PLOS GENETICS

RESEARCH ARTICLE

Genome-wide analysis of transcription-coupled repair reveals novel transcription events in *Caenorhabditis elegans*

Cansu Kose[1], Laura A. Lindsey-Boltz[1], Aziz Sancar[1]*, Yuchao Jiang[2,3,4]*

1 Department of Biochemistry and Biophysics, University of North Carolina School of Medicine, Chapel Hill, North Carolina, United States of America, **2** Department of Statistics, College of Arts and Sciences, Texas A&M University, College Station, Texas, United States of America, **3** Department of Biology, College of Arts and Sciences, Texas A&M University, College Station, Texas, United States of America, **4** Department of Biomedical Engineering, College of Engineering, Texas A&M University, College Station, Texas, United States of America

* aziz_sancar@med.unc.edu (AS); yuchaojiang@tamu.edu (YJ)

Citation: Kose C, Lindsey-Boltz LA, Sancar A, Jiang Y (2024) Genome-wide analysis of transcription-coupled repair reveals novel transcription events in *Caenorhabditis elegans*. PLoS Genet 20(7): e1011365. https://doi.org/10.1371/journal.pgen.1011365

Editor: Quanjiang Ji, ShanghaiTech University, CHINA

Received: November 16, 2023

Accepted: July 8, 2024

Published: July 19, 2024

Data Availability Statement: XR-seq and RNA-seq data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database with accession number GSE245181 and GSE262486. All code used in this paper is available at https://github.com/yuchaojiang/damage_repair/tree/master/C_elegans.

Funding: This work was supported by National Institute of health grants R35 GM118102 (A.S.), R01 ES033414 (A.S.) and R35 GM138342 (Y.J.). The funders had no role in study design, data

Abstract

Bulky DNA adducts such as those induced by ultraviolet light are removed from the genomes of multicellular organisms by nucleotide excision repair, which occurs through two distinct mechanisms, global repair, requiring the DNA damage recognition-factor XPC (xeroderma pigmentosum complementation group C), and transcription-coupled repair (TCR), which does not. TCR is initiated when elongating RNA polymerase II encounters DNA damage, and thus analysis of genome-wide excision repair in XPC-mutants only repairing by TCR provides a unique opportunity to map transcription events missed by methods dependent on capturing RNA transcription products and thus limited by their stability and/or modifications (5'-capping or 3'-polyadenylation). Here, we have performed eXcision Repair-sequencing (XR-seq) in the model organism *Caenorhabditis elegans* to generate genome-wide repair maps in a wild-type strain with normal excision repair, a strain lacking TCR (*csb-1*), and a strain that only repairs by TCR (*xpc-1*). Analysis of the intersections between the *xpc-1* XR-seq repair maps with RNA-mapping datasets (RNA-seq, long- and short-capped RNA-seq) reveal previously unrecognized sites of transcription and further enhance our understanding of the genome of this important model organism.

Author summary

Most RNA sequencing technologies are based on capturing transcripts to define the RNA expression profile of cell lines or organisms. These methods suffer from RNA's inherent instability in revealing a complete view of RNA Polymerase II (RNAPII) transcription. Here, we used a *C. elegans xpc-1* mutant which only performs transcription-coupled repair (TCR) to generate a complete RNAPII transcription profile by using the eXcision Repair-sequencing (XR-seq) technology to analyze repair of UV-induced DNA damage. XR-seq captures all the excision products from RNAPII-transcribed regions in the form of 21–28 nucleotide-long DNA oligomers and thus in a sense converts the unstable RNA signal to a

collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

more stable DNA signal enabling the generation of the most complete RNAPII transcript profile to date including heretofore unknown transcripts.

Introduction

Transcription of eukaryotic genomes by RNA polymerase II (RNAPII) produces both protein-coding mRNAs and diverse non-coding RNAs (ncRNAs), including enhancer RNAs (eRNAs), long intergenic non-coding RNAs (lincRNAs), and Piwi-interacting RNAs (piRNAs) [1]. Most ncRNAs are rapidly degraded making them difficult to detect, and therefore, likely to have not been fully mapped [2]. However, proper mapping of transient RNAs is an important first step towards understanding the function of these ncRNAs [3]. Many methods have been developed to capture and sequence RNAPII transcripts including those captured through RNA modifications, such as 3'-polyadenylation (poly(A)), used in RNA-seq [4], and 5'-capping used in capped RNA-seq [5,6]. Incorporation of an RNA size-selection step in the later technique to specifically capture short- or long-capped RNAs of less than 100 nucleotides (nt) or greater than 200 nt, respectively, has been beneficial in identifying different classes of ncRNAs and revealed many novel sites of transcription [5,6]. Here, we describe a unique way of identifying RNAPII transcription, which is independent of capturing the RNA products, but instead, harnesses the mechanistic properties of nucleotide excision repair and a sensitive method for sequencing whole-genome excision repair events called XR-seq, for eXcision Repair-sequencing [7] (**Fig 1A**).

In eukaryotes, nucleotide excision repair removes a wide range of helix-distorting DNA lesions from the genome, including UV-induced cyclobutane pyrimidine dimer (CPDs), by

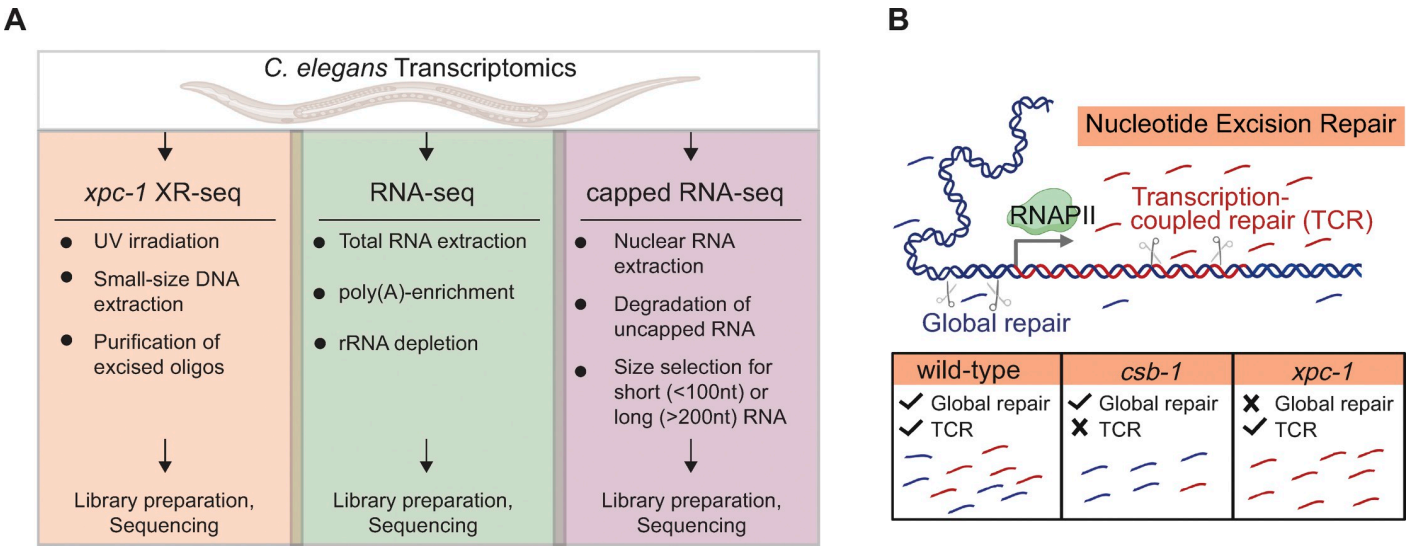

Fig 1. Overview of Study Design. (A) The illustration highlights key properties of the three comparative transcriptomic techniques (XR-seq, RNA-seq, capped RNA-seq) analyzed in this study for their capacity to identify genome-wide transcription in *C. elegans*. **(B)** Nucleotide excision repair removes DNA damage through two different mechanisms: global repair and transcription-coupled repair (TCR). Global repair depends on the XPC protein and occurs throughout the genome, whereas TCR is independent of XPC and only occurs when elongating RNA polymerase II encounters damage during transcription and recruits the CSB protein. This study uses XR-seq to map nucleotide excision repair at single-nucleotide resolution throughout the whole-genome in three strains of *C. elegans*: wild-type, *csb-1*, and *xpc-1*. Because *xpc-1* worms lack global repair, analysis of XR-seq data from this strain provides a unique opportunity to map transcription genome-wide independent of RNA capture. Fig 1 was created with BioRender.com.

https://doi.org/10.1371/journal.pgen.1011365.g001

concerted dual incision of the phosphodiester bonds bracketing the lesion at a somewhat precise distance from the damage (~19 nt 5′ and ~6 nt 3′ to the dimer) to generate excision products that are ~27-nt damage-containing single-stranded DNA oligos [8,9]. Nucleotide excision repair is carried out in most eukaryotes by the six excision repair factors XPA through XPG, originally identified by complementation assays using cells from Xeroderma Pigmentosum (XP) patients, which exhibit a hereditary condition characterized by extreme sun-sensitivity and skin cancer incidence [10]. *Caenorhabditis elegans (C. elegans)* have homologs of all the human XP excision repair factors except for XPE (DDB2) [11]. In addition to these factors, two additional proteins, which are also conserved in *C. elegans* [12], CSA and CSB, were subsequently identified in patients with a related human genetic disorder called Cockayne Syndrome (CS), exhibiting neurodegenerative symptoms and photosensitivity resulting from deficient transcription-coupled repair (TCR), which is defined as excision repair of DNA damage specifically within the transcribed strand of actively transcribed DNA [13]. Nucleotide excision repair occurs by two mechanistically distinct pathways: global repair, that depends on XPA through XPG, and TCR, that depends on these same factors excluding XPC [10]. TCR is initiated when damage in the template strand is encountered by elongating RNAPII, which subsequently recruits CSB, CSA, and additional factors. *C. elegans* have been shown to repair UV-induced DNA damage by both global repair and TCR pathways [11,12,14–17].

We conducted the current study in the *C. elegans* model organism because of its nearly completely annotated nuclear genome, which is approximately 1/30 the size of the human genome, and the availability of a range of omics data. To avoid complications of conducting experiments on mixed populations of whole animals at different developmental stages, many *C. elegans* study designs employ collecting L1-stage larvae state of developmental arrest and uniformly stimulating them into progression upon feeding in order to gather sizable cohorts of animals at a singular developmental phase. *C. elegans* studies of DNA repair have also been performed using L1-stage worms [11,12,14–17], and there are a multitude of available omics data sets examining epigenetic markers, chromatin states, and RNA expression at this stage [5,6,18], so we chose to conduct the current study at this stage as well. We performed XR-seq and RNA-seq in three previously characterized strains of *C. elegans*: wild-type (WT) exhibiting both global repair and TCR, *csb-1*, which only repairs by global repair, and *xpc-1*, which only has TCR (**Fig 1B**). We provide evidence demonstrating the utility of *xpc-1* XR-seq data set for detecting RNAPII transcription and identifying new transcripts. The integration of epigenetic markers, chromatin states, and ncRNA annotations including eRNAs, lincRNAs, and piRNAs all support the robust detection of intergenic RNAPII transcription by *xpc-1* XR-seq. Overall, our results provide a comprehensive view of the transcription-coupled repair landscape in *C. elegans*, highlighting its potential contribution to our understanding of DNA repair mechanisms and non-coding RNA biology.

Results

XR-seq repair maps of UV-induced DNA damage in wild-type, *csb-1*, and *xpc-1* strains of *C. elegans*

The XR-seq next generation sequencing method (S1 Fig) was developed to capture and identify DNA damage-containing excised oligomers to map repair at single-nucleotide resolution throughout the human genome [19]. Recently we modified the method to analyze excision repair of UV-induced CPD photoproducts in *C. elegans* and demonstrated that excision repair in *xpa-1* mutants was near background (that of unirradiated WT worms) [20]. Here we have extended our study to include two additional previously characterized repair-deficient *C. elegans* strains, *xpc-1* and *csb-1* [12,17,21] (see **S1 Table** for detailed sample information). **Fig 2A**

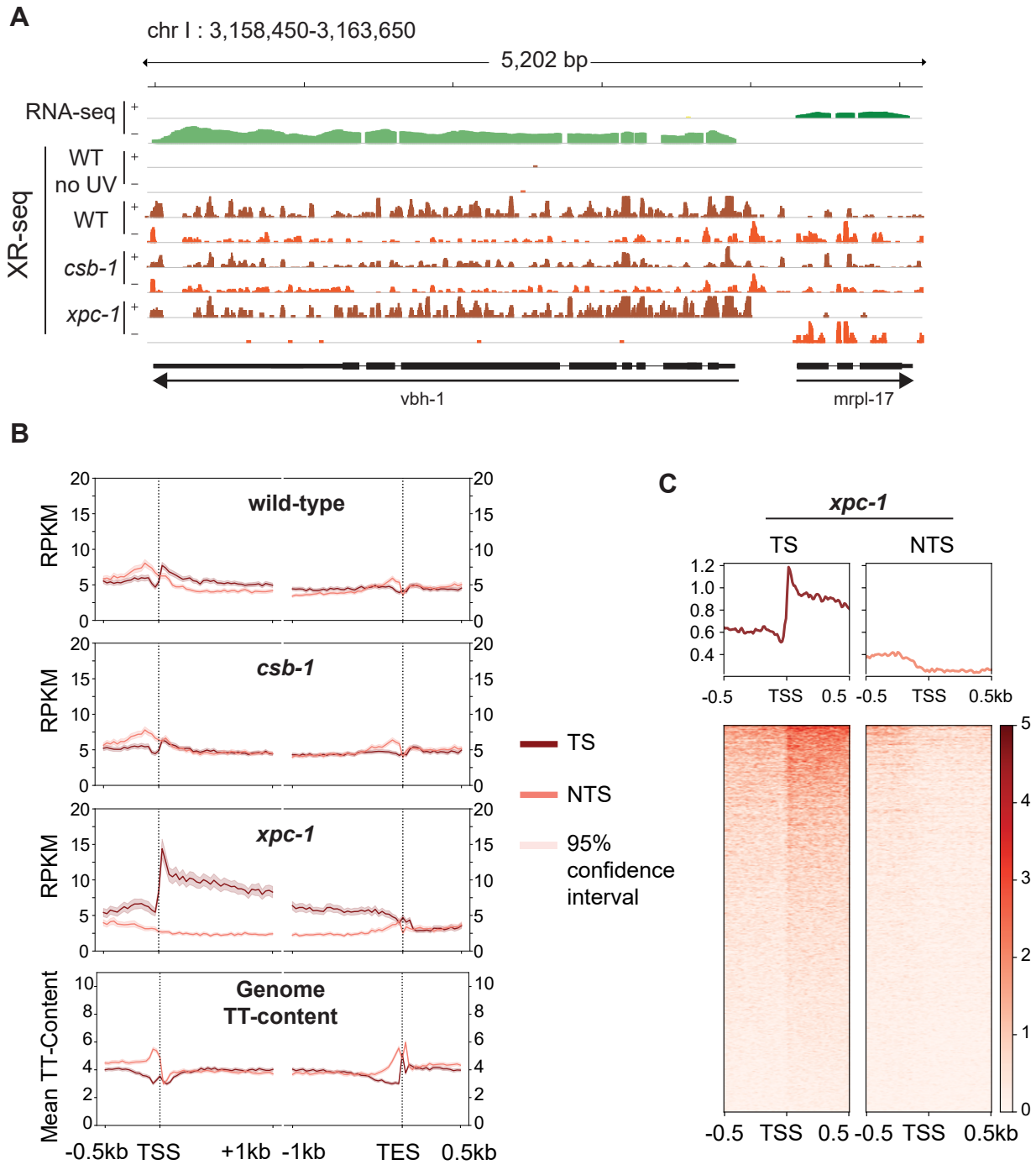

Fig 2. Detection of Transcription-Coupled Repair by XR-seq. (A) Browser view of the distribution of *C. elegans* high throughput sequencing reads separated by strand over a representative 5.2 kb region from chromosome I. RNA-seq reads (green) from wild-type (WT) worms is shown on top to illustrate the opposite direction of transcription of the genes *vbh-1* and *mrpl-17*. The strand distribution of XR-seq reads (orange) 1 hour after UV clearly demonstrates the occurrence of transcription-coupled repair within the body of both genes in WT and *xpc-1* worms but absent in *csb-1*. **(B)** To analyze transcription-coupled repair genome-wide, XR-seq reads on transcribed strand (TS) and non-transcribed strand (NTS) in the indicated strains at 1 hour repair time is plotted with mean reads per kilobase per million mapped reads (RPKM) (y-axis) along the 500 bp upstream and 1 kb downstream of transcription start site (TSS), and 1 kb upstream and 500 bp downstream of transcription end site (TES) (x-axis) for 2,142 genes selected for length > 2 kb and no overlaps with a distance of at least 500 bp between genes. The TT-distribution, as mean TT content (y-axis) was determined for the same gene set and is plotted at the bottom as a measure of expected DNA damage sites. **(C)** Profile plots and heatmaps of TS and NTS XR-seq reads from the *xpc-1* strain at 1 hour repair time spanning the best represented half of 16,588 TSSs > 1 kb apart indicate divergent transcription at promoters.

https://doi.org/10.1371/journal.pgen.1011365.g002

shows a representative Integrative Genomics Viewer (IGV) screenshot of a 5.2 kilobase (kb) region of the genome containing two genes in opposite orientations illustrating levels of transcription as measured by RNA-seq (top two rows) and repair as measured by XR-seq (remaining rows) in WT, *csb-1* and *xpc-1* strains. Reads are mapped to the two strands of the genome as shown, and for the *vbh-1* gene, the transcribed strand (TS) is the + strand and for the *mrpl-17* gene, the TS is the - strand. The results illustrate preferential repair of the TS due to TCR in both WT and *xpc-1*, and there is no strand preference observed in the *csb-1* strain. We used the ratio of read counts from the TS to those from both the TS and non-transcribed strand (NTS) the quantitate TCR genome-wide and the results are shown in **S2 Table**. The percentage of TS/(TS + NTS) for the *vbh-1* and *mrpl-17* genes are, respectively, 78% and 77% in WT, which has both global and TCR; 99% and 94% in *xpc-1*, which only has TCR; and 59% and 43% in *csb-1*, which only has global repair. As previously shown, the unirradiated control (WT no UV) results in extremely low read-numbers (0.003% of UV-irradiated WT) that are not specific [20].

The XR-seq data were then analyzed to assess repair on the TS and NTS of all non-overlapping genes greater than 2 kb in length (**Fig 2B**). Again, such analysis clearly illustrates the presence of TCR in the WT strain (top), which is partially masked due to global repair products mapping to both strands. There was no observed strand difference in repair within gene bodies in the *csb-1* mutant (middle), which lacks TCR. Notably, the differences observed upstream of the transcription start site (TSS) and transcription end site (TES) can be attributed to TT-content of these regions of the genome (bottom) resulting in different levels of DNA damage in these areas since CPDs are primarily formed at TTs and the majority of XR-seq reads contain this dinucleotide ~6 nt from the 3'-end (**S2 and S3A Fig**). The percentage of TTs is constant with gene length and between strands in *C. elegans* (**S4 Fig**). In stark contrast to the *csb-1* mutant, the majority of repair events map to the transcribed strand in the *xpc-1* mutant where TCR is the only functional excision repair pathway (**Fig 2B**). As previously seen in humans and other organisms [19,22–25], XR-seq reads peak at the 5'-end of genes and decrease toward the 3'-end, which is consistent with the proposed TCR model [26], and the skewed pattern gradually diminishes as the repair process proceeds over time (**S3B Fig**). The 5'-peak of repair on the TS is not unique to L1-stage worms as this pattern is also observed in a mixed population of worms (**S3C Fig**).

XR-seq analysis in human XP-C cells revealed pronounced CPD repair on the non-template strand upstream of the TSS due to divergent transcription at promoters [19,25]. The *C. elegans* XR-seq analysis shown in **Fig 2B** does not exhibit this pattern even though anti-sense transcription at promoters has been reported in worms [6]. Therefore, we further analyzed repair in the region of a greater number of TSSs (all TSSs that are at least 1kb apart) at an individual basis as visualized in the heatmaps shown in **Fig 2C** and **S5 Fig**. With this analysis, we were able to detect anti-sense transcription (enrichment on NTS upstream of the TSS) in a subset of genes. The TS upstream of the TSS has much higher read-count relative to the NTS, likely due to extensive transcription of upstream eRNAs, which has been reported to be prevalent in *C. elegans* and occurs in the direction of the downstream gene in 90% of cases [6].

We next sought to identify genes that exhibit significantly differential and dynamic repair using time-series *xpc-1* XR-seq data collected at 5min, 1h, 8h, 16h, 24h, and 48h after UV-irradiation (see Materials and Methods for details). We identified 121 genes exhibiting significant dynamic repair across timepoints (**S6A Fig**) and performed gene ontology (GO) analysis of biological processes (**S6B Fig**) and cellular components (**S6C Fig**). While investigation of gene-specific excision repair has been extensively explored across various model organisms [23,24,27–33] and across different timepoints [16,34,35], our current investigation centers on the domain of intergenic transcription-coupled repair and its juxtaposition with transcriptional events detectable by RNA-seq and capped RNA-seq.

Transcription-coupled repair measured by XR-seq in *xpc-1 C. elegans* serves as an RNA-independent proxy for transcription

Since the *xpc-1* worm mutant lacks global repair, the XR-seq reads from this strain can serve as a unique measure of RNAPII transcription independent of capturing the RNA product. Fig 3A shows an IGV screenshot of a 27 kb region of the genome illustrating levels of transcription as measured by RNA-seq in WT and *xpc-1* strains, long- and short-capped RNA-seq from the WT strain, and XR-seq from the *xpc-1* strain. This representative region shows genes on either side of an intergenic region (defined as a region at least 2 kb away from an annotated gene). RNA-seq reads (top) can be seen in the areas of the annotated RefSeq Genes consistent with polyadenylated protein-coding mRNA transcripts. We do not observe obvious differences in the RNA-seq data from the two different worm strains (WT and *xpc-1*) (Spearman correlation coefficient, r = 0.94). The long-capped RNA-seq reads, which do not require poly(A) for capture, are seen in these same areas of protein-coding transcripts and are also seen in the intergenic region. This is consistent with previous reports demonstrating that this technique is useful for detecting non-coding RNAs [6, 18]. Similarly, short-capped RNA-seq reads have been reported to effectively map areas of transcription initiation, of which there are many in this screenshot. There are *xpc-1* XR-seq reads (bottom) throughout this highly transcribed 27 kb area of the genome, including the intergenic region, which illustrates the potential value of using the data set as an RNA-independent proxy for transcription.

We compared the genome-wide distribution of the reads obtained from the different sequencing methods **(Figs 3B and S7A)**. For this analysis, the genome was systematically divided into three distinct categories: intergenic regions, regions within 2 kb upstream of TSSs, and genic regions. Notably, both *xpc-1* XR-seq and capped RNA-seq techniques reveal a large proportion of transcription events occurring outside of genic regions. This analysis reveals a noteworthy distinction when comparing RNA-seq, capped RNA-seq, and XR-seq. In contrast to RNA-seq, both capped RNA-seq and *xpc-1* XR-seq generate a significantly higher number of reads that map to intergenic regions and regions located within 2 kilobases upstream of TSS. This observation underscores the capability of these methods to capture transcriptional activity in these specific genomic locations. Similarly, our investigations demonstrate a high degree of concordance between genome-wide signals obtained from XR-seq and those derived from short and long-capped RNA-seq. Conversely, there is a near-zero correlation coefficient when comparing RNA-seq to the capped RNA-seq and XR-seq datasets **(S8 Fig)**.

Epigenetic markers and chromatin states validate the intergenic transcription detected by *xpc-1* XR-seq

Expanding our investigation further, we incorporated annotation of chromatin states of *C. elegans*. As illustrated in **Figs 4A** and **S7B**, our analysis of chromatin states has unveiled intriguing distinctions among the different sequencing methods. Notably, when we examine the distribution of chromatin states, RNA-seq appears to predominantly align with 5' proximal regions, gene bodies, and exons. However, it displays relatively lower read counts in categories associated with retrotransposons, pseudogenes, and tissue-specific regions. In stark contrast, both capped RNA-seq and XR-seq exhibit notably similar chromatin state patterns, although some nuanced differences do exist between the two. A closer examination demonstrates that both short-capped RNA-seq and long-capped RNA-seq reveal genic and intergenic transcription, including intergenic enhancers. Short-capped RNA-seq indicates shorter transcripts, corresponding to transcription initiation events and enhancers shorter than 100 base pairs. In contrast, long-capped RNA-seq captures longer transcripts within the nucleus, encompassing both pre-mature and mature RNAs. These longer transcripts relate to transcription elongation,

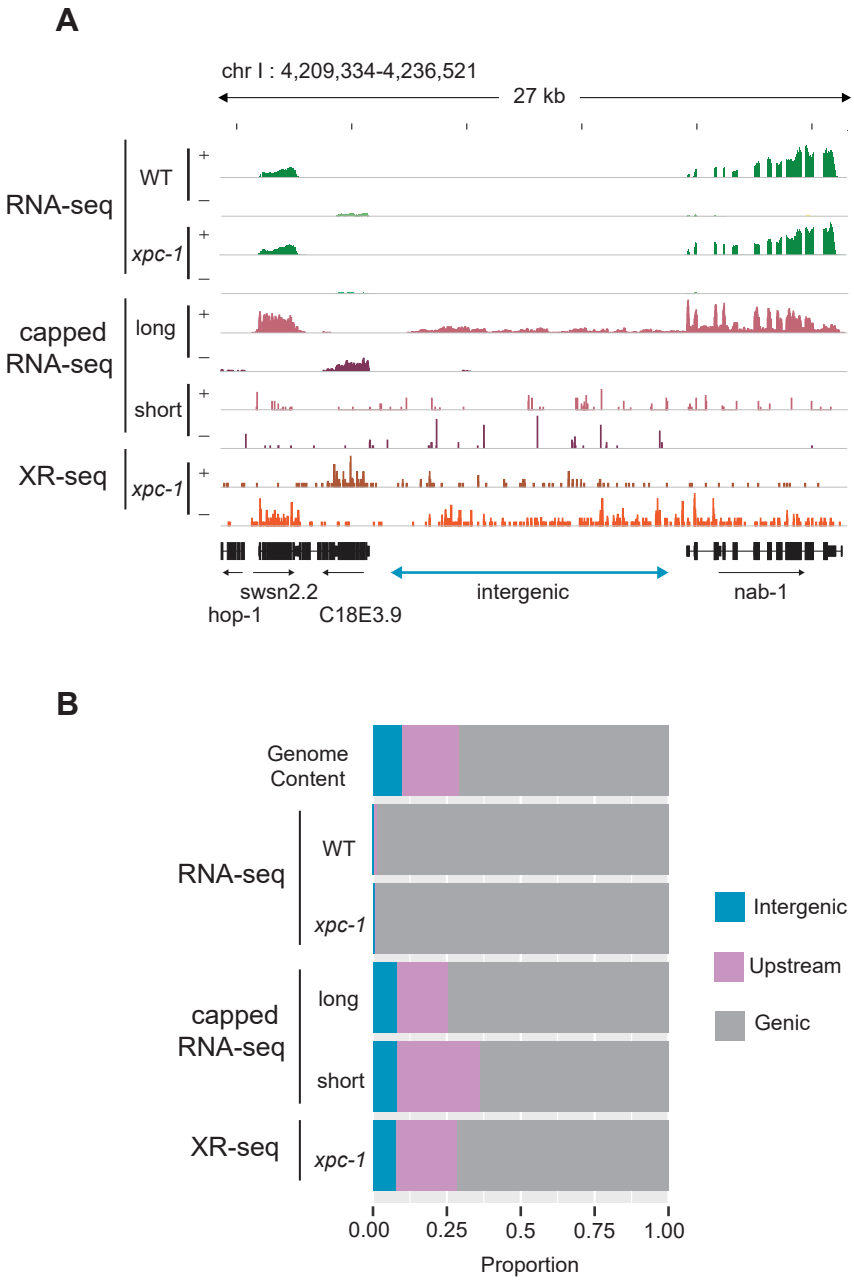

Fig 3. Transcription-Coupled Repair Reveals Transcription in Intergenic Regions. (A) Browser view of stranded read distribution from WT and *xpc-1* RNA-seq (green), capped RNA-seq (pink), and *xpc-1* XR-seq (orange) at 1 hour repair time over a representative 27 kb region from chromosome I. Both capped RNA-seq and XR-seq methods provide comprehensive coverage of the entire window, encompassing both genic and intergenic regions, in contrast to RNA-seq which only captures polyadenylated mRNAs. **(B)** Bar graphs depict the genome-wide distribution of reads obtained from the different sequencing methods, including WT and *xpc-1* RNA-seq, long-capped RNA-seq, short-capped RNA-seq, and *xpc-1* XR-seq at 1 hour repair time. Genome Content is shown on the top as a reference. Notably, both XR-seq and capped RNA-seq techniques reveal transcription events occurring outside of the defined genic regions (see Materials and Methods for details).

https://doi.org/10.1371/journal.pgen.1011365.g003

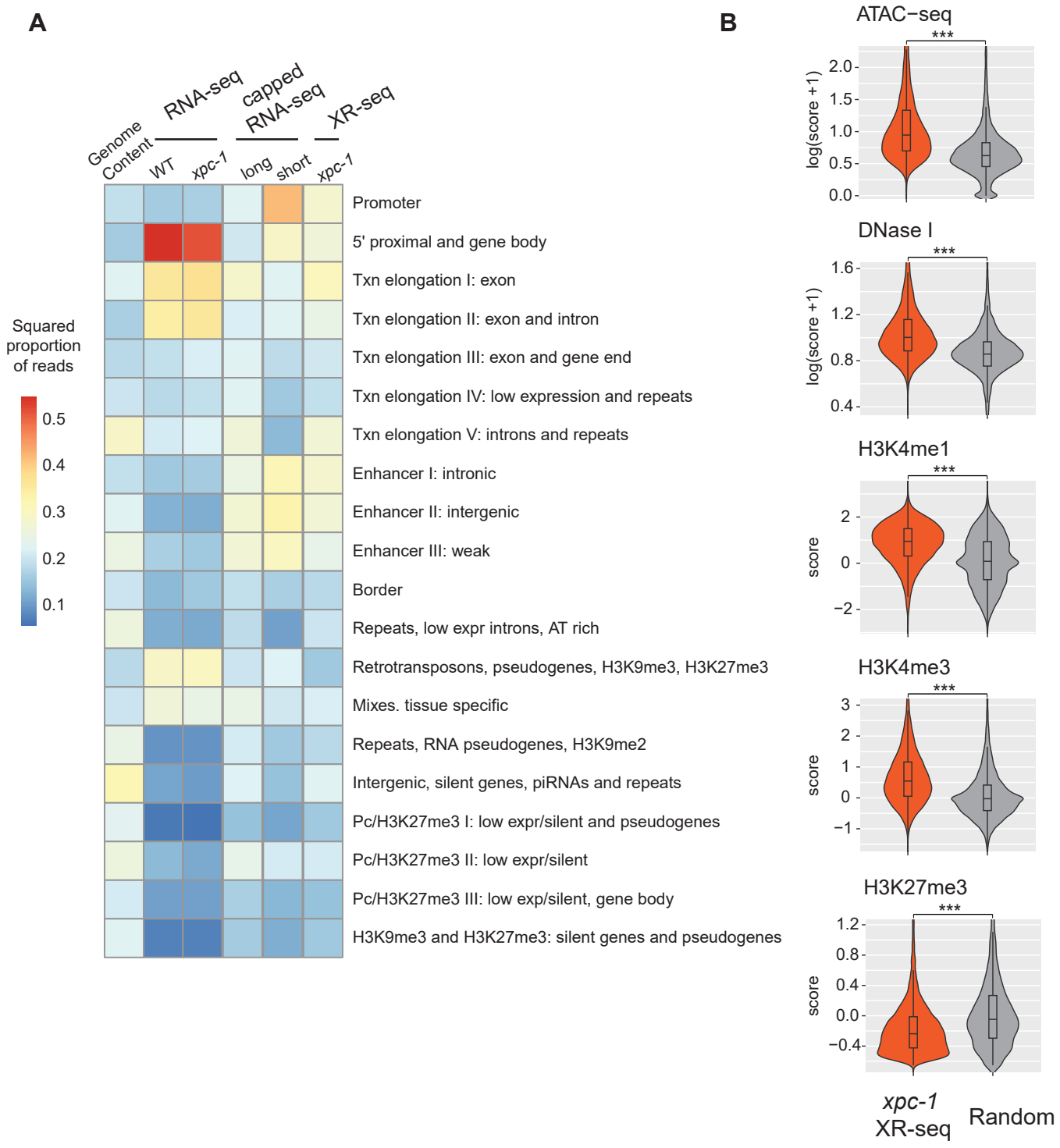

Fig 4. The Transcription-Coupled Repair in Intergenic Regions Detected by *xpc-1* XR-seq is Supported by Epigenomic Signatures. (A) Reads from XR-seq at 1 hour repair time, capped RNA-seq, and RNA-seq were analyzed for overlap with genomic intervals corresponding to 20 distinct predicted chromatin states in *C. elegans*. The proportion of reads was computed for each of the annotated chromatin states and the square root of the proportion is visualized as a heatmap. **(B)** Examination of intergenic XR-seq reads, which are undetectable by RNA-seq, in association with ATAC-seq, DNase-seq, H3K4me1, H3K4me3, and H3K27me3 peaks. XR-seq reads exhibit a strong correlation with active transcription markers, in contrast to the repressive marker H4K27me3, when compared to randomly selected genomic regions. All p-values obtained are highly significant (< 2.2e-16) according to nonparametric Wilcoxon rank sum tests.

https://doi.org/10.1371/journal.pgen.1011365.g004

enhancer regions, and tissue-specific transcription. Furthermore, categories that align with XR-seq encompass a combination of short- and long-capped RNA-seq signals, indicating the concordance between XR-seq and capped RNA-seq in capturing transcriptional events.

In our comprehensive analysis of transcribed intergenic regions specifically identified by *xpc-1* XR-seq (but not detected by RNA-seq), we focused on histone markers and chromatin accessibility (Fig 4B). When compared to randomly selected genomic regions spanning the entire genome, the regions uniquely pinpointed by *xpc-1* XR-seq exhibited distinct epigenomic signatures. Specifically, these regions displayed significantly heightened chromatin accessibility, indicating a more open chromatin structure conducive to transcription. Additionally, we observed increased overlap with histone markers such as H3K4me1 and H3K4me3, typically associated with promoters and enhancers. Conversely, there were less reads overlapping with regions with histone marker H3K27me3, associated with gene repression. These corroborating epigenomic signatures serve as compelling evidence reaffirming the existence of intergenic transcription detected by *xpc-1* XR-seq. Furthermore, they underscore the utility of transcription-coupled repair as a proxy for uncovering previously elusive intergenic transcriptional events within the genome.

Novel intergenic transcription identified with *xpc-1* XR-seq

We next examined the read density of RNA-seq, long- and short-capped RNA-seq, and XR-seq within three classes of annotated intergenic ncRNAs: enhancer RNAs (eRNAs) (**Figs** 5A and S9A), long intergenic non-coding RNAs (lincRNAs) (**Figs** 5B and S9B), and Piwi-interacting RNAs (piRNAs) (**Figs** 5C and S9C). Heatmaps (left) display normalized read counts for the individual annotated intergenic ncRNAs segregated by chromosomes and the bar graphs (right) summarize the log-normalized read counts for the class of ncRNA. Our findings reveal that the RNA-seq method shows limited ability to detect any of these intergenic ncRNA transcripts. This is likely attributed to the lack of poly(A) tailing of ncRNAs, which prevent them from being captured by the conventional RNA-seq technique. Both eRNAs and lincRNAs are very well-represented in the data obtained from *xpc-1* XR-seq and long- and short-capped RNA-seq. Interestingly, read density at piRNAs is high for both long-cap RNA seq and *xpc-1* XR-seq, but not short-capped RNA seq. The findings from the read density analysis of these three major classes of known *C. elegans* intergenic ncRNAs demonstrate the utility of mapping such transcripts with transcription-coupled repair.

To assess all intergenic regions (annotated and unannotated) to determine the degree of coverage and overlap between the three methods, we divided the intergenic regions into 85,418 bins and identified those containing *xpc-1* XR-seq, RNA-seq, or capped RNA-seq reads (**S10 Fig**). The results depicted in the Venn diagram presented in **Fig 6A** show several compelling insights. First, our analysis demonstrates that the transcription-coupled repair in the intergenic regions identified by *xpc-1* XR-seq exhibit similar coverage and remarkable concordance with capped RNA-seq, with both exhibiting ~83% bin-coverage and 80% overlap between the two datasets. Second, as observed with the analyses above, RNA-seq has low coverage in intergenic regions relative to XR-seq and capped RNA-seq. Third, 10% of the bins contain reads unique to *xpc-1* XR-seq. Taken together, these results underscore the sensitivity of transcript-detection by *xpc-1* XR-seq.

We further investigated the location and identity of the 7,903 bins that were only detected in the *xpc-1* XR-seq data set and **Fig 6B** shows a pie chart summarizing the results. Of the *xpc-1* XR-seq-unique bins, 34.7% were annotated (dark blue) and the remaining 65.3% have not been annotated (light blue). Of the bins overlapping the 2,722 annotations, 76.8% of those are annotated as piRNAs (**Fig 6C**) and are primarily found on chromosome IV (**Fig 6D**).

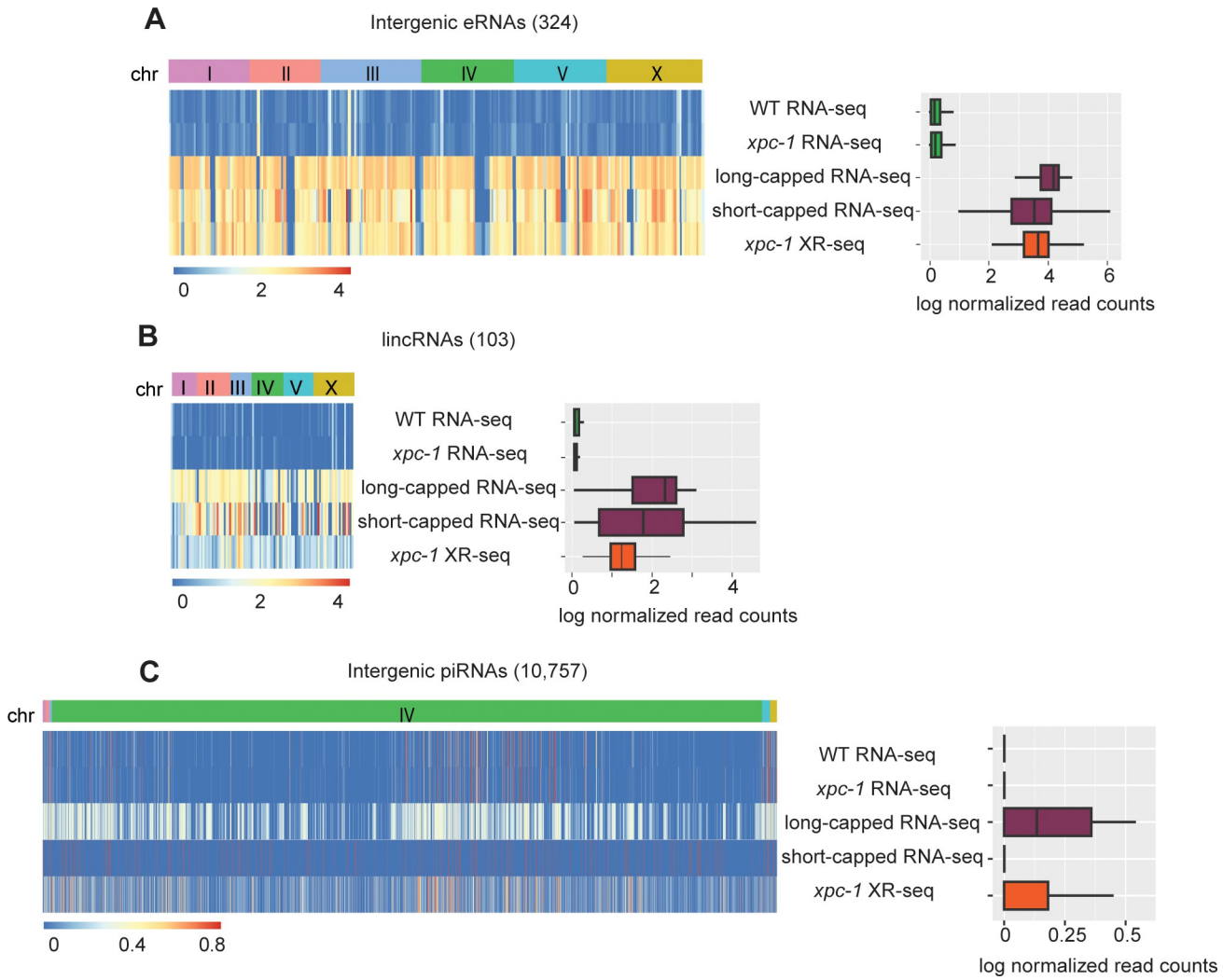

Fig 5. XR-seq Reveals Transcription-Coupled Repair in Intergenic eRNAs, lincRNAs, and piRNAs. (A) Heatmaps (left) display normalized RNA expression and transcription-coupled repair for intergenic enhancer RNAs (eRNAs) segregated by chromosomes. Normalization by log(x+1) was carried out, where x is library-size-adjusted read count. Bar graphs (right) represent log-normalized read counts for eRNA. Data are presented for WT and *xpc-1* RNA-seq, WT long- and short-capped RNA-seq, and time-course combined *xpc-1* XR-seq dataset (5min, 1h, 8h, 16h, 24h, and 48h). **(B, C)** Heatmaps and bar graphs as in A, for long intergenic non-coding RNAs (lincRNAs) and intergenic Piwi-interacting RNAs (piRNAs), respectively.

https://doi.org/10.1371/journal.pgen.1011365.g005

Interestingly, 26% of the unannotated bins also map to chromosome IV (Fig 6C) and we hypothesize that these may be novel piRNAs or piRNA precursors. In summary, the *xpc-1* XR-seq data set is a useful tool for detecting RNAPII transcription and identifying new transcripts in the previously unannotated intergenic regions of *C. elegans*.

Materials and methods

Biological resources

The *C. elegans* wild-type (N2 ancestral), *csb-1* (RB1801) and *xpc-1* (TG2226) strains were obtained from the *Caenorhabditis* Genetics Center and were cultured under standard conditions at room temperature on nematode growth media (NGM) agar plates with *E. coli* strain OP50.

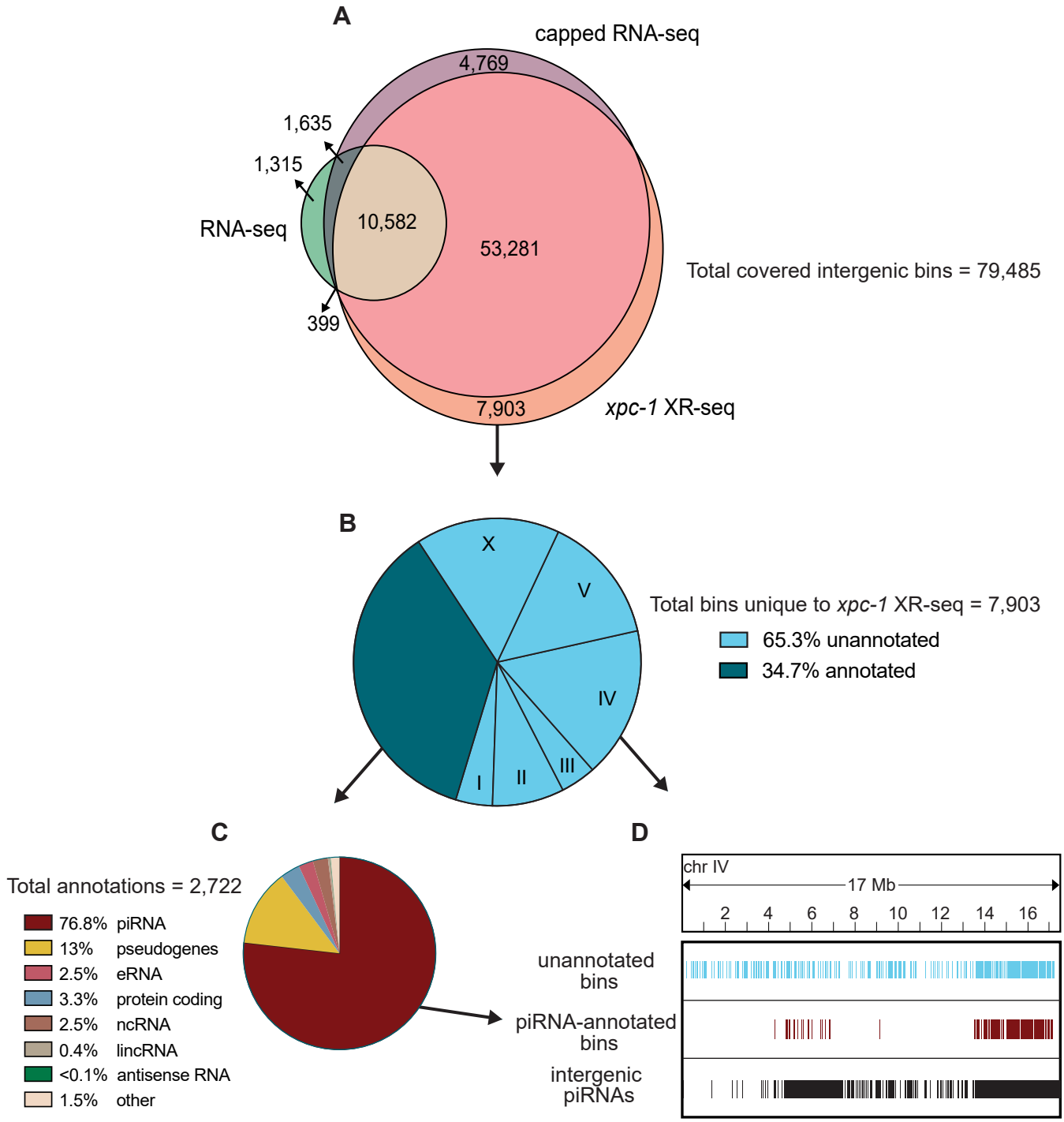

Fig 6. XR-seq identifies intergenic transcription-coupled repair in high concordance with intergenic transcription identified by capped RNA-seq and reveals novel sites of transcription. For 85,418 intergenic bins, we identified regions with non-zero read counts by short- or long-capped RNA-seq, RNA-seq, and time-course (5mins, 1h, 8h, 16h, 24h, and 48h repair times) *xpc-1* XR-seq. **(A)** Venn diagram of intergenic bins detected by capped RNA-seq, conventional RNA-seq, and *xpc-1* XR-seq. To reduce the number of call sets, we required non-zero read counts to be detected: (i) in both replicates for *xpc-1* XR-seq; (ii) in both WT and *xpc-1* RNA-seq, as they are highly correlated; and (iii) by either short-capped or long-capped RNA-seq, as they are complementary. **(B)** Pie chart summary of the 7,903 bins unique to *xpc-1* XR-seq. 34.7% have been annotated (dark blue) and the remaining 65.3% have not been annotated (light blue) according to the WormBase WS282 annotations. The distribution of chromosomal locations (I-X) is indicated for the unannotated bines. The 68% of annotated bins map to chromosome IV which is not indicated. **(C)** Pie chart summary of the bins overlapping 2,722 annotations unique to *xpc-1* XR-seq dataset. The majority of the unique annotated bins contain piRNAs from chromosome IV, with the remainder consisting of pseudogenes, protein coding regions, eRNAs, lincRNAs, and nRNAs. The 'other' category consists of RNAs excluded from the capped RNA-seq dataset (snRNA, tRNAs,

rRNAs) only contains 1.5% of bins. **(D)** Bin distribution along chromosome IV of unique to the *xpc-1* XR-seq dataset-unannotated bins (top in blue), unique to the *xpc-1* XR-seq dataset-bins with piRNA annotations (middle in burgundy), and intergenic piRNAs from WormBase WS282 annotations (bottom in black).

https://doi.org/10.1371/journal.pgen.1011365.g006

XR-seq

To obtain L1 larvae, eggs were collected from adult animals by hypochlorite treatment, and kept in M9 buffer at 22˚C for 16 hours with gentle rotation. Arrested L1 larvae were placed on NGM agar plates with OP50, fed with bacteria for 3 to 4 hours to eliminate the effect of starvation, then exposed to 400 kJ/cm^2 of UVB radiation (313 nm). The worms were collected in M9 buffer at 5 minutes, 1 hour, 8 hours, 16 hours, 24 hours, and 48 hours after irradiation, and washed until the supernatant became clear. Similarly, mixed-stage worms were exposed to 400kj/cm^2 of UVB radiation, then collected 1 hour after UVB. The pelleted *C. elegans* (~50 μl for each) were then incubated for 2 hours at 62˚C with 450 μl of Worm Hirt Lysis Buffer (0.15M Tris pH 8.5, 0.1M NaCl, 5mM EDTA, 1% SDS) and 20 μl of Proteinase K (NEB, cat. no. P8107S). Subsequently, 120 μl of 5M NaCl was added, and the mixture was inverted to ensure proper mixing, followed by an overnight incubation and one hour centrifugation at 4˚C. Supernatants were processed for XR-seq assay as described previously [20]. In brief, supernatants were incubated with 5μL RNase A and then 5μL Proteinase K, purified, and then immunoprecipitated with anti-CPD antibody. Immunoprecipitations were ligated to the adaptors, purified with anti-CPD antibody, and DNA damage was reversed by CPD photolyase. After PCR amplification, the library was sequenced with either Illumina HiSeq 4000 or NextSeq 2000 platforms.

RNA-seq

We followed existing protocol [36] for total RNA extraction in *C. elegans*. Briefly, L1 stage wild-type (WT) and *xpc-1 C. elegans* were collected in M9 and washed until the supernatant was clear, followed by incubation with TRizol and chloroform. After centrifugation at 14,000g for 15min at 4˚C, the aqueous phase was mixed with an equal volume of isopropanol. Following centrifugation, the RNA pellet was washed several times and then resuspended in RNase-free water. Quality control, followed by stranded and poly(A) enriched library preparation and sequencing, was performed by Novogene.

Bioinformatic processing

For XR-seq, cutadapt was used to trim reads with adaptor sequence TGGAATTCTCGGGTGCCAAGGAACTCCAGTNNNNNNACGATCTCGTATGCCGTCTTCTGCTTG at the 3′-end and to discard untrimmed reads [37]. Bowtie 2 was used for read alignment to the ce11 reference genome, followed by filtering, sorting, deduplication, and indexing [38]. Post-alignment filtering steps were adopted using Rsamtools (http://bioconductor.org/packages/Rsamtools). We only kept reads that: (i) have mapping quality greater than 20; (ii) are from chromosome I, II, III, IV, V, and X; and (iii) are of length 19–24 bp. Summary statistics of the XR-seq data that we generated are in **S1 Table**. For RNA-seq, reads were aligned using STAR, followed by a filtering step to remove unmapped reads, reads with unmapped mates, reads that do not pass quality controls, reads that are unpaired, and reads that are not properly paired [39]. We only kept the first read from the mate pair to ensure independent measures. Read counts for each gene were obtained using FeatureCounts [40].

Quality control and data normalization

For gene-specific XR-seq and RNA-seq measurements, we used RPKM for within-sample normalization, since the number of TT dinucleotides are highly correlated with the gene lengths from both the transcribed (TS) and non-transcribed (NTS) strands (**S4 Fig**). To investigate the relationship between gene expression, chromatin states and excision repair, we adopted a stringent quality control (QC) procedure and only retained 26,058 genes that: had at least ten TT dinucleotides in the TS or the and had at least ten reads in total across all XR-seq samples. We observed a robust correlation in repair patterns across the genome between the two replicates collected at each timepoint, underscoring the high reproducibility of our findings (**S11 Fig**). Moreover, pairwise correlation analysis of transcription-coupled repair patterns revealed sample clustering and temporal ordering of samples collected at different time intervals (**S12 Fig**).

To assess excision repair and transcription from non-coding intergenic regions, we generated consecutive and non-overlapping genomic bins of 200 bp long for a total of 501,436 bins. We then removed bins that overlap with annotated genes (gene bodies + 2 kb upstream of the TSS) and those that overlap with blacklist regions in the ce11 genome, resulting in 85,418 bins [41]. For XR-seq, RNA-seq, and short- and long-capped RNA-seq, we adjusted for library size (total number of reads divided by 10^6) for each bin. When times-series XR-seq data were reported in a combined fashion, we took the median repair across all timepoints to get the CPD repair in replicate 1 and replicate 2, respectively.

Repair profiles of TS and NTS

For plotting strand-based average repair profiles of the genes in **Figs** 2A and S3, we used WormBase WS282 genome annotations, and filtered 2,142 genes longer than 2 kilobase (kb) pair, situated at least 500 base pairs (bp) away from neighboring genes. For each gene, the region spanning from 500 bp upstream of the TSS to 1 kb downstream was divided into 50 bins. Similarly, the region from 1 kb upstream to 500 bp downstream of the transcription end site (TES) was also divided into 50 bins, resulting in a total of 100 bins per gene. Bed files of the reads were intersected to the 100 bin-divided-gene list by Bedtools intersect with the following commands -c -wa -F 0.5 -S or -s for TS and NTS, respectively[31]. Summary statistics for TCR, measured by TS/(TS+NTS) are represented in **S2 Table**.

To visualize repair around TSS in **Figs** 2C and S5, we filtered 16,588 TSS from WormBase WS291 annotations, which are at least 1 kb apart from each other. We intersected XR-seq reads over 500 bp downstream and upstream of TSS in a strand specific manner. RPKM normalized bigWig files used to create a matrix with the computeMatrix module of deepTools with the following commands reference-point -b 500 -a 500 –missingDataAsZero, and heatmap generated by plotHeatmap module of deepTools [42].

Identification of dynamic repair using time-course XR-seq data

We next seek to identify genes that exhibit significantly differential and dynamic repair using the time-series XR-seq data of *xpc-1* mutants at 5min, 1h, 8h, 16h, 24h, and 48h in **S6 Fig**. We used Trendy to carry out a breakpoint analysis, allowing for at most two breakpoints and three segments and at least one sample per segment [43]. We used a permutation-based approach with shuffled timepoints to determine the threshold of R^2 (i.e., percentage of total variance that is explained from fitting the time-series model). For the identified significant genes that exhibit dynamic repair across timepoints, we further carried out gene ontology (GO) analysis to identify significantly enriched terms in both biological processes and cellular components [44].

Capped RNA-seq and epigenomic data

Capped RNA-seq captures nuclear RNAs that are with or without poly(A) tails and is thus much more sensitive in detecting non-coding RNAs compared to RNA-seq. We took advantage of short- and long-capped RNA-seq data of wildtype L1 *C. elegans* that are strand-specific [5]. Additionally, we accessed and cross-compared publicly available epigenomic profiles of L1 *C. elegans*, including chromatin accessibility by ATAC-seq, DNase I hypersensitivity by DNase-seq, and histone modifications (H3K4me1, H3K4me3, and H3K27me3) by ChIP-seq [5]. All data were downloaded as processed bigWig files (**S3 Table**) and lifted over to ce11 when necessary. Regions from the bigWig files were overlapped with the genomic bins, and scores from the bigWig files were averaged, weighted by region widths, to yield the capped RNA-seq and epigenetic measurements for each intergenic region.

Chromatin state, eRNA, lincRNA, and piRNA annotations

The genic and intergenic regions of *C. elegans* (ce11) were annotated using the GenomicFeatures R package in conjunction with the TxDb.Celegans.UCSC.ce11.refGene annotation package. Chromatin states in the L3 stage of *C. elegans* were previously inferred, consisting of 20 distinct states as detailed in **Figs 4A** and **S7, and** a high degree of similarity in autosomal chromatin states between the embryonic and L3 larval stages of the worms was reported [45]. observed. This conservation of chromatin configuration allowed us to confidently use the chromatin state data from the L3 stage for intersection with our L1 stage data, without compromising the integrity of our analysis [45]. Each annotated chromatin region was mapped from ce10 to ce11 and intersected with RNA-seq, capped RNA-seq, and XR-seq reads. For eRNAs, 90% of which are bidirectionally transcribed, non-polyadenylated and unspliced, we retrieved 505 annotated eRNAs in *C. elegans* from the eRNAdb database [46,47]. We removed eRNAs that overlap with either annotated genes or blacklist regions, resulting in a total of 324 eRNAs, which are presented in **Figs 5A** and **S9A**. Similarly, we obtained 170 long intergenic non-coding RNAs (lincRNAs) in *C. elegans* from existing annotations [48]. After lifting over the coordinates from ce6 to ce11 and filtering out ones that overlap with genes or blacklist regions, we were left with 103 lincRNAs, which are visualized in the **Figs 5B** and **S9B**. We obtained 15,363 piRNAs in *C. elegans* from existing WormBase WS282 annotations. Removing the piRNAs that overlap with genes or blacklist regions results in 10,757 intergenic piRNAs, which are shown in **Figs 5C** and **S9C**.

Discussion

Transcription-coupled repair appears to be universal in cellular organisms ranging from bacteria to humans and has been studied in several model organisms [10,22,24,49–54]. Multiple methodologies have been developed to unravel the intricate mechanisms and required repair factors [13]. Among these methods, XR-seq, distinguished by its whole-genome analysis at single-nucleotide resolution, has been applied across a spectrum of organisms, including bacteria, yeast, flies, plants, and mammals [13]. A previous study employing a qPCR assay, indicated the existence of transcription-coupled repair in *C. elegans* [16], nevertheless, our study stands as a single-nucleotide-resolution genome-wide UV-damage transcription-coupled repair map of this important model organism. Furthermore, our investigation distinguishes itself by employing transcription-coupled repair as a proxy for RNAPII transcription, and thus *xpc-1* XR-seq data effectively complements RNA-seq and capped RNA-seq datasets to offer a more comprehensive view of transcription.

Leveraging the unique properties of XR-seq data, we aimed to delve into the realm of intergenic transcription, a domain that has posed persistent challenges for conventional RNA-seq

methods. Based on the RNAPII disassociation model in response to UV-induced damage, RNAPII encounters transcription blockage and initiates a process of transcription-coupled repair. During this repair process, RNAPII dissociates from the DNA strand, facilitating the sequential removal of lesions from the template in the 5' to 3' direction. This concerted repair mechanism eventually leads to the clearance of adducts from the template, thereby enabling the synthesis of full-length transcripts [26,55]. To comprehensively investigate these intricate transcription dynamics, we conducted XR-seq at six distinct timepoints, ranging from 5 minutes to 48 hours following UV treatment. As a result, our dataset encompasses both transcription initiation and elongation events, providing a comprehensive view of the entire transcriptional process.

Detection of non-coding RNAs has long been a formidable task due to their relatively low abundance and inherent instability. The development of cutting-edge technologies, such as RNA polymerase II chromatin immunoprecipitation coupled with high-throughput sequencing (RNAPII ChIP-seq), Global Run-On sequencing (GRO-seq), Precision Run-On Sequencing (PRO-seq), and a variety of methods for sequencing the 5'-anchored RNAs, has been driven by the desire to discern nascent RNAs and ncRNAs with heightened precision [5,6,18,56–59]. A comprehensive evaluation of the strengths and limitations of these methods has been described elsewhere [60], and in the context of *C. elegans* research, efforts to specifically target ncRNAs and identify TSS have utilized 5'-capped RNA-sequencing methods, as reported in previous studies [6,18,45,61–63].

XR-seq presents a noteworthy advantage in its ability to directly detect transcription events at the DNA level, thus circumventing the inherent limitations associated with indirect transcription detection techniques such as RNA sequencing. These conventional methods are prone to challenges stemming from the low abundance and instability of RNA molecules. Furthermore, RNA sequencing is susceptible to sequence bias resulting from early transcriptional events that introduce differences between RNA and DNA sequences [64,65]. XR-seq, conversely, by its nature of sequencing transcribed DNA, effectively eliminates this sequence bias, ensuring a more accurate representation of transcriptional activity. An additional advantage of XR-seq is its applicability to prokaryotic organisms, mirroring its utility in eukaryotes, a distinction not shared by nascent RNA sequencing methods.

Our findings demonstrate the efficacy of XR-seq in capturing transcription events within both genic and intergenic regions. While RNA-seq detects only 17.5% of intergenic transcription, our data reveal that up to 80% of the overall intergenic transcription landscape is covered and shared between XR-seq and capped RNA-seq. Notably, XR-seq exhibits sensitivity comparable to that of capped RNA-seq in detecting annotated intergenic enhancer RNAs (eRNAs) and long intergenic non-coding RNAs (lincRNAs), but is superior at detecting intergenic Piwi-interacting RNAs (piRNAs). In *C. elegans*, piRNAs are transcribed from >15,000 discrete genomic loci by RNAPII, resulting in 28-nt short-capped piRNA precursors that play key roles in germline development, genome integrity, and other biological processes [66–69]. The majority of piRNAs are localized to two ~3 Mb cluster regions on chromosome IV [70], and we found 2,090 annotated piRNAs and 1,341 unannotated intergenic regions unique to *xpc-1* XR-seq on this chromosome. We hypothesize that many of these unannotated intergenic regions on chromosome may either be transient piRNA precursor transcripts not captured by other methods or that they are UV-induced piRNAs. Future studies using methods that efficiently capture piRNAs, such as CIP-TAP [18], CAGE [70] or capped RNA-seq [71], after exposing worms to UV could be very informative. In conclusion, our findings provide valuable insights into nascent transcription dynamics and the intricate interplay between transcription-coupled repair and intergenic transcription in *C. elegans*, and this knowledge will be valuable when translated to the human genome and other organisms with large unmapped intergenic content.

Supporting information

S1 Fig. The Excision Repair-sequencing (XR-seq) Method. Excised oligonucleotides isolated by Hirt lysis from lysed worms, purified with anti-CPD specific antibodies, ligated to adapters, and then purified with anti-CPD specific antibodies to remove excess adaptors. The damage was then reversed with CPD photolyase and then PCR was performed to generate libraries for high throughput sequencing. S1 Fig was created with BioRender.com.
(EPS)

S2 Fig. Nucleotide distribution in XR-seq reads of 19–30 nt from wild-type, *csb-1*, and *xpc-1* at 1 hour repair time. The enrichment of TT is observed at a fixed distance, 6 nt from 3' end, indicating that CPD-damage carrying oligonucleotides are successfully represented by XR-seq.
(EPS)

S3 Fig. Nucleotide distribution and genome-wide repair of transcribed strand (TS) and non-transcribed strands (NTS) in *xpc-1* XR-seq. (A) Time-course of *xpc-1* XR-seq, shows TT enrichment 6 nt from 3' end in reads of 24 nt. **(B)** Genome-wide transcribed and non-transcribed strand repair in time-course *xpc-1* XR-seq is plotted as average RPKM (y-axis) along the x-axis 500 bp upstream and 1 kb downstream of transcription start sites (TSS), and 1 kb upstream and 500 bp downstream of transcription end site (TES) for 2,142 genes selected for length $>$ 2 kb and no overlaps with a distance of at least 500 bp between genes. **(C)** XR-seq from *xpc-1* mixed stage worms 1h after UV showing nucleotide distribution and genome-wide transcribed and non-transcribed strand repair as in A and B above.
(EPS)

S4 Fig. TT dinucleotides in the transcribed strand (TS) and non-transcribed strand (NTS). The number of TT dinucleotides in the TS is highly correlated with that in the NTS. Natural logarithms of the numbers of TT dinucleotides from the TS and NTS were computed, respectively. Gene length is a good proxy for the number of TT dinucleotides, and thus we use RPKM for normalization for XR-seq repair read counts.
(EPS)

S5 Fig. Extended Fig 2C showing repair in the TS and NTS around TSSs in the wild-type (L1), *csb-1* (L1) and *xpc-1* (L1 and mixed) XR-seq, 1 h after UV. In wild-type, upstream of TSS has more repair on NTS, in contrast to more repair in TS downstream of TSS. In *csb-1*, despite more repair on the NTS upstream of TSS, TSS downstream repair does not show a strand preference. In *xpc-1*, repair in L1 worms and mixed stage worms exhibit similar profiles, proving that the repair preference in TS at TSS and its immediate downstream is not unique to the L1 worms. Near background repair at NTS versus efficient repair at TS is additional evidence of lacking global repair in *xpc-1*. Although profile plots (top) mask the anti-sense transcription-coupled repair upstream of TSS, a subset of TSSs exhibits upstream TCR on the non-template strand. Genome-wide TT content (right) across the same selected TSSs shows a dip in both strands at TSS. There are more TT dinucleotides on the NTS than TS upstream of TSSs, and therefore more theoretical damages which result in more repair reads in wild-type and *csb-1* in that region.
(TIF)

S6 Fig. Time-course differential-repaired genes. Two XR-seq replicates of *xpc-1* collected at 5min, 1h, 8h, 16h, 24h, and 48h were included in the breakpoint analysis. **(A)** Heatmap of 121 significant genes that show dynamic repair patterns across timepoints. Genes cluster into two clades that exhibit early (112 genes) and late repair (9 genes). Early-repair genes were tested

for gene-ontology enrichment, with significantly enriched terms shown in **(B)** biological process and **(C)** cellular component.
(EPS)

S7 Fig. Genome-wide distribution and epigenetic signatures of the reads from RNA-seq, capped RNA-seq, and XR-seq. (A) Extended Fig 3B bar graphs depicting the genome-wide distribution of reads obtained from various sequencing methods, including wild-type (WT) and *xpc-1* RNA-seq, long-capped RNA-seq, short-capped RNA-seq, and WT, *csb-1*, *xpc-1* XR-seq. Notably, both XR-seq and capped RNA-seq techniques reveal transcription events occurring outside of genic regions. **(B)** Extended Fig 4A heatmap with reads from wild-type (WT) and *xpc-1* RNA-seq, long-capped RNA-seq, short-capped RNA-seq, and WT, *csb-1*, *xpc-1* XR-seq were overlapped with genomic intervals corresponding to 20 distinct chromatin states predicted for the autosomes of *C. elegans*. Proportion of reads were computed for each of the annotated chromatin states; square root of the proportion is visualized as a heatmap.
(EPS)

S8 Fig. Repair events captured by XR-seq are highly correlated with the capped RNA-seq transcription events. XR-seq repair signals correlate with short- and long-capped RNA-seq signals much stronger than conventional RNA-seq. Pairwise smooth scatterplots are shown on the lower triangle, where color corresponds to smoothed data density; Spearman correlation coefficients are shown on the upper triangle, with text size proportionate to the absolute value of the coefficient. Library-size-adjusted read counts from the filtered genomic bins are plotted on the original scale; XR-seq replicates were merged by taking the average, and the 1h time-point for *xpc-1* was used.
(TIF)

S9 Fig. Extended Fig 5 analysis of read numbers in intergenic eRNA, linc RNA, and intergenic piRNAs. (A) Heatmaps (left) display normalized reads for intergenic enhancer RNAs (eRNAs) segregated by chromosomes. Normalization by log(x+1) was carried out, where x is library-size-adjusted read count. Bar graphs (right) represent log-normalized read counts for eRNA. Data are presented for WT and *xpc-1* RNA-seq, WT long- and short-capped RNA-seq, and 2 replicates each of XR-seq from WT no UV, 1 hour after UV in WT and *csb-1*, and *xpc-1* combined time-course (5min, 1h, 8h, 16h, 24h, and 48h). **(B, C)** Heatmaps and bar graphs as in A, for long intergenic non-coding RNAs (lincRNAs) and intergenic Piwi-interacting RNAs (piRNAs), respectively.
(TIF)

S10 Fig. Extended Fig 6A showing intergenic repair in the wild-type (L1), *csb-1* (L1) and *xpc-1* (L1 and mixed) XR-seq, 1 h after UV. For the 85,418 intergenic bins, we identified regions with non-zero read counts by short- or long-capped RNA-seq, RNA-seq, and CPD XR-seq, respectively. **(A)** Upset plot to show the intergenic bins detected by capped RNA-seq, conventional RNA-seq, and XR-seq. To reduce the number of call sets, we required non-zero read counts to be detected: (i) in both replicates for XR-seq; (ii) in both WT and *xpc-1* RNA-seq, as they are highly correlated; and (iii) by either short-capped or long-capped RNA-seq, as they are complementary. **(B)** We used experimental results from short- and long-capped RNA-seq as ground truths and calculated sensitivity, specificity, and F measure (geometric mean of sensitivity and specificity as a joint metric) for the other data types and genotypes. RNA-seq has the lowest sensitivity; *csb-1* XR-seq has the highest sensitivity due to the pervasive global repair detected from intergenic regions, although it also suffers from low specificality.
(EPS)

S11 Fig. High reproducibility between each pair of XR-seq replicates. Normalized gene-specific repair is shown as each dot. Spearman correlation coefficient is shown.
(TIF)

S12 Fig. Pairwise correlation and principal component analysis of XR-seq reads. (A) Spearman correlation coefficient is calculated between each pair of the XR-seq samples using the normalized read counts. Heatmap is generated to visualize the symmetric matrix of correlation coefficient. **(B)** First two principal components from the principal component analysis. Each set of repeats clustered closely, and the WT and *csb-1* sets clustered closer to each other than either *xpc-1* or no UV. Between different *xpc-1* timepoints temporal changes were observed.
(EPS)

S1 Table. XR-seq sample information. Summary of *C. elegans* CPD XR-seq samples across different timepoints and replicates. Total_mapped: total mapped reads. Dedup: deduplicated reads. Mapq: reads with mapping quality > 20 (reads that are equally mapped to multiple genomic locations are removed with this QC). Chr: reads mapped to chrI, II, III, IV, V, X. Qwidth: reads with lengths 19–24. GenebodyPromoter: reads mapped to genes and 2 Kb upstream of transcription start sites (i.e., promoters). Genebody: reads mapped to genes.
(XLSX)

S2 Table. Transcription-coupled repair (TCR) measured by XR-seq. The ratio of read counts from the TS to those from both the TS and NTS serves as a proxy for TCR. *xpc-1* exemplified the strongest TCR, while csb-1 mutants showed depleted TCR as expected. TCR in WT samples was mixed with global repair, while TCR in WT samples without UV treatment reflected background noise.
(XLSX)

S3 Table. Epigenomic and capped RNA-seq data of L1 *C. elegans* adopted in this study.
(XLSX)

Acknowledgments

Figs 1 and S1 were created with BioRender.com. The authors thank Dr. Shawn Ahmed for helpful discussions and comments. Portions of this research were conducted with the advanced computing resources provided by UNC-Chapel Hill and Texas A&M High Performance Research Computing.

Author Contributions

Conceptualization: Cansu Kose, Laura A. Lindsey-Boltz, Aziz Sancar.

Data curation: Cansu Kose, Laura A. Lindsey-Boltz, Yuchao Jiang.

Formal analysis: Cansu Kose, Yuchao Jiang.

Funding acquisition: Aziz Sancar, Yuchao Jiang.

Investigation: Cansu Kose, Laura A. Lindsey-Boltz, Aziz Sancar, Yuchao Jiang.

Methodology: Cansu Kose, Laura A. Lindsey-Boltz, Aziz Sancar, Yuchao Jiang.

Project administration: Aziz Sancar, Yuchao Jiang.

Resources: Aziz Sancar, Yuchao Jiang.

Software: Yuchao Jiang.

Supervision: Laura A. Lindsey-Boltz, Aziz Sancar, Yuchao Jiang.

Validation: Cansu Kose, Laura A. Lindsey-Boltz, Aziz Sancar, Yuchao Jiang.

Visualization: Cansu Kose, Laura A. Lindsey-Boltz, Yuchao Jiang.

Writing – original draft: Cansu Kose, Aziz Sancar, Yuchao Jiang.

Writing – review & editing: Cansu Kose, Laura A. Lindsey-Boltz, Aziz Sancar, Yuchao Jiang.

References

1. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. Cell. 2014; 157(1): 77–94. https://doi.org/10.1016/j.cell.2014.03.008 PMID: 24679528.
2. Schmid M, Jensen TH. Controlling nuclear RNA levels. Nat Rev Genet. 2018; 19(8): 518–29. https://doi.org/10.1038/s41576-018-0013-2 PMID: 29748575.
3. Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen LL, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol. 2023; 24(6): 430–47. https://doi.org/10.1038/s41580-022-00566-8 PMID: 36596869.
4. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10(1): 57–63. https://doi.org/10.1038/nrg2484 PMID: 19015660.
5. Janes J, Dong Y, Schoof M, Serizay J, Appert A, Cerrato C, et al. Chromatin accessibility dynamics across C. elegans development and ageing. Elife. 2018; 7. https://doi.org/10.7554/eLife.37344 PMID: 30362940.
6. Chen RA, Down TA, Stempor P, Chen QB, Egelhofer TA, Hillier LW, et al. The landscape of RNA polymerase II transcription initiation in C. elegans reveals promoter and enhancer architectures. Genome Res. 2013; 23(8): 1339–47. https://doi.org/10.1101/gr.153668.112 PMID: 23550086.
7. Hu J, Li W, Adebali O, Yang Y, Oztas O, Selby CP, et al. Genome-wide mapping of nucleotide excision repair with XR-seq. Nat Protoc. 2019; 14(1): 248–82. https://doi.org/10.1038/s41596-018-0093-7 PMID: 30552409.
8. Reardon JT, Sancar A. Nucleotide excision repair. Prog Nucleic Acid Res Mol Biol. 2005; 79: 183–235. https://doi.org/10.1016/S0079-6603(04)79004-2 PMID: 16096029.
9. Huang JC, Svoboda DL, Reardon JT, Sancar A. Human nucleotide excision nuclease removes thymine dimers from DNA by incising the 22nd phosphodiester bond 5' and the 6th phosphodiester bond 3' to the photodimer. Proc Natl Acad Sci U S A. 1992; 89(8): 3664–8. https://doi.org/10.1073/pnas.89.8.3664 PMID: 1314396.
10. Sancar A. Mechanisms of DNA Repair by Photolyase and Excision Nuclease (Nobel Lecture). Angew Chem Int Ed Engl. 2016; 55(30): 8502–27. https://doi.org/10.1002/anie.201601524 PMID: 27337655.
11. Lans H, Vermeulen W. Nucleotide Excision Repair in Caenorhabditis elegans. Mol Biol Int. 2011; 2011: 542795. https://doi.org/10.4061/2011/542795 PMID: 22091407.
12. Lopes AFC, Bozek K, Herholz M, Trifunovic A, Rieckher M, Schumacher B. A C. elegans model for neurodegeneration in Cockayne syndrome. Nucleic Acids Res. 2020; 48(19): 10973–85. https://doi.org/10.1093/nar/gkaa795 PMID: 33021672.
13. Selby CP, Lindsey-Boltz LA, Li W, Sancar A. Molecular Mechanisms of Transcription-Coupled Repair. Annu Rev Biochem. 2023; 92: 115–44. https://doi.org/10.1146/annurev-biochem-041522-034232 PMID: 37001137.
14. Lans H, Marteijn JA, Schumacher B, Hoeijmakers JH, Jansen G, Vermeulen W. Involvement of global genome repair, transcription coupled repair, and chromatin remodeling in UV DNA damage response changes during development. PLoS Genet. 2010; 6(5): e1000941. https://doi.org/10.1371/journal.pgen.1000941 PMID: 20463888.
15. Boyd WA, Crocker TL, Rodriguez AM, Leung MC, Lehmann DW, Freedman JH, et al. Nucleotide excision repair genes are expressed at low levels and are not detectably inducible in Caenorhabditis elegans somatic tissues, but their function is required for normal adult life after UVC exposure. Mutat Res. 2010; 683(1–2): 57–67. https://doi.org/10.1016/j.mrfmmm.2009.10.008 PMID: 19879883.
16. Meyer JN, Boyd WA, Azzam GA, Haugen AC, Freedman JH, Van Houten B. Decline of nucleotide excision repair capacity in aging Caenorhabditis elegans. Genome Biol. 2007; 8(5): R70. https://doi.org/10.1186/gb-2007-8-5-r70 PMID: 17472752.
17. van der Woude M, Lans H. C. elegans survival assays to discern global and transcription-coupled nucleotide excision repair. STAR Protoc. 2021; 2(2): 100586. https://doi.org/10.1016/j.xpro.2021.100586 PMID: 34151304.

18. Gu W, Lee HC, Chaves D, Youngman EM, Pazour GJ, Conte D Jr., et al. CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. Cell. 2012; 151(7): 1488–500. https://doi.org/10.1016/j.cell.2012.11.023 PMID: 23260138.

19. Hu J, Adar S, Selby CP, Lieb JD, Sancar A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. Genes Dev. 2015; 29(9): 948–60. https://doi.org/10.1101/gad.261271.115 PMID: 25934506.

20. Kose C, Cao X, Dewey EB, Malkoc M, Adebali O, Sekelsky J, et al. Cross-species investigation into the requirement of XPA for nucleotide excision repair. Nucleic Acids Res. 2023. https://doi.org/10.1093/nar/gkad1104 PMID: 37994737.

21. Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, et al. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. Genome Res. 2014; 24(10): 1624–36. https://doi.org/10.1101/gr.175547.114 PMID: 25030888.

22. Akkose U, Kaya VO, Lindsey-Boltz L, Karagoz Z, Brown AD, Larsen PA, et al. Comparative analyses of two primate species diverged by more than 60 million years show different rates but similar distribution of genome-wide UV repair events. BMC Genomics. 2021; 22(1): 600. https://doi.org/10.1186/s12864-021-07898-3 PMID: 34362292.

23. Oztas O, Selby CP, Sancar A, Adebali O. Genome-wide excision repair in Arabidopsis is coupled to transcription and reflects circadian gene expression patterns. Nat Commun. 2018; 9(1): 1503. https://doi.org/10.1038/s41467-018-03922-5 PMID: 29666379.

24. Deger N, Yang Y, Lindsey-Boltz LA, Sancar A, Selby CP. Drosophila, which lacks canonical transcription-coupled repair proteins, performs transcription-coupled repair. J Biol Chem. 2019; 294(48): 18092–8. https://doi.org/10.1074/jbc.AC119.011448 PMID: 31624146.

25. Lindsey-Boltz LA, Yang Y, Kose C, Deger N, Eynullazada K, Kawara H, et al. Nucleotide excision repair in Human cell lines lacking both XPC and CSB proteins. Nucleic Acids Res. 2023; 51(12): 6238–45. https://doi.org/10.1093/nar/gkad334 PMID: 37144462.

26. Chiou YY, Hu J, Sancar A, Selby CP. RNA polymerase II is released from the DNA template during transcription-coupled repair in mammalian cells. J Biol Chem. 2018; 293(7): 2476–86. https://doi.org/10.1074/jbc.RA117.000971 PMID: 29282293.

27. Deger N, Cao X, Selby CP, Gulec S, Kawara H, Dewey EB, et al. CSB-independent, XPC-dependent transcription-coupled repair in Drosophila. Proc Natl Acad Sci U S A. 2022; 119(9). https://doi.org/10.1073/pnas.2123163119 PMID: 35217627.

28. Adebali O, Yang Y, Neupane P, Dike NI, Boltz JL, Kose C, et al. The Mfd protein is the transcription-repair coupling factor (TRCF) in Mycobacterium smegmatis. J Biol Chem. 2023; 299(3): 103009. https://doi.org/10.1016/j.jbc.2023.103009 PMID: 36775124.

29. Yimit A, Adebali O, Sancar A, Jiang Y. Differential damage and repair of DNA-adducts induced by anticancer drug cisplatin across mouse organs. Nat Commun. 2019; 10(1): 309. https://doi.org/10.1038/s41467-019-08290-2 PMID: 30659176.

30. Yang Y, Lindsey-Boltz LA, Vaughn CM, Selby CP, Cao X, Liu Z, et al. Circadian clock, carcinogenesis, chronochemotherapy connections. J Biol Chem. 2021; 297(3): 101068. https://doi.org/10.1016/j.jbc.2021.101068 PMID: 34375638.

31. Yang Y, Liu Z, Selby CP, Sancar A. Long-term, genome-wide kinetic analysis of the effect of the circadian clock and transcription on the repair of cisplatin-DNA adducts in the mouse liver. J Biol Chem. 2019; 294(32): 11960–8. https://doi.org/10.1074/jbc.RA119.009579 PMID: 31217280.

32. Yang Y, Adebali O, Wu G, Selby CP, Chiou YY, Rashid N, et al. Cisplatin-DNA adduct repair of transcribed genes is controlled by two circadian programs in mouse tissues. Proc Natl Acad Sci U S A. 2018; 115(21): E4777–E85. https://doi.org/10.1073/pnas.1804493115 PMID: 29735688.

33. Yang Y, Hu J, Selby CP, Li W, Yimit A, Jiang Y, et al. Single-nucleotide resolution analysis of nucleotide excision repair of ribosomal DNA in humans and mice. J Biol Chem. 2019; 294(1): 210–7. https://doi.org/10.1074/jbc.RA118.006121 PMID: 30413533.

34. Li W, Liu W, Kakoki A, Wang R, Adebali O, Jiang Y, et al. Nucleotide excision repair capacity increases during differentiation of human embryonic carcinoma cells into neurons and muscle cells. J Biol Chem. 2019; 294(15): 5914–22. https://doi.org/10.1074/jbc.RA119.007861 PMID: 30808711.

35. Jiang Y, Li W, Lindsey-Boltz LA, Yang Y, Li Y, Sancar A. Super hotspots and super coldspots in the repair of UV-induced DNA damage in the human genome. J Biol Chem. 2021; 296: 100581. https://doi.org/10.1016/j.jbc.2021.100581 PMID: 33771559.

36. Green MR, Sambrook J. Total RNA Extraction from Caenorhabditis elegans. Cold Spring Harb Protoc. 2020; 2020(9): 101683. https://doi.org/10.1101/pdb.prot101683 PMID: 32873731.

37. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011; 17(1): 10–2. https://doi.org/10.14806/ej.17.1.200

38. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9: 357–9. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

39. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29(1): 15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886.

40. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30(7): 923–30. https://doi.org/10.1093/bioinformatics/btt656 PMID: 24227677.

41. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep. 2019; 9(1): 9354. https://doi.org/10.1038/s41598-019-45839-z PMID: 31249361.

42. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014; 42(Web Server issue): W187–91. https://doi.org/10.1093/nar/gku365 PMID: 24799436.

43. Bacher R, Leng N, Chu LF, Ni Z, Thomson JA, Kendziorski C, et al. Trendy: segmented regression analysis of expression dynamics in high-throughput ordered profiling experiments. BMC Bioinformatics. 2018; 19(1): 380. https://doi.org/10.1186/s12859-018-2405-x PMID: 30326833.

44. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25(1): 25–9. https://doi.org/10.1038/75556 PMID: 10802651.

45. Evans KJ, Huang N, Stempor P, Chesney MA, Down TA, Ahringer J. Stable Caenorhabditis elegans chromatin domains separate broadly expressed and developmentally regulated genes. Proc Natl Acad Sci U S A. 2016; 113(45): E7020–E9. https://doi.org/10.1073/pnas.1608162113 PMID: 27791097.

46. Jin W, Jiang G, Yang Y, Yang J, Yang W, Wang D, et al. Animal-eRNAdb: a comprehensive animal enhancer RNA database. Nucleic Acids Res. 2022; 50(D1): D46–D53. https://doi.org/10.1093/nar/gkab832 PMID: 34551433.

47. Sartorelli V, Lauberth SM. Enhancer RNAs are an important regulatory layer of the epigenome. Nat Struct Mol Biol. 2020; 27(6): 521–8. https://doi.org/10.1038/s41594-020-0446-0 PMID: 32514177.

48. Nam JW, Bartel DP. Long noncoding RNAs in C. elegans. Genome Res. 2012; 22(12): 2529–40. https://doi.org/10.1101/gr.140475.112 PMID: 22707570.

49. Ogrunc M, Becker DF, Ragsdale SW, Sancar A. Nucleotide excision repair in the third kingdom. J Bacteriol. 1998; 180(21): 5796–8. https://doi.org/10.1128/JB.180.21.5796-5798.1998 PMID: 9791138.

50. Li W, Adebali O, Yang Y, Selby CP, Sancar A. Single-nucleotide resolution dynamic repair maps of UV damage in Saccharomyces cerevisiae genome. Proc Natl Acad Sci U S A. 2018; 115(15): E3408–E15. https://doi.org/10.1073/pnas.1801687115 PMID: 29581276.

51. Selby CP, Lindsey-Boltz LA, Yang Y, Sancar A. Mycobacteria excise DNA damage in 12- or 13-nucleotide-long oligomers by prokaryotic-type dual incisions and performs transcription-coupled repair. J Biol Chem. 2020; 295(50): 17374–80. https://doi.org/10.1074/jbc.AC120.016325 PMID: 33087442.

52. Canturk F, Karaman M, Selby CP, Kemp MG, Kulaksiz-Erkmen G, Hu J, et al. Nucleotide excision repair by dual incisions in plants. Proc Natl Acad Sci U S A. 2016; 113(17): 4706–10. https://doi.org/10.1073/pnas.1604097113 PMID: 27071131.

53. Mellon I, Spivak G, Hanawalt PC. Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene. Cell. 1987; 51(2): 241–9. https://doi.org/10.1016/0092-8674(87)90151-6 PMID: 3664636.

54. Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. Nat Rev Mol Cell Biol. 2008; 9(12): 958–70. https://doi.org/10.1038/nrm2549 PMID: 19023283.

55. Nicholson C, Odom D, Aitken S, Taylor M. DNA lesion bypass and the stochastic dynamics of transcription coupled repair. Proc Natl Acad Sci U S A. 2024; In Press. https://doi.org/10.1073/pnas.2403871121 PMID: 38717857

56. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet. 2014; 46(12): 1311–20. https://doi.org/10.1038/ng.3142 PMID: 25383968.

57. Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc. 2016; 11(8): 1455–76. https://doi.org/10.1038/nprot.2016.086 PMID: 27442863.

58. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. 2010; 8(5): e1000384. https://doi.org/10.1371/journal.pbio.1000384 PMID: 20485488.

59. Morioka MS, Kawaji H, Nishiyori-Sueki H, Murata M, Kojima-Ishiyama M, Carninci P, et al. Cap Analysis of Gene Expression (CAGE): A Quantitative and Genome-Wide Assay of Transcription Start Sites. Methods Mol Biol. 2020; 2120: 277–301. https://doi.org/10.1007/978-1-0716-0327-7_20 PMID: 32124327.

60. Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat Rev Genet. 2016; 17(4): 207–23. https://doi.org/10.1038/nrg.2016.4 PMID: 26948815.

61. Cecere G, Hoersch S, O'Keeffe S, Sachidanandam R, Grishok A. Global effects of the CSR-1 RNA interference pathway on the transcriptional landscape. Nat Struct Mol Biol. 2014; 21(4): 358–65. https://doi.org/10.1038/nsmb.2801 PMID: 24681887.

62. Cecere G, Hoersch S, Jensen MB, Dixit S, Grishok A. The ZFP-1(AF10)/DOT-1 complex opposes H2B ubiquitination to reduce Pol II transcription. Mol Cell. 2013; 50(6): 894–907. https://doi.org/10.1016/j.molcel.2013.06.002 PMID: 23806335.

63. Saito TL, Hashimoto S, Gu SG, Morton JJ, Stadler M, Blumenthal T, et al. The transcription start site landscape of C. elegans. Genome Res. 2013; 23(8): 1348–61. https://doi.org/10.1101/gr.151571.112 PMID: 23636945.

64. Wang IX, Core LJ, Kwak H, Brady L, Bruzel A, McDaniel L, et al. RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase II. Cell Rep. 2014; 6(5): 906–15. https://doi.org/10.1016/j.celrep.2014.01.037 PMID: 24561252.

65. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, et al. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011; 333(6038): 53–8. https://doi.org/10.1126/science.1207018 PMID: 21596952.

66. McEnany J, Meir Y, Wingreen NS. piRNAs of Caenorhabditis elegans broadly silence nonself sequences through functionally random targeting. Nucleic Acids Res. 2022; 50(3): 1416–29. https://doi.org/10.1093/nar/gkab1290 PMID: 35037068.

67. Price IF, Wagner JA, Pastore B, Hertz HL, Tang W. C. elegans germ granules sculpt both germline and somatic RNAome. Nat Commun. 2023; 14(1): 5965. https://doi.org/10.1038/s41467-023-41556-4 PMID: 37749091.

68. Sturm A, Saskoi E, Hotzi B, Tarnoci A, Barna J, Bodnar F, et al. Downregulation of transposable elements extends lifespan in Caenorhabditis elegans. Nat Commun. 2023; 14(1): 5278. https://doi.org/10.1038/s41467-023-40957-9 PMID: 37644049.

69. Huang X, Wang C, Zhang T, Li R, Chen L, Leung KL, et al. PIWI-interacting RNA expression regulates pathogenesis in a Caenorhabditis elegans model of Lewy body disease. Nat Commun. 2023; 14(1): 6137. https://doi.org/10.1038/s41467-023-41881-8 PMID: 37783675.

70. Beltran T, Pahita E, Ghosh S, Lenhard B, Sarkies P. Integrator is recruited to promoter-proximally paused RNA Pol II to generate Caenorhabditis elegans piRNA precursors. EMBO J. 2021; 40(5): e105564. https://doi.org/10.15252/embj.2020105564 PMID: 33340372.

71. Weick EM, Sarkies P, Silva N, Chen RA, Moss SM, Cording AC, et al. PRDE-1 is a nuclear factor essential for the biogenesis of Ruby motif-dependent piRNAs in C. elegans. Genes Dev. 2014; 28(7): 783–96. https://doi.org/10.1101/gad.238105.114 PMID: 24696457.