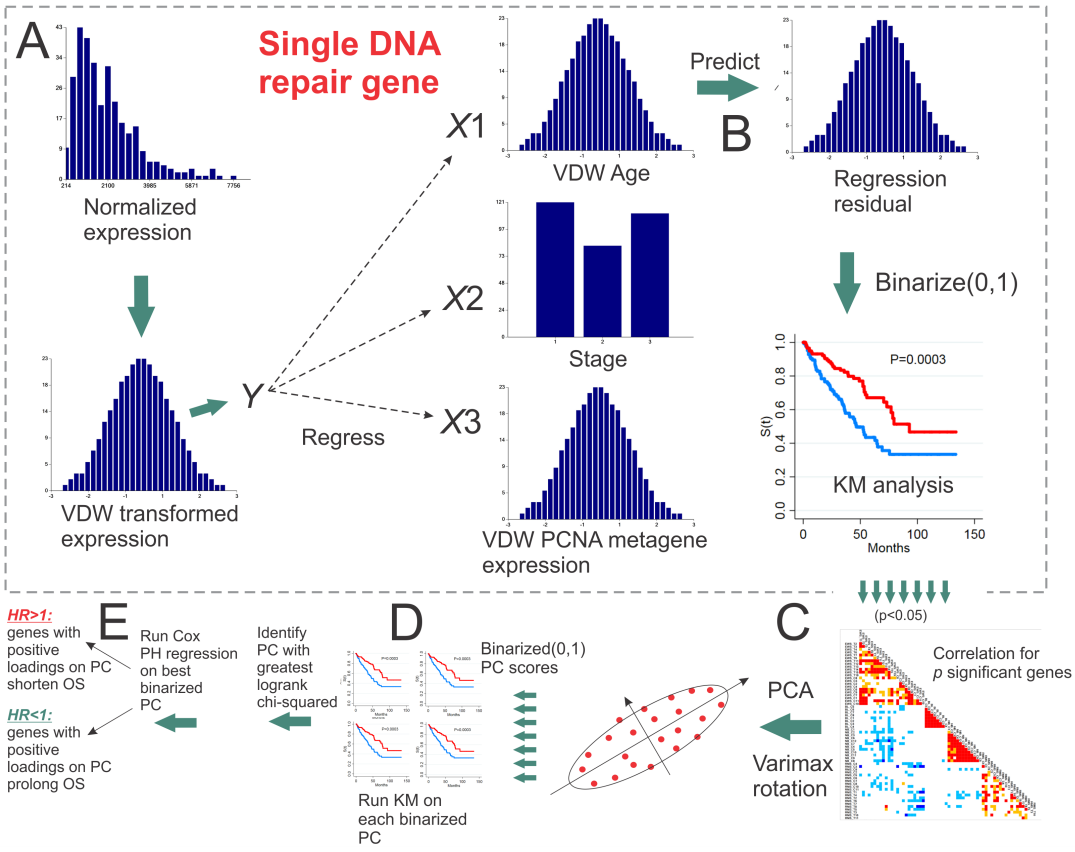

Figure 2. Workflow for identifying the "best binarized PC" for a set of significant genes from univariate Kaplan-Meier analyses. A. van der Waerden (VDW) scores of log-transformed expression values for each DNA repair gene are regressed on the VDW scores for age at diagnosis, stage, and the PCNA metagene. B. The residual values from each linear fit (i.e., expression with the effect of age at diagnosis, stage, and PCNA metagene removed) are then binarized (0,1) and input as the univariate grouping value during Kaplan-Meier (KM) analysis of overall survival (OS). C. PCA with Varimax orthogonal rotation is performed on the correlation matrix of *p* significant binarized residual vectors. D. Each of the *p* principal component (PC) score vectors is binarized (0,1) and input into univariate KM analysis. E. The PC resulting in the greatest KM chi-squared value is selected as the "best binarized PC," and univariate Cox PH regression is then run on the best binarized PC to determine if positive (negative) PC score values are associated with prolonged (shortened) OS.

5. Conclusions

In conclusion, our hypothesis-driven focus to target DNA repair gene expression adjusted for the PCNA metagene as a means of predicting OS in various cancers resulted in statistically significant sets of DNA repair genes. We also identified that AML, colorectal, and renal papillary cancers may be potentially more opportunistic for inhibition therapy because of less confounding in the form of lower rates of mutations, deletions, and amplifications in DNA repair genes which predict OS in these cancers. The most opportunistic cancer for DNA repair inhibition therapy appears to be AML, since the TCGA cases harbored the lowest rates of somatic mutations, deletions, and amplifications in DNA repair genes.

Supplementary Materials: S1 of the Supplemental Information lists the DNA repair genes used which were available in TCGA expression data. The various DNA repair mechanisms and pathways for repair genes used are described below. Table S2 of the Supplemental Information lists the cancer-specific driver genes used, which we previously reported [37]. Driver gene selection was based on the top 20 driver genes identified by at

least 2 tools in the DRIVERDB database [36]. Table S3 of the Supplemental Information lists the DNA repair genes which were excluded from survival analysis because of duplicate listing in the PCNA gene list or list of driver genes for the given cancer, and also lists driver genes excluded from GER analysis because of duplicate listing in the PCNA gene list. Figures S1-S36 show Kaplan-Meier and kernel density plots for acute myelogenous leukemia, bladder, breast, cervical, colorectal, glioblastoma multiforme, head and neck, low grade gliomas, liver, lung sq. cell, lung, melanoma, ovarian, renal clear cell, renal papillary, sarcoma, stomach, and cancers of the uterine.

Author Contributions: LEP wrote the manuscript, developed the workflow, and performed the majority of computational modeling, while TK reviewed and confirmed the statistical procedures.

Funding: Research was supported by NASA Grant NNX-12AO52A.

Conflicts of Interest: The authors declare no conflict of interest. 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.

References

1. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000, 100(1):57-70.
2. Curtin NJ. DNA repair dysregulation from cancer driver to therapeutic target. Nat Rev Cancer 2012, 12(12):801-817.
3. Cannan WJ, Pederson DS. Mechanisms and Consequences of Double-Strand DNA Break Formation in Chromatin. J Cell Physiol 2016, 231(1):3-14.
4. Forment JV, Kaidi A, Jackson SP. Chromothripsis and cancer: causes and consequences of chromosome shattering. Nat Rev Cancer 2012, 12(10):663-670.
5. Bryant HE, Schultz N, Thomas HD, Parker KM, Flower D, Lopez E, Kyle S, Meuth M, Curtin NJ, Helleday T. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. Nature 2005, 434(7035):913-917.
6. Farmer H, McCabe N, Lord CJ, Tutt AN, Johnson DA, Richardson TB, Santarosa M, Dillon KJ, Hickson I, Knights C et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature 2005, 434(7035):917-921.
7. Shaheen M, Allen C, Nickoloff JA, Hromas R. Synthetic lethality: exploiting the addiction of cancer to DNA repair. Blood 2011, 117(23):6074-6082.
8. Nickoloff JA, Jones D, Lee SH, Williamson EA, Hromas R. Drugging the Cancers Addicted to DNA Repair. J Natl Cancer Inst 2017, 109(11).
9. Budzowska M, Kanaar R. Mechanisms of dealing with DNA damage-induced replication problems. Cell Biochem Biophys 2009, 53(1):17-31.
10. Allen C, Ashley AK, Hromas R, Nickoloff JA. More forks on the road to replication stress recovery. J Mol Cell Biol 2011, 3(1):4-12.
11. Ashworth A, Lord CJ. Synthetic lethal therapies for cancer: what's next after PARP inhibitors? Nat Rev Clin Oncol 2018, 15(9):564-576.
12. Dedes KJ, Wilkerson PM, Wetterskog D, Weigelt B, Ashworth A, Reis-Filho JS. Synthetic lethality of PARP inhibition in cancers lacking BRCA1 and BRCA2 mutations. Cell Cycle 2011, 10(8):1192-1199.
13. Rehman FL, Lord CJ, Ashworth A. Synthetic lethal approaches to breast cancer therapy. Nat Rev Clin Oncol 2010, 7(12):718-724.
14. Gavande NS, VanderVere-Carozza PS, Hinshaw HD, Jalal SI, Sears CR, Pawelczak KS, Turchi JJ. DNA repair targeted therapy: The past or future of cancer treatment? Pharmacol Ther 2016, 160:65-83.
15. Essers J, Theil AF, Baldeyron C, van Cappellen WA, Houtsmuller AB, Kanaar R, Vermeulen W. Nuclear dynamics of PCNA in DNA replication and repair. Mol Cell Biol 2005, 25(21):9350-9359.
16. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with Breast cancer outcome. Plos Comput. Biology. 2011;7(10):e1002240.
17. Moldovan GL, Pfander B, Jentsch S. PCNA, the maestro of the replication fork. Cell. 2007;129(4):665–679.
18. Ge X, Yamamoto S, Tsutsumi S, Midorikawa Y, Ihara S, Wang SM, Aburatani H. Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. Genomics. 2005;86(2):127–141.

19. Shimoni Y. Association between expression of random gene sets and survival is evident in multiple cancer types and may be explained by sub-classification. PLoS Comput Biol 2018, 14(2):e1006026.
20. Davidson D, Amrein L, Panasci L, Aloyz R: Small Molecules, Inhibitors of DNA-PK, Targeting DNA Repair, and Beyond. Front Pharmacol 2013, 4:5.
21. O'Cearbhaill RE: Using PARP Inhibitors in Advanced Ovarian Cancer. Oncology (Williston Park) 2018, 32(7):339-343.
22. Waks Z, Weissbrod O, Carmeli B, Norel R, Utro F, Goldschmidt Y. Driver gene classification reveals a substantial overrepresentation of tumor suppressors among very large chromatin-regulating proteins. Scientific Reports. 2016;6:38988.
23. NCI and the NHGRI. The Cancer Genome Atlas, 2005.
24. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne, CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery. 2012;2:401.
25. Gao et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 2013;6:11.
26. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC 2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology. 2011;12(4):R41.
27. Wood RD, Mitchell M, Sgouros JG, Lindahl T. Human DNA Repair Genes. Science 2001:291:1284.
28. Wood RD, Mitchell M, Lindahl T. Human DNA Repair Genes, 2005. Mutation Res. 2005;577:275.
29. Friedberg EC, Walker GC, Siede W, Wood RD, Schulz RA, Ellenberger T. DNA Repair and Mutagenesis, 2nd edition. ASM Press, Washington, D.C., 2006.
30. Lange SS, Takata K, Wood RD. DNA Polymerases and Cancer. Nature Reviews Cancer. 2011; 11:96.
31. Ronen A, Glickman BW. Human DNA repair genes. Environ. Mol. Mutagen. 2001;37:241.
32. Eisen JA, Hanawalt PC. A phylogenomic study of DNA repair genes, proteins, and processes. Mutat. Res. DNA Repair. 1999;435:171
33. Aravind L, Walker DR, Koonin EV. Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res. 1999;27:1223.
34. Strand M, Prolla TA, Liskay RM, Petes TD. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 1993, 365(6443):274-276.
35. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell 2018, 173(2):400-416 e411.
36. Cheng WC, Chung IF, Chen CY, Sun HJ, Fen JJ, Tang WC, et al. DriverDB: an exome sequencing database for cancer driver gene identification. Nucleic Acids Res. 2014;42(Database issue):D1048-54.
37. Peterson LE, Kovyrshina T. Progression inference for somatic mutations in cancer. Heliyon. 2017;3(4):00277.
38. Weber AM, Ryan AJ. ATM and ATR as therapeutic targets in cancer. Pharmacology and Therapeutics. 2015;149:124–138.
39. Choi S, Gamper AM, White JS, Bakkenist CJ. Inhibition of ATM kinase activity does not phenocopy ATM protein disruption: Implications for the clinical utility of ATM kinase inhibitors. Cell Cycle. 2010;9:4052–4057.
40. Erasimus H, Gobin M, Niclou S, Van Dyck E. DNA repair mechanisms and their clinical impact in glioblastoma. Mutation Res. 2016;769: 19–35.
41. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013, 499(7457):214-218.
42. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E et al. COSMIC: the Catalogue of Somatic Mutations In Cancer. Nucleic Acids Res 2018.
43. Li JF, Konstantinopoulos PA, Matulonis UA. PARP inhibitors in ovarian cancer: current status and future promise. Gynecol. Oncol. 2014;133(2):362–369.
44. Dizdar O, Arslan C, Altundag K. Advances in PARP inhibitors for the treatment of breast cancer. Expert. Opinion Pharmacother. 2015;16(18):2751–2758.

45. Sonnenblick A, de Azambuja E, Azim HA, Piccart M. An update on PARP inhibitors – moving to the adjuvant setting. Nat. Rev. Clin. Oncol. 2015;12(1):27–41.
46. Bonneville R, Krook MA, Kautto EA, Miya J, Wing MR, Chen HZ, Reeser JW, Yu L, Roychowdhury S: Landscape of Microsatellite Instability Across 39 Cancer Types. JCO Precis Oncol 2017, 2017.