NRC Research Press

ARTICLE

Microsatellite landscape evolutionary dynamics across 450 million years of vertebrate genome evolution

Richard H. Adams, Heath Blackmon, Jacobo Reyes-Velasco, Drew R. Schield, Daren C. Card, Audra L. Andrew, Nyimah Waynewood, and Todd A. Castoe

Abstract: The evolutionary dynamics of simple sequence repeats (SSRs or microsatellites) across the vertebrate tree of life remain largely undocumented and poorly understood. In this study, we analyzed patterns of genomic microsatellite abundance and evolution across 71 vertebrate genomes. The highest abundances of microsatellites exist in the genomes of ray-finned fishes, squamate reptiles, and mammals, while crocodilian, turtle, and avian genomes exhibit reduced microsatellite landscapes. We used comparative methods to infer evolutionary rates of change in microsatellite abundance across vertebrates and to highlight particular lineages that have experienced unusually high or low rates of change in genomic microsatellite abundance. Overall, most variation in microsatellite content, abundance, and evolutionary rate is observed among major lineages of reptiles, yet we found that several deeply divergent clades (i.e., squamate reptiles and mammals) contained relatively similar genomic microsatellite compositions. Archosauromorph reptiles (turtles, crocodilians, and birds) exhibit reduced genomic microsatellite content and the slowest rates of microsatellite evolution, in contrast to squamate reptile genomes that have among the highest rates of microsatellite evolution. Substantial branch-specific shifts in SSR content in primates, monotremes, rodents, snakes, and fish are also evident. Collectively, our results support multiple major shifts in microsatellite genomic landscapes among vertebrates.

Key words: comparative genomics, microsatellite seeding, repeat elements, tandem repeats, simple sequence repeats.

Résumé : La dynamique évolutive des séquences simples répétées (SSR ou microsatellites) au sein des vertébrés demeure largement non-documentée et mal connue. Dans ce travail, les auteurs ont analysé l'abondance et l'évolution des microsatellites chez 71 génomes de vertébrés. Les abondances de microsatellites les plus grandes ont été rencontrées au sein des génomes des poissons dotés de nageoires à rayons, des reptiles à écailles et des mammifères, tandis que les génomes des crocodiliens, des tortues et des oiseaux présentaient le moins de microsatellites. Les auteurs ont employé des méthodes comparatives pour inférer les taux de changements évolutifs en matière d'abondance des microsatellites parmi les vertébrés et pour mettre en évidence des lignages qui ont connu des taux exceptionnellement élevés ou faibles. Globalement, la majeure partie de la variation en matière de contenu, d'abondance et de taux d'évolution des microsatellites a été observée au sein des grands groupes de reptiles, et pourtant, les auteurs ont noté que certains clades très distants (p. ex. les reptiles à écailles et les mammifères) présentaient des compositions génomiques semblables en matière de microsatellites. Les reptiles archosauromorphes (tortues, crocodiliens et oiseaux) affichaient une teneur réduite en microsatellites et les plus faibles taux évolutifs au sein des microsatellites, tandis que les génomes des reptiles à écailles avaient des taux parmi les plus élevés. Des variations substantielles limitées à certaines branches spécifiques ont également été observées chez les primates, les monotrèmes, les rongeurs, les serpents et les poissons. Ensemble, ces résultats appuient l'hypothèse de multiples changements importants dans le paysage des microsatellites chez les vertébrés. [Traduit par la Rédaction]

Mots-clés : génomique comparée, ensemencement de microsatellites, éléments répétés, répétitions en tandem, séquences simples répétées.

Received 16 September 2015. Accepted 13 February 2016.

Corresponding Editor: V. Katju.

R.H. Adams, J. Reyes-Velasco,* D.R. Schield, D.C. Card, A.L. Andrew, N. Waynewood, and T.A. Castoe. Department of Biology, 501 S. Nedderman Dr., University of Texas at Arlington, TX 76019, USA.
H. Blackmon. Department of Ecology, Evolution & Behavior, 1987 Upper Buford Cir., University of Minnesota, Saint Paul, MN 55108-6097, USA.
Corresponding author: Todd A. Castoe (email: todd.castoe@uta.edu).
*Present address: Department of Biology, New York University Abu Dhabi, Saadiyat Island, UAE.

Published at www.nrcresearchpress.com/gen on 1 March 2016.

Introduction

Microsatellites, also known as simple sequence repeats (SSRs), are 2–6 base pair (bp) nucleotide motifs that are tandemly repeated to form larger DNA stretches usually between 20 and 100 bp in length (Beckmann and Weber 1992). These non-coding elements are abundant in most eukaryotic genomes and may be found in both coding and non-coding regions (Tóth et al. 2000), although they appear to be comparatively rare in coding regions (Hancock 1995). Due to their high mutation rates and genome-wide distribution, microsatellites may significantly impact genome evolution through a number of mechanisms (Charlesworth et al. 1994), including potential effects on recombination (Pardue et al. 1987; Richard and Pâques 2000), regulation of gene expression (Martin et al. 2005; Moxon et al. 1994), gene conversion (Balaresque et al. 2014), and chromosomal organization (Pardue et al. 1987; reviewed in Li et al. 2002).

Since their discovery in the 1970s, the rapid mutation rates of microsatellites have made them an important workhorse for population genetic inferences and fine-scale estimation of relatedness among individuals, and in population biology in general (Bruford and Wayne 1993; Goldstein et al. 1999; Jarne and Lagoda 1996; Slatkin 1995). They are used as standard markers for human DNA profiling and are applied extensively in forensic sciences (Butler 2006). As molecular markers, microsatellites have increased our understanding of quantitative trait evolution (McCouch et al. 1997) and have been used in numerous association studies to map the genomic location of linked functional mutations in various diseases, including Alzheimer's and Parkinson's diseases (Kimpara et al. 1997). Additionally, microsatellites have received considerable attention for their causative associations with several cancers (Arzimanoglou et al. 1998; Wooster et al. 1994) and a number of human neurodegenerative disorders (La Spada et al. 1994; Ranum and Day 2002).

Microsatellites are inherently unstable and highly polymorphic at both species and population scales. This high rate of polymorphism is due to their high mutation rates, between 10^{-2} and 10^{-6} mutations per site per generation, which are several orders of magnitude higher than background nucleotide mutation rates (Schlötterer 2000). DNA polymerase strand slippage, which involves strand dissociation followed by misaligned re-association of the strands during DNA replication, has been proposed as the primary mechanism that generates mutation and variability in microsatellite loci (Richards and Sutherland 1994; Schlötterer 2000). The resulting misaligned sequences can produce volatile intrastrand structures that impact sequence stability and may even result in genetic diseases (Usdin 1998). Ectopic recombination can also impact microsatellite variability, leading to expansion or contraction of microsatellite length (Payseur and Nachman 2000; Schug et al. 1998). Additionally, microsatellites can be associated with transposable elements (TEs), often appearing as flanking sequences or tandem-arrays that are endogenous to the TE itself (Ramsay et al. 1999). Active TEs with associated microsatellites may facilitate microsatellite proliferation to new sites across the genome through a process known as microsatellite seeding (Arcot et al. 1995; Castoe et al. 2011*b*).

Current evidence suggests microsatellite content and abundance tends to be relatively consistent within major lineages of vertebrates, with most differences in microsatellite content arising between major lineages (Neff and Gross 2001; Tóth et al. 2000). Mammalian, fish, and squamate reptile genomes appear to be relatively microsatellite rich (Alföldi et al. 2011; Castoe et al. 2011*a*, 2011*b*; Chistiakov et al. 2006; Edwards et al. 1998), while bird, crocodilian, and turtle genomes are comparatively depauperate in microsatellites (Card et al. 2014; Primmer et al. 1997; Shedlock et al. 2007). Several studies have demonstrated microsatellite locus conservation across considerable scales of evolutionary time in specific lineages (FitzSimmons et al. 1995; Pepín et al. 1995; Rico et al. 1996; Schlotterëer et al. 1991), while other studies have noted substantial divergence in microsatellite content even between closely related species (Clisson et al. 2000). Previous analysis of microsatellite content in vertebrate genomes suggests that dinucleotide (2mer) repeats are the most abundant motif in vertebrate genomes, with nearly 1.5-fold greater abundance than that of longer microsatellite motifs (Benson et al. 2012). Nearly all previous studies of genomic microsatellite content in vertebrates, however, have been based on reduced or biased representation of vertebrate lineages, and have therefore lacked meaningful comparisons within and among major vertebrate lineages.

Here we used 71 complete vertebrate genomes to identify and characterize major compositional and evolutionary trends in the genomic microsatellite landscape of vertebrates. In our analysis of vertebrate microsatellite landscapes, we addressed the following questions: (*i*) How does variation in genomic microsatellite abundance among major vertebrate lineages compare to within-lineage variation? (*ii*) Do particular branches of the vertebrate tree appear to have experienced exceptionally high (or low) rates of change in microsatellite abundance? And (*iii*) do particular microsatellite classes or motifs, or characteristics of microsatellite loci (e.g., locus lengths), exhibit notable fluctuations across the vertebrate tree of life?

Materials and methods

Seventy-one vertebrate genome dataset

We downloaded all currently available assembled vertebrate genomes from the Ensembl database (release 75; Hubbard et al. 2002) representing a total of 61 vertebrate species. An additional 10 vertebrate genomes that were

not available via Ensembl were obtained to supplement sampling from several major vertebrate lineages. These included four snake species: Burmese python (*Python molurus bivittatus*; Castoe et al. 2013), king cobra (*Ophiophagus hannah*; Vonk et al. 2013), speckled rattlesnake (*Crotalus mitchellii*; Gilbert et al. 2014), and boa constrictor (*Boa constrictor*; Bradnam et al. 2013), two turtle species: green sea turtle (*Chelonia mydas*; Wang et al. 2013) and painted turtle (*Chrysemys picta*; Shaffer et al. 2013), and four crocodilians: American alligator (*Alligator mississippiensis*), saltwater crocodile (*Crocodylus porosus*), gharial (*Gavialis gangeticus*, St John et al. 2012), and Chinese alligator (*Alligator sinensis*, Wan et al. 2013). Our complete dataset consisted of assembled genomes from 41 mammals, 4 crocodilians, 5 birds, 3 turtles, 5 squamate reptiles, 1 amphibian (African clawed frog; *Xenopus tropicalis*), 1 lobe-finned fish (coelacanth; *Latimeria chalumnae*), 10 ray-finned fishes, and 1 jawless fish (lamprey; *Petromyzon marinus*; see Table S1[1] for taxon information and accession details). Hereafter, for the sake of brevity, we refer to species by their generic name only, unless the specific epithet is important for differentiation of two species in the same genus.

Microsatellite identification and quantification

We used Pal_finder v.0.02.03 (Castoe et al. 2010, 2012; Palfinder hereafter) to identify microsatellites across vertebrate genomes. Microsatellites were defined in Palfinder as perfect dinucleotide (2mer), trinucleotide (3mer), tetranucleotide (4mer), pentanucleotide (5mer), and hexanucleotide (6mer) tandem repeats. Default Palfinder parameters were used to identify microsatellite loci if they were tandemly repeated for a total length of at least 12 bp for 2–4mers, (i.e., 6 tandemly repeated 2mers, 4 repeated 3mers, and 3 repeated 4mers) or were repeated at least 3 times in the case of 5mers and 6mers (>15 bp in length). Microsatellite locus identification thresholds were based on previous estimates of genomic microsatellite content (Castoe et al. 2010, 2012). A custom Python script was used to summarize Palfinder output files and to calculate two measures of observed genomic microsatellite frequencies (loci/Mbp and bp/Mbp) for each species. We quantified frequencies of microsatellite loci and microsatellite base pair occupancy by dividing all observed counts of loci and all observed counts of base pair abundances by the total number of unambiguous (i.e., non-'N' base calls) megabases analyzed per genome. Loci/Mbp and bp/Mbp frequencies were calculated for all motifs and all microsatellite length classes (2–6mers, and total microsatellite content) and summarized per genome. Additionally, average bp/locus lengths were computed for each motif within each species using these Palfinder output files.

Time-calibrated phylogeny for 71 vertebrate species

To provide an evolutionary context to our genomic sampling, we used the species tree for the 61 Ensembl species as a backbone topology, and used mitochondrial genome sequences to add the remaining 10 species and to estimates divergence times. We used custom Python scripts to download and parse 12 mitochondrial-encoded protein-coding genes (excluding *ND6*, which is transcribed in the opposite direction of all other mitochondrial genes) from the mitochondrial genome of each vertebrate species with an available mitochondrial genome on GenBank (Benson et al. 2012). For species without available mitochondrial genomes, we used the mitochondrial genome for the most closely related species with a complete mitochondrial genome sequence, obtained from either GenBank or Ensembl (see Table S2[1] for GenBank and Ensembl accessions). Each of the 12 protein coding genes were then aligned individually using MUSCLE v3.8.31 (Edgar 2004) and were concatenated using the SuperMatrix function provided in the R package EvobiR v.1.1 (Blackmon and Adams 2015).

The final concatenated mitochondrial alignment was divided into three partitions (codon positions 1, 2, and 3) for each of the 12 protein coding genes for a total of 36 partitions. We used PartitionFinder v1.1.1 (Lanfear et al. 2012) to estimate the best-fit nucleotide substitution models for each partition. In all cases the best-fit model was a GTR model with gamma-distributed among-site variation. We estimated divergence times using BEAST v.2.01 (Drummond and Rambaut 2007) with a Yule model of speciation and a log-normal relaxed clock model for among-lineage variation. The tree topology was almost constrained to the Ensembl tree topology, therefore only leaving the 10 non-Ensembl species free to vary in phylogenetic position. To infer an ultrametic tree with branch lengths relative to time, we constrained divergence times at eight nodes (seven constraints obtained from the date-a-clade database Benton and Donoghue 2007 and an additional node constraint for alethinophidian snakes based on Pyron et al. 2013). A list of all calibration time points used in the analysis and priors used in the BEAST 2 analyses are given in the supporting materials (Table S3[1]). Two independent BEAST 2 analyses were run for 60 million generations each, and MCMC chains were sampled every 1000 generations. We used the program Tracer v.1.6 (Rambaut and Drummond 2014) to confirm that the MCMC chains had reached convergence based on likelihood and parameter value stationarity. We conservatively discarded 25% of the generations as burn-in, based on evidence that the likelihood and parameter values were stationary by around 10% of sampling. To account for uncertainty in divergence time estimates in our analyses of evolutionary rates, we randomly sampled

[1]Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/gen-2015-0124.

100 trees from the posterior distribution of our BEAST 2 analysis (these trees are referred to as the 100 sampled trees). These 100 trees were used to conduct censored rate tests for testing different rates of microsatellite evolution. Additionally, we used the program TreeAnnotator 1.7.4 (Rambaut and Drummond 2013) to estimate a consensus tree with branch lengths equal to the median length across the posterior distribution for visualizing relative evolutionary rates (see Fig. 2 for consensus tree).

Testing for multiple rates evolution in microsatellite abundance across lineages

We were interested in testing how overall evolutionary rates of change in microsatellite abundance (broken down by length class) differed across major clades in the vertebrate tree. To test this hypothesis and infer shifts in the rates of evolutionary change in microsatellite abundance, we conducted censored rate tests based on a Brownian motion model for continuous trait evolution of genomic microsatellite class content expressed as loci/Mbp (O'Meara et al. 2006). Our hypothesis testing was based on a null model in which microsatellites length class (2–6mers) abundance evolve at a single rate on all branches compared to an alternative model in which each of the six major vertebrate lineages represented by more than a single species (i.e., ray-finned fishes, turtles, crocodilians, squamate reptiles, birds, and mammals) have an independent rate of change in microsatellite abundance (for all motif sizes from 2 to 6mers, and for total microsatellite content). For all tests, we used the topology and branch lengths derived from the 100 randomly sampled trees obtained from the posterior distribution inferred from our BEAST 2 divergence dating analysis. Censored rates tests were performed using the R v.3.1.3 (R Development Core Team 2012) package phytools v.0.4-60 (Revell 2012) with 1000 simulations on each of the 100 sampled trees using the restricted maximum likelihood approach (REML), which yields unbiased estimates of the evolutionary rate parameter (σ^2). For each clade and test, rate parameter (σ^2) estimates were computed for each motif size (2–6mers and for the total microsatellite loci/Mbp). P values for each of the censored rate tests were calculated based on the 1000 simulations and were considered statistically significant if $P < 0.05$.

Ancestral state reconstruction of genomic microsatellite frequencies

To further identify specific branches of the vertebrate tree with evidence of particularly high or low rates of change in microsatellite abundance, absolute changes in observed frequencies of microsatellite content (expressed as loci/Mbp) between parental (ancestral nodes) and daughter nodes (|ΔSSR| hereafter) were modeled as a linear function of branch length. For each microsatellite class (2–6mer and total microsatellite content), ancestral state frequencies were reconstructed using the tree and branch lengths derived from the median consensus tree from our divergence-dating analyses in BEAST 2. Using the R package APE v.3.3 (Paradis et al. 2004), ancestral states for each interior node were estimated using the restricted maximum likelihood method (REML) under a Brownian motion model of evolution. These estimated ancestral values were used to calculate the absolute difference between each daughter and respective parental node across the tree. We fit simple linear regression models between |ΔSSR| estimates and branch length for each microsatellite class. Predicted values were obtained for each regression model using the predict function provided in the base statistics package in R. Deviations from predicted |ΔSSR| values based on the linear regression models were used to highlight branches with exceptionally high or low estimated rates of change in |ΔSSR| loci/Mbp frequencies (values outside both predicted and confidence intervals).

Identification of patterns of microsatellite landcape differentiation among species

We conducted a phylogenetically informed PCA using the phylo.pca function in phytools on the motif loci/Mbp frequencies of each species; this approach accounts for phylogenetic relationships when calculating eigenvalues, eigenvectors, and component loadings. The first two principle components (PC1 and PC2) were plotted to reflect the main factors effecting variation within microsatellite length types across the dataset. Phylogenetic PCAs were conducted for all microsatellite length classes using the observed loci/Mbp frequency of each microsatellite sequence motif. To further identify patterns of differentiation at the level of sequence-specific microsatellite motifs across species, we produced heat maps of the loci/Mbp frequencies and average bp/locus for each sequence-specific motif (for 2–4mers) or for the 25 most frequent sequence motifs per length class (4–6mers; based on average frequencies across species).

Results

Trends in genomic microsatellite content across 71 verbrate taxa

We identified considerable variation in genomic microsatellite content across the vertebrate tree of life (Figs. 1–4). On average, 4mer repeats are found at the highest loci/Mbp densities across the 71 taxa, followed by 2mers, 3mers, 5mers, and 6mers. When comparing microsatellite bp/Mbp densities, 2mers are the most abundant microsatellite, followed by 4mers, 3mers, 5mers, and 6mers. Below we focus on the average abundances of each microsatellite class and each of the six vertebrate clades that currently have more than a single representative genome (ray-finned fishes, squamates, crocodilians, turtles, birds, and mammals).

Total genomic microsatellite frequencies

Ray-finned fish genomes contain the highest observed average microsatellite loci/Mbp frequencies (mean = 716.86 loci/Mbp) of all vertebrate clades, followed closely by squamate reptiles (mean = 628.26 loci/Mbp) and

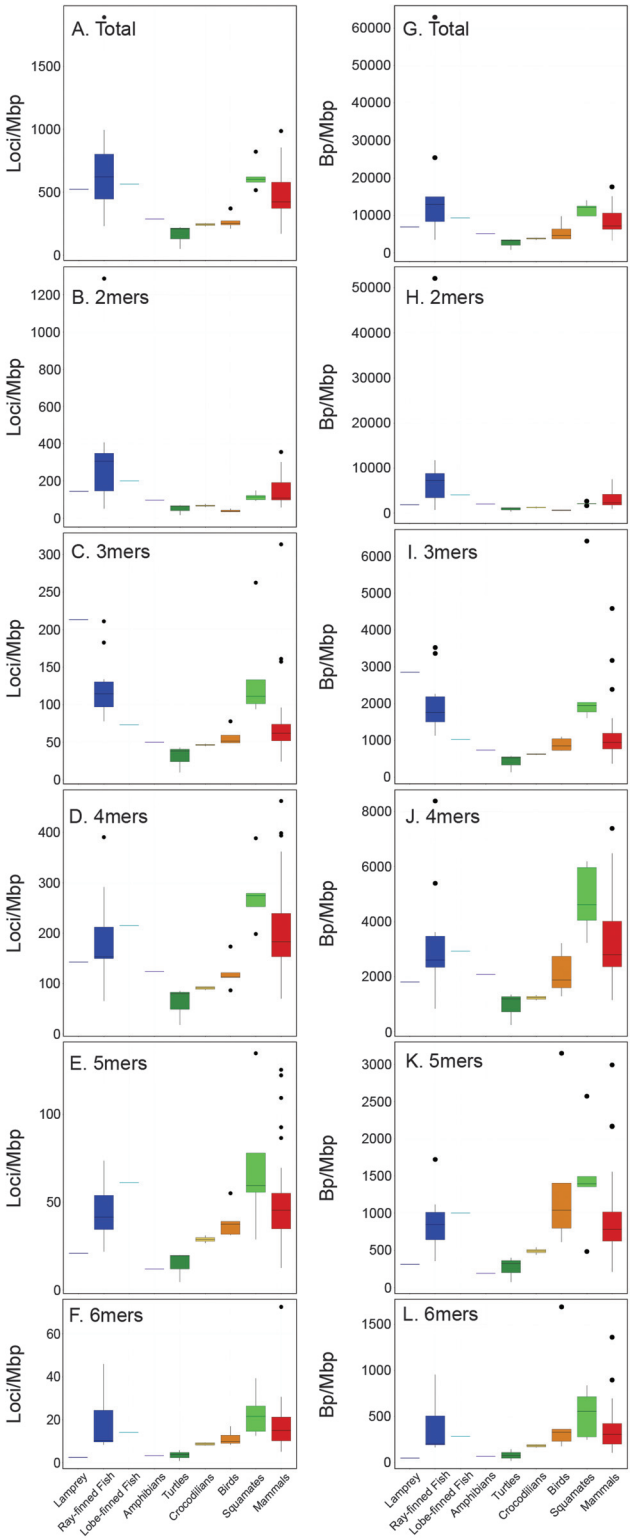


Fig. 1. Observed distributions of genomic microsatellite content for nine major vertebrate clades. Raw microsatellite frequency distributions for nine major vertebrate lineages (71 genomes), including lampreys (1 species), ray-finned fishes (10 sp.), lobe-finned fishes (1 sp.), amphibians (1 sp.), turtles (3 sp.), crocodilians (4 sp.), birds (5 sp.), squamate reptiles (5 sp.), and mammals (41 sp.). Abundances in loci/Mbp for total, 2mer, 3mer, 4mer, 5mer, and 6mer repeats are shown in panels A–F, respectively. And frequencies in bp/Mbp for total, 2mer, 3mer, 4mer, 5mer, and 6mer microsatellites are shown in panels G–L, respectively. Horizontal lines represent the average frequency for each taxon, and the 25th and 75th percentiles are indicated by the box edges. Whiskers extend to the most extreme value within 1.5 times the inter-quartile (distance between 1st and 3rd quartile) range. Any values outside these whiskers are considered as an outlier and are display as black points.

mammals (mean = 491.23 loci/Mbp; Fig. 1A). The high average abundance of microsatellite loci in ray-finned fishes is largely due to the extremely high frequency of microsatellites in the Atlantic cod (*Gadus*) genome (1890.86 loci/Mbp; Fig. 2A and Table S4[1]). On average, the genomes of turtles, crocodilians, and birds contain similarly low total abundances of microsatellites (mean = 159.78, 242.42, and 268.94 loci/Mbp, respectively) with little variance in abundance within each clade (SD = 95.57, 11.91, 60.90 loci/Mbp, respectively; Fig. 1A). Similar to trends observed in loci/Mbp frequencies, ray-finned fish also have the highest average total bp/Mbp frequencies (mean = 17 045.93 bp/Mbp), followed by squamate reptiles (mean = 11 675.68 bp/Mbp) and mammals (mean = 8610.06 bp/Mbp; Fig. 1G). Ray-finned fishes again exhibit the highest variation in total microsatellite bp/Mbp frequencies (SD = 17 223.56 bp/Mbp), due largely to the SSR-rich genome of Atlantic cod (*Gadus*; total = 62 912.41 bp/Mbp; Fig. 2A and Table S4[1]).

Dinucleotide repeat abundances

Ray-finned fish genomes have the highest average density of 2mer loci (343.14 loci/Mbp), followed by mammals (146.06 loci/Mbp) and squamates (114.76 loci/Mbp; Fig. 1B). Turtles, crocodilians, and birds have similar and lower average loci/Mbp densities (means = 49.53, 67.02, and 39.43 loci/Mbp, respectively; Figs. 1B and 2B). The average 2mer bp/Mbp frequency for ray-finned fish genomes is over three-fold that of every other lineage average (10 492.61 bp/Mbp), followed by dinucleotide densities of mammals (average = 3048.60 bp/Mbp) and squamates (average = 2123.13 bp/Mbp; Figs. 1B and 2B). As with total microsatellite abundance, extreme patterns associated with ray-finned fish 2mers stems from the Atlantic cod (*Gadus*; total dinucleotides = 52 103.99 bp/Mbp), although the clade also exhibits the largest variance in 2mer frequencies even when excluding the Atlantic cod genome (Fig. 2B).

Phylogenetic PCA across 2mer sequence motif frequencies indicate that that PC1 is driven predominantly by variation in the frequency of the AC motif (Fig. 3A), consistent with the overall high abundance and length of the motif AC in most vertebrate genomes, with several notable exceptions (Fig. 3B). For example, AT is the most abundant 2mer motif in the frog (*Xenopus*) and most bird genomes, while TC is the most abundant 2mer found in several mammalian genomes, including that of the dog,

Fig. 2. Observed microsatellite loci/Mbp frequencies and their lineage-specific evolutionary rates across 71 vertebrate taxa. Horizontal bar plots represent the observed microsatellite loci/Mbp frequencies for each vertebrate genome. Branches on the time-calibrated consensus phylogeny are colored according to the estimate rate of microsatellite evolution, with dark blue indicating slower evolution and red indicating more rapid evolution. Taxa and coinciding bars are consistently colored and ordered based on major vertebrate clades. Results are shown for (A) total genomic microsatellite content, (B) 2mer, (C) 3mer, (D) 4mer, (E) 5mer, and (F) 6mer loci/Mbp frequencies.

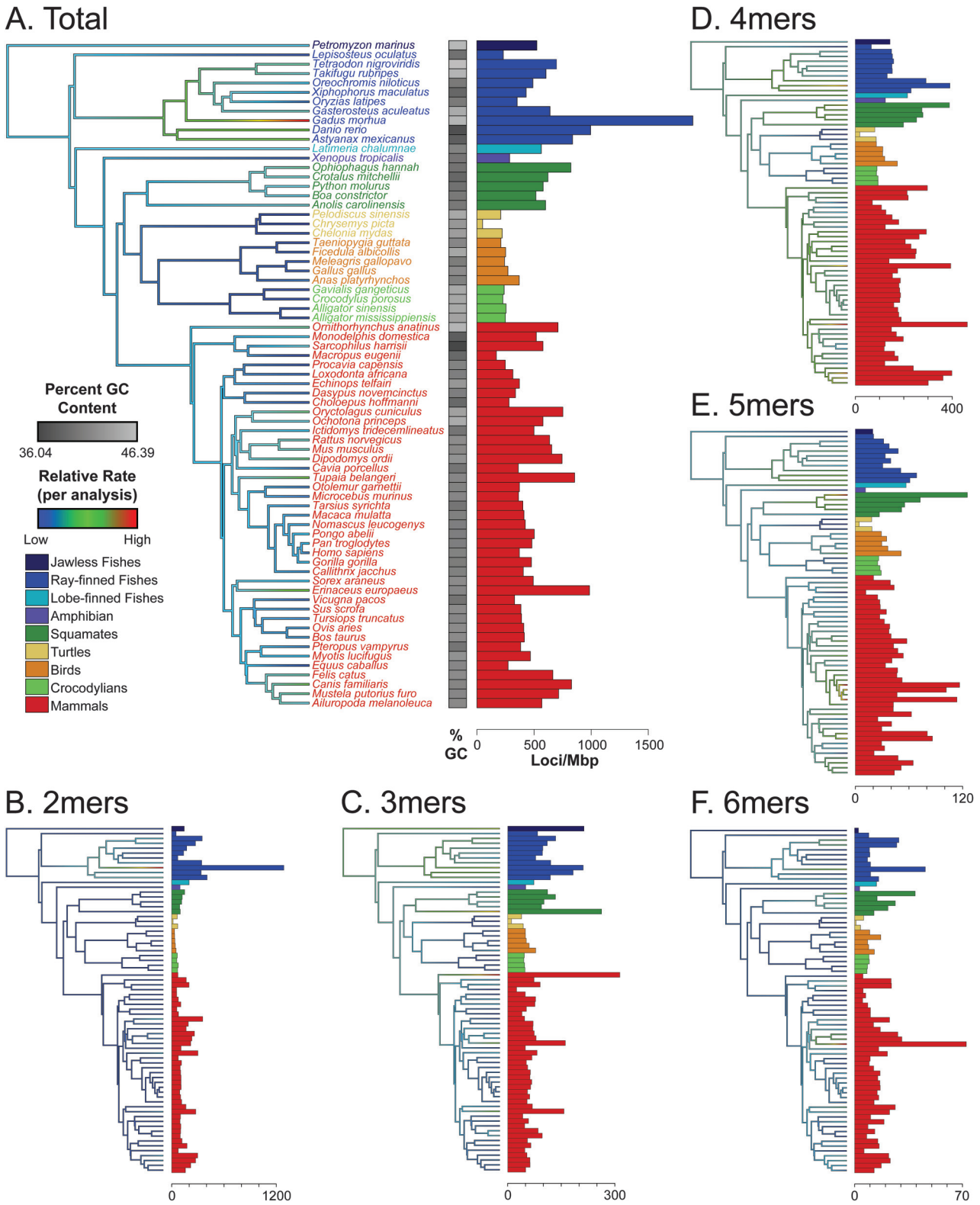
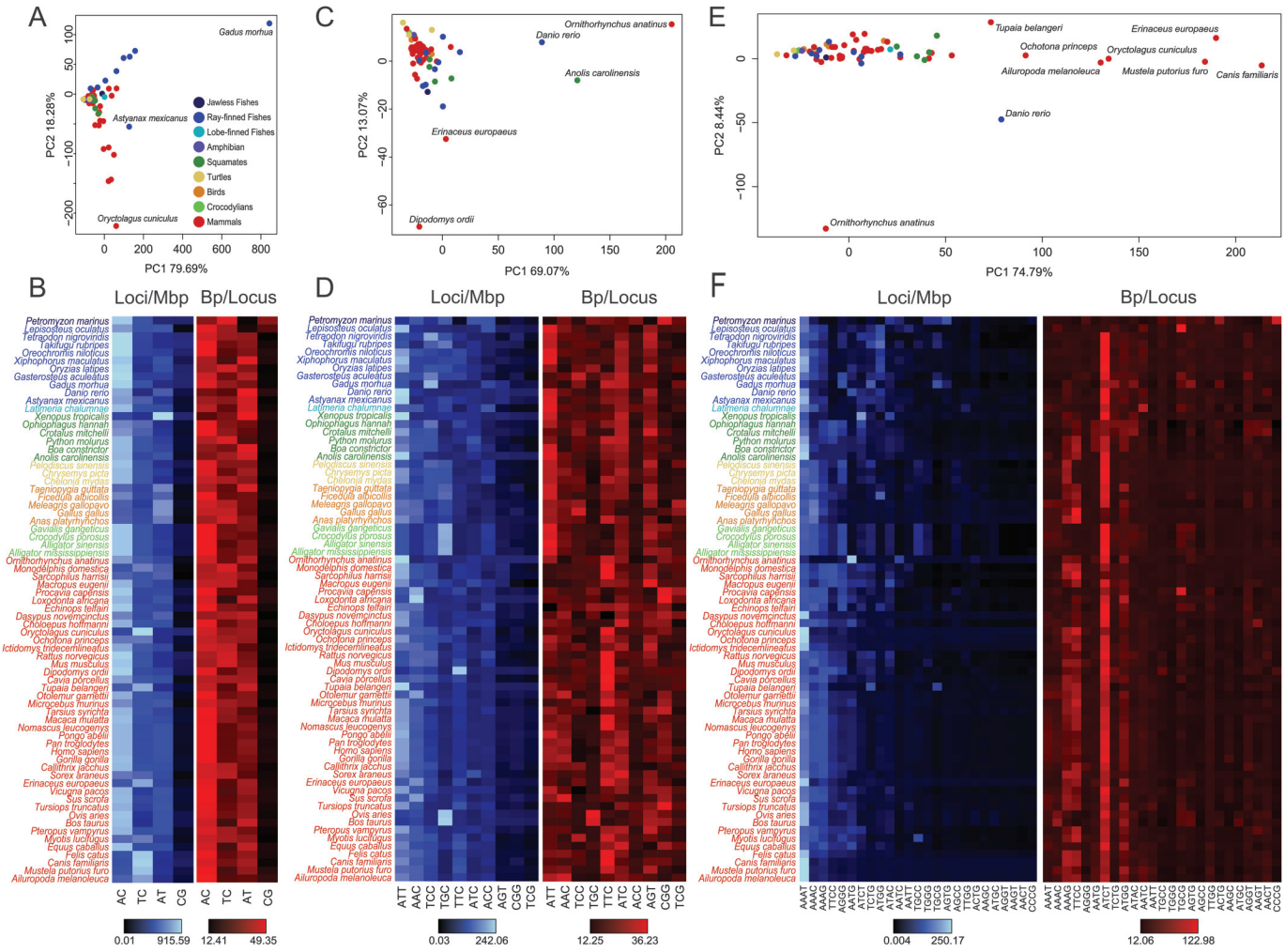


Fig. 3. Phylogenetic PCA, loci/Mbp heatmaps and bp/locus heatmaps for 2mer (A–B), 3mer (C–D), and 4mer (E–F) microsatellite motifs. Phylogenetic PCA was conducted using loci/Mbp frequencies for all motif types (2–3mers) and the top 25 most abundant motifs in 4mer microsatellites. Labeled points represent samples that deviated substantially from the major clusters. Loci/Mbp (blue) and bp/locus (red) heatmaps depict frequencies for all 2 and 3mer motifs and the top 25 most abundant loci for 4mers, sorted from the highest average abundance (left) to least highest average abundance (right).



cat, panda, rabbit, and others. Another interesting difference in 2mer motif properties among lineages is that increases in densities of the AT motif in several fish (*Danio*, *Astyanax*, *Latimeria*) and the frog (*Xenopus*) genome appear to have been driven by the greater lengths of AT loci in these lineages (Fig. 3B).

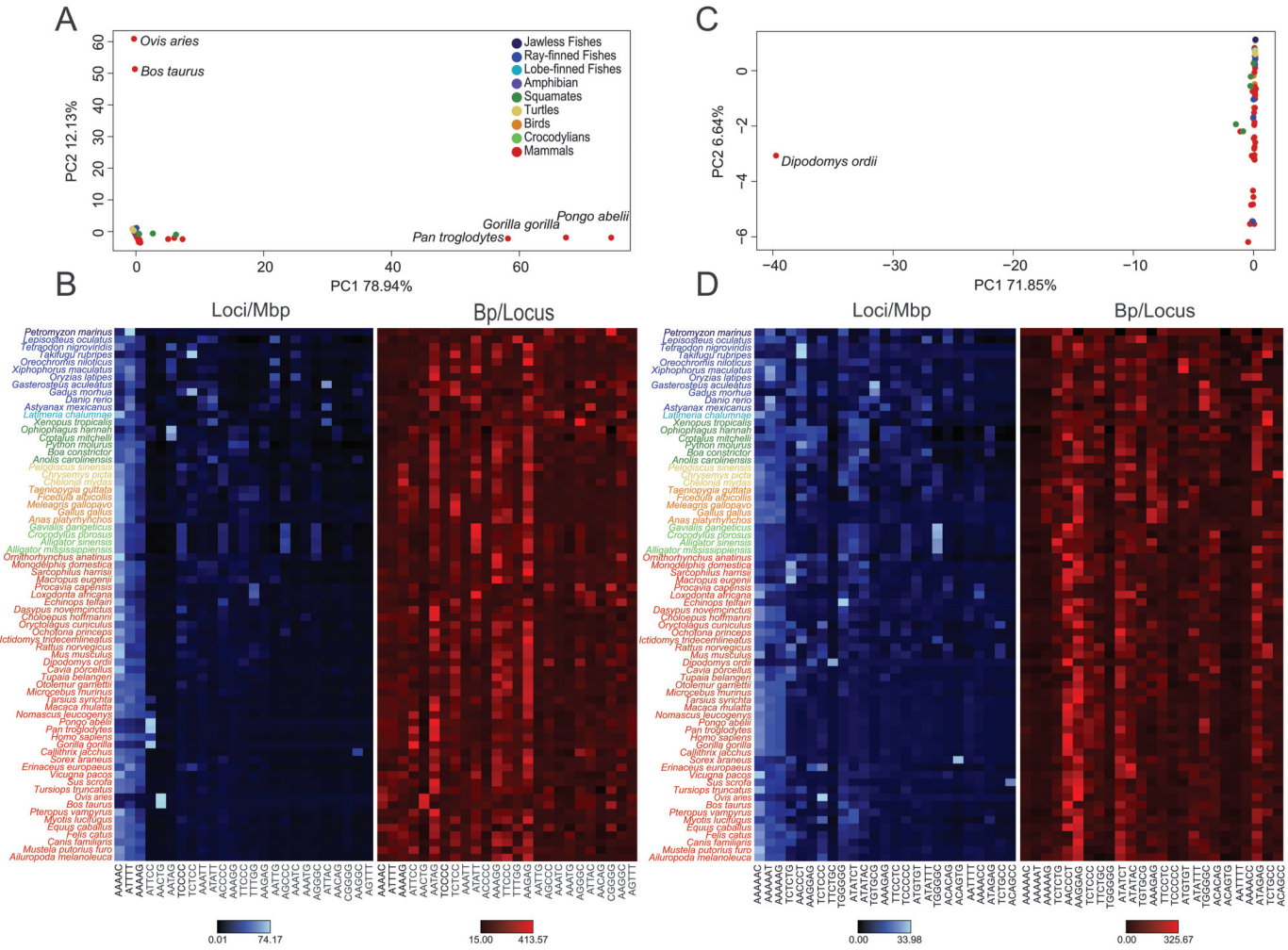
Trinucleotide repeat abundances

Genomic trinucleotide (3mers) repeat densities are on average the most frequent and variable in squamate reptile genomes (means = 140.53 loci/Mbp and 2753.61 bp/Mbp), followed by ray-finned fishes and mammals (Figs. 1C and 1I). In contrast, non-squamate reptiles (including birds) are comparatively low and consistent in average 3mer loci/Mbp and length (Figs. 1C and 1I). Phylogenetic PCA of sequence-specific 3mer motif frequencies indicated that PC1 was most strongly driven by variation in ATT motif frequencies, and the genomes of the zebra fish (*Danio*), anole lizard (*Anolis*), and platypus (*Ornithorhynchus*) were separated from other vertebrates due to their unusually high ATT loci/Mbp frequency (Fig. 3C). In contrast to these ATT-rich genomes, the TGC motif was the most abundant 3mer in all crocodilians and two mammalian genomes (*Ovis* and *Bos*; Fig. 3D). Other outlying patterns include the frequent AAC motif in two mammalian genomes (*Provacia* and *Loxodontia*), and the motif TTC being most frequent in the kangaroo rat (*Dipodemys*) genome (Fig. 3D). Interestingly, despite these divergent patterns of microsatellite motif density across species, the motif TTC has the highest average length per locus of all 3mer microsatellites in most genomes, with particularly high average locus lengths of this motif in rodent species (Fig. 3D).

Tetranucleotide repeat abundances

Squamate reptile genomes harbor the highest average densities of tetranucleotides (4mers; means = 278.75 loci/Mbp and 4813.90 bp/Mbp; Figs. 1D and 1J). Conversely, tetra-

Fig. 4. Phylogenetic PCA, loci/Mbp heatmaps, and bp/locus heatmaps for 5mer and 6mer motifs. Phylogenetic PCA was conducted using loci/Mbp frequencies of the top 25 most abundant motif types for both 5mer (A–B) and 6mer (C–D) microsatellites. Labeled points represent samples that deviated substantially from the major clusters. Loci/Mbp (blue) and bp/locus (red) heatmaps depict frequencies of the top 25 most abundant 5 and 6mer microsatellite motifs found across the 71 vertebrate genomes.



nucleotides are relatively low and consistent in abundance within the birds and the other reptile clades (turtles, crocodilians). However, ray-finned fishes exhibit the second highest average bp/Mbp content (mean = 3306.05 bp/Mbp), which is slightly higher than the average mammalian tetranucleotide bp/Mbp frequency (mean = 3189.50 bp/Mbp; Fig. 1J). Phylogenetic PCA of 4mer motifs revealed that overall 4mer variation is largely explained by AAAT loci/Mbp variation, which corresponds with the high yet variable abundance of AAAT repeats in most vertebrate genomes (Fig. 3E), although both AAAC and AAAG 4mers are also abundant in many vertebrate genomes (Fig. 3F). Among the top 25 most abundant 4mers, the motif ATCT has a particularly high average bp/locus length across most species compared to other motifs (Fig. 3F).

Pentanucleotide repeat abundances

As with 3–4mer motifs, squamate reptiles have the highest average pentanucleotide (5mer) loci/Mbp microsatellite density (mean = 71.33 loci/Mb), followed by mammals and ray-finned fishes (Fig. 1E). The average abundance of 5mer bp/Mbp in birds is the second highest of all vertebrate clades after squamates (mean = 1400.60 bp/Mbp), despite birds having among the lowest observed loci/Mbp densities (38.97 loci/Mbp); these findings indicate an expansion in the length of 5mer repeats, rather than the generation of new loci, which has led to high 5mer abundances within avian lineages (Fig. 1K). In particular, the genome of the collard flycatcher (*Ficedula*) has undergone what appears to be extreme expansion of long pentanucleotide repeats (total 5mer frequency = 3146.65 bp/Mbp) that is higher than the 5mer content observed in any other vertebrate genomes present in this study (Fig. 2E and Table S4[1]).

In the PCA of variation in 5mer motif frequencies, PC1 is largely driven by variation in the ATTCC motif, which is particularly frequent in the Great Apes (minus humans) yet otherwise quite rare across vertebrates

(Figs. 4A–4B). In most sampled vertebrate genomes, the AAAAC motif is the most abundant 5mer based on loci/Mbp frequencies (Fig. 4B). Other motifs of interest include the highly abundant AATAG motif in the king cobra (*Ophiophagus*) and TCTCC in the Atlantic cod (*Gadus*) and Japanese puffer fish (*Takifugu*; Fig. 4B). Additionally, the AACTG is the most abundant 5mer microsatellite in the domestic cow (*Bos*) and sheep (*Ovis*) genomes, which underlies their differentiation from other species in PC2 of the PCA (Fig. 4A).

Hexanucleotide repeat abundances

Observed hexanucleotide (6mer) loci/Mbp averages are highest in squamate reptile genomes (average = 22.89 loci/Mbp), followed by ray-finned fishes and mammals (average = 17.52 loci/Mbp and 16.56 loci/Mbp, respectively; Fig. 1F). Bird genomes, which have among the lowest observed loci/Mbp frequencies, have the highest average base pair density (556.73 bp/Mbp) and are the most variable (SD = 637. 17 bp/Mbp; Figs. 1F and 1L). This result is due primarily to the extreme nature of 6mer repeats found in the collared flycatcher (*Ficedula*) genome, which contains the highest observed 6mer base pair frequencies of all species surveyed (6mer content = 1688.67 bp/Mbp; Fig. 2F). Phylogenetic PCA of 6mer motifs indicates that PC1 corresponds largely to the variation in TTCTGC and AAGGAG frequencies, which when phylogenetically corrected are particularly high in the kangaroo rat (*Dipodemys*) genome (Fig. 4C). PC2 separates most other vertebrates based largely on the motifs AAAAAC and TCTCTG, which are both common and variable across species (Figs. 4C–4D). Unlike other vertebrates, the motif TCTCCC is the most abundant 6mer in the sheep (*Ovis*) genome, and the motif TGGGGG is the most abundant 6mer in the lesser hedgehog tenrec (*Echinops*) genome. In the four crocodilian genomes (*A. mississippiensis*, *A. sinensis*, *Gavialis*, and *Crocodylus*), AGGGGC is the most abundant 6mer (Fig. 4D), while the most abundant 6mer motifs overall (AAAAAC, AAAAAT, AAAAAG) have among the lowest average bp/locus lengths in these genomes (Fig. 4D).

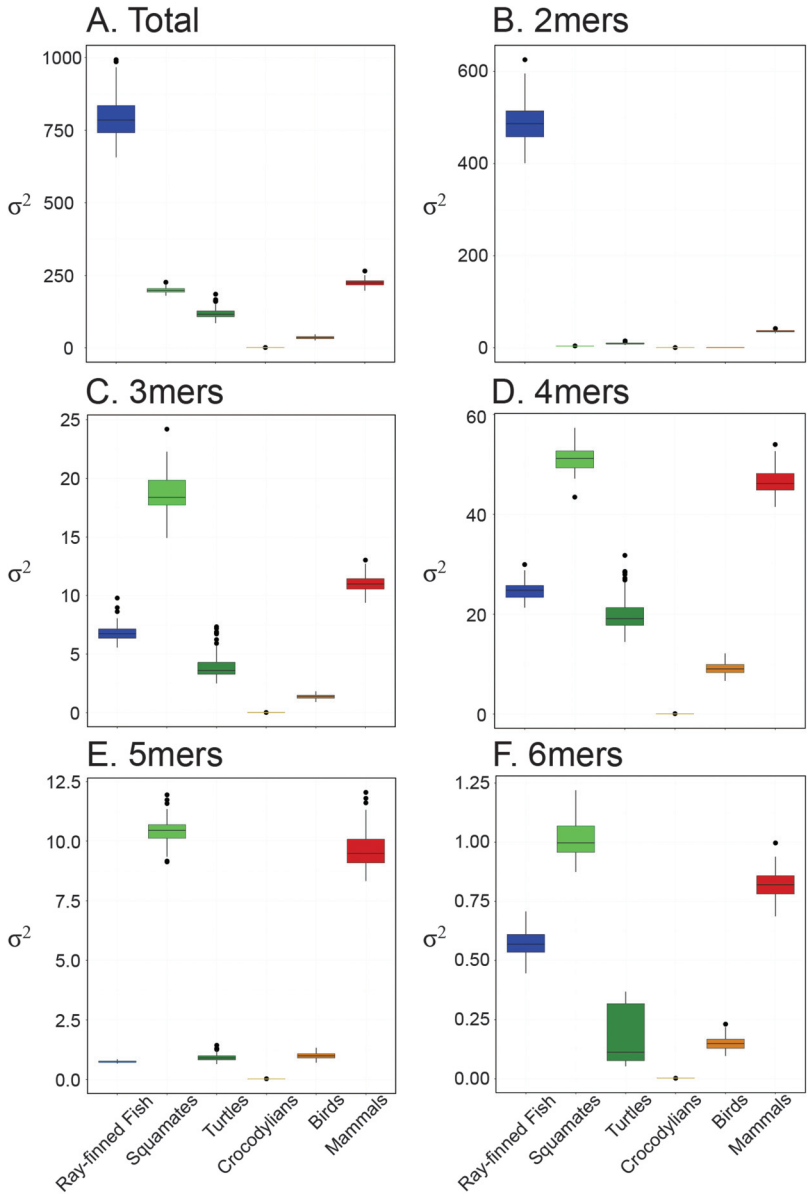
Clade-specific rates of evolution in microsatellite genomic density

All censored rate tests (100 tests for each 2–6mer and total microsatellite loci/Mbp frequencies) reject the null model (a single rate of evolution in all lineages), indicating significantly different, lineage-specific rates of change in microsatellite abundance among the six major multi-genome vertebrate clades ($P < 0.01$ for all 600 tests). Total microsatellite loci/Mbp abundances evolved most rapidly in ray-finned fishes (mean = $792.50\sigma^2$) and slowest in turtles, crocodilians, and birds (means = 120.20, 0.93, and $35.32\sigma^2$, respectively), while squamate reptiles and mammals both exhibit similar high rates of change (means = $199.10\sigma^2$ and $224.40\sigma^2$, respectively; Fig. 5A). Dinucleotide rates of evolution are also highest in ray-finned fishes (mean $\sigma^2 = 486.59$) and lowest in reptiles (mean turtles = $9.68\sigma^2$, crocodilians = $0.49\sigma^2$, birds = $0.49\sigma^2$, and squamates = $3.92\sigma^2$; Fig. 5B). For 3–6mers, rates of change in abundance of microsatellites were highest in squamate reptiles, and second highest in mammals (Figs. 5C–5F).

Branch-specific rates of change in microsatellite abundance

For all microsatellite length classes (2–6mer and complete microsatellite content), changes in microsatellite loci/Mbp frequencies between parental and daughter lineages (|ΔSSR|) are only very weakly to moderately correlated with branch lengths (scaled to time; MAX $R^2 = 0.353$, $P < 0.01$ across all comparisons), suggesting that rates of microsatellite evolution vary widely across the vertebrate tree (Figs. 6A–6F). Changes in total genomic microsatellite loci/Mbp (total |ΔSSR|) are greatest along the branches leading to the Atlantic cod (*Gadus*), European hedgehog (*Erinaceus europaeus*), and Northern treeshrew (*Tupaia*; Fig. 6A). The |ΔSSR| estimate on the branch leading to the single lamprey genome (*Petromyzon*) is the lowest rate of evolution predicted (Fig. 6A). For dinucleotide repeats (2mer), only the |ΔSSR| value for the terminal branch leading to the Atlantic cod (*Gadus*) fell far outside the predicted intervals (Fig. 6B), further illustrating the extreme rate of dinucleotide expansion within this genome. For 3mers, the platypus (*Ornithorhynchus*), anole lizard (*Anolis*), European hedgehog (*Erinaceus*), kangaroo rat (*Dipodomys*), and Atlantic cod (*Gadus*) represent |ΔSSR| estimates above predicted |ΔSSR| intervals, while the spotted gar (*Lepisosteus*) is below the predicted |ΔSSR| intervals for its branch length (Fig. 6C). For 4mers, branches exhibiting |ΔSSR| values above predicted ranges include the European hedgehog (*Erinaceus*), Northern treeshrew (*Tupaia*), zebra fish (*Danio*), tammar wallaby (*Macropus*), domestic dog (*Canis*), and horse (*Equus*), while the lamprey (*Petromyzon*) |ΔSSR| estimate is below the predicted intervals (Fig. 6D). 5mer microsatellites appear to change rapidly across several primate branches (*Homo*, *Pongo*, *Gorilla*, *Pan*, and *Nomascus*) reflecting high rates of 5mer repeat change within this clade. Again, this trend of increased 5mer rates of change appears to result from the dynamic evolution of the motif ATTCC in primates (Fig. 5B). Additionally, the king cobra (*Ophiophagus*), alpaca (*Vicugna*), and both an ancestral squamate and mammalian branch exhibit 5mer |ΔSSR| estimates above predicted intervals (Fig. 6E). This high rate of change for 5mer microsatellites in the king cobra genome appears to result from the expansion of AATAG when compared to its sister reptile lineages (Fig. 4B). For 6mer repeats, the kangaroo rat (*Dipodomys*), Atlantic cod (*Gadus*), and king cobra (*Ophiophagus*) terminal branches exhibit |ΔSSR| values outside predicted intervals (Fig. 6F).

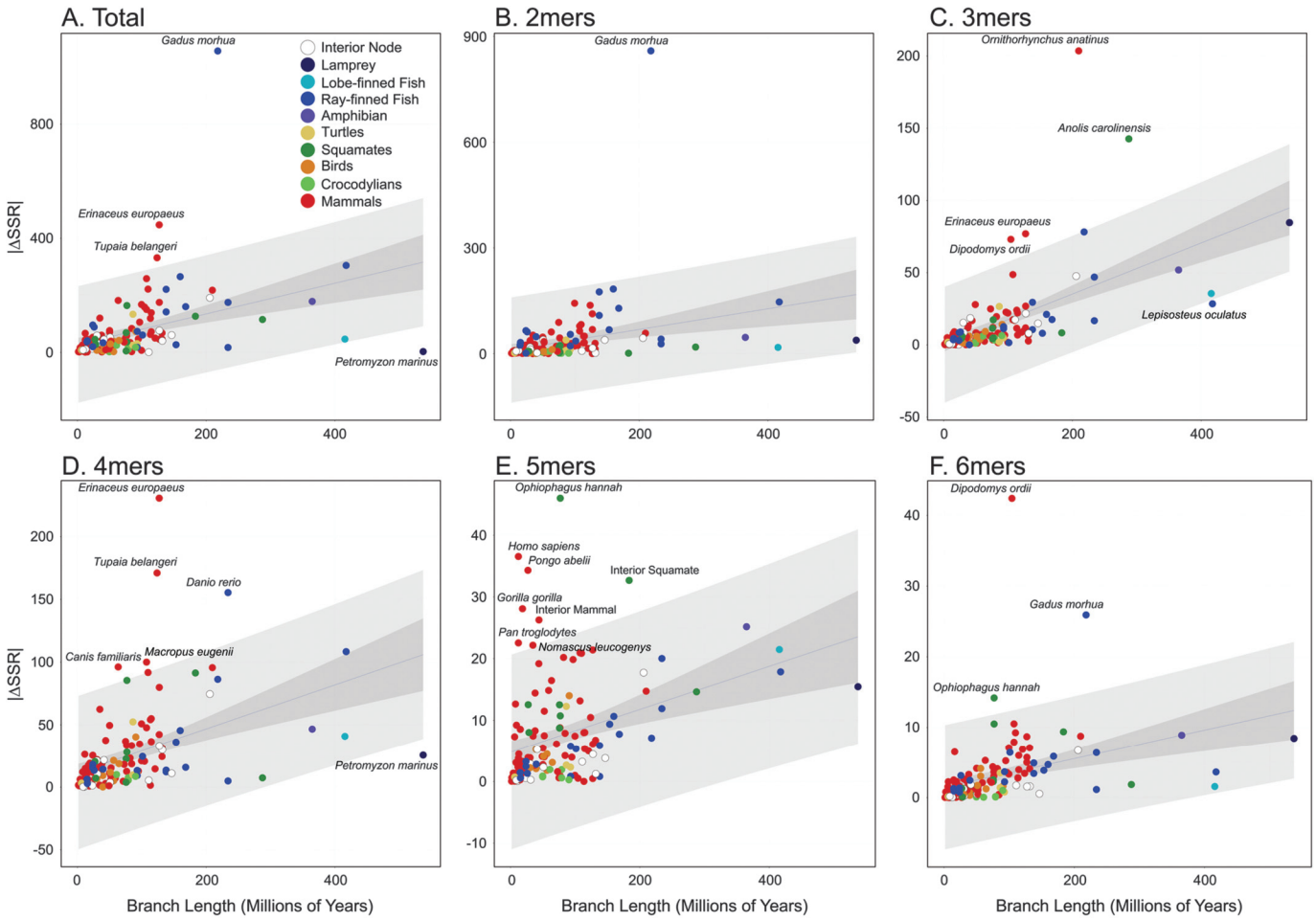
Fig. 5. Censored rate test results for lineage-specific rates of microsatellite evolution across the six major vertebrate clades with two or more genomes represented on Ensembl. Box pots represent the rate parameter (σ^2) estimates obtained across 100 trees sampled from the inferred posterior distribution for the six major clades represented by more than a single genome (n = 68), including ray-finned fishes (10 sp.), turtles (3 sp.), crocodilians (4 sp.), birds (5 sp.), squamate reptiles (5 sp.), and mammals (41 sp.). Results are shown for (A) total microsatellite content, (B) 2mer, (C) 3mer, (D) 4mer, (E) 5mer, and (F) 6mer loci/Mbp frequencies.



Discussion

Our results provide the first perspective obtained from analysis of complete genomes representing multiple members of most major vertebrate clades. Most apparent from analyses of these data are extreme shifts in genomic microsatellite content among species and among lineages, as well as evidence for multiple major shifts in the evolutionary tempo of change of microsatellite landscapes across the vertebrate tree. We also find a number of interesting contrasts and convergences, in which closely related lineages (e.g., reptiles) contain species with drastically different microsatellite landscapes, yet distantly related lineages (e.g., squamate reptiles and mammals) are found to share many characteristics of microsatellite landscapes and evolution. Factors driving these shifts in the clade-specific patterns and tempos of microsatellite evolution are likely numerous and diverse, and our analyses highlight previously unappreciated similarities between such disparate lineages that require further investigation to address what common features may underlie these similarities and differences in microsatellite landscape composition and evolution.

Fig. 6. Estimated change in microsatellite content as a function of branch length general linear models describing the change in microsatellite loci/Mbp (|ΔSSR|) frequencies between parent taxa (ancestral node) and each respective daughter taxon using the time-calibrated consensus tree. Prediction intervals (light gray) and confidence intervals (dark gray) were constructed using the predict function using the base statistics package in R for each regression model. For (A) total microsatellite content, (B) 2mers, (D) 4mers, (E) 5mers, and (F) 6mers, simple linear regression analyses indicate weak relationships between |ΔSSR| and branch lengths (R^2 = 0.163, 0.098, 0.187, 0.120, and 0.130, respectively; $P < 0.01$ for all models). For (C) 3mer microsatellites, simple linear regression supports a weak to moderate relationship between |ΔSSR| and branch lengths (R^2 = 0.353; $P < 0.0001$).



Dinucleotide repeat proliferation in fishes

Limited taxon sampling of jawless and lobe-finned fishes prevented meaningful inferences about these two clades, which is unfortunate given their position at the root of vertebrates and tetrapods. In some microsatellite measurements, these two clades have microsatellite landscapes that are similar to ray-finned fishes, while in other situations they appear to have microsatellite content more similar to that of other vertebrate clades. The massive expansion of dinucleotide repeats in several ray-finned fish genomes collectively constitute the highest abundances of any microsatellite class in any of the sampled vertebrate taxa. For example, observed frequencies (loci/Mbp and bp/Mbp) of AC dinucleotide repeats in the Atlantic cod (*Gadus*) genome rival total genomic microsatellite content of most other vertebrate species. The expansions of shorter microsatellite motifs (2–3mers) in ray-finned fishes are further illustrated by their exceptionally high estimated rates of evolution. Indeed, inferred rates of 2mer evolution are twice that of any other lineage, and 3mer rates are greater than 10-fold that of any other lineage. Further, the Atlantic cod (*Gadus*) and zebrafish (*Danio*) branches both contain |ΔSSR| values above predicted values for total microsatellite content, 2mers (Atlantic cod), 3mers (Atlantic cod), 4mers (zebrafish), 5mers (zebrafish), and 6mers (Atlantic cod). Divergence times between lineages of ray-finned fishes are quite ancient, providing an extensive timeframe over which such massive shifts in microsatellite composition may have occurred.

Amphibians represent a large gap in our understanding of vertebrate microsatellite landscapes

Like the jawless and lobe-finned fishes, our understanding of microsatellite landscapes, and the evolution of these landscapes, in amphibians is currently limited by the paucity of complete genomes for the group. In general, our results suggest that the genome of *Xenopus*

contains relatively low abundances of microsatellites when compared to other vertebrate genomes. As we completed data analysis for this study, a second amphibian genome, that of the Tibetan frog (*Nanorana parkeri*; Sun et al. 2015), was released. The microsatellite characteristics of this second frog genome appear to be comparable to that of *Xenopus*. It is notable that, among amphibians, frog genomes tend to be relatively small, as salamander genome range in size from ~15 to >100 Gbp (Gregory 2015). Given this massive variance in genome size in amphibians, it is reasonable to expect that microsatellite landscapes also vary substantially across amphibians. Indeed, preliminary studies that have sample sequenced salamander genomes indicate microsatellites may contribute relatively little to large genome size in salamanders (Sun and Mueller 2014). Until a greater diversity of amphibians genomes are available, the diversity of amphibian microsatellite landscapes will remain a major gap in our understanding of microsatellite evolution across the vertebrate tree.

High microsatellite composition and variability in mammals

Nearly all of our analytical approaches demonstrate that the genomic microsatellite content of the mammalian lineage ranks among the highest among vertebrates in abundance and variability among species. This could be a result of the moderately large average mammalian genome size (~2.3 Gbp) which may facilitate microsatellite proliferation, although correlations between microsatellite abundance and genome size remain relatively weak and only generally demonstrated (Hancock 1996; Primmer et al. 1997). Modeling changes in microsatellite content (|ΔSSR|) over branch lengths highlighted several interesting trends in mammalian microsatellite evolution. Most apparent of these trends is the clustering of primate branches, including the orangutan (*Pongo*), gorilla (*Gorilla*), and human (*Homo*), that far exceed predicted |ΔSSR| values for 5mer repeats. This highly dynamic profile of 5mer repeats in primates appears to result from exceptionally high frequencies of the motif ATTCC in all primates excluding humans, which conversely exhibit a greatly reduced presence of this motif compared to their ancestral node and sister lineages (indicating a decrease in this motif within the human lineage). The expansion of particular transposable elements (such as *Alu* retrotransposons) in primate lineages is proposed to have facilitated the proliferation of associated microsatellites through seeding process of either flanking or endogenous AT-rich microsatellites and may explain these patterns (Arcot et al. 1995). However, if *Alu* proliferation was a major contributor to microsatellite expansion in the human genome, the substantial reduction of ATTCC microsatellites (and all 5mer repeats) is particularly surprising, given that ~20% of all microsatellite loci shared between the human and chimpanzee genome lie within *Alu* elements (Kelkar et al. 2008). Changes in 3mer repeats in both the platypus (*Ornithorhynchus*) and kangaroo rat (*Dipodomys*) place these lineages well outside predicted |ΔSSR| intervals due predominantly to extraordinarily high ATT motif frequencies in the platypus genome, and TGC in the kangaroo rat genome (Fig. S1B[1]). Interestingly, our results support nearly identical 3mer motif frequencies between the platypus and anole lizard (*Anolis*), which is also highlighted outside predicted intervals. The high rate of change in dinucleotide repeats on the branch leading to the European rabbit (*Oryctolagus*) is indicative of their divergent 2mer landscape, which is dominated entirely by TC motifs rather than AC motifs, the later of which is often the most common 2mer found in most other vertebrates. This finding is notable, as regional assemblages of TC repeat arrays are thought to be linked to increased rates of recombination (Benet et al. 2000), potentially indicating a link between this motif frequency shift and recombination in rabbits. Additionally, SSRs are thought to be mechanistically involved with complex recombination dynamics in dogs (Wayne and Ostrander 2007), and we find that the domestic dog (*Canis*) has unexpectedly high total microsatellite and 3mer |ΔSSR| composition.

Divergent patterns of microsatellite composition in reptilian lineages

Recent studies indicate bimodal patterns of microsatellite content and evolution in reptile genomes and our results provide the first formal and comprehensive evidence describing this strong differentiation in reptilian microsatellite landscape characteristics. Archosauromorpha exhibit nearly homogenous genomic compositions of repetitive elements, while the highly variable microsatellite landscapes of squamate reptiles parallel the abundant and highly variable landscapes found in mammalian genomes (Castoe et al. 2011*b*; Primmer et al. 1997; Shedlock et al. 2007). The consistent dominance of squamate reptiles in nearly all measures of abundance and variation in microsatellite landscapes contrasts starkly with the relatively microsatellite-depauperate genomes of turtles, crocodilians, and birds. Furthermore, 4–6mer microsatellite in squamate reptiles exhibit almost parallel rates of evolution with their respective mammalian counterparts, while squamate 3mer evolutionary rates are estimated to be twice that of any other lineage. How microsatellite proliferation has occurred despite relatively small and constant genome sizes in squamate reptiles (averaging ~1.9 Gbp; Shaney et al. 2014) remains an open question, although evidence suggests that the expansions of particular microsatellite motifs in snake lineages may be linked to microsatellite seeding by certain families of LINEs (Castoe et al. 2013, 2011*b*). Results of this study further illustrate the extreme nature of the expansion of AT-rich 5mer and 6mer nucleotide repeats (especially AATAG), and the extent of variation in microsatellite composition in snakes compared to other expansions observed in other vertebrates.

The abundant and dynamic microsatellite landscapes of squamate reptiles contrast strongly with slowly evolving and reduced microsatellite landscape of the archosauromorph lineages. Since the sequencing of the chicken genome (Hillier et al. 2004), it has been known that bird genomes contain low levels of microsatellites compared to other vertebrates. Our results also support early (pre-genome) inferences that crocodilian and turtle genomes contained minimal microsatellite landscapes (Shedlock et al. 2007). Despite deep divergences among lineages sampled, we found that all crocodilian species contain nearly identical microsatellite landscapes, which further emphasizes the relatively slow rates of molecular evolution in crocodilians – thought to be among the slowest of all vertebrate lineages (Green et al. 2014).

Potential mechanisms and implications for vertebrate evolution

Although we know a fair amount about the processes that drive microsatellite establishment and evolution at a locus-specific scale, we know very little about global genome-wide drivers of microsatellite abundance and evolution. Lineage-specific rates of polymerase strand slippage may contribute to lineage-specific differences in microsatellite landscape evolution among species. In addition to polymerase stand slippage, a number of genomic features may direct microsatellite evolution in a lineage dependent manner. Nucleotide biases (such as high GC content) are also likely important factors that may influence rates of microsatellite evolution across vertebrates and differences in centromeric sequence evolution may also influence the proliferation and extension of specific microsatellite types (Bachtrog et al. 1999; Grady et al. 1992; Melters et al. 2013). Interspecific variation in local recombination rates may contribute to repeat instability (Pearson et al. 2005). Characteristics of DNA repair mechanisms, flanking sequences, and proteins involved in chromatin remodeling may also direct lineage-specific rates and genome abundance of microsatellites (Glenn et al. 1996; Mellon et al. 1996). Due to their high abundance and homopolymeric tracts, several types of TEs (such as *Alu* SINEs and CR1 LINEs) are known to facilitate the expansion of microsatellites through microsatellite seeding in which TE activity can result in duplications of both flanking and endogenous microsatellite arrays. Thus, genomes with TE-rich landscapes often display proportionately high amounts of microsatellite and could explain the relatively high abundances of microsatellites exemplified in both mammalian and squamate genomes (Castoe et al. 2011*b*; Cordaux and Batzer 2009; Janes et al. 2010; Primmer et al. 1997).

Results of this study provide a foundational comparative perspective for understanding microsatellite landscape characteristics and their evolution in vertebrate genomes. Our results highlight a tremendous degree of change across major vertebrate lineages and even among closely related species in genomic microsatellite landscapes, and further show evidence for rapid shifts in the tempo and patterns of evolutionary change among lineages. Because genome assemblies are inherently estimates of the true genome, our inferences of microsatellite landscapes based on these genome assemblies are also estimates, and subject to biases and errors of inferred genome assemblies. In general, we would expect our inference to be robust as typical microsatellite loci are often short enough to be readily incorporated into genome assemblies, but note that it is possible that centromeric microsatellite regions that were not effectively assembled may be absent from our analyses. Taxonomic sampling imbalances for complete genomes across the vertebrate tree also limit our current resolution to understand the ranges of variation, and tempos of evolution in microsatellite landscapes for many lineages outside mammals. As more non-mammalian vertebrate genome sequences become available, especially from squamate reptiles, fish, and amphibians, it is likely that even greater magnitudes of variation in microsatellite landscapes will be found than reported here.

Acknowledgements

Support was provided from startup funds from the University of Texas at Arlington to T.A.C.

References

Alföldi, J., Di Palma, F., Grabherr, M., Williams, C., Kong, L., Mauceli, E., et al. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature, **477**(7366): 587–591. doi:10.1038/nature10390. PMID:21881562.

Arcot, S.S., Wang, Z., Weber, J.L., Deininger, P.L., and Batzer, M.A. 1995. *Alu* repeats: a source for the genesis of primate microsatellites. Genomics, **29**(1): 136–144. doi:10.1006/geno.1995.1224. PMID:8530063.

Arzimanoglou, I.I., Gilbert, F., and Barber, H.R. 1998. Microsatellite instability in human solid tumors. Cancer, **82**(10): 1808–1820. doi:10.1002/(SICI)1097-0142(19980515)82:10<1808::AID-CNCR2>3.0.CO;2-J. PMID:9587112.

Bachtrog, D., Weiss, S., Zangerl, B., Brem, G., and Schlötterer, C. 1999. Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. Mol. Biol. Evol. **16**(5): 602–610. doi:10.1093/oxfordjournals.molbev.a026142. PMID:10335653.

Balaresque, P., King, T.E., Parkin, E.J., Heyer, E., Carvalho-Silva, D., Kraaijenbrink, T., et al. 2014. Gene conversion violates the stepwise mutation model for microsatellites in Y-chromosomal palindromic repeats. Hum. Mutat. **35**(5): 609–617. doi:10.1002/humu.22542. PMID:24610746.

Beckmann, J.S., and Weber, J.L. 1992. Survey of human and rat microsatellites. Genomics, **12**(4): 627–631. doi:10.1016/0888-7543(92)90285-Z.

Benet, A., Mollà, G., and Azorín, F. 2000. d(GA·TC)$_n$ microsatellite DNA sequences enhance homologous DNA recombination in SV40 minichromosomes. Nucleic Acids Res. **28**(23): 4617–4622. doi:10.1093/nar/28.23.4617. PMID:11095670.

Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., and Sayers, E.W. 2012. GenBank. Nucleic Acids Res. **40**: D48–D53. doi:10.1093/nar/gkr1202. PMID:22144687.

Benton, M.J., and Donoghue, P.C. 2007. Paleontological evidence to date the tree of life. Mol. Biol. Evol. **24**(1): 26–53. PMID:17047029.

Blackmon, H., and Adams, R.A. 2015. EvobiR: Tools for comparative analyses and teaching evolutionary biology. doi:10.5281/zenodo.30938.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience, **2**(1): 10–31. doi:10.1186/2047-217X-2-10. PMID:23870653.

Bruford, M.W., and Wayne, R.K. 1993. Microsatellites and their application to population genetic studies. Curr. Opin. Genet. Dev. **3**(6): 939–943. doi:10.1016/0959-437X(93)90017-J. PMID:8118220.

Butler, J.M. 2006. Genetics and genomics of core short tandem repeat loci used in human identity testing. J. Forensic Sci. **51**(2): 253–265. doi:10.1111/j.1556-4029.2006.00046.x. PMID:16566758.

Card, D.C., Schield, D.R., Reyes-Velasco, J., Fujita, M.K., Andrew, A.L., Oyler-McCance, S.J., et al. 2014. Two low coverage bird genomes and a comparison of reference-guided versus de novo genome assemblies. PLoS ONE, **9**(9): e106649. doi:10.1371/journal.pone.0106649. PMID:25192061.

Castoe, T.A., Poole, A.W., Gu, W., Jason de Koning, A., Daza, J.M., Smith, E.N., and Pollock, D.D. 2010. Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. Mol. Ecol. Res. **10**(2): 341–347. doi:10.1111/j.1755-0998.2009.02750.x. PMID:21565030.

Castoe, T.A., de Koning, A.J., Hall, K.T., Yokoyama, K.D., Gu, W., Smith, E.N., et al. 2011*a*. Sequencing the genome of the Burmese python (*Python molurus bivittatus*) as a model for studying extreme adaptations in snakes. Genome Biol. **12**(7): 406. doi:10.1186/gb-2011-12-7-406. PMID:21801464.

Castoe, T.A., Hall, K.T., Mboulas, M.L.G., Gu, W., de Koning, A.J., Fox, S.E., et al. 2011*b*. Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing. Genome Biol. Evol. **3**: 641–653. doi:10.1093/gbe/evr043. PMID:21572095.

Castoe, T.A., Streicher, J.W., Meik, J.M., Ingrasci, M.J., Poole, A.W., de Koning, A., et al. 2012. Thousands of microsatellite loci from the venomous coral snake *Micrurus fulvius* and variability of select loci across populations and related species. Mol. Ecol. Res. **12**(6): 1105–1113. doi:10.1111/1755-0998.12000. PMID:22938699.

Castoe, T.A., de Koning, A.J., Hall, K.T., Card, D.C., Schield, D.R., Fujita, M.K., et al. 2013. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proc. Natl. Acad. Sci. U.S.A. **110**(51): 20645–20650. doi:10.1073/pnas.1314475110.

Charlesworth, B., Sniegowski, P., and Stephan, W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature, **371**: 215–220. doi:10.1038/371215a0. PMID:8078581.

Chistiakov, D.A., Hellemans, B., and Volckaert, F.A. 2006. Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics. Aquaculture, **255**(1): 1–29. doi:10.1016/j.aquaculture.2005.11.031.

Clisson, I., Lathuilliere, M., and Crouau-Roy, B. 2000. Conservation and evolution of microsatellite loci in primate taxa. Am. J. Primatol. **50**(3): 205–214. PMID:10711534.

Cordaux, R., and Batzer, M.A. 2009. The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. **10**(10): 691–703. doi:10.1038/nrg2640. PMID:19763152.

Drummond, A.J., and Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7**(1): 214. doi:10.1186/1471-2148-7-214. PMID:17996036.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(5): 1792–1797. doi:10.1093/nar/gkh340. PMID:15034147.

Edwards, Y.J., Elgar, G., Clark, M.S., and Bishop, M.J. 1998. The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, *Fugu rubripes*: perspectives in functional and comparative genomic analyses. J. Mol. Biol. **278**(4): 843–854. doi:10.1006/jmbi.1998.1752. PMID:9614946.

FitzSimmons, N.N., Moritz, C., and Moore, S.S. 1995. Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. Mol. Biol. Evol. **12**(3): 432–440. PMID:7739385.

Gilbert, C., Meik, J., Dashevsky, D., Card, D., Castoe, T., and Schaack, S. 2014. Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. Proc. R. Soc. B Biol. Sci. **281**(1791): 20141122. doi:10.1098/rspb.2014.1122.

Glenn, T.C., Stephan, W., Dessauer, H.C., and Braun, M.J. 1996. Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability. Mol. Biol. Evol. **13**(8): 1151–1154. doi:10.1093/oxfordjournals.molbev.a025678. PMID:8865669.

Goldstein, D.B., Roemer, G.W., Smith, D.A., Reich, D.E., Bergman, A., and Wayne, R.K. 1999. The use of microsatellite variation to infer population structure and demographic history in a natural model system. Genetics, **151**(2): 797–801. PMID:9927470.

Grady, D.L., Ratliff, R.L., Robinson, D.L., McCanlies, E.C., Meyne, J., and Moyzis, R.K. 1992. Highly conserved repetitive DNA sequences are present at human centromeres. Proc. Natl. Acad. Sci. U.S.A. **89**(5): 1695–1699. doi:10.1073/pnas.89.5.1695.

Green, R.E., Braun, E.L., Armstrong, J., Earl, D., Nguyen, N., Hickey, G., et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science, **346**(6215): 1254449. doi:10.1126/science.1254449. PMID:25504731.

Gregory, T.R. 2015. Animal genome size database. Available from http://www.genomesize.com.

Hancock, J.M. 1995. The contribution of slippage-like processes to genome evolution. J. Mol. Evol. **41**(6): 1038–1047. PMID:8587102.

Hancock, J.M. 1996. Simple sequences and the expanding genome. Bioessays, **18**(5): 421–425. doi:10.1002/bies.950180512. PMID:8639165.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature, **432**(7018): 695–716. doi:10.1038/nature03154. PMID:15592404.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., et al. 2002. The Ensembl genome database project. Nucleic Acids Res. **30**(1): 38–41. doi:10.1093/nar/30.1.38. PMID:11752248.

Janes, D.E., Organ, C.L., Fujita, M.K., Shedlock, A.M., and Edwards, S.V. 2010. Genome evolution in Reptilia, the sister group of mammals. Annu. Rev. Genom. Hum. Genet. **11**: 239–264. doi:10.1146/annurev-genom-082509-141646. PMID:20590429.

Jarne, P., and Lagoda, P.J. 1996. Microsatellites, from molecules to populations and back. Trends Ecol. Evol. **11**(10): 424–429. doi:10.1016/0169-5347(96)10049-5. PMID:21237902.

Kelkar, Y.D., Tyekucheva, S., Chiaromonte, F., and Makova, K.D. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res. **18**(1): 30–38. PMID:18032720.

Kimpara, T., Takeda, A., Watanabe, K., Itoyama, Y., Ikawa, S., Watanabe, M., et al. 1997. Microsatellite polymorphism in the human heme oxygenase-1 gene promoter and its application in association studies with Alzheimer and Parkinson disease. Hum. Genet. **100**(1): 145–147. doi:10.1007/s004390050480. PMID:9225984.

La Spada, A.R., Paulson, H.L., and Fischbeck, K.H. 1994. Trinucleotide repeat expansion in neurological disease. Ann. Neurol. **36**(6): 814–822. doi:10.1002/ana.410360604. PMID: 7998766.

Lanfear, R., Calcott, B., Ho, S.Y., and Guindon, S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. **29**(6): 1695–1701. doi:10.1093/molbev/mss020. PMID:22319168.

Li, Y.C., Korol, A.B., Fahima, T., Beiles, A., and Nevo, E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol. Ecol. **11**(12): 2453–2465. doi:10.1046/j.1365-294X.2002.01643.x. PMID:12453231.

Martin, P., Makepeace, K., Hill, S.A., Hood, D.W., and Moxon, E.R. 2005. Microsatellite instability regulates transcription factor binding and gene expression. Proc. Natl. Acad. Sci. U.S.A. **102**(10): 3800–3804. doi:10.1073/pnas.0406805102.

McCouch, S.R., Chen, X., Panaud, O., Temnykh, S., Xu, Y., Cho, Y.G., et al. 1997. Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant Mol. Biol. **35**(1–2): 89–99. doi:10.1023/A:1005711431474. PMID:9291963.

Mellon, I., Rajpal, D.K., Koi, M., Boland, C.R., and Champe, G.N. 1996. Transcription-coupled repair deficiency and mutations in human mismatch repair genes. Science, **272**(5261): 557–560. doi:10.1126/science.272.5261.557. PMID:8614807.

Melters, D.P., Bradnam, K.R., Young, H.A., Telis, N., May, M.R., Ruby, J.G., et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. **14**(1): R10. doi:10.1186/gb-2013-14-1-r10. PMID:23363705.

Moxon, E.R., Rainey, P.B., Nowak, M.A., and Lenski, R.E. 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. **4**(1): 24–33. doi:10.1016/S0960-9822(00)00005-1. PMID:7922307.

Neff, B.D., and Gross, M.R. 2001. Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. Evolution, **55**(9): 1717–1733. doi:10.1554/0014-3820(2001)055[1717:MEIVIF]2.0.CO;2. PMID:11681728.

O'Meara, B.C., Ané, C., Sanderson, M.J., and Wainwright, P.C. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution, **60**(5): 922–933. doi:10.1554/05-130.1. PMID:16817533.

Paradis, E., Claude, J., and Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics, **20**(2): 289–290. doi:10.1093/bioinformatics/btg412. PMID:14734327.

Pardue, M., Lowenhaupt, K., Rich, A., and Nordheim, A. 1987. $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ sequences have evolutionarily conserved chromosomal locations in *Drosophila* with implications for roles in chromosome structure and function. EMBO J. **6**(6): 1781–1789. PMID:3111846.

Payseur, B.A., and Nachman, M.W. 2000. Microsatellite variation and recombination rate in the human genome. Genetics, **156**(3): 1285–1298. PMID:11063702.

Pearson, C.E., Edamura, K.N., and Cleary, J.D. 2005. Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. **6**(10): 729–742. doi:10.1038/nrg1689. PMID:16205713.

Pepín, L., Amigues, Y., Lépingle, A., Berthier, J.-L., Bensaid, A., and Vaiman, D. 1995. Sequence conservation of microsatellites between *Bos taurus* (cattle), *Capra hircus* (goat) and related species. Examples of use in parentage testing and phylogeny analysis. Heredity, **74**(1): 53–61. PMID:7852099.

Primmer, C.R., Raudsepp, T., Chowdhary, B.P., Møller, A.P., and Ellegren, H. 1997. Low frequency of microsatellites in the avian genome. Genome Res. **7**(5): 471–482. PMID:9149943.

Pyron, R.A., Burbrink, F.T., and Wiens, J.J. 2013. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. BMC Evol. Biol. **13**(1): 93. doi:10.1186/1471-2148-13-93. PMID:23627680.

R Development Core Team. 2012. R: a language and environment for statistical computing.

Rambaut, A., and Drummond, A. 2013. TreeAnnotator v1.7.0. Available from http://beast.bio.ed.ac.uk/.

Rambaut, A., and Drummond, A. 2014. Tracer v1.6. Available from http://beast.bio.ed.ac.uk/.

Ramsay, L., Macaulay, M., Cardle, L., Morgante, M., Ivanissevich, S.d., Maestri, E., et al. 1999. Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. Plant J. **17**(4): 415–425. doi:10.1046/j.1365-313X.1999.00392.x. PMID:10205898.

Ranum, L.P., and Day, J.W. 2002. Dominantly inherited, noncoding microsatellite expansion disorders. Curr. Opin. Genet. Dev. **12**(3): 266–271. doi:10.1016/S0959-437X(02)00297-6. PMID: 12076668.

Revell, L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. **3**(2): 217–223. doi:10.1111/j.2041-210X.2011.00169.x.

Richard, G.F., and Pâques, F. 2000. Mini- and microsatellite expansions: the recombination connection. EMBO Rep. **1**(2): 122–126. doi:10.1093/embo-reports/kvd031. PMID:11265750.

Richards, R.I., and Sutherland, G.R. 1994. Simple repeat DNA is not replicated simply. Nat. Genet. **6**(2): 114–116. doi:10.1038/ng0294-114. PMID:8162063.

Rico, C., Rico, I., and Hewitt, G. 1996. 470 million years of conservation of microsatellite loci among fish species. Proc. R. Soc. B Biol. Sci. **263**(1370): 549–557. doi:10.1098/rspb.1996.0083.

Schlötterer, C. 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma, **109**(6): 365–371. doi:10.1007/s004120000089. PMID:11072791.

Schlotteröer, C., Amos, B., and Tautz, D. 1991. Conservation of polymorphic simple sequence loci in cetacean species. Nature, **354**(6348): 63–65. doi:10.1038/354063a0. PMID:1944571.

Schug, M.D., Hutter, C.M., Noor, M.A., and Aquadro, C.F. 1998. Mutation and evolution of microsatellites in *Drosophila melanogaster*. Genetica, **102**: 359–367. PMID:9720288.

Shaffer, H.B., Minx, P., Warren, D.E., Shedlock, A.M., Thomson, R.C., Valenzuela, N., et al. 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol. **14**(3): R28. doi:10.1186/gb-2013-14-3-r28. PMID:23537068.

Shaney, K.J., Card, D.C., Schield, D.R., Ruggiero, R.P., Pollock, D.D., Mackessy, S.P., and Castoe, T.A. 2014. Squamate reptile genomics and evolution. *In* Venom Genomics and Proteomics. *Edited by* P. Gopalakrishnakone and J.J. Calvete. Springer. pp. 1–18.

Shedlock, A.M., Botka, C.W., Zhao, S., Shetty, J., Zhang, T., Liu, J.S., et al. 2007. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. Proc. Natl. Acad. Sci. U.S.A. **104**(8): 2767–2772. doi:10.1073/pnas.0606204104.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics, **139**(1): 457–462. PMID:7705646.

St John, J.A., Braun, E.L., Isberg, S.R., Miles, L.G., Chong, A.Y., Gongora, J., et al. 2012. Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. Genome Biol. **13**(1): 415. doi:10.1186/gb-2012-13-1-415. PMID:22293439.

Sun, C., and Mueller, R.L. 2014. Hellbender genome sequences shed light on genomic expansion at the base of crown salamanders. Genome Biol. Evol. **6**(7): 1818–1829. doi:10.1093/gbe/evu143. PMID:25115007.

Sun, Y.-B., Xiong, Z.-J., Xiang, X.-Y., Liu, S.-P., Zhou, W.-W., Tu, X.-L., et al. 2015. Whole-genome sequence of the Tibetan

frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes. Proc. Natl. Acad. Sci. U.S.A. **112**(11): E1257–E1262. doi:10.1073/pnas.1501764112.

Tóth, G., Gáspári, Z., and Jurka, J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. **10**(7): 967–981. doi:10.1101/gr.10.7.967. PMID:10899146.

Usdin, K. 1998. NGG-triplet repeats form similar intrastrand structures: implications for the triplet expansion diseases. Nucleic Acids Res. **26**(17): 4078–4085. doi:10.1093/nar/26.17.4078. PMID:9705522.

Vonk, F.J., Casewell, N.R., Henkel, C.V., Heimberg, A.M., Jansen, H.J., McCleary, R.J., et al. 2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. Proc. Natl. Acad. Sci. U.S.A. **110**(51): 20651–20656. doi:10.1073/pnas.1314702110.

Wan, Q.-H., Pan, S.-K., Hu, L., Zhu, Y., Xu, P.-W., Xia, J.-Q., et al. 2013. Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator. Cell Res. **23**(9): 1091–1105. doi:10.1038/cr.2013.104. PMID:23917531.

Wang, Z., Pascual-Anaya, J., Zadissa, A., Li, W., Niimura, Y., Huang, Z., et al. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat. Genet. **45**(6): 701–706. doi:10.1038/ng.2615. PMID:23624526.

Wayne, R.K., and Ostrander, E.A. 2007. Lessons learned from the dog genome. Trends Genet. **23**(11): 557–567. doi:10.1016/j.tig.2007.08.013. PMID:17963975.

Wooster, R., Cleton-Jansen, A.-M., Collins, N., Mangion, J., Cornelis, R., Cooper, C., et al. 1994. Instability of short tandem repeats (microsatellites) in human cancers. Nat. Genet. **6**(2): 152–156. doi:10.1038/ng0294-152. PMID:8162069.