5

Shared Regulatory Motifs in Promoters of Human DNA Repair Genes

Lonnie R. Welch[1], Laura M. Koehly[2] and Laura Elnitski[3]
[1]*School of Electrical Engineering and Computer Science, Ohio University, Athens, Ohio*
[2]*Social and Behavioral Research Branch,*
National Human Genome Research Institute, NIH, Bethesda, Maryland
[3]*Genome Technology Branch*
National Human Genome Research Institute, NIH, Bethesda, Maryland
USA

1. Introduction

This manuscript presents methods used to test, and resulting evidence to support the hypothesis that specialized transcription factor binding sites coordinate the expression of DNA repair genes. Building on the seminal work of the Elnitski laboratory (Yang et al. 2007), which identified the most complete set of human transcripts under the control of bidirectional promoters and identified the first putative regulatory networks that make use of the bidirectional promoter structure, the authors present additional details of these regulatory networks.

Much of the work regarding the regulation of DNA repair proteins is aimed at the level of protein-protein interactions and post-translational processing events (Hurley et al. 2007, Jensen et al. 2011, Shibata et al. 2010). However, transcriptional activation of DNA repair genes is likely to utilize shared factors, especially in cases of induced activation, which have not been thoroughly evaluated. Yang, Koehly and Elnitski reported the discovery and characterization of 5,653 bidirectional promoters in the human genome (Yang et al. 2007). Prior to that date, bidirectional promoters were annotated only for protein-coding genes, and only 1,352 examples had been reported in the human genome. The work of Yang et al. included evidence from all noncoding-RNA genes, as well. Each bidirectional promoter regulates the expression of two genes, oriented in opposite directions with transcription start sites within 1000 bp of one another. The authors developed a novel approach to map all bidirectional promoters by analyzing the public expressed-sequence-tag (EST) data. The prevalence of this promoter structure led the authors to explore the hypothesis that it plays a role in regulation of certain classes of genes. They discovered that many more DNA repair genes have bidirectional promoters than previously reported and that many genes with somatic mutations in cancer have bidirectional promoters. The relevance of DNA repair genes to cancers (Kinsella et al. 2009, Liang et al. 2009, Smith et al. 2010, Kelley et al. 2008, Li et al. 2009, Bellizii et al. 2009, Naccarati et al. 2007, Berwick et al. 2000)) and the association of bidirectional promoters with DNA repair genes suggested that bidirectional promoters might indicate a higher-order type of regulatory structure that could be detected through common features at the DNA sequence level. If true, these features should discriminate bidirectional promoters and unidirectional promoters of genes with DNA repair functions.

Thus, this chapter presents additional evidence of these regulatory networks. Specifically, this chapter provides evidence that there are distinct regulatory signatures for (1) genes involved in certain types of cancers, (2) bidirectional versus unidirectional promoters and (3) specific DNA repair pathways. The authors have identified transcription factor binding sites in bidirectional promoters of genes implicated in breast and ovarian (B/O) cancers. Additionally, they have discovered novel transcription factor binding sites that may serve as regulatory elements to distinguish DNA repair genes with bidirectional promoters from DNA repair genes with unidirectional promoters. Applications of this work extend to a collection of novel transcription factor binding sites shared among genes acting as checkpoint factors of DNA repair pathways. These findings have important implications – as evidence of novel regulatory mechanisms, and new insights into cancer biology (i.e., genomic elements relevant to transcriptional regulation) are gained.

2. Regulatory features of genes implicated in breast and ovarian cancers

This section provides evidence to support the hypothesis that there are distinct regulatory control systems among bidirectional and unidirectional promoters. Additionally, this section presents transcription factor binding sites discovered in bidirectional promoters of genes implicated in breast and ovarian cancers.

As reported in Yang et al. 2007, we identified transcription factor binding sites for known factors in genes implicated in B/O cancers. The enrichment of bidirectional promoters in several cancer genes, and in additional genes having functions in DNA repair, suggests common mechanisms of regulation. We used expression clustering and enrichment of genes with bidirectional promoters to group the cancer genes into expression groups from the full genome to address features common among the clusters that might indicate the presence of regulatory networks. The cancer-related genes that were identified and studied are listed below, along with their descriptions from GeneCards (Safran et al. 2010). The Elnitski group was the first to report that this set of genes has bidirectional promoters.

All genes were assessed for the top most related gene expression profiles in the genome using the gene sorter tool at the UCSC Genome Browser and expression data from the Novartis GNF Atlas2 (containing expression profiles for 96 tissues). Each cluster was then compared to all the others to identify intersection points (by gene names) among the lists of co-expressed genes. Using a process of multidimensional scaling, the gene lists were compared and a putative regulatory network was generated (Figure 1). The *MLH1* gene appeared in several co-expression clusters and therefore occupied a central location with connections to 7 other genes (*BARD1, FANCA, BRCA1, CHK2, BRCA2, TP53* and *FANCF*). Two additional genes co-occupied the central position with *MLH1*. *COMMD3* (an uncharacterized protein) and *ITGB3BP*, a regulator of apoptosis in breast cancer cells.

2.1 Network visualization

The bidirectional promoters that are associated with the breast and ovarian cancer genes were considered an affiliation network or a bipartite graph. In this example nodes represent the genes in the co-expression clusters and edges connect the genes appearing in more than one list. The higher the number of appearances of any gene from the ten co-expression lists, the more central its position in the network. Geodesic distances between genes were computed (e.g. length of the shortest path between genes through promoters, and the geodesic distance matrix was scaled using a metric multidimensional scaling (MDS)

Gene	Description from GeneCards (Safran 2010)
BARD1	This gene encodes a protein which interacts with the N-terminal region of *BRCA1*.
BRCA1	This gene encodes a nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor.
BRCA2	Inherited mutations in *BRCA1* and this gene, *BRCA2*, confer increased lifetime risk of developing breast or ovarian cancer.
CHK2	In response to DNA damage and replication blocks, cell cycle progression is halted through the control of critical cell cycle regulators. The protein encoded by this gene is a cell cycle checkpoint regulator and putative tumor suppressor.
ERBB2	This gene encodes a member of the epidermal growth factor (EGF) receptor family of receptor tyrosine kinases.
TP53	This gene encodes tumor protein *p53*, which responds to diverse cellular stresses to regulate target genes that induce cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism.
FANCA	DNA repair protein that may operate in a post-replication repair or a cell cycle checkpoint function. May be involved in inter-strand DNA cross-link repair and in the maintenance of normal chromosome stability.
FANCB	DNA repair protein required for *FANCD2* ubiquitination.
FANCD2	Required for maintenance of chromosomal stability. Promotes accurate and efficient pairing of homologs during meiosis. Involved in the repair of DNA double-strand breaks, both by homologous recombination and single-strand annealing. May participate in S phase and G2 phase checkpoint activation upon DNA damage. Promotes *BRCA2/FANCD1* loading onto damaged chromatin.
FANCF	DNA repair protein that may operate in a postreplication repair or a cell cycle checkpoint function. May be implicated in interstrand DNA cross-link repair and in the maintenance of normal chromosome stability.

Table 1. The B/O cancer-related genes that were studied.

algorithm (in UCINET 6; Borgatti et al., 2002). The distance between the 10 B/O cancer genes represents their similarity based on the number of shared genes found in the co-expression clusters. Genes in the center of the network were present in the largest number of gene clusters, seven out of 10, indicating that co-expression clusters intersect through common regulatory nodes.

2.2 Transcription factor binding site analysis

A systematic search of transcription factor binding sites in the list of bidirectional promoters was used to assess regulatory connections at the DNA level, and revealed several in common (using a motif finding algorithm we searched for the motifs reported in (Xie et al. 2005)). Notably, identical *ELK1* binding sites were located at the same distance from *ERBB2*, *FANCD2*, and *BRCA2* transcription start sites (Yang et al. 2007). *ETS* factor binding sites were present as a trio with SP1 and *PAX4/RXR* binding sites in the majority of the promoters. The transcription factors for which binding motifs were found in all of the promoters along with their descriptions from GeneCards (Safran et al. 2010) are reported in Table 2.

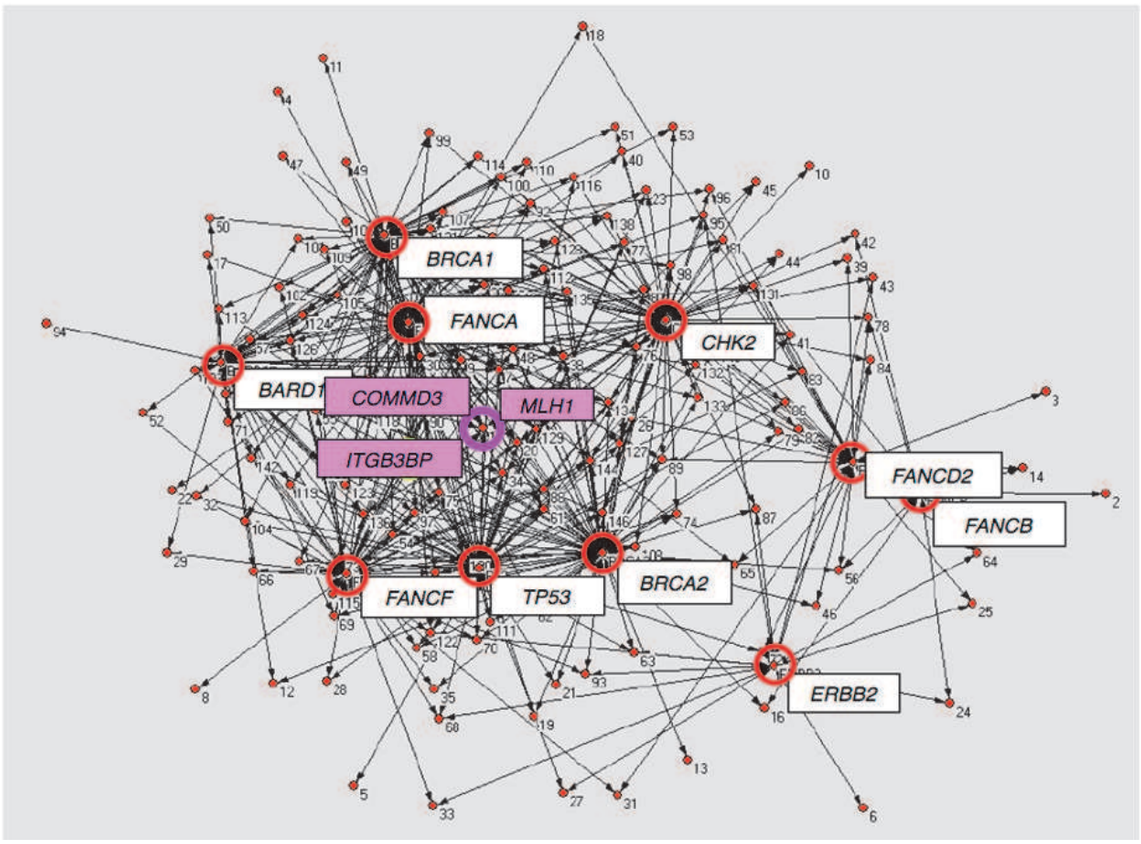


Fig. 1. Co-expression clustering analysis of 10 DNA repair genes finds intersecting nodes.

Transcription Factor	**Description from GeneCards**
Sp1	Transcription factor that can activate or repress transcription in response to physiological and pathological stimuli. Regulates the expression of a large number of genes involved in a variety of processes such as cell growth, apoptosis, differentiation and immune responses. May have a role in modulating the cellular response to DNA damage.
NFAT	The nuclear factor of activated T-cells family of transcription factors.
EGR-1	The protein encoded by this gene belongs to the EGR family of *C2H2*-type zinc-finger proteins. It is a nuclear protein and functions as a transcriptional regulator. Studies suggest this is a cancer suppressor gene.
PAX4	This gene is a member of the paired box (*PAX*) family of transcription factors. These genes play critical roles during fetal development and cancer growth.
ELK1	*ELK1* is a member of *ETS* oncogene family. The protein encoded by this gene is a nuclear target for the ras-raf-MAPK signaling cascade.

Table 2. Transcription factor binding sites in the promoters of the B/O cancer genes.

3. Unbiased assessment of transcription factor binding sites in two subgroups of genes from DNA repair pathways

The research reported in (Yang et al. 2007) provides strong evidence that a unique set of regulatory proteins control genes that contain bidirectional promoters by comparing co-expression clusters of genes enriched for bidirectional promoters versus those depleted for bidirectional promoters. This section reports on a study that identified transcription factor binding sites that are specific to genes in DNA repair pathways (Lichtenberg et al. 2009). The promoters of genes from the DNA repair pathways were partitioned into two groups, those that are bidirectional (32 promoters) and those that are unidirectional (42 promoters).

3.1 Assessment of individual sites

Each group of promoters was analyzed to discover putative transcription factor binding sites. The analysis was performed with WordSeeker motif discovery software (Lichtenberg et al. 2010), which employs high performance supercomputer-based algorithms to perform motif enumeration and to construct Markov models. Our analysis revealed that the average nucleotide G+C content of the bidirectional promoters was slightly higher than the unidirectional promoters, 59.87% versus 50.84%, respectively. These differences were rigorously controlled by the use of the Markov model, which examines background frequencies of each nucleotide in the collection of sequences. Unique sets of binding sites were identified for each group, some of which represent novel binding sites.

A statistical analysis of the promoters of the DNA repair genes revealed a number of significant DNA binding site motifs. Some of the discovered motifs correspond to recognition sequences of known proteins. These are listed in Table 3, along with their *p*-values and the corresponding transcription factors known to bind to the motifs (as determined by the TRANSFAC database (Wingender et al. 2000) and the JASPAR database (Bryne et al. 2008)). In addition, novel motifs, representing uncharacterized transcription factor binding sites, were discovered in the bidirectional and unidirectional promoters from DNA repair pathway genes (see Table 4 for the motifs and their *p*-values).

Motif (bidirectional promoters)	***p*-Value**	**Transcription Factor**	**Motif (unidirectional promoters)**	***p*-Value**	**Transcription Factor**
AGGGCCGT	0.04142	*MYB*	ACCCGCCT	0.00656	*SP1*
CAGGGGCC	0.02841	*V$WT1_Q6*	AGGAAACA	0.03295	*NFAT*
CGTGGGGG	0.04701	*E2F*	ATTAAAAT	0.05372	*OCT1*
GGCCCGCC	0.06682	*SP1*	CGGAAACC	0.04210	*AREB6*
TCCCGGCT	0.05408	*ELK1*	GCAGGGCG	0.07134	*PF0096*
TCCCGGGA	0.06861	*STAT5A*	GGGGAGTA	0.03321	*FOXC1*
TCGCGCCA	0.01539	*PF0112*	GGGGCTGC	0.06212	*LRF*
TCTGAGGA	0.01350	*TFIIA*	TGGGCGGA	0.06334	*GC*

Table 3. Enriched motifs matching characterized transcription factor binding sites discovered in the bidirectional promoters (columns 1 and 2) and in the unidirectional promoters (columns 3 and 4).

Motif (bidirectional promoters)	P-Value	**Motif (unidirectional promoters)**	P-Value
ACTCCAGC	0.06212	AGCCGGCT	0.05007
AGAAAAGA	0.02756	ATTCCCAG	0.05599
AGGGAGGG	0.07159	CCTCTTTA	0.03381
CAGCAGCC	0.10540	CGCCCCTT	0.11386
CGACTCCG	0.02756	CGGCGGCG	0.04742
CGCGGCCG	0.03377	CTCCCGCT	0.05998
CGGGCCGA	0.06548	CTTCTTTC	0.03773
GCCCCTCC	0.07021	GCGCCGCG	0.09760
GCCGGCGA	0.03662	GGGCGCCC	0.08390
GGCAGGGA	0.10334	GTGCGTTT	0.06286
GGGCCAGG	0.09632	TCCGCCGG	0.05794
GGGGCCGG	0.05265	TCTCCCCT	0.07881
TCTGGGAT	0.01466	TCTTCTTC	0.04649
TGAAGCCA	0.05699	TGCGCCGA	0.04148
TGCCCGCG	0.08277	TTGGTCTC	0.08543
TGCGGAAT	0.02132	TTTCTCCA	0.06840
TGCTGAGA	0.03377	TTTTTTGA	0.04742

Table 4. Uncharacterized motifs discovered in the promoters of DNA repair genes. Words are ordered alphabetically.

3.2 Assessment of paired binding sites

To identify putative regulatory modules (co-acting regulatory elements), we identified statistically overrepresented pairs of DNA motifs in each set of promoters. Motif pairs are shown in Table 5. The motif pair scores are computed as the product of (1) the number sequences, S, in which the pair occurs and (2) the natural log of the ratio of S and the expected value of S, E_s; i.e., the score is $S \cdot \ln(S/E_s)$. The genomic signatures (significant DNA motifs and motif pairs) of the bidirectional promoters were virtually non-overlapping with the signatures of unidirectional promoters. This provides strong support for the hypothesis that the regulatory mechanisms of bidirectional promoters are unique. Additionally, this work contributes a significant enhancement to the available knowledge about transcriptional regulation of genes involved in DNA repair pathways, and implicates the presence of a regulatory network.

4. Unbiased assessment of transcription factor binding sites of checkpoint factor genes from DNA repair pathways

We have performed a focused, detailed characterization of the checkpoint factors in DNA repair pathways (Elnitski et al. 2010). The checkpoint factors (Kanehisa et al. 2008, Wood 2005, Helleday et al. 2008) are activated upon detection of DNA damage, resulting in halting the cell cycle so that subsequent DNA repair pathways can mend the damage. In addition to examining the most recognized promoter in each gene (the 5′ end of the full-length transcription unit), we assessed alternative start sites for each checkpoint factor gene as independent regulatory units, to discover putative transcription factor binding sites. In this

Co-Occurring Motif Pair (Bidirectional Promoters)		Score	Co-Occurring Motif Pair (Unidirectional Promoters)		Score
TCTGAGGA	TCGCGCCA	12.1158	GTTCATTC	TCCGCCGG	11.2184
ACTCCAGC	TCGCGCCA	11.8387	CTGTGTGC	TGCGCCGA	11.1966
GCCCAGCC	TCCGCCGC	11.1827	TGACGCGA	CTCCCGCT	10.9997
GCCCAGCC	CGGAGCGC	10.8711	AGCCGGCT	GGGGAGTA	10.0590
TGCCCGCG	TCCCGGGA	10.7404	ATTGCAGG	ATTCTCTC	9.5459
GGCAGGGA	GGGCCAGG	9.8609	GGGGAGTA	AGGAAACA	9.3177
TCCCGGGA	TCGCGCCA	9.8112	CTGGGAGC	GTTCATTC	9.0337
AGCCTGTC	TCCCGGGA	9.7646	CCTTCCGA	CTGGGAGC	8.8439
GGAGGCTG	TCGCGCCA	9.7250	TGGGCGGA	ACCCGCCT	8.7895
TCCGCCGC	GCCCCTCC	9.6830	TTTCTCCA	CGGAAACC	8.6446
AGAAAAGA	TCGCGCCA	9.4042	CCCCCGCG	ACCCGCCT	8.5339
GCCCAGCC	GCCCCTCC	9.2808	TCCGCCGG	GGGGCTGC	7.7522
TGCCAAAA	GCCGGCGA	9.2604	AGCTGGCT	CCAGGCTG	7.7192
CAGCAGCC	TGCGGAAT	9.1297	TTGGTCTC	AGGAAACA	7.6068
AGGGCCGT	TCCCGGCT	9.1249	CTGGGAGC	TCCGCCGG	7.3021

Table 5. Putative transcription factor binding modules discovered in promoters of DNA repair genes.

section we report the DNA motifs that were discovered, along with several clusters of related genes and promoters. We hypothesize that these similar components implicate regulatory networks responsible for co-regulation of the checkpoint factor genes.

We studied fourteen checkpoint factor genes, which are listed in Table 6. The number of alternative promoters per gene, shown in parentheses, varied for each gene. Because most of the genes have alternative promoters, we analyzed a total of thirty promoters. The complete set of alternative promotes is shown in Table 7. Alternative promoters were identified using annotations of genes in the UCSC Human Genome Browser. Transcription start sites of transcript isoforms served as the coordinates around which 900 bp upstream and 100 bp downstream were defined as the putative promoter region. Alternative promoters with significant overlap were truncated or removed from the analysis. DNA sequences were obtained for the forward and reverse strands of the genome to ensure coverage of words that might have biased nucleotide content and be subject to omission during the Markov model analysis stage.

Gene	Description from GeneCards (Safran 2010)
ATM (5)	The protein encoded by this gene (ataxia telangiectasia mutated) belongs to the PI3/PI4-kinase family. This protein functions as a regulator of a wide variety of downstream proteins, including *p53*, *BRCA1, CHK2, RAD17, RAD9*, and *NBS1*. This protein and the closely related kinase ATR are thought to be master controllers of cell cycle checkpoint signaling pathways, required for cell response to DNA damage and for genome stability.
ATR (2)	The protein encoded by this gene (ataxia telangiectasia and Rad3 related) belongs the PI3/PI4-kinase family, and is most closely related to ATM. Both proteins share similarity with

	Schizosaccharomyces pombe rad3, a cell cycle checkpoint gene required for cell cycle arrest and DNA damage repair in response to DNA damage. This kinase has been shown to phosphorylate *CHK1*, *RAD17*, and *RAD9* and *BRCA1*. Transcript variants utilizing alternative polyA sites exist.
ATRIP (1)	The product of this gene (ATR interacting protein) is an essential component of the DNA damage checkpoint, and binds to single-stranded DNA coated with replication protein A that accumulates at sites of DNA damage. The encoded protein interacts with the ataxia telangiectasia and Rad3 related protein, a checkpoint kinase, resulting in accumulation of the kinase at intranuclear foci induced by DNA damage. Multiple transcript variants encoding different isoforms have been found for this gene.
CHEK1 (3)	Required for checkpoint mediated cell cycle arrest in response to DNA damage or the presence of unreplicated DNA. May also negatively regulate cell cycle progression during unperturbed cell cycles. Binds to and phosphorylates *CDC25A, CDC25B* and *CDC25C*. Binds to and phosphorylates *RAD51.* Binds to and phosphorylates *TLK1.* May also phosphorylate multiple sites within the C-terminus of *TP53*, which promotes activation of *TP53* by acetylation and enhances suppression of cellular proliferation.
CHEK2 (2)	The protein encoded by this gene is a cell cycle checkpoint regulator and putative tumor suppressor. It contains a forkhead-associated protein interaction domain essential for activation in response to DNA damage and is rapidly phosphorylated in response to replication blocks and DNA damage. This protein interacts with and phosphorylates *BRCA1*, allowing *BRCA1* to restore survival after DNA damage. Three transcript variants encoding different isoforms have been found for this gene.
CLK2 (2)	This gene encodes a member of the *CLK* family of dual specificity protein kinases. *CLK* family members have been shown to interact with, and phosphorylate, serine- and arginine-rich (SR) proteins of the spliceosomal complex, which is a part of the regulatory mechanism that enables the SR proteins to control RNA splicing.
HUS1 (1)	The protein encoded by this gene is a component of an evolutionarily conserved, genotoxin-activated checkpoint complex that is involved in the cell cycle arrest in response to DNA damage. This protein forms a heterotrimeric complex with checkpoint proteins *RAD9* and *RAD1*. DNA damage induced chromatin binding has been shown to depend on the activation of the checkpoint kinase ATM, and is thought to be an early checkpoint signaling event.
MDC1 (2)	The protein encoded by this gene (mediator of DNA-damage checkpoint) is required to activate the intra-S phase and G2/M phase cell cycle checkpoints in response to DNA damage. This nuclear protein interacts with phosphorylated histone H2AX near sites of DNA double-strand breaks through its *BRCT* motifs, and facilitates

	recruitment of the ATM kinase and meiotic recombination 11 protein complex to DNA damage foci.
NBS1 (1)	The encoded protein is a member of the *MRE11/RAD50* double-strand break repair complex which consists of 5 proteins. This gene product is thought to be involved in DNA double-strand break repair and DNA damage-induced checkpoint activation.
P53/TP53 (3)	This gene encodes tumor protein *p53*, which responds to diverse cellular stresses to regulate target genes that induce cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism.
PER1 (1)	This gene is a member of the Period family of genes and is expressed in a circadian pattern in the suprachiasmatic nucleus, the primary circadian pacemaker in the mammalian brain. Genes in this family encode components of the circadian rhythms of locomotor activity, metabolism, and behavior. The specific function of this gene is not yet known. Alternative splicing has been observed in this gene; however, these variants have not been fully described.
RAD1 (2)	This gene encodes a component of a heterotrimeric cell cycle checkpoint complex, known as the 9-1-1 complex, that is activated to stop cell cycle progression in response to DNA damage or incomplete DNA replication. The 9-1-1 complex is recruited by *RAD17* to affected sites where it may attract specialized DNA polymerases and other DNA repair effectors. Alternatively spliced transcript variants of this gene have been described.
RAD17 (3)	The protein encoded by this gene is highly similar to the gene product of Schizosaccharomyces pombe rad17, a cell cycle checkpoint gene required for cell cycle arrest and DNA damage repair in response to DNA damage. This protein recruits the *RAD1-RAD9-HUS1* checkpoint protein complex onto chromatin after DNA damage,. The phosphorylation of this protein is required for the DNA-damage-induced cell cycle G2 arrest, and is thought to be a critical early event during checkpoint signaling in DNA-damaged cells. Eight alternatively spliced transcript variants of this gene, which encode four distinct proteins, have been reported.
RAD9A (2)	This gene product is highly similar to Schizosaccharomyces pombe rad9, a cell cycle checkpoint protein required for cell cycle arrest and DNA damage repair in response to DNA damage. This protein is found to possess 3' to 5' exonuclease activity, which may contribute to its role in sensing and repairing DNA damage. It forms a checkpoint protein complex with *RAD1* and *HUS1*. This complex is recruited by checkpoint protein *RAD17* to the sites of DNA damage, which is thought to be important for triggering the checkpoint-signaling cascade. Use of alternative polyA sites has been noted for this gene.

Table 6. The checkpoint factors genes that were studied. The number of alternative promoters is shown in parentheses next to each gene name.

Checkpoint Factors	Alternative promoters (hg18 coordinates)
ATM	*(ATM_5)* chr11:107662328-107663378_+ *(ATM_5)* chr11:107662328-107663378_- *(ATM_2)* chr11:107643346-107644396_+ *(ATM_2)* chr11:107643346-107644396_- *(ATM_3)* chr11:107597768-107598818_+ *(ATM_3)* chr11:107597768-107598818_- *(ATM_4)* chr11:107671910-107672960_+ *(ATM_4)* chr11:107671910-107672960_- *(ATM_5)* chr11:107679611-107680661_+ *(ATM_5)* chr11:107679611-107680661_-
ATR	*(ATR_1)* chr3:143780308-143781358_+ *(ATR_1)* chr3:143780308-143781358_- *(ATR_2)* chr3:143671051-143672101_+ *(ATR_2)* chr3:143671051-143672101_-
ATRIP	chr3:48462221-48463271_+ chr3:48462221-48463271_-
CHEK1	*($CHEK1_3$)* chr11:125000333-125001383_+ *($CHEK1_3$)* chr11:125000333-125001383_- *($CHEK1_2$)* chr11:125018185-125019235_+ *($CHEK1_2$)* chr11:125018185-125019235_- *($CHEK1_3$)* chr11:124999245-125000295_+ *($CHEK1_3$)* chr11:124999245-125000295_-
CHEK2	*($CHEK2_2$)* chr22:27467772-27468822_+ *($CHEK2_2$)* chr22:27467772-27468822_- *($CHEK2_2$)* chr22:27460665-27461715_+ *($CHEK2_2$)* chr22:27460665-27461715_-
CLK2	*($CLK2_2$)* chr1:153509855-153510905_+ *($CLK2_2$)* chr1:153509855-153510905_- *($CLK2_2$)* chr1:153514075-153515125_+ *($CLK2_2$)* chr1:153514075-153515125_-
HUS1	chr7:47985721-47986771_+ chr7:47985721-47986771_-

MDC1	*($MDC1_2$)* chr6:30792781-30793831_+ *($MDC1_2$)* chr6:30792781-30793831_- *($MDC1_2$)* chr6:30789060-30790110_+ *($MDC1_2$)* chr6:30789060-30790110_-
NBS1	chr8:91066025-91067075_+ chr8:91066025-91067075_-
P53 (TP53)	*($TP53_3$)* chr17:7519486-7520536_+ *($TP53_3$)* chr17:7519486-7520536_- *($TP53_2$)* chr17:7531538-7532588_+ *($TP53_2$)* chr17:7531538-7532588_- *($TP53_3$)* chr17:7520612-7521662_+ *($TP53_3$)* chr17:7520612-7521662_-
PER1	chr17:7996377-7997427_+ chr17:7996377-7997427_-
RAD1	*($RAD1_2$)* chr5:34954089-34955139_+ *($RAD1_2$)* chr5:34954089-34955139_- *($RAD1_2$)* chr5:34951438-34952488_+ *($RAD1_2$)* chr5:34951438-34952488_-
RAD17	*($RAD17_3$)* chr5:68699879-68700929_+ *($RAD17_3$)* chr5:68699879-68700929_- *($RAD17_2$)* chr5:68723716-68724766_+ *($RAD17_2$)* chr5:68723716-68724766_- *($RAD17_3$)* chr5:68701287-68702337_+ *($RAD17_3$)* chr5:68701287-68702337_-
RAD9A	*($RAD9A_2$)* chr11:66918716-66919766_+ *($RAD9A_2$)* chr11:66918716-66919766_- *($RAD9A_2$)* chr11:66914998-66916048_+ *($RAD9A_2$)* chr11:66914998-66916048_-

Table 7. Alternative promoters, indicated by their genomic coordinates, of genes involved in cell-cycle checkpoint factor pathways.

Statistical analysis of thirty promoters found several interesting DNA words, which predict DNA elements that participate in the regulation of the DNA repair checkpoint factors. The most significant words discovered are listed in Table 8. Words that are shared among the gene sets identify regulatory relationships. Reverse complement words are reported separately, as internal verification on the process. Words without a reverse complement example indicate a particular bias in the nucleotide content.

Word	Promoters	Sln(S/Es)
ACAGCCAT	ATM_2	5.41
	$CHEK2_2$	
	$CLK2_1$	
ATGGCTGT	ATM_2	5.41
	$CHECK2_2$	
	$CLK2_1$	
GCCTGGGA	ATR_1	5.40
	$CHEK2_1$	
	$CLK2_2$	
	$MDC1_1$	
	$MDC1_2$	
	$RAD1_2$	
TCCCAGGC	ATR_1	5.40
	$CHEK2_1$	
	$CLK2_2$	
	$MDC1_1$	
	$MDC1_2$	
	$RAD1_2$	
ACTCCCTA	ATM_3	5.29
	$CHEK2_1$	
	$RAD17_2$	
TAGGGAGT	ATM_3	5.29
	$CHEK2_1$	
	$RAD17_2$	
AGCGGCCA	ATR_1	5.24
	ATR_2	
	$CHEK1_1$	
TGGCCGCT	ATR_1	5.24
	ATR_2	
	$CHEK1_1$	
GAAATGAA	ATM_2	5.24
	ATM_3	
	ATR_2	
	$CLK2_2$	
	$HUS1$	
	$MDC1_1$	
TTCATTTC	ATM_2	5.24
	ATM_3	
	ATR_2	
	$CLK2_2$	
	$HUS1$	
	$MDC1_1$	
AATGCAGG	$RAD1_1$	4.97
	$TP53_1$	
	$TP53_2$	
	$TP53_3$	

CCTGCATT	$RAD1_1$	4.97
	$TP53_1$	
	$TP53_2$	
	$TP53_3$	
ATCCCTGA	*ATRIP*	4.73
	$CHEK1_3$	
	$RAD1_2$	
	$RAD17_1$	
TCAGGGAT	*ATRIP*	4.73
	$CHEK1_3$	
	$RAD1_2$	
	$RAD17_1$	
GTATTTTA	ATM_4	4.58
	$CHEK1_2$	
	NBS1	
	$RAD17_1$	
	ATR_2	
	$TP53_1$	
	HUS1	
	$RAD17_1$	

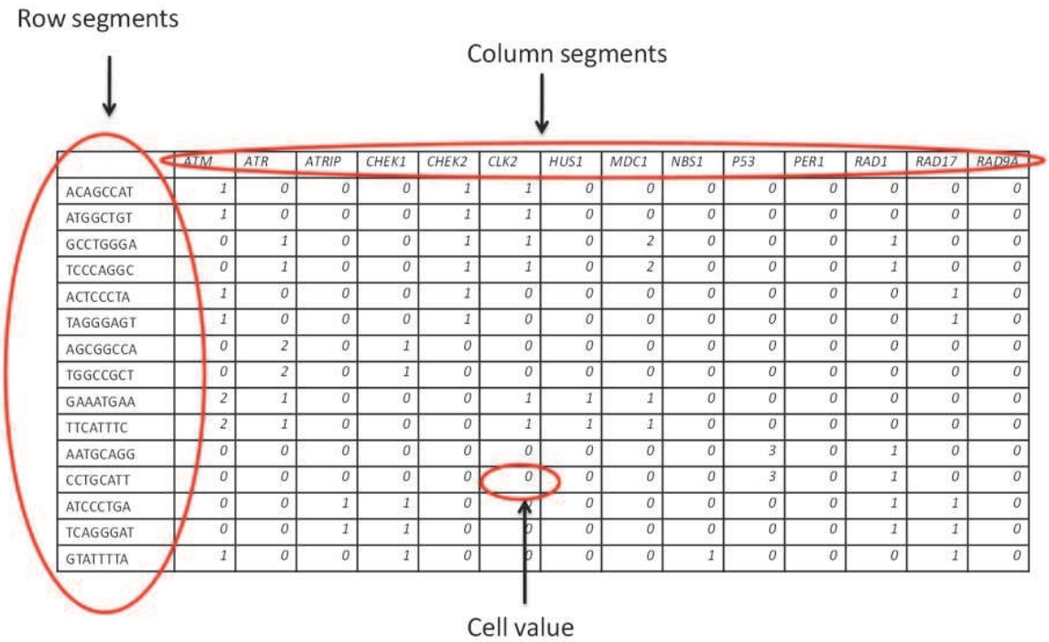
Table 8. Top 15 enumerated DNA words, based on the $S \cdot \ln(S/E_S)$ overrepresentation score, and the alternative promoters, identified by subscript.

5. Visualization and interpretation of data

Shared words among the checkpoint factor genes suggested the presence of regulatory networks. We assessed the relationships by generating network depictions in the form of interaction networks (Figure 2) and a circos diagram (Figure 3) constructed from the summary data in Table 9. To derive Figure 2, a metric MDS was conducted on the affiliation network defined in Table 9. The resulting graph was then spring-embedded, with node repulsion, to facilitate visualization (Borgatti, 2002). The interaction network depicts the distribution of the DNA words among the genes (note that each gene appears once, representing all alternative promoters as a single node). Genes are denoted by blue squares and words are represented with red circles. Bold lines indicate multiple occurrences of a word. Reverse complement words are shown independently.

The circos diagram represents the information in a closed circular space, wherein connections between words on one side of the diagram extend to genes on the other side. The putative nodes of the regulatory networks are defined by multiple edges, representing a characterized transcription factor or a novel DNA binding site, or a checkpoint factor gene.

Some of the discovered words correspond to known binding sites for transcription factors, reported in the JASPAR and TRANSFAC databases of transcription factors (see Table 10). The relationships between the top fifteen words and the transcription factors are depicted in the circos diagram in Figure 4. Note that multiple binding site motifs were discovered for many of the transcription factors, and that several of the sites match the binding patterns of more than one transcription factor.



	ATM	ATR	ATRIP	CHEK1	CHEK2	CLK2	HUS1	MDC1	NBS1	P53	PER1	RAD1	RAD17	RAD9A
ACAGCCAT	1	0	0	0	1	1	0	0	0	0	0	0	0	0
ATGGCTGT	1	0	0	0	1	1	0	0	0	0	0	0	0	0
GCCTGGGA	0	1	0	0	1	1	0	2	0	0	0	1	0	0
TCCCAGGC	0	1	0	0	1	1	0	2	0	0	0	1	0	0
ACTCCCTA	1	0	0	0	1	0	0	0	0	0	0	0	1	0
TAGGGAGT	1	0	0	0	1	0	0	0	0	0	0	0	1	0
AGCGGCCA	0	2	0	1	0	0	0	0	0	0	0	0	0	0
TGGCCGCT	0	2	0	1	0	0	0	0	0	0	0	0	0	0
GAAATGAA	2	1	0	0	0	1	1	1	0	0	0	0	0	0
TTCATTTC	2	1	0	0	0	1	1	1	0	0	0	0	0	0
AATGCAGG	0	0	0	0	0	0	0	0	0	3	0	1	0	0
CCTGCATT	0	0	0	0	0	0	0	0	0	3	0	1	0	0
ATCCCTGA	0	0	1	1	0	0	0	0	0	0	0	1	1	0
TCAGGGAT	0	0	1	1	0	0	0	0	0	0	0	1	1	0
GTATTTTA	1	0	0	1	0	0	0	0	1	0	0	0	1	0

Table 9. The top ranked words (rows of the table), based on statistical significance ($S \cdot \ln(S/E_s)$), and the number of occurrences of each word in the promoter regions of genes (columns).

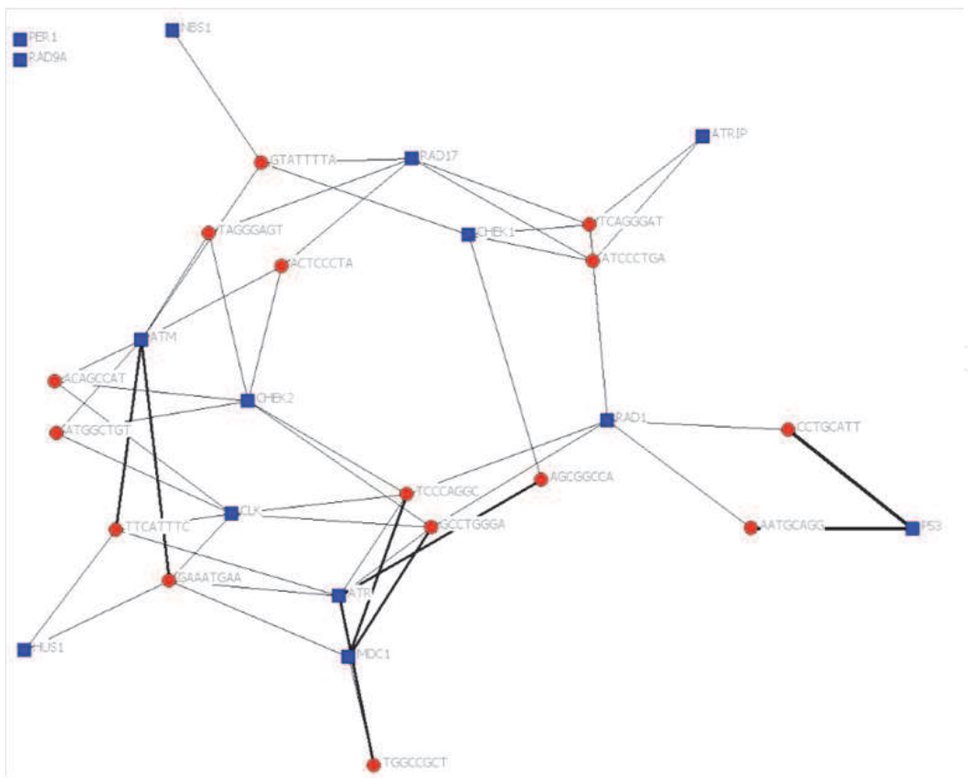


Fig. 2. Model of the checkpoint regulatory network using multidimensional scaling.

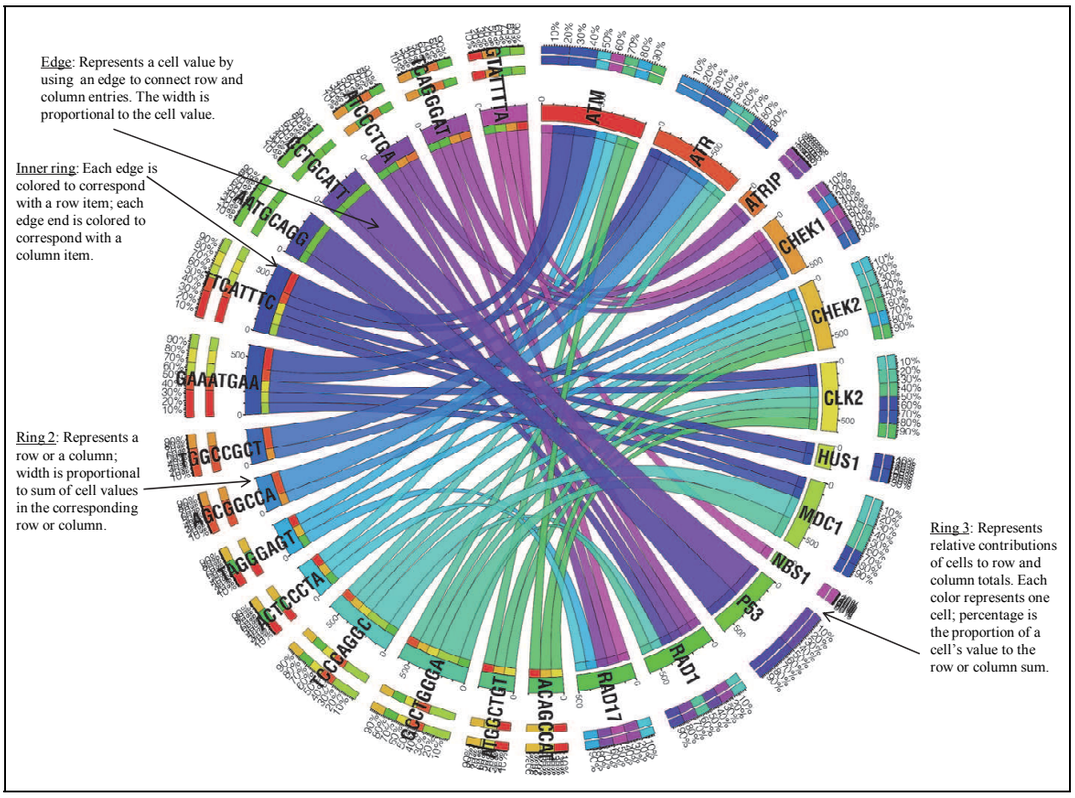


Fig. 3. Circos 2diagram of the top 15 words, based on statistical significance, and their occurrences in gene promoter regions.

TFBS	$S \cdot \ln(S/E_s)$	TF
ACCCCCAC	3.76	*PF0091, Pax-4*
ACTCCCTA	4.67	*Helios A, p300*
ATGGCTGT	5.42	*Cap*
ATTAAAGA	3.72	*Pax-2*
CGGAGCCC	3.95	*LF-A1*
CTGAAATT	3.80	*STAT1, STAT6*
CTTTTGAA	3.83	*TCF-4*
GAAAAATT	3.76	*CIZ*
GCACCTGC	3.68	*PF0035, AP-4, cap, Lmo2 complex*
GTGGCTGC	3.64	*cap*
TACTTTTT	3.82	*FOXC, CIZ, RUSH-1alpha1*
TATATTTA	3.82	*FOXL1, PF0028, PF0054*
TCCTTTCT	3.70	*Pax-2*
TTTTTATA	3.64	*FOXL1*

Table 10. Known transcription factor binding sites (with significance scores and corresponding transcription factor) discovered in the promoters of the checkpoint factors genes.

Additional insight into the regulatory network for the checkpoint factors can be seen in Figure 5, which replaces the DNA binding site motifs with the names of implicated transcription factors for each DNA repair gene. The diagram indicates the discovery of specific transcription factors involved in the control of each gene and shared among multiple genes. Up to seven transcription factors were discovered for each gene.

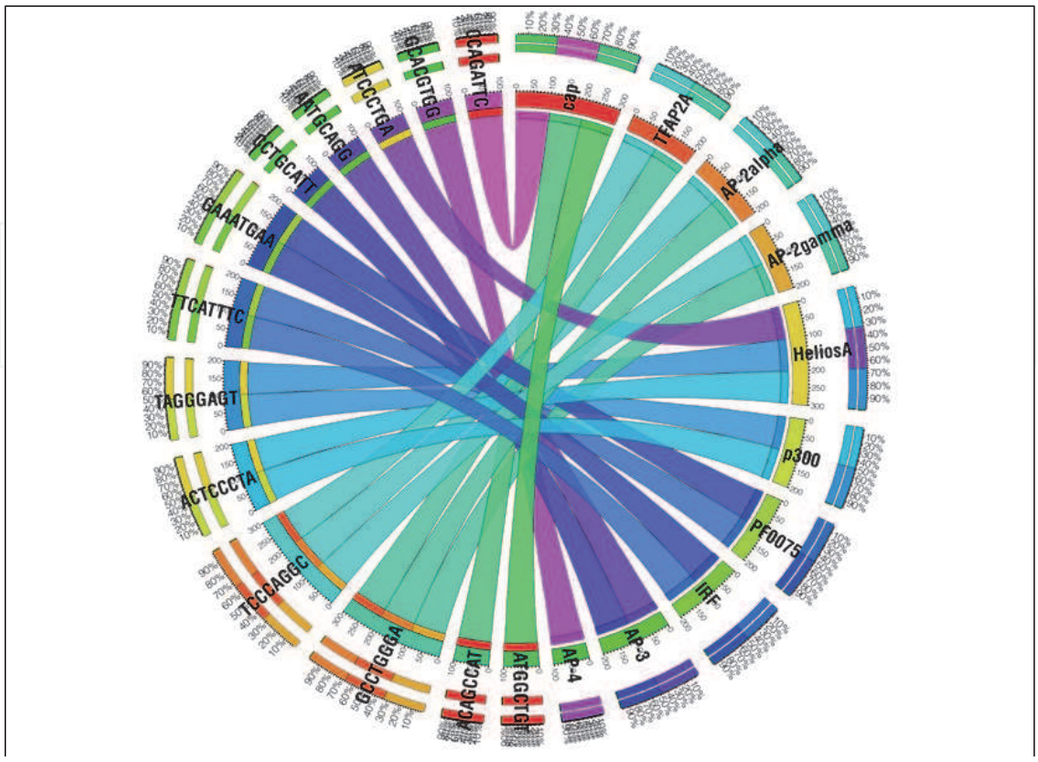


Fig. 4. Circos diagram showing the top 15 DNA motifs found in promoters of checkpoint factor genes and their related transcription factors (number of occurrences are multiplied by 100).

6. Conclusions

This chapter provides a summary of research into transcriptional regulatory networks controlling DNA repair pathways, bidirectional versus unidirectional promoters of DNA repair genes, and bidirectional promoters of breast and ovarian cancer genes. DNA words are shared among these promoters, and these words represent both known and unknown binding sites for transcription factors. When possible, we report the highest scoring assignment of transcription factor to DNA word. Our research represents a novel approach to identifying factors involved in transcriptional regulation of DNA repair genes. Many of these proteins have dual roles in transcription and DNA repair. Although many of the regulatory relationships are characterized at the level of protein-protein interactions, little research is available on the transcriptional regulatory networks that control DNA repair gene expression. We present evidence that regulatory networks exist among these genes, and support the claim that bidirectional promoters (implicated in B/O cancers) have a distinct network from unidirectional promoters. The identification of putative binding sites provides the first step in the elucidation of higher-order interdependencies among DNA repair genes in the cell. We also report preliminary findings on pairs of binding sites that represent regulatory modules. Furthermore, we show that there is much overlap among promoters of DNA repair genes, and that shared DNA binding motifs can be distributed among a collection of alternative promoters, each having distinct combinations of regulatory elements. The complex nature of the data can be simplified for visual interpretation using visualization techniques such as network modeling and circos diagrams.

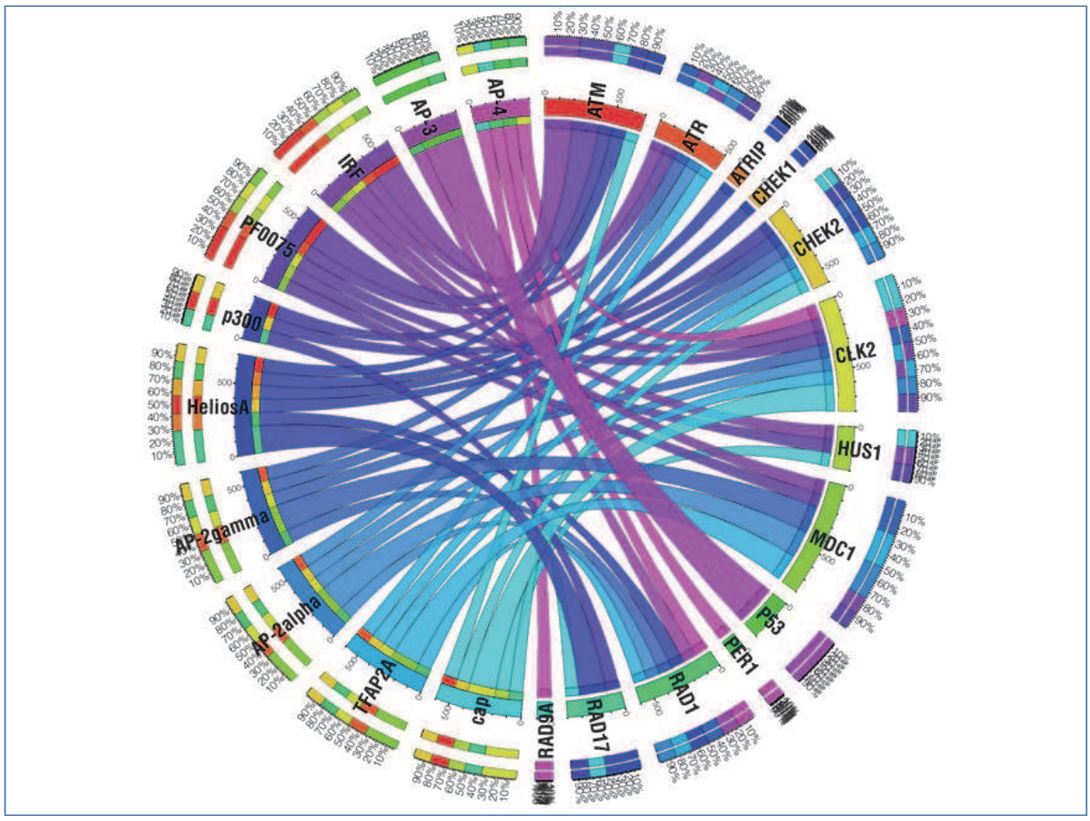


Fig. 5. Relationships between genes and transcription factors.

7. Acknowledgments

This work was supported by the Intramural Research Program of NHGRI (LE and LK) and by the Ohio Plant Biotechnology Consortium, the Choose Ohio First Program of the University System of Ohio, and the Ohio University Graduate Research and Education Board (LW).

8. References

Bellizzi, A.M., & Frankel, W.L. Colorectal cancer due to deficiency in DNA mismatch repair function: a review. *Adv Anat Pathol*. 2009 Nov;16(6):405-17.

Berwick, M. & Vineis, P. Markers of DNA repair and susceptibility to cancer in humans: an epidemiologic review. *J Natl Cancer Inst*. 2000 Jun 7;92(11):874-97.

Borgatti, S.P., Everet, M.G., and Freeman, L.C. 2002. Ucinet 6 for Windows: Software for Social Network Analysis. Harvard: Analytic Technologies.

Bryne, J.C., & Valen, E., Tang, M.H., Marstrand, T., & Winther, O. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* 2008, 36:D102-106.

Elnitski, L., Lichtenberg, J., & Welch, L.R.Regulatory network nodes of checkpoint factors in DNA repair pathways. *BCB '10: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, 529-536. 2010.

Helleday, T., Petermann, E., Lundin, C., Hodgson B., & Sharma, R.A. DNA repair pathways as targets for cancer therapy, *Nature Reviews Cancer* (2008) 8 (3): 193-204.

Hurley, P.J. & Bunz, F. ATM and ATR: components of an integrated circuit. *Cell Cycle*. 2007 Feb 15;6(4):414-7.

Jensen, N.M., Dalsgaard, T., Jakobsen, M., Nielsen, R.R., Sørensen, C.B., Bolund, L., & Jensen, T.G. An update on targeted gene repair in mammalian cells: methods and mechanisms. *J Biomed Sci*. 2011 Feb 2;18:10.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., & Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Research* (2008) 36:D480-D484.

Kelley, M.R. & Fishel, M.L. DNA repair proteins as molecular targets for cancer therapeutics. *Anticancer Agents Med Chem*. 2008 May;8(4):417-25.

Kinsella, T.J. Understanding DNA damage response and DNA repair pathways: applications to more targeted cancer therapeutics. *Semin Oncol*. 2009 Apr;36(2 Suppl 1):S42-51.

Li, C., Wang, L.E., & Wei, Q. DNA repair phenotype and cancer susceptibility--a mini review. *Int J Cancer*. 2009 Mar 1;124(5):999-1007.

Liang, Y., Lin, S.Y., Brunicardi, F.C., Goss, J., & Li, K. DNA damage response pathways in tumor suppression and cancer treatment. *World J Surg*. 2009 Apr;33(4):661-6.

Lichtenberg, J., Jacox, E., Welch, J.D., Kurz, K., Liang, X., Yang, M.Q., Drews, F., Ecker, K., Lee, S.S., Elnitski, L., & Welch, L.R. Word-based characterization of promoters involved in human DNA repair pathways. *BMC Genomics* 10(Suppl 1):S18. 2009.

Lichtenberg, J., Kurz, K., Liang, X., Al-ouran, R., Neiman, L. Nau, L.J., Welch, J.D., Jacox, E., Bitterman, T., Ecker, K., Elnitski, L., Drews, F., Lee, S.S., & Welch, L.R. WordSeeker: Concurrent bioinformatics software for discovering genome-wide patterns and word-based genomic signatures. *BMC Bioinformatics*, 2010, 11(Suppl 12):S6.

Naccarati, A., Pardini, B., Hemminki, K., &Vodicka, P. Sporadic colorectal cancer and individual susceptibility: a review of the association studies investigating the role of DNA repair genetic polymorphisms. *Mutat Res*. 2007 May-Jun;635(2-3):118-45.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., & Lancet, D. GeneCards Version 3: the human gene integrator. *Database* (Oxford). 2010 Aug 5;2010:baq020.

Shibata, A., Barton, O., Noon, A. T., Dahm, K., Deckbar D., Goodarzi, A.A., Löbrich, M., & Jeggo, P.A. Role of ATM and the damage response mediator proteins 53BP1 and MDC1 in the maintenance of G(2)/M checkpoint arrest. *Mol Cell Biol*. 2010 Jul;30(13):3371-83.

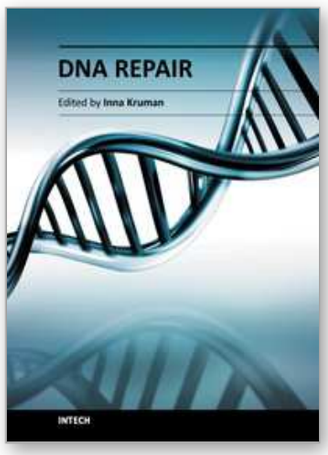
Smith, J., Tho, L.M., Xu, N., & Gillespie, D.A. The ATM-Chk2 and ATR-Chk1 pathways in DNA damage signaling and cancer. *Adv Cancer Res*. 2010;108:73-112.

Wingender, E., Chen, X., Hehl, R., Karas, H., & Liebich, I. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 2000, 28:316-319.

Wood R.D., Mitchell, M., & T. Lindahl, Human DNA repair genes, 2005, *Mutation Res* (2005) 577 (1-2): 275-83.

Xie, X., Lu J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., & Kellis, M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*. 2005 Mar 17;434(7031):338-45.

Yang, M.Q., Koehly, L.M., & Elnitski, L.L. Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes. *PLoS Comput Biol*. 2007 Apr 20;3(4):e72.



DNA Repair

Edited by Dr. Inna Kruman

ISBN 978-953-307-697-3
Hard cover, 636 pages
Publisher InTech
Published online 07, November, 2011
Published in print edition November, 2011

The book consists of 31 chapters, divided into six parts. Each chapter is written by one or several experts in the corresponding area. The scope of the book varies from the DNA damage response and DNA repair mechanisms to evolutionary aspects of DNA repair, providing a snapshot of current understanding of the DNA repair processes. A collection of articles presented by active and laboratory-based investigators provides a clear understanding of the recent advances in the field of DNA repair.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Lonnie R. Welch, Laura M. Koehly and Laura Elnitski (2011). Shared Regulatory Motifs in Promoters of Human DNA Repair Genes, DNA Repair, Dr. Inna Kruman (Ed.), ISBN: 978-953-307-697-3, InTech, Available from: http://www.intechopen.com/books/dna-repair/shared-regulatory-motifs-in-promoters-of-human-dna-repair-genes

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen