

1 **Title: Intraspecific Variation in Microsatellite Mutation Profiles in *Daphnia magna***

2 **Authors:** Eddie K. H. Ho¹, Fenner Macrae¹, Leigh C. Latta IV^{1,2}, Maia J. Benner¹,

3 Cheng Sun³, Dieter Ebert⁴, and Sarah Schaack^{1*}

4

5 **Author affiliations:**

6 ¹Department of Biology, Reed College, Portland, OR 97202

7 ²Division of Natural Sciences and Mathematics, Lewis-Clark State College, Lewiston,

8 ID, USA 83501 (current address)

9 ³Institute of Apicultural Research, Chinese Academy of Agricultural Sciences, Beijing

10 100093, China

11 ⁴University of Basel, Department of Environmental Sciences, Zoology, 4051 Basel,

12 Switzerland

13

14 ***Corresponding Author email:** schaack@reed.edu

15

16 **Running title:** Microsatellite mutations in *Daphnia magna*

17

18 **Keywords:** tandem repeats, mutation rate variation, mutation accumulation, waterfleas,

19 Cladocera

20

21

22 **Abstract:** Microsatellite loci (tandem repeats of short nucleotide motifs) are highly
23 abundant in eukaryotic genomes and are often used as genetic markers because they
24 can exhibit variation both within and between populations. Although widely recognized
25 for their mutability and utility, the mutation rates of microsatellites have only been
26 empirically estimated in a few species and have rarely been compared across
27 genotypes and populations and intraspecific differences in overall microsatellite content
28 have rarely been explored. To investigate the accumulation of microsatellite DNA over
29 long- and short-time periods, we quantified the abundance and genome-wide mutation
30 rates in whole-genome sequences of 47 mutation accumulation (MA) lines and 12 non-
31 MA lines derived from six different genotypes of the crustacean *Daphnia magna*
32 collected from three populations (Finland, Germany, and Israel). Each genotype
33 possessed a distinctive microsatellite profile and clustered according to their population
34 of origin. During the period of mutation accumulation, we observed very high
35 microsatellite mutation rates (a net change of -0.19 to 0.33 per copy per generation),
36 which surpass rates reported from a closely-related congener, *D. pulex*, by an order of
37 magnitude. Rates vary between microsatellite motifs and among genotypes, with those
38 starting with high microsatellite content exhibiting greater losses and those with low
39 microsatellite content exhibiting greater gains. Our results show that microsatellite
40 mutation rates depend both on characteristics of the microsatellites and the genomic
41 background. These context-dependent mutation dynamics may, in conjunction with
42 other evolutionary forces that may differ among populations, explain the differential
43 accumulation of repeat content in the genome over long time periods.
44

45 **Introduction**

46 Microsatellite loci, also known as short tandem repeats, are repetitive regions of the
47 genome known for their propensity to mutate rapidly (e.g., Sun et al., 2012). Although
48 exact definitions of microsatellites vary, they typically involve tandem arrays of short
49 motifs (typically, 1-6 bp long, although longer motifs can also be found in tandem
50 arrays). Microsatellites can be located inside or outside coding regions of the genome,
51 and have been shown to influence a range of phenotypes from gene expression to
52 genetic disease (Feupe Fotsing et al., 2018). Previous reports of microsatellite mutation
53 rates (MMRs) have consistently shown them to be higher than substitution rates in
54 unique sequence, often by several orders of magnitude (reviewed in Ellegren, 2004).
55 Because of their mutability, microsatellites have frequently been used in population
56 genetics studies and there is increasing interest in the role they may play in adaptation,
57 plasticity, and disease (Haasl and Payseur, 2013; Hannan, 2018).

58

59 There are three mechanisms of mutation that have been proposed to explain the
60 patterns of higher mutation rates at microsatellite loci: retrotransposition, unequal
61 crossing over, and DNA slippage. Retrotransposition, in particular, could explain the
62 frequent observation that microsatellites tend to be A-rich, although it is less clear how
63 retrotransposition would impact mutation rates of microsatellites once they are formed.
64 Unequal crossing over is thought to increase in frequency at repeat-rich loci and can,
65 potentially, lead to the expansion or contraction of tandem arrays with equal probability.
66 The most often discussed mechanisms of microsatellite mutation is strand slippage
67 during DNA replication and repair (Kornberg et al., 1964), whereby the array of repeats

68 can cause potential mispairing between template and nascent strands of DNA. If
69 uncorrected by DNA repair mechanisms, slippage can lead to the expansion or
70 contraction of a tandem array and may do so in a motif- or length-dependent manner
71 (Eckert and Hile, 2009). When substitutions occur at microsatellite repeats, they result
72 in the loss (or 'death') of the repeat, in addition to loss or contraction due to deletions or
73 contractions during slippage (Kelkar et al., 2011). A given microsatellite locus can
74 experience any of a number of different types of mutation (e.g., insertions, deletions,
75 duplications, slippage, and substitutions) which can result in either an expansion or
76 contraction of that tandem array, the interruption of the array, or the increase or
77 decrease in copy number of the array. Because all these mutation types will contribute
78 to overall copy number for any given motif (referred to as a kmer, hereafter), genome-
79 wide analyses of microsatellite mutation rates can benefit from looking at rates of copy
80 number increase and decrease as a global metric of the impact of mutation at these
81 loci.

82
83 Microsatellite mutation rate (MMR) variation based on the composition of the motif
84 (AT/GC content), length of the motif (unit length; e.g., dinucleotide versus trinucleotide
85 repeats), and the length of the array (e.g., the number of repeats occurring in tandem at
86 a given locus) has been the focus of previous studies in a variety of systems (reviewed
87 in Bhargava and Fuentes, 2010). Theoretically, mutation rates would be expected (A)
88 to be higher in AT-rich regions (due to the lower number of hydrogen bonds between
89 base pairs), (B) to decrease as a function of unit length based on the strand slippage
90 model of mutation, and (C) to increase as a function of array length, given the increased

91 number of targets for mutation (reviewed in Eckert and Hile 2009). Indeed, empirical
92 studies have shown that microsatellites with high AT-content tend to mutate at higher
93 rates than those that are GC-rich and that di-nucleotide rates are higher than tri-
94 nucleotide repeats (Chakraborty et al., 1997). Rates of expansion versus contraction,
95 however, have been shown to depend on starting length, with shorter arrays tending to
96 increase in length and longer arrays tending to decrease in length (Lai and Sun, 2003;
97 Seyfert et al., 2008). Indeed, if MMRs vary based on any of these factors, one could
98 make predictions about the accumulation of microsatellites across the genome over
99 long time periods based on starting composition of the repeat content.

100

101 As with most types of mutations, mutation rate estimates are typically performed on one
102 or a few genotypes for a representative model species, and then used to extrapolate
103 mutation rate estimates for congeners, or even more widely, despite a lack of evidence
104 for generalizing to this degree (e.g., mutation rate estimates for *Drosophila*
105 *melanogaster* are routinely used as a proxy for all insects, despite known variation in
106 rates estimates between genotypes (Haag-Liautard et al., 2007)). The degree to which
107 microsatellite mutation rates and patterns of microsatellite accumulation vary between
108 genotypes and populations, intraspecifically, or among closely-related species with
109 similar lifespans, physiologies, and life histories has remained largely unexplored.
110 Given that the rate of mutation itself is a trait that can evolve, knowing the level of
111 intraspecific variation upon which evolutionary forces can act to increase or decrease
112 the rate over time (Lynch, 2010), as well as knowing what factors influence rate
113 differences, is of major interest to biologists (Baer et al., 2007). Most recently, it has

114 been proposed that mutation rates across species hover near a “drift barrier”, meaning
115 that they are only driven down by selection to the extent possible based on the effective
116 population size, at which point they can not be lowered further due to the relative power
117 of genetic drift which permits mutations that increase (or maintain) the rate (Lynch,
118 2010). Knowing the level of intraspecific variation in mutation rates is essential for
119 assessing the potential of a drift barrier to explain mutation rate variation within and
120 between species. Mutation accumulation (MA) experiments provide the least biased
121 estimates of mutation rates available (Halligan and Keightley, 2008), although they can
122 only be conducted in organisms that can be reared in a controlled environment with
123 short generation times.

124
125 Here, we present data from 6 genotypes of *Daphnia magna*—2 each from three
126 populations (Finland, Germany and Israel), and compare our results to previously
127 published estimates of MMR in the congener, *D. pulex* (Flynn et al., 2017). *D. magna* is
128 an important model organism for ecology, evolutionary biology, and genomics studies
129 (Miner Brooks E. et al., 2012; Schaack, 2008) . The cyclical parthenogenetic nature of
130 *Daphnia* makes them an ideal organism to use in MA experiments because clonal
131 reproduction facilitates their long-term maintenance in the lab. Our goal is to
132 characterize both the microsatellite landscape and mutational profiles across these 6
133 genotypes in order to determine if there is a relationship between the two, and to assess
134 the degree to which they may vary among genotypes, populations, and closely-related
135 species. In addition, we report the microsatellites that are most abundant and most
136 mutable to determine if there are features of individual microsatellites (unit length or

137 content) that determine differences in mutation dynamics among loci. Identifying
138 patterns using mutation accumulation data collected on experimental time-scales where
139 selection is minimized and contrasting such data with patterns of microsatellite
140 accumulation over long time-periods can reveal the degree to which evolutionary forces
141 are shaping microsatellite landscapes in nature.

142

143 **Methods**

144 *Study System*

145 The *D. magna* genotypes used in this experiment were collected along a latitudinal
146 gradient that captures a range of environmental variation including temperature and
147 photoperiod. Specifically, two unique genotypes from each of three populations
148 (Finland, Germany, and Israel) were used to initiate the control and mutant lines. The
149 stock cultures for each genotype were maintained in 250 mL beakers containing 175-
150 200 mL of Aachener Daphnien Medium (ADaM; Klüttgen et al., 1994) under a constant
151 photoperiod (16L:8D) and temperature (18 °C), and fed the unicellular green alga
152 *Scenedesmus obliquus* ad libitum (2-3 times per week).

153

154 Two types of controls were used in this experiment. First, we established and
155 maintained populations of large size from descendants of the same genotypes used to
156 initiate the MA lines. In these large population controls, mutations may occur, but are
157 more likely to be purged by selection due to competition among clones (relative to the
158 MA lines, where clones are reared individually and experience no competition). At the
159 end of the mutation accumulation period, these large populations are sampled and DNA

160 is extracted (referred to hereafter as ‘extant controls’ or ECs). Although mutations can
161 occur in these lineages, the paucity of mutations observed is consistent with the idea
162 that the MA protocol (bottlenecking lineages each generation by transferring a single
163 individual) does, indeed, minimize selection (see below). The second set of control
164 lines sequenced was from tissue harvested from immediate descendants of the
165 individual from which progeny were used to establish the MA lines (referred to hereafter
166 as ‘starting controls’ or SCs) at the beginning of the mutation accumulation period.

167

168 *Mutation Accumulation Experiment*

169 Starting control (SC), extant control (EC), and mutation accumulation (MA) lines were
170 initiated from clonally-produced offspring of a single asexual female isolated from the
171 stock cultures of each of the 6 genotypes described above. Tissue samples for SCs
172 consisted of 5-20 individual *Daphnia* collected from each genotype beginning two weeks
173 after initiation of the MA experiment. Individuals were placed in a 1.5 mL
174 microcentrifuge tube, frozen in liquid nitrogen, and stored at -80 °C until DNA extraction.
175 Two EC lines were initiated from each SC, and were sampled at the end of the
176 experiment (approximately 2.5 years after initiation of the experiment). Two EC lines
177 were maintained in separate 3 L jars containing 2 L of ADaM, under constant
178 temperature (18 °C) and photoperiod (16L:8D), and fed the unicellular green alga *S.*
179 *obliquus* ad libitum. The media in the jars was replaced every 2-3 weeks, and
180 individuals in the two replicate jars were mixed to maintain as much genetic
181 homogeneity among the jars as possible. The maintenance of EC lines in large jars
182 ensures that population densities, which varied between several hundred to a few

183 thousand individuals, were high enough that new mutations with deleterious effects
184 should be efficiently eliminated from the populations by purifying selection.

185

186 In addition to the control lines, between 10-15 MA lines were initiated from each of the 6
187 starting genotypes. The MA protocol used here has been described previously (Eberle
188 et al., 2018). Briefly, MA lines were initiated by placing a single clonally-produced
189 female in a 250 mL beaker containing 100 mL of ADaM supplemented with *S. obliquus*
190 at a concentration of 600,000 cells/mL. All MA lines were maintained in environmental
191 conditions identical to the control lines (16L:8D, 18 °C). The food/media mixture in each
192 beaker was replaced once per week, and each line was fed a prescribed volume of
193 concentrated *S. obliquus* three days after the media replacement to reset the algal cell
194 concentration in the beaker to 600,000 cells/mL. From generation to generation, each
195 MA line was propagated via single progeny descent by taking a single juvenile offspring
196 from the second clutch of the previous generation. A series of backups were
197 maintained in parallel with the focal lineages in the event that the single individual
198 intended to be used to establish the next generation died before reproduction, or was a
199 male. In the event that the focal lineage and all backup lineages died before
200 reproduction or were all males, the lineage was declared extinct and a new MA line was
201 established from the ongoing EC lines. Tissue samples for each of the MA lines were
202 isolated every 5 generations, and at the end of the MA experiment the samples taken
203 from lines with the greatest number of generations of mutation accumulation were used
204 for DNA extraction and sequencing (2-26 generations, with an average of 12.4
205 generations per line).

206

207 *DNA Extraction and Sequencing*

208 Five clonal individuals from each MA line and controls (1 starting control [SC] and 2
209 extant controls [EC] per genotype) were flash frozen for DNA extractions. DNA was
210 extracted (2 extractions per line with 5 daphnia each) using the Zymo Quick-DNA
211 Universal Solid Tissue Prep Kit (No. D4069) following the manufacturer's protocol (DNA
212 from a few samples was also extracted with the Qiagen DNeasy Blood and Tissue Kit,
213 No. 69504). DNA quality was assessed by electrophoresis on 3% agarose gels and
214 DNA concentration was determined by dsDNA HS Qubit Assay (Molecular Probes by
215 Life Technologies, No. Q32851). The Center for Genome Research and Biocomputing
216 at Oregon State University generated 94 Wafergen DNA 150bp paired-end libraries
217 using the Biosystems Apollo 324 NGS library prep system. Quality was assessed using
218 a Bioanalyzer 2100 (Agilent Technologies, No. G2939BA) and libraries were pooled
219 based on qPCR concentrations across 16 lanes (2 runs). Libraries were sequenced on
220 an Illumina Hiseq 3000 (150 bp PE reads) with an average insert size of ~380bp to
221 generate approximately 50x coverage genome-wide for each sample (Table S7).

222

223 *Tandem repeat quantification*

224 Sequenced reads from all lines were trimmed of adapters and decontaminated to
225 remove mitochondrial sequences. Overlapping reads were merged with BBmerge
226 (Bushnell et al., 2017). To quantify tandem repeats, the reads were input into the
227 program k-seek (Wei et al., 2014). The program k-seek detects tandem repeats (kmers)
228 of 1-20 bp, requiring that the kmers repeat tandemly over at least 50 bp on a given read,

229 allowing for one base pair mismatch per repeat unit. Offsets and reverse complements
230 of each kmer are combined and the output is the total count across all reads for each
231 kmer. Because k-seek has the requirement that tandem repeats span at least 50 bp,
232 the threshold number of repeat units required for detection decreases as the length of
233 the repeat unit increases (e.g., 1-mers require at least 50 repeats to be detected while
234 5-mers only require a minimum of 10 repeats to be detected). On the other hand, the
235 maximum number of repeat units on a single read decreases as the kmer length
236 increases. We do not observe a detection bias towards kmers with longer or shorter
237 lengths (Table 2), which suggest that kmer length is not the main determinant of kmer
238 detection.

239
240 To compare across samples, we normalized kmer counts by dividing copy number
241 counts by the median sequence depth matched by the GC-content of the kmer. We first
242 constructed *de novo* reference genomes for each of the six *D. magna* starting
243 genotypes using Spades (Bankevich et al., 2012). Reads from each line were mapped
244 to their corresponding reference genome using BWA default settings (Li and Durbin,
245 2009). Following Flynn et al. (2017), output BAM files were input into a custom script to
246 calculate the coverage depth at each base and the GC-content of their nearby region
247 (<https://github.com/jmf422/Daphnia-MA-lines>). We then group each base pair based on
248 their nearby GC-content in the following bins: {0-0.3, 0.3-0.35, 0.35-0.4, 0.4-0.45, 0.45-
249 0.5, 0.5-0.55, 0.55-0.6, 0.6-1}; we used wider bins for GC-content ≤ 0.3 and > 0.6
250 because there were many fewer sites containing very low and high GC-content. For
251 each GC-content bin, we then determined the median base pair depth for use as the

252 normalization factor. For each line, we normalize the total count of each kmer by
253 dividing the total count by the normalization factor that corresponds with the GC-content
254 of that kmer. This normalization approximates the copy number per 1x coverage of each
255 kmer, which for simplicity we will refer to as the 'copy number'.

256

257 Note that, after normalization, the total base pairs covered by a particular kmer in a
258 particular line (i.e., kmer length * normalized copy number) can fall below 50 bp even
259 though k-seek required tandem repeats span at least 50 bp. This is due to an over-
260 correction by our normalization method (e.g., for a 1-mer that spans exactly 50 bp in the
261 genome, there may be many reads that encompass the whole 50 bp array, but also
262 many reads that only encompass a portion of the array). In this case, k-seek will not
263 count the number of 1-mer repeats in reads that do not contain the full 50 bp array,
264 because it falls below its threshold array length requirement. However, those reads are
265 still counted towards our normalization factor. Due to this, the normalized copy number
266 can fall below 50 for these 1-mers. This over-correction due to normalizing total copy
267 number by the average (or median) coverage is present in all previous analyses that
268 utilized k-seek (Flynn et al., 2017, 2018; Wei et al., 2014). Overall, this would cause an
269 underestimation of the total kmer content in the genome.

270

271 *Mutation rate estimation*

272 k-seeks outputs the total count of a given kmer across all locations in the genome as
273 long as the requirements mentioned above are met. Thus, for our estimation of mutation
274 rates, we define mutation as the change in the total copy number of a kmer, which could

275 have occurred at one or more locations in the genome. For each genotype, we restrict
276 our mutation rate analysis to kmers where the SC line had at least six copies and each
277 of the MA lines had at least two copies. This allowed us to estimate mutation rates for
278 71 kmers, on average, for each of the six genotypes. We define the *genomic mutation*
279 *rate* of kmer j in MA line i as $U_{i,j} = (c_{i,j} - c_{SC,j})/g_i$, where $c_{i,j}$ and $c_{SC,j}$ represents the copy
280 number of kmer j at MA line i and the SC line, respectively, and g_i represent the number
281 of MA generations for MA line i . We found that the absolute value of the genomic
282 mutation rate was strongly correlated with the abundance of the kmer in the SC lines
283 (Figure S1). This was not surprising because highly abundant kmers likely represent a
284 larger mutational target. To account for differences in the initial abundance of kmers, we
285 define the *per copy mutation rate* of kmer j in MA line i as $u_{i,j} = U_{i,j} / c_{SC,j}$. We define the
286 overall genomic and per copy mutation for kmer j of a genotype as U_j and u_j ,
287 respectively, which is calculated by taking the average $U_{i,j}$ and $u_{i,j}$ across all MA lines of
288 the genotype. We calculated mutation rates for EC lines in the same way as we did for
289 MA lines. To estimate the number of generations that each of the EC lines were
290 maintained, we divided the length of the experiment (830 days) by their estimated
291 generation time.

292

293 *Comparison to D. pulex*

294 Throughout our study, we compare our *D. magna* microsatellite results to previously
295 published results based on a dataset from *D. pulex* MA lines (Flynn et al., 2017). Briefly,
296 Flynn et al. (2017) examined the microsatellite content from 28 MA lines and six non-
297 MA lines that were all initially generated from a single ancestral genotype. Next

298 generation sequencing was done following a Illumina Nextera library preparation (10x
299 coverage, 100 bp PE reads). They analyzed kmers copies using k-seek and normalized
300 copy number estimates as we described above (we used their study as a guide for our
301 copy number normalization steps). In addition to shorter read lengths and lower
302 coverage depths of sequencing, another difference in the study is the controls: they did
303 not sequence their initial ancestral genotype, but used the average copy number of the
304 six non-MA lines as a proxy for the ancestral state.

305

306 *Data Availability Statement*

307 The authors affirm that all data necessary for confirming the conclusions of the article
308 are present within the article, figures, and tables and that all sequence data generated
309 will be submitted to GenBank upon acceptance of the article.

310

311 **Results**

312

313 *Microsatellite copy number profiles in D. magna*

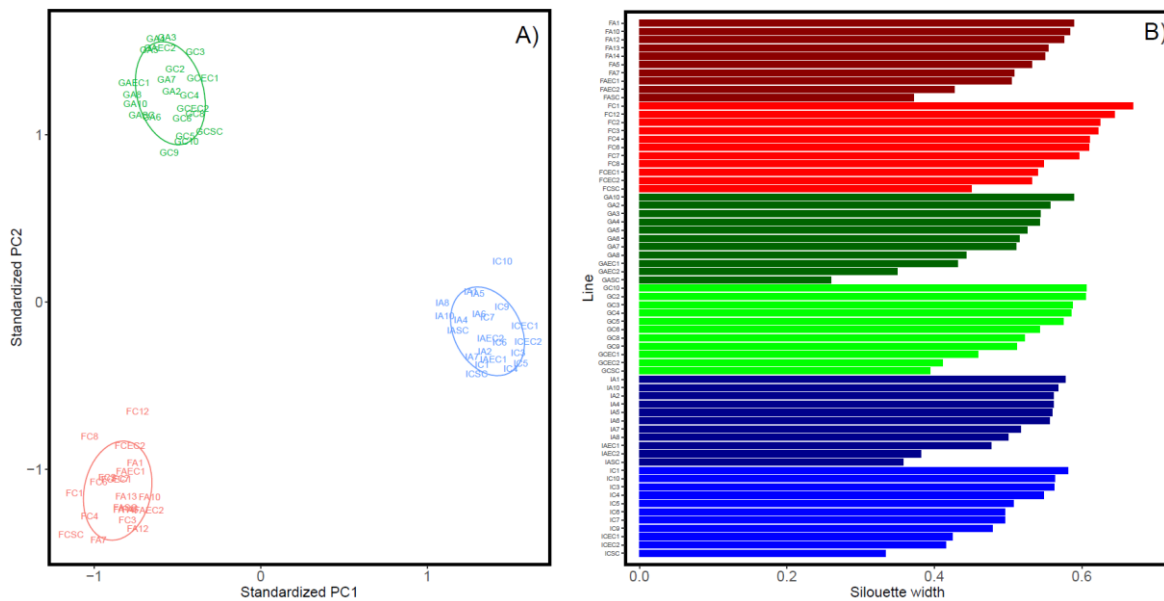
314 We scanned for kmers of lengths up to 20 bp across the genome of individuals
315 sequenced from 47 mutation accumulation lines (MA) derived from 6 starting genotypes
316 (“starting controls” [SC]) and 12 lines (2 per starting genotype) maintained in large
317 population sizes in parallel with the MA lines sampled at the end of the experimental
318 period (“extant controls” [EC]). After normalization by depth of coverage, the total
319 number of base pairs (per 1x coverage) composed of kmers ranged from 97 to 145 Kb
320 across our six *D. magna* starting genotypes, which represented 0.07 to 0.1% of the 141

321 Mb *D. magna* reference genome (Figure S2). Across all SC, EC and MA lines, the
322 median number of base pairs composed of kmers was 121 kb (0.085% of the genome).
323 In contrast, the median kmer content of *D. pulex* was 1.2 MB (0.6% of the estimated
324 200 Mb *D. pulex* genome), which is an order of magnitude higher than in *D. magna*
325 (Flynn et al., 2017). The kmer content in our *D. magna* lines was more similar to that in
326 *Chlamydomonas reinhardtii*, which contains an average of 180 Kb (0.15% of the
327 genome) (Flynn et al., 2018).

328
329 We performed a principal components analysis (PCA) using the copy number of the 100
330 kmers (average repeat unit length of 10.9) with non-zero copy numbers across all 65
331 lines in order to look for distinctive patterns of the microsatellite landscape across the 6
332 genotypes. On the first and second principle components axes, the lines clearly
333 clustered based on their population of origin (Figure 1). We additionally performed a k-
334 medoids analysis using the first 10 principle components (these 10 PCs explained 83%
335 of the variation in copy number). We found that six clusters maximized the average
336 silhouette across lines. Each of these six clusters contained the SC line, all of its
337 descendent MA lines, and the EC of that genotype (Figure 1). Overall, the kmer copy
338 number profiles distinguished lines based on their population and genotype.

339
340 Our PCA results are conservative because we only examine kmers shared across all
341 lines (including the kmers unique to each population would only increase the degree of
342 clustering observed). We observed 92, 91 and 127 kmers, respectively, that only exist in
343 the lines from Finland, Germany and Israel. Unsurprisingly, the average repeat unit

344 length of these population-specific kmers were 13.8, 14.1 and 12.8 for Finland,
 345 Germany, and Israel, respectively, which are larger than the average of the 100 shared
 346 kmers. The vast majority of population-specific kmers are low in abundance, with an
 347 average copy number below 25. The exceptions are AATAGC and ACTCCT with
 348 average copy numbers of 130 in IA and 87 in IC, respectively (but which are each still
 349 present but rare in the other genotype from that region).
 350



351
 352 **Figure 1.** Population structure using the 100 kmers with non-zero copy number across
 353 all 65 lines. (A) Each line is plotted based on the first and second principle components
 354 axis. Lines from Finland, Germany and Israel are coloured red, green and blue,
 355 respectively. (B) k-medoids analysis using the first 10 PCs of the principal components
 356 analysis. Each cluster only contained one starting genotype (SC) and all of its
 357 descendant MA and EC lines. Dark red, red, dark green, green, dark blue and blue
 358 represents lines from genotypes FA, FC, GA, GC, IA, IC, respectively.

359

360 We found a range of 104 to 148 kmers that appeared at least twice in all SC, EC and
361 MA lines of a particular genotype and observed 283 unique kmers across all genotypes.
362 There were 13 highly abundant kmers with an average copy number ≥ 100 across the
363 SC lines of all genotypes (Table 1) which ranged in length from 1 to 6 bp. In contrast, *D.*
364 *pulex* has 39 repeats with an average copy number ≥ 100 (Flynn et al., 2017) and these
365 kmers ranged in length from 1 to 20 bp. Of the highly abundant *D. magna* kmers, 12 out
366 of 13 exist in the *D. pulex* dataset, while only 19 of the 39 highly abundant *D. pulex*
367 kmers exist in our *D. magna* dataset. In both species, the most abundant kmer was the
368 1-mer A and followed by the 5-mer AACCT, but the copy number was much higher in *D.*
369 *pulex* for both (Table 1). As noted previously, AACCT is likely an ancestral telomeric
370 repeat in Arthropods that is present in several crustaceans (*D. pulicaria*, *Gammarus*
371 *pulex* and *Penaeus semisulcatus*) and insects (Okazaki et al., 1993; Sahara et al.;
372 Schumpert et al., 2015). On average, 30% of the total kmer base pairs were composed
373 of AACCT in *D. magna* and 26% in *D. pulex* (Flynn et al., 2017). Kmers C, AAC and
374 AAG had similar copy numbers between the two species, while the remaining eight high
375 abundance kmers in *D. magna* had lower copy numbers or were absent in *D. pulex*
376 (Table 1).

377

378

379 **Table 1.** Normalized copy number of highly abundant kmers for each SC line from six
 380 genotypes of *D. magna* collected from three locations, Finland (F), Germany (G) and Israel (I).

kmer	FA	FC	GA	GC	IA	IC	<i>D. pulex</i>*
A	21847	22096	19529	22181	12965	18481	85440
AACCT	9767	8083	4399	6347	6799	11390	48905
AAG	7325	6333	5047	8817	5178	7292	1893
C	7246	4892	2661	5796	2495	2870	4982
AG	3900	2528	4231	3001	3179	4687	376
ACTAT	2136	2462	1569	1678	1316	1956	-
AAAAC	988	1376	627	1236	1296	676	8
AC	622	567	604	624	948	971	269
AAC	285	278	324	360	354	440	279
AGC	177	168	209	180	167	150	55
AAT	190	150	188	169	134	142	4
AT	151	159	113	165	149	134	2
AACAGG	53	105	171	110	128	232	31

381 *Copy number for *D. pulex* represent the mean across their six non-MA lines

382

383 For the kmers with at least two copies across all lines of a genotype, the distribution of

384 repeat unit lengths was similar across the six genotypes (Table 2). Kmers with short

385 lengths tend to have higher copy number than longer kmers. We observed an
 386 abundance of kmers with lengths divisible by three (i.e. 3-, 6-, 9-, 12-, 15- and 18-mers)
 387 and an abundance of 5-mers. Kmers with lengths 5, 6, 12 and 15 were also very
 388 common in *D. pulex*. However, *D. pulex* contained an abundance of 10- and 20-mers
 389 (15 and 50, respectively), which we did not observe in *D. magna*.

390

391 **Table 2.** Count and average copy number for kmers of different lengths (k) found in each
 392 genotype of *D. magna* assayed in this experiment (also with *D. pulex* data from Flynn et al.
 393 (2017)).

<i>D. magna</i>							<i>D. pulex</i>		
k	# kmers						Mean copy number	# kmers	Mean copy number
	FA	FC	GA	GC	IA	IC		All lines	
1	2	2	2	2	2	2	13578	2	42954
2	3	3	3	3	3	3	1361	2	326
3	9	9	8	9	9	9	668	5	2159
4	8	6	3	3	5	6	31	4	267
5	11	13	9	10	12	10	938	10	8953
6	14	15	8	18	17	19	31	10	148
7	3	1	1	3	5	5	45	1	189
8	2	2	2	3	1	0	7	3	202

9	3	3	6	9	9	8	10	4	11
10	5	4	2	3	5	5	9	15	875
11	6	3	3	2	2	3	21	1	10
12	25	21	17	22	24	19	6	19	128
13	4	4	2	3	8	7	12	5	156
14	2	2	1	1	3	2	6	3	16
15	26	24	16	16	19	21	6	13	15
16	1	1	1	1	0	0	6	1	7
17	1	1	2	2	0	0	4	6	175
18	16	18	16	16	18	16	6	6	17
19	5	5	1	1	2	2	5	2	16
20	2	2	1	1	2	2	5	50	439

394 *Mean copy number was averaged across all lines for all kmers of repeat unit length k

395

396 *Microsatellite mutation rate profiles of D. magna*

397 For each of the six genotypes, we estimated mutation rates for kmers with at least six

398 copies in the SC line and at least two copies in each of the EC and MA lines. The

399 number of kmers with estimated mutation rates ranged from 60 to 79 across the six

400 genotypes, which totaled to 144 unique kmers. Across all six genotypes, 31 kmers

401 were present in all populations, while 20, 22 and 29 kmers were unique to genotypes of

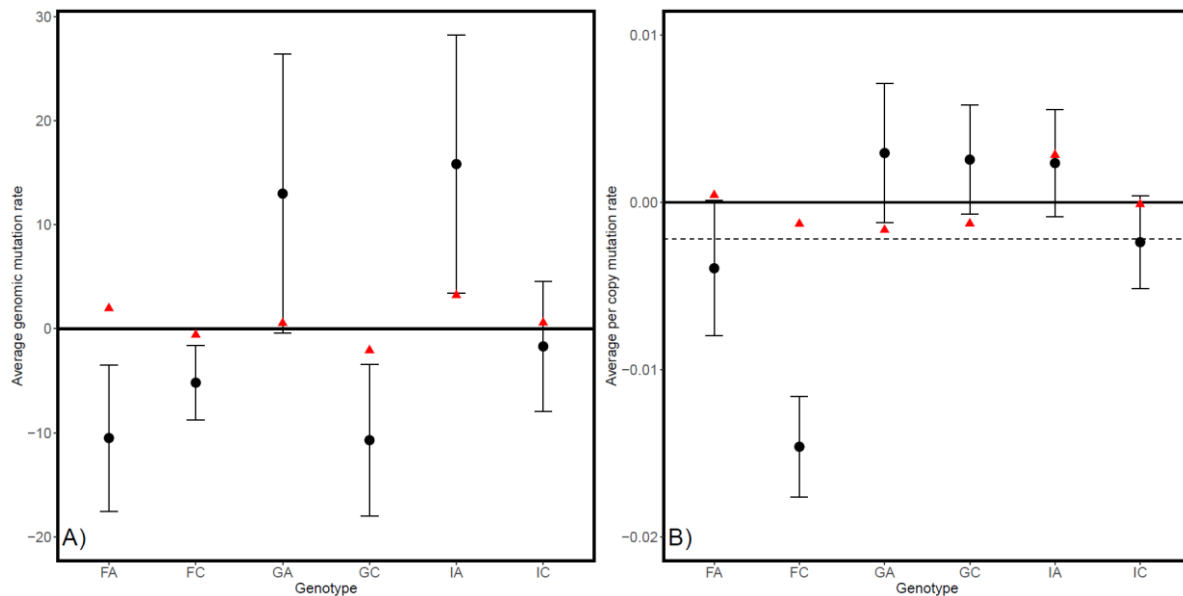
402 Finland, Germany and Israel, respectively. We observed that the absolute value of the

403 mutation rate, $|U_{i,j}|$ was strongly positively correlated with the initial copy number of the
404 kmer in the SC line for each genotype (average correlation = 0.75, Figure S1). This was
405 expected because kmers with higher representation in the genome likely represents a
406 larger mutational target. To remove this correlation, we divided the mutation rate of
407 each kmer by its initial abundance to obtain an estimate of the per copy mutation rate,
408 $u_{i,j}$ (average correlation = 0.0031, Figure S1). It is important to note, the program we
409 used (kseek) estimates the copy number across all arrays of a particular kmer and thus
410 our mutation rates is an estimate of the net change in copy number due to increases
411 and decreases at all arrays, rather than an estimate of array length changes (see
412 Methods for details). Thus, a positive (negative) value for the genome wide or per copy
413 mutation rate does not mean that the particular kmer only experienced increases or
414 expansions (decreases or contractions) in copy number, rather it means that the net
415 effect of mutation was to increase (decrease) copy number.

416
417 We observed high levels of variation in microsatellite mutation rates for *D. magna*,
418 ranging from negative (net decrease in copy number for a given kmer) to positive (net
419 increase in copy number for a given kmer), even among lines from the same genotypes.
420 Across all MA lines and kmers, the genome-wide mutation rate ($U_{i,j}$) ranged between -
421 1103 to 2370 copies per generation and the per copy mutation rate ($u_{i,j}$) ranged
422 between -0.19 to 0.33 copies per initial copy number per generation. We found that
423 mutation rates varied considerably across the six genotypes, but not consistently
424 between genotypes of the same population. Figure 2 shows the kmer mutation rates
425 (both U_j , u_j) averaged across all kmers for each genotype. For both U_j and u_j , FA, FC

426 and IC had negative mutation rates (meaning a decrease in copy number), while GA
427 and IA had positive mutation rates, on average. GC had negative U_j , but positive u_j , on
428 average, which indicates that there was one or more kmers that possessed low
429 negative per copy mutation rates (u_j), but were abundant enough to cause the average
430 of U_j to be negative (which weights u_j by the abundance of kmer j).

431



432

433 **Figure 2.** Mean (+/- SE) genomic mutation rate (A) and per copy mutation rate (B) for each
434 genotype from six genotypes of *D. magna* collected from three locations, Finland (F), Germany
435 (G) and Israel (I). Black circles and red triangles represent MA and EC lines, respectively. The
436 dashed line represents the mutation rate of MA lines averaged across all genotypes.

437

438 We used the absolute value of the per copy mutation rates ($|u_{i,j}|$) to examine the
439 magnitude of kmer copy number change. Across all MA lines and kmers, the average
440 absolute per copy mutation rate ranged from 0.0000042 to 0.33 and had a mean of
441 0.029. EC lines had a lower rate of kmer copy number change with an average 0.0049,
442 suggesting that selection indeed constrained the rate of kmer copy number change in
443 these large population controls.

444

445 Overall, kmer mutations rates were higher in magnitude and more variable in *D. magna*
446 than in *D. pulex*. To compare to the *D. pulex* dataset, we applied a similar filter
447 (requiring that each kmer considered have at least two copies in all MA lines and at
448 least six copies across all non-MA lines (as a proxy for the ancestor)). For the 121
449 kmers for which we were able to estimate absolute per copy mutation rates, $|u_{i,j}|$, the
450 mean was 0.004 copies per copy per generation (ranging from 0 to 0.053), which is an
451 order of magnitude lower than the average rate for *D. magna* MA lines. Only 21 of the
452 121 *D. pulex* kmers were present in *D. magna* (Figure S3), and the average $|u_{i,j}|$ for
453 these 21 kmers in *D. pulex* and *D. magna* was 0.0041 and 0.031 copies per copy per
454 generation, respectively. Furthermore, the coefficients of variation in $|u_{i,j}|$ for these 21
455 kmers were consistently lower in *D. pulex* than in the *D. magna* genotypes (Figure S3).

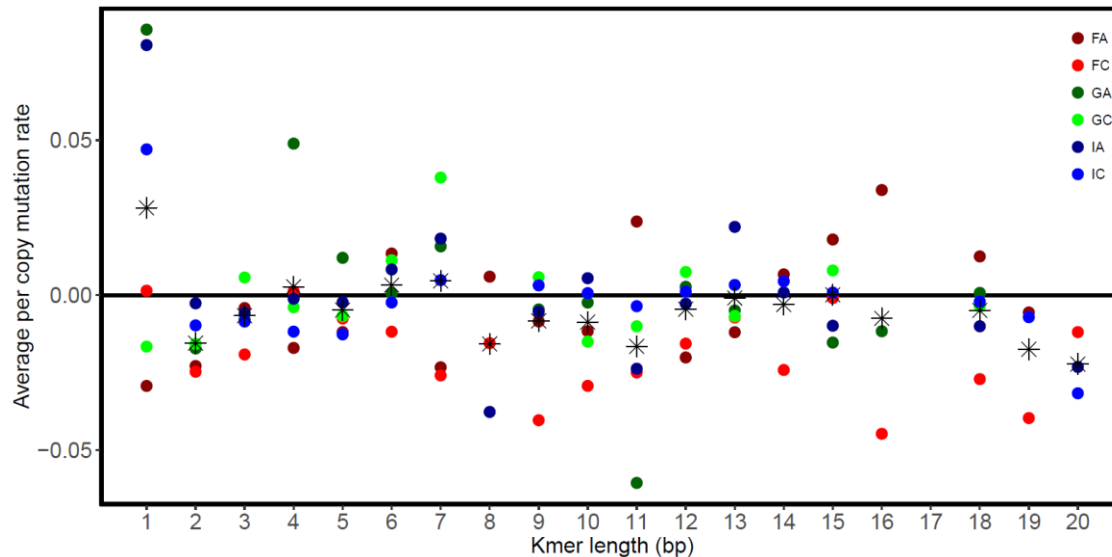
456

457 *Mutation rate variation based on features of the kmer*

458 Per copy mutation rates of individual kmers, u_j , varied greatly between kmers of different
459 lengths (Figure 3, Figure S4). Figure 3 shows the average value of u_j across kmers of
460 the same length (k) for each genotype, which we define as $u_j(k)$. This value $[u_j(k)]$ can
461 be positive or negative at most kmer lengths, depending on the particular genotype. Per
462 copy mutation rates were most positive in 1-mers and tend to be more negative in
463 kmers with $k \geq 8$. However, fitting a linear model ($\text{lm}(u_j \sim k)$), for each genotype did not
464 show that per copy mutation rate was significantly correlated with kmer length (Table
465 S1, Figure S4), likely because of the considerable variation in mutation rates even

466 among kmers of the same length. Indeed, Kruskal-Wallis tests across kmers of the
467 same length show significant variation in u_j within genotypes for most kmer lengths
468 (Table S2, Figure S4), in addition to the variation in mutation rates at each kmer length
469 observed between genotypes illustrated in Figure 3.

470



471
472 **Figure 3.** Means of kmer per copy mutation rate for each genotype and length of kmer from six
473 genotypes of *D. magna* collected from three locations, Finland (F), Germany (G) and Israel (I).
474 The asterisk symbol represents the average value across the six genotypes.

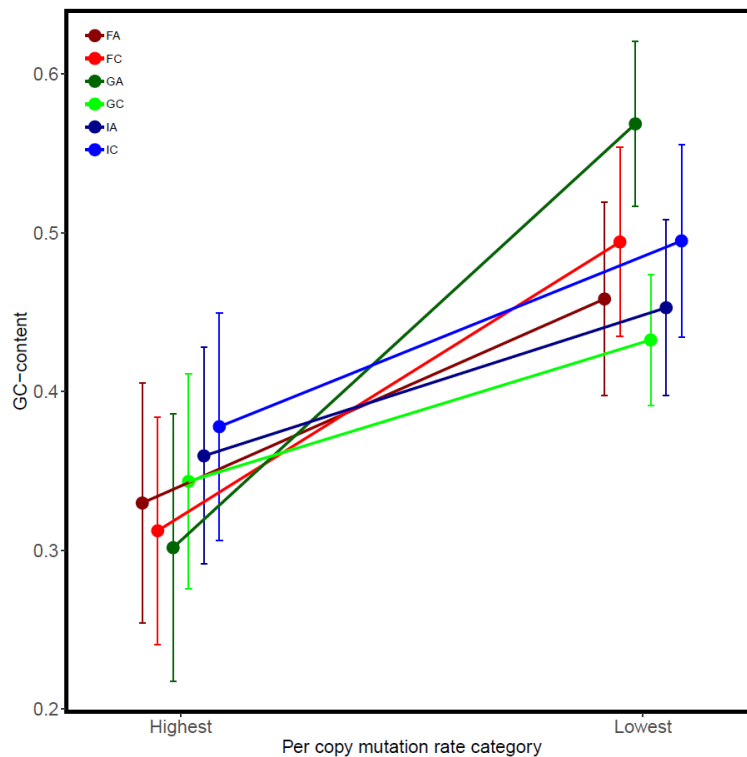
475

476 We also examined whether GC-content may have affected mutation rates. Kmers
477 containing higher levels of GC may have a lower propensity to undergo mutation
478 because GC pairs forms a more stable bond than AT pairs. To measure the propensity
479 for mutation, we calculated the absolute values of per copy mutation rates ($|u_j|$), which
480 combines the rates of kmer copy increases and decreases, for kmers with repeat unit
481 lengths longer than three base pairs. For each genotype, we examined the GC-content
482 of the ten kmers with the highest and the ten with the lowest absolute per copy rates
483 ($|u_j|$; Figure 4, Table S5). We excluded kmers less than three base pairs long because
484 these kmers will have extreme values of GC-content; including the 1-mers and 2-mers

485 did qualitatively change our results. We observed that average $|u_j|$ for the kmers with the
486 highest rates were at least four times higher than kmers with the lowest rates (Table
487 S4). As predicted, the kmers with lower $|u_j|$ tend to possess higher GC-content across
488 all genotypes. A two-way ANOVA for GC-content with genotype and mutation rate
489 category (high vs low) as factors revealed that GC-content significantly differed between
490 kmers with the highest and lowest mutation rates (Table S6).

491

492



493

494 **Figure 4.** GC-content (mean +/- SE) of kmers with the top highest and lowest absolute per copy
495 mutation rates, $|u_j|$, from six genotypes of *D. magna* collected from three locations, Finland (F),
496 Germany (G) and Israel (I).

497

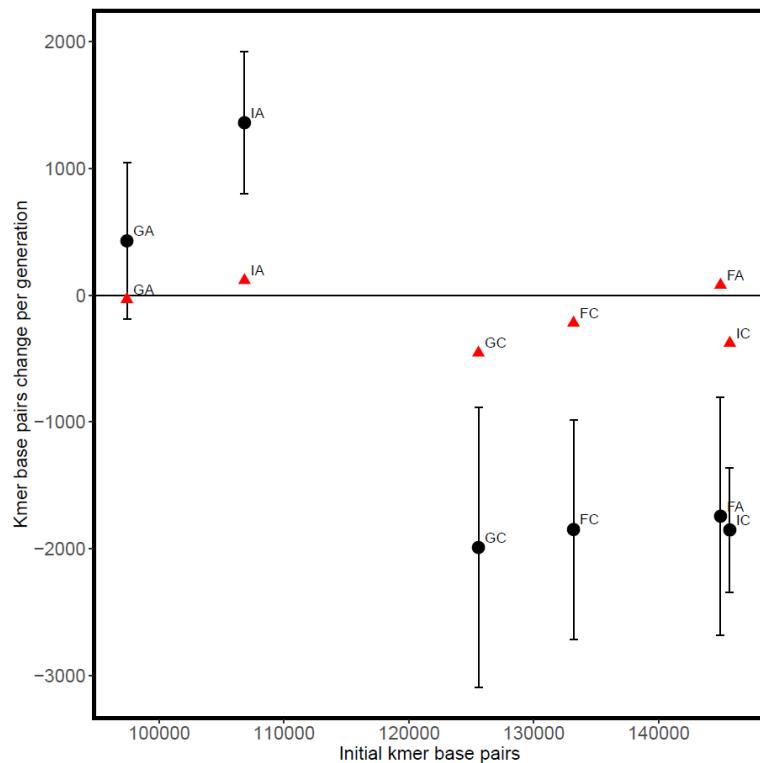
498 *Linking variation in microsatellite landscapes and microsatellite mutation rates*

499 The total amount of change in kmer content during mutation accumulation is not

500 consistent within or between populations (Figure 5). Both genotypes from Finland (FA,

501 FC) experienced a reduction in kmer content per generation, on average, while one
502 genotype from Germany and Israel (GA and IA) experienced an increase in kmer
503 content while the other experienced a decrease (GC and IC). Since the MA lines of GA
504 and IA experienced the greatest increases in kmer base pairs, on average, we expected
505 these two genotypes would contain the highest amount of kmer content, overall, but the
506 opposite is true (Figure 5). The SC lines of GA and IA contain the lowest kmer content
507 initially, but exhibit the highest rates of increase due to mutation. In contrast, the SC
508 lines of FA, FC, GC and IA contained the highest kmer content and showed the greatest
509 declines in kmer content during mutation accumulation. We tested if the change in GC-
510 content of the microsatellite portion of the genome also varied based on starting GC-
511 content level, but observed no significant relationship (Figure S5). Thus, instead of
512 exhibiting strong differences based on population of origin (Figure S7) or features of
513 individual kmers, the major differences in mutation rate profiles appear to depend on
514 features of the genome-wide kmer content, overall.

515

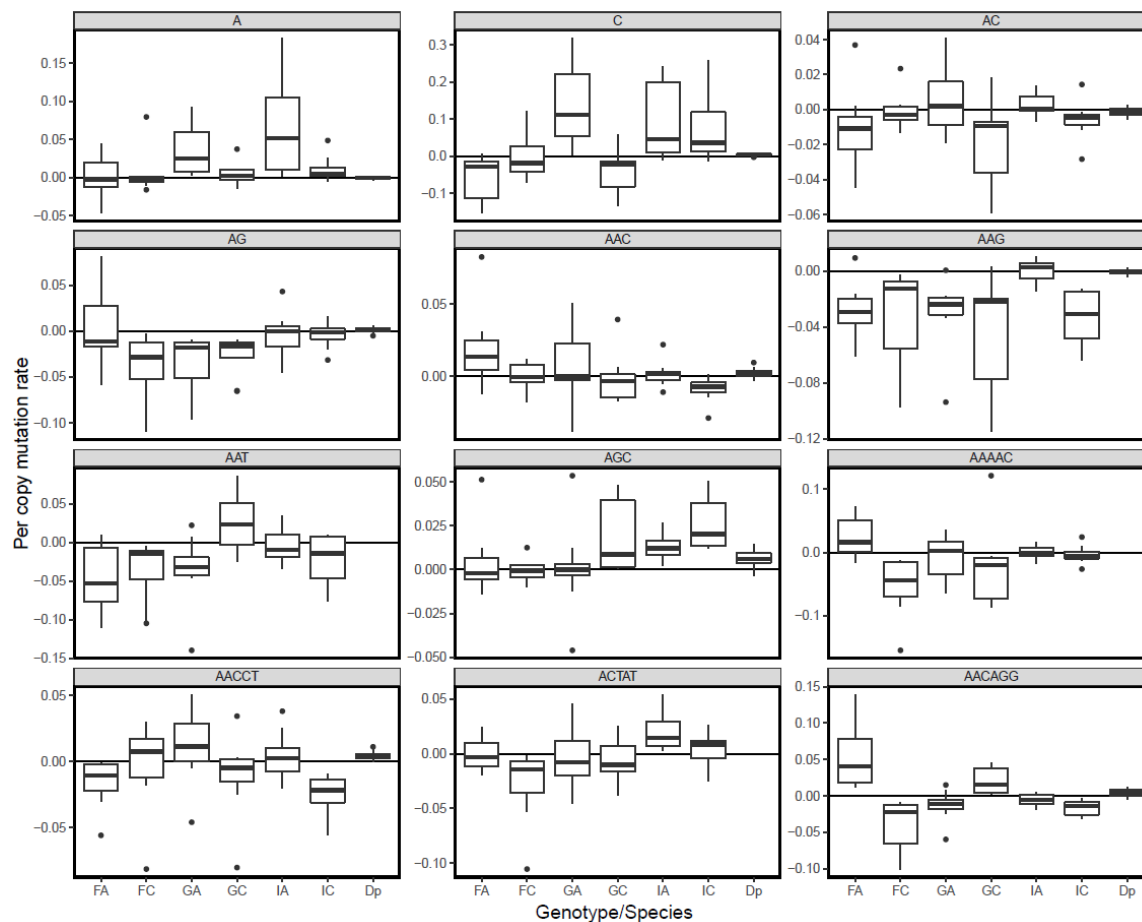


516
517 **Figure 5.** Mean (+/- SE) kmer base pair change per generation for six genotypes of *D. magna*
518 collected from three locations, Finland (F), Germany (G) and Israel (I). Black circles and red
519 triangles represent MA and EC lines, respectively.

520
521 As alluded to previously, average kmer mutation rates ranged from positive to negative
522 and varied between genotypes without being consistent within populations (Figure 2).
523 We can examine this in more detail by focusing on the 31 kmers with mutation rate
524 estimates across all six genotypes. Kruskal-Wallis tests show significant variation in u_j
525 across genotypes for all but two of the kmers (Figure 6, Table S3). If kmer mutation
526 profiles were similar within populations, we would expect high correlations in u_j for
527 genotypes from the same population, however this was not observed (Table S6). IA and
528 IC possessed a relatively strong positive correlation (0.67) in u_j , but they also shared a
529 strong correlation with GA (IA-GA: 0.60, IC-GA: 0.63). Furthermore, this correlation was
530 mainly driven by their shared high positive u_j for the kmer C (Figure 6). Removing the
531 kmer C reduced the pairwise correlations (IA-IC: 0.33, IA-GA: 0.47, IC-GA: 0.40). We

532 also performed a principal components analysis (PCA) using the u_j of the 31 shared
533 kmers and did not find evidence of clustering by population-of-origin based on principle
534 components one and two (Figure S6). Including all 144 kmers (i.e., genotype-specific
535 kmers) in the PCA improved clustering slightly (Figure S7).

536



537
538 **Figure 6.** Per copy mutation rate for the 12 kmers with the highest copy number across for six
539 genotypes of *D. magna* collected from three locations, Finland (F), Germany (G) and Israel (I).
540 Dp represents the per copy mutation rate for *D. pulex*. Points indicate lines with per copy
541 mutation rates less than $Q1-1.5*IQR$ or larger than $Q3+1.5*IQR$. Q1, Q3 and IQR represents
542 the first quartile, third quartile and the interquartile range, respectively.

543

544

545 **Discussion**

546 Repetitive regions of the genome, once overlooked, are now known to be a large and
547 dynamic component of the genome, often responsible for large proportions of the
548 genetic variation among individuals and species. Microsatellite loci, in particular, are
549 known to exhibit elevated mutation rates compared to unique sequences, and have
550 been shown to be important components of the genome in a variety of functional
551 contexts ranging from disease risk to speciation (Gemayel et al., 2012; Hannan, 2018;
552 Shah et al., 2010). The goal of this study was to quantify intraspecific and interspecific
553 variation in the microsatellite landscape and microsatellite mutational dynamics in
554 *Daphnia*. To do this, we analyzed kmers with unit lengths of 1-20 bp residing within
555 arrays spanning at least 50 bp for six *D. magna* starting control lines (SC), 47 mutation
556 accumulation lines (MA) and 12 non-MA existing control lines (EC). We were able to
557 characterize the kmer profile of our *D. magna* lines based on 283 kmers, and were able
558 to estimate mutation rates for 144 kmers. Our analysis differs from the many previous
559 studies that examine microsatellite mutation rates because we use total kmer counts
560 across all loci spanning at least 50 base pairs, rather than examining individual
561 microsatellite arrays independently. Using a mutation accumulation experiment and
562 genome-wide approach, our estimates of kmer mutation rates provide a lower-bound
563 estimate of the net change in kmer copy number due to mutations of all types across all
564 microsatellite loci.

565

566 *Microsatellite landscapes are distinct among genotypes, populations, and congeners*

567 Our results clearly show distinctive microsatellite landscapes among the 6 genotypes
568 from the three different populations of *D. magna* samples (Figure 1A and B). Using the

569 abundance of 100 kmers presents in all lines, we observed clustering of genotypes by
570 their population of origin (Figure 1A). In addition, within each population, we observed
571 that MA and EC lines formed distinct clusters based on the genotype (Figure 2). These
572 results are conservative and including kmers unique to genotypes would only
573 strengthen the clustering. For the kmers with presence/absence polymorphism across
574 genotypes, the vast majority were low in copy number suggesting that they arose
575 relatively recently. In contrast, microsatellite analysis for 6 genotypes of
576 *Chlamydomonas reinhardtii* found many kmers with hundreds of copies in some
577 genotypes but absent or rare in others (Flynn et al., 2018). Our analyses reveal that the
578 kmer content of *D. magna* is highly dynamic and can cause high levels of intraspecific
579 variation, even within populations.

580
581 The microsatellite profile of *D. magna* is distinct from that of the only previously
582 examined congener, *D. pulex* (Flynn et al., 2017), which has a much higher proportion
583 of microsatellite content in its genome (Tables 1 and 2). In *D. pulex*, the most abundant
584 kmers occurring in the genome tend to be shorter repeat units, with the exception of
585 some longer repeats, such as the known arthropod telomeric sequence (AACCT)_n
586 (Okazaki et al., 1993). However, there are many kmers that are unique to each species
587 and, for kmers that are shared, many differ greatly in copy number (Table 1, 2). *D.*
588 *magna* is enriched for kmers with unit lengths that are multiples of three (Table 2). It is
589 possible that these kmer lengths are more tolerated by selection because they are less
590 likely to cause frameshift mutations within coding regions (Metzgar et al., 2000). *D.*
591 *pulex*, on the other hand, is enriched for kmers with unit lengths that are multiples of

592 five, as has also been reported in *Drosophila melanogaster* (Wei et al., 2014). The
593 distinctive microsatellite landscapes observed both within and between these species
594 invites the question—do microsatellite mutation dynamics vary widely and thus explain
595 the accumulated differences observed among genotypes, populations and species over
596 long time periods?

597

598 *Microsatellite mutation rates vary among genotypes and between species*

599 Mutation rates (both genome-wide increases and decreases in copy number across
600 kmers [U_{ij}] and per copy adjusted rates [u_{ij}]) vary widely among genotypes (Figure 2).
601 For per copy adjusted rates, the two genotypes collected from Finland exhibit declines
602 in average copy number with mutation accumulation, while those from Germany exhibit
603 increases, and the two genotypes collected from Israel split, with one genotype having
604 an overall positive per copy mutation rate and one having a negative rate. This level of
605 intraspecific variation in rates has not been reported previously, although this could be
606 an artifact of most studies being conducted on only one or a few genotypes based on
607 the assumption that mutation rate estimates can be generalized across closely-related
608 species. One of the major take-home messages of this study is that intraspecific
609 variation in microsatellite mutation rates is substantial, with some genotypes
610 experiencing increases in kmer copy number and others exhibiting a decrease in kmer
611 copies, overall. Across all kmers and genotypes, on average, copy number change was
612 more often negative than positive (Figure 3) in *D. magna*, which is the opposite of the
613 pattern observed in *D. pulex* (Figure S8).

614

615 *Microsatellite mutation rates as a function of kmer length and kmer content*

616 We examined the variation in kmer abundance based on features of the kmers
617 themselves—both length and GC-content. There was no relationship between length
618 and copy number change (Figure 3), with one major exception-- 1mers exhibit the
619 highest positive mutation rate, on average. We observed that kmers with high GC-
620 content tend to have lower mutation rates (Figure 4 and Supplemental Table S4). This
621 is not a surprise in that the three hydrogen bonds holding GC pairs together might be
622 less prone to mutation than regions that are AT-rich, given there are only two hydrogen
623 bonds between As and Ts (Calabrese and Durrett, 2003; Fan and Chu, 2007).

624

625 *The relationship between microsatellite landscape and mutation rates*

626 We explored the relationship between initial kmer content in the genome and the
627 mutation profiles for each genotype to determine if the microsatellite landscape could be
628 explained by the patterns of mutation accumulation observed in the laboratory. Given
629 that we observed a strong positive correlation between microsatellite abundance and
630 absolute mutation rates (leading to the calculation of per copy mutation rates for our
631 subsequent analyses), we were surprised to find that genotypes with high initial
632 genome-wide kmer content exhibit greater decreases in microsatellite content (total
633 change in bp) as a result of mutation than genotypes with low initial kmer content, which
634 exhibit greater increases in the bp contributed by microsatellites during mutation
635 accumulation (Figure 5). The context-dependency of microsatellite mutation dynamics
636 have been reported previously, for example previous studies have shown that longer
637 arrays tend to decrease in length whereas those with shorter arrays tend to increase

638 (Lai and Sun, 2003; Xu et al., 2000). However, the dependency of a mutation bias
639 towards increasing or decreasing microsatellite content on the initial total amount of
640 microsatellites DNA has not yet been reported to our knowledge.

641
642 If starting microsatellite content does, indeed, determine the direction of copy number
643 change as observed here *within* a species, we would predict that the extremely high
644 microsatellite content (10-fold higher than in *D. magna*) reported for *D. pulex* in Flynn et
645 al. (2017) would correspond with declines in copy number during mutation
646 accumulation. This is, in fact, the opposite of what was reported-- *D. pulex*, overall,
647 shows a bias towards copy number increases (Flynn et al., 2017), while *D. magna*
648 shows an overall bias towards decreasing copy number (illustrated by the asterisks in
649 Figure 3; Figure S8). This observation, combined with the observation that *D. magna*
650 exhibit a ten-fold higher overall rate of microsatellite mutation presents a genomic
651 puzzle. It is possible that the copy number increase bias, combined with lower mutation
652 rates, has led to a slow but tolerable accumulation of higher kmer content in the *D.*
653 *pulex* genome over time. A similar explanation has been posited for plant versus animal
654 mitochondrial genomes, where low mutation rates and a mutation bias towards
655 insertions may have led to the tolerable accumulation of non-coding DNA resulting in,
656 typically, much larger organellar genomes than in animals (Lynch et al., 2006).

657

658 *Comparison between D. pulex and D. magna*

659 As mentioned, there were a few differences that limit our ability to make a direct
660 comparison between our results for *D. magna* and those in Flynn et al. (2017) for *D.*

661 *pulex* without some caveats (i.e., coverage differences and read length differences).
662 Since kseek only counts tandem repeats spanning at least 50 bp on a read, the shorter
663 reads and lower coverage in Flynn et al. (2017) may make it more difficult to detect
664 kmer copies, especially of longer kmers, in their study. However, kmer content reported
665 in *D. pulex* was actually much higher than in *D. magna* and there was no obvious bias
666 towards detecting shorter kmers (Table 2). In fact, Flynn et al. (2017) detected many
667 more 10-mers and 20-mers in *D. pulex* than we found in *D. magna*. An additional
668 difference was that Flynn et al. (2017) used the average copy number of kmers in non-
669 MA lines as a proxy for kmer estimates for the ancestral line as a baseline for
670 calculating rates. In our dataset, non-MA lines experienced kmer content change at a
671 much slower rate than MA lines, suggesting that while non-MA lines serve as a
672 relatively good proxy for ancestral lines (Figure 5, S8), this could contribute to a slight
673 underestimate of rates. Although we would not expect this to explain the order of
674 magnitude difference in mutation rates between *D. pulex* and *D. magna* (which we
675 observed even when only comparing their 21 shared kmers), and the differences
676 between species in overall kmer content, the differences in the studies likely affect the
677 sensitivity of each analysis.

678

679 *Conclusions*

680 We observe major differences in the microsatellite landscapes accumulated over long
681 time periods between genotypes and populations of *D. magna*, and between this
682 species and the previously studied congener, *D. pulex*. High levels of differentiation in
683 repeat landscapes were also previously reported, among both populations of *Drosophila*

684 *melanogaster* (Wei et al., 2014) and among species of Caenorhabditid worms (Subirana
685 et al., 2015). Given microsatellite landscapes are shaped not only by mutational inputs,
686 but also by selection and drift, this is not a major surprise. Our results beg the question
687 whether mutation rate differences or differential impacts of evolutionary forces play a
688 greater role in explaining these observed differences.

689

690 While we observe high levels of variation in the mutation rates among genotypes and
691 kmers in *D. magna*-- with some exhibiting net increases in copy number and others
692 exhibiting net decreases in copy number-- the variation does not mirror the differences
693 seen in landscapes over long time periods. In fact, genotypes with the lowest kmer
694 content had the highest rates of copy number increase, and vice versa. Thus, it is clear
695 the differences in microsatellite landscapes within *D. magna* are not being driven purely
696 by mutational inputs, but instead likely reflect the interplay of mutation, selection, and
697 drift, potentially resulting in an equilibrium with respect to individual loci (Kruglyak et al.,
698 1998) or overall repeat content in the genome (Petrov, 2002). Overall, genotype- and
699 kmer-specific variation in mutation rates (Figure 6) reveals a large range in terms of
700 mutation rates in this species, and suggests that there is abundant variation upon which
701 natural selection could act to shape mutation rates within *D. magna*.

702

703 In addition to investigating intraspecific variation and the degree to which long-term
704 patterns of mutation accumulation would correspond to short-term mutation rates,
705 another goal of this study was to assess the degree to which mutation rates are
706 consistent between closely-related species. Overall, we observe much higher (10-fold)

707 absolute microsatellite mutation rates in *D. magna* (regardless of bias towards
708 increasing or decreasing kmer copies), than those reported for *D. pulex* (Flynn et al.,
709 2017). Importantly, we see a mutation bias towards decreasing copy number in *D.*
710 *magna* (reflected by the lower overall kmer content in this species), relative to the
711 increase bias reported for *D. pulex*, which corresponds to the much higher level of kmer
712 content reported for that species (Flynn et al., 2017). While it is possible that a higher
713 effective population size (N_e) of *D. pulex* (estimated to be approximately double that of
714 *D. magna* by (Haag et al., 2009)) allows selection to more efficiently lower the mutation
715 rate in this species, it seems unlikely that this difference could explain the order of
716 magnitude difference in mutation rate observed. Alternatively, the bias towards a
717 decrease in copy number observed in *D. magna* may make the high mutation rates
718 more tolerable, even under a similar selective regime, assuming increasing kmer
719 content in the genome is deleterious.

720
721 It will be interesting to investigate other aspects of the mutational profile (e.g., base
722 substitution and indel rates) of these two species in order to determine what other major
723 differences in mutation dynamics are observed. Future studies examining the rates of
724 contraction and expansion across kmers, as well as the differential mutability of
725 microsatellites within, near, or far from protein-coding regions will also yield insights into
726 the intragenomic variability in mutation rates. Although it has been common heretofore
727 to generalize empirical estimates of mutation rates from experiments using a single
728 genotype from a model species, the level of intra- and interspecific variation reported
729 here suggests caution should be taken when doing so. Once we have a more complete

730 picture of the rates, directionality, and consequences of mutation within and among
731 species and across the genome, we will be able to better predict adaptive potential,
732 frequencies of genetic disease, and rates of evolution for individuals and across taxa.

733

734

735 **Acknowledgments**

736 We would like to thank Dee Denver and Dana Howe and the Center for Genome
737 Research and Bioinformatics staff for their invaluable assistance with library prep and
738 sequencing. We would also like to gratefully acknowledge our funding sources, which
739 include an Institutional Development Award (IDeA) from the National Institute of General
740 Medical Sciences of the National Institutes of Health under Grant #P20GM103408 (LL),
741 Reed College sabbatical fellowship (SS), the M.J. Murdock Charitable Trust (SS), and
742 NSF Award MCB-1150213 (SS).

743 **Supplementary Tables**

744 **Table S1.** Effect of repeat unit length (k) on mutation rate from a linear model, $\ln(u_j \sim k)$ for six
745 genotypes of *D. magna* collected from three locations, Finland (F), Germany (G) and Israel (I).

Genotype	D.f.	Coefficient	P-value
FA	70	0.00132	0.111
FC	75	-0.00066	0.232
GA	58	-0.00140	0.079
GC	64	0.00034	0.648
IA	71	-0.00103	0.154
IC	77	0.00002	0.969

746

747

748 **Table S2.** P-value from Kruskal-Wallis tests for the effects of kmer repeat unit length on per
 749 copy mutation rate for each unit length and genotype for six genotypes of *D. magna* collected
 750 from three locations, Finland (F), Germany (G) and Israel (I).

k	FA	FC	GA	GC	IA	IC
1	0.0639	0.3446	0.0587	0.0460	0.6744	0.0587
2	0.0256	0.0221	0.0297	0.1831	0.1725	0.7926
3	< 0.0001	< 0.0001	0.0017	< 0.0001	0.0012	< 0.0001
4	0.0056	0.0140	0.1791	0.0006	0.0002	0.0290
5	0.0027	0.0002	0.4714	0.0367	< 0.0001	0.0030
6	< 0.0001	< 0.0001	0.0032	< 0.0001	< 0.0001	< 0.0001
7	0.0018	-	-	-	0.0005	0.0005
9	-	-	0.0660	0.0159	0.0317	0.4399
10	0.0023	0.0005	0.0117	0.0147	< 0.0001	0.0023
11	0.0254	-	-	-	-	0.0008
12	0.0336	< 0.0001	0.0354	< 0.0001	0.0020	< 0.0001
13	0.9491	-	-	-	0.0823	0.0056
14	-	0.2936	-	-	-	-
15	0.0283	< 0.0001	0.0003	0.3305	0.3570	0.0010
18	0.0649	0.0190	0.0039	0.0003	0.5784	0.0001
19	-	0.2936	-	-	-	-

751 *Only shows Kruskal-Wallis test results when there at least two kmers with length k in a
 752 genotype.

753
 754
 755

756 **Table S3.** Kruskal-Wallis test of the 31 kmers with mutation rate estimates across all six
 757 genotypes of *D. magna* in this study.

kmer	k	Mean copy number	Mean per copy mutation rate	P-value
A	1	22123	0.0213	0.0116
C	1	5001	0.0374	0.0002
AC	2	716	-0.0043	0.0618
AG	2	3235	-0.0179	0.0298
AT	2	120	-0.0239	0.0609
AAC	3	337	0.0028	0.0497
AAG	3	5132	-0.0281	0.0114
AAT	3	137	-0.0194	0.0322
ACG	3	48	0.0095	< 0.0001
AGC	3	190	0.0108	0.0041
ATC	3	21	-0.0206	0.0672
AAAG	4	82	0.0197	0.0358
AAAT	4	19	-0.0029	0.1654
AGAT	4	13	0.0023	0.0001
AAAAC	5	923	-0.0109	0.0039

AACCT	5	7206	-0.0066	0.0167
AAGAT	5	19	-0.0064	0.0011
ACTAT	5	1827	-0.0017	0.0161
AACAGG	6	121	-0.0013	< 0.0001
AACTAC	6	95	-0.0243	0.6727
AAGGCG	6	12	0.0077	0.0479
ATCGCC	6	67	0.0016	< 0.0001
ATATCCC	7	72	0.0016	< 0.0001
AACTGCATC	9	24	-0.0066	< 0.0001
AAATAATAAT	10	9	-0.0367	0.0287
AAGGAGGTAG	10	13	0.0216	0.0251
AAGACTGACTG	11	49	-0.0253	0.0224
ACCACTACTCCG	12	10	0.0263	0.0364
AACTACTATATAG	13	42	0.0057	0.0002
ACCAGCCTACCCCGC	15	29	0.0016	0.0055
ACATCGTCCACGGATCC G	18	8	0.0032	< 0.0001

758 **Mean copy number' represents the mean copy number of the kmer across all SC, EC and MA
759 lines lines. 'P-value, represents the p-value from performing the Kruskal-Wallis test.

760 **Table S4.** Statistics for the kmers with ten highest and ten lowest $|u_j|$ for each of six genotypes
 761 of *D. magna* collected from three locations, Finland (F), Germany (G) and Israel (I).

Genotype	Category	Mean k	Mean GC content	Mean $ u_j $
FA	Highest	7.6	0.33	0.071
FA	Lowest	10.4	0.46	0.013
FC	Highest	8.1	0.31	0.056
FC	Lowest	9.4	0.49	0.012
GA	Highest	7.1	0.30	0.064
GA	Lowest	12.5	0.57	0.015
GC	Highest	6.7	0.34	0.055
GC	Lowest	8.7	0.43	0.012
IA	Highest	7.8	0.36	0.050
IA	Lowest	7.3	0.45	0.009
IC	Highest	8.0	0.38	0.044
IC	Lowest	11.4	0.49	0.010

762 *k represents the kmer length, GC represents the proportion of base pairs that are GC in the
 763 kmer

764
 765
 766
 767
 768
 769
 770

771 **Table S5.** Two-way ANOVA results testing the relationship between genotype and absolute per
 772 copy mutation rate category (high vs low) and GC-content of kmers in *D. magna*.

Factor	D.f.	SumSq	MeanSq	F-value	P-value
Genotype	5	0.043	0.0085	0.202	0.9612
Category	1	0.641	0.6405	15.177	0.0002
Genotype:Category	5	0.115	0.0231	0.547	0.7405
Residuals	108	4.558	0.0422		

773 *Category high and low represents kmers with the 10 highest and 10 lowest absolute
 774 per copy mutation rate, $|u_j|$, respectively.

775

776 **Table S6.** Pairwise correlations of per copy mutation rates among the 31 kmers shared across
 777 the six genotypes of *D. magna* collected from three locations, Finland (F), Germany (G) and
 778 Israel (I).

	FA	FC	GA	GC	IA	IC
FA	1	0.29	0.00	0.17	-0.13	0.08
FC		1	0.45	0.15	0.41	0.43
GA			1	-0.15	0.60	0.63
GC				1	-0.07	0.01
IA					1	0.67
IC						1

779

780

781

782

783 **Table S7. Total number of reads for each line**

784

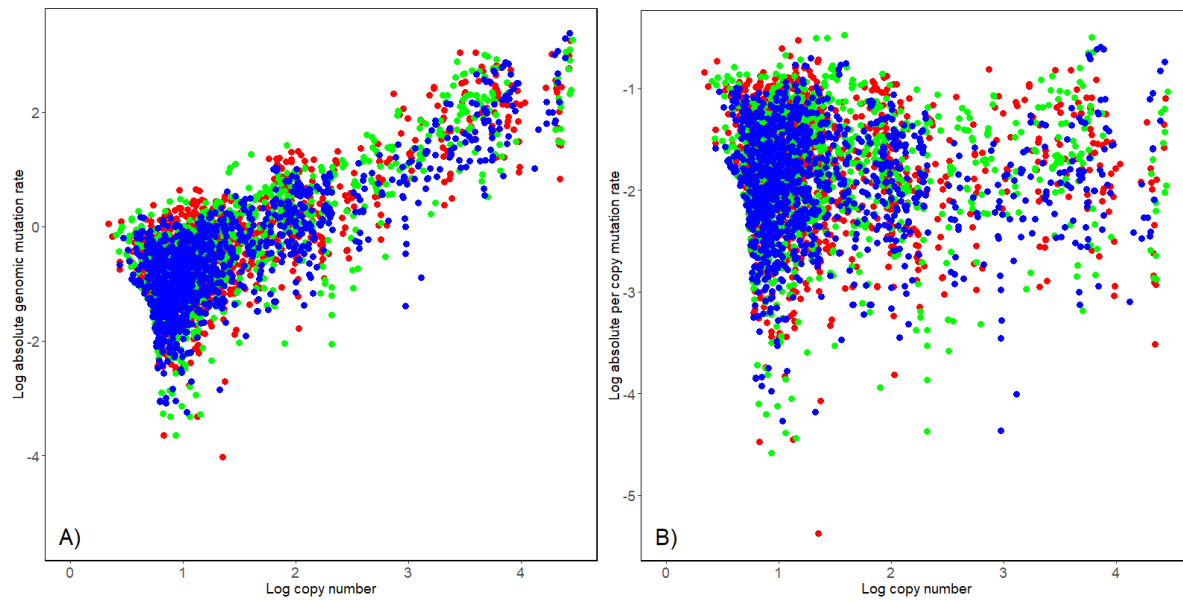
Line	Num. of reads	Line	Num. of reads	Line	Num. of reads
FA10	117276453	GA10	67872281	IA10	77469618
FA12	111118502	GA2	74933433	IA1	76339146
FA13	98944469	GA3	78929079	IA2	82762562
FA14	118671293	GA4	78740281	IA4	94985444
FA1	100759904	GA5	79320460	IA5	110594565
FA5	108817148	GA6	71148348	IA6	100880878
FA6	92034278	GA7	76557733	IA7	78383275
FASC	88590180	GA8	58644700	IA8	90086901
FC12	68037364	GASC	72730208	IASC	76126994
FC1	67658931	GC10	75779178	IC10	83523929
FC2	70000714	GC2	96690080	IC1	84683921
FC3	87568787	GC3	74165418	IC3	79042669
FC4	72157642	GC4	100073584	IC4	89897718
FC6	70810578	GC5	107328336	IC5	64317095
FC7	82224489	GC6	92860991	IC6	67608868
FC8	71891429	GC8	106121415	IC7	67236965
FCSC	93661924	GC9	105658620	IC9	70519081
FAEC1	92815980	GCSC	71010993	ICSC	64756748
FAEC2	101739760	GAEC1	83641911	IAEC1	77480400
FCEC1	79655226	GAEC2	70749011	IAEC2	84786933
FCEC2	92606026	GCEC1	119412606	ICEC1	89041032
		GCEC2	107606136	ICEC2	89387923

785

786

787 **Supplementary Figures**

788



789

790

791 **Figure S1.** Absolute genomic mutation rate (A) and absolute per copy mutation rate (B) plotted
792 against initial copy number for each kmer from each genotype. Red, green and blue represents
793 genotypes from Finland, Germany and Israel, respectively.

794

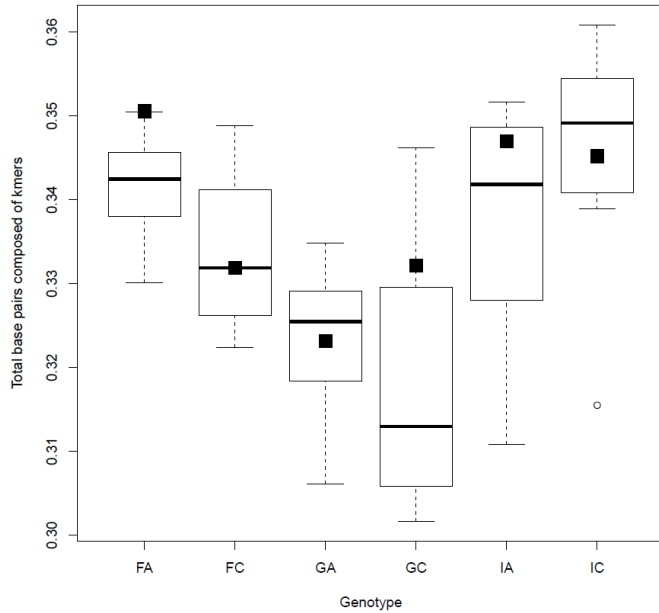
795

796

797

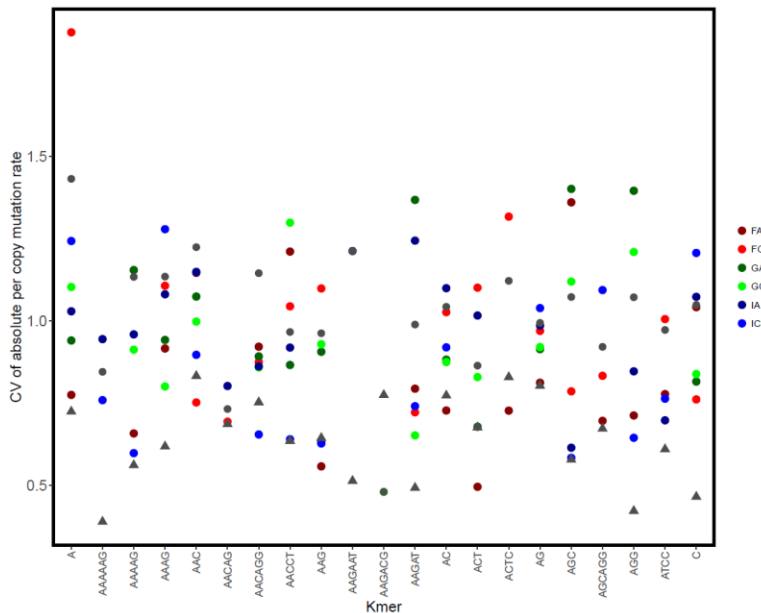
798

799



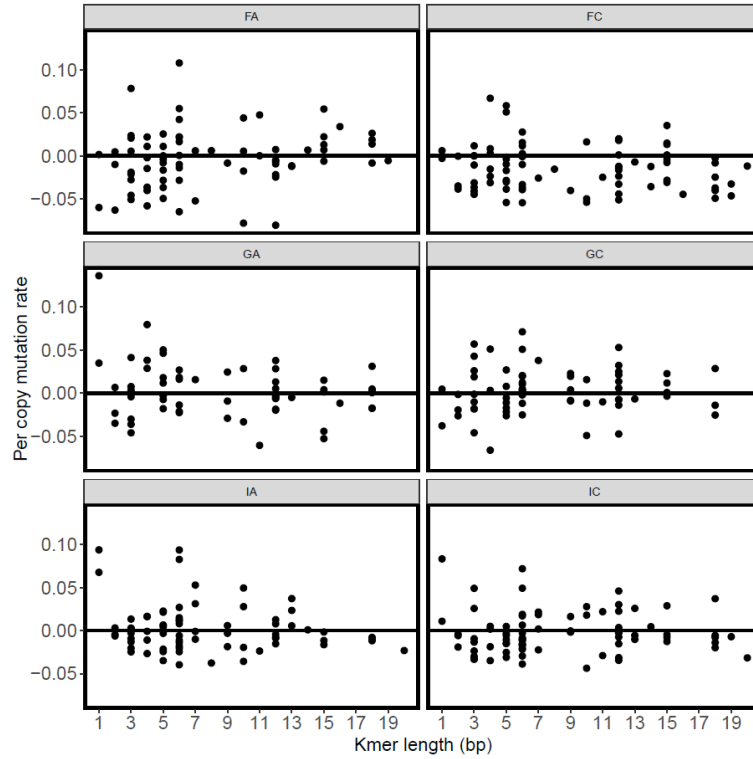
800
801
802
803
804
805
806
807

Figure S2. Total base pairs composed of kmers for SC and MA lines of each genotype. Total base pairs of MA lines represented by the boxplots; white circles represent lines outside 1.5 times of the interquartile range. Total base pairs for SC lines represented by the black squares. Data shown for each of six genotypes of *D. magna* collected from three locations, Finland (F), Germany (G) and Israel (I).



808
809
810
811
812
813

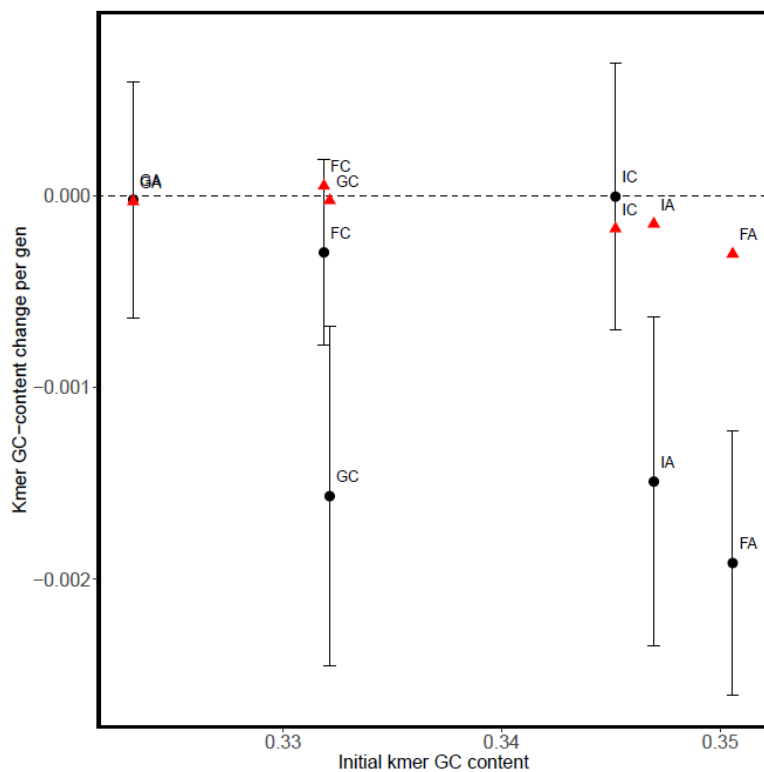
Figure S3. Coefficient of variation in $|u_{i,j}|$ for each *D. magna* genotype (circle) and for *D. pulex* (grey triangle) for the 21 kmers shared across species. Grey circles represent the coefficient of variation across all six genotypes of *D. magna* collected from three locations, Finland (F), Germany (G) and Israel (I).



814
815
816
817

Figure S4. Per copy mutation rate of kmers j (u_j) plotted against kmer lengths for six genotypes of *D. magna* collected from three locations, Finland (F), Germany (G) and Israel (I).

818

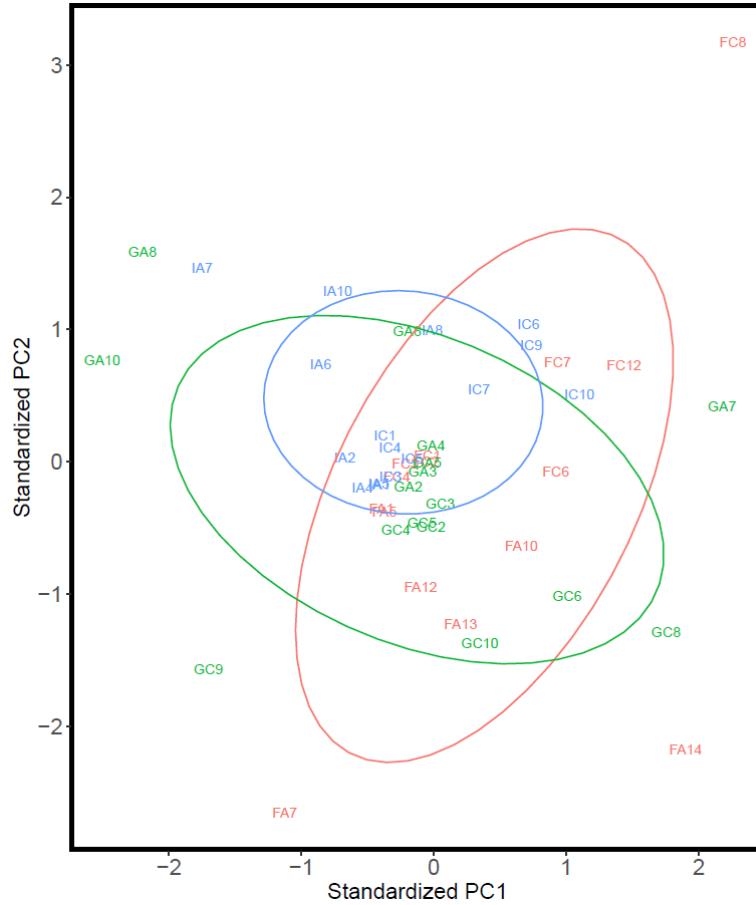


819

820 **Figure S5.** Mean (+/- SE) kmer GC-content change plotted against initial kmer GC-content for
821 all six genotypes of *D. magna* collected from three locations, Finland (F), Germany (G) and
822 Israel (I). Black circles and red triangles represent MA and EC lines, respectively.

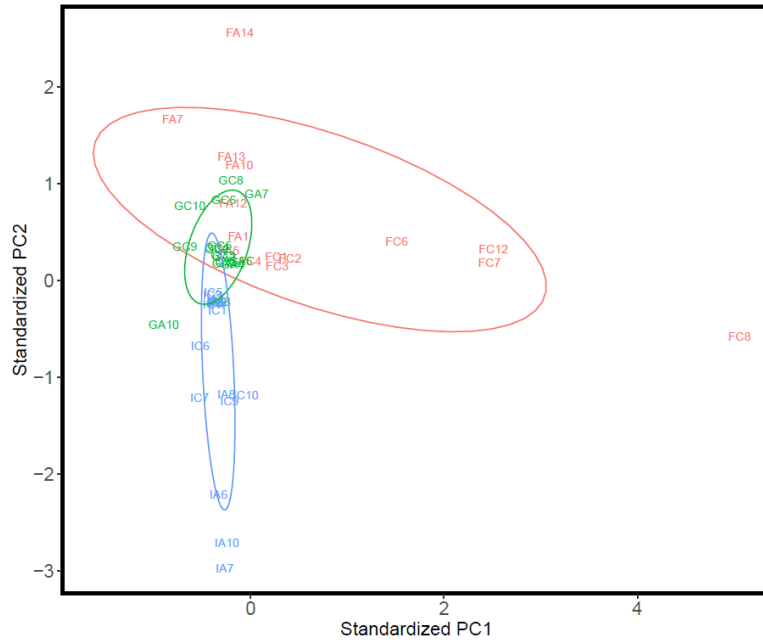
823

824



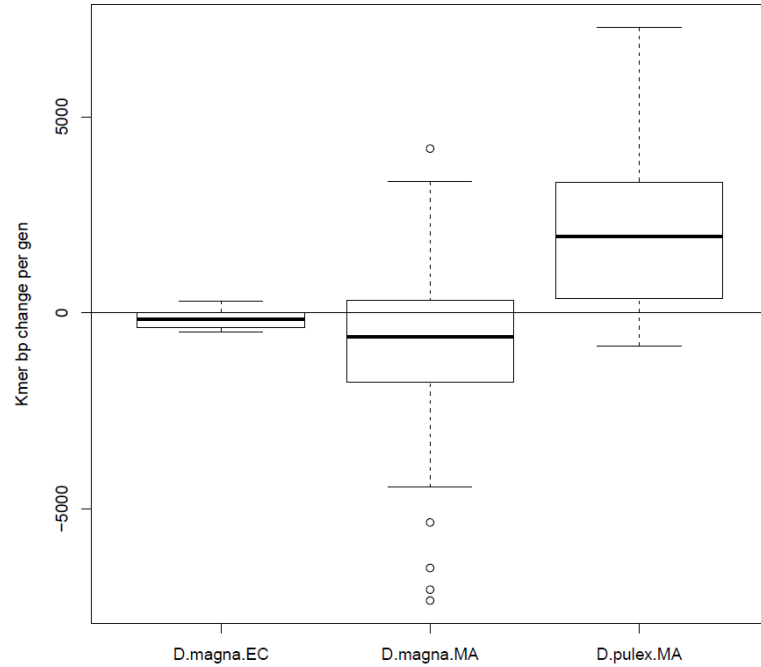
825
826
827
828
829

Figure S6. Population structure using u_j for the 31 kmers with mutation rate estimates for all six genotypes of *D. magna* collected from three locations, Finland (F; red), Germany (G; green) and Israel (I; blue). Each MA line is plotted based on the first and second principal components axis.



830
831
832
833
834
835
836
837

Figure S7. Population structure using u_j for all 144 kmers with mutation rate estimates for all six genotypes of *D. magna* collected from three locations, Finland (F; red), Germany (G; green) and Israel (I; blue). Each MA line is plotted based on the first and second principal components axis. If there was no mutation rate estimate for a kmer in a particular genotype, we set u_j as 0. Each MA line is plotted based on the first and second principle components axis.



838
839
840
841

Figure S8. Change in total kmer content (bp) per generation for *D. magna* EC lines, *D. magna* MA lines and *D. pulex* MA lines.

842 References

- 843 1. Baer, C.F., Miyamoto, M.M., and Denver, D.R. (2007). Mutation rate variation in
844 multicellular eukaryotes: causes and consequences. *Nat. Rev. Genet.* 8, 619–
845 631.
- 846 2. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S.,
847 Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A
848 New Genome Assembly Algorithm and Its Applications to Single-Cell
849 Sequencing. *J. Comput. Biol.* 19, 455–477.
- 850 3. Bhargava, A., and Fuentes, F.F. (2010). Mutational Dynamics of Microsatellites.
851 *Mol. Biotechnol.* 44, 250–266.
- 852 4. Bushnell, B., Rood, J., and Singer, E. (2017). BBMerge – Accurate paired
853 shotgun read merging via overlap. *PLOS ONE* 12, e0185056.
- 854 5. Calabrese, P., and Durrett, R. (2003). Dinucleotide Repeats in the *Drosophila*
855 and Human Genomes Have Complex, Length-Dependent Mutation Processes.
856 *Mol. Biol. Evol.* 20, 715–725.
- 857 6. Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J., and Deka, R. (1997).
858 Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc.*
859 *Natl. Acad. Sci.* 94, 1041–1046.
- 860 7. Eckert, K.A., and Hile, S.E. (2009). Every microsatellite is different: Intrinsic DNA
861 features dictate mutagenesis of common microsatellites present in the human
862 genome. *Mol. Carcinog.* 48, 379–388.
- 863 8. Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution.
864 *Nat. Rev. Genet.* 5, 435–445.
- 865 9. Fan, H., and Chu, J.-Y. (2007). A Brief Review of Short Tandem Repeat
866 Mutation. *Genomics Proteomics Bioinformatics* 5, 7–14.
- 867 10. Feupe Fotsing, S., Wang, C., Saini, S., Shleizer-Burko, S., Goren, A., and
868 Gymrek, M. (2018). Multi-tissue analysis reveals short tandem repeats as
869 ubiquitous regulators of gene expression and complex traits. *BioRxiv* 495226.
- 870 11. Flynn, J.M., Caldas, I., Cristescu, M.E., and Clark, A.G. (2017). Selection
871 Constrains High Rates of Tandem Repetitive DNA Mutation in *Daphnia pulex*.
872 *Genetics* 207, 697–710.

- 873 12. Flynn, J.M., Lower, S.E., Barbash, D.A., and Clark, A.G. (2018). Rates and
874 Patterns of Mutation in Tandem Repetitive DNA in Six Independent Lineages of
875 *Chlamydomonas reinhardtii*. *Genome Biol. Evol.* 10, 1673–1686.
- 876 13. Gemayel, R., Cho, J., Boeynaems, S., and Verstrepen, K.J. (2012). Beyond
877 Junk-Variable Tandem Repeats as Facilitators of Rapid Evolution of Regulatory
878 and Coding Sequences. *Genes* 3, 461–480.
- 879 14. Haag, C.R., McTaggart, S.J., Didier, A., Little, T.J., and Charlesworth, D. (2009).
880 Nucleotide Polymorphism and Within-Gene Recombination in *Daphnia magna*
881 and *D. pulex*, Two Cyclical Parthenogens. *Genetics* 182, 313–323.
- 882 15. Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D.L.,
883 Charlesworth, B., and Keightley, P.D. (2007). Direct estimation of per nucleotide
884 and genomic deleterious mutation rates in *Drosophila*. *Nature* 445, 82–85.
- 885 16. Haasl, R.J., and Payseur, B.A. (2013). Microsatellites as Targets of Natural
886 Selection. *Mol. Biol. Evol.* 30, 285–298.
- 887 17. Halligan, D.L., and Keightley, P.D. (2008). Spontaneous Mutation Accumulation
888 Studies in Evolutionary Genetics. *Annu. Rev. Ecol. Evol. Syst.* 40, 151–172.
- 889 18. Hannan, A.J. (2018). Tandem repeats mediating genetic plasticity in health and
890 disease. *Nat. Rev. Genet.* 19, 286–298.
- 891 19. Kelkar, Y.D., Eckert, K.A., Chiaromonte, F., and Makova, K.D. (2011). A matter
892 of life or death: How microsatellites emerge in and vanish from the human
893 genome. *Genome Res.* 21:2038-2048.
- 894 20. Klüttgen, B., Dülmer, U., Engels, M., and Ratte, H.. (1994). ADaM, an artificial
895 freshwater for the culture of zooplankton. *Water Res.* 28, 743–746.
- 896 21. Kornberg, A., Bertsch, L.L., Jackson, J.F., and Khorana, H.G. (1964). Enzymatic
897 Synthesis Of Deoxyribonucleic Acid, Xvi. Oligonucleotides As Templates And
898 The Mechanism Of Their Replication. *Proc. Natl. Acad. Sci. USA* 51, 315–323.
- 899 22. Kruglyak, S., Durrett, R.T., Schug, M.D., and Aquadro, C.F. (1998). Equilibrium
900 distributions of microsatellite repeat length resulting from a balance between
901 slippage events and point mutations. *Proc. Natl. Acad. Sci. USA* 95, 10774–
902 10778.
- 903 23. Lai, Y., and Sun, F. (2003). The relationship between microsatellite slippage

- 904 mutation rate and the number of repeat units. *Mol. Biol. Evol.* 20, 2123–2131.
- 905 24. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with
906 Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.
- 907 25. Lynch, M. (2010). Evolution of the mutation rate. *Trends Genet.* 26, 345–352.
- 908 26. Lynch, M., Kroskella, M., and Schaack, S. (2006). Mutation Pressure and the
909 Evolution of Organelle Genomic Architecture. *Science* 311, 1727–1730.
- 910 27. Metzgar, D., Bytof, J., and Wills, C. (2000). Selection Against Frameshift
911 Mutations Limits Microsatellite Expansion in Coding DNA. 10(1):72-80.
- 912 28. Miner Brooks E., De Meester Luc, Pfrender Michael E., Lampert Winfried, and
913 Hairston Nelson G. (2012). Linking genes to communities and ecosystems:
914 *Daphnia* as an ecogenomic model. *Proc. R. Soc. B Biol. Sci.* 279, 1873–1882.
- 915 29. Okazaki, S., Tsuchida, K., Maekawa, H., Ishikawa, H., and Fujiwara, H. (1993).
916 Identification of a Pentanucleotide Telomeric Sequence, (TTAGG)_n, in the
917 Silkworm *Bombyx mori* and in Other Insects. *Mol. Cell Biol.* 13, 9.
- 918 30. Petrov, D.A. (2002). Mutational Equilibrium Model of Genome Size Evolution.
919 *Theor. Popul. Biol.* 61, 531–544.
- 920 31. Sahara, K., Marec, F., and Traut, W. (1999) TTAGG Telomeric Repeats in
921 Chromosomes of Some Insects and Other Arthropods. *Chromosome Res* 7: 449.
- 922 32. Schaack, S. (2008). *Daphnia* comes of age: an ecological model in the genomic
923 era. *Mol. Ecol.* 17, 1634–1635.
- 924 33. Schumpert, C., Nelson, J., Kim, E., Dudycha, J.L., and Patel, R.C. (2015).
925 Telomerase Activity and Telomere Length in *Daphnia*. *PLOS ONE* 10, e0127196.
- 926 34. Seyfert, A.L., Cristescu, M.E.A., Frisse, L., Schaack, S., Thomas, W.K., and
927 Lynch, M. (2008). The rate and spectrum of microsatellite mutation in
928 *Caenorhabditis elegans* and *Daphnia pulex*. *Genetics* 178, 2113–2121.
- 929 35. Shah, S.N., Hile, S.E., and Eckert, K.A. (2010). Defective Mismatch Repair,
930 Microsatellite Mutation Bias, and Variability in Clinical Cancer Phenotypes.
931 *Cancer Res.* 70, 431–435.
- 932 36. Subirana, J.A., Albà, M.M., and Messeguer, X. (2015). High evolutionary turnover
933 of satellite families in *Caenorhabditis*. *BMC Evol. Biol.* 15.
- 934 37. Sun, J.X., Helgason, A., Masson, G., Ebenesersdóttir, S.S., Li, H., Mallick, S.,

- 935 Gnerre, S., Patterson, N., Kong, A., Reich, D., et al. (2012). A direct
936 characterization of human mutation based on microsatellites. *Nat. Genet.* *44*,
937 1161–1165.
- 938 38. Wei, K.H.-C., Grenier, J.K., Barbash, D.A., and Clark, A.G. (2014). Correlated
939 variation and population differentiation in satellite DNA abundance among lines
940 of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* *111*, 18793–18798.
- 941 39. Xu, X., Peng, M., Fang, Z., and Xu, X. (2000). The direction of microsatellite
942 mutations is dependent upon allele length. *Nat. Genet.* *24*, 396–399.
- 943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958