Global annotation and molecular evolutionary analysis of genomic repeats in the Painted Turtle, *Chrysemys picta*.

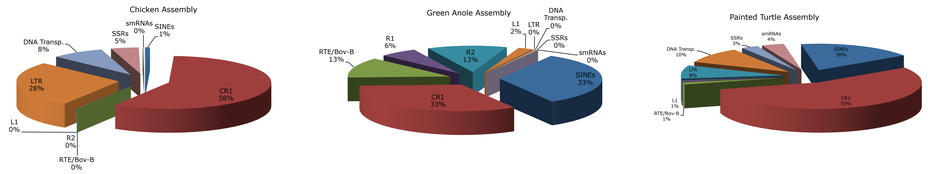
Quan Fang[1,2], Christopher Botka[2] and Andrew Shedlock[3]

[1]Boston University, Bioinformatics Graduate Program, [2]Harvard Medical School, Research Computing, [3]College of Charleston, Biology Department, MUSC-HML

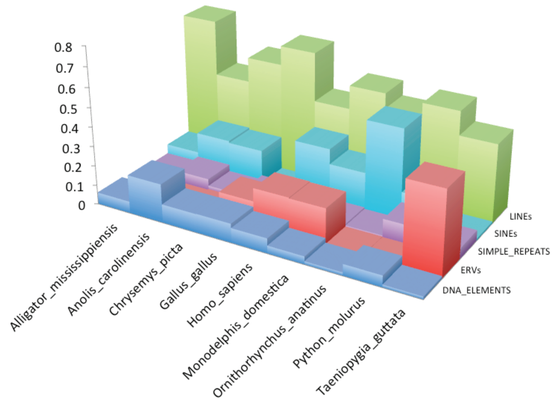
Introduction

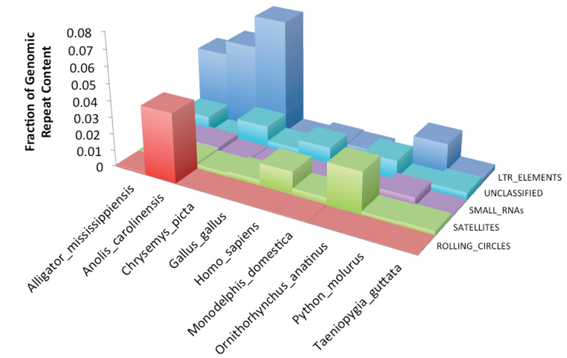
The analysis of genome-scale information from the sister group of mammals, Reptilia, is presently closing a large gap in the study of vertebrate biology. Turtles are of central importance to this effort in that they have been difficult to place accurately in the amniote tree of life, they exhibit a number of unique phenotypic and life history traits among vertebrates, and are threatened globally by a variety of human impacts on the environment. In particular, annotating the abundance and age of different classes of mobile DNA repeats and DNA divergence values provides a basis for testing alternative models of molecular evolution which describe forces that have shaped our own genomes uniquely as compared to other amniote species. We annotated global repeat content for the full set of pre-published genomic sequence reads generated for the painted turtle, C. picta, the first turtle species now having its complete genome assembled under NIH sponsorship. The Repeatmasker software program, Perl scripting and Excel spreadsheets were used to extract, sort, and chart comparative statistics for specific genomic repeat classes. Comparison of sequence divergence values based on original turtle query sequence aligned against a comprehensive reference database for vertebrate repeats reveals differential age distributions among major repeat subfamilies. A broad range of values was recovered for all types examined with lower values indicating the presence of relatively young active elements still driving turtle genomic diversity and distributions skewed toward higher mean values suggesting numerous distinct lineages of older inactive repeats that can be viewed as molecular fossils embedded in turtle chromosomes. The present study suggests that alternative modes of evolution are likely operating in different turtle repeat lineages whose proliferation and persistence cannot be easily explained by a strict vertical inheritance of single source genes, such as seen for the dominant form of interspersed repeats in the human genome and that contrasts with the largely inactive compact repetitive DNA landscape of birds.

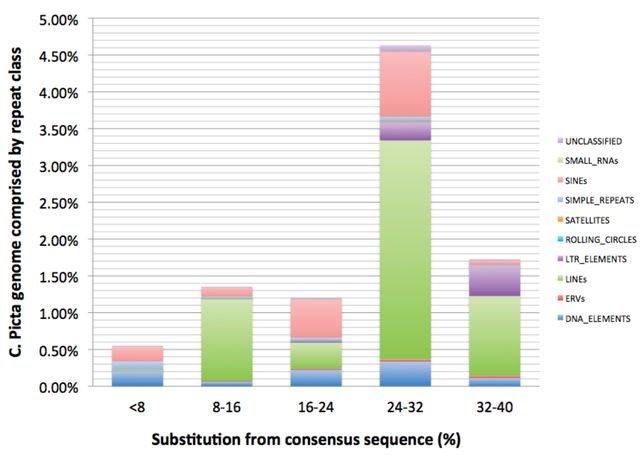
Percentages of the repetitive landscape for three complete amniote genome assemblies by major repeat families.



Genomic fraction of major repeat types among 9 amniote genome assemblies



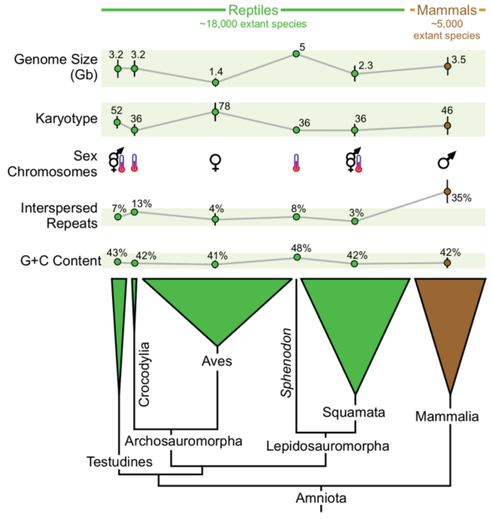




	Thousands of elements	Mb of Sequence occupied	Percent of Genome	Percent of Repeat Content
Retroelements	714.17	208.27	8.04	81.89
SINEs	320.73	45.22	1.75	17.78
Penelope	3.27	0.48	0.02	0.19
LINEs	331.72	143.05	5.52	56.24
L2_CR1_Rex	311.76	135.85	5.25	53.41
R1_LOA_Jockey	2.86	1.15	0.04	0.45
RTE_Bov-B	7.18	3.31	0.13	1.30
L1_CIN4	6.02	2.18	0.08	0.86
LTR_elements	61.72	20.00	0.77	7.86
Gypsy_DIRS1	41.24	17.45	0.67	6.86
Retroviral	16.89	2.15	0.08	0.85
DNA Transposons	173.11	24.08	0.93	9.47
hobo-Activator	87.70	6.85	0.26	2.70
Tc1-IS630-Pogo	16.65	2.39	0.09	0.94
PiggyBac	13.64	2.52	0.10	0.99
Tourist_Harbinger	8.62	3.20	0.12	1.26
	0.00	0.00		
Unclassified	22.74	3.50	0.14	1.38
Total Interspersed Repeats	0.00	235.85	9.11	92.73
Small_RNA	48.95	8.67	0.33	3.41
Satellites	4.48	0.53	0.02	0.21
Simple_Repeats	133.87	6.67	0.26	2.62
Low_Complexity	244.33	11.17	0.43	4.39
Bases Masked		254.34	9.82	100.00
Total Length		2,589.74		

Discussion

A basic assumption (with caveats) in molecular evolution is that sequence non-identity between aligned query vs. reference sequence provides an index of relative age based on accumulated nucleotide substitutions over time and this is supported by empirical studies. Patterns of turtle repeat divergence reveal a mix of both young active and old inactive elements whose proliferation and persistence cannot be easily explained by a strict vertical inheritance of single source genes, such as seen for L1s in mammals. Instead, turtle LINEs such as CR1 and L1-2 are likely operating under a multiple source gene model where copies of original elements remain active and diversify from multiple loci over time. The present study reveals alternative modes of evolution are likely operating in different turtle repeat lineages and skewed distributions. The comprehensive analysis of turtle repeats paints a portrait of genomic complexity in both pattern and process that contrasts with mammalian, avian and lizard of amniote genomic diversification. We anticipate further work on the locus-specific profiles of turtle repeat families such as CR1 will provide valuable new insights into the vertebrate genotype-phenotype relationship.



Sparkline plots of amniote genomic diversity summarizing data published in Shedlock et al., 2007, Janes et al. 2010. The width of clades on the tree is proportional to species diversity (~300 species in Testudines, 23 species of crocodilians, ~9800 species of birds, 3 species of Sphenodon, and ~7500 species of squamates). Genome size in gigabases (Gb) is reported as the clade-wide average. Karyotype is reported as the average number of chromosomes, including micro, macro, and sex chromosomes. For sex chromosomes, female heterogamety is denoted by the female symbol, male heterogamety by the male symbol, and temperature-dependent sex determination by the thermometer symbol. Interspersed repeats and GC content are reported as a percentage of the total genome (these estimates are based on a relatively small sample size for reptiles: 1 turtle, 1 crocodilian, 4 birds, Sphenodon, and 1 squamate). Bars behind the data points are standard deviation and the light green background marks the data range in reptiles. Note that the phylogenetic position of turtles (Testudines) remains contentious based on conflict among morphological and molecular characters. The position indicated here is based on a consensus of views summarized by Shedlock and Edwards, 2009.

fangquan@bu.edu | botka@hms.harvard.edu | shedlockam@cofc.edu

Acknowledgements

Turtle, Snake Alligator, Genome Working Groups at UCSC (DH), WUStL (WW), UCDen (DP), MSU (DR), the Laboratory of Scott Edwards, Harvard OEB MCZ, Manfred Grabherr, K. Lindblad-Toh, Broad Ins titute of MIT and Harvard, Craig Harms Lab, NCSU Vet School and CMAST, Alex Feltus, Clemson Genetics & Biochem & CUGI, Chris Amemiya, NIH & Benaroya Research Institute,The Orchestra High Performance Compute Cluster at Harvard Medical School was used for data analysis for this project.

Funding: NSF • COFC-MUSC-HML • Harvard Medical School

References

1. Shedlock, A. M., and S. V. Edwards. 2009. Amniotes (Amniota). Pp. 373-380 In: The Timetree of Life (S. B. Hedges and S. Kumar, Eds.), Oxford University Press, New York.
2. Shedlock, A. M. 2006. Phylogenomic investigation of CR1 LINE diversity in reptiles. Systematic Biology 55(6):902-911. [cover article]
3. Shedlock, A. M., C. W. Botka, S. Zhao, J. Shetty,T. Zhang, J. S. Liu, P. J. Deschavanne, and S. V. Edwards. 2007. Phylogenomics of non-avian reptiles and the structure of the ancestral amniote genome. Proc. Natl. Acad. Sci. USA. 104:2767-2772.
4. Shedlock, A. M. and N. Okada. 2000. SINE insertions: Powerful tools for molecular systematics. Bioessays 22:148-160.
5. Shedlock, A. M., K. Takahashi, and N. Okada. 2004. SINEs of speciation: tracking lineages with retroposons. Trends in Ecology and Evolution. 19(10): 545-553.
6. Smit AFA, Hubley R, Green P (2004) RepeatMasker Open-3.0.5. Available at: www.repeatmasker.org
7. Jurka, J., V. V. Kapitonov,A. Pavlicek, P. Klownowski, O. Kohany, and J. Walichiewicz. 2005. Repbase Update, a database of eukaryotic repetitive elements (www.girinst.org/repbase/index.html). Cytogenet. Genome Res. 110:462–467
8. Shedlock et al. 2004. TRENDS in Ecology and Evolution. Vol 19 No 10. p 545.
9. Y.Cao, M.D. Sorenson, Y. Kumazawa, D.P. Mindell, M. Hasegawa, Gene 259, 139 (2000).
10. R. Zardoya, A. Meyer, Proc.Natl. Acad. Sci. U.S.A. 95, 14226 (1998).
11. S.B. Hedges, S. Kumar, Trends Genet. 20, 242 (2004).
12. O. Rieppel, R.R. Reisz, Ann. Rev. Ecol. Syst. 30, 1 (1999).