genes

MDPI

Article

Mode and tempo of microsatellite evolution across 300 million years of insect evolution

Michelle Jonika[1,2], Johnathan Lo[1], and Heath Blackmon[1,2]*

[1] Department of Biology, Texas A&M University, College Station, Texas, USA
[2] Genetics Interdisciplinary Program, Texas A&M University, College Station, Texas, USA
* Correspondence: coleoguy@gmail.com

Received: date; Accepted: date; Published: date

Abstract: Microsatellites are short, repetitive DNA sequences that can rapidly expand and contract due to slippage during DNA replication. Despite their impacts on transcription, genome structure, and disease, relatively little is known about the evolutionary dynamics of these short sequences across long evolutionary periods. To address this gap in our knowledge we have performed comparative analyses of 304 available insect genomes. We have investigated the impact of sequence assembly methods and assembly quality on the inference of microsatellite content and explored the influence of chromosome type and number on the tempo and mode of microsatellite evolution across one of the most speciose clades on the planet. Diploid chromosome number had no impact on the rate of microsatellite evolution or the amount of microsatellite content in genomes. We find that centromere type (holocentric or monocentric) is not associated with a difference in the amount of microsatellite content, but in those species with monocentric chromosomes microsatellite content tends to evolve faster than in species with holocentric chromosomes.

Keywords: Microsatellite evolution; insects; repetitive DNA; chromosome evolution; genome size; centromere

1. Introduction

Genomes contain a variety of sequence classes many of which are repetitive in nature. The smallest of these are microsatellites, simple sequence repeats that have a 2-6 base pair repeating motif. Microsatellites are highly polymorphic sequences that are most commonly found within non-coding portions of the genome, however, they can also be located in regulatory or intronic regions as well (Edwards *et al.* 1998; Metzgar *et al.* 2000). The location of microsatellites within the genome can have strong impacts on their stability. For instance, microsatellites occurring in regulatory or protein-coding regions tend to be highly conserved (Moore *et al.* 1999; Dokholyan *et al.* 2000). Similarly, conserved noncoding microsatellites which occur in the 5′ flanking regions of some protein coding genes in plants, are as their name suggests conserved among species possibly for their role in gene regulation (Fujimori *et al.* 2003; Zhang *et al.* 2006). In contrast, microsatellites found in regions without regulatory or coding functions (e.g. intergenic and some intronic regions) are likely to have little impact on organism fitness and thus their frequency and distribution should reflect the underlying mutation processes (Ellegren 2000).

Microsatellites have been useful to biologists as easily accessed genetic markers and some microsatellites have fundamental impacts on organismal functioning. Because of their relative neutrality microsatellites have been useful in inferences of population demography as well as genetic diversity (Slatkin 1995; Criscione *et al.* 2011). The large quantity and variability in microsatellites even within a species makes them useful for forensics (Ballantyne *et al.* 2010), kinship analysis (Blouin 2003), and medical profiling (Highnam *et al.* 2012). Microsatellites also play an important role in chromatin organization (Field D 1998), DNA structure (Pearson and Sinden 1996) and centromeres and telomere function (Schmidt and Heslop-Harrison 1996). However, the

Deleted: 303

Deleted: s

most studied effect of microsatellites is in the regulation of gene activity, where microsatellites can impact transcription (Hoffman *et al.* 1990), gene expression (Chamberlain *et al.* 1994), protein binding (Lue *et al.* 1989), and translation (Sandberg and Schalling 1997) thus leading to diseases (Rubinsztein *et al.* 1995b).

The primary mechanism for expansion and contraction of microsatellites is slippage that occurs during DNA replication (Eisen 1999; Klintschar *et al.* 2004). However, differential abundance of repeats in exonic, intronic, and intergenic regions among taxa may suggest that strand slippage alone is insufficient to explain microsatellite distribution (Tóth *et al.* 2000). While strand slippage can account for the expansion and contraction of microsatellite content, it cannot account for large shifts in the relative abundance of types of microsatellites in closely related species (e.g., a shift from AC repeats to TA repeats being the most common).

While previous studies have focused largely on specific classes of microsatellites in one or a handful of species, few studies have examined the dynamics of all microsatellite content across large clades (Bell and Ecker 1994; Neff and Gross 2001; but see Adams *et al.* 2016). Adams et al. showed that ray-finned fish, squamate reptiles, and mammalian genomes had higher microsatellite content, than crocodilian, turtle, and avian genomes. Additionally, some lineages had unusually high rates of change in microsatellite content providing support for multiple major shifts in the microsatellite genomic landscape. The goal of our study was to determine whether microsatellites evolve differently in different clades of insects and evaluate the impact of chromosome number, genome size, and centromere type (i.e. holocentric and monocentric) on both the content and rate of microsatellite evolution. Our analyses reveal that chromosome number has no impact on either content or rate of microsatellite evolution, and that centromere type has no impact on total microsatellite content. However, our study shows that different insect orders have significantly different rates and that the rate of microsatellite evolution is different among species with monocentric and holocentric chromosomes. Additionally, we find that genome size correlates with total content and rate of microsatellite evolution.

2. Materials and Methods

Sequence Data: We downloaded all available insect genome assemblies from NCBI, ENSMBL, and Baylor HGC (accession numbers and site addresses in Supplemental Table 1, accessed August 2018). A total of 304 genomes were available spanning 18 of the 24 insect orders. Six orders were represented by single species while Diptera and Hymenoptera were the most frequently sequenced orders with 116 and 71 species, respectively. In all cases, the most recent assembly with no masking was downloaded (accession numbers in Supplemental Table 1).

Assembly quality: Repetitive sequences are one of the central challenges in genome assembly, and because of this, it is possible that poorly assembled genomes or genomes assembled with shorter read technology will lead to inaccurate inference of microsatellite content. We took two approaches to assess and control for this possibility. First, we reanalyzed data from a survey of microsatellites across 71 vertebrates (Adams *et al.* 2016). In this analysis we categorized each genome assembly by the sequencing platform (short, Sanger, and long) and tested whether genomes in these three classes had significantly different microsatellite content. If mixed data was available for a genome assembly (e.g. long read sequencing with short read polishing) this was classified as long read for our categories. Second, we also evaluated the correlation between scaffold and contig N50 and total microsatellite content. Based on results from this analysis (described below) we chose to include all insect genomes regardless of sequencing platform or N50 statistics.

Preliminary inspection of our insect genomes suggested that some were highly incomplete (e.g., assembly size 2% of expected genome size). Because of this, we performed a second quality assessment comparing BUSCO scores and total microsatellite content in all insect genomes (Simão *et al.* 2015). We used default settings for BUSCO in conjunction with the insect gene set. Scores were calculated as the proportion of genes searched that were found as complete genes (in either single or duplicate copies). These scores were then compared to the total microsatellite content in each genome

Deleted: and

Deleted: repair errors

Deleted: 303

Deleted: complete

to determine if there was a cutoff below which microsatellites were poorly inferred. Based on this approach we reduced the number of genomes examined to 231.

Phylogenetic data: For downstream comparative analyses, we downloaded sequences for the four most frequently sequenced mitochondrial genes (12S, COI, COII, and cytochrome b) and four nuclear genes (18S, 28S, elongation factor 1, and arginine kinase). This yielded a dataset of 221 operational taxonomic units (OTUs) representing members of 12 of the 24 insect orders. All sequences were downloaded from GenBank (accession numbers in Supplemental Table 2). The sequences were aligned in MAFFT v.7 using default settings (Katoh *et al.* 2019). We used Gblocks 0.91b to remove ambiguously aligned sites from 12S, 18S, and 28S alignments, using options for less stringent selection including allowing smaller final blocks, gap positions within blocks, and less strict flanking requirements (Talavera and Castresana 2007). This resulted in alignments for 12s, 18s, and 28s of 346 bp, 1442 bp, and 253 bp in length, respectively. We used MEGA to manually adjust the alignments of protein coding genes (COI, COII, elongation factor 1, cytochrome b, and arginine kinase) to ensure that the reading frame was maintained (Tamura *et al.* 2011). These alignments were 1463 bp, 683 bp, 1064 bp, 409 bp, and 1019 bp in length, respectively. For tree inferences in RAxML, alignments were concatenated (total length 6686 bp) while each gene was kept separate for Bayesian tree inference.

Rogue taxa, or taxa which are placed inconsistently with equal probability during phylogenetic inference due to insufficient or erroneous data, will often lead to overestimation of rates of trail evolution, similar to what is seen in supertrees (Thomson and Shaffer 2010; Rabosky 2015). To avoid this problem we inferred 100 rapid bootstrap trees using RAxML-HPC v.8 on XSEDE using the CIPRES Science Gateway (Miller *et al.* 2010; Stamatakis 2014). Using these trees, we calculated taxon instability index with Mesquite v 3.6 (Maddison and Maddison 2007). A high index value, indicates a taxon has variable placement among trees. By visual inspection of this distribution, we found that 92% of taxa have indices less than 5000 but that above this, instability quickly increases (Supplemental Figure 1). To ensure that our estimate of rates was conservative, we removed the 18 taxa with scores higher than this cutoff. Filtering our dataset based on BUSCO scores (discussed below), gene sequence availability, and taxonomic instability index scores led to a final dataset of 201 taxa for our Bayesian analysis.

We used BEAST v2.5.2 for the inference of time-calibrated phylogenies (Drummond and Rambaut 2007). For a starting tree, we selected the best maximum likelihood tree from RAxML which we converted to an ultrametric tree using nonparametric rate smoothing implemented in the function chronos in the R package APE (Paradis *et al.* 2004). We assumed a relaxed log-normal clock, a GTR substitution model with among site rate variation modeled with a gamma distribution, and a birth-death branching model. We estimated nucleotide substitution model parameters independently across four partitions: protein codon positions 1, 2, 3, and the ribosomal positions. To calibrate divergence time estimation, we placed eight priors on node ages in the tree. Normal distributions with means and standard deviations were chosen to represent previous estimates of the ages of the root of the tree and the origin of: Lepidoptera, Diptera, Hymenoptera, Coleoptera, Blattodea/Phasmatodea, Hemiptera, and Ephemeroptera/Odonata (Supplemental Table 3) (Misof *et al.* 2014). Two independent BEAST runs were completed.

Microsatellite and other trait data: We used micRocounter v.1.1.0 to characterize the microsatellite content within the insect genome assemblies (Lo *et al.* 2019). We recorded the number of dinucleotide (2mer), trinucleotide (3mer), tetranucleotide (4mer), pentanucleotide (5mer) and hexanucleotide (6mer) repetitive sequences. Default micRocounter settings for all parameters were used (2mers required 6 repeats, 3mers required 4 repeats, while 4-6mers required 3 repeats). We used a publicly available dataset to gather centromere type (holocentric or monocentric) and chromosome number for as many species as possible in our study (Tree of Sex 2014; Blackmon *et al.* 2017). We additionally gathered available genome size estimates from the Animal Genome Size Database (Gregory 2020).

Estimating rates of microsatellite evolution: We fit Brownian motion models to estimate rates of microsatellite evolution at several levels. All rate estimates described were generated using the restricted maximum likelihood approach using the ace function in the R package APE (Paradis *et al.* 2004). This function takes observed microsatellite content and the phylogeny, and returns an

Deleted: Both MCMC chains converged on a parameter space with equal likelihood by 100 million generations. The chains were then allowed to run for an additional 100 million generations to ensure neither chain discovered an area of higher likelihood. Convergence was evaluated using Tracer v1.7.1 (Rambaut *et al.* 2018). The first 75% of each MCMC was discarded as burnin, and 50 trees were randomly sampled from the post-burnin portion of each MCMC. These 100 trees represent our posterior distribution of trees and were used in all downstream analyses. The phylogenies inferred were consistent with a comprehensive order level phylogeny (Misof *et al.* 2014). ¶

ancestral state estimate for every node in the tree and the maximum likelihood estimate for the rate of evolution. For comparison, we fit the same model using the fitContinuous function in the R package Geiger v2.0.6.4 (Harmon *et al.* 2008). Rate estimates between these two approaches were qualitatively identical.

First, we fit a model where we assumed a single rate of microsatellite evolution across the entire phylogeny. Next, we estimated rates individually for each order that had at least 10 species in our dataset (for both of these analyses we fit our model based on microsatellite bp per Mbp. Finally, we calculated tip rates. Using the ancestral state estimates from our combined analysis of all data, tip rates were estimated by taking the difference in microsatellite content of a species and the ancestral state estimate for the node from which it descends. This value represents the change since the last speciation event sampled on our phylogeny (this was calculated based on the total bp of microsatellites estimated for each tip in the tree). This value was then divided by the branch length since that speciation event providing an estimate for the recent rate of evolution in a species lineage.

The impact of centromere type on microsatellite content and evolution: We first tested whether species with holocentric and monocentric chromosomes have significantly different microsatellite content. We analyzed the quantity of each microsatellite size class (2-6mer), total microsatellite content, and microsatellite content per Mbp using a phylogenetic ANOVA implemented in Geiger (Harmon *et al.* 2008). The phylogenetic ANOVA was repeated for each tree from the posterior distribution. To calculate p-values the observed F-statistic was compared to a null distribution generated from 100 simulations.

In addition to differences in microsatellite content, type of centromere may also affect the rate at which microsatellite content evolves. We tested for a difference in the rate of microsatellite evolution in species with holocentric and monocentric chromosomes using a censored rate test implemented in the brownieREML function in phytools v0.6-99 (Revell 2012). This allows us to compare models where the continuous trait (microsatellite content) evolves at a single rate on all branches to a model where each state has independent rates of evolution (O'Meara et al. 2006). We used the function make.simmap in phytools to generate the stochastic maps (holocentric vs monocentric states) that are used in brownieREML (Revell 2012). In construction of the stochastic map, we used a Markov model and allowed rates of transition between holocentric and monocentric to differ. To account for uncertainty in ancestral states, we repeated our analysis across 100 stochastic maps.

Comparing rates and content to chromosome number and genome size: We hypothesized that if microsatellites are common in structural elements of chromosomes then those species with more chromosomes would be expected to have greater microsatellite content. We analyzed the data using a phylogenetic linear model where microsatellite content in bp of microsatellite content/Mbp of genome was the response variable and chromosome number was the predictor variable. We also fit a phylogenetic linear model where the tip rate as described above was the response variable and chromosome number was the predictor variable. Both of these models were fit using the function phylolm in the R package phylolm v2.6 and used all 100 posterior distribution trees with 100 bootstraps for each tree (Ho *et al.* 2018).

We also assessed genome size as a predictor of microsatellite evolution. For this analysis, we used genome size in Mbp. We analyzed the data using a phylogenetic linear model where microsatellite content in Mbp was the response variable and genome size was the predictor variable. This analysis used the 100 posterior distribution trees with 100 bootstraps for each tree. We also fit a phylogenetic linear model where the tip rate as described above was the response variable and genome size as the predictor variable. Again, this analysis used the 100 posterior distribution trees with 100 bootstraps for each tree. Both of these models were fit using the function phylolm in the R package phylolm. (Ho *et al.* 2018). All analyses were completed in R version 3.6.3 (R Core Team 2019). All tests were considered significant at $\alpha = 0.05$. All data and code necessary for our analyses are available on Dryad (available upon acceptance).

3. Results

3.1 Data quality and collection

Deleted: Tip

3.1.1 Assembly Quality: First, repetitive sequences were reanalyzed from a prior study with 71 vertebrates to determine the effect of sequencing platform (short, Sanger, and long) on microsatellite inference. We hypothesized that shorter sequencing may not be able to fully span repetitive regions generating a pattern of estimates of lower microsatellite content in genomes assembled from short read sequencing platforms. However, we found no significant difference in microsatellite content among genomes assembled with these three classes of platforms (Supplemental Figure 2). We also investigated the impact of N50 for both contigs and scaffolds on estimated microsatellite content. Using a linear model, we found no significant effect of either of these measures on microsatellite content (p-value for N50 contig = 0.70 and p-value for N50 scaffold = 0.18). Finally, we fit a linear model where microsatellite content was the response variable and genome size was the predictor variable and found no significance between them (p-value = 0.24; Supplemental Figure 3). These results suggest no strong biases in estimates of microsatellite content among these vertebrate genomes. While this suggests that it may be acceptable to use existing genome assemblies for comparative analyses, we feared that some of the 304 insect genomes we downloaded may be more poorly assembled than the vertebrate genomes we examined. For this reason, we investigated the relationship between BUSCO scores and estimated microsatellites content. We found an unexpected pattern where some genomes with very low BUSCO scores (e.g. less than 0.05) had the highest microsatellite content (Supplemental Figure 4). These low scoring genome assemblies also exhibited both some of the smallest and largest genome assemblies in our collection of genomes. Genomes with BUSCO scores greater than 0.1 do not appear to have any consistent bias in microsatellite content; however, we chose to conservatively retain only those genomes with BUSCO scores of at least 0.90 reasoning that these genomes were most likely the most well assembled in our dataset. Using this threshold, we discarded 83 genomes and all downstream analyses were performed on the remaining 221 genomes. As mentioned above this was further reduced to 201 for all analyses involving our phylogeny due to elevated taxonomic instability scores during tree inference.

3.1.2 Phylogenetic reconstruction: Both MCMC chains converged on a parameter space with equal likelihood by 100 million generations. The chains were then allowed to run for an additional 100 million generations to ensure neither chain discovered an area of higher likelihood. Convergence was evaluated using Tracer v1.7.1 (Rambaut *et al.* 2018). The first 75% of each MCMC was discarded as burnin, and 50 trees were randomly sampled from the post-burnin portion of each MCMC. These 100 trees represent our posterior distribution of trees and were used in all downstream analyses. The phylogenies inferred were consistent with a comprehensive order level phylogeny (Misof *et al.* 2014).

3.1.3 Microsatellite Content and rates**:** Microsatellite content for each type of microsatellite (2-mer, 3-mer, 4-mer, 5-mer, 6-mer) was measured for each of the species in our dataset (Figure 1). The highest total microsatellite content was found in *Blatella germanica* with a total of 19.03 Mbp of microsatellite content. The lowest microsatellite content was found in *Belgica antarctica* with a total of 0.17 Mbp of microsatellite content. Across all insects, we found that 2-mers are the most abundant type of microsatellite accounting for an average of 1.06 Mbp of the average insect genome assembly.

Our estimates of order rates revealed striking variation in rates of evolution (Figure 2A). Coleoptera exhibited the lowest rates of microsatellite evolution (σ^2=0.006x10^5) while Diptera and Hemiptera exhibited the highest rates of microsatellite evolution (σ^2=1.278x10^5 and 1.157x10^5 respectively). Next, we estimated tip rates of microsatellite evolution. Tip rates (in units of bp change per million years) for most species were normally distributed around zero. However, two

Deleted: Because of these peculiarities and because we have such a large number of genomes available, we chose to conservatively base all downstream analyses only on those genomes with a BUSCO score of 0.90 or higher (Supplemental Figure 4).

Deleted: are

hemipterans, *Pseudococcus longispinus* and *Paracoccus marginatus*, both exhibited strikingly negative tip rate values (-3.3x10^{-6} and -3.7x10^{-6} respectively). While alone these values may not seem striking, these numbers are 45 and 51 times larger respectively than the relative mean tip rate observed in our dataset. These two species also exhibited a considerably smaller genome size than is typical for hemipterans (average = 490 Mbp) with a genome size of 285 Mbp for *Pseudococcus longispinus* and 191 Mbp for *Paracoccus marginatus*.

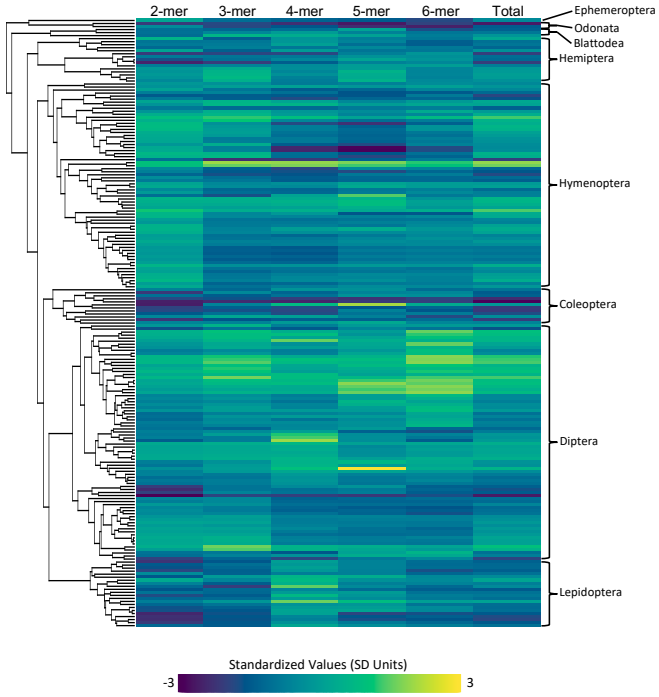

Figure 1. The inferred microsatellite counts for each type (2-mer, 3-mer, 4-mer, 5-mer, 6-mer, and total) with cool colors representing lower microsatellite content and warm colors representing higher microsatellite content. The values are standardized across all types. The phylogeny depicting the row that corresponds to each species can be seen on the left, and the orders that each of these species encompass is on the right.

3.2 Comparative analyses

3.2.1 Order: We first tested whether different orders of insects had different microsatellite content using both a standard and a phylogenetically corrected ANOVA. The response variables were the raw count of bp of each microsatellite type (2-6mer), the total sum of all microsatellites, and the proportion of the genome that is composed of microsatellite content. This proportion was calculated by taking the total raw microsatellite content divided by the assembly size. While we found that standard ANOVAs returned significant results for six of the seven response variables none of these

were significant after correction for phylogenetic history (Supplemental Table 4). This result is not unexpected when we examine the distribution of microsatellite proportions (Figure 2B).

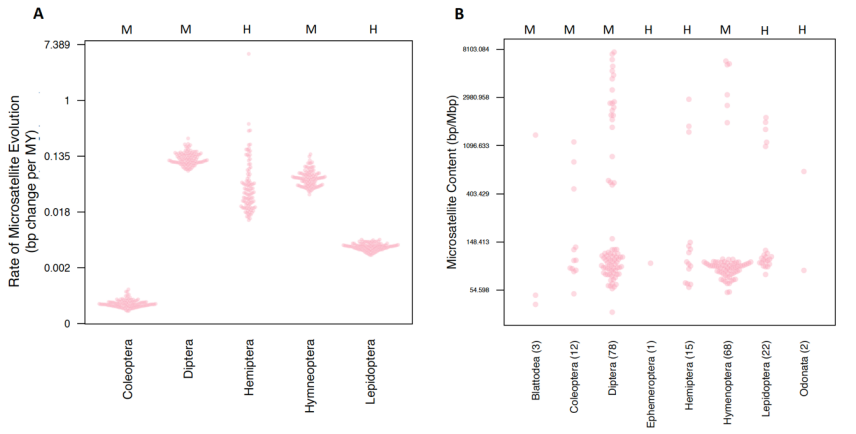

Figure 2. Comparing microsatellite content and rates of evolution among orders. Both y-axes are measured in a log scale. The centromere type present in an order is indicated with an H or M at the top of the plot for holocentric and monocentric respectively. Orders are indicated on the horizontal axis. **(a)** The rate of microsatellite evolution for all orders with at least ten representatives. For each order 100 estimates derived from each of the 100 trees is plotted. **(b)** Microsatellite content for all species included in comparative analyses.

3.2.2 Centromere Type: We reasoned that due to the repetitive nature of centromeric sequences, species with different centromere types may have distinct tempos and modes of evolution. Using stochastic maps of centromere type evolution, we performed a censored rate test to determine if rates of microsatellite evolution were significantly different in lineages with these two types of centromeres. Out of the 100 posterior distribution trees, 99 favored a two-rate model. The rate estimates for microsatellite evolution were higher in lineages with monocentric chromosomes for all trees including the one tree that did not support a two-rate model as significantly better (Figure 3A). We also compared the mean microsatellite content (2-6mer and total content) of lineages with monocentric and holocentric chromosomes using a standard and phylogenetically corrected ANOVA (Figure 3B). In no cases were monocentric and holocentric lineages significantly different (Supplemental Table 5).

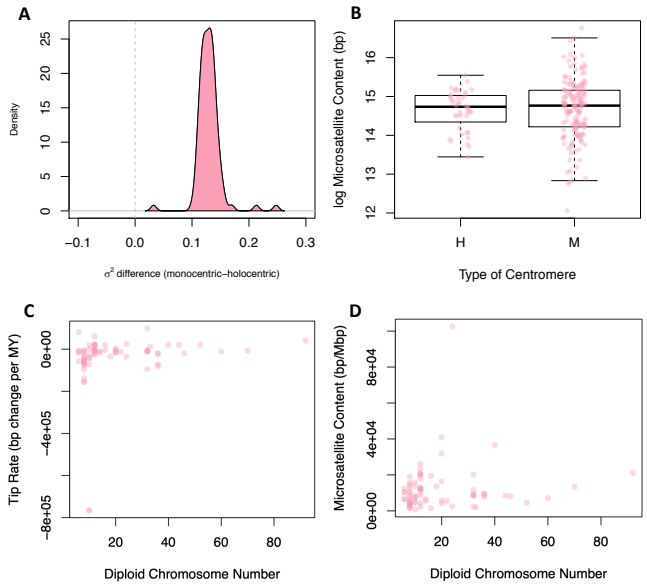

Figure 3. Comparing rates and content to centromere type and chromosome number. **(a)** The difference between the monocentric and holocentric rate predicted under a Brownian motion model for the 100 posterior distribution trees. **(b)** The difference between the microsatellite content in base pairs between holocentric and monocentric species. The y-axis is in the log scale. **(c)** The relationship between microsatellite evolution rates and the diploid chromosome number for each species. **(d)** The relationship between the microsatellite content in bp/Mbp and the diploid chromosome number for each of the species.

3.2.3 Chromosome Number and Genome Size: To test for a contingency between these traits we fit linear models where chromosome number or genome size were predictor variables and tip rates of microsatellite evolution or microsatellite content were response variables; in all cases, linear models were fit while correcting for phylogeny. We found no significant relationship between chromosome number and rates of microsatellite evolution (Figure 3c) or microsatellite content (Figure 3d). In contrast, when we tested genome size as the predictor variable for microsatellite content 96 of 100 phylogenies produced a significant result (Figure 4a). For 99 of the models, the slope of this relationship is positive indicating increased genome size is associate with increased microsatellite content. Likewise, when the response variable was rate of microsatellite evolution 53 out of 100 trees produced a significant result (Figure 4b). However, in this case, the increased genome size was associated with a very small decrease in the rate of microsatellite evolution in 84 of the 100 trees.

Deleted: 94

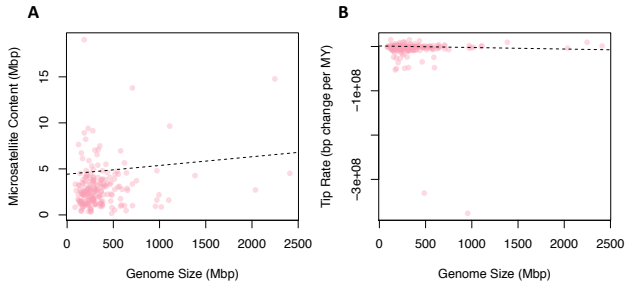

Figure 4. Comparing rates and content to genome size. **(a)** The relationship between microsatellite content in Mbp and the genome size in Mbp for each species. **(b)** The relationship between microsatellite evolution rates and the genome size in Mbp for each species.

4. Discussion

While most other studies have focused on microsatellite content and rates of evolution across specific taxa, we have investigated several possible predictors that may account for different microsatellite dynamics among orders of insects. We investigated the impact of chromosome number, centromere type, and genome size on microsatellite content and rates of evolution. We find significant differences in the microsatellite content among taxa, even those closely related. We also find that across all sequenced genomes microsatellite content scales with genome size. Finally, we show that species with monocentric chromosomes have significantly higher rates of microsatellite evolution.

The impacts of genome assembly quality: One of our initial questions in using genome assemblies to understand microsatellite evolution was the potential impact of assembly quality on downstream analyses. We addressed this question using two datasets. First, we analyzed data from a previous microsatellite study using vertebrate genomes (ADAMS *et al.* 2016). Then, we evaluated the inference of microsatellites across all sequenced insect genomes. Our results demonstrate that for the vertebrate data, microsatellite content inference is not biased by any quality metric we analyzed. In contrast, for insect genomes, we found that assemblies with very low BUSCO scores exhibited an exceedingly wide range of microsatellites. This pattern could be a result of some genomes with exceptionally large amounts of repetitive elements being both difficult to assemble and greatly enriched for microsatellites. However, the fact that genomes with these low BUSCO scores exhibit both higher and lower microsatellite content than is typical of well-assembled genomes suggests that this may be an artifact of a poor assembly. Furthermore, these species do not all have larger genomes than is typical for their clades which would be expected if the pattern is driven by a massive expansion of repetitive elements. Based on this logic, we excluded genomes with low BUSCO scores and suggest that future studies ensure that any genome used for investigations of microsatellite content have a BUSCO score in excess of 90 to ensure that the results reflect biological reality rather than the poor quality of the assembly.

Traits

Centromere type: Theoretical work has shown that microsatellites should expand in regions such as the centromere where there is suppression of recombination and weak selective pressure on the

array length (STEPHAN 1986). Two primary types of centromeres are present across the tree of life and within insects (MELTERS *et al.* 2012; BLACKMON *et al.* 2017). Holocentric chromosomes where centromere function is diffuse across the entire length of the chromosome and monocentric chromosomes where a single region of the chromosome functions as the centromere. These two types of centromeres lead to fundamentally different behavior with regard to resolution of chiasma that form during recombination. While monocentric chromosomes can segregate with multiple chiasma even in a single arm, holocentric chromosomes with more than one or two chiasma per chromosome fail to segregate properly (NOKKALA *et al.* 2004). Furthermore, centromeric regions in both types of chromosomes exhibit reduced recombination (CUACOS *et al.* 2015).This difference could lead to more regions of low recombination in holocentric species and greater opportunity for expansion of microsatellites relative to monocentric species. This might suggest that holocentric species would have larger genomes due to the proliferation of microsatellites and other repetitive sequences. However, analyses across a range of taxa do not support a difference in genome size between holo- and monocentric species (MANDRIOLI AND CARLO MANICARDI 2012). Another potential cause of differences in monocentric and holocentric species could be differences in DNA replication mechanics. Some evidence suggests that the distribution of translation initiation sites may vary based on centromere type (HECKMANN *et al.* 2013).

In some plants and nematodes, studies have suggested a higher satellite content in species with holocentric chromosomes (HECKMANN *et al.* 2013; SUBIRANA AND MESSEGUER 2013). However, the most taxonomically broad analysis of satellite content comparing holocentric and monocentric species supports lower satellite content in holocentric species (MELTERS *et al.* 2013). These studies have focused on all satellite content and the results are largely driven by mini- or macrosatellites. Our results suggest that within insects there is no significant difference in microsatellite content when comparing holocentric and monocentric species. This suggests that microsatellites do not play a central role in defining centromeric regions and that the selective forces constraining microsatellite expansion may be similar regardless of the centromere type.

Based on the similarity in microsatellite content that we observe among holocentric and monocentric lineages, we hypothesized that they would also demonstrate similar rates of evolution. However, when we fit a Brownian motion model of evolution to microsatellite content, we infer consistently higher rates in monocentric species than in holocentric species. Examining the total microsatellite content of all species in our study we can see that the monocentric orders Hymenoptera and Diptera both exhibit a range of microsatellite content that exceeds all holocentric orders combined (Figure 2B). We suggest that these orders likely drive much of the signal for higher rates of microsatellite evolution in monocentric species. However, not all monocentric clades exhibit high rates of microsatellite evolution; Coleoptera actually exhibits the lowest rate of microsatellite evolution of any order that we studied (Figure 2A). This highlights one of the dangers of comparative approaches that test for differences in rates of continuous trait evolution in two states of a discrete trait—a small portion of a phylogeny may contain such a strong signal that any binary trait mapped onto that portion of the phylogeny will be positively correlated with high rates. This is similar to the source of inflated rates of false positives that have recently caused upheaval in attempts to detect differential diversification under BiSSE models (RABOSKY AND GOLDBERG 2015). More broadly, the magnitude of variation we observe is striking and was not fully explained by any of the explanatory

variables that we examined. This suggests that additional factors must play an important role in determining the observed differences among species.

Chromosome number: Based on detailed analyses of Diptera genomes, microsatellites appear to be rare in heterochromatic portions of the genome (LOWENHAUPT *et al.* 1989; BACHTROG *et al.* 1999). However, heterochromatic regions are generally enriched for a variety of repetitive sequences (YUNIS AND YASMINEH 1971; CHARLESWORTH *et al.* 1994). Centromeres and telomeres as well as the regions adjacent to them are often heterochromatic and the number of these structural regions will scale with the number of chromosomes a genome contains. We might also expect chromosome number to correlate with rate of evolution where more recombination occurs in species with many chromosomes. With each recombination event, there is an opportunity for misalignment of repeat units which would lead to a longer or shorter locus in the resulting gametes. In our analysis, we find no significant correlation between chromosome number and either microsatellite content or microsatellite rate of evolution. We interpret this as evidence that these regions are not a "hot spot" for microsatellite accumulation in most insect species, and that changes in microsatellite length due to recombination errors are rare. However, we note that centromeric and telomeric regions are difficult to assemble regions of the genome and may become more difficult to assemble as the number of chromosomes increases. As such, the use of whole genome assemblies rather than raw reads may reduce our ability to detect a concentration of microsatellites in these regions. More broadly the results that we have presented are likely most applicable to the tempo and mode of microsatellite evolution in euchromatic portions of the genome.

Deleted: Centromeric

Deleted: telomeric

Deleted: both normally

Deleted:

Genome size: There is large variation in genome size among eukaryotes; even closely related species often have strikingly different genome sizes (HARTL 2000). This variation in eukaryotes is not predictive of complexity, ploidy level, or the number of protein-coding genes (MIRSKY AND RIS 1951; LYNCH AND CONERY 2003). The evolution of genome size can be directly affected by a variety of processes (e.g. insertions, deletions, polysomy, proliferation of transposons; reviewed in: PETROV 2001). A correlation between microsatellite content and genome size may be produced in two distinctly different fashions. First, microsatellite expansion or contraction may lead directly to changes in genome size, or broader deletion or insertion processes may drive a global change in genome size that impacts microsatellites as a byproduct. The relationship between microsatellite content and genome size has been confirmed in many species (PRIMMER *et al.* 1997; FIELD AND WILLS 1998; KUBIS *et al.* 1998). Although our study cannot distinguish among the possible drivers of this correlation, we do find a signal for contingency among genome size and microsatellite content (Figure 4A). Furthermore, we find a correlation between the rate of microsatellite evolution and genome size. We interpret this as support for the proportional model of genome size evolution (OLIVER *et al.* 2007). This hypothesis suggests that variation in rates of genome size evolution (and in turn genome size) is driven by broad rate differences that are a function of the genome size such that species with large genomes also have higher rates of genome size evolution.

Clades

Content: Our analysis demonstrates that microsatellite content is highly variable across species (Figure 1). However, differences among orders are not statistically significant once we correct for the phylogenetic history. Instead, we find a pattern where often even closely related species exhibit striking differences in microsatellite content. For instance, within the fly family Tephritidae, *Ceratitis capitata* has 10x more 5-mer repeats than the four *Bactrocera* species. *Ceratitis capitata* even has three

times more 5-mer repeats than *Rhagoletis zephyria* despite the *Rhagoletis* species having 30% more total microsatellite content. This pattern suggests that microsatellite content can evolve rapidly and confirms previous studies that have suggested that species often have unique microsatellite landscapes (RUBINSZTEIN *et al*. 1995a).

Rates: Although we found that clades with monocentric chromosomes have higher rates of microsatellite evolution, they also exhibited striking variation in rates of evolution. In fact, among all orders evaluated the monocentric orders Diptera and Coleoptera had the highest and lowest rate of microsatellite evolution respectively (Supplemental Figure 5). In contrast, both holocentric orders had intermediate rates. The broad range of rates estimated within monocentric clades suggests that monocentricity is unlikely to be the driving force responsible for rate differences among clades. Instead, we suggest that there must be other traits that are present in some monocentric clades that contribute to a higher rate estimate in lineages with this type of chromosome.

5. Conclusions

The molecular mechanism by which microsatellite content evolves is well-understood; however, the drivers of variation in patterns of microsatellite evolution across large clades is not. This study is to our knowledge the first to test how centromere type, chromosome number, and genome size impact clade-level microsatellite content and rates of evolution. Our results show that there is large variation both in microsatellite content and type of microsatellite repeats within and among orders. Furthermore, the rates at which this microsatellite content evolves differs among orders and centromere type. Based on our study, we suggest that Coleoptera and Diptera are particularly good clades to compare as they exhibit the largest difference in rates of evolution. The advent of long read sequencing technology coupled with approaches that provide genome wide scaffolding will lead to a vast increase in the number and completeness of genomes that are publicly available. These new assemblies will have the potential to evaluate the evolution of microsatellites across the entire genome rather than being concentrated in euchromatic portions as is currently the case. Approaches like those we have used that allow for an evolving trait to impact the rate of evolution of a second trait could leverage these genomes to reveal the impact of a broad range of characters (e.g. TEs, structural elements, codon bias, recombination rates) on the evolution of microsatellite content.

Deleted: The molecular mechanism by which microsatellite content evolves is well-understood; however, the drivers of variation in patterns of microsatellite evolution across large clades is not.

Deleted: .Our results show that there is large variation both in microsatellite content and type of microsatellite repeats within and among orders. Furthermore, the rates at which this microsatellite content evolves differs among orders. Based on our study, we suggest that Coleoptera and Diptera are particularly good clades to compare as they exhibit the largest difference in rates of evolution. As more insect genome assemblies become available and the quality of assemblies increases, future studies could use approaches similar to ours to understand the impact of a variety of genomic characters (i.e., codon bias, transposable elements, recombination rates) on microsatellite landscapes.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Taxonomic instability indices; Figure S2: Microsatellite inference and the type of sequencing; Figure S3: Measures of the quality of genomes; Figure S4: BUSCO scores; Figure S5: Tip rates and Table S1: Genome accession numbers; Table S2: Gene accession numbers Table S3: Node constraints; Table S4: ANOVA results microsatellite content by order; Table S5 ANOVA results microsatellite content by centromere type.

Author Contributions: All authors contributed to all phases of this study and have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Institute of General Medical Sciences at the National Institutes of Health, grant number R35GM138098.

Acknowledgments: We thank members of the Blackmon Lab and Claudio Casola for discussions of this work.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

518 Adams, R. H., H. Blackmon, J. Reyes-Velasco, D. R. Schield, D. C. Card *et al.*, 2016 Microsatellite landscape
519 evolutionary dynamics across 450 million years of vertebrate genome evolution. Genome 59**:** 295-310.
520 Bachtrog, D., S. Weiss, B. Zangerl, G. Brem and C. Schlötterer, 1999 Distribution of dinucleotide microsatellites
521 in the Drosophila melanogaster genome. Molecular Biology and Evolution 16**:** 602-610.
522 Ballantyne, K. N., M. Goedbloed, R. Fang, O. Schaap, O. Lao *et al.*, 2010 Mutability of Y-chromosomal
523 microsatellites: rates, characteristics, molecular bases, and forensic implications. The American Journal
524 of Human Genetics 87**:** 341-353.
525 Bell, C. J., and J. R. Ecker, 1994 Assignment of 30 microsatellite loci to the linkage map of Arabidopsis. Genomics
526 19**:** 137-144.
527 Blackmon, H., L. Ross and D. Bachtrog, 2017 Sex Determination, Sex Chromosomes, and Karyotype Evolution
528 in Insects. Journal of Heredity 108**:** 78-93.
529 Blouin, M. S., 2003 DNA-based methods for pedigree reconstruction and kinship analysis in natural populations.
530 Trends in Ecology & Evolution 18**:** 503-511.
531 Chamberlain, N. L., E. D. Driver and R. L. Miesfeld, 1994 The length and location of CAG trinucleotide repeats
532 in the androgen receptor N-terminal domain affect transactivation function. Nucleic Acids Research 22**:**
533 3181-3186.
534 Charlesworth, B., P. Sniegowski and W. Stephan, 1994 The evolutionary dynamics of repetitive DNA in
535 eukaryotes. Nature 371**:** 215-220.
536 Criscione, C. D., R. Vilas, E. Paniagua and M. S. Blouin, 2011 More than meets the eye: detecting cryptic
537 microgeographic population structure in a parasite with a complex life cycle. Molecular Ecology 20**:**
538 2510-2524.
539 Cuacos, M., H. Franklin, F. Chris and S. Heckmann, 2015 Atypical centromeres in plants—what they can tell us.
540 Frontiers in plant science 6**:** 913.
541 Dokholyan, N. V., S. V. Buldyrev, S. Havlin and H. E. Stanley, 2000 Distributions of dimeric tandem repeats in
542 non-coding and coding DNA sequences. Journal of Theoretical Biology 202**:** 273-282.
543 Drummond, A. J., and A. Rambaut, 2007 BEAST: Bayesian evolutionary analysis by sampling trees. BMC
544 evolutionary biology 7**:** 214.
545 Edwards, Y. J. K., G. Elgar, M. S. Clark and M. J. Bishop, 1998 The identification and characterization of
546 microsatellites in the compact genome of the japanese pufferfish, Fugu rubripes: perspectives in
547 functional and comparative genomic analyses. Journal of Molecular Biology 278**:** 843-854.
548 Eisen, J. A., 1999 Mechanistic basis for microsatellite instability. Microsatellites : evolution and applications**:** 34-
549 48.
550 Ellegren, H., 2000 Microsatellite mutations in the germline:: implications for evolutionary inference. Trends in
551 Genetics 16**:** 551-558.
552 Field D, W. C., 1998 Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different
553 distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutation
554 pressures and a variety of selective forces. Proceedings of the National Academy of Sciences of the
555 United States of America (USA) 95**:** 1647-1652.
556 Field, D., and C. Wills, 1998 Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the
557 different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong
558 mutation pressures and a variety of selective forces. Proceedings of the National Academy of Sciences
559 95**:** 1647-1652.

560 Fujimori, S., T. Washio, K. Higo, Y. Ohtomo, K. Murakami *et al.*, 2003 A novel feature of microsatellites in plants:
561 a distribution gradient along the direction of transcription. FEBS Letters 554**:** 17-22.
562 Gregory, T. R., 2020 Animal Genome Size Database. 2020. See http://www. genomesize. com.
563 Harmon, L. J., J. T. Weir, C. D. Brock, R. E. Glor and W. Challenger, 2008 GEIGER: investigating evolutionary
564 radiations. Bioinformatics 24**:** 129-131.
565 Hartl, D. L., 2000 Molecular melodies in high and low C. Nature Reviews Genetics 1**:** 145-149.
566 Heckmann, S., J. Macas, K. Kumke, J. Fuchs, V. Schubert *et al.*, 2013 The holocentric species L uzula elegans
567 shows interplay between centromere and large-scale genome organization. The Plant Journal 73**:** 555-
568 565.
569 Highnam, G., C. Franck, A. Martin, C. Stephens, A. Puthige *et al.*, 2012 Accurate human microsatellite genotypes
570 from high-throughput resequencing data using informed error profiles. Nucleic Acids Research 41**:** e32-
571 e32.
572 Ho, L. S. T., C. Ane, R. Lachlan, K. Tarpinian, R. Feldman *et al.*, 2018 Package 'phylolm', pp.
573 Hoffman, E. K., S. P. Trusko, M. Murphy and D. L. George, 1990 An S1 nuclease-sensitive
574 homopurine/homopyrimidine domain in the c-Ki-ras promoter interacts with a nuclear factor.
575 Proceedings of the National Academy of Sciences 87**:** 2705.
576 Katoh, K., J. Rozewicki and K. D. Yamada, 2019 MAFFT online service: multiple sequence alignment, interactive
577 sequence choice and visualization. Briefings in bioinformatics 20**:** 1160-1166.
578 Klintschar, M., E. M. Dauber, U. Ricci, N. Cerri, U. D. Immel *et al.*, 2004 Haplotype studies support slippage as
579 the mechanism of germline mutations in short tandem repeats. Electrophoresis 25**:** 3344-3348.
580 Kubis, S., T. Schmidt and J. S. Heslop-Harrison, 1998 Repetitive DNA elements as a major component of plant
581 genomes. Annals of Botany 82**:** 45-55.
582 Lo, J., M. M. Jonika and H. Blackmon, 2019 micRocounter: Microsatellite Characterization in Genome
583 Assemblies. G3: Genes, Genomes, Genetics 9**:** 3101-3104.
584 Lowenhaupt, K., A. Rich and M. Pardue, 1989 Nonrandom distribution of long mono-and dinucleotide repeats
585 in Drosophila chromosomes: correlations with dosage compensation, heterochromatin, and
586 recombination. Molecular and Cellular Biology 9**:** 1173-1182.
587 Lue, N. F., A. R. Buchman and R. D. Kornberg, 1989 Activation of yeast RNA polymerase II transcription by a
588 thymidine-rich upstream element in vitro. Proceedings of the National Academy of Sciences 86**:** 486.
589 Lynch, M., and J. S. Conery, 2003 The origins of genome complexity. science 302**:** 1401-1404.
590 Maddison, W., and D. Maddison, 2007 Mesquite: a modular system for evolutionary analysis. 2011. See
591 http://mesquiteproject. org.
592 Mandrioli, M., and G. Carlo Manicardi, 2012 Unlocking holocentric chromosomes: new perspectives from
593 comparative and functional genomics? Current genomics 13**:** 343-349.
594 Melters, D. P., K. R. Bradnam, H. A. Young, N. Telis, M. R. May *et al.*, 2013 Comparative analysis of tandem
595 repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biology
596 14**:** R10.
597 Melters, D. P., L. V. Paliulis, I. F. Korf and S. W. Chan, 2012 Holocentric chromosomes: convergent evolution,
598 meiotic adaptations, and genomic analysis. Chromosome Research 20**:** 579-593.
599 Metzgar, D., J. Bytof and C. Wills, 2000 Selection against frameshift mutations limits microsatellite expansion in
600 coding DNA. Genome Research 10**:** 72-80.
601 Miller, M. A., W. Pfeiffer and T. Schwartz, 2010 Creating the CIPRES Science Gateway for inference of large
602 phylogenetic trees, pp. 1-8 in *2010 gateway computing environments workshop (GCE)*. Ieee.

603 Mirsky, A., and H. Ris, 1951 The desoxyribonucleic acid content of animal cells and its evolutionary significance.
604 The Journal of general physiology 34**:** 451.
605 Misof, B., S. Liu, K. Meusemann, R. S. Peters, A. Donath *et al.*, 2014 Phylogenomics resolves the timing and
606 pattern of insect evolution. Science 346**:** 763-767.
607 Moore, H., P. W. Greenwell, C.-P. Liu, N. Arnheim and T. D. Petes, 1999 Triplet repeats form secondary
608 structures that escape DNA repair in yeast. Proceedings of the National Academy of Sciences 96**:** 1504.
609 Neff, B. D., and M. R. Gross, 2001 Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats.
610 Evolution 55**:** 1717-1733.
611 Nokkala, S., V. Kuznetsova, A. Maryanska-Nadachowska and C. Nokkala, 2004 Holocentric chromosomes in
612 meiosis. I. Restriction of the number of chiasmata in bivalents. Chromosome Research 12**:** 733-739.
613 Oliver, M. J., D. Petrov, D. Ackerly, P. Falkowski and O. M. Schofield, 2007 The mode and tempo of genome size
614 evolution in eukaryotes. Genome research 17**:** 594-601.
615 Paradis, E., J. Claude and K. Strimmer, 2004 APE: analyses of phylogenetics and evolution in R language.
616 Bioinformatics 20**:** 289-290.
617 Pearson, C. E., and R. R. Sinden, 1996 Alternative structures in duplex DNA formed within the trinucleotide
618 repeats of the myotonic dystrophy and fragile X loci. Biochemistry 35**:** 5041-5053.
619 Petrov, D. A., 2001 Evolution of genome size: new approaches to an old problem. TRENDS in Genetics 17**:** 23-
620 28.
621 Primmer, C. R., T. Raudsepp, B. P. Chowdhary, A. P. Møller and H. Ellegren, 1997 Low frequency of
622 microsatellites in the avian genome. Genome Research 7**:** 471-482.
623 R Core Team, 2019 R: A Language and Environment for Statistical Computing, pp. R Foundation for Statistical
624 Computing Vienna, Austria.
625 Rabosky, D. L., 2015 No substitute for real data: a cautionary note on the use of phylogenies from birth–death
626 polytomy resolvers for downstream comparative analyses. Evolution 69**:** 3207-3216.
627 Rabosky, D. L., and E. E. Goldberg, 2015 Model inadequacy and mistaken inferences of trait-dependent
628 speciation. Syst Biol 64**:** 340-355.
629 Rambaut, A., A. J. Drummond, D. Xie, G. Baele and M. A. Suchard, 2018 Posterior Summarization in Bayesian
630 Phylogenetics Using Tracer 1.7. Systematic Biology 67**:** 901-904.
631 Revell, L. J., 2012 phytools: an R package for phylogenetic comparative biology (and other things). Methods in
632 ecology and evolution 3**:** 217-223.
633 Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, S. Jain *et al.*, 1995a Microsatellite evolution—evidence for
634 directionality and variation in rate between species. Nature genetics 10**:** 337-343.
635 Rubinsztein, D. C., J. Leggo, G. A. Coetzee, R. A. Irvine, M. Buckley *et al.*, 1995b Sequence variation and size
636 ranges of CAG repeats in the Machado-Joseph disease, spinocerebellar ataxia type 1 and androgen
637 receptor genes. Human Molecular Genetics 4**:** 1585-1590.
638 Sandberg, G., and M. Schalling, 1997 Effect of in vitro promoter methylation and CGG repeat expansion on FMR-
639 1 expression. Nucleic Acids Research 25**:** 2883-2887.
640 Schmidt, T., and J. S. Heslop-Harrison, 1996 The physical and genomic organization of microsatellites in sugar
641 beet. Proceedings of the National Academy of Sciences 93**:** 8761.
642 Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva and E. M. Zdobnov, 2015 BUSCO: assessing
643 genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31**:** 3210-
644 3212.

Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics 139**:** 457-462.

Stamatakis, A., 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30**:** 1312-1313.

Stephan, W., 1986 Recombination and the evolution of satellite DNA. Genetics Research 47**:** 167-174.

Subirana, J. A., and X. Messeguer, 2013 A satellite explosion in the genome of holocentric nematodes. PLoS One 8.

Talavera, G., and J. Castresana, 2007 Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic biology 56**:** 564-577.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei *et al.*, 2011 MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular biology and evolution 28**:** 2731-2739.

Thomson, R. C., and H. B. Shaffer, 2010 Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Systematic biology 59**:** 42-58.

Tóth, G., Z. Gáspári and J. Jurka, 2000 Microsatellites in different eukaryotic genomes: survey and analysis. Genome research 10**:** 967-981.

Tree of Sex, C., 2014 Tree of Sex: A database of sexual systems. Scientific Data.

Yunis, J. J., and W. G. Yasmineh, 1971 Heterochromatin, satellite DNA, and cell function. Science 174**:** 1200-1209.

Zhang, L., K. Zuo, F. Zhang, Y. Cao, J. Wang *et al.*, 2006 Conservation of noncoding microsatellites in plants: implication for gene regulation. BMC Genomics 7**:** 323.