4-2019

Plant mitochondrial genome evolution and structure has been shaped by double-strand break repair and recombination

Emily Wynn
University of Nebrask-Lincoln, emilywynn6@gmail.com

Plant mitochondrial genome evolution and structure has been shaped by double-strand break repair and recombination

by

Emily Wynn

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Biological Sciences

(Genetics, Cellular and Molecular Biology)

Under the Supervision of Professor Alan C. Christensen

Lincoln, Nebraska

April, 2019

Plant mitochondrial genome evolution and structure has been shaped by double-strand break repair and recombination

Emily Wynn, Ph.D.

University of Nebraska, 2019

Adviser: Alan C. Christensen

Plant mitochondrial genomes are large but contain a small number of genes. These genes have very low mutation rates, but genomes rearrange and expand at significant rates. We propose that much of the apparent complexity of plant mitochondrial genomes can be explained by the interactions of double-strand break repair, recombination, and selection. One possible explanation for the disparity between the low mutation rates of genes and the high divergence of non-genes is that synonymous mutations in genes are not truly neutral. In some species, *rps14* has been duplicated in the nucleus, allowing the mitochondrial copy to become a pseudogene. By measuring the synonymous substitution rate of *rps14* genes and the total substitution rate of *Ψrps14* pseudogenes we inferred that synonymous mutations in plant mitochondrial genes are not truly neutral. Plant mitochondrial genomes contain many repeated sequences and little is known about their evolution. We wrote a Python script that utilizes BLAST to identify and organize repeated sequences in DNA. Using this program on a large number of species from many different lineages of plants, we found that large repeats above 1kb are found only in the tracheophytes, and repeats larger than 10kb are unique to angiosperms. We proposed that the creation and maintenance of these repeats may be a side effect of the DNA repair pathways necessary to survive desiccation during seed or spore formation. To test our hypothesis that double-strand break repair is a generalized DNA

repair pathway in plant mitochondria, we examined an *Arabidopsis thaliana* uracil DNA N-glycosylase (UNG) mutant, which cannot repair uracil in DNA through the base excision repair pathway. We set up a mutation-accumulation study, growing independent *ung* mutant lines for 10 generations and sequencing the mitochondrial genome with next-generation sequencing. No mutations had reached fixation in any of the sequenced lines, and the rate of heteroplasmic mutation accumulation was not different from wild-type. Using RT-PCR, we found that genes involved in double strand break repair were transcriptionally elevated. Clearly double strand break repair is an effective and generalized form of DNA repair in plant mitochondria.

ACKNOWLEDGEMENTS

This work would not have been possible without the support of so many people. I'd like to thank my advisor Dr. Alan Christensen for being a kind and insightful mentor, and for protecting my RNA extractions with a barbed-wire fence.

Thank you to the members of my committee, Dr. Kristi Montooth, Dr. Jeffrey Mower, Dr. Heriberto Cerutti, and Dr. Thomas Clemente, for asking the tough questions and keeping me on track.

I'd like to thank Dr. Etsuko Moriyama and Dr. Chad Brassil for answering my questions about bioinformatics and statistics, respectively. Thanks to Julien Gradnigo for help learning Python and the Linux command line.

This would have been so much more difficult without all of the help from Emma Purfeerst for keeping the lab running smoothly and to all the undergraduate researchers, especially Emily Jezewski, who helped me get experiments done.

Of course none of this would be possible without the love of my family. Thank you to my dad, Clyde Wynn, my mom, Linda Urban, my stepdad, Jerry Urban, my sister Allie Wynn, and my niblings: Tahjaia, Aniyah, Khyler, Marley, and Madden.

All the gratitude I can muster goes to my wife, Maya Khasin. Their love and support makes anything possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

DNA REPAIR, RECOMBINATION, AND GENOME STRUCTURE IN PLANT MITOCHONDRIA

Introduction

Plant mitochondrial genomes are weird. Animal mitochondrial genomes are known for their small size, circular structures, and high mutation rates. In contrast, plant genomes can be extremely large, are made up of overlapping linear, branched and circular molecules (Bendich 1993), and have very low mutation rates in genes (Wynn and Christensen 2015). Despite the low mutation rate in coding sequences, the genomes expand and rearrange at appreciable rates (Palmer and Herbon 1988). Here we review the literature of plant mitochondrial DNA repair and genome dynamics and propose a model to explain the seeming complexity and contradictions of plant mitochondrial genomes by the interactions of a few simple forces, namely double-strand break repair, recombination, and selection.

Plant mitochondria and animal mitochondria are likely derived from the same endosymbiotic event (Gray 1999). However, their evolution has produced very different strategies for genome maintenance. For comparison, the human mitochondrial genome is around 16.5 kilobases (kb) and contains 37 genes (Taanman 1999), whereas the *Arabidopsis thaliana* mitochondrial genome is almost 367 kb and contains 57 genes (Unseld *et al.* 1997). The *Arabidopsis* mitochondrial genome is more than 20 times the size of the human mitochondrial genome, but contains not even twice the number of genes. The mitochondrial genome of the plant *Silene conica* is over 11.3 megabases (Mb)

in length, but contains only 30 genes (Sloan *et al.* 2012). Clearly plant mitochondrial genomes have grown exceptionally large by accumulating non-coding DNA, not by acquiring new genes.

One explanation for the differences in sizes between plant and animal genomes is the mutational hazard hypothesis (Lynch *et al.* 2006). The premise of this hypothesis is that mutations are more likely to be deleterious than beneficial, even in non-coding DNA. One strategy to minimize genetic risk and maximize fitness when mutation rates are high is to reduce genome size and thus reduce the number of targets for a deleterious mutation. This hypothesis can explain the dynamics of animal mitochondrial genomes quite nicely: animal mitochondrial genomes have high mutation rates and therefore have selective pressure to reduce the target for potentially deleterious mutations by reducing genome size as much as possible. It may appear that the mutational hazard hypothesis can explain plant mitochondrial genomes as well: plant mitochondrial genomes have very low mutation rates, alleviating the selective pressure to maintain small genomes. Plant mitochondrial genomes have low mutation rates in genes, however it is difficult to calculate mutation rates in non-gene regions because rearrangements and large indels make it difficult to align sequences of common origin among taxa (Christensen 2013). Some plants, however, such as *Silene conica* described above, have large and expanded mitochondrial genomes as well as increased mutation rates in genes (Mower *et al.* 2007, Sloan *et al.* 2012). The mutational hazard hypothesis is then insufficient to explain plant mitochondrial genome dynamics.

Another hypothesis is that the DNA repair pathways have shaped the evolution of plant mitochondrial genomes (Christensen 2014). This hypothesis argues that, in contrast

to animal mitochondria, double-strand break repair is the predominant method of DNA repair in plant mitochondria, and that this reliance on double-strand break repair has shaped genome evolution and structure. Double-strand break repair can have three different outcomes: long-homology based repair which is accurate, short-homology based repair which can be accurate but can also introduce expansions or rearrangements, and non-homologous repair which is inaccurate and introduces expansions or rearrangements. There is strong selective pressure to complete accurate repair in genes, however the non-coding regions can be free to expand, contract, rearrange, and diverge.

DNA repair in plant mitochondria

The powerhouse of the cell is a dangerous place to store DNA. Reactive oxygen species (ROS), which can damage DNA (Cadet and Wagner 2013), are byproducts of the electron transport that drives oxidative ATP formation in mitochondria (Murphy 2009). Despite the high potential for damage to mtDNA and the low mutation rate of plant mitochondrial genes, we have only a partial understanding of the mechanisms of DNA repair in plant mitochondria. Many of the proteins responsible for essential steps in DNA repair have not been characterized or confirmed to act in the mitochondria.

Plant mitochondria have evolved specific repair pathways for some types of DNA lesions. The base excision repair pathway is known to be active in plant mitochondria (Boesch *et al.,* 2009). This pathway is initiated by a DNA glycosylase that binds to a specific DNA lesion and excises the damaged base, leaving an abasic site. Several different DNA glycosylases have been characterized in plant mitochondria: OGG1 recognizes and excises 8-oxo-guanine (Dany and Tissier 2001, García-Ortiz *et al.* 2001), Neil1/2 recognizes and excises 5-hydroxy deoxyuracil (Ferrando *et al.* 2018), and UNG

recognizes and excises uracil in DNA (Boesch *et al.* 2009). The DNA backbone at the abasic site can then be cleaved by an apurinic/apyrimidinic endonuclease, allowing a new, undamaged, nucleotide to be polymerized in its place (see Figure 1.1).

Many commonly occurring DNA repair pathways are apparently absent in plant mitochondria. There is currently no evidence for the existence of nucleotide excision repair (NER) for bulky adducts, nor for the photoreactivation of pyrimidine dimers. MSH1 is a homolog of the *E.coli* mismatch repair enzyme MutS. MSH1 is targeted to plant mitochondria and plastids (Christensen *et al.* 2005). However, there is no evidence that MSH1 initiates mismatch repair through the canonical MutS-catalyzed pathway (Abdelnoor *et al.* 2003). The plant MSH1 has a mismatch-recognition domain and DNA-binding domain similar to MutS, but it also has a novel GIY-YIG endonuclease not found in animal, fungal or bacterial MutS homologs. It has recently been shown that the GIY-YIG endonuclease domain of MSH1 binds to branched DNA structures such as D-loops and Holliday junctions and shows no endonuclease activity by itself (Fukui *et al.* 2018). However, MSH1 also contains the MutS domain I, which recognizes distortions of the DNA backbone agnostic to the specific interactions of base pairs causing the distortions. This mismatch-recognition domain then binds to the distorted double-stranded DNA as a dimer (Lamers *et al.* 2000). Some GIY-YIG endonucleases act as homodimers to cut DNA (Liu *et al.* 2013), while others act in complex with other protein subunits (Gaur *et al.* 2015). MSH1 therefore may have the ability to recognize and bind to non-specific DNA lesions as a dimer, and then to create a double strand break either as a dimer or as part of a protein complex.

We hypothesize that the DNA lesions normally repaired through the "missing" repair pathways, NER and MMR, are instead recognized by the mismatch-recognition domain of MSH1 and shunted into the double-strand break repair pathway by a double-strand break made by the GIY-YIG endonuclease domain of MSH1 (see Figure 1.2).

Double-strand breaks, whether initiated by MSH1 or directly induced by DNA damage, can be repaired by several different pathways (see Figure 1.3). If a homologous template is available, a double-strand break can be repaired by homologous recombination (Jasin and Rothstein 2013). In this pathway, the 5' ends at the site of the double-strand break are resected, leaving 3' overhangs. The single-strand binding proteins WHY2 (Zaegel *et al.* 2006) and OSB (Maréchal *et al.* 2009) bind to these overhangs and protect the ssDNA and allow for the recruitment of the RECA proteins, RECA2 and RECA3 (Miller-Messmer *et al.* 2012). The RECA proteins aid in strand invasion of the homologous template, allowing the 3' overhang to anneal to the template and provide a priming site for DNA polymerase to begin DNA synthesis. Once the gap has been filled in, ligation can occur and Holliday junctions will be resolved. This pathway will accurately repair the DNA at a double strand break, but requires a significant region of sequence homology between the broken DNA and the repair template.

If there is a homologous template that is long enough for one side of a double strand break to do a single strand invasion, but short enough that the other side of the break cannot find significant homology, then a double-strand break can be repaired by break-induced replication (BIR) (Kramara *et al.* 2018). This situation may occur near the intermediate repeats of a plant mitochondrial genome. In *Arabidopsis*, these repeats are

between 50-600bp long and are present at 2-4 copies in the genome (Arrieta-Monteil *et al.* 2009). If a double-strand break were to occur in or near one of these repeats, one end of the break could find homology in another copy of the repeat, while the other end of the break would not. Similar to homologous recombination, BIR begins with the resection of the 5' and the creation of a 3' overhang. This overhang will invade a homologous sequence and provide a primer for DNA synthesis. Unlike homologous recombination, the other end of the double-strand break cannot anneal to a homologous template or prime DNA synthesis. This means that a Holliday junction will not form and DNA synthesis of the annealed end will continue unabated. The final outcome of BIR is a chimeric molecule containing both one end of the double-strand break and the region of DNA that the broken end annealed to, leading to a duplication of that region. This type of repair is rare in wild-type mitochondria, but can result in many rearrangements when enzymes involved in recombination are impaired or absent. While these events may be rare in population, this process may account for the expansion and rearrangements of mitochondrial genomes over evolutionary time.

If no homologous template is available, a double-strand break can be repaired through microhomology-mediated end-joining (MMEJ) or non-homologous end-joining (NHEJ). In MMEJ, 3' overhangs at a double-strand break can transiently anneal at short homologies or bind to the ssDNA binding protein SSB. Transient annealing or SSB binding can allow a DNA polymerase to bind and begin DNA synthesis (García-Medel *et al.* 2019). This can result in accurate repair, small indels, or rearrangements depending on the nature of the double-strand break and the available microhomology. If a double-strand break has blunt ends, it can simply be ligated together in the process of NHEJ. In

many organisms, NHEJ is mediated by homologs of Ku70/80 (Davis and Chen 2013), but no such homologs have been characterized in plant mitochondria. Non-homologous end joining can be accurate if the two ends have been blunted and no nucleotides have been lost, but can also cause deletions if ends have been damaged or processed, or can cause rearrangements if two ends from different regions of the genome are joined.

A double-strand break anywhere in the genome can be repaired by any of these mechanisms, but inaccurate repair of essential genes will be heavily selected against. In contrast, inaccurate repair of double-strand breaks may persist in non-conserved regions of the genome that are neutrally evolving via the process of genetic drift. The prevalence of double strand break repair in plant mitochondrial genomes followed by different outcomes of selection in coding versus non-coding regions provides a model that can explain many of the seemingly anomalous aspects of plant mitochondrial dynamics.

Plant Mitochondrial Genome Structure

Plant mitochondrial genomes can be aligned to circles and are often displayed in the literature as circular molecules. However, pulse-field gel electrophoresis and electron microscopy have shown that these genomes consist of linear, branched, and sigma shaped molecules of different sizes (Bendich 1996; Backert 1996). These molecules contain overlapping DNA sequences, allowing the genomes to be mapped as master circles, but there is no evidence that these master circles exist as molecules in the mitochondria of vascular plants. The presence of sigma shaped molecules and branched rosettes indicates that vascular plant mitochondrial genomes likely replicate by recombination-dependent or rolling-circle replication, similar to bacteriophage T4 (Backert and Börner 2000)(see Figure 1.4). The mitochondria of some non-vascular plants, such as the liverwort

Marchantia polymorpha have circular, genome sized molecules (Oda *et al.* 1992). Even in these species that contain circular, genome sized molecules, most mitochondrial DNA is contained in smaller linear molecules, with large circles being around 5% of the total DNA content (Oldenburg and Bendich 2001).

All post-embryogenesis cell division in a plant occurs in meristematic tissue. Plant growth in mature, differentiated tissue is the result of cell elongation and expansion, not of cell division. In the shoot apical meristem (SAM), and by extension the floral meristem that develops from the SAM, the mitochondria undergo extensive fusion to form a large, cage-like mitochondrion that surrounds the nucleus and can divide and segregate during meristematic cell division (Seguí-Simarro *et al.* 2008). The centralization of many copies of the mitochondrial genome within the cage-like fused mitochondrion ensures that there are ample templates for DNA repair by homology-based mechanisms, providing the raw material necessary to accurately repair damage. The mitochondrial genomes in the female gametes of a flowering plant originate in the SAM, so there is strong selection to maintain an undamaged mitochondrial genome to pass on to the next generation. In contrast, the mitochondria within the root apical meristem (RAM) do not fuse to form a large, cage-like mitochondrion and instead remain small and "sausage-shaped" like the mitochondria of terminally differentiated cells (Seguí-Simarro and Staehelin 2009). None of the mitochondrial genomes derived from the RAM will be passed on to the next generation, so the selection for accurate repair is much weaker. Despite the absence of a large mitochondrial fusion, cells in the RAM perform sufficient mtDNA replication and segregation to provide each root cell with a population of functioning mitochondria. Clearly, mtDNA replication and repair is still possible by the

small-scale fusion of mitochondria to find homologous sequence in subgenomic molecules.

In terminally differentiated cells, mtDNA degrades over time, likely due to DNA damage caused by ROS generated during respiration. Extremely long PCR products (~11kb) are amplified less in older cells relative to smaller PCR products, but *in vitro* treatment of DNA from these older cells with DNA repair enzymes restores the ability to amplify long PCR products (Kumar *et al.* 2014). This indicates that much of the mtDNA damage accumulated as cells age is due to DNA lesions, not from structural damage or loss of subgenomes. In a physiologically active cell, why would a plant allow it's organellar genomes to degrade? One explanation is that in mature leaves, mitochondria transition from a state of high respiration and ATP production, to a state in which their primary function is to detoxify glycolate generated during photorespiration of the chloroplasts (Oldenburg *et al.* 2013). The gene products encoded in the mitochondria are those focused on respiration, while gene products encoded in the nucleus and targeted to the mitochondria are involved in more diverse biochemistry, such as glycolate detoxification. However, it has been shown that mature leaves continue to perform respiration at significant rates up until senescence of the leaf (Hardwick *et al.* 1968). Another explanation for this phenomenon is that the mRNAs that code for the respiration proteins may be exceptionally stable. Chloroplasts also undergo a degradation of DNA in mature leaves, but it has been shown that the mRNA of the chloroplast gene *psbA* becomes more stable and persists longer in mature leaves (Klaff and Gruissen 1991), allowing cpDNA to degrade without a loss of translational ability. One final explanation that has not yet been addressed is the possibility of transcription-coupled repair in mature

leaves. In mature leaves when mitogenomes are dispersed throughout the cell, the stalling of an RNA polymerase during transcription and the recruitment of a transcription-coupled repair complex could signal a mitochondrion to initiate fusion and find a template for repair. Repairing only the transcribed regions of the genome would allow a mature leaf to maintain its respiratory machinery, but the non-gene regions would accumulate damage and degrade. RNA-seq data from several different ages and tissue types of *Arabidopsis thaliana* shows that there are transcripts of mitochondrial mRNA available in mature leaves, including in senescing leaves (personal analysis of SRA Bioproject PRJNA314076, Klepikova *et al.* 2016), indicating that increases in mRNA stability or transcription coupled repair are plausible. Regardless of why mtDNA is able to degrade in mature leaves, it is clear that there is a fundamental difference in both the methods and outcomes of DNA repair between meristematic and differentiated cells.

In fused meristematic mitochondria, recombination between subgenomic molecules to initiate replication can produce differing isoforms of the genome. Southern blot analysis and Pac-Bio long-sequencing reads have revealed the existence of multiple different isoforms at repeated regions in the genome (Dawson *et al.* 1986, Kozik *et al.* 2019). These different isoforms can be formed due to recombination at repeated sequences, if the repeats are long enough to provide sufficient homology. Isomerization of the genome can result in loss of large regions of DNA by genetic drift if an subgenomic unit does not contain a region under selective pressure (Wu and Sloan 2019, Chang *et al.* 2013). In *Arabidopsis thaliana*, there are two large repeats that can produce the major isoforms of the mitochondrial genome (Klein *et al.* 1994). In addition to these large repeats, there are dozens of smaller repeats between 50 and 500 base pairs. These

intermediate repeats can recombine if there is a deficiency in the mitochondrial recombination machinery (Arrieta-Montiel *et al.* 2009), but do not commonly recombine in wild-type plants. MSH1 has been implicated in homology surveillance during recombination (Shedge *et al.* 2007). Thus, the rarity of recombination at the intermediate repeats allows us to infer the length of homology necessary to successfully perform recombination. Homologies that are the length of the intermediate repeats or shorter are apparently too short to escape the homology surveillance of MSH1, while the 4kb length of Large Repeat 2 is long enough. The length of homology necessary to initiate recombination can determine what happens to the DNA ends at a double-strand break and which of the double-strand break repair pathways are used to repair the break

RESEARCH GOALS

This dissertation is an investigation of the hypothesis that double strand break repair and recombination have shaped the evolution of plant mitochondrial genomes. The following chapters will detail different avenues of examining the predictions and implications of this hypothesis.

In chapter 2 we will show that synonymous substitutions in plant mitochondrial genes are not truly neutral. Due to heavy selection against deleterious mutations in genes, gene conversion by homologous recombination can cause a selective sweep of nearby synonymous mutations, lowering the mutation rate of synonymous substitutions in genes.

In chapter 3 we analyze the intermediate repeats of plant mitochondrial genomes. We show that these repeats emerge and become common in land plants. We hypothesize that the desiccation that occurs during seed or spore formation caused an increase in

double strand breaks compared to bryophytes and algae, providing more opportunities for break induced replication to create duplications of the genome. We also show that the pathways that create repeats by DNA duplication are rare, but can occur suddenly and become fixed in a population during speciation.

In chapter 4 we examine a line of *Arabidopsis thaliana* that is deficient in base excision repair and we test the hypothesis that these DNA lesions will now be repaired by MSH1 initiated double strand breaks.

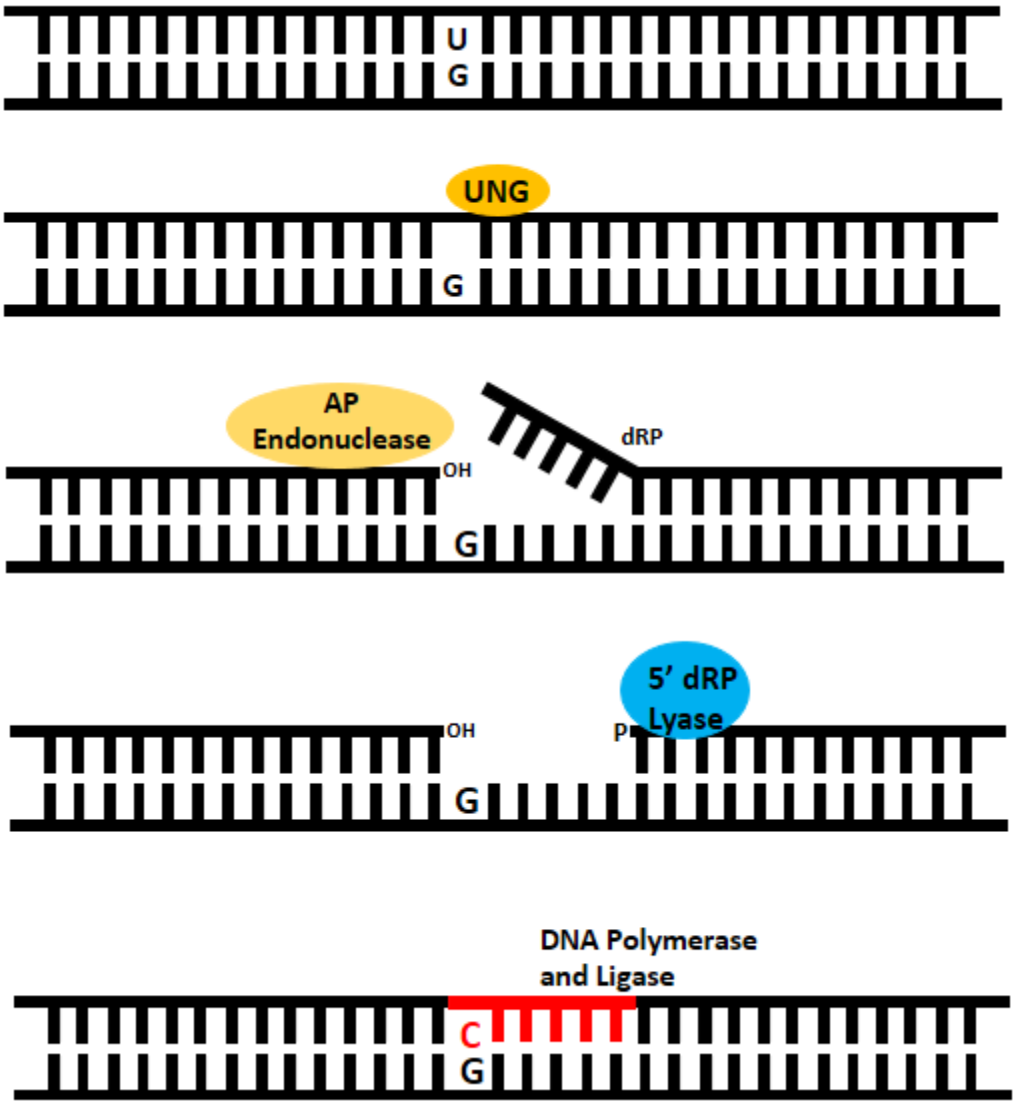

Figure 1.1: The Base Excision Repair Pathway. During Base Excision Repair, a DNA glycosylase such as UNG will recognize and bind to a specific DNA lesion and excise the base, leaving an abasic site. An AP endonuclease will cut the DNA backbone at the abasic site and a patch of DNA will be removed by a lyase. DNA polymerase can then bind and synthesize new DNA, correcting the former lesion, and the newly synthesized DNA will be ligated to the rest of the DNA molecule.

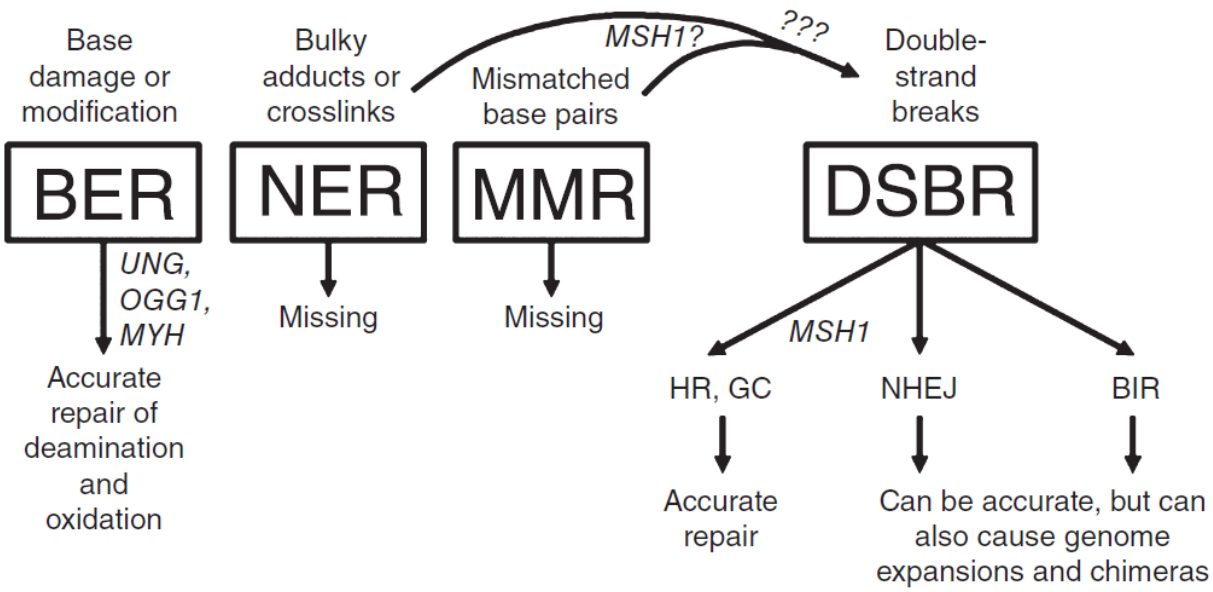

Figure 1.2 [Reproduced from (Christensen 2018)]: A model of plant mitochondrial genome evolution by double-strand break repair. DNA lesions may be recognized by MSH1 and shunted into the DSBR pathway. Double-strand breaks can be repaired by homologous recombination (HR), gene conversion (GC), non-homologous end-joining (NHEJ), or break-induced replication (BIR). Selection ensures accurate repair in conserved regions, while non-conserved "junk" can expand and rearrange.

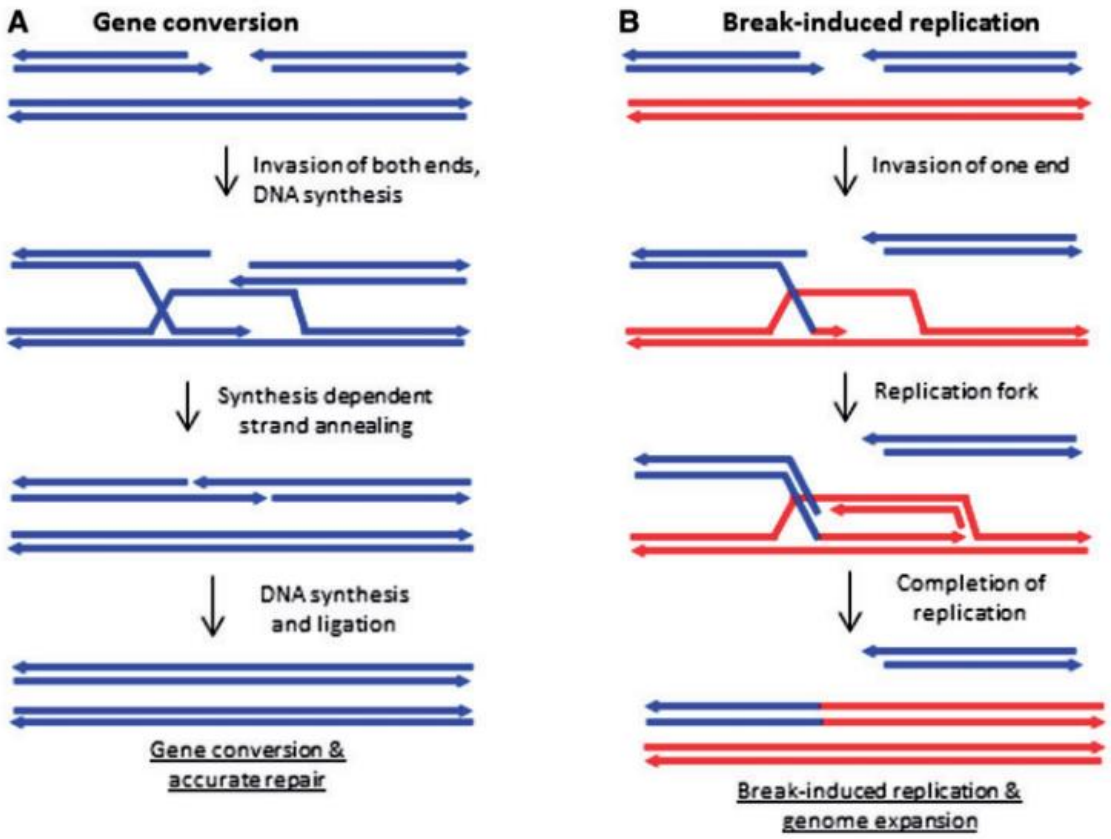

Figure 1.3 [Reproduced from (Christensen 2013)]: Double-strand break repair by homologous recombination of break-induced replication. (A) If a long homologous template is available, a double-strand break can be accurately repaired by gene conversion by homologous recombination. (B) If only one end of a double-strand break can find a homologous template, break-induced replication can occur, leading to rearrangements and expansions.

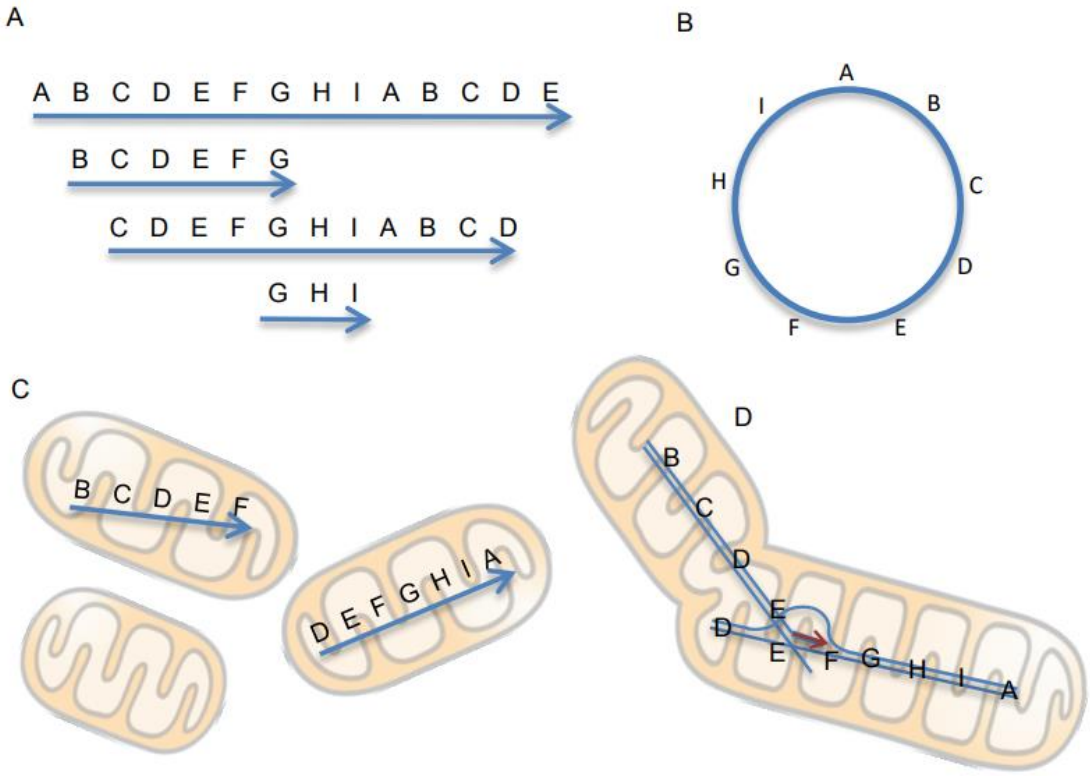

Figure 1.4 [Reproduced from (Arimura 2018)]: Structure of mitochondrial DNA and genome replication by recombination dependent replication. (A) Plant mitochondrial genomes consist of many subgenomic linear molecules with overlapping sequences. (B) The overlapping sequences of the linear molecules allows the construction of master circles, but no such molecules exist *in vivo*. (C) Individual mitochondria can contain different subgenomic molecules, or no DNA at all. (D) Mitochondrial fusion brings subgenomic molecules together, providing a template for recombination dependent replication or DNA repair.

REFERENCES

Abdelnoor RV, Yule R, Elo A, Christensen AC, Meyer-Gauen G, Mackenzie SA. 2003. Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. Proc Natl Acad Sci USA. 100(10): 5968-5973

Arimura SI. 2018. Fission and Fusion of Plant Mitochondria, and Genome Maintenance. Plant Physiology. 176(1): 152-161

Arrieta-Montiel MP, Shedge V, Davila J, Christensen AC, Mackenzie SA. 2009. Diversity of the Arabidopsis Mitochondrial Genome Occurs via Nuclear-Controlled Recombination Activity. Genetics. 183(4): 1261-1268

Backert S, Dörfel P, Lurz R, Börner T. 1996. Rolling-Circle Replication of Mitochondrial DNA in the Higher Plant Chenopodium album (L.). Molecular and Cellular Biology. 16(11): 6285-6294

Backert S, Börner T. 2000. Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant Chenopodium album (L.). Curr Genet. 37: 304-314.

Bendich AJ. 1993. Reaching for the ring: the study of mitochondrial genome structure. Curr Genet. 24(4): 279-290

Bendich AJ. 1996. Structural Analysis of Mitochondrial DNA Molecules from Fungi and Plants Using Moving Pictures and Pulsed-field Gel Electrophoresis. Journal of Molecular Biology. 255(4): 564-588

Boesch P, Ibrahim N, Paulus F, Cosset A, Tarasenko V, Dietrich A. 2009. Plant mitochondria possess a short-patch base excision DNA repair pathway. Nucleic Acids Research. 37(17): 5690-5700

Cadet J, Wagner JR. 2013. DNA Base Damage by Reactive Oxygen Species, Oxidizing Agents, and UV Radiation. Cold Spring Harb Perspect Biol. 5(2):a012559

Chang S, Wang Y, Lu J, Gai J, Li J, Chu P, Guan R, Zhao T. 2013. The Mitochondrial Genome of Soybean Reveals Complex Genome Structures and Gene Evolution at Intercellular and Phylogenetic Levels. PLoS ONE 8(2): e56502.

Christensen AC, Lyznik A, Mohammed S, Elowsky CG, Elo A, Yule R, Mackenzie SA. 2005. Dual-Domain, Dual-Targeting Organellar Protein Presequences in Arabidopsis Can Use Non-AUG Start Codons. The Plant Cell. 17(10) 2805-2816

Christensen AC. 2013. Plant Mitochondrial Genome Evolution Can Be Explained by DNA Repair Mechanisms. Genome Biol Evol. 5(6): 1079-1086

Christensen AC. 2014. Genes and Junk in Plant Mitochondria – Repair Mechanisms and Selection. Genome Biol Evol. 6(6): 1448-1453

Christensen AC. 2018. Mitochondrial DNA Repair and Genome Evolution. Annual Plant Reviews. 50, 11-32

Dany AL, Tissier A. 2001. A functional OGG1 homologue from Arabidopsis thaliana. Mol Genet Genomics. 265(2): 293-301

Davis AJ, Chen DJ. 2013. DNA double strand break repair via non-homologous end-joining. Transl Cancer Res. 2(3): 130-143

Dawson AJ, Hodge TP, Isaac PG, Leaver CJ, Lonsdale DM. 1986. Location of the genes for cytochrome oxidase subunits I and II, apocytochrome *b*, α-subunit of the F_1 ATPase and the ribosomal RNA genes on the mitochondrial genome of maize (*Zea mays* L.). Curr Genet. 10: 561-564

Ferrando B, Furlanetto ALDM, Gredilla R, Havelund JF, Hebelstrup KH, Møller IM, Stevnsner T. 2018. DNA repair in plant mitochondria – a complete base excision repair pathway in potato tuber mitochondria. Physiologia Plantarum. doi:10.1111/ppl.12801

Fukui K, Harada A, Wakamatsu T, Minobe A, Ohshita K, Ashiuchi M, Yano T. 2018. The GIY-YIG endonuclease domain of Arabidopsis MutS homolog 1 specifically binds to branched DNA structures. FEBS Letters. 592: 4066-4077

García-Medel PL, Baruch-Torres N, Peralta-Castro A, Trasviña-Arenas CH, Torres-Larios A, Brieba LG. 2019. Plant organellar DNA polymerases repair double-stranded breaks by microhomology-mediated end-joining. Nucleic Acids Research. doi: 10.1093/nar/gkz039

García-Ortiz MV, Ariza RR, Roldán-Arjona T. 2001. An OGG1 orthologue encoding a functional 8-oxoguanine DNA glycosylase/lyase in Arabidopsis thaliana. Plant Mol Biol. 47(6): 795-804.

Gaur V, Wyatt HDM, Komorowska W, Szczepanowski RH, de Sanctis D, Gorecka KM, West SC, Nowotny M. 2015. Structural and Mechanistic Analysis of the Slx1-Slx4 Endonuclease. Cell Rep. 10(9): 1467-1476

Gray MW. 1999. Evolution of organellar genomes. Curr Opin Genet Dev. 9:678–687.

Hardwick K, Wood M, Woolhouse HW. 1968. Photosynthesis and Respiration in Relation to Leaf Age in Perilla frutescens (L.) Britt. New Phytol. 67: 79-86

Jasin M, Rothstein R. 2013. Repair of Strand Breaks by Homologous Recombination. Cold Spring Harb Perspect Biol. 5(11): a012740

Klaff P, Gruissem W. 1991. Changes in Chloroplast mRNA Stability during Leaf Development. The Plant Cell. 3(5): 517-539.

Klein M, Eckert-Ossenkop U, Schmiedeberg I, Brandt P, Unseld M, Brennicke A, Schuster W. 1994. Physical mapping of the mitochondrial genome of Arabidopsis thaliana by cosmid and YAC clones. The Plant Journal. 6(3): 447-455

Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. 2016. A high resolution map of the Arabidopsis thalina developmental transcriptome based on RNA-seq profiling. The Plant Journal. 88: 1058-1070

Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, Christensen AC. 2019 The alternative reality of plant mitochondrial DNA. BioRxiv. doi: 10.1101/564278

Kramara J, Osia B, Malkova A. 2018. Break-Induced Replication: The Where, The Why, and The How. Trends in Genetics. 34(7) 518-531

Kumar RA, Oldenburg DJ, Bendich AJ. 2014. Changes in DNA damage, molecular integrity, and copy number for plastid DNA and mitochondrial DNA during maize development. J Exp Bot. 65(22): 6425-6439.

Lamers MH, Perrakis A, Enzlin JH, Winterwerp HHK, de Wind NW, Sixma TK. 2000. The crystal structure of DNA mismatch repair protein MutS binding to a G·T mismatch. Nature. 407: 711-717.

Liu X, Feng Y, Liu JZ, Chen Y, Pham K, Deng H, Hirschi KD, Wang X, Cheng N. 2013. Structural insights into the N-terminal GIY-YIG endonuclease activity of Arabidopsis glutaredoxin AtGRXS16 in chloroplasts. PNAS. 110(23): 9565-9570.

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. Science. 24;311(5768):1727-1730

Maréchal A, Parent JS, Véronneau-Lafortune F, Joyeux A, Lang BF, Brisson N. 2009. Whirly proteins maintain plastid genome stability in Arabidopsis. PNAS. 106(34): 14693-14698

Miller-Messmer M, Kühn K, Bichara M, Le Ret M, Imbault P, Gualberto JM. 2012. RecA-Dependent DNA Repair Results in Increased Heteroplasmy of the Arabidopsis Mitochondrial Genome. Plant Physiol. 158(1): 211-226

Morales-Ruiz T, Birincioglu M, Jaruga P, Rodriguez H, Roldan-Arjona T, Dizdaroglu M. 2003. Arabidopsis thaliana Ogg1 Protein Excises 8-Hydroxyguanine and 2,6-Diamino-4-hydroxy-5-formamidopyrimidine from Oxidatively Damaged DNA Containing Multiple Lesions. Biochemistry. 42(10) 3089-3095

Mower JP, Touzet P, Gummow JS, Delph LS, Palmer JD. 2007. Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol Biol. 7:135.

Murphy MP. 2009. How mitochondria produce reactive oxygen species. Biochem J. 417(1):1-13

Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, Akashi K, Kanegae T, Ogura Y, Kohchi T, *et al.* 1992. Gene organization deduced from the complete sequence of liverwort Marchantia polymorpha mitochondrial DNA. A primitive form of plant mitochondrial genome. J Mol Biol. 223(1): 1-7

Oldenburg DJ, Bendich AJ. 2001. Mitochondrial DNA from the liverwort *Marchantia polymorpha*: circularly permuted linear molecules, head-to-tail concatemers, and a 5' protein. J Mol Biol. 310(3): 549-562

Oldenburg DJ, Kumar RA, Bendich AJ. 2013. The amount and integrity of mtDNA in maize decline with development. Planta. 237: 603-617

Palmer JD, Herbon LA. 1988.Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol. 28:87-97

Seguí-Simarro JM, Coronado MJ, Staehelin LA. 2008. The Mitochondrial Cycle of Arabidopsis Shoot Apical Meristem and Leaf Primordium Meristematic Cells Is Defined by a Perinuclear Tentaculate/Cage-Like Mitochondrion. Plant Physiology. 148: 1380-1393

Seguí-Simarro JM, Staehelin LA. 2009. Mitochondrial reticulation in shoot apical meristem cells of Arabidopsis provides a mechanism for homogenization of mtDNA prior to gamete formation. Plant Signaling & Behavior. 4(3): 168-171

Shedge V, Arrieta-Montiel M, Christensen AC, Mackenzie SA. 2007. Plant Mitochondrial Recombination Surveillance Requires Unusual RecA and MutS Homologs. Plant Cell. 19(4): 1251-1264

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. 2012. Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. PLoS Biol. 10(1):e1001241

Taanman JW. 1999. The mitochondrial genome: structure, transcription, translation, and replication. Biochimica et Biophysica Acta – Bioenergetics. 1410(2):103-123

Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides. Nat Genet. 15(1):57-61

Wang DY, Zhang Q, Liu Y, Lin ZF, Zhang SX, Sun MX, Sodmergen. 2010. The Levels of Male Gametic Mitochondrial DNA are Highly Regulated in Angiosperms with Regard to Mitochondrial Inheritance. The Plant Cell. 22: 2402-2416

Wu Z, Sloan DB. 2019. Recombination and intraspecific polymorphism for the presence and absence of entire chromosomes in mitochondrial genomes. Heredity. 122:647-659

Zaegel V, Guermann B, Le Ret M, Andrés C, Meyer D, Erhardt M, Canaday J, Gualberto JM, Imbault P. 2006. The Plant-Specific ssDNA Binding Protein OSB1 Is Involved in the Stoichiometric Transmission of Mitochondrial DNA in Arabidopsis. The Plant Cell. 18: 3548-3563

CHAPTER 2

ARE SYNONYMOUS SUBSTITUTIONS IN PLANT MITOCHONDRIA NEUTRAL?

This chapter has been published: Wynn EL, Christensen AC. 2015. Are Synonymous Substitutions in Flowering Plant Mitochondria Neutral? J Mol Evol. 81(3-4): 131-135. doi: 10.1007/s00239-015-9704-x.

ABSTRACT

Angiosperm mitochondrial genes appear to have very low mutation rates, while non-gene regions expand, diverge, and rearrange quickly. One possible explanation for this disparity is that synonymous substitutions in plant mitochondrial genes are not truly neutral and selection keeps their occurrence low. If this were true, the explanation for the disparity in mutation rates in genes and non-genes needs to consider selection as well as mechanisms of DNA repair. *Rps14* is co-transcribed with *cob* and *rpl5* in most plant mitochondrial genomes, but in some genomes, *rps14* has been duplicated to the nucleus leaving a pseudogene in the mitochondria. This provides an opportunity to compare neutral substitution rates in pseudogenes with synonymous substitution rates in the orthologs. Genes and pseudogenes of *rps14* have been aligned among different species and the mutation rates have been calculated. Neutral substitution rates in pseudogenes and synonymous substitution rates in genes are significantly different, providing evidence that synonymous substitutions in plant mitochondrial genes are not completely neutral. The non-neutrality is not sufficient to completely explain the exceptionally low mutation rates in land plant mitochondrial genomes, but selective forces appear to play a small role.

INTRODUCTION

Synonymous substitution rates in angiosperm mitochondrial genes are about 10-fold lower than in the nuclear genes (Drouin *et al.* 2008; Richardson *et al.* 2013; Wolfe *et al.* 1987) and approximately 100-fold lower than in animal mitochondria (Palmer and Herbon 1988). This low rate appears to be a derived trait in land plants (Smith 2015). Synonymous substitutions are often used to calculate mutation rates in genes under the assumption that they are selectively neutral (Nei *et al.* 2010). It might also be expected that mutations in non-coding or nonessential regions would also be neutral, and this could provide an interesting comparison to synonymous substitution rates. However, the non-gene regions of land plant mitochondrial genomes expand and rearrange so quickly, and to such an extent, that it is difficult to align the non-gene regions outside of very closely related species (Christensen 2013, 2014; Darracq *et al.* 2010; Kubo and Newton 2008; Mower *et al.* 2007; Palmer and Herbon 1988; Richardson *et al.* 2013; Sloan *et al.* 2012; Smith and Keeling 2015). If the mutation rate in plant mitochondrial genomes is truly low, then why do the non-gene regions diverge so quickly? One possible part of the explanation may be that synonymous substitutions in angiosperm mitochondria are not selectively neutral, and therefore underestimate the mutation rate. If so, the explanation for the paradox of low mutation rates in genes and high mutation rates in junk may need to be explained not just by DNA repair and maintenance mechanisms, but by a further understanding of the role of selection on synonymous substitutions.

This possibility has been addressed (Sloan and Taylor 2010) using patterns of codon usage in mitochondrial genes. Their study concluded that selection on synonymous sites was neutral or nearly neutral, and that selective effects on synonymous sites were

too weak to explain the reduced substitution rates. They also identified a bias toward A-T bases and pyrimidines at synonymous sites, but this non-randomness is not fully understood. More recently, presumably neutral mutation rates in mitochondrial insertions of plastid DNA were measured, but were not able to be directly compared to homologous sequences under selection in mitochondria (Sloan and Wu 2014). Thus, the substitution rates of synonymous sites have never been directly compared to truly neutral substitution rates, such as the rates of homologous non-selected sequences. Such a comparison would provide a direct way of confirming that synonymous substitutions are truly neutral; however, the highly divergent nature of non-gene regions prevents proper alignment among lineages, and thus, there are very few opportunities for direct comparisons across diverse species.

Ribosomal protein small subunit 14 (*rps14*) is co-transcribed in many plant mitochondrial genomes (see Fig. 2.1) with ribosomal protein large subunit 5 (*rpl5*) and cytochrome b (*cob*) (Hoffmann *et al.* 1999; Quinones *et al.* 1996). In some lineages, a copy of *rps14* has been relocated to the nucleus and the protein is imported by mitochondria. In these lineages, the mitochondrial copy of *rps14* has become a pseudogene (Aubert *et al.* 1992; Figueroa *et al.* 1999; Ong and Palmer 2006). These pseudogenes accumulate frameshift mutations so are clearly non-functional and not under selection for protein coding capacity. Because both *rps14* genes and pseudogenes are co-transcribed with and located between *rpl5* and *cob*, large rearrangements of the area will be selected against, as *cob* would lose its promoter. These *rps14* pseudogenes are thus a unique example of a non-coding sequence that can still be aligned to homologous coding sequences across very diverse lineages. Therefore, *rps14* is a perfect candidate to

measure neutral mutation rates. In lineages with functional *rps14* genes, the synonymous substitution rate can be measured, while in lineages with *ψrps14* pseudogenes, the total substitution rate is the neutral mutation rate. These rates can be compared to find out if synonymous substitutions in plant mitochondrial genes are selectively neutral.

METHODS

Accession numbers of all sequences used are listed in Table 2.1. In a few species, the synteny of *cob* with *rpl5* and *rps14* was disrupted, but it was still possible to identify *rps14* or *ψrps14* just downstream of *rpl5*. The *ψrps14* pseudogenes were confirmed by the presence of internal stop codons or frameshifts. Four multiple alignments were used in this analysis: an alignment of the *rps14* sequences in all species analyzed, an alignment of the concatenated sequences of atp4, *rpl5*, and *cob* in all species analyzed, an alignment of the functional *rps14* sequences, and an alignment of the concatenated sequences of atp4, *rpl5*, and *cob* in only those species with a functional *rps14* gene.

There is also RNA editing by pentatricopeptide repeat (PPR) proteins in the analyzed genes in several of these species (Uchida *et al.* 2011). A PPR protein binds to an mRNA and edits a cytosine to a uracil. These edits may change the amino acid encoded. A mutation at an edited site, or in the binding sequence of the PPR protein, may appear synonymous at the DNA level, but change the final protein, or may appear non-synonymous at the DNA level but leave the protein sequence unchanged. To avoid confounding the analysis, edited codons and the 18 upstream nucleotides representing potential PPR binding sites under selection have been deleted from analysis.

Two phylogenetic trees were constructed: one using the concatenated sequences of atp4, *rpl5*, and *cob* from all species analyzed, and one using the concatenated

sequences of atp4, *rpl5*, and *cob* from only those species with a functional *rps14* gene. The atp4 gene was chosen because it is independently transcribed (Forner *et al.* 2007). All alignments and phylogenetic trees were constructed with Mega5 (Tamura *et al.* 2011).

Analysis of functional *rps14* genes was done using CodeML in PAML 4.8 implemented in PAMLX (Yang 2007). Branch lengths were calculated using synonymous substitutions, and the phylogenetic tree of the concatenated sequences of atp4, *rpl5*, and *cob* was used to set the topology. This was done separately using the multiple alignment of the *rps14* sequence including only species with functional *rps14* genes (Fig. 2.2A) and the multiple alignment of the concatenated sequences of atp4, *rpl5*, and *cob* including only species with a functional *rps14* gene (Fig. 2.2B). Taking the branch length of each terminal branch leading to a lineage on the *rps14* tree and dividing it by the length of the same branch on the atp4, *rpl5*, and *cob* tree provides a ratio of the synonymous substitution rate of *rps14* genes compared to the synonymous substitution rate of the other three genes.

Analysis of *ψrps14* pseudogenes was done using BaseML in PAML 4.8 implemented in PAMLX (Yang 2007), branch lengths were calculated using total substitutions, and the phylogenetic tree of the concatenated sequences of atp4, *rpl5*, and *cob* was used to set the topology. This was done using the multiple alignment of the *rps14* sequence including all species (Fig. 2.2C). A phylogenetic tree using CodeML as described above was made using the multiple alignment of the concatenated sequences of atp4, *rpl5*, and *cob* including all species analyzed (Fig. 2.2D). Taking the branch length of each terminal branch leading to a lineage with an *ψrps14* pseudogene on the *rps14* tree

and dividing it by the length of the same branch on the atp4, *rpl5*, and *cob* tree provides a ratio of the total substitution rate of the *ψrps14* pseudogene compared to the synonymous substitution rate of the other three genes. Species with functional *rps14* genes were included in these trees to avoid counting as much divergence before the pseudogenes became pseudogenes as possible. Indels were counted in all *rps14* sequences. Indel rates per site were calculated.

All alignments and tree files can be found at https://link.springer.com/article/10.1007%2Fs00239-015-9704-x#SupplementaryMaterial

RESULTS

If synonymous substitutions in plant mitochondria are not neutral, then the synonymous substitution rate would erroneously underestimate the neutral mutation rate. In this event, we would expect *rps14* genes to have a significantly lower synonymous substitution rate than the total substitution rate in an *ψrps14* pseudogene. Alignments were done for *ψrps14* of the chosen species as well as *rps14* genes for the chosen species. Alignments were also done for the concatenated sequences of atp4, *rpl5*, and *cob* for all chosen species in order to generate the trees shown in Fig. 2. Following alignments, we calculated both rates.

Terminal branch lengths for the genes were calculated using PAML 4.8 (Yang 2007), and are shown in Fig. 2.2. For *rps14* genes, the normalized neutral mutation rate is calculated by dividing the terminal branch length of the *rps14* tree by the terminal branch length of the atp4, *rpl5*, and *cob* tree, both calculated using synonymous substitutions per synonymous site. For *ψrps14* pseudogenes, the normalized neutral mutation rate is calculated by dividing the terminal branch length of the *rps14* tree (calculated using total

substitutions per site) by the terminal branch length of the atp4, *rpl5*, and *cob* tree (calculated using synonymous substitutions per synonymous site).

The neutral mutation rates normalized with the atp4, *rpl5* and *cob* genes are shown in Table 2.2 and Fig. 2.3. The average normalized neutral mutation rate of the functional *rps14* genes is 0.276, and the average normalized neutral mutation rate of the *ψrps14* pseudogenes is 1.32. Using a Student's t test, these rates are significantly different ($p = 0.0099$). One species, *Citrullus lanatus*, had branch lengths of zero for both *ψrps14* and atp4, *rpl5*, *cob*, and was excluded from analysis. Despite having no lineage specific substitutions when compared to neighboring species, *C. lanatus* differed by several indels.

In addition to substitutions, we also measured indel rates. Indels should be strongly selected against in functional genes, but neutral in pseudogenes. The *ψrps14* pseudogenes had an average indel rate of 0.011 indels per site. The *rps14* genes had an average indel rate of 0 indels per site. These rates are significantly different ($p = 0.00043$), as expected.

DISCUSSION

Because there is no selective pressure on a non-functional pseudogene, substitutions will be neutral. The availability of both genes and alignable pseudogenes of *rps14* allowed us to measure the neutral substitution rate directly and compare it to the synonymous substitution rate, often used as a proxy for the neutral rate. The normalized synonymous substitution rate of the *rps14* genes is significantly different from the neutral substitution rate of the *ψrps14* pseudogenes (Fig. 2.3; Table 2.2). Therefore, it can be

inferred that the number of observable synonymous substitutions in plant mitochondria is lower than we would expect in the absence of any selection.

One possible explanation for the apparent selection on synonymous substitutions is RNA stability and translation efficiency. If synonymous substitutions affect the stability of mitochondrial RNA or the association with the translation machinery, then there will be selective pressure to repair them even without a difference in the encoded protein. Another possibility is that mutational processes may be responsible for the A-T and pyrimidine biases in codon usage observed by Sloan and Taylor (2010), as well as the A-T bias in mutations of neutral insertions of plastid DNA in mitochondrial genomes (Sloan and Wu 2014). In other systems, it has been estimated that the rate of cytosine deamination which causes G-C to A-T transitions is at least 50-fold higher than deamination reactions that could cause A-T to G-C transitions (Friedberg *et al.* 2006). The oxidation of guanine to 8-oxo-guanine, which can result in G-C to T-A transversions, appears to occur in plant mitochondria as well (Christensen 2013; Markkanen *et al.* 2012; van Loon *et al.* 2010). These two processes may skew the overall mutational spectrum toward an A-T bias, resulting in the non-randomness at synonymous sites previously observed (Sloan and Taylor 2010; Sloan and Wu 2014).

Another possible explanation for the apparent selection on synonymous substitutions is that synonymous substitutions might be repaired simultaneously with non-synonymous substitutions via gene conversion if gene conversion tracts are long enough. In genes, the selective pressure on deleterious mutations is very high, so repaired mutations should be frequent. In the pseudogene, there will not be selection to repair

mutations, so nearby neutral mutations will not be repaired as a result of a selective sweep.

The low mutation rate in land plant mitochondrial genes compared to non-genes does not appear to be due to differences in repair processes available, but is likely due to differences in selection on the repaired products (Christensen 2013, 2014). Gross rearrangements or even small indels would be strongly selected against in gene sequences, while they would not be selected against in non-genes, including pseudogenes. These events appear to be common on evolutionary timescales, explaining the large divergence of non-coding sequences.

This study is the first direct comparison of plant mitochondrial synonymous substitution rates with a neutral substitution rate in homologous pseudogenes. Although we have found that synonymous substitutions are not completely neutral, we still concur with the conclusion of Sloan and Taylor (2010) that the non-neutrality is not sufficient to explain the large disparity between the low mutation rates in genes and the much higher mutation, rearrangement, and expansion rates of the non-coding sequences in plant mitochondria.

Figure 2.1: A map showing the three co-transcribed mitochondrial genes, *rpl5*, *rps14*, and *cob*. These three genes are syntenic in all the species of angiosperm examined. A single promoter has been identified in several species (Forner *et al.* 2007; Hoffmann *et al.* 1999; Quinones *et al.* 1996) indicated at left

Table 2.1: Accession numbers for sequences used in analysis.

Species	*atp4*	*rpl5*	*cob*	*rps14*
Citrullus lanatus	>gi|295311632:c365914-365318	>gi|295311632:274340-274897	>gi|295311632:275820-276992	>gi|37896208|gb|AY305267.1|
Cucurbita pepo	>gi|295311672:471138-471713	>gi|295311672:83088-83645	>gi|295311672:84568-85740	>gi|37896209|gb|AY305268.1|
Cucumis sativus	>gi|346683357:419518-420114	>gi|346683357:c1190589-1189999	>gi|346683357:c378504-377338	>gi|31322689|gb|AY258274.1|
Vigna angularis	>gi|501594995:214689-215273	>gi|501594995:263314-263871	>gi|501594995:265075-266256	>gi|501594995:263875-264177
Lotus japonicus	>gi|387866040:c98133-97546	>gi|387866040:c276862-276305	>gi|387866040:c274684-273500	>gi|387866040:c276301-276061
Mimulus guttatus	>gi|391348915:c374983-374393	>gi|391348915:c146313-145759	>gi|391348915:c144016-142829	>gi|391348915:c145757-145455
Carica papaya	>gi|224020948:c290564-28998	>gi|224020948:c322118-321561	>gi|224020948:c319915-318734	>gi|224020948:c321559-321257
Brassica napus	>gi|112253843:c41887-41309	>gi|112253843:202531-203088	>gi|112253843:204898-206079	>gi|1524184|emb|X63653.1|
Arabidopsis thaliana	>gi|13984|emb|X67105.1|	>gi|26556996:57774-58331	>gi|26556996:60235-61416	>gi|14340|emb|X65123.1|
Vitis vinifera	>gi|224365609:274514-275110	>gi|224365609:175173-175724	>gi|224365609:177407-178588	>gi|224365609:175672-176028
Spirodela polyrhiza	>gi|387164694:119785-120315	>gi|387164694:97522-98076	>gi|387164694:99277-100458	>gi|387164694:98082-98340
Phoenix dactylifera	>gi|372450205:c249115-248528	>gi|372450205:10114-10680	>gi|372450205:11928-13109	>gi|372450205:10661-10984
Triticum aestivum	>gi|81176508:c197162-196584	>gi|81176508:c30347-29778	>gi|81176508:c64318-63122	>gi|27803141|emb|AJ535507.1|
Oryza sativa	>gi|89280701:c18996-18403	>gi|194033210:c343465-342899	>gi|89280701:c306002-304809	>gi|89280701:c424972-424696
Amborella trichopoda	>gb|KF754803.1|:652621-653210	>gb|KF754803.1|:2010747-2011304	>gb|KF754803.1|:2013306-2014494	>gb|KF754803.1|:2011306-2011608
Cycas taitungensis	>gi|166895601:c309686-309093	>gi|166895601:81967-82545	>gi|166895601:83720-84916	>gi|166895601:82550-82852

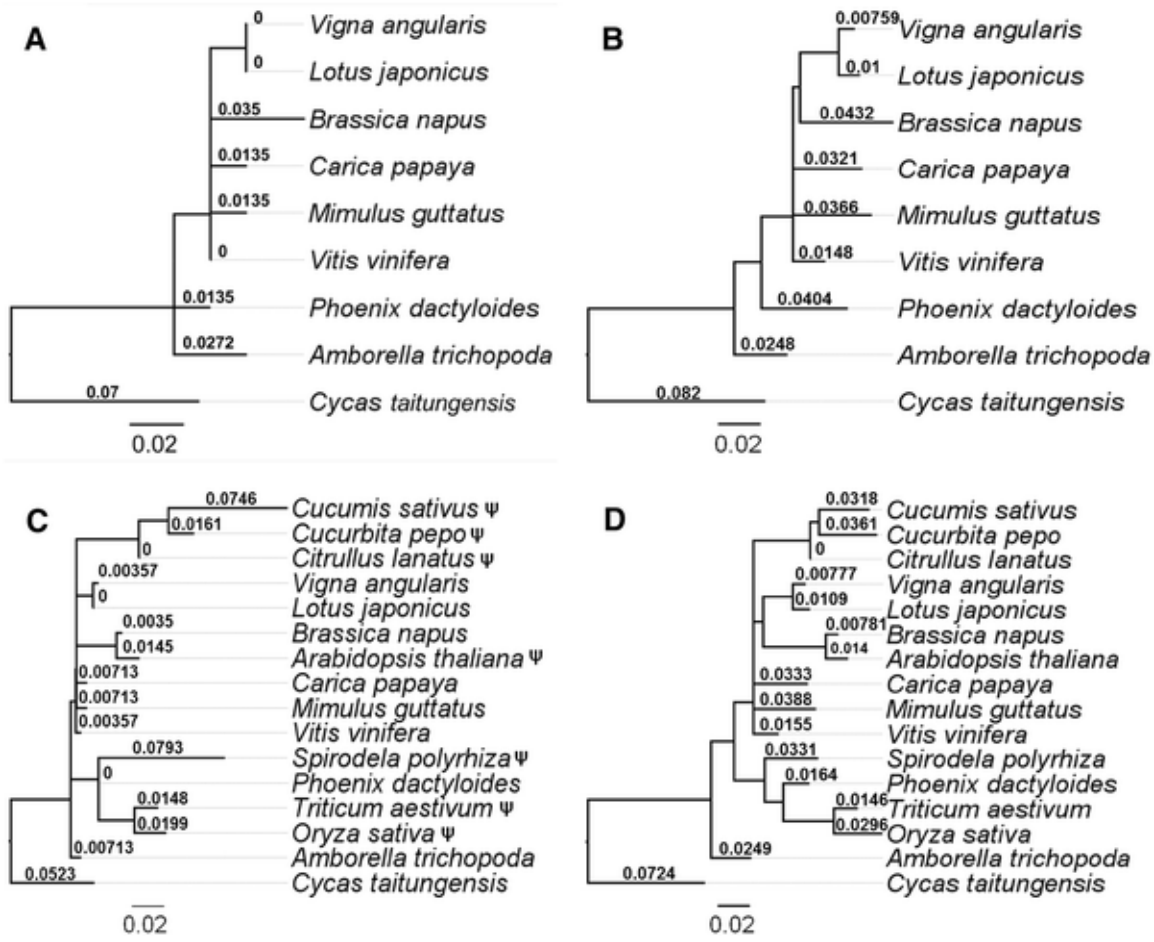

Figure 2.2: Phylogenetic trees with terminal branch lengths calculated using PAML as described in methods. A) Phylogenetic tree of functional *rps14* genes using synonymous substitutions. B) Phylogenetic tree of atp4, *rpl5*, and *cob* using synonymous substitutions and including only species with functional *rps14* genes. C) Phylogenetic tree of both *rps14* genes and *ψrps14* pseudogenes using total substitutions. Species with *ψrps14* pseudogenes are labeled with a *ψ*. D) Phylogenetic tree of atp4, *rpl5*, and *cob* using synonymous substitutions and including all species. For all trees, topology is based on an initial tree of atp4, *rpl5*, and *cob* sequences

Table 2.2: Synonymous substitution rates in *rps14* genes and substitution rates in *ψrps14* pseudogenes, relative to synonymous substitution rates in atp4, *rpl5*, and *cob* in the same species. Rates were calculated as described in methods, using the terminal branch lengths shown in Fig. 2.2. *Citrullus lanatus* was excluded from analysis because both branch lengths were zero

Species	Rate
rps14 genes	
Brassica napus	0.81
Carica papaya	0.421
Lotus Japonicus	0
Vigna angularis	0
Vitis vinifera	0
Phoenix dactyloides	0.334
Mimulus guttatus	0.369
Mean ± standard error	0.276 ± 0.106
ψrps14 pseudogenes	
Oryza sativum	0.672
Triticum aestivum	1.01
Cucurbita pepo	0.446
Citrullus lanatus	N/A
Arabidopsis thaliana	1.04
Cucumis sativus	2.35
Spirodela polyrhiza	2.4
Mean ± standard error	1.32 ± 0.315

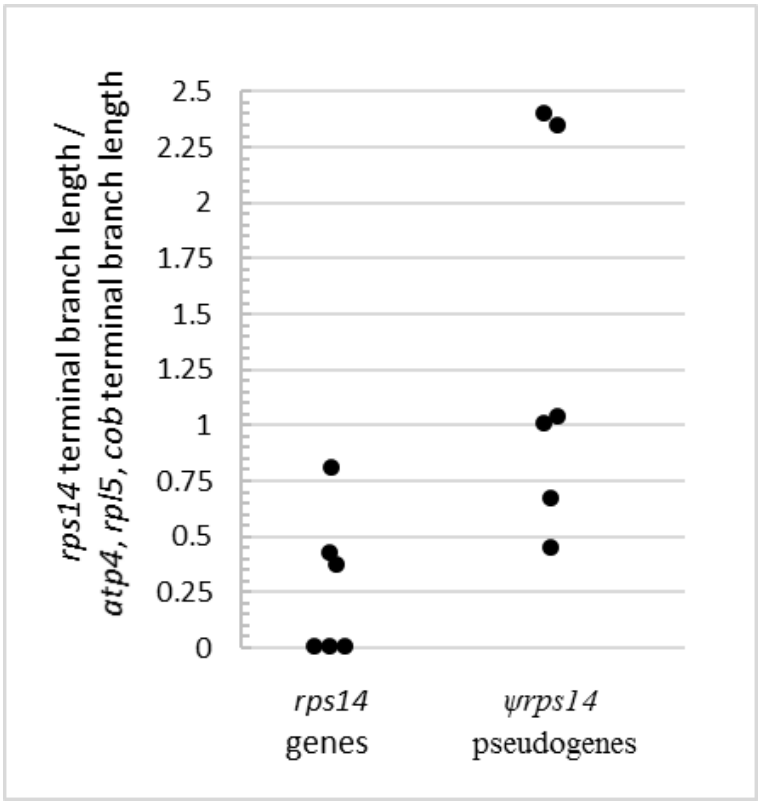

Figure 2.3: Comparison of the neutral mutation rate of species with functional *rps14* genes and species with *ψrps14* pseudogenes. Rates are from Table 2.2. This figure has been edited from the original publication for clarity.

REFERENCES

Aubert D, Bisanz-Seyer C, Herzog M (1992) Mitochondrial *rps14* is a transcribed and edited pseudogene in Arabidopsis thaliana. Plant Mol Biol 20:1169

Christensen AC (2013) Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. Genome Biol Evol 5:1079

Christensen AC (2014) Genes and junk in plant mitochondria-repair mechanisms and selection. Genome Biol Evol 6:1448

Darracq A, Varre JS, Touzet P (2010) A scenario of mitochondrial genome evolution in maize based on rearrangement events. BMC Genomics 11:233

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

Figueroa P, Gomez I, Carmona R, Holuigue L, Araya A, Jordana X (1999) The gene for mitochondrial ribosomal protein S14 has been transferred to the nucleus in Arabidopsis thaliana. Mol Gen Genet 262:139

Forner J, Weber B, Thuss S, Wildum S, Binder S (2007) Mapping of mitochondrial mRNA termini in Arabidopsis thaliana: t-elements contribute to 5′ and 3′ end formation. Nucleic Acids Res 35:3676

Friedberg EC, Walker GC, Siede W, Wood RD, Schultz RA, Ellenberger T (2006) DNA repair and mutagenesis. ASM Press, Washington, DC

Hoffmann M, Dombrowski S, Guha C, Binder S (1999) Cotranscription of the *rpl5-rps14-cob* gene cluster in pea mitochondria. Mol Gen Genet 261:537

Kubo T, Newton KJ (2008) Angiosperm mitochondrial genomes and mutations. Mitochondrion 8:5

Markkanen E, Hubscher U, van Loon B (2012) Regulation of oxidative DNA damage repair: the adenine:8-oxo-guanine problem. Cell Cycle 11:1070

Mower JP, Touzet P, Gummow JS, Delph LF, Palmer JD (2007) Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol Biol 7:135

Nei M, Suzuki Y, Nozawa M (2010) The neutral theory of molecular evolution in the genomic era. Annu Rev Genomics Hum Genet 11:265

Ong HC, Palmer JD (2006) Pervasive survival of expressed mitochondrial *rps14* pseudogenes in grasses and their relatives for 80 million years following three functional transfers to the nucleus. BMC Evol Biol 6:55

Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol 28:87

Quinones V, Zanlungo S, Moenne A, Gomez I, Holuigue L, Litvak S, Jordana X (1996) The *rpl5-rps14-cob* gene arrangement in Solanum tuberosum: *rps14* is a transcribed and unedited pseudogene. Plant Mol Biol 31:937

Richardson AO, Rice DW, Young GJ, Alverson AJ, Palmer JD (2013) The "fossilized" mitochondrial genome of Liriodendron tulipifera: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. BMC Biol 11:29

Sloan DB, Taylor DR (2010) Testing for selection on synonymous sites in plant mitochondrial DNA: the role of codon bias and RNA editing. J Mol Evol 70:479

Sloan DB, Wu Z (2014) History of plastid DNA insertions reveals weak deletion and AT mutation biases in angiosperm mitochondrial genomes. Genome Biol Evol 6:3210

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol 10:e1001241

Smith DR (2015) Mutation rates in plastid genomes: they are lower than you might think. Genome Biol Evol 7:1227

Smith DR, Keeling PJ (2015) Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. Proc Natl Acad Sci USA 112:10177

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731

Uchida M, Ohtani S, Ichinose M, Sugita C, Sugita M (2011) The PPR-DYW proteins are required for RNA editing of *rps14*, cox1 and nad5 transcripts in Physcomitrella patens mitochondria. FEBS Lett 585:2367

van Loon B, Markkanen E, Hubscher U (2010) Oxygen as a friend and enemy: how to combat the mutational potential of 8-oxo-guanine. DNA Repair (Amst) 9:604

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

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586

CHAPTER 3

REPEATS OF UNUSUAL SIZE IN PLANT MITOCHONDRIAL GENOMES: IDENTIFICATION, INCIDENCE AND EVOLUTION

This chapter has been published: Wynn EL, Christensen AC. 2019. Repeats of Unusual Size in Plant Mitochondrial Genomes: Identification, Incidence and Evolution. G3. 9(2): 549-559. doi: 10.1534/g3.118.200948

ABSTRACT

Plant mitochondrial genomes have excessive size relative to coding capacity, a low mutation rate in genes and a high rearrangement rate. They also have abundant non-tandem repeats often including pairs of large repeats which cause isomerization of the genome by recombination, and numerous repeats of up to several hundred base pairs that recombine only when the genome is stressed by DNA damaging agents or mutations in DNA repair pathway genes. Early work on mitochondrial genomes led to the suggestion that repeats in the size range from several hundred to a few thousand base pair are underrepresented. The repeats themselves are not well-conserved between species, and are not always annotated in mitochondrial sequence assemblies. We systematically identified and compared these repeats, which are important clues to mechanisms of DNA maintenance in mitochondria. We developed a tool to find and curate non-tandem repeats larger than 50bp and analyzed the complete mitochondrial sequences from 157 plant species. We observed an interesting difference between taxa: the repeats are larger and more frequent in the vascular plants. Analysis of closely related species also shows that plant mitochondrial genomes evolve in dramatic bursts of breakage and rejoining,

complete with DNA sequence gain and loss. We suggest an adaptive explanation for the existence of the repeats and their evolution.

INTRODUCTION

It has long been known that plant mitochondrial genomes are much larger than those of animals (Ward, B. L. *et al.* 1981) and include significant amounts of non-coding DNA (Schuster, W. and A. Brennicke 1994). These genomes also often have repeats of several kb, leading to multiple isomeric forms of the genome (Folkerts, O. and M. R. Hanson 1989; Klein, M. *et al.* 1994; Palmer, J. D. and L. A. Herbon 1988; Palmer, J. D. and C. R. Shields 1984; Siculella, L. *et al.* 2001; Sloan, D. B. *et al.* 2010; Stern, D. B. and J. D. Palmer 1986). Plant mitochondrial genomes have very low mutation rates, but paradoxically have such high rearrangement rates that there is virtually no conservation of synteny (Drouin, G. *et al.* 2008; Palmer, J. D. and L. A. Herbon 1988; Richardson, A. O. *et al.* 2013; Wolfe, K. *et al.* 1987).

In addition to the large, frequently recombining repeats, there are often other repeated sequences in the size range of 1kb and lower (Arrieta-Montiel, M. P. *et al.* 2009; Forner, J. *et al.* 2005). Ectopic recombination between these non-tandem repeats has been shown to increase when double-strand breakage is increased, or in plants mutant for DNA maintenance genes (Abdelnoor, R. V. *et al.* 2003; Shedge, V. *et al.* 2007; Wallet, C. *et al.* 2015). Understanding the repeats is critical to understanding the mechanisms of DNA maintenance and evolution in plant mitochondria, yet they have never been systematically identified and analyzed. In addition to being infrequently and inconsistently annotated and described in mitochondrial genome sequences, repeats are often described as long, short and intermediate-length (Arrieta-Montiel, M. P. *et al.* 2009;

Davila, J. I. *et al.* 2011; Miller-Messmer, M. *et al.* 2012). The repeats are sometimes thought to be distributed into two size classes (one of up to several hundred bp and another of several kb), but this is derived from early studies of *Arabidopsis* and a few other species in which repeats were described and annotated (Alverson, A. J. *et al.* 2011b; Andre, C. *et al.* 1992; Arrieta-Montiel, M. P. *et al.* 2009; Davila, J. I. *et al.* 2011; Folkerts, O. and M. R. Hanson 1989; Sugiyama, Y. *et al.* 2005).

The most likely hypothesis that explains the peculiar characteristics of plant mitochondrial genomes is that double-strand break repair (DSBR) is abundantly used in plant mitochondria, perhaps to the exclusion of nucleotide excision and mismatch repair pathways (Christensen, A. C. 2014; Christensen, A. C. 2018). Double-strand break repair is very accurate when the repair is template-based, accounting for the low mutation rate in genes, but the nonhomologous end-joining or break-induced-replication pathways can account for the creation of repeats and chimeric genes, expansions, and loss of synteny through rearrangements.

The inconsistent reporting and annotation of repeated sequences leads to a number of questions. What is the best way to discover and characterize them? Is the size distribution really bimodal in angiosperms? Are there repeats in the mitochondria of other groups of green plants? How do they differ between groups? Can they be followed through evolutionary lineages like genes? Are the repeats themselves somehow adaptive, or are they a side-effect of DSBR that is neutral or nearly neutral? The availability in recent years of complete mitochondrial genome sequences across a wide variety of taxa of green plants allows us to begin addressing these questions. We describe a computational strategy for finding non-tandem repeats within plant mitochondrial

genomes. Using this tool, we describe the phylogenetic distribution of repeats in both size classes, examine their evolution in a family of closely related angiosperms, and propose an hypothesis for the evolutionary significance of the repeats and the DSBR processes that produce them.

MATERIALS AND METHODS

Sequence Data and Manipulation

Table 1 lists the mitochondrial genome sequences that were downloaded as FASTA format files from GenBank (https://www.ncbi.nlm.nih.gov/genbank/). BLAST searches (Altschul, S. F. *et al.* 1990) were done using version 2.7.1 on a Linux-based machine. In addition to the sequences shown in Table 1, mitochondrial genomes from several Brassica species were used to compare close relatives. These sequences are as follows: *Brassica carinata*; JF920287, *Brassica rapa*; JF920285, *Brassica oleracea fujiwase*; AP012988, *Brassica napus polima*; FR715249, *Brassica juncea*; JF920288. Alignments were done using the clustalW implementation in the VectorNTI 11.5 software package (ThermoFisher).

Repeat Analysis

Custom Python scripts are in Supplementary Materials. The script ROUSFinder.py (Appendix A1) uses blastn to perform a pairwise ungapped comparison of a sequence with itself, both strands separately, using a word size of 50, E value of 10,000, reward for a match +1, penalty for a mismatch -20. The script then concatenates the two output files and the full length self-identity is deleted. Alignments are then sorted and compared to identify and remove duplicate repeats, and an output file containing

each distinct repeat in fasta format is created. The sorting by size allows the script to automate the curation of the repeats by comparing the query start and end coordinates of each identified repeated sequence with the subject start and end coordinates of repeated sequences of the same size. When there are more than two copies of a repeat, BLAST does not report every pairwise hit, so the output file of FASTA-formatted sequences of repeats is then used as a query with the entire genome as subject to locate every copy of that repeat, create a table, and a table of binned sizes. The output can also be formatted for GenBank annotation. MultipleRepeats.py (Appendix A2) automates running ROUSFinder.py on every sequence within a directory.

Prior work identifying repeats, especially in *Arabidopsis thaliana* (Arrieta-Montiel, M. P. *et al.* 2009) showed that although BLAST is very useful, it has some characteristics that make it difficult to automate curation of the identified matches of non-tandem DNA sequences. For example, if there are a number of mismatches in a repeat, sometimes BLAST will identify subsets of the repeat sequence, or not give the same alignments when two imperfect repeats are used as queries of the entire genome. When there are three or more copies of a repeat, BLAST will also not identify every possible pairwise alignment, giving a subset instead. When examining the repeats in a single species, these problems can be solved by additional manual curation and inspection of the species. However, for automated curation of sequences from multiple species, some compromises have to be made. The simplest way to curate the repeated sequences is to ensure that the sizes of each repeat in a pair are the same. This ensures that repeats can be matched with each other by examining the coordinates of each copy to find all copies of the repeated sequence. In Arabidopsis thaliana, an ungapped blastn search with a match

reward of +1 and a mismatch penalty of -18 or lower ensured that different copies of the repeats were the same length, allowing automated curation. We therefore set the penalty parameter to -20 to make the automated curation reliable and fast. In order to more carefully examine the repeats in a single species, the script ROUSFinder2.py (Appendix A3) allows the user to set the match reward and mismatch penalty parameters on the command line.

After an initial analysis of sequences available in early 2018, we added additional species to the data in late 2018. These additional species are indicated in Table 1 by an asterisk. These include two hornworts, two liverworts, 3 bryophytes and 14 angiosperms. These new species do not change the patterns or conclusions compared to the earlier analysis, providing additional validation of the use of BLAST and the curation methods described.

Data Availability

The authors state that all data necessary for confirming the conclusions presented in this article are represented fully within the article, including python scripts in Appendices and accession numbers of DNA sequences shown in Table 3.1.

RESULTS

Repeats in Plant Mitochondrial Genomes

The existence of large non-tandem repeats in plant mitochondrial genomes is well known by now, but they have not been systematically identified and analyzed. Prior studies used variations of BLAST (Altschul, S. F. *et al.* 1990) to find repeats (Alverson, A. J. *et al.* 2011a; Alverson, A. J. *et al.* 2010; Alverson, A. J. *et al.* 2011b; Liu, Y. *et al.*

2014) or REPuter (Hecht, J. *et al.* 2011; Kurtz, S. and C. Schleiermacher 1999). Other available software packages specifically identify tandem repeats, or repeats matching known repetitive sequences. Due to the ready availability of BLAST and the flexibility of its use, and because most prior work used it, we wrote and used a Python script called ROUSFinder.py that uses BLAST to identify non-tandem repeats within mitochondrial genomes. The parameters for identification of a sequence repeat were relatively stringent and included a blastn word size of 50, and match/mismatch scores of +1/-20. Any choice of parameters will necessarily identify some false positives and false negatives. These parameters were chosen in order to find duplicate copies of sequence that were either recently created or recently corrected by gene conversion. As described in the Methods, they were also chosen to enable automated curation of the repeats that are found in the first iteration of BLAST. A duplication longer than 100 bases that has several mismatches or a gap in the center of the repeat unit will be identified as two different repeats by this script. However, mismatches in the center of one copy of a repeat are indicative of either two independent events producing the two parts of the repeat, or mutation and drift that have escaped gene conversion. Because we are concerned with the recombination behavior of the repeats we therefore chose to call these two different repeats. To analyze and identify repeats in a single sequence for further study or annotation would require additional manual curation of the output. The word size parameter of 50 is chosen to make the output more manageable. Reducing the word size identifies numerous smaller repeats, but the smaller the repeats get, the more complex the curation task of distinguishing identical sequences from similar ones of the same size – a task that at this point still needs to be done manually. Reducing the word size does not change the

conclusions about any of the potentially recombinogenic repeats that also distinguish the major groups in the plant kingdom. Previous work in *Arabidopsis thaliana* has shown that crossover products between non-tandem repeats of 556bp or smaller in accession Col-0 and 204bp or smaller in accession Ws are undetectable by PCR (Wallet, C. *et al.* 2015), similar to prior results using Southern blots (Arrieta-Montiel, M. P. *et al.* 2009; Davila, J. I. *et al.* 2011; Sakamoto, W. *et al.* 1996)

The species we used represent a significant subset of the complete mitochondrial genome sequences from green plants in GenBank and are shown in Table 3.1. Sequences available on GenBank are not a random sample across taxa (food crops are very over-represented, for example), so to reduce sampling bias somewhat we used only one species per genus. Incomplete sequences or sequences with gaps or wildcard characters (such as N, R, Y, etc.) are not handled well by BLAST without further curation, so these were not used. Species with multiple distinct chromosomes were also not used because of the additional layer of complexity from inter- and intra-chromosomal repeats. The full output is in Supplemental Table S1. The repeats seen in plant mitochondrial genomes are much larger than those found in random sequence (data not shown), suggesting that they arise from specific biological processes and are not stochastic.

Phylogenetic Clustering

The distribution of repeat sizes forms distinct clusters between broad phylogenetic groups (see Figure 3.1). Because there are different numbers of species in each group, and some species have an order of magnitude more total repeats than others, we represent the data as the fraction of species within that group that have at least one repeat within a given size range. Within the chlorophytes, repeats of greater than 200bp are rare. The

exceptions are the prasinophytes (discussed below) and a few interesting cases. *Chlamydomonas reinhardtii* has a 532 bp inverted repeat at the termini of its linear chromosome. *Dunaliella salina, Kirchneriella aperta* and *Polytoma uvella* have novel structures at a small number of loci that consist of overlapping and nested repeats and palindromes (Smith, D. R. *et al.* 2010). The function of these structures is unknown, but they are unusual and not common in the chlorophytes. The prasinophyte group resembles the rest of the chlorophytes in having no non-tandem repeats greater than 200bp but many of them include two copies of a single very large repeat between 9.5 and 14.4 kb. This is similar to many chloroplast genomes and it is possible that this structure is involved in replication (Bendich, A. J. 2004). The bryophytes generally resemble the chlorophytes; there are no repeats longer than 200bp.

In contrast to the chlorophytes and bryophytes, the Marchantiophyta (liverworts) and Anthocerotophyta (hornworts) have repeats greater than 200bp in size, but none bigger than 1131bp. The other lineages of streptophytic green algae (referred to as charophytes in GenBank) resemble the chlorophytes albeit with a slightly higher upper limit. In this group the largest repeat is found in Chlorokybus atmophyticus and is 291bp.

The ferns and lycophytes are strikingly different from the previous groups. Unfortunately, the number of species sequenced is low. They have large numbers of repeats and the repeat sizes range well above 200bp, up to 10 kb. Some members of these groups, such as *Huperzia*, are similar to the bryophytes, but others are large and have significant repeat content (Guo, W. *et al.* 2017). These groups are underrepresented among available mitochondrial sequences, in part due to the complexity caused by the

repetitive nature of lycophyte and fern mitochondrial genomes (Grewe, F. *et al.* 2009), but the patterns are noticeably different from the nonvascular plants described above.

The angiosperms are represented very well in the sequence databases. Only one member of this group does not have any repeats larger than 200bp (*Medicago truncatula*). A small number of angiosperms lack repeats larger than 1 kbp, and approximately half include repeats larger than 9 kbp. *Silene conica*, a species with multiple large chromosomes not included in our dataset has a nearly 75kb sequence found in both chromosomes 11 and 12 (Sloan, D. B. *et al.* 2012). Gymnosperms are also underrepresented, but appear to be similar to the other vascular plants. Interestingly, the gymnosperms *Ginkgo biloba* and *Welwitschia mirabilis* resemble angiosperms, while *Cycas taitungensis* is more similar to ferns. The *C. taitungensis* mitochondrion has numerous repeats, including many that are tandemly repeated. Five percent of this genome consists of the mobile Bpu element, a remarkable level of repetitiveness (Chaw, S. M. *et al.* 2008).

It is only in the vascular plants that the number and size of repeated sequences in mitochondrial genomes has been expanded. The vascular plants generally only have mitochondrial genomes a few times larger than the bryophytes, liverworts and hornworts, but the repeats are expanded well beyond proportionality to size. Some taxa, such as the Geraniaceae, Plantago, and *Silene* include species with significantly expanded mitochondrial DNA (Park, S. *et al.* 2015; Parkinson, C. L. *et al.* 2005; Sloan, D. B. *et al.* 2012). These species are outliers in the mitochondrial genome sizes and the number of repeats, but the underlying DNA replication, recombination and repair processes are

likely to be the same. There appears to have been a significant change in mitochondrial DNA maintenance mechanisms coincident with the origin of the vascular plants.

Repeat Sizes and Frequency in Angiosperms

Large repeats of several kilobases have been identified in several species and shown to be recombinationally active, isomerizing angiosperm mitochondrial genomes (Folkerts, O. and M. R. Hanson 1989; Klein, M. *et al.* 1994; Palmer, J. D. and L. A. Herbon 1988; Palmer, J. D. and C. R. Shields 1984; Siculella, L. *et al.* 2001; Sloan, D. B. 2013; Stern, D. B. and J. D. Palmer 1986). A few species have been reported to lack such structures (Palmer, J. D. 1988). The first comprehensive catalog of repeated sequences shorter than 1000 base pairs was done in *Arabidopsis thaliana*, and they were shown to be recombinationally active in some mutant backgrounds, but not generally in wild type (Arrieta-Montiel, M. P. *et al.* 2009; Davila, J. I. *et al.* 2011; Miller-Messmer, M. *et al.* 2012; Shedge, V. *et al.* 2007). Is the spectrum of repeat sizes in *Arabidopsis*, and its bimodality, typical for angiosperms? Figure 3.2 illustrates the presence of repeats in the size range of 50bp to over 10,000 bp in 72 angiosperms, sorted by the class and order of the species. While individual species often have a bimodal distribution of sizes, there is no size range that is universally absent from the distribution. Thirteen of the 72 species have no repeats larger than 600bp, leaving open the question of whether those particular mitochondrial genomes isomerize through recombination. All of the other species have a large repeat of somewhere between 600bp and 65kbp. There is no pattern of repeat size distribution or total size with the phylogenetic group or total mitochondrial genome size, suggesting that these are not produced by stochastic processes, and suggesting that they occur and change faster than speciation does.

Alignment of Repeats Within the Brassicales

In order to test the hypothesis that the repeated sequences change rapidly compared to speciation events, leading to the lack of pattern in the Angiosperm orders, we analyzed 6 closely related species in the *Brassica* genus. Within the *Brassica* genus there are three diploid species: *Brassica rapa*, *Brassica nigra* and *Brassica oleracea*, and three allotetraploid species (Cheng, F. *et al.* 2017). The diploid nuclear genomes are called the A, B and C genomes, respectively. Based on both nuclear and mitochondrial sequences it appears that *Brassica carinata* has the *B. nigra* and *B. oleracea* nuclear genomes (BBCC) and the *B. nigra* mitochondrial genome, while *Brassica juncea* has the *B. nigra* and *B. rapa* nuclear genomes (BBCC) and the *B. rapa* mitochondrial genome. *Brassica napus* has two subspecies, *polima* and *napus*. Both have the *B. oleracea* and *B. rapa* nuclear genomes (AACC), but *B. napus polima* appears to have the *B. rapa* mitochondrial genome and *B. napus napus* has the *B. oleracea* mitochondrial genome (Chang, S. *et al.* 2011; Franzke, A. *et al.* 2011; Grewe, F. *et al.* 2014). Thus it appears that the hybridization event between *B. oleracea* and *B. rapa* occurred at least twice, with each species being the maternal parent. In the analysis below we use the *B. napus polima* mitochondrial genome. We compared these *Brassica* species to *Raphanus sativus* and *Sinapis arvensis* as outgroups. These species are the closest relatives of the *Brassicas* with complete mitochondrial genome sequences (Grewe, F. *et al.* 2014). Several of these species were mapped prior to genomic sequencing, and repeated sequences and mitochondrial genome isomerization was observed (Palmer, J. D. 1988; Palmer, J. D. and L. A. Herbon 1986).

All eight of these species include one pair of long repeats, ranging in length from 1.9kb to 9.7kb. However, these species show an interesting pattern. *B. nigra*, *B. carinata*, *R. sativus* and *S. arvensis*, hereafter referred to as group A, each have two copies of a 6.5 to 9.7kb repeat that is only present as single copy sequence in the mitochondria of *B. rapa*, *B. oleracea*, *B. napus* and *B. juncea*, herafter referred to as group B (see Figure 3.3). The group B species each have two copies of a long repeat 1.9kb long that is present as single-copy sequence in group A. Figure 3.3 shows these repeated sequences, aligned only to each other and placed onto the known phylogenetic tree of the Brassicales (Grewe, F. *et al.* 2014). The longest repeats are aligned, and the genes flanking them are shown. Part A shows the long repeat and neighboring sequences from the A group and the homologous single-copy sequences from the B group. Part B compares the long repeat from the B group to the single-copy homologous region from the A group.

Grewe *et al.* examined the synonymous substitution rates in genes of Brassicales mitochondrial genomes (Grewe, F. *et al.* 2014) and found them to be very low, consistent with most land plants. However, the presence of repeats allows mutations in non-coding DNA to be examined qualitatively. The long repeats in the A group differ by large block substitutions and insertion/deletions. Where two copies are present in a species there are very few difference between copies, and they are generally near the boundaries of the repeats. Although significant differences can arise during speciation events, both copies of a repeat within a species remain identical. This supports the hypothesis that copies of repeated DNA are maintained as identical sequence by frequent recombination and gene conversion.

The long repeat of *B. nigra* and *B. carinata* underwent massive change in the lineage leading to the B group of Brassica species (see Figure 3.3). The first 1.6kb and the last 1.7kb of the long repeat in the A group are conserved in the B group, and the *ccmB* gene still flanks the repeat on one side. However, the last 1.7kb are inverted and separated from the first 1.6kb by 3.3kb of a sequence of unknown origin. An additional difference is seen in *B. oleracea* wherein *rps7* now flanks the repeat rather than *ccmB.* Other major changes appear to have occurred in the time since *B. nigra* diverged from the ancestor of *B. oleracea* and *B. rapa*; a comparison of the complete mitochondrial genomes of *B. rapa* and *B. nigra* reveal at least 13 segments of DNA that have been rearranged. No major rearrangements have occurred between *B. nigra* and *B. carinata*, nor between *B. rapa*, *B. juncea* and *B. napus polima*. *B. oleracea* differs from *B. rapa* by approximately six rearrangement events (Grewe, F. *et al.* 2014).

At the same time that the long repeat of the A group was being dramatically altered in the lineage leading to *B. rapa* and *B. oleracea*, a new long repeat appeared in the B group, which includes the coding sequence of the *cox2* gene. This new long repeat is maintained throughout this group of four species, and the flanking genes are also conserved (see Figure 3.3). The *cox2* gene is single copy in the A group and is in a nearly syntenic arrangement with neighboring genes.

All alignments used in this analysis can be found at http://www.g3journal.org/content/9/2/549.supplemental

DISCUSSION

The availability of complete mitochondrial genome sequences from many taxa of green plants allows us to compare the abundance and size distribution of non-tandem repeats across taxa. Although such repeats have been known for some time, their functions (if any) and evolution are largely mysterious. It has been suggested that their existence and maintenance are outgrowths of double-strand break repair events such as nonhomologous end-joining (NHEJ), break-induced replication (BIR) and gene conversion (Christensen, A. C. 2018). We describe here a Python script that uses BLAST (Altschul, S. F. *et al.* 1990) to find non-tandem repeats within sequences, and use it to analyze plant mitochondrial DNA. In addition, comparison of repeats between closely related species within the Brassicales showed that repeat differences between species were largely due to rearrangements and block substitutions or insertions, which could be due to NHEJ and BIR, while the two copies of the repeat were identical within a species, suggesting continuing repair by gene conversion or homologous recombination.

Repeats in mitochondria appear to be more abundant and larger in the vascular plants than in the non-vascular taxa. This suggests that the first vascular common ancestor of lycophytes, ferns, gymnosperms and angiosperms acquired new mechanisms of mitochondrial genome replication and repair that led to a proliferation of repeats and increases in repeat size and mitochondrial genome size. Complete sequences of more species, particularly in the lycophytes and ferns, is necessary to add clarity but the ancestor of vascular plants evidently made a transition to increased use of double-strand break repair in their mitochondria, leading to the genomic gymnastics seen today.

The analysis of repeats in the *Brassica* species suggests that mitochondrial genomes can remain relatively static for long periods of time, but can also diverge very rapidly by rearrangements, sequence loss, and gain of sequences of unknown origin. This pattern resembles punctuated equilibrium (Gould, S. J. and N. Eldredge 1977). The mechanisms and frequency are unknown, but it suggests that a lineage can experience a burst of genome recombination, breakage and rejoining, dramatically rearranging and altering the mitochondrial genome, as if it had been shattered and rebuilt. These events occur on a time scale that is faster than that of speciation, leading to high levels of divergence, and loss of synteny.

Qualitative differences have been described between the repeats shorter and longer than about 1kb (Arrieta-Montiel, M. P. *et al.* 2009; Klein, M. *et al.* 1994; Mower, J. P. *et al.* 2012). In general, the largest repeats within a species have been found to recombine constitutively, leading to isomerization of the genome into multiple major forms. The shortest repeats (less than 50bp) may be involved in homologous recombination events only rarely, while those of intermediate size, generally in the 100s of base pairs, can recombine in response to genome damage or in DNA maintenance mutants, but do not normally do so in unstressed, non-mutant plants, as noted above. The intermediate size repeats have been primarily analyzed in *Arabidopsis thaliana*, and have been found to recombine in abnormal conditions. In plants treated with ciprofloxacin (which induces mitochondrial double-strand breaks), or in mutants of the mitochondrial *recG* homolog, repeats of 452, 249, 204 and 126bp were seen to recombine (Wallet, C. *et al.* 2015). In mutants of *msh1* (which results in high levels of ectopic recombination), there was some recombination seen between repeats as small as 70bp, but none in repeats

of 50bp or smaller (Davila, J. I. *et al.* 2011). This suggests a changing spectrum of function and activity correlated with size, which could also vary by species.

Functional analysis of repeat recombination can be done by analyzing clones big enough to include the repeats (Klein, M. *et al.* 1994), by long read sequencing (Shearman, J. R. *et al.* 2016), PCR (Wallet, C. *et al.* 2015) or by Southern blotting (Arrieta-Montiel, M. P. *et al.* 2009; Sakamoto, W. *et al.* 1996). Functional analysis of the large repeats is an important step in understanding plant mitochondrial genome structure and evolution (Guo, W. *et al.* 2016; Guo, W. *et al.* 2017; Sloan, D. B. 2013) and may reveal different patterns of recombination between species, which would reveal important differences in the replication and repair machinery and dynamics.

We doubt that there is an adaptive advantage to large size and abundant rearrangements in the genomes of plant mitochondria. We suggest that these are correlated traits accompanying the adaptive advantage of a greatly increased reliance on double-strand break repair. DNA repair is critically important because damage is more likely in mitochondria than the nucleus due to the presence of reactive oxygen species. In animals, the mitochondrial mutation rate is high, but the reduced mitochondrial genome size minimizes the number of potential mutational targets (Lynch, M. *et al.* 2006; Smith, D. R. 2016). However, with multiple copies of mitochondrial DNA in each cell, an alternative strategy in a high DNA damage environment is to increase the use of template DNA in repair. The accuracy of double-strand break repair when a template is used is accompanied by the creation of chimeras, rearrangements and duplications when templates are not identical or cannot be found by the repair enzymes. Dramatic expansions, rearrangements and losses, accompanied by low substitution rates in genes is

characteristic of flowering plant mitochondria. Selection on gene function maintains the genes, while the expansions and rearrangements must be nearly neutral. Once mitochondria evolved very efficient double-strand break repair, and a mechanism for inducing double-strand breaks at the sites of many types of damage, more primitive mechanisms, such as nucleotide excision repair can and have been lost (Gualberto, J. M. *et al.* 2014; Gualberto, J. M. and K. J. Newton 2017) without obvious evolutionary cost.

The adaptive value of increased and efficient double-strand break repair is probably to avoid mutations in the essential genes of mitochondria, and is possible because of the abundance of double-stranded template molecules in each cell. However, this mechanism of repair has an additional correlated trait. There are bacterial species, such as *Deinococcus radiodurans*, that excel at double-strand break repair and can rebuild even significantly fragmented genomes (Krisko, A. and M. Radman 2013) while also being able to minimize radiation-induced damage (Sharma, A. *et al.* 2017). While *D. radiodurans* is notoriously resistant to ionizing radiation, the adaptive value is thought to be desiccation resistance, because dehydration is more likely to have been experienced than extreme radiation in the history of the lineage, and also produces double-strand breaks (Mattimore, V. and J. R. Battista 1996). Radiation resistant bacteria in unrelated phylogenetic groups show more genome rearrangements and loss of synteny than their radiation sensitive relatives (Repar, J. *et al.* 2017), suggesting that abundant double-strand break repair is the cause of both the resistance to significant double-strand breakage and the loss of synteny. An interesting possibility is that very efficient double-strand break repair in plant mitochondria also confers desiccation resistance as a correlated trait. Because mitochondria are metabolically active immediately upon

imbibition of seeds, DNA damage must be repaired very efficiently and rapidly (Paszkiewicz, G. *et al.* 2017). Efficient repair of desiccation-mediated damage in all cellular compartments is a prerequisite to being able to produce seeds or spores for reproduction. It is possible that the DNA repair strategy of plant mitochondria was one of several factors (including desiccation resistance of the nuclear and plastid genomes, presumably by distinct mechanisms) that are beneficial to vascular plants. The evidence of the repeats suggests that the transition to double-strand break repair in mitochondria occurred at approximately the same time as the transition to vascularity in plants, and it may have been one of several traits that enabled their success. In addition, once the life cycles of land plants included periods of desiccation in spores and seeds, double-strand breakage would have increased, accompanied by increases in rearrangements, expansions, and chimeras. The mechanisms of double-strand break repair continue to be important for understanding the evolution of plant mitochondrial genomes.

Group	Subgroup	# of species	50-199	200-499	500-999	1000-2499	2500-4999	5000-7499	7500-9999	≥ 10000
Chlorophytes	Chlorophyta	26	0.88	0.19	0.08	0.00	0.00	0.00	0.00	0.00
Chlorophytes	Prasinophytes	8	0.63	0.00	0.00	0.00	0.00	0.00	0.50	0.25
Anthocerotophyta	Anthocerotophyta	4	1.00	0.50	1.00	0.25	0.00	0.00	0.00	0.00
Marchantiophyta	Marchantiophyta	6	1.00	1.00	0.67	0.00	0.00	0.00	0.00	0.00
Bryophytes	Bryophytes	26	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Streptophyta	Charophyta	8	0.88	0.25	0.00	0.00	0.00	0.00	0.00	0.00
Tracheophyta	Fern	2	1.00	1.00	1.00	1.00	1.00	0.00	1.00	0.00
Tracheophyta	Lycophyte	2	1.00	1.00	1.00	1.00	1.00	0.00	0.50	0.00
Tracheophyta	Gymnosperm	3	1.00	1.00	0.33	0.67	0.33	0.33	0.00	0.00
Tracheophyta	Angiosperm	72	1.00	0.96	0.47	0.49	0.46	0.36	0.25	0.43

Figure 3.1: Size distributions of repeats in groups of species. The number of species represented in each group is shown. Headings indicate the bins of repeat sizes and the numbers indicate the fraction of species in that group that have at least one repeat of that size. Heat map color coding is blue for one and yellow for zero.

Class or Subclass	Order	Species	Genome_size	50-99	100-199	200-299	300-399	400-499	500-599	600-699	700-799	800-899	900-999	1000-1499	1500-1999	2000-2499	2500-2999	3000-3499	3500-3999	4000-4499	4500-4999	5000-5499	5500-5999	6000-6499	6500-6999	7000-7499	7500-7999	8000-8499	8500-8999	9000-9499	9500-9999	≥ 10000
ANA Grade	Nymphaeales	*Nymphaea colorata**	617195	123024	61729	10130	2393	705	126	32	12	0	0	4	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Magnoliids	Magnoliales	*Liriodendron tulipifera*	553721	190	35	6	0	4	0	4	0	2	0	2	2	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	2
Monocots	Alismatales	*Butomus umbellatus*	450826	54	28	4	4	0	0	0	0	3	0	3	2	5	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	0
Monocots	Asparagales	*Allium cepa*	316363	67	8	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Monocots	Arecales	*Cocos nucifera*	678653	110	38	19	0	0	0	0	0	0	0	4	2	4	2	0	0	2	0	0	0	0	3	3	0	0	0	0	4
Monocots	Arecales	*Phoenix dactylifera*	715001	64	18	0	4	0	0	0	0	0	0	2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Monocots	Poales	*Triticum aestivum*	452528	87	29	6	3	0	0	0	0	0	0	2	3	4	0	0	2	3	0	2	0	4	0	2	0	0	0	0	0	0
Monocots	Poales	*Sorghum bicolor*	468628	75	17	7	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	4
Monocots	Poales	*Aegilops speltoides*	476091	64	24	4	3	3	0	0	0	0	0	0	2	0	0	4	0	0	4	4	3	0	0	2	0	0	2	0	0	0
Monocots	Poales	*Oryza sativa Indica*	491515	152	17	3	0	2	0	0	0	0	0	3	2	0	0	0	0	3	0	0	0	2	0	2	0	0	2	0	2	6
Monocots	Poales	*Hordeum vulgare*	525599	76	20	5	7	4	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	2	0	2	4
Monocots	Poales	*Zea mays strain NB*	569630	77	15	6	0	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	6
Monocots	Poales	*Lolium perenne*	678580	101	33	4	2	12	8	8	4	2	0	14	6	4	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	6
Monocots	Poales	*Tripsacum dactyloides*	704100	8356	1261	7	5	0	0	0	0	0	0	0	2	0	3	0	0	0	0	0	0	0	0	0	0	2	0	0	4
Eudicots	Proteales	*Nelumbo nucifera**	524797	1908	326	53	12	10	12	6	4	8	4	8	2	0	4	8	0	2	0	0	3	2	2	0	0	0	0	0	0	2
Superrosids	Saxifragales	*Heuchera parviflora*	542954	69	34	7	4	0	3	0	0	0	0	3	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	2
Superrosids	Vitales	*Vitis vinifera complete*	773279	219	35	6	4	2	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Fabids	Malpighiales	*Ricinus communis*	502773	204	4	4	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
Rosids/Fabids	Malpighiales	*Salix suchowensis*	644437	54	12	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	0
Rosids/Fabids	Malpighiales	*Populus tremula*	783442	78	22	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Fabids	Fabales	*Medicago truncatula*	271618	40	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Fabids	Fabales	*Lotus japonicus*	380861	60	33	4	7	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	2
Rosids/Fabids	Fabales	*Glycine max*	402558	79	127	11	3	0	0	0	0	0	0	0	9	0	7	2	2	0	2	0	3	2	4	0	0	0	0	0	0	0
Rosids/Fabids	Fabales	*Vigna angularis*	404466	27	50	6	4	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Fabids	Fabales	*Millettia pinnata*	425718	54	9	2	0	0	0	3	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2
Rosids/Fabids	Fabales	*Ammopiptanthus mongolicus**	475396	67	68	4	4	0	0	2	0	0	0	0	2	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Fabids	Fabales	*Styphnolobium japonicum**	484916	61	10	4	0	2	0	0	2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Fabids	Fabales	*Senna tora**	566589	74	36	5	5	2	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	9
Rosids/Fabids	Fabales	*Vicia faba*	588000	58	55	2	6	5	0	3	0	0	0	0	5	3	2	2	0	0	0	0	2	0	0	2	2	0	0	2	0	11
Rosids/Fabids	Fabales	*Leucaena trichandra**	722009	344	76	16	4	0	3	0	0	0	0	2	4	0	0	4	0	2	0	2	0	0	0	2	0	0	2	0	4
Rosids/Fabids	Rosales	*Rosa chinensis**	313448	65	10	6	4	5	0	0	2	2	0	0	2	2	2	4	0	0	0	0	0	4	2	0	0	0	0	0	0	0
Rosids/Fabids	Rosales	*Ziziphus jujuba*	365190	41	12	4	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Fabids	Rosales	*Cannabis sativa*	415602	76	15	4	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Fabids	Cucurbitales	*Citrullus lanatus*	379236	62	36	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Rosids/Fabids	Fagales	*Betula pendula*	581505	444	44	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Malvids	Geraniales	*Geranium maderense*	737091	305	55	17	6	2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Malvids	Myrtales	*Lagerstroemia indica*	333948	1152	21	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Malvids	Sapindales	*Citrus sinensis**	640906	100	31	8	2	2	0	0	0	0	0	2	4	2	0	0	0	0	0	0	0	0	0	2	2	0	0	0	2	
Rosids/Malvids	Malvales	*Hibiscus cannabinus*	569915	108	32	6	5	5	2	0	0	2	0	0	4	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Rosids/Malvids	Malvales	*Bombax ceiba**	594390	75	43	18	4	2	2	4	0	4	0	6	8	0	4	0	2	0	2	0	0	0	0	2	2	0	0	2	0	0
Rosids/Malvids	Malvales	*Gossypium hirsutum*	668584	81	70	38	25	11	2	2	0	0	0	2	0	0	0	0	0	0	2	0	0	2	0	2	0	0	0	2	0	4
Rosids/Malvids	Malvales	*Corchorus capsularis*	1999602	665	125	8	2	0	2	2	0	0	0	0	0	2	2	0	0	0	0	0	0	0	2	0	0	0	0	0	4	
Rosids/Malvids	Brassicales	*Brassica nigra*	232407	70	8	5	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
Rosids/Malvids	Brassicales	*Sinapis arvensis*	240024	59	23	10	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
Rosids/Malvids	Brassicales	*Eruca vesicaria*	247696	84	35	5	7	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2
Rosids/Malvids	Brassicales	*Schrenkiella parvula*	255773	39	22	4	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Malvids	Brassicales	*Raphanus sativus*	258426	50	11	7	4	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
Rosids/Malvids	Brassicales	*Arabis alpina**	323159	68	20	4	0	2	0	0	0	0	0	2	0	0	0	4	0	0	0	0	2	0	0	0	0	0	0	0	0	0
Rosids/Malvids	Brassicales	*Arabidopsis thaliana*	367808	91	34	14	6	2	4	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0
Rosids/Malvids	Brassicales	*Batis maritima*	403930	41	23	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rosids/Malvids	Brassicales	*Carica papaya*	476890	46	17	4	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Superasterids	Santalales	*Viscum album*	565432	286	34	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Superasterids	Caryophyllales	*Silene latifolia*	253413	8	4	0	0	0	0	0	0	0	0	35	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Superasterids	Caryophyllales	*Spinacia oleracea*	329613	181	9	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
Superasterids	Caryophyllales	*Beta vulgaris*	501020	94	18	3	13	3	0	0	0	3	0	3	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	4
Asterids/Lamiids	Solanales	*Ipomoea nil*	265768	174	43	4	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asterids/Lamiids	Solanales	*Nicotiana tabacum*	430597	83	34	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	5
Asterids/Lamiids	Solanales	*Hyoscyamus niger*	501401	1435	138	5	10	5	2	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	2	0	2	0	0	2
Asterids/Lamiids	Solanales	*Capsicum annuum*	511530	91	21	0	2	0	0	0	0	0	0	3	2	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	2	
Asterids/Lamiids	Gentianales	*Rhazya stricta*	548608	53	9	5	0	0	3	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Asterids/Lamiids	Lamiales	*Ajuga reptans*	352069	18	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asterids/Lamiids	Lamiales	*Castilleja paramensis*	495499	33	13	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
Asterids/Lamiids	Lamiales	*Salvia miltiorrhiza*	499236	77	18	4	2	0	0	2	0	0	0	0	0	2	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Asterids/Lamiids	Lamiales	*Boea hygrometrica*	510519	41	8	4	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asterids/Lamiids	Lamiales	*Mimulus guttatus*	525671	41	12	9	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
Asterids/Lamiids	Lamiales	*Utricularia reniformis**	857234	113	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Asterids/Campanulids	Asterales	*Chrysanthemum boreale**	211002	31	8	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asterids/Campanulids	Asterales	*Diplostephium hartwegii*	277718	52	10	8	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asterids/Campanulids	Asterales	*Helianthus annuus*	301004	42	17	7	0	2	0	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2
Asterids/Campanulids	Asterales	*Platycodon grandiflorus**	1249593	64	19	8	2	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Asterids/Campanulids	Apiales	*Daucus carota subsp. sativus*	281132	55	21	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	5
Asterids/Campanulids	Apiales	*Bupleurum falcatum**	463792	83	27	2	6	0	2	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	2

Figure 3.2: Distribution of repeat sizes among angiosperms. Species are sorted by the phylogenetic groups as described by the Angiosperm phylogeny group (Cole *et al.* 2017). The number of repeats of each size class is shown. Blue shading indicates a number greater than zero.

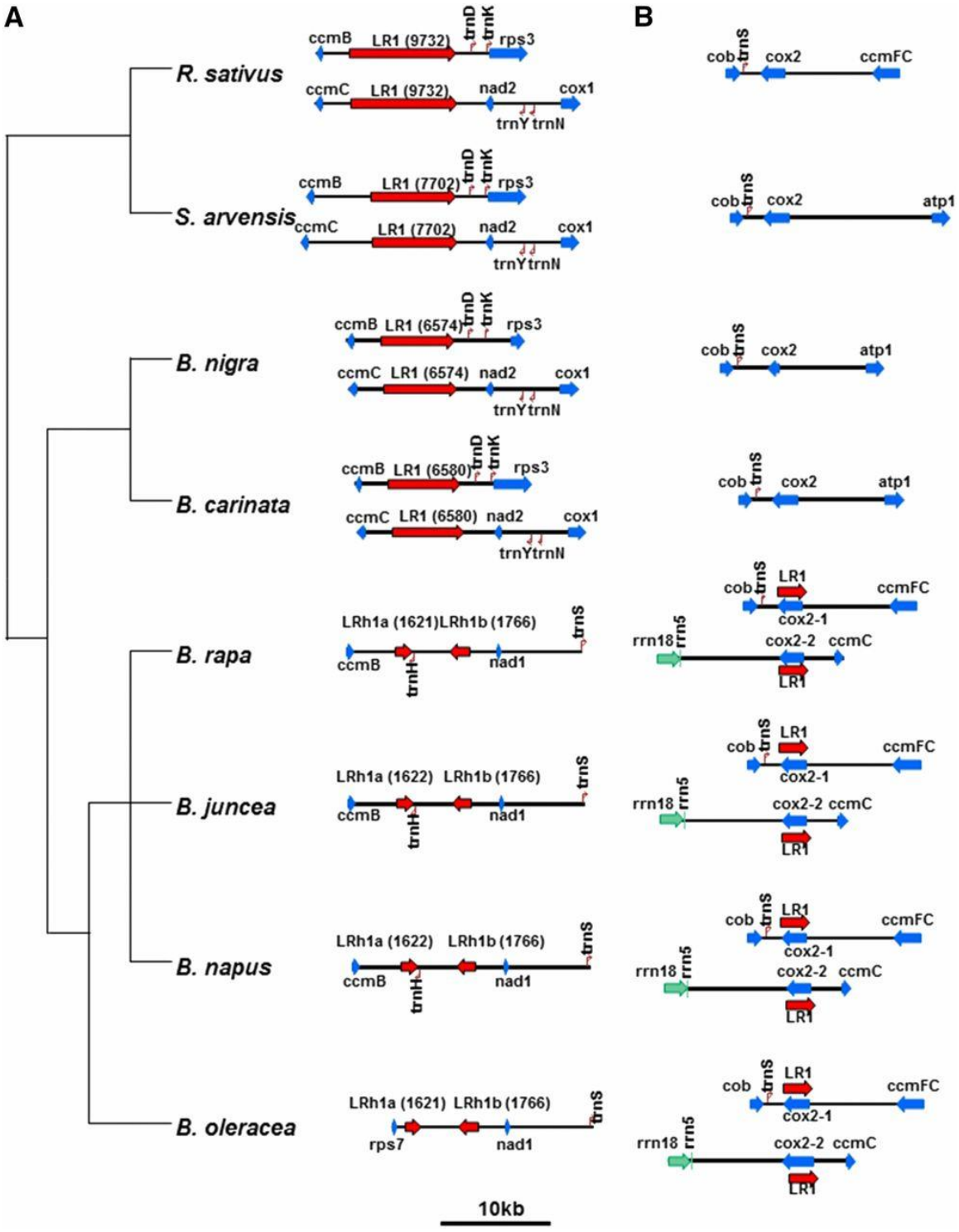
A
B
R. sativus
S. arvensis
B. nigra
B. carinata
B. rapa
B. juncea
B. napus
B. oleracea
ccmB
LR1 (9732)
trnD
trnK
rps3
ccmC
nad2
cox1
trnY
trnN
LR1 (7702)
LR1 (6574)
LR1 (6580)
LRh1a (1621)
LRh1b (1766)
trnS
trnH
rps7
LRh1a (1622)
cob
cox2
ccmFC
atp1
LR1
cox2-1
rrn18
rrn5
cox2-2
10kb

Figure 3.3: Alignment of long repeats in the Brassicales. A phylogenetic tree is shown at left, derived from Grewe *et al.* (Grewe, F. *et al.* 2014). Part A aligns the longest repeat in Group A (*R. sativus, S. arvensis, B. nigra* and *B. carinata*) and shows the genes flanking them. The homologous single-copy sequence from *B. rapa, B. napus, B. juncea* and *B. oleracea* is also shown. Part B aligns the longest repeat in Group B (*B. rapa, B. napus, B. juncea* and *B. oleracea*), and shows the homologous single-copy region in Group A. Red arrows indicate the long repeats that were used to align all sequences in the two parts of the figure. Blue indicates genes in the flanking regions that may or may not be conserved or rearranged. Green indicates rRNA genes and small arrows represent tRNA genes. Branch lengths in the tree are not to scale. The sequences are depicted at the scale shown in the figure.

REFERENCES

Abdelnoor R. V., Yule R., Elo A., Christensen A. C., Meyer-Gauen G., *et al.*, 2003 Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. Proc. Natl. Acad. Sci. USA 100: 5968–5973. doi:10.1073/pnas.1037651100

Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J., 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410. doi:10.1016/S0022-2836(05)80360-

Alverson A. J., Rice D. W., Dickinson S., Barry K., Palmer J. D., 2011a Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. Plant Cell 23: 2499–2513. doi:10.1105/tpc.111.087189

Alverson A. J., Wei X., Rice D. W., Stern D. B., Barry K., *et al.*, 2010 Insights into the evolution of mitochondrial genome size from complete sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae). Mol. Biol. Evol. 27: 1436–1448. doi:10.1093/molbev/msq029

Alverson A. J., Zhuo S., Rice D. W., Sloan D. B., Palmer J. D., 2011b The mitochondrial genome of the legume Vigna radiata and the analysis of recombination across short mitochondrial repeats. PLoS One 6: e16404. doi:10.1371/journal.pone.0016404

Andre C., Levy A., Walbot V., 1992 Small repeated sequences and the structure of plant mitochondrial genomes. Trends Genet. 8: 128–132.

Arrieta-Montiel M. P., Shedge V., Davila J., Christensen A. C., Mackenzie S. A., 2009 Diversity of the Arabidopsis Mitochondrial Genome Occurs via Nuclear-Controlled Recombination Activity. Genetics 183: 1261–1268. doi:10.1534/genetics.109.108514

Bendich A. J., 2004 Circular Chloroplast Chromosomes: The Grand Illusion. Plant Cell 16: 1661–1666. doi:10.1105/tpc.160771

Chang S., Yang T., Du T., Huang Y., Chen J., *et al.*, 2011 Mitochondrial genome sequencing helps show the evolutionary mechanism of mitochondrial genome formation in Brassica. BMC Genomics 12: 497. doi:10.1186/1471-2164-12-497

Chaw S. M., Shih A. C., Wang D., Wu Y. W., Liu S. M., *et al.*, 2008 The mitochondrial genome of the gymnosperm Cycas taitungensis contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol. Biol. Evol. 25: 603–615. doi:10.1093/molbev/msn009

Cheng F., Liang J., Cai C., Cai X., Wu J., *et al.*, 2017 Genome sequencing supports a multi-vertex model for Brassiceae species. Curr. Opin. Plant Biol. 36: 79–87. doi:10.1016/j.pbi.2017.01.006

Christensen A. C., 2014 Genes and Junk in Plant Mitochondria—Repair Mechanisms and Selection. Genome Biol. Evol. 6: 1448–1453. doi:10.1093/gbe/evu115

Logan D. C., Christensen A. C., 2018 Mitochondrial DNA repair and genome evolution, pp. 11–31 in Annual Plant Reviews, Plant Mitochondria, Ed. 2, edited by Logan D. C., Wiley-Blackwell, New York, NY

Cole, T. C. H., H. H. Hilger, and P. F. Stevens, 2017 Angiosperm phylogeny poster (APP) – Flowering plant systematics, 2017. PeerJ Preprints 5:e2320v4. https://doi.org/10.7287/peerj.preprints.2320v4.

Davila J. I., Arrieta-Montiel M. P., Wamboldt Y., Cao J., Hagmann J., *et al.*, 2011 Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. BMC Biol. 9: 64. doi:10.1186/1741-7007-9-64

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

Folkerts O., Hanson M. R., 1989 Three copies of a single recombination repeat occur on the 443 kb master circle of the Petunia hybrida 3704 mitochondrial genome. Nucleic Acids Res. 17: 7345–7357. doi:10.1093/nar/17.18.7345

Forner J., Weber B., Wietholter C., Meyer R. C., Binder S., 2005 Distant sequences determine 5′ end formation of cox3 transcripts in Arabidopsis thaliana ecotype C24. Nucleic Acids Res. 33: 4673–4682. doi:10.1093/nar/gki774

Franzke A., Lysak M. A., Al-Shehbaz I. A., Koch M. A., Mummenhoff K., 2011 Cabbage family affairs: the evolutionary history of Brassicaceae. Trends Plant Sci. 16: 108–116. doi:10.1016/j.tplants.2010.11.005

Gould S. J., Eldredge N., 1977 Punctuated Equilibria: The Tempo and Mode of Evolution Reconsidered. Paleobiology 3: 115–151. doi:10.1017/S0094837300005224

Grewe F., Edger P. P., Keren I., Sultan L., Pires J. C., *et al.*, 2014 Comparative analysis of 11 Brassicales mitochondrial genomes and the mitochondrial transcriptome of Brassica oleracea. Mitochondrion. 19: 135–143. doi:10.1016/j.mito.2014.05.008

Grewe F., Viehoever P., Weisshaar B., Knoop V., 2009 A trans-splicing group I intron and tRNA-hyperediting in the mitochondrial genome of the lycophyte Isoetes engelmannii. Nucleic Acids Res. 37: 5093–5104. doi:10.1093/nar/gkp53

Gualberto J. M., Mileshina D., Wallet C., Niazi A. K., Weber-Lotfi F., *et al.*, 2014 The plant mitochondrial genome: dynamics and maintenance. Biochimie 100: 107–120. doi:10.1016/j.biochi.2013.09.016

Gualberto J. M., Newton K. J., 2017 Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation. Annu. Rev. Plant Biol. 68: 225–252. doi:10.1146/annurev-arplant-043015-112232

Guo W., Grewe F., Fan W., Young G. J., Knoop V., *et al.*, 2016 Ginkgo and Welwitschia Mitogenomes Reveal Extreme Contrasts in Gymnosperm Mitochondrial Evolution. Mol. Biol. Evol. 33: 1448–1460. doi:10.1093/molbev/msw024

Guo W., Zhu A., Fan W., Mower J. P., 2017 Complete mitochondrial genomes from the ferns Ophioglossum californicum and Psilotum nudum are highly repetitive with the largest organellar introns. New Phytol. 213: 391–403. doi:10.1111/nph.14135

Hecht J., Grewe F., Knoop V., 2011 Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of Selaginella moellendorffii mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. Genome Biol. Evol. 3: 344–358. doi:10.1093/gbe/evr027

Klein M., Eckert-Ossenkopp U., Schmiedeberg I., Brandt P., Unseld M., *et al.*, 1994 Physical mapping of the mitochondrial genome of Arabidopsis thaliana by cosmid and YAC clones. Plant J. 6: 447–455. doi:10.1046/j.1365-313X.1994.06030447.x

Krisko A., Radman M., 2013 Biology of extreme radiation resistance: the way of Deinococcus radiodurans. Cold Spring Harb. Perspect. Biol. 5: a012765. doi:10.1101/cshperspect.a012765

Kurtz S., Schleiermacher C., 1999 REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15: 426–427. doi:10.1093/bioinformatics/15.5.426

Liu Y., Medina R., Goffinet B., 2014 350 my of mitochondrial genome stasis in mosses, an early land plant lineage. Mol. Biol. Evol. 31: 2586–2591. doi:10.1093/molbev/msu199

Lynch M., Koskella B., Schaack S., 2006 Mutation Pressure and the Evolution of Organelle Genomic Architecture. Science 311: 1727–1730. doi:10.1126/science.1118884

Mattimore V., Battista J. R., 1996 Radioresistance of Deinococcus radiodurans: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation. J. Bacteriol. 178: 633–637. doi:10.1128/jb.178.3.633-637.1996

Miller-Messmer M., Kuhn K., Bichara M., Le Ret M., Imbault P., *et al.*, 2012 RecA-dependent DNA repair results in increased heteroplasmy of the Arabidopsis mitochondrial genome. Plant Physiol. 159: 211–226. doi:10.1104/pp.112.194720

Mower J. P., Case A. L., Floro E. R., Willis J. H., 2012 Evidence against equimolarity of large repeat arrangements and a predominant master circle structure of the mitochondrial genome from a monkeyflower (Mimulus guttatus) lineage with cryptic CMS. Genome Biol. Evol. 4: 670–686. doi:10.1093/gbe/evs042

Palmer J. D., 1988 Intraspecific variation and multicircularity in Brassica mitochondrial DNAs. Genetics 118: 341–351.

Palmer J. D., Herbon L. A., 1986 Tricircular mitochondrial genomes of Brassica and Raphanus: reversal of repeat configurations by inversion. Nucleic Acids Res. 14: 9755–9764. doi:10.1093/nar/14.24.9755

Palmer J. D., Herbon L. A., 1988 Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J. Mol. Evol. 28: 87–97. doi:10.1007/BF02143500

Palmer J. D., Shields C. R., 1984 Tripartite structure of the Brassica campestris mitochondrial genome. Nature 307: 437–440. doi:10.1038/307437a0

Park S., Grewe F., Zhu A., Ruhlman T. A., Sabir J., *et al.*, 2015 Dynamic evolution of Geranium mitochondrial genomes through multiple horizontal and intracellular gene transfers. New Phytol. 208: 570–583. doi:10.1111/nph.13467

Parkinson C. L., Mower J. P., Qiu Y. L., Shirk A. J., Song K., *et al.*, 2005 Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. BMC Evol. Biol. 5: 73. doi:10.1186/1471-2148-5-73

Paszkiewicz G., Gualberto J. M., Benamar A., Macherel D., Logan D. C., 2017 Arabidopsis Seed Mitochondria Are Bioenergetically Active Immediately upon Imbibition and Specialize via Biogenesis in Preparation for Autotrophic Growth. Plant Cell 29: 109–128. doi:10.1105/tpc.16.00700

Repar J., Supek F., Klanjscek T., Warnecke T., Zahradka K., *et al.*, 2017 Elevated Rate of Genome Rearrangements in Radiation-Resistant Bacteria. Genetics 205: 1677–1689. doi:10.1534/genetics.116.196154

Richardson A. O., Rice D. W., Young G. J., Alverson A. J., Palmer J. D., 2013 The "fossilized" mitochondrial genome of Liriodendron tulipifera: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. BMC Biol. 11: 29. doi:10.1186/1741-7007-11-29

Sakamoto W., Kondo H., Murata M., Motoyoshi F., 1996 Altered mitochondrial gene expression in a maternal distorted leaf mutant of Arabidopsis induced by chloroplast mutator. Plant Cell 8: 1377–1390. doi:10.1105/tpc.8.8.1377

Schuster W., Brennicke A., 1994 The Plant Mitochondrial Genome: Physical Structure, Information Content, RNA Editing, and Gene Migration to the Nucleus. Annu. Rev. Plant Physiol. Plant Mol. Biol. 45: 61–78. doi:10.1146/annurev.pp.45.060194.000425

Sharma A., Gaidamakova E. K., Grichenko O., Matrosova V. Y., Hoeke V., *et al.*, 2017 Across the tree of life, radiation resistance is governed by antioxidant Mn(2+), gauged by paramagnetic resonance. Proc. Natl. Acad. Sci. USA 114: E9253–E9260. doi:10.1073/pnas.1713608114

Shearman J. R., Sonthirod C., Naktang C., Pootakham W., Yoocha T., *et al.*, 2016 The two chromosomes of the mitochondrial genome of a sugarcane cultivar: assembly and recombination analysis using long PacBio reads. Sci. Rep. 6: 31533. doi:10.1038/srep31533

Shedge V., Arrieta-Montiel M., Christensen A. C., Mackenzie S. A., 2007 Plant mitochondrial recombination surveillance requires unusual RecA and MutS homologs. Plant Cell 19: 1251–1264. doi:10.1105/tpc.106.048355

Siculella L., Damiano F., Cortese M. R., Dassisti E., Rainaldi G., *et al.*, 2001 Gene content and organization of the oat mitochondrial genome. Theor. Appl. Genet. 103: 359–365. doi:10.1007/s001220100568

Sloan D. B., 2013 One ring to rule them all? Genome sequencing provides new insights into the 'master circle' model of plant mitochondrial DNA structure. New Phytol. 200: 978–985. doi:10.1111/nph.12395

Sloan D. B., Alverson A. J., Chuckalovcak J. P., Wu M., McCauley D. E., *et al.*, 2012 Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 10: e1001241. doi:10.1371/journal.pbio.1001241

Sloan D. B., Alverson A. J., Storchova H., Palmer J. D., Taylor D. R., 2010 Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm Silene latifolia. BMC Evol. Biol. 10: 274. doi:10.1186/1471-2148-10-274

Smith D. R., 2016 The mutational hazard hypothesis of organelle genome evolution: 10 years on. Mol. Ecol. 25: 3769–3775. doi:10.1111/mec.13742

Smith D. R., Lee R. W., Cushman J. C., Magnuson J. K., Tran D., *et al.*, 2010 The Dunaliella salina organelle genomes: large sequences, inflated with intronic and intergenic DNA. BMC Plant Biol. 10: 83. doi:10.1186/1471-2229-10-83

Stern D. B., Palmer J. D., 1986 Tripartite mitochondrial genome of spinach: physical structure, mitochondrial gene mapping, and locations of transposed chloroplast DNA sequences. Nucleic Acids Res. 14: 5651–5666. doi:10.1093/nar/14.14.5651

Sugiyama Y., Watase Y., Nagase M., Makita N., Yagura S., *et al.*, 2005 The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial

genome: comparative analysis of mitochondrial genomes in higher plants. Mol. Genet. Genomics 272: 603–615. doi:10.1007/s00438-004-1075-8

Wallet C., Le Ret M., Bergdoll M., Bichara M., Dietrich A., *et al.*, 2015 The RECG1 DNA Translocase Is a Key Factor in Recombination Surveillance, Repair, and Segregation of the Mitochondrial DNA in Arabidopsis. Plant Cell 27: 2907–2925. doi:10.1105/tpc.15.00680

Ward B. L., Anderson R. S., Bendich A. J., 1981 The mitochondrial genome is large and variable in a family of plants (cucurbitaceae). Cell 25: 793–803. doi:10.1016/0092-8674(81)90187-

Wolfe K., Li W., Sharp P., 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial,chloroplast and nuclear DNAs. Proc. Natl. Acad. Sci. USA 84: 9054–9058. doi:10.1073/pnas.84.24.9054

CHAPTER 4

MITOCHONDRIAL DNA REPAIR IN AN *ARABIDOPSIS THALIANA* URACIL DNA N-GLYCOSYLASE MUTANT

ABSTRACT

Substitution rates in plant mitochondrial genes are extremely low, indicating strong selective pressure as well as efficient repair. Plant mitochondria possess base excision repair pathways, however, many repair pathways such as nucleotide excision repair and mismatch repair appear to be absent. In the absence of these pathways, many DNA lesions must be repaired by a different mechanism. To test the hypothesis that double-strand break repair (DSBR) is that mechanism, we maintained independent self-crossing lineages of plants deficient in uracil-N-glycosylase (UNG) for 10 generations to determine the repair outcomes when that pathway is missing. Surprisingly, no single nucleotide polymorphisms (SNPs) were fixed in any line in generation 10. The pattern of heteroplasmic SNPs was also unaltered through 10 generations. Clearly DNA maintenance in reproductive meristem mitochondria is effective in the absence of UNG. In mature leaves, there is evidence of aborted DSBR at short homologies, indicating an increase in double strand breaks. In young leaves there is no evidence of aborted DSBR, indicating that mitochondria in meristematic tissue have access to full homologous repair templates. These results indicate that double strand break repair is a general system of repair in plant mitochondria. The existence of this general system may explain the seemingly anomalous differences in plant mitochondria between low mutation rates in genes and rearrangements in non-genes.

INTRODUCTION

Plant mitochondrial genomes have very low base substitution rates, while also expanding and rearranging rapidly (Wolfe *et al.*, 1987, Palmer and Herbon, 1988, Drouin *et al.*, 2008, Richardson *et al.*, 2013). The low substitution rate and the high rearrangement rate of plant mitochondria can be explained by selection and the specific DNA damage repair mechanisms available. These mechanisms can also account for the observations of genome expansion found in land plant mitochondria. The low nonsynonymous substitution rates in protein coding genes indicates that selective pressure to maintain the genes is high, and the low synonymous substitution rates indicate that the DNA repair mechanisms are very accurate (Sloan and Taylor, 2010, Wynn and Christensen 2015). Despite the low mutation rate of mitochondrial genes over evolutionary time, mitochondrial genomes in mature cells accumulate DNA damage that is not repaired (Kumar *et al.* 2014). This indicates that there are fundamental differences between DNA maintenance in genomes meant to be passed on to the next generation and genomes that are not. In meristematic cells, where cell division occurs, mitochondria fuse together to form a large mitochondrion (Seguí-Simarro and Staehelin 2008). This fusion brings mitochondrial genomes together for genome replication, but also ensures that there is a homologous template available for DNA repair. These meristematic cells eventually produce the reproductive tissue of a plant; from embryogenesis to egg cell production, the mitochondrial genomes inherited from parents and passed down to offspring will have homologous templates available to them (Seguí-Simarro and Staehelin 2009).

However, little is known about the multiple pathways of DNA repair in plant mitochondria. So far, there is no evidence of nucleotide excision repair (NER), nor

mismatch repair (MMR) in plant mitochondria (Boesch *et al.*, 2009, Gualberto and Newton, 2017). It has been hypothesized that in plant mitochondria, the types of DNA damage that are usually repaired through NER and MMR are repaired through double-strand break repair (DSBR) (Christensen, 2014, Christensen, 2018). Plant mitochondria have the nuclear-encoded base excision repair (BER) pathway enzyme Uracil DNA glycosylase (UNG) (Boesch *et al.*, 2009). UNG is an enzyme that can recognize and bind to uracil in DNA and begin the process of base excision repair by enzymatically excising the uracil (U) residue from single stranded or double stranded DNA (Cordoba-Cañero *et al.*, 2010). Uracil can appear in a DNA strand due to the spontaneous deamination of cytosine, or by the misincorporation of dUTP during replication (Krokan *et al.*, 1997). Unrepaired uracil in DNA can lead to G-C to A-T transitions within the genome.

Few pathways of repair besides BER and DSBR are known in plant mitochondria, and it is possible that many lesions, including mismatches, are repaired by creating double-strand breaks and using a template to repair both strands. Our hypothesis is that DSBR accounts for most of the repair in meristematic plant mitochondria, and both error-prone and accurate subtypes of DSBR lead to the observed patterns of genome evolution (Christensen, 2013). One way of testing this is to eliminate the pathway of uracil base excision repair and ask if the G-U mispairs that occur by spontaneous deamination are repaired, and if so are instead repaired by DSBR. In this work we examine an *Arabidopsis thaliana* UNG knockout line and investigate the effects on the mitochondrial genome over many generations. To further disrupt the genome, we express the cytidine deaminase APOBEC3G in the *Arabidopsis* mitochondria to increase the rate of cytosine deamination and accelerate DNA damage.

One of the hallmarks of DSBR in plant mitochondria is the effect on the non-tandem repeats that exist in virtually all plant mitochondria (Wynn and Christensen2019). The *Arabidopsis thaliana* mitochondrial genome contains two pairs of very large repeats (4.2 and 6.6kb) that commonly undergo recombination (Palmer and Shields, 1984, Klein *et al.*, 1994, Unseld *et al.*, 1997) producing multiple isoforms of the genome. The mitochondrial genome also contains many smaller repeats between 50 and 1000 base pairs, (Unseld *et al.*, 1997, Arrieta-Montiel *et al.*, 2009, Davila *et al.*, 2011, Wynn and Christensen, 2019). In wild type plants, these intermediate-size repeats recombine at very low rates. However, these repeats have been shown to recombine with ectopic repeat copies at higher rates in several mutants in these DSBR genes, such as *msh1* and *reca3* (Abdelnoor *et al.*, 2003, Shedge *et al.*, 2007, Miller-Messmer *et al.*, 2012). Thus genome dynamics around intermediate repeats can be an indicator of increased DSBs. In this work we show that a loss of uracil base excision repair leads to alterations in repeat dynamics.

Numerous proteins known to be involved in the processing of plant mitochondrial DSBs have been characterized. Plants lacking the activity of mitochondrially targeted recA homologs have been shown to be deficient in DSBR (Odahara *et al.*, 2007, Miller-Messmer *et al.*, 2012). In addition, it has been hypothesized that the plant MSH1 protein may be involved in binding to DNA lesions and initiating DSBs (Christensen, 2014, Christensen, 2018). The MSH1 protein contains a mismatch binding domain fused to a GIY-YIG type endonuclease domain which may be able to make DSBs (Abdelnoor *et al.*, 2006, Kleinstiver *et al.*, 2013). In this work we provide evidence that in the absence of

mitochondrial UNG activity, several genes involved in DSBR, including *MSH1*, are transcriptionally upregulated, providing a possible explanation for the increased DSBR.

RESULTS

Lack of UNG activity in mutants

It has previously been reported that cell extracts of the *Arabidopsis thaliana ung* T-DNA insertion strain used in this experiment, GK-440E07 (ABRC seed stock CS308282), shows no uracil glycosylase activity (Boesch 2009). To increase the rate of cytosine deamination in the mitochondrial genome and show that effects of the UNG knockout on mitochondrial mutation rates could be detected, the human APOBEC3G – CTD 2K3A cytidine deaminase (A3G) (Chen *et al.*, 2007) was expressed in both wild-type and *ung* Arabidopsis thaliana lines and targeted to the mitochondria by an amino-terminal fusion of the 62 amino acid mitochondrial targeting peptide (MTP) from Alternative Oxidase (AOX1A). Fluorescence microscopy of Arabidopsis thaliana expressing an MTP-A3G-GFP fusion shows that the MTP-A3G construct is expressed and targeted to the mitochondria (See Figure 4.1).

We expected that in the absence of UNG there would be an increase in G-C to A-T substitution mutations. To test this prediction, we sequenced both a wild-type *Arabidopsis* plant expressing the MTP-A3G construct (Col-0 MTP-A3G) and a *ung* plant expressing the MTP-A3G construct (*ung* MTP-A3G) using an Illumina Hi-Seq4000 system. Mitochondrial sequences from these plants were aligned to the Columbia-0 reference genome using BWA-MEM (Li, 2013) and single nucleotide polymorphisms were identified using VarDict (Lai *et al.*, 2016).

There were no SNPs that reached fixation (an allele frequency of 1) in either plant. Mitochondrial genomes are not diploid; each cell can have many copies of the mitochondrial genome. Therefore, it is possible that an individual plant could accumulate low frequency mutations in some of the mitochondrial genomes in the cell. VarDict was used to detect heteroplasmic SNPs at allele frequencies as low as 0.05. VarDict's sensitivity in calling low frequency SNPs scales with depth of coverage and quality of the sample, so it is not possible to directly compare heteroplasmic mutation rates in samples with different depths of coverage. However, because the activity of the UNG protein is specific to uracil, the absence of the UNG protein should not have any effect on mutation rates other than G-C to A-T transitions. We therefore considered heteroplasmic mutations that are not G-C to A-T transitions to be the background rate of heteroplasmic SNP accumulation in plant mitochondria. We therefore compared the numbers of G-C to A-T transitions to all other mutations. If the *ung* MTP-A3G line is accumulating G-C to A-T transitions at a faster rate than the Col-0 MTP-A3G line, we would expect to see that as an increased ratio of G-C to A-T transitions compared to other mutation types. The Col-0 MTP-A3G plant had a heteroplasmic GC-AT/total SNPs ratio of 0.59, while the *ung* MTP-A3G plant had a heteroplasmic GC-AT/total SNPs ratio of 0.68 (Table 4.1). When the rate of cytosine deamination is increased by the activity of APOBEC3G, the *ung* plant accumulates heteroplasmic GC-AT SNPs at a faster rate than wild-type, and our computational pipeline is able to detect this difference.

Mutation accumulation in the absence of UNG

To determine the effects of the UNG knockout across multiple generations under normal conditions, without the presence of APOBEC3G in the mitochondria. We

performed a mutation accumulation study (Halligan and Keightley, 2009). We chose 23 different *ung* homozygous plants derived from one hemizygous parent. These 23 plants were designated as generation 1 *ung* and were allowed to self-cross. The next generation was derived by single-seed descent from each line, and this was repeated until generation 10 *ung* plants were obtained. Leaf tissue and progeny seeds from each line were kept at each generation.

The leaf tissue from generation 10 of the 23 *ung* mutation accumulation lines and a wild-type Col-0 were sequenced and analyzed with VarDict as described above. Similar to the MTP-A3G plants, there were no SNPs in any of our *ung* mutation accumulation lines that had reached fixation (an allele frequency of one). In contrast, there is little difference in the ratios of GC-AT/total SNPs between the *ung* lines and Col-0 (see Table 4.1). Because detection of low frequency SNPs depends on read depth, we only analyzed the 7 *ung* samples with an average mitochondrial read depth above 125x for this comparison. In the absence of a functional UNG protein and under normal greenhouse physiological conditions, plant mitochondria do not accumulate cytosine deamination mutations at an increased rate.

Nuclear Mutation Accumulation

UNG is the only Uracil Glycosylase in *Arabidopsis thaliana* and may be active in the nucleus as well as the mitochondria. To test for nuclear mutations due to the absence of UNG, sequences were aligned to the Columbia-0 reference genome using BWA-MEM and single nucleotide polymorphisms were identified using Bcftools Call (Li, 2011). No increase in GC-AT transitions was detected in any line (Table 4.2)

Increased Double-strand break repair

If most DNA damage in plant mitochondria is repaired by double-strand break repair (DSBR), supplemented by base excision repair (Boesch *et al.*, 2009), then in the absence of the Uracil-N-glycosylase (UNG) pathway we predict an increase in DSBR. To find evidence of this we used quantitative PCR (qPCR) to assay crossing over between identical non-tandem repeats, which increases when DSBR is increased (Shedge *et al.*, 2007, Miller-Messmer *et al.*, 2012, Wallet *et al.*, 2015). Different combinations of primers in the unique sequences flanking the repeats allow us to determine the relative copy numbers of parental-type repeats and low frequency recombinants (Figure 4.2A). The mitochondrial genes *cox2* and *rrn18* were used to standardize relative amplification between lines. We and others (Davila *et al.*, 2011, Wallet *et al.*, 2015) have found that some of the intermediate repeats are well-suited for qPCR analysis and are sensitive indicators of ectopic recombination, increasing in repair-defective mutants and when drugs are used to increase double-strand breaks. We analyzed the three repeats known as Repeats B, D, and L (Arrieta-Montiel *et al.*, 2009) in both young leaves and mature leaves. In young leaves, there is no significant difference in the amounts of parental or recombinant forms between *ung* lines and Col-0 (Figure 4.2B). In mature leaves, all three repeats show significant reductions in the parental 2/2 form, while repeat B also shows a reduction in the parental 1/1 form (unpaired T-test $p<0.05$, Figure 4.2C).

Alternative Repair Pathway Genes

Because the *ung* mutants show increased double-strand break repair but not an increase of G-C to A-T transition mutations, we infer that the inevitable appearance of uracil in the DNA is repaired via conversion of a G-U pair to a double-strand break and

efficiently repaired by the DSBR pathway. If this is true, genes involved in the DSBR processes of breakage, homology surveillance and strand invasion in mitochondria will be up-regulated in *ung* mutants. To test this hypothesis, we assayed transcript levels of several candidate genes known to be involved in DSBR (Abdelnoor *et al.*, 2003, Khazi *et al.*, 2003, Edmondson *et al.*, 2005, Odahara *et al.*, 2007, Shedge *et al.*, 2007, Arrieta-Montiel *et al.*, 2009, Miller-Messmer *et al.*, 2012, Gualberto *et al.*, 2014, Wallet *et al.*, 2015, Gualberto and Newton, 2017) in *ung* lines compared to wild-type using RT-PCR. *MSH1* and *RECA2* were significantly upregulated in *ung* lines (*MSH1*: 5.60-fold increase, unpaired T-test $p<0.05$. *RECA2*: 3.19-fold increase, unpaired T-test $p<0.05$ – see Figure 4.3). The single-strand binding protein gene *OSB1* was also measurably upregulated in *ung* lines (3.07-fold increase, unpaired T-test $p=0.053$). *RECA3*, *SSB*, and *WHY2* showed no differential expression compared to wild-type (unpaired T-test $p>0.05$).

DISCUSSION

In the mitochondrion as well as in the nucleus and chloroplast, cytosine is subject to deamination to uracil. This could potentially lead to transition mutations, and is dealt with by a specialized base excision repair pathway. The first step in this pathway is hydrolysis of the glycosidic bond by the enzyme Uracil-N-glycosylase (UNG), leaving behind an abasic site (Cordoba-Cañero *et al.*, 2010). An AP endonuclease can then cut the DNA backbone, producing a 3’ OH and a 5’ dRP. Both DNA polymerases found in *Arabidopsis* mitochondria, POL1A and POL1B, exhibit 5’-dRP lyase activity, allowing them to remove the 5’ dRP and polymerize a new nucleotide replacing the uracil (Trasviña-Arenas *et al.*, 2018). In the absence of functional UNG protein, cytosine will still be deaminated in plant mitochondrial genomes, so efficient removal of uracil must be

through a different repair mechanism, most likely DSBR (Christensen, 2014, Christensen, 2018). We have found that in *ung* mutant lines, there are significant changes in the relative abundance of parental and recombinant forms of intermediate repeats, as well as an increase in the expression of genes known to be involved in DSBR, consistent with this hypothesis.

We have shown that when cytosine deamination is increased by the expression of the APOBEC3G cytidine deaminase in plant mitochondria, *ung* lines accumulate more G-C to A-T transitions than wild-type. Surprisingly, we have also found that under normal cellular conditions, without the added deamination activity of APOBEC3G, *ung* lines do not accumulate G-C to A-T transition mutations at a higher rate than wild-type. This finding is particularly surprising given the presumed bottlenecking of mitochondrial genomes during female gametogenesis, and given the deliberate bottleneck in the experimental design of single-seed descent for 10 generations. This finding supports the hypothesis that plant mitochondria have a very efficient alternative damage surveillance system that can prevent G-C to A-T transitions from becoming fixed in the mitochondrial population.

The angiosperm MSH1 protein consists of a DNA mismatch binding domain fused to a double-stranded DNA endonuclease domain (Abdelnoor *et al.*, 2006, Kleinstiver *et al.*, 2013). Although mainly characterized for its role in recombination surveillance (Shedge *et al.*, 2007), MSH1 is a good candidate for a protein that may be able to recognize and bind to various DNA lesions and make DSBs near the site of the lesion, thus funneling these types of damage into the DSBR pathway. With many mitochondria and many mitochondrial genomes in each cell there are numerous available

templates to accurately repair DSBs through homologous recombination, making this a plausible mechanism of genome maintenance. Here we show that in *ung* lines, *MSH1* is transcriptionally upregulated more than 5-fold compared to wild-type. This further supports the hypothesis that MSH1 initiates repair in plant mitochondria by creating a double-strand break at G-U pairs, and possibly other mismatches and damaged bases.

Several other proteins involved in processing plant mitochondrial DSBs have been characterized. The RECA homologs, RECA2 and RECA3, are homology search and strand invasion proteins (Xu and Marians, 2002, Khazi *et al.*, 2003, McGrew and Knight, 2003, Odahara *et al.*, 2007, Shedge *et al.*, 2007, Rowan *et al.*, 2010, Miller-Messmer *et al.*, 2012). The two mitochondrial RECAs share much sequence similarity, however RECA2 is dual targeted to both the mitochondria and the chloroplast, while RECA3 is found only in the mitochondria (Shedge *et al.*, 2007, Miller-Messmer *et al.*, 2012). RECA3 also lacks a C-terminal motif present on RECA2 and most other homologs. This motif has been shown to modulate the ability of RECA proteins to displace competing ssDNA binding proteins in E. coli (Eggler *et al.*, 2003). *Arabidopsis reca2* mutants are seedling lethal and both *reca2* and *reca3* lines show increased ectopic recombination at intermediate repeats (Miller-Messmer *et al.*, 2012). Arabidopsis RECA2 has functional properties that RECA3 cannot perform, such as complementing a bacterial *recA* mutant during the repair of UV-C induced DNA lesions (Khazi *et al.*, 2003). Here we show that in *ung* lines, *RECA2* is transcriptionally upregulated more than 3-fold compared to the wild-type. However, *RECA3* is not upregulated in *ung* lines. Responding to MSH1-initiated DSBs may be one of the functions unique to RECA2. The increased expression

of *RECA2* in the absence of a functional UNG protein is further evidence that uracil arising in DNA may be repaired through the mitochondrial DSBR pathway.

The ssDNA binding protein OSB1 is upregulated over 3-fold. At a double strand break, OSB1 competitively binds to ssDNA and recruits the RECA proteins to promote the repair of a double strand break by a homologous template and avoid the error-prone microhomology-mediated end-joining pathway (García-Medel *et al.* 2019).

We also tested the differential expression of other genes known to be involved in processing mitochondrial DSBs. The single stranded binding protein genes *WHY2* and *SSB* were not found to be differentially expressed at the transcript level compared to wild-type. The presence of different ssDNA binding proteins influences which pathway of DSBR a break is repaired by (García-Medel *et al.* 2019). Increased amounts of WHY2 and SSB may not be needed to accurately repair induced DSBs in the *ung* lines.

The specific patterns of recombination at mitochondrial intermediate repeats are different between wild-type, *ung* mutants, and DSBR mutants. In *msh1* lines, there is an increase in repeat recombination likely due to relaxed homology surveillance in the absence of the MSH1 protein (Shedge *et al.*, 2007). In mutant lines of ssDNA binding proteins involved in DSBR, such as *recA2, recA3,* and *osb1* (Miller-Messmer *et al.* 2012, Zaegel *et al.* 2006), there is an increase in repeat recombination due to differences in the way DNA ends are handled in the absence of the ssDNA binding proteins. In *ung* lines, the mitochondrial recombination machinery is still intact, so any differences in repeat recombination between *ung* lines and wild-type are not due to differences in processing the DSB, but due to the increase in the amount of DSBs in the absence of UNG.

Plant mitochondrial genomes likely replicate by recombination-dependent replication (RDR) (Backert and Börner, 2000). Most organellar genome replication occurs in meristematic tissue, where mitochondria fuse together to form a large, reticulate mitochondrion (Seguí-Simarro *et al.* 2008). This mitochondrial fusion provides a means to homogenize mtDNA by gene conversion, and repair lesions through homologous recombination (Rose and McCurdy, 2017). As cells differentiate and age, organellar genomes degrade (Bendich 2003). Clearly there is a difference in mitochondrial DNA maintenance in mature cells compared to young cells, either due to a lack of DNA repair in mature mitochondria, or a difference in DNA repair mechanism.

In young leaves, there is no significant difference in recombination at intermediate repeats between *ung* lines and wild-type. In meristematic cells, mitochondrial fusion brings many copies of the mitochondrial genome together, providing many possible templates for the accurate repair of Uracil by homologous recombination. In mature leaves, *ung* lines show a reduction in parental type repeats compared to wild-type. This indicates that there is an increase in double strand breaks and an increase in attempted DSBR by break-induced replication at intermediate repeats. However, MSH1 aborts recombination at the heteroduplexes that form during recombination at intermediate repeats (Shedge *et al.*, 2007)(see Figure 4.4). The dispersal of subgenomic molecules into individual mitochondria during cell maturation and differentiation increases the difficulty of finding a long homology for DNA repair, leading to an increase in aborted recombination at intermediate repeats and may help explain the degradation of mtDNA in mature cells.

To determine the outcomes of genomic uracil in the absence of a functional UNG protein, we sequenced the genomes of several *ung* lines. No fixed mutations of any kind were found in *ung* lines, even after 10 generations of self-crossing. Low frequency heteroplasmic SNPs were found in both wild-type and *ung* lines, but *ung* lines showed no difference in the ratio of G-C to A-T transitions to other mutation types when compared to wild-type.

Clearly the double-strand break repair pathway in plant mitochondria can repair uracil in DNA sufficiently to prevent mutation accumulation in the absence of the UNG protein. Why then has the BER pathway been conserved in plant mitochondria while NER and MMR have apparently been lost? DSBR protects the genome efficiently from mutations in plants growing under ideal conditions, but cannot successfully repair all lesions when the rate of cytosine deamination is increased (see Table 4.1). Throughout the evolutionary history of *Arabidopsis thaliana* and into the present, wild growing plants are exposed to a range of growth conditions and stresses that experimental plants in a greenhouse avoid. The rate of spontaneous cytosine deamination increases with increasing temperature (Drake and Baltz, 1976, Lewis *et al.*, 2016), so DSBR alone may not be able repair the extent of uracil found in DNA across the range of temperatures a wild plant would experience, providing the selective pressure to maintain a distinct BER pathway in plant mitochondria.

Here we have provided evidence that in the absence of a dedicated BER pathway, plants growing in greenhouse growth chamber conditions do not accumulate mitochondrial SNPs at an increased rate. Instead, DNA damage is accurately repaired by double-strand break repair which also causes an increase in ectopic recombination at

identical non-tandem repeats. It has recently been shown that mice lacking a different mitochondrial BER protein, oxoguanine glycosylase, also do not accumulate mitochondrial SNPs (Kauppila *et al.*, 2018). Here we show that in plants base-excision repair by UNG is similarly unnecessary to prevent mitochondrial mutations in growth chamber conditions. Perhaps a generalized system of DNA repair also exists in mammalian mitochondria similar to the broad capacity of DSBR to repair different lesions in plant mitochondria. Clearly DSBR is efficient and accurate, and the presence of the UNG pathway reduces ectopic recombination slightly and can successfully repair uracil in DNA even if the rate of cytosine deamination is increased. Double strand break repair and recombination are important mechanisms in the evolution of plant mitochondrial genomes, but many key enzymes and steps in the repair pathway are still unknown. Further identification and characterization of these missing steps is sure to provide additional insight into the unique evolutionary dynamics of plant mitochondrial genomes.

METHODS

Plant growth conditions

Arabidopsis thaliana Columbia-0 (Col-0) seeds were obtained from Lehle Seeds (Round Rock, TX, USA). UNG (AT3G18630) T-DNA insertion hemizygous lines were obtained from the Arabidopsis Biological Resource Center, line number CS308282. Hemizygous T-DNA lines were self-crossed to obtain homozygous lines (Genotyping primers: wild-type 5’-TGTCAAAGTCCTGCAATTCTTCTCACA-3’ and 5’-TCGTGCCATATCTTGCAGACCACA-3’,

ung 5’-ATAATAACGCTGCGGACATCTACATTTT-3’ and

5'-ACTTGGAGAAGGTAAAGCAATTCA-3'). All plants were grown in walk-in growth chambers under a 16:8 light:dark schedule at 22^{o}C. Plants grown on agar were surface sterilized and grown on 1x Murashige and Skoog Basal Medium (MSA) with Gamborg's vitamins (Sigma) with 5µg/mL Nystatin Dihydrate to prevent fungal contamination.

Vector construction

The APOBEC3G gene was synthesized by Life Technologies Gene Strings (Carpenter et al 2010) using *Arabidopsis thaliana* preferred codons and including the 62 amino acid mitochondrial targeting peptide (MTP) from Alternative Oxidase on the N-terminus of the translated protein. The MTP-APOBEC3G construct was cloned into the vector pUB-DEST (NCBI:taxid1298537) driven by the ubiquitin (UBQ10) promoter and transformed into wild-type and *ung Arabidopsis thaliana* plants by the *Agrobacterium* floral dip method (Clough and Bent 1998). To ensure proper mitochondrial targeting of the MTP-APOBEC3G construct, the construct was cloned into pK7FWG2 with a C-terminal GFP fusion (Karimi *et al.* 2002). *Arabidopsis thaliana* plants were again transformed by the *Agrobacterium* floral dip method and mitochondrial fluorescence was confirmed with fluorescence microscopy.

RT-PCR

RNA was extracted from young leaves of plants grown in soil during *ung* generation ten (Onate-Sanchez and Vicente-Carbajosa, 2008). Reverse transcription using Bio-Rad iScript was performed and the resulting cDNA was used as a template for qPCR to measure relative transcript amounts. Quantitative RT-PCR data was normalized using *UBQ11* as a housekeeping gene control. Reactions were performed in a Bio-Rad CFX96 thermocycler using 96 well plates and a reaction volume of 20µL/well. SYBRGreen

mastermix (Bio-Rad) was used in all reactions. Three biological and three technical replicates were used for each amplification. Primers are listed in Table S2. The MIQE guidelines were followed (Bustin *et al.*, 2009) and primer efficiencies are listed in Table S3. The thermocycling program for all RT-qPCR was a ten-minute denaturing step at 95^{o} followed by 45 cycles of 10s at 95^{o}, 15s at 60^{o}, and 13s at 72^{o}. Following amplification, melt curve analysis was done on all reactions to ensure target specificity. The melt curve program for all RT-qPCR was from 65^{o}-95^{o} at 0.5^{o} increments for 5s each.

Repeat recombination qPCR

DNA was collected from the mature leaves of Columbia-0 and generation ten *ung* plants using the CTAB DNA extraction method (Allen *et al.*, 2006). qPCR was performed using primers from the flanking sequences of the intermediate repeats. Primers are listed in Table S1. Using different combinations of forward and reverse primers, either the parental or recombinant forms of the repeat can be selectively amplified (see Figure 1A). The mitochondrially-encoded *COX2* and *RRN18* genes were used as standards for analysis. Reactions were performed in a Bio-Rad CFX96 thermocycler using 96 well plates with a reaction volume of 20µL/well. SYBRGreen mastermix (Bio-Rad) was used in all reactions. Three biological and three technical replicates were used for each reaction. The thermocycling program for all repeat recombination qPCR was a ten-minute denaturing step at 95^{o} followed by 45 cycles of 10s at 95^{o}, 15s at 60^{o}, and a primer specific amount of time at 72^{o} (extension times for each primer pair can be found in Table S1). Following amplification, melt curve analysis was done on all reactions to ensure target specificity. The melt curve program for all qPCR was from 65^{o}-95^{o} at 0.5^{o} increments for 5s each.

DNA sequencing

DNA extraction from frozen young leaves of Columbia-0, generation 10 *ung*, and APOBEC3G plants was done by a modification of the SPRI magnetic beads method of Rowan *et al* (Rowan *et al.*, 2015, Rowan *et al.*, 2017). Genomic libraries for paired-end sequencing were prepared using a modification of the Nextera protocol (Caruccio, 2011), modified for smaller volumes following Baym et al (Baym *et al.*, 2015). Following treatment with the Nextera Tn5 transpososome 14 cycles of amplification were done. Libraries were size-selected to be between 400 and 800bp in length using SPRI beads (Rowan *et al.*, 2017). Libraries were sequenced with 150bp paired-end reads on an Illumina HiSeq 4000 by the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley.

Reads were aligned using BWA-MEM v0.7.12-r1039 (Li, 2013). The reference sequence used for alignment was a file containing the improved Columbia-0 mitochondrial genome (accession BK010421.1) (Sloan *et al.*, 2018) as well as the TAIR 10 *Arabidopsis thaliana* nuclear chromosomes and chloroplast genome sequences (Berardini *et al.*, 2015). Using Samtools v1.3.1 (Li *et al.*, 2009), bam files were sorted for uniquely mapped reads for downstream analysis.

Organellar variants were called using VarDict (Lai *et al.*, 2016). To minimize the effects of sequencing errors, SNPs called by VarDict were filtered by the stringent quality parameters of Allele Frequency ≥ 0.05, Qmean ≥ 30, MQ ≥ 30, NM ≤ 3, Pmean ≥ 8, Pstd $= 1$, AltFwdReads ≥ 3, and AltRevReads ≥ 3. The mitochondrial reference genome positions corresponding to *RRN18* and *RRN26* were excluded from analysis because they have similarity to bacterial 16S and 23S ribosomal RNAs, respectively. Sequencing reads

from contaminating soil bacteria can be misaligned to these positions and falsely called as low frequency SNPs. No other mitochondrial sequences show enough similarity to bacterial genes to be misaligned by BWA MEM.

Nuclear variants were called using Samtools mpileup (v. 1.3.1) and Bcftools call (v. 1.2) and filtered for a call quality of 30. To avoid false positives, a 5 Mb region of each chromosome was used for analysis, avoiding centromeric and telomeric regions.

Accession Numbers

Fastq files generated from Illumina sequencing of *ung* lines and wild-type control are available from the Sequence Read Archive, BioProject ID PRJNA492503.

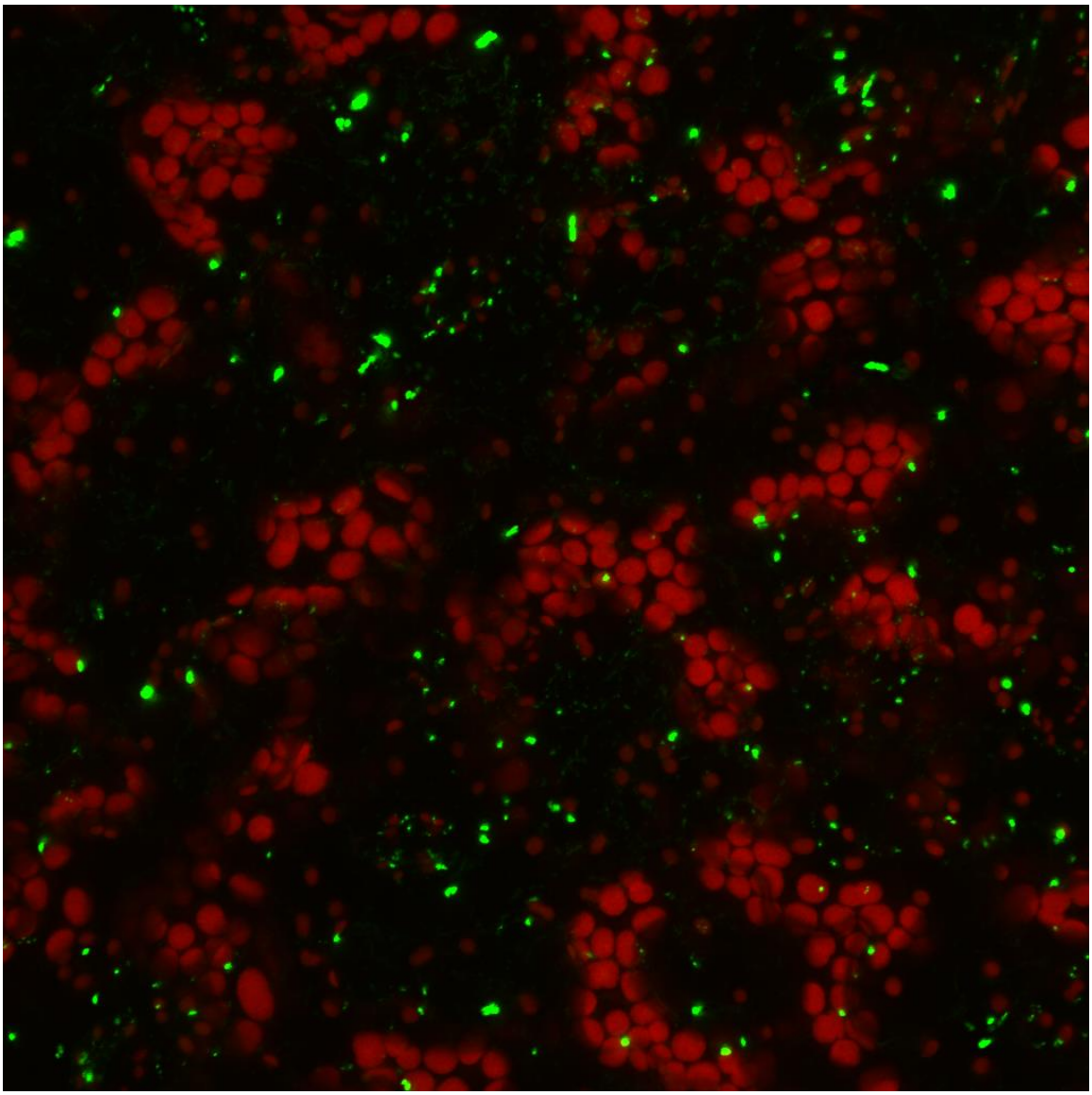

Figure 4.1: Mitochondrial targeting of a GFP labeled MTP-APOBEC3G construct. Fluorescence microscopy of an *Arabidopsis thaliana* plant transformed with an MTP-APOBEC3G-GFP construct. Green mitochondria indicate the proper expression and targeting of the construct. Autofluorescence of chloroplasts can be seen in red.

Table 4.1: Heteroplasmic mitochondrial SNPs in Col-0 wild-type, *ung* mutant lines, Col-0 MTP-A3G, and *ung* MTP-A3G. SNPs were called using VarDict as described in Methods. SNP counts are shown for the entire mitochondrial genome, sorted by the type of change. Only lines with average mitochondrial depth greater than 125x are used in this analysis

	GC-AT	GC-TA	GC-CG	AT-GC	AT-TA	AT-CG	Total	GC-AT/total
Col-0	41	8	6	6	7	3	71	0.577465
ung115	19	6	3	4	3	3	38	0.5
ung159	20	5	3	5	5	1	39	0.512821
ung163	38	10	6	5	7	3	69	0.550725
ung176	37	10	4	6	6	3	66	0.560606
ung198	31	7	5	12	10	4	69	0.449275
ung201	37	7	5	12	8	3	72	0.513889
ung203	28	5	3	18	4	2	60	0.466667
Col-0 MTP-A3G	44	11	4	3	9	3	74	0.594595
ung MTP-A3G	81	7	4	20	5	2	119	0.680672

Table 4.2: Nuclear SNPs in Col-0 wild-type, *ung* mutant lines, Col-0 MTP-A3G, and *ung* MTP-A3G. SNPs were called using Bcftools Call as described in Methods. SNP counts are shown for 5Mb regions of each chromosome.

	GC-AT	Total SNPs	Ratio
Col-0	287	2173	0.132075
ung115	1207	10967	0.110057
ung159	1396	12676	0.110129
ung163	260	2281	0.113985
ung176	650	6427	0.101136
ung198	1301	11679	0.111397
ung201	1311	13713	0.095603
ung203	1313	12702	0.10337
Col-0 MTP-A3G	334	2756	0.12119
ung MTP-A3G	888	7310	0.121477

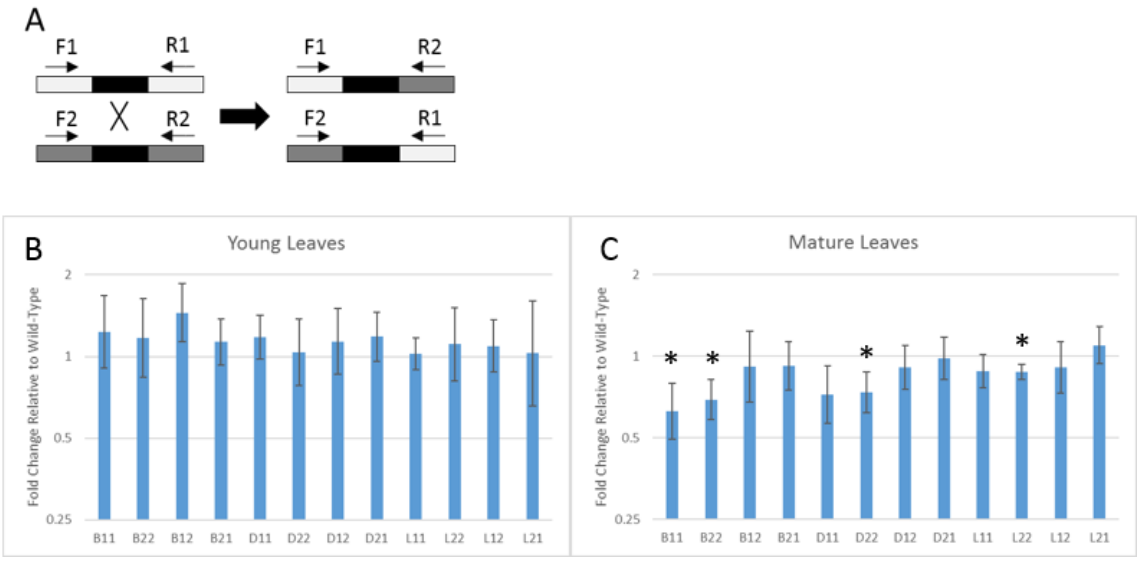

Figure 4.2: qPCR analysis of intermediate repeat recombination in *ung* lines compared to wild-type. Recombination at intermediate repeats is an indicator of increased double strand breaks in plant mitochondrial genomes. **A)** Primer scheme for detecting parental and recombinant repeats. Using different combinations of primers that anneal to the unique sequence flanking the repeats, either parental type (1/1 and 2/2) or recombinant type (1/2 and 2/1) repeats can be amplified **B)** Fold change of intermediate repeats in young leaves of *ung* lines relative to wild-type. Error bars are standard deviation of three biological replicates. **C)** Fold change of intermediate repeats in mature leaves of *ung* lines relative to wild-type. Error bars are standard deviation of three biological replicates. B1/1, B2/2, D2/2, and L2/2 show significant reduction in copy number (unpaired, 2-tailed Student's t-test, * indicates $p<0.05$)

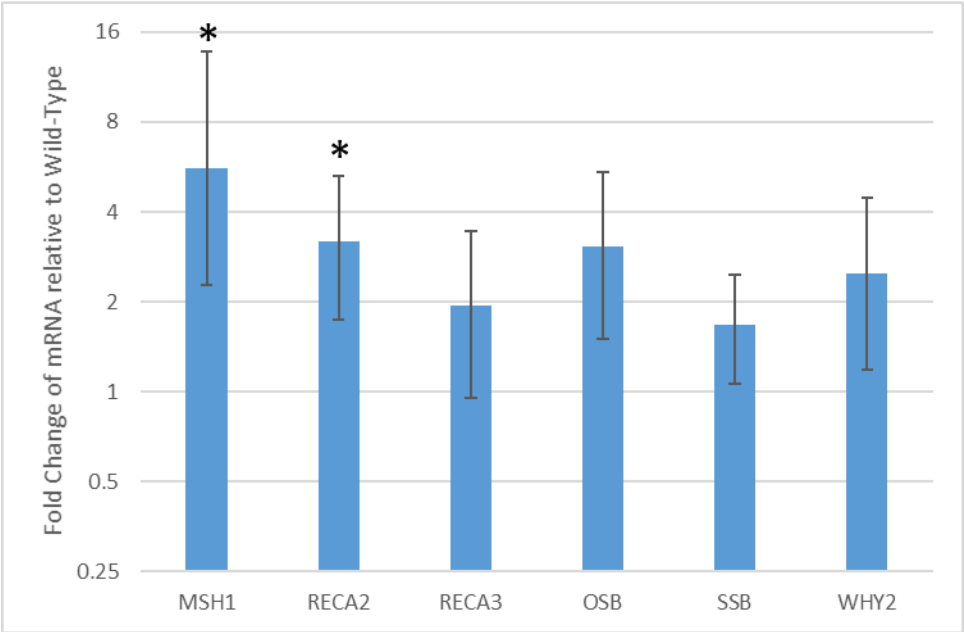

Figure 4.3: Quantitative RT-PCR assays of enzymes involved in DSBR in *ung* lines relative to wild-type. Fold change in transcript level is shown on the Y-axis. Error bars are standard deviation of three biological replicates. *MSH1* and *RECA2* are significantly transcriptionally upregulated in *ung* lines relative to wild-type (5.60-fold increase and 3.19-fold increase, respectively. Unpaired, 2-tailed Student's t-test, * indicates $p<0.05$). *OSB1* is nearly significantly upregulated in *ung* lines relative to wild-type (3.07-fold increase. Unpaired T-test $p=0.053$).

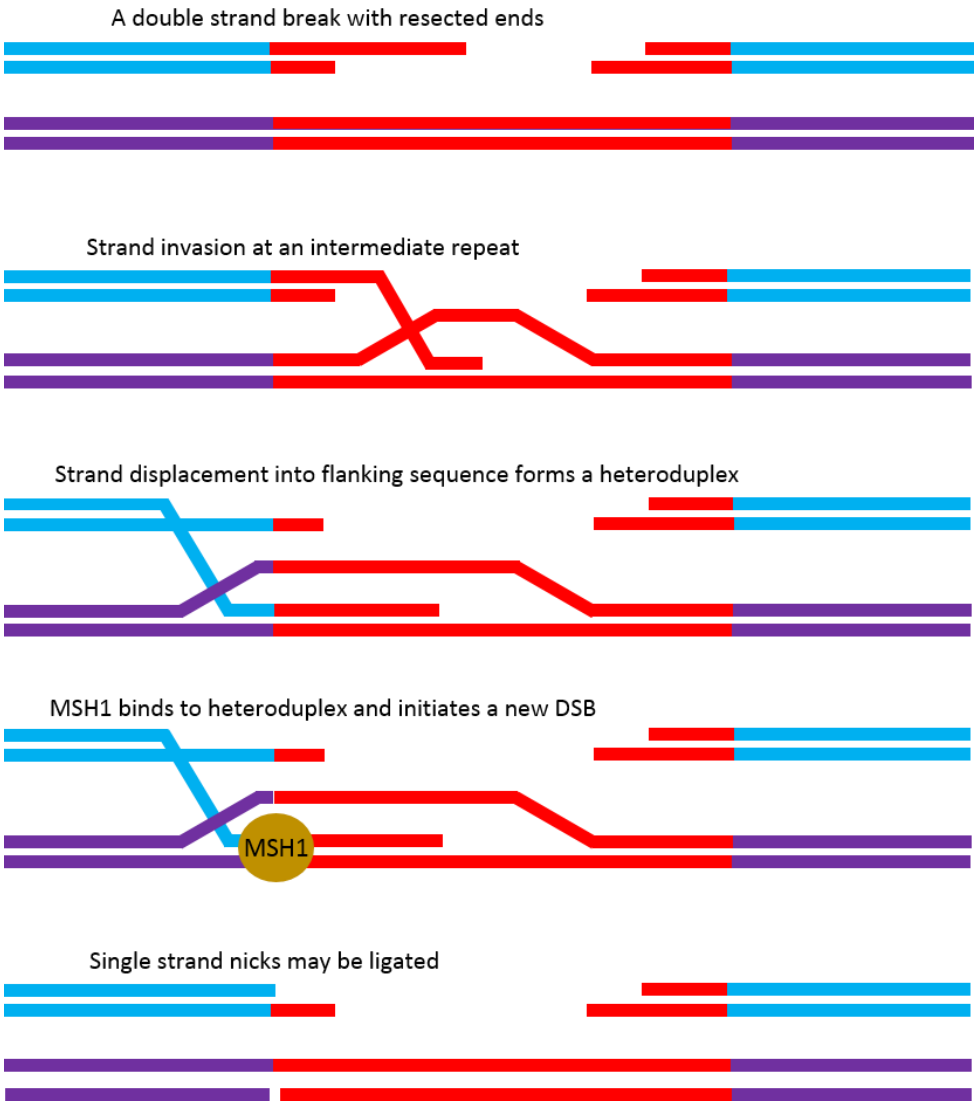

Figure 4.4: Model for the loss of intermediate repeats due to aborted base excision repair. Red strands represent short homologous sequences at an intermediate repeat, blue and purple strands represent flanking regions. As part of the homology surveillance system, *MSH1* binds to heteroduplexes that form at the margins of recombining intermediate repeats. Endonuclease activity at the heteroduplex creates a break between the annealing strands. The invaded strand (purple) can be ligated back together, while the invading strand (red) remains broken.

REFERENCES:

Abdelnoor, R.V., Christensen, A.C., Mohammed, S., Munoz-Castillo, B., Moriyama, H. and Mackenzie, S.A. (2006) Mitochondrial genome dynamics in plants and animals: Convergent gene fusions of a MutS homolog. J. Molec. Evol., 63, 165-173.

Abdelnoor, R.V., Yule, R., Elo, A., Christensen, A.C., Meyer-Gauen, G. and Mackenzie, S.A. (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.

Allen, G.C., Flores-Vergara, M.A., Krasynanski, S., Kumar, S. and Thompson, W.F. (2006) A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nature protocols, 1, 2320-2325.

Arrieta-Montiel, M.P., Shedge, V., Davila, J., Christensen, A.C. and Mackenzie, S.A. (2009) Diversity of the Arabidopsis Mitochondrial Genome Occurs via Nuclear-Controlled Recombination Activity. Genetics, 183, 1261-1268.

Baym, M., Kryazhimskiy, S., Lieberman, T.D., Chung, H., Desai, M.M. and Kishony, R. (2015) Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One, 10, e0128036.

Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E. and Huala, E. (2015) The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis (New York, N.Y. : 2000), 53, 474-485.

Boesch, P., Ibrahim, N., Paulus, F., Cosset, A., Tarasenko, V. and Dietrich, A. (2009) Plant mitochondria possess a short-patch base excision DNA repair pathway. Nucleic Acids Res, 37, 5690-5700.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J. and Wittwer, C.T. (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clinical chemistry, 55, 611-622.

Caruccio, N. (2011) Preparation of next-generation sequencing libraries using Nextera technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition. Methods Mol Biol, 733, 241-255.

Christensen, A.C. (2013) Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. Genome biology and evolution, 5, 1079-1086.

Christensen, A.C. (2014) Genes and Junk in Plant Mitochondria—Repair Mechanisms and Selection. Genome biology and evolution, 6, 1448-1453.

Christensen, A.C. (2018) Mitochondrial DNA Repair and Genome Evolution. In Annual Plant Reviews, 2nd Edition, Plant Mitochondria (Logan, D.C. ed. New York, NY, USA: Wiley-Blackwell, pp. 11-31.

Clough SJ, Bent AF. 1998. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. The Plant Journal. 16(6): 735-743

Cordoba-Cañero, D., Dubois, E., Ariza, R.R., Doutriaux, M.P. and Roldan-Arjona, T. (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.

Davila, J.I., Arrieta-Montiel, M.P., Wamboldt, Y., Cao, J., Hagmann, J., Shedge, V., Xu, Y.Z., Weigel, D. and Mackenzie, S.A. (2011) Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. BMC biology, 9, 64.

Drake JW, Baltz RH. 1976. The Biochemistry of Mutagenesis. Annu Rev Biochem. 45:11-37

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

Edmondson, A.C., Song, D., Alvarez, L.A., Wall, M.K., Almond, D., McClellan, D.A., Maxwell, A. and Nielsen, B.L. (2005) Characterization of a mitochondrially targeted single-stranded DNA-binding protein in Arabidopsis thaliana. Mol. Genet. Genomics, 273, 115-122.

Eggler, A.L., Lusetti, S.L. and Cox, M.M. (2003) The C terminus of the Escherichia coli RecA protein modulates the DNA binding competition with single-stranded DNA-binding protein. J Biol Chem, 278, 16389-16396.

Gualberto, J.M., Mileshina, D., Wallet, C., Niazi, A.K., Weber-Lotfi, F. and Dietrich, A. (2014) The plant mitochondrial genome: dynamics and maintenance. Biochimie, 100, 107-120.

Gualberto, J.M. and Newton, K.J. (2017) Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation. Annu Rev Plant Biol.

Halligan, D.L. and Keightley, P.D. (2009) Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. Annual Review of Ecology, Evolution, and Systematics, 40, 151-172.

Karimi, M., Inzé, D. and Depicker, A. 2002. GATEWAY vectors for Agrobacterium-mediated plant transformation. Trends in Plant Science 7(5): 193-195

Kauppila, J.H.K., Bonekamp, N.A., Mourier, A., Isokallio, M.A., Just, A., Kauppila, T.E.S., Stewart, J.B. and Larsson, N.G. (2018) Base-excision repair deficiency alone or combined with increased oxidative stress does not increase mtDNA point mutations in mice. Nucleic Acids Res, 46, 6642-6669.

Khazi, F.R., Edmondson, A.C. and Nielsen, B.L. (2003) An Arabidopsis homologue of bacterial RecA that complements an E. coli recA deletion is targeted to plant mitochondria. Mol Gen Genet, 269, 454-463.

Klein, M., Eckert-Ossenkopp, U., Schmiedeberg, I., Brandt, P., Unseld, M., Brennicke, A. and Schuster, W. (1994) Physical mapping of the mitochondrial genome of Arabidopsis thaliana by cosmid and YAC clones. Plant J., 6, 447-455.

Kleinstiver, B.P., Wolfs, J.M. and Edgell, D.R. (2013) The monomeric GIY-YIG homing endonuclease I-BmoI uses a molecular anchor and a flexible tether to sequentially nick DNA. Nucleic Acids Res, 41, 5413-5427.

Krokan, H.E., Standal, R. and Slupphaug, G. (1997) DNA glycosylases in the base excision repair of DNA. The Biochemical journal, 325 (Pt 1), 1-16.

Kumar RA, Oldenburg DJ, Bendich AJ. 2014. Changes in DNA damage, molecular integrity, and copy number for plastid DNA and mitochondrial DNA during maize development. J Exp Bot. 65(22): 6425-6439.

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J.C. and Dry, J.R. (2016) VarDict: a novel and

versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res, 44, e108.

Lewis CA Jr, Crayle J, Zhou S, Swanstrom R, Wolfenden R. 2016. Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. Proc Natl Acad Sci USA. 19;113(29):8194-8199

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 27(21): 2987-2993

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078-2079.

McGrew, D. and Knight, K. (2003) Molecular design and functional organization of the RecA protein. Critical Reviews in Biochemistry and Molecular Biology, 38, 385-432.

Miller-Messmer, M., Kuhn, K., Bichara, M., Le Ret, M., Imbault, P. and Gualberto, J.M. (2012) RecA-dependent DNA repair results in increased heteroplasmy of the Arabidopsis mitochondrial genome. Plant Physiol, 159, 211-226.

Odahara, M., Inouye, T., Fujita, T., Hasebe, M. and Sekine, Y. (2007) Involvement of mitochondrial-targeted RecA in the repair of mitochondrial DNA in the moss, Physcomitrella patens. Genes & genetic systems, 82, 43-51.

Onate-Sanchez, L. and Vicente-Carbajosa, J. (2008) DNA-free RNA isolation protocols for Arabidopsis thaliana, including seeds and siliques. BMC research notes, 1, 93.

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

Palmer, J.D. and Shields, C.R. (1984) Tripartite structure of the Brassica campestris mitochondrial genome. Nature, 307, 437.

Richardson, A.O., Rice, D.W., Young, G.J., Alverson, A.J. and Palmer, J.D. (2013) The "fossilized" mitochondrial genome of Liriodendron tulipifera: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. BMC biology, 11, 29.

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

Rowan, B.A., Patel, V., Weigel, D. and Schneeberger, K. (2015) Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. G3 (Bethesda, Md.), 5, 385-398.

Rowan, B.A., Seymour, D.K., Chae, E., Lundberg, D.S. and Weigel, D. (2017) Methods for Genotyping-by-Sequencing. In Genotyping: Methods and Protocols (White, S.J. and Cantsilieris, S. eds). New York, NY: Springer New York, pp. 221-242.

Seguí-Simarro JM, Coronado MJ, Staehelin LA. 2008. The Mitochondrial Cycle of Arabidopsis Shoot Apical Meristem and Leaf Primordium Meristematic Cells Is Defined by a Perinuclear Tentaculate/Cage-Like Mitochondrion. Plant Physiology. 148: 1380-1393

Seguí-Simarro JM, Staehelin LA. 2009. Mitochondrial reticulation in shoot apical meristem cells of Arabidopsis provides a mechanism for homogenization of mtDNA prior to gamete formation. Plant Signaling & Behavior. 4(3): 168-171

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

Sloan, D.B. and Taylor, D.R. (2010) Testing for selection on synonymous sites in plant mitochondrial DNA: the role of codon bias and RNA editing. J Mol Evol, 70, 479-491.

Sloan, D.B., Wu, Z. and Sharbrough, J. (2018) Correction of Persistent Errors in Arabidopsis Reference Mitochondrial Genomes. The Plant Cell, 30, 525-527.

Trasviña-Arenas, C.H., Baruch-Torres, N., Cordoba-Andrade, F.J., Ayala-Garcia, V.M., Garcia-Medel, P.L., Diaz-Quezada, C., Peralta-Castro, A., Ordaz-Ortiz, J.J. and Brieba, L.G. (2018) Identification of a unique insertion in plant organellar DNA polymerases responsible for 5'-dRP lyase and strand-displacement activities: Implications for Base Excision Repair. DNA Repair (Amst), 65, 1-10.

Unseld, M., Marienfeld, J.R., Brandt, P. and Brennicke, A. (1997) The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides. Nat. Genet., 15, 57-61.

Wallet, C., Le Ret, M., Bergdoll, M., Bichara, M., Dietrich, A. and Gualberto, J.M. (2015) The RECG1 DNA Translocase Is a Key Factor in Recombination Surveillance, Repair, and Segregation of the Mitochondrial DNA in Arabidopsis. Plant Cell, 27, 2907-2925.

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

Wynn, E.L. and Christensen, A.C. (2015) Are Synonymous Substitutions in Flowering Plant Mitochondria Neutral? J. Molec. Evol, 81, 131-135.

Xu, L. and Marians, K. (2002) A dynamic RecA filament permits DNA polymerase-catalyzed extension of the invading strand in recombination intermediates. J. Biol. Chem., 277, 14321-14328.

Zaegel V, Guermann B, Le Ret M, Andrés C, Meyer D, Erhardt M, Canaday J, Gualberto JM, Imbault P. 2006. The Plant-Specific ssDNA Binding Protein OSB1 Is Involved in the Stoichiometric Transmission of Mitochondrial DNA in Arabidopsis. The Plant Cell. 18: 3548-3563

APPENDICES

Appendix A1: Rousfinder.py

```
#! /usr/bin/env python
import sys, math, os, argparse, csv
csv.field_size_limit(sys.maxsize)

# January 16, 2018 version 1.1
# Find dispersed repeated sequences in genomes.
# Designed for plant mitochondrial genomes of up to a few Mbp.
# May be very slow with larger genomes.
# Blast can also sometimes give odd results with large or highly repetetive genomes.

# Gaps, or runs of 'N's in the sequence will definitely give weird results.
# The program assumes there aren't any, and that the longest repeat will be the full
 sequence to itself.
# If there are long repeats in the output that are listed as being only at one locat
ion, this is probably what happened.
# If there are a lot of repeats within repeats the results can also be odd.
# Copyright Alan C. Christensen, University of Nebraska, 2018
# No guarantees, warranties, support, or anything else is implicit or explicit.
# Input is a fasta format file of a sequence. Genbank format works but generates lot
s of error messages in stdout.
# Output is a list of unique, ungapped repeated sequences, fasta formatted.
# The names are in the format '>Repeat/ROUS_name_start_end_length'.
# Percent identity is limited to >=99%, to allow for sequencing errors of <1%.
# A table of repeats with the coordinates of each one is generated.
# A list of repeat name, length and copy number is generated.
# A binned table of the total number of repeats in size ranges is generated.
#
# PARAMETERS
#   REQUIRED:
#      input file in fasta format
#   Optional
#      -o output file name
#      -m minimum length of exact matches to keep
#      -b path to blastn (default is /usr/bin/)
#      -k keep temp files
#      -gb to write the repeats to a genbank format file

parser = argparse.ArgumentParser(description='Find repeats in a fasta sequence file'
)
parser.add_argument('infile', action='store', help='Input .fasta file')
parser.add_argument('-
o', action='store', dest='outfile', help='Output file name seed, default is input_re
peats', default='default')
parser.add_argument('-
m', action='store', dest='minlen', help='Minimum length of matches to keep, default=
24', default='24')
parser.add_argument('-
b', action='store', dest='blast_path', help='Path to blastn program, default is /usr
/bin/', default='/usr/bin/')
parser.add_argument('-
k', action='store_true', dest='keep', help='True to keep temp files', default=False)
```

```
parser.add_argument('-
gb', action='store_true', dest='genbank', help='True to write GenBank format file',
default=False)
results = parser.parse_args()
infile = results.infile
outfile = results.outfile
minlen = int(results.minlen)
blast_path = results.blast_path
keep = results.keep
genbank = results.genbank

# It might be useful to define the wordsize as something less than minlen, so both v
ariables are used.
# Wordsize smaller than minlen would give smaller core identical sequences in the mi
ddle of repeats.
# An example might be to change this to wordsize = str(int(minlen/2)).
wordsize = str(minlen)

# If no output file seed is specified, make one by stripping leading directory infor
mation
# and stripping trailing .fa or .fasta from the input file name and using that.
if outfile == 'default':
    outfile = infile
    if outfile.count('/') > 0:
        for i in range(outfile.count('/')):
            index = outfile.index('/')
            outfile = outfile[index+1:]
    if outfile.endswith('.fa') or outfile.endswith('.fasta'):
        outfile = outfile.rstrip('fasta')
    outfile = outfile.rstrip('.')
outfa = outfile+'_rep.fasta'
outtab = outfile+'_rep_table.txt'
outbin = outfile+'_binned.txt'
outcount = outfile+'_rep_counts.txt'
outgb = outfile+'_repeats.gb.txt'
tempblast = outfile+'_tempblast.txt'
temprepeats = outfile+'_temprepeats.txt'
tempparse = outfile+'_sequence_parsing.txt'

# Get sequence name and length from fasta file.
seq = open(infile, 'r')
seqname = seq.readline()
seqname = seqname.lstrip('> ')
seqname = seqname.rstrip()
seqlen = 0
for line in seq:
    if(line[0] == ">"):
        continue
    seqlen += len(line.strip())
seq.close()

# run blastn with query file plus strand (removing first line which is full length s
equence), minus strand, and concatenate
print 'Performing self-blastn comparison with '+seqname
os.system(blast_path+'blastn -query '+infile+' -strand plus -subject '+infile+' -
word_size '+wordsize+' -reward 1 -penalty -20 -ungapped -dust no -
soft_masking false -evalue 1000 -
outfmt "10 qstart qend length sstart send mismatch sstrand qseq" | tail -
n+2 > tempblast1.txt')
os.system(blast_path+'blastn -query '+infile+' -strand minus -subject '+infile+' -
word_size '+wordsize+' -reward 1 -penalty -20 -ungapped -dust no -
```

```
soft_masking false -evalue 1000 -
outfmt "10 qstart qend length sstart send mismatch sstrand qseq" > tempblast2.txt')

os.system('cat tempblast1.txt tempblast2.txt > '+tempblast)
os.system('rm tempblast1.txt tempblast2.txt')

# open tempblast.txt, convert to list of lists, and sort by length and position desc
ending
# This is necessary because blastn does not output every possible pair of hits when
there are more than 2 copies of a repeat

print 'Sorting alignments...'
f = open(tempblast, 'r')
reader = csv.reader(f)
alignments = list(reader)
f.close()
alignments = sorted(alignments, key=lambda x: (-1*int(x[2]), -1*int(x[0])))
alignments.append(['1','1','1','1','1','0','A','X'])

# New list of uniques
# Text file '_sequence_parsing.txt' includes the information on how duplicates were
found.
# Start at row 0. Compare to subsequent rows.
# If repeat length is different from the next row, it has passed all the tests, writ
e it to the file.
# If query or subject coordinates are the same as the query or subject or reversed c
oordinates
# of a subsequent row, it is not unique, so go to the next row and do the comparison
s again.
# Thanks to Alex Kozik for repeatedly testing and finding bugs in the algorithm.
print 'Finding unique repeats...'
uniques = []
sp = open(tempparse, 'w')
for row in range(len(alignments)):
    sp.write('row '+str(row)+'\n')

    if int(alignments[row][2]) < minlen:
        # This won't happen unless the word_size is defined as something other than
minlen.
        # That could be useful under some circumstances.
        sp.write('row '+str(row)+' is less than minlength')
        break
    else:

        for compare in range(row+1,len(alignments)):
            if alignments[row][2] != alignments[compare][2]:
                uniques.append(alignments[row])
                sp.write('\tadding row '+str(row)+' to unique list\n')
                break
            else:
                sp.write('\tcomparing to '+str(compare)+'\n')

                if alignments[row][0] == alignments[compare][0] and alignments[row][
1] == alignments[compare][1]:
                    sp.write('\tqstart and qend of row '+str(row)+' and '+str(compar
e)+' are the same\n')
                    break
                elif alignments[row][0] == alignments[compare][1] and alignments[row
][1] == alignments[compare][0]:
                    sp.write('\tqstart and qend of row '+str(row)+' is the same as q
end and qstart of '+str(compare)+'\n')
```

```
                    break
                elif alignments[row][0] == alignments[compare][3] and alignments[row
][1] == alignments[compare][4]:
                    sp.write('\tqstart and qend of row '+str(row)+' is the same as s
start and send of '+str(compare)+'\n')
                    break
                elif alignments[row][0] == alignments[compare][4] and alignments[row
][1] == alignments[compare][3]:
                    sp.write('\tqstart and qend of row '+str(row)+' is the same as s
end and sstart of '+str(compare)+'\n')
                    break
                elif alignments[row][3] == alignments[compare][0] and alignments[row
][4] == alignments[compare][1]:
                    sp.write('\tsstart and send of row '+str(row)+' is the same as q
start and qend of '+str(compare)+'\n')
                    break
                elif alignments[row][3] == alignments[compare][1] and alignments[row
][4] == alignments[compare][0]:
                    sp.write('\tsstart and send of row '+str(row)+' is the same as q
end and qstart of '+str(compare)+'\n')
                    break
                elif alignments[row][3] == alignments[compare][3] and alignments[row
][4] == alignments[compare][4]:
                    sp.write('\tsstart and send of row '+str(row)+' is the same as s
start and send of '+str(compare)+'\n')
                    break
                elif alignments[row][3] == alignments[compare][4] and alignments[row
][4] == alignments[compare][3]:
                    sp.write('\tsstart and send of row '+str(row)+' is the same as s
end and sstart of '+str(compare)+'\n')
                    break
                else:
                    sp.write('\t'+str(row)+' is different\n')

sp.close()

# Write uniques into output file
# Start list for copy number table
rous_count = 0
g = open(outfa, 'w')
repcopies = []

for i in range(len(uniques)):
    qstart = uniques[i][0]
    qend = uniques[i][1]
    length = uniques[i][2]
    seq = uniques[i][7]

    rous_count += 1
    g.write('>Repeat_'+str(rous_count)+'\n'+seq+'\n')
    repcopies.append(['Repeat_'+str(rous_count),length])

if rous_count == 0:
    print "\tRepeats of unusual size? I don't think they exist"
g.close()
print 'Repeat fasta file is done, as you wish.'

# Now find each copy of each repeat. Again, this is because the blastn output file d
oes not have every possible alignment.
# It is also because the information on locations and strand is not organized well i
n the blastn output.
```

```
# In addition, this subroutine eliminates duplicates of nested repeats.

print "Finding all copies of repeats..."
g = open(outfa, 'r')
os.system(blast_path+'blastn -query '+outfa+' -strand both -subject '+infile+' -
word_size '+wordsize+' -reward 1 -penalty -20 -ungapped -dust no -
soft_masking false -evalue 1000 -
outfmt "10 qseqid length sstart send sstrand qcovhsp" > '+temprepeats)
g.close()

tempr = open(temprepeats, 'r')
reader = csv.reader(tempr)
replist = list(reader)
tempr.close()

print "Making a table of the repeats..."
sum_rep_len = 0
bin_dict = {}
binned = [seqname,seqlen,0]

# defining the bins
i = 0
j = 50
while j < 1000:
    bin_dict[i] = j
    binned.append(0)
    i += 1
    j += 50
while j <= 10000:
    bin_dict[i] = j
    binned.append(0)
    i +=1
    j += 250

# make list for entire sequence, set each position as 0
posit = []
for n in range(seqlen):
    posit.append(0)

# Thanks to Emily Wynn for suggesting qcovhsp for this loop.
# if qcovhsp is >98%, write to the file
# write tab separated values of repeat name, length, start, end, strand to outtab
# make list for genbank file
# Keep stats on lengths
rt = open(outtab, 'w')
rt.write(seqname+'\t'+str(seqlen)+'\n')
templist = []
gblist =[]

# look at each repeat in turn
for i in range(len(replist)):
    # if repeat is good (>98% identical to another one), write it to the file, and p
ut the name in a list
    if int(replist[i][5])>98:
        rt.write(str(replist[i][0])+'\t'+str(replist[i][1])+'\t'+str(replist[i][2])+
'\t'+str(replist[i][3])+'\t'+str(replist[i][4])+'\n')
        if replist[i][4] == 'minus':
            location = 'complement('+replist[i][3]+'..'+replist[i][2]+')'
        else:
            location = replist[i][2]+'..'+replist[i][3]
```

```
        gblist.append('     repeat_region   '+location+'\n                     /rpt_
type=dispersed\n                     /label='+replist[i][0]+'\n')
        templist.append(replist[i][0])
        # then write 1's at every position in the sequence covered by that repeat
        # these can then be summed to get total bases of repeats
        # bases in overlapping repeats are only counted once
        for n in range(int(replist[i][2]), int(replist[i][3])):
            posit[n] = 1
        # then scan through bin sizes and if a repeat is greater than the
        # bin_dict size cutoff, add one to the bin
        for j in range(len(binned)-4, -1, -1):
            if int(replist[i][1]) >= bin_dict[j]:
                binned[j+3] +=1
                break
sum_rep_len = posit.count(1)
binned[2] = sum_rep_len
rt.close()
if genbank == True:
    gb = open(outgb, 'w')
    for i in range(len(gblist)):
        gb.write(gblist[i])
    gb.close()

# write tab separated values of repeat name, length, copy number to outcount
# first two lines are also a table of stats on repeats
rc = open(outcount,'w')
rc.write('Sequence\tGenome_size\tNumRepeats\tAvgSize\tAvgCopyNum\n')

numrous = 0
sizerous = 0
copyrous = 0

for i in range(len(repcopies)):
    repname = repcopies[i][0]
    replen = float(repcopies[i][1])
    repcop = float(templist.count(repname))

    numrous += 1
    sizerous += replen
    copyrous += repcop

if numrous == 0:
    avsizerous = 'NA'
    avcopyrous = 'NA'
else:
    avsizerous = sizerous/numrous
    avcopyrous = copyrous/numrous


rc.write(seqname+'\t'+str(seqlen)+'\t'+str(numrous)+'\t'+str(avsizerous)+'\t'+str(av
copyrous)+'\n')

for i in range(len(repcopies)):
    rc.write(repcopies[i][0]+'\t'+repcopies[i][1]+'\t'+str(templist.count(repcopies[
i][0]))+'\n')

rc.close()

# Write binned table headers, then stats for this sequence.
binfile = open(outbin, 'w')
binfile.write('Sequence\tSeq_len\tRep_len\t')
```

```
for i in range(len(bin_dict)):
    binfile.write(str(bin_dict[i])+'\t')
binfile.write('\n')
for i in range(len(binned)):
    binfile.write(str(binned[i])+'\t')
binfile.write('\n')
binfile.close()
print "Repeat tables are done, as you wish."

# Removing temp files if necessary
if keep == False:
    os.system('rm '+tempblast+' '+temprepeats+' '+tempparse)

# Rachael Schulte, William Goldman and Rob Reiner inspired this section of code
quote_dict = {0:"48656c6c6f2e204d79206e616d6520697320496e69676f204d6f6e746f79612e205
96f75206b696c6c6564206d7920666174686572e20507265706172652074 6f206469652e", 1:"57686
56e20492077617320796f7572206167652c2074656c65766973696f6e207761732063616c6c656420626
f6f6b732e", 2:"486176652066756e2073746f726d696e27206461206361737 46c6521", 3:"4d79207
761792773206e6f74207665727920737 06f7274736d616e6c696b652e", 4:"596f75206b65657020757
3696e6720746861742077 6f72642e204920646f206e6f74207468696e6b206974206d65616e732077686
17420796f75207468696e6b206974206d65616e732e", 5:"4d7572646572656420627920706972617465
57320697320676f6f642e",6:"496e636f6e6365697661626c6521", 7:"5468657265277320612062696
72064696666657265 6e6365206265747765656e206d6f73746c79206465616420616e6420616c6c2064
6561642e", 8:"596f752072757368206120 6d697261636c65206d616e2c20796f752067657420726f74
74656e206d697261636c65732e", 9:"476f6f64206e696768742c20576573746c65792e20476f6f6420
776f726b2e20536c6565702077656c6c2e2049276c6c206d6f7374206c696b656c79206b696c6c20796f
7520696e20746865206d6f726e696e672e",10:"4e6f206d6f7265207268796d65732c2049206d65616e
2069742120416e79626f64792077616e742061207065616e75743f"}
import random, binascii
z = random.randint(0,10)
print binascii.unhexlify(quote_dict[z])+'\n'
```

Appendix A2: MultipleRepeats.py

```
#! /usr/bin/env python
import sys, math, os, argparse

# Usage: -din directory of files to find repeats in
#        -word word_size

parser = argparse.ArgumentParser(description='Find repeats in a directory of fasta s
equence files')
parser.add_argument('-
din', action='store', dest='din', help='Input .fasta directory')
parser.add_argument('-
word', action='store', dest='word', help='Word size for blast')
results = parser.parse_args()
din = results.din
word = results.word

li = os.listdir(din)
inputs = filter(lambda x: '.fasta' in x, li)
inputs.sort()

for i in range(len(inputs)):
    infile = str(inputs[i])
    os.system("/home/alan/applications/ROUSFinder.py -m "+word+" "+din+infile)
```

Appendix A3: Rousfinder2.py

```
#! /usr/bin/env python
import sys, math, os, argparse, csv
csv.field_size_limit(sys.maxsize)

# Version 2.0, November 21, 2018
# Changes: uses variable parameters
# Find dispersed repeated sequences in genomes.
# Designed for plant mitochondrial genomes of up to a few Mbp.
# May be very slow with larger genomes.
# Blast can also sometimes give odd results with large or highly repetetive genomes.

# Gaps, or runs of 'N's in the sequence will definitely give weird results.
# The program assumes there aren't any, and that the longest repeat will be the full
 sequence to itself.
# If there are long repeats in the output that are listed as being only at one locat
ion, this is probably what happened.
# If there are a lot of repeats within repeats the results can also be odd.
# Copyright Alan C. Christensen, University of Nebraska, 2018
# No guarantees, warranties, support, or anything else is implicit or explicit.
# Input is a fasta format file of a sequence. Genbank format works but generates lot
s of error messages in stdout.
# Output is a list of unique, ungapped repeated sequences, fasta formatted.
# The names are in the format '>Repeat/ROUS_name_start_end_length'.
# A table of repeats with the coordinates of each one is generated.
# A list of repeat name, length and copy number is generated.
# A binned table of the total number of repeats in size ranges is generated.
#
# PARAMETERS
#   REQUIRED:
#      input file in fasta format
#   Optional
#      -o output file name
#      -m minimum length of exact matches to keep
#      -b path to blastn (default is /usr/bin/)
#      -k keep temp files
#      -gb to write the repeats to a genbank format file
#      -rew reward for match (default is 1)
#      -pen penalty for mismatch (default is 20)

parser = argparse.ArgumentParser(description='Find repeats in a fasta sequence file'
)
parser.add_argument('infile', action='store', help='Input .fasta file')
parser.add_argument('-
o', action='store', dest='outfile', help='Output file name seed, default is input_re
peats', default='default')
parser.add_argument('-
m', action='store', dest='minlen', help='Minimum length of matches to keep, default=
50', default='50')
parser.add_argument('-
b', action='store', dest='blast_path', help='Path to blastn program, default is /usr
/bin/', default='/usr/bin/')
parser.add_argument('-
k', action='store_true', dest='keep', help='True to keep temp files', default=False)

parser.add_argument('-
gb', action='store_true', dest='genbank', help='True to write GenBank format file',
default=False)
parser.add_argument('-
rew', action='store', dest='reward', help='Reward for match', default='1')
```

```
parser.add_argument('-
pen', action='store', dest='penalty', help='Penalty for mismatch', default='20')
results = parser.parse_args()
infile = results.infile
outfile = results.outfile
minlen = int(results.minlen)
blast_path = results.blast_path
keep = results.keep
genbank = results.genbank
reward = results.reward
penalty = results.penalty

# It might be useful to define the wordsize as something less than minlen, so both v
ariables are used.
# Wordsize smaller than minlen would give smaller core identical sequences in the mi
ddle of repeats.
# An example might be to change this to wordsize = str(int(minlen/2)).
wordsize = str(minlen)

# If no output file seed is specified, make one by stripping leading directory infor
mation
# and stripping trailing .fa or .fasta from the input file name and using that.
if outfile == 'default':
    outfile = infile
    if outfile.count('/') > 0:
        for i in range(outfile.count('/')):
            index = outfile.index('/')
            outfile = outfile[index+1:]
    if outfile.endswith('.fa') or outfile.endswith('.fasta'):
        outfile = outfile.rstrip('fasta')
    outfile = outfile.rstrip('.')
outfa = outfile+'_rep.fasta'
outtab = outfile+'_rep_table.txt'
outbin = outfile+'_binned.txt'
outcount = outfile+'_rep_counts.txt'
outgb = outfile+'_repeats.gb.txt'
tempblast = outfile+'_tempblast.txt'
temprepeats = outfile+'_temprepeats.txt'
tempparse = outfile+'_sequence_parsing.txt'

# Get sequence name and length from fasta file.
seq = open(infile, 'r')
seqname = seq.readline()
seqname = seqname.lstrip('> ')
seqname = seqname.rstrip()
seqlen = 0
for line in seq:
    if(line[0] == ">"):
        continue
    seqlen += len(line.strip())
seq.close()

# run blastn with query file plus strand (removing first line which is full length s
equence), minus strand, and concatenate
print 'Performing self-blastn comparison with '+seqname
os.system(blast_path+'blastn -query '+infile+' -strand plus -subject '+infile+' -
word_size '+wordsize+' -reward '+reward+' -penalty -'+penalty+' -ungapped -dust no -
soft_masking false -evalue 10  -
outfmt "10 qstart qend length sstart send mismatch sstrand qseq" | tail -
n+2 > tempblast1.txt')
```

```
os.system(blast_path+'blastn -query '+infile+' -strand minus -subject '+infile+' -
word_size '+wordsize+' -reward '+reward+' -penalty -'+penalty+' -ungapped -dust no -
soft_masking false -evalue 10 -
outfmt "10 qstart qend length sstart send mismatch sstrand qseq" > tempblast2.txt')

os.system('cat tempblast1.txt tempblast2.txt > '+tempblast)
os.system('rm tempblast1.txt tempblast2.txt')

# open tempblast.txt, convert to list of lists, and sort by length and position desc
ending
# This is necessary because blastn does not output every possible pair of hits when
there are more than 2 copies of a repeat

print 'Sorting alignments...'
f = open(tempblast, 'r')
reader = csv.reader(f)
alignments = list(reader)
f.close()
alignments = sorted(alignments, key=lambda x: (-1*int(x[2]), -1*int(x[0])))
alignments.append(['1','1','1','1','1','0','A','X'])

# New list of uniques
# Text file '_sequence_parsing.txt' includes the information on how duplicates were
found.
# Start at row 0. Compare to subsequent rows.
# If repeat length is different from the next row, it has passed all the tests, writ
e it to the file.
# If query or subject coordinates are the same as the query or subject or reversed c
oordinates
# of a subsequent row, it is not unique, so go to the next row and do the comparison
s again.
# Thanks to Alex Kozik for repeatedly testing and finding bugs in the algorithm.
print 'Finding unique repeats...'
uniques = []
sp = open(tempparse, 'w')
for row in range(len(alignments)):
    sp.write('row '+str(row)+'\n')

    if int(alignments[row][2]) < minlen:
        # This won't happen unless the word_size is defined as something other than
minlen.
        # That could be useful under some circumstances.
        sp.write('row '+str(row)+' is less than minlength')
        break
    else:

        for compare in range(row+1,len(alignments)):
            if alignments[row][2] != alignments[compare][2]:
                uniques.append(alignments[row])
                sp.write('\tadding row '+str(row)+' to unique list\n')
                break
            else:
                sp.write('\tcomparing to '+str(compare)+'\n')

                if alignments[row][0] == alignments[compare][0] and alignments[row][
1] == alignments[compare][1]:
                    sp.write('\tqstart and qend of row '+str(row)+' and '+str(compar
e)+' are the same\n')
                    break
                elif alignments[row][0] == alignments[compare][1] and alignments[row
][1] == alignments[compare][0]:
```

```
                    sp.write('\tqstart and qend of row '+str(row)+' is the same as q
end and qstart of '+str(compare)+'\n')
                    break
                elif alignments[row][0] == alignments[compare][3] and alignments[row
][1] == alignments[compare][4]:
                    sp.write('\tqstart and qend of row '+str(row)+' is the same as s
start and send of '+str(compare)+'\n')
                    break
                elif alignments[row][0] == alignments[compare][4] and alignments[row
][1] == alignments[compare][3]:
                    sp.write('\tqstart and qend of row '+str(row)+' is the same as s
end and sstart of '+str(compare)+'\n')
                    break
                elif alignments[row][3] == alignments[compare][0] and alignments[row
][4] == alignments[compare][1]:
                    sp.write('\tsstart and send of row '+str(row)+' is the same as q
start and qend of '+str(compare)+'\n')
                    break
                elif alignments[row][3] == alignments[compare][1] and alignments[row
][4] == alignments[compare][0]:
                    sp.write('\tsstart and send of row '+str(row)+' is the same as q
end and qstart of '+str(compare)+'\n')
                    break
                elif alignments[row][3] == alignments[compare][3] and alignments[row
][4] == alignments[compare][4]:
                    sp.write('\tsstart and send of row '+str(row)+' is the same as s
start and send of '+str(compare)+'\n')
                    break
                elif alignments[row][3] == alignments[compare][4] and alignments[row
][4] == alignments[compare][3]:
                    sp.write('\tsstart and send of row '+str(row)+' is the same as s
end and sstart of '+str(compare)+'\n')
                    break
                else:
                    sp.write('\t'+str(row)+' is different\n')

sp.close()

# Write uniques into output file
# Start list for copy number table
rous_count = 0
g = open(outfa, 'w')
repcopies = []

for i in range(len(uniques)):
    qstart = uniques[i][0]
    qend = uniques[i][1]
    length = uniques[i][2]
    seq = uniques[i][7]

    rous_count += 1
    g.write('>Repeat_'+str(rous_count)+'\n'+seq+'\n')
    repcopies.append(['Repeat_'+str(rous_count),length])

if rous_count == 0:
    print "\tRepeats of unusual size? I don't think they exist"
g.close()
print 'Repeat fasta file is done, as you wish.'

# Now find each copy of each repeat. Again, this is because the blastn output file d
oes not have every possible alignment.
```

```
# It is also because the information on locations and strand is not organized well i
n the blastn output.
# In addition, this subroutine eliminates duplicates of nested repeats.

print "Finding all copies of repeats..."
g = open(outfa, 'r')
os.system(blast_path+'blastn -query '+outfa+' -strand both -subject '+infile+' -
word_size '+wordsize+' -reward 1 -penalty -20 -ungapped -dust no -
soft_masking false -evalue 1000 -
outfmt "10 qseqid length sstart send sstrand qcovhsp" > '+temprepeats)
g.close()

tempr = open(temprepeats, 'r')
reader = csv.reader(tempr)
replist = list(reader)
tempr.close()

print "Making a table of the repeats..."
sum_rep_len = 0
bin_dict = {}
binned = [seqname,seqlen,0]

# defining the bins
i = 0
j = 50
while j < 1000:
    bin_dict[i] = j
    binned.append(0)
    i += 1
    j += 50
while j <= 10000:
    bin_dict[i] = j
    binned.append(0)
    i +=1
    j += 250

# make list for entire sequence, set each position as 0
posit = []
for n in range(seqlen):
    posit.append(0)

# Thanks to Emily Wynn for suggesting qcovhsp for this loop.
# if qcovhsp is >98%, write to the file
# write tab separated values of repeat name, length, start, end, strand to outtab
# make list for genbank file
# Keep stats on lengths
rt = open(outtab, 'w')
rt.write(seqname+'\t'+str(seqlen)+'\n')
templist = []
gblist =[]

# look at each repeat in turn
for i in range(len(replist)):
    # if repeat is good (>98% identical to another one), write it to the file, and p
ut the name in a list
    if int(replist[i][5])>98:
        rt.write(str(replist[i][0])+'\t'+str(replist[i][1])+'\t'+str(replist[i][2])+
'\t'+str(replist[i][3])+'\t'+str(replist[i][4])+'\n')
        if replist[i][4] == 'minus':
            location = 'complement('+replist[i][3]+'..'+replist[i][2]+')'
        else:
```

```
            location = replist[i][2]+'..'+replist[i][3]
        gblist.append('     repeat_region   '+location+'\n                     /rpt_
type=dispersed\n                     /label='+replist[i][0]+'\n')
        templist.append(replist[i][0])
        # then write 1's at every position in the sequence covered by that repeat
        # these can then be summed to get total bases of repeats
        # bases in overlapping repeats are only counted once
        for n in range(int(replist[i][2]), int(replist[i][3])):
            posit[n] = 1
        # then scan through bin sizes and if a repeat is greater than the
        # bin_dict size cutoff, add one to the bin
        for j in range(len(binned)-4, -1, -1):
            if int(replist[i][1]) >= bin_dict[j]:
                binned[j+3] +=1
                break
sum_rep_len = posit.count(1)
binned[2] = sum_rep_len
rt.close()
if genbank == True:
    gb = open(outgb, 'w')
    for i in range(len(gblist)):
        gb.write(gblist[i])
    gb.close()

# write tab separated values of repeat name, length, copy number to outcount
# first two lines are also a table of stats on repeats
rc = open(outcount,'w')
rc.write('Sequence\tGenome_size\tNumROUS\tAvgSize\tAvgCopyNum\n')

numrous = 0
sizerous = 0
copyrous = 0

for i in range(len(repcopies)):
    repname = repcopies[i][0]
    replen = float(repcopies[i][1])
    repcop = float(templist.count(repname))

    numrous += 1
    sizerous += replen
    copyrous += repcop

if numrous == 0:
    avsizerous = 'NA'
    avcopyrous = 'NA'
else:
    avsizerous = sizerous/numrous
    avcopyrous = copyrous/numrous


rc.write(seqname+'\t'+str(seqlen)+'\t'+str(numrous)+'\t'+str(avsizerous)+'\t'+str(av
copyrous)+'\n')

for i in range(len(repcopies)):
    rc.write(repcopies[i][0]+'\t'+repcopies[i][1]+'\t'+str(templist.count(repcopies[
i][0]))+'\n')

rc.close()

# Write binned table headers, then stats for this sequence.
binfile = open(outbin, 'w')
```

```
binfile.write('Sequence\tSeq_len\tRep_len\t')
for i in range(len(bin_dict)):
    binfile.write(str(bin_dict[i])+'\t')
binfile.write('\n')
for i in range(len(binned)):
    binfile.write(str(binned[i])+'\t')
binfile.write('\n')
binfile.close()
print "Repeat tables are done, as you wish."

# Removing temp files if necessary
if keep == False:
    os.system('rm '+tempblast+' '+temprepeats+' '+tempparse)

# Rachael Schulte, William Goldman and Rob Reiner inspired this section of code
quote_dict = {0:"48656c6c6f2e204d79206e616d6520697320496e69676f204d6f6e746f79612e205
96f75206b696c6c6564206d7920666174686572e20507265706172652074f206469652e", 1:"57686
56e2049207761732079f7572206167652c2074656c65766973696f6e207761732063616c6c656420626
f6f6b732e", 2:"486176652066756e2073746f726d696e272064612063617374 6c6521", 3:"4d79207
761792773206e6f742076657279207 3706f7274736d616e6c696b652e", 4:"596f75206b65657020757
3696e6720746861742077 6f72642e204920646f206e6f74207468696e6b206974206d65616e732077686
17420796f75207468696e6b206974206d65616e732e", 5:"4d7572646572656420627920706972617461
57320697320676f6f642e",6:"496e636f6e6365697661626c6521", 7:"5468657265277320612062696
72064696666657265e6365206265747765656e206d6f73746c79206465616420616e6420616c6c2064
6561642e", 8:"596f752072757368206120 6d697261636c65206d616e2c20796f752067657420726f74
74656e206d697261636c65732e", 9:"476f6f64206e696768742c20576573746c65792e20476f6f6420
776f726b2e20536c6565702077656c6c2e2049276c6c206d6f7374206c696b656c79206b696c6c20796f
7520696e20746865206d6f726e696e672e",10:"4e6f206d6f7265207268796d65732c2049206d65616e
2069742120416e79626f64792077616e742061207065616e75743f"}
import random, binascii
z = random.randint(0,10)
print binascii.unhexlify(quote_dict[z])+'\n'
```