Selection Constrains High Rates of Tandem Repetitive DNA Mutation in *Daphnia pulex*

Jullien M. Flynn,* Ian Caldas,[†] Melania E. Cristescu,[‡] and Andrew G. Clark^{*,†}

*Department of Molecular Biology and Genetics and [†]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, and [‡]Department of Biology, McGill University, Montreal, Quebec, Canada H3A 1B1

ORCID ID: 0000-0001-5999-4781 (J.M.F.)

ABSTRACT A long-standing evolutionary puzzle is that all eukaryotic genomes contain large amounts of tandemly-repeated DNA whose sequence motifs and abundance vary greatly among even closely related species. To elucidate the evolutionary forces governing tandem repeat dynamics, quantification of the rates and patterns of mutations in repeat copy number and tests of its selective neutrality are necessary. Here, we used whole-genome sequences of 28 mutation accumulation (MA) lines of *Daphnia pulex*, in addition to six isolates from a non-MA population originating from the same progenitor, to both estimate mutation rates of abundances of repeat sequences and evaluate the selective regime acting upon them. We found that mutation rates of individual repeats were both high and highly variable, ranging from additions/deletions of 0.29–105 copies per generation (reflecting changes of 0.12–0.80% per generation). Our results also provide evidence that new repeat sequences are often formed from existing ones. The non-MA population isolates showed a signal of either purifying or stabilizing selection, with 33% lower variation in repeat copy number on average than the MA lines, although the level of selective constraint was not evenly distributed across all repeats. The changes between many pairs of repeats were correlated, and the pattern of correlations was significantly different between the MA lines and the non-MA population. Our study demonstrates that tandem repeats can experience extremely rapid evolution in copy number, which can lead to high levels of divergence in genome-wide repeat composition between closely related species.

KEYWORDS tandem repeats; mutation accumulation (MA); stabilizing selection; satellite evolution

UP to half of the genome of higher eukaryotes is composed of tandem arrays of simple repetitive motifs that can be contained within satellite DNA and span kilobases to megabases (Platero *et al.* 1998; Padeken *et al.* 2015). These long arrays of short tandem repeats had initially been thought to carry no useful function, and because they posed a replication burden, became known as “junk DNA” (Orgel and Crick 1980). They also have the potential to be harmful because they can cause deleterious genomic rearrangements facilitated by recombination between similar motifs on different chromosomes (Bzymek and Lovett 2001). Thus, evolutionary biologists have long wondered why repeated sequences accumulate and are maintained at such abundance in the genome. Satellite DNA is

typically found in heterochromatic regions where expression is silenced and recombination is low, in Y chromosomes, and in centromeric and telomeric regions (Charlesworth and Charlesworth 2000; Henikoff *et al.* 2001). One hypothesis is that being situated in heterochromatin, where recombination is suppressed, could both minimize the cost of repeated DNA and reduce the efficacy of selection against it, allowing its accumulation (Charlesworth *et al.* 1994). Moreover, the observation that genome-wide, short tandem repeats undergo rapid turnover, with nearly complete replacement of these repeat sequences even between closely related species, poses challenging questions about the evolutionary forces that shape their evolution (Wei *et al.* 2014; Subirana *et al.* 2015). The study of tandem repetitive DNA evolution on the timescales of diverged populations and species has revealed interesting patterns, but understanding the forces that generated these patterns requires direct observation on a shorter timescale.

Early evolutionary models did not consider selection to be influencing satellite sequence evolution, except perhaps

Copyright © 2017 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.117.300146>

Manuscript received June 27, 2017; accepted for publication August 11, 2017; published Early Online August 15, 2017.

Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.300146/-/DC1.

[†]Corresponding author: Cornell University, Biotechnology-Rm 227, 526 Campus Rd, Ithaca, NY 14853. E-mail: jmf422@cornell.edu

selection on overall genome size (MacGregor and Sessions 1986; Charlesworth *et al.* 1994; Stephan and Cho 1994). Stephan and Cho (1994) proposed a model where tandemly-repeated sequences could be generated and expanded from mutation, recombination, and replication slippage, combined with selection to maintain the genome size. Another model, the “library hypothesis,” attempts to explain the differences in satellite composition by posing that the common ancestor of related species contained a library of many satellites, which are then differentially amplified within each lineage as they diverge (Fry and Salser 1977). However, the rate at which short repeat sequences expand and contract is difficult to quantify. Most models pose evolutionary neutrality of satellite repeats, but we now know that satellite sequences also carry vital cellular functions. Since many tandem repeats are contained in satellite DNA near centromeric regions, there is evidence that selection acts on the sequence motifs themselves, as they serve to bind centromere and histone proteins for centromere maintenance and chromosome separation (Henikoff *et al.* 2001; Malik 2009). These functions have implications for genome stability, and changes in satellite DNA sequences have been shown in some cases to be drivers of speciation by inducing chromosome rearrangements leading to karyotype divergence (Paço *et al.* 2015). Moreover, some satellite sequences are transcribed and may be involved in the regulation of heterochromatin formation (Palomeque and Lorite 2008; Plohl *et al.* 2008). Even more intriguing is the observation that perturbation of satellite content can alter gene expression genome-wide (Lemos *et al.* 2008). At present, it is not known whether these examples of functional importance of satellite DNA are the exceptions or the rule and to what extent selection governs the evolution of short tandem repetitive DNA.

Here, we focus on tandem repeats of unit length 1–20 bp that span at least 50 bp to capture genome-wide tandem repeats, especially those in long arrays of satellite DNA. Several potential and nonmutually exclusive selection regimes could be operating on arrays of tandem repeats: negligible selection where repetitive DNA content is primarily governed by mutation and drift, stabilizing selection to maintain a particular “optimal” composition of copy number and genomic arrangement, negative selection to purge repeated DNA, and positive selection for rapidly generating new repeat motifs. Studying satellite DNA changes over an evolutionary time-scale showed clear differences across geographic subpopulations of a single species and almost complete turnover between species, confirming that they evolve rapidly (Wei *et al.* 2014; Subirana *et al.* 2015). However, quantifying the relative contributions of mutation, genetic drift, and selection is difficult since all these forces are at play in influencing genetic variation in natural populations. Additionally, it has been a challenge to quantify the genome-wide repetitive DNA composition since it is problematic to sequence and assemble and is thus underrepresented in genome assemblies (Hoskins *et al.* 2007). Early studies describing satellite composition in different species have focused on single satellites or a small family of satellite sequences (Lohe and Brutlag 1987).

Obtaining a genome-wide view of the rate of mutation in tandem repeat sequences and how selection shapes their evolution would facilitate better understanding of the long-standing puzzles of satellite DNA evolution.

Mutation accumulation (MA) experiments reduce selection to a minimum by enforcing bottlenecks every generation and reducing the effective population size so that both neutral and deleterious mutations can accumulate almost neutrally (Simmons and Crow 1977). An MA experiment combined with observations from a population with the same genetic background where selection is not removed allows the study of the effects of mutation alone vs. the effects of mutation combined with selection (Flynn *et al.* 2017). *Daphnia pulex* is an ideal organism in which to study tandem repeat mutation, first because MA studies can be conducted effectively under asexual reproduction and single-progeny bottlenecks. Short microsatellite arrays have been studied in *D. pulex* (Colbourne *et al.* 2004; Sung *et al.* 2010), but the sequence motifs and abundances of genome-wide tandem repeats have not been characterized, although 25% of the genome is estimated to be heterochromatic (Colbourne *et al.* 2011). *D. pulex* has a potentially dynamic genome, with evidence of high rates of deletion and duplication (Keith *et al.* 2016). The lineage we use reproduces exclusively asexually via apomixis, meaning changes in satellite content represent spontaneous mutations in the germline without the opportunity for meiotic drive to play a role (Malik 2009; Wei *et al.* 2017).

In this study, we compare the tandem repeat copy number of MA lines (without selection) to individuals raised in a competitive non-MA population (with selection) of the identical *D. pulex* genetic background. We were interested in exploring the mutational dynamics of tandem repetitive DNA that spans long arrays and can often not be reliably mapped to the reference genome. We quantify the repeat composition using a mapping-independent method to detect and quantify all tandem repeat sequences or “kmers” up to 20 bp unit length detectable from short-read data, without introducing biases from the reduced representation of repetitive sequences in the reference genome. Our approach allows us to (1) estimate the repeat mutation rates, including expansions, contractions, complete loss of kmers, and generation of new kmers; and (2) evaluate the type of selection (if any) acting on the tandem repeat arrays. Given the observed natural variation in satellite DNA content among populations and species, we expect expansions and contractions to occur at high rates in the MA lines. If tandem repetitive DNA is under selective constraint, we expect there to be less variation in repeat copy number in the population evolving under selection than the neutrally-evolving MA lines.

Materials and Methods

Daphnia line setup and DNA sequencing

The current study includes a total of 34 genomes: 24 MA line genomes, 6 population isolate genomes that were sequenced

in Flynn *et al.* (2017), and 4 additional previously unpublished MA genomes. The asexually reproducing MA lines and a non-MA population were initiated from a single progenitor [see Flynn *et al.* (2017) for a detailed description of the MA experiment]. MA lines were propagated for an average of 82 generations before whole-genome sequencing with single-individual bottlenecks between generations prior to DNA isolation and whole-genome sequencing. In contrast, the non-MA population was maintained without inducing bottlenecks for 46 months in a 15 liter tank, before six individuals were isolated for sequencing. The approximate census population size was estimated to be 100–250. Overlapping generations occurred so the exact number of generations that the population isolates progressed could not be recorded. Thus, we used a life history experiment and estimated the slowest and fastest moving lineages to calculate that the population underwent at least 62 generations (Supplemental Material, File S3). Prior to sequencing, *Daphnia* individuals were subjected to a brief antibiotic treatment and fed with sterile Sephadex beads to reduce contaminants before sequencing (Fields *et al.* 2015).

Sequencing libraries were prepared with Illumina Nextera procedures in two batches and all genomes were sequenced to approximately the same coverage (10×). To ensure reproducibility between multiple library preparations, we performed technical replication on two MA lines, such that two independent library preparations and Illumina sequencing runs were done for the MA lines C01 and C35 (in a completely separate, third sequencing batch). We analyzed independently all technical replicates to ensure that variation produced by different library preparations is smaller than the variation due to biological expansion/contraction mutations. Libraries were sequenced on an Illumina HiSeq2500 instrument at Genome Quebec of McGill University with 100 bp paired-end reads.

Tandem repeat quantification

To remove redundant sequences, adapters were trimmed and overlapping reads were merged with SeqPrep (<https://github.com/jstjohn/SeqPrep>). To identify and quantify repeat sequences, the resulting unmapped read files were used as input for the program k-Seek (Wei *et al.* 2014, <https://github.com/weikevinhc/k-seek>). The current version of k-Seek detects repeat units or kmers of length 1–20 bp that are repeated tandemly to cover at least 50 bp within the same read. The program allows for one single-nucleotide mismatch per repetition of each kmer [see Wei *et al.* (2014) for details]. All offsets and reverse complements of each kmer are compiled and the total of all individual unit copies are summed across all reads. Kmer sequences are presented as the strand/offset that is alphabetically ordered (*i.e.*, A's will be at the start of the sequences if possible). This method of tandem repeat detection and quantification has been found to be reproducible across different library preparations (Wei *et al.* 2014).

To obtain a quantitative comparison between samples, we normalized the kmer counts by both individual library

sequencing depth and GC content, as PCR-based library preparations are known to show a bias in the GC content of the fragments amplified and sequenced (Benjamini and Speed 2012). This is especially problematic for repetitive DNA analysis since, by nature, many of the repeated sequences may have extreme GC contents. First, reads were mapped against the *D. pulex* reference genome (version 1, Colbourne *et al.* 2011) using BWA (Li and Durbin 2009) v0.7.10 with default settings. Output BAM files spanning the whole genome were given as input to a custom shell script (https://github.com/jmf422/Daphnia-MA-lines/tree/master/GC_correction) to calculate correction factors following Benjamini and Speed (2012). The correction factors produced for high and low GC contents were extreme and variable, which was likely due to low read counts giving unstable estimates as well as mapping biases to the reference genome (Flynn *et al.* 2017). Thus, to smooth the correction, we used a Python script to bin values of GC content together, employing wider bins for GC content < 0.25 and > 0.60 (https://github.com/jmf422/Daphnia-MA-lines/tree/master/GC_correction). The correction factor was then applied to kmer counts according to which GC bin their content belonged.

Mutation and interspersal metrics

We define mutations in repeats as any change in the number of copies of each specific kmer sequence. The rate of mutation is then the observed change in the repeat's copy number per generation. Each repeat sequence may be in tandem arrays found in one or in several locations across the genome, and our method sums the counts of each repeat unit copy across all genomic regions. A mutation rate for a given repeat reflects the sum of changes in copy number across these potential multiple loci. To calculate the mutation rates, we used the mean copy number of the repeat in the non-MA population isolates as a proxy for the ancestral abundance using the following equation:

$$u_{i,m} = \frac{m_i - \bar{P}_i}{G_m},$$

where $u_{i,m}$ is the mutation rate of kmer i in MA line m , m_i is the abundance of kmer i in MA line m , $\bar{P}_i$ is the mean abundance in the population for kmer i , and G_m is the number of generations propagated for MA line m . u could be negative (for contractions) or positive (for expansions). We used the same equation to calculate the absolute rate of change of each kmer, except we took the absolute value of the numerator.

As mutation rates could be correlated with copy number, we also calculated the absolute mutation rate normalized by initial copy number using the following equation:

$$u_{i,m} = \frac{|m_i - \bar{P}_i|}{G_m \times \bar{P}_i}.$$

We used the same equations for calculating the “realized” repeat mutation rates (mutations that made it through selection in the population isolates), and we used the conservative estimate of 62 for the number of generations.

We also searched for new repeat sequences generated *de novo* in the MA experiment. To be considered as a new repeat, putative new repeats had to have at least three copies in at least one MA line and be completely absent from all other lines. New shared repeats had to have at least three copies in one MA line and at least two copies in a second line. Since k-Seek has a detection limit for repeat arrays that are at least 50 bp long, we checked if the putative new repeats were present in shorter arrays (~25 bp) in the other MA lines to determine if the repeat sequence was generated truly *de novo* or if it expanded from an already present repeated motif. Similarly, we searched for repeats that had been lost in individual and pairs of MA lines throughout the course of the experiment. Our criteria for identifying lost repeats were that they had to have 0 copies in the affected line or pair of lines, and at least two copies in all other lines. We checked if putative lost repeats were completely undetectable or if they were still present in short arrays (~25 bp), below the detection threshold of k-Seek.

Since we have paired-end reads, we have the potential to detect the level to which tandem repeats are present on both reads of the pair (on the same genomic fragment). To quantify this interspersion level, we used the metric I , as in Wei *et al.* (2014):

$$I = \left(n_{ij} / \sqrt{n_i n_j} \right),$$

where n_{ij} is the number of read pairs containing kmer i and kmer j , and n_i and n_j are the number of read pairs containing kmer i and kmer j , respectively.

Data availability

Raw sequence files have been deposited in the National Center for Biotechnology Information Sequence Read Archive BioProject ID: PRJNA341529 BioSamples SAMN05725816, SAMN05725817, SAMN07417200, and SAMN07417201.

All analyses were performed using RStudio (V 0.99.903), Python V2, and Perl V5 scripts. Html files showing the code and output to all the analyses are available in the Supplemental Material, and all scripts as well as the required processed input files are available at <https://github.com/jmf422/Daphnia-MA-lines>. File S1 contains descriptions of all supplemental files, as well as the supplemental table and figure legends.

Results

Library and GC correction

Over 713 million unmapped reads across all samples were scanned for short tandemly-repeated sequences or kmers with k-Seek. Of these, over 5.85 million reads (0.82%) were found to be composed of kmers of unit length 1–20 bp and encompassed at least 50 bp of the read. Although the second library batch had more reads overall, a similar proportion of the total reads contained repeats in both libraries (Figure S1 in File S2).

Unless otherwise noted, the copy number of each repeat is presented in copies per $1 \times$ read depth after GC normalization.

We found that our qualitative conclusions were robust whether or not we applied a correction factor to normalize the repeat counts based on each repeat sequence's GC content. Our results were also robust across a wide range of parameters used for the GC correction. As library preparations involving PCR have been shown to result in a biased representation of fragments based on GC content (Benjamini and Speed 2012), we present results after GC correction as described in the *Materials and Methods*.

Since the libraries were prepared in two separate batches, we were concerned that this could confound our results. A Principal Components Analysis was able to separate samples based on library batch along PC2, but PC1 appeared to separate samples based on biological mutation patterns (Figure S2 in File S2). There were no consistent differences across all repeat sequences or GC contents of repeats that could be consistently corrected for (File S5). To ensure that library batch would not be a confounding factor, we performed a technical replicate of two MA lines and showed that the abundances of each repeat between the library prep batches were highly similar (File S5). Although there was minor variation between technical replicates in some repeats, their overall mutational patterns were highly similar, shown by their high clustering on PC1 (Figure S2 in File S2). Additionally, we tested for differences between the mean copy number of the repeats between the MA lines prepared in the first batch and the MA lines prepared in the second batch. The mean copy number of most of the repeats (28 out of 39) in the second-batch MA lines were within the 99% C.I.s from 1000 subsampling replicates of four first-batch MA lines. Thirty of the repeats had mean copy number abundances in the second batch within the range of observed abundances of the first-batch MA lines (File S5). For these reasons, we did not consider library preparation batch to be a significant confounding factor.

Description of short tandem repeat content in *D. pulex*

We first sought to describe and quantify the genome-wide repeat content in *D. pulex* and make inferences about repeat sequence origins. There were 162 repeats that had at least two copies per $1 \times$ coverage in all our *D. pulex* lines. There were 39 repeats that had an average abundance of at least 100 copies, and 12 that had an average abundance of at least 1000 copies after normalization (Table S1). We chose to focus most of our analysis on the 39 kmers that had at least 100 copies after normalization (Figure 1). The most abundant kmer, the poly-A repeat, was present at an average of 79,528 copies in each genome. The second most abundant kmer, present at an average of 62,281 copies, was AACCT. This 5-mer is known to be the ancestral telomere repeat in Arthropods (Sahara *et al.* 1999) and was previously found in *Daphnia* (Colbourne *et al.* 2011). Runs of poly-C were also found to be in the “top 39 kmers,” having an average of 6379 copies. Two of the four possible 2-mers were included,

Figure 1 Copy number abundance of the kmers that contain at least 100 counts after normalization. Kmers are arranged in order of length, from 1 to 20 mers.

AG and AC, having an average of 408 and 244 copies, respectively. The most abundant repeat sizes were 5-mers (there were six 5-mers of the most abundant 39 kmers), 10-mers (7/39), and 20-mers (7/39). No 15-mers were abundant (> 100 copies) in the genome, but there were 13 among all 162 kmers. Overall, 20-mers had the highest representation with 50 repeat sequences of the 162. Adding up the total simple tandem repeat content per genome in base pairs, we found the median to be 1.20 Mb per $1\times$ coverage, which represents 0.6% of the estimated 200 Mb genome.

We also compared the short tandem repeat composition in *D. pulex* to that in *Drosophila melanogaster*, the only other arthropod that has had its genome-wide tandem repeat content characterized to the same level of detail (Wei *et al.* 2014). This comparison might indicate the extent of repeat diversity across arthropods and functional conservation of some repeat sequences. Comparing the most abundant repeats (at least on average 100 copies normalized) between these two species, we found that 10 short kmers (mostly derivations of AG repeats) were present in both species: A, C, AC, AG, AAC, AAG, ACT, AAAAG, AAGAG, and AAGGAG. We found several short (3–5 bp) motifs to be highly represented in the top 39 kmers, and among the 162 kmers (Table 1). Most of these motifs were completely absent or rare within the top 108 *D. melanogaster* kmer sequences, except for AG-type motifs (Table 1).

To understand the origins and dynamics of tandem repeat sequences in *Daphnia*, we inspected the repeats for sequence

similarities. We found that many repeats belonged to families of related sequences separated by a small number of potential mutational steps. Figure 2 shows a network diagram highlighting the potential mutational relationships between 29 of the 39 of the top repeats. The most striking example of a family of related repeats is the 9 kmers of length 10–20 bp shown in the top left corner of Figure 2. Ten repeats were not included in the network, either because they were mono-, di-, or trinucleotides whose sequence similarities were not meaningful (six repeats: A, C, AC, AG, AGG, and ACT), or because they did not have a clear mutational relationship with any of the other top repeats (four repeats: AGCCTG, ACAGC, ATCC, and AACGGTACGG).

Extremely high rate of mutation in short tandem repeats

Because we could not call single-nucleotide changes reliably without mapping to the reference genome, we were not able to estimate mutation rates for single-base changes in the repeats and restrict our attention to changes in copy number. To estimate the copy number mutation rates, we used the mean abundance of each kmer in the population isolates as a proxy for the ancestral abundance of the kmer in the MA progenitor. This should be a reasonable estimation assuming that individuals in the population roughly maintained the ancestral repeat content (see below). In fact, all repeats were similar in abundance between the MA lines and the population isolates: all kmer copy numbers in the population were within two SD

Table 1 Number of kmers that contain certain sequence motifs

Motif sequence	In top 39 kmers	In top 162 kmers	In <i>D. melanogaster</i> kmers
AAAA	4	13	12
AAC	7	32	16
AAG	11	49	45
ACGC	8	52	0
AGC	12	69	4
AGG	11	41	14
AGGAG	3	15	5
GCCAG	7	48	0
TAGG	4	10	0
TCCAG	3	14	0

The top 39 kmers are characterized by having an average of at least 100 copies per 1× coverage, and the top 162 kmers have at least two copies per 1× coverage in all samples. The *D. melanogaster* kmers are the 108 kmers that have at least 100 copies normalized (K. Wei, S. Lower, T. J. Sless, D. A. Barbash, A. G. Clark, unpublished results).

of the copy numbers in the MA lines, and 69% were within one SD. This is consistent with the MA lines both gaining and losing kmer copies. The absolute mutation rates (summing expansions and contractions) per kmer sequence ranged from 0.29 to 105 copies/generation, with a median of 1.26 copies/generation. This is not including the telomere repeat, which had the highest rate of change at 199 copies/generation. It is likely that the changes affecting the telomere repeats are from a different, nonmutational process (*i.e.*, caused by a mechanism involving telomerase, which is not transgenerationally inherited). The suppression of meiosis in these *Daphnia* lines may have resulted in relaxed selection on telomere maintenance, but we have no direct evidence in support of this.

We found that the mutation rates were correlated with the repeat's initial copy number abundance, proxied by the mean copy number in the population isolates, and that the best fit of this relationship is linear ($r^2 = 0.88$, $P < 2.2 \times 10^{-16}$). Thus, to compare mutation rates of individual repeats, we normalized the mutation rates by the initial unit copy number to give mutation rates in copies per generation per original copy. Even after this normalization, we still found mutation rates to be exceptionally high—on the order of 10^{-3} per copy per generation—compared to previous estimates of other types of mutation such as microsatellite mutations. Normalized mutation rates ranged from 1.23×10^{-3} to 7.97×10^{-3} , with a mean of 2.74×10^{-3} copies per generation per original copy. Expansions were more frequent than contractions, with more MA lines experiencing expansions in most repeats, and the overall rate of change was in a positive direction for 124/162 kmers (76%) (Figure 3, A and B). Expansion and contraction rates were not correlated with either GC content or repeat unit size (Figure S3 in File S2).

A multi-generational MA experiment provides the potential opportunity to observe completely new kmer sequences arising. We searched for repeats that were generated *de novo* during the experiment, in single lines and shared across pairs of MA lines. We found five repeats gained in single MA lines and three gained in three independent pairs of MA lines (Table 2). These MA lines did not share

single-nucleotide mutations within uniquely mapping regions of the genome (Flynn *et al.* 2017). Five of the eight new repeats seemed to originate from a single-base substitution of an existing repeat present in high abundance (Table 2). Two of the other new repeat sequences contained motifs that were abundant in existing repeats. Only three out of eight putative new repeats were present in shorter arrays (~25 bp, half of k-Seek's requirement) in the ancestral lineage (Table 2). We also searched for repeats that were lost from individual and pairs of MA lines. There were seven repeats lost uniquely and two lost in a pair of lines (Table 3). Three of the seven lost repeats were lost from MA line C40. C40 is the MA line that was found to have ~100 kb of homozygous deletions of nonrepetitive sequence mapped to the reference genome on one chromosome (Flynn *et al.* 2017), so it is possible that these lost repeats were present exclusively in the deleted regions. Of the lost kmers, five out of seven were still detectable in shorter arrays (~25 bp) in the affected MA line (Table 3).

MA lines vs. population with selection

Comparing the changes in repeat copy number between the MA lines and the population isolates, which experienced selection, reveals the potential influences of selection on tandemly-repeated DNA. Since we already have evidence that purifying selection was acting in the population in both single-nucleotide mutations (Flynn *et al.* 2017) and copy number variants (F. J. J. Chain, J. M. Flynn, J. K. Bull, and M. E. Cristescu, unpublished results), this sets up the expectation for selection to be potentially operating on tandem repeat copy changes as well. The population did not experience a recent bottleneck and in fact the most recent common ancestor of the isolates analyzed here was the progenitor of the experiment (Flynn *et al.* 2017). We also made a conservative estimate of the number of generations the population underwent to compare mutation rates (File S3). First, we compared the total genome-wide amount of tandem repeats by multiplying the length of the kmer unit in base pairs with its number of copies in each genome. The MA lines diverged by a factor of 1.67 in genome-wide repeat amount, from 1.00 to 1.67 Mb total, with a median of 1.23 Mb. The population isolates had both a lower genomic amount of repeats and a narrower range among isolates, ranging from 0.97 to 1.22 Mb, with a median of 1.05 Mb (Figure 4A). A Levene's test did not detect a significant difference in the variances between these two groups ($P = 0.12$). However, the mean total amount (but not the variance) of repeats in the population was below the 5% C.I. produced from 1000 replicates of six randomly-sampled MA lines (File S4).

The MA lines underwent expansions across many repeats, with some MA lines experiencing a considerable increase in repeated DNA content in their genome over the course of the MA experiment compared to the non-MA population isolates (Figure 4B). Most notable was MA line C20, which experienced expansions in 34 of the 39 top kmers, with an overall increase in 580.2 kb in repeat content over 81 generations.

Figure 2 Network of relatedness of 29 of the 39 most abundant kmers. Networks show possible mutational steps between related kmers. Kmers are numbered based on their copy number abundance rank. The sizes of the circles also correspond to the kmer abundance. Indel, insertion/deletion.

On the other hand, population isolates deviated less from an overall balance of expansions and contractions (Figure 4B).

To perform a contrast in the variation of individual kmers between the MA lines and the population, we calculated the coefficient of variation (CV, the SD divided by the mean) for each kmer for both the MA lines and the population. We found that the CV was higher in the MA lines than the population for 37 of the top 39 kmers (Figure 4C, $P < 2.84 \times 10^{-9}$, sign test). We considered the possibility that the CV of the MA lines was artificially inflated if the mean abundances of some repeats in the second library preparation MA lines were different to those of the first library preparation MA lines. We also considered that the CV may be greater in the MA lines partially because the MA lines may have undergone more generations than the population. To ensure these factors were not founding our results, we also calculated the CV of the MA lines comparing only the ones that were prepared in the same library and normalized the CV of the MA lines by the difference in the generation number between the MA lines and the population. After this conservative calculation, the CV was higher than the population isolates across 25 of the 39 top kmers (File S4).

Next, we compared the realized mutation rate (the mutations that made it through the “filter” of selection in the non-MA population) to the mutation rate in the MA lines. We used the conservative calculation of 62 generations propagated in the population (File S3). The realized mutation rates in the population were lower than the MA lines for 32 of the top 38 kmers not including the telomere repeat (Figure 4D, $P < 2.43 \times 10^{-5}$, sign test). Since the sample

size of the non-MA population isolates is lower than the sample size of the MA lines (6 vs. 28), we performed 1000 subsampling replicates for each repeat by sampling six random MA lines and calculating the mean mutation rate. We found that 22 out of 38 repeats had a lower realized mutation rate in the population than the 5% quantile of the 1000 MA line subsample replicates (File S4). Moreover, some repeats seemed to be more constrained than others. We roughly quantified constraint by the difference between the mutation rate of the MA lines to the realized mutation rate of the population for each repeat. Five repeats experienced high levels of constraint by this measure, with the realized mutation rate being at least 75% lower in the population than the MA lines: AGG, C, AGCCTG, AAGCCAGTGCAGC, and AATCTGGAATGGAATGG. All of these highly constrained repeats had statistically significant lower realized mutation rates by the subsampling test above. On the other hand, nine repeats seem to have experienced little constraint in their expansions and contractions, with their realized mutation rates in the population being either slightly higher in the population or very close to that of the MA lines (Figure 4D and File S4). The level of constraint was not related to the copy number of the repeat (File S4).

Wei *et al.* (2014) noted strong patterns of correlation of kmer abundances across different populations of *D. melanogaster*, suggesting some sort of evolutionary nonindependence. These correlation patterns may suggest constraints on the mutational process, or they may be driven by selection. To investigate correlations in the expansions/contractions between repeats, we computed a correlation matrix (using

Figure 3 Kmer mutation rates. (A) Boxplot of mutation rates in the MA (mutation accumulation) lines ($n = 28$). The top 20 kmers are shown, in descending order of abundance. Negative values indicate contractions and positive values indicate expansions. (B) Plot of mean mutation rate across MA lines for each of 162 kmers with normalized abundance of at least two copies per $1\times$ coverage. Each point is a kmer, and indexes indicate descending order of abundance. The red line indicates an average balanced mutation rate of 0.

the Pearson correlation) between each repeat for the MA lines and the population. This would indicate correlations between changes in repeat copy number inherent to mutation as well as correlations caused by similar selective regimes among repeats. We found that some (mainly positive) correlations existed in the MA lines, but the correlations between repeat copy changes in the population were more common and stronger for both positive and negative correlations (Figure 5, A and B). A Mantel test indicated that these correlation

matrices were significantly different ($P < 0.001$), indicating a role of selection in interactions between repeat sequences.

For both the MA lines and the population, strong correlations between kmers that had GC contents between 0.6 and 0.7 were apparent. Although it is possible that the correlations reflect technical bias, we think this is unlikely because these strong correlations were between the nine repeats closely related in sequence (Figure 2). These nine 10–20 bp kmers have a GC content between 0.6 and 0.7, and their deviations

Table 2 New kmers generated *de novo* during the MA experiment

New kmer	Line(s)	Normalized abundance	Present in shorter arrays	Similar kmer, or common motifs	Abundance of similar kmer (copy number)
AAATG	C18	4	No	AATGG or AAAAG	18,853 or 471
ACCAG	C02	11	No	AACAG	6,846
AAACCCTAGTGGGTTGT	C02	14	No	AAC, TAGG, AGG	—
AAGCCACGC	C02	3	No	AAG, AGC, ACGC	—
ACAT	C35	3	Yes	—	—
AAACT	C13, C21	2, 3	No	AACCT	62,281
AATG	C12, C34	3, 5	Yes	AAAG or AAGG	708 or 11
AGAT	C20, CC8	3, 2	Yes	AGAA (AAAG)	708

Many new kmers are potentially generated from single-nucleotide mutations or rearrangements of existing kmers. It is indicated if the putative new kmer was present in shorter arrays (~25 bp) in the ancestral lineage.

were in fact strongly positively correlated (Figure 5C). Aggregations of repeats that are related in sequence might easily explain some of the patterns of correlation among repeats that we identified. To follow up on this, we analyzed the paired-end reads to measure the level of interspersion between different repeats and among the repeat with itself. All repeats were interspersed with themselves to some extent. On average across the top 39 kmers, 40% of paired-end reads contained tandem repeats by k-Seek's conditions on both mate pairs (File S6). This indicates that a high proportion of the repeat arrays we are studying are in large tandem blocks spanning at least the length of the sequenced fragments (~250 bp). The per kmer mutation rate was not related to the proportion of paired-end reads having repeats on both mate pairs for that particular kmer (File S4). Several groups of kmers were interspersed with each other (File S6). The telomere repeat, unsurprisingly, was not interspersed with any other of the top 39 kmers. The group of nine related repeats mentioned above were the most mutually interspersed as a group. Eight of the nine were interspersed with at least one other in the group, and the more abundant repeats of the group were interspersed with up to six other repeats within the group.

Discussion

To our knowledge, this is the first study that assays the mutation rates of genome-wide tandemly-repeated DNA using MA lines. The inclusion of data from a population that experienced selection also enables us to begin to detect whether selective forces act on long arrays of short tandem repeats. We found that mutation rates in tandem repeats are high, but they appear to be constrained by selection. We assumed that the population isolates maintained the ancestral repeat copy number when calculating mutation rates, which was reasonable given that the means of the population and MA lines were similar for most repeat sequences, and the population isolates showed narrow changes in repeat abundance. The high variance in repeat copy number across the MA lines resulted in high estimated mutation rates. We are confident that these results are biological and not due to technical biases because we normalized kmer counts by the GC-normalized

read coverage (see *Materials and Methods*) and, after normalization, there was no longer a correlation between sequencing depth and repeat copy number (Figure S4 in File S2). All MA lines and population isolates contained a similar proportion of kmer-containing reads to total reads (Figure S1 in File S2). There was also a high level of concordance in the kmer content between technical replicates. Another line of support that these data are reliable for this type of study is that the analysis of copy number variation of mapped reads revealed mutation rate estimates close to those of previous studies (F. J. J. Chain, J. M. Flynn, J. K. Bull, and M. E. Cristescu, unpublished results). Going forward, using PCR-free library preparation methods can reduce batch-based biases in sequencing results (K. Wei, S. Lower, T. J. Sless, D. A. Barbash, A. G. Clark, unpublished results). The tandem repeats that we are targeting here consist of long repetitive arrays, many of which are likely contained in satellite DNA in heterochromatic regions. In contrast, previous studies have used PCR primers to amplify short arrays of microsatellite loci (26–94 bp, Seyfert *et al.* 2008). Microsatellites have been thoroughly characterized in *D. pulex* using both cloning and reference genome based methods (Colbourne *et al.* 2004; Sung *et al.* 2010), but nearly all of the microsatellites identified were 1–3-mers repeated < 20 times. Since k-Seek requires tandem repeats to encompass at least 50 bp, most of these microsatellites were not included, setting our study apart from previous work. Moreover, on average 40% of reads originated from tandem arrays at least as long as the length of the sequenced fragment.

High mutation rates can explain the rapid turnover of short tandem repeats

The rates of mutation in tandem repeats in our *D. pulex* MA lines were extremely high, ranging from 0.29 to 105 copy changes per generation for a given kmer sequence. Our mutation rates include the sum of changes across all loci of a particular kmer, thus can include both loss/gain of entire repeat arrays (*i.e.*, from recombination) and changes in the number of repeats within individual loci (*i.e.*, from replication slippage). This also makes our study distinct from previous studies of microsatellite mutation rates that have focused on single loci (Seyfert *et al.* 2008). Here, mutation

Table 3 Kmers that were lost during the course of the MA experiment

Lost kmer	Line(s)	Present in shorter arrays	Normalized mean abundance in other lines (copy number)
AAGGAATGG	C40	No	61.6
AAAAAGAAGGAGATAG	C36	No	47.6
AATAG	C25	Yes	16.3
ACGCCCGAGC	C40	Yes	13.8
AATCT	C20	Yes	7.2
ACACCGACCACTACT	C40	Yes	5.6
AAAGGCAACAACAGT	C35	Yes	4.2
AATGT	C02, C34	Yes	12.4
AAAAAAAC	C39, C44	Yes	7.5

It is indicated if the putatively lost kmer is present in short arrays at low frequencies in the mutation accumulation (MA) line in question.

rates estimated on a per-copy basis for short tandem repeats that we examined were on the order of 10^{-3} copies/generation/copy. Seyfert *et al.* (2008) estimated the per locus mutation rate of *D. pulex* dinucleotide microsatellites 13–47 repeats long to be on the order of 10^{-5} to 10^{-4} copies/locus/generation. This indicates not only that the tandem repeat arrays we investigated are present as multiple loci and long arrays, but that they mutate at much higher rates than microsatellites previously studied. One contributing factor could be that the short microsatellite arrays are more constrained since they are often found in introns, coding regions, or within ribosomal gene spacer sequences (Colbourne *et al.* 2004; Ellegren 2004). Past studies of microsatellites have found that rates of mutation are positively correlated with the number of repeats, but limited data points prevented the determination of a linear relationship (Wierdl *et al.* 1997; Brinkmann *et al.* 1998). We had mutation rate estimates for 39 repeat sequences ranging in abundance from 100 to 50,000 copies and found a significant positive linear correlation between copy number and mutation rate. This suggests a simple relationship between the number of copy units present and the potential for copy units to be gained or lost, consistent with replication slippage being the dominant mechanism of mutation (Schlötterer 2000). We also found that expansions are more common than contractions, indicating that tandem repetitive DNA in *Daphnia* has the intrinsic tendency to expand, which is in agreement with the consensus finding in microsatellites across taxa [reviewed in Ellegren (2004)].

This high mutation rate demonstrates that there is a high potential for rapid turnover in tandem repeat sequences between species, and even substantial differences in abundance among diverged populations of the same species. Wei *et al.* (2014) used k-Seek to characterize the tandem repeat sequences in five different populations of *D. melanogaster*, distributed globally. They found that the overall repeat content varied by a factor of 2.5, accounting for about a 4 Mb difference. This initially surprising finding makes sense in light of our results that the MA lines that were only 82 generations diverged differed in their genome-wide repeat content by a factor of 1.67, accounting for ~0.67 Mb. Even the population isolates, which showed a reduction of variation in repeat content compared to the MA lines, varied in

their overall repeat amount by a factor of 1.27. This highlights the great potential for expansions and contractions in tandemly-repeated sequences, which can explain the high levels of differentiation in genome-wide repeat compositions between populations and related species (Wei *et al.* 2014; Subirana *et al.* 2015; K. Wei, S. Lower, T. J. Sless, D. A. Barbash, A. G. Clark, unpublished results).

Daphnia and *Drosophila* are both arthropods, although they are estimated to have diverged ~400 MYA (Rehm *et al.* 2011). Given the high differentiation observed between closely related species, the paucity of shared sequences and motifs between these taxa is not surprising. However, it is worth noting that the AGG and AAG motifs, as well as a couple of AG-rich repeats, were conserved between the two species. This conservation could be random or it could indicate a conserved functional role for AG-rich repeats; for example, involvement in binding conserved proteins. The GAGA factor is known to bind to AG-rich satellite sequences in *Drosophila*, although it has been hypothesized to have evolved in the ancestor of Diptera and Hymenoptera and is not thought to be present in *Daphnia* (Heger *et al.* 2013). However, it is possible that other proteins could be playing a role in binding to AG-rich satellites in *Daphnia*, similarly to *Drosophila*. Another possibility is that AG-rich repeat sequences result in a favorable DNA curvature (Palomeque and Lorite 2008). *D. melanogaster* populations were found to be enriched in kmers that were multiples of 5 bp long (*i.e.*, 5, 10, 15, and 20-mers) (Wei *et al.* 2014). Interestingly, we also found that *Daphnia* were enriched in kmers 5, 10, and 20 bp long. Although there were no 15-mers at high abundance (≥ 100 copies), 15-mers were enriched when considering the data set of 162 kmers. This pattern, conserved from *Drosophila* to *Daphnia*, could indicate conservation in the periodicity of DNA wrapping around histones (Lohe and Brutlag 1987).

Based on the new repeats generated *de novo* in the MA experiment and the relatedness of existing repeats, we suggest that novel kmers typically result from a point mutation in an already abundant kmer, followed by expansion of the new repeat. This is in agreement with the unpublished result (K. Wei, S. Lower, T. J. Sless, D. A. Barbash, and A. G. Clark), which found new repeats originating in different *Drosophila*

Figure 4 Reduced variation in kmer content in the population compared to the MA lines. (A) Total absolute amount of tandemly-repeated DNA sequences with length 1–20 bp in the MA lines vs. the population, including all 162 kmers with at least two copies per $1\times$ coverage per line. (B) Histogram of the cumulative copy changes across the top 39 kmers in the MA lines and population. (C) The CV across the top 39 kmers for the MA lines and the population. Kmers are ordered by copy number abundance. (D) The absolute normalized mutation rate (sum of expansions and contractions), divided by the initial copy number of the kmer. Data from the MA lines are shown in red and the population in blue. CV, coefficient of variation; MA, mutation accumulation.

species that also differed by only a single nucleotide from preexisting repeats. Additionally, arrays of tandem repeats can be generated when they expand from a sequence motif that is already present in only a few copies, as we suggest could have occurred for some of our new repeats (Table 2). Concordant with these hypotheses, we found that two repeats were gained independently in two independent MA lines, demonstrating the high potential for new repeat sequences to be generated from existing ones or from an existing motif, and that parallel independent gains of the same repeat sequence are possible. We detected the newly-generated repeats at low abundance, but after many more generations they would presumably have the potential to expand and achieve higher abundance in the genome. Most of the lost repeats were not completely lost, but contracted to be part of a short enough array to not be detected by k-Seek (Table 3). Some of the repeat losses we found likely resulted from deletions of segments of chromosomes. Flynn *et al.* (2017) found that the MA line C40 experiences homozygous deletions totaling ~ 100 Mb on chromosome 11. Here, we found that C40 lost three different repeat sequences, so we

think it is likely that these tandem repeats were present on the deleted regions of chromosome 11 (although these repeat sequences were not present in tandem in the genome assembly). In fact, C40 experienced a complicated recombination event that resulted in these deletions and complete loss of heterozygosity across the chromosome (Flynn *et al.* 2017). It is possible that the tandem repetitive DNA composition on this chromosome causes it to be a hotspot for structural rearrangements. Nevertheless, if most of our repeat losses resulted from deletions of large segments also containing functionally important regions, these would likely be purged from natural populations and thus would not be detectable in population data.

Selection on short tandem repeat content

In contrast to the MA lines, which accumulate mutations with minimal regard to their phenotypic consequences, we found that most repeats experienced constraint in their evolution in the non-MA population. The constraint was disproportionately distributed among repeat sequences, with some being unconstrained in these conditions and others being

Figure 5 Correlated changes in kmer copy number in (A) the MA (mutation accumulation) lines, and (B) the population. (C) shows the correlations between the family of 9 kmers that are closely related by sequence and GC content, which are situated at the bottom corner of the matrix in (B). Matrices were computed from a pairwise Pearson correlation matrix between kmers in the deviation of their copy number from the inferred ancestral copy number. Kmers are ordered by their GC content.

highly constrained. Both the coefficient of variation and mutation rates were lower for many repeats in the population. The population reproduced asexually and it had a census population size of 100–250, indicating that selection would have to be substantially strong to produce the signal that we observed. Our results suggest that tandem repeats have the intrinsic tendency to expand in the *Daphnia* genome, but the magnitude of expansions are reduced under a selection regime. It is possible that selection enforces an upper limit to the length of a tandem array (Stephan and Cho 1994). Although this study does not investigate specific functional roles of satellites, we recognize the potential for some satellites to have specific roles (e.g., Fry and Salser 1977; Palomeque and Lorite 2008; Plohl *et al.* 2008). This may be true for AG-rich satellites, which are known to bind proteins and are conserved in both *Drosophila* species and *D. pulex*

(Raff *et al.* 1994). The selection we infer might stem from conservation of a specific role for some satellites and/or deleterious effects of changing the copy number of repeats.

In calculating mutation rates, we assumed that the population maintained the ancestral kmer copy number. Concorant with this assumption, there was little variation in the total genome-wide amount of repeat sequences in the population. This would be consistent with the operation of stabilizing selection on overall repeated DNA content (implying that there is an optimal copy number and/or genomic arrangement). However, we note that it is possible that negative selection (purging repetitive DNA) could have been acting in the population and the lower realized mutation rates in the population was a result of mutation-selection balance (*i.e.*, mutation pressure to increase repeat content, selection to reduce it). Under stabilizing selection, the reduced realized

kmer mutation rates in the population could stem from selection on overall genome size or chromosome organization. Previous studies have shown the importance of satellites for maintaining chromatin organization and that genetic variation in satellite content can affect genome stability and chromosome function (Kim *et al.* 2009; Aldrup-MacDonald *et al.* 2016). If the genome has evolved to accommodate tandem repetitive arrays, altering the length of an array could alter the chromatin organization and disrupt nearby functional regions, causing deleterious effects. This could be especially important for *Daphnia*, which has a compact genome with little intergenic space (Colbourne *et al.* 2011). If selection to remove repeated DNA was the dominant form of selection, this would imply that the repetitive arrays were mostly junk and that the genomes would be in a constant state of purging tandem repetitive DNA. Moreover, it seems that the level of selection varies across repeats. We found five repeats to be under very little constraint, shown by the lack of difference between their CV and mutation rates between the MA lines and population (Figure 4, C and D).

Correlations in repeat copy number changes

Across the MA lines, we found mainly positive correlations between the kmer sequences in their expansions and contractions. These could have arisen randomly or been driven by physical location/linkage. We found that repeats that were closely related to each other by mutation were strongly correlated in their changes in abundance, which was especially evident for the family of 10–20 bp kmers with the GCCAG motif (Figure 2, top left). Concordant with the hypothesis that new repeat sequences come from point mutations in existing ones, we found that these closely related repeats were highly interspersed within each other. Since these positive correlations were present in both the MA lines and the population, it is likely to be a consequence of the mutational process. Segmental duplications and deletions, which are common in *Daphnia* (Keith *et al.* 2016), could explain the observed positive correlations. When only the population isolates were considered in the correlations, we identified more correlations of stronger effect, both positive and negative. The positive correlations could arise from the same selective regime acting on pairs of repeats, *i.e.*, if expansions of both repeats of the pair in question are both selected against. On the other hand, the negative correlations may arise if different repeat sequences are in conflict (Wei *et al.* 2014). Specifically, there could be selection on a chromosome to maintain its organization such that if one repeat happens to expand, another will have to contract. Repeats localized on the same chromosomal region may be the ones negatively correlated with each other; however, we were not able to test this thoroughly with our data because of the unmappable nature of repeats.

Conclusions

We were able to gain a picture of genome-wide short tandem repeat expansions and contractions, and provide evidence

that tandemly-repeated DNA is not always neutrally evolving but can experience strong selection. We also show evidence that new repeat sequences are often generated from existing ones, and there is a complex interaction structure between individual repeat sequences that differs if selection is at play. Future studies using different sequencing technologies, especially with longer reads that can capture longer repeat units and span the boundaries of unique and repeated sequences, will provide further insights into satellite DNA evolution.

Acknowledgments

We would like to thank D. Barbash and F. Chain for useful discussions, and K. Wei for providing scripts for the kmer interspersed analysis. J. Bull provided summary data from the life history experiment that allowed us to calculate the number generations in the non-mutation accumulation (MA) population. We would also like to thank S. Lower for helpful discussions on the analysis of the data and comments that helped improve the quality of the manuscript. The original MA lines and sequencing work was supported by a Natural Sciences and Engineering Research Council (NSERC) Discovery grant (341423-2012) to M.E.C. and the kmer analysis was supported by a National Institutes of Health grant R01 GM119125 to A.G.C. and D. Barbash. J.M.F. was supported by a NSERC Postgraduate Scholarship-Doctoral scholarship.

Literature Cited

- Aldrup-MacDonald, M. E., M. E. Kuo, L. L. Sullivan, K. Chew, and B. A. Sullivan, 2016 Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Res.* 26: 1301–1311.
- Benjamini, Y., and T. P. Speed, 2012 Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 40: e72.
- Brinkmann, B., M. Klitsch, F. Neuhuber, J. Hühne, and B. Rolf, 1998 Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.* 62: 1408–1415.
- Bzymek, M., and S. T. Lovett, 2001 Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. USA* 98: 8319–8325.
- Charlesworth, B., and D. Charlesworth, 2000 The degeneration of Y chromosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 355: 1563–1572.
- Charlesworth, B., P. Sniegowski, and W. Stephan, 1994 The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371: 215–220.
- Colbourne, J., B. Robison, K. Bogart, and M. Lynch, 2004 Five hundred and twenty-eight microsatellite markers for ecological genomic investigations using *Daphnia*. *Mol. Ecol. Resour.* 4: 485–490.
- Colbourne, J. K., M. E. Pfrender, D. Gilbert, W. K. Thomas, A. Tucker *et al.*, 2011 The ecoresponsive genome of *Daphnia pulex*. *Science* 331: 555–561.
- Ellegren, H., 2004 Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 5: 435–445.
- Fields, P. D., C. Reisser, M. Dukic, C. R. Haag, and D. Ebert, 2015 Genes mirror geography in *Daphnia magna*. *Mol. Ecol.* 24: 4521–4536.

- Flynn, J. M., F. J. Chain, D. J. Schoen, and M. E. Cristescu, 2017 Spontaneous mutation accumulation in *Daphnia pulex* in selection-free vs. competitive environments. *Mol. Biol. Evol.* 34: 160–173.
- Fry, K., and W. Salser, 1977 Nucleotide sequences of HS- α satellite DNA from kangaroo rat *Dipodomys ordii* and characterization of similar sequences in other rodents. *Cell* 12: 1069–1084.
- Heger, P., R. George, and T. Wiehe, 2013 Successive gain of insulator proteins in arthropod evolution. *Evolution* 67: 2945–2956.
- Henikoff, S., K. Ahmad, and H. S. Malik, 2001 The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293: 1098–1102.
- Hoskins, R. A., J. W. Carlson, C. Kennedy, D. Acevedo, M. Evans-Holm *et al.*, 2007 Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* 316: 1625–1628.
- Keith, N., A. E. Tucker, C. E. Jackson, W. Sung, J. I. Lucas Lledo *et al.*, 2016 High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. *Genome Res.* 26: 60–69.
- Kim, J. H., T. Ebersole, N. Kouprina, V. N. Noskov, J. Ohzeki *et al.*, 2009 Human gamma-satellite DNA maintains open chromatin structure and protects a transgene from epigenetic silencing. *Genome Res.* 19: 533–544.
- Lemos, B., L. O. Araripe, and D. L. Hartl, 2008 Polymorphic Y chromosomes harbor cryptic variation with manifold functional consequences. *Science* 319: 91–93.
- Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.
- Lohe, A. R., and D. L. Brutlag, 1987 Identical satellite DNA sequences in sibling species of *Drosophila*. *J. Mol. Biol.* 194: 161–170.
- MacGregor, H. C., and S. K. Sessions, 1986 The biological significance of variation in satellite DNA and heterochromatin in newts of the genus *Triturus*: an evolutionary perspective. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 312: 243–259.
- Malik, H. S., 2009 The centromere-drive hypothesis: a simple basis for centromere complexity, pp. 33–52 in *Centromere*, edited by Đ. Ugarković. Springer, Berlin.
- Orgel, L. E., and F. H. Crick, 1980 Selfish DNA: the ultimate parasite. *Nature* 284: 604–607.
- Paço, A., F. Adegá, N. Meštrović, M. Plohl, and R. Chaves, 2015 The puzzling character of repetitive DNA in Phodopus genomes (Cricetidae, Rodentia). *Chromosome Res.* 23: 427–440.
- Padeken, J., P. Zeller, and S. M. Gasser, 2015 Repeat DNA in genome organization and stability. *Curr. Opin. Genet. Dev.* 31: 12–19.
- Palomeque, T., and P. Lorite, 2008 Satellite DNA in insects: a review. *Heredity* 100: 564–573.
- Platero, J. S., A. K. Csink, A. Quintanilla, and S. Henikoff, 1998 Changes in chromosomal localization of heterochromatin-binding proteins during the cell cycle in *Drosophila*. *J. Cell Biol.* 140: 1297–1306.
- Plohl, M., A. Luchetti, N. Meštrović, and B. Mantovani, 2008 Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero) chromatin. *Gene* 409: 72–82.
- Raff, J. W., R. Kellum, and B. Alberts, 1994 The *Drosophila* GAGA transcription factor is associated with specific regions of heterochromatin throughout the cell cycle. *EMBO J.* 13: 5977–5983.
- Rehm, P., J. Borner, K. Meusemann, B. M. von Reumont, S. Simon *et al.*, 2011 Dating the arthropod tree based on large-scale transcriptome data. *Mol. Phylogenet. Evol.* 61: 880–887.
- Sahara, K., F. Marec, and W. Traut, 1999 TTAGG telomeric repeats in chromosomes of some insects and other arthropods. *Chromosome Res.* 7: 449–460.
- Schlötterer, C., 2000 Evolutionary dynamics of microsatellite DNA. *Chromosoma* 109: 365–371.
- Seyfert, A. L., M. E. Cristescu, L. Frisse, S. Schaack, W. K. Thomas *et al.*, 2008 The rate and spectrum of microsatellite mutation in *Caenorhabditis elegans* and *Daphnia pulex*. *Genetics* 178: 2113–2121.
- Simmons, M. J., and J. F. Crow, 1977 Mutations affecting fitness in *Drosophila* populations. *Annu. Rev. Genet.* 11: 49–78.
- Stephan, W., and S. Cho, 1994 Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. *Genetics* 136: 333–341.
- Subirana, J. A., M. M. Albà, and X. Messegue, 2015 High evolutionary turnover of satellite families in *Caenorhabditis*. *BMC Evol. Biol.* 15: 1.
- Sung, W., A. Tucker, R. D. Bergeron, M. Lynch, and W. K. Thomas, 2010 Simple sequence repeat variation in the *Daphnia pulex* genome. *BMC Genomics* 11: 691.
- Wei, K. H., J. K. Grenier, D. A. Barbash, and A. G. Clark, 2014 Correlated variation and population differentiation in satellite DNA abundance among lines of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* 111: 18793–18798.
- Wei, K. H., H. M. Reddy, C. Rathnam, J. Lee, D. Lin *et al.*, 2017 A pooled sequencing approach identifies a candidate meiotic driver in *Drosophila*. *Genetics* 206: 451–465.
- Wierdl, M., M. Dominska, and T. D. Petes, 1997 Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* 146: 769–779.

Communicating editor: W. Stephan