

The Role of ZBTB48 In Regulating Biological Aging

Introduction

Aging is the continuous process of functional decay, which is a result of the atrophy of tissues and organs, as well as the cumulative damage to biomolecules such as DNA. Many mechanisms have been proposed to explain why aging occurs in nature, but the most compelling is that aging is a consequence of the accumulation of DNA damage¹. Therefore, the role of DNA repair is vital in curtailing the effects of senescence. The regulation of DNA repair enzymes could hold clues to the mechanisms involved in maintaining genomic stability and offsetting the consequences of aging. Of particular interest is the transcription factors involved in regulating the expression of genes involved in DNA damage repair, and how their ability to silence or enhance the expression of these genes could affect the aging process. The zinc-finger protein ZBTB48 is responsible for regulating telomere length but has also been shown to act as a transcription factor². Due to its involvement in DNA maintenance and the well-known ability of zinc finger proteins to bind to DNA, we speculate that ZBTB48 also acts as a transcriptional regulator for DNA repair genes. Here we will use techniques such as RNA-seq and ChIP-seq to determine the effect ZBTB48 has on the expression of DNA repair enzymes and look for evidence of its ability to act as a transcription factor for proteins involved in DNA repair pathways.

Methods

To perform the necessary analysis, we used the GSE96778 dataset. This dataset contains sequence data for both RNA-seq and ChIP-seq analysis. The dataset is characterized by an experimental design where researchers used wild-type HeLa cells as the treatment and a ZBTB48 knockout as the negative control. This is ideal, as we can observe how the expression of DNA repair enzymes is affected as the activity of ZBTB48 is eliminated. Knocking out ZBTB48 can also allow us to observe its binding activity, since we can compare the binding in the wild-type to that of the knockout treatment, enabling us to find to which genes it bound to. However, one caveat about this analysis is that since we are dealing with cancer cells, there maybe some DNA repair mechanisms that are present or absent when compared to normal human cells. Nevertheless, we hope this analysis can give us insight into the ability of ZBTB48 to act as a transcription factor.

First, we will describe how we analyzed the RNA-seq data to understand how ZBTB48 affects the expression of DNA repair enzymes. The GSE96778 dataset is downloaded from the GEO database using the SRA toolset. The downloaded data corresponds to five wild-type replicates and five ZBTB48 knockout samples:

```
fastq-dump --split-files --gzip SRR5356290
fastq-dump --split-files --gzip SRR5356291
fastq-dump --split-files --gzip SRR5356292
fastq-dump --split-files --gzip SRR5356293
fastq-dump --split-files --gzip SRR5356294
fastq-dump --split-files --gzip SRR5356295
fastq-dump --split-files --gzip SRR5356296
fastq-dump --split-files --gzip SRR5356297
fastq-dump --split-files --gzip SRR5356298
fastq-dump --split-files --gzip SRR5356299
```

We then run kallisto for each sample to calculate the transcript abundances for each gene:

```
kallisto quant -i  
./Genomics2020/data/human/Homo_sapiens.GRCh38.cdna.all.release-94_k31.idx -o  
./seaceved/kallisto_ZBTB48/SRR5356293_GRCh38  
./seaceved/FinalProject/SRR5356293_1.fastq.gz  
./seaceved/FinalProject/SRR5356293_2.fastq.gz
```

```
kallisto quant -i  
./Genomics2020/data/human/Homo_sapiens.GRCh38.cdna.all.release-94_k31.idx -o  
./seaceved/kallisto_ZBTB48/SRR5356294_GRCh38  
./seaceved/FinalProject/SRR5356294_1.fastq.gz  
./seaceved/FinalProject/SRR5356294_2.fastq.gz
```

```
kallisto quant -i  
./Genomics2020/data/human/Homo_sapiens.GRCh38.cdna.all.release-94_k31.idx -o  
./seaceved/kallisto_ZBTB48/SRR5356295_GRCh38  
./seaceved/FinalProject/SRR5356295_1.fastq.gz  
./seaceved/FinalProject/SRR5356295_2.fastq.gz
```

```
kallisto quant -i  
./Genomics2020/data/human/Homo_sapiens.GRCh38.cdna.all.release-94_k31.idx -o  
./seaceved/kallisto_ZBTB48/SRR5356296_GRCh38  
./seaceved/FinalProject/SRR5356296_1.fastq.gz  
./seaceved/FinalProject/SRR5356296_2.fastq.gz
```

```
kallisto quant -i  
./Genomics2020/data/human/Homo_sapiens.GRCh38.cdna.all.release-94_k31.idx -o  
./seaceved/kallisto_ZBTB48/SRR5356297_GRCh38  
./seaceved/FinalProject/SRR5356297_1.fastq.gz  
./seaceved/FinalProject/SRR5356297_2.fastq.gz
```

```
kallisto quant -i  
./Genomics2020/data/human/Homo_sapiens.GRCh38.cdna.all.release-94_k31.idx -o  
./seaceved/kallisto_ZBTB48/SRR5356298_GRCh38  
./seaceved/FinalProject/SRR5356298_1.fastq.gz  
./seaceved/FinalProject/SRR5356298_2.fastq.gz
```

```
kallisto quant -i  
./Genomics2020/data/human/Homo_sapiens.GRCh38.cdna.all.release-94_k31.idx -o  
./seaceved/kallisto_ZBTB48/SRR5356299_GRCh38  
./seaceved/FinalProject/SRR5356299_1.fastq.gz  
./seaceved/FinalProject/SRR5356299_2.fastq.gz
```

After importing the abundance.tsv files in R, the data is transformed into an acceptable format for further downstream analysis using DESeq2:

```
SRR5356290 <- read.delim("SRR5356290_GRCh38.tsv")
SRR5356291 <- read.delim("SRR5356291_GRCh38.tsv")
SRR5356292 <- read.delim("SRR5356292_GRCh38.tsv")
SRR5356293 <- read.delim("SRR5356293_GRCh38.tsv")
SRR5356294 <- read.delim("SRR5356294_GRCh38.tsv")
SRR5356295 <- read.delim("SRR5356295_GRCh38.tsv")
SRR5356296 <- read.delim("SRR5356296_GRCh38.tsv")
SRR5356297 <- read.delim("SRR5356297_GRCh38.tsv")
SRR5356298 <- read.delim("SRR5356298_GRCh38.tsv")
SRR5356299 <- read.delim("SRR5356299_GRCh38.tsv")
cts <- data.frame(
  WT_repA = round(SRR5356290$est_counts),
  WT_repB = round(SRR5356291$est_counts),
  WT_repC = round(SRR5356292$est_counts),
  WT_repD = round(SRR5356293$est_counts),
  WT_repE = round(SRR5356294$est_counts),
  ZBTB48_KO_repA = round(SRR5356295$est_count),
  ZBTB48_KO_repB = round(SRR5356296$est_count),
  ZBTB48_KO_repC = round(SRR5356297$est_count),
  ZBTB48_KO_repD = round(SRR5356298$est_count),
  ZBTB48_KO_repE = round(SRR5356299$est_count)
)
rownames(cts) <- SRR5356290$target_id
coldata <- data.frame(
  treatment = c(rep("WT",5), rep("ZBTB48_KO",5))
)
rownames(coldata) <- colnames(cts)
```

After importing the necessary libraries such as DESeq and pheatmap, we analyzed those genes whom have statistically significant expression changes (adjusted p-value < 0.1) and have the largest log fold changes. This was to ensure we analyzed the genes that were most likely affected by the absence of ZBTB48. We constructed a heatmap to allow us to more clearly identify the most differentially expressed genes:

```
library("DESeq2")
dds <- DESeqDataSetFromMatrix(countData = cts,colData = coldata,
design = ~treatment)
```

```

dds <- DESeq(dds)
res <- results(dds)
library("pheatmap")
mat = assay(vsd)[head(order(res$padj),20), ]
mat = mat - rowMeans(mat)
df = as.data.frame(colData(vsd)[c("treatment")])
colnames(df) = "condition"
rownames(df) = colnames(mat)
pheatmap(mat, annotation_col=df)

```

For the ChIP-seq analysis, we first ran alignments using bwa for three different treatments: two wild-type HeLa cells and one ZBTB48 knockout sample both treated with the ZBTB48 antibody:

```

bwa mem hg38.fa.gz /project/meisel/seaceved/FinalProject/SRR5356232.fastq.gz >
/project/meisel/seaceved/FinalProject/bwa_ZBTB48/SRR5356232_hg38.sam

```

```

bwa mem hg38.fa.gz /project/meisel/seaceved/FinalProject/SRR5356236.fastq.gz >
/project/meisel/seaceved/FinalProject/bwa_ZBTB48/SRR5356236_hg38.sam

```

```

bwa mem hg38.fa.gz /project/meisel/seaceved/FinalProject/SRR5356237.fastq.gz >
/project/meisel/seaceved/FinalProject/bwa_ZBTB48/SRR5356237_hg38.sam

```

We then converted the resulting .sam files into .bam files using samtools to minimize storage:

```

samtools view -S -b SRR5356232_hg38.sam > SRR5356232_hg38.bam
samtools view -S -b SRR5356236_hg38.sam > SRR5356236_hg38.bam
samtools view -S -b SRR5356237_hg38.sam > SRR5356237_hg38.bam

```

We then performed peak calling using MACS against the knockout control for the two different treatment samples:

```

macs2 callpeak -t /project/meisel/seaceved/FinalProject/bwa_ZBTB48/SRR5356236_hg38.bam -c
/project/meisel/seaceved/FinalProject/bwa_ZBTB48/SRR5356232_hg38.bam -f BAM -n
SRR5356236 -g hs --keep-dup 10 -B -m 10 100 --outdir
/project/meisel/seaceved/FinalProject/MACS_Revised

```

```

macs2 callpeak -t /project/meisel/seaceved/FinalProject/bwa_ZBTB48/SRR5356237_hg38.bam -c
/project/meisel/seaceved/FinalProject/bwa_ZBTB48/SRR5356232_hg38.bam -f BAM -n
SRR5356237 -g hs --keep-dup 10 -B -m 10 100 --outdir
/project/meisel/seaceved/FinalProject/MACS_Revised

```

This was followed by a data transformation to allow for compatibility with the UCSC genome browser. We added headers to both the .narrowPeaks and the .bedGraph files that were produced by MACS:

```
echo "track type=narrowPeak name= SRR5356236_narrowPeak" >>
SRR5356236_narrowPeak_header.txt
cat SRR5356236_narrowPeak_header.txt SRR5356236_peaks.narrowPeak >
SRR5356236.narrowPeak.header
rm SRR5356236_narrowPeak_header.txt
```

```
echo "track type=narrowPeak name=SRR5356237_narrowPeak" >>
SRR5356237_narrowPeak_header.txt
cat SRR5356237_narrowPeak_header.txt SRR5356237_peaks.narrowPeak >
SRR5356237.narrowPeak.header
rm SRR5356237_narrowPeak_header.txt
```

```
echo "track type=bedGraph name= SRR5356236.bdg" >> SRR5356236_treat_pileup_header.txt
cat SRR5356236_treat_pileup_header.txt SRR5356236_treat_pileup.bdg > SRR5356236.bedGraph
rm SRR5356236_treat_pileup_header.txt
gzip SRR5356236.bedGraph
```

```
echo "track type=bedGraph name= SRR5356237.bdg" >> SRR5356237_treat_pileup_header.txt
cat SRR5356237_treat_pileup_header.txt SRR5356237_treat_pileup.bdg > SRR5356237.bedGraph
rm SRR5356237_treat_pileup_header.txt
gzip SRR5356237.bedGraph
```

Using the genome browser allowed us to locate binding sites for the ZBTB48 protein and more specifically look for binding to DNA maintenance genes. To extract the motif sequences from the peaks near DNA repair enzymes, we first centered the peaks on a 100 bp window of both treatment replicates using bedtools:

```
awk '{ $2 = $2 - 50; print }' SRR5356236_summits.bed | awk -v OFS='\t' '{ $3 = $3 + 50; print }' >
SRR5356236.window.bed
awk '{ $2 = $2 - 50; print }' SRR5356237_summits.bed | awk -v OFS='\t' '{ $3 = $3 + 50; print }' >
SRR5356237.window.bed
```

We then take the intersect of the peaks centered on a 100 bp window to produce a set of high confidence peaks:

```
intersectBed -a SRR5356236.window.bed -b SRR5356237.window.bed > shared_summits.bed
```

We then selected the summits that overlapped by at least 20 bp:

```
awk '($3 - $2) > 20 { print $0 }' shared_summits.bed > shared20bp_summits.bed
```

Next, we took a random sample of peaks along with selecting the 1000 highest peaks. This is for comparison purposes, as we can compare the sequences extracted from the random peaks to the most intense peaks to prevent obtaining any false positive motifs:

```
perl -MList::Util -e 'print List::Util::shuffle <>' shared20bp_summits.bed | head -1000 > ran.1K.bed
```

```
sort -k5 -r -n shared20bp_summits.bed | head -1000 > shared.top1K.bed
```

Next, we can use bedtools to extract the sequences from the human genome fasta file, hg38.fa.

```
fastaFromBed -fi /project/meisel/Genomics2019/data/ChIPseq/hg38.fa -bed ran.1K.bed -fo ran.1K.fasta
```

```
fastaFromBed -fi /project/meisel/Genomics2019/data/ChIPseq/hg38.fa -bed shared.top1K.bed -fo shared.top1K.fasta
```

We then run MEME and fimo to extract the motif sequences, which was followed by querying the motif against a known database using TOMTOM:

```
meme ran.1K.fasta -oc ran.1K.meme -dna -nmotifs 2 -minw 5 -maxw 15 -revcomp
```

```
meme shared.top1K.fasta -oc shared.top1K.meme -dna -nmotifs 2 -minw 5 -maxw 15 -revcomp
```

```
fimo -oc shared.top1K.fimo shared.top1K.meme/meme.txt  
/project/meisel/Genomics2020/data/ChIPseq/hg38.fa
```

```
tomtom -no-ssc -oc shared.top1K.tomtom -thresh 0.01 -eps shared.top1K.meme/meme.txt  
/project/meisel/Genomics2020/data/ChIPseq/me-14-1130-2.txt
```

Results

We looked at the expression of genes which have statistically significant log fold changes. Figure 1 compares the expression of the first 20 genes in this group. More specifically figure 1 shows the expression of the top 20 most differentially expressed genes, defined here as those with the lowest adjusted p-values. The expression of each gene was normalized by subtracting from the mean expression of all samples. These results seem to suggest that ZBTB48 is more likely to act as an inducer than a repressor. Extending the analysis by looking at the expression of the first 50 genes, we observed that there were many more genes that showed higher expression in the wild-type HeLa cells. As such, we will focus mostly on the inductive capabilities of ZBTB48 for the remainder of our analysis.

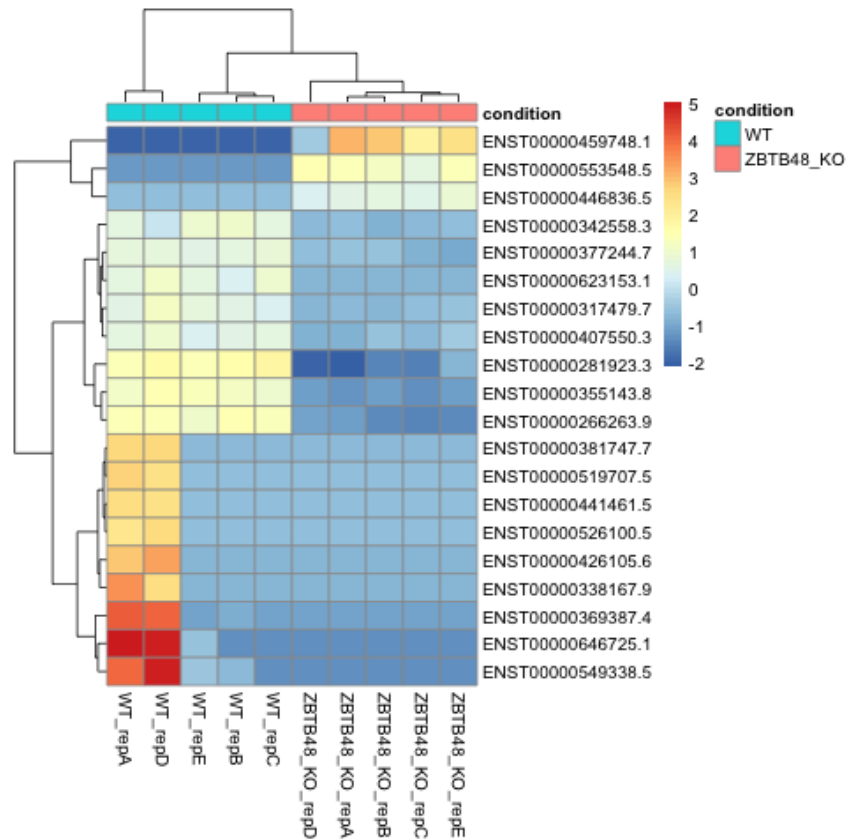
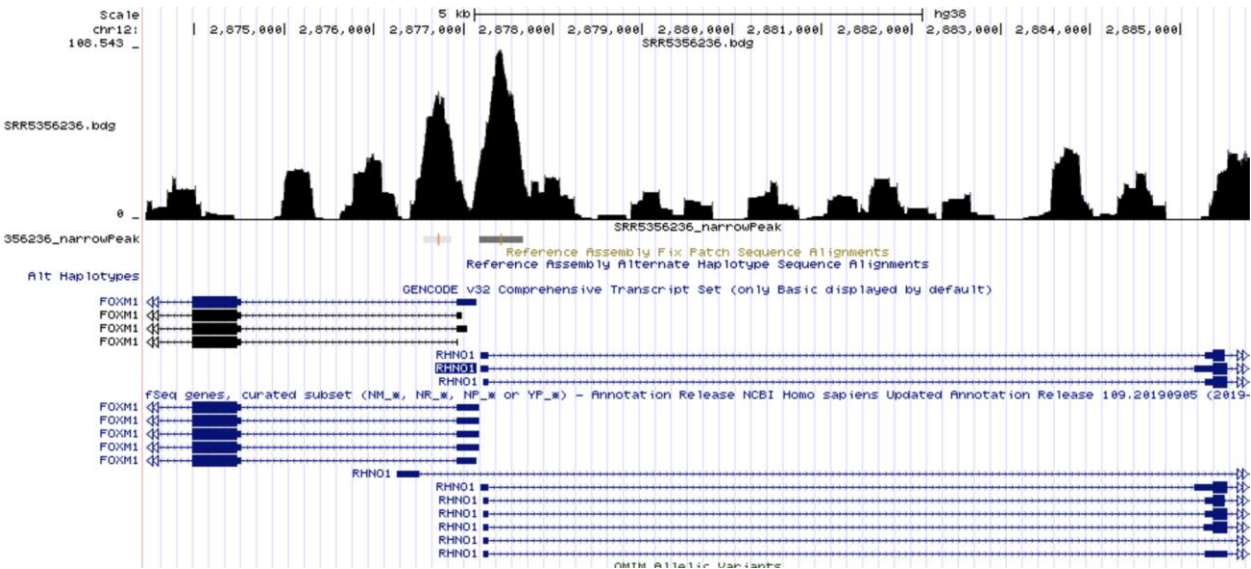


Figure 1. The first 20 most differentially expressed genes (lowest adjusted p-values). The expression of each gene was normalized by subtracting from the mean expression of all samples.

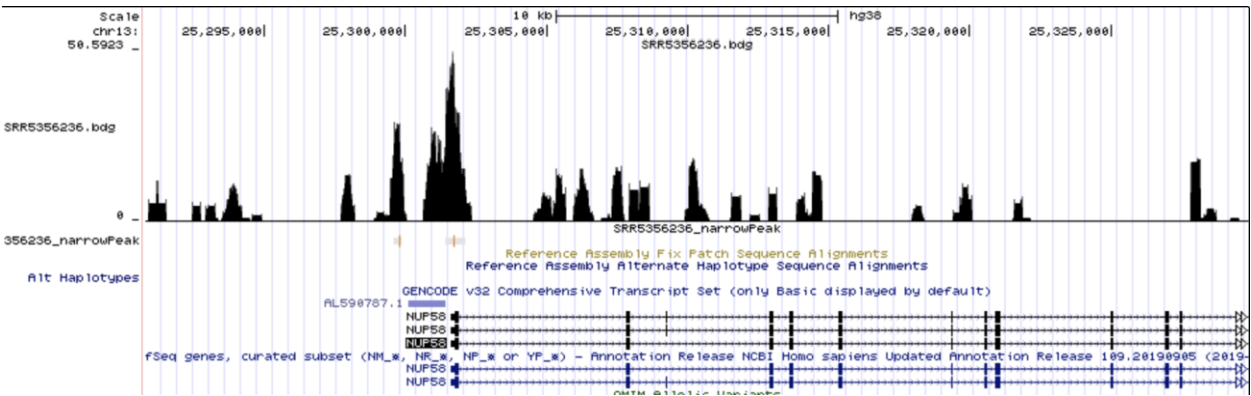
We found several genes that are involved in DNA maintenance processes, many of which function in response to environmental damage, such as radiation, most commonly UV light. For example, the RHNO1 gene (ENST00000623153.1) codes for a protein involved in the DNA damage pathway and is activated in response to exposure to ionizing radiation³. Based on our results, it seems that ZBTB48 may upregulate RHNO1. From our ChIP-seq analysis, we found evidence that ZBTB48 binds to RHNO1 (Figure 2). This might indicate that the ZBTB48 protein acts as a transcription factor for RHNO1. Other proteins that we found that have been identified as radioprotecting include NUP58 (ENST00000381747.7), HNRNPH1 (ENST00000519707.5), and CTNNB1 (ENST00000646725.1). Like RHNO1, these proteins show high expression in the presence of ZBTB48, and ZBTB48 shows some binding activity to their respective genes (Figure 1 and 2). Many of these proteins are involved in the repair of double stranded breaks that result from UV light exposure⁴⁻⁶. Of particular interest to researchers may be the ability of the ZBTB48 protein to regulate CTNNB1 gene, which produces beta catenin, a protein that is involved in the Wnt/beta-catenin signaling pathway. This pathway is important in mediating resistance to radiation and is shown to be associated with radioresistance in many mouse somatic cells⁶. Other proteins that have been observed to be involved in DNA repair pathways is RPL36A (ENST00000459748.1) and APLP2 (ENST00000338167.9)^{7,8}. Like the proteins discussed previously, APLP2 shows high expression when ZBTB48 is present, and appears to be regulated by the protein. However, RPL36A is downregulated in the presence of ZBTB48, indicating some repression by the transcription factor. This is further supported by the binding activity we found by ZBTB48 to the RPL36A gene (Figure

1 and 2). We could not isolate a proper sequence motif after performing the analysis with MEME (Figure 3). However, the query motifs are rich in adenine, which are indicative of dual-coding genes⁹. Dual coding genes code for different proteins in overlapping reading frames¹⁰. As it stands, it is unclear what this could indicate. Alternatively, the enrichment of adenine observed could be due to the TATA box found in promoter regions of many genes¹¹.

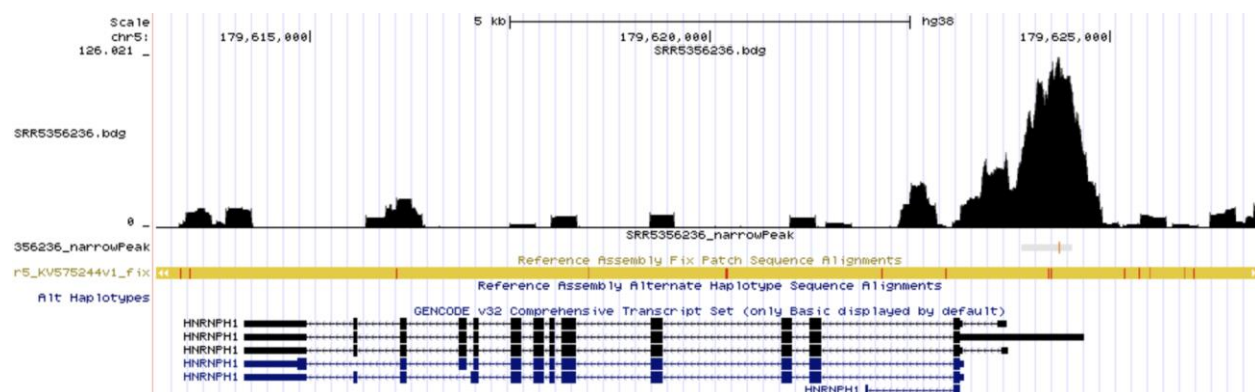
RHNO1



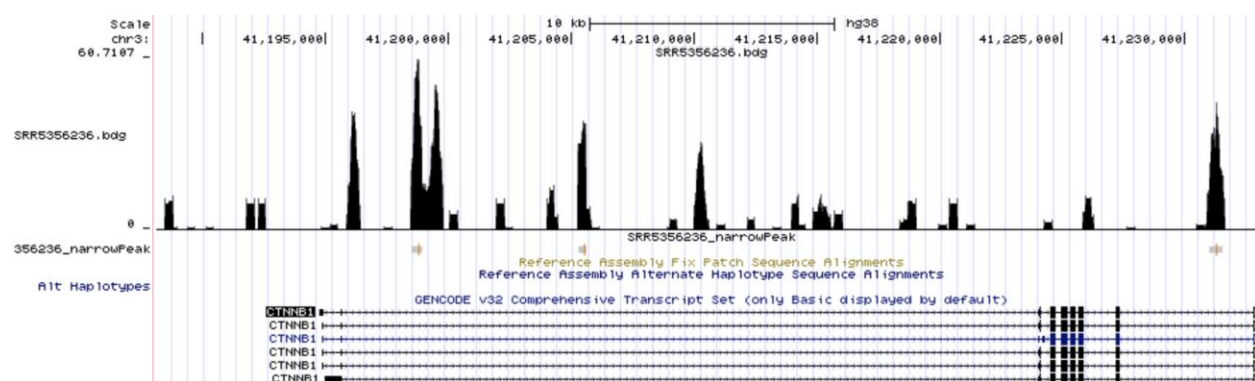
NUP58



HNRNPH1



CTNNB1



RPL36A

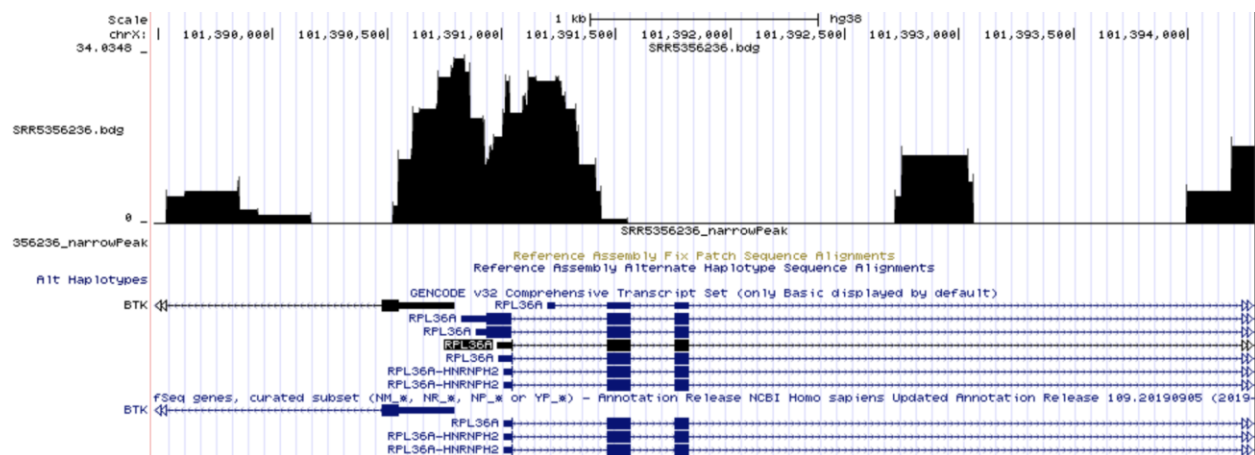


Figure 2. ChIP-Seq for several genes, most of which are involved in environmental DNA damage repair pathways.

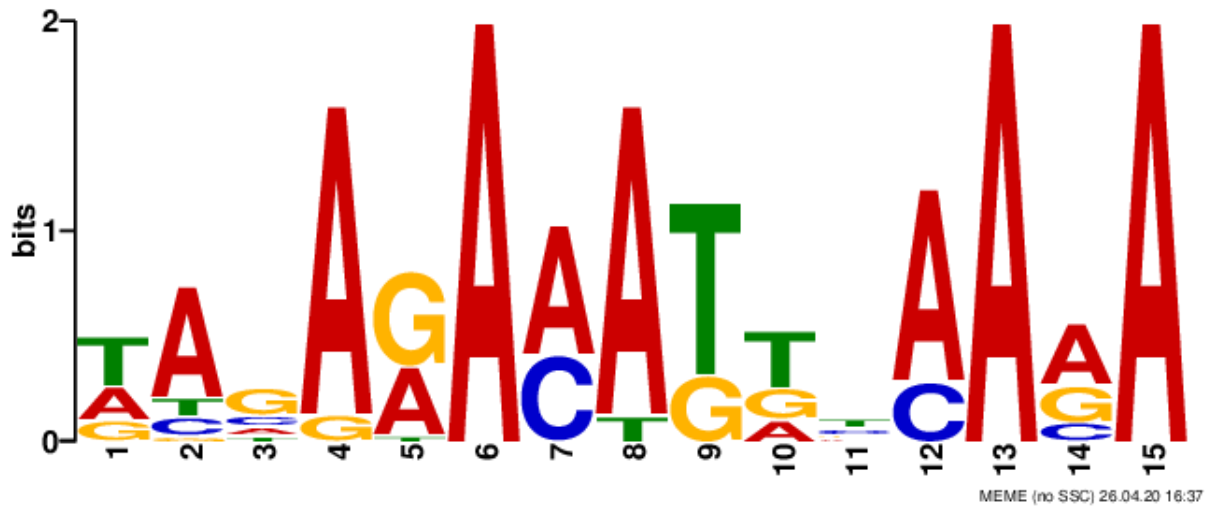


Figure 3. Extracted motif from ChIP-seq peaks. No motif matches were obtained by querying the sequence to a database.

Discussion

Our findings indicate that the ZBTB48 protein is indeed involved in the regulation of DNA repair enzymes. Of particular interest is its involvement in environmental DNA damage repair pathways. We identified several genes involved in the repair of double stranded breaks that result from exposure to UV light or other radiation. We also found evidence of ZBTB48 acting as a transcription factor for many of these genes, which further strengthens our argument that it may play a role in regulating the cell's DNA repair mechanisms. Though we cannot comment on the extent of the regulatory activity of ZBTB48, we hope our findings help encourage new research into this area. We acknowledge the fact that using a cancer cell line such as HeLa could exaggerate the role ZBTB48 plays in DNA repair regulation, since these types of cells suffer from extensive DNA damage. Nonetheless, our findings and the following discussion could be valuable in understanding what classes of proteins could be involved in DNA repair.

The contribution of ZBTB48 to the aging process is not exactly clear, but here we wish to offer some plausible hypotheses. There has been some discussion into how double stranded breaks can contribute to the process of senescence¹². Double stranded breaks can be induced through exposure to ionizing radiation such as ultraviolet light. Once a double stranded break occurs in the cell, a signaling cascade known as the DNA damage response activates and initiates the DNA repair process as well as preventative measures to maintain the proper function and survival of the cell¹³. The majority of double stranded breaks that occur in the cell are repaired, however any remaining genomic lesions could cause a plethora of problems. One genomic lesion that has been studied is the ability of double stranded breaks to fuse with other DNA ends, particularly the ends of chromosomes. This can occur especially during telomeric dysfunction, where the telomeric ends of chromosomes are rather short and do not provide the necessary protection against further DNA damage. This results in the shortened telomeres to fuse with regions affected by double stranded breaks. Aging cells that fail to initiate senescence mechanisms will allow more opportunities for fusion between short telomeres and double stranded

breaks, which can result in genomic instability. Such constant genomic damage can promote the production of cancer cells and other irregularities. Therefore, although genomic instability arises as a consequence of aging, double stranded breaks due to radiation exposure can further compromise genomic stability in older cells¹². We propose that ZBTB48 may aid in diminishing the effects of aging by not only elongating telomeres, but regulating the DNA repair machinery responsible for preventing lesions that result from the fusion of short telomeres and genomic regions suffering from double stranded breaks.

The regulatory effects of ZBTB48 can transcend beyond aging. Understanding how ZBTB48 regulates the DNA repair machinery can be instructive in understanding the progression of disease, especially rare disorders such as Cockayne syndrome and Xeroderma pigmentosum. These disorders are characterized by an abnormal sensitivity to sunlight¹⁴. The photosensitivity of these disorders is due to inefficient nucleotide excision repair mechanisms. The nucleotide excision repair pathway is primarily responsible for fixing thymine dimers, a genomic lesion that results from exposure to ultraviolet light¹⁵. Though the repair enzymes we have found in our analysis primarily focus on repair double stranded breaks, its possible they are involved in other repair processes such as nucleotide excision repair. An open question would be whether or not the products of the genes we identified moonlight as excision repair enzymes. We hope our findings guide researchers to further understand the capabilities of ZBTB48, including its possible involvement in other pathways besides those of DNA repair.

References

1. Da Silva, P. F. L. & Schumacher, B. DNA damage responses in ageing. *Open Biol.* **9**, (2019).
2. Jahn, A. *et al.* ZBTB 48 is both a vertebrate telomere-binding protein and a transcriptional activator . *EMBO Rep.* **18**, 929–946 (2017).
3. Day, M. *et al.* BRCT domains of the DNA damage checkpoint proteins TOPBP1/Rad4 display distinct specificities for phosphopeptide ligands. *Elife* **7**, 1–18 (2018).
4. Rodriguez-Berriguete, G. *et al.* Nucleoporin 54 contributes to homologous recombination repair and post-replicative DNA integrity. *Nucleic Acids Res.* **46**, 7731–7746 (2018).
5. Haley, B., Paunesku, T., Protic, M. & Woloschak, G. E. Response of heterogeneous ribonuclear proteins (hnRNP) to ionising radiation and their involvement in DNA damage repair. *Int. J. Radiat. Biol.* **85**, 643–655 (2009).
6. Dong, Z., Zhou, L., Han, N., Zhang, M. & Lyu, X. Beteiligung des Wnt/ β -Catenin-Signalwegs an der strahleninduzierten Invasion von U87-Glioblastomzellen. *Strahlentherapie und Onkol.* **191**, 672–680 (2015).
7. Provost, E. *et al.* The tumor suppressor rpl36 restrains KRASG12V-induced pancreatic cancer. *Zebrafish* **11**, 551–559 (2014).
8. Stante, M. *et al.* Fe65 is required for Tip60-directed histone H4 acetylation at DNA strand breaks. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 5093–5098 (2009).
9. Abrahams, L. & Hurst, L. D. Adenine enrichment at the fourth CDS residue in bacterial genes is consistent with error proofing for 11 frameshifts. *Mol. Biol. Evol.* **34**, 3064–3080 (2017).
10. Liang, H. Decoding the dual-coding region: Key factors influencing the translational potential of a

two-ORF-containing transcript. *Cell Res.* **20**, 508–509 (2010).

11. Smale, S. T. & Kadonaga, J. T. The RNA Polymerase II Core Promoter. *Annu. Rev. Biochem.* **72**, 449–479 (2003).
12. Hernández, L. *et al.* Aging and radiation: Bad companions. *Aging Cell* **14**, 153–161 (2015).
13. White, R. R. & Vijg, J. Do DNA Double-Strand Breaks Drive Aging? *Mol. Cell* **63**, 729–738 (2016).
14. Rapin, I., Lindenbaum, Y., Dickson, D. W., Kraemer, K. H. & Robbins, J. H. Cockayne syndrome and xeroderma pigmentosum. *Neurology* **55**, 1442–9 (2000).
15. Rastogi, R. P., Richa, Kumar, A., Tyagi, M. B. & Sinha, R. P. Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair. *J. Nucleic Acids* **2010**, (2010).