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, NC, 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, [4]Department of Biomedical Engineering, College of Engineering, Texas A&M University, College Station, TX 77843

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

KEYWORDS: nucleotide excision repair; eXcision Repair-Sequencing (XR-Seq); Xeroderma Pigmentosum complementation group C (XPC); Cockayne Syndrome complementation group B (CSB); transcriptome; non-coding RNAs (ncRNAs)

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 the eXcision Repair-sequencing (XR-seq) in the model organism *Caenorhabditis elegans* to generate genome-wide repair maps from a wild-type strain with normal excision repair, a strain lacking TCR (*csb-1*), or one 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.

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 that harness RNA capture through their 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 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 ~27-nt damage-containing oligonucleotide excision products [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 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 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* worm model organism because of its nearly completely annotated nuclear genome, which is approximately 1/30 the size of the human genome, and because of 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 (Fig S1) 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** 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. 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 to the coding strand of the gene. 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. Quantitatively, we used the ratio of read counts from the TS to those from both the TS and non-transcribed strand (NTS) as a proxy for TCR, with genome-wide results 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

114 previously shown, the unirradiated control (WT no UV) results in extremely low read-numbers
115 (0.003% of UV-irradiated WT) that are not specific [20].
116
117 The XR-seq data were then analyzed to assess repair on the TS and NTS of all non-
118 overlapping genes greater than 2 kb in length (**Fig 2B**). Again, such analysis clearly illustrates the
119 presence of TCR in the WT strain (top), which is partially masked due to global repair products
120 mapping to both strands. There was no observed strand difference in repair within gene bodies in
121 the *csb-1* mutant (middle), which lacks TCR. Notably, the differences observed upstream of the
122 transcription start site (TSS) and transcription end site (TES) can be attributed to TT-content of
123 these regions of the genome (bottom) resulting in different levels of DNA damage in these areas
124 since CPDs are primarily formed at TTs and the majority of XR-seq reads contain this dinucleotide
125 ~6 nt from the 3'-end (**S2, S3A Fig)**. Within gene bodies the TT-content does not significantly vary
126 with gene length or between strands in *C. elegans* (**S4 Fig)**. In stark contrast to the *csb-1* mutant,
127 the majority of repair events map to the transcribed strand in the *xpc-1* mutant where TCR is the
128 only functional excision repair pathway **(Fig 2B)**. As previously seen in humans and other
129 organisms [19, 22-25], XR-seq reads peak at the 5'-end of genes and decrease toward the 3'-end,
130 which is consistent with the TCR model proposed by Chiou *et al.* [26], and the skewed pattern
131 gradually diminishes as the repair process proceeds over time (**S3B Fig)**. The 5'-peak of repair on
132 the TS is not unique to L1-stage worms as this pattern is also observed in a mixed population of
133 worms (**S3C Fig)**.
134
135 XR-seq analysis in human XP-C cells revealed pronounced CPD repair on the non-
136 template strand upstream of the TSS due to divergent transcription at promoters [19, 25]. The *C.*
137 *elegans* XR-seq analysis shown in **Fig 2B** does not exhibit this pattern even though anti-sense
138 transcription at promoters has been reported in worms [6]. Therefore, we further analyzed repair
139 in the region of a greater number of TSSs (all TSSs that are at least 1kb apart) at an individual
140 basis as visualized in the heatmaps shown in **Fig 2C** and **S5 Fig**. With this analysis, we were able
141 to detect anti-sense transcription (enrichment on NTS upstream of the TSS) in a subset of genes.
142 The TS upstream of the TSS has much higher read-count, likely due to extensive transcription of
143 upstream eRNAs, which has been reported to be prevalent in *C. elegans* and occur in the direction
144 of the downstream gene in 90% of cases [6].
145
146 We next sought to identify genes that exhibit significantly differential and dynamic repair
147 using time-series *xpc-1* XR-seq data collected at 5min, 1h, 8h, 16h, 24h, and 48h after UV-
148 irradiation (see Materials and Methods for details). We identified 121 genes exhibiting significant
149 dynamic repair across timepoints **(S6A Fig)** and performed gene ontology (GO) analysis of
150 biological processes **(S6B Fig)** and cellular components **(S6C Fig)**. While investigation of gene-
151 specific excision repair has been extensively explored across various model organisms [23, 24, 27-
152 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 **(Fig 3B, S7A Fig)**. 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.

192 Expanding our investigation further, we incorporated annotation of chromatin states of *C. elegans*.
193 As illustrated in **Fig 4A** and **S7B Fig**, our analysis of chromatin states has unveiled intriguing
194 distinctions among the different sequencing methods. Notably, when we examine the distribution
195 of chromatin states, RNA-seq appears to predominantly align with 5' proximal regions, gene
196 bodies, and exons. However, it displays relatively lower read counts in categories associated with
197 retrotransposons, pseudogenes, and tissue-specific regions. In stark contrast, both capped RNA-
198 seq and XR-seq exhibit notably similar chromatin state patterns, although some nuanced
199 differences do exist between the two. A closer examination demonstrates that both short-capped
200 RNA-seq and long-capped RNA-seq reveal genic and intergenic transcription, including intergenic
201 enhancers. Short-capped RNA-seq indicates shorter transcripts, corresponding to transcription
202 initiation events and enhancers shorter than 100 base pairs. In contrast, long-capped RNA-seq
203 captures longer transcripts within the nucleus, encompassing both pre-mature and mature RNAs.
204 These longer transcripts relate to transcription elongation, enhancer regions, and tissue-specific
205 transcription. Furthermore, categories that align with XR-seq encompass a combination of short-
206 and long-capped RNA-seq signals, indicating the concordance between XR-seq and capped RNA-
207 seq in capturing transcriptional events.
208
209 In our comprehensive analysis of transcribed intergenic regions specifically identified by
210 *xpc-1* XR-seq (but not detected by RNA-seq), we focused on histone markers and chromatin
211 accessibility (**Fig 4B**). When compared to randomly selected genomic regions spanning the entire
212 genome, the regions uniquely pinpointed by *xpc-1* XR-seq exhibited distinct epigenomic
213 signatures. Specifically, these regions displayed significantly heightened chromatin accessibility,
214 indicating a more open chromatin structure conducive to transcription. Additionally, we observed
215 increased overlap with histone markers such as H3K4me1 and H3K4me3, typically associated
216 with promoters and enhancers. Conversely, there were less reads overlapping with regions with
217 histone marker H3K27me3, associated with gene repression. These corroborating epigenomic
218 signatures serve as compelling evidence reaffirming the existence of intergenic transcription
219 detected by *xpc-1* XR-seq. Furthermore, they underscore the utility of transcription-coupled repair
220 as a proxy for uncovering previously elusive intergenic transcriptional events within the genome.
221
222 ***Novel intergenic transcription identified with xpc-1 XR-seq.***
223 We next examined RNA-seq, XR-seq, and long- and short-capped RNA-seq read density
224 specifically within three classes of annotated intergenic ncRNAs: enhancer RNAs (eRNAs) (**Fig**
225 **5A** and **S9A Fig**), long intergenic non-coding RNAs (lincRNAs) (**Fig 5B** and **S9B Fig**), and Piwi-
226 interacting RNAs (piRNAs) (**Fig5C** and **S9C Fig**). Heatmaps (left) display normalized read counts
227 for the individual annotated intergenic ncRNAs segregated by chromosomes and the bar graphs
228 (right) summarize the log-normalized read counts for the class of ncRNA. Our findings reveal that
229 the RNA-seq method shows limited ability to detect any of these intergenic ncRNA transcripts.
230 This is likely attributed to the lack of poly(A) tailing of ncRNAs, which prevent them from being
231 captured by the conventional RNA-seq technique. Both eRNAs and lincRNAs are very well-

232 represented in the data obtained from *xpc-1* XR-seq and long- and short-capped RNA-seq.
233 Interestingly, read density at piRNAs is high for both long-cap RNA seq and *xpc-1* XR-seq, but
234 not short-capped RNA seq. The findings from the read density analysis of these three major classes
235 of known *C. elegans* intergenic ncRNAs demonstrate the utility of mapping such transcripts with
236 transcription-coupled repair.
237
238 To assess all intergenic regions (annotated and unannotated) to determine the degree of
239 coverage and overlap between the three methods, we divided the intergenic regions into 85,418
240 bins and identified those containing *xpc-1* XR-seq, RNA-seq, or capped RNA-seq reads (**S10 Fig)**.
241 The results depicted in the Venn diagram presented in **Fig 6A** show several compelling insights.
242 First, our analysis demonstrates that the transcription-coupled repair in the intergenic regions
243 identified by *xpc-1* XR-seq exhibit similar coverage and remarkable concordance with capped
244 RNA-seq, with both exhibiting ~83% bin-coverage and 80% overlap between the two datasets.
245 Second, as observed with the analyses above, RNA-seq has low coverage in intergenic regions
246 relative to XR-seq and capped RNA-seq. Third, 10% of the bins contain reads unique to *xpc-1* XR-
247 seq. Taken together, these results underscore the sensitivity of transcript-detection by *xpc-1* XR-
248 seq.
249
250 We further investigated the location and identity of the 7,903 bins that were only detected
251 in the *xpc-1* XR-seq data set and **Fig 6B** shows a pie chart summarizing the results. Of the *xpc-1*
252 XR-seq-unique bins, 34.7% were annotated (dark blue) and the remaining 65.3% have not been
253 annotated (light blue). Of the bins overlapping the 2,722 annotations, 76.8% of those are annotated
254 as piRNAs **(Fig 6C)** and are primarily found on chromosome IV **(Fig 6D)**. Interestingly, 26% of
255 the unannotated bins also map to chromosome IV **(Fig 6C)** and we hypothesize that these may be
256 novel piRNAs or piRNA precursors. In summary, the *xpc-1* XR-seq data set is a useful tool for
257 detecting RNAPII transcription and identifying new transcripts in the previously unannotated
258 intergenic regions of *C. elegans*.
259
260 **MATERIALS AND METHODS**
261 ***Biological Resources***
262 The *C. elegans* wild-type (N2 ancestral), *csb-1* (RB1801) and *xpc-1* (TG2226) strains were
263 obtained from the *Caenorhabditis* Genetics Center and were cultured under standard conditions at
264 room temperature on nematode growth media (NGM) agar plates with *E. coli* strain OP50.
265
266 ***XR-seq***
267 To obtain L1 larvae, eggs were collected from adult animals by hypochlorite treatment, and kept
268 in M9 buffer at 22°C for 16 hours with gentle rotation. Arrested L1 larvae were placed on NGM
269 agar plates with OP50, fed with bacteria for 3 to 4 hours to eliminate the effect of starvation, then
270 exposed to 400 kJ/cm^2 of UVB radiation (313 nm). The worms were collected in M9 buffer at 5
271 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 either anti-CPD antibody. Immunoprecipitations were ligated to the adaptors, purified with the antibody used in the first purification, and DNA damage was reversed by either 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 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 TGGAATTCTCGGGTGCCAAGGAACTCCAGTNNNNNNACGATCTCGTATGCCGTCTTC TGCTTG 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 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. 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: (i) had at least ten TT 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. 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 **Fig 2A** and **S3 Fig**, 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 1 kb upstream of the TSS to 500 bp 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 **Fig 2C** and **S5 Fig**, 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

352 timepoints to determine the threshold of R^2 (i.e., percentage of total variance that is explained
353 from fitting the time-series model). For the identified significant genes that exhibit dynamic repair
354 across timepoints, we further carried out gene ontology (GO) analysis to identify significantly
355 enriched terms in both biological processes and cellular components [44].
356
357 ***Capped RNA-seq and epigenomic data***
358 Capped RNA-seq captures nuclear RNAs that are with or without poly(A) tails and is thus much
359 more sensitive in detecting non-coding RNAs compared to RNA-seq. We took advantage of short-
360 and long-capped RNA-seq data of wildtype L1 *C. elegans* that are strand-specific [5]. Additionally,
361 we accessed and cross-compared publicly available epigenomic profiles of L1 *C. elegans*,
362 including chromatin accessibility by ATAC-seq, DNase I hypersensitivity by DNase-seq, and
363 histone modifications (H3K4me1, H3K4me3, and H3K27me3) by ChIP-seq [5]. All data were
364 downloaded as processed bigWig files (**S3 Table**) and lifted over to ce11 when necessary. Regions
365 from the bigWig files were overlapped with the genomic bins, and scores from the bigWig files
366 were averaged, weighted by region widths, to yield the capped RNA-seq and epigenetic
367 measurements for each intergenic region.
368
369 ***Chromatin state, eRNA, lincRNA, and piRNA annotations***
370 The genic and intergenic regions of *C. elegans* (ce11) were annotated using the GenomicFeatures
371 R package in conjunction with the TxDb.Celegans.UCSC.ce11.refGene annotation package.
372 Chromatin states in the L3 stage of *C. elegans* were previously inferred, consisting of 20 distinct
373 states as detailed in **Fig 4A** and **S7 Fig** [45]. Evans *et al*. observed a high degree of similarity in
374 autosomal chromatin states between the embryonic and L3 larval stages of the worms. This
375 conservation of chromatin configuration allowed us to confidently use the chromatin state data
376 from the L3 stage for intersection with our L1 stage data, without compromising the integrity of
377 our analysis [45]. Each annotated chromatin region was mapped from ce10 to ce11 and intersected
378 with RNA-seq, capped RNA-seq, and XR-seq reads. For eRNAs, 90 % of which are bidirectionally
379 transcribed, non-polyadenylated and unspliced, we retrieved 505 annotated eRNAs in *C. elegans*
380 from the eRNAdb database [46, 47]. We removed eRNAs that overlap with either annotated genes
381 or blacklist regions, resulting in a total of 324 eRNAs, which are presented in **Fig 5A** and **S9A Fig**.
382 Similarly, we obtained 170 long intergenic non-coding RNAs (lincRNAs) in *C. elegans* from
383 existing annotations [48]. After lifting over the coordinates from ce6 to ce11 and filtering out ones
384 that overlap with genes or blacklist regions, we were left with 103 lincRNAs, which are visualized
385 in the **Fig 5B** and **S9B Fig**. We obtained 15,363 piRNAs in *C. elegans* from existing WormBase
386 WS282 annotations. Removing the piRNAs that overlap with genes or blacklist regions results in
387 10,757 intergenic piRNAs, which are shown in **Fig 5C** and **S9C Fig**.
388
389 **DISCUSSION**
390 Transcription-coupled repair appears to be universal in cellular organisms ranging from bacteria
391 to humans and has been studied in several model organisms [10, 22, 24, 49-54]. Multiple

392 methodologies have been developed to unravel the intricate mechanisms and required repair
393 factors [13]. Among these methods, XR-seq, distinguished by its whole-genome analysis at single-
394 nucleotide resolution, has been applied across a spectrum of organisms, including bacteria, yeast,
395 flies, plants, and mammals [13]. A previous study employing a qPCR assay, indicated the existence
396 of transcription-coupled repair in *C. elegans* [16], nevertheless, our study stands as a single-
397 nucleotide-resolution genome-wide UV-damage transcription-coupled repair map of this
398 important model organism. Furthermore, our investigation distinguishes itself by employing
399 transcription-coupled repair as a proxy for RNAPII transcription, and *xpc-1* XR-seq data to
400 effectively complement RNA-seq and capped RNA-seq datasets to offer a more comprehensive
401 view of transcription.
402
403 Leveraging the unique properties of XR-seq data, we aimed to delve into the realm of
404 intergenic transcription, a domain that has posed persistent challenges for conventional RNA-seq
405 methods. Based on the RNAPII disassociation model in response to UV-induced damage, RNAPII
406 encounters transcription blockage and initiates a process of transcription-coupled repair. During
407 this repair process, RNAPII dissociates from the DNA strand, facilitating the sequential removal
408 of lesions from the template in the 5' to 3' direction. This concerted repair mechanism eventually
409 leads to the clearance of adducts from the template, thereby enabling the synthesis of full-length
410 transcripts [26, 55]. To comprehensively investigate these intricate transcription dynamics, we
411 conducted XR-seq at six distinct timepoints, ranging from 5 minutes to 48 hours following UV
412 treatment. As a result, our dataset encompasses both transcription initiation and elongation events,
413 providing a comprehensive view of the entire transcriptional process.
414
415 Detection of non-coding RNAs has long been a formidable task due to their relatively low
416 abundance and inherent instability. The development of cutting-edge technologies, such as RNA
417 polymerase II chromatin immunoprecipitation coupled with high-throughput sequencing (RNAPII
418 ChIP-seq), Global Run-On sequencing (GRO-seq), Precision Run-On Sequencing (PRO-seq), and
419 a variety of methods for sequencing the 5'-anchored RNAs, has been driven by the desire to discern
420 nascent RNAs and ncRNAs with heightened precision [5, 6, 18, 56-59]. A comprehensive
421 evaluation of the strengths and limitations of these methods has been described elsewhere [60],
422 and in the context of *C. elegans* research, efforts to specifically target ncRNAs and identify TSS
423 have utilized 5'-capped RNA-sequencing methods, as reported in previous studies [6, 18, 45, 61-
424 63].
425
426 XR-seq presents a noteworthy advantage in its ability to directly detect transcription events
427 at the DNA level, thus circumventing the inherent limitations associated with indirect transcription
428 detection techniques such as RNA sequencing. These conventional methods are prone to
429 challenges stemming from the low abundance and instability of RNA molecules. Furthermore,
430 RNA sequencing is susceptible to sequence bias resulting from early transcriptional events that
431 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 short capRNA-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.

AUTHOR CONTRIBUTIONS

A.S. and L.L.-B. 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 and GSE262486. 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/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 helpful discussions and comments.

REFERENCES

1. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. Cell. 2014;157(1):77-94. doi: 10.1016/j.cell.2014.03.008. PubMed PMID: 24679528.
2. Schmid M, Jensen TH. Controlling nuclear RNA levels. Nat Rev Genet. 2018;19(8):518-29. doi: 10.1038/s41576-018-0013-2. PubMed 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. Epub 20230103. doi: 10.1038/s41580-022-00566-8. PubMed PMID: 36596869; PubMed Central PMCID: PMCPMC10213152.
4. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63. doi: 10.1038/nrg2484. PubMed PMID: 19015660; PubMed Central PMCID: PMCPMC2949280.
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. Epub 20181026. doi: 10.7554/eLife.37344. PubMed PMID: 30362940; PubMed Central PMCID: PMCPMC6231769.
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. Epub 20130402. doi: 10.1101/gr.153668.112. PubMed PMID: 23550086; PubMed Central PMCID: PMCPMC3730107.
7. Hu J, Li W, Adebali O, Yang Y, Oztas O, Selby CP, Sancar A. Genome-wide mapping of nucleotide excision repair with XR-seq. Nat Protoc. 2019;14(1):248-82. doi: 10.1038/s41596-018-0093-7. PubMed PMID: 30552409; PubMed Central PMCID: PMCPMC6429938.
8. Reardon JT, Sancar A. Nucleotide excision repair. Prog Nucleic Acid Res Mol Biol. 2005;79:183-235. Epub 2005/08/13. doi: 10.1016/S0079-6603(04)79004-2. PubMed 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. doi: 10.1073/pnas.89.8.3664. PubMed PMID: 1314396; PubMed Central PMCID: PMCPMC48929.
10. Sancar A. Mechanisms of DNA Repair by Photolyase and Excision Nuclease (Nobel Lecture). Angew Chem Int Ed Engl. 2016;55(30):8502-27. doi: 10.1002/anie.201601524. PubMed PMID: 27337655.

510 11. Lans H, Vermeulen W. Nucleotide Excision Repair in Caenorhabditis elegans. Mol Biol
511 Int. 2011;2011:542795. Epub 20110817. doi: 10.4061/2011/542795. PubMed PMID: 22091407;
512 PubMed Central PMCID: PMCPMC3195855.
513 12. Lopes AFC, Bozek K, Herholz M, Trifunovic A, Rieckher M, Schumacher B. A C. elegans
514 model for neurodegeneration in Cockayne syndrome. Nucleic Acids Res. 2020;48(19):10973-85.
515 doi: 10.1093/nar/gkaa795. PubMed PMID: 33021672; PubMed Central PMCID:
516 PMCPMC7641758.
517 13. Selby CP, Lindsey-Boltz LA, Li W, Sancar A. Molecular Mechanisms of Transcription-
518 Coupled Repair. Annu Rev Biochem. 2023; 92:115-44. Epub 20230331. doi: 10.1146/annurev-
519 biochem-041522-034232. PubMed PMID: 37001137.
520 14. Lans H, Marteijn JA, Schumacher B, Hoeijmakers JH, Jansen G, Vermeulen W.
521 Involvement of global genome repair, transcription coupled repair, and chromatin remodeling in
522 UV DNA damage response changes during development. PLoS Genet. 2010;6(5):e1000941. doi:
523 10.1371/journal.pgen.1000941. PubMed PMID: 20463888; PubMed Central PMCID:
524 PMCPMC2865526.
525 15. Boyd WA, Crocker TL, Rodriguez AM, Leung MC, Lehmann DW, Freedman JH, et al.
526 Nucleotide excision repair genes are expressed at low levels and are not detectably inducible in
527 Caenorhabditis elegans somatic tissues, but their function is required for normal adult life after
528 UVC exposure. Mutat Res. 2010;683(1-2):57-67. doi: 10.1016/j.mrfmmm.2009.10.008. PubMed
529 PMID: 19879883; PubMed Central PMCID: PMCPMC2799044.
530 16. Meyer JN, Boyd WA, Azzam GA, Haugen AC, Freedman JH, Van Houten B. Decline of
531 nucleotide excision repair capacity in aging Caenorhabditis elegans. Genome Biol. 2007;8(5):R70.
532 doi: 10.1186/gb-2007-8-5-r70. PubMed PMID: 17472752; PubMed Central PMCID:
533 PMCPMC1929140.
534 17. van der Woude M, Lans H. C. elegans survival assays to discern global and transcription-
535 coupled nucleotide excision repair. STAR Protoc. 2021;2(2):100586. Epub 20210608. doi:
536 10.1016/j.xpro.2021.100586. PubMed PMID: 34151304; PubMed Central PMCID:
537 PMCPMC8192855.
538 18. Gu W, Lee HC, Chaves D, Youngman EM, Pazour GJ, Conte D, Jr., Mello CC. CapSeq
539 and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA
540 precursors. Cell. 2012;151(7):1488-500. doi: 10.1016/j.cell.2012.11.023. PubMed PMID:
541 23260138; PubMed Central PMCID: PMCPMC3581324.
542 19. Hu J, Adar S, Selby CP, Lieb JD, Sancar A. Genome-wide analysis of human global and
543 transcription-coupled excision repair of UV damage at single-nucleotide resolution. Genes Dev.
544 2015;29(9):948-60. doi: 10.1101/gad.261271.115. PubMed PMID: 25934506; PubMed Central
545 PMCID: PMCPMC4421983.
546 20. Kose C, Cao X, Dewey EB, Malkoc M, Adebali O, Sekelsky J, et al. Cross-species
547 investigation into the requirement of XPA for nucleotide excision repair. Nucleic Acids Res. 2023.
548 Epub 20231122. doi: 10.1093/nar/gkad1104. PubMed PMID: 37994737.

549 21. Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, et al. C. elegans whole-
550 genome sequencing reveals mutational signatures related to carcinogens and DNA repair
551 deficiency. Genome Res. 2014;24(10):1624-36. Epub 20140716. doi: 10.1101/gr.175547.114.
552 PubMed PMID: 25030888; PubMed Central PMCID: PMCPMC4199376.
553 22. Akkose U, Kaya VO, Lindsey-Boltz L, Karagoz Z, Brown AD, Larsen PA, et al.
554 Comparative analyses of two primate species diverged by more than 60 million years show
555 different rates but similar distribution of genome-wide UV repair events. BMC Genomics.
556 2021;22(1):600. Epub 2021/08/08. doi: 10.1186/s12864-021-07898-3. PubMed PMID: 34362292;
557 PubMed Central PMCID: PMCPMC8349011.
558 23. Oztas O, Selby CP, Sancar A, Adebali O. Genome-wide excision repair in Arabidopsis is
559 coupled to transcription and reflects circadian gene expression patterns. Nat Commun.
560 2018;9(1):1503. doi: 10.1038/s41467-018-03922-5. PubMed PMID: 29666379; PubMed Central
561 PMCID: PMCPMC5904149.
562 24. Deger N, Yang Y, Lindsey-Boltz LA, Sancar A, Selby CP. Drosophila, which lacks
563 canonical transcription-coupled repair proteins, performs transcription-coupled repair. J Biol
564 Chem. 2019;294(48):18092-8. doi: 10.1074/jbc.AC119.011448. PubMed PMID: 31624146;
565 PubMed Central PMCID: PMCPMC6885609.
566 25. Lindsey-Boltz LA, Yang Y, Kose C, Deger N, Eynullazada K, Kawara H, Sancar A.
567 Nucleotide excision repair in Human cell lines lacking both XPC and CSB proteins. Nucleic Acids
568 Res. 2023;51(12):6238–45. Epub 20230505. doi: 10.1093/nar/gkad334. PubMed PMID:
569 37144462.
570 26. Chiou YY, Hu J, Sancar A, Selby CP. RNA polymerase II is released from the DNA
571 template during transcription-coupled repair in mammalian cells. J Biol Chem. 2018;293(7):2476-
572 86. doi: 10.1074/jbc.RA117.000971. PubMed PMID: 29282293; PubMed Central PMCID:
573 PMCPMC5818198.
574 27. Deger N, Cao X, Selby CP, Gulec S, Kawara H, Dewey EB, et al. CSB-independent, XPC-
575 dependent transcription-coupled repair in Drosophila. Proc Natl Acad Sci U S A. 2022;119(9). doi:
576 10.1073/pnas.2123163119. PubMed PMID: 35217627; PubMed Central PMCID:
577 PMCPMC8892495.
578 28. Adebali O, Yang Y, Neupane P, Dike NI, Boltz JL, Kose C, et al. The Mfd protein is the
579 transcription-repair coupling factor (TRCF) in Mycobacterium smegmatis. J Biol Chem.
580 2023;299(3):103009. Epub 20230211. doi: 10.1016/j.jbc.2023.103009. PubMed PMID: 36775124;
581 PubMed Central PMCID: PMCPMC10023983.
582 29. Yimit A, Adebali O, Sancar A, Jiang Y. Differential damage and repair of DNA-adducts
583 induced by anti-cancer drug cisplatin across mouse organs. Nat Commun. 2019;10(1):309. doi:
584 10.1038/s41467-019-08290-2. PubMed PMID: 30659176; PubMed Central PMCID:
585 PMCPMC6338751.
586 30. Yang Y, Lindsey-Boltz LA, Vaughn CM, Selby CP, Cao X, Liu Z, et al. Circadian clock,
587 carcinogenesis, chronochemotherapy connections. J Biol Chem. 2021;297(3):101068. Epub

2021/08/11. doi: 10.1016/j.jbc.2021.101068. PubMed PMID: 34375638; PubMed Central PMCID: PMCPMC8403766.

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. doi: 10.1074/jbc.RA119.009579. PubMed PMID: 31217280; PubMed Central PMCID: PMCPMC6690688.

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. doi: 10.1073/pnas.1804493115. PubMed PMID: 29735688.

33. Yang Y, Hu J, Selby CP, Li W, Yimit A, Jiang Y, Sancar A. Single-nucleotide resolution analysis of nucleotide excision repair of ribosomal DNA in humans and mice. J Biol Chem. 2019;294(1):210-7. doi: 10.1074/jbc.RA118.006121. PubMed PMID: 30413533; PubMed Central PMCID: PMCPMC6322885.

34. Li W, Liu W, Kakoki A, Wang R, Adebali O, Jiang Y, Sancar A. 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. Epub 20190226. doi: 10.1074/jbc.RA119.007861. PubMed PMID: 30808711; PubMed Central PMCID: PMCPMC6463700.

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. Epub 2021/03/28. doi: 10.1016/j.jbc.2021.100581. PubMed PMID: 33771559; PubMed Central PMCID: PMCPMC8081918.

36. Green MR, Sambrook J. Total RNA Extraction from Caenorhabditis elegans. Cold Spring Harb Protoc. 2020;2020(9):101683. Epub 20200901. doi: 10.1101/pdb.prot101683. PubMed PMID: 32873731.

37. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17(1):10-2. Epub 2011-08-02. doi: 10.14806/ej.17.1.200.

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

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. Epub 20121025. doi: 10.1093/bioinformatics/bts635. PubMed PMID: 23104886; PubMed Central PMCID: PMCPMC3530905.

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. Epub 20131113. doi: 10.1093/bioinformatics/btt656. PubMed 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. Epub 20190627. doi:

627 10.1038/s41598-019-45839-z. PubMed PMID: 31249361; PubMed Central PMCID:
628 PMCPMC6597582.
629 42. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for
630 exploring deep-sequencing data. Nucleic Acids Res. 2014;42(Web Server issue):W187-91. Epub
631 20140505. doi: 10.1093/nar/gku365. PubMed PMID: 24799436; PubMed Central PMCID:
632 PMCPMC4086134.
633 43. Bacher R, Leng N, Chu LF, Ni Z, Thomson JA, Kendziorski C, Stewart R. Trendy:
634 segmented regression analysis of expression dynamics in high-throughput ordered profiling
635 experiments. BMC Bioinformatics. 2018;19(1):380. Epub 20181016. doi: 10.1186/s12859-018-
636 2405-x. PubMed PMID: 30326833; PubMed Central PMCID: PMCPMC6192113.
637 44. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology:
638 tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.
639 doi: 10.1038/75556. PubMed PMID: 10802651; PubMed Central PMCID: PMCPMC3037419.
640 45. Evans KJ, Huang N, Stempor P, Chesney MA, Down TA, Ahringer J. Stable Caenorhabditis
641 elegans chromatin domains separate broadly expressed and developmentally regulated genes. Proc
642 Natl Acad Sci U S A. 2016;113(45):E7020-E9. Epub 20161025. doi: 10.1073/pnas.1608162113.
643 PubMed PMID: 27791097; PubMed Central PMCID: PMCPMC5111720.
644 46. Jin W, Jiang G, Yang Y, Yang J, Yang W, Wang D, et al. Animal-eRNAdb: a comprehensive
645 animal enhancer RNA database. Nucleic Acids Res. 2022;50(D1):D46-D53. doi:
646 10.1093/nar/gkab832. PubMed PMID: 34551433; PubMed Central PMCID: PMCPMC8728245.
647 47. Sartorelli V, Lauberth SM. Enhancer RNAs are an important regulatory layer of the
648 epigenome. Nat Struct Mol Biol. 2020;27(6):521-8. Epub 20200608. doi: 10.1038/s41594-020-
649 0446-0. PubMed PMID: 32514177; PubMed Central PMCID: PMCPMC7343394.
650 48. Nam JW, Bartel DP. Long noncoding RNAs in C. elegans. Genome Res.
651 2012;22(12):2529-40. Epub 20120615. doi: 10.1101/gr.140475.112. PubMed PMID: 22707570;
652 PubMed Central PMCID: PMCPMC3514682.
653 49. Ogrunc M, Becker DF, Ragsdale SW, Sancar A. Nucleotide excision repair in the third
654 kingdom. J Bacteriol. 1998;180(21):5796-8. Epub 1998/10/29. doi: 10.1128/JB.180.21.5796-
655 5798.1998. PubMed PMID: 9791138; PubMed Central PMCID: PMCPMC107647.
656 50. Li W, Adebali O, Yang Y, Selby CP, Sancar A. Single-nucleotide resolution dynamic repair
657 maps of UV damage in Saccharomyces cerevisiae genome. Proc Natl Acad Sci U S A.
658 2018;115(15):E3408-E15. doi: 10.1073/pnas.1801687115. PubMed PMID: 29581276; PubMed
659 Central PMCID: PMCPMC5899493.
660 51. Selby CP, Lindsey-Boltz LA, Yang Y, Sancar A. Mycobacteria excise DNA damage in 12-
661 or 13-nucleotide-long oligomers by prokaryotic-type dual incisions and performs transcription-
662 coupled repair. J Biol Chem. 2020;295(50):17374-80. Epub 2020/10/23. doi:
663 10.1074/jbc.AC120.016325. PubMed PMID: 33087442.
664 52. Canturk F, Karaman M, Selby CP, Kemp MG, Kulaksiz-Erkmen G, Hu J, et al. Nucleotide
665 excision repair by dual incisions in plants. Proc Natl Acad Sci U S A. 2016;113(17):4706-10. doi:

666 10.1073/pnas.1604097113. PubMed PMID: 27071131; PubMed Central PMCID:
667 PMCPMC4855589.
668 53. Mellon I, Spivak G, Hanawalt PC. Selective removal of transcription-blocking DNA
669 damage from the transcribed strand of the mammalian DHFR gene. Cell. 1987;51(2):241-9. Epub
670 1987/10/23. doi: 10.1016/0092-8674(87)90151-6. PubMed PMID: 3664636.
671 54. Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and
672 surprises. Nat Rev Mol Cell Biol. 2008;9(12):958-70. doi: 10.1038/nrm2549. PubMed PMID:
673 19023283.
674 55. Nicholson C, Odom D, Aitken S, Taylor M. DNA lesion bypass and the stochastic dynamics
675 of transcription coupled repair. Proc Natl Acad Sci U S A. 2024;In Press.
676 56. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA
677 identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat
678 Genet. 2014;46(12):1311-20. doi: 10.1038/ng.3142. PubMed PMID: 25383968; PubMed Central
679 PMCID: PMCPMC4254663.
680 57. Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, et al. Base-pair-
681 resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on
682 (PRO-seq). Nat Protoc. 2016;11(8):1455-76. Epub 20160721. doi: 10.1038/nprot.2016.086.
683 PubMed PMID: 27442863; PubMed Central PMCID: PMCPMC5502525.
684 58. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, et al. A large fraction of
685 extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. 2010;8(5):e1000384.
686 Epub 20100511. doi: 10.1371/journal.pbio.1000384. PubMed PMID: 20485488; PubMed Central
687 PMCID: PMCPMC2867938.
688 59. Morioka MS, Kawaji H, Nishiyori-Sueki H, Murata M, Kojima-Ishiyama M, Carninci P,
689 Itoh M. Cap Analysis of Gene Expression (CAGE): A Quantitative and Genome-Wide Assay of
690 Transcription Start Sites. Methods Mol Biol. 2020;2120:277-301. doi: 10.1007/978-1-0716-0327-
691 7_20. PubMed PMID: 32124327.
692 60. Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: recent
693 insights and future perspectives. Nat Rev Genet. 2016;17(4):207-23. Epub 20160307. doi:
694 10.1038/nrg.2016.4. PubMed PMID: 26948815.
695 61. Cecere G, Hoersch S, O'Keeffe S, Sachidanandam R, Grishok A. Global effects of the CSR-
696 1 RNA interference pathway on the transcriptional landscape. Nat Struct Mol Biol.
697 2014;21(4):358-65. Epub 20140330. doi: 10.1038/nsmb.2801. PubMed PMID: 24681887;
698 PubMed Central PMCID: PMCPMC4068146.
699 62. Cecere G, Hoersch S, Jensen MB, Dixit S, Grishok A. The ZFP-1(AF10)/DOT-1 complex
700 opposes H2B ubiquitination to reduce Pol II transcription. Mol Cell. 2013;50(6):894-907. doi:
701 10.1016/j.molcel.2013.06.002. PubMed PMID: 23806335; PubMed Central PMCID:
702 PMCPMC3784254.
703 63. Saito TL, Hashimoto S, Gu SG, Morton JJ, Stadler M, Blumenthal T, et al. The transcription
704 start site landscape of C. elegans. Genome Res. 2013;23(8):1348-61. Epub 20130501. doi:
705 10.1101/gr.151571.112. PubMed PMID: 23636945; PubMed Central PMCID: PMCPMC3730108.

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. Epub 20140220. doi: 10.1016/j.celrep.2014.01.037. PubMed PMID: 24561252; PubMed Central PMCID: PMCPMC4918108.

65. 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. 2011;333(6038):53-8. Epub 20110519. doi: 10.1126/science.1207018. PubMed PMID: 21596952; PubMed Central PMCID: PMCPMC3204392.

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. doi: 10.1093/nar/gkab1290. PubMed PMID: 35037068; PubMed Central PMCID: PMCPMC8860604.

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. Epub 20230925. doi: 10.1038/s41467-023-41556-4. PubMed PMID: 37749091; PubMed Central PMCID: PMCPMC10520050.

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. Epub 20230829. doi: 10.1038/s41467-023-40957-9. PubMed PMID: 37644049; PubMed Central PMCID: PMCPMC10465613.

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. Epub 20231002. doi: 10.1038/s41467-023-41881-8. PubMed PMID: 37783675; PubMed Central PMCID: PMCPMC10545829.

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. Epub 20201219. doi: 10.15252/embj.2020105564. PubMed PMID: 33340372; PubMed Central PMCID: PMCPMC7917550.

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. doi: 10.1101/gad.238105.114. PubMed PMID: 24696457; PubMed Central PMCID: PMCPMC4015492.

FIGURE LEGENDS

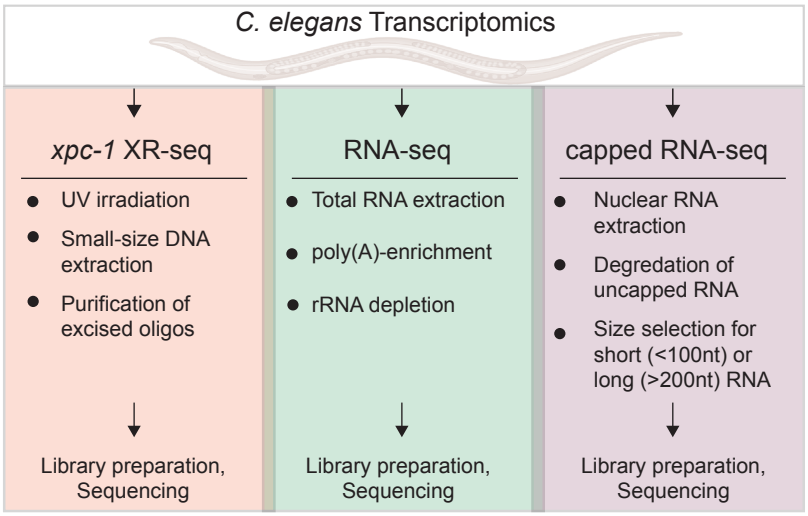
Fig 1. Overview of Study Design. (A) The illustration highlights key properties of the three comparative transcriptomic techniques (XR-seq, RNA-seq, capped RNA-seq) analyzed in this study for their capacity to identify genome-wide transcription in *C. elegans*. **(B)** Nucleotide excision repair removes DNA damage through two different mechanisms: global repair and transcription-coupled repair (TCR). Global repair depends on the XPC protein and occurs

746 throughout the genome, whereas TCR is independent of XPC and only occurs when elongating
747 RNA polymerase II encounters damage during transcription and recruits the CSB protein. This
748 study uses XR-seq to map nucleotide excision repair at single-nucleotide resolution throughout the
749 whole-genome in three strains of *C. elegans*: wild-type, *csb-1*, and *xpc-1*. Because *xpc-1* worms
750 lack global repair, analysis of XR-seq data from this strain provides a unique opportunity to map
751 transcription genome-wide independent of RNA capture.
752
753 **Fig 2. Detection of Transcription-Coupled Repair by XR-seq. (A)** Browser view of the
754 distribution of *C. elegans* high throughput sequencing reads separated by strand over a
755 representative 5.2 kb region from chromosome I. RNA-seq reads (green) from wild-type (WT)
756 worms is shown on top to illustrate the opposite direction of transcription of the genes *vbh-1* and
757 *mrpl-17*. The strand distribution of XR-seq reads (orange) 1 hour after UV clearly demonstrates
758 the occurrence of transcription-coupled repair within the body of both genes in WT and *xpc-1*
759 worms but absent in *csb-1*. **(B)** To analyze transcription-coupled repair genome-wide, XR-seq
760 reads on transcribed strand (TS) and non-transcribed strand (NTS) in the indicated strains at 1 hour
761 repair time is plotted with mean reads per kilobase per million mapped reads (RPKM) (y-axis)
762 along the 500 bp upstream and 1 kb downstream of transcription start site (TSS), and 1 kb upstream
763 and 500 bp downstream of transcription end site (TES) (x-axis) for 2,142 genes selected for length >
764 2 kb and no overlaps with a distance of at least 500 bp between genes. The TT-distribution, as
765 mean TT content (y-axis) was determined for the same gene set and is plotted at the bottom as a
766 measure of expected DNA damage sites. **(C)** Profile plots and heatmaps of TS and NTS XR-seq
767 reads from the *xpc-1* strain at 1 hour repair time spanning the best represented half of 16,588 TSSs >
768 1 kb apart indicate divergent transcription at promoters.
769
770 **Fig 3. Transcription-Coupled Repair Reveals Transcription in Intergenic Regions. (A)**
771 Browser view of stranded read distribution from WT and *xpc-1* RNA-seq (green), capped RNA-
772 seq (pink), and *xpc-1* XR-seq (orange) at 1 hour repair time over a representative 27 kb region
773 from chromosome I. Both capped RNA-seq and XR-seq methods provide comprehensive coverage
774 of the entire window, encompassing both genic and intergenic regions, in contrast to RNA-seq
775 which only captures polyadenylated mRNAs. **(B)** Bar graphs depict the genome-wide distribution
776 of reads obtained from the different sequencing methods, including WT and *xpc-1* RNA-seq, long-
777 capped RNA-seq, short-capped RNA-seq, and *xpc-1* XR-seq at 1 hour repair time. Notably, both
778 XR-seq and capped RNA-seq techniques reveal transcription events occurring outside of the
779 defined genic regions (see Materials and Methods for details).
780
781 **Fig 4. The Transcription-Coupled Repair in Intergenic Regions Detected by *xpc-1* XR-seq is**
782 **Supported by Epigenomic Signatures. (A)** Reads from XR-seq at 1 hour repair time, capped
783 RNA-seq, and RNA-seq were analyzed for overlap with genomic intervals corresponding to 20
784 distinct predicted chromatin states in *C. elegans*. The proportion of reads was computed for each
785 of the annotated chromatin states and the square root of the proportion is visualized as a heatmap.

786 **(B)** Examination of intergenic XR-seq reads, which are undetectable by RNA-seq, in association
787 with ATAC-seq, DNase-seq, H3K4me1, H3K4me3, and H3K27me3 peaks. XR-seq reads exhibit
788 a strong correlation with active transcription markers, in contrast to the repressive marker
789 H4K27me3, when compared to randomly selected genomic regions. All p-values obtained are
790 highly significant (< 2.2e-16) according to nonparametric Wilcoxon rank sum tests.
791
792 **Fig 5. XR-seq Reveals Transcription-Coupled Repair in Intergenic eRNAs, lincRNAs, and**
793 **piRNAs. (A)** Heatmaps (left) display normalized RNA expression and transcription-coupled repair
794 for intergenic enhancer RNAs (eRNAs) segregated by chromosomes. Normalization by log(x+1)
795 was carried out, where x is library-size-adjusted read count. Bar graphs (right) represent log-
796 normalized read counts for eRNA. Data are presented for WT and *xpc-1* RNA-seq, WT long- and
797 short-capped RNA-seq, and time-course combined *xpc-1* XR-seq dataset (5min, 1h, 8h, 16h, 24h,
798 and 48h). **(B, C)** Heatmaps and bar graphs as in A, for long intergenic non-coding RNAs
799 (lincRNAs) and intergenic Piwi-interacting RNAs (piRNAs), respectively.
800
801 **Fig 6. XR-seq identifies intergenic transcription-coupled repair in high concordance with**
802 **intergenic transcription identified by capped RNA-seq and reveals novel sites of**
803 **transcription.** For 85,418 intergenic bins, we identified regions with non-zero read counts by
804 short- or long-capped RNA-seq, RNA-seq, and time-course (5mins, 1h, 8h, 16h, 24h, and 48h
805 repair times) *xpc-1* XR-seq. **(A)** Venn diagram of intergenic bins detected by capped RNA-seq,
806 conventional RNA-seq, and *xpc-1* XR-seq. To reduce the number of call sets, we required non-
807 zero read counts to be detected: (i) in both replicates for *xpc-1* XR-seq; (ii) in both WT and *xpc-1*
808 RNA-seq, as they are highly correlated; and (iii) by either short-capped or long-capped RNA-seq,
809 as they are complementary. **(B)** Pie chart summary of the 7,903 bins unique to *xpc-1* XR-seq. 34.7%
810 have been annotated (dark blue) and the remaining 65.3% have not been annotated (light blue)
811 according to the WormBase WS282 annotations. The distribution of chromosomal locations (I-X)
812 is indicated for the unannotated bines. The 68% of annotated bins map to chromosome IV which
813 is not indicated. **(C)** Pie chart summary of the bins overlapping 2,722 annotations unique to *xpc-1*
814 XR-seq dataset. The majority of the unique annotated bins contain piRNAs from chromosome IV,
815 with the remainder consisting of pseudogenes, protein coding regions, eRNAs, lincRNAs, and
816 nRNAs. The 'other' category consists of RNAs excluded from the capped RNA-seq dataset
817 (snRNA, tRNAs, rRNAs) only contains 1.5% of bins. **(D)** Bin distribution along chromosome IV
818 of unique to the *xpc-1* XR-seq dataset-unannotated bins (top in blue), unique to the *xpc-1* XR-seq
819 dataset-bins with piRNA annotations (middle in burgundy), and intergenic piRNAs from
820 WormBase WS282 annotations (bottom in black).

Fig. 1

A


C. elegans Transcriptomics
xpc-1 XR-seq
UV irradiation
Small-size DNA extraction
Purification of excised oligos
Library preparation, Sequencing
RNA-seq
Total RNA extraction
poly(A)-enrichment
rRNA depletion
Library preparation, Sequencing
capped RNA-seq
Nuclear RNA extraction
Degredation of uncapped RNA
Size selection for short (<100nt) or long (>200nt) RNA
Library preparation, Sequencing

B

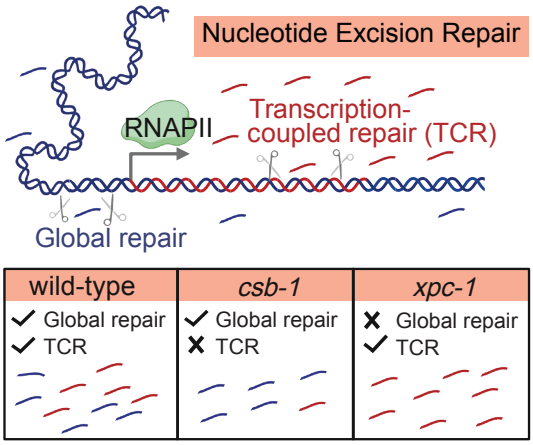

Nucleotide Excision Repair
RNAPII
Transcription-coupled repair (TCR)
Global repair
wild-type
✓ Global repair
✓ TCR
csb-1
✓ Global repair
✗ TCR
xpc-1
✗ Global repair
✓ TCR

Fig. 2

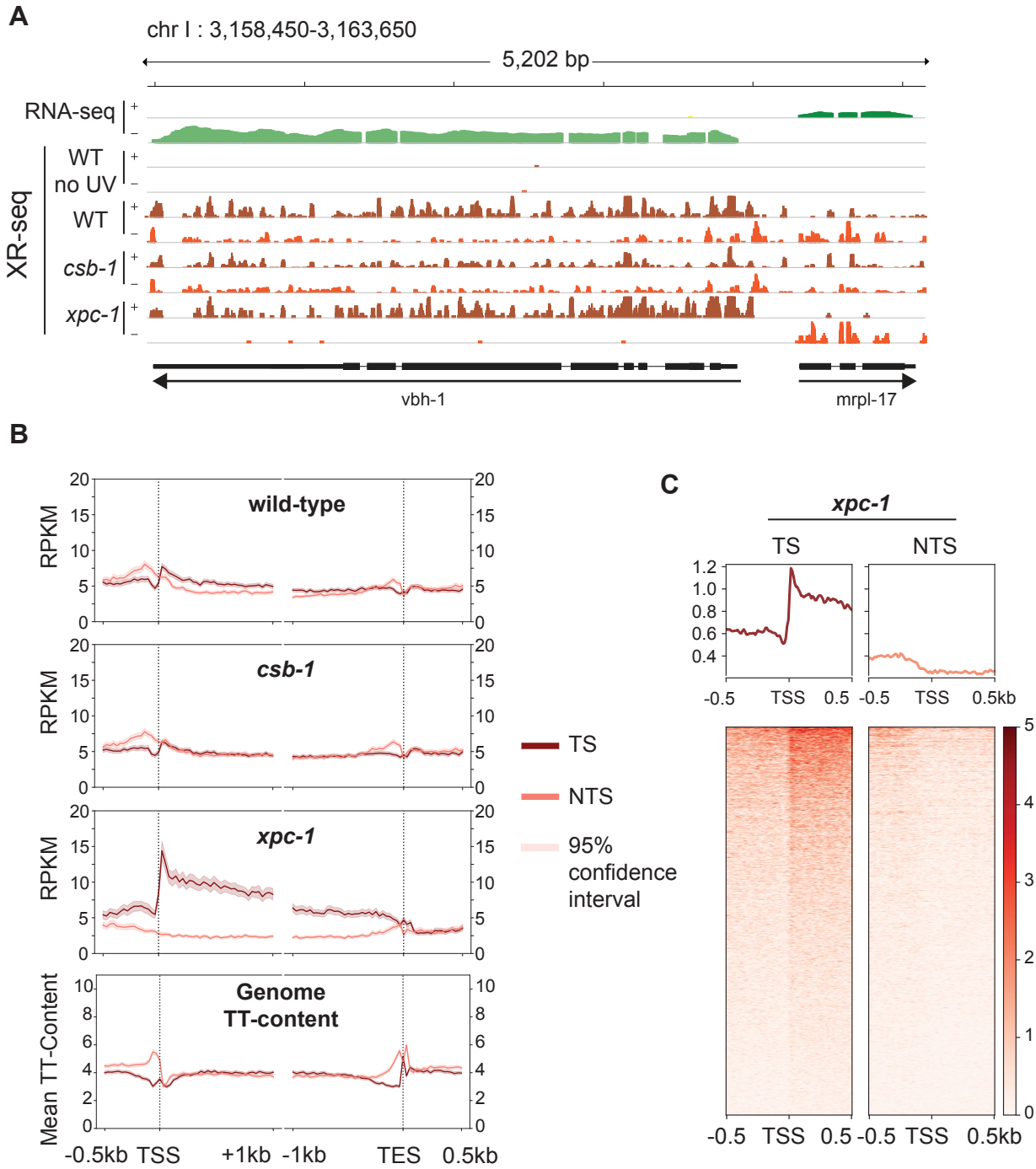
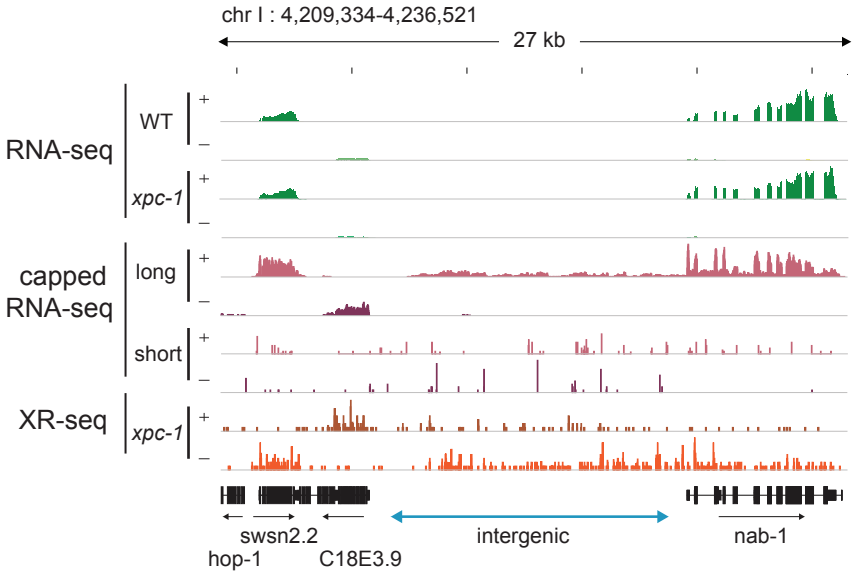

A
chr I : 3,158,450-3,163,650
5,202 bp
RNA-seq
XR-seq
WT no UV
WT
csb-1
xpc-1
vbh-1
mrpl-17
B
wild-type
csb-1
xpc-1
RPKM
Genome TT-content
Mean TT-Content
-0.5kb TSS
+1kb -1kb
TES 0.5kb
TS
NTS
95% confidence interval
C
xpc-1
TS
NTS
-0.5 TSS 0.5
-0.5 TSS 0.5kb

Fig. 3

A


chr I : 4,209,334-4,236,521
27 kb
RNA-seq
WT
xpc-1
capped RNA-seq
long
short
XR-seq
xpc-1
+
−
swsn2.2
hop-1
C18E3.9
intergenic
nab-1

B

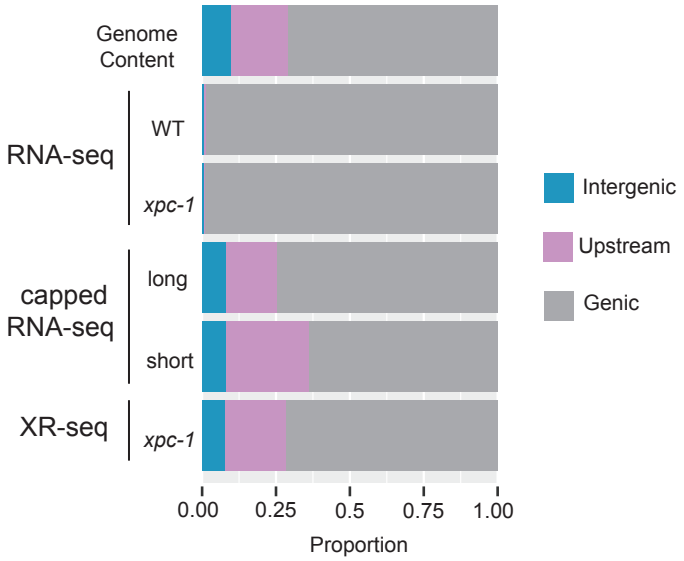
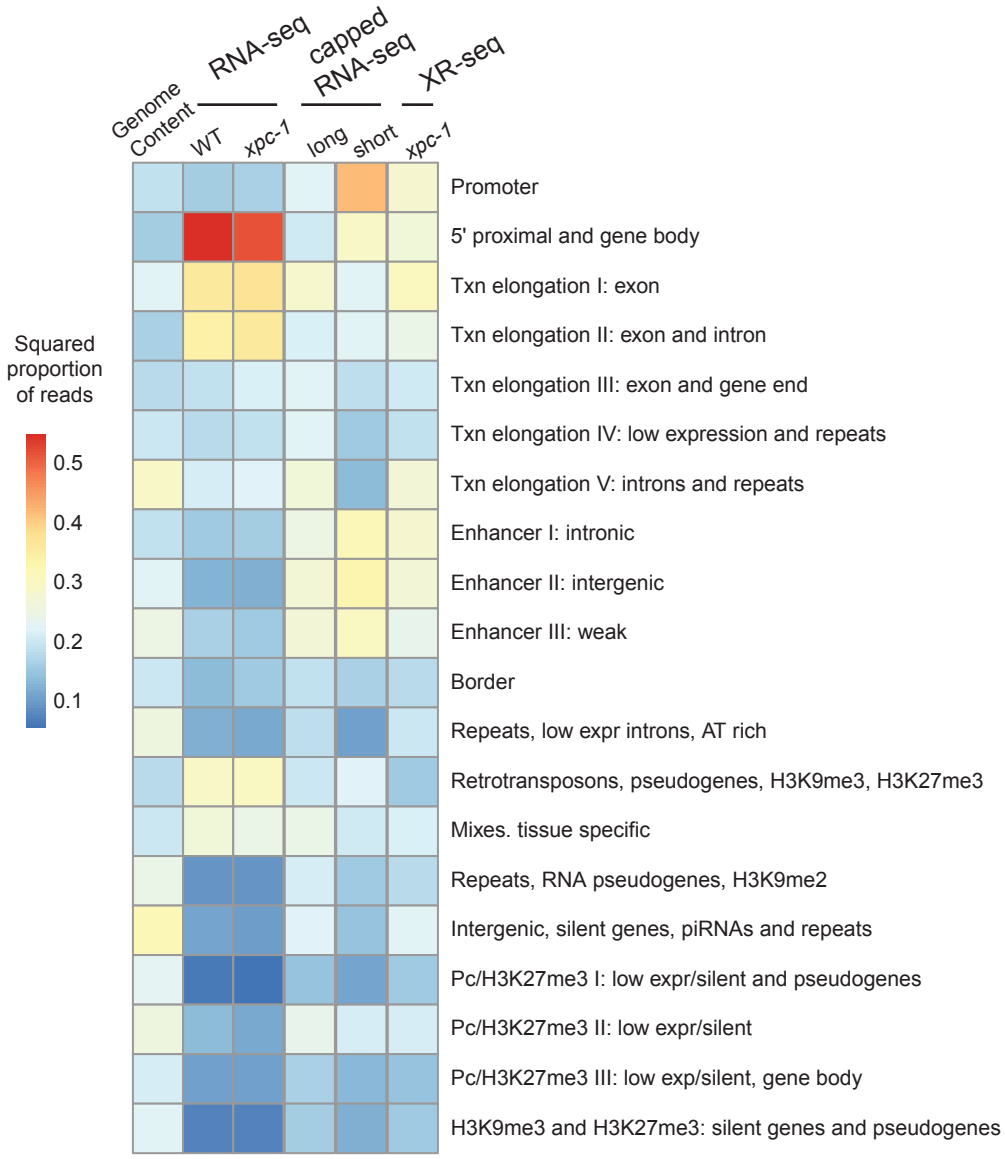

Genome Content
RNA-seq
WT
xpc-1
capped RNA-seq
long
short
XR-seq
xpc-1
Intergenic
Upstream
Genic
0.00
0.25
0.5
0.75
1.00
Proportion

Fig. 4

A


Genome Content
RNA-seq
capped RNA-seq
XR-seq
WT
xpc-1
long
short
xpc-1
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

B

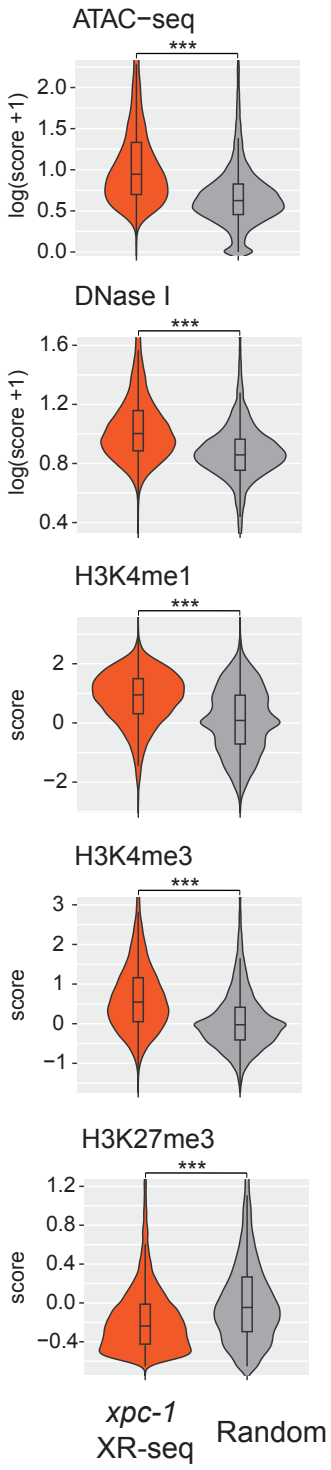

ATAC−seq
***
log(score +1)
2.0
1.5
1.0
0.5
0.0
DNase I
***
log(score +1)
1.6
1.2
0.8
0.4
H3K4me1
***
score
2
0
−2
H3K4me3
***
score
3
2
1
0
−1
H3K27me3
***
score
1.2
0.8
0.4
0.0
−0.4
xpc-1
XR-seq
Random

Fig. 5

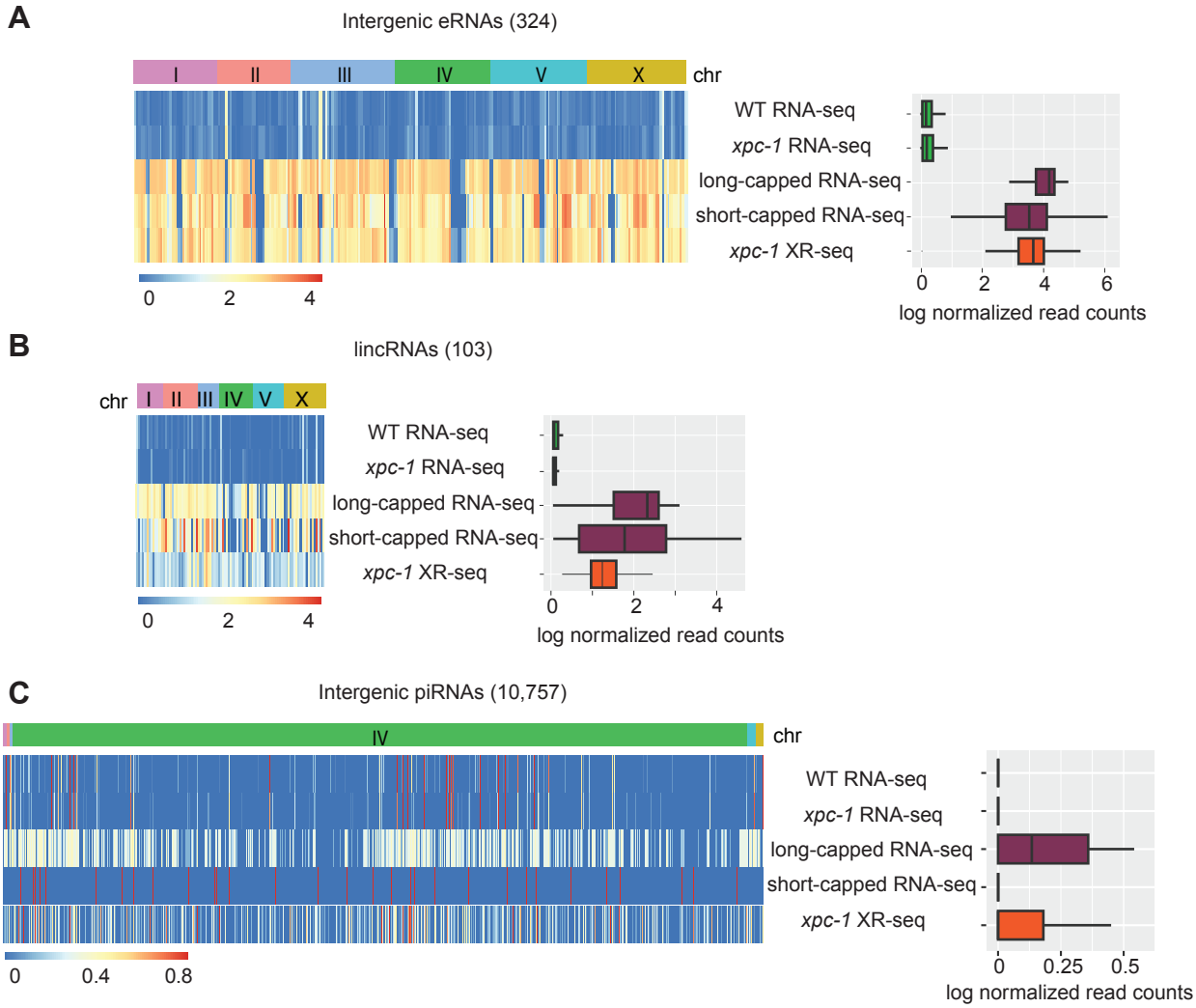
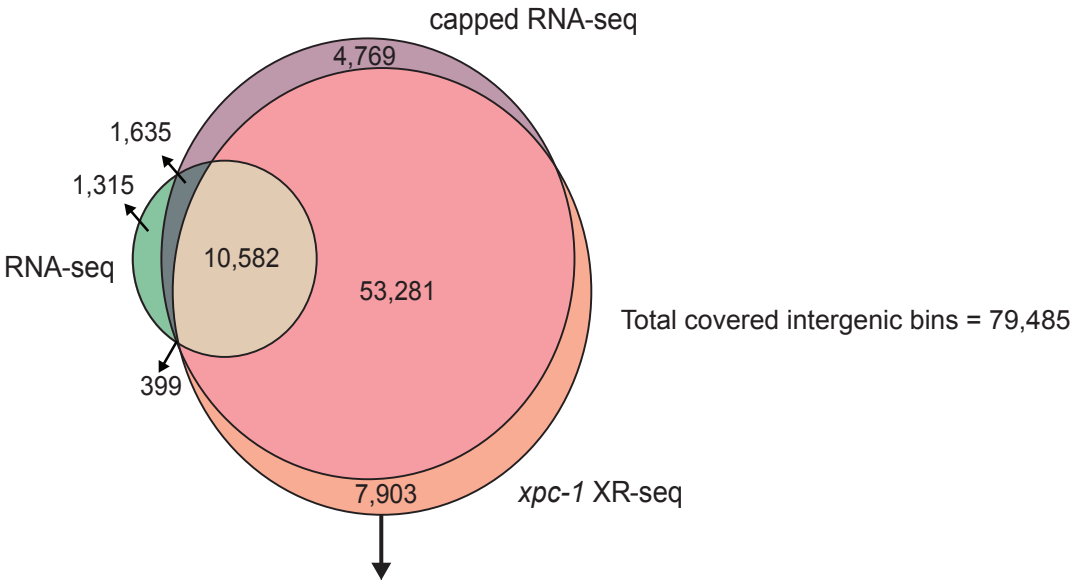
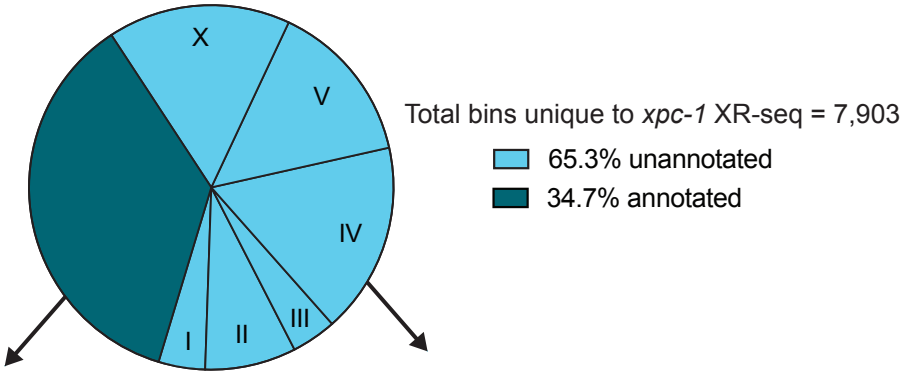

A
Intergenic eRNAs (324)
I
II
III
IV
V
X
chr
0
2
4
WT RNA-seq
xpc-1 RNA-seq
long-capped RNA-seq
short-capped RNA-seq
xpc-1 XR-seq
0
2
4
6
log normalized read counts
B
lincRNAs (103)
chr
I
II
III
IV
V
X
0
2
4
WT RNA-seq
xpc-1 RNA-seq
long-capped RNA-seq
short-capped RNA-seq
xpc-1 XR-seq
0
2
4
log normalized read counts
C
Intergenic piRNAs (10,757)
IV
chr
0
0.4
0.8
WT RNA-seq
xpc-1 RNA-seq
long-capped RNA-seq
short-capped RNA-seq
xpc-1 XR-seq
0
0.25
0.5
log normalized read counts

Fig. 6

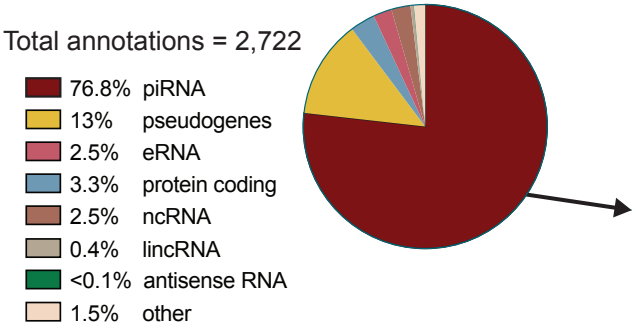
A


capped RNA-seq
4,769
1,635
1,315
RNA-seq
10,582
53,281
399
7,903
xpc-1 XR-seq
Total covered intergenic bins = 79,485

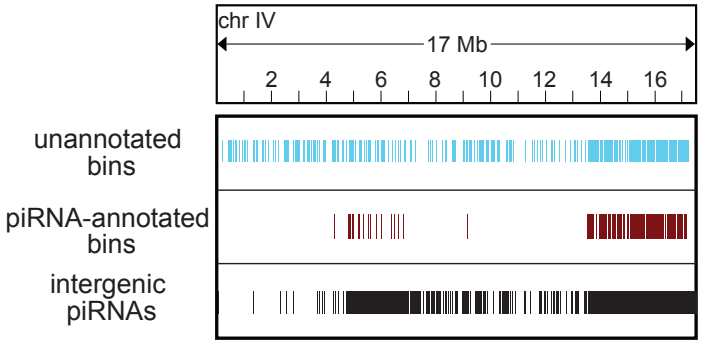
B


X
V
IV
III
II
I
Total bins unique to *xpc-1* XR-seq = 7,903
65.3% unannotated
34.7% annotated

C


Total annotations = 2,722
76.8% piRNA
13% pseudogenes
2.5% eRNA
3.3% protein coding
2.5% ncRNA
0.4% lincRNA
<0.1% antisense RNA
1.5% other

D


chr IV
17 Mb
2
4
6
8
10
12
14
16
unannotated bins
piRNA-annotated bins
intergenic piRNAs