Regulations of DNA Damage Response via Gene Copy Number Variations

by

Sihan Li

Bachelor of Science, Fudan University, 2017

Submitted to the Graduate Faculty of

School of Pharmacy

of the requirements for the degree of

Master of Science

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF PHARMACY

This thesis/dissertation was presented

by

Sihan Li

It was defended on

March 27, 2019

and approved by

Da Yang, Assistant Professor, School of Pharmacy

Xiaochao Ma, Associate Professor, School of Pharmacy

Christian A Fernandez, Assistant Professor, School of Pharmacy

Regulations of DNA Damage Response via Gene Copy Number Variations

Sihan, BS

University of Pittsburgh, 2019

Both endogenous and exogenous factors may damage DNA. DNA damage response genes work together in different pathways are responsible for the repair of the damaged genome. DNA damage deficiency is one of the most important reasons for cancer initiation and progression. To date, studies have demonstrated that gain of function alteration of DNA damage genes and pathways could cause therapeutic resistance to the genome-instability drugs. In the first section, based on TCGA and GDSC database, the landscape of the copy number alterations of the DDR genes was first depicted and we demonstrated that their overexpression is mainly driven by the copy number amplification. Then we delineated reduced mutation burdens/signatures and poorer survival correlate with the amplification of DDR genes. At last, we showed the landscape of the correlation between cell line genome-instability drug response and DDR gene copy number. This study gave us confidence that DDR gene amplification could serve as a biomarker for the clinical outcome and possible mechanism of chemotherapy resistance. In the second section, focusing on one DNA damage response gene, TP53, and a specific cancer type, colorectal cancer, we performed a genome-wide correlation analysis to identify p53 protein regulating gene. Multi-omics data including TCGA gene expression, copy number variations, and RPPA protein level were integrated into this analysis. After identifying several candidate genes in 8q24.21, *in vitro* study and clinical relevance analysis further demonstrated a p53 protein inhibiting lncRNA, named PiHL.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

1.1 DNA damage repair and cancer

Both endogenous and exogenous factors may damage DNA. Exogenous factor polycyclic aromatic hydrocarbons in smoke could create adducts to bases and crosslinking of DNA, and UV may cause pyrimidine dimers. Endogenous reactive oxygen species (ROS), a natural byproduct of the normal metabolism of oxygen and have essential roles in cell signaling and homeostasis [1], may cause DNA base lesions. DNA is also easy to be damaged when one has a long-time smoking. All the aforementioned molecules or processes can cause DNA structure alterations. Even natural DNA replication may induce errors.

Thanks to DNA damage repair/response (DDR) pathways, most errors and base/structure lesions can be repaired. 9 main DDR pathways have been identified, which are involved in different scenarios of damage and cause different repaired characteristics. For example, homology directed repair (HDR) pathway is error-free and mainly functions on double strand break repair, while some other pathways like microhomology-mediated end joining (MMEJ) is error-prone. Genome stability and normal functions of cells and body can be maintained by these pathways. Cells with DDR pathway malfunction are possible to undergo a harmful change in its DNA and may not be in a normal status. The alterations in DNA may be passed to the daughter cells, and new damage may gain again and again. Cells with unstable genome may gain some cancer-like features such as unregulated proliferation [2]. These cells are probably to become cancerous cells and cancer is initialized in the human body.

1.2 Mutation signatures

Somatic mutations found in cancer genomes may be the consequence of the slight intrinsic infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA, or defective DNA repair [3]. In some cancer types, a substantial proportion of somatic mutations are known to be generated by exposures, for example, tobacco smoking in lung cancers and ultraviolet light in skin cancers [4], or by abnormalities of DNA maintenance, for example, defective DNA mismatch repair in some colorectal cancers [5]. However, our understanding of the mutational processes that cause somatic mutations in most cancer classes is limited.

Different combinations of different mutation types with different weights could be generated by different mutational factors, termed 'signatures' [6]. Mutational signatures in human cancer have been explored through a small number of frequently mutated cancer genes, notably TP53 [7]. Thousands of somatic mutations can now be identified in a single cancer sample through recent advances in sequencing technology, offering the possibility of deciphering mutational signatures even when several mutational processes are operative. Moreover, because most mutations in cancer genomes are 'passengers' [3], they do not bear strong imprints of selection.

1.3 Copy number variation and cancer

Copy number variation (CNV) is a phenomenon in which sections of the genome are repeated and the number of repeats in the genome varies between individuals in the human population [8]. Humans are diploid organisms, thus, in most cases, genes should have 2 copies,

except for some genes in sex chromosomes. However, the genome may contain some duplication or deletion in some regions, causing some genes may have more than two copies. Based on the size of the copy number variation region, this type of CNV is structural variation (long repeats) in the chromosome. Another type of CNV is short repeats, which is nucleotide repeating (e.g., C-G-C-G-C-G…) in some focal regions [9].

Gene copy number variation is one of the genomic profiles, which has been reported to be important in cancer. For example, the copy number amplification of a lncRNA FAL1 was found to be associated with clinical outcomes of patients with ovarian cancer [10]. Focusing on copy number variation of genes in cancer is a possible and significant way to identify biomarkers or treatment targets.

1.4 LncRNA and its mechanisms of regulation

Non-coding RNAs (ncRNAs) are the RNAs that lack protein-coding potential, including microRNA, shRNA, etc. Long non-coding RNA (lncRNA) is defined as the ncRNA whose length is > 200 bp. According to their location in the genome, lncRNAs can be mainly classified as natural antisense transcripts, long intergenic ncRNAs, promoter-associated transcripts, and pseudogenes [11].

To date, studies have demonstrated that lncRNA transcripts are approximately 10-fold less abundant than mRNAs in human cells [12]. This may be explained by the dramatic expression variance of lncRNAs in one cell or amongst different cells, which could be higher than that of protein-coding genes. Besides, the majority of lncRNAs are tissue-specific, but only around 20%

of the protein-coding genes are tissue-specific [13]. This indicates tissue-specific characteristics is also significant for lncRNA and results in the overall low abundance.

However, emerging evidence still shows that lncRNAs are biologically functional, especially in cancer. Some lncRNAs are transcribed from the antisense strand of the well-defined transcriptional units and, they may be involved in *cis* or *trans* regulation of genes located in their vicinity or at distant loci [14]. For example, ANRASSF1 (antisense intronic non-coding RASSF1) is a capped and polyadenylated unspliced long non-coding RNA, with nuclear localization, which is transcribed in the antisense direction relative to the protein-coding mRNAs of the *RASSF1* gene locus. Due to the complementary feature, its interaction with genomic DNA, forming an RNA/DNA hybrid, leads to downregulation of the sense gene at the pre-transcriptional level [15]. Due to its relatively long length, some lncRNAs may form a secondary structure and could be involved in its biological regulation. For instance, a highly conserved uracil-rich region in lncRNA metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) contributes to RNA stability through the formation of a triple helix [16]. The tumor suppressor function of the lncRNA MEG3 can be attributed to two secondary fold motifs [17]. LncRNAs are also reported to be functional through RNA-DNA, RNA-RNA, and RNA-protein interactions.

1.4.1 LncRNA in epigenetic regulation

LncRNA shows strong epigenetic regulation function by the direct or indirect interaction with other molecules. LncRNAs exhibit epigenetic characteristics that are similar to protein-coding genes, such as DNA methylation and histone modification activity. *HOTAIR*, an oncogenic lncRNA which is expressed in different cancer cells such as breast and colorectal cancer, can

interact with both Polycomb Repressive Complex 2 (PRC2) and lysine-specific histone demethylase 1A (LSD1). PRC2 is a histone methyltransferase that implements epigenetic silencing during different processes including cancer development [18]. LSD1 is involved in demethylation of histone H3 at lysine 4 [19]. Specifically, PRC2 binds to a 5' domain and LSD1 to a 3' domain of HOTAIR, and HOTAIR coordinates their functions for chromatin modification. Through these functions, HOTAIR regulates the expression of multiple genes involved in various pathways [20]. Therefore, the expression level of HOTAIR may serve as a potential diagnostic and therapeutic biomarker for some cancers such as gastric cancer [21] and colorectal cancer [22].

1.4.2 LncRNA in transcriptional regulation

Gene transcription is a tightly and precisely regulated process, in which various units work together to achieve it, including lncRNA. LncRNA could interact with different components of the transcription, including transcriptional factors, RNA polymerase, and even DNA strand. DNA-damage-activated tumor suppressor protein p53 induces the transcription of the lncRNA DINO (Damage Induced Noncoding RNA under the DNA damage pressure. In turn, DINO binds and stabilizes p53, promoting the binding of this transcription factor to the p53-response elements (PRE) of target genes [23]. SOX2, a transcription factor known to regulate neural fate, and lncRNA rhabdomyosarcoma 2-associated transcript (RMST) coregulate a large pool of downstream genes implicated in neurogenesis. RMST is required for the binding of SOX2 to promoter regions of neurogenic transcription factors, which indicates the role of RMST as a transcriptional coregulator of SOX2 and a key player in the regulation of neural stem cell fate [24].

1.4.3 LncRNA in post-transcriptional regulation

In addition to the epigenetic regulation and transcription, lncRNAs are also functional in post-transcriptional regulation such including mRNA processing and further translation. LncRNAs could target mRNA with the complementary base pairing, which may hide key elements in the mRNAs, influencing some processes such as splicing or binding to other functional proteins. Alternative splicing of premature mRNA is utilized by higher eukaryotes to achieve increased transcriptome and proteomic complexity. LncRNA MALAT1 has been reported to be involved in alternative splicing regulation, which interacts with splicing factors and influences the distribution of them in the nucleus [25].

1.5 LncRNA in cancer initiation and progression

LncRNA is a class of ncRNA transcripts that have not been well studied now. However, based on versatile studies so far, we've already known that lncRNAs participated in various aspects of cellular pathways and their potential in disease etiology, especially cancer. Due to the large regulatory network of lncRNA, cancer initiation, progression, and metastasis are believed to be influenced by different lncRNAs, during which the cancer diagnosis, treatment and prognosis are also contributed by them.

As an essential transcription factor, MYC regulates the expression of many genes including cell growth, apoptosis, and differentiation, and is considered as an oncogene. LncRNA PVT1 has been reported to reduce its level in HCT116 human colorectal cancer cell line and then show effects

to the expression of downstream genes [26]. Besides, prostate cancer associated lncRNA transcript 1 (PCAT-1) has been identified as an oncogenic lncRNA in some solid tumors, including prostate cancer. PCAT-1 overexpression promoted prostate cancer cell proliferation, migration, invasion and inhibited apoptosis [27]. A recent study demonstrated that lncRNA MALAT1 could bind and inactivate the prometastatic transcription factor TEAD, preventing TEAD from associating with its co-activator YAP and target gene promoters. MALAT1 level is negatively correlated with breast cancer progression and metastatic ability [28].

Collectively, these examples and the aforementioned examples of lncRNA regulation mechanism emphasize the importance that lncRNA is a force to be reckoned with in cancer. The imbalance of the oncogenic and tumor-suppressing effect may cause the initiation of cancer and influence the whole downstream events. These ideas suggest the potentials of lncRNAs to be the biomarker for cancer diagnosis and therapy.

1.6 Pharmacogenomics data

1.6.1 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) is a database comprising 2.5 petabytes of multidimensional genomic and epigenetic data for more than 11,000 cancer patients across 33 cancer types. This informative database has dramatically facilitated the cancer research for decades in understanding the cancer initiation, progression, and therapeutics.

National Institutes of Health (NIH) launched the TCGA project to comprehensively explore the landscape of genomic alterations in human tumors on December 13, 2015. Since then,

taking the advantage of the high-speed development of next-generation sequencing technologies, researches have created a huge amount of patient data including DNA-sequencing (DNA-seq), RNA-sequencing (RNA-seq), methylation, copy number variation, SNPs and clinical information (such as age at diagnosis, survival, tumor residues, drug treatment profiles, drug response, and prognostic metrics). These data enabled both the TCGA network and independent researchers to explore the association between individual or sets of genes and various cancer disease.

1.6.2 Cancer Cell Line Encyclopedia (CCLE)

The Cancer Cell Line Encyclopedia (CCLE) [29] project is a combination of gene expression, copy number variation, and mutation profiles from 947 human cancer cell lines. Coupling drug response profiles for 24 anti-cancer drugs across 479 of the cell lines, this database enabled researchers to identify predictor or biomarker of drug resistance through different genomic profiles.

1.6.3 Genomics of Drug Sensitivity in Cancer (GDSC)

The Genomics of Drug Sensitivity in Cancer (GDSC) [30] database is one of the largest resources for the information of drug response of anti-cancer drugs in different cancer cell lines. This database now is still under maintenance and being updated, which makes it more and more comprehensive. Currently, this database contains drug sensitivity data for approximately 75,000 experiments, covering response profiles of 265 anti-cancer drugs across more than 1,000 cancer cell lines, including 48 clinical drugs, 76 drugs in clinical development and 141 experimental compounds.

Most cell line drug sensitivity data are integrated with large genomic database obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC) [31] database. These genomic profiles include information on gene expression, copy number variation, somatic mutation, and tissue type. Combining the genotypes with drug sensitivity phenotypes, this database could be considered as an extension of CCLE database due to more drugs covered in it, which provides researchers opportunities to identify new biomarkers for cancer precision therapies.

1.7 Multidimensional multi-omics data analysis

Next-generation sequencing (NGS) has already facilitated depicting the cancer genome and transcriptome profiles in a high speed and large scale. The multidimensional multi-omics data generated from NGS has promoted our understanding on cancer in gene level. Efforts have been taken to create and optimize different methods to do data mining effectively. The relationship between an individual gene and cancer can be studied by simple correlation analysis or statistic test. Gene Set Enrichment Analysis (GSEA), clustering analysis, linear regression, and machine learning methods make it possible that explore the relationship between cancer and a large gene set and identify the inner relation among genes in a gene set.

2.0 Gain of Function Alterations of DNA Damage Repair Pathways Potentiate Therapeutic Resistance in Cancer

DNA damage surveillance and repair are the guardian of human genome integrity, which play pivotal roles in an individual's cancer risk and response to anti-cancer therapy. Insufficient DNA damage repair (DDR) leads to accumulated instability in the cellular genome by generating replication stress together with errors that can be passed into daughter generations of cells. Therefore, loss of the DDR ability will significantly increase the cell's predisposition to become a cancer cell. The loss-of-function (LoF) alterations of DDR genes, including nonsense/deleterious mutations, chromosomal aberrations, and epigenetic silencing in cancer patients have been intensively studied [32, 33], demonstrating that these LoF alterations of DDR genes are "driver" molecular events during cancer development [34, 35].

Moreover, numerous studies have revealed that LoF alterations of DDR genes induce "synthetic lethality" in tumors when treated by a number of genome-instability targeting chemotherapies [36]. The focus of identifying "synthetic lethality" therapeutics in DDR deficient patients has prompted fruitful achievements for cancer therapy [37, 38]. Poly ADP ribose polymerase (PARP) inhibitor (PARPi) is the most remarkable example to induce the synthetic lethality to tumor harboring "BRACness" phenotype [38, 39], which showed success in both *in vitro* studies and cancer patients [18-20]. This led to the FDA's consecutive approval of Olaparib (2014) [40], Rucaparib (2016) [41], Niraparib (2017) [42], and Talazoparib (2018) [43], for the treatment of advanced ovarian cancer and metastatic breast cancer patients with the germline BRCA mutation.

On the other side, increased DDR function is recently demonstrated to introduce the therapeutic resistance of chemotherapy drugs. Previous reports revealed that restoring the homology-dependent recombination (HDR) deficiency (e.g., alterations of 53BP111 or REV712) introduce Cisplatin or Olaparib resistance in BRCA mutated breast cancer patients. However, how frequently the gain-of-function (GoF) alterations in DDR pathways occur in cancer, and to what extent they affect the DNA damage repair and even drug response remain elusive. In this study, we aimed to characterize the GoF alterations landscape of nine DDR pathways in cancer by integrating the multi-dimensional genomic data from primary cancer samples and cancer cell lines across 32 cancer types. By further integrating the DDR gene GoF landscape in cancer with tumor mutation burden, mutation signatures and clinical data of tumor patients, we sought to determine the DDR gene GoF alterations' impacts on the tumor genome instability, patient prognosis, and drug responses. Finally, we systematically scrutinized the association of DDR gene GoF alterations and the pharmacological data of 37 genome-instability targeting drugs in 505 cancer cell lines, which characterized a panorama of DDR genes-cancer drugs interactions and identified novel therapeutic targets or biomarkers for the chemotherapy resistance of cancer

2.1 Methods

2.1.1 Characterization of DDR gene copy number amplification and overexpression across 32 cancer types

RNA-Seq gene expression, somatic mutation and somatic copy number alteration (SCNA) of 80 "core-list" from 276 "full-list" DNA Damage Repair (DDR) genes [44] in 10,489 primary

tumors were obtained from the TCGA PanCancerAtlas cohort consisting of tumor patients across 32 cancer types. The copy number segmentation data (SCNA score) were obtained from the Circular Binary Segmentation (CBS) algorithm [45], and the GISTIC calls comprising -2 (deletion), -1 (loss), 0 (diploid), 1 (gain), and 2 (amplification) were made using GISTIC2.0 [46]. mRNA expressions and copy number alterations of the 80 core DDR genes across 1 005 cancer cell lines were downloaded from Genomics of Drug Sensitivity in Cancer (GDSC) [47]. Genes with over 5% of samples harboring GISTIC call = -2 or 2 in more than two cancer types were defined as recurrently copy number deleted or amplified. A pathway is labeled as amplified in one sample if at least one gene in the pathway showed amplification in the sample. Spearman's rank correlation coefficient was used to detect the correlation between the gene expression and copy number alteration for each gene in the cell lines and patient samples respectively. Gene Set Enrichment Analysis (GSEA) [48] was performed based on the protein-coding gene list ranked by the signed log transformed Spearman's rank correlation p-values using different DDR pathway gene sets from the DDR "full-list" to further interpret the association between the DDR gene amplification and mRNA overexpression.

2.1.2 Assessment of the relationship between tumor genome stability and DDR copy number amplification in the TCGA patient samples

The tumor genome stability information, including mutation burden and mutation signature scores for the PanCancerAtlas tumor patients, was obtained from The Cancer Genome Atlas (TCGA) database. Two-sample t-test was used to show the difference in mutation burden/mutation signature scores between samples containing copy number amplifications (GISTIC call =2) of a specific gene vs. the other samples. Wilcoxon rank-sum test was done for each gene to compare

the mutation burden between amplified samples (GISTIC score = 2) and other samples in a cancer-specific manner. Then GSEA was performed based on the protein-coding gene list ranked by the signed log transformed Wilcoxon rank-sum test p-values using the full-list DDR pathway genes.

2.1.3 Summary for DDR Pathway Alteration across PanCanAtlas

Recurrently amplified/deleted DDR genes: Genes with over 5% of samples harboring GISTIC score = -2 or 2 in more than two cancer types were defined as recurrently copy number deleted or amplified. Pathway level DDR amplification: A pathway is called amplified in one sample if at least one gene in the pathway showed amplification in the sample.

2.1.4 Chemotherapy and radiotherapy information of tumor patients and survival analysis

The raw clinical data of 10,237 TCGA patients across 33 cancer types were obtained from the Genomic Data Commons (GDC). The chemotherapy, radiotherapy and patient survival information [49] were extracted from the raw TCGA clinical data as we previously reported [50] (see Supplementary Methods in Supplement 1). The overall survival rates were estimated by Kaplan-Meier curves between patients with or without specific gene copy number amplification/gain (CNAmp, GISTIC calls = 2 or 1) versus others and compared in the specific cancer types using a Cox regression model stratified by the DDR gene SNCA score.

2.1.5 Association analysis between DDR gene copy number alteration and cell line drug response

Drug response data of 37 genome-instability targeting drugs across 1,005 cancer cell lines were downloaded from the GDSC database (see Supplementary Methods in Supplement 1) and processed as in our previous report [50]. 505 cell lines with multi-dimensional pharmacogenomics data available are retained for the following analysis. A logarithmic transformed half maximal inhibitory concentration (IC50) value was used to indicate the drug response in each cell line. Spearman's rank correlation coefficient was used to between gene copy number alteration and treatment response to each drug. The difference between drug responses in the cell lines bearing different DDR gene copy numbers was determined by Wilcoxon rank-sum test.

2.2 Results

2.2.1 A systematic analysis revealed that DDR genes are significantly amplified and overexpressed in cancer patients

We focused on scrutinizing the gain-of-function (GoF) alterations of 80 “core” DNA damage repair (DDR) genes [44] composing nine major DDR pathways, including Base Excision Repair (BER), Direct Repair (DR), Fanconi Anemia (FA), Homology-Dependent Recombination (HDR), Mismatch Repair (MMR), Non-homologous End Joining (NHEJ), Nucleotide Excision Repair (NER), Translesion Synthesis (TLS), Damage Sensors and Others.

Intriguingly, we observed recurrent DDR gene copy number amplifications/gains (CNAmps) among the 10,489 TCGA cancer samples across 32 tumor types (Figure 1, Table 1). Among the nine DDR pathways, the HDR, which is responsible for the error-free double-strand break repair, was amplified in 76.8% of the pan-cancer cohort and ranked the most frequently amplified DDR pathway. Eighteen out of the 32 cancer types showed an HDR CNAmp frequency over 80% and the top five HDR CNAmp enriched cancer types are ovarian serous cystadenocarcinoma (OV, 544 [96.8%] of 562), esophageal carcinoma (ESCA, 176 [96.7%] of 182), uterine carcinosarcoma (UCS, 54 [96%] of 56), lung squamous cell carcinoma (LUSC, 455 [93.4%] of 487) and rectum adenocarcinoma (READ, 142 [91.6%] of 155).

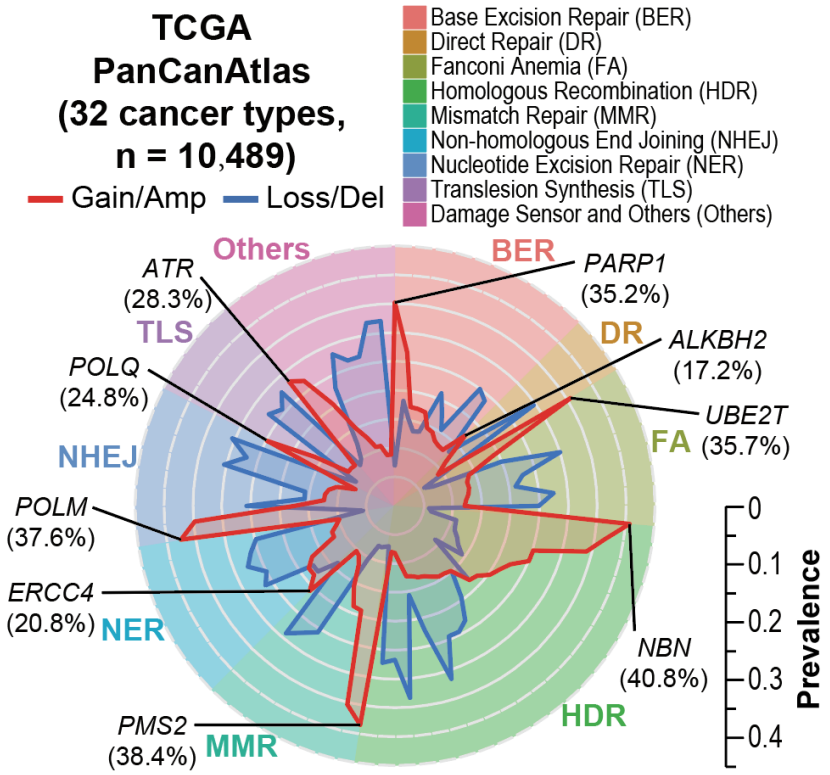


Figure 1 Landscape of the prevalence of DDR gene copy number variations

The radar plot of the prevalence of copy number gain/amplification (red line) and loss/deletion (blue line) events in the 80 "core" DDR genes across nine DDR pathways among the 10,489 TCGA pan-cancer tumors. The most prevalent amplified gene in each DDR pathway is marked in the italic alias with the pan-cancer prevalence. The prevalence of the copy number alteration events was indicated by the scale bar.

On the individual gene level, 13 of the 80 core DDR genes showed significantly recurrent amplification among multiple cancer types. In contrast, only 3 DDR genes showed significantly recurrent deletion under the same criteria (Table 2). The most frequently amplified genes across the pan-cancer are NBN (n = 4,275, 40.8%), EXO1 (n = 3,714, 35.4%), PARP1 (n = 3,695, 35.2%), PRKDC (n = 3,650, 34.8%) and POLB (n = 2,794, 26.6%). All the 13 recurrently amplified DDR genes exhibited significant overexpression in the amplified tumors ($p < 10^{-20}$, Table 2). GSEA analysis revealed that all the 9 DDR pathways' overexpression in the tumors is significantly driven by their CNAmp (FDR < 0.1) (Figure 2). To further validate the CNAmp and overexpression of

DDR genes in cancer, we investigated the DNA copy number and mRNA expression data of 1,005 cancer cell lines from the GDSC databases. The overexpression of the 13 recurrently amplified DDR genes was all significantly driven by CNAmp in the cancer cell lines (Figure 2).

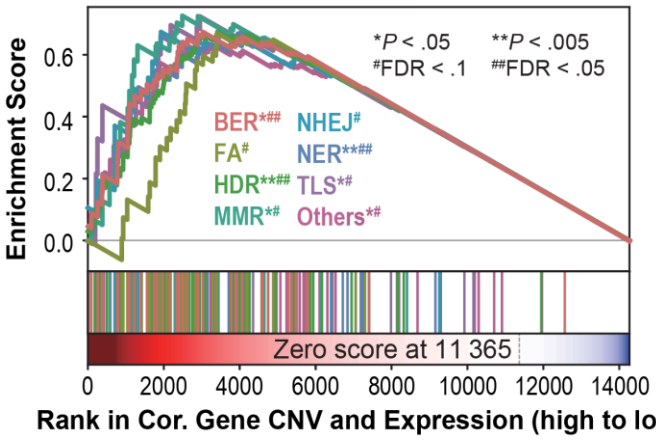


Figure 2 DDR pathways' overexpression in the tumors is significantly driven by their CNAmp

First, we did correlation analysis between gene expression and CNV only in the samples with CNV > 0, then the genes were ranked by the correlation significance and GSEA was done based on the ranking list. DDR pathways are significantly positive correlated, which indicates DDR pathways' overexpression was highly driven by its CNAmp.

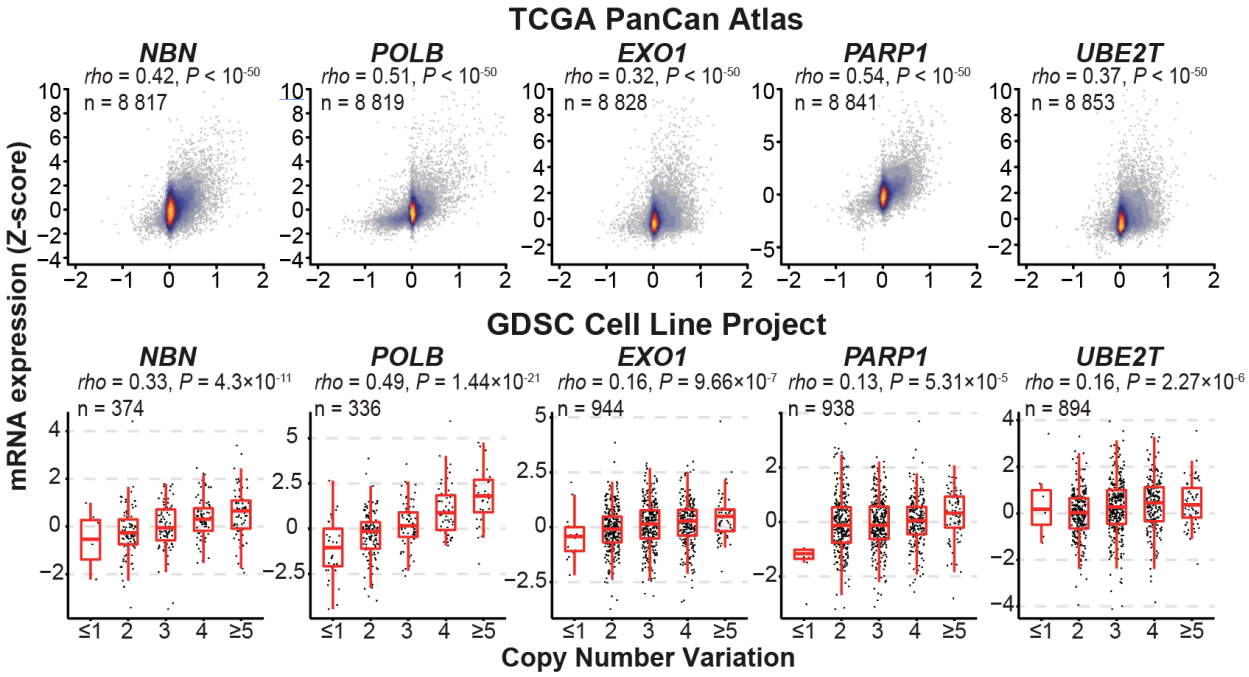


Figure 3 Overexpression of the recurrently amplified DDR genes were driven by copy number amplification

Overexpression of the recurrently amplified DDR genes was significantly driven by copy number amplification in TCGA patient database (upper panel) and GDSC cell line database (lower panel). Gene expression and copy number alterations are represented in y-axis and x-axis, respectively. Spearman's rank correlation analysis was used to see the correlation between gene expression and copy number alterations. The coefficients, p-values, and sample size are plotted under the gene name.

2.2.2 Tumors with DDR gene CNAmp exhibit decreased tumor genome instability and reduced mutational signature scores

With the observation of the significantly recurrent overexpression and amplification of DDR genes in both primary tumors and cancer cell lines, we wonder if these GoF alterations of DDR genes would increase the DNA damage repair function in tumor cells. In this regard, we investigated the mutation burdens between the tumors with or without DDR gene CNAmp. This analysis revealed tumors harboring the CNAmp of 11 individual DDR genes (4 of which are recurrently amplified among multiple cancer types, UBE2T, PARP1, PRKDC, and RAD52) exhibited significantly reduced mutation burden (Figure 4a), suggesting the amplification of DDR genes might lead to an increased DNA damage repair function in those tumors. For example, the amplification of BER pathway, including gene UNG, POLE, TDG, and PARP1, is prominently correlated with genome stability in the OVs, as tumors with stable genome were significantly enriched in the BER pathway gene amplified sample set (NES = 1.802, FDR = .007) (Figure 4b).

Cancer type specific mutation burden analysis at the gene level further confirmed that the association between DDR gene CNAmp and increased tumor genome stability (Figure 4c, Table 3). For instance, POLE is a gene involved in the BER pathway. Its loss-of-function mutations have been established to cause hyper-mutator phenotype in multiple cancer types [51, 52]. In our study,

we found POLE amplified tumor samples exhibited 50% reduced mutation burden than the tumors without POLE amplification in the bladder urothelial carcinoma (BLCA) and OV (Figure 4c), suggesting that the GoF POLE CNAmp associate with an increased DDR function in cancer.

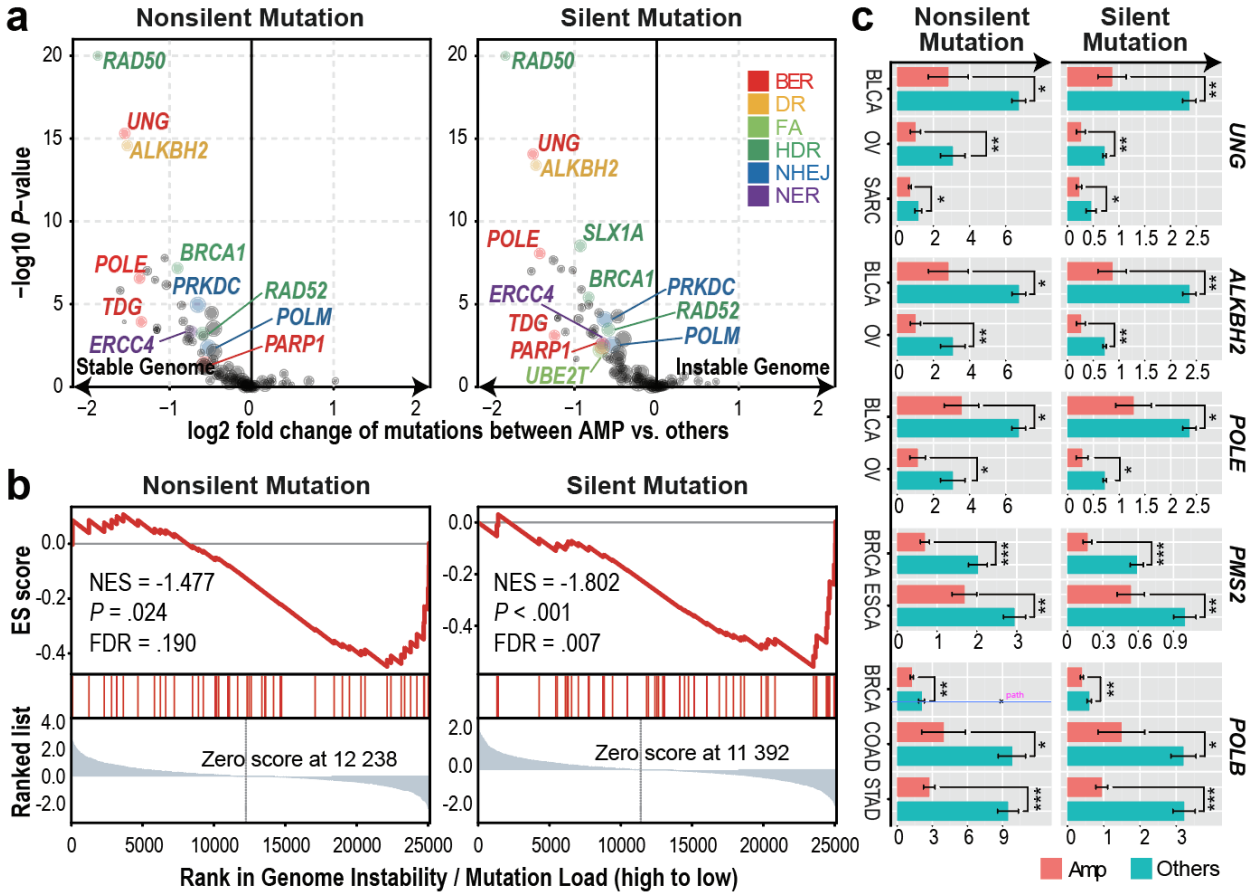


Figure 4 Tumors with DDR gene amplification showed decreased mutation load.

(a) DDR gene amplification is associated with reduced mutation burdens (pan-cancers). Student's t-test results showed significant differences in both the silent and non-silent mutation burdens between tumors with or without amplification of an individual DDR gene. Log2-transformed fold changes of the mutation burden scores between the amplified samples versus others and their corresponding negative log10-transformed t-test p-values were shown in the x-axis and y-axis of the volcano plot.

(b) The GSEA analysis revealed BER pathway amplification is associated with reduced tumor mutation burden in TCGA ovarian cancers. The genes are ranked based on the mutation burden differences between ovarian cancer tumors with or without each gene's CNAmp.

(c) Significant cancer specific silent/non-silent mutation burden reduction in the tumors with amplification (pink) of a specific gene compared to tumors without the amplification (green) by the Student's t-test. Error bars indicate mean ± SEM. *: $p < 0.05$, **: $p < 0.005$, ***: $p < 0.0005$.

Besides using the mutation burden as an indicator for global genome instability, we further compared the mechanistic specific mutation spectrum between the DDR gene amplified tumors and other tumor samples to determine if the DDR gene CNAmp could alter the recurrent mutations accumulated under specific mutation sources and mutation mechanisms [6]. When considering all the 21 previously defined mutation signatures, including smoking-, UVB exposure-, and POLE deficiency-induced mutation signatures, we observed the tumors with DDR gene amplification have significantly lower aforementioned DNA damages (Figure 5). For example, LUSCs, COADs, and OVs with PARP1 amplification have a significantly reduction in smoking-induced mutation signature (LUSC, fold change $< 10^{-3}$, $p = 3.09\times10^{-23}$; COAD, fold change $< 10^{-3}$, $p = 1.74\times10^{-19}$; OV, fold change = 0.45, p = .01) (Figure 5b), which indicates the critical role of the error-free BER pathway genes in amending the smoking-induced DNA lesions [53]. Tumors bearing the amplification of FA pathway genes (FANCB, FANCC and FANCM in lower grade glioma [LGG] and UBE2T in glioblastoma multiforme [GBM]) and HDR genes (ATM, CHECK1, MRE11A in GBM and GEN1 in skin cutaneous melanoma [SKCM]) showed significantly reduced temozolomide-induced signature score (Figure 5d), indicating the critical role of double-strand breakage (DSB)-associated recombinational repair in attenuating the alkylating agent-induced genome lesions [54]. These observations suggest that the GoF alterations of DDR genes would restore the DNA damage repair function in tumor cells, thus alleviate the genome lesions and maintain the genome stability in the tumor.

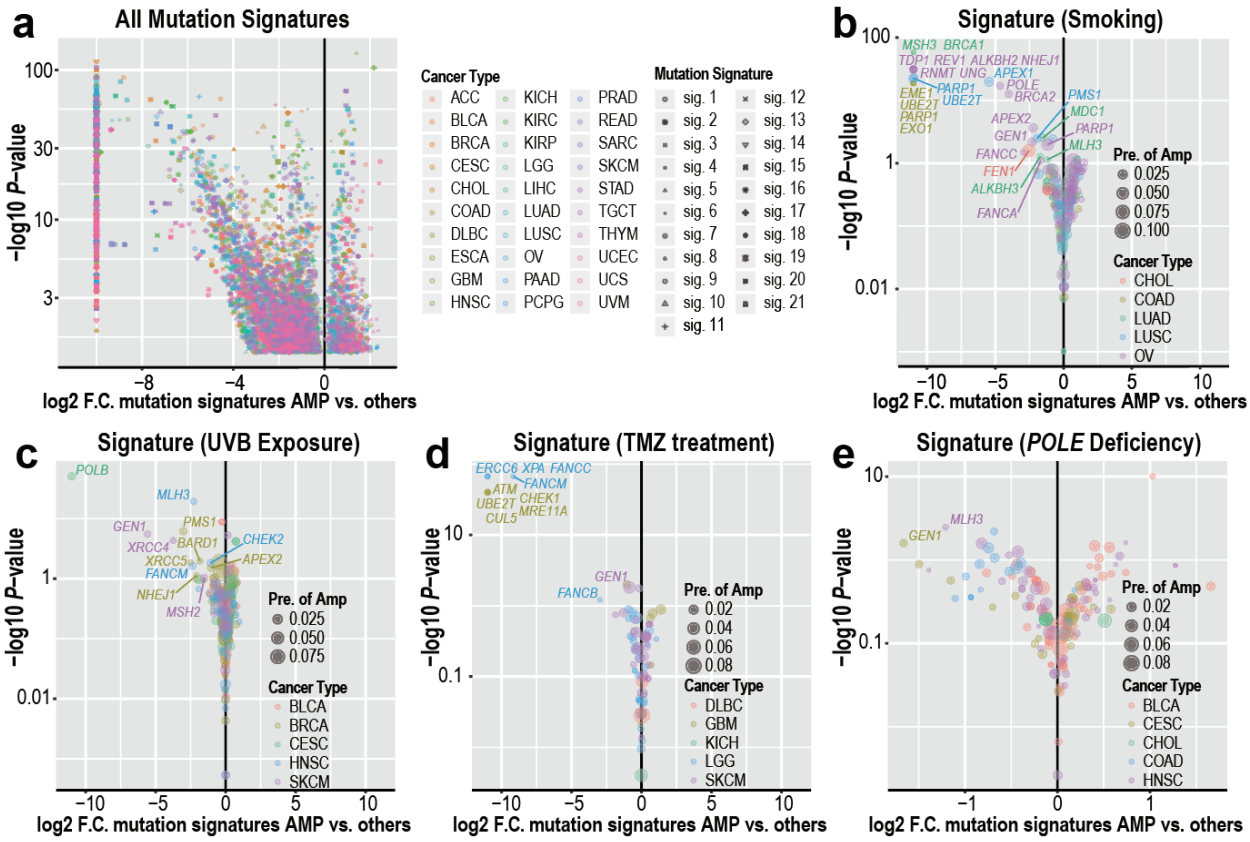


Figure 5 Tumors with DDR gene amplification showed decreased mutation signatures

(a) DDR gene amplification is significantly associated with reduced mutation signature scores of 21 different mechanisms (shapes) in each cancer type (colors) by Student's t-test. Log2-transformed fold change of each mutational signature score between samples with or without amplification of an individual DDR gene and its negative log10-transformed t-test p-values were shown in the x-axis and y-axis of the volcano plot respectively.

(b-e) Mechanism-specific, such as smoking- (b), UVB exposure- (c), Temozolomide treatment- (d), and POLE deficiency- (e) induced mutation signatures decreased in the DDR gene-amplified tumors in different cancer types. The size of each circle indicates cancer-specific gene amplification prevalence in the pan-cancer samples and colors specify diverse cancer types.

2.2.3 DDR gene CNAmp in the tumor is significantly correlated with poor cancer survival

To investigate whether the DDR gene amplification is clinically relevant, we analyzed the correlation between patient overall survival and DDR gene CNAmp in each cancer type. As shown in Figure 6a, the CNAmp of core DDR genes exhibited a broad correlation with unfavorable survival of tumor patients. For example, PMS2 is a critical component of the MutL alpha heterodimer for the initiation of mismatch repair [55]. The amplification of PMS2 gene is frequently found in GBM (454 [79.5%] of 571), LGG (134 [22.4%] of 510), HNSC (205 [39.7%] of 517) and OV (132 [23.5%] of 562). Those patients have significantly shorter survival comparing to non-CNAmp patients in each cancer type (GBM, HR = 1.53, 95 CI 1.22 to 1.92, $p = 1.98\times10^{-4}$; LGG, HR = 2.32, 95% CI 1.69 to 3.19, p = 2.33×10-7; HNSC, HR = 1.58, 95% CI 1.24 to 2.02, $p = 2.22\times10^{-4}$; OV, HR = 1.42, 95% CI 1.12 to 1.79, $p = 3.48\times10^{-3}$) (Figure 6b). Another recurrently amplified DDR gene, POLM, plays an essential role in the NHEJ repair for DSB [56]. Poor survival was observed in patients with POLM amplification in multiple cancer types (HNSC: CNAmp frequency = 36.4% [188 of 517], HR = 1.40, 95% CI 1.09 to 1.81, $p = 9.17\times10^{-3}$; LGG: CNAmp frequency = 23.5% [120 of 510], HR = 2.47, 95% CI 1.78 to 3.44, $p = 6.83\times10^{-3}$; and LUAD: CNAmp frequency = 51.9% [265 of 511], HR = 1.52, 95% CI 1.15 to 2.03, $p = 3.83\times10^{-3}$) (Figure 6c). PRKDC's amplification also significantly correlated with poor patient survival in multiple cancer types (SARC: CNAmp frequency = 35.2% [89 of 253], HR = 2.15, 95% CI 1.46 to 3.18, $p = 1.12\times10^{-4}$; UCEC: CNAmp frequency = 31.5% [165 of 523], HR = 1.72, 95% CI 1.21 to 2.44, $p = 2.50\times10^{-3}$) (Figure 6d).

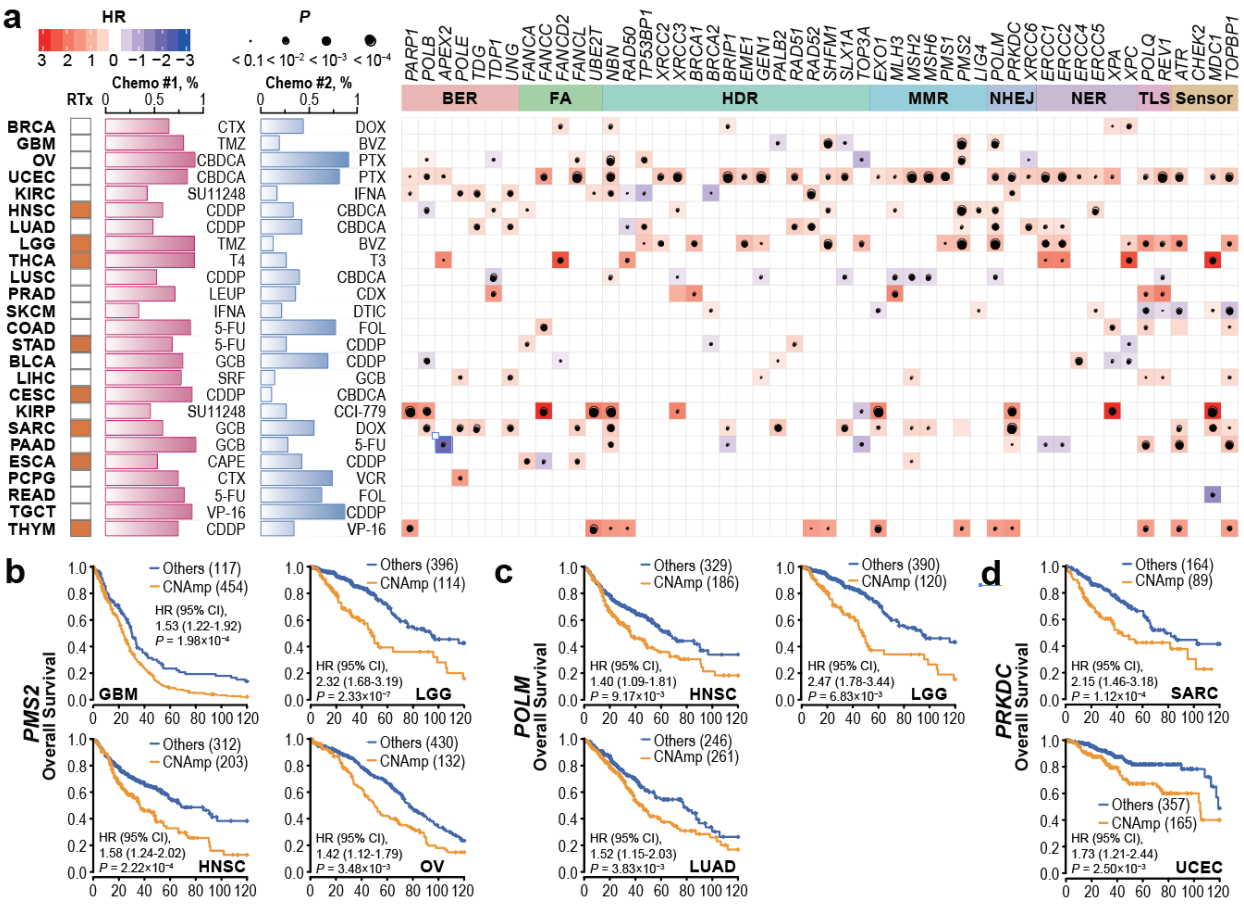


Figure 6 DDR gene amplification in the tumor correlates with poor patient survival

(a) Tumor patients with specific DDR gene CNAmps (GISTIC calls = 1 and 2) showed poor overall survival comparing to the rest of the patients by the Cox regression model. The radiotherapy and chemotherapy background of each cancer type is shown in the left panel. RTx: radiotherapy; Yellow: radiotherapy is commonly applied. Chemo #1, Chemo #2: the top 2 commonly used chemotherapy agents; percentage of patients received each agent was indicated by the bar chart. CTX: Cyclophosphamide; DOX: Doxorubicin; TMZ: Temozolomide; BVZ: Bevacizumab; CBDCA: Carboplatin; PTX: Paclitaxel; SU11248: Sunitinib; IFNA: Interferon A; CDDP: Cisplatin; T4: Levothyroxine; T3: Liothyronine; LEUP: Leuprolide; CDX: Bicalutamide; DTIC: Dacarbazine; 5-FU: Fluorouracil; FOL: Leucovorin; GCB: Gemcitabine; SRF: Sorafenib; CCI-779: Temsirolimus; CAPE: Capecitabine; VCR: Vincristine; VP-16: Etoposide. Right panel: Heatmap for blue and red represented negative and positive hazard ratios respectively and the p-value is denoted by the dot size. DDR genes with CNAmp in over 35% of pan-cancer patients were presented here.

(b-d) PMS2 (b), POLM (c), and PRKDC (d) amplification is significantly associated with poor survival in multiple cancer types. The overall survival rates were estimated by Kaplan-Meier curves between patients with or without specific gene CNAmp versus others and compared in the specific cancer types.

2.2.4 Pharmacogenomics analysis unveiled an overall significant correlation between the DDR gene CNAmps and the genome-instability targeting drugs

Previous studies have demonstrated that restored DDR function leads to chemotherapy resistance and thus poor patient survival [57]. The observation of the significant positive correlations between DDR gene CNAmp with reduced tumor mutation burden, mechanism specific mutation signatures, and poor patient survival lead to our hypothesis that these DDR gene GoF alterations may cause poor patient survival by augmenting DNA damage repair function and consequently chemotherapy resistance in the tumor. Among the recurrent DDR gene CNAmps, NBN CNAmp is the most prominent molecular event that occurs in over 40% patients across 16 cancer types. The NBN's overexpression is highly driven by its CNAmp in both primary tumors (TCGA database, $p = 2.50\times10^{-60}$) and cancer cell lines (GDSC database, $p = 3.94\times10^{-5}$). Moreover, NBN CNAmp is most prominently correlated with poor overall survival in OV patients (HR = 1.36, 95% CI 1.13 to 1.64, $p = 9.62\times10^{-4}$). We validated NBN's CNAmp and overexpression in an independent cohort including 22 serous ovarian cancer samples. These assays independently confirmed that the NBN protein overexpression is highly associated with its CNAmp (Fisher's exact test, $p = 0.019$). We also demonstrated that the NBN CNAmp induced genome-instability drugs Cisplatin and Olaparib resistance *in vitro* (data not shown).

The resistance to chemotherapy is one of the major barriers to improving cancer survival. One-third of the current FDA approved anti-cancer therapy drugs are targeting genome instability

or DNA replication [58]. Encouraged by the experimental validation that NBN CNAmp leads to Cisplatin and Olaparib resistance, we wondered if other DDR gene CNAmps associate to poor patient prognosis through prompting the resistance to genome instability targeted chemotherapy. In this regard, we further integrated the copy number alterations data across 505 cancer cell lines and their responses to 37 genome-instability targeting drugs (i.e., 23 DNA-damaging drugs and 14 cell cycle/TP53 targeted agents) in GDSC. This analysis revealed the landscape of DDR gene CNAmps and response to genome-instability targeting drugs including 468 significant associations between the 80 DDR gene CNAmps and the 37 genome-instability targeting drugs (Figure 7). Among the 468 significant associations, 430 (92%) significant positive correlations indicated DDR CNAmp lead to drug resistance, suggesting GoF alterations of DDR gene might induce the resistant phenotype to chemotherapy targeting genome-instability.

Gene-level analysis revealed the CNAmp of FANCM, the pivotal component of Fanconi anemia (FA) pathway that relieves the DNA inter-strand cross-link [59], has significant correlations with the cellular responses to the DSB inducing agents such as topoisomerase inhibitor (Camptothecin, $p = 5.54\times10^{-3}$), CHECK1/2 inhibitor (AZD7762, $p = 1.90\times10^{-5}$), and PARP inhibitors (Olaparib: $p = 3.34\times10^{-3}$, Veliparib: $p = 1.29\times10^{-4}$), which is consistent with the previous studies showing FANCM is intensively involved in DDR and regulates the PARPi sensitivity [60].

On the pathway level, the cancer cell lines with CNAmp of DDR genes across BER (5 out of 10 genes), FA (3 out of 8 genes) and HDR (7 out of 21 genes) pathways exhibited significant higher IC50s to Camptothecin ($p < 0.05$).

Consistent with our previous observation, HDR pathway CNAmp is associated with increased IC50 of PARPi. 6 (29% of 21) HDR genes (NBN, GEN1, BARD1, RAD50, BRCA1, and BRIP1) showed significant positive correlations between the gene copy number alterations

and cell line responses to Veliparib and Olaparib treatments respectively ($p < 0.05$). Together with our previous experimental phenotype validation of NBN overexpression in the cancer cell lines, this study further suggested that the observed significant correlation between DDR gene CNAmp and poor patient survival may be attributed to the increased DNA damage repair function and chemotherapy resistance in tumors with DDR gene CNAmps.

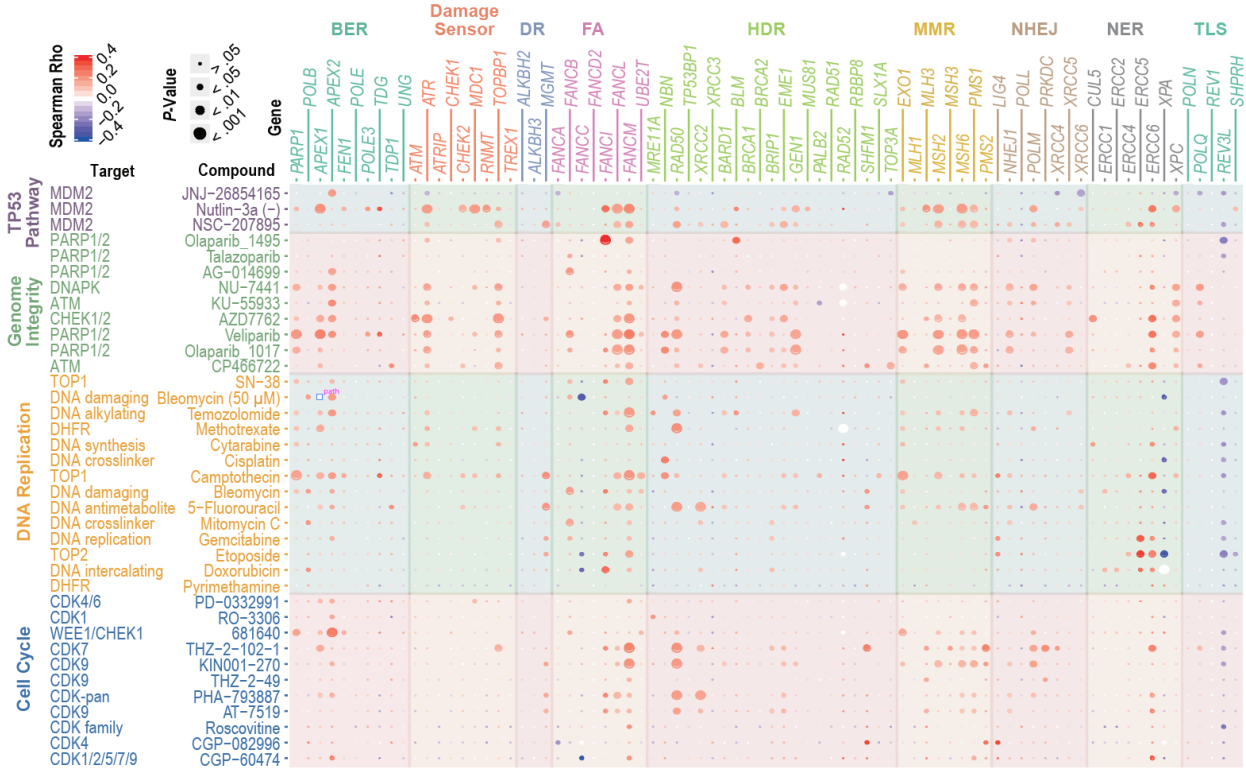


Figure 7 Landscape of the correlation between DDR gene copy number drug response in GDSC

DDR gene copy number is highly associated with the in vitro response of drugs targeting genome-instability. The Spearman's rank correlations test was used to determine the correlation between the copy number alteration of each of the 80 core DDR genes and the log-transformed IC50 of the 37 genome-instability targeting drugs across the cancer cell lines. The color and size of each bubble indicate the Spearman's rank correlation coefficients and p values, respectively.

2.3 Discussion

In the current study, by integrating multi-dimensional genomics and clinical data in tumor patients and cancer cell lines, we demonstrated that DNA damage repair (DDR) genes' copy number amplification/gain (CNAmp) and overexpression not only recurrently occur across 32 cancer types but also correlates with increased DNA damage repair capabilities in the pan-cancer tumor samples. By integrating the clinical follow-up data, we found DDR gene CNAmp is significantly associated with reduced patient survival, which could be majorly attributed to the chemotherapy resistance induced by DDR genes' alleviation of the genomic scars in cancer accumulated during drug treatments. To determine the causal relationship between DDR gene CNAmp, elevated DNA repair capacity, and increased chemotherapy resistance, we have experimentally validated that augmented NBN expression leads to increased homology-dependent recombination (HDR) efficiency. We have further shown that the HDR gene (i.e., NBN) CNAmps can induce drug resistance (Cisplatin and Olaparib) in breast cancer and ovarian cancer cells.

DDR pathway deficiencies have been well-established as drug-actionable targets for cancer therapy. However, the copy number amplifications of DDR genes, which may play critical roles in the therapeutic resistance of cancer, have long been neglected in the cancer study and treatment. Previous studies reported the restoration of HDR function by somatic reversion of germline BRCA1/2 mutations confers platinum and PARPi resistance in ovarian cancer [53, 79]. In our study, we have unveiled a general correlation between the DDR gene CNAmps and the drug treatment resistance of 37 genome-instability targeting drugs by integrating the pharmacogenomics data across 505 cancer cell lines. Our study not only provides a landscape of DDR genes and anti-cancer drugs interactions but also a novel molecular mechanism for the intrinsic resistance of genome-instability targeting chemotherapy. Note that all the 10,489 tumor

samples and 505 cancer cell lines in this study are chemotherapy naïve, which means the DDR gene CNAmps have already existed in some tumors before chemotherapy. It is possible that some patients have a subclone of cancer cells with DDR gene CNAmp in their primary tumors. Those patients will initially respond to DDR targeting drugs, and later on develop resistance to the drug after DDR gene CNAmp subclone start to expand and become the major clone under the drug selecting pressure [80, 81]. One limitation of our study was that we were not able to determine if DDR gene CNAmp occurs in all cancer cells or only a subclone within a tumor. Further investigation harnessing single-cell sequencing technology [82-84], and clinical follow-ups are required to comprehensively decipher the roles of DDR gene GoF events in cancer development and drug resistance.

The DDR gene CNAmps may also serve as reliable biomarkers for cancer precision therapy. Tremendous efforts have been invested to develop clinical actionable methods to measure DDR function and thus predict clinical outcome and chemotherapy response in various cancer types. These findings suggest the copy number quantification of DDR genes at DNA level can be a promising biomarker for the beneficiary identification, response evaluation and prediction of the genome-instability targeted therapy, although its clinical feasibility awaits validation from further clinical trials.

3.0 Systematic correlation analysis identified PiHL as a wild-type p53 suppressing lncRNA in colorectal cancer

To facilitate aberrant proliferation and cell survival during tumor progression, a number of genetic alterations are typically selected for in cancerous cells [61]. Among these alterations, somatic copy number variants (CNVs) play important roles in various cancers, including colorectal cancer (CRC) [62]. Although mammalian genomes are widely transcribed, the vast majority of these transcripts are non-coding RNAs (ncRNAs), among which are long non-coding RNAs (lncRNAs) with a length of over 200 bases. Studies have pointed to the emerging roles of lncRNAs locating at these aberrant chromosome regions in tumor development. For example, the copy number amplification of lncRNA FAL1 was found to be associated with clinical outcomes in patients with ovarian cancer [10]. Therefore, linking cancer-associated CNVs to lncRNAs will provide independent support for functional implications and lead to a greater understanding of cancer pathogenesis.

In its wild-type state, p53 is an important tumor suppressor and the p53 pathway is activated in the presence of cellular stress, such as DNA damage and oncogenic signaling, and in turn, coordinates the transcriptional response of hundreds of genes [63]. As a haplo-insufficient gene, a relatively small decrease of p53 level or activity can largely impact tumorigenesis [64]. P53 activation can initiate multiple pathways that lead to a temporary pause at a cell-cycle checkpoint to allow for DNA repair, permanent growth arrest (senescence), or cell death (apoptosis) [65]. Recently, Several molecules have been implicated in regulating p53 protein synthesis including translation initiation factors [66], RNA-binding proteins (RBPs) [67] and MDM2 [68]. LncRNAs have been implicated in post-translational regulation of p53. For example,

p53-induced lncRNA DINO can bind to the p53 protein and promote its stabilization, regulating cell cycle arrest and apoptosis in response to DNA damage [23]. While lncRNAs are known to be involved in p53 pathways, the role of lncRNAs in regulating the p53 protein remains mostly unknown.

LncRNAs with a high level of copy number amplification/deletion in cancer, a high level of expression/methylation discrepancy between cancer and normal tissue, or a strong correlation with some well-defined tumor regulating gene are usually tumor driven lncRNA. We integrated our multi-omics data including lncRNA expression and copy number variation, protein expression, and mutation status data to identify p53 regulating lncRNA in CRC. p53 is a strong tumor suppressor and TP53 mutation is common in tumor, we hypothesized that p53 protein levels in TP53 wild-type and mutated patients are different, and we found it is higher in TP53 wild-type samples. Considering there is often a loss of function of TP53 protein product (p53), our analysis mainly focused on the TP53 wild-type CRC samples and took TP53 mutated samples as a negative control.

In this study, we identified and characterized a novel long intergenic non-coding RNA PiHL (RP11-382A18.2). PiHL's copy number amplification is significantly concurred with p53 protein downregulation without influencing its mRNA level. PiHL is upregulated in CRC and is associated with poor prognosis of CRC patients. Functional screening study revealed PiHL's role in maintaining CRC cell proliferation and inhibiting apoptosis *in vitro* and *in vivo* in p53 wild-type cancer cells.

3.1 Methods

3.1.1 Collection and processing of genomics data

Gene expression, GISTIC copy number alteration, RPPA, and whole-exome mutation data were downloaded from TCGA Pan-Cancer Project. 23,117 genes, including 1,025 long non-coding intergenic RNAs and 18,706 protein-coding genes, were annotated in 589 TCGA colorectal patient samples by GENCODE (v22, GRCh38).

3.1.2 Genome-wide copy number prevalence and correlation analysis

Colorectal cancer patient samples were first divided into two groups, TP53 wild-type samples and TP53 mutant samples. In each group, mRNA expression data is logarithmic transformed first. Spearman's rank correlation analysis was used to understand the correlation between the CNV and TP53 mRNA expression or sp53 protein level. Copy number frequencies of amplification or deletion were defined as the percentage of the samples with CNV > 1 or CNV < -1.

3.1.3 Identification of candidate gene regulating p53 protein.

Integrating previous correlation and copy number prevalence analysis, regions with high copy number alteration prevalence were first taken into consideration. Regions containing genes with high correlation with TP53 mRNA level but weak correlation with p53 protein level was considered as regions regulating TP53 mRNA. Regions containing genes with high correlation

with p53 protein level but weak correlation with TP53 mRNA level was considered as regions regulating p53 protein. In the candidate regions regulating p53 protein, Wilcoxon's rank-sum test was done to show the differential expression of each gene between 644 tumors and 51 normal samples. Spearman's rank correlation was done to delineate the association between gene expression and its own CNV. We set 2 for the filter of the fold change of the expression level between tumors and normal tissues, which means the genes we identified as p53 regulating genes are upregulated in tumor. We also set 10^{-12} as the significant cutoff of the correlation analysis between gene expression and its own CNV, considering a possible multi-test problem and sample-size problem. After the identification of candidate genes, *in vitro* functional screening was done to further identify the p53 regulating gene. (Figure 8)

3.1.4 Other data analyses.

Integrative Genome Browser (IGV) was used to delineate the copy number alterations in different regions. Kaplan-Meier survival analysis was utilized to compare the patients with high PiHL expression level and low PiHL expression level. Student's t-test was conducted to compare the expression level between tumors and normal tissues of the clinical cohort patients.

3.1.5 In situ RNA hybridization (ISH) for PiHL

Tissue microarray (TMA) chips of 100 paired colorectal cancer patients' tumor and normal tissues were incubated with a double-DIG-labeled custom LNA probe for PiHL (5DigN-TTGGACACTGCATCAATAGTT-3DigN, Exiqon, Denmark) and detected with polyclonal anti-DIG Fab fragments (Roche, USA) and alkaline phosphatase conjugated secondary antibody

(Invitrogen) using NBT-BCIP as the substrate. TMA were then counterstained with nuclear fast red staining solution (Sigma Chemical Co, USA). High-resolution images were captured with an Aperio Scan Scope AT Turbo (Aperio, Vista, CA, USA) equipped with Aperio Image Scope software (Aperio). Assessment of the staining was based on the staining intensity and the percentage of positively stained cells using Image-Pro Plus 6.0 software (Media Cybernetics, Inc., Silver Spring, MD, USA). The median signal of PiHL positive staining was defined as cutoff value.

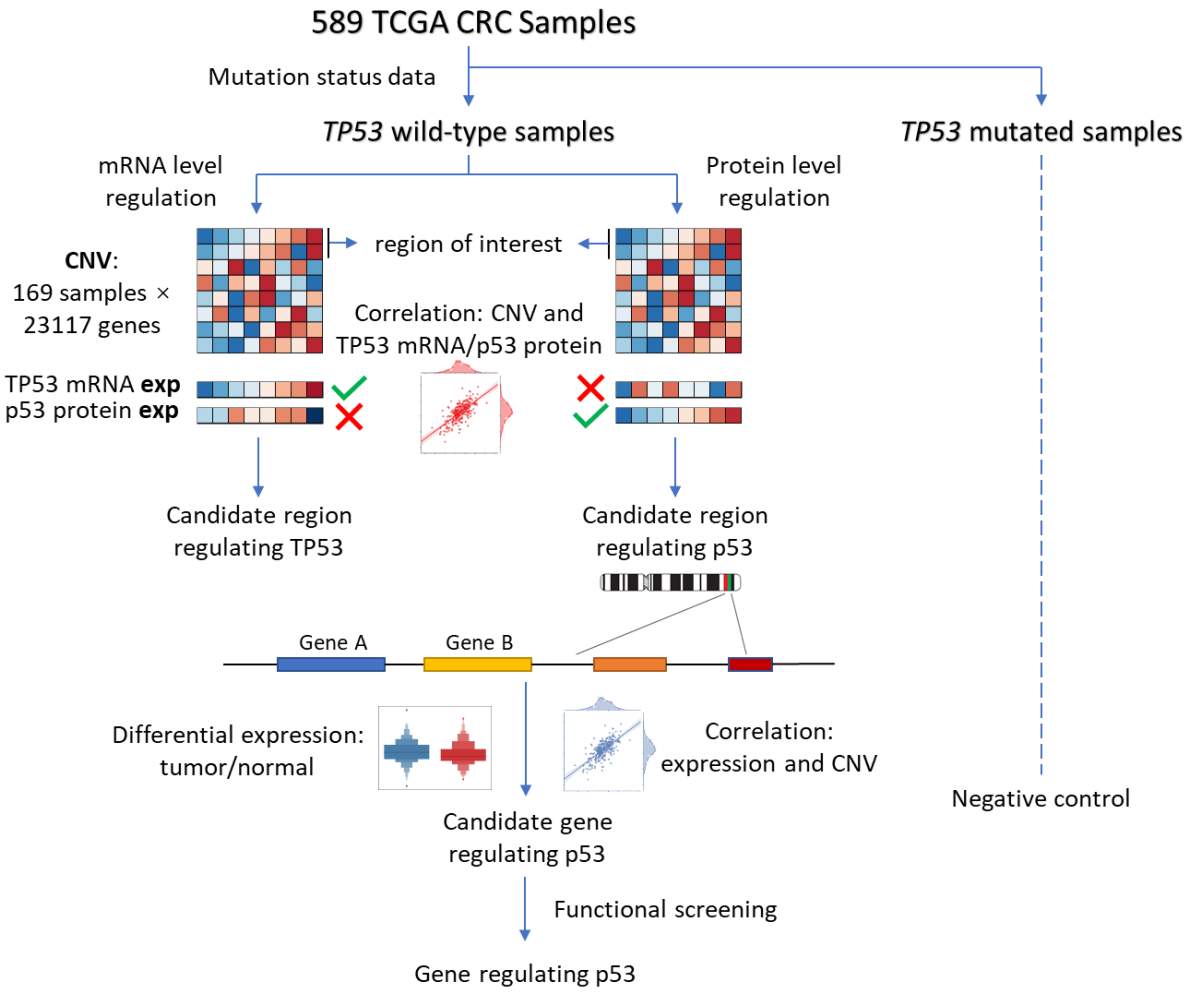


Figure 8 Schematics of the identification of gene regulating p53

3.2 Results

3.2.1 Overview of CNV prevalence in CRC patients

We overviewed the CNV prevalence in TCGA CRC patient (Figure 9). The CNV frequency of each gene is the percentage of samples with this lncRNA amplification (GISTIC call >= 1) or deletion (GISTIC call <= -1) versus all TP53 CRC samples. Some regions have a higher frequency of copy number amplification, and some have a higher frequency of copy number deletion. Regions with a higher frequency of copy number amplification (or deletion) are usually not harboring higher frequency of deletion (or amplification), indicating a mutual exclusivity of these two events. Consisting with previous report, TP53 in chr17 harbors a high level of copy number deletion (59.5%) in our analysis.

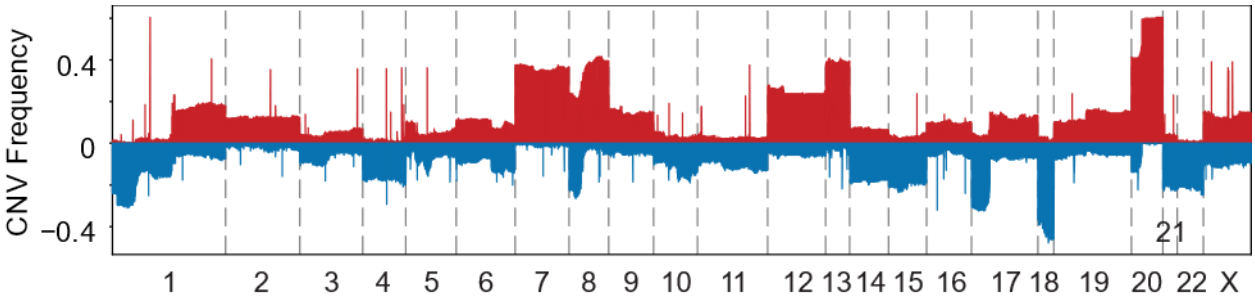


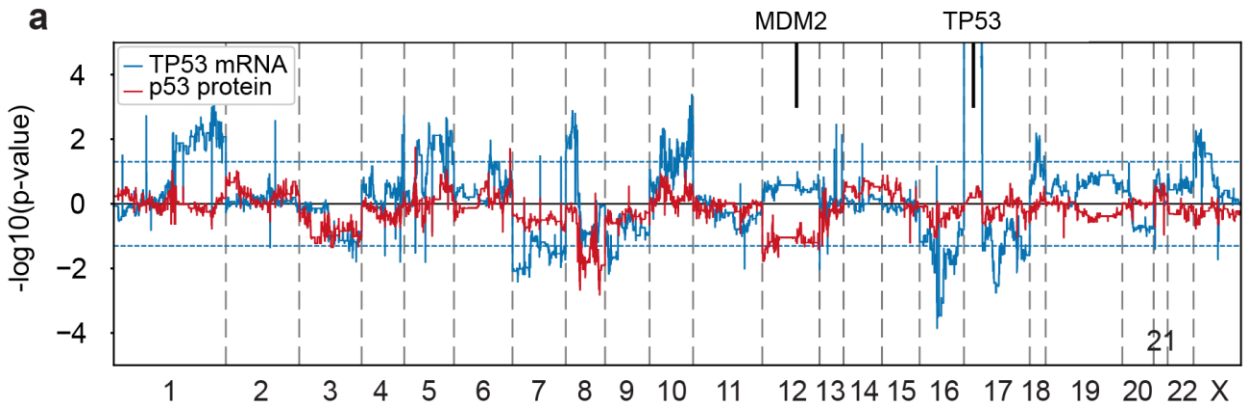
Figure 9 Landscape of CNV frequency of the whole genome

Prevalence of gene copy number amplification (red) and deletion (blue).

3.2.2 Correlation analysis between p53 protein/TP53 mRNA level and gene copy number variations

High level CNV of a gene in tumor is usually a driven event for cancer, and gene CNV may correlate with some tumor biomarker or tumor regulators, such as p53. We reasoned that if

the copy number alterations of some genes can influence p53 protein expression without changing TP53 mRNA levels, these genes may be involved in regulating p53 protein stability. Correlation analyses of the copy number altered lesions with p53 protein or mRNA expression in 169 p53 wild-type CRC tumors identified 24 regions correlated with p53 protein levels and 137 regions correlated with TP53 transcription. The most prominent regions regulating TP53 transcription is 17p13.1, where TP53 gene is located (Figure 10a). Consistent with previous report [69], TP53 is deleted in 49 (29%) p53 wild-type CRC tumors and tumors with TP53 deletion have significantly lower TP53 mRNA expression ($p = 1.34\times10^{-7}$, Figure 10b, d). The coding gene of MDM2 protein, a well-established E3 ligase of p53 protein ubiquitination [70], is located in one of the top regions regulating p53 protein identified by our analysis. We observed that MDM2 was amplified in 47 (27.8%) CRC tumors, and tumors with MDM2 amplification have significantly higher MDM2 mRNA expression ($p = 5.05\times10^{-7}$, Figure 10c, e). These data confirmed the capability of our approach to identify genomic copy number gains or losses in CRC, and to discover potential p53 regulators as the basis for further analysis.



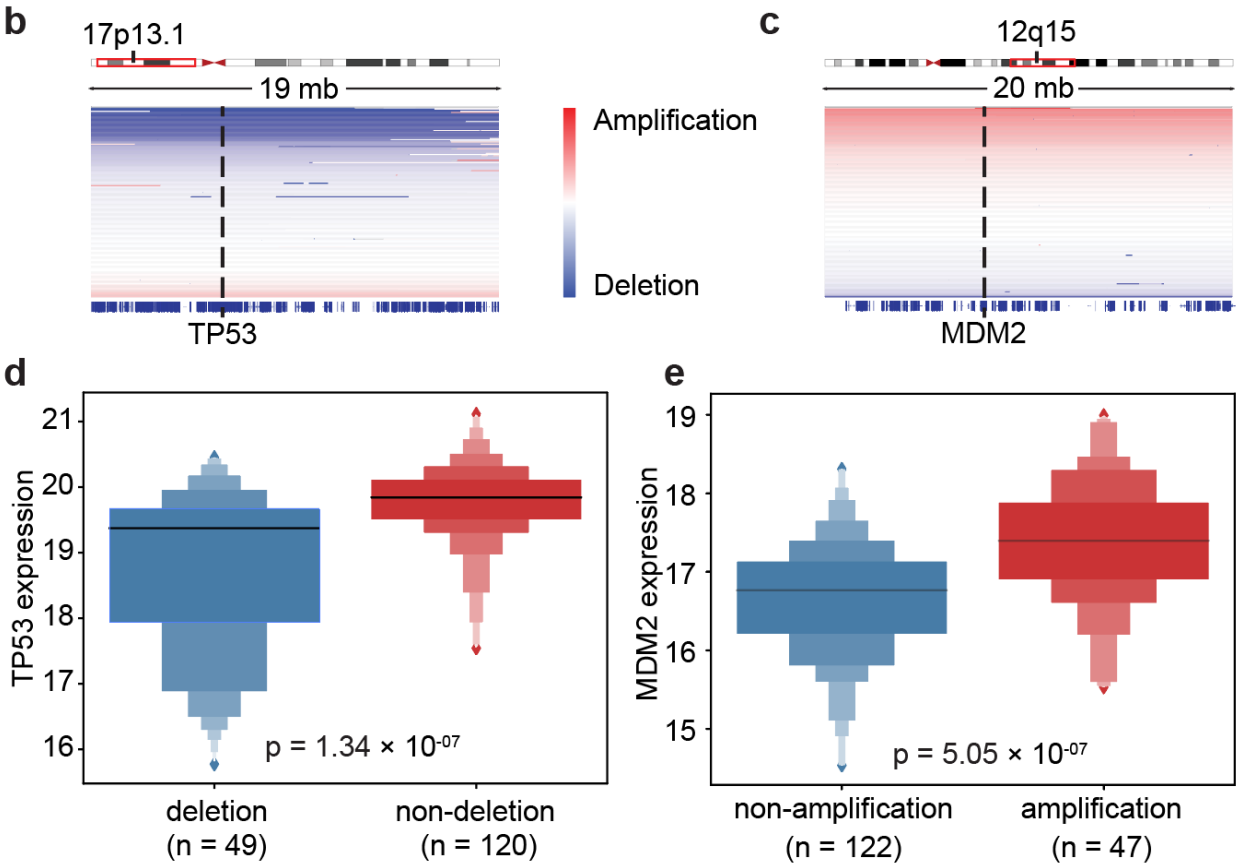


Figure 10 A robust genome-wide correlation analysis identifying p53 regulating gene

(a) Correlation between genome-wide gene CNV and TP53 mRNA expression/p53 protein level in TP53 wild-type samples. For correlation analysis results, blue and red line correspond to the correlation with TP53 mRNA and p53 protein level, respectively.

(b, c) IGV figures showing the copy number alterations of the regions around TP53 and MDM2 in wild-type samples. Deletion: copy number deletion; Amplification: copy number amplification.

(d, e) TP53 mRNA differential expression between TP53/MDM2 deletion and other samples.

3.2.3 Differential expression profiles of genes in 8q24.21 region identified a novel lncRNA

Intriguingly, one of the most prominent regions negatively correlated with p53 protein in p53 wild-type but not p53 mutant samples is chromosome 8q24.21. Chromosome 8q24.21 was

reported to be frequently amplified in CRC and includes genes such as PVT1 and CCAT1. However, none of these genes have been reported to regulate p53 protein stability. RNA-Seq data were utilized to analyze the expression changes of genes located in 8q24.21 in CRC samples compared to normal samples. Among upregulated genes, the expression of CCAT1, PVT1, and a lncRNA RP11-382A18.2 were found to be significantly correlated with their copy number alteration (Figure 11).

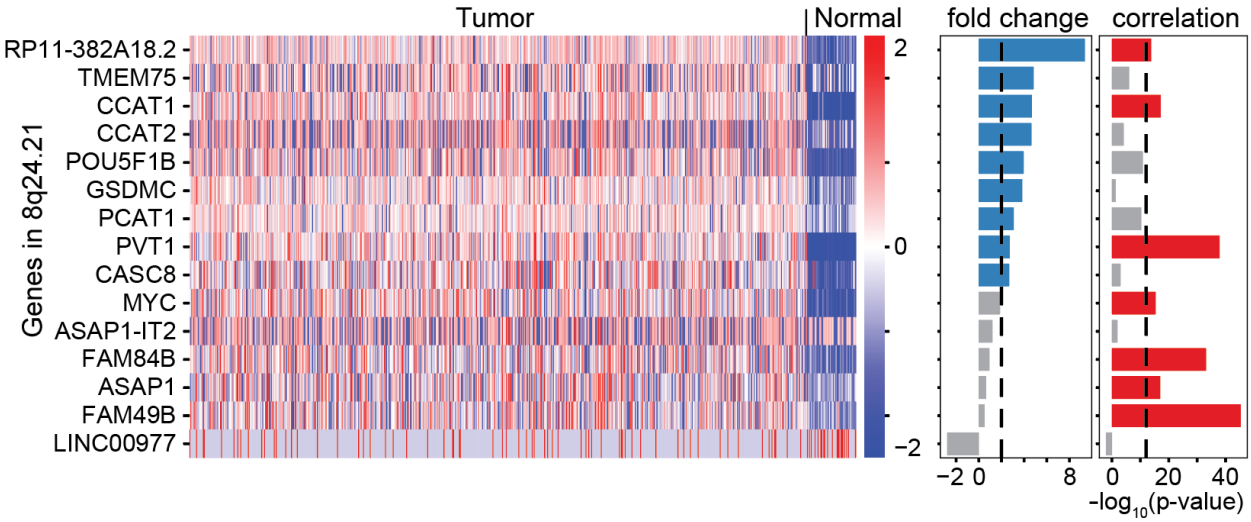


Figure 11 Differential expression of genes in 8q24.21

Heatmap showing the gene expression in 466 tumors and 51 normal samples. Fold change of the gene expression of tumor versus normal and correlation of gene expression and its copy number are also plotted on the right of the heatmap. The dash lines indicate the filter cutoff 2 and 10^{-12} for the of the fold change and correlation between gene expression and CNV, separately.

3.2.4 *In vitro* confirmed inhibition on p53 protein by lncRNA PiHL

Knockdown of these three lncRNAs using siRNAs revealed that only silencing of RP11-382A18.2 strongly upregulated p53 protein levels but not mRNA expression in p53 wild-type HCT116 cells (Figure 12a, b, d). Thus, we named this lncRNA PiHL (P53 inHibiting LncRNA). To further confirm PiHL's regulation on p53 protein, we activated the endogenous transcription

of PiHL gene using the CRISPR synergistic activation mediator (SAM) system [71]. Consistently, activation of PiHL downregulated p53 protein but not p53 mRNA levels in HCT116 cells (Figure 12b). In aggregate, genomic, transcriptomic and proteomic analyses with functional screening identified lncRNA PiHL as a potential p53 protein regulator.

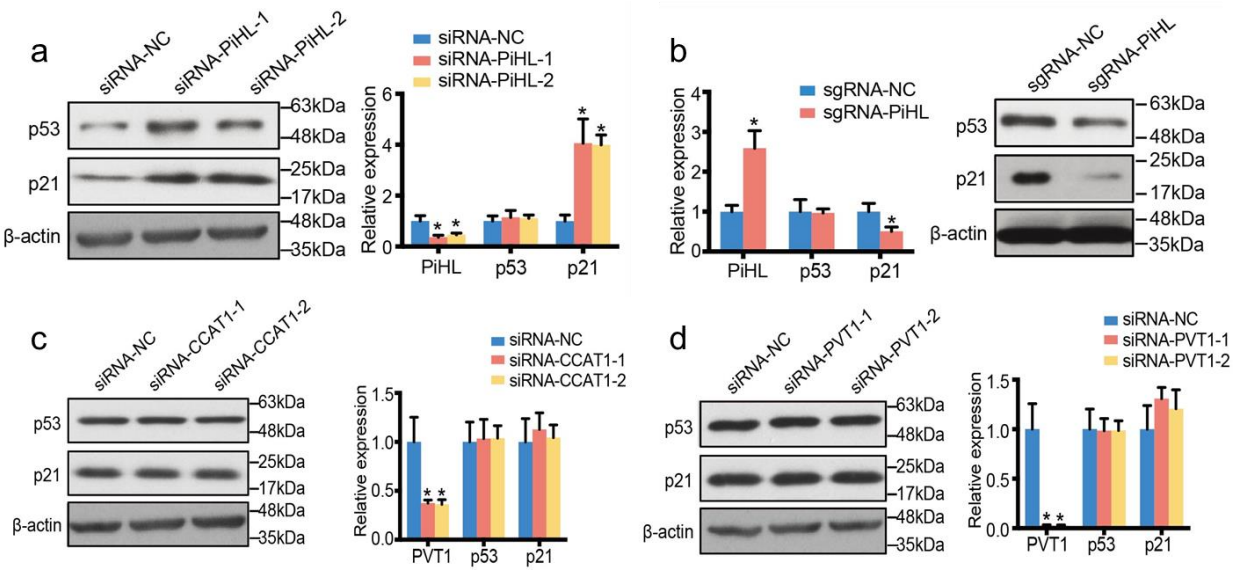


Figure 12 *In vitro* study confirmed that PiHL inhibits p53 protein but not TP53 mRNA level

(a, c, d) Western blot and qRT-PCR analysis of p53, p21, and PiHL (a), CCAT1 (c) or PVT1 (d) expression. β-actin served as the control. Data are shown as mean ± s.e.m.; Student's t-test.

(b) Western blot and qRT-PCR analysis of p53, p21 and PiHL expression upon single guided RNAs (sgRNAs) transfection with the SAM system in human HCT116 cells.

3.2.5 LncRNA PiHL is clinically relevant in CRC

We next validated the upregulation of PiHL in 87 CRC tissues and paired adjacent normal tissues by qRT-PCR (Cohort 1, Figure 13a). *In situ* hybridization analyses of 100 independent paraffin-embedded CRC specimens confirmed the overexpression of PiHL in CRC tissues (Cohort 2, Figure 13b). Next, we analyzed the association between PiHL and clinicopathologic status in CRC patients from Cohort 1. A significant correlation was found between high levels of PiHL and poor

tumor differentiation (p = 0.034), and large tumor size (p = 0.011) (Table 4). A similar correlation between high levels of PiHL and large tumor size was observed in Cohort 2 (p = 0.027) (Table 5). Furthermore, Kaplan-Meier survival analysis of Cohort 2 revealed that higher PiHL expression was significantly associated with poorer overall survival (OS) (p = 0.002; Figure 13c). Multivariate Cox regression analysis suggested that PiHL expression was independently correlated with CRC OS. Taken together, these data suggest that increased PiHL levels are correlated with a poor prognosis in CRC patients.

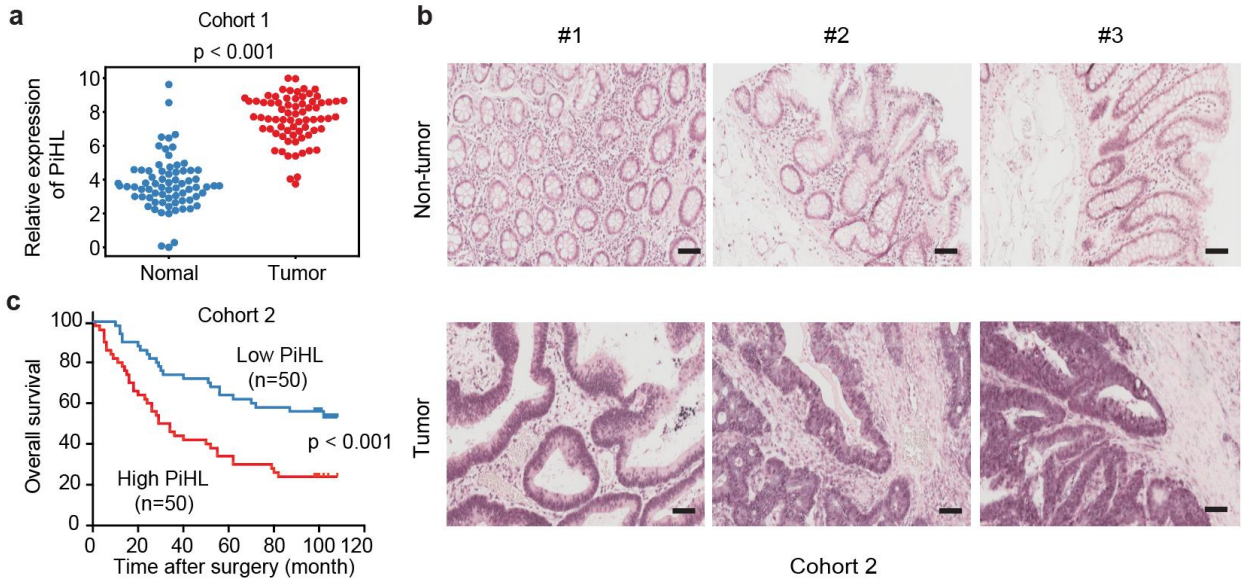


Figure 13 LncRNA PiHL is clinically relevant in CRC

(a) PiHL levels were quantified in 87 pairs of CRC tissues and adjacent normal tissues in cohort 1 using qRT-PCR, two-tailed Student's t-test.

(b) Representative images of PiHL expression in CRC and adjacent tissues using ISH analysis in cohort 2. Scale bar, 100 μm.

(3) Kaplan–Meier analyses of the correlation between PiHL RNA levels and overall survival in cohort 2.

3.3 Discussion

In current study, we developed a strategy using TCGA multi-omics data to identify copy number abnormalities that correlate with p53 mRNA or protein expression. This strategy successfully recapitulated the well-established copy number regulation of p53 expression including TP53 gene deletion and MDM2 gene amplification. We also identified chromosome 8q24.21 as a region strongly correlated with p53 protein levels. Further differential expression analysis showed genes in this region are significantly upregulated in tumor versus normal tissues. Integrated with correlation analysis between the gene expression and its own copy number alterations and further functional screening, we narrowed down the candidates to a lncRNA, PiHL, whose copy number amplification negatively correlates p53 protein stability. Using two independent clinical cohorts, we have shown that PiHL expression is significantly associated with tumor size, tumor differentiation, and CRC prognosis. These studies demonstrated this novel PiHL can be a p53 protein regulating lncRNA. We also need to note that the half-life p53 protein is relatively short (around 20 min), which weaken the reliability of the data. But we utilized several steps and 3 independent cohorts (TCGA database and two clinical cohorts) to depict genomic characteristics and clinical relevance of PiHL. *In vitro* studies validated PiHL's inhibition on p53 protein but not the TP53 mRNA level. Later, cellular localization, *in vivo* study and functional mechanisms need to be further demonstrated and validated.

Appendix

Table 1 Details of the genes in each DDR pathway

Base Excision Repair (BER)	TDG, UNG, PARP1, TDP1, APEX1, APEX2, POLE3, POLB, FEN1, POLE
Direct Repair (DR)	ALKBH2, ALKBH3, MGMT
Fanconi Anemia (FA)	FANCA, FANCB, FANCC, FANCL, FANCM, UBE2T, FANCD2, FANCI
Homology-Dependent Recombination (HDR)	MRE11A, RAD50, NBN, BRCA1, BARD1, BRIP1, RBBP8, TP53BP1, EXO1, PALB2, BRCA2, RAD52, SHFM1, RAD51, XRCC2, XRCC3, BLM, EME1, TOP3A, MUS81, GEN1, SLX1A
Mismatch Repair (MMR)	MLH1, MLH3, PMS1, PMS2, MSH2, MSH3, MSH6, EXO1, LIG4
Non-homologous End Joining (NHEJ)	XRCC5, XRCC6, PRKDC, POLL, POLM, XRCC4, LIG4, NEHJ1
Nucleotide Excision Repair (NER)	CUL5, XPC, XPA, ERCC1, ERCC2, ERCC4, ERCC5, ERCC6, POLE, POLE3
Translesion Synthesis (TLS)	REV1, POLQ, REV3L, SHPRH, POLN
Damage Sensors and Others	ATM, ATR, ATRIP, TOPBP1, CHEK1, CHEK2, MDC1, RNMT, TREX1

Table 2 Recurrently amplified or deleted DDR genes in TCGA PanCanAtlas cohort (n = 10,489)

Gene Symbol	Recurrently_Amplified_ in_Cancer_Types	p-value_mRNA _Amp_vs_Others[a]
NBN	UCS, BRCA, PRAD, LIHC, BLCA	2.50×10^{-60}
SHFM1	ESCA, STAD, UCS	3.20×10^{-43}
UBE2T	BRCA, LIHC, CHOL	2.17×10^{-20}
EXO1	BRCA, LIHC, OV, CHOL	6.80×10^{-22}
PARP1	BRCA, CHOL, LIHC, UCS	1.86×10^{-44}
PRKDC	UCS, LIHC, BRCA	1.20×10^{-34}
ATR	LUSC, ESCA, CESC OV, HNSC	1.89×10^{-56}
TOPBP1	LUSC, CESC	5.67×10^{-41}
POLB	UCS, BRCA, ESCA, LUSC, BLCA	2.16×10^{-111}
RAD52	UCS, OV	1.84×10^{-59}
POLQ	LUSC, CESC	7.33×10^{-32}
EME1	UCS, BRCA	2.78×10^{-28}
RBBP8	PAAD, ESCA, STAD	1.08×10^{-41}
	Recurrently_Deleted_ in_Cancer_Types	**p-value_mRNA_ Del_vs_Others[b]**
CHEK1	TGCT, UVM	0.027
SHPRH	UVM, DLBC	5.99×10^{-10}
REV3L	DLBC, PRAD, UVM	9.18×10^{-17}

a. p-value for mRNA expression comparison between samples with specific DDR gene amplification (GISTIC score = 2) and the other copy numbers (Wilcoxon rank-sum test)

b. p-value for mRNA expression comparison between samples with specific DDR gene deletion (GISTIC score = -2) and the other copy numbers (Wilcoxon rank-sum test)

Table 3 DDR CNAmp associates with reduced mutation load

Gene	Cancer Type	tStat, Silent Mutation Load	p-value, Silent Mutation Load	tStat, nonSilent Mutation Load	p-value, nonSilent Mutation Load
UNG	BLCA	-4.9558	5.77E-03	-3.3514	3.14E-02
UNG	OV	-4.9874	1.95E-03	-2.8151	5.49E-03
UNG	PRAD	-0.0738	9.42E-01	-0.2947	7.69E-01
UNG	SARC	-2.1070	3.88E-02	-2.2826	2.39E-02
CUL5	LUAD	-0.5312	6.47E-01	-0.1613	8.86E-01
ATR	BLCA	-6.1410	1.25E-04	-8.0198	6.76E-08
ATR	LIHC	-2.7937	5.44E-02	-2.9307	5.37E-02
ERCC5	LGG	-1.7294	8.44E-02	-1.3016	1.94E-01
TOP3A	LUSC	-3.2572	6.70E-02	-4.2672	3.16E-02
TOP3A	UCEC	-7.6717	8.49E-14	-7.7300	5.84E-14
POLE3	UCEC	-7.8759	2.02E-14	-8.0100	7.67E-15
TDP1	OV	-3.5904	1.43E-02	-2.0035	4.79E-02
TDP1	UCEC	-7.1376	2.91E-11	-7.1002	2.42E-10
ERCC2	UCS	-1.0449	3.01E-01	-1.1904	2.39E-01
ATM	STAD	-6.6264	1.00E-07	-7.7246	3.86E-13
RAD50	BRCA	-5.8071	6.09E-04	-6.8721	7.27E-09
TDG	BLCA	0.1242	9.12E-01	0.0245	9.83E-01
TDG	OV	-4.5170	6.52E-03	-2.9686	3.75E-03
SLX1A	BLCA	-3.7573	8.66E-04	-3.1658	4.52E-03
SLX1A	BRCA	-1.3084	1.95E-01	-1.7948	7.48E-02
SLX1A	COAD	-4.7441	4.49E-03	-5.0257	9.19E-04
SLX1A	OV	-3.8770	9.59E-03	0.9778	3.84E-01
SLX1A	UCEC	-7.8527	2.40E-14	-7.9826	9.45E-15
POLM	BLCA	-0.9523	3.90E-01	-0.4810	6.54E-01
POLM	GBM	-2.0707	3.91E-02	-2.1825	2.97E-02
POLM	UCEC	-7.4886	4.53E-13	-7.6209	2.03E-13
TOPBP1	BLCA	-3.1406	6.69E-02	-5.6251	5.19E-03
TOPBP1	ESCA	-3.6422	7.05E-04	-3.4734	8.71E-04
TOPBP1	OV	-0.2133	8.35E-01	-1.0887	2.77E-01
TOPBP1	PRAD	-1.4623	1.47E-01	-0.8867	3.76E-01
TOPBP1	UCEC	-6.3461	6.91E-07	-6.4970	4.24E-07
XRCC6	PAAD	-1.0510	2.95E-01	-0.9391	3.49E-01
XPC	UCEC	-7.8825	1.90E-14	-7.9472	1.20E-14
REV1	UCEC	-7.7635	5.09E-14	-7.6805	1.29E-13
ALKBH2	BLCA	-4.9558	5.77E-03	-3.3514	3.14E-02
ALKBH2	OV	-4.9874	1.95E-03	-2.8151	5.49E-03
XRCC3	HNSC	-3.0594	4.36E-02	-2.5476	9.38E-02

XRCC3	UCEC	-7.8190	3.04E-14	-7.9608	1.10E-14
RAD51	UCEC	0.0729	9.46E-01	-0.0198	9.85E-01
ATRIP	UCEC	-7.8394	2.60E-14	-8.0457	5.88E-15
MRE11A	BLCA	-1.1215	3.72E-01	-0.9666	4.31E-01
MRE11A	ESCA	-2.2272	9.52E-02	-3.3046	1.34E-02
FEN1	BLCA	-2.2704	9.55E-02	-2.9148	4.72E-02
FEN1	HNSC	-1.5262	2.53E-01	-2.6614	9.75E-02
FEN1	LIHC	-4.4083	1.84E-03	-4.8227	3.34E-04
FEN1	STAD	-5.2682	9.87E-04	-5.3791	7.99E-04
FEN1	UCEC	-3.1666	1.50E-02	-4.1735	1.80E-03
RNMT	LIHC	-6.0002	7.62E-04	-8.1415	2.04E-14
RNMT	LUAD	-7.9232	1.13E-03	-7.5365	2.03E-03
RNMT	LUSC	0.0065	9.95E-01	0.1866	8.64E-01
RNMT	OV	-9.3758	7.46E-06	0.9895	4.27E-01
RNMT	PAAD	-0.8663	3.88E-01	-0.9073	3.65E-01
RNMT	STAD	-4.7487	3.26E-04	-5.4287	2.04E-05
RNMT	UCEC	-7.1066	1.92E-11	-7.2238	1.18E-11
APEX1	ESCA	-1.2014	3.15E-01	-2.5835	2.07E-02
APEX1	UCEC	-2.7928	2.65E-02	-3.5687	6.20E-03
SHFM1	BLCA	-0.2750	8.01E-01	-0.4328	6.93E-01
SHFM1	STAD	-5.0701	1.32E-06	-5.0346	1.72E-06
SHFM1	UCEC	-5.3580	2.86E-04	-5.0727	1.06E-03
SHFM1	UCS	-0.7573	4.58E-01	-1.0175	3.14E-01
POLN	UCEC	-7.7080	6.60E-14	-7.8397	2.61E-14
ERCC6	HNSC	-2.7225	3.43E-02	-3.0467	1.95E-02
POLQ	ESCA	-3.7659	1.31E-03	-3.3425	2.26E-03
POLQ	OV	0.0275	9.79E-01	-1.4438	1.50E-01
POLQ	SARC	-1.9223	6.54E-02	-2.0457	4.49E-02
POLQ	UCEC	-7.7913	3.66E-14	-7.9374	1.29E-14
APEX2	OV	-2.1926	6.02E-02	-1.9277	5.63E-02
APEX2	UCEC	-7.6424	1.35E-13	-7.8814	2.21E-14
ERCC1	UCEC	-7.3107	1.09E-12	-7.5768	1.83E-13
ERCC1	UCS	-1.0449	3.01E-01	-1.1904	2.39E-01
MDC1	UCEC	-3.5130	5.19E-03	-4.3141	6.74E-04
MDC1	UVM	-0.5128	6.25E-01	-0.4072	6.86E-01
EME1	BLCA	-0.7938	4.81E-01	-0.6706	5.47E-01
EME1	OV	-0.7277	4.97E-01	-1.7552	8.22E-02
EME1	PAAD	-1.0935	2.76E-01	-1.0264	3.06E-01
EME1	UCEC	-7.8379	2.62E-14	-7.8065	3.48E-14
PMS2	BRCA	-6.2250	9.93E-08	-5.0982	1.18E-06
PMS2	ESCA	-3.0741	7.12E-03	-3.0077	7.25E-03

PMS2	LUSC	0.0614	9.56E-01	0.4357	7.04E-01
PMS2	UCEC	-7.6806	8.52E-14	-7.7216	7.44E-14
GEN1	GBM	-2.3325	2.02E-02	-2.3121	2.13E-02
GEN1	OV	-0.4592	6.61E-01	-0.8666	3.96E-01
GEN1	UCEC	-7.7514	4.88E-14	-7.8428	2.57E-14
MUS81	STAD	-7.0509	8.73E-10	-7.0941	7.98E-11
MUS81	UCEC	-5.3890	1.72E-06	-6.2643	9.75E-09
MUS81	UCS	-1.2818	2.05E-01	-1.1160	2.69E-01
RBBP8	BRCA	-1.3089	2.96E-01	-1.3740	2.61E-01
LIG4	ESCA	-1.9155	1.00E-01	-3.1222	2.52E-03
LIG4	LGG	-1.3486	1.78E-01	-0.9621	3.37E-01
LIG4	READ	-2.2095	2.90E-02	-1.9051	5.91E-02
FANCD2	ESCA	-0.9709	4.18E-01	-0.6229	5.86E-01
FANCD2	LGG	-1.0385	3.00E-01	-1.0657	2.87E-01
FANCD2	OV	-0.5049	6.47E-01	-0.7199	4.85E-01
PRKDC	HNSC	-1.6304	1.12E-01	-0.8740	3.89E-01
PRKDC	PRAD	-1.3306	1.84E-01	-1.1466	2.52E-01
PRKDC	READ	-1.9448	5.41E-02	-1.8334	6.92E-02
PRKDC	SARC	-2.3623	2.37E-02	-1.2138	2.50E-01
PRKDC	STAD	-6.5652	7.26E-09	-6.2989	3.02E-08
PRKDC	UCEC	-7.7388	5.33E-14	-7.9494	1.19E-14
PRKDC	UCS	-1.2465	2.18E-01	-1.2191	2.29E-01
POLB	BRCA	-2.7803	5.63E-03	-2.9146	3.66E-03
POLB	COAD	-2.3473	2.79E-02	-2.6475	1.34E-02
POLB	ESCA	-1.2123	2.35E-01	-1.3961	1.71E-01
POLB	LIHC	-0.8253	4.40E-01	-0.6744	5.24E-01
POLB	OV	0.4799	6.38E-01	-0.0663	9.47E-01
POLB	STAD	-6.5748	1.48E-09	-6.6506	1.04E-09
POLB	UCEC	-7.6108	1.33E-13	-7.7985	3.56E-14
POLB	UCS	-1.4091	1.65E-01	-1.2437	2.20E-01
CHEK2	HNSC	-1.6095	1.71E-01	-1.7718	1.41E-01
CHEK2	LUSC	-0.2101	8.40E-01	-0.8391	4.29E-01
BRIP1	MESO	-1.1112	3.13E-01	-0.2166	8.35E-01
BRIP1	PAAD	-0.9720	3.32E-01	-0.9309	3.53E-01
BRIP1	UCEC	-7.7821	3.89E-14	-7.9769	9.67E-15
BRCA2	LGG	-0.7288	4.79E-01	-0.5911	5.59E-01
BRCA2	OV	-2.7052	4.61E-02	-1.7908	8.33E-02
BRCA2	STAD	-6.6352	2.15E-10	-6.5769	3.76E-10
MLH3	HNSC	-0.3963	7.29E-01	-0.0193	9.86E-01
MLH3	LUAD	0.4557	6.79E-01	0.3540	7.47E-01
MLH3	OV	-2.4499	1.18E-01	-2.2314	3.34E-02

FANCL	OV	-0.1687	8.70E-01	-1.0429	3.01E-01
MSH2	BLCA	0.5757	5.85E-01	0.2919	7.80E-01
MSH2	UCEC	-7.6023	1.43E-13	-7.7702	4.27E-14
MSH2	UCS	-1.4155	1.63E-01	-1.3670	1.77E-01
MSH6	BLCA	0.5757	5.85E-01	0.2919	7.80E-01
MSH6	HNSC	-0.0326	9.77E-01	0.0026	9.98E-01
MSH6	UCEC	-7.1643	3.45E-11	-7.5001	1.74E-12
ERCC4	BRCA	-1.4207	1.60E-01	-1.8911	6.09E-02
ERCC4	LUAD	-0.4177	7.16E-01	-0.5308	6.47E-01
ERCC4	OV	0.3243	7.55E-01	-0.9093	3.67E-01
ERCC4	PRAD	-1.1061	2.69E-01	-1.5068	1.34E-01
ERCC4	SARC	-1.3672	2.06E-01	-1.2529	2.47E-01
BLM	LUAD	-1.4067	2.90E-01	-3.0749	7.80E-02
BLM	SARC	-1.9882	4.91E-02	-1.5065	1.38E-01
BLM	STAD	-5.7343	3.95E-08	-5.6638	5.72E-08
FANCA	BLCA	-0.5331	6.12E-01	-0.0711	9.45E-01
MGMT	OV	0.2464	8.10E-01	-0.6803	4.98E-01
MGMT	PCPG	-0.2766	8.08E-01	-0.9754	4.27E-01
TREX1	UCEC	-7.8394	2.60E-14	-8.0457	5.88E-15
UBE2T	LIHC	-0.7937	4.36E-01	-1.1881	2.47E-01
UBE2T	LUAD	-3.3430	3.02E-03	-3.4569	2.23E-03
UBE2T	OV	-2.2366	3.94E-02	-1.6877	9.26E-02
UBE2T	PCPG	-0.8218	4.94E-01	0.2159	8.48E-01
UBE2T	UCEC	-5.6119	1.73E-07	-6.2799	2.79E-09
NBN	ESCA	-1.4564	1.75E-01	-0.7641	4.67E-01
NBN	HNSC	-1.8479	8.14E-02	-1.1983	2.50E-01
NBN	LGG	-0.4442	6.65E-01	-0.4633	6.45E-01
NBN	OV	-1.2175	2.42E-01	-1.3882	1.66E-01
NBN	STAD	-1.3117	2.10E-01	-1.4921	1.57E-01
NBN	UCEC	-7.7982	3.51E-14	-7.9739	1.01E-14
NBN	UCS	-1.3339	1.88E-01	-1.2052	2.34E-01
BRCA1	PCPG	-0.1267	9.11E-01	1.3333	2.96E-01
NHEJ1	OV	-2.5498	1.16E-01	-1.9788	7.87E-02
NHEJ1	UCEC	-6.0460	8.67E-08	-6.5918	1.58E-09
PALB2	BLCA	0.7588	4.67E-01	0.6952	5.04E-01
PALB2	BRCA	-1.3190	1.92E-01	-1.8707	6.42E-02
FANCI	LUAD	-1.9938	1.33E-01	-4.2078	1.48E-02
FANCI	STAD	-6.6687	1.18E-10	-6.5631	2.46E-10
FANCI	UCEC	0.5354	6.29E-01	0.4248	6.99E-01
POLE	BLCA	-2.9425	4.37E-02	-3.0802	3.66E-02
POLE	LGG	-1.3217	1.87E-01	-1.0443	2.97E-01

POLE	OV	-3.6987	1.06E-02	-2.4170	1.89E-02
POLE	UCEC	-7.4980	3.28E-13	-7.7475	5.28E-14
PARP1	LIHC	-0.7798	4.45E-01	-1.1321	2.71E-01
PARP1	LUAD	-1.8526	7.98E-02	-1.7729	9.22E-02
PARP1	OV	-0.1074	9.15E-01	-0.9635	3.36E-01
PARP1	PCPG	-0.8062	4.75E-01	0.9708	4.00E-01
PARP1	SARC	-1.2352	2.56E-01	-1.5063	1.72E-01
PARP1	THYM	-1.1021	2.97E-01	-1.0700	3.07E-01
PARP1	UCEC	-4.6412	5.55E-05	-5.4157	1.62E-06
PARP1	UCS	-1.0370	3.05E-01	-1.1915	2.39E-01
ALKBH3	BRCA	-3.0981	5.50E-03	-3.2443	3.16E-03
ALKBH3	HNSC	-1.5185	1.60E-01	-2.0940	6.38E-02
ALKBH3	LUAD	-0.2997	7.84E-01	-0.2560	8.14E-01
ALKBH3	OV	-1.1019	3.24E-01	-1.0079	3.21E-01
ALKBH3	SARC	-2.6246	1.24E-02	-1.7635	1.11E-01
ALKBH3	STAD	-3.9915	2.40E-03	-4.5355	6.02E-04
ALKBH3	UCEC	-7.6645	9.35E-14	-7.8446	2.49E-14

In this table, tStat and p-value indicate the t-test statistics and p-values of of the silent mutation load or non-silent mutation load between samples with and without DDR gene amplification.

Table 4 Details of colorectal cancer cohort 1

Clinicopathological Feature	No. of Patients	CLIP expression High (n=42)	CLIP expression Low (n=41)	p-value
Gender				0.76
Male	54	28	26	
Female	29	14	15	
Age				0.234
<65	37	16	21	
≥65	46	26	20	
location				0.223
colon	36	21	15	
rectum	47	21	26	
Differentiation				0.034*
Well and Moderately	31	11	20	
Poorly and others	52	31	21	
Tumor size (cm^3)				0.011*
<20	35	12	23	
≥20	48	30	18	
Lymph node metastasis				0.328
absent	42	19	23	
present	41	23	18	
venous metastasis				0.7
absent	74	37	37	
present	9	5	4	
TNM stage				0.158
I-II	22	14	8	
III-IV	61	28	33	

Table 5 Details of colorectal cancer cohort 2

Clinicopathological Feature	No. of Patients	CLIP expression High (n=50)	CLIP expression Low (n=50)	p-value
Gender				1
Male	54	27	27	
Female	46	23	23	
Age				0.466
<65	21	9	12	
≥65	79	41	38	
Tumor size (cm^3)				0.027*
<20	45	17	28	
≥20	55	33	22	
Pathological stage				0.693
I-II	50	26	24	
III-IV	50	24	26	
Lymph node metastasis				0.31
absent	61	28	33	
present	39	22	17	
TNM stage				0.419
I-II	60	28	32	
III-IV	40	22	18	

Bibliography

1. Devasagayam, T.P., et al., *Free radicals and antioxidants in human health: current status and future prospects.* J Assoc Physicians India, 2004. **52**: p. 794-804.
2. Feitelson, M.A., et al., *Sustained proliferation in cancer: Mechanisms and novel therapeutic targets.* Semin Cancer Biol, 2015. **35 Suppl**: p. S25-S54.
3. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-24.
4. Pfeifer, G.P., *Environmental exposures and mutational patterns of cancer genomes.* Genome Med, 2010. **2**(8): p. 54.
5. Pena-Diaz, J., et al., *Noncanonical mismatch repair as a source of genomic instability in human cells.* Mol Cell, 2012. **47**(5): p. 669-80.
6. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
7. Olivier, M., M. Hollstein, and P. Hainaut, *TP53 mutations in human cancers: origins, consequences, and clinical use.* Cold Spring Harb Perspect Biol, 2010. **2**(1): p. a001008.
8. McCarroll, S.A. and D.M. Altshuler, *Copy-number variation and association studies of human disease.* Nat Genet, 2007. **39**(7 Suppl): p. S37-42.
9. Hastings, P.J., et al., *Mechanisms of change in gene copy number.* Nat Rev Genet, 2009. **10**(8): p. 551-64.
10. Hu, X., et al., *A functional genomic approach identifies FAL1 as an oncogenic long noncoding RNA that associates with BMI1 and represses p21 expression in cancer.* Cancer Cell, 2014. **26**(3): p. 344-357.
11. St Laurent, G., C. Wahlestedt, and P. Kapranov, *The Landscape of long noncoding RNA classification.* Trends Genet, 2015. **31**(5): p. 239-51.
12. Kornienko, A.E., et al., *Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans.* Genome Biol, 2016. **17**: p. 14.
13. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes Dev, 2011. **25**(18): p. 1915-27.
14. Guil, S. and M. Esteller, *Cis-acting noncoding RNAs: friends and foes.* Nat Struct Mol Biol, 2012. **19**(11): p. 1068-75.
15. Beckedorff, F.C., et al., *The intronic long noncoding RNA ANRASSF1 recruits PRC2 to the RASSF1A promoter, reducing the expression of RASSF1A and increasing cell proliferation.* PLoS Genet, 2013. **9**(8): p. e1003705.
16. Brown, J.A., et al., *Formation of triple-helical structures by the 3'-end sequences of MALAT1 and MENbeta noncoding RNAs.* Proc Natl Acad Sci U S A, 2012. **109**(47): p. 19202-7.
17. Zhang, X., et al., *Maternally expressed gene 3 (MEG3) noncoding ribonucleic acid: isoform structure, expression, and functions.* Endocrinology, 2010. **151**(3): p. 939-47.
18. Davidovich, C., et al., *Promiscuous RNA binding by Polycomb repressive complex 2.* Nat Struct Mol Biol, 2013. **20**(11): p. 1250-7.
19. Li, L., et al., *Targeted disruption of Hotair leads to homeotic transformation and gene derepression.* Cell Rep, 2013. **5**(1): p. 3-12.

20. Tsai, M.C., et al., *Long noncoding RNA as modular scaffold of histone modification complexes.* Science, 2010. **329**(5992): p. 689-93.
21. Elsayed, E.T., et al., *Plasma long non-coding RNA HOTAIR as a potential biomarker for gastric cancer.* Int J Biol Markers, 2018: p. 1724600818760244.
22. Xiao, Z., et al., *LncRNA HOTAIR is a Prognostic Biomarker for the Proliferation and Chemoresistance of Colorectal Cancer via MiR-203a-3p-Mediated Wnt/ss-Catenin Signaling Pathway.* Cell Physiol Biochem, 2018. **46**(3): p. 1275-1285.
23. Schmitt, A.M., et al., *An inducible long noncoding RNA amplifies DNA damage signaling.* Nat Genet, 2016. **48**(11): p. 1370-1376.
24. Ng, S.Y., et al., *The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis.* Mol Cell, 2013. **51**(3): p. 349-59.
25. Tripathi, V., et al., *The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation.* Mol Cell, 2010. **39**(6): p. 925-38.
26. Tseng, Y.Y., et al., *PVT1 dependence in cancer with MYC copy-number increase.* Nature, 2014. **512**(7512): p. 82-6.
27. Xu, W., et al., *Long non-coding RNA PCAT-1 contributes to tumorigenesis by regulating FSCN1 via miR-145-5p in prostate cancer.* Biomed Pharmacother, 2017. **95**: p. 1112-1118.
28. Kim, J., et al., *Long noncoding RNA MALAT1 suppresses breast cancer metastasis.* Nat Genet, 2018. **50**(12): p. 1705-1715.
29. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
30. Yang, W., et al., *Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.* Nucleic Acids Res, 2013. **41**(Database issue): p. D955-61.
31. Tate, J.G., et al., *COSMIC: the Catalogue Of Somatic Mutations In Cancer.* Nucleic Acids Res, 2019. **47**(D1): p. D941-D947.
32. Jeggo, P.A., L.H. Pearl, and A.M. Carr, *DNA repair, genome stability and cancer: a historical perspective.* Nat Rev Cancer, 2016. **16**(1): p. 35-42.
33. Bouwman, P. and J. Jonkers, *The effects of deregulated DNA damage signalling on cancer chemotherapy response and resistance.* Nat Rev Cancer, 2012. **12**(9): p. 587-98.
34. Setlow, R.B., et al., *Evidence that xeroderma pigmentosum cells do not perform the first step in the repair of ultraviolet damage to their DNA.* Proc Natl Acad Sci U S A, 1969. **64**(3): p. 1035-41.
35. Miki, Y., et al., *A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.* Science, 1994. **266**(5182): p. 66-71.
36. Audeh, M.W., et al., *Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial.* Lancet, 2010. **376**(9737): p. 245-51.
37. Bryant, H.E., et al., *Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase.* Nature, 2005. **434**(7035): p. 913-7.
38. Lord, C.J. and A. Ashworth, *PARP inhibitors: Synthetic lethality in the clinic.* Science, 2017. **355**(6330): p. 1152-1158.
39. Lord, C.J. and A. Ashworth, *BRCAness revisited.* Nat Rev Cancer, 2016. **16**(2): p. 110-20.

40. Kim, G., et al., *FDA Approval Summary: Olaparib Monotherapy in Patients with Deleterious Germline BRCA-Mutated Advanced Ovarian Cancer Treated with Three or More Lines of Chemotherapy.* Clin Cancer Res, 2015. **21**(19): p. 4257-61.
41. Balasubramaniam, S., et al., *FDA Approval Summary: Rucaparib for the Treatment of Patients with Deleterious BRCA Mutation-Associated Advanced Ovarian Cancer.* Clin Cancer Res, 2017. **23**(23): p. 7165-7170.
42. Scott, L.J., *Niraparib: First Global Approval.* Drugs, 2017. **77**(9): p. 1029-1034.
43. Mittica, G., et al., *PARP Inhibitors in Ovarian Cancer.* Recent Pat Anticancer Drug Discov, 2018. **13**(4): p. 392-410.
44. Knijnenburg, T.A., et al., *Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas.* Cell Rep, 2018. **23**(1): p. 239-254 e6.
45. Olshen, A.B., et al., *Circular binary segmentation for the analysis of array-based DNA copy number data.* Biostatistics, 2004. **5**(4): p. 557-72.
46. Mermel, C.H., et al., *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol, 2011. **12**(4): p. R41.
47. Iorio, F., et al., *A Landscape of Pharmacogenomic Interactions in Cancer.* Cell, 2016. **166**(3): p. 740-754.
48. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
49. Liu, J., et al., *An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.* Cell, 2018. **173**(2): p. 400-416 e11.
50. Wang, Y., et al., *Systematic identification of non-coding pharmacogenomic landscape in cancer.* Nat Commun, 2018. **9**(1): p. 3192.
51. Palles, C., et al., *Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas.* Nat Genet, 2013. **45**(2): p. 136-44.
52. Shinbrot, E., et al., *Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication.* Genome Res, 2014. **24**(11): p. 1740-50.
53. Besaratinia, A., et al., *A multi-biomarker approach to study the effects of smoking on oxidative DNA damage and repair and antioxidative defense mechanisms.* Carcinogenesis, 2001. **22**(3): p. 395-401.
54. Chen, C.C., T. Taniguchi, and A. D'Andrea, *The Fanconi anemia (FA) pathway confers glioma resistance to DNA alkylating agents.* J Mol Med (Berl), 2007. **85**(5): p. 497-509.
55. Prolla, T.A., et al., *Tumour susceptibility and spontaneous mutation in mice deficient in Mlh1, Pms1 and Pms2 DNA mismatch repair.* Nat Genet, 1998. **18**(3): p. 276-9.
56. Martin, M.J., R. Juarez, and L. Blanco, *DNA-binding determinants promoting NHEJ by human Polmu.* Nucleic Acids Res, 2012. **40**(22): p. 11389-403.
57. Edwards, S.L., et al., *Resistance to therapy caused by intragenic deletion in BRCA2.* Nature, 2008. **451**(7182): p. 1111-5.
58. Kinch, M.S., *An analysis of FDA-approved drugs for oncology.* Drug Discov Today, 2014. **19**(12): p. 1831-5.
59. Yan, Z., et al., *A histone-fold complex and FANCM form a conserved DNA-remodeling complex to maintain genome stability.* Mol Cell, 2010. **37**(6): p. 865-78.

60. Stoepker, C., et al., *DNA helicases FANCM and DDX11 are determinants of PARP inhibitor sensitivity.* DNA Repair (Amst), 2015. **26**: p. 54-64.
61. Tang, Y.C. and A. Amon, *Gene copy-number alterations: a cost-benefit analysis.* Cell, 2013. **152**(3): p. 394-405.
62. Cancer Genome Atlas, N., *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.
63. Vousden, K.H. and C. Prives, *Blinded by the Light: The Growing Complexity of p53.* Cell, 2009. **137**(3): p. 413-31.
64. Bond, G.L., et al., *A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans.* Cell, 2004. **119**(5): p. 591-602.
65. Levine, A.J. and M. Oren, *The first 30 years of p53: growing ever more complex.* Nat Rev Cancer, 2009. **9**(10): p. 749-58.
66. Yang, D.Q., M.J. Halaby, and Y. Zhang, *The identification of an internal ribosomal entry site in the 5'-untranslated region of p53 mRNA provides a novel mechanism for the regulation of its translation following DNA damage.* Oncogene, 2006. **25**(33): p. 4613-9.
67. Zhang, J., et al., *Translational repression of p53 by RNPC1, a p53 target overexpressed in lymphomas.* Genes Dev, 2011. **25**(14): p. 1528-43.
68. Yin, Y., et al., *p53 Stability and activity is regulated by Mdm2-mediated induction of alternative p53 translation products.* Nat Cell Biol, 2002. **4**(6): p. 462-7.
69. Liu, Y., et al., *TP53 loss creates therapeutic vulnerability in colorectal cancer.* Nature, 2015. **520**(7549): p. 697-701.
70. Wade, M., Y.C. Li, and G.M. Wahl, *MDM2, MDMX and p53 in oncogenesis and cancer therapy.* Nat Rev Cancer, 2013. **13**(2): p. 83-96.
71. Bester, A.C., et al., *An Integrated Genome-wide CRISPRa Approach to Functionalize lncRNAs in Drug Resistance.* Cell, 2018. **173**(3): p. 649-664 e20.