DISSERTATION

UNCOVERING MECHANISMS DRIVING VARIATION IN MUTATION RATES IN ORGANELLAR AND NUCLEAR GENOMES

Submitted by

Gus Waneka

Department of Biology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2024

Doctoral Committee:

Advisor: Daniel B. Sloan

Rachel Lockridge Mueller
Juan Lucas Argueso
Mark Stenglein

ABSTRACT

UNCOVERING MECHANISMS DRIVING VARIATION IN MUTATION RATES IN ORGANELLAR AND NUCLEAR GENOMES

Mutations are changes to DNA sequences which drive evolution by supplying raw genetic variation for natural selection to act upon. At the same time, mutations tend to have negative fitness consequences and are the source of genetic diseases. Such costs and benefits of mutation create opposing forces of selection on mutation rate modifiers, which are alleles (typically in DNA repair genes) responsible for increases or decreases to mutation rates.

Essentially all eukaryotes possess at least two genomes: the nuclear genome (nucDNA) and the endosymbiotically derived mitochondrial genome (mtDNA). Photosynthetic plants and algae additionally possess endosymbiotically derived plastid genomes (cpDNAs). Together, the mtDNA and cpDNA are referred to as organellar genomes. Chapter 1 of this dissertation provides a framework for understanding how DNA repair machinery and mutation rates have evolved in complex eukaryotic cells. Chapters 2 and 3 focus on specific repair pathways active in organellar genomes. Finally, Chapter 4 shifts focus to understand how environmental perturbations in expression level impact mutation in plant nuclear genomes.

Repair of organellar genomes is conducted by nuclear-encoded genes that are translated in the cytosol and targeted to the organelles. In terms of evolutionary history, organellar repair machinery is a mosaic network of bacterial-like repair genes, which came into the cell with the organelles, and nuclear repair genes that have been co-opted for organellar function. In some cases, repair proteins are targeted to both the nucDNA and mtDNA (and/or cpDNA in plants) to

perform similar functions. This is the case as for many base excision repair (BER) proteins, which identify and remove of chemically damaged bases. In contrast, organellar repair arsenals are thought to lack canonical mismatch-repair (MMR) and nucleotide excision repair (NER), which are both important repair pathways in nuclear genomes.

Instead, the diverse eukaryotic lineages have adopted unique strategies for organellar genome maintenance. As a result, there is a tremendous diversity in mtDNA mutation rates, which show over a 4000-fold variation across eukaryotes. Interestingly, much of this variation is driven by the extremely low point mutation rates plant mtDNAs. In addition, plant organellar genomes are more recombinationally active and plant mtDNAs are structurally unstable compared to the mtDNAs of other eukaryotes.

Chapter 2 of this dissertation explores the mechanistic basis of low point mutation rates and recombination-mediated repair in plant organellar genomes. We performed high-fidelity Duplex Sequencing on a panel of *Arabidopsis thaliana* lines lacking specific organellar genome repair genes. We report large point mutation increases in mutant lines lacking *MSH1*, a *mutS* homolog that has been proposed to induce double-stranded breaks at the site of DNA mismatches, effectively shuttling such lesions into homologous recombination (HR) pathways that play important roles in plant organellar genome replication and repair. We see smaller point mutation increases in other mutant lines lacking *RADA*, *RECA1* and *RECA3*. In addition, we generated long-read Oxford Nanopore sequencing to characterize repeat-mediated recombination in several of the mutants in the panel. Our findings provide valuable insights into the mechanisms driving the fascinating patterns of organellar genome evolution in land plants.

The aforementioned lack of NER in organellar genomes is surprising given the importance of NER as a ‘catchall’ for repair of a variety of bulky DNA lesions in nuclear

genomes. Chapter 3 focuses in on the fate of photodamage (an important type of bulky DNA damage) in organellar DNA. To do so, we leverage publicly available XR-seq datasets, which were generated to quantify and map active NER excision products in nuclear genomes following UV exposure. The taxonomic scope of chapter is expanded from plants to also include fungi (the brewer's yeast *Saccharomyces cerevisiae*) and animals (the model fruit fly *Drosophila melanogaster*). We find that mtDNA-derived XR-seq reads in *A. thaliana* and *S. cerevisiae* have distinct and repeatable patterns in terms of length and internal positioning of pyrimidine dimers (known targets of photodamage formation). These data mirror established patterns of NER-derived reads originating from the nuclear genomes, raising the exciting possibility that NER-like repair pathways may exist in for repair of photodamage in organellar genomes.

The focus of chapter 4 shifts to understanding how environmental changes impact mutation in plant nuclear genomes. The textbook view of mutation and adaptation is that mutations occur randomly with respect to their environment-specific fitness consequences. However, this view of random mutation is challenged as evidence increasingly establishes a correlation between increased expression and decreased mutation via the coordination of transcription and DNA repair machinery at the molecular level. As a result of this correlation, intragenomic mutation rates likely vary with changing environments given that expression levels are environmentally labile. Therefore, certain genes may be predisposed to higher or lower mutation rates depending on the environment, though the magnitude and importance of this effect remains largely untested. A technical challenge in addressing these questions is that large scale plant mutation datasets are time and resource intensive to generate. A recent plant study relied on low frequency somatic calls from Illumina based shotgun libraries to generate a large

number of mutations, but others report that most of these inferred mutations are sequencing errors.

To overcome these challenges, we took a novel approach to measuring somatic mutations by using Duplex Sequencing to quantify mutations in targeted regions of the *A. thaliana* nucDNA. We identified a set of differentially expressed genes in plants grown at different temperatures, which we then targeted for mutation detection using hybrid capture. In addition to wild type (WT) lines, we also studied mutant lines deficient in BER and MMR to test if either of these pathways are responsible for the correlation between expression and mutation in plants. We found large point mutation increases in the MMR mutants compared to WT plants, which displayed surprisingly few mutations at either temperature. Though the small number of WT mutations precluded a meaningful comparison of expression and mutation in a WT background, this result if nonetheless valuable for establishing that the true frequency of somatic mutations in plants is indeed very low suggesting that previous estimates likely conflated Illumina based sequencing artifacts with mutations.

Mutation rates vary by over three orders of magnitude across the tree of life. Much of this variation is captured in mitochondrial mutation rates. The chapters of this dissertation provide valuable insights into the molecular processes that drive mutation rate variation in eukaryotic genomes. Such mechanistic understandings are critical for advancing the broader field of mutation rate evolution.

ACKNOWLEDGEMETS

I am extremely proud of the work presented in this dissertation, and none of it would have been possible without my wonderful scientific and personal support systems. At the helm of my scientific support network is my advisor, Dan Sloan. Working with Dan has been one of the major highlights of my graduate school experience. I do not know how Dan manages to oversee and co-ordinate all the diverse research interests in his lab. Despite his many commitments, he has been incredibly present whenever I have needed guidance. Over the course of my PhD, Dan has schooled me in topics ranging from effective science communication, technical lab protocols, big picture evolutionary principals, and everything in between. Yet the most valuable lesson I've learned from Dan is that when someone brings genuine interest and passion to their work, they can be tremendously successful, elevate those around them, and enjoying themselves in the process.

I am also thankful to the professors on my committee and other collaborators who have enriched my PhD. I am especially grateful to have had nearly weekly interactions with my committee member Rachel Mueller at the Sloan/Mueller joint lab meetings. I have appreciated her positive outlook on science and the excitement she brings to mentoring and learning ever since I had the privilege to take a class from her as an undergrad. Mark Stenglein and Lucas Argueso were also a joy to have on my committee, and I really appreciate their unique perspectives and constructive suggestions.

I would also like to thank other members of the Sloan lab who have provided valuable feedback and stimulating conversations throughout my PhD. Amanda Broz has been a fantastic scientist to work with and I equally appreciate her mentorship in the lab, her questions at

seminars, and her perspectives on life in general. I'm also grateful to the two prior Sloan lab PhDs; Jessica Warren and Alissa Williams, who set a high bar and provided a model for how to navigate the various challenges associated with this process.

Given the innumerable interests and perspectives I share with Evan Forsythe, I consider it to be a stroke of luck that he joined the Sloan lab as a postdoc around the same time I started my PhD. Evan is a fantastic researcher and science mentor and I appreciate his feedback on countless research ideas and practice presentations. Moreover, I will cherish all the lasting memories we made fishing, running, camping, rafting, barbequing, and playing rock-and-roll together. Evan was the out-of-towner when we first started in the Sloan lab, but thanks to his gregarious personality he instantly hit it off with my friends and quickly became the hub of multiple overlapping friends' groups in Fort Collins. As with the rest of the Sloan lab alumni, I am excited to see where Evan's research takes him in the future.

Much of my social life through grad school revolved around seeing concerts and playing in different music projects and bands. Music was a fantastic outlet to balance my scientific pursuits and I have always particularly enjoyed connecting with my friends at band practice. Thank you to Maxwell, Adam, Hans, Jerrell, Nash, Lana, McKenzie, Evan, Mason, Tobias, and Andy. You are all incredible friends and musicians, and I am so lucky to have had the chance to share the stage with you at one point over the last 5 years. I'm also grateful to my non-band friends who were always there to cheer us on. Cole, Nat, Isaac, Sidney, Joshua, Ashley, Saksham, Mike, Andrew, Julie, Marion, Pat, Erika, Taylor, Tony, Alyssa, Tana, Bridgette, Brye and countless others – you all are the real rockstars. Thanks for being there and for giving my bands a reason to practice. Of course, it wasn't all music with these great friends, and I am also

grateful for all the hilarious game-nights, barbeques, egg-races, camping trips and other shenanigans.

I am so grateful for the constant love and support I have received from my family, not just during this doctorate, but throughout my life. Lana is my rock, and I would like to thank her for supporting me in every way. She provided comfort and homecooked meals when the road was rocky and was there to celebrate even the smallest of successes. Most of all, she helped to bring balance to my life as a scientist and provided a constant motivation to complete my work goals so that I could move onto whatever fun adventure we had planned. I am so grateful for all the hikes, runs, meals, boardgames, road trips, international vacations, dogs, fish, roommates, house projects, bike rides, pottery classes, concerts, dinner parties and other experiences I have shared with Lana – the memories we made are the highlight of my time at CSU. I am also grateful to Lana's wonderful parents, Kris and Bob, who have always been extremely supportive and have provided a remarkable model for how to make the most out of life, even in the face of adversity.

I am fortunate to have grown up with three wonderful older siblings who have all been awesome role models for me. My siblings and their partners constantly build me up and help me to have confidence in myself. Lastly, my parents are two of the most supportive, selfless, hardworking people in the world. I am so grateful for all the sacrifices they made to open doors and opportunities which allowed me to walk this path. In addition, there were many times in the last 5 years when they came through in a big way to help us with a house project or a homecooked meal. I owe my work ethic to them, and without that I would not have been able to complete this dissertation.

TABLE OF CONTENTS

CHAPTER 1: THE EVOLUTION OF MUTATION RATES

Natural selection drives evolution by favoring alleles that improve organismal fitness. Over time, selection spreads beneficial alleles through a given population. As such, natural selection often gets the credit for the tremendous diversity of life on earth, but evolution would not be possible without another process – mutation, which is most simply defined as a change to a DNA sequence. Mutation is the ultimate source of biological diversity since it creates genetic variation, thus providing natural selection different alleles to 'select'. Without mutation, evolution as we know it would come screeching to a halt. At the same time, most new mutations negatively impact organismal fitness. In some cases, a single inherited mutation can have fatal consequences (de Vernejoul and Kornak 2010). *De-novo* mutations that arise during an organism's life can also be lethal, as is the case with many types of cancer (Veltman and Brunner 2012; Martínez-Jiménez *et al.* 2020).

Mutations come in many varieties, ranging from simple events like single nucleotide substitutions to large scale events such as chromosomal fusions (Ventura *et al.* 2012). In many cases, mutations occur in a stepwise manner where initially DNA is damaged, and then the damage is turned into a mutation when the genome is replicated. In between these steps, damage may be targeted for repair by cellular machinery. The individual DNA repair proteins that make up this machinery are themselves encoded in the genome. Interestingly, repair genes can also incur mutations, thereby creating variation in the efficacy of the DNA repair machinery. This variation is itself subject to natural selection, which means mutation rates are an evolving trait with an optimum fitness state (Sniegowski *et al.* 2000). This led the early geneticist Alfred

Sturtevant (Sturtevant 1937) to pose the famous question, 'Why does the mutation rate not become reduced to zero'?

In the intervening decades, biologists have amassed droves of mutation rate estimates in diverse organisms across the tree of life, aided in recent years by DNA sequencing advances. Several potential answers to Sturtevant's question have emerged. One idea, termed the drift barrier hypothesis (DBH), is rooted in the population genetics principle that natural selection's ability to drive a trait towards a fitness optimum is positively related to the selective advantage of a trait and inversely related to the effective population size (N_e) of a given population (Kimura 1967; Lynch 2008; Sung *et al.* 2012; Lynch *et al.* 2016). The DBH assumes that since most new mutations have negative fitness consequences (Eyre-Walker and Keightley 2007), natural selection should always favor anti-mutator alleles, which are defined as alleles that lower the mutation rate. Non-zero mutation rates therefore arise from natural selection's inability to fix anti-mutator alleles because they have a small effect (i.e. negligible decreases to mutation rates) relative to the sampling variance associated with small N_e (i.e. genetic drift; (Lynch *et al.* 2016).

The utility of the DBH as a unifying theory for the evolution of mutation rates is demonstrated by the strikingly strong correlation between N_e and mutation rates (per base-pair per generation) across the entire tree of life (Lynch 2010). The mutation rates of viruses, bacteria and multicellular eukaryotes can all be reasonably estimated as a function of N_e (Lynch *et al.* 2016; Bergeron *et al.* 2023). Yet, there are notable exceptions; rulebreakers that do not conform to the DBH which provide valuable insights for understanding mutation rate evolution more broadly.

The DBH predicts that mutation rates in mitochondrial genomes (mtDNAs) should be higher than in nuclear genomes (nucDNAs) since 1) mtDNAs have a decreased N_e because they

are uniparentally inherited and non-recombining and 2) they are orders of magnitude smaller and therefore have far fewer sites under selection than nuclear genomes, reducing the potential selective advantage of anti-mutator alleles (Lynch *et al.* 2006, 2016). For many eukaryotes, including most metazoans, this prediction of the DBH is satisfied and mtDNA point mutation rates are often about an order of magnitude higher than those of nuclear genomes (Wolfe *et al.* 1987; Havird and Sloan 2016). In land plants, however, the pattern is reversed, as mtDNA point mutation rates are often about an order of magnitude lower than those of nucDNAs (Wolfe *et al.* 1987; Drouin *et al.* 2008). Land plant plastid genome (cpDNA) mutation rates are similarly low (Smith and Keeling 2015). A mechanistic understanding of the repair machinery protecting land plant organellar genomes is important for reconciling organellar mutation rate evolution with the DBH.

Low point mutation rates are not the only feature differentiating land plant mtDNA evolution from other eukaryotes. Land plants mtDNAs, with an average size of 395 kb (Wu *et al.* 2022), are much larger than mtDNAs of other eukaryotes, which are typically less than 20 kb in metazoans and range from ~20 to 235 kb in fungi (Sandor *et al.* 2018). In addition, plant mtDNAs are recombinationally active and display extreme structural instability, as evidenced by the lack of shared synteny in the mtDNAs of closely related plant species (Handa 2003; Kubo and Newton 2008; Skippington *et al.* 2017; Gualberto and Newton 2017; Brieba 2019; Chevigny *et al.* 2020). It has been hypothesized that the seemingly disparate features of land plant mtDNA evolution (low point mutation and rapid recombination rates) may be explained by an unusual DNA repair strategy where DNA mismatches (the progenitors of point mutations) are converted into double stranded breaks and resolved via homology directed repair (Christensen 2014, 2018; Ayala-García *et al.* 2018; Broz *et al.* 2022). A tradeoff where low point mutation rates are

achieved at the expense of increased recombination explains how land plant organellar genomes seem to defy the DBH, since the DBH is narrowly focused on point mutations while other types of mutation, including structural rearrangement, tend to dominate land plant organellar genomes.

Characterization of the repair mechanisms responsible for low point mutations in plant organellar genomes has been hampered by the technical challenge of obtaining direct measurements of mutation, given the high error-rate of Illumina sequencing (Schaack *et al.* 2020). Fortunately several high-fidelity approaches have been developed which use unique molecular identifiers to create consensus sequences from PCR duplicates (Sloan *et al.* 2018), the most accurate of which is Duplex Sequencing (Schmitt *et al.* 2012). In chapter 2 of this dissertation, we implemented our optimized Duplex Sequencing protocol (Wu *et al.* 2020) to characterize the relationship between point mutation and recombination in a panel of mutant *Arabidopsis thaliana* lines with deficiencies in organellar homologous recombination pathways.

Another important question impacting plant organellar mutation rates is determining how photodamage is processed or repaired in organellar genomes (Gualberto and Newton 2017). Photodamage is a type of bulky DNA damage induced by ultra-violet (UV) radiation and is known to be a potent mutagen capable of disrupting numerous biological processes including transcription and DNA replication (Sinha and Häder 2002). In bacterial and nuclear genomes, well characterized nucleotide excision repair (NER) pathways are responsible for removing photodamage, as well as other types of bulky DNA damage. NER pathways evolved independently in prokaryotes and eukaryotes but share common mechanistic features. First, NER machinery recognizes bulky DNA damage and then endonucleases make single-stranded incisions upstream and downstream of the damaged site. The damage containing single-stranded

oligo is then excised and the resulting gap is filled in using the complementary strand as a template (Sancar 1996).

Attempts to identify an analogous repair pathway for repair of bulky DNA damage in mtDNAs have so far been unsuccessful (Clayton *et al.* 1974; Waters and Moustacchi 1974; Prakash 1975; LeDoux *et al.* 1992; Hunter *et al.* 2010; Saki and Prakash 2017). Photolyases, which can directly reverse photodamage, have been shown to play a role in organellar genome repair in some eukaryotes (Prakash 1975; Takahashi *et al.* 2011). However, while NER is a 'catch-all' for the many different classes of bulky DNA damage, photolyases are only capable of repairing photodamage (an important subset of bulky DNA damage) and even then, they are damage specific. For example, entirely different photolyases are required to repair the two most common types of photodamage: cyclobutane pyrimidine dimers (CPDs) and pyrimidine-pyrimidone (6-4) photoproducts ((6-4)PPs; (Lucas-Lledó and Lynch 2009). Compared to other repair machinery which tends to be broadly conserved, photolyases are missing from many eukaryotic lineages (Mei and Dvornyk 2015). For example, mammals lack photolyases completely and, while the some fungi have a complete photolyase arsenal, the brewer's yeast *Saccharomyces cerevisiae* possesses a CPD-photolyase but lacks a (6-4)PP photolyase (Sancar 2004).

In the absence of NER-like repair or photolyase protection it has been proposed that photodamaged mtDNA may simply be degraded and destroyed (Clayton *et al.* 1974; Waters and Moustacchi 1974). However, mechanistic details of mtDNA degradation pathways remain uncharacterized and it is unclear how new mtDNA copies could be synthesized in instances where every mtDNA copy in a cell receives photodamage (Takami *et al.* 2018; Zhao 2019). A historic barrier in understanding the fate of photodamage in organellar DNA is the difficulty in

distinguishing signals of damage and repair from the nuclear vs. organellar genomes (Zhao and Sumberaz 2020).

A recently developed technique called excision repair sequencing (XR-Seq for short) provides a solution to this challenge. XR-Seq utilizes anti-damage antibodies to capture and sequence the short oligos that are excised during NER, providing nucleotide specific snapshots of active repair locations following UV exposure (Hu *et al.* 2015). UV response experiments utilizing XR-Seq have been performed in a number of model eukaryotic species (Li *et al.* 2018; Oztas *et al.* 2018; Deger *et al.* 2019) but the organellar originating sequencing products have remained largely unexplored.

Chapter 3 of this dissertation leverages these datasets to gain insights into the fate of photodamage in organellar DNA. The taxonomic scope of chapter 3 is expanded from plants (Oztas *et al.* 2018) to also include fungi (*S. cerevisiae*; Li *et al.* 2018)) and animals (the model fruit fly *Drosophila melanogaster*; Deger *et al.* 2019). Together, chapters 2 and 3 provide mechanistic insights into how organellar genomes are maintained and repaired, informing broader questions about organellar genome evolution, including how low point mutations in plant organellar genomes can be reconciled with the DBH.

Another answer to Sturtevant's famous question, 'Why does the mutation rate not become reduced to zero?' is that increases to mutation rates can in some cases be adaptive, resulting in the selective spread of mutator alleles through a population (Agashe *et al.* 2023). The selective advantage of mutators directly challenges the DBH assumption that selection always favors decreased mutation rates (Lynch *et al.* 2016). Mutators are not directly beneficial, but spread by hitch-hiking with the new, beneficial alleles they introduce (Sniegowski *et al.* 2000; Raynes and Sniegowski 2014). Given that sexual recombination disrupts the linkage between

beneficial alleles and mutators, it is not surprising that the experimental studies documenting mutator hitch-hiking have primarily been conducted with clonal bacteria (Chao and Cox 1983; Giraud *et al.* 2001; Shaver *et al.* 2002; Labat *et al.* 2005; Gentile *et al.* 2011; Wiser *et al.* 2013) and asexual yeast (Thompson *et al.* 2006; Raynes *et al.* 2011, 2012, 2014).

Longer-term experimental evolution studies have established that increased mutation rates associated with the spread of mutators are often transient (Raynes and Sniegowski 2014). As genetic load accumulates in individuals with high mutations rates they suffer fitness declines (Thompson *et al.* 2006; Raynes *et al.* 2011) and are eventually outcompeted by organisms with decreased genetic load and associated anti-mutator alleles, resulting in a reversion of the mutation rate back toward WT levels (Wielgoss *et al.* 2013; Turrientes *et al.* 2013). Alleles that perturb existing mutation biases (but do not necessarily increase mutation rates) can confer organisms with increased adaptive potential without increasing genetic load (Agashe *et al.* 2023; Tuffaha *et al.* 2023). Two common examples of mutational biases which are fairly widespread across the tree of life are higher transition than transversion rates and higher deletion than insertion rates (Hodgkinson and Eyre-Walker 2011; Quiroz *et al.* 2023; Horton and Taylor 2023). The conclusion of these studies is that in changing environments, alleles which modify mutational biases are adaptive as long as they increase the supply of previously under-sampled mutational classes (Sane *et al.* 2023).

Exposure to novel or stressful environments can also cause perturbations to mutation rates and biases, as demonstrated in diverse taxa from bacteria (Maharjan and Ferenci 2017) to plants (Jiang *et al.* 2014; Belfield *et al.* 2021; Lu *et al.* 2021). Such perturbations inevitably increase the supply of under-sampled mutational classes, raising the somewhat surprising

possibility that stressful environments may predispose organisms to mutational rates and biases with increased adaptive potential (Monroe *et al.* 2022; Quiroz *et al.* 2023; Monroe 2023).

The evolutionary implications of environmentally labile mutation profiles are profound (Luria and Delbrück 1943) and are explored in detail in chapter 4. It is equally interesting to consider how bias shifts unfold at the molecular level. In prokaryotes, certain nutrient deficient mediums have been shown to trigger the downregulation of high-fidelity machinery and the upregulation of error-prone repair pathways (Foster 2007; Agashe 2017; Maharjan and Ferenci 2017). In eukaryotes similar results may be achieved through the coupling of DNA repair and transcription machinery. In transcription-coupled NER, RNA polymerases that have become stalled on bulky DNA lesions are used by repair machinery to identify areas in need of repair. As a result, template DNA strands bound by RNA polymerases are repaired at a much higher rate than coding strands (Hu *et al.* 2015; Oztas *et al.* 2018; Yang *et al.* 2018; Deger *et al.* 2019). Similarly, in plants it has been proposed that mismatch repair (MMR) may provide preferential protection to actively transcribed genes based on the differential histone modifications of these regions (Quiroz *et al.* 2022). Therefore, when genes are downregulated in stressful environments, they likely receive less protection from TC-NER and MMR machinery, potentially increasing opportunities for adaptive mutations to occur.

The fourth and final chapter of this dissertation is designed to test mechanisms of stress induced mutagenesis and adaptation in *A. thaliana.* We identified a set of differential expressed nucDNA genes in plants grown at different temperatures, which we then targeted for mutation detection using hybrid capture and Duplex Sequencing. In addition to WT lines, we also studied mutant lines deficient in base excision repair (BER) and MMR to test if either of these pathways are responsible for the correlation between expression and mutation in plants.

Throughout the coming pages I explore mechanisms of damage and repair, placing each distinct study into the relevant evolutionary framework. Chapters 2 and 3 explore big questions concerning the tremendous diversity of organellar genome sizes, structures, and mutation rates. Chapter 2 provides novel insights into how low point mutations are achieved through recombination mediated repair in plant organellar genomes. Chapter 3 reframes a previously dismissed repair pathway in organellar genomes through the creative analysis of publicly available sequencing datasets. The exciting patterns described in chapter 3 raise the possibility of that an NER-like repair pathway is functional in organellar genomes. Chapter 4 shifts focus to the mechanisms responsible for generating adaptive mutations under changing environments. This dissertation is an important step toward understanding the mechanisms that drive mutation across the multiple genomic compartments of eukaryotic cells.

LITERATURE CITED

Agashe D., 2017 The road not taken: Could stress-specific mutations lead to different evolutionary paths? PLoS Biol. 15: e2002862.

Agashe D., M. Sane, and S. Singhal, 2023 Revisiting the Role of Genetic Variation in Adaptation. Am. Nat. 202: 486–502.

Ayala-García V. M., N. Baruch-Torres, P. L. García-Medel, and L. G. Brieba, 2018 Plant organellar DNA polymerases paralogs exhibit dissimilar nucleotide incorporation fidelity. FEBS J. 285: 4005–4018.

Belfield E. J., C. Brown, Z. J. Ding, L. Chapman, M. Luo, *et al.*, 2021 Thermal stress accelerates Arabidopsis thaliana mutation rate. Genome Res. 31: 40–50.

Bergeron L. A., S. Besenbacher, J. Zheng, P. Li, M. F. Bertelsen, *et al.*, 2023 Evolution of the germline mutation rate across vertebrates. Nature 615: 285–291.

Brieba L. G., 2019 Structure-Function Analysis Reveals the Singularity of Plant Mitochondrial DNA Replication Components: A Mosaic and Redundant System. Plants 8. https://doi.org/10.3390/plants8120533

Broz A. K., A. Keene, M. F. Gyorfy, M. Hodous, I. G. Johnston, *et al.*, 2022 Sorting of mitochondrial and plastid heteroplasmy in *Arabidopsis* is extremely rapid and depends on MSH1 activity. Proceedings of the National Academy of Sciences 119: e2206973119.

Chao L., and E. C. Cox, 1983 Competition Between High and Low Mutating Strains of Escherichia coli. Evolution 37: 125–134.

Chevigny N., D. Schatz-Daas, F. Lotfi, and J. M. Gualberto, 2020 DNA Repair and the Stability of the Plant Mitochondrial Genome. Int. J. Mol. Sci. 21. https://doi.org/10.3390/ijms21010328

Christensen A. C., 2014 Genes and Junk in Plant Mitochondria—Repair Mechanisms and Selection. Genome Biol. Evol. 6: 1448–1453.

Christensen A. C., 2018 Mitochondrial DNA Repair and Genome Evolution. Annual Plant Reviews online 11–32.

Clayton D. A., J. N. Doda, and E. C. Friedberg, 1974 The absence of a pyrimidine dimer repair mechanism in mammalian mitochondria. Proc. Natl. Acad. Sci. U. S. A. 71: 2777–2781.

Deger N., Y. Yang, L. A. Lindsey-Boltz, A. Sancar, and C. P. Selby, 2019 Drosophila, which lacks canonical transcription-coupled repair proteins, performs transcription-coupled repair. J. Biol. Chem. 294: 18092–18098.

Drouin G., H. Daoud, and J. Xia, 2008 Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol. Phylogenet. Evol. 49: 827–831.

Eyre-Walker A., and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8: 610–618.

Foster P. L., 2007 Stress-induced mutagenesis in bacteria. Crit. Rev. Biochem. Mol. Biol. 42: 373–397.

Gentile C. F., S.-C. Yu, S. A. Serrano, P. J. Gerrish, and P. D. Sniegowski, 2011 Competition between high- and higher-mutating strains of Escherichia coli. Biol. Lett. 7: 422–424.

Giraud A., I. Matic, O. Tenaillon, A. Clara, M. Radman, *et al.*, 2001 Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. Science 291: 2606–2608.

Gualberto J. M., and K. J. Newton, 2017 Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation. Annu. Rev. Plant Biol. 68: 225–252.

Handa H., 2003 The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (Brassica napus L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana. Nucleic Acids Res. 31: 5907–5916.

Havird J. C., and D. B. Sloan, 2016 The Roles of Mutation, Selection, and Expression in Determining Relative Rates of Evolution in Mitochondrial versus Nuclear Genomes. Mol. Biol. Evol. 33: 3042–3053.

Hodgkinson A., and A. Eyre-Walker, 2011 Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. 12: 756–766.

Horton J. S., and T. B. Taylor, 2023 Mutation bias and adaptation in bacteria. Microbiology 169. https://doi.org/10.1099/mic.0.001404

Hu J., S. Adar, C. P. Selby, J. D. Lieb, and A. Sancar, 2015 Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. Genes Dev. 29: 948–960.

Hunter S. E., D. Jung, R. T. Di Giulio, and J. N. Meyer, 2010 The QPCR assay for analysis of mitochondrial DNA damage, repair, and relative copy number. Methods 51: 444–451.

Jiang C., A. Mithani, E. J. Belfield, R. Mott, L. D. Hurst, *et al.*, 2014 Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. Genome Res. 24: 1821–1829.

Kimura M., 1967 On the evolutionary adjustment of spontaneous mutation rates*. Genet. Res. 9: 23–34.

Kubo T., and K. J. Newton, 2008 Angiosperm mitochondrial genomes and mutations. Mitochondrion 8: 5–14.

Labat F., O. Pradillon, L. Garry, M. Peuchmaur, B. Fantin, *et al.*, 2005 Mutator phenotype confers advantage in Escherichia coli chronic urinary tract infection pathogenesis. FEMS Immunol. Med. Microbiol. 44: 317–321.

LeDoux S. P., G. L. Wilson, E. J. Beecham, T. Stevnsner, K. Wassermann, *et al.*, 1992 Repair of mitochondrial DNA after various types of DNA damage in Chinese hamster ovary cells. Carcinogenesis 13: 1967–1973.

Li W., O. Adebali, Y. Yang, C. P. Selby, and A. Sancar, 2018 Single-nucleotide resolution dynamic repair maps of UV damage in Saccharomyces cerevisiae genome. Proc. Natl. Acad. Sci. U. S. A. 115: E3408–E3415.

Lu Z., J. Cui, L. Wang, N. Teng, S. Zhang, *et al.*, 2021 Genome-wide DNA mutations in Arabidopsis plants after multigenerational exposure to high temperatures. Genome Biol. 22: 160.

Lucas-Lledó J. I., and M. Lynch, 2009 Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. Mol. Biol. Evol. 26: 1143–1153.

Luria S. E., and M. Delbrück, 1943 Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics 28: 491–511.

Lynch M., B. Koskella, and S. Schaack, 2006 Mutation pressure and the evolution of organelle genomic architecture. Science 311: 1727–1730.

Lynch M., 2008 The cellular, developmental and population-genetic determinants of mutation-rate evolution. Genetics 180: 933–943.

Lynch M., 2010 Evolution of the mutation rate. Trends Genet. 26: 345–352.

Lynch M., M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, *et al.*, 2016 Genetic drift, selection and the evolution of the mutation rate. Nat. Rev. Genet. 17: 704–714.

Maharjan R. P., and T. Ferenci, 2017 A shifting mutational landscape in 6 nutritional states: Stress-induced mutagenesis as a series of distinct stress input-mutation output relationships. PLoS Biol. 15: e2001477.

Martínez-Jiménez F., F. Muiños, I. Sentís, J. Deu-Pons, I. Reyes-Salazar, *et al.*, 2020 A compendium of mutational cancer driver genes. Nat. Rev. Cancer 20: 555–572.

Mei Q., and V. Dvornyk, 2015 Evolutionary History of the Photolyase/Cryptochrome Superfamily in Eukaryotes. PLoS One 10: e0135940.

Monroe J. G., T. Srikant, P. Carbonell-Bejerano, C. Becker, M. Lensink, *et al.*, 2022 Mutation bias reflects natural selection in Arabidopsis thaliana. Nature 602: 101–105.

Monroe G., 2023 Are mutations random? Evolution 77: 2522–2527.

Oztas O., C. P. Selby, A. Sancar, and O. Adebali, 2018 Genome-wide excision repair in Arabidopsis is coupled to transcription and reflects circadian gene expression patterns. Nat. Commun. 9: 1503.

Prakash L., 1975 Repair of pyrimidine dimers in nuclear and mitochondrial DNA of yeast irradiated with low doses of ultraviolet light. J. Mol. Biol. 98: 781–795.

Quiroz D., K. Zhao, P. Carbonell-Bejerano, and G. Monroe, 2022 Biased mutagenesis and H3K4me1-targeted DNA repair in plants

Quiroz D., M. Lensink, D. J. Kliebenstein, and J. G. Monroe, 2023 Causes of Mutation Rate Variability in Plant Genomes. Annu. Rev. Plant Biol. 74: 751–775.

Raynes Y., M. R. Gazzara, and P. D. Sniegowski, 2011 Mutator dynamics in sexual and asexual experimental populations of yeast. BMC Evol. Biol. 11: 158.

Raynes Y., M. R. Gazzara, and P. D. Sniegowski, 2012 Contrasting dynamics of a mutator allele in asexual populations of differing size. Evolution 66: 2329–2334.

Raynes Y., A. L. Halstead, and P. D. Sniegowski, 2014 The effect of population bottlenecks on mutation rate evolution in asexual populations. J. Evol. Biol. 27: 161–169.

Raynes Y., and P. D. Sniegowski, 2014 Experimental evolution and the dynamics of genomic mutation rate modifiers. Heredity 113: 375–380.

Saki M., and A. Prakash, 2017 DNA damage related crosstalk between the nucleus and mitochondria. Free Radic. Biol. Med. 107: 216–227.

Sancar A., 1996 DNA EXCISION REPAIR. Annu. Rev. Biochem. 65: 43–81.

Sancar A., 2004 Photolyase and cryptochrome blue-light photoreceptors. Adv. Protein Chem. 69: 73–100.

Sandor S., Y. Zhang, and J. Xu, 2018 Fungal mitochondrial genomes and genetic polymorphisms. Appl. Microbiol. Biotechnol. 102: 9433–9448.

Sane M., G. D. Diwan, B. A. Bhat, L. M. Wahl, and D. Agashe, 2023 Shifts in mutation spectra enhance access to beneficial mutations. Proc. Natl. Acad. Sci. U. S. A. 120: e2207355120.

Schaack S., E. K. H. Ho, and F. Macrae, 2020 Disentangling the intertwined roles of mutation, selection and drift in the mitochondrial genome. Philos. Trans. R. Soc. Lond. B Biol. Sci. 375: 20190173.

Schmitt M. W., S. R. Kennedy, J. J. Salk, E. J. Fox, J. B. Hiatt, *et al.*, 2012 Detection of ultra-rare mutations by next-generation sequencing. Proc. Natl. Acad. Sci. U. S. A. 109: 14508–14513.

Shaver A. C., P. G. Dombrowski, J. Y. Sweeney, T. Treis, R. M. Zappala, *et al.*, 2002 Fitness evolution and the rise of mutator alleles in experimental Escherichia coli populations. Genetics 162: 557–566.

Sinha R. P., and D. P. Häder, 2002 UV-induced DNA damage and repair: a review. Photochem. Photobiol. Sci. 1: 225–236.

Skippington E., T. J. Barkman, D. W. Rice, and J. D. Palmer, 2017 Comparative mitogenomics indicates respiratory competence in parasitic Viscum despite loss of complex I and extreme sequence divergence, and reveals horizontal gene transfer and remarkable variation in genome size. BMC Plant Biol. 17: 1–12.

Sloan D. B., A. K. Broz, J. Sharbrough, and Z. Wu, 2018 Detecting Rare Mutations and DNA Damage with Sequencing-Based Methods. Trends Biotechnol. 36: 729–740.

Smith D. R., and P. J. Keeling, 2015 Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proc. Natl. Acad. Sci. U. S. A. 112: 10177–10184.

Sniegowski P. D., P. J. Gerrish, T. Johnson, and A. Shaver, 2000 The evolution of mutation rates: separating causes from consequences. Bioessays 22: 1057–1066.

Sturtevant A. H., 1937 Essays on Evolution. I. On the Effects of Selection on Mutation Rate. Q. Rev. Biol. 12: 464–467.

Sung W., M. S. Ackerman, S. F. Miller, T. G. Doak, and M. Lynch, 2012 Drift-barrier hypothesis and mutation-rate evolution. Proc. Natl. Acad. Sci. U. S. A. 109: 18488–18492.

Takahashi M., M. Teranishi, H. Ishida, J. Kawasaki, A. Takeuchi, *et al.*, 2011 Cyclobutane pyrimidine dimer (CPD) photolyase repairs ultraviolet-B-induced CPDs in rice chloroplast and mitochondrial DNA. Plant J. 66: 433–442.

Takami T., N. Ohnishi, Y. Kurita, S. Iwamura, M. Ohnishi, *et al.*, 2018 Organelle DNA degradation contributes to the efficient use of phosphate in seed plants. Nat Plants 4: 1044–1055.

Thompson D. A., M. M. Desai, and A. W. Murray, 2006 Ploidy controls the success of mutators and nature of mutations during budding yeast evolution. Curr. Biol. 16: 1581–1590.

Tuffaha M. Z., S. Varakunan, D. Castellano, R. N. Gutenkunst, and L. M. Wahl, 2023 Shifts in Mutation Bias Promote Mutators by Altering the Distribution of Fitness Effects. Am. Nat. 202: 503–518.

Turrientes M.-C., F. Baquero, B. R. Levin, J.-L. Martínez, A. Ripoll, *et al.*, 2013 Normal mutation rate variants arise in a Mutator (Mut S) Escherichia coli population. PLoS One 8: e72963.

Veltman J. A., and H. G. Brunner, 2012 De novo mutations in human genetic disease. Nat. Rev. Genet. 13: 565–575.

Ventura M., C. R. Catacchio, S. Sajjadian, L. Vives, P. H. Sudmant, *et al.*, 2012 The evolution of African great ape subtelomeric heterochromatin and the fusion of human chromosome 2. Genome Res. 22: 1036–1049.

Vernejoul M.-C. de, and U. Kornak, 2010 Heritable sclerosing bone disorders: presentation and new molecular mechanisms. Ann. N. Y. Acad. Sci. 1192: 269–277.

Waters R., and E. Moustacchi, 1974 The fate of ultraviolet-induced pyrimidine dimers in the mitochondrial DNA of Saccharomyces cerevisiae following various post-irradiation cell treatments. Biochim. Biophys. Acta 366: 241–250.

Wielgoss S., J. E. Barrick, O. Tenaillon, M. J. Wiser, W. J. Dittmar, *et al.*, 2013 Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc. Natl. Acad. Sci. U. S. A. 110: 222–227.

Wiser M. J., N. Ribeck, and R. E. Lenski, 2013 Long-term dynamics of adaptation in asexual populations. Science 342: 1364–1367.

Wolfe K. H., W. H. Li, and P. M. Sharp, 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. U. S. A. 84: 9054–9058.

Wu Z., G. Waneka, A. K. Broz, C. R. King, and D. B. Sloan, 2020 MSH1 is required for maintenance of the low mutation rates in plant mitochondrial and plastid genomes. Proc. Natl. Acad. Sci. U. S. A. 117: 16448–16455.

Wu Z.-Q., X.-Z. Liao, X.-N. Zhang, L. R. Tembrock, and A. Broz, 2022 Genomic architectural variation of plant mitochondria—A review of multichromosomal structuring. J. Syst. Evol. 60: 160–168.

Yang Y., O. Adebali, G. Wu, C. P. Selby, Y.-Y. Chiou, *et al.*, 2018 Cisplatin-DNA adduct repair of transcribed genes is controlled by two circadian programs in mouse tissues. Proc. Natl. Acad. Sci. U. S. A. 115: E4777–E4785.

Zhao L., 2019 Mitochondrial DNA degradation: A quality control measure for mitochondrial genome maintenance and stress response. Enzymes 45: 311–341.

Zhao L., and P. Sumberaz, 2020 Mitochondrial DNA Damage: Prevalence, Biological Consequence, and Emerging Pathways. Chem. Res. Toxicol. 33: 2491–2502.

CHAPTER 2: DISRUPTION OF RECOMBINATION MACHINERY ALTERS THE MUTATIONAL LANDSCAPE IN PLANT ORGANELLAR GENOMES

Summary

Land plant organellar genomes have extremely low rates of point mutation yet also experience high rates of recombination and genome instability. Characterizing the molecular machinery responsible for these patterns is critical for understanding the evolution of these genomes. While much progress has been made towards understanding recombination activity in land plant organellar genomes, the relationship between recombination and point mutation rates remains uncertain. The organellar targeted *mutS* homolog MSH1 has previously been shown to suppress point mutations as well as non-allelic recombination between short repeats. We therefore implemented high-fidelity Duplex Sequencing to test if other genes that function in recombination and maintenance of genome stability also affect point mutation rates. We found small to moderate increases in the frequency of single nucleotide variants (SNVs) and indels in mitochondrial and/or plastid genomes of mutant lines lacking *radA*, *recA1*, or *recA3*. In contrast, *osb2* and *why2* mutants did not exhibit an increase in point mutations compared to WT controls. In addition, we analyzed the distribution of SNVs in previously generated Duplex Sequencing data from *A. thaliana* organellar genomes and found unexpected strand asymmetries and large effects of flanking nucleotides on mutation rates in wild type plants and *msh1* mutants. Finally, using long-read Oxford Nanopore sequencing, we characterized structural variants in organellar genomes of the mutant lines and show that different short repeat sequences become recombinationally active in different mutant backgrounds. Together, these complementary

sequencing approaches shed light on how recombination may impact the extraordinarily low point mutation rates in plant organellar genomes.

INTRODUCTION

Nearly all eukaryotes rely on genes encoded in endosymbiotically derived mitochondrial genomes (mtDNAs) for cellular respiration. Plants and algae additionally rely on the endosymbiotically derived plastid genome (cpDNA) for photosynthesis. In several regards, land plant organellar genome evolution is atypical compared to mtDNA evolution in other eukaryotes (Smith and Keeling 2015). For one, plant organellar genomes have exceptionally low nucleotide substitution rates relative to those in plant nuclear genomes and to those of many other eukaryotic mtDNAs. The low substitution rates of plant organellar genomes extend even to synonymous sites, which likely experience very little purifying selection, suggesting that the cause of the low substitution rates is low point mutation rates (Wolfe *et al.* 1987; Drouin *et al.* 2008).

Compared to the small mtDNAs typical in metazoans (generally below 20 kb) and in algae and fungi (with sizes ranging from approximately 13 to 96 kb and ~20 to 235 kb, respectively), land plant mtDNAs are much larger with sequenced mtDNAs averaging 395 kb (Wu *et al.* 2022) and a known range extending from 70 kb to over 10 Mb (Boore 1999; Sloan *et al.* 2012; Skippington *et al.* 2017; Gualberto and Newton 2017; Sandor *et al.* 2018; Chen *et al.* 2019). Very little of this size variation stems from differences in coding capacity, as plant mtDNAs generally contain a subset of the same 41 protein-coding genes (Mower *et al.* 2012). Instead, the fluctuations in total mtDNA size primarily result from the acquisition and loss of noncoding DNA. Even closely related species possess very little shared noncoding sequences

(Kubo and Newton 2008; Skippington *et al.* 2017). For example, a comparative analysis of the mtDNAs of two species within the Brassicaceae, *Arabidopsis thaliana* (367 kb) and *Brassica napus* (222 kb), revealed a mere 78 kb of shared sequence, most of which is coding (Handa 2003). Though size variation of cpDNAs is less extreme than in plant mtDNAs, variation still exists with 98.7% of sequenced land plant cpDNAs ranging from 100-200 kb in size (Xiao-Ming *et al.* 2017).

Plant organellar genomes also experience exceptionally high rates of structural mutation and rearrangement (Palmer and Herbon 1988). As a result, there is virtually no conservation of synteny of plant mtDNAs, as evidenced by the extensive rearrangements in alignments of mtDNAs from Col-0 and Ler ecotypes of *A. thaliana* (Stupar *et al.* 2001; Huang *et al.* 2005; Davila *et al.* 2011; Pucker *et al.* 2019). The structural instability in plant mtDNAs is partly explained by the presence of repeats of various lengths, which recombine frequently and give rise to multiple isomeric subgenomes with circular, linear and/or branched structures (Palmer and Herbon 1988; Alverson *et al.* 2011; Wynn and Christensen 2019). In fact, plant mtDNAs lack origins of replication, which help coordinate genome replication in many other eukaryotes, and are instead thought to replicate through break induced recombination (Gualberto and Newton 2017; Chevigny *et al.* 2020). Land plant cpDNAs are also recombinationally active but usually remain structurally conserved, albeit with some significant exceptions (Smith and Keeling 2015).

The seemingly disparate features of plant organellar evolution (i.e. high rates of recombination and low rates of point mutation) may be unified through a DNA repair mechanism reliant on recombination (Christensen 2014). This hypothesized mechanism hinges on the activity of the *mutS* homolog MSH1 (Abdelnoor *et al.* 2003), which is dual-targeted to mitochondria and plastids and has long been known to suppress non-allelic recombination

between intermediate-sized repeats (50 to 600 bps) in the *A. thaliana* mtDNA (Martínez-Zapater *et al.* 1992; Arrieta-Montiel *et al.* 2009; Davila *et al.* 2011; Zou *et al.* 2022). Plant MSH1 is a chimeric fusion of a *mutS* gene with a GIY-YIG endonuclease domain (Abdelnoor *et al.* 2006) that has been proposed to introduce breaks in organellar DNA at the site of mismatches, which would then be repaired through homologous recombination (Christensen 2014, 2018; Ayala-García *et al.* 2018; Broz *et al.* 2022). Assays conducted on purified MSH1 *in vitro* have found that it has DNA binding and endonuclease activity with affinity for displacement loops (D-loops) (Peñafiel-Ayala *et al.* 2023).

We previously found support for a MSH1-mediated link between recombination and point mutations by using a high-fidelity Duplex Sequencing technique (Kennedy *et al.* 2014) to screen for single nucleotide variants (SNVs) and indels in *msh1* mutants (Wu *et al.* 2020). In that study, we also included a panel of mutants lacking functional copies of other genes involved in organellar DNA replication, recombination, and/or repair, including the recombination protein RECA3, the paralogous organellar DNA polymerases POLIA and POLIB, and the glycosylases UNG, FPG, and OGG (Wu *et al.* 2020). Compared to wild type (WT) lines, *msh1* mutants incurred SNVs at a ~10-fold increase in mtDNA and a ~100-fold increase in cpDNA and increases in indel frequencies were even greater. In contrast, *recA3* mutants showed only a small (and marginally significant) increase in mtDNA mutation, and none of the other lines in the mutant panel showed a significant increase in SNVs or indels compared to WT plants (Wu *et al.* 2020).

Here, we investigate additional organellar genome repair proteins (WHY2, RADA, RECA1, OSB2) known to play a role in the suppression of non-allelic recombination in the *A. thaliana* organellar genomes. WHY2 is a mitochondrially targeted whirly protein that binds

single-stranded DNA to inhibit recombination between small repeated sequences via micro-homology mediated end joining (Cappadocia *et al.* 2010) and is also the most abundant protein in mitochondrial nucleoids (as measured in *A. thaliana* cell culture; Fuchs *et al.* 2020). RADA is a dual-targeted DNA helicase, which has been shown to accelerate the processing of recombination intermediates and promote mtDNA stability in *A. thaliana* (Chevigny *et al.* 2022). RECA1 is a plastid-targeted protein that has been proposed to act synergistically with plastid whirly proteins to promote plastid genome integrity either by facilitating polymerase lesion bypass or by reversing stalled replication forks (Rowan *et al.* 2010; Zampini *et al.* 2015). OSB2 is a plastid-targeted single-stranded DNA binding protein that has been shown to hamper microhomology-mediated end joining *in vitro* (García-Medel *et al.* 2021). Given that we previously saw a weak signal of increased mtDNA mutation in *recA3* mutants (Wu *et al.* 2020), we included another *recA3* mutant allele in this study. In addition to these newly generated mutant lines, we also present an extended analysis of Duplex Sequencing data from Wu *et al.* (2020) to understand how SNVs are distributed among genomic regions, strand (template vs. non-template) of genic regions, and trinucleotide contexts. Finally, we also performed long-read Oxford Nanopore sequencing on the mutant lines, allowing us to study structural mutations and rearrangements. Collectively, these analyses provide a detailed characterization of the effects of numerous recombination-related genes on point mutations and structural variants in plant organellar genomes.

METHODS

Generation and analysis of Duplex Sequencing libraries for SNV and indel detection

We obtained seeds for *A. thaliana osb2*, *radA*, *recA1*, *recA3*, and *why2* mutants from the Arabidopsis Biological Resource Center (Table S1). The generation of Duplex Sequencing data from mutants and matched WT controls (including crossing, plant growth, organelle isolation, DNA extraction, and library preparation) closely followed our previously described protocols (Wu *et al.* 2020). For each gene of interest, homozygous mutants were used as the paternal pollinators in crosses against WT maternal plants, which introduced 'clean' organellar genomes (i.e. never exposed to a mutant background) into the resulting heterozygous F1s. The presence of one WT allele in the F1 heterozygotes should be sufficient for WT-like organelle genome maintenance since the mutant alleles of the repair genes of interest are thought to act recessively. The heterozygous F1s were then allowed to self-cross and we identified three homozygous mutant and three homozygous WT F2s, which were also allowed to self-cross. Families of F3 seeds were grown together to obtain sufficient leaf tissue for organelle isolation and mutation detection via Duplex Sequencing.

The only notable differences between the methods in this study compared to Wu et al. 2020 were 1) we only isolated organelles for which the protein of interest is targeted (plastid: *OSB2*, *RADA,* and *RECA1*; mitochondrial: *RADA, RECA3, and WHY2*), whereas in Wu *et al.,* (2020) we isolated both organelles regardless of targeting. 2) Due to the shift to single-organelle isolation for some genes the mitochondrial fraction was discarded for the chloroplast-only extractions and vice versa. 3) We adjusted our Duplex Sequencing library construction protocol to obtain larger inserts by ultrasonicating the DNA for only 60 seconds (three bouts of 20 seconds, with 15 second pauses between each) and size selecting libraries with a 2% gel on a BluePippin (Sage Science), using a specified target range of 400-700 bp. 4) We implemented a new approach to filter spurious variant calls resulting from nuclear insertions of mitochondrial

DNA (NUMTs) by comparing putative mutations directly against the new assembly of the large NUMT on chromosome 2 (Fields *et al.* 2022), replacing the *k*-mer based NUMT filtering approach described in Wu *et al.* (2020).

Generation and analysis of nanopore sequencing libraries for structural variant detection

Nanopore libraries were produced from the same DNA samples that were used for Duplex Sequencing. Sequencing libraries were created following the protocol outlined in the Oxford Nanopore Technologies Rapid Barcoding Kit 96 (SQK-RBK110-96) manual (v110 Mar 24, 2021 revision) and were sequenced on MinION flow cells (FLO-MIN106) under the control of MinKNOW software v22.08.4 or 22.08.9. Multiplexed libraries from cpDNA samples were pooled and run on a single flow cell, whereas pooled mtDNA libraries were run on two flow cells. All runs were conducted for 72 hrs with a minimum read length of 200 bp. Data were processed using the Guppy Basecalling Software v6.3.4+cfaa134.

We sequenced three mutant replicates and one matched WT control for each gene of interest. Mutant lines for the chloroplast extractions included *msh1* (CS3246)*, osb2, recA1,* and *radA* (only two *radA* mutants were sequenced due to a lack of DNA in mutant replicate 2), while mutant lines for the mitochondrial extractions included *msh1* (CS3246)*, recA3, why2,* and *radA*. The total sequencing yield (3.72 Gb) in our initial run of 15 cpDNA samples was an order of magnitude higher than our subsequent run with the 16 mtDNA samples (0.33 Gb). To increase mtDNA coverage we re-sequenced 12 of those mtDNA samples (all but the *msh1* mutants and matched WT control) in a third run, which had a similar low yield (0.42 Gb) to the second run. In all cases, samples were run on fresh flow cells as opposed to flow cells that had been washed for a second run. Because the *msh1* and *radA* mtDNA samples produced very little data (Table S4),

we used the mtDNA contamination in the *msh1* and *radA* cpDNA samples in downstream analyses of the nanopore data.

To calculate mitochondrial and plastid read depth, we aligned the nanopore reads to the organellar genomes with minimap2 (version 2.24; Li 2018) and tabulated depth at each position with bedtools (version 2.30.0; Quinlan and Hall 2010). We calculated the average depth in 1000-bp sliding windows tiling the organellar genomes and plotted depth as a normalized mutant:WT ratio.

The nanopore reads were analyzed with HiFiSr (https://github.com/zouyinstein/hifisr), a software tool developed to identify structural variants using BLASTn alignments of long reads in plant organellar genomes (Zou *et al.* 2022). Because the tool was originally developed for PacBio HiFi reads, which are more accurate than nanopore reads, we required at least two independent nanopore reads to support putative indels. We compared the breakpoints of putative recombination events with the repeats in the *A. thaliana* organellar genomes, which are reported in Tables S10 (mtDNA) and S28 (cpDNA) by Zou *et al.* (2022) and restricted the analyses of overall recombination frequency to repeats that showed a total of at least ten mtDNA recombination reads across all replicates. Because cpDNA recombination events were much less common, we lowered the threshold to a minimum of three recombining reads per repeat for calculating recombinaiton frequencies. The matched WT controls were pooled for these analyses because we only sequenced one WT control for each gene of interest.

RESULTS

Duplex Sequencing coverage

We generated Duplex Sequencing libraries from DNA extracted from isolated organelles to test if genes involved in recombination-suppression also impact accumulation of SNVs and short indels in *A. thaliana* organellar genomes. Duplex Sequencing libraries were sequenced on a NovaSeq 6000 to produce between 30.6 to 139.1 million paired-end reads (2×150 nt) per library (Table S2). Processing the Duplex Sequencing libraries to collapse Illumina reads into consensus sequences and map them to organellar genomes resulted in coverage of 94.2 to 816.3× in the mitochondrial libraries (*radA, recA3,* and *why2*) and 234.2 to 1176.6× in the plastid libraries (*radA, recA1,* and *osb2*; Table S2).

Increased SNV and indel frequency in *radA*, *recA1*, and *recA3* mutants

We compared variant frequencies of each mutant to the matched WT controls (two-tailed *t*-test) and found significant increases in SNV and indel frequencies in the *radA* mutants (p-values reported in Fig. 2.1). We also observed significant indel and weakly significant SNV increases in the *recA3* and *recA1* mutants in the mtDNA and cpDNA, respectively. We analyzed our previously generated *recA3* mutant from Wu *et al.,* (2020), which represents an independent mutant allele of *recA3*, and similarly found significant indel and weakly significant SNV increases in mtDNA (Fig. S2.1). Dinucleotide mutations involve neighboring sites both experiencing a substitution at the same time and are increasingly being recognized as an important type of mutation (Kaplanis *et al.* 2019). We assessed whether these mutations increase in frequency in any of the analyzed mutant backgrounds but found no significant differences relative to WT controls (Wilcoxon signed rank test, $p > 0.05$, Fig. S2.2).

Decreased frequency of CG→TA transitions in the mtDNA of newly generated WT lines

The mutant lines assayed in both this study and in Wu *et al.* (2020) were sequenced with matched WT controls. Surprisingly, pooled WT SNV frequencies generated in the current study were lower than the pooled WT SNV frequencies from the Wu *et al.* (2020) dataset (2.8×10^{-8} vs. 1.7×10^{-7}, *t*-test, $p = 8.85\times10^{-12}$), driven by a decrease in CG→TA transitions (*t*-test, $p = 2.2\times10^{-10}$; Fig. 2.2). To understand if the decreased SNV rate in the newly generated WT libraries (Fig. 2.2) resulted from the changes we made to our library preparation protocol, we created a Duplex Sequencing library following our new protocol using one of the original WT DNA samples from Wu *et al.,* (2020). This new library had SNV rate of 1.97×10^{-7} which is in line with the SNV rates observed in the WT libraries from the 2020 study (Fig 2.2). In fact, the new SNV rate for this DNA sample was actually slightly higher than that of the original library (1.36×10^{-7}). Given that the newly created libraries were all size selected on a BluePippin, which involves mixing the libraries with fluorescein labeled DNA as an internal standard for gauging DNA migration speed, we re-sequenced two stored libraries from Wu *et al.,* (2020) with and without size selection on the BluePippin. The inclusion of the sample without size selection on the BluePippin served as a control for the sample processed on the BluePippin and also as an independent test to understand if changes in the sequencing platform could be responsible (all samples were sequenced on a NovaSeq 6000, but the chemistry of the flow cells has been updated). These re-sequenced libraries had SNV rates typical of the old WT libraries of 1.57×10^{-7} (size selected library) and 1.47×10^{-7} (not size selected). Again, these values were slightly higher than the SNV rates from the original round of sequencing (1.36×10^{-7} and 1.39×10^{-7}, respectively). Therefore, it seems highly unlikely that the decreased SNV rate in the new WT libraries is associated with the changes we made to our library preparation protocol. Instead, these appear to be genuine

differences in the DNA samples, perhaps due to unaccounted for variation in the growth conditions or DNA extraction procedures between the two batches.

SNV frequencies are similar among different genomic regions

To gain a deeper understanding of mutational process in the organellar genomes, we next turned our attention to the distribution of SNVs, focusing primarily on the *msh1* mutants and the pooled WT libraries from the Wu *et al.* (2020) study, given the larger number of mutations in those datasets. First, we assessed if the SNVs in *msh1* mutants and pooled WT libraries from Wu *et al.* (2020) are evenly distributed between intergenic, protein-coding (CDS), intronic, rRNA, and tRNA regions (Fig. 2.3) and found no significant differences among genomic regions (Kruskal-Wallis test, $p > 0.05$, Table S3) except in the WT plastid comparison, which is likely not biologically meaningful, given the small number of observed WT plastid SNVs (Fig. 2.2). Given that the vast majority of mtDNA SNVs in the Wu *et al.* (2020) WT dataset are CG→TA transitions, we separately tested if this class of substitutions is evenly distributed across regions and found significant differences (Kruskal-Wallis test, $p = 0.0295$), driven by a decrease in tRNA genes compared to intergenic sequences (pairwise comparisons with Wilcoxon rank sum test, $p=0.0013$). However, tRNA genes make up a small fraction of the genome and, thus, are subject to higher sampling variance, precluding any confident conclusions about whether they actually accumulate fewer CG→TA transitions than intergenic sequence.

C→T substitutions are more common on the template strand in genic regions

Next, we performed a strand asymmetry analysis to understand if the SNVs in these datasets are evenly distributed on non-template vs. template strands in the CDS, intronic, rRNA,

and tRNA regions of the organellar genomes. The analysis of the CG→TA transitions from the Wu *et al.* (2020) WT dataset revealed that G→A substitutions are significantly enriched on the non-template strand of the DNA (paired Wilcoxon signed-rank test; $p < 0.05$ for CDS, rRNA and tRNA genes). Conversely, C→T substitutions predominately occur on the template strand, which is read by RNA polymerases during transcription (Fig. 2.4). This asymmetry is most striking in rRNA and tRNA genes, where every C→T substitution occurred on the template strand (25 in rRNA and 7 in tRNA). CG→TA transitions were also asymmetrically distributed between strands in genic regions of the Wu *et al.* (2020) *msh1* mutants (Fig. 2.5), though only in certain regions of the mtDNA (Fig. 2.5 top right facet), and not in the cpDNA (Fig. 2.5 bottom right facet). We also investigated strand asymmetries in the AT→GC transitions of the Wu *et al.* (2020) *msh1* mutants and found a trend toward more C→T substitutions on the template strand of plastid genes (Fig. 2.5 left panels). We did not investigate strand asymmetries for the other substitution classes in WT or *msh1* mutants because the small number of data points precludes meaningful comparisons between strands (see Fig. 2.5 of Wu *et al.* 2020).

CG→TA transition frequencies vary depending on trinucleotide context

To understand how surrounding nucleotides impact SNV accumulation in plant organellar genomes, we performed a trinucleotide analysis, again focusing on CG→TA transitions in WT and both transition types in *msh1* mutants, due to a lack of data in other substitution classes. In the WT dataset (Wu *et al.* 2020), we find that CG→TA transitions are 8.4-fold more common in the mtDNA and 3.7-fold more common in the cpDNA when the C is 3′ of a pyrimidine (Fig. 2.6). Interestingly, this same trinucleotide context (5′ pyrimidine) is not enriched for CG→TA transitions in the *msh1* mutant data. Instead CG→TA transitions are 3.0-fold more common when

the C is 5′ of a G in the *msh1* mutants (Fig. 2.7 right facets). Meanwhile AT→GC transitions are 1.8-fold more common when the A is 5′ of a C (Fig. 2.7 left facets). In all cases, these trinucleotide mutation frequencies are normalized by the total coverage of a given trinucleotide context so that the values are not inflated in trinucleotides that are relatively common in the mtDNA.

Chloroplast extractions produced an order of magnitude more nanopore sequencing data than mitochondrial extractions

We next generated long-read Oxford Nanopore libraries to gain a deeper understanding of how the genes in our panel impact plant organellar genome stability. Unexpectedly, the libraries produced from the mitochondrial isolations sequenced poorly compared to the plastid-derived libraries (see methods), so we investigated cross-organelle contamination (mtDNA molecules in the plastid-derived samples and cpDNA molecules in mitochondrially derived samples) to understand if poor mtDNA sequencing performance was inherit to the mtDNA or associated with differences in the organellar isolation methods. The level of mtDNA contamination in the plastid-derived nanopore libraries is similar to the level of contamination in the Duplex Sequencing libraries (Fig. S2.3). The average median read length of the mitochondrial derived nanopore libraries is about 2.5-fold higher than the average median read length of the plastid derive libraries (2.48 kb vs. 1.08 kb, respectively). In the plastid derived nanopore libraries, the median lengths of the contaminating mtDNA reads tend to be slightly longer than the median lengths of native cpDNA reads (average median lengths of 1.17 kb vs 0.98 kb, respectively), though there is substantial variation between samples (Fig. S2.4). In the mitochondrially derived libraries, the

contaminating cpDNA and native mtDNA median read lengths show more correlation (average median lengths of 2.41 kb and 2.56 kb, respectively; Fig. S2.4).

These analyses suggest that the difference in yields for the different nanopore runs is likely related to differences in the organellar isolation methods. One unique feature of the mitochondrial isolation protocol is the use of a DNase I treatment to remove contaminating nuclear and plastid DNA molecules (Wu *et al.* 2020). It is possible that this treatment results in nicking of the mtDNA that interrupts the molecules as they are threaded through the nanopore in a single-stranded fashion. Such nicking would not be expected to disrupt Duplex Sequencing library creation since the first step of making Duplex Sequencing libraries is to break DNA into small fragments via ultrasonication. This explanation is somewhat inconsistent with the 2.5-fold greater median read length in the mitochondrially derive nanopore libraries.

Repeat-mediated recombination drives distinct patterns of mtDNA instability in *msh1*, *radA,* and *recA3 mutants*

To assess whether recombination in any of the mutants perturbs organellar genome copy number and structural stability, we analyzed the ratio of mutant coverage to WT coverage (Fig. 2.8). We see distinct variation patterns in the mtDNA coverage in *msh1, radA* and *recA3* mutants, consistent with the expected structural effects of these genes (Fig. 2.8, left side). In contrast, the *why2* coverage does not deviate from WT coverage, suggesting there is not a substantial and consistent structural effect of losing *why2.* In *recA3*, the nanopore and Duplex Sequencing lines are tightly correlated, while the nanopore data tends to show greater variance in the *msh1, radA*, and *why2* plots, perhaps because of the lower nanopore coverage in those samples (Table S5; Figs. S2.5 and S2.6). Interestingly, *radA* and *recA3* share many major coverage peaks and

valleys, suggesting genome structure is perturbed in similar ways in these mutants (Fig. 2.8, Figs S2.5 and S2.6). Compared to the mitochondrial samples, the cpDNA samples display much less coverage variation (Fig. 2.8. Right side), with a notable exception in the *recA1* nanopore data. However, inspection of the coverage in the individual cpDNA replicates (Fig. S2.7) reveals depth irregularities in the WT control compared to the other WT samples. Regardless, the *recA1* Duplex Sequencing data does not show any depth variation along the cpDNA, so the nanopore result does not appear to reflect a biological effect on cpDNA structure. One other intriguing pattern in the cpDNA plots is an apparent correlation in peaks and valleys in *radA* and *osb2* in the Duplex Sequencing data (most notable is the shared valley at 112 kb). However, inspection of the individual *recA1* mutant and matched WT control replicates (Fig. S2.8) reveals all samples have a dip at 112 kb and the dip is more pronounced in one or more of the *osb2* and *radA* mutants. Given the large number of PCR cycles used to amplify the Duplex Sequencing libraries (19 cycles) the unified movement of all replicates is likely explained in part by amplification bias in AT or GC rich regions. Therefore, variation in amplification bias may result in lower coverage of AT or GC rich regions, so these patterns are likely not biological.

We analyzed the nanopore reads for evidence of repeat mediated recombination. To do so, we calculated recombination frequencies for each repeat pair as the count of nanopore reads that recombined at a given repeat (according the BLASTn alignments generated by HiFiSr (Zou *et al.* 2022)) divided by the total number of reads that mapped to the repeat. We then calculated total recombination frequencies for the mtDNA by summing across repeats with at least 10 recombining reads. The threshold was lowered to repeats with at least three recombining reads in the cpDNA given the smaller number of recombining reads observed in the cpDNA.

We found significant differences in the frequency of mtDNA rearrangements among the WT and mutant lines (one-way ANOVA, $p = 1.5\times10^{-8}$, Fig. 2.9), which were driven by increases in recombination frequency in *msh1, radA* and *recA3* compared to WT (Tukey pairwise comparison, $p = 3\times10^{-7}$, 2×10^{-7}, and 0.02, respectively). In contrast, there was no mtDNA recombination frequency difference between *why2* mutants and WT samples (Tukey pairwise comparison, $p = 0.99$). We found that different repeats apparently become active in different mutant background as evidenced by a two-way ANOVA with a significant interactionn between genotype and repeat ($p < 2\times10^{-16}$). Consistent with previous characterization of repeat mediated recombination in plant mtDNAs (Arrieta-Montiel *et al.* 2009; Davila *et al.* 2011; Miller-Messmer *et al.* 2012; Chevigny *et al.* 2022; Zou *et al.* 2022), we found that repeat length and percent identity are also predictive of recombination frequency through a three-way ANCOVA with repeat length and percent identity as continuous variables ($p = 1.77\times10^{-12}$ and 1.35×10^{-6}, respectively) and genotype as a categorical variable ($p = 2\times10^{-22}$). There were no significant differences in repeat-mediated recombination between any of the cpDNA mutants (*msh1, radA, osb2, recA1*) compared to the WT samples (one-way ANOVA, $p = 0.849$; Fig. 2.9). We identified no insertions in the HiFiSr variant calls (after requiring at least two nanopore reads to support a putative insertion) and only a single cpDNA deletion of 106 bp in *msh1* mutant replicate 2, which was supported by 18 independent nanopore reads (cpDNA position 148490-148596).

DISCUSSION

Potential causes of elevated organellar mutation rates in lines with disrupted recombination machinery

By utilizing highly accurate Duplex Sequencing for point mutation detection and long-read Oxford Nanopore sequencing for structural variant detection, we have characterized the overall organellar mutational dynamics in *A. thaliana* lines lacking genes with roles in organellar genome recombination. The increases in point mutations we observed in *radA, recA3,* and *recA1* are much smaller than the effects previously observed in *msh1* mutants (Wu *et al.* 2020) where mutants experience 6.0-fold and 116.5-fold increases in SNVs (in mtDNA and cpDNA, respectively) and 86.6-fold and 790.6-fold increases in indels (in mtDNA and cpDNA, respectively). In contrast, *radA* mutants incurred 2.6-fold and 12.6-fold more mtDNA and cpDNA SNVs (respectively) and 5.1-fold and 3.1-fold more mtDNA and cpDNA indels (respectively) than the matched WT controls. The point mutation increases in *recA3* and *recA1* were even smaller than in the *radA* mutants. One complication with directly comparing the mutant vs. WT fold changes across the newly generated genes compared to those generated in Wu *et. al.,* (2020) is the decrease in WT mutation rates in the new genes (Fig. 2.2). Because of the shift in the baseline WT rates, the numbers cited above may actually underestimate the gap in effect size between *msh1* and the newly analyzed genes.

The point mutation increases in *msh1* mutants have clear mechanistic explanations which were first predicted based on the MSH1 mismatch recognition and GIY-YIG endonuclease domains (Christensen 2014; Wu *et al.* 2020). In contrast, given that *RADA, RECA3* and *RECA1* are all thought to function in the resolution of recombination intermediates, it is unclear how to explain the mechanisms responsible for increased point mutations in these lines. One possibility is that in the absence of one recombination pathway, recombining molecules are shuttled into an alternative, less faithful recombination pathway. In mutant lines deficient in homologous recombination (HR), double-stranded breaks (DSBs) may be repaired via error prone non-

homologous or microhomology mediated end joining (NHEJ or MMEJ, respectively), which could drive increases in indels and SNVs (Waters *et al.* 2014; García-Medel *et al.* 2019). Evidence suggests that *RADA* functions as the principal branch migration factor in a *RECA2*-dependent homologous recombination (HR) pathway, while *RECA3* promotes a partially redundant and less utilized alternative HR pathway (Chevigny *et al.* 2022). The larger SNV and indel increases in the *radA* mutants than in the *recA3* mutants may reflect the relative utilization (and importance) of these two partially redundant HR pathways (Chevigny *et al.* 2022). Similarly, previous studies have documented increased NHEJ and MMEJ in cpDNA of *recA1* mutants, which is consistent with the significant increase in indels and marginally significant increase in SNVs reported here (Fig. 2.1).

Another possibility is that the rise in point mutations is an indirect effect of increased repeat-mediated recombination and its associated harm to organelle function. Increased recombination between short repeat sequences may disrupt genes, organellar genome stoichiometry, and genome organellar replication, which is recombination-dependent in plants (Shedge *et al.* 2007; Rowan *et al.* 2010; Chevigny *et al.* 2020). Plant organellar genomes encode proteins necessary for the electron transport chains of respiration and photosynthesis and disruption of these pathways can result in the excess production of DNA damaging reactive oxygen species (ROS; Liu *et al.* 2021). Although a direct link between ROS-mediated damage to DNA and mutation rates remains contentious (Kennedy *et al.* 2013; Itsara *et al.* 2014; Broz *et al.* 2021; Waneka *et al.* 2021; Sanchez-Contreras *et al.* 2021), ROS molecules have been shown to indirectly affect point mutation rates by impairing proofreading capabilities via damage to the metazoan mtDNA polymerase (Pol γ; Anderson *et al.* 2020). Impairment if organellar function is

also consistent with phenotypic growth defects in *radA*, which included retarded development and distorted leaves with chlorotic sectors (Chevigny *et al.* 2022).

Potential explanations of mutational biases based on DNA strand asymmetry and flanking nucleotides

We found that SNVs in the *msh1* mutants and WT plants from Wu *et al.,* (2020) had biased distributions in terms of strand (non-template vs. template) and trinucleotide context. Such patterns are useful for understanding the underlying mechanisms driving mutation formation (Haradhvala *et al.* 2016; Sun *et al.* 2018; Moeckel *et al.* 2023). For example, CG→TA strand asymmetries documented in diverse metazoan mtDNAs have been proposed to result from the two DNA strands experiencing unequal time in single-stranded states during mtDNA replication, since single-stranded DNA is more vulnerable to cytosine deamination (a primary driver of CG→TA transitions) (Kennedy *et al.* 2013; Itsara *et al.* 2014; Arbeithuber *et al.* 2020; Waneka *et al.* 2021; Sanchez-Contreras *et al.* 2021). In mammals, C→T substitutions are ~10-fold more common than G→A substitution on the mtDNA heavy strand (H-strand), which likely spends more time in a single-stranded state as the mtDNA is copied via a strand-asynchronous replication mechanism (Kennedy *et al.* 2013; Arbeithuber *et al.* 2020). Further, the C→T substitutions form two gradients starting at the two H-strand origins of replication, consistent with the regions closest to the origin being single stranded for longer (Sanchez-Contreras *et al.* 2021).

The substantial CG→TA strand asymmetries we observed in the mtDNA of the Wu *et al.,* (2020) WT libraries are unlikely to be explained by replication mechanisms given that plants mtDNAs lack discrete origins of replication or dedicated ‘leading and lagging’ strands

(alternatively referred to as light and heavy strands, respectively, in some systems) and instead rely on recombination-mediated replication (Gualberto and Newton 2017; Brieba 2019; Chevigny *et al.* 2020). Instead, our strand asymmetry analysis focused on genic regions, motivated by well-established patterns of more C→T than G→A substitution on non-template strands which spend more time in exposed single-stranded during transcription (Haradhvala *et al.* 2016; Vöhringer *et al.* 2021; Moeckel *et al.* 2023). Surprisingly, we found an opposite pattern with template strands exhibiting far more C→T than G→A substitutions (Fig. 2.4). This effect was especially pronounced in rRNA and tRNA genes where the C→T substitutions occurred on the template strand in all 32 observed CG→TA transitions. An enrichment of C→T substitutions on template strands also occurred in the mtDNA (but not the cpDNA) of the *msh1* mutants, though there was less power for detecting statistically significant effects (Fig. 2.5). The overabundance of A→G compared to T→C substitutions in *msh1* mutant cpDNA template strands also occurs in the opposite direction of predicted effects given that the non-template strand is again expected to experience increased adenine deamination (which proceeds A→G substitutions; Mugal *et al.* 2009; Sanchez-Contreras *et al.* 2021).

Enrichment of C→T and A→G substitutions on template strands is puzzling, and to our knowledge there are no other instances where this widespread transcriptional asymmetry has been reversed (Mugal *et al.* 2009; Moeckel *et al.* 2023). Reversals in strand asymmetries have been reported in metazoan mitochondrial genomes, but in these cases the asymmetries are replication based, and the reversals are proceeded by an inversion of the origin of replication, effectively switching the leading and lagging strands (Wei *et al.* 2010). It is notable that the WT CG→TA asymmetries are most pronounced in the rRNA and tRNA genes (Fig. 2.4), which are likely more highly expressed than the protein coding genes. Increases in transcription have been

shown to drive genomic instability in the *A. thaliana* cpDNA due to the increased formation of R-loops (RNA/DNA hybrids formed by displacement of the other DNA strand) which stall replication forks and lead to DSBs (Pérez Di Giorgio *et al.* 2019). It is possible that increased mtDNA expression also leads to the formation of R-loops and DSBs which may then be repaired through error prone NHEJ and MMEJ. However, it is not clear how this would drive strand asymmetric mutation. Further, such a mechanism is not consistent with the relatively even distribution of SNVs across intergenic vs. transcribed regions of the genome (Fig. 2.3). The magnitude of the CG→TA asymmetries is decreased in the *msh1* mutants (roughly 2-fold averaging across all genic sequences) compared to in the WT controls (roughly 6-fold). Without a clear mechanism to explain the WT asymmetries it is difficult to interpret why the *msh1* asymmetries decreased as the overall mutation SNV frequency increased.

The CG→TA transitions in the WT lines and both transitions in the *msh1* mutants were also impacted by the identity of neighboring nucleotides (Figs. 2.6 and 2.7). Trinucleotide effects have previously been implicated to bias mutation distribution in the *A. thaliana* nuclear genome (Lu *et al.* 2021) as well as in the mtDNAs of various metazoans (Itsara *et al.* 2014; Arbeithuber *et al.* 2020; Waneka *et al.* 2021; Sanchez-Contreras *et al.* 2021). It is noteworthy that the specific trinucleotides associated with CG→TA transitions differ between WT and *msh1* mutants. The 5′ YCN signature (where Y is any pyrimidine and N is any nucleotide) in the WT lines is similar to that induced by APOBEC3-mediated cytosine deamination in human cell lines (Carpenter *et al.* 2023), though plants lack APOBEC enzymes so the relevance of this shared pattern in unclear. Meanwhile, the 5′ NCG signature in the *msh1* mutants is consistent with spontaneous water mediated cytosine deamination (Carpenter *et al.* 2023).

Patterns of repeat-mediated recombination differs among mutant lines

The repeat mediated mtDNA recombination activity we documented in the *msh1*, *radA* and *recA3* mutants is consistent with the previously documented recombination increases of these mutant backgrounds (Shedge *et al.* 2007; Arrieta-Montiel *et al.* 2009; Rowan *et al.* 2010; Davila *et al.* 2011; Miller-Messmer *et al.* 2012; Zampini *et al.* 2015; Wu *et al.* 2020; Chevigny *et al.* 2022; Zou *et al.* 2022). Likewise, the absence of an effect in the *why2* mutants is consistent with a previous study which showed *why2* mutants become more recombinationally active than WT under increased genotoxic stress (ciprofloxacin treatment) but showed no recombinational difference from WT under 'normal' growth conditions (Cappadocia *et al.* 2010).

Though *msh1*, *radA* and *reca3* are all required for the suppression of repeat-mediated recombination in mtDNA, these proteins likely function in independent HR pathways or in different ways. As, noted, the HR pathway coordinated by RECA3 is thought to relatively minor compared to the one coordinated by RECA2 (Chevigny *et al.* 2020, 2022), which is a highly abundant protein in the mitochondria (Fuchs *et al.* 2020) and necessary for plant survival (*recA2* mutants die at the seedling stage; Miller-Messmer *et al.* 2012). Previous studies of *recA3/msh1* and *recA3/radA* double mutants have shown the double mutants are more recombinationally active than *recA3* single mutants (Shedge *et al.* 2007), supporting the hypothesis that *RECA3*-mediated HR is in part independent of the more utilized *RECA2*-dependent pathway in which *radA* mediates branch migration (Miller-Messmer *et al.* 2012; Chevigny *et al.* 2022). Meanwhile, MSH1 has been proposed to suppress non-allelic recombination by recognizing and rejecting mismatches in the invading strand during heteroduplex formation (Christensen 2018; Broz *et al.* 2022)

Given that recombination is activated differently between the mutants (Fig. 2.8), the high degree of repeatability between replicates is fascinating (Figs. S2.5, S2.6, S2.7, S2.8). These repeatable patterns rely on consistent activation of distinct repeat pairs and/or consistent maintenance/replication of certain recombination product. Understanding why different repeats become active and how these patterns relate to the increase in point mutations reported here remains an important unanswered question in the field of plant organellar genome maintenance.

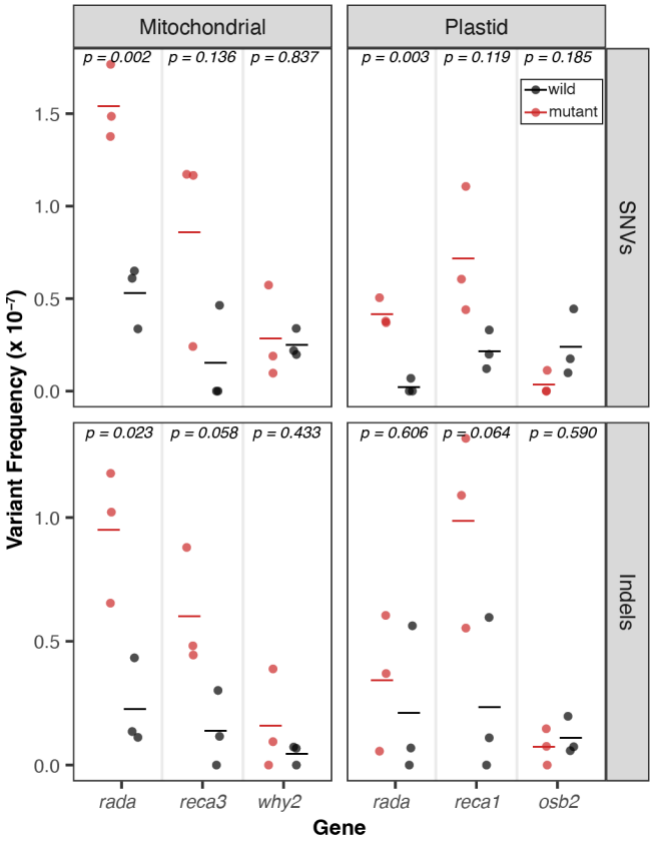

Figure 2.1. *De novo* point mutations measured with Duplex Sequencing. For each gene of interest (x-axis), mutant lines are plotted in red, and matched WT controls are plotted in black. The individual biological replicates are plotted as circles and group averages are plotted as dashes. Facets separate the data by genome; left column: Mitochondria and right column: Plastid, and by point mutation type; top row: SNVs and bottom row: indels. Variant frequencies (y-axis) were calculated as the total number of SNVs/total Duplex Sequencing coverage. P-values show the result of a two-tailed *t*-test comparing WT vs mutant mutation frequencies for each gene of interest.

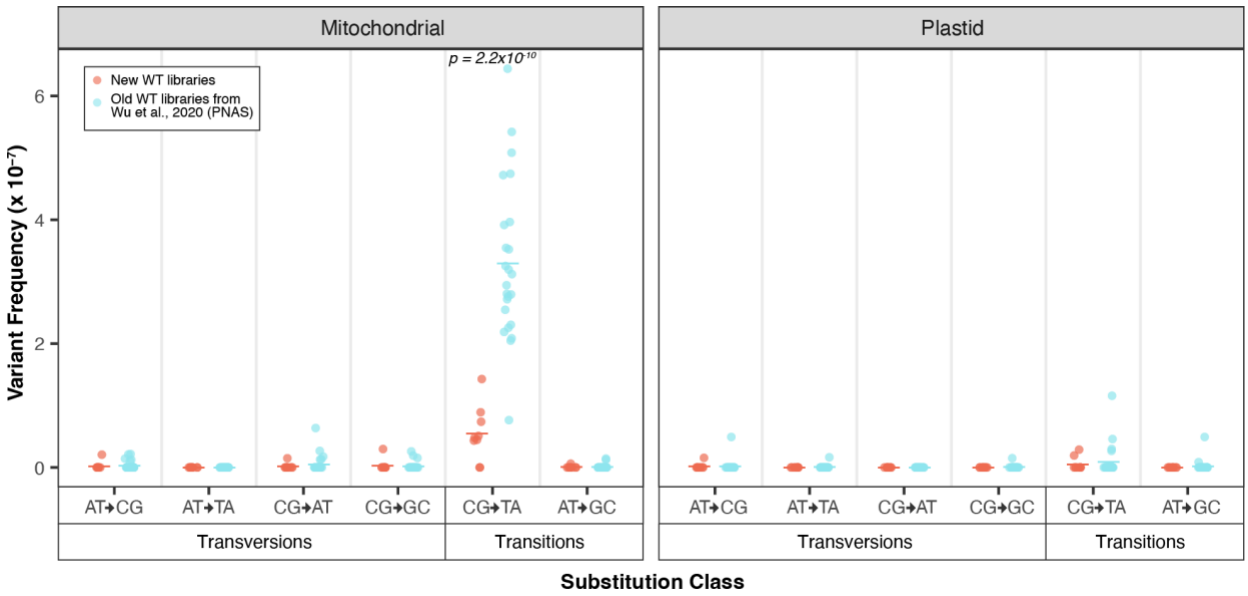

Figure 2.2. Comparison of the mutational spectrum of pooled WT controls from the current study (orange) vs. the WT controls from Wu *et al.* 2020 (blue). The two facets show the mitochondrial and plastid data and the x-axis separates substitutions type by transversions vs. transitions and further by the six types of substitutions. Individual biological replicates are plotted as circles while group averages are plotted as dashes. Only CG>TA transitions showed a significant increase in the old data set (two tailed t-test; $p=2.2\times10^{10}$).

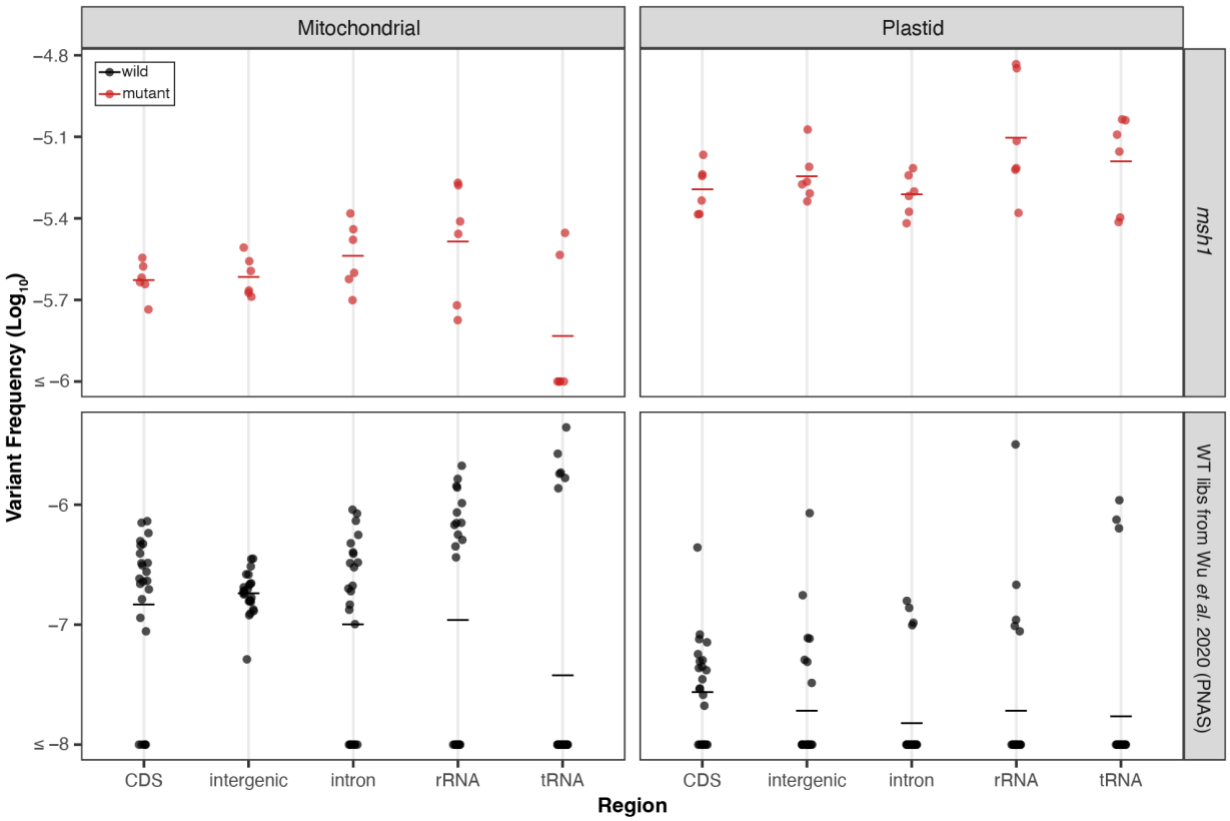

Figure 2.3. Distribution of WT (black) and *msh1* (red) SNVs (from Wu et al., 2020) across genomic region. The individual biological replicates are plotted as circles and group averages are plotted as dashes. Facets separate the data by genome; left column: Mitochondria and right column: Plastid, and by genotype with *msh1* mutants on top and WT on the bottom. Note the difference in y-axis scale for *msh1* mutants and WT. For each of the four facets, we performed a Kruskal-Wallis test and found no significant difference between genomic regions except the WT plastid facet ($p = 0.022$) where comparisons between regions are likely not biologically meaningful given the low number of WT plastid mutations. Note that for this and subsequent analyses of the *msh1* Duplex Sequencing data, we pooled the two null *msh1* alleles to increase statistical power.

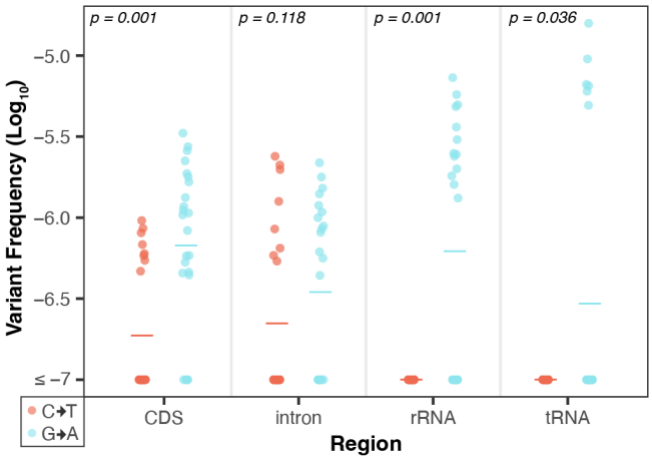

Figure 2.4. Strand asymmetry analysis of CG>TA transitions in the WT mtDNA Duplex Sequencing data from Wu *et al.* (2020). Shown are the log-transformed SNV frequencies (y-axis) of C>T (red) vs. G>A (blue) mutations on the non-template strand of all genes, separated by genomic region (x-axis). The individual biological replicates are plotted as circles and group averages are plotted as dashes. P-values show the result of paired Wilcoxon tests comparing the complementary substitution classes in each genomic region. In all but intronic regions, G>A substitutions are significantly higher on the non-template than strand (conversely, C>T substitutions are significantly higher on the template strand). Strikingly, in all of the observed CG>TA transitions in the rRNA and tRNA genes the C>T substitution occurred on the template strand.

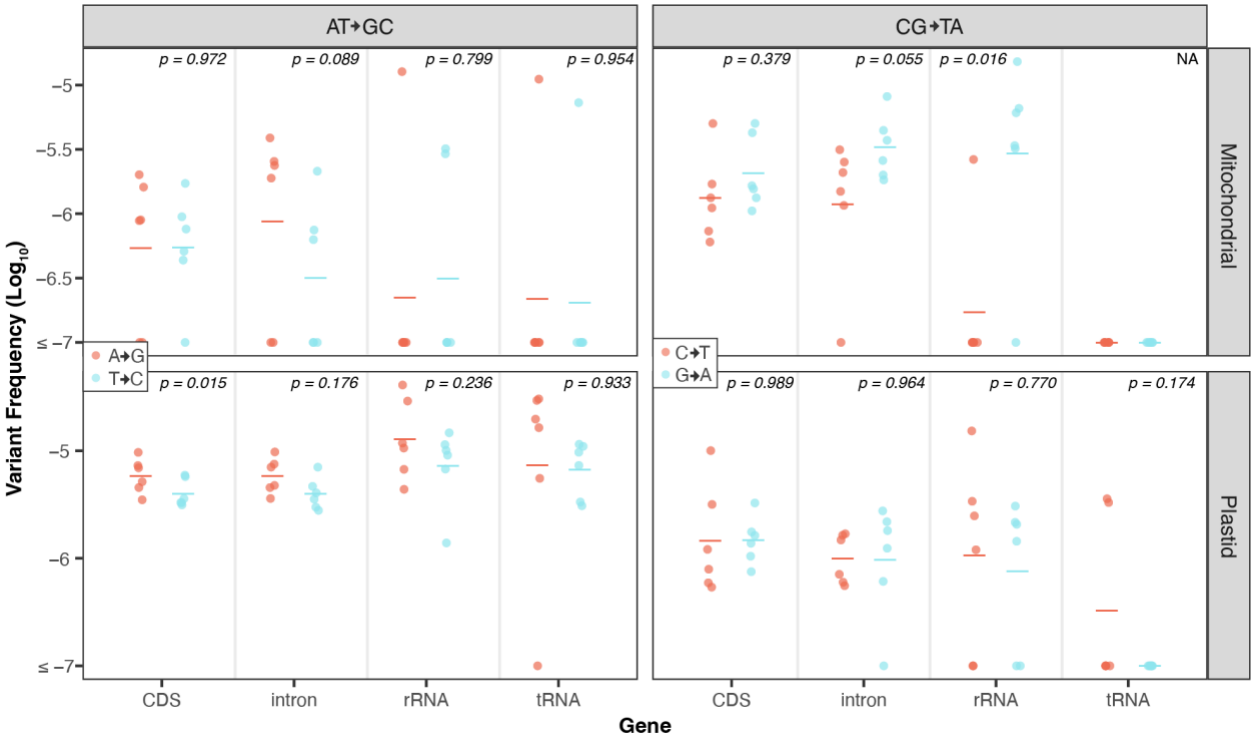

Figure 2.5. Strand asymmetry analysis of CG>TA and AT>GC transitions in the *msh1* Duplex Sequencing data from Wu *et al.* (2020). Shown are the log-transformed SNV frequencies (y-axis) of mutations on the non-template strands of all genes with complementary substitution types designated by color (see figure legends for colors of specific substitution types). The individual biological replicates are plotted as circles, and group averages are plotted as dashes. The facets divide the data by transition type, with AT>GC transitions on the left and CG>AT transitions shown on the right, and by genome, with mitochondrial data on the top and plastid data on the bottom. Transversions were not analyzed because there were relatively few observed mutations of this type in the *msh1* duplex data. P-values show the result of paired *t*-tests comparing the complementary substitution classes in each genomic region.

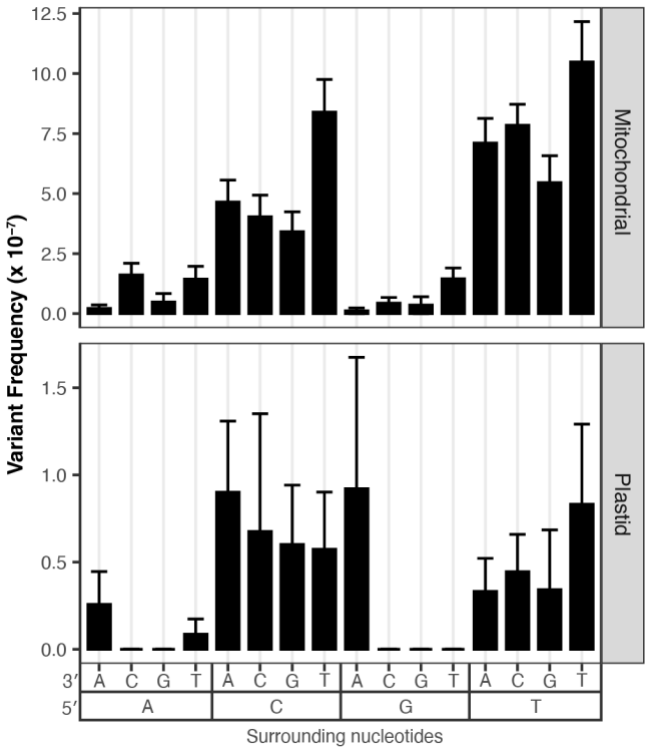

Figure 2.6. Analysis of surrounding nucleotides on C>T transition frequencies in the WT Duplex Sequencing data from Wu *et al.* (2020). The facets divide the data based on genome with mitochondrial data on the top and plastid data on the bottom, note the difference in the y-axis scale, as CG>TA were less frequent in the plastid. The x-axis captures the trinucleotide context with downstream nucleotides displayed next to the 3′ and upstream nucleotides display next to the 5′. Complementary trinucleotide contexts (corresponding to G>A substitutions) are not shown but are accounted for and collapsed to simplify the plot in terms of C>T substitutions. The data suggest that trinucleotide contexts with upstream pyrimidines (5′ CCN 3′ and 5′ TCN 3′, where N is any nucleotide) have increased frequencies of C>T substitutions.

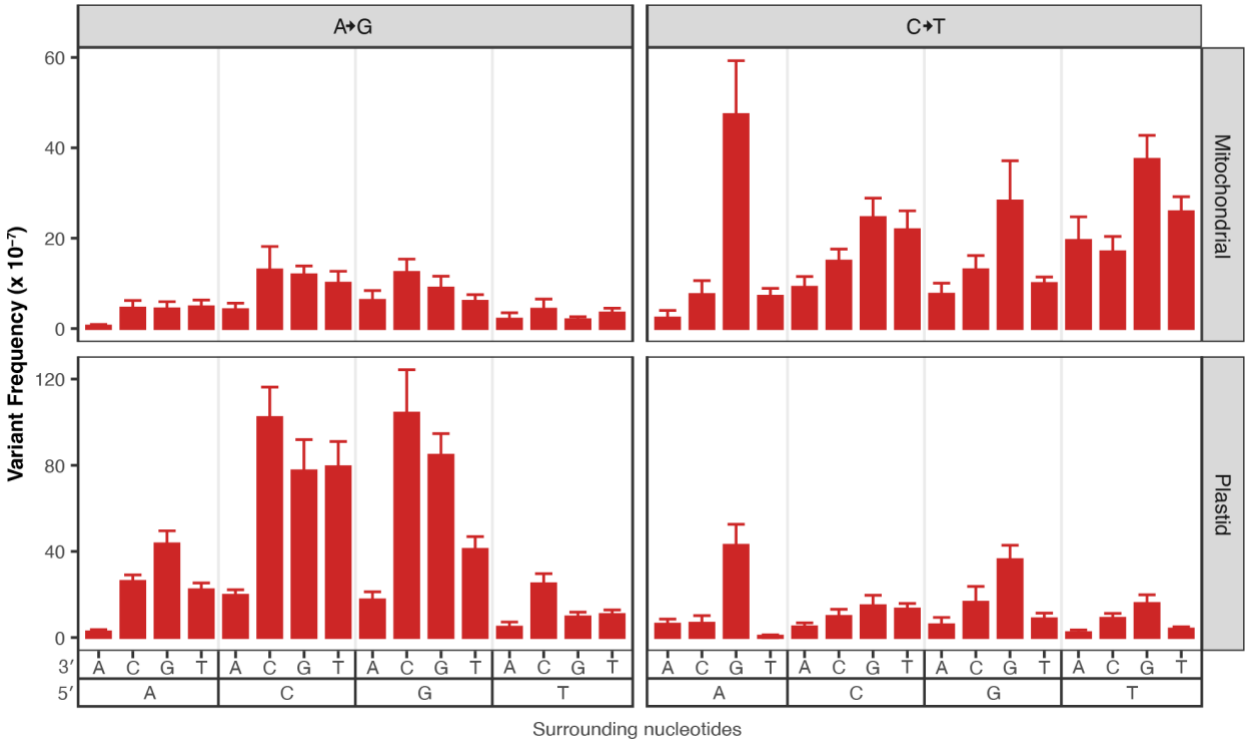

Figure 2.7. Analysis of surrounding nucleotides on A>G and C>T transition frequencies in the *msh1* Duplex Sequencing data from Wu *et al.* (2020). The facets divide the data based on substitution type A>G substitutions on the left and C>T substitutions on the right and by with genome with mitochondrial data on the top and plastid data on the bottom. The x-axis captures the trinucleotide context with downstream nucleotides displayed next to the 3′ and upstream nucleotides display next to the 5′. Complementary substitutions and trinucleotide contexts are collapsed to simplify the plot in terms of A>G or C>T substitutions. The A>G data suggest that trinucleotide contexts with downstream Cs (5′ NAC 3′) have increased frequencies of A>G substitutions. The C>T data suggest that trinucleotide contexts with downstream Gs (5′ NCG 3′) have increased frequencies of C>T substitutions.

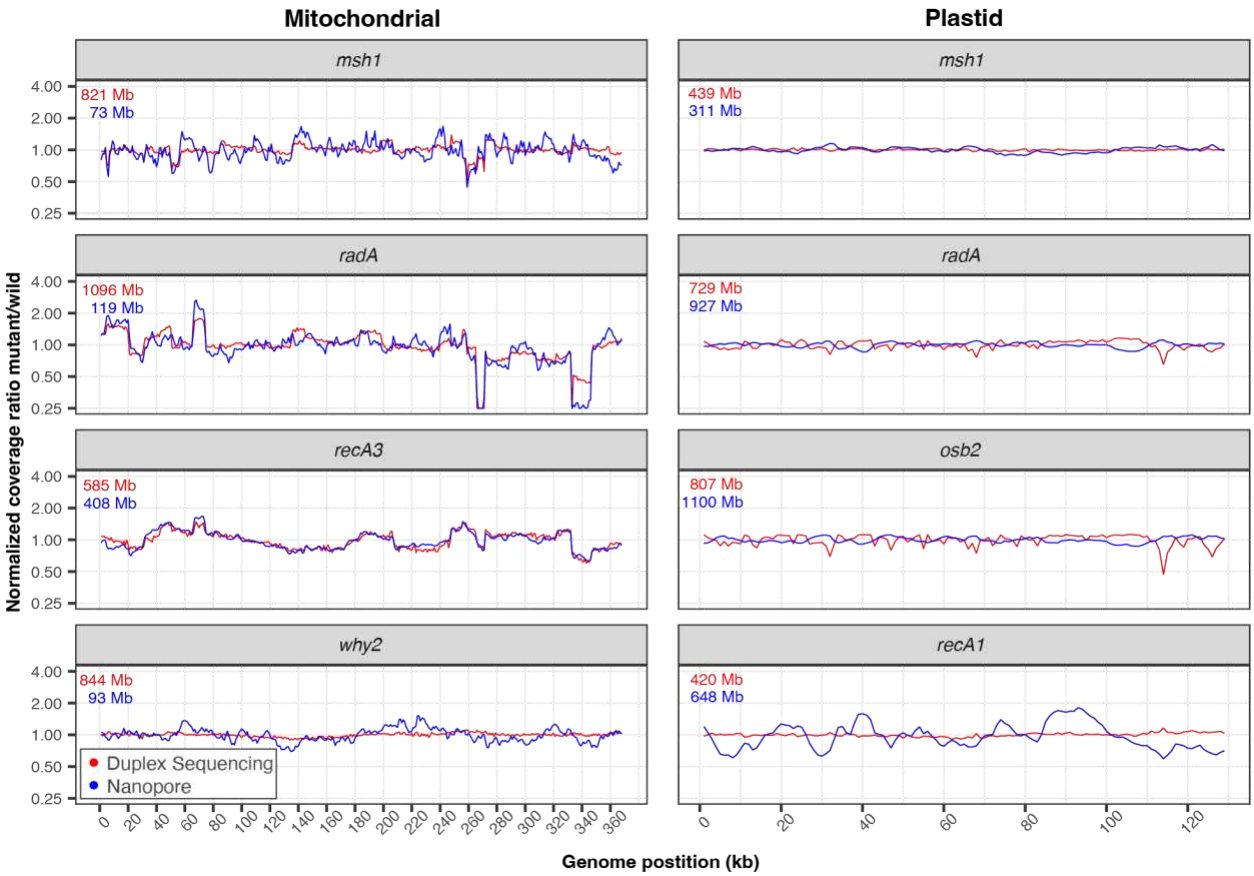

Figure 2.8. Normalized coverage of mitochondrial (left) and plastid (right) genomes. Coverage of each Duplex Sequencing (red) or nanopore (blue) library was calculated in 1000-bp windows. Mutant coverage was pooled and divided by WT coverage and the resulting ratios were normalized to 1 for plotting. The total amount of sequencing data used to generate each plot is shown in the top left corner of each facet (red=Duplex Sequencing and blue=nanopore) and is included to highlight the instances where disagreement between the Duplex Sequencing and nanopore lines may be explained by increased variance in the nanopore sample due to lower mtDNA coverage.

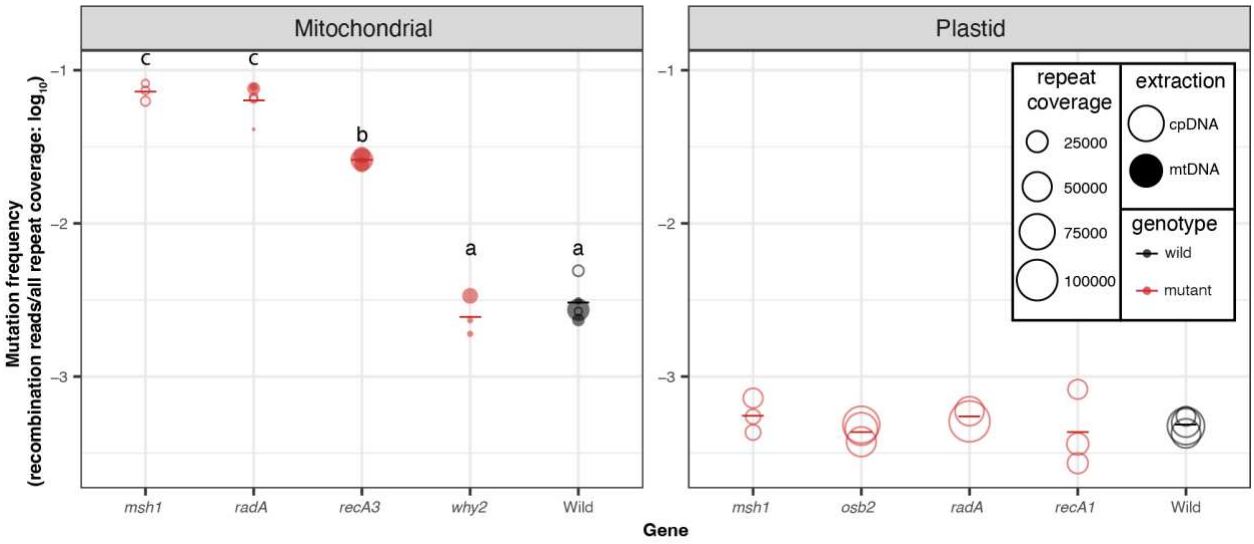

Figure 2.9. Frequency of repeat-mediated structural variants in the nanopore data. The individual biological replicates are plotted as circles with the size of the circle scaled by the number of repeats that are covered in the nanopore alignments. Closed circles are the libraries from mitochondrial extractions, while the open circles are libraries from the plastid extractions. In some cases, cpDNA extractions were used to harvest contaminating mtDNA-mapping reads because of low yield from direct sequencing of the mtDNA extractions. Group averages are plotted as dashes. Mutants are plotted in red, while WT samples are plotted in black. Letters represent statistically significant groupings according to Tukey pairwise comparisons on a one-way ANOVA ($p<0.001$). There were no differences among plastid genotypes.

LITERATURE CITED

Abdelnoor R. V., R. Yule, A. Elo, A. C. Christensen, G. Meyer-Gauen, *et al.*, 2003 Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. Proc. Natl. Acad. Sci. U. S. A. 100: 5968–5973.

Abdelnoor R. V., A. C. Christensen, S. Mohammed, B. Munoz-Castillo, H. Moriyama, *et al.*, 2006 Mitochondrial genome dynamics in plants and animals: convergent gene fusions of a MutS homologue. J. Mol. Evol. 63: 165–173.

Alverson A. J., S. Zhuo, D. W. Rice, D. B. Sloan, and J. D. Palmer, 2011 The mitochondrial genome of the legume Vigna radiata and the analysis of recombination across short mitochondrial repeats. PLoS One 6: e16404.

Anderson A. P., X. Luo, W. Russell, and Y. W. Yin, 2020 Oxidative damage diminishes mitochondrial DNA polymerase replication fidelity. Nucleic Acids Res. 48: 817–829.

Arbeithuber B., J. Hester, M. A. Cremona, N. Stoler, A. Zaidi, *et al.*, 2020 Age-related accumulation of de novo mitochondrial mutations in mammalian oocytes and somatic tissues. PLoS Biol. 18: e3000745.

Arrieta-Montiel M. P., V. Shedge, J. Davila, A. C. Christensen, and S. A. Mackenzie, 2009 Diversity of the Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity. Genetics 183: 1261–1268.

Ayala-García V. M., N. Baruch-Torres, P. L. García-Medel, and L. G. Brieba, 2018 Plant organellar DNA polymerases paralogs exhibit dissimilar nucleotide incorporation fidelity. FEBS J. 285: 4005–4018.

Boore J. L., 1999 Animal mitochondrial genomes. Nucleic Acids Res. 27: 1767–1780.

Brieba L. G., 2019 Structure-Function Analysis Reveals the Singularity of Plant Mitochondrial DNA Replication Components: A Mosaic and Redundant System. Plants 8. https://doi.org/10.3390/plants8120533

Broz A. K., G. Waneka, Z. Wu, M. Fernandes Gyorfy, and D. B. Sloan, 2021 Detecting de novo mitochondrial mutations in angiosperms with highly divergent evolutionary rates. Genetics 218. https://doi.org/10.1093/genetics/iyab039

Broz A. K., A. Keene, M. F. Gyorfy, M. Hodous, I. G. Johnston, *et al.*, 2022 Sorting of mitochondrial and plastid heteroplasmy in *Arabidopsis* is extremely rapid and depends on MSH1 activity. Proceedings of the National Academy of Sciences 119: e2206973119.

Cappadocia L., A. Maréchal, J.-S. Parent, E. Lepage, J. Sygusch, *et al.*, 2010 Crystal structures of DNA-Whirly complexes and their role in Arabidopsis organelle genome repair. Plant Cell 22: 1849–1867.

Carpenter M. A., N. A. Temiz, M. A. Ibrahim, M. C. Jarvis, M. R. Brown, *et al.*, 2023 Mutational impact of APOBEC3A and APOBEC3B in a human cell line and comparisons to breast cancer. PLoS Genet. 19: e1011043.

Chen C., Q. Li, R. Fu, J. Wang, C. Xiong, *et al.*, 2019 Characterization of the mitochondrial genome of the pathogenic fungus Scytalidium auriculariicola (Leotiomycetes) and insights into its phylogenetics. Sci. Rep. 9: 17447.

Chevigny N., D. Schatz-Daas, F. Lotfi, and J. M. Gualberto, 2020 DNA Repair and the Stability of the Plant Mitochondrial Genome. Int. J. Mol. Sci. 21. https://doi.org/10.3390/ijms21010328

Chevigny N., F. Weber-Lotfi, A. Le Blevenec, C. Nadiras, A. Fertet, *et al.*, 2022 RADA-dependent branch migration has a predominant role in plant mitochondria and its defect leads to mtDNA instability and cell cycle arrest. PLoS Genet. 18: e1010202.

Christensen A. C., 2014 Genes and Junk in Plant Mitochondria—Repair Mechanisms and Selection. Genome Biol. Evol. 6: 1448–1453.

Christensen A. C., 2018 Mitochondrial DNA Repair and Genome Evolution. Annual Plant Reviews online 11–32.

Davila J. I., M. P. Arrieta-Montiel, Y. Wamboldt, J. Cao, J. Hagmann, *et al.*, 2011 Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. BMC Biol. 9: 64.

Drouin G., H. Daoud, and J. Xia, 2008 Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol. Phylogenet. Evol. 49: 827–831.

Fields P. D., G. Waneka, M. Naish, M. C. Schatz, I. R. Henderson, *et al.*, 2022 Complete Sequence of a 641-kb Insertion of Mitochondrial DNA in the Arabidopsis thaliana Nuclear Genome. Genome Biol. Evol. 14. https://doi.org/10.1093/gbe/evac059

Fuchs P., N. Rugen, C. Carrie, M. Elsässer, I. Finkemeier, *et al.*, 2020 Single organelle function and organization as estimated from Arabidopsis mitochondrial proteomics. Plant J. 101: 420–441.

García-Medel P. L., N. Baruch-Torres, A. Peralta-Castro, C. H. Trasviña-Arenas, A. Torres-Larios, *et al.*, 2019 Plant organellar DNA polymerases repair double-stranded breaks by microhomology-mediated end-joining. Nucleic Acids Res. 47: 3028–3044.

García-Medel P. L., A. Peralta-Castro, N. Baruch-Torres, A. Fuentes-Pascacio, J. A. Pedroza-García, *et al.*, 2021 Arabidopsis thaliana PrimPol is a primase and lesion bypass DNA polymerase with the biochemical characteristics to cope with DNA damage in the nucleus, mitochondria, and chloroplast. Sci. Rep. 11: 20582.

Gualberto J. M., and K. J. Newton, 2017 Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation. Annu. Rev. Plant Biol. 68: 225–252.

Handa H., 2003 The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (Brassica napus L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana. Nucleic Acids Res. 31: 5907–5916.

Haradhvala N. J., P. Polak, P. Stojanov, K. R. Covington, E. Shinbrot, *et al.*, 2016 Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell 164: 538–549.

Huang C. Y., N. Grünheit, N. Ahmadinejad, J. N. Timmis, and W. Martin, 2005 Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol. 138: 1723–1733.

Itsara L. S., S. R. Kennedy, E. J. Fox, S. Yu, J. J. Hewitt, *et al.*, 2014 Oxidative stress is not a major contributor to somatic mitochondrial DNA mutations. PLoS Genet. 10: e1003974.

Kaplanis J., N. Akawi, G. Gallone, J. F. McRae, E. Prigmore, *et al.*, 2019 Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations. Genome Res. 29: 1047–1056.

Kennedy S. R., J. J. Salk, M. W. Schmitt, and L. A. Loeb, 2013 Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet. 9: e1003794.

Kennedy S. R., M. W. Schmitt, E. J. Fox, B. F. Kohrn, J. J. Salk, *et al.*, 2014 Detecting ultralow-frequency mutations by Duplex Sequencing. Nat. Protoc. 9: 2586–2606.

Kubo T., and K. J. Newton, 2008 Angiosperm mitochondrial genomes and mutations. Mitochondrion 8: 5–14.

Li H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.

Liu Y., B. Zhou, A. Khan, J. Zheng, F. U. Dawar, *et al.*, 2021 Reactive Oxygen Species Accumulation Strongly Allied with Genetic Male Sterility Convertible to Cytoplasmic Male Sterility in Kenaf. Int. J. Mol. Sci. 22. https://doi.org/10.3390/ijms22031107

Lu Z., J. Cui, L. Wang, N. Teng, S. Zhang, *et al.*, 2021 Genome-wide DNA mutations in Arabidopsis plants after multigenerational exposure to high temperatures. Genome Biol. 22: 160.

Martínez-Zapater J. M., P. Gil, J. Capel, and C. R. Somerville, 1992 Mutations at the Arabidopsis CHM locus promote rearrangements of the mitochondrial genome. Plant Cell 4: 889–899.

Miller-Messmer M., K. Kühn, M. Bichara, M. Le Ret, P. Imbault, *et al.*, 2012 RecA-dependent DNA repair results in increased heteroplasmy of the Arabidopsis mitochondrial genome. Plant Physiol. 159: 211–226.

Moeckel C., A. Zaravinos, and I. Georgakopoulos-Soares, 2023 Strand asymmetries across genomic processes. Comput. Struct. Biotechnol. J. 21: 2036–2047.

Mower J. P., D. B. Sloan, and A. J. Alverson, 2012 Plant Mitochondrial Genome Diversity: The Genomics Revolution, pp. 123–144 in *Plant Genome Diversity Volume 1: Plant Genomes, their Residents, and their Evolutionary Dynamics*, edited by Wendel J. F., Greilhuber J., Dolsezel J., Leitch I. J. Springer Vienna, Vienna.

Mugal C. F., H.-H. von Grünberg, and M. Peifer, 2009 Transcription-induced mutational strand bias and its effect on substitution rates in human genes. Mol. Biol. Evol. 26: 131–142.

Palmer J. D., and L. A. Herbon, 1988 Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J. Mol. Evol. 28: 87–97.

Peñafiel-Ayala A., A. Peralta-Castro, J. Mora-Garduño, P. García-Medel, A. G. Zambrano-Pereira, *et al.*, 2023 Plant organellar MSH1 is a displacement loop specific endonuclease. Plant Cell Physiol. https://doi.org/10.1093/pcp/pcad112

Pérez Di Giorgio J. A., É. Lepage, S. Tremblay-Belzile, S. Truche, A. Loubert-Hudon, *et al.*, 2019 Transcription is a major driving force for plastid genome instability in Arabidopsis. PLoS One 14: e0214552.

Pucker B., D. Holtgräwe, K. B. Stadermann, K. Frey, B. Huettel, *et al.*, 2019 A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. PLoS One 14: e0216233.

Quinlan A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Rowan B. A., D. J. Oldenburg, and A. J. Bendich, 2010 RecA maintains the integrity of chloroplast DNA molecules in Arabidopsis. J. Exp. Bot. 61: 2575–2588.

Sanchez-Contreras M., M. T. Sweetwyne, B. F. Kohrn, K. A. Tsantilas, M. J. Hipp, *et al.*, 2021 A replication-linked mutational gradient drives somatic mutation accumulation and influences germline polymorphisms and genome composition in mitochondrial DNA. Nucleic Acids Res. 49: 11103–11118.

Sandor S., Y. Zhang, and J. Xu, 2018 Fungal mitochondrial genomes and genetic polymorphisms. Appl. Microbiol. Biotechnol. 102: 9433–9448.

Shedge V., M. Arrieta-Montiel, A. C. Christensen, and S. A. Mackenzie, 2007 Plant mitochondrial recombination surveillance requires unusual RecA and MutS homologs. Plant Cell 19: 1251–1264.

Skippington E., T. J. Barkman, D. W. Rice, and J. D. Palmer, 2017 Comparative mitogenomics indicates respiratory competence in parasitic Viscum despite loss of complex I and extreme sequence divergence, and reveals horizontal gene transfer and remarkable variation in genome size. BMC Plant Biol. 17: 1–12.

Sloan D. B., A. J. Alverson, J. P. Chuckalovcak, M. Wu, D. E. McCauley, *et al.*, 2012 Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 10: e1001241.

Smith D. R., and P. J. Keeling, 2015 Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proc. Natl. Acad. Sci. U. S. A. 112: 10177–10184.

Stupar R. M., J. W. Lilly, C. D. Town, Z. Cheng, S. Kaul, *et al.*, 2001 Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proc. Natl. Acad. Sci. U. S. A. 98: 5099–5103.

Sun S., Q. Li, L. Kong, and H. Yu, 2018 Multiple reversals of strand asymmetry in molluscs mitochondrial genomes, and consequences for phylogenetic inferences. Mol. Phylogenet. Evol. 118: 222–231.

Vöhringer H., A. Van Hoeck, E. Cuppen, and M. Gerstung, 2021 Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. Nat. Commun. 12: 3628.

Waneka G., J. M. Svendsen, J. C. Havird, and D. B. Sloan, 2021 Mitochondrial mutations in Caenorhabditis elegans show signatures of oxidative damage and an AT-bias. Genetics 219. https://doi.org/10.1093/genetics/iyab116

Waters C. A., N. T. Strande, J. M. Pryor, C. N. Strom, P. Mieczkowski, *et al.*, 2014 The fidelity of the ligation step determines how ends are resolved during nonhomologous end joining. Nat. Commun. 5: 4286.

Wei S.-J., M. Shi, X.-X. Chen, M. J. Sharkey, C. van Achterberg, *et al.*, 2010 New views on strand asymmetry in insect mitochondrial genomes. PLoS One 5: e12708.

Wolfe K. H., W. H. Li, and P. M. Sharp, 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. U. S. A. 84: 9054–9058.

Wu Z., G. Waneka, A. K. Broz, C. R. King, and D. B. Sloan, 2020 MSH1 is required for maintenance of the low mutation rates in plant mitochondrial and plastid genomes. Proc. Natl. Acad. Sci. U. S. A. 117: 16448–16455.

Wu Z.-Q., X.-Z. Liao, X.-N. Zhang, L. R. Tembrock, and A. Broz, 2022 Genomic architectural variation of plant mitochondria—A review of multichromosomal structuring. J. Syst. Evol. 60: 160–168.

Wynn E. L., and A. C. Christensen, 2019 Repeats of Unusual Size in Plant Mitochondrial Genomes: Identification, Incidence and Evolution. G3 9: 549–559.

Xiao-Ming Z., W. Junrui, F. Li, L. Sha, P. Hongbo, *et al.*, 2017 Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. Sci. Rep. 7: 1555.

Zampini É., É. Lepage, S. Tremblay-Belzile, S. Truche, and N. Brisson, 2015 Organelle DNA rearrangement mapping reveals U-turn-like inversions as a major source of genomic instability in Arabidopsis and humans. Genome Res. 25: 645–654.

Zou Y., W. Zhu, D. B. Sloan, and Z. Wu, 2022 Long-read sequencing characterizes mitochondrial and plastid genome variants in Arabidopsis msh1 mutants. Plant J. 112: 738–755.

CHAPTER 3: UV DAMAGE INDUCES PRODUCTION OF MITOCHONDRIAL DNA FRAGMENTS WITH SPECIFIC LENGTH PROFILES

Summary

UV light is a potent mutagen that induces bulky DNA damage in the form of cyclobutane pyrimidine dimers (CPDs). Photodamage and other bulky lesions occurring in nuclear genomes can be repaired through nucleotide excision repair (NER), where incisions on both sides of a damaged site precede the removal of a single-stranded oligonucleotide containing the damage. Mitochondrial genomes (mtDNAs) are also susceptible to damage from UV light, but current evidence suggests that the only way to eliminate bulky mtDNA damage is through mtDNA degradation. Damage-containing oligonucleotides excised during NER can be captured with anti-damage antibodies and sequenced (XR-seq) to produce high resolution maps of active repair locations following UV exposure. We analyzed previously published datasets from *Arabidopsis thaliana, Saccharomyces cerevisiae*, and *Drosophila melanogaster* to identify reads originating from the mtDNA (and plastid genome in *A. thaliana*). In *A. thaliana* and *S. cerevisiae*, the mtDNA-mapping reads have unique length distributions compared to the nuclear-mapping reads. The dominant fragment size was 26 nt in *S. cerevisiae* and 28 nt in *A. thaliana* with distinct secondary peaks occurring in regular intervals. These reads also show a nonrandom distribution of di-pyrimidines (the substrate for CPD formation) with TT enrichment at positions 7-8 of the reads. Therefore, UV damage to mtDNA appears to result in production of DNA fragments of characteristic lengths and positions relative to the damaged location. The mechanisms producing these fragments are unclear, but we hypothesize that they result from a previously uncharacterized DNA degradation pathway or repair mechanism in mitochondria.

INTRODUCTION

Mitochondria are vital organelles involved in energy production and cellular metabolism. Due to the endosymbiotic origins of mitochondria, they retain their own genomes that are replicated, repaired, and inherited independently of nuclear DNA (nucDNA). Mitochondrial genome (mtDNA) mutation rates show over a 4000-fold variation across eukaryotes (Wolfe *et al.* 1987; Drouin *et al.* 2008; Smith *et al.* 2012; Havird and Sloan 2016), which likely reflects a wide range of mtDNA replication and repair mechanisms. However, significant gaps in our understanding of mtDNA repair mechanisms still remain (Rong *et al.* 2021).

The existence of multiple mtDNA copies within a cell (St. John 2016) led to the hypothesis that DNA repair mechanisms might not be necessary because damaged mtDNA could be degraded without undergoing repair and undamaged mtDNA could act as a template for mtDNA synthesis (Clayton *et al.* 1974; Druzhyna *et al.* 2008). This idea was bolstered by the observation in metazoans that mtDNA mutation rates are much higher than nucDNA mutation rates (Wolfe *et al.* 1987) and the fact that mitochondria are an abundant source of DNA damaging reactive oxygen species (Harman 1972; Murphy 2009). In subsequent decades, however, researchers have determined that mtDNA repair is an important component of mtDNA maintenance and have begun to work out the mechanisms of various mtDNA repair pathways (Saki and Prakash 2017).

With only one known exception (Muthye and Lavrov 2021), mtDNA repair enzymes are encoded in the nucDNA, translated in the cytosol, and targeted to the mitochondria (Mower *et al.* 2012; Zardoya 2020). In some cases, mtDNA repair pathways are highly similar to nucDNA repair pathways, often utilizing enzymatic machinery that is dual-targeted to the nucleus and the mitochondria (Kazak *et al.* 2012). For example, chemically modified mtDNA and nucDNA bases

are both removed through base excision repair (BER), which is perhaps the most ubiquitous and best studied mtDNA repair pathway (Szczesny *et al.* 2008). In contrast, mtDNAs appear to lack canonical mismatch repair (MMR), the principal pathway for correcting mismatches that arise through erroneous base incorporation during DNA replication in nucDNA (Modrich 2006). Instead, various novel/non-canonical mismatch repair pathways may fill this role, with a piecemeal, taxon-specific distribution. For example, the Y-box binding protein YB-1 has been shown to play a role in mismatch elimination in human cell lines, primarily through mismatch recognition and binding (de Souza-Pinto *et al.* 2009). Meanwhile, plants appear to utilize a non-canonical mismatch repair pathway reliant on homologous recombination, facilitated by *MSH1* (Wu *et al.* 2020b).

Nucleotide excision repair (NER) is the major nucDNA repair pathway for bulky DNA damage, a broad class of lesions that occur on one strand of DNA and are characterized by the covalent attachment of large chemical moieties or compounds (Wood 1999). Diverse types of bulky lesions can result from the binding of various chemicals, metabolites, or environmental agents to DNA, leading to structural distortions and functional impairment. NER pathways have evolved independently in bacteria and eukaryotes, with distinct variations in the protein components and regulatory mechanisms. However, both systems follow the same general mechanism in which single-stranded incisions are made both upstream and downstream of a damaged site, followed by the removal of a damage-containing oligonucleotide ranging from ~10-13 (bacterial NER) or ~23-30 (eukaryotic nuclear NER) nt in length. A polymerase fills the resulting gap using the opposite strand as a template, and ligation completes the NER process (Sancar 1996). As is the case for MMR, mtDNAs are thought to lack a conventional NER pathway. Because there are no known alternative pathways for repair of bulky DNA damage in

mtDNAs, it is generally assumed that it leads to mtDNA degradation (Clayton *et al.* 1974; Prakash 1975; Kazak *et al.* 2012; Saki and Prakash 2017), but open questions remain regarding the molecular components of mtDNA degradation, how such degradation would be coordinated, and how new mtDNA molecules could be recovered (Sakamoto and Takami 2018; Zhao 2019; Zhao *et al.* 2023).

Degradation of damaged mtDNAs has been documented in metazoan and yeast cells in response to a variety of DNA damaging agents including UV (Bess *et al.* 2012, 2013), acrolein (Wang *et al.* 2017), gamma irradiation (Dan *et al.* 2020), H_2O_2 (Shokolenko *et al.* 2016), and enzymatically induced double-stranded breaks (Moretton *et al.* 2017). The timelines of mtDNA degradation exhibit considerable variation depending on the organism, cell type, and DNA damaging agents involved; however, it typically proceeds slowly (taking as long as 72 hours in some cases; Bess *et al.* 2012, 2013). MtDNA degradation is frequently associated with mitochondrial fission and mitochondrial-specific autophagy, known as mitophagy (Wang *et al.* 2017; Dan *et al.* 2020). Mitophagy increases during genotoxic stress, but it also occurs in unperturbed cells as part of normal mitochondrial turnover and cellular energetics (Urbina-Varela *et al.* 2020), and defects in mitophagy are associated with multiple human diseases (Springer and Macleod 2016; Doblado *et al.* 2021).

UV light is a potent mutagen capable of causing multiple bulky lesions, predominately in the form of cyclobutane pyrimidine dimers (CPDs; ~80% occurrence) but also as pyrimidine–pyrimidone (6–4) photoproducts ((6–4)-PPs; ~20% occurrence) (Emmerich *et al.* 2020). In addition to repair through NER, some organisms possess photolyases for the direct chemical reversal of photodamage. Photolyases are damage-specific, meaning a CPD photolyase can only repair CPDs and (6-4)PP photolyases can only repair (6-4)PPs. All photolyases use blue light as

an energy source, and they tend to have a spotty distribution across the tree of life. Roughly half of bacteria, a quarter of archaea, most plants and fungi, and most vertebrates possess CPD photolyases; (6-4)PP photolyases are generally not as common (Goosen and Moolenaar 2008; Lucas-Lledó and Lynch 2009; Mei and Dvornyk 2015). Photolyases have also been shown to repair photodamage in mtDNAs of some plants (Takahashi *et al.* 2011) and some fungi (Prakash 1975; Yasui *et al.* 1992). For other groups, such as mammals, there is no known mechanism for the repair of photodamage in mtDNA.

A handful of studies aimed at detecting NER in mtDNA have yielded negative results (Clayton *et al.* 1974; Waters and Moustacchi 1974; Prakash 1975; Ledoux *et al.* 1992; Hunter *et al.* 2010; Takahashi *et al.* 2011). The earliest experiments leveraged the CPD nicking T4 endonuclease V to measure the amount of CPDs in mtDNAs of UV exposed cells. Irradiated mammalian cells given time for dark repair (NER) or light repair (photolyase) showed the same amount of mtDNA CPDs as irradiated cells given no time for repair, suggesting there is a complete lack of photodamage repair in mammalian mtDNA (Clayton *et al.* 1974). Similar studies found that the yeast *Saccharomyces cerevisiae* also lacks dark repair of CPDs in mtDNA but does exhibit light repair (Waters and Moustacchi 1974; Prakash 1975), and subsequent work established that a dual-targeted CPD photolyase protects both nuclear and mitochondrial DNA in *S. cerevisiae* (Yasui *et al.* 1992). Tests for NER in mtDNA using qPCR in rice (Takahashi *et al.* 2011) and zebrafish (Hunter *et al.* 2010) found no reduction in the number of polymerase blocking lesions after irradiated organisms were given periods of dark repair. qPCR studies with mice cells did detect a decrease in frequency of polymerase blocking lesions in mtDNA after long periods of repair (8 to 24 hours), but this was attributed to the repair of non-pyrimidine dimer polymerase blocking lesions, which can also be induced through UV irradiation

(Kalinowski *et al.* 1992). It therefore remains unclear if and how eukaryotes repair pyrimidine dimers in mtDNA. While photolyases may fill this role for some eukaryotes, they are missing entirely in some groups (mammals) or are only partially represented, such as in *S. cerevisiae*, which lacks a photolyase for the repair of the (6-4)PPs (Sancar 2004).

In recent years, a series of DNA sequencing techniques leveraging antibodies that specifically recognize CPDs or (6-4)PPs have been developed to characterize pyrimidine dimer formation and repair on genome-wide scales (Hu *et al.* 2015, 2017; Mao *et al.* 2016; Alhegaili *et al.* 2019). One technique called DDIP-seq uses anti-damage antibodies to capture and sequence damage-containing molecules from samples of sonicated DNA (~100 to 300 bp) (Amente *et al.* 2019). A DDIP-seq study with human HaCaT cells (keratinocyte cell line) and anti-CPD antibodies showed that CPD damage occurs at a high rate in mtDNA immediately following UV exposure. Surprisingly, after 24 hours allowing for repair, as much as 50% of the mtDNA damage had disappeared (Alhegaili *et al.* 2019), contrasting with previous reports documenting no CPD repair in mammalian mtDNA (Clayton *et al.* 1974; Ledoux *et al.* 1992). Anti-damage antibodies can also be used to detect excision oligos directly in excision assays, where damage containing oligos are captured with anti-damage antibodies, 3′ radiolabeled and visualized on high-density polyacrylamide gels (Hu *et al.* 2013).

Another technique called XR-sequencing (XR-seq) has been particularly useful for understanding repair dynamics (Hu *et al.* 2015). XR-seq uses anti-damage antibodies to capture the oligonucleotides that are excised during NER (Fig. 3.1). These oligonucleotides are then subject to adaptor ligation, treated with photolyases, and sequenced on Illumina platforms. Sequenced reads can be aligned to reference genomes, yielding maps of active repair locations following UV exposure at single-nucleotide resolution. The technique achieves an extremely

high sensitivity through the combined action of multiple filtering steps built into the library preparation (Hu *et al.* 2019). First, the antibodies have a high specificity for their damage targets (Mori *et al.* 1991), as evidenced by control immunoprecipitations with unirradiated cells, which yield no detectable DNA on polyacrylamide gels (Oztas *et al.* 2018). The anti-CPD and anti-(6-4)PP antibodies may bind damage in both ssDNA and dsDNA (Mori *et al.* 1991). However, dsDNAs containing CPDs should not receive adaptors, which anneal to ssDNA through overhanging, random 5-nt sequences. XR--seq experiments have been performed with cells or tissue samples from *Homo sapiens* (Hu *et al.* 2015), *Mus musculus* (Yang *et al.* 2018), *Microcebus murinus* (Akkose *et al.* 2020), *Drosophila melanogaster* (Deger *et al.* 2019), *S. cerevisiae* (Li *et al.* 2018), and *Arabidopsis thaliana* (Oztas *et al.* 2018). The mtDNA-mapping reads from these datasets remain largely unexplored.

It is possible that previous attempts to detect NER in mtDNA may have failed because of a relatively weak signal of mtDNA repair compared to dominant signal of NER in nucDNA (Zhao and Sumberaz 2020). We reasoned that the high sensitivity of XR-seq would provide increased power for detecting previously uncharacterized repair activity in mtDNA. If there is no repair pathway for excision of photodamage or other bulky DNA lesions in mtDNA (as is generally thought) and instead such lesions lead to mtDNA degradation and turnover, the XR-seq data can still provide valuable insights into fate of photodamage during degradation and whether degradation is ordered or localized to certain regions of the genome. Published mammalian XR-seq datasets are unsuitable for such mtDNA analysis because they include an initial immunoprecipitation against TFIIH, a nuclear-localized protein complex that associates with excised oligonucleotides in mammalian NER (Fig. 3.1) (Lindsey-Boltz *et al.* 2023). Therefore, in this study, we analyzed the mtDNA-mapping reads from published *S. cerevisiae*, *A. thaliana* and

D. melanogaster datasets, in which the extracted small DNA molecules were immediately immunoprecipitated with anti-damage antibodies (anti-CPD or anti-(6-4)PP) without an initial TFIIH immunoprecipitation (Fig. 3.1).

METHODS

XR-seq datasets

The XR-seq datasets from *S. cerevisiae, A. thaliana* and *D. melanogaster* were generated in previous experiments (Li *et al.* 2018; Oztas *et al.* 2018; Deger *et al.* 2019; respectively). The methods used to generate those data sets are briefly summarized here. In the *A. thaliana* experiment, plants were irradiated with 120 J/m^2 UVC at eight different times (spaced 3 hours apart) throughout a 24 hour day-night cycle and given 30 minutes of 'dark repair' time (Oztas *et al.* 2018). In the *S. cerevisiae* experiment, cells were grown to late log phase and then irradiated with 120 J/m^2 UVC and given either 5, 20 or 60 minutes of 'dark repair' time (Li *et al.* 2018). In the *D. melanogaster* experiment, S2-DGRC cells were grown to 25-80 % confluence and then irradiated with 20 J/m^2 UVC and given either 0.16, 0.5, 8, 16 or 24 hours of 'dark repair' time (Deger *et al.* 2019). In all three experiments, two biological replicates were included for each timepoint. The library preparation protocols were similar in all experiments, though there were differences in the methods of DNA extraction. Specifically, for *S. cerevisiae* and *D. melanogaster,* cells were disrupted through bead beating and the excised DNA was enriched by Hirt lysis, where salt is used to precipitate away the chromatin fraction of the cell lysate, and through G-50 column filtration, which further depletes the chromatin fraction (Li *et al.* 2018; Deger *et al.* 2019; Hu *et al.* 2019). For *A. thaliana*, whole leaves were frozen in liquid nitrogen and ground into a powder before they were vortexed with glass beads (Oztas *et al.* 2018). In all

three preparations, DNA was extracted through ethanol precipitation and damage-containing products were immunoprecipitated with anti-CPD or anti-(6-4)PP antibodies. Adaptors were ligated onto the excised oligomers before a second immunoprecipitation was performed to further enrich damage-containing molecules (Li *et al.* 2018; Oztas *et al.* 2018; Deger *et al.* 2019; Hu *et al.* 2019). In all three preparations, the adaptor-ligated products were then treated with photolyases (either CPD- or (6-4)PP-specific, depending on the library) before the samples were amplified and sequenced using 50-nt single-read Illumina chemistry.

Alignment

Raw XR-seq reads were downloaded from NCBI BioProjects (*A. thaliana*: PRJNA429185, *D. melanogaster*: PRJNA577587, *S. cerevisiae*: PRJNA434118) via the SRA Toolkit fastq-dump command (ver 2.8.0; Andrews 2010). Adaptor sequences (reported in original publications: Li *et al.* 2018; Oztas *et al.* 2018; Deger *et al.* 2019) were removed with cutadapt (version 1.18; Martin 2011), using the discard untrimmed reads option. Reads were aligned to reference genomes (*A. thaliana*: TAIR10, *D. melanogaster*: dm6_UCSC, *S. cerevisiae*: sacCer3), which included the organellar genomes (*A. thaliana* mtDNA: NC_037304.1, *A. thaliana* plastid DNA (ptDNA): NC_000932.1, *D. melanogaster* mtDNA: NC_024511.2, *S. cerevisiae* mtDNA: NC_001224.1), using bowtie2 (ver 2.3.5; Langmead and Salzberg 2012) with the –phred33 flag (Oztas *et al.* 2018).

Alignment filtering and XR-seq analysis

Nuclear insertions of mtDNA or ptDNA (termed NUMTs and NUPTs, respectively) warrant special consideration in this analysis because repair of organelle-derived nuclear DNA

through conventional NER could result in the false mapping of XR-seq reads to organelle genomes. To ensure that reads mapping to the organelle genomes truly originated from the organelle genomes, we used samtools (ver 1.9; Li *et al.* 2009) to discard reads with MAPQ scores of less than 30, effectively removing all reads which map equally well to multiple locations. As a result of this filtering step, NUMTs/NUPTs that are correctly assembled in the nuclear reference (and any homologous sequences present in the assemblies) are 'unmappable' to either copy (organellar or nuclear). The *A. thaliana* ptDNA contains a large, inverted repeat (~26 kb). Since both copies of the repeat would be 'unmappable' after filtering out reads with MAPQ scores of less than 30, we removed the second copy of the repeat (positions 128214-154478) from the reference genome and divided all read counts in the first copy of the repeat by two when calculating coverage statistics. A 641-kb NUMT on chromosome 2 of the *A. thaliana* reference genome contains more than an entire copy of the mitochondrial genome (Fields *et al.* 2022), which introduces a potential bias as only the identical portions of the NUMT and the mitochondrial genome will be 'unmappable' using a MAPQ cutoff of 30. We therefore used a modified reference where the NUMT (positions 3239038-3509765 of chromosome 2) was manually removed. While interpreting the *A. thaliana* dataset, it is therefore important to remember that some mtDNA mapping reads may be nuclear-derived. After MAPQ filtering, we used custom scripts to remove reads with mismatches (all scripts used in this study are available via https://github.com/dbsloan/mtDNA_UV_damage).

We used custom scripts to calculate the read length distributions, nucleotide frequencies and di-pyrimidine frequencies of the mtDNA mapping reads and compared them to equivalent analyses from the nuclear mapping reads, which were previously reported (Li *et al.* 2018; Oztas *et al.* 2018; Deger *et al.* 2019). We analyzed the differences in read coverage (reads per kilobase

per million mapped reads; RPKM) between organellar and nuclear genomes and between different genomic regions (i.e. intergenic, intronic, protein coding (CDS), rRNA genes, and tRNA genes) of the organellar genomes. In genic regions, we compared the XR-seq read coverage of the template vs the coding strand.

Excision assay

To study mtDNA-derived DNA fragments with a method independent of the XR-seq data, we performed an excision assay with *S. cerevisiae* cells exposed to UV light. To isolate mtDNA-derived DNA fragments, we produced a NER-deficient line, which in theory should be unable to produce nucDNA-derived excision oligonucleotides. Specifically, we created a deletion of the *RAD14* gene, which encodes a subunit of nucleotide excision repair factor 1 (NEF1) complex that binds to damaged DNA during NER (Guzder *et al.* 2006). Deletions were generated through homologous recombination-mediated integration of the *NatMX4* nourseothricin resistance cassette (Goldstein and McCusker 1999) in strain FY86 (*MATa, ura3-52, leu2D1, his3D200*; Winston *et al.* 1995), which is isogenic with the S288c reference genome background. We amplified *NatMX4* from pAG25 using primers JAO2397 and JAO2398 (reported in Table S1) to generate a PCR product flanked by 42-bp homologous regions (upper case in primer sequences), targeting integration to each side of the *RAD14* open reading frame. We screened transformants and confirmed the presence of the *rad14D::NatMX4* deletion in two independently generated clones, using PCR with primers flanking both sides of the insertion site (primers JAO2399 and JAO2401, reported in Table S1).

Yeast growth, UV exposure, DNA extraction, immunoprecipitation with an anti-CPD antibody, radiolabeling and DNA visualization all followed previously described protocols (Hu *et*

al. 2019), with these exceptions; 1) UV exposure was performed in a CL-1000 UV crosslinker, which was placed on a shake plate rotating at 120 rpm to ensure even UV administration, 2) we radiolabeled the 3′ ends of the putative damaging containing DNA fragments with GTP [α-^{32}P] (Boulé *et al.* 2001) instead of ^{32}P-Cordycepin due to changes in product availability, and 3) we added 5% glycerol to the 11% acrylamide gel mix and electrophoresis running buffering solutions in an attempt to reduce gel shattering while drying at 80 °C (Altschuler *et al.* 2013). Following UV exposure, all work was conducted in the dark or under yellow light to avoid the activation of photolyases. We included wild type (WT) and *rad14D* replicates that were not exposed to UV as controls, and UV-exposed strains were given 20 minutes of repair time in YPD at 30 °C. For each of the four treatments (WT vs mutant with or without UV exposure), we included two technical replicates for a total of eight samples.

RESULTS AND DISCUSSION

Pre-processing of existing XR-seq datasets

We analyzed the mtDNA mapping reads from *S. cerevisiae, A. thaliana* and *D. melanogaster* XR-seq datasets to gain insights into what happens to photodamaged mtDNA. In the *A. thaliana* dataset, we also investigated ptDNA-derived reads. Due to the short length of excised oligonucleotides in NER, nuclear-derived XR-seq sequences may map incorrectly to organellar genomes during alignment. To ensure such mapping artifacts are not interpreted as organellar derived DNA fragments, we filtered our alignments to retain only uniquely mapping reads with no mismatches. We assessed the impact of this filtering step by comparing XR-seq coverage of the filtered and unfiltered alignment files and found that filtering renders 5 to 13% of organellar

genomes ‘unmappable’. The fraction of each genome retained for downstream analyses, broken down by genomic region, can be found in Table 1.

XR-seq coverage of organellar vs. nuclear genomes

We next compared the depth of XR-seq coverage (computed as reads per kilobase of mapped genome; RPKM) of the organellar and nuclear genomes (Table 2). In the *S. cerevisiae* and *A. thaliana* datasets, organellar XR-seq coverage was roughly one-third to two-fold that of the nuclear genome, while in the *D. melanogaster* data coverage of the mtDNA was over 50-fold that of the nuclear genome. Note that these estimates should not be directly interpreted as measures of the relative rates of degradation or repair in nuclear vs. organellar DNA because they do not adjust for differences in organellar genome copy per nuclear genome, a parameter known to be highly variable under different life stages (Shen *et al.* 2019), tissue and cell types (Herbers et al. 2019; O’Hara et al. 2019), and physiological conditions (Göke *et al.* 2020). The relative rates of pyrimidine dimer formation in organellar vs. nuclear DNA will also impact rates of repair, and estimates of the relative damage rates vary among species (Kalinowski *et al.* 1992; Ledoux *et al.* 1992; Takahashi *et al.* 2011) and depend on methods of detection (Gonzalez-Hunt *et al.* 2018).

Unique length distributions of organellar-mapping reads

We next analyzed the length of XR-seq fragments mapping to organellar and nuclear genomes. For all datasets, the organellar-mapping reads contain unique length distributions compared to those mapping to the nuclear genomes. As reported in the initial publication (Li *et al.* 2018), there are two peaks in the *S. cerevisiae* nucDNA mapping reads (in both anti-CPD and anti-(6-

4)PP datasets), one derived from the primary excision products (23 nt) and the other (~16 nt) presumably resulting from the 5′ degradation of the primary excision products (Figs. 3.2, S3.1, S3.6). The *S. cerevisiae* CPD and 6-4(PP) mtDNA-mapping reads show distinct peaks at read lengths of 26, 24, 22 and 20 nt (Figs. 3.2, S3.2, S3.7). The largest mtDNA peak of 26 nt is longer than the peak length in the nuclear mapping reads of 23 nt. The *A. thaliana* mtDNA read length distributions also differ from the nucDNA read length distributions (Figs. 3.3, S3.11). In the *A. thaliana* mtDNA read length distribution there is a cluster of reads 36-39 nt in length, with additional distinct peaks in read lengths of 32, 28, 24, 20 and 16 nt (Figs. 3.3, S3.12). Therefore, the patterns in these datasets were similar, but the peaks were spaced at different intervals (2-nt in *S. cerevisiae* and 4-nt in *A. thaliana*).

The *A. thaliana* ptDNA read length distribution lacks distinct peaks occurring at regular intervals and instead contains a single, less extreme peak comprised of reads 24 nt in length (Figs. 3.3, S3.13). The *D. melanogaster* mtDNA-mapping reads have a different read length distribution compared to the nuclear-mapping reads (Figs. 3.4, S3.19, S3.20), but the mtDNA-mapping reads lack the discrete peaks we observed in *S. cerevisiae* and *A. thaliana* organellar reads.

The origins of the distinct peaks in the XR-seq read length distributions of the *S. cerevisiae* and *A. thaliana* datasets are unclear. It is possible that these DNA fragments are derived from a mitochondrial-specific repair pathway for photodamage repair. Alternatively, the abundance of reads of certain lengths could arise from mtDNA degradation, which would represent a previously uncharacterized mechanism of damage-induced mtDNA degradation that results in fragments of specific lengths. One possibility is that the regular packaging of mitochondrial nucleoids could result in areas that are exposed to initial DNA damage or to

subsequent incision/degradation, thereby affecting the length profile of mtDNA fragments after UV exposure. For example, the mtDNA binding protein ABF2 has been shown to protect yeast mtDNA from oxidative damage (O'Rourke *et al.* 2002) and the mammalian homolog TFAM is known to bind mtDNA every ~16 bps (Kukat *et al.* 2011).

XR-seq experiments in *E. coli* reveal somewhat similar patterns in the sense that there are a few read lengths that account for most of the reads in the length distribution, except that in *E. coli* most reads are of 10 or 13 nt in length (Adebali *et al.* 2017b). Interestingly, there is only a single peak of 13 nt in excision assays with *E. coli* mutants lacking *UvrD*, presumably because the primary 13mer oligonucleotide is unable to dissociate from the UvrB-UvrC heterodimer without the activity of the UvrD helicase and is therefore inaccessible to the exonucleases that degrade the oligonucleotide from the 3′ end (Adebali *et al.* 2017a). In the *S. cerevisiae* nucDNA derived reads, TT peaks consistently occur 6 nt from the 3′ ends of reads, including in reads less than 23 nt (the length of the primary excision product in nucDNA NER), suggesting that nucDNA-derived reads are degraded from the 5′ ends. Given that secondary excision products have been shown to arise through exonuclease degradation of a primary oligonucleotide in *E. coli* and in the *S. cerevisiae* nucDNA NER, we hypothesize that the 26-nt peak in the *S. cerevisiae* mtDNA may be a 'primary' product, with the less abundant 24, 22 and 20 nt oligonucleotides arising through degradation.

Attempts to visualize and validate the read length distributions observed in XR-seq data with a conventional excision assay found that the *S. cerevisiae* mtDNA signal was undetectable above background, even when the nucDNA signal was reduced by using a nuclear NER-deficient mutant strain background (*rad14D*) (Fig. S3.24). It is likely that the signal of repair or degradation from the mtDNA is relatively weak compared to the noise of the assay (see faint

gray smear in every lane of Fig. S3.24). Future efforts to identify mtDNA fragments in excision assays may benefit from increased sample volumes and from physically isolating mitochondria from the cell suspensions before immunoprecipitation with anti-damage antibodies.

Preferential positioning of pyrimidines within organellar-mapping XR-seq reads

We analyzed the nucleotide and di-pyrimidine frequencies of all the organellar mapping reads, focusing especially on the dominant read lengths in the *S. cerevisiae* and *A. thaliana* mtDNA datasets (shown in Fig. 3.2 and Fig. 3.3, respectively). In the 26-nt *S. cerevisiae* mtDNA mapping reads (the most frequent length class) in the CPD dataset, adjacent thymines (TTs) are most abundant at position 7-8, with additional smaller peaks spaced at 2-nt intervals, starting at positions 10 (left panel of Fig. 3.5; di-pyrimidine frequencies of all the (6-4)PP-mapping reads including rare size classes are shown in Fig. S3.3). The 24-nt reads show a similar TT peak pattern, though it is shifted forward two positions compared to the pattern in the 26-nt reads (i.e., a peak at position 5-6, followed by secondary peaks starting at positions 8, 10, and 12). In the 22-nt reads the TT peaks are shifted forward four positions. Therefore, the TT peaks fall in the same position when the 26, 24 and 22-nt reads are 3′ or right aligned as they are in Fig. 3.5. In the (6-4)PP dataset, the TT peaks also fall in similar positions when the most common reads (26, 24, 22 and 20 nt) are 3’ aligned (right panel of Fig. 3.5, di-pyrimidine frequencies of all the (6-4)PP mapping reads including rare size classes are shown in Fig. S3.8), though the TT peak at position 7-8 does not rise above the null expectation derived from the frequency of TTs in the *S. cerevisiae* mtDNA.

The observed patterns in the mtDNA-derived XR-seq reads may arise through mtDNA degradation or through an incision-based repair process. In either scenario, we propose a

potential mechanism in which 'primary' 26-nt DNA fragments may be degraded in 2-nt intervals from the 5′ end to produce 24, 22 and 20-nt products (Fig. 3.7, right panel). Alternatively, incisions of 6, 4 or 2 nt upstream of a CPD could yield the 26, 24 and 22-nt products, respectively (Fig. 3.7, left panel). Under the model that these DNA fragments arise through wholesale mtDNA degradation, rather than a specific incision-based pathway, we hypothesize that TT dimers inhibit mtDNA degrading nucleases from accessing upstream or downstream nucleotides, resulting in the enrichment of di-pyrimidines at internal locations in the DNA fragments, though we are not aware of examples of exonucleases stalling at such distances from pyrimidine-dimers in the literature. Instead, previous efforts to understand the fate of excised oligonucleotides generated during NER in nucDNA have identified multiple exonucleases that can remove nucleotides up to a dimer (Kemp and Sancar 2012; Hu *et al.* 2013; Adebali *et al.* 2017a; Kim *et al.* 2022).

DNA fragments with dimers on the end are difficult to study because they can be recalcitrant to elongation by terminal transferase enzymes necessary for radiolabeling and likely to ligation of adaptors necessary for XR-seq (Kim *et al.* 2022). Therefore, it is possible that 'dimer-capped' mtDNA molecules generated through exonuclease activity up to the dimer would be undetectable in XR-seq datasets. Previous XR-seq studies have also suggested that adaptor ligation biases may drive variation in nucleotide composition within reads (Li *et al.* 2017), especially at read-ends where adaptors are ligated. We cannot rule out the possibility that adaptor ligation biases are responsible for the enrichment of TT dinucleotides at specific positions (e.g., position 7-8 in 26-nt *S. cerevisiae* CPD reads). It is also possible that the anti-damage antibodies favor binding to certain sequence motifs, which could lead to the preferential positioning of pyrimidine peaks within reads. However, we view ligation or antibody biases as unlikely

explanations given that the enrichment patterns differ greatly across species (e.g., *S. cerevisiae* vs. *D. melanogaster*) and across genomes of the same species (*S. cerevisiae* nucDNA vs. mtDNA or *A. thaliana* ptDNA vs mtDNA) despite use of the same antibodies and adaptors. Ligation biases seem especially unlikely given that the TT enrichment is internal to the fragments and not directly at the ligated ends and often outside the random 5-nt sequence used for adaptor annealing. Further, the enriched positions shift relative to the 5′ end depending on the length of read (Fig. 3.5, Fig. 3.7), whereas if ligation biases were responsible, we would expect a consistent position of pyrimidines relative to the ends of reads.

As in the *S. cerevisiae* dataset, the TT peaks in the *A. thaliana* mtDNA mapping reads fell in the same position when the most frequent read lengths are aligned (in this case reads 32, 28, 24, 20 and 16 nt long). For the *A. thaliana* dataset, this pattern holds regardless of whether the reads are left (5′) aligned (as in Fig. 3.6; di-pyrimidine frequencies of all the mtDNA-mapping reads including rare size classes are shown in Fig S14) or right (3′) aligned (not shown). If these DNA fragments are arising through targeted incisions, we posit that these incisions occur primarily either 6, 10, 14, 18 or 22 nt upstream of a CPD, and either 8, 12, 16, 20 or 24 nt downstream of a CPD. Alternatively, the 4-nt spacing of pyrimidine dimers could be explained by regular degradation of a primary excision product of undetermined length. Yet another possibility might be that if these fragments are arising through mtDNA degradation, it would again appear that degrading nucleases are unable to access DNA within a certain distance of pyrimidine dimers.

The *A. thaliana* ptDNA and *D. melanogaster* mtDNA reads lack obvious di-pyrimidine patterns (Figs. S3.15, S3.21). Interestingly, the *D. melanogaster* mtDNA mapping reads show extreme nucleotide biases at both the 5′ and 3′ ends of reads (Fig. S3.22). Such biases may be

driven by biased composition of the overhanging Ns that allow for adaptor annealing. However, nuclear mapping XR-seq reads do not display extreme nucleotide biases at read ends (see Figure 1c; Deger et al. 2019). End biases are also mostly absent from the *S. cerevisiae* and *A. thaliana* organellar-mapping reads (Figures S3.4, S3.9, S3.16, S3.17; respectively), which were created with the same or similar adaptors. Another explanation could be that in the *D. melanogaster* mtDNA, distance from a CPD is not important in determining upstream or downstream incision sites, and instead local sequence contexts drive incision locations. Such a phenomenon would also explain why the *D. melanogaster* read length distribution lacks discrete peaks (Fig. S3.20) and why the *D. melanogaster* mtDNA-mapping reads lack an enriched localization of di-pyrimidines (Fig. S3.21).

Variation in the distribution of XR-seq reads among genomic regions

We determined the location of the organellar mapping reads as either intergenic, CDS (protein coding), intronic, tRNA coding, or rRNA coding. In both *S. cerevisiae* datasets (CPD and (6-4)PP), we find elevated coverage of genic regions (CDS, rRNA and tRNA) compared to coverage in intergenic regions (Figures S3.5, S3.10; left panel). This pattern is consistent with trends in the *S. cerevisiae* nuclear genome (Li *et al.* 2018), where increased genic XR-seq coverage is attributed to transcription-coupled NER (TC-NER). Another feature of TC-NER is increased coverage of the template DNA strand compared to the coding DNA strand. After controlling for differences in the numbers of dipyrimidines on the template DNA strand compared to the coding DNA strand we find elevated coverage of the template strand compared to the coding strand in the tRNA coding regions of the genome in both *S. cerevisiae* datasets (CPD and (6-4)PP; Figures S3.5, S3.10; right panel). However, CDS, intronic, and rRNA regions

show no difference in coverage of the template vs. coding strand or slightly elevated coverage of the coding strand compared to the template strand, which is inconsistent with expectations of TC-NER.

In the *A. thaliana* mtDNA, we see slightly elevated XR-seq coverage of the CDS compared to the intergenic regions of the genome, but rRNA and tRNA genes, which are typically expressed more highly than CDS regions (Belozerova *et al.* 2011; Di Giorgio *et al.* 2019), have XR-seq coverage below or near the level of intergenic sequence (Fig. S3.18, top left panel). This suggests that increases in expression may not correlate with increased levels of incisions or repair activity as is observed in the *A. thaliana* nucDNA due to TC-NER (Oztas *et al.* 2018). In the *A. thaliana* ptDNA, we see relatively even levels of CDS and intergenic coverage but decreased coverage of rRNAs and tRNAs (Fig. S3.18, top right panel), again opposite of the expectations under a TC repair model where more highly expressed genes receive increased NER protection. If the organellar-derived DNA fragments arise through organellar genome degradation rather than by uncharacterized repair pathways, variation in XR-seq read depth across genomic compartments may provide a snapshot of variation in damage formation. There are no large effect asymmetries in coding vs template strand in the *A. thaliana* data (Fig. S3.18, bottom panels) except for in mtDNA and ptDNA rRNA genes, especially in the ptDNA where template coverage is roughly 2-fold that of the coding strand. It is difficult to know whether these asymmetries arise through variation in damage formation, uncharacterized repair pathways, or asymmetrical DNA degradation. Therefore, overall, we find very little support for the possibility that the observed DNA fragments produced in response to UV damage are dependent on transcriptional activity.

Differences in XR-seq coverage between genomic regions may also arise from different levels of pyrimidine dimer formation, which has been shown to vary across nucDNAs due to variation in local sequence motifs and nucleosome density (Mao *et al.* 2016). Organellar DNA lacks nucleosomes and is instead packaged in nucleoids, which can vary in protein components based on developmental and physiological status of a given organelle, but are generally assumed to confer many of the same protective benefits as nucleosomes (Bogenhagen 2012; Sakamoto and Takami 2018; Zhao 2019).

In the *D. melanogaster* mtDNA, we find a drastic reduction in coverage of the intergenic portion of the genome compared to the CDS, rRNA and tRNA genes (Fig. S3.23, top panel). Metazoan mtDNAs are extremely gene dense, so essentially all of the 'intergenic' sequence in the *D. melanogaster* mtDNA is located in the AT-rich region of the genome, which serves as the mtDNA replication origin and termination sites. Given the preponderance of thymines in this region, one might expect an increase in CPD formation compared to other regions of the genome, making the lack of XR-seq read in this region intriguing. However, AT rich sequences also experience negative amplification biases during the PCR stages of library construction (Daniel *et al.* 2011; Wu *et al.* 2020a; Waneka *et al.* 2021), so comparisons of XR-seq coverage between regions of varied AT content must be made cautiously. Coverage is lower in on template strand than the coding strand in all genomic regions in the *D. melanogaster* mtDNA, opposite of expectations given TC-NER (Fig. S3.23, bottom panel).

CONCLUSIONS

Early studies that found no repair of UV-damage mtDNAs in human and yeast cells (Clayton *et al.* 1974; Prakash 1975) helped shape the notion that mitochondria lack DNA repair altogether

and that damaged mtDNA molecules are simply degraded, with undamaged copies serving as templates for new mtDNA synthesis (Druzhyna *et al.* 2008). While subsequent investigations have unveiled that specific types of mtDNA base damage such as deamination, simple alkylation, and oxidation can indeed be effectively repaired within the mitochondria, it is still generally accepted that all eukaryotes lack any pathway for repair of bulky DNA damage in mtDNAs (Kazak *et al.* 2012; Stein and Sia 2017; Alencar *et al.* 2019; Chevigny *et al.* 2020). MtDNA damage has been demonstrated to lead to mtDNA degradation in a variety of instances (Bess *et al.* 2012, 2013; Shokolenko *et al.* 2016; Moretton *et al.* 2017; Wang *et al.* 2017; Dan *et al.* 2020), but this process remains enigmatic, with open questions as to how damaged mtDNAs are distinguished from healthy mtDNAs, how damaged mtDNAs promote fusion and or mitophagy (Bess *et al.* 2013; Wang *et al.* 2017; Doblado *et al.* 2021), and which enzymes actually degrade the mtDNA (Moretton *et al.* 2017; Matic *et al.* 2018; Peeva *et al.* 2018).

Our analysis of XR-seq experiments shows that mitochondrially derived DNA fragments of characteristic length and nucleotide composition are produced following mtDNA photodamage in both *S. cerevisiae* and *A. thaliana*. As we have laid out, we envision two potential mechanisms that could be responsible for productions of these DNA fragments: 1) an uncharacterized repair pathway functioning in mitochondria, or 2) a previously uncharacterized programmed degradation of damaged mtDNA. Either of these possibilities point to the exciting prospect of novel maintenance or processing in response to exogenous damage. A key next step in differentiating between these and other possible models will be to be identify the specific molecular machinery that produces the observed DNA fragments in response to UV damage**.**

Table 1: Fraction of each organellar genome retained after filtering to remove multi-mapping reads. Retained fractions are the averages of all replicates for each dataset.

	S. cerevisiae mtDNA: CPD	*S. cerevisiae* mtDNA: (6-4)PP	*A. thaliana* mtDNA: CPD	*A. thaliana* ptDNA: CPD	*D. melanogaster* mtDNA: CPD
intergenic	0.88	0.88	0.88	0.95	0.49
intron	0.88	0.88	0.97	0.97	Not applicable
CDS	0.94	0.94	0.91	0.96	0.997
rRNA	0.91	0.91	0.97	0.91	0.995
tRNA	0.97	0.97	0.61	0.88	0.9996
total	0.90	0.90	0.89	0.95	0.87

Table 2: Organellar vs. nuclear XR-seq coverage (as RPKM)

	S. cerevisiae mtDNA: CPD	*S. cerevisiae* mtDNA: (6-4)PP	*A. thaliana* mtDNA: CPD	*A. thaliana* ptDNA: CPD	*D. melanogaster* mtDNA: CPD
organellar RPKM	25.0	72.1	11.7	13.6	424.9
nuclear RPKM	82.7	82.3	8.4	8.4	6.9
ratio: org/nuc RPKM	0.30	0.88	1.39	1.62	61.58

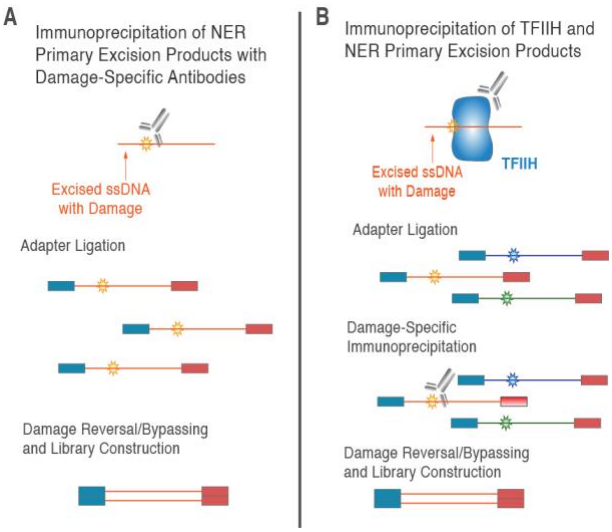

Figure 3.1. Overview of XR-seq protocol. Panel A shows direct capture of damage containing excised oligomers as performed in experiments with S. cerevisiae, A. thaliana, and D. melanogaster. After immunoprecipitation with damage specific antibodies, adaptors are attached to excised and the damaged sites are repaired by a photolyase before the molecules are amplified and sequenced. Panel B shows the alternative XR-seq approach, which includes an initial immunoprecipitation against TFIIH (the enzymatic complex that associates with excised oligomers in mammalian cells). Then, adaptors are ligated to the ssDNA fragments before a second immunoprecipitation with anti-damage antibodies, photolyase damage reversal, and library amplification/sequencing.

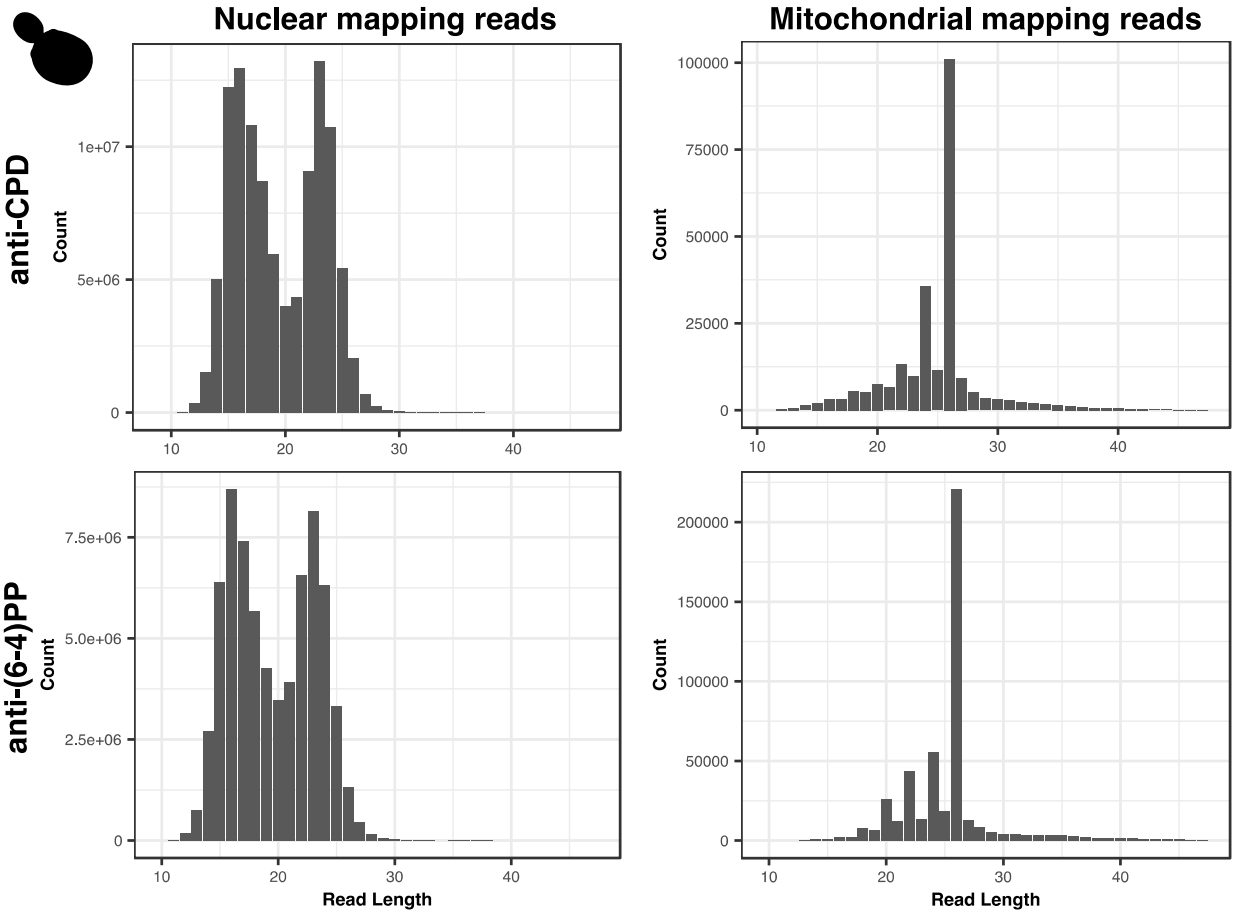

Figure 3.2. Read length distributions of nuclear and mitochondrial reads from anti-CPD and anti-(6-4)PP libraries from *S. cerevisiae.* These distributions exhibited a high degree of repeatability across samples and conditions. Pearson's correlation analyses reveal significant correlations between the anti-CPD and anti-(6-4)PP read length distributions (R=0.9555, p=1.6E-11) as well as between anti-CPD (5 min vs 20 min; R=0.9951, p=2.2E-16) and anti-(6-4)PP (5 min vs 20 min; R=0.9765, p=3.92E-14) timepoints.

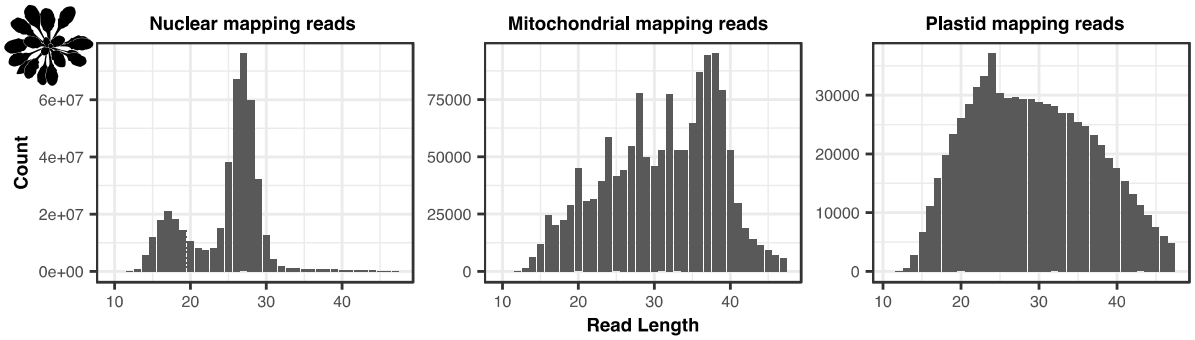

Figure 3.3. Read length distributions of nuclear, mitochondrial and plastid mapping reads from the *A. thaliana* anti-CPD libraries. Pearson's correlation analyses reveal significant correlations between in the mtDNA read length distributions between time points (2 hours vs. 5 hours: R= 0.9899, p=2.2E-16).

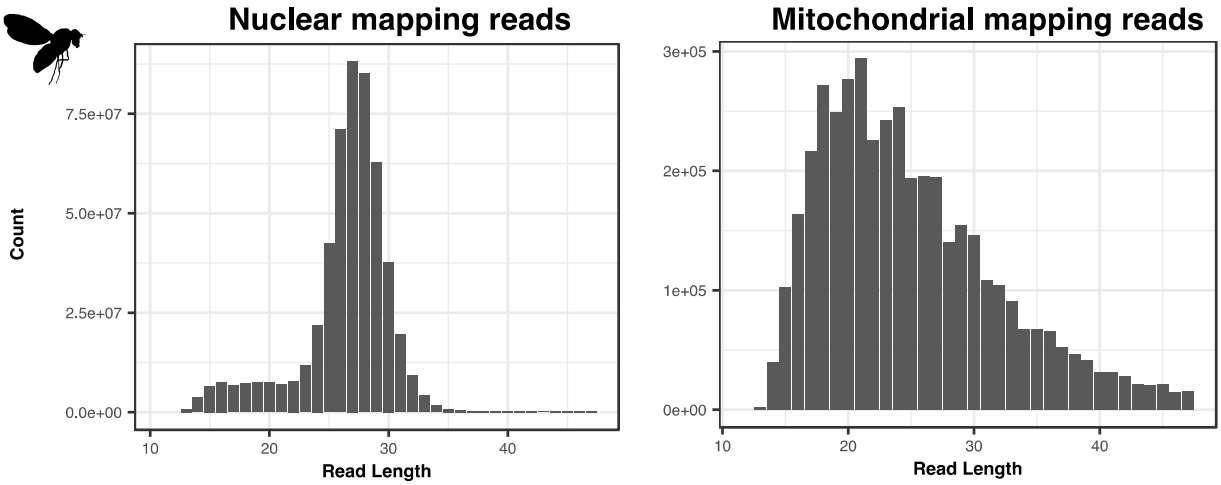

Figure 3.4. Read length distributions of nuclear and mitochondrial mapping reads from the *D. melanogaster* anti-CPD libraries.

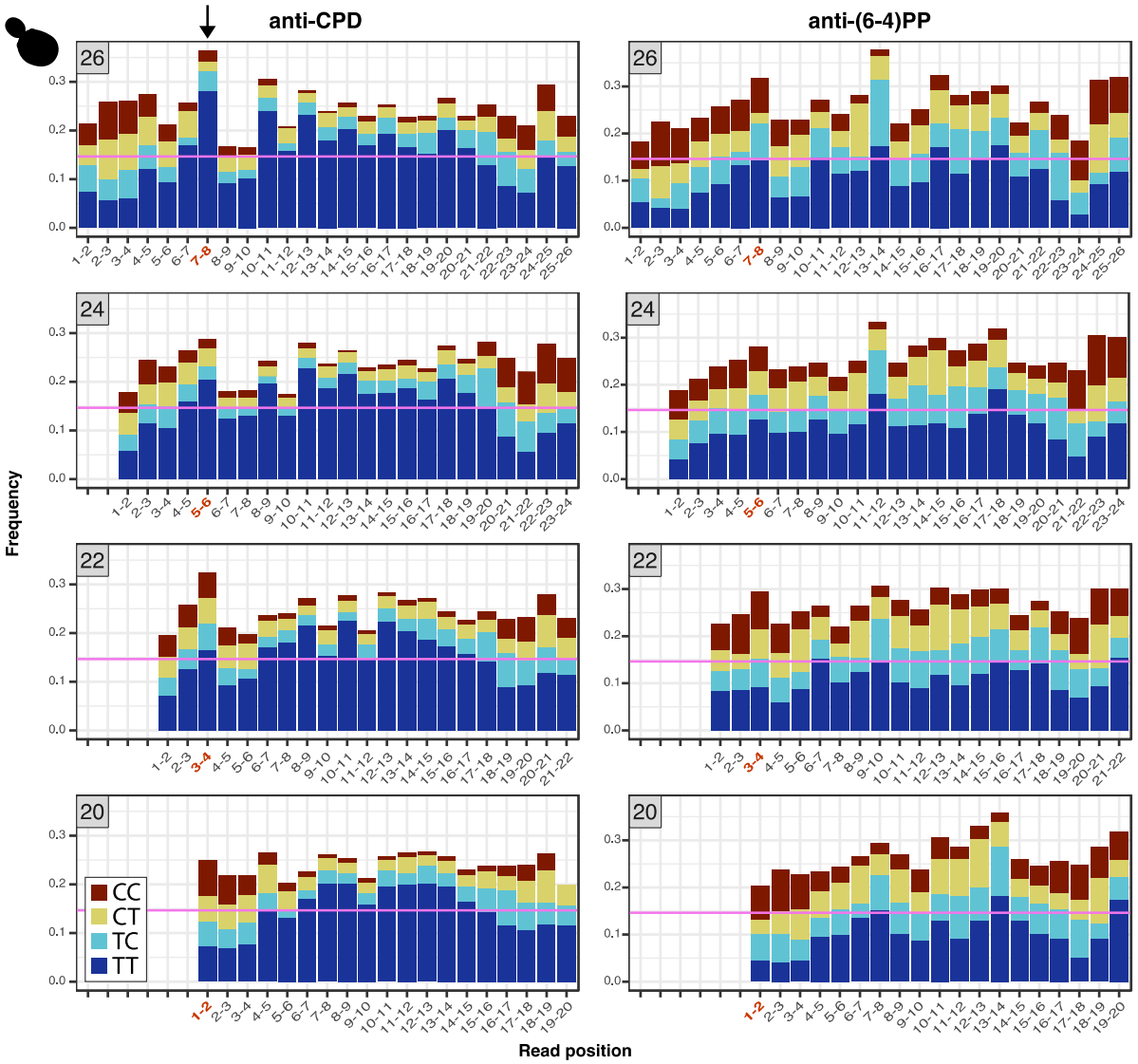

Figure 3.5. Di-pyrimidine frequencies in the most abundant read-length classes (26, 24, 22 and 20 nt) from the *S. cerevisiae* anti-CPD libraries. Read lengths are denoted in the gray boxes at the top left of each panel. The pink horizontal lines show the frequency of TT dinucleotides in the *S. cerevisiae* mtDNA, providing a null expectation for TT dinucleotide frequencies in the XR-seq reads. Positions with TT peaks in the 26 nt reads are in red, and the equivalent positions in the 3′ aligned (right aligned) 24, 22 and 20 nt reads are also in red. We approximated the 95% confidence interval as 2 times the standard error of the expected TT frequency given the number of reads included for each di-pyrimidine calculation. Given the large number of reads analyzed, 95 % confidence intervals are very small, ranging from 0.1472 ± 0.0022 for the CPD 26-nt reads to 0.1472 ± 0.0081 for the CPD 20-nt reads. As a result, all blue bars that appear above the pink line in the figure represent a significant statistical enrichment relative to the expectation and its 95% confidence interval.

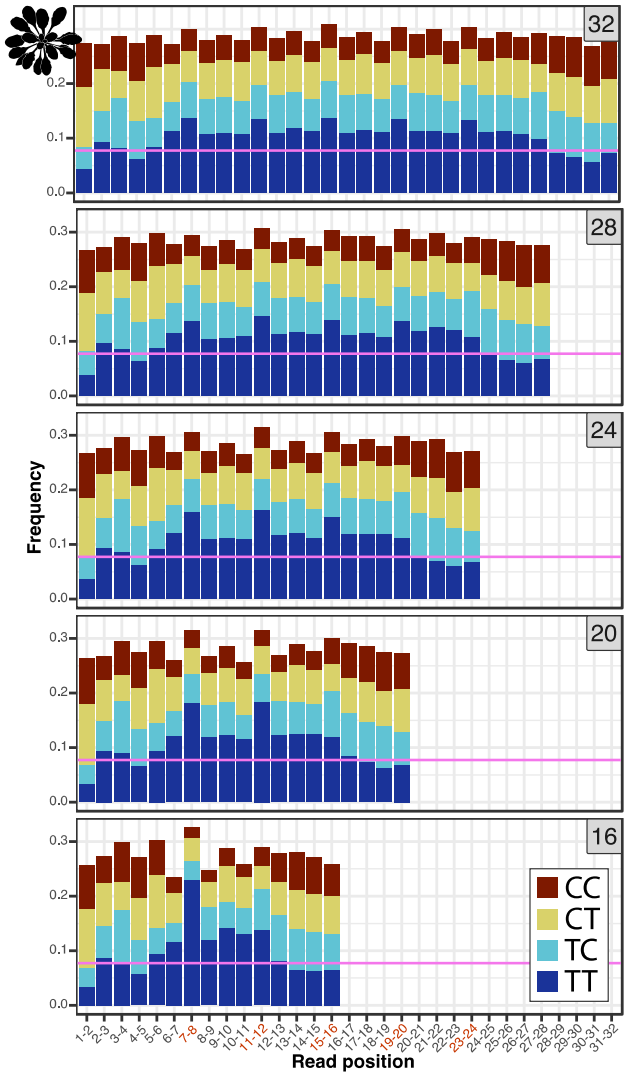

Figure 3.6. Di-pyrimidine frequencies in the most abundant read-length classes (32, 28, 24, 20 and 16 nt) from the *A. thaliana* mtDNA mapping reads. Read lengths are denoted in the gray boxes at the top right of each panel. The pink horizontal lines show the frequency of TT dinucleotides in the *A. thaliana* mtDNA, providing a null expectation for TT dinucleotide frequencies in the XR-seq reads. See Fig. 3.5 for a description of calculating 95% confidence intervals around this expectation. These confidence intervals were very small due to the large number of reads, ranging from 0.0743 ± 0.0033 for the 16 nt reads to 0.0743 ± 0.0018 for the 32 nt reads.

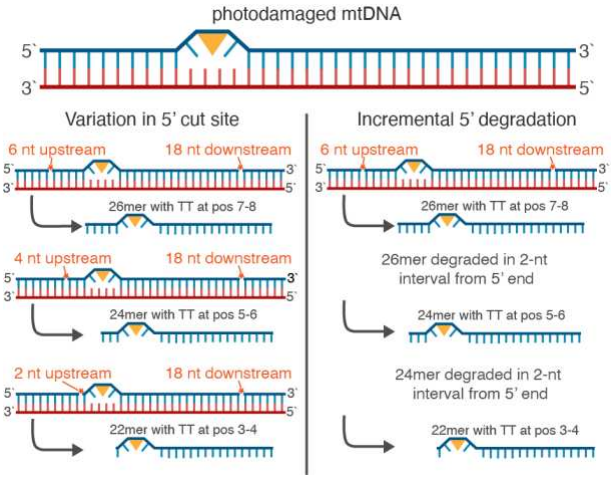

Figure 3.7. Proposed alternative explanations for unique read length distributions and dinucleotide composition patterns in the *S. cerevisiae* anti-CPD and anti (6-4)PP datasets.

LITERATURE CITED

Adebali O., Y. Y. Chiou, J. Hu, A. Sancar, and C. P. Selby, 2017a Genome-wide transcription-coupled repair in Escherichia coli is mediated by the Mfd translocase. Proc. Natl. Acad. Sci. 114: E2116–E2125. https://doi.org/10.1073/pnas.1700230114

Adebali O., A. Sancar, and C. P. Selby, 2017b Mfd translocase is necessary and sufficient for Transcription-coupled repair in Escherichia coli. J. Biol. Chem. 292: 18386–18391. https://doi.org/10.1074/jbc.C117.818807

Akkose U., V. O. Kaya, L. Lindsey-Boltz, Z. Karagoz, A. D. Brown, *et al.*, 2020 Comparative analyses of two primate species diverged by more than 60 million years show different rates but similar distribution of genome-wide UV repair events. BMC Genomics 22: 1–13. https://doi.org/10.1101/2020.04.06.027201

Alencar R. R., C. M. P. F. Batalha, T. S. Freire, and N. C. de Souza-Pinto, 2019 *Enzymology of mitochondrial DNA repair*. Elsevier Inc.

Alhegaili A. S., Y. Ji, N. Sylvius, M. J. Blades, M. Karbaschi, *et al.*, 2019 Genome-wide adductomics analysis reveals heterogeneity in the induction and loss of cyclobutane thymine dimers across both the nuclear and mitochondrial genomes. Int. J. Mol. Sci. 20. https://doi.org/10.3390/ijms20205112

Altschuler S. E., K. A. Lewis, and D. S. Wuttke, 2013 Practical strategies for the evaluation of high-affinity protein/nucleic acid interactions. J. Nucleic Acids Investig. 4: 3. https://doi.org/10.4081/jnai.2013.e3

Amente S., G. Di Palo, G. Scala, T. Castrignanò, F. Gorini, *et al.*, 2019 Genome-wide mapping of 8-oxo-7,8-dihydro-2′-deoxyguanosine reveals accumulation of oxidatively-generated

damage at DNA replication origins within transcribed long genes of mammalian cells. Nucleic Acids Res. 47: 221–236. https://doi.org/10.1093/nar/gky1152

Andrews S., 2010 Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data. URL https//www. bioinformatics. babraham. ac. uk/projects/fastqc.

Belozerova N. S., E. S. Pozhidaeva, A. G. Shugaev, and V. V. Kusnetsov, 2011 Run-on transcription as a method for the analysis of mitochondrial genome expression. Russ. J. Plant Physiol. 58: 164–168. https://doi.org/10.1134/S1021443711010031

Bess A. S., T. L. Crocker, I. T. Ryde, and J. N. Meyer, 2012 Mitochondrial dynamics and autophagy aid in removal of persistent mitochondrial DNA damage in Caenorhabditis elegans. Nucleic Acids Res. 40: 7916–7931. https://doi.org/10.1093/nar/gks532

Bess A. S., I. T. Ryde, D. E. Hinton, and J. N. Meyer, 2013 UVC-Induced Mitochondrial Degradation via Autophagy Correlates with mtDNA Damage Removal in Primary Human Fibroblasts. J. Biochem. Mol. Toxicol. 27: 28–41. https://doi.org/10.1002/jbt.21440

Bogenhagen D. F., 2012 Mitochondrial DNA nucleoid structure. Biochim. Biophys. Acta - Gene Regul. Mech. 1819: 914–920. https://doi.org/10.1016/j.bbagrm.2011.11.005

Boulé J. B., F. Rougeon, and C. Papanicolaou, 2001 Terminal Deoxynucleotidyl Transferase Indiscriminately Incorporates Ribonucleotides and Deoxyribonucleotides. J. Biol. Chem. 276: 31388–31393. https://doi.org/10.1074/jbc.M105272200

Chevigny N., D. Schatz-Daas, F. Lotfi, and J. M. Gualberto, 2020 DNA repair and the stability of the plant mitochondrial genome. Int. J. Mol. Sci. 21. https://doi.org/10.3390/ijms21010328

Clayton D. A., J. N. Doda, and E. C. Friedberg, 1974 The absence of a pyrimidine dimer repair mechanism in mammalian mitochondria. Proc. Natl. Acad. Sci. 71: 2777–2781. https://doi.org/10.1073/pnas.71.7.2777

Dan X., M. Babbar, A. Moore, N. Wechter, J. Tian, *et al.*, 2020 DNA damage invokes mitophagy through a pathway involving Spata18. Nucleic Acids Res. 48: 6611–6623. https://doi.org/10.1093/nar/gkaa393

Daniel A., R. Michael, Chen Wei-Sheng, Danielsson Maxwell, Fennell Timothy, *et al.*, 2011 Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12: 1–14.

Deger N., Y. Yang, L. A. Lindsey-Boltz, A. Sancar, and C. P. Selby, 2019 Drosophila, which lacks canonical transcription-coupled repair proteins, performs transcription-coupled repair. J. Biol. Chem. 294: 18092–18098. https://doi.org/10.1074/jbc.AC119.011448

Doblado L., C. Lueck, C. Rey, A. K. Samhan-arias, I. Prieto, *et al.*, 2021 Mitophagy in human diseases. Int. J. Mol. Sci. 22: 1–51. https://doi.org/10.3390/ijms22083903

Drouin G., H. Daoud, and J. Xia, 2008 Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol. Phylogenet. Evol. 49: 827–831. https://doi.org/10.1016/j.ympev.2008.09.009

Druzhyna N. M., G. L. Wilson, and S. P. LeDoux, 2008 Mitochondrial DNA repair in aging and disease. Mech. Ageing Dev. 129: 383–390. https://doi.org/10.1016/j.mad.2008.03.002

Emmerich H. J., M. Saft, L. Schneider, D. Kock, A. Batschauer, *et al.*, 2020 A topologically distinct class of photolyases specific for UV lesions within single-stranded DNA. Nucleic Acids Res. 48: 12845–12857. https://doi.org/10.1093/nar/gkaa1147

Fields P. D., G. Waneka, M. Naish, M. C. Schatz, I. R. Henderson, *et al.*, 2022 Complete Sequence of a 641-kb Insertion of Mitochondrial DNA in the Arabidopsis thaliana Nuclear Genome. Genome Biol. Evol. 14: 1–9. https://doi.org/10.1093/gbe/evac059

Giorgio J. A. P. Di, É. Lepage, S. Tremblay-Belzile, S. Truche, A. Loubert-Hudon, *et al.*, 2019

Transcription is a major driving force for plastid genome instability in Arabidopsis. PLoS One 14: 1–30. https://doi.org/10.1371/journal.pone.0214552

Göke A., S. Schrott, A. Mizrak, V. Belyy, C. Osman, *et al.*, 2020 Mrx6 regulates mitochondrial DNA copy number in Saccharomyces cerevisiae by engaging the evolutionarily conserved Lon protease Pim1. Mol. Biol. Cell 31: 527–545. https://doi.org/10.1091/MBC.E19-08-0470

Goldstein A. L., and J. H. McCusker, 1999 Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae. Yeast 15: 1541–1553. https://doi.org/10.1002/(SICI)1097-0061(199910)15:14<1541::AID-YEA476>3.0.CO;2-K

Gonzalez-Hunt C. P., M. Wadhwa, and L. H. Sanders, 2018 DNA damage by oxidative stress: Measurement strategies for two genomes. Curr. Opin. Toxicol. 7: 87–94. https://doi.org/10.1016/j.cotox.2017.11.001

Goosen N., and G. F. Moolenaar, 2008 Repair of UV damage in bacteria. DNA Repair (Amst). 7: 353–379. https://doi.org/10.1016/j.dnarep.2007.09.002

Guzder S. N., C. H. Sommers, L. Prakash, and S. Prakash, 2006 Complex Formation with Damage Recognition Protein Rad14 Is Essential for Saccharomyces cerevisiae Rad1-Rad10 Nuclease To Perform Its Function in Nucleotide Excision Repair In Vivo . Mol. Cell. Biol. 26: 1135–1141. https://doi.org/10.1128/mcb.26.3.1135-1141.2006

Harman D., 1972 The Biologic Clock: The Mitochondria? J. Am. Geriatr. Soc. 20: 145–147.

Havird J. C., and D. B. Sloan, 2016 The roles of mutation, selection, and expression in determining relative rates of evolution in mitochondrial versus nuclear genomes. Mol. Biol. Evol. 33: 3042–3053.

Herbers E., N. J. Kekäläinen, A. Hangas, J. L. Pohjoismäki, and S. Goffart, 2019 Tissue specific

differences in mitochondrial DNA maintenance and expression. Mitochondrion 44: 85–92. https://doi.org/10.1016/j.mito.2018.01.004

Hu J., J. H. Choi, S. Gaddameedhi, M. G. Kemp, J. T. Reardon, *et al.*, 2013 Nucleotide excision repair in human cells: Fate of the excised oligonucleotide carrying dna damage in vivo. J. Biol. Chem. 288: 20918–20926. https://doi.org/10.1074/jbc.M113.482257

Hu J., S. Adar, C. P. Selby, J. D. Lieb, and A. Sancar, 2015 Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. Genes Dev. 29: 948–960. https://doi.org/10.1101/gad.261271.115

Hu J., O. Adebali, S. Adar, and A. Sancar, 2017 Dynamic maps of UV damage formation and repair for the human genome. Proc. Natl. Acad. Sci. 114: 6758–6763. https://doi.org/10.1073/pnas.1706522114

Hu J., W. Li, O. Adebali, Y. Yang, O. Oztas, *et al.*, 2019 Genome-wide mapping of nucleotide excision repair with XR-seq. Nat. Protoc. 14: 248-282f. https://doi.org/10.1038/s41596-018-0093-7

Hunter S. E., D. Jung, R. T. Di Giulio, and J. N. Meyer, 2010 The QPCR assay for analysis of mitochondrial DNA damage, repair, and relative copy number. Methods 51: 444–451. https://doi.org/10.1016/j.ymeth.2010.01.033

John J. C. St., 2016 Mitochondrial DNA copy number and replication in reprogramming and differentiation. Semin. Cell Dev. Biol. 52: 93–101. https://doi.org/10.1016/j.semcdb.2016.01.028

Kalinowski D. P., S. Illenye, and B. van Houten, 1992 Analysis of DNA damage and repair in murine leukemia L1210 cells using a quantitative polymerase chain reaction assay. Nucleic Acids Res. 20: 3485–3494. https://doi.org/10.1093/nar/20.13.3485

Kazak L., A. Reyes, and I. J. Holt, 2012 Minimizing the damage: Repair pathways keep mitochondrial DNA intact. Nat. Rev. Mol. Cell Biol. 13: 659–671. https://doi.org/10.1038/nrm3439

Kemp M. G., and A. Sancar, 2012 DNA excision repair: Where do all the dimers go? Cell Cycle 11: 2997–3002. https://doi.org/10.4161/cc.21126

Kim S. H., G. H. Kim, M. G. Kemp, and J. H. Choi, 2022 TREX1 degrades the 3′ end of the small DNA oligonucleotide products of nucleotide excision repair in human cells. Nucleic Acids Res. 50: 3974–3984. https://doi.org/10.1093/nar/gkac214

Kukat C., C. A. Wurm, H. Spåhr, M. Falkenberg, N. G. Larsson, *et al.*, 2011 Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently contain a single copy of mtDNA. Proc. Natl. Acad. Sci. 108: 13534–13539. https://doi.org/10.1073/pnas.1109263108

Langmead B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.

Ledoux S. P., G. L. Wilson, E. J. Beecham, T. Stevnsner, K. Wassermann, *et al.*, 1992 Repair of mitochondrial DNA after various types of DNA damage in chinese hamster ovary cells. Carcinogenesis 13: 1967–1973.

Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li W., J. Hu, O. Adebali, S. Adar, Y. Yang, *et al.*, 2017 Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene. Proc. Natl. Acad. Sci. 114: 6752–6757. https://doi.org/10.1073/pnas.1706021114

Li W., O. Adebali, Y. Yang, C. P. Selby, and A. Sancar, 2018 Single-nucleotide resolution

dynamic repair maps of UV damage in Saccharomyces cerevisiae genome. Proc. Natl. Acad. Sci. 115: E3408–E3415. https://doi.org/10.1073/pnas.1801687115

Lindsey-Boltz L. A., Y. Yang, C. Kose, N. Deger, K. Eynullazada, *et al.*, 2023 Nucleotide excision repair in Human cell lines lacking both XPC and CSB proteins. Nucleic Acids Res. 51: 6238–6245. https://doi.org/10.1093/nar/gkad334

Lucas-Lledó J. I., and M. Lynch, 2009 Evolution of mutation rates: Phylogenomic analysis of the photolyase/cryptochrome family. Mol. Biol. Evol. 26: 1143–1153.

Mao P., M. J. Smerdon, S. A. Roberts, and J. J. Wyrick, 2016 Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. Proc. Natl. Acad. Sci. 113: 9057–9062. https://doi.org/10.1073/pnas.1606667113

Martin M., 2011 Cutadapt removes adapter sequences from high-throuoghput sequencing reads. EMBnet.jouonal 17: 10–12.

Matic S., M. Jiang, T. J. Nicholls, J. P. Uhler, C. Dirksen-Schwanenland, *et al.*, 2018 Mice lacking the mitochondrial exonuclease MGME1 accumulate mtDNA deletions without developing progeria. Nat. Commun. 9. https://doi.org/10.1038/s41467-018-03552-x

Mei Q., and V. Dvornyk, 2015 Evolutionary history of the photolyase/cryptochrome superfamily in eukaryotes. PLoS One 10: 1–20. https://doi.org/10.1371/journal.pone.0135940

Modrich P., 2006 Mechanisms in eukaryotic mismatch repair. J. Biol. Chem. 281: 30305–30309. https://doi.org/10.1074/jbc.R600022200

Moretton A., F. Morel, B. Macao, P. Lachaume, L. Ishak, *et al.*, 2017 Selective mitochondrial DNA degradation following double-strand breaks. PLoS One 12: 1–17. https://doi.org/10.1371/journal.pone.0176795

Mori T., M. Nakane, T. Hattori, T. Matsunaga, M. Ihara, *et al.*, 1991 Simultaneous Establishment

of Monoclonal Antibodies Specific for Either Cyclobutane Pyrimidine Dimer or (6-4)Photoproduct From the Same Mouse Immunized With Ultraviolet-Irradiated Dna. Photochem. Photobiol. 54: 225–232. https://doi.org/10.1111/j.1751-1097.1991.tb02010.x

Mower J. P., D. B. Sloan, and A. J. Alverson, 2012 *Plant mitochondrial genome diversity*.

Murphy M. P., 2009 How mitochondria produce reactive oxygen species. Biochem. J. 417: 1–13.

Muthye V., and D. V. Lavrov, 2021 Multiple Losses of MSH1, Gain of mtMutS, and Other Changes in the MutS Family of DNA Repair Proteins in Animals. Genome Biol. Evol. 13: 1–8. https://doi.org/10.1093/gbe/evab191

O'Hara R., E. Tedone, A. Ludlow, E. Huang, B. Arosio, *et al.*, 2019 Quantitative mitochondrial DNA copy number determination using droplet digital PCR with single-cell resolution. Genome Res. 29: 1878–1888. https://doi.org/10.1101/gr.250480.119

O'Rourke T. W., N. A. Doudican, M. D. Mackereth, P. W. Doetsch, and G. S. Shadel, 2002 Mitochondrial Dysfunction Due to Oxidative Mitochondrial DNA Damage Is Reduced through Cooperative Actions of Diverse Proteins. Mol. Cell. Biol. 22: 4086–4093. https://doi.org/10.1128/mcb.22.12.4086-4093.2002

Oztas O., C. P. Selby, A. Sancar, and O. Adebali, 2018 Genome-wide excision repair in Arabidopsis is coupled to transcription and reflects circadian gene expression patterns. Nat. Commun. 9: 1–8. https://doi.org/10.1038/s41467-018-03922-5

Peeva V., D. Blei, G. Trombly, S. Corsi, M. J. Szukszto, *et al.*, 2018 Linear mitochondrial DNA is rapidly degraded by components of the replication machinery. Nat. Commun. 9: 1–11. https://doi.org/10.1038/s41467-018-04131-w

Prakash L., 1975 Repair of pyrimidine dimers in nuclear and mitochondrial DNA of yeast irradiated with low doses of ultraviolet light. J. Mol. Biol. 98: 781–795.

https://doi.org/10.1016/S0022-2836(75)80010-6

Rong Z., P. Tu, P. Xu, Y. Sun, F. Yu, *et al.*, 2021 The Mitochondrial Response to DNA Damage. Front. Cell Dev. Biol. 9: 1–10. https://doi.org/10.3389/fcell.2021.669379

Sakamoto W., and T. Takami, 2018 Chloroplast DNA dynamics: Copy number, quality control and degradation. Plant Cell Physiol. 59: 1120–1127. https://doi.org/10.1093/pcp/pcy084

Saki M., and A. Prakash, 2017 DNA damage related crosstalk between the nucleus and mitochondria. Free Radic. Biol. Med. 107: 216–227. https://doi.org/10.1016/j.freeradbiomed.2016.11.050

Sancar A., 1996 DNA excision repair. Annu. Rev. Biochem. 65: 43–81.

Sancar A., 2004 Photolyase and cryptochrom blue-light photoreceptors. Adv. Protein Chem. 69: 73–100.

Shen J., Y. Zhang, M. J. Havey, and W. Shou, 2019 Copy numbers of mitochondrial genes change during melon leaf development and are lower than the numbers of mitochondria. Hortic. Res. 6. https://doi.org/10.1038/s41438-019-0177-8

Shokolenko I. N., G. L. Wilson, and M. F. Alexeyev, 2016 The “fast” and the “slow” modes of mitochondrial DNA degradation. Mitochondrial DNA 27: 490–498. https://doi.org/10.3109/19401736.2014.905829

Smith D. R., J. Hua, R. W. Lee, and P. J. Keeling, 2012 Relative rates of evolution among the three genetic compartments of the red alga Porphyra differ from those of green plants and do not correlate with genome architecture. Mol. Phylogenet. Evol. 65: 339–344. https://doi.org/10.1016/j.ympev.2012.06.017

Souza-Pinto N. C. de, P. A. Mason, K. Hashiguchi, L. Weissman, J. Tian, *et al.*, 2009 Novel DNA mismatch-repair activity involving YB-1 in human mitochondria. DNA Repair (Amst). 8:

704–719. https://doi.org/10.1016/j.dnarep.2009.01.021

Springer M. Z., and K. F. Macleod, 2016 In Brief: Mitophagy: mechanisms and role in human disease. J. Pathol. 240: 253–255. https://doi.org/10.1002/path.4774

Stein A., and E. A. Sia, 2017 Mitochondrial DNA repair and damage tolerance. Front. Biosci. - Landmark 22: 920–943. https://doi.org/10.2741/4525

Szczesny B., A. W. Tann, M. J. Longley, W. C. Copeland, and S. Mitra, 2008 Long patch base excision repair in mammalian mitochondrial genomes. J. Biol. Chem. 283: 26349–26356. https://doi.org/10.1074/jbc.M803491200

Takahashi M., M. Teranishi, H. Ishida, J. Kawasaki, A. Takeuchi, *et al.*, 2011 Cyclobutane pyrimidine dimer (CPD) photolyase repairs ultraviolet-B-induced CPDs in rice chloroplast and mitochondrial DNA. Plant J. 66: 433–442. https://doi.org/10.1111/j.1365-313X.2011.04500.x

Urbina-Varela R., N. Castillo, L. A. Videla, and A. Del Campo, 2020 Impact of mitophagy and mitochondrial unfolded protein response as new adaptive mechanisms underlying old pathologies: Sarcopenia and non-alcoholic fatty liver disease. Int. J. Mol. Sci. 21: 1–27. https://doi.org/10.3390/ijms21207704

Waneka G., J. M. Svendsen, J. C. Havird, and D. B. Sloan, 2021 Mitochondrial mutations in Caenorhabditis elegans show signatures of oxidative damage and an AT-bias . Genetics 219. https://doi.org/10.1093/genetics/iyab116

Wang H. T., J. H. Lin, C. H. Yang, C. H. Haung, C. W. Weng, *et al.*, 2017 Acrolein induces mtDNA damages, mitochondrial fission and mitophagy in human lung cells. Oncotarget 8: 70406–70421. https://doi.org/10.18632/oncotarget.19710

Waters R., and E. Moustacchi, 1974 The fate of ultraviolet-induced pyrimidine dimers in the

mitochondrial DNA of Saccharomyces cerevisiae following various post-irradiation cell treatments. BBA Sect. Nucleic Acids Protein Synth. 366: 241–250. https://doi.org/10.1016/0005-2787(74)90282-2

Winston F., C. Dollard, and S. L. Ricupero-Hovasse, 1995 Construction of a set of convenient saccharomyces cerevisiae strains that are isogenic to S288C. Yeast 11: 53–55. https://doi.org/10.1002/yea.320110107

Wolfe K. H., W. H. Li, and P. M. Sharp, 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. 84: 9054–9058.

Wood R. D., 1999 DNA damage recognition during nucleotide excision repair in mammalian cells. Biochimie 81: 39–44. https://doi.org/10.1016/S0300-9084(99)80036-4

Wu Z., G. Waneka, and D. B. Sloan, 2020a The tempo and mode of angiosperm mitochondrial genome divergence inferred from intraspecific variation in arabidopsis thaliana. G3 Genes, Genomes, Genet. 10: 1077–1086. https://doi.org/10.1534/g3.119.401023

Wu Z., G. Waneka, A. K. Broz, C. R. King, and D. B. Sloan, 2020b MSH1 is required for maintenance of the low mutation rates in plant mitochondrial and plastid genomes. Proc. Natl. Acad. Sci. 117: 16448–16455. https://doi.org/10.1101/2020.02.13.947598

Yang Y., O. Adebali, G. Wu, C. P. Selby, Y. Y. Chiou, *et al.*, 2018 Cisplatin-DNA adduct repair of transcribed genes is controlled by two circadian programs in mouse tissues. Proc. Natl. Acad. Sci. 115: E4777–E4785. https://doi.org/10.1073/pnas.1804493115

Yasui A., H. Yajima, T. Kobayashi, A. P. M. Eker, and A. Oikawa, 1992 Mitochondrial DNA repair by photolyase. Mutat. Res. Repair 273: 231–236.

Zardoya R., 2020 Recent advances in understanding mitochondrial genome diversity.

F1000Research 9. https://doi.org/10.12688/f1000research.21490.1

Zhao L., 2019 *Mitochondrial DNA degradation: A quality control measure for mitochondrial genome maintenance and stress response*. Elsevier Inc.

Zhao L., and P. Sumberaz, 2020 Mitochondrial DNA Damage: Prevalence, Biological Consequence, and Emerging Pathways. Chem. Res. Toxicol. 33: 2491–2502. https://doi.org/10.1021/acs.chemrestox.0c00083

Zhao W., L. Kong, W. Guan, J. Liu, H. Cui, *et al.*, 2023 Yeast UPS1 deficiency leads to UVC radiation sensitivity and shortened lifespan. Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol. 116: 773–789. https://doi.org/10.1007/s10482-023-01847-8

CHAPTER 4: INVESTIGATING LOW FREQUENCY SOMATIC MUTATIONS IN ARABIDOPSIS WITH DUPLEX SEQUENCING

Summary

Mutations are the source of novel diversity but can also lead to disease and maladaptation. The conventional view is that mutations occur randomly with respect to their environment-specific fitness consequences. However, intragenomic mutation rates can vary dramatically due to transcription coupled repair and local epigenomic modifications, which are non-uniformly distributed across genomes. One sequence feature associated with decreased mutation is higher expression level, which can vary depending on environmental cues. To understand whether the association between expression level and mutation rate creates a systematic relationship with environment-specific fitness effects, we perturbed expression through a heat treatment in *Arabidopsis thaliana*. We quantified gene expression to identify differentially expressed genes, which we then targeted for mutation detection using Duplex Sequencing. This approach provided a highly accurate measurement of the frequency of rare somatic mutations in vegetative plant tissues, which has been a recent source of uncertainty in plant mutation research. We included mutant lines lacking mismatch repair (MMR) and base excision repair (BER) capabilities to understand how repair mechanisms may drive biased mutation accumulation. We found wild type (WT) and BER mutant mutation frequencies to be very low (mean variant frequency 1.8×10^{-8} and 2.6×10^{-8}, respectively), while MMR mutant frequencies were significantly elevated (1.13×10^{-6}). These results show that somatic variant frequencies are extremely low in WT plants, indicating that larger datasets will be needed to address the fundamental evolutionary question as to whether environmental change leads to gene-specific changes in mutation rate

INTRODUCTION

Mutations in DNA sequences accumulate over time and produce the variation that allows populations to adapt to novel or changing environments. In this sense, mutation is the ultimate source of evolutionary innovation. At the same time, mutations are often deleterious (Eyre-Walker and Keightley 2007), and somatic mutations can cause disease, setting up a dynamic where selection may favor alleles that lower mutation rates, even though mutational input is required for adaptation and evolution (Zhang 2023).

The textbook view of mutation and adaptation is that mutations occur randomly with respect to their environment-specific fitness consequences. This principle was established in early investigations by Max Delbrück and Salvador Luria, who found that mutations in bacteria that confer phage resistance were equally likely to occur regardless of whether bacteria were grown in the presence of phage (Luria and Delbrück 1943). In other words, a phage-containing environment creates selection for genetic variants responsible for resistance but does not induce mutations to specifically occur at those loci. After subsequent decades of study, mutations are still widely considered to be random in this respect even though both the type and location of mutations are now known to have non-uniform distributions across genomes. For example, transition substitutions are far more common than transversions in most organisms across the tree of life. This bias in the mutation spectrum arises through the simple properties of DNA bases and chemical damage, but it has important consequences for the relationship between fitness effects and the probability of mutations. Due to the structure of the genetic code, transversions are more likely than transitions to be nonsynonymous (i.e. result in amino acid changes) and, therefore, have harmful fitness effects. As such, the average fitness effect of mutations is lower

than it would be if all types of nucleotide substitutions occurred with equal probability (Eyre-Walker and Keightley 2007).

Mutation rates can also vary depending on genomic location. For example, mutational gradients arise in mammalian mitochondrial genomes because regions near replication origins are single-stranded (and more vulnerable to mutation causing damage) for longer periods during DNA replication (Sanchez-Contreras *et al.* 2021). Variation in intragenomic mutation rates can also occur at smaller scales, such is in *Arabidopsis thaliana* where mutations are enriched in intergenic sequences compared to genes (Ossowski *et al.* 2010; Belfield *et al.* 2018; Weng *et al.* 2019) and in introns compared to exons (Monroe *et al.* 2022, 2023a; Quiroz *et al.* 2023; Staunton *et al.* 2023). Because mutations in coding sequences are more likely to have functional consequences, this biased distribution of mutations should again result in lower average fitness effects than if mutations were uniformly distributed across the genome.

The probability of a mutation, therefore, cannot be considered independent of the fitness consequences of that mutation. However, to challenge the textbook view that mutations occur randomly with respect to environment-specific fitness effects, gene-specific mutational biases would have to systematically vary with changes in the environment. One potential mechanism that could create such a relationship between environment and mutation bias is the coupling of DNA repair surveillance with transcription machinery, which results in lower mutation rates for highly expressed genes (Supek and Lehner 2017; Oztas *et al.* 2018; Huang *et al.* 2018; Huang and Li 2018; Gonzalez-Perez *et al.* 2019; Monroe *et al.* 2022). Therefore, environmental changes that increase a gene's expression level should lower its mutation rate. In addition, highly expressed genes are known to experience stronger selection (Zhang and Yang 2015), so genes may be most protected from mutation in environments where they are most functionally

important. Alternatively, transcription may be mutagenic, as increased DNA damage associated with exposure of single-stranded DNA to mutagens can potentially overpower the increased protection of actively transcribed genes (Kim *et al.* 2007; Jinks-Robertson and Bhagwat 2014; Seplyarskiy *et al.* 2023).

A challenge associated with addressing how local mutation rates vary with environment is the difficulty of measuring mutations in experimental settings. Historical estimates of mutation relied on comparisons of synonymous substitutions between populations or species. Because these substitutions do not result in a change in amino acid, they are expected to experience minimal selection and thus approximate mutational input, though in reality synonymous sites do experience selection due to codon usage bias (Grantham *et al.* 1980; Hershberg and Petrov 2008) and other mechanisms (Bailey *et al.* 2021). It is inherently difficult to measure mutation rates more directly in large multicellular organisms because their long generations require many individuals and/or large amounts of time for sufficient mutations to occur, making methods such as mutation accumulation lines and parent-offspring trio sequencing (Lynch *et al.* 2016; Tatsumoto *et al.* 2017) expensive and time-consuming.

An alternative and potentially complementary approach to mutation accumulation and trio sequencing studies is to detect the mutations that accumulate in an organism's somatic tissues (Gundry and Vijg 2012; Moore *et al.* 2021; Monroe *et al.* 2022; Quiroz *et al.* 2023; Schmitt *et al.* 2023; Staunton *et al.* 2023; Satake *et al.* 2023; Goel *et al.* 2024). This approach benefits from the fact that many more cell lineages can be tracked than just the germline. Inclusion of somatic (vegetative) mutations in recent *Arabidopsis* studies led to the identification of thousands of mutations, which increased power to test for relationships between local mutation rates and various sequence features, such as GC content, DNA methylation, histone

modifications and expression level (Monroe *et al.* 2022). However, this approach appears to have been inaccurate because low frequency somatic variants can be difficult to distinguish from sequencing errors, and reanalysis of the somatic mutation calls showed that many of the putative mutations arose from technical artefacts (Liu and Zhang 2022; Monroe *et al.* 2023a; Wang *et al.* 2023; Monroe *et al.* 2023b). Therefore, the actual frequency of somatic mutations in vegetative plant tissue remains an open question.

Measurements of low frequency somatic mutations can be obtained using a high-fidelity sequencing technology to distinguish mutational signal from noise (Sloan *et al.* 2018). For example, Duplex Sequencing is an Illumina-based method in which unique molecular identifiers (UMIs) are included in adaptors and attached to both ends of DNA fragments before library amplification (Schmitt *et al.* 2012; Kennedy *et al.* 2014). After sequencing, the UMIs are used to cluster families of reads that originated from each strand of a given DNA fragment so that a double-stranded consensus sequence can be created that is virtually error free ($< 5\times10^{-8}$ errors per base pair; Kennedy *et al.* 2014).

Our goal in this study was to test if the pattern of local mutation rate variation across a genome depends on environmental effects on gene expression levels. We also wanted to determine whether low-frequency somatic mutations in plant tissues could provide a robust signal for addressing this type of question. Therefore, we perturbed gene expression by growing *Arabidopsis* under different temperatures. We identified differentially expressed (DE) genes with RNA-seq, which we then targeted for low-frequency somatic mutation detection using Duplex Sequencing coupled with hybrid capture. We included mutant lines *msh2* and *ung*, which respectively lack mismatch repair (MMR) and base excision repair (BER) capabilities, in order to understand how repair mechanisms may drive biased mutation accumulation (Cordoba-Canero

et al. 2010; Belfield *et al.* 2018). We also included *hsp70-16* mutant lines, which are deficient for a key heat shock protein, as a means to endogenously manipulate gene expression and potentially interact with our temperature treatment (Ran *et al.* 2020). As expected, we found significant increases in variant frequencies in the MMR deficient lines. In wild type (WT) lines and other mutant lines, measured mutation frequencies were too low to quantify relationships between mutation rates and environment-specific gene expression levels. Therefore, our results support the conclusion that earlier estimates of somatic variant frequencies were inflated (Monroe *et al.* 2023a; Wang *et al.* 2023) and indicate that much larger datasets will be needed to test for environment-specific changes in mutation biases.

RESULTS

To test if environment specific changes in gene expression impact mutation, we performed mutation detection on a targeted set of *Arabidopsis* genes that were DE in plants grown at 20°C vs. 30°C. We first generated and analyzed RNA-seq data to identify genes in six categories: 1) increased expression at 30°C compared to 20°C in WT plants, 2) increased expression at 20°C compared to 30°C in WT plants, 3) constitutively high expression in WT plants at both 20°C and 30°C, 4) constitutively low expression in WT plants at both 20°C and 30°C, 5) genes that had increased expression at 30°C vs. 20°C in WT plants (like category 1) and also had an interaction between WT and *hsp70-16*, and 6) genes that had increased expression at 30°C vs. 20°C in WT plants (like category 2) and also had an interaction between WT and *hsp70-16* (Table S4.1). The sequences of the DE genes were used to create a custom probe-set for hybrid capture of Duplex Sequencing libraries.

Duplex Sequencing coverage of the genes and 250 bp of flanking sequence in the probe-set ranged from 74.7× to 109.4× (Fig. S4.1), and the average probe-set coverage across all libraries was 193.1-fold higher than the genome background. In total, we obtained 1.89 Gb of Duplex Sequencing coverage of our region of interest across the 24 libraries (Table S4.2)

We then looked for the presence of single nucleotide variants (SNVs) and short indels within the 339 genes covered in the probe-set. Mutant alleles already present in the parents of the assayed sets of full-sib plants have the potential to bias estimates of *de novo* mutation frequencies but should be readily identifiable. For a homozygous parent, they would be present in all Duplex Sequencing reads of all the replicates of a given genotype. For a heterozygous parent, they would segregate in a 1:2:1 Mendelian ratio and account for roughly 50% of the reads for all replicates of a given genotype (as each replicate represents a pool of five sibling plants). We identified just three apparent fixed SNVs (Table S4.3), which were removed for downstream analyses. In contrast, we identified 41 fixed indels, over half of which were in the *msh2* background (Table S4.4). One gene (AT5G39190) had five sites that appeared to be segregating SNVs in all 24 replicates. We suspected this might be caused by a cryptic gene duplication which was not captured in the TAIR 10.2 reference genome (Jaegle *et al.* 2023). Indeed, when we realigned the reads to the improved Col-CC genome (Reiser *et al.* 2023), the mutation calls in AT5G39190 were absent. As such, reads mapping to AT5G39190 were disregarded in downstream analyses. The rest of the SNVs we identified were unique to each replicate and all were present at a frequency of no more than 17.64% (the average variant frequency across all mutations was 2.27%), suggesting that these are low frequency somatic variants that arose during the experiment and were present in a subset of the sampled vegetative tissue.

Among the six WT biological replicates, we detected a single indel and just six SNVs, one in each replicate (Fig. 4.1). As such, there was very limited statistical power to test for the effects of temperature or expression level on mutation frequency in WT plants. Similarly, we detected few or no SNVs and indels in the *hsp70-16* and the *ung* mutant lines (Fig. 4.1). In contrast, variant frequencies were significantly elevated in the *msh2* mutant lines (compared to WT plants), where we detected 271 indels and 180 SNVs (Fig. 4.1; two-way ANOVA with Tukey's test, $p < 0.0001$). The mutations in the *msh2* lines were distributed relatively evenly across the temperature treatments, as we found that temperature did not influence either SNV or indel frequency (Fig. 4.1; two-way ANOVA, $p = 0.99$). In the *msh2* lines, deletions were 8.5-fold more common than insertions (Table S4.5; two-way ANOVA, $p < 0.0001$). We observed significant differences among SNV classes in *msh2* SNV spectrum (Fig. 4.2; two-way ANOVA, $p < 0.0001$), which was dominated by CG→TA transitions. The next most common types of substitutions were AT→GC transitions and CG→AT transversions. We compared the *msh2* mutation frequencies in the constitutively lowly expressed (group 3 in Table S4.1) vs constitutively highly expressed (group 4 in Table S4.1) genes and found no significant differences (paired t-test; Table S4.6), though we did observe a trend towards higher indel frequencies in constitutively highly expressed genes at 30°C. We did not analyze the SNV spectra or indel bias in WT, *ung*, *or hsp70-16* lines because the small number of sampled mutations precluded a statistically meaningful comparison.

DISCUSSION

In this study we took a novel approach to studying plant mutation by utilizing high fidelity Duplex Sequencing to measure low-frequency somatic variants in a targeted region of the

A. thaliana nuclear genome. Variants in unopened floral bud tissue of WT plants were present at very low frequencies (Fig. 4.1), which were near the detection threshold of Duplex Sequencing (Kennedy *et al.* 2014; Wu *et al.* 2020). Although we did not have enough power to address our prediction that increases in gene expression would correlate with decreases in mutation rates in WT plants, the results are nonetheless of interest given recent debates about the frequency of somatic mutations in plant tissues (Monroe *et al.* 2022; Liu and Zhang 2022; Monroe *et al.* 2023a; Wang *et al.* 2023; Monroe *et al.* 2023b). Our results support the conclusion that the high error rate of Illumina short-read sequencing makes it difficult to reliably discern sequencing errors from extremely rare WT somatic mutations. That said, we are skeptical of directly comparing the variant frequencies we measured in unopened floral buds with those obtained in differentiated leaves (Monroe *et al.* 2022, 2023a) given recent evidence showing substantial variation in somatic mutation rates depending on plant tissue (Goel *et al.* 2024).

We also surveyed variant frequencies in *ung* mutant plants and did not observe a difference between WT and *ung* lines. Given that *ung* plants have previously been shown to accumulate more uracil in DNA (presumably to the loss of base-excision repair activity on deaminated cytosines) than WT plants (Cordoba-Canero *et al.* 2010), we interpret the lack of a difference between WT and *ung* lines as evidence that actual WT mutation frequencies may be below the detection threshold of Duplex Sequencing. However, it is also possible that the similarly low mutation rates in WT and *ung* reflect the lack of a true biological difference, which may be possible if redundant pathways exist that prevent uracils in DNA from becoming CG→TA transitions.

In contrast, we found significantly elevated variant frequencies in *msh2* mutants compared to WT lines (Fig. 4.1). MSH2 is known to function in mismatch repair (MMR) and

mutation accumulation experiments with *msh2* mutant lines have established that the germline SNV rate is 132 to 204-fold greater than the WT SNV rate (Ossowski *et al.* 2010; Jiang *et al.* 2014; Belfield *et al.* 2018). Here, we found that the average *msh2* SNV frequency was 27-fold greater than the average WT SNV frequency (Fig. 4.1). Though somatic variant frequencies measured with Duplex Sequencing are not directly comparable to germline mutation rates assayed with mutation accumulation experiments, the smaller magnitude of the difference between *msh2* vs. WT in our dataset may be interpreted as further evidence that the actual WT variant frequency is beneath the detection threshold of Duplex Sequencing. Alternatively, the smaller difference between WT and *msh2* reported here could be evidence that MMR is particularly important for buffering against mutation in germline plant tissues, which is supported by elevated expression of *MSH2* and other mismatch repair genes in meristematic tissues (Klepikova *et al.* 2016).

Variant frequencies in the *msh2* mutant lines showed no significant difference in plants grown at 20°C vs. 30°C. This finding contrasts with a recent mutation accumulation study that found elevated germline mutation rates in WT plants grown at 29°C compared to those grown at 23°C (Belfield *et al.* 2021) and another study that documented increases at 28°C and 32°C compared to 23°C (Lu *et al.* 2021). One potential explanation of this result is that heat stress may be mutagenic in WT plants *because* it impairs MMR since in the absence of MMR there is no apparent heat effect. However, this interpretation would be at odds with the fact that the genome-wide distribution of mutations in the heat-stressed plants mirrors the distribution of WT plants grown at standard temperature, not of mismatch repair mutants (see Fig. 3 of (Belfield *et al.* 2021). The Duplex Sequencing variant frequencies in the *msh2* mutant lines also did not vary significantly between lowly expressed vs. highly expressed genes at either 20°C or 30°C (Fig.

4.1). This result is consistent with the model that MMR provides special protection to actively transcribed genes (Belfield *et al.* 2018; Huang *et al.* 2018; Huang and Li 2018). However, we present this interpretation cautiously in the absence of WT data to test for an impact of expression when MMR is functional.

In summary, we took a novel approach to studying plant mutations by using Duplex Sequencing and hybrid capture to obtain a highly accurate snapshot of somatic variants in targeted regions of the *A. thaliana* genome. We designed our experiment to test if environmental conditions alter mutation rates in a gene-specific fashion. However,
the low rate of mutations in WT plants prevented testing for how expression levels impact mutation rates. Nonetheless, the link between increased expression and decreased mutation in plants is well documented (Oztas *et al.* 2018; Monroe *et al.* 2022; Quiroz *et al.* 2023), as is the fact that gene expression is environmentally determined (Richards *et al.* 2012), so by logical extension environmental conditions must drive mutation rates and related fitness consequences. However, whether the magnitude of such an effect is biologically meaningful in shaping mutation and evolution remains an important, unanswered question. Though mutation accumulation and parent-offspring sequencing are time- and resource-intensive experiments, they are both increasingly feasible due to continued declines in the cost of DNA sequencing (Ossowski *et al.* 2010; Weng *et al.* 2019; Monroe *et al.* 2022). Conducting such experiments under contrasting environments (Jiang *et al.* 2014; Belfield *et al.* 2021; Lu *et al.* 2021) to measure the correlation between expression and mutation seems to be the key to understanding how environments impact the types of mutations that organisms accumulate.

MATERIALS AND METHODS

All plants were grown in environmentally controlled growth chambers (75% humidity) under a long-day photoperiod (16 hrs light, 8 hrs dark) with irradiance of 185 μmol m^{-2} sec^{-1} at constant temperatures (either 20°C or 30°C, as specified below). Prior to planting, seeds were stratified for 5 days in sterile ddH20. *Arabidopsis thaliana* ecotype Col-0 was used as the WT line. Existing mutant lines were obtained from the Arabidopsis Biological Resource Center (Table S4.7) and seedlings were screened with allele-specific PCR markers to identify plants that were homozygous for the mutant alleles used in this study (*msh2*, *ung, hsp70-16;* Table S4.8).

Sibling plants (roughly 35 for each genotype and each temperature treatment) were planted in 2.5-inch pots. Both temperature treatments were initiated in chambers (Convarion models PGR15 (20°C) and PGCFLEX (30°C)) at 20°C because elevated ambient temperatures (30°C) can inhibit seed germination (Silva-Correia *et al.* 2014). After 5 days, the temperature was turned up for the 30°C treatment and kept at 20°C for the other treatment. When the plants had reached stage 6.5 of development (where ~50 % of flowers have opened) (Boyes *et al.* 2001), we performed DNA and RNA extractions on unopened floral buds from laterally branching florets. The 30°C plants reached developmental stage 6.5 at 31 days while the 20°C plants reached developmental stage 6.5 at 41 days, consistent with faster plant development at elevated ambient temperatures (Silva-Correia *et al.* 2014).

For the RNA extractions, plant material was collected from the unopened floral buds of 3 laterally branching florets from 3 WT and 3 *hsp70-16* plants in each temperature treatment. The harvested tissues were immediately placed into liquid nitrogen and homogenized for 10 seconds at 30 beats/sec with the Qiagen TissueLyser, before being processed with the Qiagen RNeasy Plant Mini Kit, according to manufacturer's instructions. The RNA samples were then sent to Novogene and RNA-Seq libraries were made using the NEBNext Ultra II Directional RNA

Library Prep Kit with the NEBNext Poly(A) mRNA Magnetic Isolation Module. The RNA-Seq libraries were sequenced on a NovaSeq 6000 using the PE150 strategy to generate 29 to 54 million read pairs per library (see Table S4.9).

Tissue was harvested for DNA sequencing and mutation detection at the same time as the tissue for RNA extraction, from siblings of the plants used for RNA extraction. For each replicate in the DNA extractions, plant material was pooled from 5 siblings from the unopened floral buds of 3 laterally branching florets from 5 plants per each replicate, with 3 replicates per genotype (WT, *hsp70-16, msh2, ung*) per temperature treatment. The floret tissue was homogenized for 10 seconds at 30 beats/sec with the Qiagen TissueLyser, before being processed with the DNeasy Plant Mini Kit from Qiagen.

The RNA-seq reads were analyzed to detect DE genes at 20°C vs. 30°C. First, the adaptors were removed with Cutadapt version 4.0 with Python 3.9.16 (Martin 2011). Then the reads were mapped to the TAIR10.2 reference genome with HISAT2 (version 2.2.1; (Kim *et al.* 2019). Read counts were generated with HTSeq-count version 2.0.2 (Anders *et al.* 2014), and DESeq2 models (Love *et al.* 2014) were implemented to identify genes that were differentially expressed or constitutively highly or lowly expressed.

We created a custom probe-set to enrich the sequences of DE genes via hybrid capture so that we could perform mutation detection with Duplex Sequencing. We sent the sequences of 400 DE genes (plus 250 nt of flanking sequence on the end of each gene) to the probe design team at Arbor Bioscience, which flagged 61 of the genes as unsuitable for hybrid capture because they were > 25 % soft-masked for repeats in a BLAST search against the Arbor Biosciences eudicot database. The remaining 339 genes and flanking sequences spanned a total length of 855,123 nt. Sets of 80-nt probes were 2× tiled across the target sequence at approximately every 40 nt. The

probes were biotinylated so that probe-bound library molecules can be captured with streptavidin-coated magnetic beads.

We created Duplex Sequencing libraries from the 24 DNA samples (3 replicates × 4 genotypes × 2 temperature treatments), following our previously described library preparation protocols (Wu *et al.* 2020; Waneka *et al.* 2021), except that in this case the amount of input DNA was increased to 500 ng because the target sequence comprises a small fraction (< 1%) of the total-cellular DNA sample. Once DNA samples had been fragmented via ultrasonication, end-repaired, A-tailed, adaptor-ligated, and treated with a cocktail of damage removal enzymes (Wu *et al.* 2020), we amplified 0.73 ng of DNA (per reaction) for 13 PCR cycles with New England Biolabs Q5 High-Fidelity Polymerase and dual-indexed primers. We then created 3 pools by combining 350 ng of each amplified library as the Arbor Biosciences hybrid-capture reactions have enough capacity for 8 libraries in each pool. We performed the overnight hybrid-capture reaction at 65°C, according to the manufacturer's instructions (Arbor Biosciences MyBaits Kit Manual v. 5.02). We assessed enrichment efficiency and library concentrations through qPCR (as previously described; (Waneka *et al.* 2021)) before amplifying the enriched pools for an additional 9 cycles to obtain sufficient library amounts for sequencing.

Duplex Sequencing libraries were sequenced with PE150 reads on an Illumina NovaSeq 6000 S4 Lane (Novogene) to generate 87 to 123 million read pairs per library (Table S4.10). Processing of the Duplex Sequencing reads to was performed with our previously described pipeline (Wu *et al.* 2020), which trimmed adaptor sequences, created duplex consensus sequences based on the presence of shared barcodes, mapped the consensus sequences to the entire TAIR10.2 reference genome. Each duplex consensus sequences is composed of at least 6 Illumina reads (at least 3 originating from each strand of a DNA fragment). Alignment files were

then parsed to identify duplex consensus sequences that contain SNVs and short indels. Since Duplex Sequencing is highly accurate ($< 5\times10^{-8}$ errors per base pair; Kennedy *et al.* 2014) we require just a single duplex consensus to support a putative mutation. Comparisons of coverage in the probe-set vs. outside the probe-set were performed with Samtools version 1.6 (Li *et al.* 2009). For variant frequency calculations, we excluded the first or last 10 bps of a read because we have previously identified elevated mutation frequencies at read ends (Wu *et al.* 2020).

DATA AVAILABILITY

The raw reads are available via the NCBI Sequence Read Archive under accessions SRR27564102-SRR27564113 (RNA-seq libraries) and SRR27693810-SRR27693833 (Duplex Sequencing libraries). Duplex Sequencing datasets were processed with a previously published pipeline (https://github.com/dbsloan/duplexseq) (Wu *et al.* 2020).

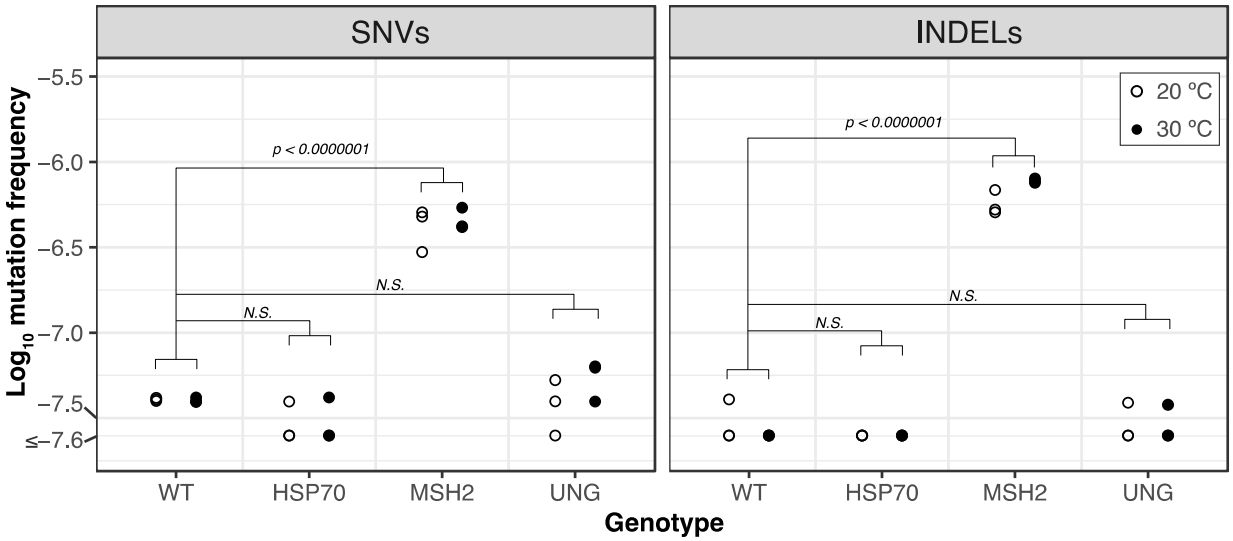

Figure 4.1. Mutation frequencies in WT vs mutant lines at 20°C and 30°C. Log_{10} mutation frequencies for single nucleotide variants (SNVs) and insertions/deletions (INDELs) calculated as the number of events (SNVs or INDELs) divided by the duplex sequencing coverage of the probe-set. A floor of 2.5×10^{-8} was applied to the y-axis for data visualization. *P-values* are from a Tukey's test on a two-way ANOVA performed in R with the emmeans package (version 1; (Lenth *et al.* 2021).

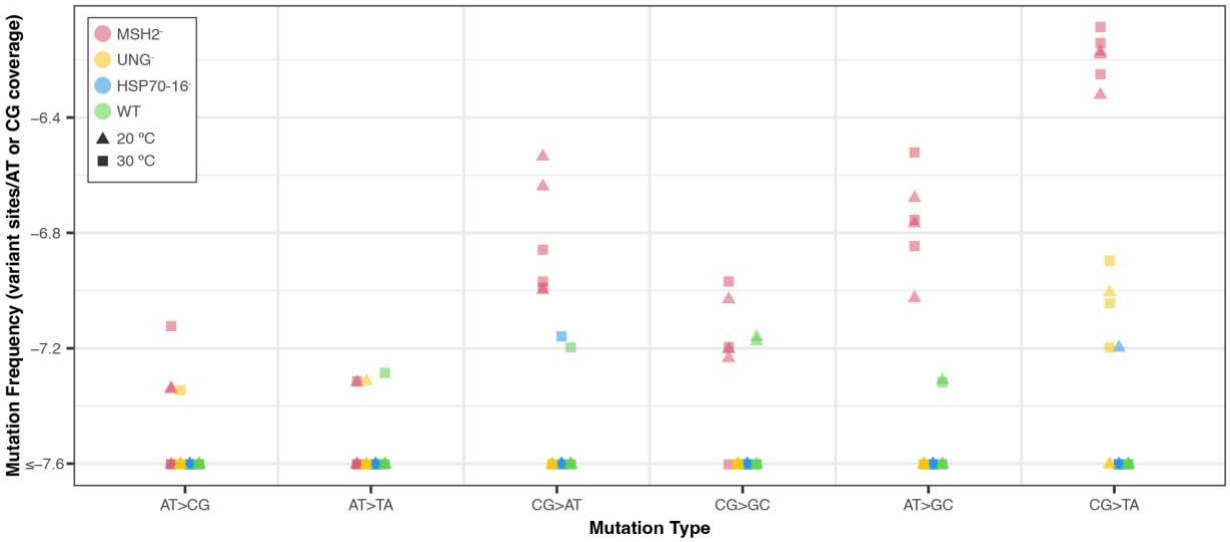

Figure 4.2. Mutation spectrum for WT and mutant plants at 20 °C and 30 °C. Log_{10} mutation frequencies for different types of single nucleotide variants were calculated as the number of events divided by the nucleotide-specific duplex sequencing coverage of the probe-set. A floor of 2.5×10^{-8} was applied to the y-axis for data visualization.

LITERATURE CITED

Anders, S., P. T. Pyl, and W. Huber, 2014 HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166–169.

Bailey, S. F., L. A. Alonso Morales, and R. Kassen, 2021 Effects of Synonymous Mutations beyond Codon Bias: The Evidence for Adaptive Synonymous Substitutions from Microbial Evolution Experiments. Genome Biol. Evol. 13:.

Belfield, E. J., C. Brown, Z. J. Ding, L. Chapman, M. Luo *et al.*, 2021 Thermal stress accelerates Arabidopsis thaliana mutation rate. Genome Res. 31: 40–50.

Belfield, E. J., Z. J. Ding, F. J. C. Jamieson, A. M. Visscher, S. J. Zheng *et al.*, 2018 DNA mismatch repair preferentially protects genes from mutation. Genome Res. 28: 66–74.

Boyes, D. C., A. M. Zayed, R. Ascenzi, A. J. McCaskill, N. E. Hoffman *et al.*, 2001 Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. Plant Cell 13: 1499–1510.

Cordoba-Canero, D., E. Dubois, R. R. Ariza, M.-P. Doutriaux, and T. Roldán-Arjona, 2010 Arabidopsis uracil DNA glycosylase (UNG) is required for base excision repair of uracil and increases plant sensitivity to 5-fluorouracil. J. Biol. Chem. 285: 7475–7483.

Eyre-Walker, A., and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8: 610–618.

Goel, M., J. A. Campoy, K. Krause, L. C. Baus, A. Sahu *et al.*, 2024 The majority of somatic mutations in fruit trees are layer-specific. bioRxiv 2024.01.04.573414.

Gonzalez-Perez, A., R. Sabarinathan, and N. Lopez-Bigas, 2019 Local Determinants of the Mutational Landscape of the Human Genome. Cell 177: 101–114.

Grantham, R., C. Gautier, M. Gouy, R. Mercier, and A. Pavé, 1980 Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8: r49–r62.

Gundry, M., and J. Vijg, 2012 Direct mutation analysis by high-throughput sequencing: from germline to low-abundant, somatic variants. Mutat. Res. 729: 1–15.

Hershberg, R., and D. A. Petrov, 2008 Selection on Codon Bias.

Huang, Y., L. Gu, and G.-M. Li, 2018 H3K36me3-mediated mismatch repair preferentially protects actively transcribed genes from mutation. J. Biol. Chem. 293: 7811–7823.

Huang, Y., and G.-M. Li, 2018 DNA mismatch repair preferentially safeguards actively transcribed genes. DNA Repair 71: 82–86.

Jaegle, B., R. Pisupati, L. M. Soto-Jiménez, R. Burns, F. A. Rabanal *et al.*, 2023 Extensive sequence duplication in Arabidopsis revealed by pseudo-heterozygosity. Genome Biol. 24: 44.

Jiang, C., A. Mithani, E. J. Belfield, R. Mott, L. D. Hurst *et al.*, 2014 Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. Genome Res. 24: 1821–1829.

Jinks-Robertson, S., and A. S. Bhagwat, 2014 Transcription-associated mutagenesis. Annu. Rev. Genet. 48: 341–359.

Kennedy, S. R., M. W. Schmitt, E. J. Fox, B. F. Kohrn, J. J. Salk *et al.*, 2014 Detecting ultralow-frequency mutations by Duplex Sequencing. Nat. Protoc. 9: 2586–2606.

Kim, N., A. L. Abdulovic, R. Gealy, M. J. Lippert, and S. Jinks-Robertson, 2007 Transcription-associated mutagenesis in yeast is directly proportional to the level of gene expression and influenced by the direction of DNA replication. DNA Repair 6: 1285–1296.

Kim, D., J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, 2019 Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37: 907–915.

Klepikova, A. V., A. S. Kasianov, E. S. Gerasimov, M. D. Logacheva, and A. A. Penin, 2016 A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant J. 88: 1058–1070.

Lenth, R., H. Singmann, J. Love, P. Buerkner, and M. Herve, 2021 Emmeans: Estimated marginal means, aka least-squares means. R Package Version 1 (2018). Preprint at.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Liu, H., and J. Zhang, 2022 Is the Mutation Rate Lower in Genomic Regions of Stronger Selective Constraints? Mol. Biol. Evol. 39:.

Love, M. I., W. Huber, and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15: 550.

Lu, Z., J. Cui, L. Wang, N. Teng, S. Zhang *et al.*, 2021 Genome-wide DNA mutations in Arabidopsis plants after multigenerational exposure to high temperatures. Genome Biol. 22: 160.

Luria, S. E., and M. Delbrück, 1943 Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics 28: 491–511.

Lynch, M., M. S. Ackerman, J.-F. Gout, H. Long, W. Sung *et al.*, 2016 Genetic drift, selection and the evolution of the mutation rate. Nat. Rev. Genet. 17: 704–714.

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10–12.

Monroe, J. G., K. D. Murray, W. Xian, T. Srikant, P. Carbonell-Bejerano *et al.*, 2023a Reply to: Re-evaluating evidence for adaptive mutation rate variation. Nature 619: E57–E60.

Monroe, J. G., T. Srikant, P. Carbonell-Bejerano, C. Becker, M. Lensink *et al.*, 2023b Author Correction: Mutation bias reflects natural selection in Arabidopsis thaliana. Nature 620: E13.

Monroe, J. G., T. Srikant, P. Carbonell-Bejerano, C. Becker, M. Lensink *et al.*, 2022 Mutation bias reflects natural selection in Arabidopsis thaliana. Nature 602: 101–105.

Moore, L., A. Cagan, T. H. H. Coorens, M. D. C. Neville, R. Sanghvi *et al.*, 2021 The mutational landscape of human somatic and germline cells. Nature 597: 381–386.

Ossowski, S., K. Schneeberger, J. I. Lucas-Lledó, N. Warthmann, R. M. Clark *et al.*, 2010 The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327: 92–94.

Oztas, O., C. P. Selby, A. Sancar, and O. Adebali, 2018 Genome-wide excision repair in Arabidopsis is coupled to transcription and reflects circadian gene expression patterns. Nat. Commun. 9: 1503.

Quiroz, D., M. Lensink, D. J. Kliebenstein, and J. G. Monroe, 2023 Causes of Mutation Rate Variability in Plant Genomes. Annu. Rev. Plant Biol. 74: 751–775.

Ran, X., X. Chen, L. Shi, M. Ashraf, F. Yan *et al.*, 2020 Transcriptomic insights into the roles of HSP70-16 in sepal's responses to developmental and mild heat stress signals. Environ. Exp. Bot. 179: 104225.

Reiser, L., E. Bakker, S. Subramaniam, X. Chen, and S. Sawant, 2023 The Arabidopsis Information Resource in 2024. bioRxiv.

Richards, C. L., U. Rosas, J. Banta, N. Bhambhra, and M. D. Purugganan, 2012 Genome-wide patterns of Arabidopsis gene expression in nature. PLoS Genet. 8: e1002662.

Sanchez-Contreras, M., M. T. Sweetwyne, B. F. Kohrn, K. A. Tsantilas, M. J. Hipp *et al.*, 2021 A replication-linked mutational gradient drives somatic mutation accumulation and influences germline polymorphisms and genome composition in mitochondrial DNA. Nucleic Acids Res. 49: 11103–11118.

Satake, A., R. Imai, T. Fujino, S. Tomimoto, K. Ohta *et al.*, 2023 Somatic mutation rates scale with time not growth rate in long-lived tropical trees. eLife.

Schmitt, S., P. Heuret, V. Troispoux, M. Beraud, J. Cazal *et al.*, 2023 Plant mutations: slaying beautiful hypotheses by surprising evidence. bioRxiv 2023.06.05.543657.

Schmitt, M. W., S. R. Kennedy, J. J. Salk, E. J. Fox, J. B. Hiatt *et al.*, 2012 Detection of ultra-rare mutations by next-generation sequencing. Proc. Natl. Acad. Sci. U. S. A. 109: 14508–14513.

Seplyarskiy, V., E. M. Koch, D. J. Lee, J. S. Lichtman, H. H. Luan *et al.*, 2023 A mutation rate model at the basepair resolution identifies the mutagenic effect of polymerase III transcription. Nat. Genet. 55: 2235–2242.

Silva-Correia, J., S. Freitas, R. M. Tavares, T. Lino-Neto, and H. Azevedo, 2014 Phenotypic analysis of the Arabidopsis heat stress response during germination and early seedling development. Plant Methods 10: 7.

Sloan, D. B., A. K. Broz, J. Sharbrough, and Z. Wu, 2018 Detecting Rare Mutations and DNA Damage with Sequencing-Based Methods. Trends Biotechnol. 36: 729–740.

Staunton, P. M., A. J. Peters, and C. Seoighe, 2023 Somatic mutations inferred from RNA-seq data highlight the contribution of replication timing to mutation rate variation in a model plant. Genetics 225:.

Supek, F., and B. Lehner, 2017 Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. Cell 170: 534-547.e23.

Tatsumoto, S., Y. Go, K. Fukuta, H. Noguchi, T. Hayakawa *et al.*, 2017 Direct estimation of de novo mutation rates in a chimpanzee parent-offspring trio by ultra-deep whole genome sequencing. Sci. Rep. 7: 13561.

Waneka, G., J. M. Svendsen, J. C. Havird, and D. B. Sloan, 2021 Mitochondrial mutations in Caenorhabditis elegans show signatures of oxidative damage and an AT-bias. Genetics 219:.

Wang, L., A. T. Ho, L. D. Hurst, and S. Yang, 2023 Re-evaluating evidence for adaptive mutation rate variation. Nature 619: E52–E56.

Weng, M.-L., C. Becker, J. Hildebrandt, M. Neumann, M. T. Rutter *et al.*, 2019 Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in Arabidopsis thaliana. Genetics 211: 703–714.

Wu, Z., G. Waneka, A. K. Broz, C. R. King, and D. B. Sloan, 2020 MSH1 is required for maintenance of the low mutation rates in plant mitochondrial and plastid genomes. Proc. Natl. Acad. Sci. U. S. A. 117: 16448–16455.

Zhang, G., 2023 The mutation rate as an evolving trait. Nat. Rev. Genet. 24: 3.

Zhang, J., and J.-R. Yang, 2015 Determinants of the rate of protein sequence evolution. Nat. Rev. Genet. 16: 409–420.

APPENDIX I: ADDITIONAL PUBLICATIONS

Several publications were produced as part of this PhD which are not included in this dissertation. These projects address similar questions and utilize some overlapping techniques to the dissertation research. This appendix summarizes and references these studies so they may be easily found and appreciated as a part of the overall contributions of this dissertation.

Given how rare plant organelle mutations are, we were concerned that it would be difficult to optimize the Duplex Sequencing protocol (used in dissertation chapters 2 and 4) in plants because we would struggle to know if low mutational measurements were arising biologically or as a shortcoming of the sequencing technique implementation. We therefore implemented Duplex Sequencing in biological systems with higher baseline mutation rates, which served as positive controls for optimizing the Duplex Sequencing protocol. The first of these positive controls used Duplex Sequencings to measure *de novo* mutations in two bacterial endosymbionts of leafhoppers and other sap feeding insects (Waneka *et al.* 2021a). We found over 100-fold more mutations in the endosymbiont which has also experienced more evolutionary instability in the form of extinction and replacement across independent insect lineages. The second positive control used Duplex Sequencing to detect *de novo* mitochondrial genome mutations in the nematode *Caenorhabditis elegans* (Waneka *et al.* 2021b). We found a relatively high frequency of CG>AT transversions, which suggests that oxidative damage may play an important role in driving *C. elegans* mitochondrial mutations.

We additionally used Duplex Sequencing in two other publications focused on plant organellar genomes. The first of these studies was performed in *Arabidopsis* and is methodologically almost identical to the Duplex Sequencing portion of chapter two of this

dissertation except that we targeted a different panel of organelle maintenance genes (Wu *et al.* 2020b). This study provided the first genetic evidence that *MSH1* is involved in maintaining low point mutations in organelle genomes, a result which is further explored in chapter 2 of this dissertation. The second study used Duplex Sequencing to compare *de novo* organellar genome mutations in several species in the genus *Silene* (Broz *et al.* 2021). We found stark differences in the overall mitochondrial mutation frequency between the different species, supporting the hypothesis that elevated rates of mitochondrial evolution in some *Silene* lineages result from increased mutation rates.

Finally, there were several additional studies that also focused on plant mutation but did not utilize Duplex Sequencing. The first was a project that used the *Arabidopsis* 1001 Genomes dataset to analyze organelle genome divergence at the population level (Wu *et al.* 2020a). The other publications were led by Peter Fields and provide genomic resources which are valuable for studying mutation, first in the form of an assembly of a large insertion of mitochondrial DNA in the *Arabidopsis* nuclear genome (Fields *et al.* 2022) and second in the assembly of the *Silene conica* nuclear genome (Fields *et al.* 2023).

CITATIONS

Broz A. K., G. Waneka, Z. Wu, M. Fernandes Gyorfy, and D. B. Sloan, 2021 Detecting de novo mitochondrial mutations in angiosperms with highly divergent evolutionary rates. Genetics 218. https://doi.org/10.1093/genetics/iyab039

Fields P. D., G. Waneka, M. Naish, M. C. Schatz, I. R. Henderson, *et al.*, 2022 Complete Sequence of a 641-kb Insertion of Mitochondrial DNA in the Arabidopsis thaliana Nuclear Genome. Genome Biol. Evol. 14. https://doi.org/10.1093/gbe/evac059

Fields P. D., M. M. Weber, G. Waneka, A. K. Broz, and D. B. Sloan, 2023 Chromosome-Level Genome Assembly for the Angiosperm Silene conica. Genome Biol. Evol. 15. https://doi.org/10.1093/gbe/evad192

Waneka G., Y. M. Vasquez, G. M. Bennett, and D. B. Sloan, 2021a Mutational Pressure Drives Differential Genome Conservation in Two Bacterial Endosymbionts of Sap-Feeding Insects. Genome Biol. Evol. 13. https://doi.org/10.1093/gbe/evaa254

Waneka G., J. M. Svendsen, J. C. Havird, and D. B. Sloan, 2021b Mitochondrial mutations in Caenorhabditis elegans show signatures of oxidative damage and an AT-bias. Genetics 219. https://doi.org/10.1093/genetics/iyab116

Wu Z., G. Waneka, and D. B. Sloan, 2020a The Tempo and Mode of Angiosperm Mitochondrial Genome Divergence Inferred from Intraspecific Variation in Arabidopsis thaliana. G3 10: 1077–1086.

Wu Z., G. Waneka, A. K. Broz, C. R. King, and D. B. Sloan, 2020b MSH1 is required for maintenance of the low mutation rates in plant mitochondrial and plastid genomes. Proc. Natl. Acad. Sci. U. S. A. 117: 16448–16455.

APPENDIX II: SUPPLEMENTAL TABLES AND FIGURES

Table S2.1. Mutant lines used in this study and primers

Gene	Salk line (all from ABRC)	Locus	Forward Primer Wild	Forward Primer Mutant	Reverse Primer Wild and Mutant
radA	SALK_097880	AT5G50340	TTTCACTTATCGAGCCAGAGC	ATTTTGCCGATTTCGGAAC	ATGCCATAATGCTTTTTGCTG
recA1	SALK_072979	AT1G79050	TAGGGTGAGATTGGAATGCAG	ATTTTGCCGATTTCGGAAC	AAGAGCTGCTGCTCATCAAAG
recA3	SALK_146388	AT3G10140	CGTTTGGTCAGTTGAAGCTTC	ATTTTGCCGATTTCGGAAC	CTCCACAAGTCACTTCTTCGG
osb2	SALK_061852	At4g20010	AGCGTGAAAGGTGAGACGTT	ATTTTGCCGATTTCGGAAC	GGGAAATAACAGTACCAGCCC
why2	SALK_118900	AT1G71260	CAGGAAGTCACTGTCAGTTAAGC	ATTTTGCCGATTTCGGAAC	ACCCATGATTTAGAAGTCTTAGAGAGG

Table S2.2. Duplex read-pairs and organellar genome coverage

Sample	Count of read-pairs (2x150)	organellar coverage per bp
mitochondrial_rada_mut_1	75863350	297.9
mitochondrial_rada_mut_2	64771627	281.8
mitochondrial_rada_mut_3	139195192	803.4
mitochondrial_rada_wild_1	70847671	127.8
mitochondrial_rada_wild_2	61998713	246.9
mitochondrial_rada_wild_3	127161861	816.4
mitochondrial_reca3_mut_1	43472074	94.2
mitochondrial_reca3_mut_2	44312097	229.3
mitochondrial_reca3_mut_3	59128403	497.3
mitochondrial_reca3_wild_1	62354817	238.2
mitochondrial_reca3_wild_2	54311915	183.2
mitochondrial_reca3_wild_3	40734051	144.8
mitochondrial_why2_mut_1	63375069	338.0
mitochondrial_why2_mut_2	76906783	284.9
mitochondrial_why2_mut_3	76221972	292.6
mitochondrial_why2_wild_1	68231709	279.8
mitochondrial_why2_wild_2	81396138	379.9
mitochondrial_why2_wild_3	86880259	408.4
plastid_osb2_mut_1	47505179	1176.6
plastid_osb2_mut_2	54307516	870.8
plastid_osb2_mut_3	59415250	898.6
plastid_osb2_wild_1	59542949	1132.8
plastid_osb2_wild_2	69408084	889.7
plastid_osb2_wild_3	67727784	668.6
plastid_rada_mut_1	76116128	1174.4
plastid_rada_mut_2	68615282	871.7
plastid_rada_mut_3	45985626	1068.7
plastid_rada_wild_1	53480887	234.2
plastid_rada_wild_2	46684396	954.5
plastid_rada_wild_3	46190084	776.4
plastid_reca1_mut_1	66804365	543.7
plastid_reca1_mut_2	38396319	594.8
plastid_reca1_mut_3	30645358	299.3
plastid_reca1_wild_1	37377457	598.2
plastid_reca1_wild_2	32420159	543.5
plastid_reca1_wild_3	33351491	331.1

Table S2.3. Results from Kruskal-Wallis test comparing SNV frequencies among genomic regions in WT and *msh1* mutant data from Wu *et al.,* (2020)

Sample	Kruskal-Wallis chi-squared value	p-value
msh1 mitochondria	6.03	0.19
msh1 plastid	5.47	0.24
WT mitochondria	6.66	0.15
WT plastid	11.35	0.02

Table S2.4. Oxford Nanopore sequencing yields for each of the three runs

Sample	Sequencing run	read count	long reads count (>500bp)	total yield (Mb)
plastid_recA1_wild_1	1	224238	131719	412.21
plastid_recA1_mut_1	1	82051	39041	121.45
plastid_recA1_mut_2	1	74341	32851	130.77
plastid_recA1_mut_3	1	72833	41137	167.30
plastid_radA_wild_1	1	127499	85335	307.34
plastid_radA_mut_1	1	111390	72395	297.02
plastid_radA_mut_3	1	186393	119090	540.66
plastid_osb2_wild_3	1	101793	66081	239.62
plastid_osb2_mut_1	1	143806	92534	407.94
plastid_osb2_mut_2	1	103151	74649	349.70
plastid_osb2_mut_3	1	109492	63761	260.72
plastid_msh1_wild_3	1	45501	29533	126.16
plastid_msh1_mut_1	1	36518	24441	111.72
plastid_msh1_mut_2	1	46533	26330	97.91
plastid_msh1_mut_3	1	47757	32369	153.13
mitochondrial_recA3_wild_1	2	8481	6019	56.73
mitochondrial_recA3_mut_1	2	1442	813	8.70
mitochondrial_recA3_mut_2	2	20256	14861	101.19
mitochondrial_recA3_mut_3	2	13261	8069	52.79
mitochondrial_radA_wild_3	2	2119	766	4.64
mitochondrial_radA_mut_1	2	13790	6079	22.39
mitochondrial_radA_mut_2	2	3675	154	0.90
mitochondrial_radA_mut_3	2	1384	681	6.32
mitochondrial_why2_wild_1	2	4720	2629	20.01
mitochondrial_why2_mut_1	2	9965	6992	50.01
mitochondrial_why2_mut_2	2	1112	411	3.34
mitochondrial_why2_mut_3	2	1279	287	2.78
mitochondrial_MSH1_wild_1	2	931	95	0.52
mitochondrial_MSH1_mut_1	2	959	151	0.64
mitochondrial_MSH1_mut_2	2	925	50	0.42
mitochondrial_MSH1_mut_3	2	471	34	0.22
mitochondrial_recA3_wild_1	3	16270	12616	120.48
mitochondrial_recA3_mut_1	3	1684	1028	11.65
mitochondrial_recA3_mut_2	3	13791	10705	97.58
mitochondrial_recA3_mut_3	3	13488	9944	82.53
mitochondrial_radA_wild_3	3	869	650	3.98
mitochondrial_radA_mut_1	3	18460	13217	53.96
mitochondrial_radA_mut_2	3	393	44	0.19
mitochondrial_radA_mut_3	3	1017	500	5.02
mitochondrial_why2_wild_1	3	3596	2496	18.99
mitochondrial_why2_mut_1	3	6147	3618	26.85

mitochondrial_why2_mut_2	3	1507	85	0.62
mitochondrial_why2_mut_3	3	508	92	0.77

Table S2.5. Sequencing depth per bp (calculated with bedtools depth) of samples in Fig 8.

sample	Sequencing protocol	mut (total cov/bp)	wt (total cov/bp)
radA_mito	nanopore	222.9	101.6
recA3_mito	nanopore	736.4	372.1
why2_mito	nanopore	170.9	82.5
msh1_mito	nanopore	161.9	36.9
radA_mito	duplex	1600.2	1377.4
recA3_mito	duplex	941.4	647.8
why2_mito	duplex	1058.7	1235.2
msh1_mito	duplex	1020.9	1209.6
msh1_plastid	nanopore	1650.1	762.0
osb2_plastid	nanopore	6862.2	1663.7
radA_plastid	nanopore	5514.3	1668.5
recA1_plastid	nanopore	2422.6	2603.1
msh1_plastid	duplex	1754.4	1645.3
osb2_plastid	duplex	3244.9	3009.4
radA_plastid	duplex	3444.6	2205.6
recA1_plastid	duplex	1606.6	1650.2

Table S3.1: Primers used for generation and confirmation of *rad14D*

Primer Name	Sequence
JAO2397	ATTATGACTTTCTTGTTATATTCTTATATACATAACCAACATCAGATCTGTTTAGCTTGC
JAO2398	AAGAGTTTGGATCTTCGTAGTGAAGGTATCGAACGTAACGCTGGCGTTAGTATCGAATCG
JAO2399	ATGCACCCAAGGAATTGATTG
JAO2401	TATAGAAGCTCTATCTACAGC

Table S4.1. Differentially expressed genes from the RNA-seq analysis identified with DESeq2

Category	Genotype	Comparison	p-value	Log fold change	Average normalized cover of each treatment	Number of genes	Included in probe-set	Genes retained after arbor repeat filtering
1	WT	Increased exp. at 30°C	0.05	> 2	Minimum coverage > 5	683	100 with greatest LFC	84
2	WT	Increased exp. at 20°C	0.05	< -2	Minimum coverage > 5	350	100 with lowest LFC	80
3	WT	Constitutive low exp.	0.05	50 genes with LFC closest to 0	50 genes with lowest coverage (ranges from 129 to 400	50	50	44
4	WT	Constitutive high exp.	0.05	50 genes with LFC closest to 0	50 genes with highest coverage (ranges from 8384 to 68053	50	50	45
5	WT vs. HSP70-16	Interaction between genotype and temp	0.05	>2	Minimum coverage > 5	106 (39 of which are also in group 1)	92 with highest LFC	81
6	WT vs HSP70-60	Interaction between genotype and temp	0.05	<-2	Minimum coverage > 5	8 (5 of which are also in group 2)	All 8	5
total							400	339

Table S4.2. Duplex Sequencing coverage for each replicate

Sample	Mean Depth of Coverage	Total Duplex Seq. Data (bp)
WT 20°C A	86.86	74273348
WT 20°C B	92.16	78809954
WT 20°C C	82.40	70459706
WT 30°C A	81.46	69660673
WT 30°C B	95.39	81571700
WT 30°C C	93.77	80187868
HSP70-16 20°C A	82.31	70384149
HSP70-16 20°C B	74.75	63917524
HSP70-16 20°C C	93.94	80328860
HSP70-16 30°C A	93.65	80085644
HSP70-16 30°C B	81.50	69690981
HSP70-16 30°C C	98.70	84396810
MSH2 20°C A	105.53	90244630
MSH2 20°C B	95.50	81667422
MSH2 20°C C	107.69	92087225
MSH2 30°C A	95.50	81666433
MSH2 30°C B	87.40	74739952
MSH2 30°C C	93.40	79871709
UNG 20°C A	98.30	84059203
UNG 20°C B	93.33	79804898
UNG 20°C C	75.23	64327096
UNG 30°C A	109.44	93588299
UNG 30°C B	93.79	80203757
UNG 30°C C	106.23	90842455

Table S4.3. Putative fixed SNVs removed before downstream analysis of Duplex Sequencing data

Genotype	Chromosome	Position	Substitution type	Shared among all replicates
ung	2	2016156	AT→GC	yes
wild-type	2	14827204	CG→AT	yes
msh2	4	14827204	CG→AT	yes

Table S4.4. Putative fixed indels removed before downstream analysis of Duplex Sequencing data

Chrom	Pos	Indel Type	Genotype	Number of Reps (of 6)	Indel Length	Indel Seq
Chrom1	2243387	I	MSH2	6	1	G
Chrom1	2243387	I	WT	6	1	G
Chrom1	2243387	I	UNG	6	1	G
Chrom1	2243387	I	HSP70	6	1	G
Chrom1	2269740	D	MSH2	6	1	A
Chrom1	2270545	D	MSH2	5	1	T
Chrom1	2437835	D	MSH2	5	1	T
Chrom1	5291180	D	MSH2	6	1	T
Chrom1	6591532	I	MSH2	6	1	A
Chrom1	6591532	I	WT	6	1	A
Chrom1	6591532	I	UNG	6	1	A
Chrom1	6591532	I	HSP70	6	1	A
Chrom1	8551177	I	MSH2	6	1	G
Chrom1	8551177	I	WT	6	1	G
Chrom1	8551177	I	UNG	6	1	G
Chrom1	8551177	I	HSP70	6	1	G
Chrom1	11646952	D	MSH2	6	1	T
Chrom1	13533273	I	MSH2	6	3	AGA
Chrom1	13533273	I	WT	6	3	AGA
Chrom1	13533273	I	UNG	6	3	AGA
Chrom1	13533273	I	HSP70	6	3	AGA
Chrom1	17886514	D	MSH2	6	1	A
Chrom1	23734915	D	MSH2	6	1	A
Chrom1	26640491	D	MSH2	6	1	A
Chrom2	11236090	D	MSH2	4	1	A
Chrom2	11567248	I	MSH2	4	1	T
Chrom2	11567248	I	WT	6	1	T
Chrom2	11567248	I	UNG	6	1	T
Chrom2	11567248	I	HSP70	6	1	T
Chrom2	17464171	D	MSH2	6	1	T
Chrom3	4833763	D	MSH2	6	1	A
Chrom3	8412456	D	MSH2	4	1	T
Chrom3	18338647	D	MSH2	6	1	T
Chrom4	13742764	D	MSH2	6	1	T
Chrom4	16470637	I	MSH2	6	1	T
Chrom4	16470637	I	WT	6	1	T
Chrom4	16470637	I	UNG	6	1	T
Chrom4	16470637	I	HSP70	6	1	T
Chrom5	2974730	D	MSH2	4	1	T
Chrom5	7718829	D	MSH2	6	1	T
Chrom5	25010019	D	MSH2	6	1	A

Table S4.5. Indel mutations in *msh2* mutant lines

Sample	Deletions	Insertions
MSH2 20°C A	44	7
MSH2 20°C B	33	2
MSH2 20°C C	33	5
MSH2 30°C A	47	4
MSH2 30°C B	43	5
MSH2 30°C C	47	6
total	247	29

Table S4.6. Paired t-test results of group 3 vs group 4 mutation rates in *msh2*$^{-}$ lines (two-tailed)

Temp	Mutation class	Group 3 ave. variant frequency	Group 4 ave. variant frequency	P value
20 °C	SNV	1.02×10^{-7}	1.04×10^{-7}	0.9771
30 °C	SNV	7.25×10^{-8}	9.47×10^{-8}	0.6815
20 °C	INDEL	1.19×10^{-7}	1.38×10^{-7}	0.1615
30 °C	INDEL	1.17×10^{-7}	1.72×10^{-7}	0.0695

Table S4.7. Mutant lines used, all sourced from ABRC

Gene	AGI	Mutant Allele	Ref
HSP70-16	AT1G11660	SALK_028829	(Ran *et al.* 2020)
MSH2	AT3G18524	SALK_002708	(Belfield *et al.* 2018)
UNG	AT3G18630	CS308297	(Cordoba-Canero *et al.* 2010)

Table S4.8. PCR primers used to identify mutant alleles in the three mutant lines

Gene/line	Fwd Primer	Rev Primer
HSP70-16 WT	TACGCACTCACTTGCATTCAC	TGTGTTATCGCAGTTGCAAAG
HSP70-16 Mut	ATTTTGCCGATTTCGGAAC	TGTGTTATCGCAGTTGCAAAG
MSH2 WT	TCACCACGATGATGTCAAGAG	AGGAGCTGTCAAAAGGAGCTC
MSH2 Mut	ATTTTGCCGATTTCGGAAC	AGGAGCTGTCAAAAGGAGCTC
UNG WT	ACTTGGAGAAGGTAAAGCAATTCA	CCATACAAAATATAATACACCACCACTC
UNG Mut	ACTTGGAGAAGGTAAAGCAATTCA	ATATTGACCATCATACTCATTGC

Table S4.9. Read counts for the 12 RNA-seq libraries

Sample	Count of read pairs
HSP70-16 20°C A	29689895
HSP70-16 20°C B	32052311
HSP70-16 20°C C	33450418
HSP70-16 30°C A	32567642
HSP70-16 30°C B	31456737
HSP70-16 30°C C	29678098
WT 20°C A	30417658
WT 20°C B	54410188
WT 20°C C	42449872
WT 30°C A	34353207
WT 30°C B	36605678
WT 30°C C	37953073

Table S4.10. Read counts for the 24 Duplex Sequencing libraries

Sample	Count of read-pairs
HSP70-16 20°C A	102214316
HSP70-16 20°C B	88105828
HSP70-16 20°C C	106355604
HSP70-16 30°C A	88061502
HSP70-16 30°C B	99506728
HSP70-16 30°C C	112263590
MSH2 20°C A	106838516
MSH2 20°C B	90724220
MSH2 20°C C	111544972
MSH2 30°C A	115206890
MSH2 30°C B	93741162
MSH2 30°C C	111444292
UNG 20°C A	113380236
UNG 20°C B	110455064
UNG 20°C C	108883106
UNG 30°C A	91537708
UNG 30°C B	87766824
UNG 30°C C	123532620
WT 20°C A	100905496
WT 20°C B	102443086
WT 20°C C	116973524
WT 30°C A	97650342
WT 30°C B	105779540
WT 30°C C	110474398

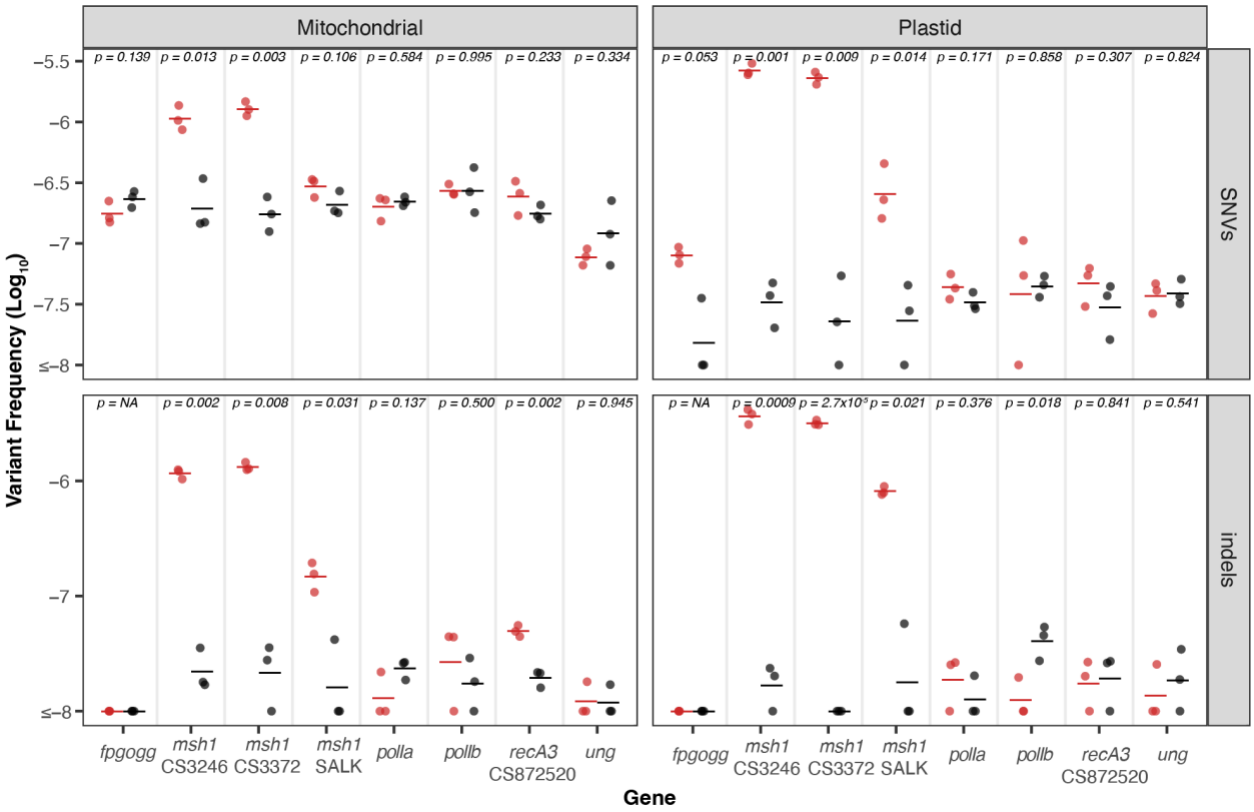

Figure S2.1. *De novo* point mutations measured with Duplex Sequencing from data generated in Wu *et al.* 2020. For each gene of interest (x axis) mutant lines are plotted in red and match WT controls are plotted in black. The individual biological replicates are plotted as circles, and group averages are plotted as dashes. Facets separate the data by genome (left column: Mitochondria and right column: Plastid) and by point mutation type (top row: SNVs and bottom row: indels). The y-axis shows the log-transformed SNV frequencies (total SNVs/total DCS coverage). P-values show the result of a two-tailed *t*-test comparing WT vs mutant mutation frequencies for each gene of interest. We found significant increases in SNV and indel frequencies in the *msh1* CS3246 and *msh1* CS3372 mutants (both genomes) but the *msh1* SALK046763 mutant, which is not a complete knockout of the *msh1* gene (Wu *et al.,* 2020) had weaker effects. In addition, we note that this *recA3* null allele is different from the *recA3* null allele than that reported in the new dataset, but both yielded similar results: significant indel and weakly significant SNV increases in mtDNA of the *recA3* mutant. Also note the marginally significant difference in *fpg*/*ogg* plastid SNVs is explained by just 5 SNVs in mutants and a single SNV in the WT controls, which we do not consider this to be a biologically meaningful difference.

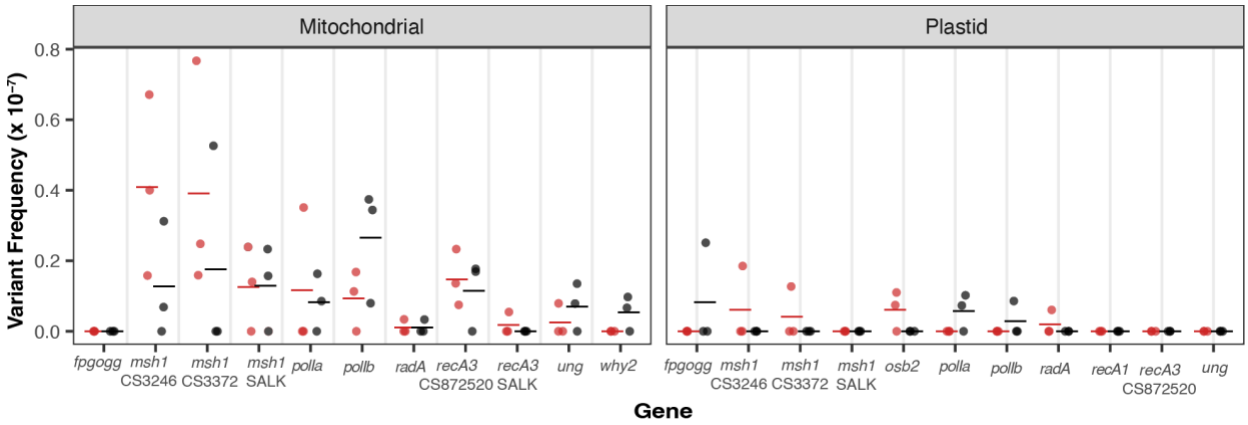

Figure S2.2. Dinucleotide mutations measured with Duplex Sequencing. For each gene of interest (x axis) mutant lines are plotted in red, and matched WT controls are plotted in black. The individual biological replicates are plotted as circles, and group averages are plotted as dashes. Facets divide the data by mitochondrial and plastid. We performed Wilcoxon rank sum tests to look for differences between mutant and matched wild-type controls and all p-values were > 0.05. Note that *recA3* CS872520 dataset was generated in Wu *et al.* (2020), and the *recA3* SALK 146388 dataset was generated in this study.

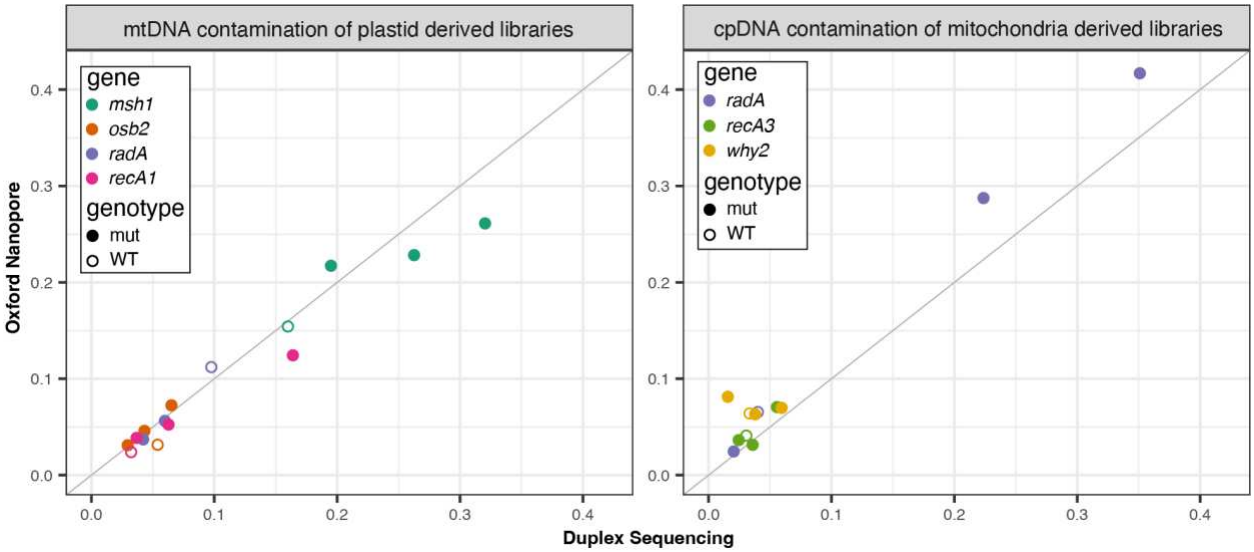

Figure S2.3. Correlation of cross-organelle contamination in Oxford Nanopore and Duplex Sequencing libraries. Contamination is calculated as the number of contaminating reads in the read alignments divided by the total number of organellar alignments. The different mutant lines are colored according to the figure legend with mutant replicates plotted using closed circles and matched WT controls plotted with open circles. The 1:1 diagonal line is show in gray. Though the level of contamination varies between different DNA samples (for example mtDNA contamination is higher in the plastid derived *msh1* libraries) the contamination levels are generally similar irrespective of sequencing technique.

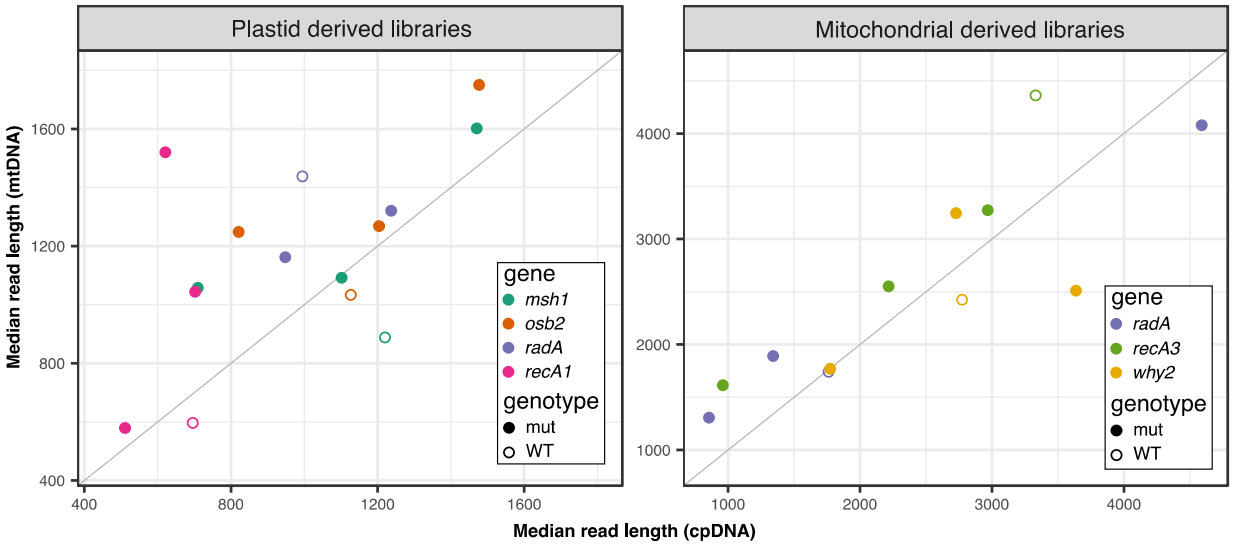

Figure S2.4. Median read length cross-organelle contaminating and native reads in the plastid and mitochondrial derived nanopore libraries. The different mutant lines are colored according to the figure legend with mutant replicates plotted using closed circles and matched WT controls plotted with open circles. The 1:1 diagonal line is show in gray.

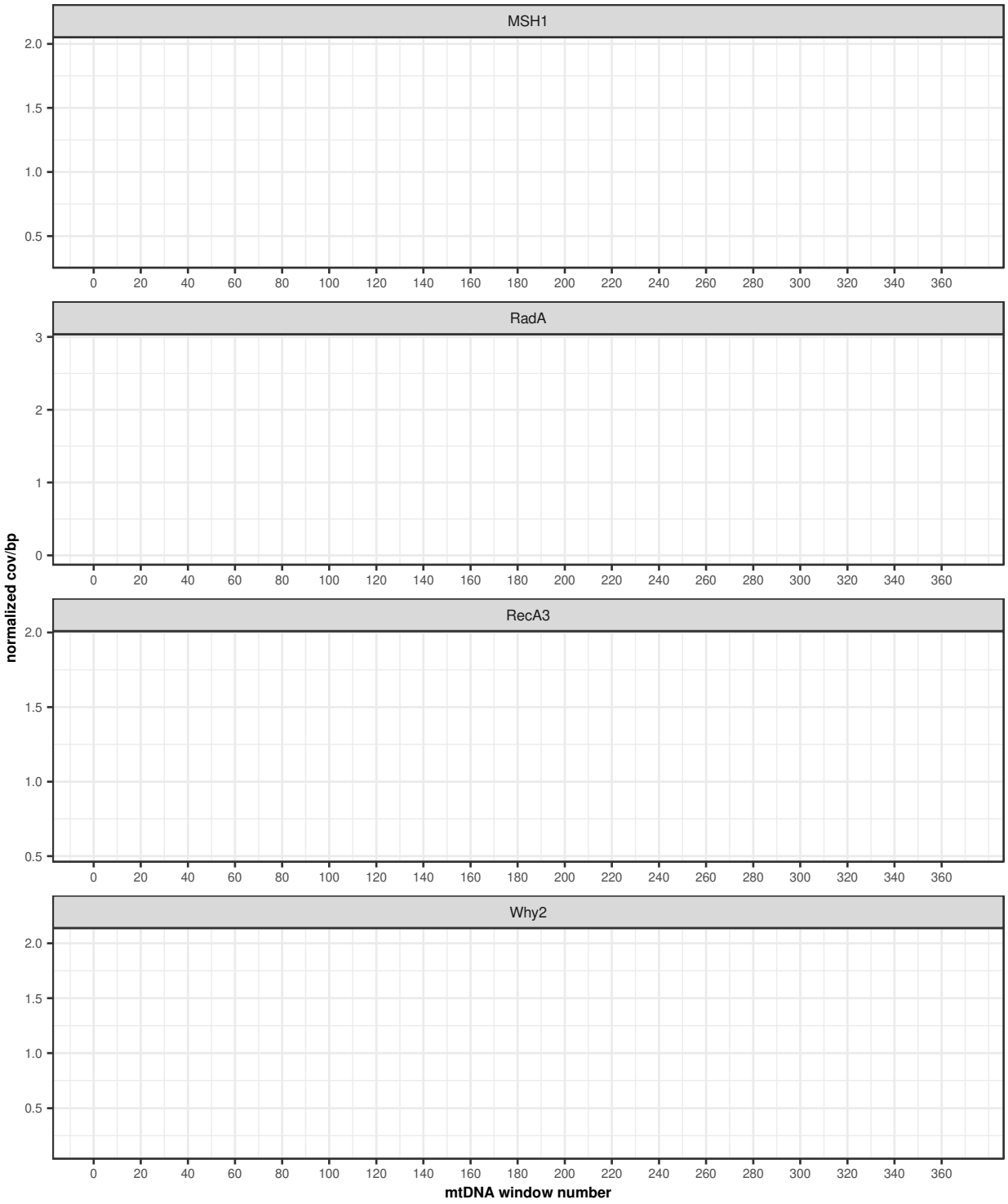

Figure S2.5. Normalized coverage of the individual nanopore mtDNA replicates (used to generate Fig. 2.8). The red and black lines show the normalized coverage of the mutant replicates and the matched WT control, respectively. Note that variation in the *why2* mutants is likely due to extremely low coverage in these samples (average coverage per bp of 157.3, 6.5 and 7.0 in mutant replicates 1, 2 and 3, respectively).

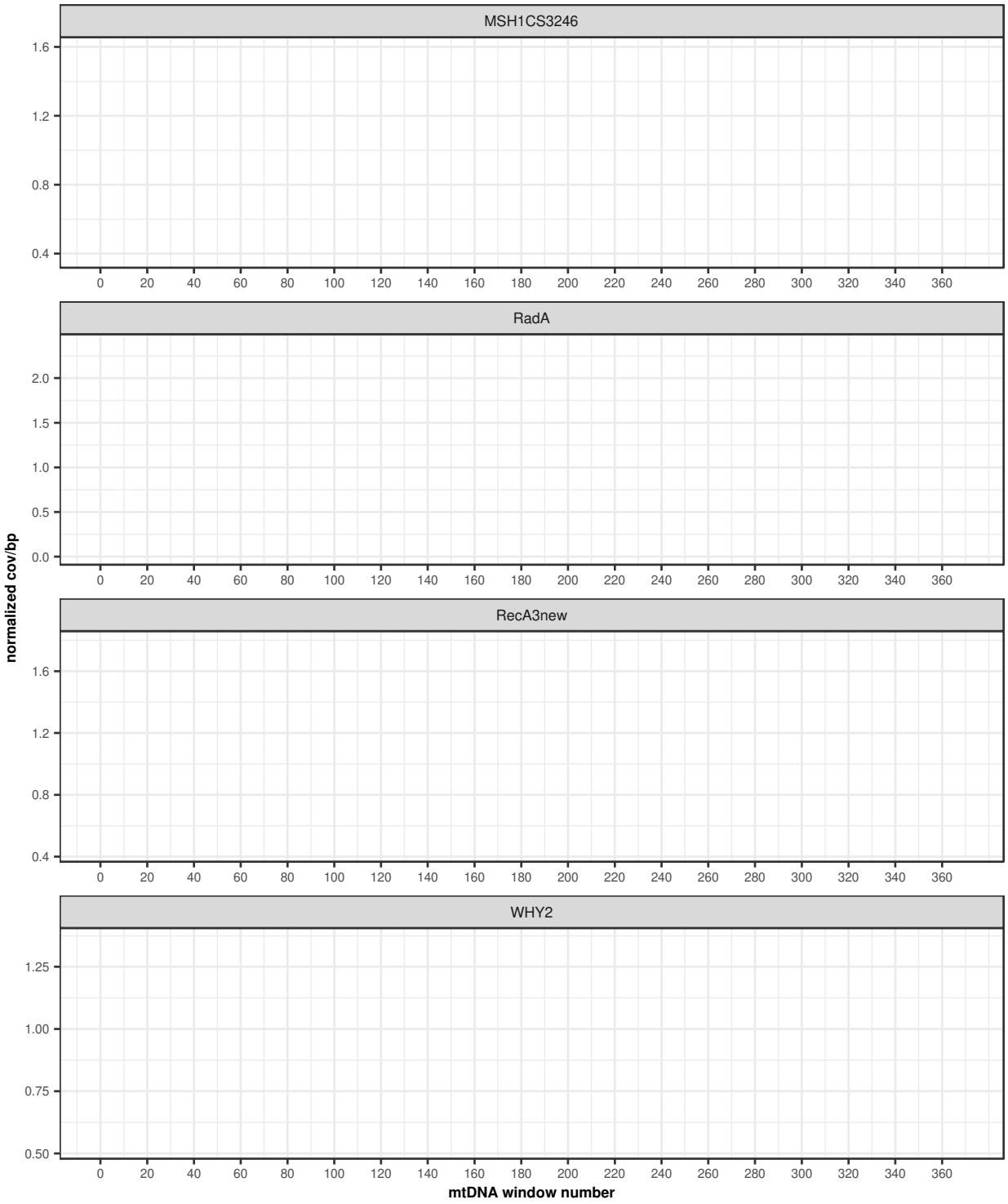

Figure S2.6. Depth of coverage of the individual Duplex Sequencing mtDNA replicates (used to generate Fig. 2.8). The red and black lines show the normalized coverage of the mutant replicates and the matched WT control, respectively

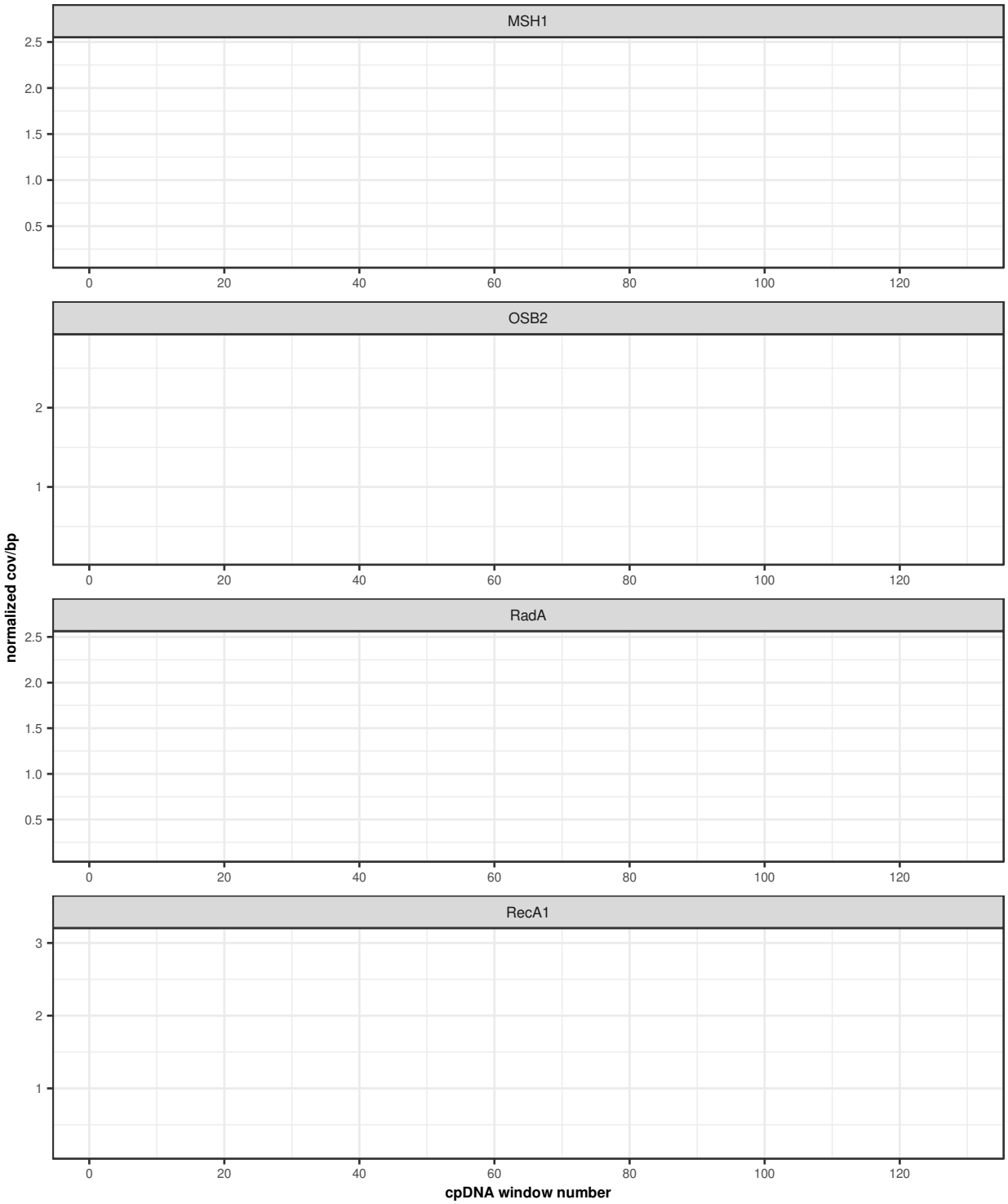

Figure S2.7. Normalized of coverage of the individual nanopore cpDNA replicates (used to generate Fig. 2.8). The red and black lines show the normalized coverage of the mutant replicates and the matched WT control, respectively

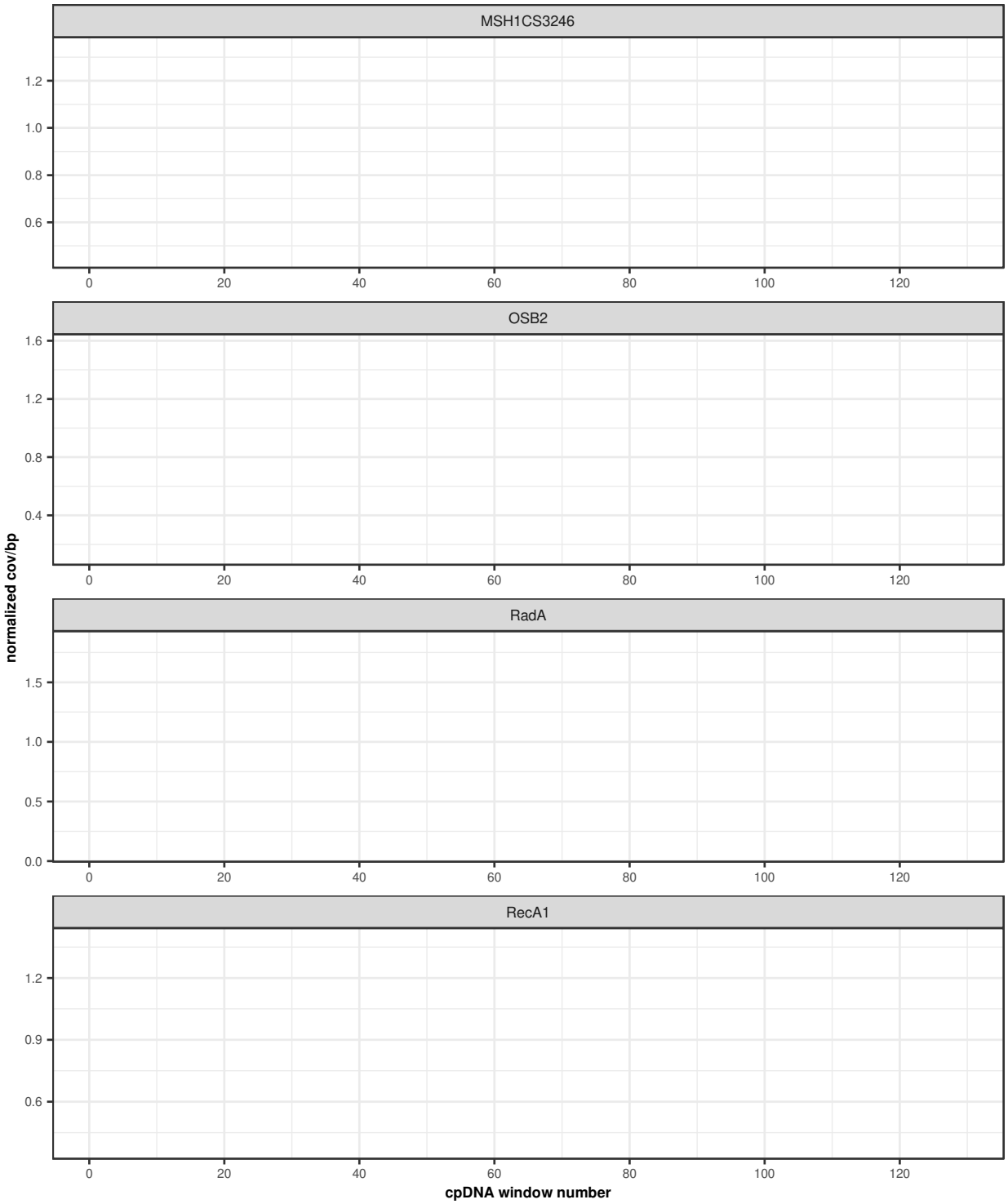

Figure S2.8. Normalized coverage of the individual Duplex Sequencing cpDNA replicates (used to generate Fig. 2.8). The red and black lines show the normalized coverage of the mutant replicates and the matched WT control, respectively.

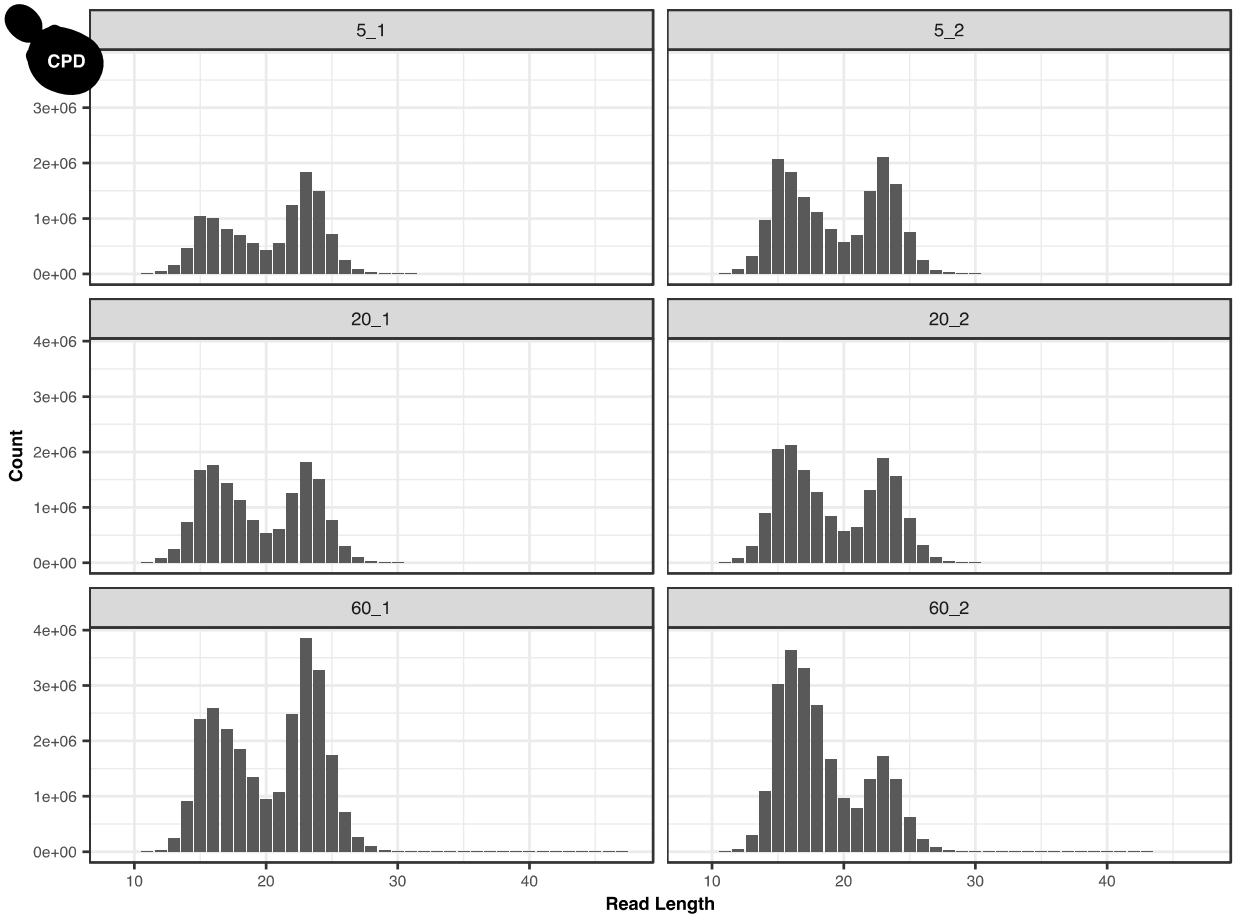

Figure S3.1. Length distributions of the nuclear-mapping reads from the *S. cerevisiae* anti-CPD libraries. The panel labels show the library ID, with the first number indicating the repair time (5, 20 or 60 minutes) and the second number indicating the replicate number (1 or 2).

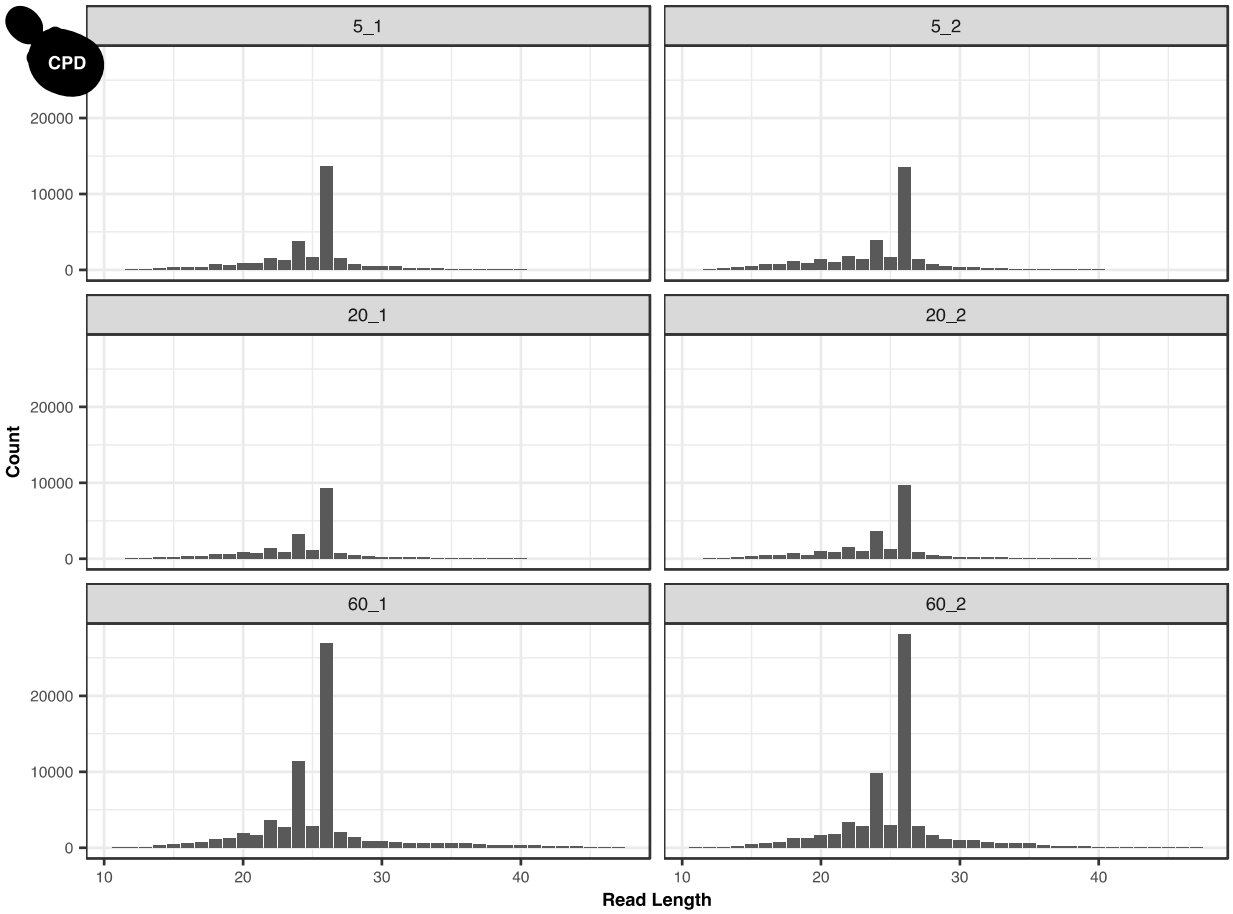

Figure S3.2. Length distributions of the mtDNA mapping reads from the *S. cerevisiae* anti-CPD libraries. The panel labels show the library ID, with the first number indicating the repair time (5, 20 or 60 minutes) and the second number indicating the replicate number (1 or 2).

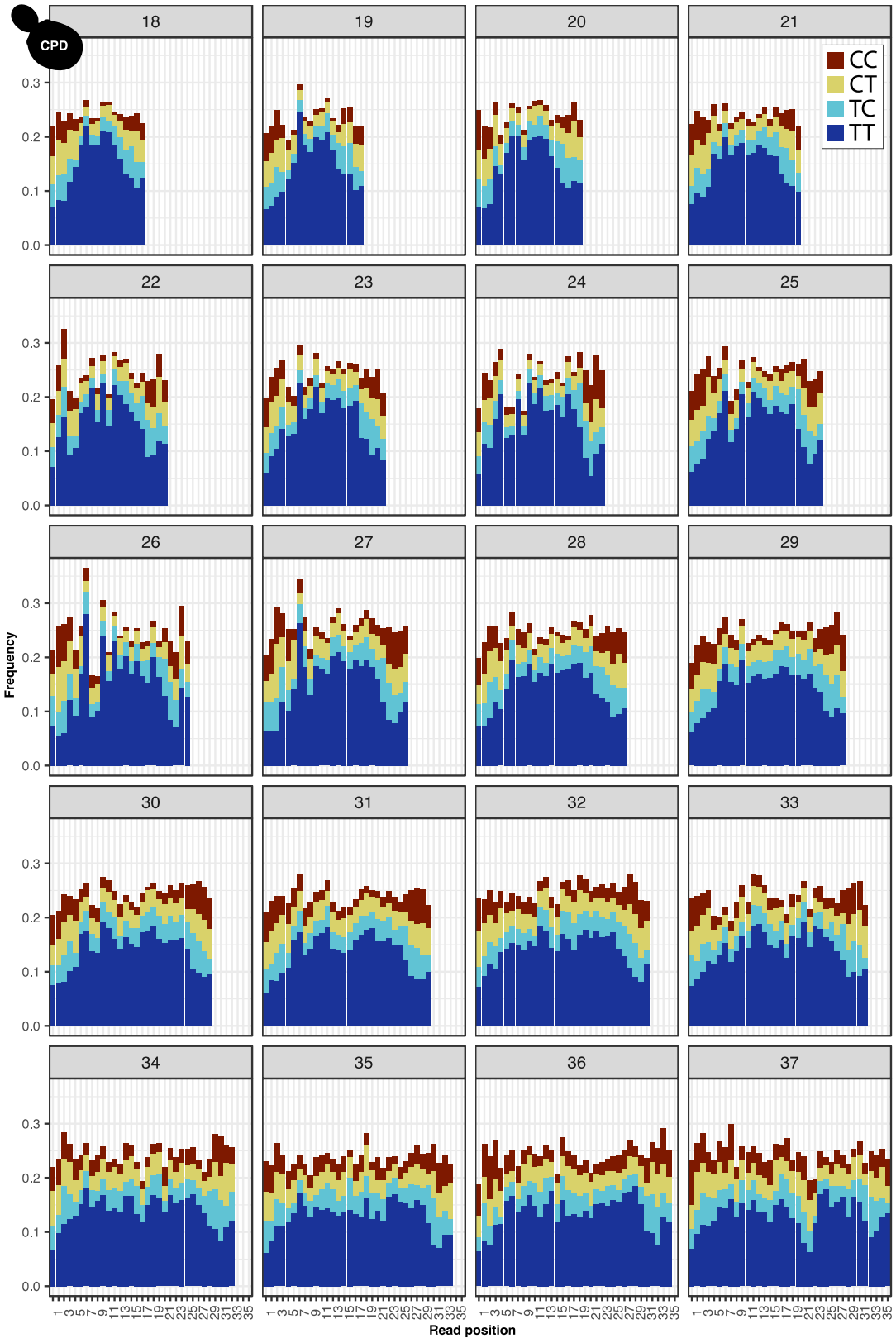

Figure S3.3. Di-pyrimidine frequencies in the mtDNA mapping reads from the *S. cerevisiae* anti-CPD libraries for all read-length classes.

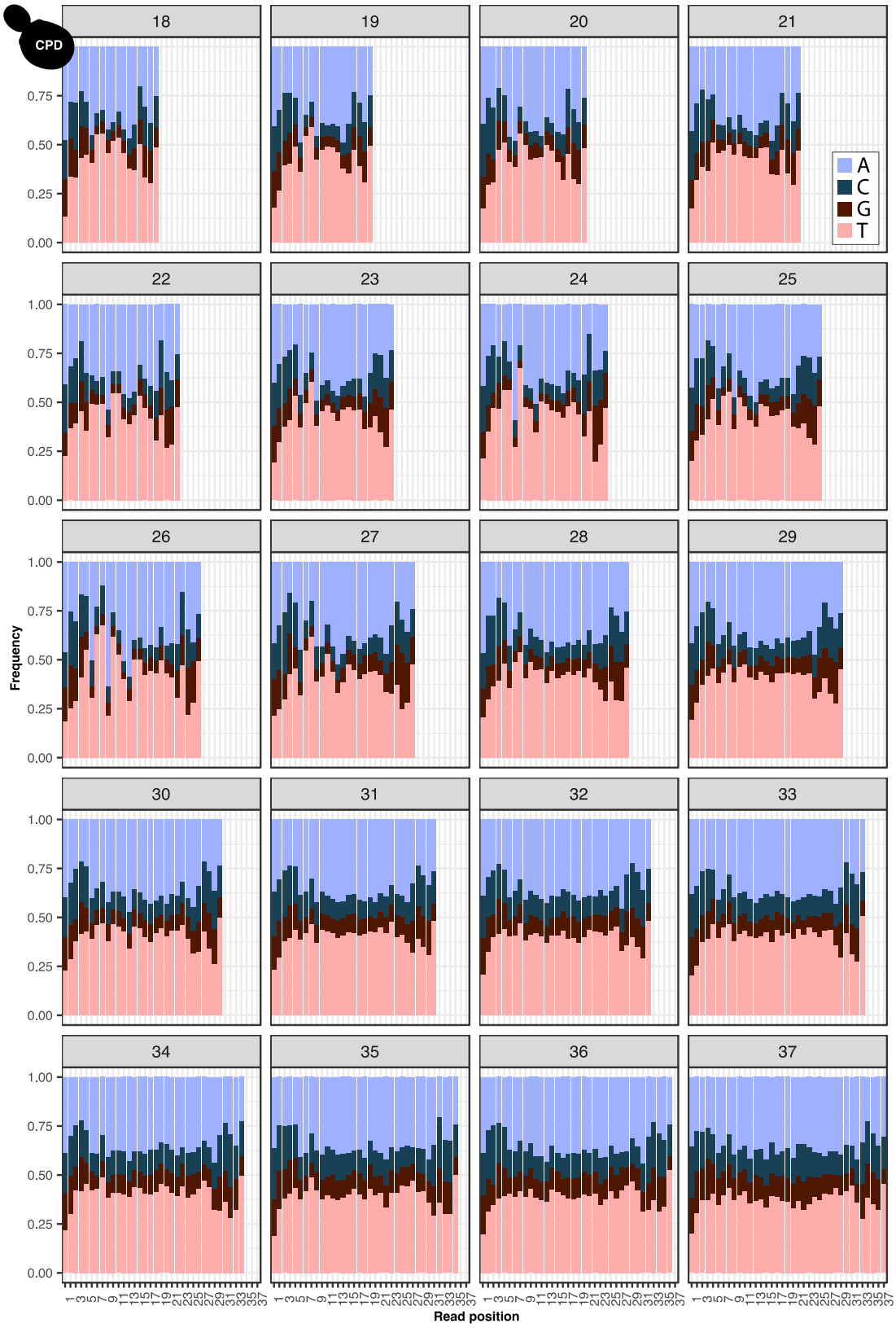

Figure S3.4. Nucleotide frequencies in the mtDNA mapping reads from the *S. cerevisiae* anti-CPD libraries for all read-length classes.

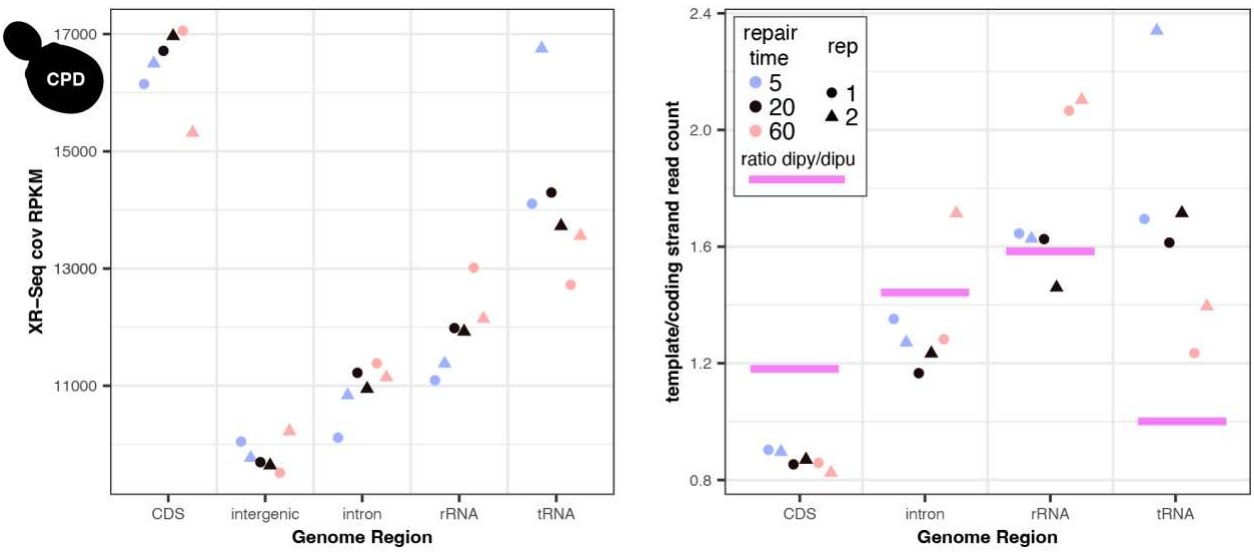

Figure S3.5. Depth of coverage (RPKM) and strand assymetry by genomic region in the mtDNA-mapping reads from the *S. cerevisiae* anti-CPD libraries. The pink bars show the ratio of dipyrimidines over dipurines on the coding strand, serving as an approximate null hypothesis for the ratio of XR-seq reads mapping to the template or coding strand.

***S. cerevisiae* anti-(6-4)PP libraries**

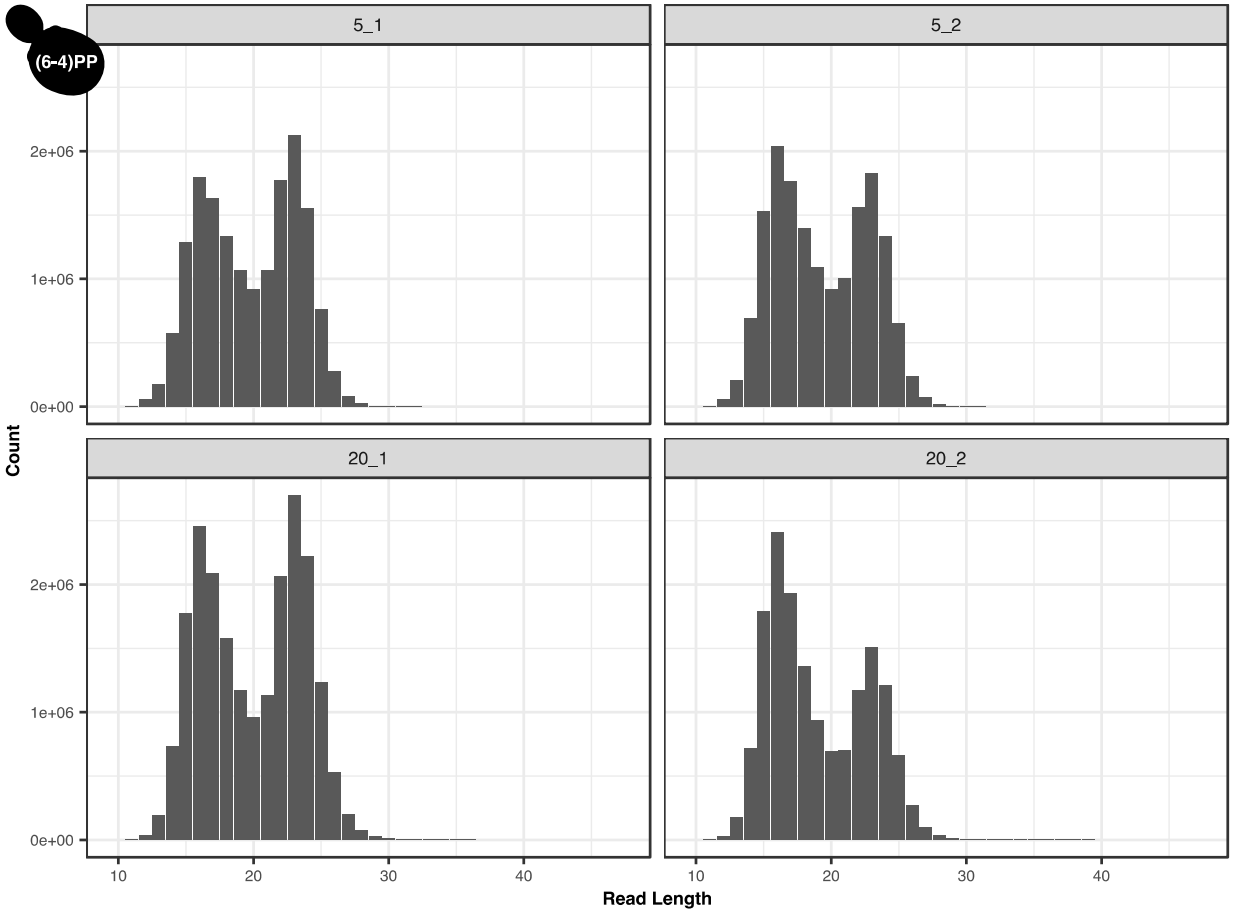

Figure S3.6. Length distributions of the nuclear-mapping reads from the *S. cerevisiae* anti-(6-4)PP libraries. The panel labels show the library ID, with the first number indicating the repair time (5 or 20 minutes) and the second number indicating the replicate number (1 or 2).

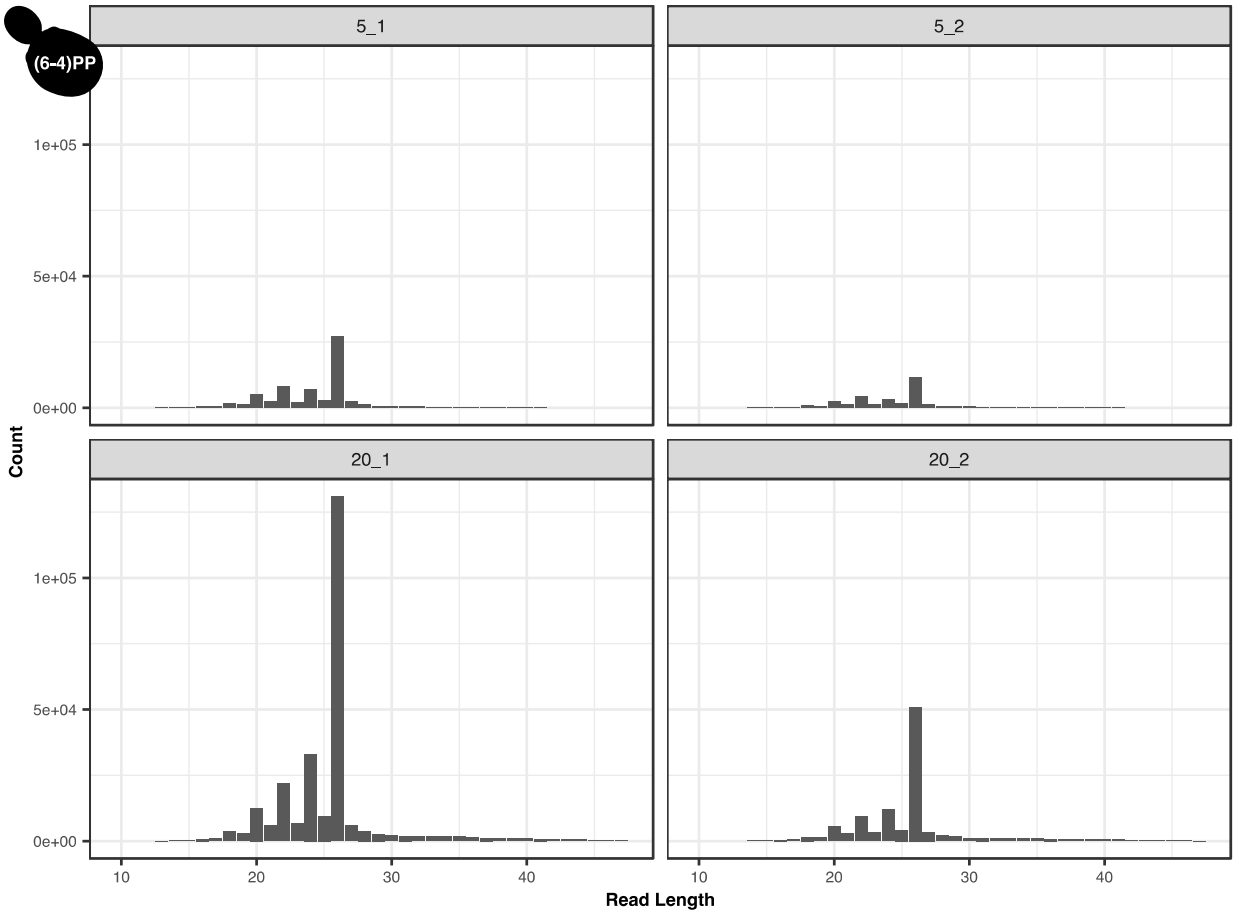

Figure S3.7. Length distributions of the mtDNA-mapping reads from the *S. cerevisiae* anti-(6-4)PP libraries. The panel labels show the library ID, with the first number indicating the repair time (5 or 20 minutes) and the second number indicating the replicate number (1 or 2).

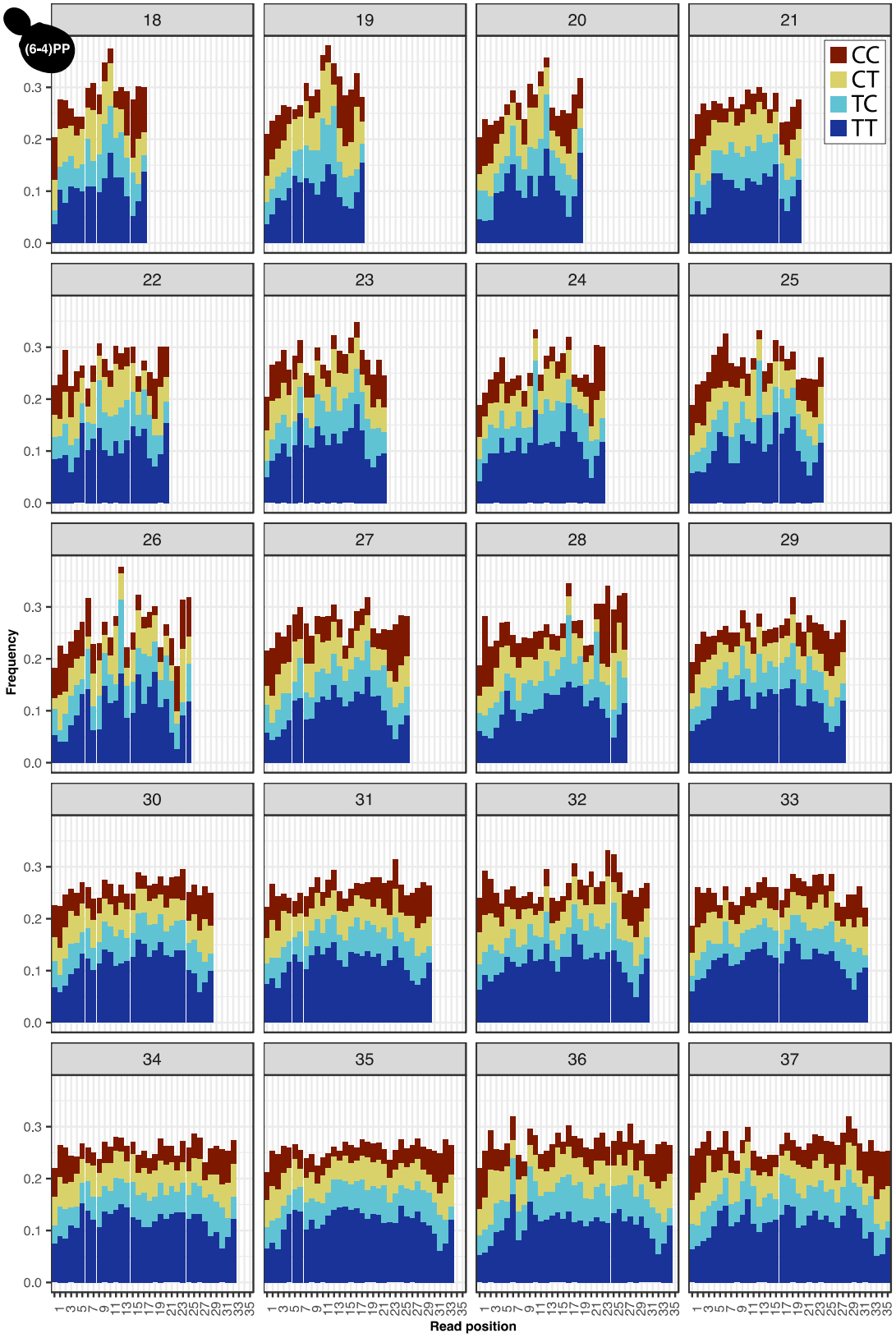

Figure S3.8. Di-pyrimidine frequencies in the mtDNA-mapping reads from the *S. cerevisiae* anti-(6-4)PP libraries for all read-length class.

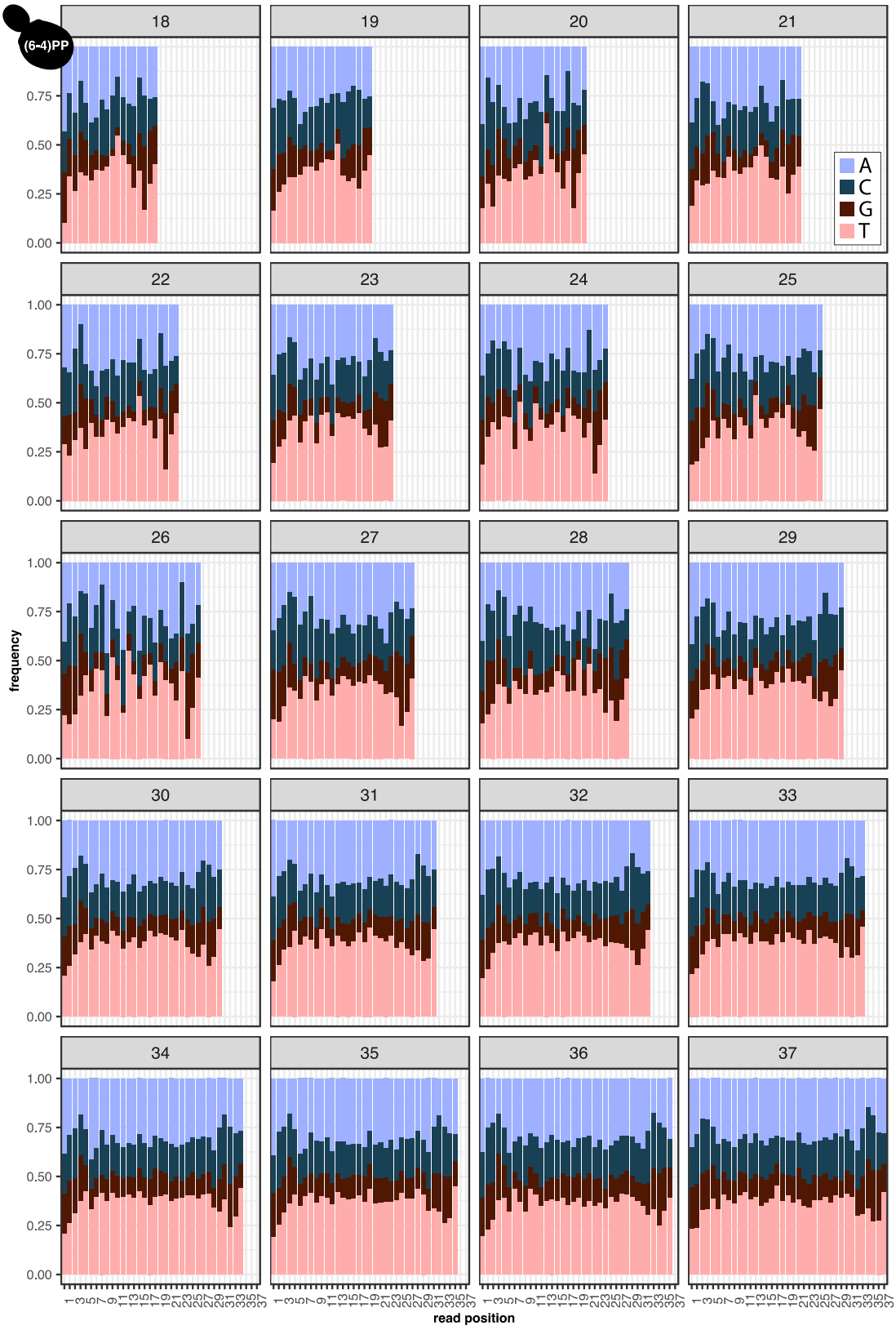

Figure S3.9. Nucleotide frequencies in the mtDNA-mapping reads from the *S. cerevisiae* anti-(6-4)PP libraries for all read-length classes.

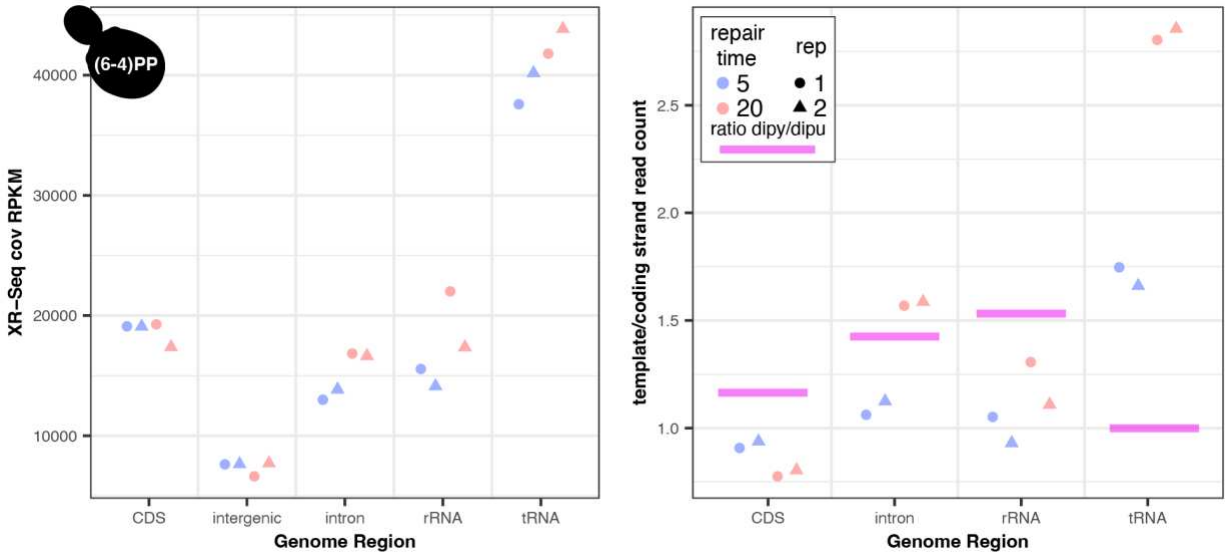

Figure S3.10. Depth of coverage (RPKM) and strand assymetry by genomic region in the mtDNA-mapping reads from the *S. cerevisiae* anti-(64)PP libraries. The pink bars show the ratio of dipyrimidines over dipurines on the coding strand, serving as an approximate null hypothesis for the ratio of XR-seq reads mapping to the template or coding strand.

***A. thaliana* anti-CPD libraries**

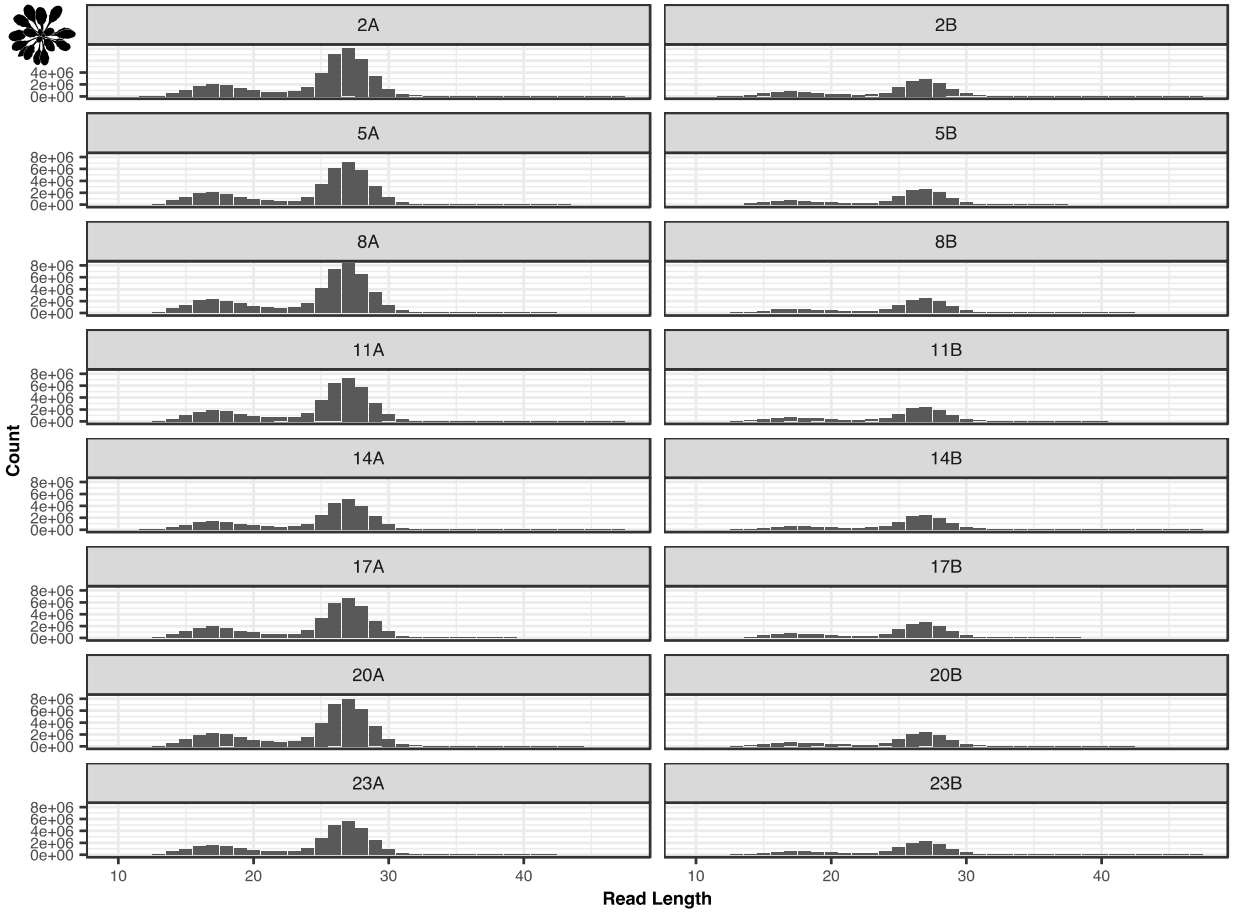

Figure S3.11. Length distributions of the nuclear-mapping reads from the *A. thaliana* anti-CPD libraries. The panel labels show the library ID, with the number indicating the time of day the plants were irradiated (2, 5, 8, 11, 14, 17, 20 or 23 hours) and the letter indicating the replicate (A or B).

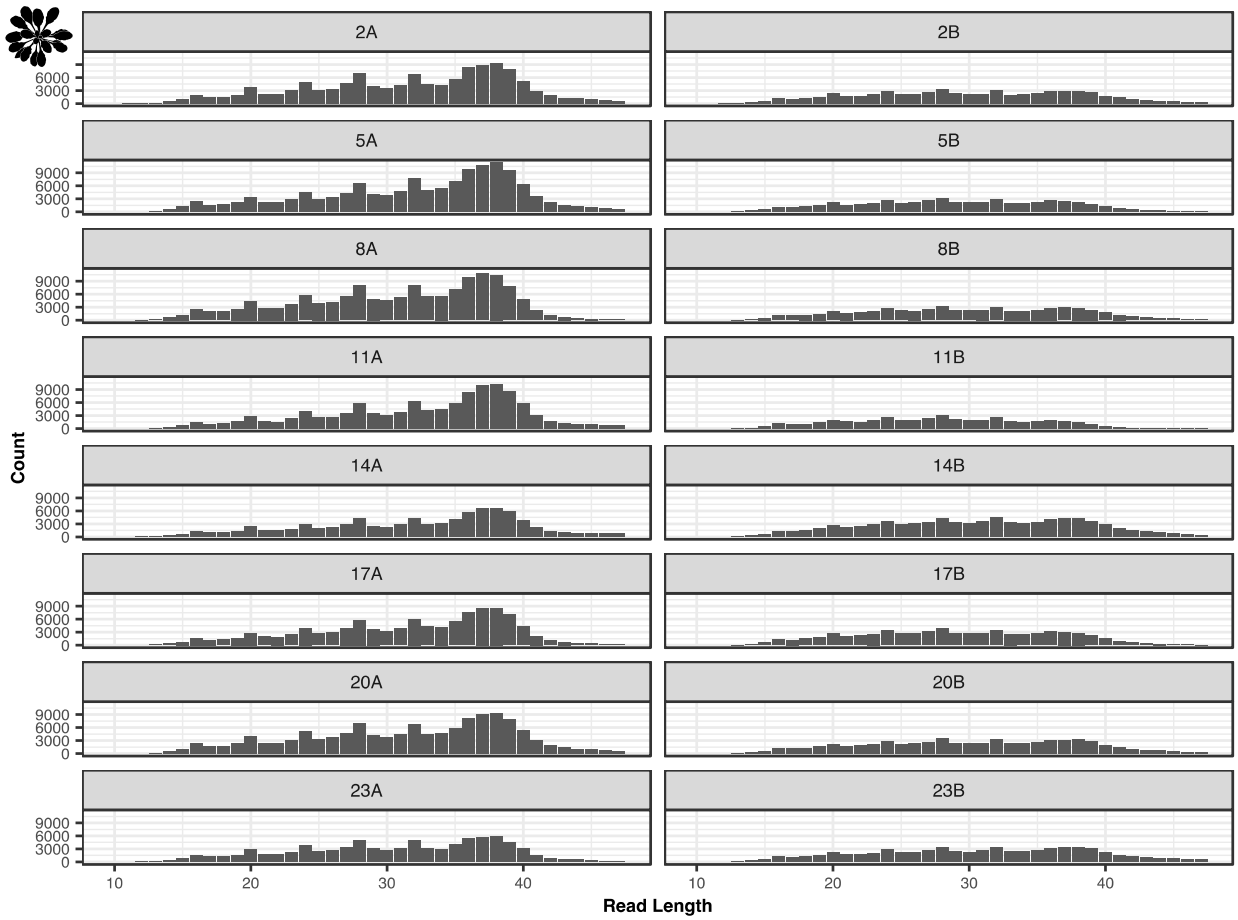

Figure S3.12. Length distributions of the mtDNA-mapping reads from the *A. thaliana* anti-CPD libraries. The panel labels show the library ID, with the number indicating the time of day the plants were irradiated (2, 5, 8, 11, 14, 17, 20 or 23 hours) and the letter indicating the replicate (A or B).

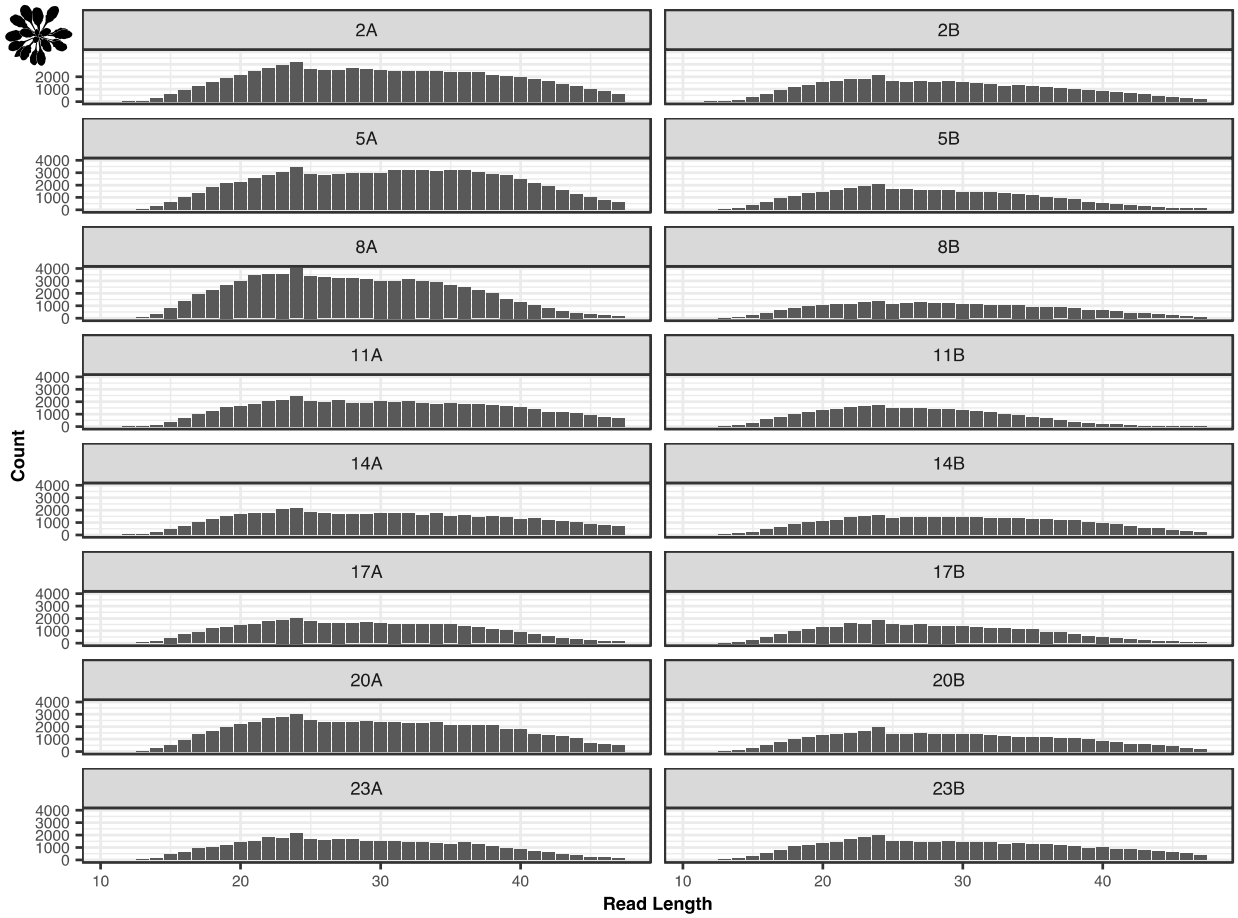

Figure S3.13. Length distributions of the ptDNA-mapping reads from the *A. thaliana* anti-CPD libraries. The panel labels show the library ID, with the number indicating the time of day the plants were irradiated (2, 5, 8, 11, 14, 17, 20 or 23 hours) and the letter indicating the replicate (A or B).

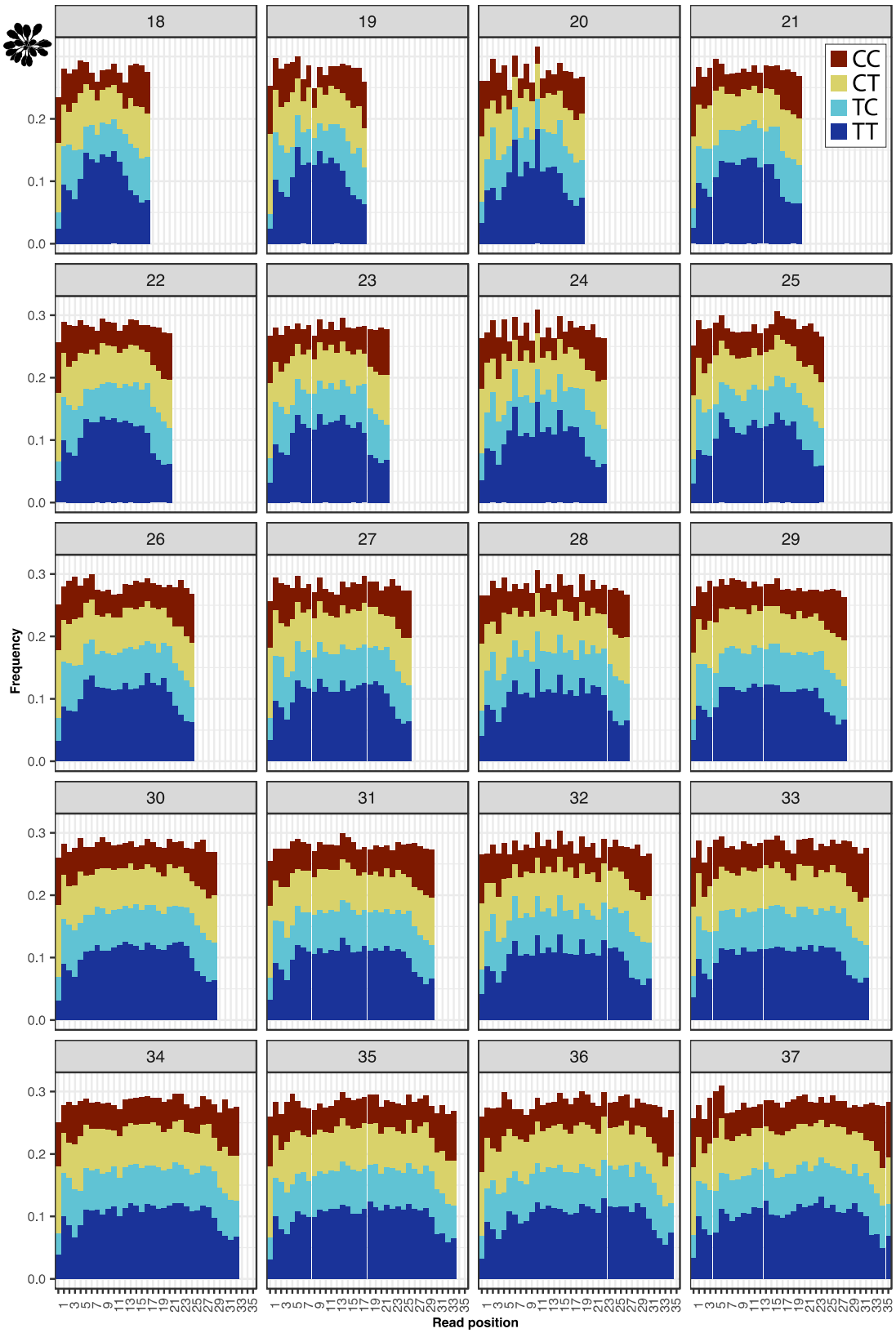

Figure S3.14. Di-pyrimidine frequencies in the mtDNA-mapping reads from the *A. thaliana* anti-CPD libraries.

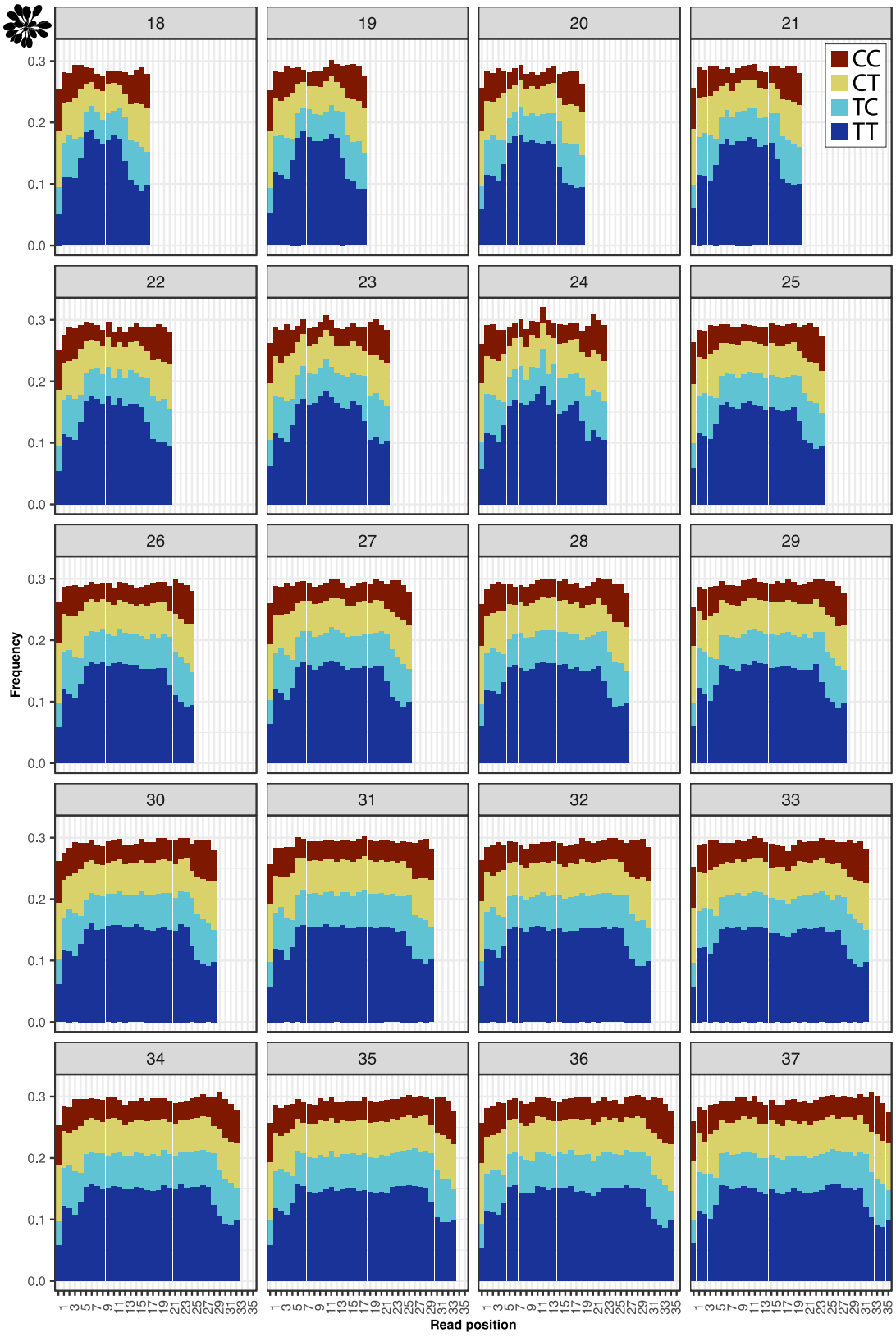

Figure S3.15. Di-pyrimidine frequencies in the ptDNA-mapping reads from the *A. thaliana* anti-CPD libraries.

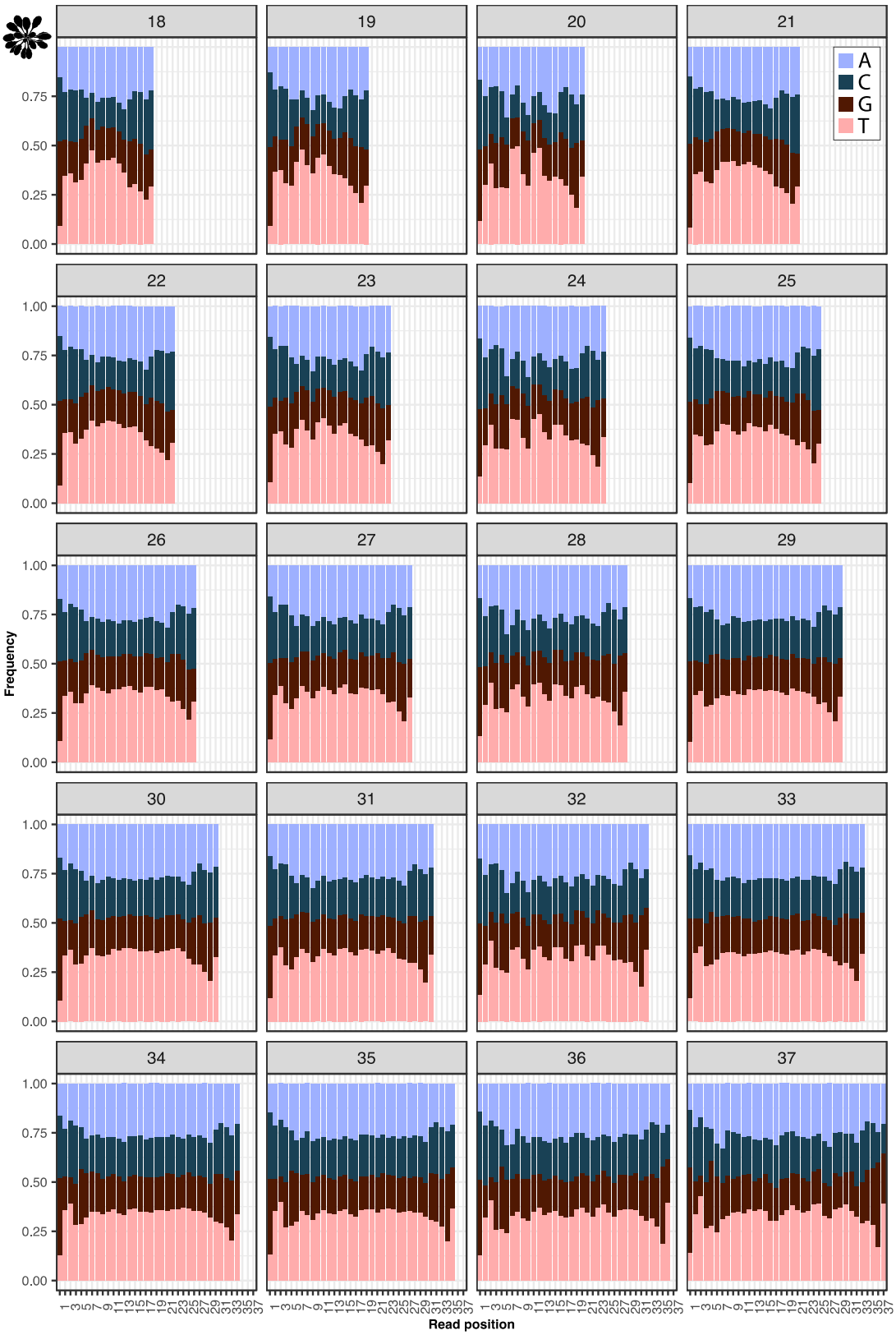

Figure S3.16. Nucleotide frequencies in the mtDNA-mapping reads from the *A. thaliana* anti-CPD libraries for all read-length classes.

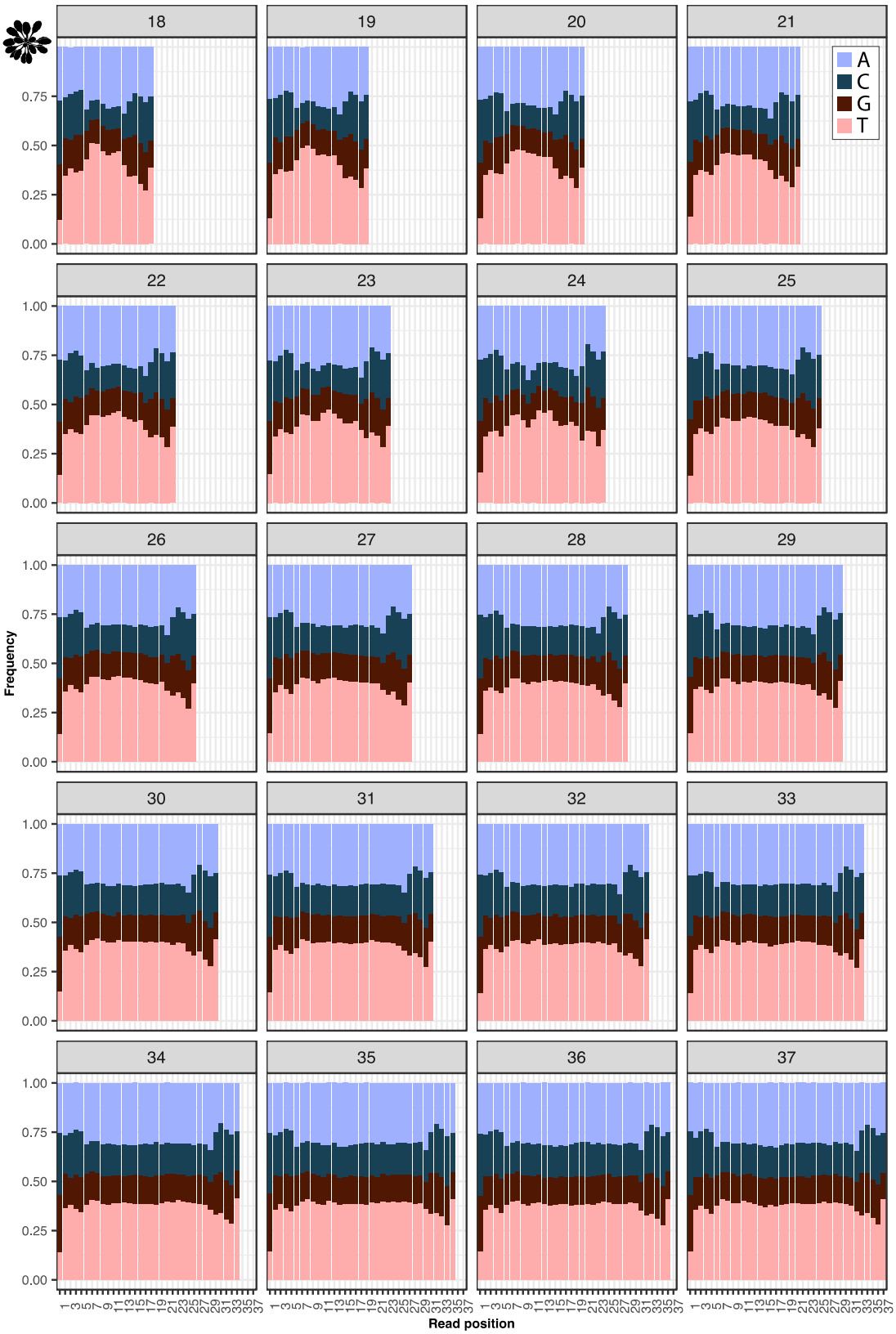

Figure S3.17. Nucleotide frequencies in the ptDNA-mapping reads from the *A. thaliana* anti-CPD libraries for all read-length classes.

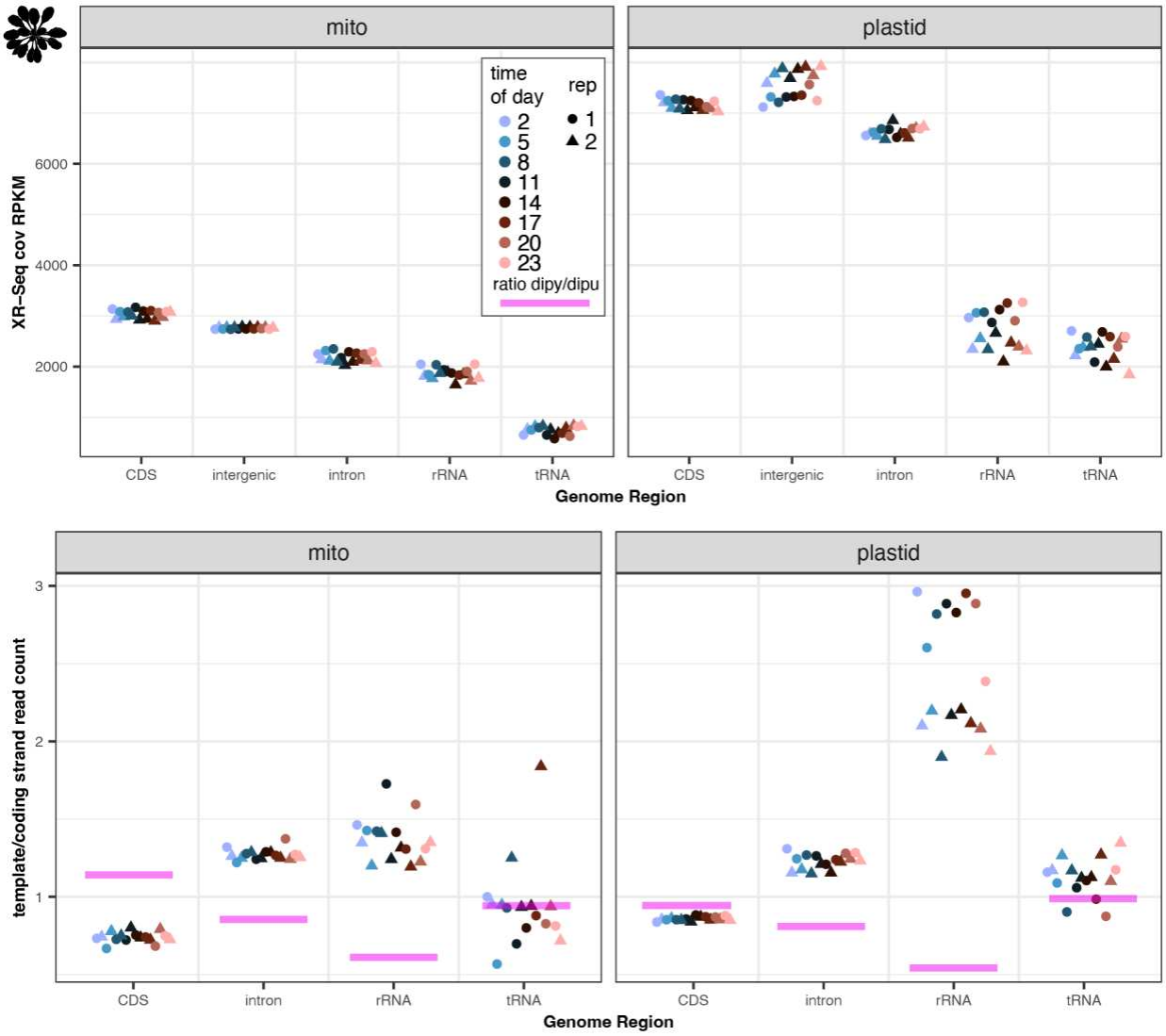

Figure S3.18. Depth of coverage (RPKM) and strand assymetry by genomic region in the mtDNA-and ptDNA-mapping reads from the *A. thaliana* anti-CPD libraries. The pink bars show the ratio of dipyrimidines over dipurines on the coding strand, serving as an approximate null hypothesis for the ratio of XR-seq reads mapping to the template or coding strand.

***D. melanogaster* anti-CPD libraries**

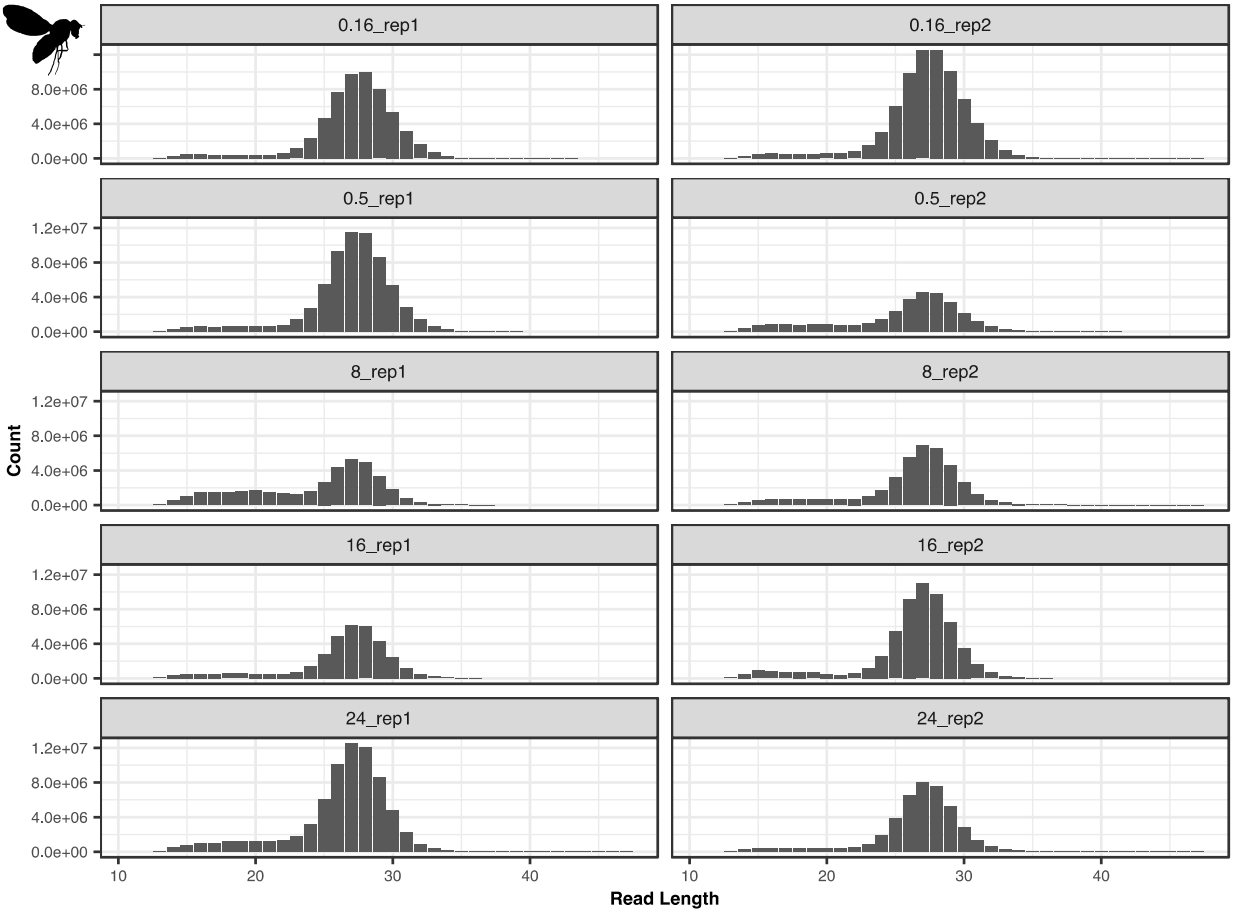

Figure S3.19. Length distributions of the nuclear-mapping reads from the *D. melanogaster* anti-CPD libraries. The panel labels show the library ID, with the number indicating the repair time (0.16, 0.5, 8, 16, or 24 hours) and the rep# indicating the replicate (1 or 2).

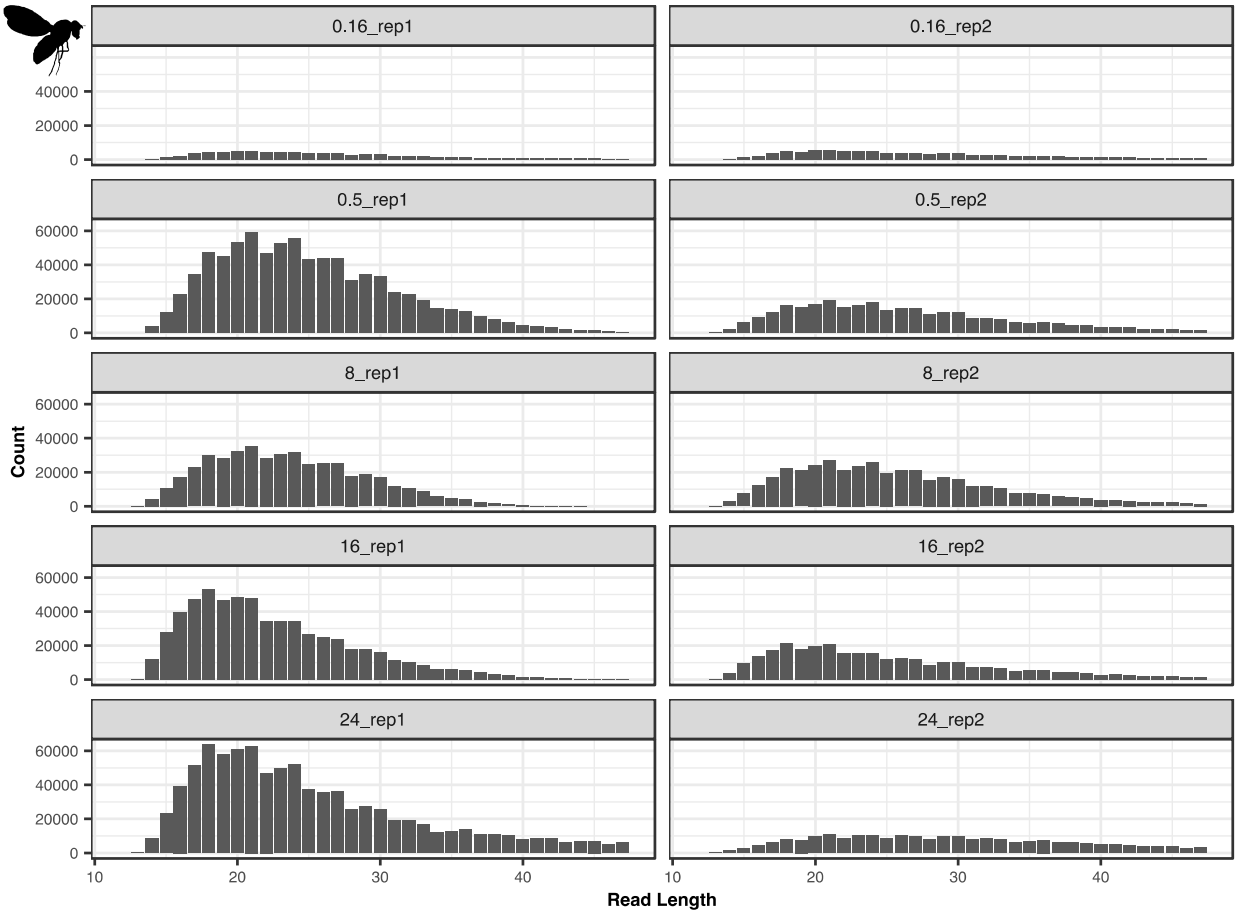

Figure S3.20. Length distributions of the mtDNA-mapping reads from the *D. melanogaster* anti-CPD libraries. The panel labels show the library ID, with the number indicating the repair time (0.16, 0.5, 8, 16, or 24 hours) and the rep# indicating the replicate (1 or 2).

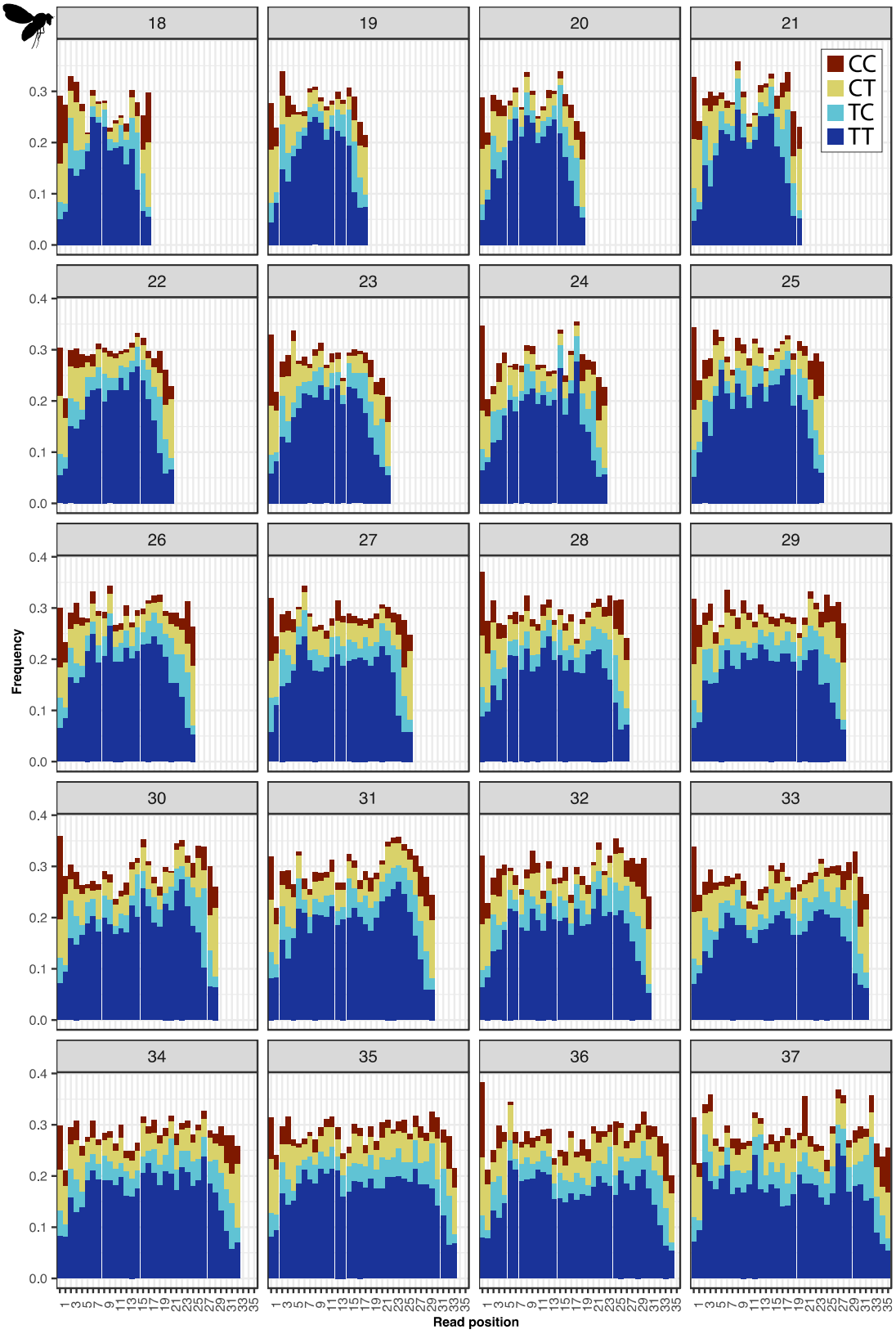

Figure S3.21. Di-pyrimidine frequencies in the mtDNA-mapping reads from the *D. melanogaster* anti-CPD libraries for all read-length classes.

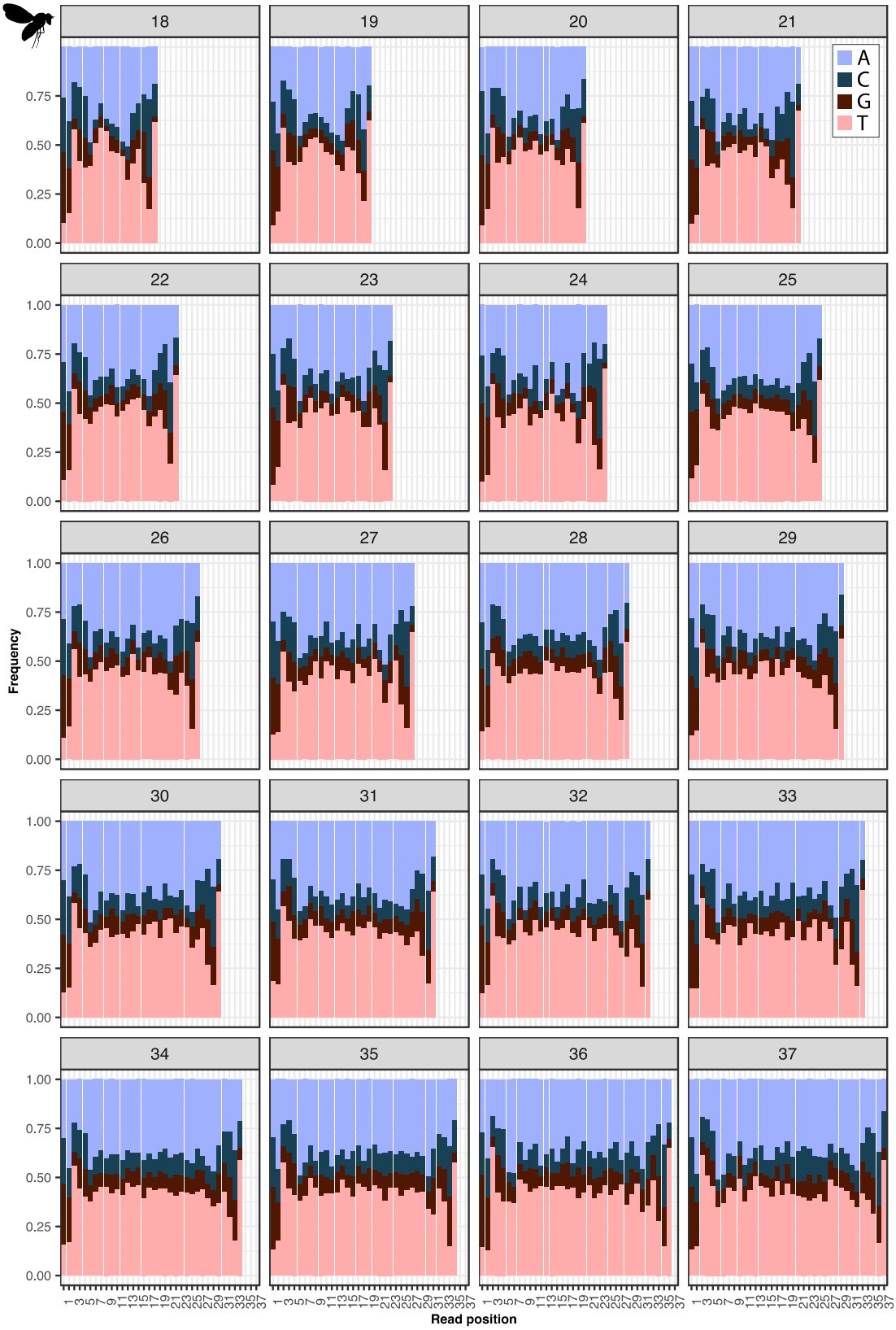

Figure S3.22. Nucleotide frequencies in the mtDNA mapping reads from the *D. melanogaster* anti-CPD libraries for all read-length classes.

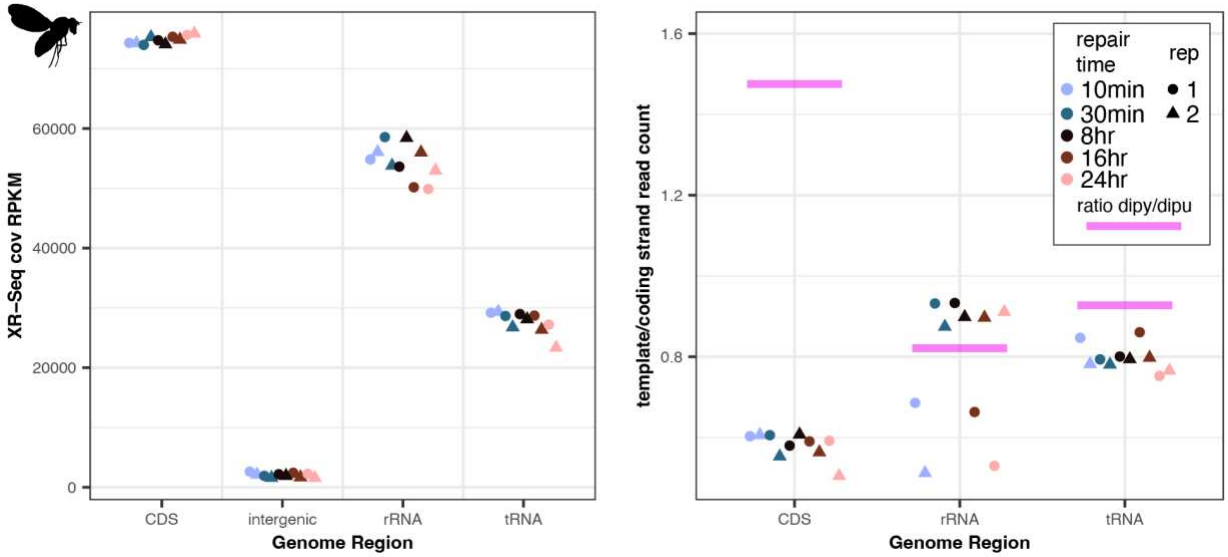

Figure S3.23. Depth of coverage (RPKM) and strand assymetry by genomic region in the mtDNA-mapping reads from the *D. melanogaster* anti-CPD libraries. The pink bars show the ratio of dipyrimidines over dipurines on the coding strand, serving as an approximate null hypothesis for the ratio of XR-seq reads mapping to the template or coding strand.

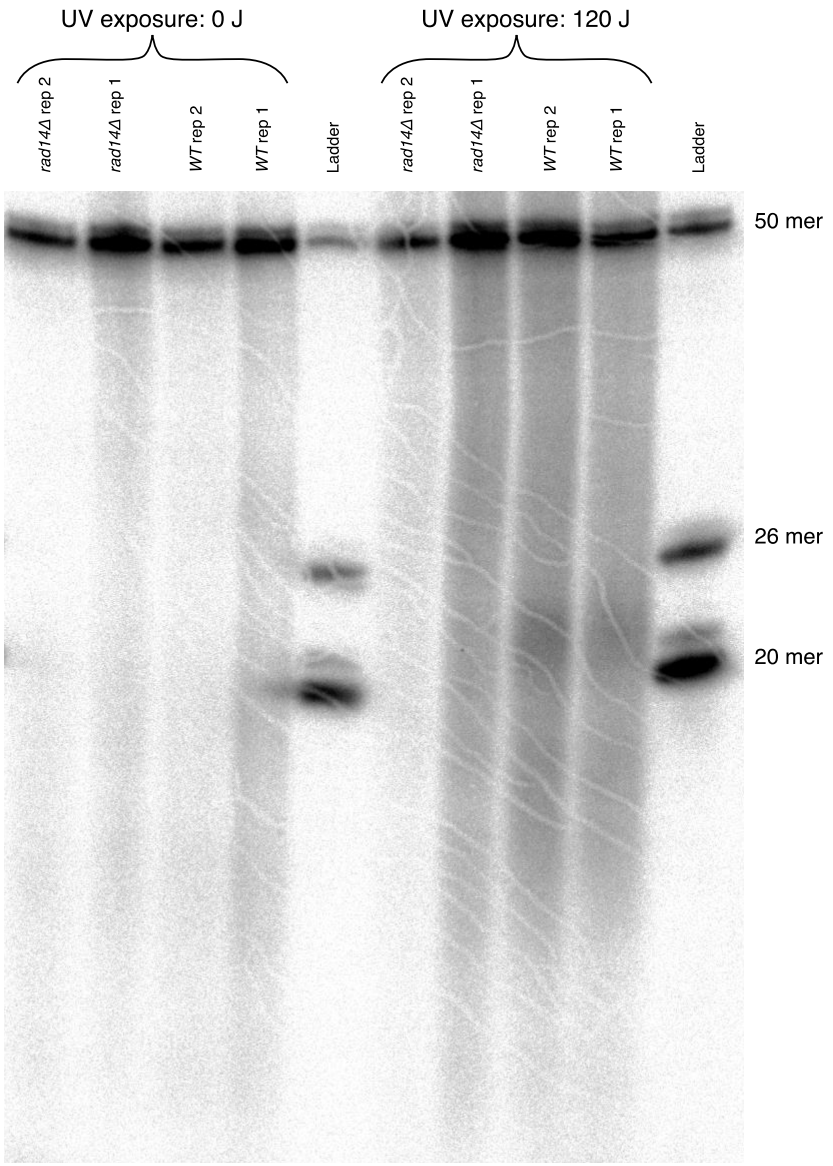

Figure S3.24. Excision assay gel with WT and *rad14D* deficient *S. cerevisiae* lines either not exposed to UV (left) or exposed to UV and given 20 minutes of dark repair (right). In conventional excision assays (Hu *et al.* 2013, 2019), 25mer and 50mer PAGE purified oligos are radiolabeled alongside samples of interest to serve as a ladder in the high density polyacrylamide electrophoresis. Under our prediction that the nuclear-NER mutant would produce only mtDNA-derived fragments, we expected to detect the 26-nt, 24-nt and 22-nt peaks that punctuate the *S. cerevisiae* read length distribution (Fig. 3.1), and thus used PAGE purified oligos of 50, 26, and 20 nt in length to make a custom ladder. Though our 11 % acrylamide gel cracked during drying, the samples and bands can still be reliably identified. In the UV-exposed WT samples, note the darker area between the 26mer and the 20mer markers, and the slightly less dark area below the 20mer marker, corresponding to the primary products and degraded products in the excsion assay shown in Fig. 3.1 (and Fig. 2 of (Li *et al.* 2018)). We predicted to detect mtDNA-derived peaks at 26, 24 and 22 nt in the *rad14D* lanes, but such bands were not apparent, likely due the high background noise of this assay.

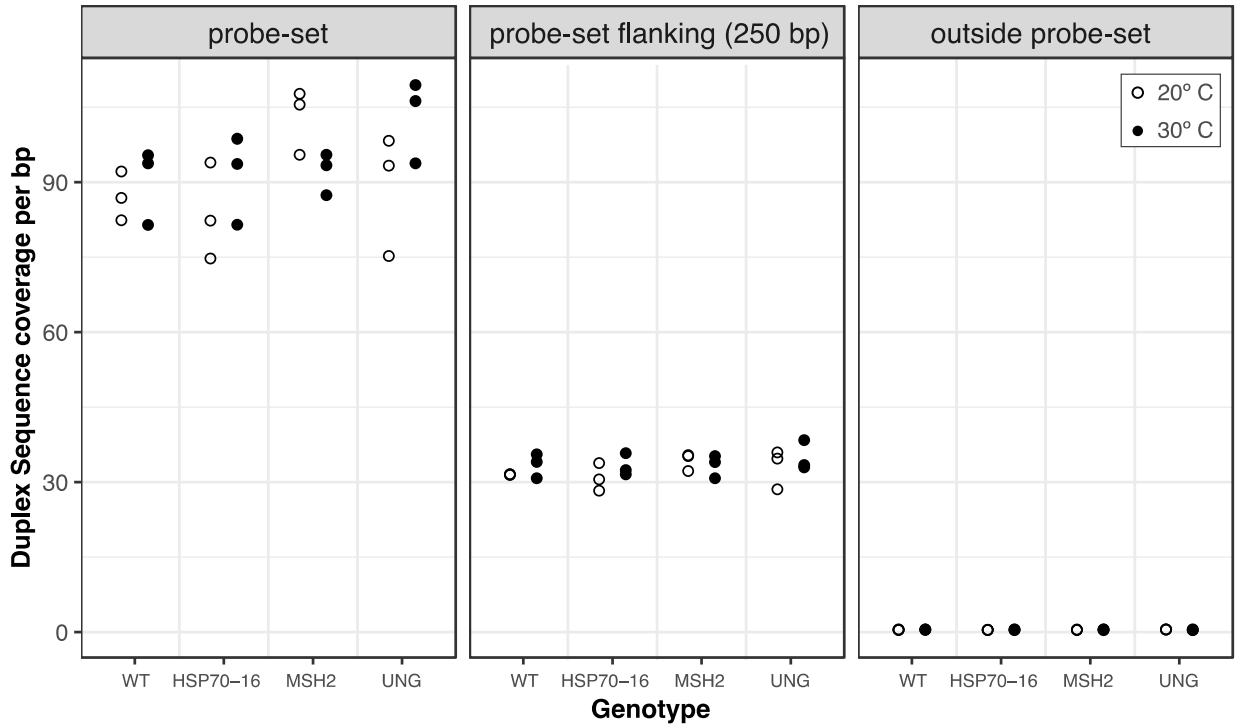

Figure S4.1. Duplex Sequencing coverage of the probe-set (panel 1), the 250 bps flanking the probe-set (panel 2) and the rest of the genome, outside of the probe-set (panel 3).