Estimation of the SNP mutation rate in two vegetatively propagating species of duckweed

George Sandler[1], Magdalena Bartkowska[1], Aneil F. Agrawal[1,2]*, Stephen I. Wright[1,2]*

[1]Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks Street, Toronto, ON M5S 3B2, Canada; [2]Center for Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks Street, Toronto, ON M5S 3B2, Canada

Author for correspondence

George Sandler

Tel: +1 416 6172717

Email: george.sandler@mail.utoronto.ca

*These authors contributed equally to this work.

Key words: Mutation rate, vegetative reproduction, duckweed, asexual, effective population size

Total word count	**6159**	**No. Figures**	**4**
Abstract	**128**	**No. Tables**	**0**
Introduction	**1007**	**No. Supporting information files**	**1**
Materials and Methods	**1926**		
Results	**1053**		
Discussion	**2173**		
Acknowledgements	**41**		

Abstract

Mutation rate estimates for vegetatively reproducing organisms are rare, despite their frequent occurrence across the tree of life. Here we report mutation rate estimates in two vegetatively reproducing duckweed species, *Lemna minor* and *Spirodela polyrhiza*. We use a modified approach to estimating mutation rates by taking into account the reduction in mutation detection power that occurs when new individuals are produced from multiple cell lineages. We estimate an extremely low per generation mutation rate in both species of duckweed and note that allelic coverage at *de novo* mutation sites is very skewed. We also find no substantial difference in mutation rate between mutation accumulation lines

propagated under benign conditions and those grown under salt stress. We discuss the implications of interpreting mutation rate estimates in vegetatively propagating organisms.

Introduction

Most research on the evolution of mutation rates has focused either on sexually reproducing eukaryotes or unicellular organisms, both of which feature a single cell phase as part of their life cycle. However, a diverse array of organisms reproduce either through clonal budding, fission or vegetative growth, whereby a single cell phase is not imposed every generation (Bell, 1982). This mode of reproduction potentially allows multiple cell lineages to be transmitted from parent to offspring, complicating the process of genotyping individuals. This happens because when individuals composed of a mosaic of cells are sequenced, the mean number of sequencing reads supporting non-reference mutations is no longer 50%. Such a skew in allelic coverage makes it harder to distinguish true mutations from sequencing errors (Cibulskis *et al.*, 2013), complicating the assessment of power when calculating per base pair mutation rates. Even if a cellular mutation rate can be calculated for an organism with multiple cell lineages, it becomes more challenging to use this parameter in population genetics analyses as mutations can potentially be lost within an organism before truly contributing to population level genetic diversity. Previous theoretical work modeling mutation load in organisms with multiple cell lineages has suggested that cell lineage selection can significantly reduce mutation load by purging deleterious mutations during somatic growth (Otto & Orive, 1995). As most new mutations are thought to be deleterious (Sturtevant, 1937; Eyre-Walker & Keightley, 2007) this type of selection might skew the level of genetic diversity observed in organisms with vegetative reproduction compared to the level expected given their per base pair mutation rates.

Previous studies have investigated the rate of somatic mutations in plants where multiple cell lineages can segregate within a generation (Watson *et al.*, 2016; Schmid-Siegert *et al.*, 2017; Plomion *et al.*, 2018; Wang *et al.*, 2019). While somatic mutations can be transmitted from generation to generation in plants (Plomion *et al.*, 2018; Wang *et al.*, 2019), if somatic growth is followed by sexual reproduction, a single cell bottleneck is nonetheless imposed on any segregating variation within the soma, removing the persistence of multiple cell lineages across generations. This is however not the case for organisms reproducing through vegetative growth, budding or fission. Despite their

frequency across the eukaryotic tree of life, almost no per-base-pair mutation rate estimates exist for organisms procreating through such modes of reproduction. One recent study in a vegetatively growing fairy-ring mushroom reported very low mutation rates per mitotic cell division (Hiltunen *et al.*, 2019). The authors of this work used simulated mutations to assess the level of power they had to detect low frequency *de novo* mutation in this dataset, improving their estimate of the fairy ring mushroom mutation rate.

Here we report mutation rate estimates in two species of duckweed (*L. minor* and *S. polyrhiza*). Both species are free-floating, facultatively sexual aquatic plants. While duckweed can produce seed though sexual reproduction, most growth occurs vegetatively via clonal budding from two pouches present in the duckweed frond (Landolt, 1986). As a result, while these species are found all across the globe and likely have enormous census population sizes, allozyme and genomic analyses have revealed low levels of genetic diversity within local populations (Cole & Voskuil, 1996; Ho, 2018; Xu *et al.*, 2019; Ho *et al.*, 2019). Work by Xu et al (2019), has estimated the per base pair mutation rate in a genotype of *S. polyrhiza* growing in the field, finding an extremely low rate of mutation. However, their analysis did not take into account the fact that duckweed individuals are likely composed of a mosaic of cell lineages during periods of asexual growth, potentially leading to an underestimate of the true mutation rate.

Studying two duckweed species allows us to contribute to three other questions in mutation rate evolution research. First, our mutation rate estimates provide another species to add to the existing set of species with mutation rate estimates that, collectively, allow for testing the theory that selection against mutators should be most efficient in species with large effective population sizes (Sung *et al.*, 2012; Lynch *et al.*, 2016). Duckweeds are useful additions to this set as previous work has suggested that the effective population size (N_e) of *S. polyrhiza* is on the order of a million individuals (Ho et al. 2019; Xu et al. 2019). Second, the inclusion of two facultatively sexual species that differ in their degree of sexuality allow us to preliminarily investigate the effect of recombination on the evolution of mutation rates. Genomic and allozyme patterns have suggested that *L. minor* undergoes bouts of sexual reproduction more often than *S. polyrhiza*, a pattern that is in line with flowering observations of these species in the field (Hicks, 1932; Landolt, 1986). Theoretical work has shown that when recombination breaks apart associations between mutator alleles (that elevate mutation rate) and the mutations they produce, mutation rates can evolve in

89 several ways. On one hand selection against mutators in more sexual populations may be
90 relaxed as they no longer remain linked to new deleterious mutations (Kimura, 1967; Leigh,
91 1970). Alternatively mutator alleles can spread when recombination is sufficiently low if
92 they hitch hike along with any beneficial mutations they produce (André & Godelle, 2006).
93 Finally, environmental stress is known to increase mutation rates in bacteria in a process
94 known as stress-induced mutagenesis (Foster, 2007). A few examples of stress increasing
95 mutation rates are known in eukaryotes (Matsuba et al. 2013; Jiang et al. 2014; Sharp and
96 Agrawal 2016; but see Saxena et al. 2019), however it is unclear how general this
97 phenomenon is. We performed our experiment both under a control and salt stress
98 treatment to test whether stress-induced mutagenesis is a common phenomenon in plants.
99 We estimated the mutation rate in 46 asexually propagated mutation accumulation
100 lines, including two genotypes of *S. polyrhiza* and one genotype of *L. minor*. We report an
101 exceptionally low rate of mutation in both species of duckweed and note a pattern of
102 skewed allelic counts at *de novo* sites that suggests the presence of multiple segregating cell
103 lineages in vegetatively reproducing duckweed.
104
105 **Materials and Methods**
106 Mutation accumulation and DNA extraction
107 MA lines were started for three genotypes in April 2014 and propagated for
108 approximately 60 generations. Two *Spirodela polyrhiza* genotypes were used: GP23 from
109 Grenadier Pond, Toronto, Canada and CC from Cowan Creek, Oklahoma, USA. A single
110 genotype of *L. minor* (GPL7) was also isolated from Grenadier Pond, Toronto, Canada. For
111 each genotype (CC, GP23 and GPL7), we established 16 MA lines. We generated each line
112 from a single maternal plant, which was started by isolating two fronds from each genotype
113 culture. Because daughter fronds are generated iteratively, we grew and isolated daughters
114 tracking pedigree until a minimum of 16 daughter fronds paired by generation were
115 available for each of the three genotypes (arising from a single starting maternal frond).
116 Daughters in each frond pairs (matched for generation relative to maternal frond) were
117 assigned to one of two growth medium treatments (salt stress and control). Daughters are
118 produced from two pockets of meristem tissue on either side of the maternal frond and
119 mature daughter fronds remain attached to the maternal plant via a stipule for a short time
120 (Landolt, 1986). To ensure that each generation was propagated with a daughter frond, we

121 separated the daughter from the maternal frond as soon as the daughter began producing
122 her own frond. Each line was checked for mature daughter fronds every two days. The first
123 daughter produced was used whenever possible. MA lines were propagated in 0.5X
124 Appenroth liquid growth medium (Appenroth *et al.*, 1996) at 24°C with 12 hours of light per
125 day. Salt stress lines were supplemented with 25mM of NaCl for *S. polyrhiza* and 50mM of
126 NaCl for *L. minor*. Prior to the start of the MA experiment, we performed growth assays to
127 establish stressful NaCl levels for both species. The chosen salinity levels caused duckweed
128 fronds to become patchy and thin but still allowed for continual asexual propagation.
129 After the termination of the MA experiment we allowed MA lines to continue
130 growing for several generations to obtain enough plant material to perform CTAB DNA
131 extractions.
132
133 Sequencing and filtering
134 We sequenced the MA lines at the McGill Innovation Centre. Illumina HISeq 2000
135 sequencing with 100bp paired end reads was used for both *S. polyrhiza* genotypes while
136 Illumina HISeq 2500 sequencing with 125bp paired end reads was used for the *L. minor*
137 genotype.
138 Paired end reads were mapped to the *S. polyrhiza* and *L. minor* reference genomes
139 (Van Hoeck *et al.*, 2015; Cao *et al.*, 2016) using the Burrows-Wheeler aligner (BWA) 0.717
140 using the BWA-MEM option (Li & Durbin, 2009). We then used Picard to remove duplicate
141 reads before calling indels using the HaploTypeCaller tool in GATK 3.7 (McKenna *et al.*,
142 2010). Next, we used the IndelRealigner tool in GATK to perform indel realignment. Finally,
143 we used BCFtools (1.6) (Li, 2011) to create mpileup files for the realigned output from GATK
144 and to call SNPs and short indels (indels no more than 10bp) .
145 We first filtered out sites with unusually high or low coverage. We did this by
146 eliminating sites that had coverage outside +/- 200x median coverage (summed across all
147 lines) in each genotype individually. We visualised relatedness between our lines using a
148 PCA plot created from heterozygous sites present in our MA lines in R (v5.3.5) (R Core Team,
149 2019) using the package SNPRelate (Zheng *et al.*, 2012). In doing so we discovered two
150 major outliers in one of our *S. polyrhiza* genotypes (CC) suggesting that these two lines were
151 cross contaminated. We subsequently removed these two lines from our analysis.

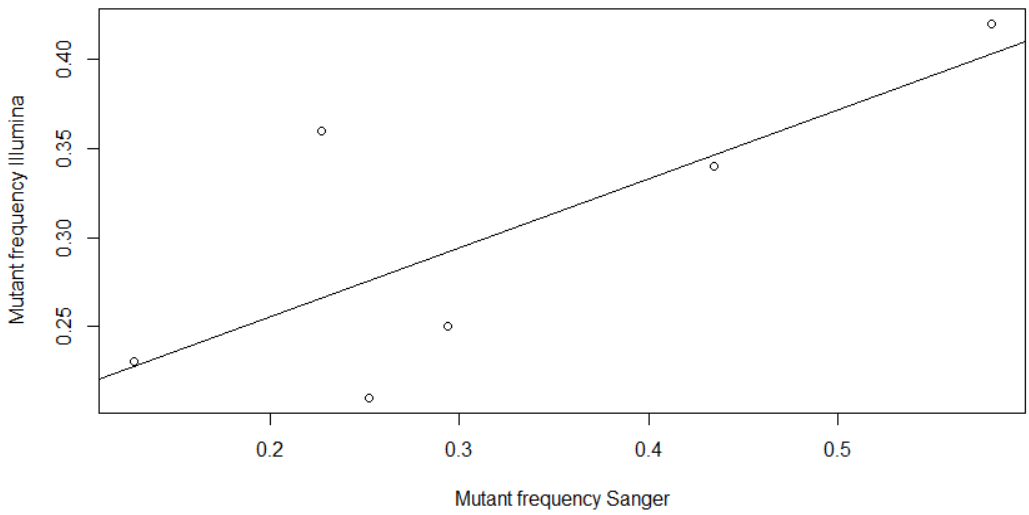
152 Our next round of filtering aimed to remove low quality regions of the genome that
153 contain unusually high amounts of in-phase heterozygous variants (e.g. Fig. S1). Such
154 regions likely represent collapsed duplications in the reference genome that map poorly to
155 an incorrect genomic coordinate. These variants are highly reference-biased in their allelic
156 coverage likely due to the poor mapping of reads that contain many differences relative to
157 the reference genome. To remove such regions, we first created a consensus genotype for
158 each set of lines; if more than one line in a given genotype supported the existence of a
159 heterozygote at a site, that site was designated as heterozygous in the consensus genotype.
160 We then performed a sliding window analysis on heterozygosity on each consensus
161 genotype. We used 1000bp windows with a 100bp step. After trying a variety of filtering
162 criteria, we decided to designate regions of the genome as callable if there existed no more
163 than 10 heterozygous calls in a 1000bp window in each *S. polyrhiza* genotype and no more
164 than 5 heterozygous calls per 1000bp window in the *L. minor* genotype. We used more
165 stringent criteria in *L. minor* due to the lower quality of the genome assembly. These cut-
166 offs represent a trade-off between eliminating problematic, variant-rich areas of the
167 genome and excluding well assembled genomic areas with higher than average diversity.
168 After filtering, around 100Mb of the genome was retained as callable in each of the three
169 genotypes. This filtering step greatly improved the allelic coverage of ancestral
170 heterozygous sites by removing suspected hidden duplications that map poorly to the
171 reference (Fig. S2).
172 Next, we called putative *de novo* mutations in the remaining callable regions. Within
173 each set of lines, we picked sites where one line had a heterozygous genotype, with at least
174 5 reads supporting the non-reference base, but all other lines supported a homozygous
175 genotype. We then extracted such sites from the mpileup file used to call genotypes. This
176 was done as the pileup file contains reads that are filtered out during genotype calling but
177 are useful in our case as they can lead to the elimination of false positive mutations. We
178 filtered putative *de novo* mutations using the mpileup file in two ways. First, if a line other
179 than the one which contained the *de novo* mutation had any reads which supported the *de*
180 *novo* base call we excluded the site. Second, if a site with a *de novo* base call contained
181 reads with more than two non-reference bases across all samples, we also excluded the site.
182 We did this to exclude sites where a high rate of sequencing errors might have occurred. We
183 used this cut-off based on the observation that at sites where all lines supported a

184 homozygous genotype, the vast majority of sites contain no more than one alternate base
185 call (again likely due to sequencing errors which can be observed in the mpileup file). The
186 remaining putative *de novo* mutations that passed these filtering criteria were visually
187 inspected in the Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011). We excluded a
188 few mutations which appeared on reads in complete linkage with other non-reference bases
189 (an indication of hidden genomic duplications) or on reads that looked like the product of
190 PCR or sequencing errors (see Fig. S3-5 for examples).
191
192 Power analysis
193 To calculate the mutation rate, we first needed to know how much power we had to
194 detect *de novo* mutations at our callable sites. To assess power, we first obtained a list of
195 sites where we knew we had non-zero power to detect mutations, this included sites where
196 all lines within a genotype supported a homozygous reference base call and no more than
197 one alternate base was present in the mpileup file (one less than in the case if a *de novo*
198 mutation was present). We then randomly sampled 500,000 such sites from each genotype
199 independently, and randomly chose a line where a mutation could have happened. We
200 randomly eliminated a third of the sites where one alternate base was present in the
201 mpileup file as our filtering criteria would eliminate true *de novo* mutations if another line
202 by chance contained a sequencing error which matched the *de novo* base call (i.e., we
203 assume the probability of this occurring is 1/3). Of the remaining sites, we assigned how
204 many reads would support the *de novo* base call by drawing from a binomial distribution
205 with a success rate of 50%, 34%, 26%, 20% and 10%. These different values were chosen to
206 represent a range of frequencies a mutation may be found due to the inheritance of
207 multiple cell lineages in asexual reproduction. This is similar to the approach taken by
208 Hiltunen et al. (2019) when calculating mutation rates in vegetatively growing fairy-ring
209 mushrooms.
210 Our estimate of power was the proportion of sites (out of the original 500,000) that
211 had at least 5 reads supporting the *de novo* mutation (for each possible binomial success
212 rate that we tested). We then multiplied our power estimates by the number of callable
213 sites in the genome. Then separately for each line, we multiplied the adjusted number of
214 callable sites by the number of MA generations and summed these values across all lines in
215 a given genotype (split by treatment). This provided us with a denominator for our mutation

216 rate calculation. Our final step was to divide the number of *de novo* mutations identified in
217 each genotype (split by treatment) by this denominator. The *de novo* mutation count was
218 adjusted for false positives identified during mutation validation.
219
220 Calling indels
221 We scanned our lines for *de novo* indels in the same way as we searched for point
222 mutations with one key modification. We used the same regions of the genome that we had
223 previously assessed as "callable"; however, we only considered *de novo* indels if they were
224 at least 2000 bp away from any other indel in any other line of the same genotype. This
225 filtering step was required to avoid false positive indel calls which appear due to spurious
226 mapping patterns in repetitive regions.
227
228 Mutation validation
229 We only had tissue from *S. polyrhiza* to perform validation on putative *de novo*
230 mutations. We extracted DNA from our MA lines using a Qiagen DNeasy kit. Afterwards, we
231 designed primers for 14 SNPs and one indel found in our two *S. polyrhiza* genotypes. We
232 performed PCR reactions using FroggaBio PCR mastermix which were then sent off for
233 Sanger sequencing at Eurofins Genomics. We inspected the Sanger sequencing
234 chromatograms in FinchTV 1.4.0 (Geospiza, Inc.; Seattle, WA, USA;
235 http://www.geospiza.com) to see if we could detect a peak suggesting the presence of a
236 mutant base in the focal MA line and checked that the base was not present in a different
237 MA line of the same genotype as a negative control. One line from the *S. polyrhiza* CC
238 genotype contained a massive overabundance of putative *de novo* mutations (of the 31
239 mutations found in 14 lines of the CC genotype, 9 were identified from this one line); all 3
240 variants that we attempted to Sanger sequence validate from this line failed. Ultimately, we
241 excluded this CC line from all the results presented here. Excluding that line, 11 total SNPs
242 that were checked with Sanger sequencing and used to correct for the false positive rate in
243 mutation detection in all our *S. polyrhiza* lines.
244
245 Statistical analysis
246 We calculated 95% confidence intervals for our mutation rates using the Agresti
247 Coull method implemented in the R package "binom" (Dorai-Raj, 2014). We tested for

248 significant differences between our salt and normal treatment lines using independent Chi-
249 square tests in each genotype. To test for differences between *L. minor* and *S. polyrhiza* we
250 merged the number of callable sites (scaled by power) and number of mutations in the two
251 *S. polyrhiza* genotypes (separately for each treatment), and also applied Chi-square tests to
252 test for significance in mutation rate variation between species and conditions.
253
254 Data availability
255 All sequencing data will be deposited to the Sequencing Read Archive at NCBI upon
256 acceptance of this manuscript for publication. Supplementary material has been uploaded
257 as a separate file.
258
259 **Results**
260
261 *De novo* mutations
262 In total, we identified 24 and 17 putative *de novo* mutations in *S. polyrhiza*
263 genotypes GP23 and CC, respectively and a further 29 mutations in *L. minor* (Supplemental
264 table 1). When inspecting the putative *de novo* mutations in IGV, we observed that most
265 mutant sites exhibited highly reference biased allelic counts (over 50% of reads support the
266 reference base call). On average, mutant sites contained 26%, 29%, and 34% mutant reads
267 in genotypes GP23, CC (*S. polyrhiza*) and GPL7 (*L. minor)* respectively. We implemented
268 several filtering and quality control steps to ensure these mapping patterns were not a
269 result of sequencing error or genome mis-assembly. First, we masked areas of the genome
270 that had odd coverage patterns or were enriched for heterozygous calls to avoid areas
271 containing hidden genomic duplications. The heterozygous variants that pass these filtering
272 criteria and are present in the ancestral genotypes (i.e. are shared by all MA lines within a
273 genotype) show relatively normal patterns of reference and non-reference allele coverage.
274 Second, we only considered *de novo* mutations at sites with minimal sequencing errors. This
275 step eliminated problematic areas of the genome prone to genotyping error. Finally, we
276 performed Sanger sequencing validation on 11 putative *de novo* mutations (in *S. polyrhiza*)
277 of which 6 were validated with both positive and negative controls. The pattern of reference
278 bias in our *de novo* mutations was also present in our Sanger sequencing chromatograms;
279 the reference base peak was generally much larger than the mutant peak. When we plotted

280 Illumina mutant base frequency vs. Sanger mutant peak height (standardized by reference
281 peak height), we observed strong concordance for our 6 validated mutations (R^2 = 0.56; Fig.
282 1) suggesting that these are not spurious mapping patterns or sequencing errors but rather
283 reflect a true bias in mutation abundance. When we inspected our PCR products with gel
284 electrophoresis we observed single, clean bands of the expected product size so we
285 consider it unlikely that multiple sites in the genome were amplified leading to odd mapping
286 patterns at our *de novo* mutation sites. One alternative is that reference biased mapping
287 patterns may be due to some bias in the PCR amplification process. We do not believe this is
288 likely however since our primers were designed for sequences that were flanking the site of
289 *de novo* mutations that should be identical whether a *de novo* mutations is present or not.



290

291 ***Fig.*** *1 Proportion of de novo bases from Illumina reads vs. relative peak height of mutant*
292 *base in Sanger sequencing data in Spirodela polyrhiza.* R^2 *= 0.57.*

293

294 The majority of mutations were C -> T transitions in both species of duckweeds (Fig.
295 2) in concordance with patterns found in previous mutation rate studies (Ossowski *et al.*,
296 2010; Thomas *et al.*, 2018). Transitions were more common than transversions, with the
297 average ti/tv ratio being 2.04 for *S. polyrhiza* and 4.5 for *L. minor* propagated in standard
298 medium. This is consistent with previous results reported in *S. polyrhiza* by Xu et al. (2019)
299 where three C -> T and one C -> A mutations were detected and validated.
300 Our analysis uncovered only one putatively *de novo* short indel that turned out to be
301 a false positive based on our Sanger sequencing validation analysis.

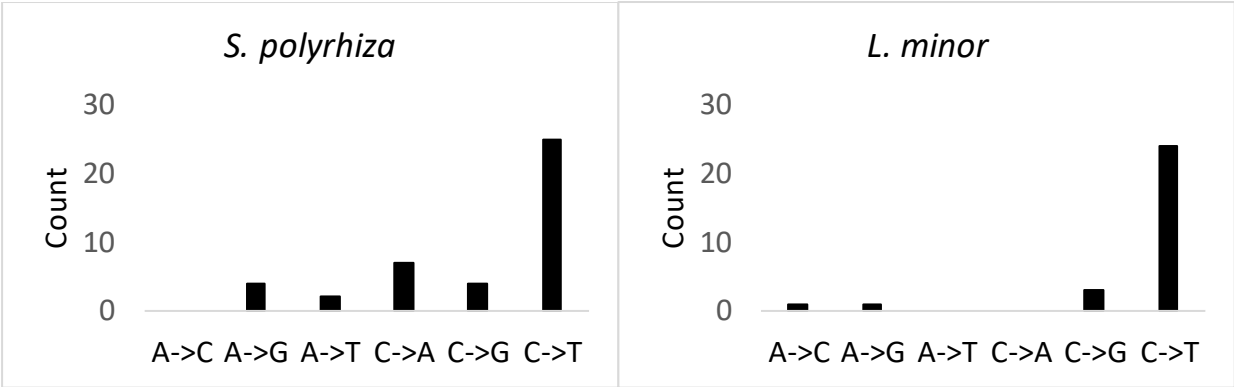


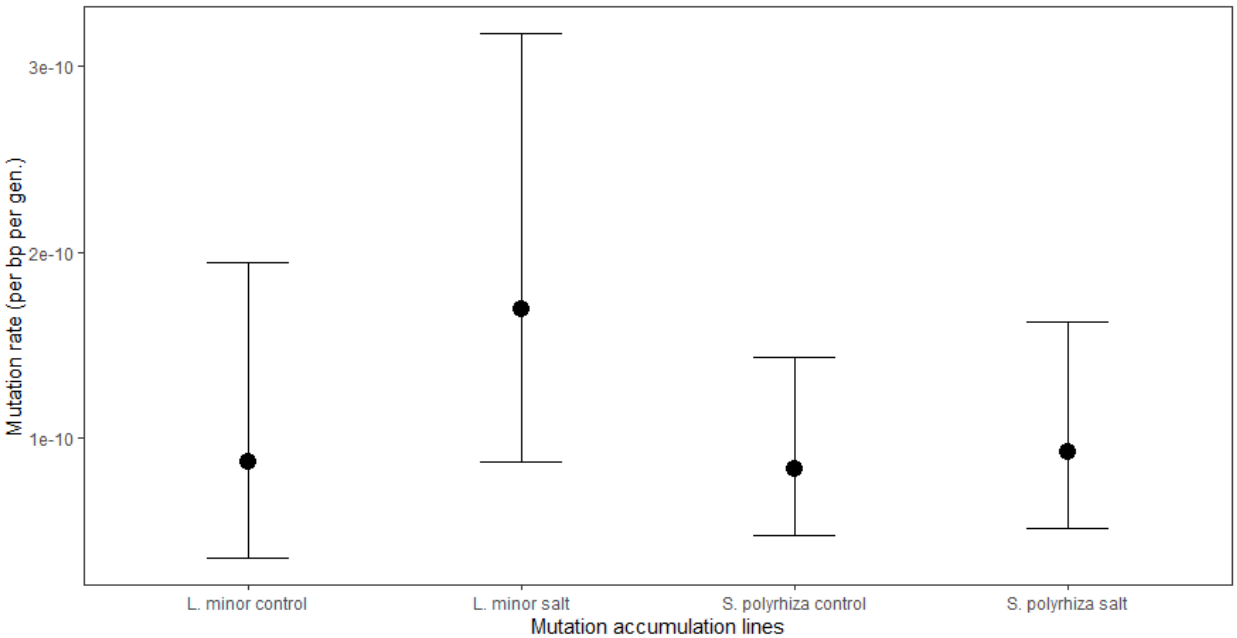
Fig. 2 *De novo mutation spectra in two species of duckweed (Spirodela polyrhiza, Lemna minor).*

Mutation rate comparisons between *S. polyrhiza* and *L. minor*

Our estimate of the per base pair SNP mutation rate highly depends on our power to detect mutations, which in turn depends on factors such as total read depth and the number of reads that support a *de novo* mutation base-call (See Supplemental tables 2-5). Heterozygous SNPs are expected to be supported by around 50% of reads in organisms with a single cell phase. In our case, on average, 27% (*S. polyrhiza*) and 34% (*L. minor*) of reads supported *de novo* heterozygous mutations. To improve our estimate of the mutation rate, we assumed that we should only expect to find mutations that are on average supported by 27% (in *S. polyrhiza)* or 34% (in *L.* minor) of reads at each mutant site (see methods for more details of power analysis). In this case, our point estimate of the mutation rate for plants grown in standard medium is 8.33E-11 for *S. polyrhiza* and 8.66E-11 for *L. minor* (Fig. 3). For plants grown in stressful salt medium our point estimates for the mutation rate are 9.20E-11 for *S. polyrhiza* and 16.9E-11 for *L. minor*.

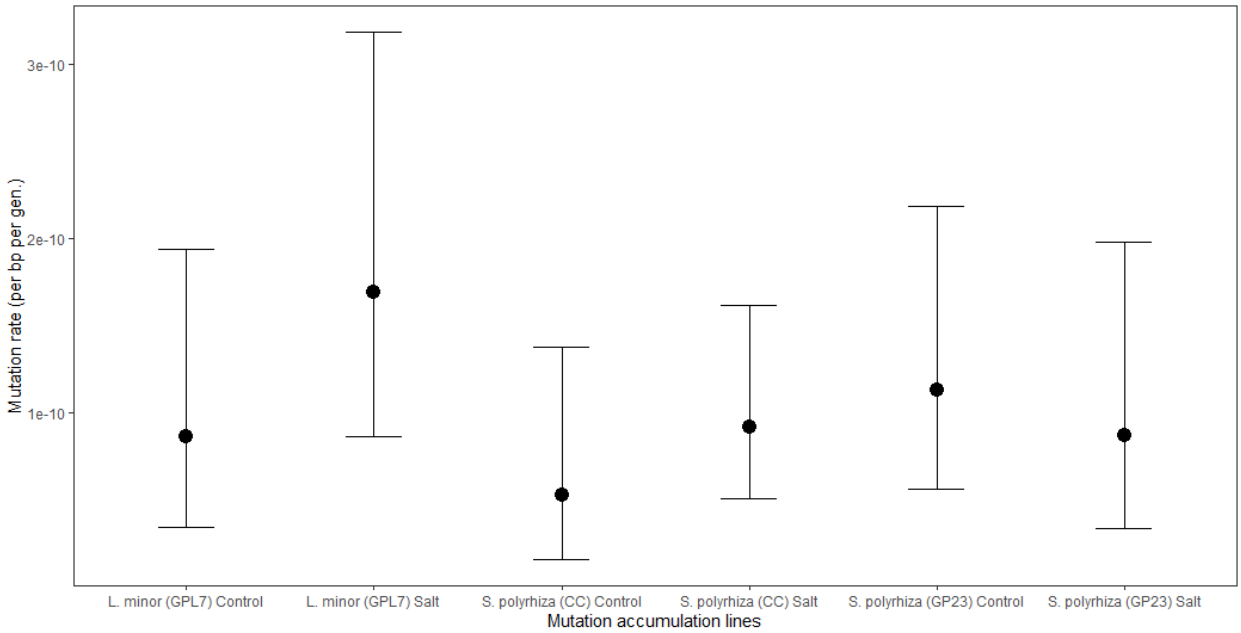
While the estimated mutation rate is higher for *L. minor* in both conditions, the difference between the two species is non-significant (chi-square test: p = 1 and p = 0.22 for control and salt stressed conditions). The difference in mutation rate in the two species is, however, very sensitive to the expected frequency of reads that support the mutant base. For example, if our power estimate is based on the assumption that *de novo* mutations are expected to constitute only 10% of reads, the mutation rate under normal conditions increases to 4.93E-10 in *S. polyrhiza* and 16.8E-10 in *L. minor*. Our mutation rate estimates were not significantly different between our two individual *S. polyrhiza* genotypes (chi-

328 square test: p = 1 and p = 0.29 for control and salt stressed conditions). Supplemental tables
329 3-4 give the inferred mutation rates with alternative assumptions for the expected
330 proportion of non-reference reads at sites harbouring *de novo mutations*.
331



332
333 ***Fig. 3*** *Mutation rate estimates in two species of duckweeds duckweed (Spirodela polyrhiza,*
334 *Lemna minor). Power adjusted to account for multiple cell lineages. Error bars show Agresti-*
335 *Coull 95% confidence intervals.*
336
337 <u>Effect of salt stress on mutation rate</u>
338 There were no significant differences in the mutation rate between lines propagated
339 in normal and salt stress medium for either of our three genotypes (Fig. 4, chi-square test: p
340 = 0.81, p = 0.51, p = 0.29 for genotypes GP23, CC, GPL7 respectively). There was also no
341 consistent trend in the change of mutation rate with the addition of salt stress as genotype
342 GP23 appeared to have a decreased mutation rate while genotypes CC and GPL7 appeared
343 to have an increase in the mutation rate with the addition of salt stress.
344 Salt stress also did not have a significant impact on the ti/tv ratio. In two of the
345 genotypes (1.33 -> 2.67 *S. polyrhiza* CC and 4.5 -> 8 *L. minor*), salt stress increased the ti/tv
346 ration. In *S. polyrhiza* genotypes GP23 salt stress decreased the ti/tv ration (2.75 -> 2.00)
347 although the difference was not significant in either of the three genotypes (chi-square test:
348 p > 0.5 all cases). These results run counter to previous work in *A. thaliana* where salt stress

349 was found to increase the de-novo mutation rate ~two-fold and lower the ti/tv ratio (Jiang
350 *et al.*, 2014).



351

352 ***Fig. 4*** *Effect of salt stress on the mutation rate in two genotypes of Spirodela polyrhiza and*
353 *one genotype of Lemna minor. Error bars show Agresti Coull 95% confidence intervals.*

354

355 **Discussion**

356 Using whole genome sequencing on 36 MA lines we report the *de novo* SNP
357 mutation rate in two species of duckweeds under both control and stressful growth
358 conditions. Our results are in agreement with previous estimates of a low mutation rate in *S.*
359 *polyrhiza* (Xu *et al.*, 2019) while also extending this observation to another duckweed
360 species, *L. minor*. An important result in our study is that *de novo* mutations appear to have
361 considerably reference biased genomic coverage in both duckweed species. We believe that
362 this pattern is not indicative of sequencing or genome assembly errors but rather is a by-
363 product of vegetative reproduction for several reasons. First, upon our inspection of the
364 data in an independent study of mutation rates in *S. polyrhiza* by Xu et al.; we noticed
365 similar levels of reference bias in Illumina short-read sequencing data at validated *de novo*
366 mutations. Second, our own validation with Sanger sequencing showed that mutations with
367 a higher reference bias in the Illumina dataset tended to have higher reference base peaks
368 in their Sanger sequencing chromatograms and vice versa. Moreover, after filtering, most
369 ancestral heterozygous sites that were shared by all lines in the three clonal genotypes were

370 not reference biased in such a manner and those that were, were found in regions highly
371 enriched for non-reference base calls. None of our putative *de novo* mutations are found in
372 such regions; due to the low genetic diversity in duckweed, *de novo* mutations were
373 generally the only heterozygous variants present in the immediate genomic area. This
374 means that mapping bias due to divergence from the reference is unlikely to cause the
375 strong allelic coverage bias we observed. Finally, a study of mutations in a vegetatively
376 growing fairy-ring mushroom also reported reference biased allelic counts at *de novo*
377 mutations (Hiltunen *et al.*, 2019). Therefore, the reference bias in our mutations is most
378 likely a signature of the segregation of multiple cell lines from generation to generation in
379 our duckweed MA experiment. Given that vegetative reproduction through clonal budding
380 is the main form of reproduction in duckweed (Landolt, 1986), it seems reasonable that
381 duckweed might not undergo a single cell phase for prolonged periods of time.
382 Simple models can be helpful in clarifying how multi-cell descent during vegetative
383 reproduction may affect mutation rates and our estimation of them. Assume that clonal
384 reproduction occurs by n parental cells forming a diploid offspring (generation 1). If a
385 mutation occurs in one of these n cells, because of multi-cell descent, the offspring will be a
386 genetic mosaic (i.e., not all cells will be genetically identical). The mutation begins at a
387 frequency of $1/2n$ in the offspring. With conventional single-cell descent, where $n = 1$, a new
388 mutation is expected to be at 50% frequency but the frequency will be lower than 50%
389 whenever $n > 1$. The n cells multiply in some (unknown) growth pattern to produce the
390 mature offspring, which then itself reproduces. In the absence of detailed knowledge of
391 developmental growth trajectories, we do not know what the representation of the original
392 n cell lineages will be in the subsequent generation.
393 We can consider two simple scenarios: If only one of these original cell lineages gives
394 rise to the next set of n cells used to produce the next generation (generation 2), then the
395 mutation will either be completely lost from the lineage (if it was not in the chosen cell
396 lineage) or it will be present in heterozygous state in all cells of future generations (if it was
397 present in the chosen cell lineage). In this first scenario, the genetic mosaicism resulting
398 from multi-cell descent persists only a single generation. Thus, in an experiment such as
399 ours, only mutations occurring in the very last generation will be affected by this issue (i.e.,
400 multi-cell descent is a trivial issue in this case). A second scenario is that all cell lineages
401 grow at equal rates and, each generation, n cells are chosen at random to form the next

402 daughter. This is conceptually similar to the process of coalescence in a population (of cells
403 here rather than individuals) of size *n*. Thinking backwards in time, all cells must eventually
404 trace their coalescent history to a single cellular ancestor, but it may take many generations
405 for this to happen (i.e., 2*n* generations, on average, but with variance proportional to n^2).
406 Thinking forwards in time, a mutation occurring in one of the *n* cells forming the generation
407 1 offspring may be present for many future generations, possibly becoming quite common
408 within individuals, before eventually being present in all future cells (with chance 1/*n*) or
409 none of them (with chance (*n* - 1)/*n*). With even modest values of *n* (e.g., *n* = 8-80), genetic
410 mosaicism can persist for many generations. Mutations that will eventually be present in all
411 lineages—as well as those that will eventually be eliminated—will thus be found below the
412 50% frequency expected in a "traditional" (i.e., non-mosaic) heterozygote. Although this
413 second scenario as formulated here is unrealistically simple (e.g., random and independent
414 choice of *n* cells for each generation), it illustrates how multi-cell descent can have
415 consequences across multiple generations. Though developmental growth trajectories in
416 duckweed are insufficient to formulate a more realistic model, we suspect that multiple cell
417 lineages persist across multiple generations in duckweed and is responsible for the clear
418 bias towards mutant SNPs being less than 50%. Recent work on segregating mutations in a
419 single *Zostera marina* seagrass clone has made similar arguments to this model (Yu et al.,
420 2020). In their study, Yu et al. uncovered a large class of reference biased SNPs that were
421 present in some but not all sampled *Z. marina* ramets. By reconstructing the genealogical
422 relationship of the sampled ramets, the authors demonstrated that such SNPs changed in
423 frequency during vegetative growth, with some reaching heterozygous fixation in specific
424 ramets. The authors argued that this data was well explained by a model of "somatic drift"
425 whereby *de novo* mutations arise at low frequency within a single cell lineage before
426 ultimately either reaching fixation or being lost.
427 Calculating mutation rates in organisms with multiple segregating cell lineages poses
428 a technical challenge due to the difficulty of assessing power when mutations are present in
429 only a subset of the cells of an individual. We implemented a method that takes into
430 account the fraction of reads we expect to support a *de novo* mutation by using observed
431 mapping patterns of putatively *de novo* mutations, similar to the approach used by Hiltunen
432 et al. (2019). This method considers the fact that we have reduced power to detect recently
433 arisen mutations that are at low frequency within their MA lines, giving a more accurate

434 estimate of the SNP mutation rate. This approach is an admittedly crude attempt to address
435 the problem. Rather than assuming the expected frequency of the mutant allele is 50%, we
436 simply choose a lower value based on the observed coverage at putative *de novo* mutations.
437 In reality, this lower value is unknown and will differ among mutations depending upon
438 when each mutation arose. Nonetheless, our approach is an improvement over completely
439 ignoring the issue. Moreover, the variation in mutation rate estimates inferred using
440 different values for the assumed expected mutant frequency provides some sense of the
441 sensitivity of these estimates to this assumption.

442 A natural consequence of the lack of a single cell phase is that from a population
443 genetics perspective, the mutation rate becomes a harder parameter to interpret. On one
444 hand, we may be interested in calculating the mutation rate that captures every new
445 mutation that has arisen in a clonal bud. Alternatively, from an evolutionary perspective, we
446 might be only be interested in mutations that will not only arise in a clonal bud but also
447 persist in a future clonal descendant such that they contribute to population level genetic
448 diversity. Aside from random inter-cell lineage "drift" one possible force that may bias which
449 mutations are eventually lost before fixing in a clonal lineage is cell lineage selection which
450 may purge deleterious *de novo* mutations (Otto & Orive, 1995). In principle, we might be
451 able to calculate an "evolutionarily relevant" mutation rate by performing long term
452 mutation accumulation experiments allowing the majority of *de novo* mutations to either be
453 lost or to have fixed within a clonal lineage such that all cells in any clone will be either fully
454 homozygous (in the case of loss) or heterozygous (in the case of fixation). However, from a
455 practical standpoint, it is hard to know *a priori* how many generations will be necessary for
456 this process to occur. In practice, we could attempt to estimate the frequency of *de novo*
457 mutations that have become sufficiently common such that they are likely to not be lost
458 before fixation leading to an estimate of an evolutionarily relevant mutation rate. For
459 example, we could assume that mutations found in at least 50% of cells (i.e. at a frequency
460 of at least 25% in a clonal bud) are more likely than not to eventually fix in their clonal
461 lineage, being found in 100% of cells in some future generation. In reality, some of these
462 mutations are still likely to be lost before they have a chance to fix while simultaneously
463 some mutations that are below 50% frequency might still reach fixation in the future. More
464 generally, the lower the frequency cut-off we use for mutations that are "likely" to fix, the

465 better the chance that we capture all of the mutations that will reach fixation with the
466 caveat that we will also be capturing more mutations that will eventually be lost.
467 Our estimate of the mutation rate in *S. polyrhiza* (9.20E-11 per bp per gen.) is similar
468 to the estimates reported by Xu et al. (7.92E-11 per bp per gen.), however there are several
469 important differences between our studies. First, our MA experiment was conducted only in
470 the lab (but under two growth conditions), while Xu et al. placed MA lines both in the lab
471 and in the field. While we observed multiple mutations in our lab MA lines, Xu et al. report
472 no *de novo* mutations in their lab MA lines, likely due to the lower number of MA
473 generations in their study. Thus, despite the similarity in our mutation rate estimates, they
474 come from very different environmental settings. Another critical difference is that Xu et al.
475 used ancestral heterozygous sites to estimate their power to detect *de novo* mutations
476 which are not strongly reference biased. This suggests that the estimate from Xu et al.,
477 while conducted in a more naturally realistic environment is likely an underestimate of the
478 mutation rate in the field.
479 The estimates of the mutation rate in *S. polyrhiza* and *L. minor* are among the lowest
480 mutation rates observed so far for multicellular eukaryotes, comparable to rates observed
481 in unicellular eukaryotes (Lynch *et al.*, 2016). A low mutation rate is consistent with efficient
482 selection against mutator alleles in highly asexual organisms such as duckweed (Kimura,
483 1967; Leigh, 1970). Our observation that the mutation rate in *L. minor* shows a trend
484 towards being marginally higher (though non-significantly) than that in *S. polyrhiza* is also
485 consistent with the role of linkage in favouring the evolution of lower mutation rates
486 (Kimura, 1967; Leigh, 1970) as genomic analyses and field observations suggest that *L. minor*
487 outcrosses more frequently than *S. polyrhiza* (Vasseur *et al.*, 1993; Ho, 2018; Xu *et al.*, 2019;
488 Ho *et al.*, 2019). This is in contrast to the theoretical prediction that strong linkage in
489 asexual genomes can allow mutators to fix in asexual populations if they hitchhike to
490 fixation with beneficial mutations they produce (André & Godelle, 2006). Population
491 genomic analyses in duckweed have shown that selection on protein coding genes is weak
492 as evidenced by elevated measures of π_N/π_S (Ho, 2018; Xu *et al.*, 2019; Ho *et al.*, 2019). This
493 suggests that strongly beneficial mutations might be too rare to allow mutators to be
494 selected for in duckweed.
495 Xu et al. inferred a global N_e for *S. polyrhiza* of ~1x10E-06 so it might also be possible
496 that the large effective population sizes of these species allow selection to be efficient

enough to lower the mutation rate more than in most multicellular eukaryotes (Sung *et al.*, 2012; Lynch *et al.*, 2016). This explanation, however, is also inconsistent with the fact that measures for the efficacy of selection suggest that selection is weak in duckweed, likely due to the predominance of asexual reproduction (Xu *et al.*, 2019; Ho *et al.*, 2019).

In conclusion, we report a very low SNP mutation rate in two species of duckweed consistent with previous results in this group. We found that *de novo* mutations appear at low frequencies within MA lines suggesting the presence of multiple segregating cell lineages. We then used an approach that allows us to estimate the mutation rate when multiple cell lineages are transmitted across generations. The low mutation rate of these duckweeds is consistent with the idea that a higher degree of asexual reproduction leads to strong selection for low mutation rates.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (AFA and SIW). We thank Mitch Cruzan for helpful suggestions with data analysis. We thank Jade Lavallee and Victor Mollov for help with maintaining MA lines.

Author contributions

SIW and AFA conceived the project. AFA, SIW, and MB collected samples. MB performed the MA experiment. GS did bioinformatics analyses. AFA, SIW, GS wrote the manuscript. All authors contributed to manuscript editing. SIW and AFA contributed equally to this work.

520 **Literature Cited:**

521 **André J-B, Godelle B. 2006.** The Evolution of Mutation Rate in Finite Asexual Populations.
522 *Genetics* **172**: 611–626.

523 **Appenroth K-J, Teller S, Horn M. 1996.** Photophysiology of turion formation and
524 germination in *Spirodela polyrhiza*. *Biologia Plantarum* **38**: 95–106.

525 **Bell G. 1982.** *The Masterpiece of Nature: The Evolution and Genetics of Sexuality*. University
526 of California, USA: Academic Press.

527 **Cao HX, Vu GTH, Wang W, Appenroth KJ, Messing J, Schubert I. 2016.** The map-based
528 genome sequence of *Spirodela polyrhiza* aligned with its chromosomes, a reference for
529 karyotype evolution. *New Phytologist* **209**: 354–363.

530 **Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson**
531 **M, Lander ES, Getz G. 2013.** Sensitive detection of somatic point mutations in impure and
532 heterogeneous cancer samples. *Nature Biotechnology* **31**: 213–219.

533 **Cole CT, Voskuil MI. 1996.** Population genetic structure in duckweed (*Lemna minor*,
534 Lemnaceae). *Canadian Journal of Botany* **74**: 222–230.

535 **Dorai-Raj S. 2014.** *binom: Binomial Confidence Intervals for Several Parameterizations. R*
536 *package*. URL https://cran.r-project.org/web/packages/binom/index.html. [accessed
537 October 5 2018]

538 **Eyre-Walker A, Keightley PD. 2007.** The distribution of fitness effects of new mutations.
539 *Nature Reviews Genetics* **8**: 610–618.

540 **Foster PL. 2007.** Stress-induced mutagenesis in bacteria. *Critical Reviews in Biochemistry*
541 *and Molecular Biology* **42**: 373–397.

542 **Hicks LE. 1932.** Flower Production in the Lemnaceae. **32**: 115–132.

543 **Hiltunen M, Grudzinska-Sterno M, Wallerman O, Ryberg M, Johannesson H. 2019.**
544 Maintenance of High Genome Integrity over Vegetative Growth in the Fairy-Ring Mushroom
545 *Marasmius oreades*. *Current Biology* **29**: 2758–2765.e6.

546 **Ho KHE. 2018.** *The Effects of Asexuality and Selfing on Genetic Diversity, the Efficacy of*
547 *Selection and Species Persistence.* PhD thesis, University of Toronto, Toronto, ON, Canada.

548 **Ho E, Bartkowska M, Wright SI, Agrawal A. 2019.** Population genomics of the facultatively
549 asexual duckweed *Spirodela polyrhiza*. *New Phytologist* **224:** 1361-1371.

550 **Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP. 2014.** Environmentally
551 responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and
552 epimutations. *Genome Research* **24**: 1821–1829.

553 **Kimura M. 1967.** On the evolutionary adjustment of spontaneous mutation rates. *Genetics*
554 *Research* **9**: 23–34.

Landolt E. 1986. *Biosystematic investigations in the family of duckweeds (Lemnaceae)*. *Vol.2: The family of Lemnaceae: a monographic study*. Zurich, Switzerland. Geobotanischen Instutites der ETH.

Leigh EG. **1970**. Natural Selection and Mutability. *The American Naturalist* **104**: 301–305.

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

Li H, Durbin R. **2009**. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL. **2016**. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics* **17**: 704–714.

Matsuba C, Ostrow DG, Salomon MP, Tolani A, Baer CF. **2013**. Temperature, stress,and spontaneous mutation in *Caenorhabditis briggsae* and *Caenorhabditis elegans*. *Biology Letters* **9**: 20120334.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. **2010**. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* **327**: 92–94.

Otto SP, Orive ME. **1995**. Evolutionary Consequences of Mutation and Selection within an Individual. *Genetics* **141**: 1173–1187.

Plomion C, Aury J-M, Amselem J, Leroy T, Murat F, Duplessis S, Faye S, Francillonne N, Labadie K, Provost GL, *et al.* 2018. Oak genome reveals facets of long lifespan. *Nature Plants* **4**: 440–452.

R Core Team. **2019**. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.r-project.org/ [Accessed September 1 2017]

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. **2011**. Integrative Genomics Viewer. *Nature biotechnology* **29**: 24–26.

Saxena AS, Salomon MP, Matsuba C, Yeh S-D, Baer CF. **2019**. Evolution of the Mutational Process under Relaxed Selection in *Caenorhabditis elegans*. *Molecular Biology and Evolution* **36**: 239–251.

590 **Schmid-Siegert E, Sarkar N, Iseli C, Calderon S, Gouhier-Darimont C, Chrast J, Cattaneo P,**
591 **Schütz F, Farinelli L, Pagni M, *et al.* 2017**. Low number of fixed somatic mutations in a long-
592 lived oak tree. *Nature Plants* **3**: 926–929.

593 **Sharp NP, Agrawal AF**. **2016**. Low Genetic Quality Alters Key Dimensions of the Mutational
594 Spectrum. *PLoS Biology* **14**.

595 **Sturtevant AH**. **1937**. Essays on Evolution. I. On the Effects of Selection on Mutation Rate.
596 *The Quarterly Review of Biology* **12**: 464–467.

597 **Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M**. **2012**. Drift-barrier hypothesis and
598 mutation-rate evolution. *Proceedings of the National Academy of Sciences* **109**: 18488–
599 18492.

600 **Thomas GWC, Wang RJ, Puri A, Harris RA, Raveendran M, Hughes DST, Murali SC, Williams**
601 **LE, Doddapaneni H, Muzny DM, *et al.* 2018**. Reproductive Longevity Predicts Mutation
602 Rates in Primates. *Current biology* **28**: 3193–3197.e5.

603 **Van Hoeck A, Horemans N, Monsieurs P, Cao HX, Vandenhove H, Blust R**. **2015**. The first
604 draft genome of the aquatic model plant Lemna minor opens the route for future stress
605 physiology research and biotechnological applications. *Biotechnology for Biofuels* **8**.

606 **Vasseur L, Aarssen LW, Bennett T**. **1993**. Allozymic Variation in Local Apomictic Populations
607 of *Lemna minor* (Lemnaceae). *American Journal of Botany* **80**: 974–979.

608 **Wang L, Ji Y, Hu Y, Hu H, Jia X, Jiang M, Zhang X, Zhao L, Zhang Y, Jia Y, *et al.* 2019**. The
609 architecture of intra-organism mutation rate variation in plants. *PLoS biology* **17**: e3000191.

610 **Watson JM, Platzer A, Kazda A, Akimcheva S, Valuchova S, Nizhynska V, Nordborg M, Riha**
611 **K**. **2016**. Germline replications and somatic mutation accumulation are independent of
612 vegetative life span in Arabidopsis. *Proceedings of the National Academy of Sciences of the*
613 *United States of America* **113**: 12226–12231.

614 **Xu S, Stapley J, Gablenz S, Boyer J, Appenroth KJ, Sree KS, Gershenzon J, Widmer A, Huber**
615 **M**. **2019**. Low genetic variation is associated with low mutation rate in the giant duckweed.
616 *Nature Communications* **10**: 1243.

617 **Yu L, Boström C, Franzenburg S, Bayer T, Dagan T, Reusch TBH. 2020.** Somatic genetic drift
618 and multilevel selection in a clonal seagrass. *Nature Ecology and Evolution*. *In press.*
619
620 **Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS**. **2012**. A high-performance
621 computing toolset for relatedness and principal component analysis of SNP data.
622 *Bioinformatics* **28**: 3326–3328.

623

624

625 **Supporting Information**

626

627 **Fig. S1** Visualization of a low-quality genomic region that was filtered from our analyses.

628

629 **Fig. S2** Allelic coverage at heterozygous sites before and after filtering.

630

631 **Fig. S3** Visualization of a validated *de novo* mutation from both Illumina short read and
632 Sanger sequencing data.

633

634 **Fig. S4** Visualisation of a putative *de novo* mutation that was identified as a likely sequencing
635 error

636

637 **Fig. S5** Visualisation of a putative *de novo* mutation that was identified as a likely genome
638 misassembly error

639

640 **Table S1** List of *de novo* mutations identified in two species of duckweed

641

642 **Table S2** Mutation rate estimates with power adjusted to reflect the observed proportion of
643 reads that support *de novo* mutations

644

645 **Table S3** Mutation rate estimates with power adjusted under several scenarios of allelic bias
646 at *de novo* mutant sites in salt stressed mutation accumulation lines

647

648 **Table S4** Mutation rate estimates with power adjusted under several scenarios of allelic bias
649 at *de novo* mutant sites in control mutation accumulation lines

650

651 **Table S5** Number and fraction of callable sites under several scenarios of allelic bias at *de*
652 *novo* mutant sites

653

654