Genome-wide transcription and repair maps of *Caenorhabditis elegans*

Cansu Kose[1], Aziz Sancar[1,*], Yuchao Jiang[2,3,4,*]

1. Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.
2. Department of Statistics, College of Arts and Sciences, Texas A&M University, College Station, TX 77843, USA.
3. Department of Biology, College of Arts and Sciences, Texas A&M University, College Station, TX 77843, USA.
4. Department of Biomedical Engineering, College of Engineering, Texas A&M University, College Station, TX 77843, USA.

*. To whom correspondence should be addressed: aziz_sancar@med.unc.edu, yuchaojiang@tamu.edu.

ABSTRACT

We have adapted the eXcision Repair-sequencing (XR-seq) method to generate single-nucleotide resolution dynamic repair maps of UV-induced cyclobutane pyrimidine dimers and (6-4) pyrimidine-pyrimidone photoproducts in the *Caenorhabditis elegans (C. elegans)* genome. We focus on the *C. elegans* ortholog of the human XPC-deficient strain (*xpc-1*) and its exclusive use of transcription-coupled repair. We provide evidence demonstrating the utility of *xpc-1* XR-seq as a remarkable tool for detecting nascent transcription and identifying new transcripts. The integration of epigenetic markers, chromatin states, enhancer RNA and long intergenic non-coding RNA annotations supports the robust detection of intergenic nascent transcription by XR-seq. Overall, our results provide a comprehensive view of the transcription-coupled repair landscape in *C. elegans*, highlighting its potential contributions to our understanding of DNA repair mechanisms and non-coding RNA biology.

INTRODUCTION

Genome integrity is a fundamental requirement for the maintenance of life. Organisms have evolved intricate mechanisms to ensure the fidelity of their genetic material[1]. One such mechanism, nucleotide excision repair, is responsible for repairing DNA lesions that distort the DNA helix, including those caused by exposure to ultraviolet (UV) radiation[2]. The solar energy in UV light can induce the formation of DNA lesions such as cyclobutane pyrimidine dimers (CPDs) and 6–4 pyrimidine-pyrimidone photoproducts ((6-4)PPs) between adjacent pyrimidine bases[3]. These aberrant DNA structures disrupt normal cellular processes, necessitating their removal.

Nucleotide excision repair operates by precisely excising damaged DNA bases through a dual incision process, creating single-stranded, damage-containing oligonucleotides. The length of these oligonucleotides varies between prokaryotes (12-13 nucleotides) and eukaryotes (24-32 nucleotides)[4,5]. In humans, the recognition of DNA damage occurs through two pathways of nucleotide excision repair: global repair and transcription-coupled repair[6]. In the global repair pathway, damage is recognized by cooperative interactions of XPC, RPA, and XPA, followed by kinetic proofreading by TFIIH to achieve high specificity[7,8]. In the transcription-coupled repair pathway, these same factors except for XPC are required, and the stalling of RNA polymerase II (Pol II) at damaged sites triggers repair, aided by CSB and CSA proteins[9]. Subsequent processes in both pathways involve the recruitment of XPG and XPF endonucleases. Excised oligonucleotides are approximately 25-30 nucleotides in length and carry the damage at 6-7 nucleotide from 3' end[10,11]. Repair is then completed through gap filling and ligation[12].

The nematode *Caenorhabditis elegans* (*C. elegans*), with its relatively small, fully sequenced genome and conservation of major cellular events with humans, serves as a valuable model organism in the field of DNA repair. Studies have demonstrated that *C. elegans* employs both global and transcription-coupled repair mechanisms, mirroring the repair processes found in humans[13–15]. To enhance our understanding of these repair mechanisms, we have adapted the eXcision Repair Sequencing (XR-seq) method to *C. elegans*.

XR-seq offers a powerful tool for mapping repair events with single-nucleotide precision[3]. In this study, we focus on the *C. elegans* ortholog of the human XPC-deficient strain (*xpc-1*) and its exclusive use of transcription-coupled repair. We provide evidence demonstrating the utility of *xpc-1* XR-seq as a remarkable tool for detecting nascent transcription and identifying new transcripts. Our results reveal that a substantial portion of repair reads aligned to intergenic regions in XR-seq exhibit significant overlap with reads from short- and long-capped RNA sequencing (RNA-seq), far surpassing the capabilities of the polyadenylated RNA-seq[16]. Furthermore, the integration of epigenetic markers, chromatin states, enhancer RNA (eRNA) and long intergenic non-coding RNA (lincRNA) annotations supports the robust detection of intergenic nascent transcription by XR-seq[16–19]. In this article, we provide comprehensive results, which shed light on the transcription-coupled repair landscape in *C. elegans* and its relevance to intergenic transcription. Finally, we discuss the implications of our findings and their potential contributions to our understanding of DNA repair mechanisms and non-coding RNA biology.

RESULTS

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

We employed XR-seq to evaluate genome-wide excision repair dynamics in *xpc-1 C. elegans* at distinct time points following UV exposure, specifically at 5 minutes, 1 hour, 8 hours, 16 hours, 24 hours, and 48 hours post-treatment (Figure 1A). UV irradiation induced the formation of CPDs and (6-4)PPs, located 6 nucleotides from the 3' terminus of the excised oligonucleotides, with lengths spanning from 19 to 28 base pairs (Supplementary Figure 1). For subsequent analyses, we judiciously selected reads in the 19-24 nucleotide length range, as they exhibited the most pronounced enrichment of dipyrimidine sequences across all samples. Following normalization through reads per kilobase per million reads (RPKM; Supplementary Figure 2), as detailed in the Materials and Methods section, we observed a robust correlation in repair patterns across the genome between the two replicates collected at each time point, underscoring the high reproducibility of our findings (Supplementary Figure 3). Moreover, pairwise correlation analysis of transcription-coupled repair patterns revealed sample clustering based on the type of DNA damage ((6-4)PP vs. CPD) as well as temporal ordering of samples collected at different time intervals (Supplementary Figure 4).

Our experimental data unequivocally affirm that *xpc-1 C. elegans* predominantly employs transcription-coupled repair to rectify DNA adducts, as evidenced by significantly higher repair of both (6-4)PP and CPD damages on the transcribed strand (TS) compared to the non-transcribed strand (NTS) (Supplementary Figure 5). Figure 1B shows an Integrative Genomics Viewer (IGV) screenshot of a 13-

kilobase region on chromosome I, featuring XR-seq, RNA-seq, and epigenomic profiles. When juxtaposed with RNA-seq, XR-seq offers more consistent and comprehensive insights into unspliced and nascent transcripts, encompassing both exons and introns. As depicted in Figure 1B, we illustrate a representative gene whose transcription is detected through long-capped RNA-seq, while simultaneously unveiling transcription-coupled repair through XR-seq. It is noteworthy that the reads acquired from XR-seq align to the template strand and are complementary to those obtained from RNA-seq, which align with the coding strand of the gene. Additionally, within the gene body, the signals derived from long-capped RNA-seq and XR-seq manifest a notably more uniform distribution compared to those obtained from RNA-seq analyses.

Intriguingly, we also observed instances of transcription-coupled repair within numerous intergenic regions, as exemplified in Figure 1C. To comprehensively explore intergenic transcription and its relationship with transcription-coupled repair, we systematically constructed consecutive genomic bins within intergenic regions and assayed their respective RNA-seq, capped RNA-seq, and XR-seq measurements (see Materials and Methods for details). Our investigations demonstrate a high degree of concordance between genome-wide signals obtained from XR-seq and those derived from capped RNA-seq, a method capable of capturing nuclear RNAs, irrespective of their polyadenylation (poly(A)) status. Conversely, conventional RNA-seq techniques primarily target RNAs with poly(A) tails, thereby falling short in capturing the entirety of intergenic transcriptional activity. Consequently, there is a near-zero correlation coefficient when comparing these conventional RNA-seq results to the capped RNA-seq and XR-seq datasets (Supplementary Figure 6). While gene-specific excision repair mechanisms have been extensively explored across various model organisms[3,20–26], 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 (Figure 1A).

Epigenetic markers and chromatin states validate the intergenic transcription detected by XR-seq.

To validate the nascent and intergenic transcription detected by XR-seq, we retrieved both genic and intergenic annotations of the *C. elegans* genome (ce11). First, the genome was systematically divided into three distinct categories: intergenic regions, regions within 2 kilobases upstream of transcription start sites (TSS), and transcript regions. Our analysis revealed a noteworthy distinction when comparing RNA-seq, capped RNA-seq, and XR-seq. Figure 2A illustrates that, in contrast to RNA-seq, both capped RNA-seq and 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 superior capability of capped RNA-seq and XR-seq in capturing transcriptional activity in these specific genomic locations.

Expanding our investigation further, we incorporated annotation of chromatin states of *C. elegans*[18]. As illustrated in Figure 2B, 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 200 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, enhancer regions, and tissue-specific transcription. Furthermore, categories that align with both (6-4)PP XR-seq and CPD XR-seq results 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 identified by XR-seq (not detected by RNA-seq), we focused on histone markers and chromatin accessibility (Figure 2C)[16,18]. When compared to randomly selected genomic regions spanning the entire genome, the regions uniquely pinpointed by 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 intensities of histone markers such as H3K4me1 and H3K4me3, typically associated with promoters and enhancers. Conversely, the intensities of histone marker H3K27me3, associated with gene repression, were diminished in these regions (Figure 2C). These corroborating epigenomic signatures serve as compelling evidence reaffirming the existence of intergenic transcription detected by XR-seq. Furthermore, they underscore the utility of XR-seq, utilizing transcription-coupled repair of DNA damage as a proxy, in uncovering previously elusive intergenic transcriptional events within the genome.

Transcription-coupled repair employs on annotated eRNA and lincRNA.

We next sought to examine the presence of transcription-coupled repair within annotated eRNAs and lincRNAs[17,19]. Previous studies, involving patients with XP-C, have provided evidence of XR-seq's capability to detect eRNA transcription[3]. Building upon this knowledge, we systematically examined both excision repair and transcription within these annotated regions. Our findings, as depicted in Figure 3, reveal that eRNAs (Figure 3 A, B) and lincRNAs (Figure 3 C, D) exhibit a notable presence in the data obtained from XR-seq, short-capped RNA-seq, and long-capped RNA-seq. In contrast, conventional RNA-seq methods show a limited ability to detect these transcripts. This discrepancy can be attributed to the intrinsic instability of eRNAs and lincRNAs, which renders them challenging to capture using conventional RNA-seq techniques. Remarkably, despite the inherent instability of eRNAs and lincRNAs, XR-seq proves to be a robust method for capturing transcription-coupled repair events within these regions, highlighting its sensitivity and utility in studying intergenic transcription.

XR-seq is a tool to detect intergenic transcription.

Upon overlaying the intergenic regions identified by (6-4)PP XR-seq, CPD XR-seq, RNA-seq, and capped RNA-seq, our observations, as meticulously depicted in the Venn diagrams presented in Figure 4, unveil compelling insights. First, our analysis demonstrates that intergenic transcription-coupled repair regions identified by (6-4)PP XR-seq and CPD XR-seq exhibit a remarkable level of concordance, with a complete overlap between these two damages. This remarkable alignment underscores the high reproducibility and accuracy of nascent transcript detection facilitated by XR-seq. Moreover, our investigations reveal an intriguing contrast when comparing XR-seq with RNA-seq. XR-seq, which distinguishes itself by employing transcription repair as a proxy for transcription, effectively complements capped RNA-seq and offers a comprehensive view of transcription in intergenic regions. In Figure 4A, we elucidate these regions detected in both replicates (representing higher specificity) show that XR-seq identifies a striking 55% additional regions beyond what RNA-seq detects. Furthermore, the regions detected in either replicate (reflecting higher sensitivity) display XR-seq's capacity to uncover 46% additional regions compared to RNA-seq alone. These findings underscore the enhanced sensitivity and specificity of XR-seq in delineating intergenic transcription compared to RNA-seq. Importantly, XR-seq's ability to capture transcription independent of RNA itself positions it as a powerful tool for investigating transcription in various genomic contexts.

MATERIALS AND METHODS

Biological Resources

The *C. elegans* wild-type (N2 ancestral) and xpc-1 (tm3886) strains were obtained from the *Caenorhabditis* Genetics Center and were cultured under standard conditions at room temperature on nematode growth media plates with *E. coli* strain OP50.

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. L1 larvae were exposed to 4,000 J/m^2 of UVB radiation (313 nm). The animals 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. 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[27]. In brief, supernatants were incubated with 5µL RNase A and then 5µL Proteinase K, purified, and then immunoprecipitated with either anti-CPD or anti-(6-4)PP antibodies. Immunoprecipitations were ligated to the adaptors, purified with the antibody used in the first purification, and DNA damage was reversed by either CPD or (6-4)PP photolyase. After PCR amplification, the library was sequenced with either Illumina HiSeq 4000 or NextSeq 2000 platforms.

RNA-seq

We followed existing protocol[28] for total RNA extracting 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[29]. Bowtie 2 was used for read alignment to the ce11 reference genome, followed by filtering, sorting, deduplication, and indexing[30]. Post-alignment filtering steps were adopted using Rsamtools (http://bioconductor.org/packages/Rsamtools). We only keep 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. For plotting strand-based average repair profiles of the genes, we selected 7061 genes longer than 1 kilobase pair, situated at least 500 base pairs away from neighboring genes. Each gene was evenly divided into 100 bins from the Transcription Start Site (TSS) to the Transcription End Site (TES), and 25 bins (2 kbp) upstream of TSS, 25 bins (2 kbp) downstream of TES. Bed files of the reads were intersected to the 150 bin-divided-gene list by Bedtools intersect with the following commands -d -wa -F 0.5 -S or -s for TS and NTS, respectively[31]. We present the descriptive properties of our data in Supplementary Table 1. 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[32]. We only kept the first read from the mate pair to ensure independent measures. Read counts for each gene were obtained using FeatureCounts[33].

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 and TC dinucleotides are highly correlated with the gene lengths from both the transcribed (TS) and non-transcribed (NTS) strands (Supplementary Figure 2). 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: (i) had at least ten TT or TC dinucleotides in the TS or the NTS; (ii) were less than 300 Kb; and (iii) had at least ten reads in total across all XR-seq samples.

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 transcription start sites) and those that overlap with blacklist regions in the ce11 genome, resulting in 85,418 bins[34]. 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 (6-4)PP and CPD repair in replicate 1 and replicate 2, respectively.

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[16]. 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[16]. All data were downloaded as processed bigwig files (Supplementary Table 2), and the regions were overlapped with the genomic regions to obtain the epigenetic measurements for each intergenic region.

Chromatin state, eRNA, and lincRNA 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 Figure 2B[18]. 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[35,19]. We removed eRNAs that overlap with either annotated genes or blacklist regions, resulting in a total of 324 eRNAs, which are presented in Figure 3 A and B. Similarly, we obtained 170 long intergenic non-coding RNAs (lincRNAs) in *C. elegans* from existing annotations[17]. 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 Figure 3 C and D.

DISCUSSION

The concept of transcription-coupled repair first surfaced in mammalian cells in 1987, and since then, a multitude of in vitro and in vivo methodologies have been developed to unravel the intricate mechanisms of repair factors and repair events[9,36,37]. Among these methods, XR-seq, distinguished by its single-nucleotide resolution, has been applied across a spectrum of organisms, including bacteria, yeast, flies, plants, and mammals[3,20–26,38]. While previous studies in *C. elegans* have suggested the existence of transcription-coupled repair through QPCR assay, our study stands as the pioneering high-resolution, genome-wide transcription-coupled repair map in response to UV damage in *C. elegans*[13]. Leveraging the precision of our 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[39]. To comprehensively investigate these intricate transcription dynamics, we conducted XR-seq at six distinct time points, 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 cap analysis gene expression (CAGE)-seq has been driven by the desire to discern transcription start sites and ncRNAs with heightened precision[16,40–45]. A comprehensive evaluation of the strengths and limitations of these methods can be found in[46].

In the context of *C. elegans* research, efforts to specifically target nascent RNAs and identify transcription start sites have utilized two primary techniques: GRO-seq and capped RNA-seq (CapSeq), as reported in previous studies [16,18,44,47–49]. Capped RNA-seq represents a modified version of CAGE-seq, where enzymatic background reduction is applied instead of affinity purification. It has been demonstrated that CapSeq exhibits greater precision in identifying transcription start sites compared to GRO-seq specifically within the *C. elegans* model[48]. Both of these methods rely on nuclei isolation, which exhibits an efficiency of approximately 50% [50]. Consequently, they necessitate a substantial amount of initial material for analysis. In the case of CapSeq, a multistep enzymatic degradation process is employed to remove uncapped RNAs, and it is important to note that this method may not detect noncanonical capped RNAs[51,52].

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 RNAPII ChIP-seq and 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[53,54]. 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 other nascent RNA sequencing methods.

Our findings demonstrate the efficacy of XR-seq in capturing transcription events within both genic and intergenic regions. Notably, XR-seq exhibits sensitivity comparable to that of capped RNA-seq in detecting annotated enhancer RNAs (eRNAs) and long intergenic non-coding RNAs (lincRNAs). While RNA-seq detects only 19-44% of intergenic transcription, our data reveal that up to 70% of the overall intergenic transcription landscape is shared between XR-seq and capped RNA-seq, highlighting the substantial overlap and providing valuable insights into nascent transcription dynamics and the intricate interplay between transcription-coupled repair and intergenic regions.

AUTHOR CONTRIBUTIONS

A.S. envisioned and initiated the study, while C.K. conducted the experiment. All authors designed and conducted the analysis, wrote, and approved the manuscript.

DATA AVAILABILITY

XR-seq and RNA-seq data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database with accession number GSE245181 (to be released after peer review). ATAC-seq, ChIP-seq, and DNase-seq are available from GEO with accession numbers GSE114439, GSE114440, and

GSE114481, respectively. All code used in this paper is available at https://github.com/yuchaojiang/damage_repair/tree/master/XPC_C_elegans.

COMPETING INTERESTS

The authors declare that they have no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by National Institute of health grants R35 GM118102 (A.S.), R01 ES033414 (A.S.) and R35 GM138342 (Y.J.). Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing. The authors thank Dr. Shawn Ahmed for help discussions and comments.

FIGURE LEGENDS

Figure 1. Detection of Transcription-Coupled Repair and Genome-Wide Transcription by XR-seq. (A) Overview of the study design illustrating the comparative analysis of RNA-seq, capped-RNA-seq, and XR-seq reads for their capacity to identify genome-wide transcription. (B) Distribution of the XR-seq signal over the 13Kb region, separated by strand, for CPD and (6-4)PP 1 hour after 4,000J/m^2 UVB treatment. Stranded *xpc-1 RNA-seq*, long and short capped RNA-seq tracks in blue (plus strand) and red (minus strand) are plotted above, and ATAC-seq (dark green), DNase (dark green), H3K4me3 (light green), H3K4me1 (light green) and H3K27me3 (gray) ChIP-seq tracks are plotted below the XR-seq tracks. Browser view of representative genes clearly demonstrates the occurrence of transcription-coupled repair within the gene body. XR-seq and long-capped RNA-seq methods provide comprehensive coverage of the entire transcript, encompassing both intronic and exonic regions, in annotated genes, in contrast to RNA-seq. The expression of these genes is further substantiated by the presence of high levels of open chromatin and expression-associated markers, including ATAC-seq, DNase-seq, and H3k4me3. The minus strand denotes the transcribed strand, depicted in brown color in the XR-seq representation. (C) Browser view of a representative intergenic region reveals transcription events detected by long-capped RNA-seq and XR-seq but not by RNA-seq. Expression in this intergenic region is corroborated by the presence of elevated levels of open chromatin and expression markers, including ATAC-seq, DNase-seq, H3k4me3, and H3Kme1.

Figure 2. Transcription-Coupled Repair in Intergenic Regions Detected by XR-seq Supported by Epigenomic Signatures. (A) Bar graphs depict the genome-wide distribution of reads obtained from various sequencing methods, including CPD XR-seq, (6-4)PP XR-seq, long-capped RNA-seq, short-capped RNA-seq, *xpc-1* RNA-seq, and wild-type (WT) RNA-seq. Notably, both XR-seq and capped RNA-seq techniques reveal transcription events occurring outside of annotated transcripts. (B) Overlapping reads from XR-seq, capped RNA-seq, and RNA-seq were analyzed within genomic intervals corresponding to 20 distinct chromatin states predicted for the autosomes of L3 stage C. elegans. Values were normalized with respect to read depth and interval length. (C) Examination of intergenic XR-seq reads, which are undetectable by RNA-seq, in association with ATAC-seq, DNase-seq, H3K4me3, H3K4me1, and H3K27me3 peaks. XR-seq reads exhibit a strong correlation with active transcription markers, contrasting with 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.

Figure 3. XR-seq Reveals Transcription-Coupled Repair in eRNAs and lincRNAs overlooked by RNA-seq. Heatmaps display log-normalized gene expression and transcription-coupled repair for annotated enhancer RNAs (eRNAs) (A) and long intergenic non-coding RNAs (lincRNAs) (C), segregated by chromosomes. Bar graphs represent log-normalized read counts for eRNA (B) and lincRNA (D). Data are presented for WT RNA-seq, *xpc-1* RNA-seq, short-capped RNA-seq, long-capped RNA-seq, and two independent replicates of (6-4)PP and CPD XR-seq experiments.

Figure 4. XR-seq identifies intergenic transcription-coupled repair, in high concordance with intergenic transcription identified by capped RNA-seq. For the 85,418 intergenic bins, we identified regions with non-zero read counts by short- or long-capped RNA-seq, RNA-seq, (6-4)PP XR-seq, and CPD XR-seq, respectively. We require non-zero read counts to be detected in both (A) or either replicate (B) and report the overlapping results separately.

REFERENCES

1. Lukas J, Lukas C, Bartek J. More than just a focus: The chromatin response to DNA damage and its role in genome integrity maintenance. Nat Cell Biol. Nature Publishing Group; 2011 Oct;13(10):1161–1169.

2. Reardon JT, Sancar A. Nucleotide excision repair. Prog Nucleic Acid Res Mol Biol. 2005;79:183–235. PMID: 16096029

3. 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 May 1;29(9):948–960. PMCID: PMC4421983

4. 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. Proceedings of the National Academy of Sciences; 1992 Apr 15;89(8):3664–3668.

5. Sancar A. DNA excision repair. Annu Rev Biochem. 1996;65:43–81. PMID: 8811174

6. Sancar A. Mechanisms of DNA Repair by Photolyase and Excision Nuclease (Nobel Lecture). Angew Chem Int Ed. 2016;55(30):8502–8527.

7. Mu D, Park CH, Matsunaga T, Hsu DS, Reardon JT, Sancar A. Reconstitution of human DNA repair excision nuclease in a highly defined system. J Biol Chem. 1995 Feb 10;270(6):2415–2418. PMID: 7852297

8. Reardon JT, Sancar A. Recognition and repair of the cyclobutane thymine dimer, a major cause of skin cancers, by the human excision nuclease. Genes Dev. 2003 Oct 15;17(20):2539–2551. PMCID: PMC218148

9. Selby CP, Lindsey-Boltz LA, Li W, Sancar A. Molecular Mechanisms of Transcription-Coupled Repair. Annu Rev Biochem. 2023;92(1):115–144. PMID: 37001137

10. Mu D, Hsu DS, Sancar A. Reaction mechanism of human DNA repair excision nuclease. J Biol Chem. 1996 Apr 5;271(14):8285–8294. PMID: 8626523

11. Evans E, Moggs JG, Hwang JR, Egly JM, Wood RD. Mechanism of open complex and dual incision formation by human nucleotide excision repair factors. EMBO J. 1997 Nov 3;16(21):6559–6573. PMCID: PMC1170260

12. Kemp MG. Damage removal and gap filling in nucleotide excision repair. The Enzymes. 2019;45:59–97. PMID: 31627883

13. 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. PMCID: PMC1929140

14. Lans H, Vermeulen W. Nucleotide Excision Repair in Caenorhabditis elegans. Mol Biol Int. 2011;2011:542795. PMCID: PMC3195855

15. 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 Nov 4;48(19):10973–10985.

16. Jänes J, Dong Y, Schoof M, Serizay J, Appert A, Cerrato C, Woodbury C, Chen R, Gemma C, Huang N, Kissiov D, Stempor P, Steward A, Zeiser E, Sauer S, Ahringer J. Chromatin accessibility dynamics across C. elegans development and ageing. Lee SS, Tyler JK, editors. eLife. eLife Sciences Publications, Ltd; 2018 Oct 26;7:e37344.

17. Nam JW, Bartel DP. Long noncoding RNAs in C. elegans. Genome Res. 2012 Dec;22(12):2529–2540. PMCID: PMC3514682

18. 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. Proceedings of the National Academy of Sciences; 2016 Nov 8;113(45):E7020–E7029.

19. Jin W, Jiang G, Yang Y, Yang J, Yang W, Wang D, Niu X, Zhong R, Zhang Z, Gong J. Animal-eRNAdb: a comprehensive animal enhancer RNA database. Nucleic Acids Res. 2022 Jan 7;50(D1):D46–D53.

20. Adebali O, Sancar A, Selby CP. Mfd translocase is necessary and sufficient for transcription-coupled repair in Escherichia coli. J Biol Chem. 2017 Nov 10;292(45):18386–18391.

21. Adebali O, Yang Y, Neupane P, Dike NI, Boltz JL, Kose C, Braunstein M, Selby CP, Sancar A, Lindsey-Boltz LA. The Mfd protein is the transcription-repair coupling factor (TRCF) in Mycobacterium smegmatis. J Biol Chem. 2023 Mar 1;299(3):103009.

22. 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 Apr 10;115(15):E3408–E3415. PMCID: PMC5899493

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. Nature Publishing Group; 2018 Apr 17;9(1):1503.

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 Nov 29;294(48):18092–18098. PMCID: PMC6885609

25. Akkose U, Kaya VO, Lindsey-Boltz L, Karagoz Z, Brown AD, Larsen PA, Yoder AD, Sancar A, Adebali O. 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 Aug 6;22(1):600.

26. Yimit A, Adebali O, Sancar A, Jiang Y. Differential damage and repair of DNA-adducts induced by anti-cancer drug cisplatin across mouse organs. Nat Commun. Nature Publishing Group; 2019 Jan 18;10(1):309.

27. Lindsey-Boltz LA, Yang Y, Kose C, Deger N, Eynullazada K, Kawara H, Sancar A. Nucleotide excision repair in Human cell lines lacking both XPC and CSB proteins. Nucleic Acids Res. 2023 Jul 7;51(12):6238–6245.

28. Green MR, Sambrook J. Total RNA Extraction from Caenorhabditis elegans. Cold Spring Harb Protoc. 2020 Sep 1;2020(9):101683. PMID: 32873731

29. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011 May 2;17(1):10–12.

30. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. Nature Publishing Group; 2012 Apr;9(4):357–359.

31. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinforma. 2014;47(1):11.12.1-11.12.34.

32. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013 Jan 1;29(1):15–21. PMCID: PMC3530905

33. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923–930.

34. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep. Nature Publishing Group; 2019 Jun 27;9(1):9354.

35. Sartorelli V, Lauberth SM. Enhancer RNAs are an important regulatory layer of the epigenome. Nat Struct Mol Biol. Nature Publishing Group; 2020 Jun;27(6):521–528.

36. Mellon I, Spivak G, Hanawalt PC. Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene. Cell. 1987 Oct 23;51(2):241–249. PMID: 3664636

37. Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. Nat Rev Mol Cell Biol. Nature Publishing Group; 2008 Dec;9(12):958–970.

38. Hu J, Selby CP, Adar S, Adebali O, Sancar A. Molecular mechanisms and genomic maps of DNA excision repair in Escherichia coli and humans. J Biol Chem. 2017 Sep 22;292(38):15588–15597.

39. 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 Feb 16;293(7):2476–2486. PMCID: PMC5818198

40. Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, Waters CT, Munson K, Core LJ, Lis JT. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc. Nature Publishing Group; 2016 Aug;11(8):1455–1476.

41. Santa FD, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G. A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. PLOS Biol. Public Library of Science; 2010 May 11;8(5):e1000384.

42. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008 Dec 19;322(5909):1845–1848. PMCID: PMC2833333

43. Morioka MS, Kawaji H, Nishiyori-Sueki H, Murata M, Kojima-Ishiyama M, Carninci P, Itoh M. Cap Analysis of Gene Expression (CAGE): A Quantitative and Genome-Wide Assay of Transcription Start Sites. Methods Mol Biol Clifton NJ. 2020;2120:277–301. PMID: 32124327

44. Gu W, Lee HC, Chaves D, Youngman EM, Pazour GJ, Conte D, Mello CC. CapSeq and CIP-TAP Identify Pol II Start Sites and Reveal Capped Small RNAs as C. elegans piRNA Precursors. Cell. Elsevier; 2012 Dec 21;151(7):1488–1500. PMID: 23260138

45. Chen RAJ, Down TA, Stempor P, Chen QB, Egelhofer TA, Hillier LW, Jeffers TE, Ahringer J. The landscape of RNA polymerase II transcription initiation in C. elegans reveals promoter and enhancer architectures. Genome Res. 2013 Aug;23(8):1339–1347. PMCID: PMC3730107

46. Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat Rev Genet. 2016 Apr;17(4):207–223. PMID: 26948815

47. 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. Nature Publishing Group; 2014 Apr;21(4):358–365.

48. 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 Jun 27;50(6):894–907.

49. Saito TL, Hashimoto S ichi, Gu SG, Morton JJ, Stadler M, Blumenthal T, Fire A, Morishita S. The transcription start site landscape of C. elegans. Genome Res. 2013 Aug;23(8):1348–1361. PMCID: PMC3730108

50. Quarato P, Cecere G. Global Run-On sequencing to measure nascent transcription in C. elegans. STAR Protoc. 2021 Dec 17;2(4):100991.

51. Doamekpor SK, Sharma S, Kiledjian M, Tong L. Recent insights into noncanonical 5′ capping and decapping of RNA. J Biol Chem. 2022 Jun 21;298(8):102171. PMCID: PMC9283932

52. Jiao X, Doamekpor SK, Bird JG, Nickels BE, Tong L, Hart RP, Kiledjian M. 5' End Nicotinamide Adenine Dinucleotide Cap in Human Cells Promotes RNA Decay through DXO-Mediated deNADding. Cell. 2017 Mar 9;168(6):1015-1027.e10. PMCID: PMC5371429

53. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. Widespread RNA and DNA Sequence Differences in the Human Transcriptome. Science. American Association for the Advancement of Science; 2011 Jul;333(6038):53–58.

54. Wang IX, Core LJ, Kwak H, Brady L, Bruzel A, McDaniel L, Richards AL, Wu M, Grunseich C, Lis JT, Cheung VG. RNA-DNA Differences Are Generated in Human Cells within Seconds after RNA Exits Polymerase II. Cell Rep. 2014 Mar 13;6(5):906–915.

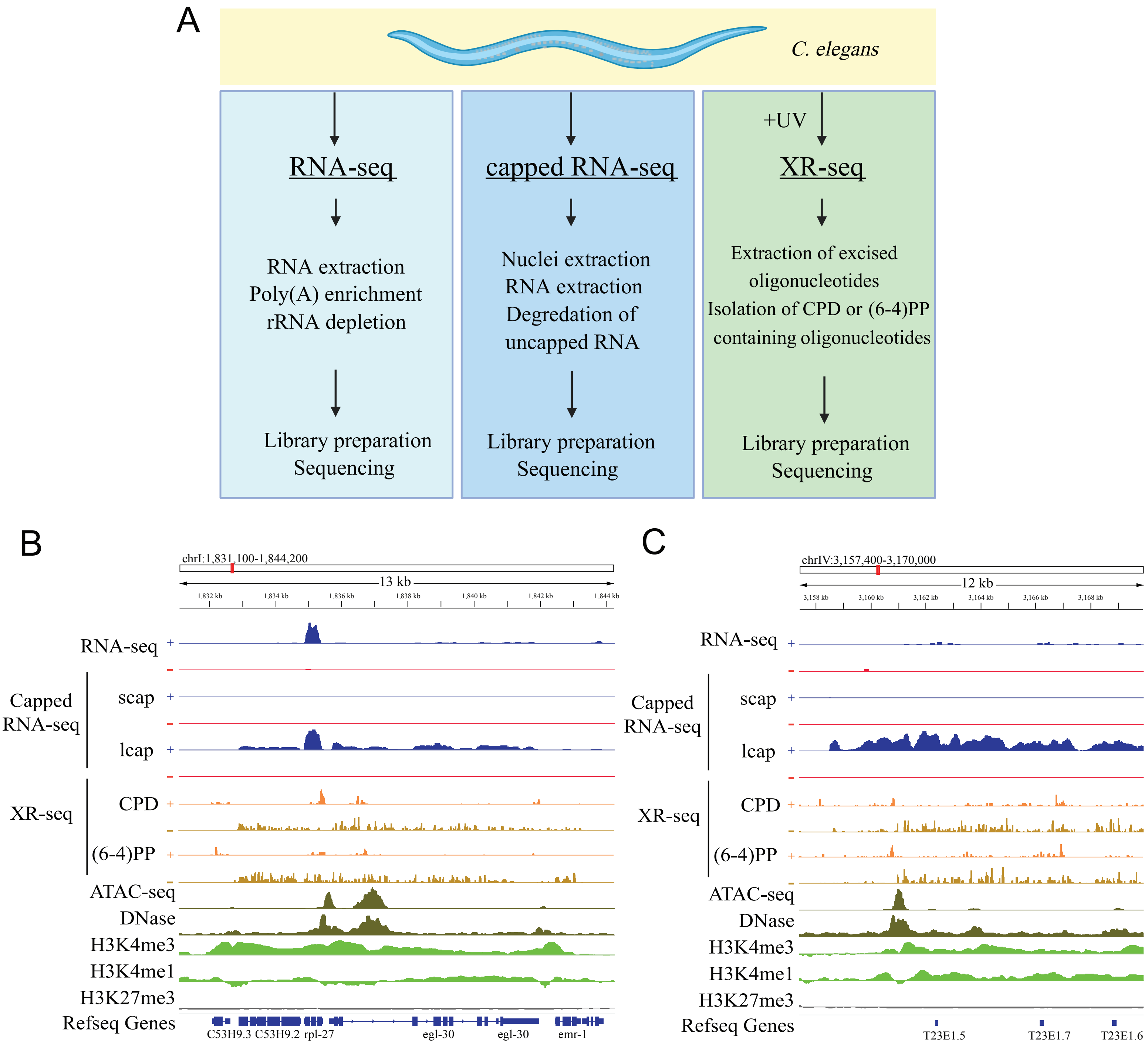
A
C. elegans
RNA-seq
RNA extraction
Poly(A) enrichment
rRNA depletion
Library preparation
Sequencing
capped RNA-seq
Nuclei extraction
RNA extraction
Degredation of
uncapped RNA
Library preparation
Sequencing
+UV
XR-seq
Extraction of excised
oligonucleotides
Isolation of CPD or (6-4)PP
containing oligonucleotides
Library preparation
Sequencing
B
chrI:1,831,100-1,844,200
13 kb
1,832 kb
1,834 kb
1,836 kb
1,838 kb
1,840 kb
1,842 kb
1,844 kb
RNA-seq
Capped
RNA-seq
scap
lcap
XR-seq
CPD
(6-4)PP
+
-
ATAC-seq
DNase
H3K4me3
H3K4me1
H3K27me3
Refseq Genes
C53H9.3
C53H9.2
rpl-27
egl-30
egl-30
emr-1
C
chrIV:3,157,400-3,170,000
12 kb
3,158 kb
3,160 kb
3,162 kb
3,164 kb
3,166 kb
3,168 kb
T23E1.5
T23E1.7
T23E1.6

Figure 1

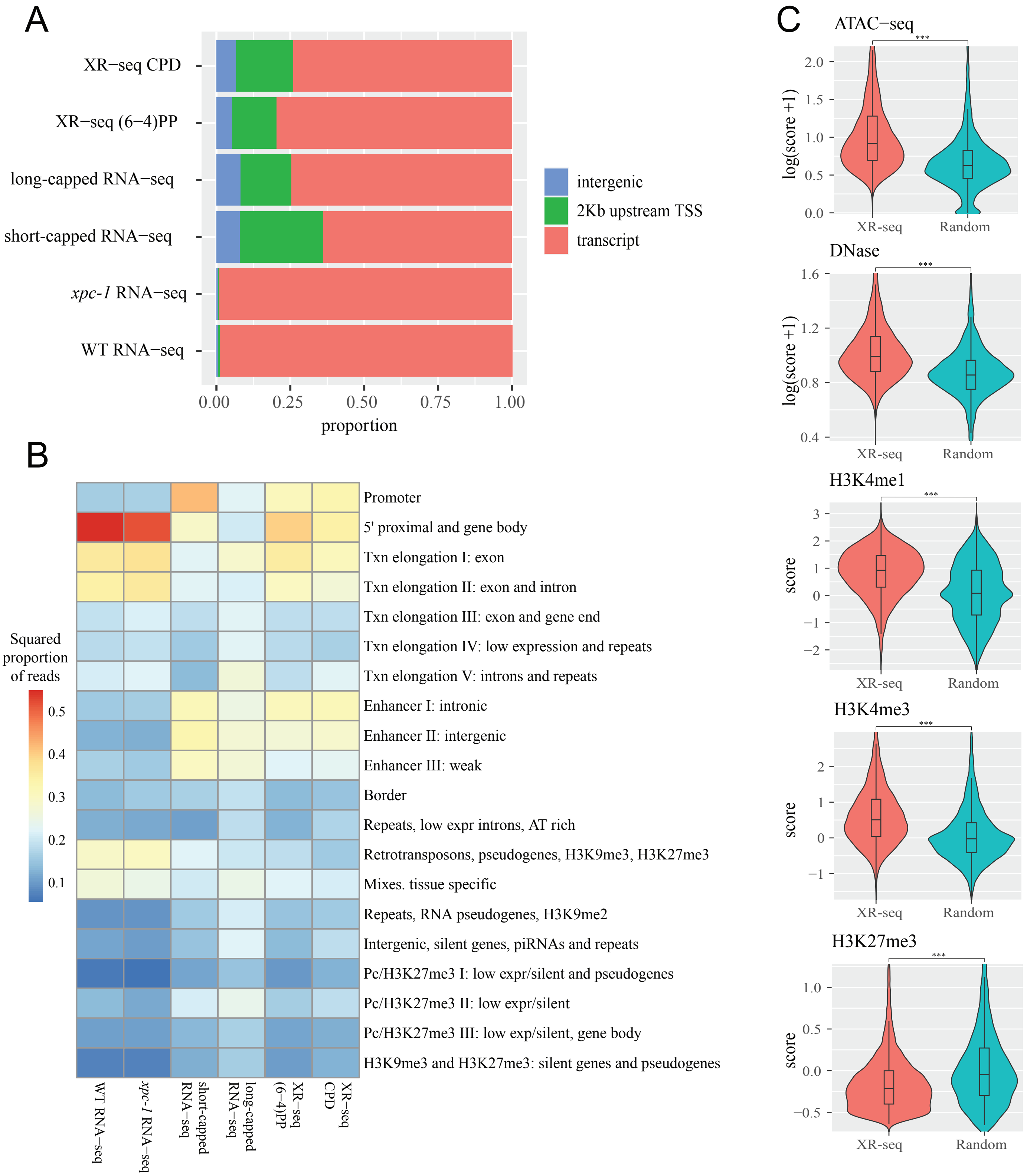
A
XR−seq CPD
XR−seq (6−4)PP
long-capped RNA−seq
short-capped RNA−seq
xpc-1 RNA−seq
WT RNA−seq
0.00
0.25
0.50
0.75
1.00
proportion
intergenic
2Kb upstream TSS
transcript
B
Squared proportion of reads
0.5
0.4
0.3
0.2
0.1
Promoter
5' proximal and gene body
Txn elongation I: exon
Txn elongation II: exon and intron
Txn elongation III: exon and gene end
Txn elongation IV: low expression and repeats
Txn elongation V: introns and repeats
Enhancer I: intronic
Enhancer II: intergenic
Enhancer III: weak
Border
Repeats, low expr introns, AT rich
Retrotransposons, pseudogenes, H3K9me3, H3K27me3
Mixes. tissue specific
Repeats, RNA pseudogenes, H3K9me2
Intergenic, silent genes, piRNAs and repeats
Pc/H3K27me3 I: low expr/silent and pseudogenes
Pc/H3K27me3 II: low expr/silent
Pc/H3K27me3 III: low exp/silent, gene body
H3K9me3 and H3K27me3: silent genes and pseudogenes
WT RNA−seq
xpc-1 RNA−seq
short-capped RNA−seq
long-capped RNA−seq
XR−seq (6−4)PP
XR−seq CPD
C
ATAC−seq
***
log(score +1)
2.0
1.5
1.0
0.5
0.0
XR-seq
Random
DNase
***
log(score +1)
1.6
1.2
0.8
0.4
XR-seq
Random
H3K4me1
***
score
3
2
1
0
−1
−2
XR-seq
Random
H3K4me3
***
score
2
1
0
−1
XR-seq
Random
H3K27me3
***
score
1.0
0.5
0.0
−0.5
XR-seq
Random

Figure 2

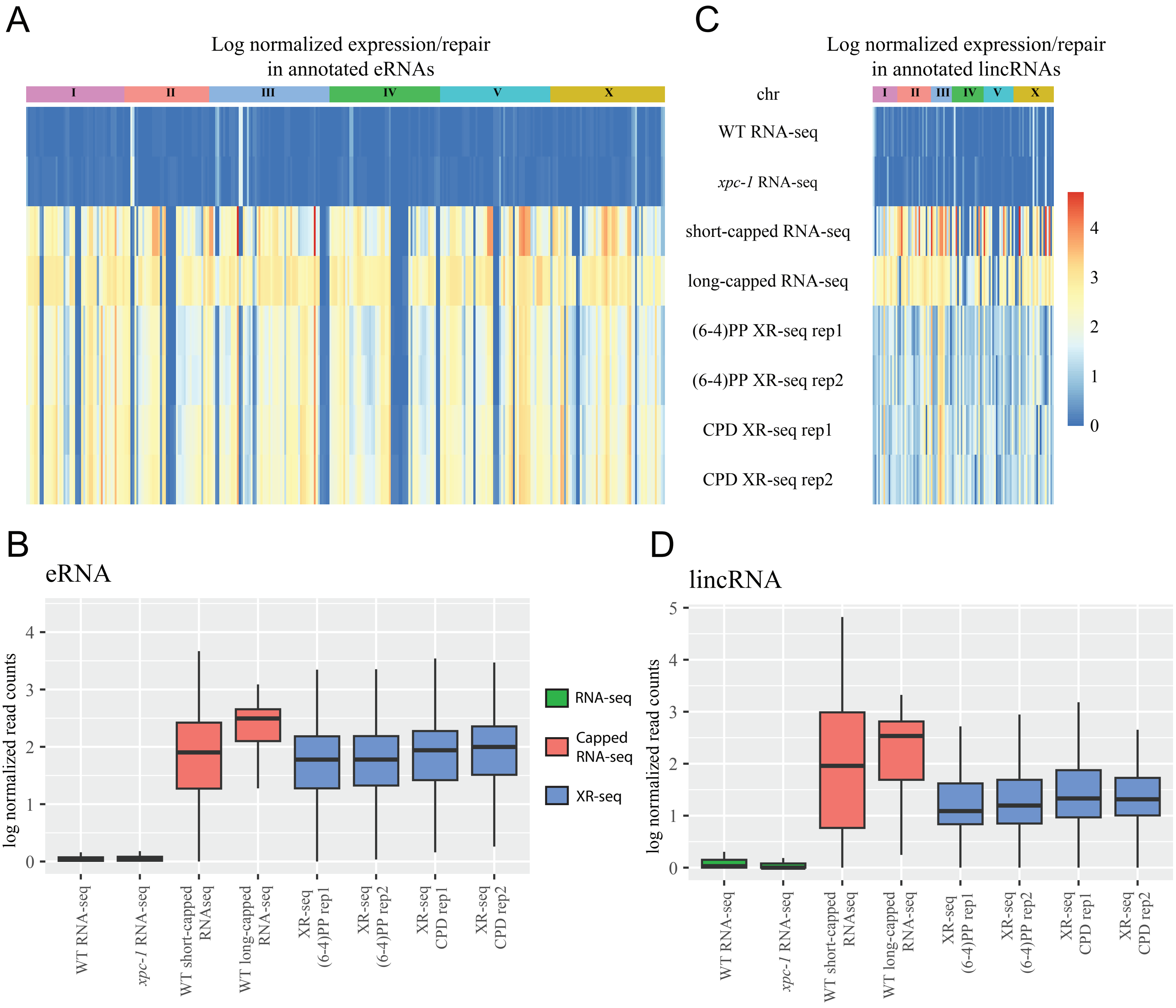
A
Log normalized expression/repair
in annotated eRNAs
I
II
III
IV
V
X
C
Log normalized expression/repair
in annotated lincRNAs
chr
WT RNA-seq
xpc-1 RNA-seq
short-capped RNA-seq
long-capped RNA-seq
(6-4)PP XR-seq rep1
(6-4)PP XR-seq rep2
CPD XR-seq rep1
CPD XR-seq rep2
4
3
2
1
0
B
eRNA
log normalized read counts
0
1
2
3
4
WT RNA-seq
xpc-1 RNA-seq
WT short-capped RNAseq
WT long-capped RNA-seq
XR-seq (6-4)PP rep1
XR-seq (6-4)PP rep2
XR-seq CPD rep1
XR-seq CPD rep2
RNA-seq
Capped RNA-seq
XR-seq
D
lincRNA
log normalized read counts
0
1
2
3
4
5
WT RNA-seq
xpc-1 RNA-seq
WT short-capped RNAseq
WT long-capped RNA-seq
XR-seq (6-4)PP rep1
XR-seq (6-4)PP rep2
XR-seq CPD rep1
XR-seq CPD rep2

Figure 3

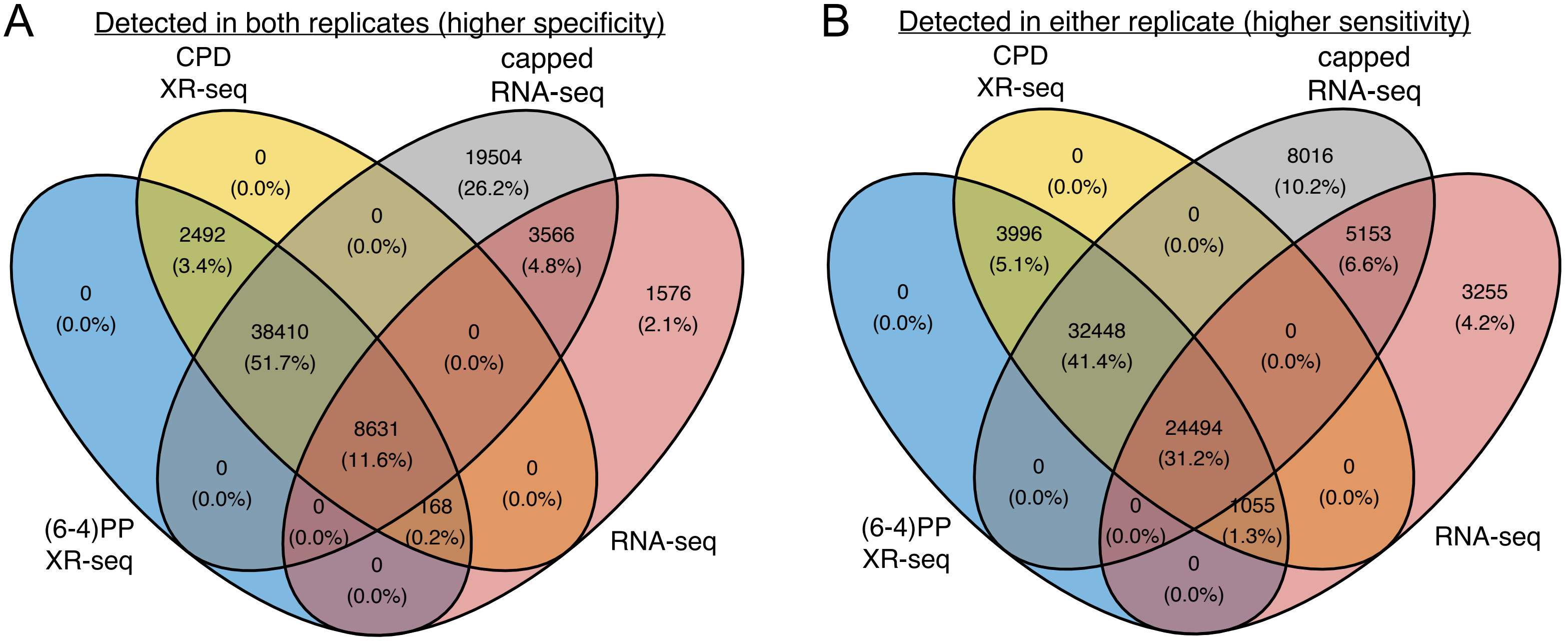
A
Detected in both replicates (higher specificity)
CPD
XR-seq
capped
RNA-seq
(6-4)PP
XR-seq
RNA-seq
0
(0.0%)
19504
(26.2%)
2492
(3.4%)
0
(0.0%)
3566
(4.8%)
0
(0.0%)
1576
(2.1%)
38410
(51.7%)
0
(0.0%)
8631
(11.6%)
0
(0.0%)
0
(0.0%)
0
(0.0%)
168
(0.2%)
0
(0.0%)
B
Detected in either replicate (higher sensitivity)
CPD
XR-seq
capped
RNA-seq
(6-4)PP
XR-seq
RNA-seq
0
(0.0%)
8016
(10.2%)
3996
(5.1%)
0
(0.0%)
5153
(6.6%)
0
(0.0%)
3255
(4.2%)
32448
(41.4%)
0
(0.0%)
24494
(31.2%)
0
(0.0%)
0
(0.0%)
0
(0.0%)
1055
(1.3%)
0
(0.0%)

Figure 4