Direct Estimate of the Spontaneous Mutation Rate Uncovers the Effects of Drift and Recombination in the *Chlamydomonas reinhardtii* Plastid Genome

Rob W. Ness,*1,2 Susanne A. Kraemer,1 Nick Colegrave,1 and Peter D. Keightley1

1Institute of Evolutionary Biology, Ashworth Laboratories, University of Edinburgh, Edinburgh, United Kingdom
2Department of Biology, University of Toronto Mississauga, Mississauga, ON, Canada

*Corresponding author:* E-mail: rob.ness@utoronto.ca.

**Abstract**

Plastids perform crucial cellular functions, including photosynthesis, across a wide variety of eukaryotes. Since endosymbiosis, plastids have maintained independent genomes that now display a wide diversity of gene content, genome structure, gene regulation mechanisms, and transmission modes. The evolution of plastid genomes depends on an input of de novo mutation, but our knowledge of mutation in the plastid is limited to indirect inference from patterns of DNA divergence between species. Here, we use a mutation accumulation experiment, where selection acting on mutations is rendered ineffective, combined with whole-plastid genome sequencing to directly characterize de novo mutation in *Chlamydomonas reinhardtii*. We show that the mutation rates of the plastid and nuclear genomes are similar, but that the base spectra of mutations differ significantly. We integrate our measure of the mutation rate with a population genomic data set of 20 individuals, and show that the plastid genome is subject to substantially stronger genetic drift than the nuclear genome. We also show that high levels of linkage disequilibrium in the plastid genome are not due to restricted recombination, but are instead a consequence of increased genetic drift. One likely explanation for increased drift in the plastid genome is that there are stronger effects of genetic hitchhiking. The presence of recombination in the plastid is consistent with laboratory studies in *C. reinhardtii* and demonstrates that although the plastid genome is thought to be uniparentally inherited, it recombines in nature at a rate similar to the nuclear genome.

**Key words:** plastid genome, mutation rate, effective population size, recombination, *Chlamydomonas reinhardtii*, chloroplast.

**Introduction**

Plastids, including the chloroplast, arose via the primary endosymbiosis of a cyanobacterium and subsequently spread laterally into other lineages through eukaryote-eukaryote endosymbioses. The genome of the original endosymbiont has been retained since the ancestor of glaucophytes, red algae, and green plants (but see Molina et al. 2014; Smith and Lee 2014). The near ubiquity, sequence conservation and high copy number of plastid DNA has made it an invaluable resource for DNA barcoding and biogeographical/phylogenetic investigations. Most of our knowledge of plastid biology and genomics is derived from a disproportionate sampling of land plant chloroplasts, but recent sampling of a broader array of taxa has revealed an extraordinary diversity of genomic architectures and modes of inheritance (reviewed in Smith and Keeling 2015). The structure and composition of plastid genomes often differ significantly from their nuclear and mitochondrial counterparts. Despite the widespread use of plastid genomes in evolutionary genetics, we know little about the forces that explain diversity and influence their molecular evolution.

Variation in the type and rate of mutation among the nuclear, mitochondrial, and plastid genomes has been posited as the source of differences in their genome structure and evolution (Allen and Raven 1996; Race et al. 1999; Lynch et al. 2006). However, despite a growing body of information on spontaneous mutation in the nucleus (e.g., Kong et al. 2012; Zhu et al. 2014) and mitochondria (e.g., Haag-Liautard et al. 2008; Lynch et al. 2008; Howe et al. 2010), we know relatively little about spontaneous mutation in plastids. Most, if not all, DNA repair and replication genes used in the plastid are nuclear-encoded, some of which are dually targeted (e.g., *PHR1*; Petersen and Small 2001), which suggests that patterns of mutation might be similar. However, biochemical processes in the plastid and mitochondria create mutagenic reactive oxygen species and may lead to a fundamentally different mutational environment for organelar genomes. Mutagenesis by oxidative stress has been implicated in cellular aging (Van Breusegem and Dat 2006), the transfer of plastid genes to the nucleus (Allen and Raven 1996), and the retention of particular proteins in the organelle genome to maintain redox balance (Race et al. 1999). Moreover, the plastid and mitochondrial genomes replicate multiple times per cell division and this could increase the mutation rate if mutations were introduced during DNA replication.

Virtually all of our knowledge of mutational processes in plastids is derived from patterns of DNA substitution...
between species. In a strictly neutral stretch of DNA, the amount of divergence between a pair of species is proportional to the mutation rate and the time since divergence (Kimura 1983). Analysis of organelle genome divergence has been a fruitful approach and yielded some general conclusions about mutation in plastid genomes. In land plants, most evidence suggests that plastid genomes have a lower mutation rate than their nuclear counterpart, but higher than the mitochondrial genome (for review see Smith and Keeling 2012). The lower plastid mutation rate may be limited to land plants, however, and the opposite appears to be true in most other taxonomic groups (Smith 2015). For example, in red algae and glaucophytes, putatively neutral substitution rates in the plastid are similar to the nuclear genome, but much lower than the rate of substitution in the mitochondria. In green algae, the focus of this study, divergence in the plastid, mitochondrial, and nuclear genomes is remarkably similar (Smith and Keeling 2012).

There are a number of difficulties with using between-species nucleotide divergence for estimating mutation rates (for a review see Sloan and Taylor 2011). Primarily, the assumption that nucleotide substitutions are neutral is unlikely to be true. Synonymous sites are frequently used as a proxy for neutral sites, and sometimes combined with intronic or intergenic sites (often referred to as “silent” sites). However, there is clear evidence that synonymous sites are under selection related to codon usage bias, and RNA secondary structure (Chamary et al. 2006). Silent sites may also be influenced by nonneutral processes, such as GC-biased gene conversion that are heterogeneous across the genome(s) and between species, complicating the interpretation of substitution rates (Lesecque et al. 2013). In addition, the base composition of synonymous sites is distinct from the genome background, and, given that the mutation rate of the four bases is variable and dependent on the nucleotide sequence surrounding a given site, synonymous sites are expected to have their own mutational properties that may vary between nuclear and organellar genomes (Ness et al. 2015). Finally, phylogenetic-based estimates of substitution need to account for multiple mutations occurring at the same site, which is especially problematic when comparing distantly related taxa. Phylogenetic methods are also prone to violations of their assumptions, such as the presence of recombination, lateral gene transfer, and intraspecific polymorphism (see Lanfear et al. 2010; Sloan and Taylor 2011). More direct measurements of the properties of spontaneous mutation are therefore needed to assess how well patterns of substitution reflect the process of mutation in plastid genomes.

One method to circumvent the complications of inferring mutation from substitution and to disentangle mutation rate from divergence time is to measure de novo mutations directly in mutation accumulation (MA) experiments. In a MA experiment populations are maintained for many generations under minimal natural selection to allow mutations to accumulate regardless of their fitness consequences. Increasing the strength of genetic drift by regularly bottlenecking the population allows random, unbiased accumulation of all but the most strongly deleterious mutations. When MA is combined with genome sequencing, we can directly study of the process of mutation (e.g., Zhu et al. 2014; Ness et al. 2015). Data from direct estimates of mutation rate in mitochondria suggest that substitution-based estimates of mutation may be as much as two orders of magnitude lower than the true rate (Denver et al. 2000; Howell et al. 2003; Haag-Liautard et al. 2008; Lynch et al. 2008). Data of this kind do not exist for plastid genomes because, at least in land plants, they are thought to have lower mutation rates than mitochondria and therefore require many generations of MA and/or sequencing of many mutated genomes to observe de novo mutations. Two previous MA experiments that estimate the mutation rate in *Chlamydomonas reinhardtii* failed to detect any mutations in the plastid genome (Ness et al. 2012; Sung et al. 2012).

Estimates of the spontaneous mutation rate can be combined with patterns of genetic variation to describe key properties of species’ history and population genetics. Following Kimura (1983), genetic divergence between two species is twice the product of the mutation rate and the number of generations (t) since speciation (2μt). A direct estimate of mutation rate can therefore be used to calculate divergence times between species, and identify regions of the genome where selection has increased or decreased the rate of molecular evolution. In addition, the amount of neutral genetic variation maintained in a population (θ) is determined by the strength of drift, measured as the effective population size (N_e), and the mutation rate via the relationship \( \theta = 2N_e\mu \) (4N_eμ in diploids). We can therefore use the mutation rate to infer N_e from natural levels of polymorphism. The effective population size is a fundamentally important property of a population that has a strong impact on the efficacy of selection (2N_eμ_s) and the breakdown of linkage disequilibrium (LD) (population recombination rate, \( \rho = 2N_e\rho_r \) where \( \rho \) is the rate of recombination per base per generation). Thus, the integration of laboratory-based estimates of mutation with population genomic sequencing from natural samples can facilitate investigation into the biology of species not possible with either approach in isolation.

In this study, we present the first direct estimate of the spontaneous mutation rate in a plastid genome by combining an MA experiment with sequencing of the plastid genome of *C. reinhardtii*. We combine our estimate of mutation rate with whole-plastid sequencing of 20 natural isolates of *C. reinhardtii* to investigate patterns of drift and recombination. *Chlamydomonas reinhardtii* is a unicellular green algae with a plastid genome of 204 kb. It can reproduce both sexually and asexually. Sex occurs by fusion of opposite mating types (mt−and mt+), after which the plastid from the mt−parent and the mitochondria from the mt+parent are eliminated, resulting in a primarily uniparental mode of organellar inheritance. Comparisons of divergence between *C. reinhardtii* and *C. incerta* genomes suggest that, unlike land plants, the plastid mutation rate of *C. reinhardtii* is similar to the rate in the nuclear and mitochondrial genomes (Hua et al. 2012). However, for the above mentioned reasons it is difficult to know whether the patterns of substitution reflect real differences in the mutation between the three genomes. In this
article, we address three questions 1) What is the mutation rate of the plastid genome of C. reinhardtii? 2) How does the rate and spectrum of plastid mutation compare with the nuclear genome? 3) Using the mutation rate, what is the effective population size of the C. reinhardtii plastid and nuclear genomes and how does this relate to recombination?

Results and Discussion

Plastid Mutation Rate

In total, across all 85 MA lines, we analyzed $1.38 \times 10^7$ plastid genome sites over an average of 929 generations (mean coverage of each line = 1,565, from 317 to 4,743). On average, 163 kb of the 204 kb plastid genome was callable in each line and the missing sites were accounted for by the two approximately 19 kb inverted repeats, where ambiguous alignments made genotypes difficult to call (fig. 1). We identified 12 de novo mutations, including ten single nucleotide mutations (SNMs) and two short insertions (fig. 1 and table 1). We confirmed nine of the ten SNMs and the two insertions using Sanger sequencing. The unconfirmed SNM appeared very well aligned in the Illumina read data (see supplementary fig. S1, Supplementary Material online), but the region in which it occurs is flanked by repetitive DNA that made PCR amplification of a single locus difficult. We therefore treated this SNM as a true mutation, yielding a total mutation rate of $\mu_{\text{total}} = 9.23 \times 10^{-10}$ (Poisson 95% CI 4.8–16.1 $\times 10^{-10}$) with an SNM rate $\mu_{\text{SNM}} = 7.69 \times 10^{-10}$ (excluding the unconfirmed SNM yields, $\mu_{\text{total}} = 8.46 \times 10^{-10}$ and $\mu_{\text{SNM}} = 6.92 \times 10^{-10}$).

Both the total and SNM plastid genome mutation rates are marginally lower, but not significantly different from, the nuclear genome mutation rate ($\mu_{\text{nuclear total}} = 11.5 \times 10^{-10}, \mu_{\text{nuclear SNM}} = 9.63 \times 10^{-10}$), estimated in the same sets of 85 MA
The overall similarity of the mutation rate between the nuclear and plastid genomes is consistent with similar levels of synonymous nucleotide divergence in the nuclear and plastid genomes between *C. reinhardtii* and *C. incerta* (Smith 2015). Thus, even if selection or other forces, such as biased gene conversion, influence silent site divergence, their effects are similar in the nuclear and plastid genomes. Unfortunately, we cannot estimate the extent to which the rate of synonymous mutations the affected gene is reported.

Note.—The strand of each mutation was converted so that ancestral state is an A or C. For intergenic mutations, the two genes flanking each mutated site are indicated, and for the nonsynonymous mutations the affected gene is reported.

*aThis mutation could not be confirmed using PCR-Sanger sequencing.*

Table 1. Table of De Novo Mutations Identified in the Plastids of 85 MA Lines of *Chlamydomonas reinhardtii*.

<table>
<thead>
<tr>
<th>Position</th>
<th>Mutation</th>
<th>Strain-MA Line</th>
<th>Annotation</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>2245</td>
<td>A→C</td>
<td>CC-2344-MA7</td>
<td>Intergenic</td>
<td>trnV−petA</td>
</tr>
<tr>
<td>1265</td>
<td>C→A</td>
<td>CC-2931-MA9</td>
<td>Intergenic</td>
<td>psbK−tufA</td>
</tr>
<tr>
<td>26027</td>
<td>A→ATGTTACATACGGAAGGAGGA</td>
<td>CC-2937-MA15</td>
<td>Intergenic</td>
<td>rpl2−rps19</td>
</tr>
<tr>
<td>29936</td>
<td>A→C</td>
<td>CC-2342-MA12</td>
<td>Intergenic</td>
<td>rpl14−rpl15</td>
</tr>
<tr>
<td>71706</td>
<td>A→AAGTTGTTATGTTAATCGGC</td>
<td>CC-2342-MA13</td>
<td>Intergenic</td>
<td>psaA−tmfE</td>
</tr>
<tr>
<td>94282</td>
<td>C→A</td>
<td>CC-2931-MA9</td>
<td>Intergenic</td>
<td>rps9−psbE</td>
</tr>
<tr>
<td>101819</td>
<td>C→T</td>
<td>CC-2344-MA9</td>
<td>Intergenic</td>
<td>rpoB1−psbF</td>
</tr>
<tr>
<td>102393</td>
<td>C→T</td>
<td>CC-2931-MA1</td>
<td>Intergenic</td>
<td>psbF−psbL</td>
</tr>
<tr>
<td>104852</td>
<td>A→C</td>
<td>CC-1952-MA4</td>
<td>Nonsynonymous</td>
<td>rps3</td>
</tr>
<tr>
<td>117587</td>
<td>A→T</td>
<td>CC-1952-MA7</td>
<td>Intergenic</td>
<td>rpoC2−trnY</td>
</tr>
<tr>
<td>178067</td>
<td>A→C</td>
<td>CC-2937-MA4</td>
<td>Nonsynonymous</td>
<td>ORF2971</td>
</tr>
<tr>
<td>191233a</td>
<td>A→C</td>
<td>CC-2931-MA13</td>
<td>Intergenic</td>
<td>trnN−rpoC1B</td>
</tr>
</tbody>
</table>

**Base Spectrum of Plastid Mutation**

There is a notable difference in base spectrum of mutations between the plastid and nuclear genomes. Although the total number of mutations was small, the base spectrum of the plastid is significantly different from the nuclear genome ($\chi^2(5,10) = 11.4, P < 0.05$). Although the nuclear genome is strongly biased toward C:G -> T:A transitions, in contrast the plastid genome SNMs show an excess of A:T -> C:G and C:G -> A:T mutations and a deficit of A:T -> G:C (from Ness et al. 2015). The excess of A:T -> C:G transversions is consistent with a study that examined mutation in three positions of an Aat II restriction site in the *C. reinhardtii* plastid 16 S sequence (GuhaMajumdar and Sears 2005). Using the population genomic data, we also examined the composition of naturally occurring polymorphisms in the plastid to assess whether their base spectrum is consistent with our inference from de novo mutations. Specifically, low frequency polymorphisms at 4-fold degenerate sites will have had less time to be influenced by selection, and should therefore reflect the base spectrum of mutations. In accordance with the de novo mutations there is an excess of A:T -> C:G singleton and doubleton polymorphisms at 4-fold degenerate sites (observed/expected = 1.5), but there is an even stronger bias towards mutations at C:G sites (observed/expected = 2.8). Both the base composition of de novo mutations and 4-fold singletons predict that the equilibrium GC content of the plastid would be 39.7 which is close to the observed value of 34.5%. This contrasts with the nuclear mutation spectrum, which would predict an equilibrium GC content of 29% (Ness et al. 2015). Interestingly, the base composition of 4-fold degenerate sites in the plastid genome is 89.2% AT, suggesting strong selection on codon usage in plastid genes.

**Integrating Mutation Rate Estimates with Population Genetics**

We can use our estimate of the mutation rate to quantify the time since divergence from other species. If substitutions are neutral, then the expected divergence between two species is $2\mu t$, where $t$ is the time in generations since the split. Based
on a divergence value of 0.30 for synonymous sites in the plastid (Hua et al. 2012) the split between *C. reinhardtii* and *C. incerta* therefore occurred approximately $1.93 \times 10^8$ generations ago. Unfortunately, we know little to nothing about the number of generations per year in nature and therefore cannot speculate on the amount of time that separates these species. As a result, it is not possible to ascertain the extent to which estimates of neutral substitution rate and spontaneous mutation rate differ.

Our estimate of the mutation rate can reveal other aspects of the population biology and evolutionary history of *C. reinhardtii*. Equilibrium neutral nucleotide diversity ($\theta$) is proportional to product of the mutation rate and effective population size ($\theta = 2N_e \mu$). The effective population size reflects the amount of genetic drift in a population, specifically the number of individuals in an idealized Wright–Fisher population that would display the observed amount of drift. $N_e$ can therefore be altered by population size fluctuations, bottlenecks, and natural selection. Among the 20 natural isolates from Quebec, plastid genome neutral genetic diversity was 7-fold lower than that of the nuclear genome (plastid $\theta_s = 0.0039$, nuclear $\theta_s = 0.0275$, bootstrap test $P < 0.001$; fig. 2). If the plastid genome is uniparentally inherited, all else being equal, it is expected to have approximately one-half the $N_e$ of the nuclear genome. The SNM mutation rate estimates from the plastid (7.8 $\times 10^{-10}$) and nuclear genomes (9.65 $\times 10^{-10}$; Ness et al. 2015) imply that $N_e$ for the plastid genome is 2.8 $\times$ lower than the predicted 2-fold reduction (plastid $N_e = 2.5 \times 10^5$, SE 1.6 $\times 10^5$, nuclear $N_e = 1.4 \times 10^7$ SE 2.0 $\times 10^5$, bootstrap test $P < 0.01$, fig. 2). Given that the plastid genome is thought to be largely non-recombining, hitchhiking, that is, by selection for advantageous alleles (sweeps) or selection against deleterious alleles (background selection), may reduce plastid $N_e$ more severely than the nuclear genome. Moreover, strong effects of hitchhiking could be observed in the presence of recombination if the plastid is subject to more frequent and stronger selection than comparable regions of the nuclear genome.

Having an estimate of $N_e$ based on the mutation rate and genetic diversity, we can use patterns of LD to test the
assumption that the rate of recombination is indeed lower in the plastid. During sexual reproduction in *C. reinhardtii*, there is a period after the gametes fuse, but before the mt—plastid is eliminated, that provides an opportunity for recombination to occur. Using the program LDHelmet, we estimated the population recombination rate $\rho = 2N_r\rho$ for each of the nuclear chromosomes and the plastid. As expected, the plastid had much stronger LD and significantly lower $\rho$ than observed in the nucleus ($\rho_{ptDNA} = 4.2 \times 10^{-4}$, $\rho_{ntDNA} = 2.7 \times 10^{-3}$, bootstrap test $P < 0.001$). With our estimate of $N_e$ for the nuclear and plastid genomes, we can estimate the rates of recombination per cell division ($r$) from $r = \rho/2N_r$. This calculation reveals that, in nature, the physical rate of recombination in the plastid is similar to the rate of recombination in the nuclear genome (Plastid $r = 8.2 \times 10^{-11}$/bp/generation, Nuclear $r = 9.2 \times 10^{-11}$/bp/generation, bootstrap test $p = 0.64$, fig. 2). Our finding seems unexpected, given the widely held view that the plastid is uniparentally inherited. However, there is clear evidence from genetic studies in *C. reinhardtii* that demonstrates nonreciprocal exchange between parental plastid genomes (e.g., Boynton et al. 1987; Lemieux and Lee 1987; Durrenberger et al. 1996). If this form of gene conversion affects only short tracts of DNA it would explain why most of the plastid genome transmitted in any one cross is inherited from the mt+ parent. However, over evolutionary time, gene conversion can break down LD, especially between nearby loci.

Given that recombination rates of the plastid and nucleus appear to be similar, explaining the reduction in plastid $N_e$ by hitchhiking would require substantially more frequent or stronger selection in the plastid than the nucleus. One possible cause of increased hitchhiking in the plastid is the high gene density of the plastid genome, that is, it contains 69 protein-coding genes and 40 RNA genes, compared with an average nuclear region of the same size carrying only 28 protein-coding genes. However, because nuclear genes are larger, the mean fraction of protein coding or RNA coding sites in the plastid genome is not substantially different from the average across the nuclear genome (plastid 43.3% vs. nuclear 35.5%). It is possible that the mutations in the plastid tend to have stronger selection coefficients causing either selective sweeps for beneficial changes (Maynard Smith and Haigh 1974) or background selection when mutations are deleterious (Charlesworth et al. 1993). Future investigation into the distribution of fitness effects of mutations across both genomes will be necessary to determine whether the reduction in $N_e$ in the plastid can be attributed to stronger effects of hitchhiking.

We have shown that rates of de novo mutation in the plastid and nuclear genomes are of similar magnitudes, in agreement with earlier studies based on DNA substitution patterns. This contrasts with the chloroplast genomes of land plants, where substitution rates are substantially lower than their nuclear counterparts. By integrating our direct, laboratory-based estimate of mutation with population genomic data, we have inferred that strong genetic drift, likely driven by hitchhiking, has reduced the amount of genetic diversity and increased LD in the plastid genome. Our finding that the recombination rates of the plastid and nuclear genomes are similar means that in terms of population genetics we should not consider the plastid to be uniparentally inherited. This implies that the reduced diversity in the plastid genome is likely the result of stronger or more frequent selective sweeps and background selection. Increasing our knowledge of the role of mutation in shaping plastid genome evolution will require more comprehensive sampling among the millions of plastid-bearing species. As technical limitations associated with detecting de novo mutations are removed, opportunities will arise to investigate the spectrum and mechanisms of mutation in diverse plastid genomes and compare these to the nuclear genome.

Materials and Methods

MA Experiment

We conducted a MA experiment in six genetically diverse strains of *C. reinhardtii* obtained from the Chlamydomonas Resource Center (chlamycollection.org). The strains were chosen to broadly cover the geographic range of known *C. reinhardtii* samples in North America (Supplementary table S1, Supplementary Material online). To initiate the MA lines, a single colony from each of the six ancestral strains was streaked out, and we randomly selected 15 individual colonies to start the replicated MA lines (90 MA lines total). We bottlenecked the MA lines at regular intervals by selecting a random colony that was streaked onto a fresh agar plate. We estimated the number of generations undergone by each MA line over the course of the experiment by measuring the number of cells in colonies grown on agar plates after a period of growth equivalent to the times between transfers in the experiment. Characterization of these MA lines was published in Morgan et al. (2014) and whole-genome sequencing and characterization of the nuclear mutation rate was published in Ness et al. (2015).

Population Genomic Sequencing

Twenty natural isolates of *C. reinhardtii* sampled from two fields approximately 75 km apart in Quebec, Canada, were acquired from the Chlamydomonas collection (www.chlamycollection.org, last accessed December 11, 2015). These *C. reinhardtii* strains are the only collection where more than two individuals from a single locality were collected. A within-location sampling strategy allowed better estimation of population genetic parameters, such as the effective rate recombination ($\rho = 2N_r\rho$) and genetic diversity ($\theta_w = 2N_r\mu$), than a species-wide sample, which may be confounded by the effects of population structure (Städler et al. 2009).

Sequencing and Alignment

DNA was extracted from frozen, concentrated cultures using a standard phenol–chloroform extraction. We conducted whole-genome resequencing on the Illumina GAI platform at the Beijing Genomics Institute (BGI-HongKong Co., Ltd, Hong Kong). We modified the library amplification to accommodate the unusually high GC content of the *C. reinhardtii*
genomes (mean GC = 63.9%) following (Aird et al. 2011) (3 min at 98°C; 10 × [80 s at 98°C, 30 s at 65°C, 30 s at 72°C]; 10 min at 72°C and slow temperature ramping 2.2°C/s). We obtained an average of 28.6 × coverage of each sampled genome (3 Gb of 100 bp paired-end sequence) and because the plastid is present at high copy number in the cell, mean coverage of the plastid was 1034 ×.

We aligned sequencing reads using Burrows-Wheeler Aligner (BWA) 0.7.4-r385 (Li and Durbin 2009) to the GATK v3.3 (Depristo et al. 2011) tool “HaplotypeCaller” (NCBI accession GU814015). To call genotypes, we used the plastid genome (NCBI accession NC_005353), mitochondrial genome does not include the organelles, we added the in silico mutation identification and population genetic analysis.

We investigated the possibility that spontaneous mutations may be heteroplasmic within an MA line. Alleles at heteroplasmic sites are not expected to segregate at 1:1, like heterozygote sites in diploids, and will not be accurately called by standard variant calling algorithms. We therefore attempted to detect heteroplasmic sites directly from raw read data, extracted using SamTools mpileup (Li et al. 2009). We applied the following filters to each group of replicate MA lines from a common ancestor:

1. One individual has ≥50× coverage, ≥5% of which carry a minor allele
2. At least two MA lines had reads aligned to the position
3. Other samples have <2 aligned reads that carry the candidate mutant allele

We estimated the accuracy of our mutation calls using Sanger sequencing of all candidate mutations including both indels and SNMs. We amplified each locus in the putative mutant MA line and a nonmutated line from the same ancestral strain. Sequences were then visually inspected in SeqTrace v0.9.0 to confirm the presence of the mutated site.

We calculated the mutation rate as, μ = mutations/(high quality sites × MA generations). To obtain accurate mutation rate estimates, it is necessary to define the number of sites at which a mutation could have been detected. Such “callable” sites were determined using the same criteria outlined above except we required all MA lines from a given ancestor share the same genotype and the minimum PHRED-scaled QUAL score was 36 to account for the difference in QUAL distributions between invariant and variant sites (see Ness et al. 2015). To analyze mutation spectrum, we treated complementary mutations (C:G and A:T) symmetrically, such that there were six distinct SNMs (AT→CG, AT→GC, AT→TA, CG→AT, CG→GC, CG→TA). To assess the base spectrum of mutation, we calculated the frequency of each mutation type relative to a random expectation adjusted for the base composition of the callable sites in the plastid genome. We also compared the plastid and nuclear mutation spectrum reported by Ness et al. (2015).

We estimated nucleotide diversity (θπ = 2Nμ) in the plastid and nuclear genomes of the 20 natural lines from Quebec. We estimated neutral diversity using silent sites (4-fold degenerate, intronic, and intergenic sites) and we estimated selective constraint as the ratio of diversity at 0-fold and 4-fold degenerate sites (θs/θπ). We inferred the effective population size (Ne) for the plastid and nuclear genomes using their respective levels of silent diversity (θπ) and mutation rates (μ) as Ne = θπ/2μ. If the plastid genome is uniparentally inherited, all else being equal, it is expected to have approximately one-half the Np of the nuclear genome. Using 4-fold degenerate sites to estimate “neutral” diversity suffers from the problems explained in our introduction, therefore our estimates of neutral diversity may be underestimated. To test whether the θπ for nuclear and plastid genomes differ, we calculated θπ for sliding windows of 10 kb in both genomes. We then bootstrapped the sliding windows 1,000 times with replacement and for each bootstrap replicate we calculated the difference in mean θπ between the plastid and nuclear genomes. The proportion of bootstrap replicates with differences in mean θπ that overlap zero indicate whether the plastid and nuclear genomes have significantly different mean levels of nucleotide diversity.

We inferred variation in the effective rate of recombination (ρ = 2Nμr) from polymorphism data with the program LDhelmet (Chan et al. 2016). LDhelmet uses a similar reverse-jump Markov Chain-Monte Carlo approach as employed in LDHat (Mcvean et al. 2004), but includes modifications to improve recombination rate estimates and computational efficiency, especially when there are missing data. We generated haplotype configurations with the
recommended window size of 50 single nucleotide polymorphisms (\(\mathrm{w}=50\)). To calculate the likelihood look-up table and Padé coefficients, we used the mean 4-fold degenerate diversity measures from the nuclear and plastid genomes for the \(\theta\) parameters (\(\mathrm{t}=0.039\) and 0.003, respectively). To estimate whether \(\rho\) varied between the nuclear and plastid genomes we calculated \(\rho\) per base in sliding windows of 10 kb. We then used the same bootstrapping method described above for nucleotide diversity to determine whether there was a significant difference between \(\rho\) in the plastid and nuclear genomes.

We estimated \(N_e\) as \(\theta_{\pi}/2\mu\) and the recombination rate \(r\) as \(\rho/2N_e\). We approximated error in our \(N_e\) estimate by resampling both the mutation rate and \(\theta_{\pi}\). For mutation rate, we drew 1,000 samples from a Poisson distribution with a mean equal to the number of observed mutations in the nuclear and plastid genomes. We then calculated \(\mu\) for each draw by dividing it by the number of sites and generations. We then estimated error in \(N_e\) by dividing the mean of each of 1,000 bootstrap samples of \(\theta_{\pi}\) by twice each sampled mutation rate. The standard error was calculated as the standard deviation of the 1,000 samples. We then used the 1,000 \(N_e\) values to estimate the standard error of the recombination rate (\( \bar{r}\)) by dividing each of the 1,000 bootstrap samples of \(\rho\) by twice each sampled value of \(N_e\).

**Supplementary Material**

Supplementary table S1 and figure S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

The work presented here was funded by a British Biotechnology and Biological Sciences Research Council (BBSRC) grant (BB/H006109/1) to P.D.K. and N.C.

**References**


