A Bioinformatic Exploration of Unique Biological Features in *Sciara Coprophila*

Jaison Jain
Advisor: Dr. Susan Gerbi
Second Reader: Dr. Sohini Ramachandran

Senior Honors Thesis in Computational Biology

April 2020

Abstract

The lower dipteran fungus fly, *Sciara coprophila*, boasts unique biological properties shared by few others in the kingdom of life. Here, two such biological features were explored: non-disjunction of the X chromosome and resistance to irradiation.

While non-disjunction is generally pathologic in species, X chromosomal non-disjunction occurs naturally during spermatogenesis in male *Sciara*. The genomic element mediating non-disjunction, termed the *controlling element*, is known to be found within ribosomal DNA (rDNA) but its exact identity remains elusive. In order to further interrogate the controlling element, the ribosomal DNA of *Sciara* was characterized. In particular, Bayesian change-point analysis was employed to annotate previously-uncharacterized rDNA segments, and a modified global alignment algorithm was used to identify conserved islands of rDNA sequence. Furthermore, differential expression analysis of putative controlling elements within the rDNA tandem array was conducted.

A second unique biological feature of *Sciara* is its marked resistance to irradiation, able to withstand nearly twice the dose of X-irradiation as its dipteran relative, *Drosophila melanogaster*. RNA-seq analysis was conducted to better characterize the radiation response in *Sciara*. Irradiated samples exhibited robust upregulation of genes involved in nucleotide excision repair and the RNA-induced silencing complex, both offering insight into possible mechanisms of radioresistance in *Sciara*.

Part 1: An Exploration of rDNA and the Controlling Element

Introduction

My exploration of the ribosomal DNA (rDNA) and surrounding genomic elements in the fungus fly, *Sciara coprophila*, is motivated by the phenomenon of X chromosome non-disjunction. While chromosomal non-disjunction is normally pathologic in species, non-disjunction of the X chromosome is a normal part of spermatogenesis in male *Sciara.* Two X chromosomes are passed to offspring by the paternal parent and one by the materal parent, producing a triploid zygote. Subsequently, one or two X chromosomes (in females and males, respectively) are eliminated in somatic cells during the 8th embryonic cleavage. The process of chromosomal non-disjunction followed by elimination is reminiscent of uniparental disomy in humans, an abnormal phenomenon that contributes to conditions such as Angelman syndrome and Prader-Willi syndrome, both of which involve chromosome 15. Furthermore, non-disjunction underlies myriad human diseases including Klinefelter's (XXY genotype), Down syndrome (trisomy 21), and genomically unstable cancers. *Sciara* therefore offers a unique opportunity to study a complex pathologic phenomenon in a tractable organism.

The genomic element(s) mediating *Sciara* X non-disjunction, termed the *controlling element* (CE), has been mapped to the rDNA tandem repeats within the proximal heterochromatin of the X chromosome, which is subdivided into heterochromomeres Hc1, Hc2 and Hc3. In particular, through a series of translocation experiments, Hc2 was found to be necessary for non-disjunction, with X failing to undergo non-disjunction in the absence of Hc2 and with autosomes gaining the capacity for non-disjunction when containing Hc2 (Crouse, 1977; Crouse, 1979). It may be the case that the controlling element is a transcribed element residing in Hc2 that affects global chromosome dynamics, akin to the *Xist* long non-coding RNA that coats and inactivates the X chromosome in humans. An alternative explanation is that the translocation chromosome with the greatest number of rDNA repeats is the one that exhibits CE activity (Abbott & Gerbi, 1981).

In order to further interrogate the identity of the controlling element, the rDNA tandem array and intervening genomic elements were studied. In particular, the "R3" insertion within rDNA has been found to be both unique to *Sciara* and within Hc2 (Kerrebrock, unpublished). In order to enable future assessments of R3 as a candidate for the controlling element, the primary, secondary, and tertiary structures of *Sciara* rRNA were determined. The location of R3 was identified within each of these structures and any putative functions noted. Furthermore, a large 50-kilobase segment of non-rDNA interrupting the rDNA tandem array within Hc2 has also been recently identified (Urban & Gerbi, unpublished). RNA-seq reads were mapped to this region, and differential expression between male and female samples was assessed to identify candidates for the controlling element.

Here, I discuss various structural studies of the rDNA, including segmentation by Bayesian changepoint analysis, primary alignment by a position-weighted scoring scheme, and visualization of 2D/3D rDNA structures with localization of R3 insertion sites. Finally, I present results from differential expression analysis of the 50 kb non-rDNA interruption.

Segmentation and annotation of rDNA

In order to conduct downstream analysis on rDNA, the consensus sequence must first be segmented into known regions. The rRNA coding regions (18S, 5.8S, 2S and 28S) are separated by transcribed spacers, which are processed out from the precursor rRNA. In particular, each rDNA repeat is composed of the following segments, in 5' to 3' order: externally transcribed spacer (ETS), 18S, internal transcribed spacer 1 (ITS1), 5.8S, (5.8S gap region,) 2S, internal transcribed spacer 2 (ITS2), 28S, and a (generally) non-transcribed intergenic spacer (IGS). Strong experimental evidence by S1 nuclease mapping and primer extension has identified 5' and 3' ends of all rDNA segments in *Drosophila* (Jordan, Jourdan, & Jacq, 1976; Mandal & Dawid, 1981). While much of this information is transferrable to related species such as *Sciara* due to sequence conservation, direct experimental evidence for endpoints of certain *Sciara* rDNA segments, including the ETS, is weak or lacking entirely.

The rDNA segments corresponding to mature rRNA (2S, 5.8S, 18S, 28S) are expected to be enriched in RNA-seq libraries relative to those that are cleaved and processed out from transcribed spacers (ETS, ITS1, ITS2), which in turn are expected to be enriched relative to non-transcribed regions (IGS), absent in RNA-seq libraries. Thus, by mapping start positions of RNA-seq reads along a contiguous sequence of rDNA, there ought to be identifiable points ("changepoints") at which total read count changes in magnitude. This can be observed in the IGS, in which a small blip of change in mapped read count is observed, presumed to correspond to the ETS (Figure 1, left panel, red enclosure).

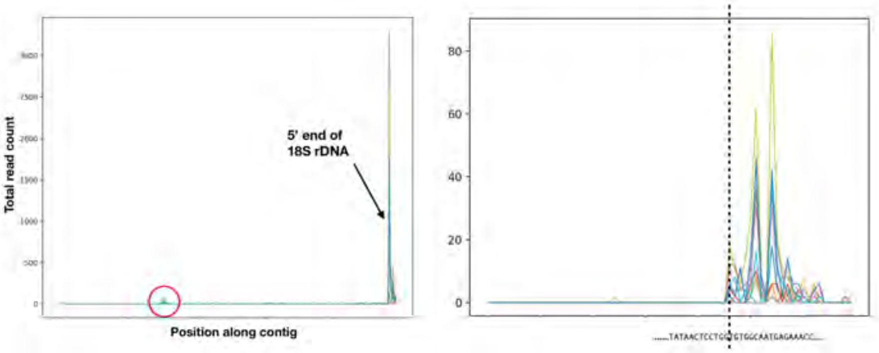


Figure 1: Mapped 5' start positions of RNA-seq reads onto IGS (left); magnified red enclosure with maximum likelihood changepoint indicated (right). Colors indicate distinct samples.

In order to infer the exact start position of the ETS, a 200 base pair window surrounding the putative ETS was selected for analysis. Given the small size of this window and knowledge about rDNA structural invariance, it was assumed *a priori* that only a single changepoint resided within this region. Standard Bayesian changepoint analysis was then conducted to obtain a posterior on the set of changepoints C (with a strong prior on $|C| = 1$). Total read count at any given starting position was modeled as a Poisson random variable, and log-likelihood of the observed read counts was maximized by an efficient dynamic programming-based implementation (Appendix).

The inferred location of the changepoint was identical across five of seven samples (Figure 1, right), with a small 1-2 nucleotide divergence for two low-read samples. Following primary sequence alignment with *Drosophila*, there was found to be little-to-no homology between the ETS of *Sciara* and *Drosophila*. Hence the results from this analysis provide the foundation for any downtream experimental analyses on the ETS segment.

Position-weighted sequence alignment

In addition to annotating transcribed spacers and mature rRNA coding regions, sequence alignment of rDNA from *Sciara* and *Drosophila* was performed in order to determine regions of homology. Sequence alignment of primary sequences will also aid in the construction of 2D *Sciara* ribosomal subunit structures from 2D *Drosophila* reference structures. A modified Needleman-Wunsch algorithm was used for global alignment of *Sciara* and *Drosophila* rDNA segments. First I describe the standard method for global alignment, followed by modifications to achieve more optimal results.

Needleman-Wunsch is a standard dynamic programming-based solution to optimal global alignment of nucleotide or protein sequences. Given two sequences A and B of lengths m and n, respectively, the algorithm computes and stores the maximum score $F_{i,j}$ of the alignment of prefixes $A_{1:i}$ and $B_{1:j}$, for all $1 \leq i \leq m$ and $1 \leq j \leq n$. Alignment is scored on the basis of a scoring matrix S, for which the entry S_{b_1,b_2} gives the penalty/score of aligned base pairs b_1 and b_2 (e.g, $S_{(A,G)} = -1$, a mismatch; $S_{(A,A)} = 1$, a match; $S_{(A,-)} = -2$, a deletion/insertion). The algorithm runs in $O(mn)$ time by taking advantage of subproblem structure. In particular, the maximum alignment score of length-i and -j prefixes is conditionally independent of all other prefix alignments given the scores of prefixes one unit shorter:

$$F_{i,j} = \max\left(F_{i-1,j-1} + S_{(A_i,B_j)}, F_{i-1,j} + S_{(A_i,-)}, F_{i,j-1} + +S_{(-,B_j)}\right)$$

One limitation of this algorithm is the invariance of the scoring matrix. That is, the scoring of matches, mismatches, and gaps are uniform throughout the sequence alignment. This may be problematic for alignment of sequences such as rDNA, in which small, highly-conserved segments are flanked by regions of intermediate or low conservation. An example has been constructed to demonstrate troublesome alignments that may result from position-uniform scoring:

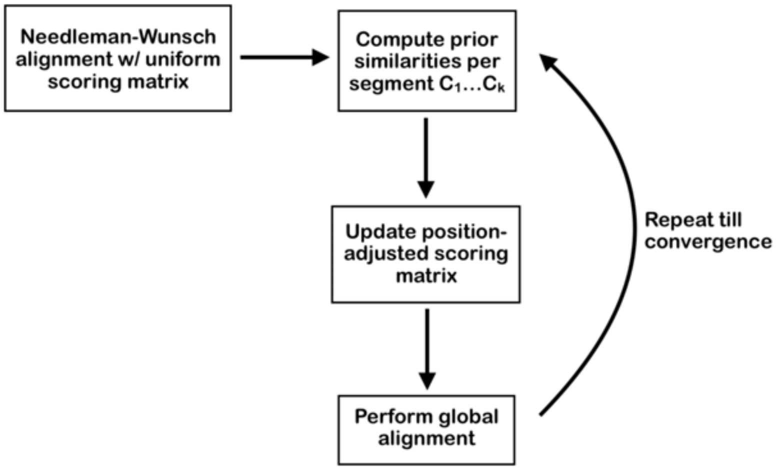
```
ATCGA---TAGT        ATCGATAGT---  ATCGATAG---T
|||||   ||||        |||||||||     ||||||||   |
ATCGATAGTAGT        ATCGATAGTAGT  ATCGATAGTAGT

Conserved reference segment
```

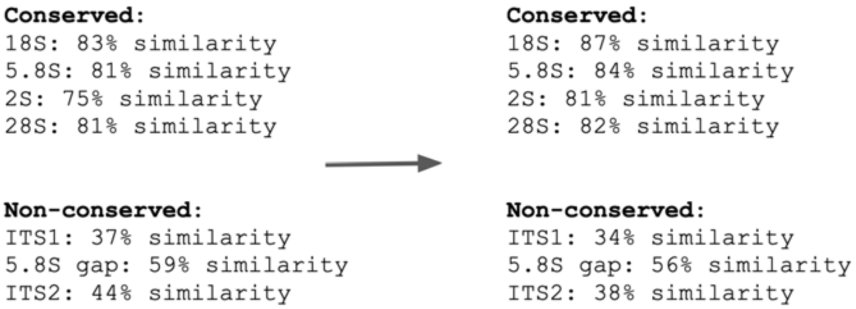
In these ambiguous alignments, those on the right retain end matches in the conserved region, whereas those on the left fail to do so. All have the same alignment score. Even alignments with lower scores may be preferable if matches are maximized in conserved end regions (especially when those end sequences have been experimentally confirmed):

```
ATCGA---TAGT        ATCGATAGT---
|||||   ||||        ||||||.||
ATCGATCGTAGT        ATCGATCGTAGT
```

In order to penalize errors more severely if they occur in regions of high similarity, similarity score $S_i(b_1, b_2)$ was made dependent on reference position i. Match scores and penalties were made larger in known conserved regions. A heuristic method was implemented, in which a standard Needleman-Wunsch alignment is first performed in order to obtain priors on the degree of conservation $C_1...C_k$ in each of k rDNA segments (e.g, 18S, 28S, 5.8S, 2S, ITS1, ITS2, and the gap region separating 5.8S from 2S). Next, the scoring matrix for each region is updated in each region r by multiplying matrix elements by a factor bC_r. This process is repeated until convergence of alignment:


Needleman-Wunsch alignment w/ uniform scoring matrix
Compute prior similarities per segment C₁...Cₖ
Update position-adjusted scoring matrix
Perform global alignment
Repeat till convergence

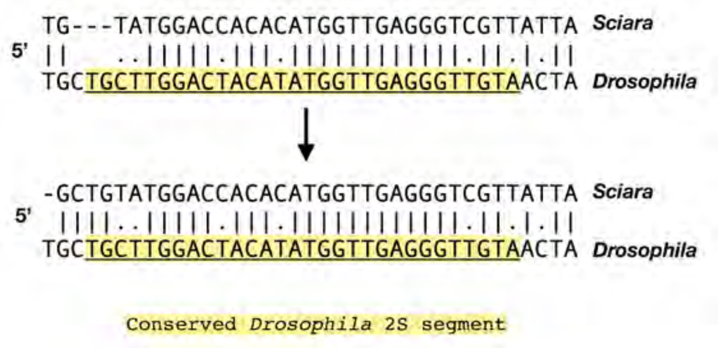
Sufficient convergence was achieved after 1-2 iterations. Employing this method, marginal increase in similarity of conserved regions was seen, accompanied by decrease in similarity of non-conserved regions:



```
Conserved:                           Conserved:
18S: 83% similarity                  18S: 87% similarity
5.8S: 81% similarity                 5.8S: 84% similarity
2S: 75% similarity                   2S: 81% similarity
28S: 81% similarity                  28S: 82% similarity

Non-conserved:                       Non-conserved:
ITS1: 37% similarity                 ITS1: 34% similarity
5.8S gap: 59% similarity             5.8S gap: 56% similarity
ITS2: 44% similarity                 ITS2: 38% similarity
```

As expected, accuracy in known conserved regions increased at the expense of known variable regions such as transcribed spacers. Percent similarity changes were on the order of 1-10%, unsurprising since this method is likely to select favorable alignments at conserved end regions where there are competing options, while retaining a majority of the alignment. As an instance of alignment improvement, one can see a reasonable change made to the 5' end of the 2S region:



```
    TG---TATGGACCACACATGGTTGAGGGTCGTTATTA Sciara
5'  ||   ..|||||.|||.||||||||||||.||.|.||
    TGCTGCTTGGACTACATATGGTTGAGGGTTGTAACTA Drosophila

                      ↓

    -GCTGTATGGACCACACATGGTTGAGGGTCGTTATTA Sciara
5'   ||||..|||||.|||.||||||||||||.||.|.||
    TGCTGCTTGGACTACATATGGTTGAGGGTTGTAACTA Drosophila

        Conserved Drosophila 2S segment
```

Visualization of 2D and 3D ribosomal subunits

In order to predict 2D ribosomal subunit structures in *Sciara*, known 2D structures for *Drosophila* (Ribovision; Bernier et al, 2014) were used to anchor the primary nucleotide sequence from our *Sciara*-*Drosophila* alignment. The reference position-adjusted scoring scheme produced more sensible 2D structure. As a representative example, the effect of the correction on the 2S 5' end is displayed in Figure 2 below.

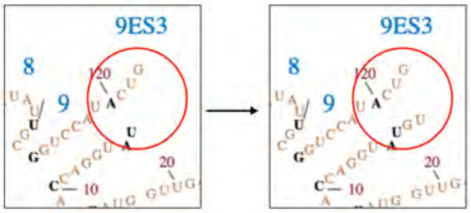


Figure 2: 2D structure of *Sciara* 2S coding region before (left) and after (right) position-adjusted alignment. ES = expansion segment.

Positions where base pair identity differs in *Sciara* as compared to *Drosophila* are marked in black. A number of compensatory base pair changes can be observed (Ext. Data Figure 1), suggesting that our 2D structure is logically consistent with evolution of ribosomal structure.

Next, we sought to determine the role of R3, an insertion in the 28S coding region of 1-2 rDNA repeats and a candidate for the controlling element of non-disjunction given that it maps to Hc2 (A. Kerrebrock, unpublished). The location of the R3 insertion site was visualized in 2D and 3D models of the large ribosomal subunit and compared to the positions of R1 and R2. (Note: Any given rDNA repeat unit would have none or just one of the three insertions (R1, R2, R3), but all three insertions are shown superimposed on the same 2D or 3D structure of 28S rRNA so that their relative positions to one another may be visualized.) In 2D space, both R3 and R2 sites were found to be located on 28S helix 69 (Ext. Data Figure 2). In 3D space, the R3 site was also observed to be near R2 (Ext. Data Figure 3), and all three insertion sites were located on the subunit interface, suggesting possible interaction with the small ribosomal subunit. Further inspection of the primary sequence revealed that the R3 and R2 sites were indeed found be located within highly-conserved eukaryotic nucleotide element 38 (eCNE 38; Doris et al, 2015), which has a putative role in the intersubunit bridges B2a, B2b, B3, and B7a. This is also in agreement with the 2D structure, as 28S helix 69 (on which R3/R2 reside) is known to participate in the formation of bridge B2a (Behrmann et al, 2016).

Comparing expression in the non-rDNA interruption

The putative controlling element of X chromosome non-disjunction may reside in a large 50-kilobase segment of non-rDNA that maps to Hc2 and is embedded within the tandem rDNA array. Since non-disjunction occurs during spermatogenesis, male pupae ought to show evidence of greater CE expression if the CE is a transcribed RNA. Therefore, mRNA reads from female pupae, male pupae, and testes isolated from male pupae were mapped to contig 230, an element of a Falcon assembly of the *Sciara* genome that contains a large portion of the 50 kb non-rDNA interruption (Ext. Data Figures 4). Total reads mapping to each 100 bp window in the non-rDNA were summed. Counts were then normalized by size factors computed by the "median ratio method," incorporating all genes represented across samples. P-values for differential expression of normalized data were then computed by *t*-test.

There was found to be no statistically significant differences in expression in non-rDNA regions. RNA-seq libraries from the male pupal testes contained markedly less rRNA contamination ($< 0.5\%$ as compared to 2-5% in other samples). This may be attributed to the different protocols used (i.e, polyA selection methods) to extract and purify RNA from these samples, rather than true biological differences.

Discussion

Primary, secondary, and tertiary structures of rDNA were successfully constructed through a position-weighted alignment scheme following by anchoring of the *Sciara* sequence to experimentally verified *Drosophila* structures. R3 was found to be located at conserved regions near the subunit interface, presumed to be involved in inter-subunit bridging. It remains unknown why the roughly 100 nt region in 28S rRNA is a hotspot for insertion of R1, R2 and/or R3. Given that R3 maps to Hc2, it constitutes a candidate for the CE, but the mechanism by which it may mediate non-disjunction is enigmatic.

Mapping of RNA-seq reads to the non-rDNA interrupt revealed no regions of differential expression between male pupal testes and whole-bodied male pupae, nor between male pupae and female pupae. The controlling element may therefore have not been present in our RNA-seq library (e.g, due to short transcript length as in small noncoding RNAs or due to lack of a polyA tail), or the CE may map to another region of Hc2 entirely (i.e, not within contig 230 of the Falcon genome assembly). Alternatively, the controlling element of non-disjunction may not be a transcribed element at all. Rather, it is possible that whichever chromosome contains the majority of rDNA undergoes non-disjunction (Abbott & Gerbi, 1981). Since 10% of rDNA repeats are found within Hc1, 50% within Hc2, and 40% within Hc3 (Crouse et al, 1977), the latter hypothesis is consistent with translocation results indicating that Hc1 + Hc2 (60% total rDNA) and Hc2 + Hc3 (90% total rDNA) have the CE activity (Crouse, 1979).

References

[1] A. Simeone E. Boncinelli. 5-Cleavage site of D. melanogaster 18 S rRNA. FEBS Lett. 1984 Feb 27;167(2):249-53.

[2] Behrmann E, Loerke J, Budkevich TV, Yamamoto K, Schmidt A, Penczek PA, Vos MR, Bürger J, Mielke T, Scheerer P, Spahn CM. Structural snapshots of actively translating human ribosomes. Cell. 2015 May 7;161(4):845-57. doi: 10.1016/j.cell.2015.03.052. PMID: 25957688; PMCID: PMC4432480.

[3] Bernier, C. R., Petrov, A. S., Waterbury, C. C., Jett, J., Li, F., Freil, L. E., Xiong, X., Wang, L., Migliozzi, B. L. R., Hershkovits, E., Xue, Y., Hsiao, C., Bowman, J. C., Harvey, S. C., Grover, M. A., Wartell, Z. J., and Williams, L. D. (2014). FD169: RiboVision Suite for Visualization and Analysis of Ribosomes. Faraday Discussions

[4] Crouse, Helen. (1960). The Controlling Element in Sex Chromosome Behavior in Sciara. Genetics. 45. 1429-43.

[5] Crouse, Helen. (1977). X heterochromatin subdivision and cytogenetic analysis in Sciara coprophila (Diptera, Sciaridae) - I. Centromere localization. Chromosoma. 63. 39-55.

[6] Crouse, Helen. (1979). X heterochromatin subdivision and cytogenetic analysis in Sciara coprophila (diptera, sciaridae) - II. The controlling element. Chromosoma. 74. 219-239.

[7] Crouse H, Gerbi S, Liang C, Magnus L, Mercer I. (1977). Localization of ribosomal DNA within the proximal X heterochromatin of Sciara coprophila (Diptera, Sciaridae). Chromosoma. 64. 305-18. 10.1007/BF00294938.

[8] de Lanversin, G., Jacq, B. Sequence and secondary structure of the central domain of Drosophila 26S rRNA: A universal model for the central domain of the large rRNA containing the region in which the central break may happen. J Mol Evol 28, 403–417 (1989). https://doi.org/10.1007/BF02603076

[9] Doris SM, Smith DR, Beamesderfer JN, Raphael BJ, Nathanson JA, Gerbi SA. Universal and domain-specific sequences in 23S-28S ribosomal RNA identified by computational phylogenetics. RNA. 2015 Oct;21(10):1719-30. doi: 10.1261/rna.051144.115. Epub 2015 Aug 17. PMID: 26283689; PMCID: PMC4574749.

[10] Gerbi S.A. (1986) Unusual Chromosome Movements in Sciarid Flies. In: Hennig W. (eds) Germ Line — Soma Differentiation. Results and Problems in Cell Differentiation (A Series of Topical Volumes in Developmental Biology), vol 13. Springer, Berlin, Heidelberg

[11] Jordan BR, Jourdan R, Jacq B. Late steps in the maturation of Drosophila 26 S ribosomal RNA: generation of 5.8 S and 2 S RNAs by cleavages occurring in the cytoplasm. J Mol Biol. 1976 Feb 15;101(1):85-105.

[12] Jordan BR, Latil-Damotte M, Jourdan R. Sequence of the 3'-terminal portion of Drosophila melanogaster 18 S rRNA and of the adjoining spacer: comparison with corresponding prokaryotic and eukaryotic sequences. FEBS Lett. 1980 Aug 11;117(1):227-31.

[13] Long, E.O., Collins, M., Kiefer, B.I. et al. Expression of the ribosomal DNA insertions in bobbed mutants of Drosophila melanogaster . Molec. Gen. Genet. 182, 377–384 (1981). https://doi.org/10.1007/BF00293925

[14] Mandal RK, Dawid IB. The nucleotide sequence at the transcription termination site of ribosomal RNA in Drosophila melanogaster. Nucleic Acids Res. 1981 Apr 24;9(8):1801-11. doi: 10.1093/nar/9.8.1801. PMID: 6264393; PMCID: PMC326804.

[15] Pavlakis GN, Jordan BR, Wurst RM, Vournakis JN. Sequence and secondary structure of Drosophila melanogaster 5.8S and 2S rRNAs and of the processing site between them. Nucleic Acids Res. 1979 Dec 20;7(8):2213-38. doi: 10.1093/nar/7.8.2213. PMID: 118436; PMCID: PMC342381.

[16] Ware VC, Renkawitz R, Gerbi SA, rRNA proceesing: removal of only nineteen bas at the gap between 28S and 28S rRNAs in Sciara coprophila, Nucleic Acids Research, Volume 13, Issue 10, 24 May 1985, Pages 3581–3597, https://doi.org/10.1093/nar/13.10.3581

Part 2: An Exploration of the Radiation Response via RNA-seq Analysis

Introduction

Sciara coprophila exhibits greater resistance to X-irradiation than related organisms. The larvae of *Drosophila melanogaster*, for instance, can withstand a 20 Gy dose of X-irradiation but will succumb to a 40 Gy dose (Jaklevic and Su, 2004; Ashburner et al, 2005). By contrast, *Sciara* larvae are able to withstand an 80 Gy dose of X-irradiation and retain full ability to pupate (though there is a decreased eclosion rate [J. Borden et al, unpublished]). Whereas *Drosophila* has long been popular as a model organism for forward genetic studies (Muller, 1928; Ashburner et al, 2005), *Sciara* is notably more resistant to mutation. While both gross and minute chromosomal abnormalities may be induced in *Sciara* upon irradiation, visible mutations are rare (Crouse, 1949). Some have attributed this phenomenon to the physical appearance of *Sciara*, whose black body and eyes make difficult the observation of phenotypic changes. Though this may be true, *Sciara* may also be likely to harbor an enhanced biochemical responses to radiation.

It is worth noting that the stated radioresistance of *Sciara* is based on comparison with species of its own class. An increased resistance to X-irradiation is generally seen with organismal simplicity. The bacterium *Deinococcus radiodurans*, for example, can survive a 5000 Gy dose of X-irradiation (Slade and Radman 2011). And while an X-irradiation dose of 1200 Gy is lethal to the budding yeast *Saccharomyces cerevisiae* (Mitchel and Morrison, 1984), the Bdelloid rotifers *Adineta vaga* and *Philodina roseola* can survive this dose (Gladyshev and Meselson, 2008). Conversely, complex multicellular organisms have been shown to be highly sensitive to X-irradiation, with 1.5 to 11 Gy being the lethal dose for humans. Furthermore, among complex organisms, insects exhibit relatively high X-irradiation resistance, with the lethal dose depending on developmental stage: eggs are more sensitive than larvae, with pupae and adults being the most resistant. Radioresistance, therefore, varies widely among organisms.

To further explore the relative radioresistance of *Sciara*, RNA-seq analysis was performed on irradiated larvae to study gene expression changes. The study of radioresistance may lead to better understanding of both general DNA damage responses and the development of radioresistance following radiation therapy for cancer malignancies.

Methods[1]

Sample treatment and library preparation

Roughly 590 larvae (*Sciara coprophila* HoLo2 strain, female, pre-eyespot) from 8 vials were distributed across three 2.2% (w/v) agar Petri plates in a sequential process (i.e, a single larva was placed on plate A, followed by plate B, plate C, and so forth). Larvae from each of these three plates were in turn distributed across four additional plates to provide replicates, making for 12 plates total. Two of the four plates in each replicate were randomly selected for irradiation, with the other two plates serving as matched controls. Approximately 50 larvae were present on each plate. Samples were irradiated for 50 minutes at 80 Gy using 137Cs γ rays emitted by a JL Shepherd irradiator. Doses in Gray (Gy) correspond to the absorption of 1 J/kg, with 1 Gy equivalent to 100 rads. A continuous dose rate of 1.7 Gy/min was administered to the larvae.

Subsequent steps were performed (or samples frozen) roughly 45 minutes after radiation. Total RNA isolation was performed by homogenization after addition of TRIzol reagent (1 ml per 50-100 mg tissue), followed by pelleting of debris through centrifugation at 12000g for 10 minutes. Subsequently, 0.2 ml chloroform was added per 1 ml of the Trizol-containing supernatant and shaken for 15 seconds followed by a 2-3 minute incubation. After another centrifugation of 12000g for 10 minutes, the RNA occupying the aqueous layer in the resulting solution was precipitated with isopropanol. RNeasy columns were used for isolation of total RNA. Subsequent library preparation involved AMPure bead cleaning, polyA selection, and NEB adaptor ligation for paired-end, strand-specific RNA sequencing.

RNA-seq analysis

RNA-seq reads were available as fastq files. Analysis pipeline involved preprocessing and quality control of fastq files (i.e, trimming of adaptors and removal of low-quality reads), mapping of reads to the reference genome or transcriptome, counting reads mapped to features of interest (i.e, genes or gene isoforms), and analysis of differential expression among samples. All computa-

[1] Prior to the present work, preparation of samples was completed by John Urban, Jacob Bliss, and Julia Borden of the Gerbi Lab with irradiation performed with assistance from Richard Shea of Brown's Environmental Safety Office. Sequencing of the *Sciara* genome and its assembly including Falcon contig 230 was completed by John Urban. RNA-seq data for pupal testes was courtesy of Christina Hodson.

tional work was done on N1 cloud machines hosted on Google Cloud Platform.

Quality control was performed using FastQC software (Babraham Bioinformatics) followed by adaptor trimming by Trimmomatic. Per base sequence quality was observed to be sufficient (Ext. Data Figure 1A), with mild degradation of signal at 3' and 5' end of reads. Loss of signal is expected due to accumulation of errors during sequencing-by-synthesis methods. Per base sequence content was roughly equivalent across a majority of read positions (Ext. Data Figure 1B), with large biases at read ends likely due to non-random hexamer priming. As noted in the literature, hexamer primer bias may persist for up to 13 bases as opposed to the expected six (Hansen et al, 2010). Per base GC content was normally distributed with a mean at 45%. Sequence duplication levels were low. Based on these evaluations of quality control metrics, samples were high-quality and additional removal of reads beyond adaptors proved unnecessary.

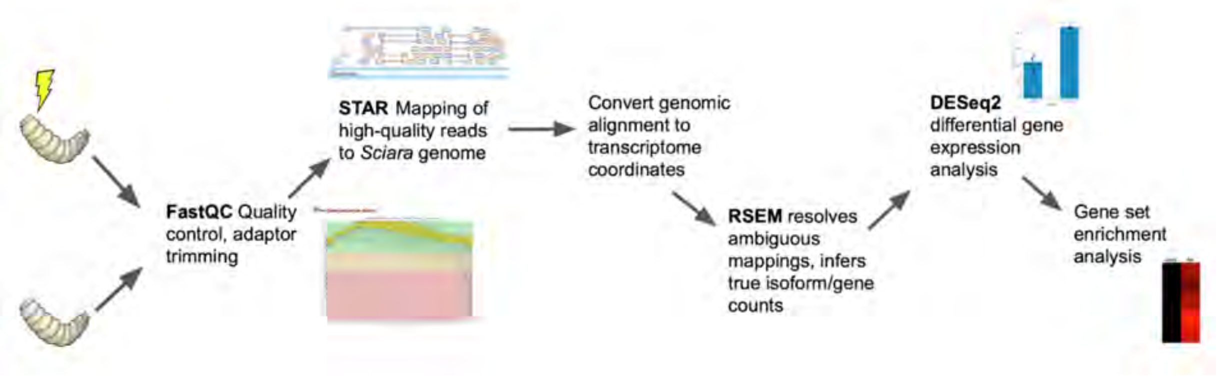

FastQC Quality control, adaptor trimming
STAR Mapping of high-quality reads to *Sciara* genome
Convert genomic alignment to transcriptome coordinates
RSEM resolves ambiguous mappings, infers true isoform/gene counts
DESeq2 differential gene expression analysis
Gene set enrichment analysis

Figure 1: RNA-seq analysis pipeline.

Next, reads were aligned to a Canu-assembled reference genome of *Sciara* by STAR software. STAR constructs a suffix array of the reference genome, thus allowing for rapid read alignment. Reads are first seeded onto the genome, followed by stitching of the alignment and determination of optimal gaps. STAR outputs .sam/.bam files with read alignments in both genomic and transcriptomic coordinates.

Following read alignment with STAR, ambiguously-mapped reads remained unresolved. 4-8% of reads mapped to multiple regions of the genome. Retaining all multi-mapping reads may bias gene/isoform counts. If reads mapping to two genes or isoforms A and B, for instance, are included in both counts, biologically true differences in expression may be masked. Discarding ambiguous reads may also lead to both bias and loss of information. Thus, ambiguous mappings were resolved using RSEM, which takes as input the genome sequence, gene annotation, and .bam alignment from STAR, and outputs expected read count distribution and a probabilistically-weighted alignment.

Expected read counts from RSEM cannot be immediately compared between sample groups. Differences in sequencing depth, for instance, will confound true biological differences in gene expression. Furthermore, simply normalizing by equating total RNA read counts may lead to misleading results, as differences in total gene expression may be biologically genuine. Thus, DESeq2 was used to infer true gene expression given observed read counts.

The process for DESeq2 will be explained briefly. First, differences in library depth are accounted for by the inclusion of a "size factor" term. Size factors for each sample are estimated by computing the "median of ratios" (that is, the median of the ratios of each gene count in a sample relative to the geometric mean of gene counts across all samples). In addition to sequencing depth, DESeq2 also models variability in gene expression within each group, via a negative-binomial dispersion parameter. This is performed on a gene-wise basis, operating under the assumption that biological variance in gene expression (across cells) ought to be similar. Dispersion is therefore first estimated for each gene individually. Subsequently, these gene-wise dispersions are fitted to a smooth function of dispersions relative to mean gene expression. This serves as a form of regularization, an essential step given that sample sizes are frequently small in RNA-seq workflows (usually less than 10 samples per condition). Finally, DESeq2 fits a negative binomial model to each gene given the estimated size factor and dispersion terms, and employs the Wald Test to determine significance of differential expression. Log-fold changes (LFC) estimates are "shrunken" using the mean LFC across all genes as a prior. This reduces the chance of observing inaccurate LFC estimates when quality is low.

After obtaining log-fold changes in gene expression by DESeq2, over-representation analysis was performed on the set of differentially expressed genes. In this analysis, the frequency of a gene set characterization in the differentially-expressed group is compared to its overall frequency among all known genes.

If this frequency is significantly higher than would be expected by random chance, the gene set is considered “over-represented.” This frequency is commonly given by the hypergeometric distribution. Gene set enrichment analysis is similar in objective to over-representation analysis, with the most important difference being the identity of the input, which is generally a ranked list of all genes rather than a set of differentially-expressed genes. Over-representation analysis of Gene Ontology (GO) terms was performed by using the Panther web interface. A diagram of the complete pipeline is given below (Figure 1).

Results

Evaluating sensibility of the RNA-seq pipeline

Normalized counts and log-fold changes were visualized for preliminary exploration of the data. Plotting variance of expression within each sample group against mean gene expression value revealed that the former is generally greater than the latter (Figure 2).

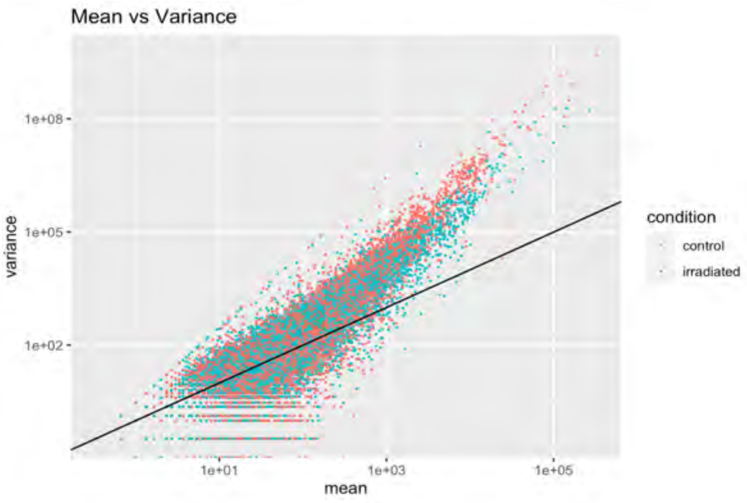


Figure 2: Mean vs variance in control and irradiated samples. Line is of slope 1 (mean = variance).

This justified our use of DESeq2, which employs a negative binomial model that allows variance to be greater than the mean. Furthermore, Figure 2 reveals a clear dependency of the variability in the variance on mean gene expression (heteroscedasticity). This justifies both the necessity of modeling dispersion, as well as the mean-based regularization of dispersion that occurs within DESeq2.

Log-fold changes were visualized both before and after shrinking (Ext. Data Figure 2). Differentially-expressed genes, marked in red, spanned much of the range of possible mean normalized counts, and thus DESeq2 appeared to be effective in removing association between gene mean and likelihood of differential expression.

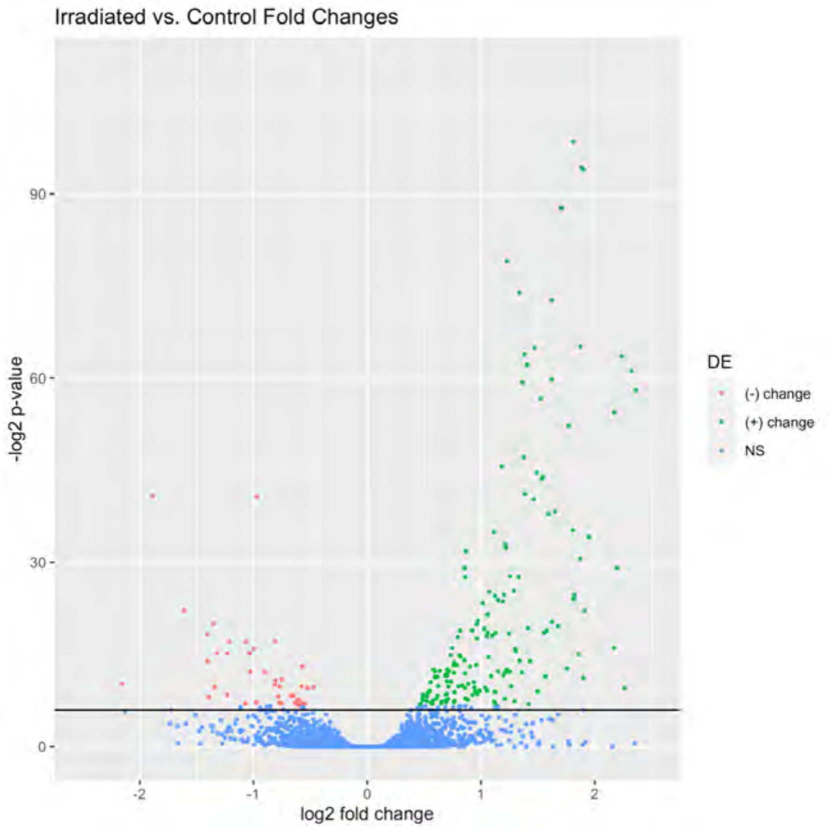


Figure 3: Volcano plot of log-fold changes against negative-log p-values for all genes Line marks p = 0.01

The radiation response is a transcriptionally active state

The biochemical response to radiation may involve both up-regulation and down-regulation of gene expression. In our samples, over 80% of differentially-expressed genes had greater expression in irradiated samples relative to controls (Figure 3). This suggests that the radiation response is a transcriptionally active state, with the predominant short-term response being up-regulation rather than down-regulation of genes. Additionally, as can be seen on the volcano plot, log-fold changes of up-regulated genes tended to be larger in magnitude than those of down-regulated changes and their p-values significantly lower. It is important to note, however, that these changes only reflect relatively immediate changes in RNA expression, as total RNA was extracted (or samples frozen) roughly 45 minutes after the culmination of radiation treatment.

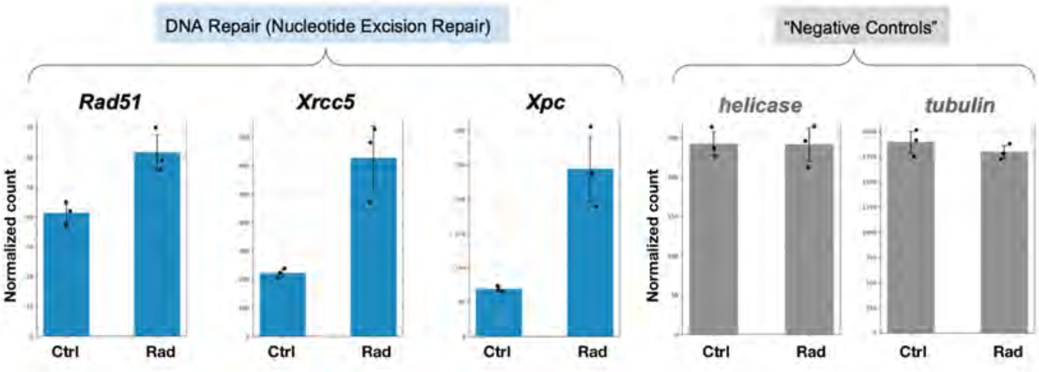
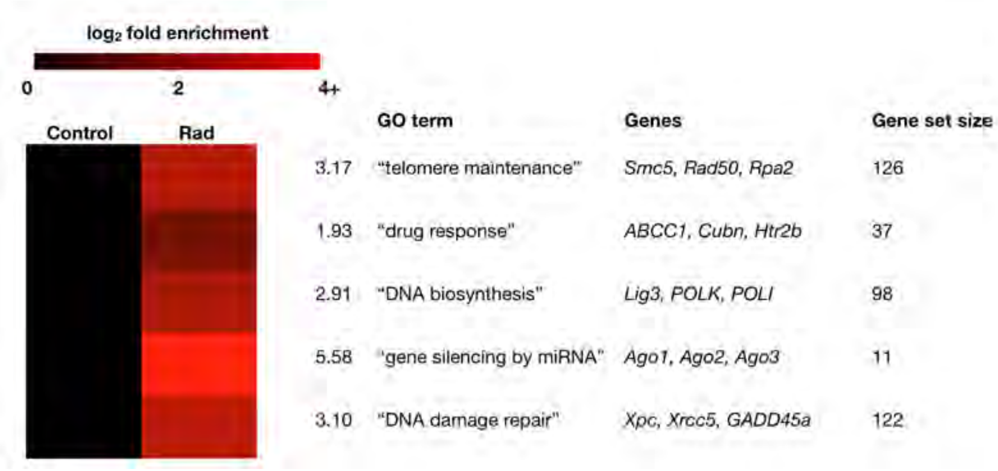


Figure 4: DESeq2-normalized counts of nucleotide excision repair genes.

DNA repair genes are up-regulated in response to radiation

DNA repair genes involved in nucleotide excision repair, including homologs of *Rad51*, *Xrcc5*, and *Xpc*, were observed to be up-regulated in irradiated samples, with 2-3 fold differences in mean normalized expression, computed using normalized counts multiplied by fold changes derived from DESeq2 (Figure 4). Upon observing these differences, expression in other DNA repair pathways was investigated. Interestingly, differences in expression of mismatch repair genes (*Mlh1*, *Msh4*) and base excision glycosylases (*Ung*) were generally not significant (with the exception of a small subset of base-excision repair genes that were down-regulated; Ext. Data Figure 3). This finding may be partially attributed to the minimal role that mismatch repair proteins tend to play in correction of gross DNA abnormalities caused by radiation.



	GO term	Genes	Gene set size
3.17	"telomere maintenance"	*Smc5, Rad50, Rpa2*	126
1.93	"drug response"	*ABCC1, Cubn, Htr2b*	37
2.91	"DNA biosynthesis"	*Lig3, POLK, POLI*	98
5.58	"gene silencing by miRNA"	*Ago1, Ago2, Ago3*	11
3.10	"DNA damage repair"	*Xpc, Xrcc5, GADD45a*	122

Figure 5: Enrichment analysis of Gene Ontology terms in differentially-expressed gene set.

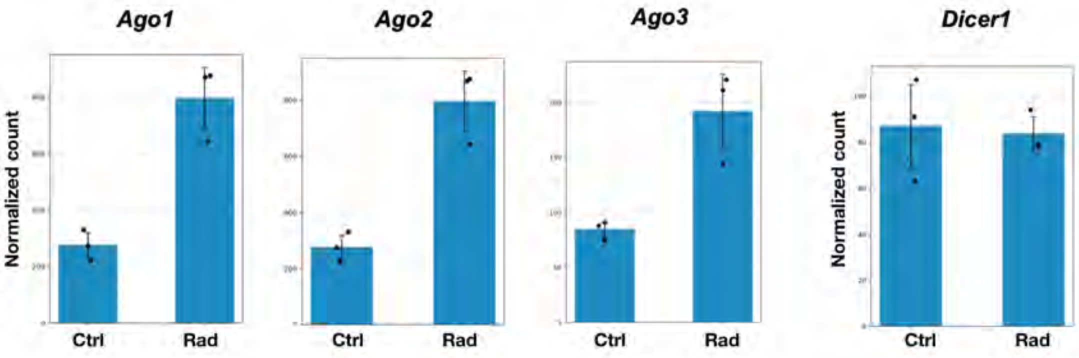


Figure 6: DESeq2-normalized counts of *Ago* and *Dicer* genes.

Enrichment analysis reveals upregulation of RISC-associated genes
Over-representation analysis of GO sets ("Gene Ontology" sets, grouped by

biological process, molecular function, or cellular component) revealed enrichment of gene sets involved in ‘telomere maintenance,’ ‘DNA damage repair,’ and ‘DNA biosynthesis.’ Interestingly, genes involved in gene silencing by miRNA were also observed to be enriched in the differentially-expressed set (Figure 5). A nearly three-fold increase in *Argonaute* mRNA expression was noted, though other RISC-associated genes such as *Dicer1* appeared to be unaffected (Figure 6).

Discussion

RNA-seq analysis revealed up-regulation of DNA repair genes as expected, including *Rad51* for homologous recombination, *Xrcc5* for non-homologous end-joining, and a variety of genes mediating nucleotide excision repair including *Xpc*. The magnitude and rapid onset (within 1 hour of irradiation) of this response may contribute to radioresistance in *Sciara*.

Furthermore, our observations of the up-regulation of *Ago* genes after radiation are consistent with findings in the literature. For instance, an increase in cell death has been observed post-radiation if *Ago*/*Dicer* are suppressed (Kraemer et al, 2011). Furthermore, there are well-documented miRNA responses to radiation (Metheetrairut & Slack, 2014), providing a biological rationale for the up-regulation of *Ago* genes observed here.

Both Dicer and AGO2 have been found to affect double-strand break (DSB) repair efficiency, suggesting an active role of these factors in this process of recruiting DSB repair proteins (Wei et al. 2012). In addition, human AGO2 has been shown to interact with RAD51 (Gao et al., 2014) and recruit this recombinase to DSB sites, thereby mediating double-stranded DNA ligation during repair by homologous recombination (HR). Because we did not find *Dicer* to be up-regulated after irradiation, it might be the case that any small RNA that interacts with Ago for radiation repair is Dicer-independent, or perhaps even RNA-independent. However, since Dicer has been reported to mediate the DNA damage response (Francia et al. 2016; Burger et al 2017), it may be more likely that sufficient Dicer is already present in cells and that it does not have to be up-regulated following X-irradiation. An alternative explanation is that *Dicer* is up-regulated on a time frame longer than one hour.

References

[1] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics, Volume 29, Issue 1, January 2013, Pages 15–21.

[2] Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[3] Ashburner M, Golic KG, and Hawley RS. Drosophila: A Laboratory Handbook, Second Edition. New York: Cold Spring Harbor Laboratory Press, 2005.

[4] Burger K, Schlackow M, Potts M, Hester S, Mohammed S, Gullerova M (2017) Nuclear phosphorylated Dicer processes double-stranded RNA in response to DNA damage. J. Cell Biol. 216 (8): 2373–2389. a

[5] Francia S, Cabrini M, Matti V, Oldani A, d'Adda di Fagagna F (2016) DICER, DROSHA and DNA damage response RNAs are necessary for the secondary recruitment of DNA damage response factors. J. Cell Sci. 129: 1468-1476.

[6] Gao M, Wei W, Li MM, Wu YS, Ba Z, Jin KX, Li MM, Liao YQ, Adhikari S, Chong Z, et al. (2014) Ago2 facilitates Rad51 recruitment and DNA double-strand break repair by homologous recombination. Cell Res 24: 532–541

[7] Gladyshev E and Meselson M (2008). Extreme resistance of Bdelloid rotifers to ionizing radiation. Proc. Natl. Acad. Sci. 105: 5139-5144.

[8] Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010 Jul;38(12):e131. doi: 10.1093/nar/gkq224. Epub 2010 Apr 14. PMID: 20395217; PMCID: PMC2896536.

[9] Jaklevic BR and Su TT (2004). Relative contribution of DNA repair, cell cycle checkpoints, and cell death to survival after DNA damage in Drosophila larvae. Current Biology. 14: 23-32.

[10] Kraemer A, Anastasov N, Angermeier M, Winkler K, Atkinson MJ, Moertl S. MicroRNA-mediated processes are essential for the cellular radiation response. Radiat Res. 2011; 176(5): 575–586.

[11] Krisko A and Radman M (2010). Protein damage and death by radiation in Escherichia coli and Deinococcus radiodurans. Proc. Natl. Acad. Sci. 107: 14373 -14377.

[12] Li, B., Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011).

[13] Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, 550.

[14] Metheetrairut C, Slack FJ. MicroRNAs in the ionizing radiation response and in radiotherapy. Curr Opin Genet Dev. 2013 Feb;23(1):12-9. doi: 10.1016/j.gde.2013.01.002. Epub

2013 Feb 28. PMID: 23453900; PMCID: PMC3617065.

[15] Mitchel REJ and Morrison DP (1984). An oxygen effect for gamma-radiation induction of radiation resistance in yeast. Radiation Research 100: 205-210.

[16] Muller HJ (1928). The production of mutations by X-Rays. Proc. Natl. Acad. Sci. 14: 714–726.

[17] Paul D. Thomas, Michael J. Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, Apurva Narechania. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res., 13: 2129-2141.

[18] Slade D and Radman M (2011). Oxidative stress resistance in Deinococcus radiodurans. Microbiol. Mol. Biol. Rev. 75: 133-191.

[19] Wei, W. et al. A role for small RNAs in DNA double-strand break repair. Cell 149, 101–112 (2012).

Appendix

Bayesian changepoint analysis

Here, I discuss standard mathematics behind Bayesian changepoint inference, details of my computational implementation, and various modeling choices.

Bayesian statistics enables the inference of parameters responsible for generation of observable data. In the case of changepoint inference, we wish to infer the position where a parameter changes value in sequential data. In the context of this work, changepoint inference enables the detection of the starting point of transcription, at which point the magnitude of mapped read counts is likely to increase.

To formalize this problem, let our sequence of observed read counts be represented by S. We model count data as Poisson with mean λ. We wish to find the set of changepoints C that maximizes the likelihood of observed read counts. The posterior distribution on C is given by:

$$\begin{aligned} P(C|S_{1:n}) &= \frac{P(S_{1:n}|C)P(C)}{\sum\limits_{C} P(S_{1:n}|C)P(C)} \\ &= \frac{P(S_{1:n}|C,\lambda)P(C)P(\lambda)}{\sum\limits_{C}\int_{\lambda} P(S_{1:n}|C,\lambda)P(C)P(\lambda)d\lambda} \end{aligned}$$

where the summation is over all sets of changepoints, and the integration is comprised of k nested integrals over $\lambda = (\lambda_1, \lambda_2 \cdots \lambda_k)$, the set of mean count parameters given that $|C| = k - 1$.

In my implementation, the space of λ parameters is discretized and clipped to a range of reasonable values:

$$\lambda \in [0, P_{0.95}(S_{1:n})]$$

Dynamic programming is used, taking advantage of optimal subproblem structure. First cumulative log-probabilities of subsequences are stored in a matrix C, where $C(m,i)$ is the log-probability of observing $S_{1:i-1}$ given that $\lambda = \lambda_m$, the m^{th} value in our discretized space:

$$\begin{aligned} C(m,i) &= \log P(S_{1:i-1}|\lambda_m) \\ &= \log\left(\prod_{j=1}^{i-1} \frac{e^{-\lambda_m}\lambda_m^{s_j}}{s_j!}\right) \\ &= \sum_{j=1}^{i-1}\left[\log\left(e^{-\lambda_m}\right) - \log\left(s_j!\right) + \log\left(\lambda_m^{s_j}\right)\right] \\ &= -(i-1)\lambda_m - \sum_{j=1}^{i-1}\log\left(s_j!\right) + \log(\lambda_m)\sum_{j=1}^{i-1} s_j \end{aligned}$$

Next, a lookup table L is constructed with values for maximum log-probability of observing any given subsequence in S, computed efficiently by making use of stored log-probabilities in C:

$$L(i,j) = \max_m \left[\log P(S_{i:j}|\lambda_m)\right] = \max_m \left[C(\lambda_m, j-1) - C(\lambda_m, i-1)\right]$$

Finally, a dynamic programming table D is computed in which entry $D(k,i)$ is the maximum likelihood of observing sequence $S_{0:i}$ given that the k^{th} changepoint is located at position i. The likelihood of the sequence up till the k^{th} changepoint is conditionally independent of all other changepoints given the $(k-1)^{st}$ changepoint. Thus, our dynamic programming table can be computed by:

$$D(k,j) = \max_{i<j} D(k-1,i) + L(i,j)$$

As this table of likelihoods is constructed, so too is a table of the λ values and previous changepoint positions associated with each maximum likelihood. The k^{th} entry in the last column of our dynamic programming (assuming 1-indexing) is the maximum likelihood of observing $S_{1:n}$ given k changepoints. These likelihoods are multiplied by an exponential prior on the number of changepoints (in log-space, a term $-\alpha k$ is added to the k^{th} entry in the last column), thus giving us a value proportional to the posterior. The sequence of changepoints and corresponding lambda values is then obtained by back-propagating through D and associated tables.

To infer locations of conserved sequence ends in rDNA, the observed data is taken to be the vicinity (e.g, 200 bp) around the approximate expected location. This amounts to setting a strong prior on $k = 1$, simplifying the above analysis to a single changepoint.

Conserved reference sequences

These are the original papers identifying 18S, 5.8S, 2S, and 28S rRNA 5' and 3' boundary sequences in *Drosophila*. Most of them used S1 nuclease mapping to locate the 5' and 3' cleavage sites (in some cases 5' ends were also confirmed with primer extension analysis).

ETS rDNA
Long, Rebbert, & Dawid, 1981
Sequence: 5' AGGTAGGCAGTGGTTGCCGACC . . .

18S rDNA
Jordan, Latil-Damotte, & Jourdan, 1980 (3' cleavage site)
Simeone & Boncinelli, 1984 (5' cleavage site)
5' ATTCTGGTTGATCCTGCCAG . . . GGAAGGATCATTA 3'

5.8S rDNA
Pavlakis et al, 1979 (determined via homology rather than mapping)
5'AACTCTAAGCG...ACGCATATCGCAGTCCATGCTG 3'

2S rDNA
Jordan, Jourdan, & Jacq, 1976; Jordan, Latil-Damotte, & Jourdan, 1980;
Pavlakis et al, 1979
5' TGCTTGGACTACATATGGTTGAGGGTTGTA 3'

28S rDNA
Mandal & Dawid, 1981
5' TTATATACAACCT . . . TTTGCTTGATGATTCGA 3'

28S rDNA gap
Ware, Renkawitz, & Gerbi, 1985 (Sciara); deLanversin & Jacq, 1989
5' AAAATGCCT . . . CACTTGAA 3'

Extended Data (Part 1)

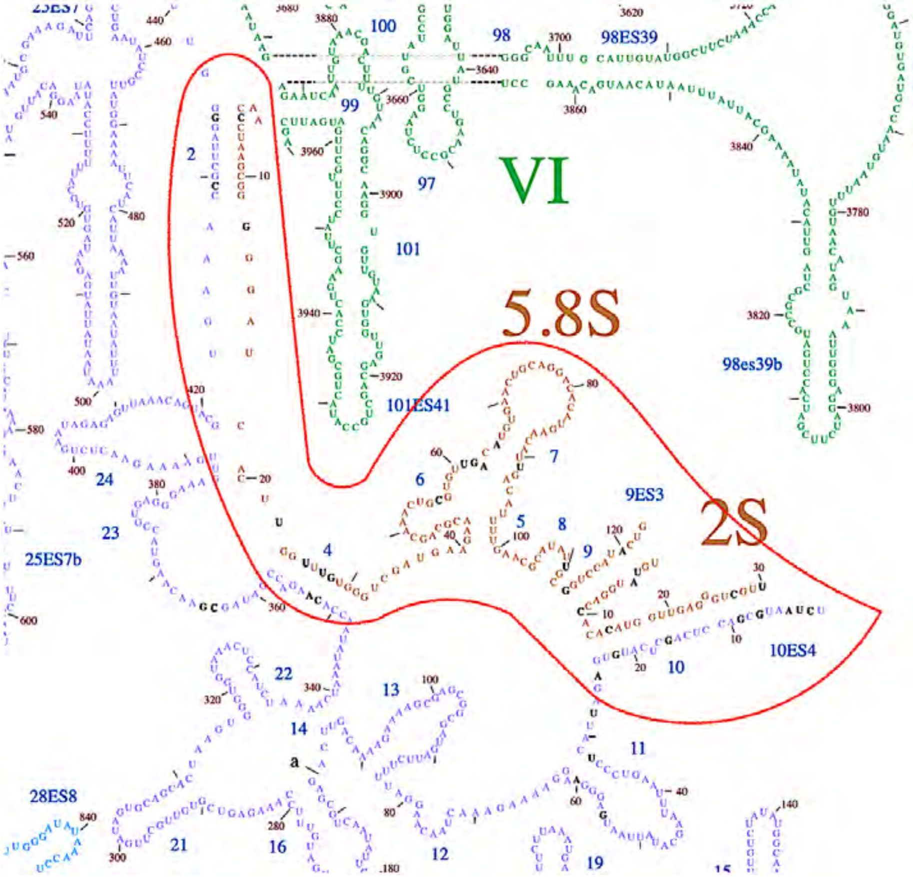


Figure 1: Partial 2D structure of *Sciara* large ribosomal subunit; base pair changes relative to *Drosophila* highlighted in black for enclosed region.

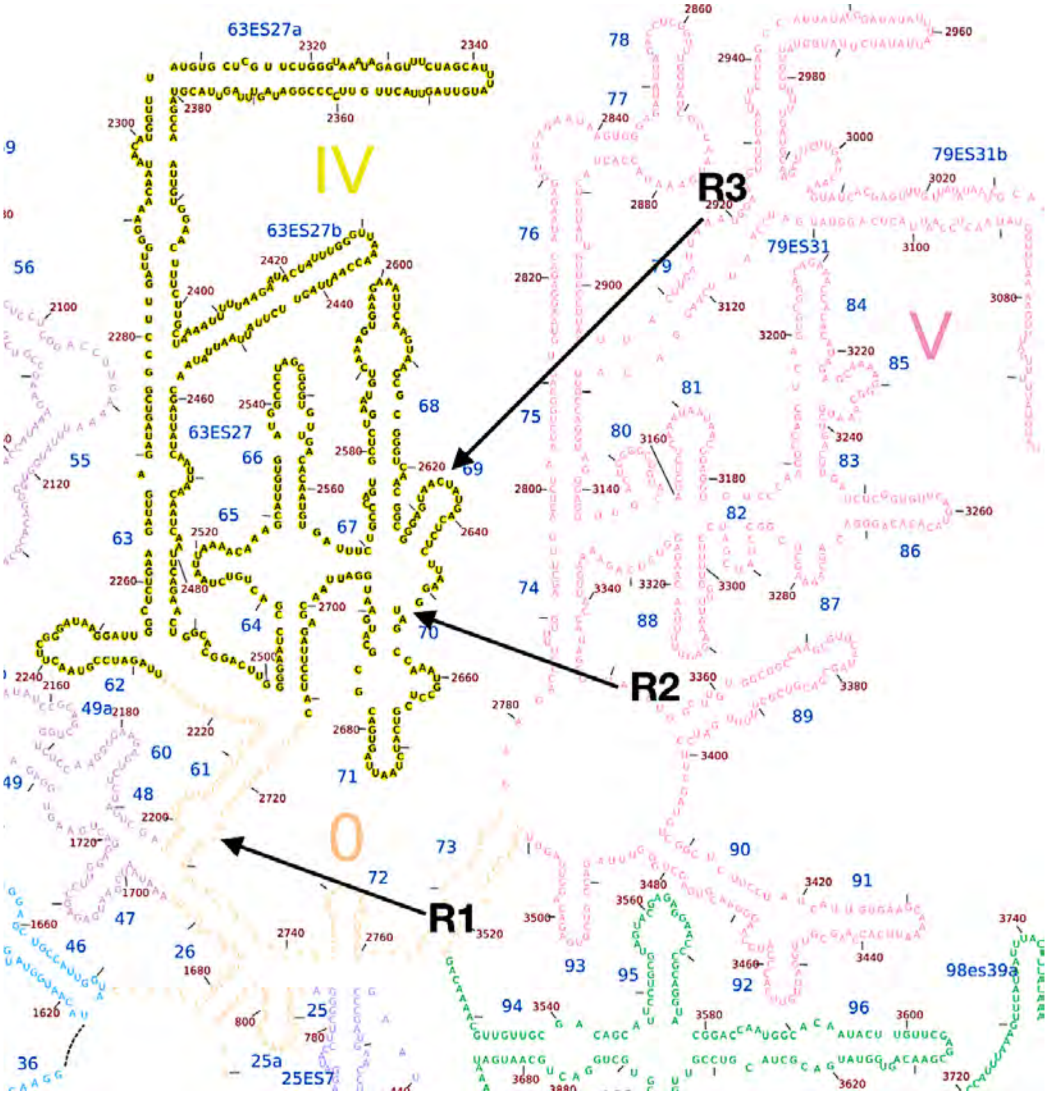
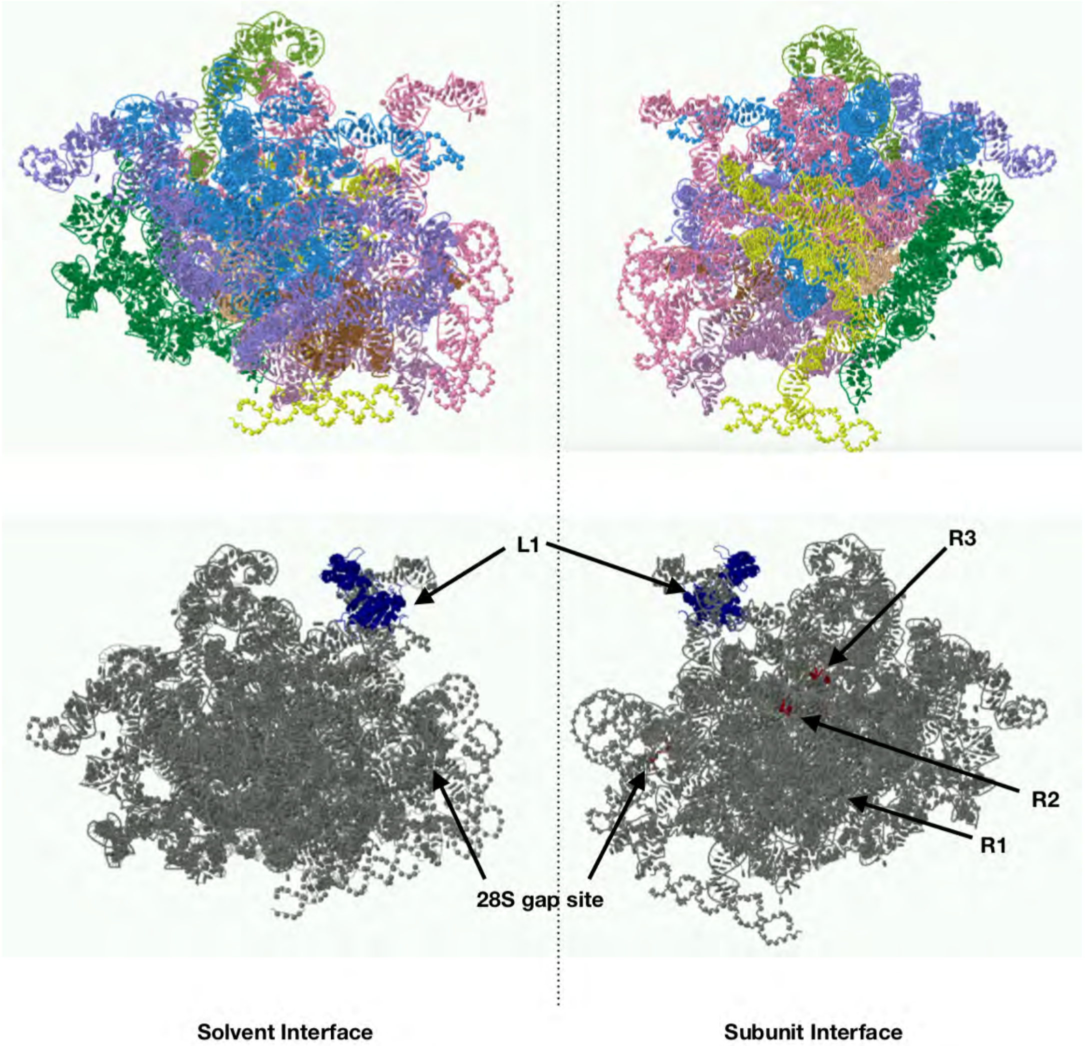


Figure 2: Partial 2D structure of *Drosophila* large ribosomal subunit (high similarity to *Sciara*), with R1/R2/R3 insertion sites denoted.



Domains: 28S 0 - Tan; I - Dark Purple; II - Blue; III - Light Purple; IV - Yellow; V - Pink; VI - Dark Green; 5S - Light Green; 2S/5.8S - Brown

Figure 3: 3D structures of *Drosophila* (high similarity to *Sciara*) large ribosomal subunit (Ribovision)

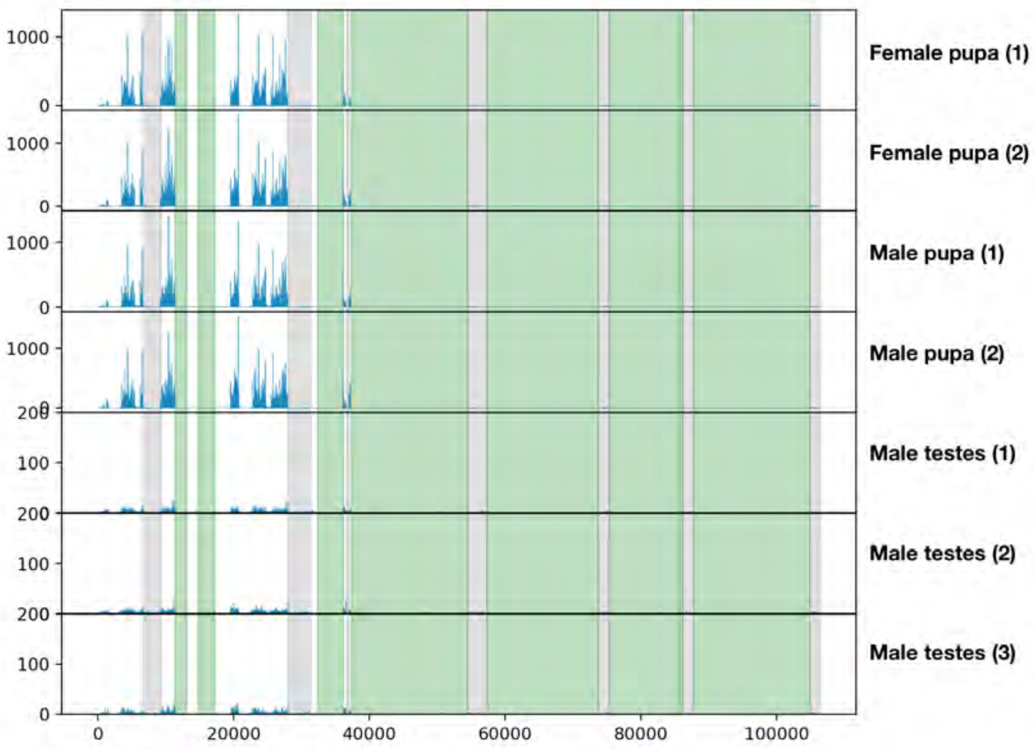


Figure 4: Total reads mapped, normalized by size factor, along a contig containing rDNA (white) interrupted by retrotransposons (gray) and unique non-rDNA (green).

Extended Data (Part 2)

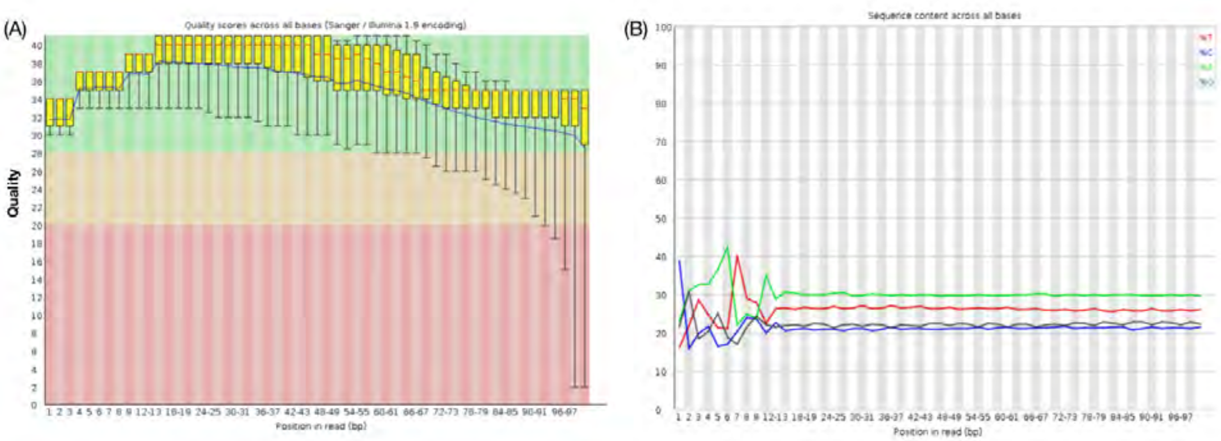


Figure 1: Quality control by FastQC. (A) Per-base sequence quality. (B) Per-base sequence content.

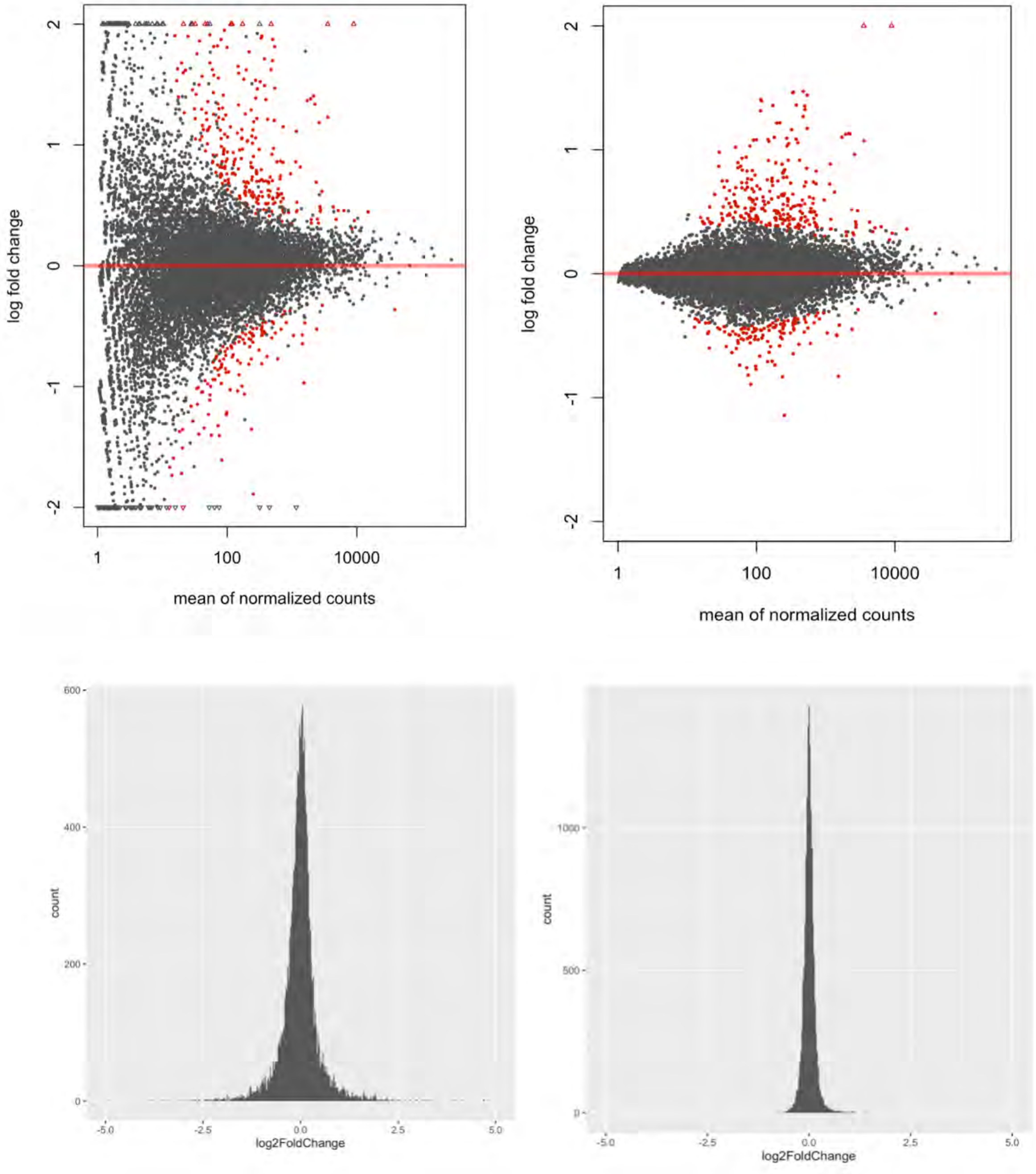


Figure 2: A comparison of shrunken (right) and non-shrunken (left) log-fold changes with respect to mean count.

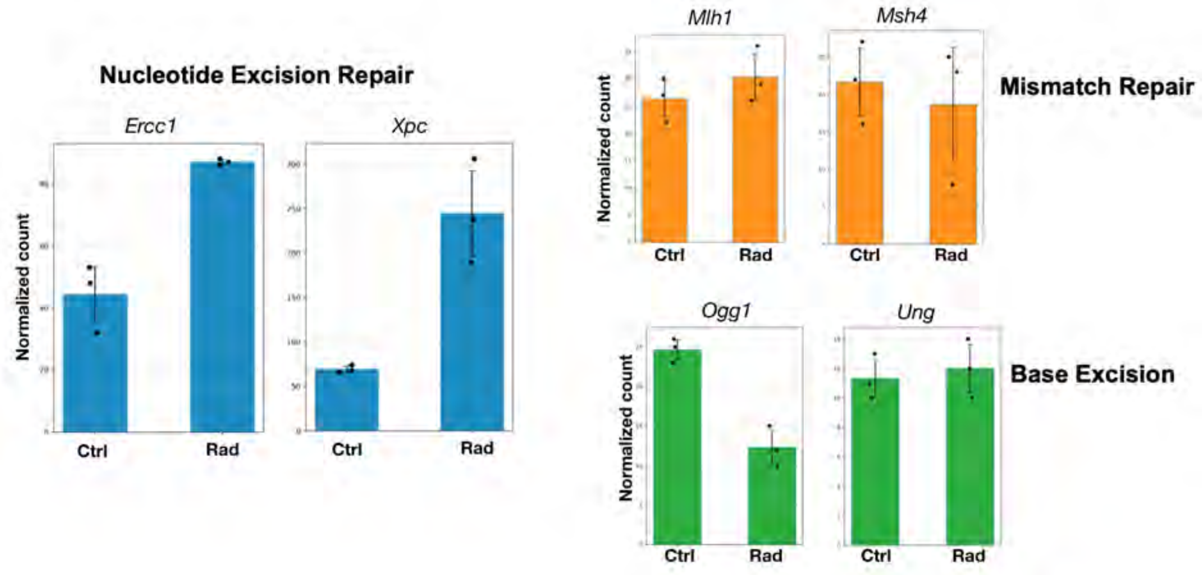


Figure 3: DESeq2-normalized counts of genes in three DNA repair pathways.