UNIVERSITY OF CALIFORNIA

Los Angeles

Quantitative analysis of phylogenetic informativeness, signal and noise in ultraconserved elements within Percomorpha and Neoaves

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Biology

by

Princess Scheran Gilbert

2017

ABSTRACT OF THE DISSERTATION

Quantitative analysis of phylogenetic informativeness, signal and noise in ultraconserved elements within Percomorpha and Neoaves

by

Princess Scheran Gilbert
Doctor of Philosophy in Biology
University of California, Los Angeles, 2017
Professor Michael Edward Alfaro, Chair

The work described herein explores the ability of UCEs to resolve clade relationships within the vertebrate tree of life, specifically percomorph fishes and neoaves birds. To do so, I used Phylogenetic Informativeness and the phylogenetic signal to noise ratios in order to calculate the ability of a marker to resolve deep clade relationships, I also developed an automated pipeline in order to calculate these statistical measures for each of the nucleotides in thousands of UCEs. UCE cores and their respective flanking regions are large and spread out across the entire genome. Thus the approaches and findings described here are the first to analyze UCEs at a fine scale (per nucleotide) and the first to assess this phylogenetic marker type using these methods. Chapter 2 has been previously published as Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. (2015) Gilbert PS, Chang J, Pan

C, Sobel EM, Sinsheimer JS, Faircloth BC, Alfaro ME. Mol Phylogenet Evol. 2015 Nov;92:140-6. doi: 10.1016/j.ympev.2015.05.027.) Chapter 3 is in preparation for submission.

The dissertation of Princess Scheran Gilbert is approved.

Janet S. Sinsheimer

Paul H. Barber

Michael Edward Alfaro, Committee Chair

University of California, Los Angeles

2017

DEDICATION

For my grandmother Athalia, who passed on before I was born and whom I never had the pleasure of meeting but whose love of animals and nature courses through my veins every moment of every day.

For Grandma Gert, who never finished elementary school but always asked if I was keeping up with my studies and who I know would be proud that I finally "finished my schoolin’ ".

And for those who rooted for me to finish and cheered me on until their very last breath; Cousin Laura, Aunt Norma, Sheri Gilbert, Uncle Henry, Neicy, Calvin and Mingo.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Chapter Two is a reproduction of:

Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. Gilbert PS, Chang J, Pan C, Sobel EM, Sinsheimer JS, Faircloth BC, Alfaro ME. Mol Phylogenet Evol. 2015 Nov; 92:140-6. doi: 10.1016/j.ympev.2015.05.027.

PSG reviewed literature, drafted and revised the manuscript and created graphics. JC, CP, and EMS assisted in computational algorithm design. JSS, BCF and MEA helped to review study design and revise drafts of the manuscript.

Chapter three is in preparation for the Journal of Molecular Phylogenetics and Evolution.

Funding was provided by

NIH Genomic Analysis and Interpretation Training Program Fellowship, GAANN (U.S. Department of Education Graduate Assistance in Areas of National Need Graduate Support Fellowship, Eugene Cota-Robles and the NSF UCLA Competitive Edge.

Thank you to my dissertation committee members. You championed my academic success over my entire nine year graduate career and I am grateful for your support. Thank you for all the late Wednesday and Friday evening meetings, field work opportunities in Bali, Indonesia, conference talk critiques, paper revisions, letters of rec., congratulatory post-orals champagne, hallway pep-talks and for approving my dissertation!

Thanks also to the Alfaro lab; both current and former members.

Thanks to the numerous post-docs who filled a special academic role in my graduate school learning. Thanks also to former Wayne lab members who mentored me during my first 4 years of my grad school; Klaus-Peter K., Borja M., Shauna P., Katherine P. , Olaf and Dorrine T. and Katy S. and more generally members of the EEB department who helped my figure out how grad school worked, Tessa V., Jocelyn Y., Deb P., Graham S., Dave J., Peggy F., Peter N., Dan B., Camille B., Adam H. F., Ryan H., Jaime C., Thea W., Doug S., and Victoria S. And thank you to my dissertation writing group for helping me cross the finish line: Veronica F., Rita R., Stephanie S., Mairin B., and D. Bird.

A special acknowledgment is due to the person who spent countless hours introducing me to ecology and evolutionary biology research and who always left me feeling inspired to go further and reach higher. John Pollinger, I can’t thank you enough and I hope you know how much of a difference you’ve made in my life.

Thanks for teaching me how to scuba dive, Mike A. Thank you for teaching me how to keep my calm Amanda, Ian and KYC friends, and thank you for normalizing my trials and tribulations GSOC, Tanya B., Nicole G., Misa D., Mzilikazi K., Veena H., Cheryl F., Ancy C., and CAPS.

Thank you to my closest friends who stood by me, and sometimes carried me, throughout the last nine years: Valerie O. Pam T., Sam W., Kayleigh P., David G., Jonathan G., Ryan E., Josh C., Tina M., Marco G., and Scott B.

And lastly Mom and Dad! Your endless support, encouragement and love motivated me more than you will ever know! Thank you. I know that you always told everyone I was doing this on my own, but I did this with you both and I don't think that I could have completed this journey without you.

"We did it!"

VITA/BIOGRAPHICAL SKETCH

DEGREES

Master of Science, Evolutionary Biology, UCLA 2012.
Advancement to Candidacy, Biology, UCLA 2011.
Bachelor of Science, Biology; **Minor**, African American Studies, UCLA 2007.

RESEARCH INTERESTS

Molecular Ecology, Phylogenomics, Phylogenetics, Biogeography, Population Genetics, High-throughput Sequencing Approaches, Genomic Analysis, Neoaves, Acanthomorpha, The Coral Triangle, Madagascar Fauna & Flora.

ACADEMIC & RESEARCH FELLOWSHIPS/ GRANTS

NIH Genomic Analysis and Interpretation Training Program Fellowship ($34,943.33)
2013-2014 Academic Year, UCLA
GAANN Graduate Support Fellowship ($26,984)
(U.S. Department of Education Graduate Assistance in Areas of National Need)
2012-2013 Academic Year, UCLA
NIH Genomic Analysis and Interpretation Training Program Fellowship ($21,180)
2011-2012 Academic Year, UCLA
NIH Genomic Analysis and Interpretation Training Program Fellowship ($21,180)
2010-2011 Academic Year, UCLA
Eugene Cota-Robles Graduate Research Mentorship Year Fellowship ($20,000)
2009-2010 Academic Year, UCLA
GAANN Graduate Support Fellowship ($17,707)
(U.S. Department of Education Graduate Assistance in Areas of National Need)
2008-2009 Academic Year, UCLA
Eugene Cota Robles Graduate Fellowship ($18,000)
2007-2008 Academic Year, UCLA
GAANN Graduate Support Fellowship ($20,919)
2007-2008 Academic Year, UCLA (offer declined)

HONORS & AWARDS

Ecology & Evolutionary Biology Travel Award ($500)
Summer 2012, UCLA
Charlotte Magnum Student Support (Covered Registration Fees for Annual SICB Meeting)
January 2012, Society for Integrative & Comparative Biology
Ecology & Evolutionary Biology Research Award ($1,000)
Summer 2010, UCLA
Carl Storm Underrepresented Minority Fellowship (Travel Funds to Gordon Research Conference)
Summer 2009, Gordon Research Conference on Evolutionary & Functional Genomics
Diversity Recruitment Funding Award (Travel Funds to Recruit at Annual SMBE Meeting)
Summer 2009, Genomic Analysis & Training Program, UCLA
6th Annual Preparing Future Faculty Summer Institute (Howard University) Attendance Nomination
June 2009, Graduate Division, UCLA
17th Annual Compact for Faculty Diversity Institute Attendance Nomination
October 2009, Graduate Division, UCLA

PUBLICATIONS & PRESENTATIONS

Gilbert, PS, Wu, J, Simon, M, Sinsheimer JS, Alfaro ME. (2017) Filtering nucleotide sites from ultraconserved elements by phylogenetic signal to noise ratio improves the precision of the avian phylogeny. *In prep*
Gilbert, PS, Chang J, Pan C, Sobel EM, Sinsheimer JS, Faircloth BC, Alfaro ME (2015). Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. Molecular Phylogenetics and Evolution 92: 140-146.

Oral Presentations & Poster Presentations

June 2012. Genomic Analysis & Training Program Research Symposium, UCLA
January 2012. Society for Integrative and Comparative Biology Annual Conference. Charleston, South Carolina.
May 2009. Diversification and Adaptive Variation in the Chameleons of Madagascar. Ecolunch Graduate Series, UCLA.
July 2012. Society of Evolution Conference, Ottawa Canada
June 2011. Genomic Analysis & Training Program Research Symposium, UCLA

TEACHING EXPERIENCE

Guest Lecturer

Introduction to Marine Science EEB109-Sea Grass, Rocky Reefs & Kelp Forests
Life Science 1-Animal Diversity

Teaching Fellow/Assistantships

Life Science 1 - Introduction to Ecology & Evolution
Molecular Evolution
Introduction to Marine Ecology/ Intro to Marine Ecology Lab

UCLA Undergraduate Research Center Workshops

How to Enroll in the Student Research Program at UCLA
How to Get Involved & Find Funding for Undergraduate Research
How to Make A Scientific Poster/ Presentation,
How to Write A Scientific Article

PROFESSIONAL SOCIETY MEMBERSHIPS

American Association for the Advancement of Science
American Society of Naturalists
Association for Black Women in Higher Education
Conservation International
Society for Conservation Biology
Society for the Study of Evolution
Society for Integrative and Comparative Biology

SERVICE

The Leadership Alliance National Symposium: Invited Graduate Student Judge
Summer 2013
The Leadership Alliance National Symposium: Invited Graduate Student Judge
Summer 2012
1st Joint Congress on Evolutionary Biology Undergraduate Diversity: Participating Mentor
Summer 2012
UCLA DNA Day Presenter/ Workshop Leader
Spring 2010, 2011 & 2012, Genomic Analysis Training Program/Human Genetics Dept., UCLA
SMBE Diversity Mentoring Program: Graduate Student Mentor
Summer 2009, Society for Molecular Biology and Evolution
URC-CARE Graduate Student Mentor, UCLA
(Undergraduate Research Center- Center for Academic & Research Excellence)
Summer 2008- December 2012
AAP Scholars Day Science Lab Tour Guide, UCLA
(Academic Advancement Program)
Spring 2006, Spring 2007, Spring 2008, Spring 2009, Spring 2010
Undergraduate Researchers Mentored
Jing Wu, Alan Kha, Daisy Carrillo, Mark Oliva

SCUBA CERTIFICATION (30 DIVES)

AAUS Open Water, Winter 2008
AAUS Scientific Diver Certification, Spring 2008

Chapter One

Introduction

1.1 Defining UCEs

Ultraconserved elements (UCEs) are short regions of DNA that highly similar (>80% identical) distantly related species (Bejarano et al., 2004; McCormack et al. 2012; Siepel et al. 2005; Stephen et al. 2008). UCEs are dispersed throughout the vertebrate genome. Immediately upstream and downstream of the highly similar region, or core, are flanking regions of DNA sequence that increase in variation as distance from the core increases. Flanking regions are where nucleotide substitution rates increase in variation and as demonstrated in Gilbert et al. (2015) this is where phylogenetic informativeness becomes highest. Interestingly, UCE core regions can include gaps in the core sequence region and therefore still carry phylogenetic signal despite their low variation (Chapter 1 Appendix Figures 1-10).

UCEs can be used for analyzing the speed of the rate of UCE evolution via comparisons of molecular clock estimates and substitution rates. Stephen et al. (2008) was able to show an increased substitution rate in amniote taxa using UCE markers, an increase that could not be detected with coding sequences alone. The contrast between substitution rates in Stephen et al.'s (2008) coding sequences and the substitution rates of UCEs highlights exactly why these markers could be uniquely appropriate for deep-time phylogenetic questions: their core region's slowly evolving nature in combination with variable-rate flanking regions.

Research has shown that UCEs can be spatially involved in gene transcription such as at splicing sites, exonic untranslated regions or UTRs, as well as near or within protein coding regions (Bejerano et al., 2004; Siepel et al., 2005; Baira et al., 2008; and Stephen et al., 2008). As such, genetic studies have largely focused on the functional aspects of UCEs and their application in human disease. However UCEs in phylogenetic analysis are becoming increasingly more common in systematic research.

1.2 UCEs in Systematics

The increasing use of coding and non-coding regions along with coding regions in comparative genomics have allowed evolutionary biologists to exploit the commonly shared genetic features of distantly related organisms (Boffelli et al. 2004; Siepel et al. 2005; Margulies and Birney, 2008). For example, a remarkable amount of sequence homology can be found among shared cis-regulatory elements in humans, mice, rats, chickens, frog and fish (Boffelli et al., 2004).

UCEs have demonstrated high phylogenetic utility and have been applied to historically difficult clades such as archosaurs, birds, bees, fish(Faircloth et al. 2013, 2014; McCormack et al. 2013; Smith et al. 2014) . However, clades have remained unresolved because either they are the result of adaptive radiations which create short internodes followed by long branches, incomplete lineage sorting or they are related to one-another deep in time and finding non-coding loci with phylogenetic signal deep in time and which have not become saturated proves difficult.

1.3. Phylogenetic Informativeness

A locus or character is phylogenetically informative if it has the power to resolve a polytomy and remain unchanged along the branches leading from that polytomy to the tips of the tree; this last stipulation is to insure homoplasy does not swamp out the informative change. (Townsend et al. 2007). There are number of ways of applying this principle in evolutionary biology and it is an active area in phylogenetic research. This dissertation is based upon and largely benefits from recent developments in PI (Su et al. 2014, 2015; Townsend et al. 2007, 2012).

Central to the study of the tree of life, systematics and phylogenetics is that of phylogenetic resolving power. Townsend has built upon this evolutionary theoretical framework to develop a quantifiable way to assess the phylogenetic resolving power of a given genetic locus (Townsend 2007). By assessing the phylogenetic informativeness of a given set of characters during a specified time epoch, one can determine exactly which markers will yield the most phylogenetic information for a given group of taxa at that node in tree (Townsend, 2007).

To start, we define what a phylogenetically informative character is. If we imagine a hypothetical star phylogeny with four tip taxa, a, b, c and d whose common ancestor occurred at time T, a character that evolved at the optimal evolutionary rate of change would resolve this star phylogeny or polytomy. A character would be able to resolve a polytomy by evolving along an internal node and not change along any of the subsequent branch lengths or tips.

Character assessment is based on approximating the optimal rate of change for a phylogenetic character or locus. However, characters never evolve at the optimal rate. Phylogenetic Informativeness (PI) is an index or relative informativeness measure; a function which is defined by ρ for a given node at time *T*, an ancestral node occurring at time t_0 and an evolutionary or substitution rate for that character, λ (Townsend 2007). For a hypothetical polytomy, it is assumed that t_0 is much smaller than *T*. Thus taking the limit as t_0 approaches 0 yields an equation that is maximized when the relative rate $\dot{\lambda}$ equals the optimal rate λ, and ρ equals 1, thus $\rho t_0(T, t_0,\lambda)=[e^{-\lambda(4T-2t0)} -e^{-\lambda(4T-t0)}]/ [e^{-1/4T(4T-2t0)}-e^{-1/4T(4T-t0)}]$.

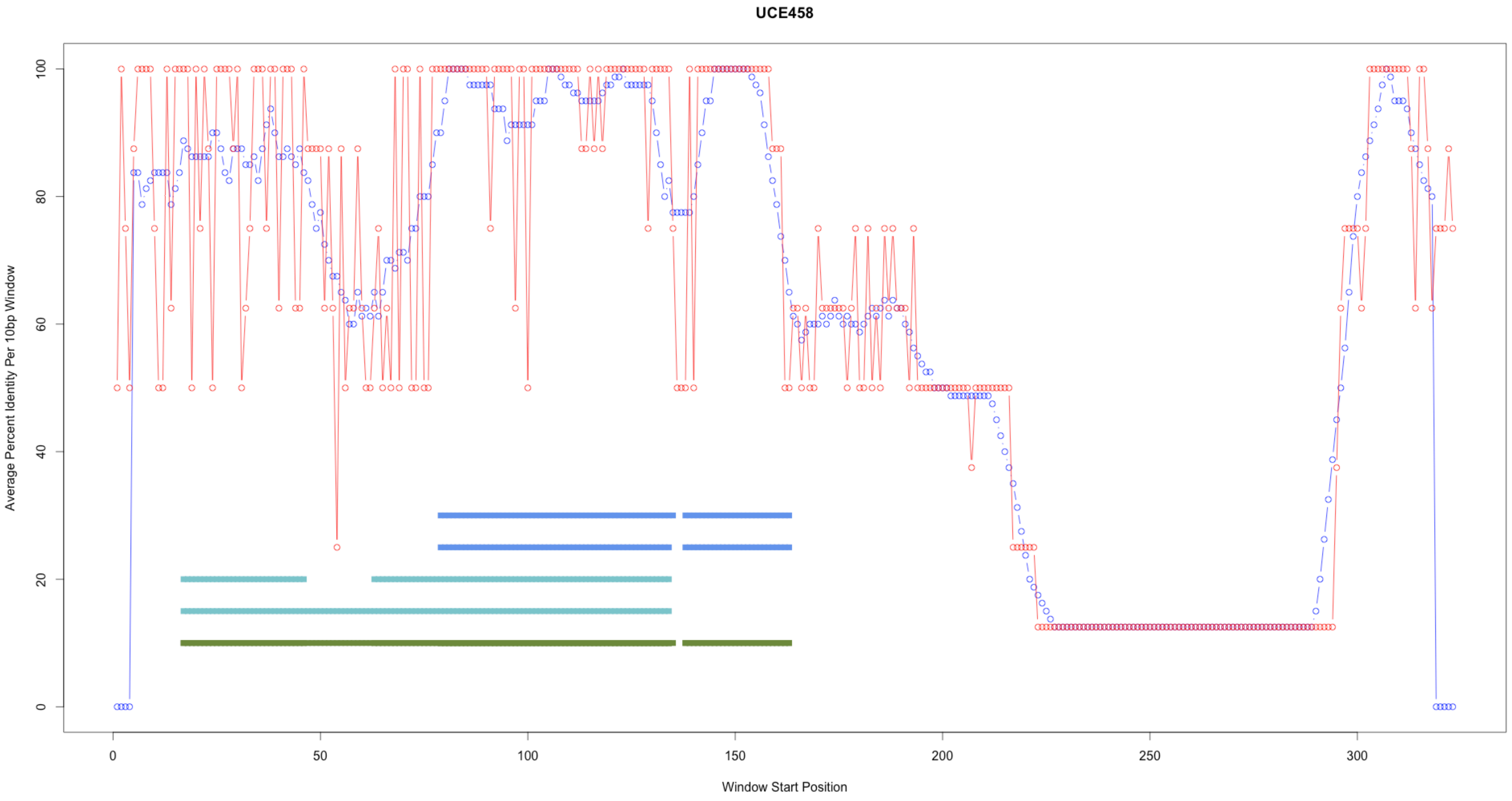
However, we might want to know the area under the curve defined by ρ_0. Integrating ρ_0 in the equation above gives smaller substitution rates a higher informativeness value, so the authors normalized ρ_0 and allowed for more than one character or for a sequence. Thus informativeness is calculated for each site in a given locus and their respective substitution rate, λ, and then is summed across all character sites in the sequence. This value can then be integrated over the time period of interest. Thus a prediction of the phylogenetic power of specific characters for explicit historical time periods can be established by $\rho(T,\lambda_1,\ldots,\lambda_n)=\Sigma 16\lambda_i^2 Te^{-4\lambda_i T}$.

1.4 The Probability of Phylogenetic Signal, Noise, and Polytomy

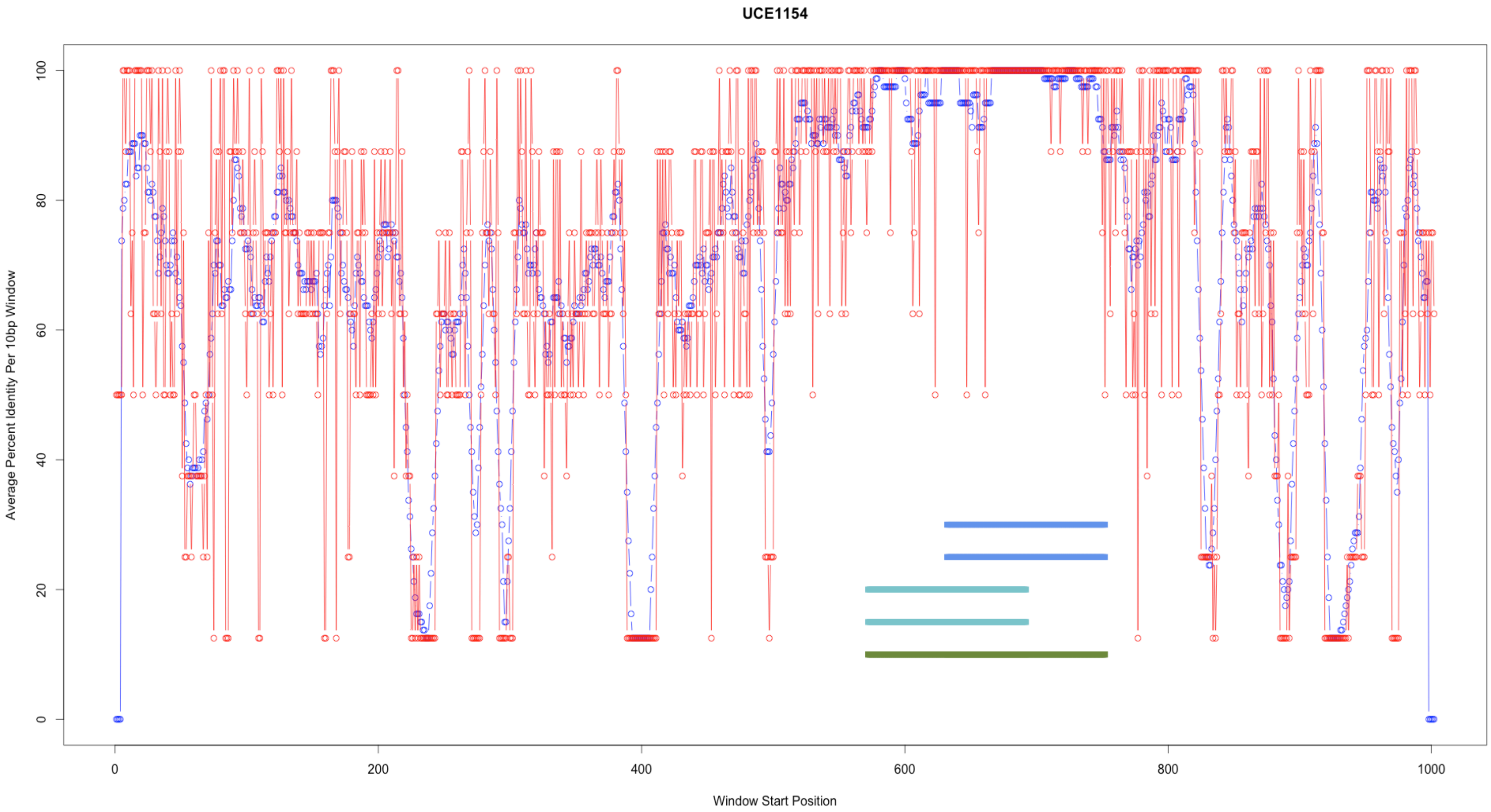
Closely tied to the concept of phylogenetic informativeness (PI) is the measure of phylogenetic signal and noise. Because PI does not account for homoplasy, Townsend et al. (2012) developed a method that can. For a four-taxon unrooted tree, a character is said to have high phylogenetic signal if the probability of observing a parsimony informative synapomorphic site pattern at the leaves (e.g. the branches extending to the tips of the tree) of the taxa is high. A character is said to have phylogenetic noise if the probability distribution function over time for homoplasious site patterns mimics the correct pattern and misleads parsimony or other analyses. Homoplasy occurs when multiple character substitutions occur at the same site after the initial character evolves. As such, there are two ways this can occur, via an extremely high evolutionary rate or convergent evolution.

Appendix

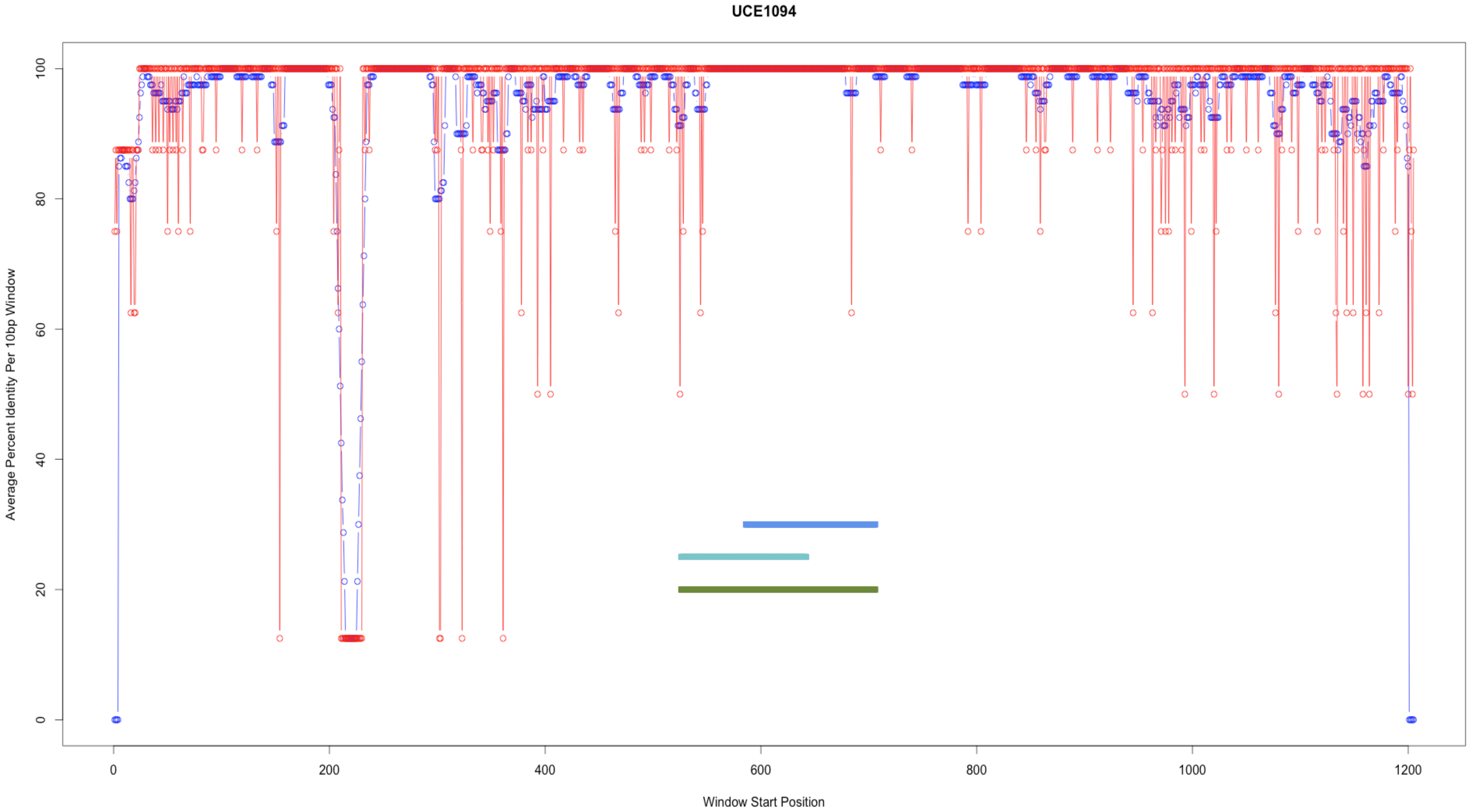
In this appendix, I illustrate the site by site sequence identity for nine randomly selected UCEs. These figures illustrate the change in percent identity as a function of position as well as illustrating how the core was identified. Interestingly the core is neither always centralized in the UCE nor continuous. Often there are gaps within the core as well as the flanking regions. I chose to include the nine following figures in order to show the diversity in core position and ways in which it can be broken up by gaps.



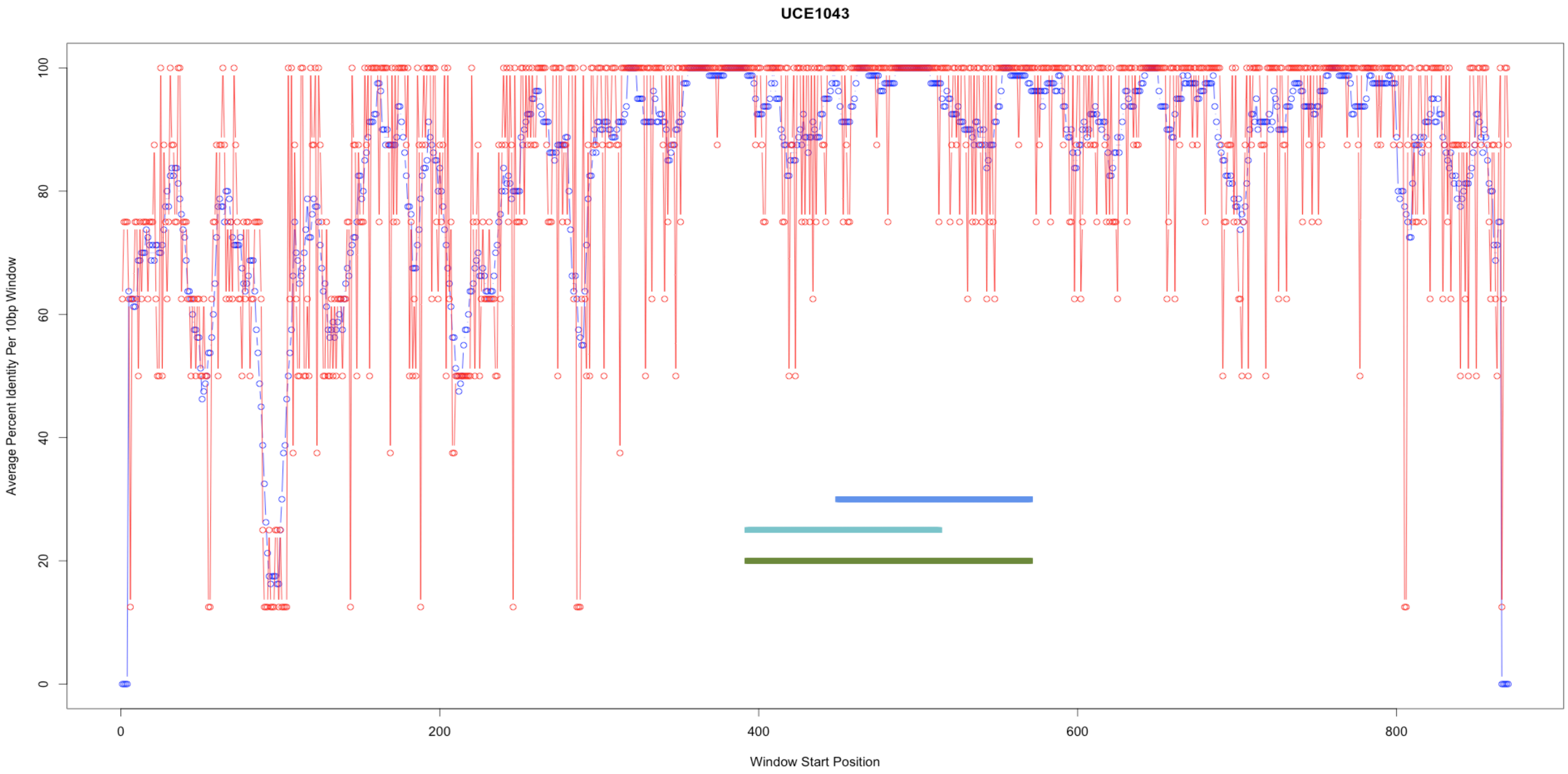
1-1 Probe 458 Ten base sliding window plot of percent sequence identity of one UCE is shown in blue. . Individual nucleotide percent sequence identity is shown in red. Percent sequence identity for each UCE alignment was used to determine a given UCE's core region. Dark blue bars indicate region of one hundred percent sequence identity with *Gasterosteus aculeatus* for probe 0 and probe 1. Light blue bars indicate one hundred percent sequence identity with *Oryzias latips* for probe 0 and probe 1. Green bar spans both dark and light blue regions and illustrates the actual UCE "Core".



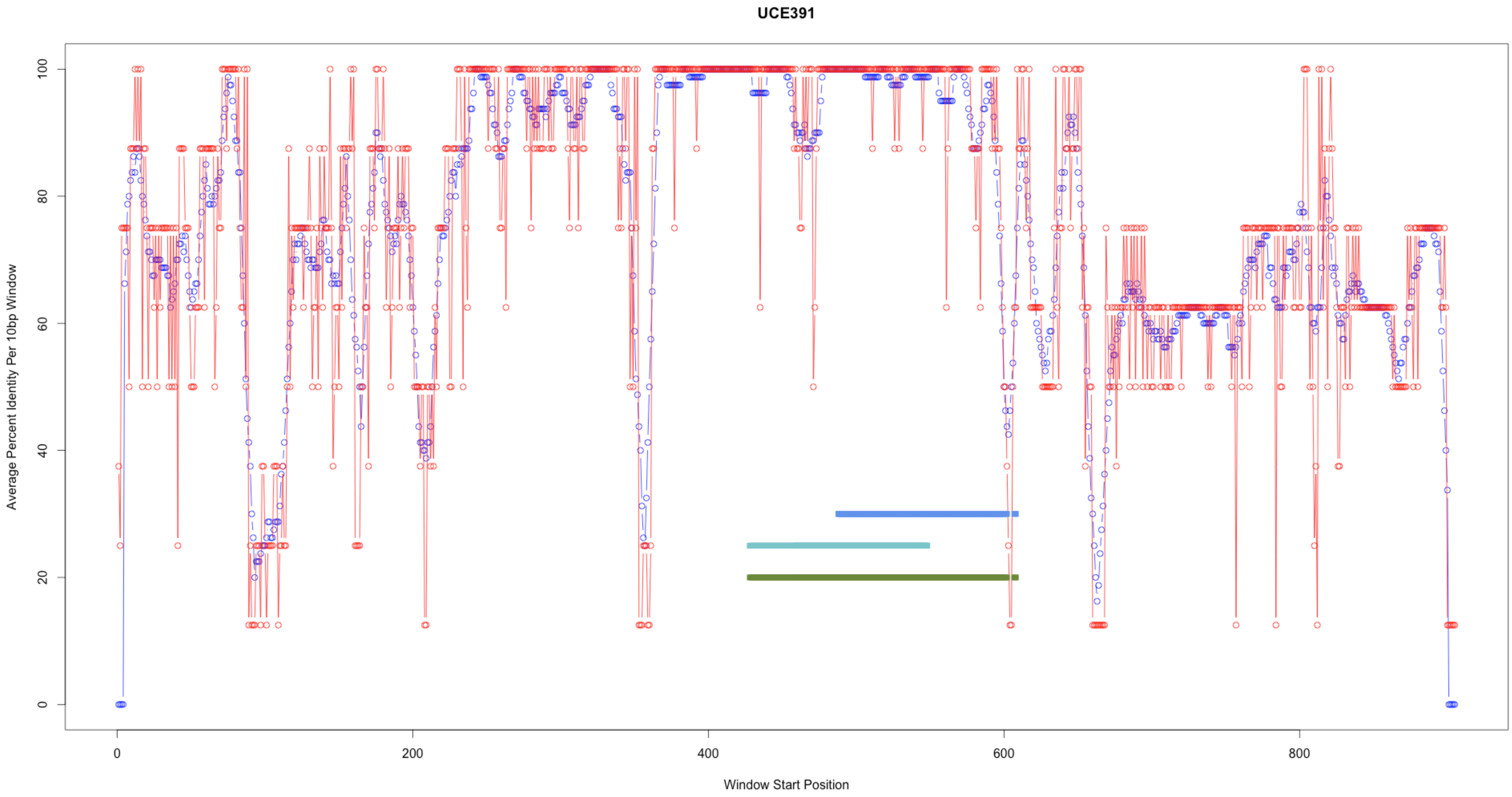
1-2 Probe 1154 Ten base sliding window plot of percent sequence identity of one UCE is shown in blue. . Individual nucleotide percent sequence identity is shown in red. Percent sequence identity for each UCE alignment was used to determine a given UCE's core region. Dark blue bars indicate region of one hundred percent sequence identity with *Gasterosteus aculeatus* for probe 0 and probe 1. Light blue bars indicate one hundred percent sequence identity with *Oryzias latips* for probe 0 and probe 1. Green bar spans both dark and light blue regions and illustrates the actual UCE "Core".



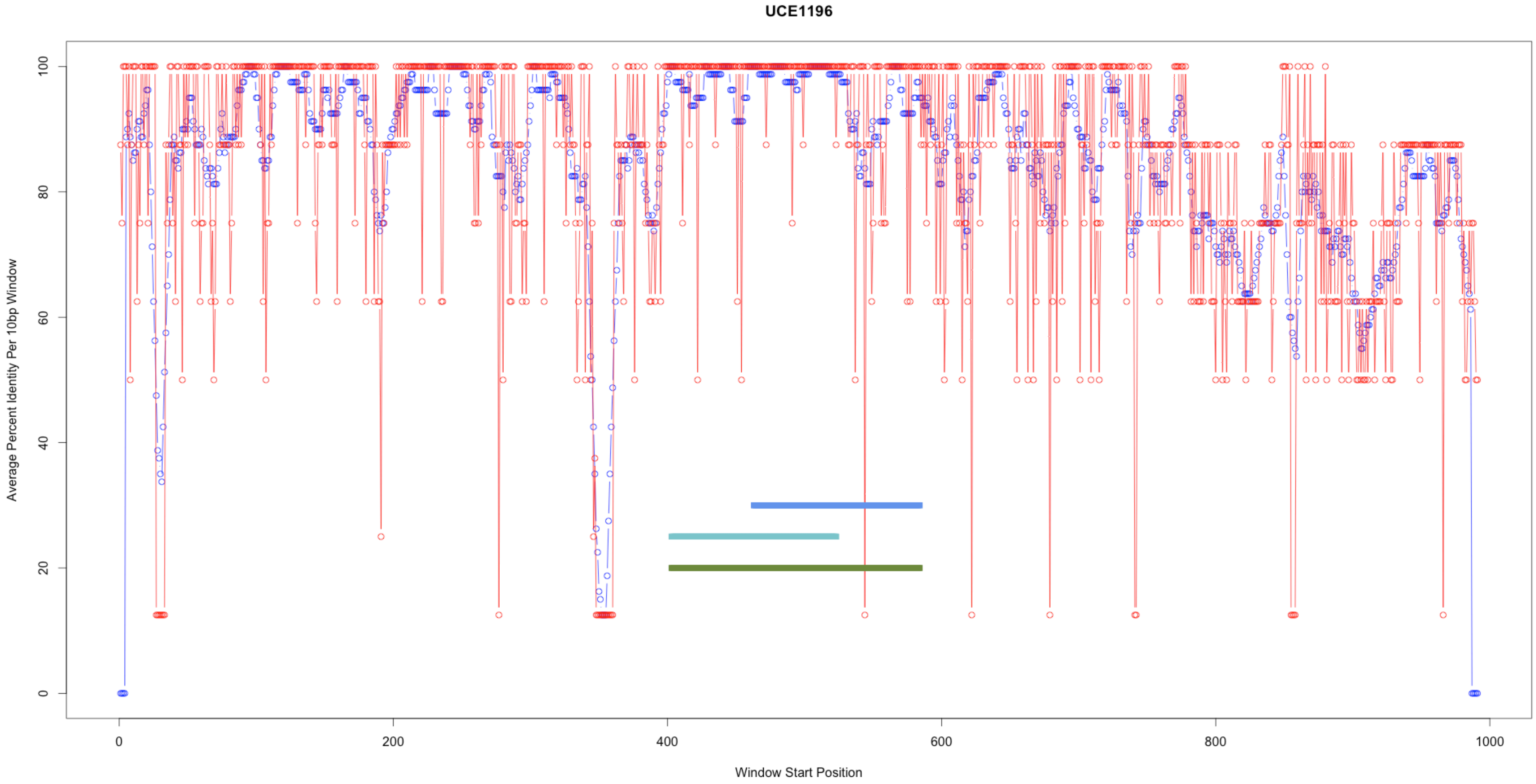
1-3 Probe 1094 Ten base sliding window plot of percent sequence identity of one UCE is shown in blue. . Individual nucleotide percent sequence identity is shown in red. Percent sequence identity for each UCE alignment was used to determine a given UCE's core region. Dark blue bars indicate region of one hundred percent sequence identity with *Gasterosteus aculeatus* for probe 0 and probe 1. Light blue bars indicate one hundred percent sequence identity with *Oryzias latips* for probe 0 and probe 1. Green bar spans both dark and light blue regions and illustrates the actual UCE "Core".



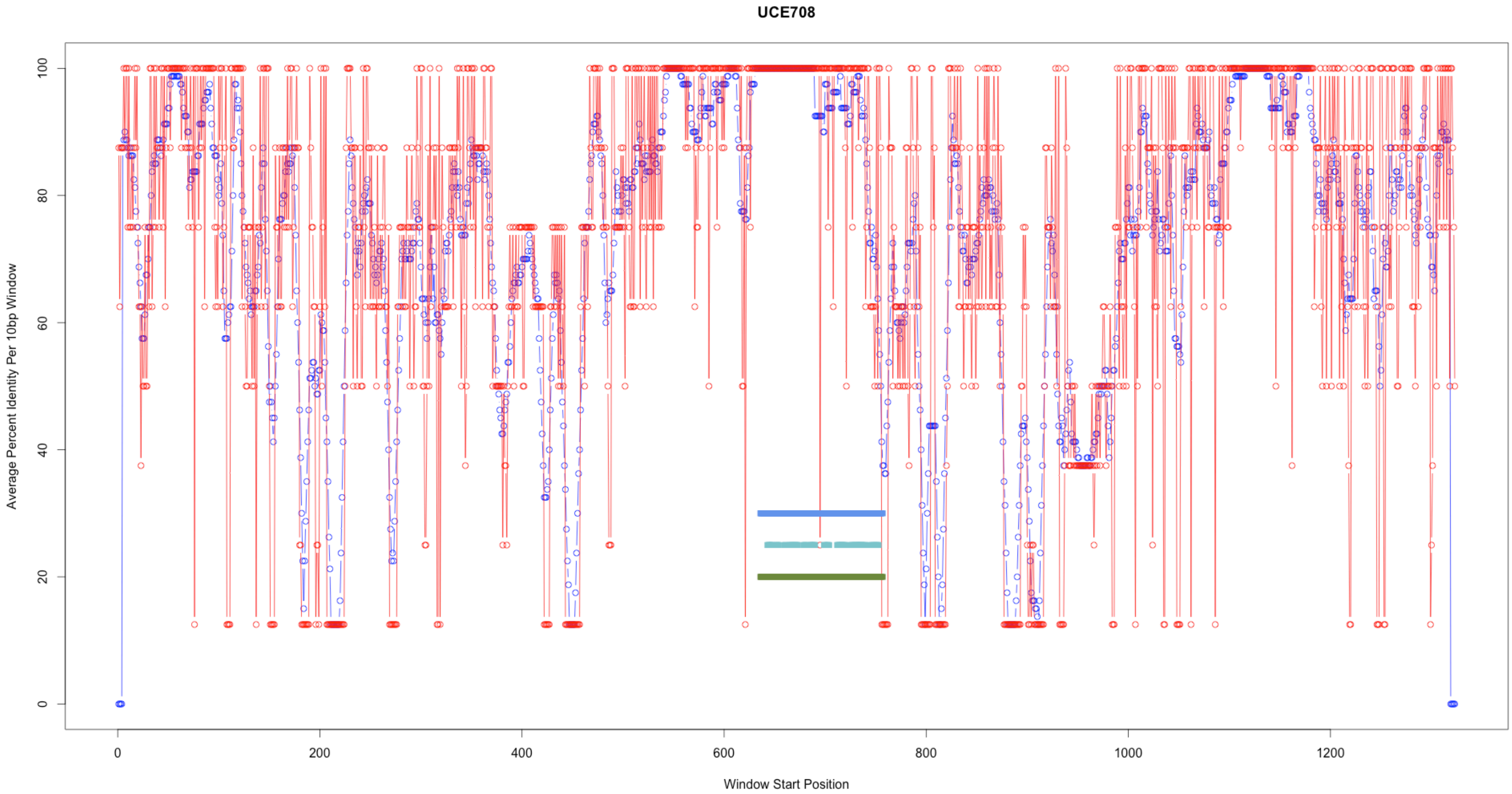
1-4 Probe 1043 Ten base sliding window plot of percent sequence identity of one UCE is shown in blue. . Individual nucleotide percent sequence identity is shown in red. Percent sequence identity for each UCE alignment was used to determine a given UCE's core region. Dark blue bars indicate region of one hundred percent sequence identity with *Gasterosteus aculeatus* for probe 0 and probe 1. Light blue bars indicate one hundred percent sequence identity with *Oryzias latips* for probe 0 and probe 1. Green bar spans both dark and light blue regions and illustrates the actual UCE "Core".



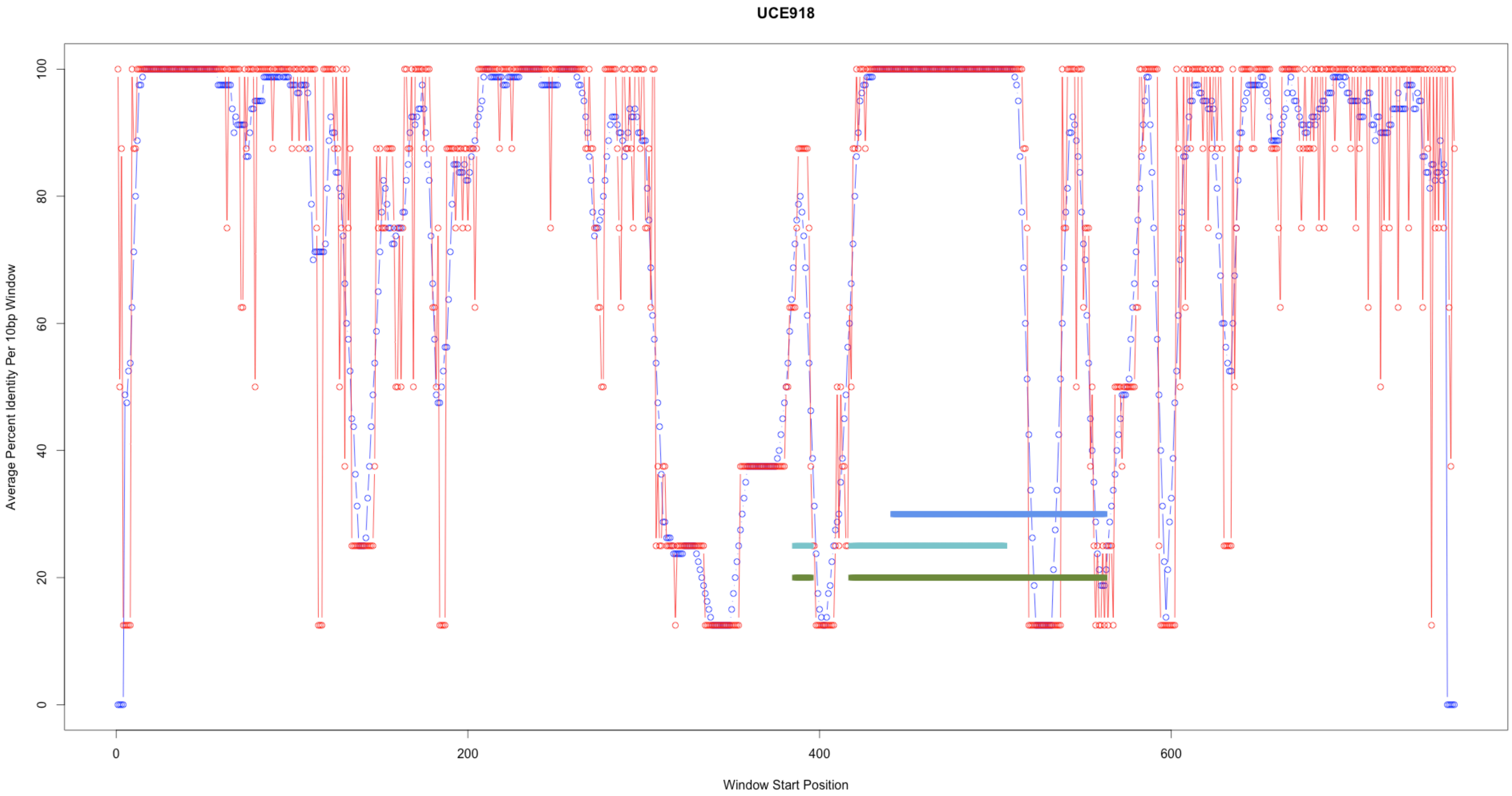
1-5 Probe 391 Ten base sliding window plot of percent sequence identity of one UCE is shown in blue. . Individual nucleotide percent sequence identity is shown in red. Percent sequence identity for each UCE alignment was used to determine a given UCE's core region. Dark blue bars indicate region of one hundred percent sequence identity with *Gasterosteus aculeatus* for probe 0 and probe 1. Light blue bars indicate one hundred percent sequence identity with *Oryzias latips* for probe 0 and probe 1. Green bar spans both dark and light blue regions and illustrates the actual UCE "Core".



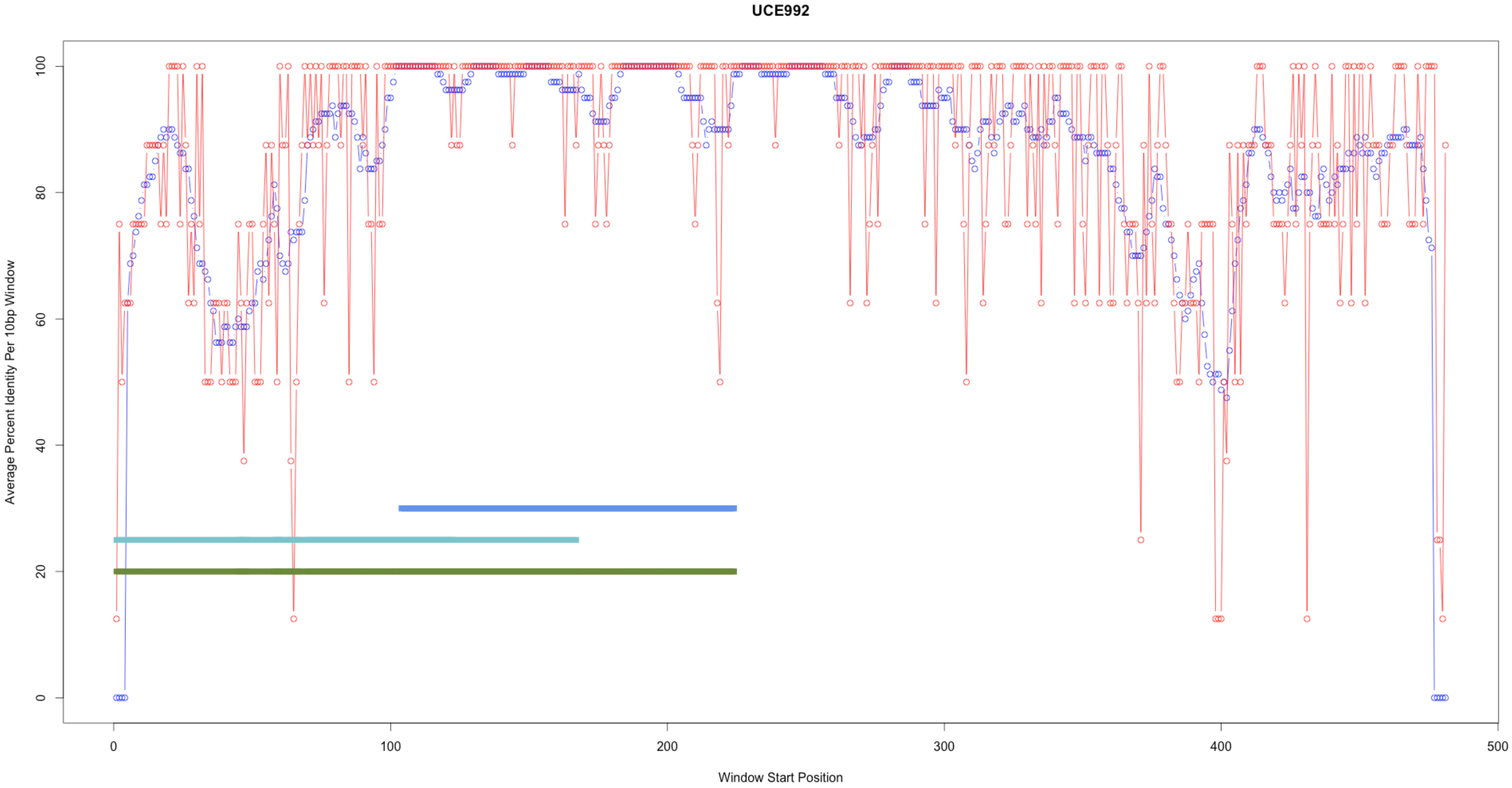
1-6 Probe 1196 Ten base sliding window plot of percent sequence identity of one UCE is shown in blue. . Individual nucleotide percent sequence identity is shown in red. Percent sequence identity for each UCE alignment was used to determine a given UCE's core region. Dark blue bars indicate region of one hundred percent sequence identity with *Gasterosteus aculeatus* for probe 0 and probe 1. Light blue bars indicate one hundred percent sequence identity with *Oryzias latips* for probe 0 and probe 1. Green bar spans both dark and light blue regions and illustrates the actual UCE "Core".



1-7 Probe 708 Ten base sliding window plot of percent sequence identity of one UCE is shown in blue. . Individual nucleotide percent sequence identity is shown in red. Percent sequence identity for each UCE alignment was used to determine a given UCE's core region. Dark blue bars indicate region of one hundred percent sequence identity with *Gasterosteus aculeatus* for probe 0 and probe 1. Light blue bars indicate one hundred percent sequence identity with *Oryzias latips* for probe 0 and probe 1. Green bar spans both dark and light blue regions and illustrates the actual UCE "Core".



1-8 Probe 918 Ten base sliding window plot of percent sequence identity of one UCE is shown in blue. . Individual nucleotide percent sequence identity is shown in red. Percent sequence identity for each UCE alignment was used to determine a given UCE's core region. Dark blue bars indicate region of one hundred percent sequence identity with *Gasterosteus aculeatus* for probe 0 and probe 1. Light blue bars indicate one hundred percent sequence identity with *Oryzias latips* for probe 0 and probe 1. Green bar spans both dark and light blue regions and illustrates the actual UCE "Core".



1-9. Probe 992 Ten base sliding window plot of percent sequence identity of one UCE is shown in blue. . Individual nucleotide percent sequence identity is shown in red. Percent sequence identity for each UCE alignment was used to determine a given UCE's core region. Dark blue bars indicate region of one hundred percent sequence identity with *Gasterosteus aculeatus* for probe 0 and probe 1. Light blue bars indicate one hundred percent sequence identity with *Oryzias latips* for probe 0 and probe 1. Green bar spans both dark and light blue regions and illustrates the actual UCE "Core.

References

Baira E., Greshock J., Coukos G., Zhang L., 2008. Ultraconserved elements: Genomics, function and disease. RNA Biol 5, 132-134

Boffelli D.A., Nobrega M.A., Rubin E.M., 2004. Comparative genomics at the vertebrate extremes. Nat. Rev. Genet. 6, 456-465.

Faircloth, B.C., Branstetter, M.G., White, N.D., Brady, S.G., 2014. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. Res. 15, 489–501. http://dx.doi.org/10.1111/1755-0998.12328.

Faircloth, B.C., Sorenson, L., Santini, F., Alfaro, M.E., 2013. A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). PLoS One 8, e65923.

Gilbert P.S., Chang J., Pan C., Sobel E.M., Sinsheimer J.S., Faircloth B.C., Alfaro M.E., 2015. Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. Mol. Phylogen. Evol. 92, 140–146.

Margulies E. H., Birney E., 2008. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. Nat. Rev. Genet. 9, 303-313.

McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn T.C., 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. Genome Res. 22, 746–754.
pmid: 22207614 doi: 10.1101/gr.125864.111.

Siepel, A., Bejerano G., Pedersen J.S., Hinrichs A. S., Hou M., Rosenbloom K., Clawson H., Spieth J., Hillier L. W., Richards S., Weinstock G. M., Wilson R.K., Gibbs RA, Kent J., Miller W., Haussler D., 2005. Evolutionary conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res., 15, 1034-1050.

Smith, B.T., Harvey, M.G., Faircloth, B.C., Glenn, T.C., Brumfield, R.T., 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol. 63, 83–95.

Stephen S., Pheasant M., Makunin I. V., Mattick J. S., 2008. Large-scale appearance of Ultraconserved elements in Tetrapod Genomes and Slowdown of the Molecular Clock. Mol. Biol. Evol., 25, 402-408.

Su Z., Townsend J. P., 2015. Utility of characters evolving at diverse rates of evolution to resolve quartet trees with unequal branch lengths: analytical predictions of long-branch effects. BMC Evol. Biol. 15, 1-13.

Su Z., Wang Z., Lopez-Giraldez F., Townsend J. P., 2014. The impact of incorporating molecular evolutionary model into predictions of phylogenetic signal and noise. Frontiers in Ecol. and Evol. 2, 1-12.

Townsend J. P., 2007. Profiling phylogenetic informativeness. Syst. Biol. 56, 222-231.

Townsend J. P., Su Z., Tekle Y.I., 2012. Phylogenetic signal and noise: Predicting the power of a data set to resolve phylogeny. Syst. Biol. 61, 835-849.

Chapter 2

Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes

Molecular Phylogenetics and Evolution 92 (2015) 140–146

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes ☆

Princess S. Gilbert [a,*], Jonathan Chang [a], Calvin Pan [b], Eric M. Sobel [d], Janet S. Sinsheimer [c,d,e], Brant C. Faircloth [f], Michael E. Alfaro [a,*]

[a] *Department of Ecology & Evolutionary Biology, University of California, Los Angeles, CA, USA*
[b] *Department of Medicine, University of California, Los Angeles, CA, USA*
[c] *Department of Biomathematics, University of California, Los Angeles, CA, USA*
[d] *Department of Human Genetics, University of California, Los Angeles, CA, USA*
[e] *Department of Biostatistics, University of California, Los Angeles, CA, USA*
[f] *Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA, USA*

ARTICLE INFO

Article history:
Received 14 February 2015
Revised 13 May 2015
Accepted 26 May 2015
Available online 12 June 2015

Keywords:
Ultraconserved elements
Next-generation sequencing
Non-coding DNA
Phylogenomics
Molecular evolution
Phylogenetic informativeness

ABSTRACT

Ultraconserved elements (UCEs) have become popular markers in phylogenomic studies because of their cost effectiveness and their potential to resolve problematic phylogenetic relationships. Although UCE datasets typically contain a much larger number of loci and sites than more traditional datasets of PCR-amplified, single-copy, protein coding genes, a fraction of UCE sites are expected to be part of a nearly invariant core, and the relative performance of UCE datasets versus protein coding gene datasets is poorly understood. Here we use phylogenetic informativeness (PI) to compare the resolving power of multi-locus and UCE datasets in a sample of percomorph fishes with sequenced genomes (genome-enabled). We compare three data sets: UCE core regions, flanking sequence adjacent to the UCE core and a set of ten protein coding genes commonly used in fish systematics. We found the net informativeness of UCE core and flank regions to be roughly ten-fold and 100-fold more informative than that of the protein coding genes. On a per locus basis UCEs and protein coding genes exhibited similar levels of phylogenetic informativeness. Our results suggest that UCEs offer enormous potential for resolving relationships across the percomorph tree of life.

1. Introduction

Ultraconserved elements (UCEs) have become increasingly popular in recent phylogenomic studies. They have been used to reconstruct phylogenies for clades as divergent as the mammals, fish, birds, turtles, and arthropods (Bejerano et al., 2004; Faircloth et al., 2014, 2013; McCormack et al., 2013; Smith et al., 2014; Sun et al., 2014). The utility of UCEs for sequence-capture approaches has been well justified on practical grounds. They are shared loci found among most, if not all vertebrate genomes (Bejerano et al., 2004; Siepel et al., 2005) and researchers can easily detect and align UCEs from divergent taxonomic groups (Miller et al., 2007). UCEs do not intersect paralogous genes (Derti et al., 2006) or have retroelement insertions (Simons et al., 2006). Stephen et al. (2008) found that most eutherian UCEs were intergenic with only 3% falling within protein coding exons and suggested splicing regulation as one of their functions. One of the most compelling phylogenetic characteristics of UCEs is that the flanking regions increase in variant sites as the distance from the UCE center increases, allowing for better resolution of nodes across a range of evolutionary timescales in a given phylogeny (Faircloth et al., 2012b). This aspect potentially allows phylogeneticists to tailor their use of UCEs by choosing those with similar evolutionary rates or selecting a subsample of UCE regions whose flanking regions optimize their analyses. However, the relative performance of UCEs compared to traditional molecular markers remains poorly understood.

Traditional markers might be expected to exhibit better phylogenetic performance than UCEs because traditional markers have been highly selected for their potential ability to resolve polytomies and they have been well curated and validated. Sets of traditional markers that yield reasonable phylogenetic results have been identified for many major sections of the tree of life. In fishes for example, Li et al. (2007) identified a cohort of 10 genes from a pool of 154 that have become widely used at various phylogenetic scales (Betancur-R et al., 2013; Li et al., 2009, 2008; Near et al.,

☆ This paper was edited by the Associate Editor Giacomo Bernardi.

* Corresponding authors at: Department of Ecology & Evolutionary Biology, 621 Charles E. Young Drive South, University of California, Los Angeles, CA 90095, USA.

E-mail addresses: ps.gilbert@ucla.edu (P.S. Gilbert), michaelalfaro@ucla.edu (M.E. Alfaro).

http://dx.doi.org/10.1016/j.ympev.2015.05.027
1055-7903/© 2015 Elsevier Inc. All rights reserved.

2012; Wainwright et al., 2012). These protein coding genes were carefully selected and validated for the purpose of reconstructing the ray-finned fish phylogeny (Li et al., 2007). In contrast, UCEs are identified by the presence of nearly invariant core regions. UCE cores are thus expected to have very low to no phylogenetic resolving power. The flanking regions of the UCE are, by definition, not invariant and should thus provide more resolving power than the core. However individual UCE loci have not generally been subjected to the same degree of scrutiny as the phylogenetic workhorse, PCR-amplified, single copy protein coding genes, and thus, on average, might be expected to perform more poorly at resolving phylogenetic problems. One resolution of this paradox would be that the greater degree of resolution obtained in recent UCE studies (Crawford et al., 2012; McCormack et al., 2012) is largely due to the sheer number of sites that are captured through high-throughput sequencing methods, as on a per locus basis the ability of UCEs to resolve polytomies is thought to be relatively poor.

UCE cores are highly conserved throughout the genome, which suggests there may be little phylogenetic informativeness in these regions. More specifically, we ask the question, what is the impact of UCE core conservation on overall phylogenetic informativeness and on the UCEs' ability to resolve hypothetical polytomies?

To better understand the utility of UCEs in a phylogenetic context, we characterize their phylogenetic informativeness (Townsend, 2007) by analyzing a dataset comprised of 1201 UCEs and 10 protein coding genes collected from eight species of percomorphs with fully sequenced genomes (genome-enabled), *Gasterosteus aculeatus*, *Oryzias latipes*, *Takifugu rubripes*, *Tetraodon nigroviridis*, *Oreochromis niloticus*, *Neolamprologus brichardi*, *Pundamila nyererei* and *Haplochromis burtoni.* We chose to examine the percomorphs because recent studies have demonstrated that this large clade has undergone recent radiations and many relationships remain unresolved, which heavily impact age estimations in the clade (Betancur-R et al., 2013; Broughton et al., 2013; Smith et al., 2007; Wainwright et al., 2012). Li et al. (2007) demonstrated that a carefully chosen set of 10 protein coding genes can successfully resolve many groups within the percomorphs. Faircloth et al. (2012b) demonstrated that UCEs successfully resolve older lineage relationships in the euteleost tree of life but they did not specifically focus on resolving polytomies within sub-clades of the percomorphs, for example the order Perciformes, and it is yet untested whether more recent radiations within the Euteleosts can be resolved using UCEs.

We chose phylogenetic informativeness (PI) to make our comparison. PI estimates the probability that a character resolves a hypothetical polytomy in a four-taxon phylogeny and then remains unchanged along the peripheral branches (Townsend, 2007). PI is a function of the rate of evolutionary change and the time to most recent common ancestor among the taxa under analysis, and it provides one estimate of the amount of phylogenetic signal relative to noise across a specified time period. Marker sets for more than four taxa can be compared using PI if a consistent topology is used across the markers. Calculation of the PI per nucleotide allows estimation of the cost-effectiveness of character sampling. Thus our study seeks to address which dataset, the UCEs or the protein coding genes, has the greatest PI so that researchers interested in clades within the percomorphs can focus on the appropriate data to best resolve the remaining polytomies.

2. Materials and methods

2.1. UCE core region design pipeline

We identified 1201 UCEs found in the eight percomorphs whose genomes were available at the start of our study, one three-spined stickleback, *G. aculeatus*, one medaka, *O. latipes*, two puffers, *T. rubripes*, *T. nigroviridis*, and four cichlids, *O. niloticus*, *N. brichardi*, *P. nyererei* and *H. burtoni.* Following Faircloth et al. (2013), we: (1) located nuclear DNA regions of 180 ± 10 base pairs (bp) where there were at least 80 contiguous bp with 100% conservation and the remainder with >80% conservation between *G. aculeatus* and *O. latipes*; (2) aligned these sequences to the genomes of the remaining six fishes (*T. rubripes*, *T. nigroviridis*, *O. niloticus*, *N. brichardi*, *P. nyererei* and *H. burtoni)* using LASTZ (Harris, 2007); and (3) required >80% sequence identity across all eight species. We defined the core as the contiguous region of the aligned sequence, which corresponds to the original 180 bp from *G. aculeatus* and *O. latipes*, and flank as all the remaining sequence 5′ or 3′ of the core. To ensure that PI is accurately calculated, we limited our analysis to UCEs with at least 50 bp flanking the 5′ or 3′ end of the core. This reduced the final count used for all further analysis to 988 UCE loci with cores of aligned lengths of 171 bp to 219 bp and flanks of aligned lengths of 144 bp to 1626 bp.

2.2. Protein coding genes

We compared the UCEs recovered in this study to ten protein coding genes identified by Li et al. (2007) (see Supplemental Table S1). We downloaded individual gene data for each of these loci across the eight genome-enabled percomorph species from the ENSEMBL Genome Browser (Hubbard et al., 2007), the UCSC genome browser (Kent et al., 2002), and NCBI GenBank (Benson et al., 2005). We translated the nucleotide sequences of the ten loci into amino acid sequences using TranslatorX (Abascal et al., 2010) and aligned amino acids using MUSCLE (Edgar, 2004). We used the DNA version of these alignments when calculating PI.

2.3. In silico phylogeny design for the PI guide tree

We constructed a time-calibrated phylogenetic framework needed for calculation of PI using divergence times from recently published phylogenetic studies to date node splits for the eight taxon tree of genome-enabled percomorph fishes (Betancur-R et al., 2013; Broughton et al., 2013; Santini et al., 2009; Wainwright et al., 2012). We provide the time-calibrated phylogeny for the eight genome-enabled species used in this study (Supplemental Fig. 1).

2.4. PI Calculations

We used the software package TAPIR (http://faircloth-lab.github.com/tapir/) to measure the PI of the UCE core regions, the flanking regions of the UCE cores and the set of ten protein coding genes. TAPIR employs a similar pipeline for estimating PI to that used in PhyDesign (Lopez-Giraldez and Townsend, 2011) although the PI computation is parallelized to work across large genomic datasets (Faircloth et al., 2012a; Pond et al., 2005). TAPIR calculates substitution rates from sequence alignment files and then uses those substitution rates to estimate the PI profile of each locus. We calculated net PI for each dataset, PI per locus per dataset, and PI per nucleotide per locus per dataset. The net PI is the sum of the individual PI's for each nucleotide across all loci in a dataset. Thus, net PI is additive and the length of each dataset contributes to its respective net PI curve. When displaying or analyzing the time of maximum PI, we removed seven UCEs whose cores were invariant across all taxa and thus had PI = 0 across the entire time-calibrated phylogeny.

2.5. Statistical analysis

We conducted statistical analyses using the R package (http://www.r-project.org/) and TAPIR (http://faircloth-lab.github.com/tapir/). We calculated the distribution of the average per nucleotide PI, the maximum nucleotide PI, and the time in millions of years (Ma) of maximum PI using plyr, gtools, and xtable libraries in R and ggplot2 (Harrell and Dupont, 2014; Team, 2014; Warnes et al., 2014; Wickham, 2009, 2011). We performed regression analyses using the lm function of R.

2.6. Verification of the percomorph phylogeny

To verify that both the UCE dataset and the protein coding gene dataset produced the expected phylogeny (Dornburg et al., 2014; Faircloth et al., 2013; Near et al., 2013; Wainwright et al., 2012)) we reconstructed the phylogeny for the eight genome-enabled species (Supplemental Table S2). We prepared our data for phylogenetic reconstruction using phyluce (https://github.com/faircloth-lab/phyluce). To estimate the best fitting locus-specific site rate substitution models we used Cloudforest (Crawford and Faircloth, 2014) and partitioned the UCEs by their best-fitting substitution models. Bayesian methods were used for phylogenetic inference as implemented in MrBayes 3.1(Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003; Ronquist et al., 2012) thus over 5,000,000 iterations we sampled trees every 500 iterations to yield 10,000 trees. Convergence was confirmed by checking Effective Sampling Size values >200 in TRACER (Rambaut et al., 2014).

3. Results

3.1. Net phylogenetic informativeness of each dataset

The UCE flanking regions outperformed the UCE core regions, which outperformed the protein coding genes, for estimates of net PI across all times scales (presented as the $\log_{10}$ of PI versus time in Ma in Fig. 1). PI for the UCE flanks rose rapidly, reached a maximum at 43 Ma and then slowly tapered off. We observed similar behavior for the PI of the UCE cores and the PI of the protein coding genes (Fig. 1).

3.2. PI per locus in each data set

The average and 95% confidence interval (CI) for the per locus PI of the UCE flanking regions, the UCE core regions, and the ten protein coding genes are shown versus time in Ma (Fig. 2a). UCE flanking regions had the highest PI per locus, surpassing both the UCE core regions and protein coding genes. The UCE core had the lowest per locus PI, reflecting that region's relative invariance.

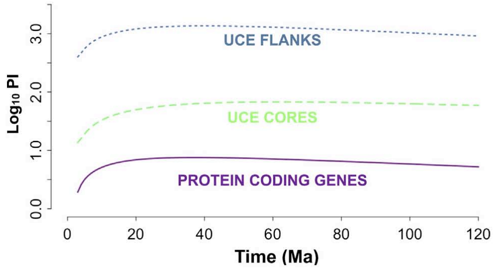


Fig. 1. The $\log_{10}$ of net phylogenetic informativeness plotted against time for each data type. The blue short dashed line shows UCE flanking regions, the green long dashed line shows the UCE core, and the purple line shows the protein coding genes chosen from Li et al. (2007).

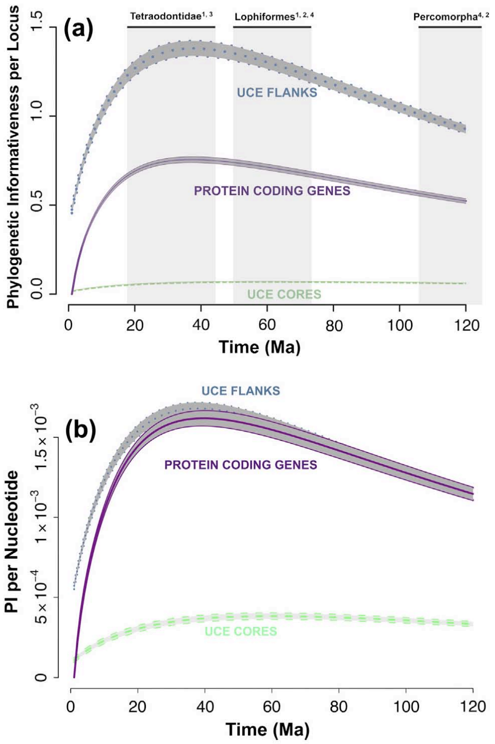


Fig. 2. (a and b) The 95% confidence interval for phylogenetic informativeness (PI) per locus (a) and per nucleotide (b) across time. Flanking regions (dotted, blue), UCE core regions (dashed, green) and protein coding genes (solid, purple) overlay a shaded gray region illustrating the average ± 2 std. errors. The central line is the average PI across all UCEs or loci for each time point. The estimate for the age of the most recent common ancestor (MRCA) of Tetraodontidae, Lophiformes and Percomorpha is plotted on the *x*-axis of (a) with grey shading. Chen et al., 2014[1]; Near et al., 2013[2]; Santini et al., 2013[3]; Betancur-R et al., 2013[4]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The ability of UCEs to resolve polytomies depends on the time of divergence from the most recent common ancestor (MRCA) of the polytomy, thus we calculated the time in Ma at which PI is maximized. Based on the average and 95% CI, we observed that the UCE flanking region PI reached its maximum at 39 ± 20 Ma (Fig. 2a), which was similar to that of the protein coding genes, suggesting UCE loci should be suitable for resolving the same polytomies as protein coding genes. Similarly, the maximum PI for UCE cores occurred at 61 ± 20 Ma (Fig. 2a), suggesting these data are suitable for resolving polytomies occurring deeper in time.

To illustrate how these maxima correspond to the age of the MRCA of the percomorphs and two key clades within the percomorphs, we included in Fig. 2a the estimates of the ages of these clades. We use the results of four previously time calibrated phylogenetic reconstructions. The estimates for the MRCA of the Tetraodontidae span from 18 Ma to 44 Ma (Chen et al., 2014; Santini et al., 2013). The maximum PI for the UCE flanking region and for the protein coding genes fall within this range therefore PIs are still driven far more by signal than noise (Townsend,

2007) (Fig. 2a). The estimates for the age of the MRCA of Lophiformes span from 50 Ma to 73 Ma (Betancur-R et al., 2013; Chen et al., 2014). At ~60 Ma, the UCE flanking region PI and the protein coding gene PI have decayed to less than 10% from their maxima indicating again that these loci are still within optimal signal for this clade. The estimates for the age of the MRCA for the percomorphs span from 106 Ma to 133 Ma (Betancur-R et al., 2013; Near et al., 2013; Chen et al., 2014). At ~120 Ma, UCE flanking region PI and the protein coding gene PI have decayed less than 33% from their maxima. These comparisons illustrate that UCE flanking regions are appropriate for resolving polytomies within Tetraodontidae and Lophiformes as well as within Percomorpha.

3.3. PI per nucleotide in each data set

The average and 95% CI for the per nucleotide PI of the UCE flanking regions, the UCE core regions and the protein coding genes are shown versus time in Ma in Fig. 2b. The UCE flanking regions had PI values that are slightly higher but similar to the protein coding genes. The UCE core regions had the lowest PI at each time point which is likely a consequence of how UCEs are chosen and the different evolutionary pressures on the UCE cores relative to the UCE flanks or the protein coding genes.

3.4. Average PI, Max PI, and time at maximum PI for the UCE core, flank and protein coding datasets

The results shown thus far provide the average UCE behavior for each point in time. When comparing the individual UCEs versus the average behavior across the set, we found that the per nucleotide PI maxima and averages were higher for the flanking regions (mean of max PI = 1.700×10^{-3}, std. dev. of max PI = 5.955×10^{-4} and mean of average PI = 1.437×10^{-3}, std. dev. of average PI = 4.636×10^{-4}) than for its corresponding core regions (mean of max PI = 4.097×10^{-4}, std. dev. of max PI = 3.103×10^{-4} and mean of average PI = 2.899×10^{-4}, std. dev. of average PI = 2.443×10^{-4}) and were better approximated by normal distributions (Fig. 3a–d and Supplemental Table S1).

For UCE core regions, the median time of maximum PI was 71 Ma (interquartile range for core = 53 Ma, 94 Ma) but the distribution was quite wide with a number of UCE cores reaching maximum PI at 120 Ma, the oldest time point included in our analysis (Fig. 3e). For the UCE flanking regions, the median time of maximum PI was 41 Ma with an interquartile range for the flank of (36 Ma, 47 Ma, Fig. 3f). For the protein coding genes, the median time of maximum PI was 32 Ma (Table 1) with an interquartile range of (28.75 Ma, 44.25 Ma).

3.5. Determinants of PI – linear regression analyses

As expected, there was a strong correlation between average per nucleotide PI and the maximum per nucleotide PI for each locus in the UCE core ($R^2 = 0.91$) and UCE flanking regions ($R^2 = 0.99$) (Supplemental Fig. S3a and S3b). We thus only present results for the average per nucleotide PI. We found a significant but weak correlation between average PI per nucleotide for UCE flanking regions and the average PI per nucleotide for the UCE core regions, $R^2 = 0.14$ (Fig. 4), indicating that if the UCE had an increased average PI for its core region, they also had an increased PI for its flanking region.

We plotted the average PI per upstream and downstream UCE flanking region against that region's length (Fig. 5). We observed an increasing trend in average PI per region as the flanking region's length increased, as would be expected as variation has been shown to increase with distance from the core (Faircloth et al., 2012b). Further, if we controlled for the average per nucleotide PI of the core, we found that total flank length was a significant predictor of average per nucleotide PI of the flank ($p < 2.2 \times 10^{-16}$, Table 2).

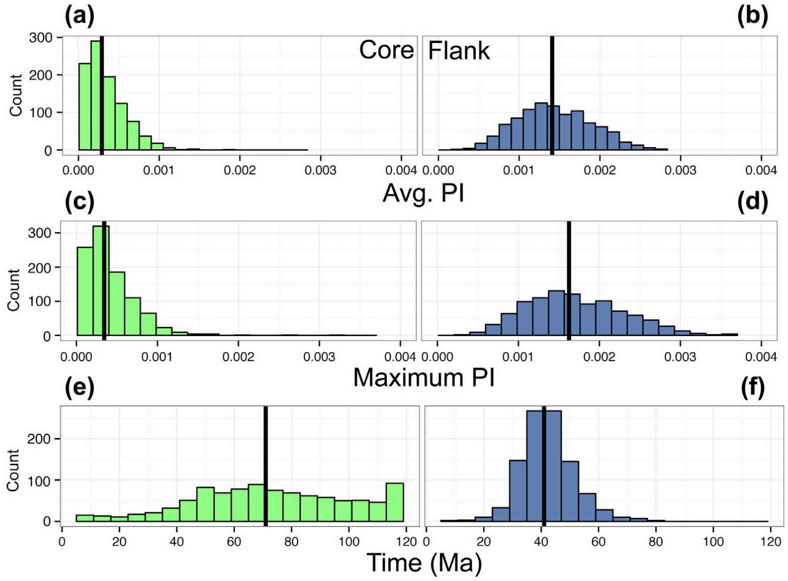


Fig. 3. (a–f) UCE core and flank dataset phylogenetic informativeness (PI) distributions. The left column of histograms shows the observed distributions of the core UCE regions. The right column of histograms shows the observed distributions of the flank UCE regions. Average PI per nucleotide for each dataset (a and b); Maximum PI per nucleotide for each dataset (c and d); The time point when PI reaches its maximum for each dataset (e and f). The black line marks the median of each histogram.

Table 1
Summary statistics for average per nucleotide PI, maximum per nucleotide and time at maximum PI for the core, flank and protein coding genes.

	UCE core $_{\text{avg. per nucleotide PI}}$	UCE flank $_{\text{avg. per nucleotide PI}}$	Protein coding genes $_{\text{avg. per nucleotide PI}}$
Median	2.889×10^{-4}	1.409×10^{-3}	1.08×10^{-3}
Average	3.406×10^{-4}	1.437×10^{-3}	1.34×10^{-3}
Std. deviation	2.443×10^{-4}	4.636×10^{-4}	5.8×10^{-4}
	UCE core $_{\text{max. per nucleotide PI}}$	UCE flank $_{\text{max. per nucleotide PI}}$	Protein coding genes $_{\text{max. per nucleotide PI}}$
Median	3.430×10^{-4}	1.625×10^{-3}	1.37×10^{-3}
Average	4.097×10^{-4}	1.700×10^{-3}	1.61×10^{-3}
Std. deviation	3.103×10^{-4}	5.955×10^{-4}	6.8×10^{-4}
	UCE core $_{\text{time at maximum PI}}$	UCE flank $_{\text{time at maximum PI}}$	Protein coding genes $_{\text{time at maximum PI}}$
Median	71 Ma	41 Ma	32 Ma
Average	72.71 Ma	41.84 Ma	34.9 Ma
Std. deviation	27.39 Ma	9.37 Ma	10.1 Ma

Note: See Section 2.5 and Fig. 3 for details.

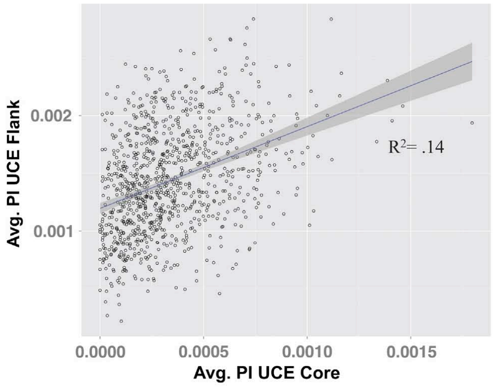


Fig. 4. Average PI per nucleotide for the UCE flanking regions versus average PI per nucleotide for the UCE core regions. Linear regression results: adjusted $R^2 = 0.14$; p-value = $<2.2 \times 10^{-16}$; slope = 0.7081; and Y-intercept = 1.196×10^{-3}.

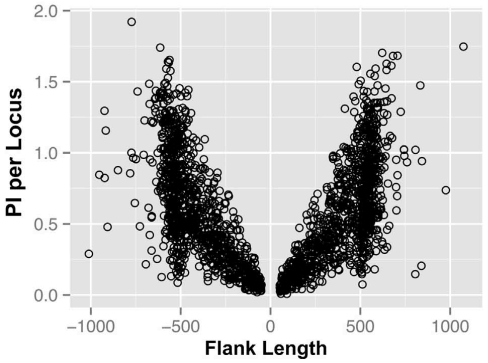


Fig. 5. Average PI for each UCE plotted against upstream and downstream flank length.

3.6. Verification of the Phylogeny

We recovered the relationships supported in the current literature (Faircloth et al., 2013; Li et al., 2007) with high posterior probabilities using either the protein coding genes or the 988 UCEs (Supplemental Fig. S2).

4. Discussion

Molecular marker choice is arguably the most important decision made before one embarks on a phylogenetic analysis. Here we explore 3 datasets: UCE core regions, UCE flanking regions and protein coding gene regions, in order to understand PI patterns. UCE flanking and core regions have higher net PI than protein coding genes (Fig. 1). This outcome was expected as there were far more UCEs than protein coding genes analyzed. Our analysis corroborates Faircloth et al. (2012b) by finding that the major source of PI for more recent splits is derived from the UCE flanking region and not its core (Faircloth et al., 2012b). Furthermore as the flanking region length increased the average per locus PI for that region increased (Fig. 5). We believe this can be attributed to the fact that longer flanking regions had greater sequence diversity and thus higher PI than shorter regions.

A second important result is that on a per nucleotide scale, the UCE flanking regions have similar PI to protein coding genes (Fig. 2b). *A priori*, we suspected that the protein coding genes would have greater PI than the UCE flanking regions on a per-nucleotide and per-locus level because the protein coding genes we used were carefully selected and validated to be useful in reconstructing the ray-finned fish phylogeny (Li et al., 2007). UCE flanking regions show more variation than the UCE cores and yet are still readily aligned among a set of taxa such as the percomorphs chosen for our analysis. Although we suspect that our results extend beyond these eight taxa, it would be interesting to determine if they hold for a larger set of fishes, birds or mammals.

Despite the low PI of UCE core regions on a per-locus or per nucleotide basis (Fig. 2a and b), the net PI of the UCE cores exceeds that of the protein coding genes (Fig. 1). Although UCEs are highly conserved, they still yield varying levels of PI. The explanation for UCE cores exceeding protein coding genes in net PI is sheer loci number. The median time when UCE cores reach its maximum PI is greater than the median time when the UCE flanks reach its maximum PI (Fig. 3 and Table 1), suggesting that UCE cores may be more useful for resolving phylogenetic relationships than previously thought, relationships that are more ancient than the radiation of the percomorphs. Therefore UCE core regions can and should be retained in a phylogenetic reconstruction along with the UCE flanking regions.

Our choice of phylogenetic informativeness as a measure of the suitability of a marker stems from a growing body of publications that demonstrate the comparative quality of PI (Lopez-Giraldez et al., 2013; Schoch et al., 2009; Townsend, 2007; Townsend and Leuenberger, 2011; Townsend et al., 2008). We believe PI holds the key to framing quantitative comparisons of marker types and gives researchers the ability to choose markers based on real data and not just hypothetical assumptions. However PI has garnered

Table 2
Multiple linear regression analysis of PI per nucleotide for the flanking region.

Coefficients	Estimate	Std. Error	*t*-value	Pr (>|*t*|)
Y-intercept	1.042×10^{3}	5.178×10^{5}	20.114	$<2 \times 10^{16}$
Average PI per nucleotide in the core region	7.322×10^{1}	5.623×10^{2}	13.022	$<2 \times 10^{16}$
Total flank length	1.795×10^{7}	5.361×10^{8}	3.349	8.41×10^{4}

Note: Residual standard error of 4.28×10^{4} on 985 degrees of freedom. Adjusted R^2 of 0.149, *F*-statistic of 86.23 on 2 and 985 degrees of freedom. *P*-value $<2.2 \times 10^{16}$.

criticism in regards to possible biases placed on fast evolving characters in a given sequence or gene and reduced applicability to real datasets with greater than four taxa (Klopfstein et al., 2010). Per Townsend and Leuenberger (2011), we limited our interpretation of PI profiles to details of the phylogeny on which we based our analyses. Detection of the phylogenetic signal, the subsequent loss of that signal and replacement with non-informative character states all depend upon the specific time epoch one is interested in studying.

In summary, our study provides preliminary evidence that the net phylogenetic informativeness of ultraconserved elements, at both flank and core regions, is superior to the phylogenetic informativeness of the set of protein coding genes recommended for resolving polytomies in the percomorphs. The improvement over the protein coding genes in net phylogenetic informativeness is made possible due to the large number of UCEs that can be detected and aligned among these taxa. It is also a novel finding of this study that UCE flanking regions and protein coding genes have similar levels of per nucleotide phylogenetic informativeness. Although a more comprehensive test with more taxa is required to insure that these results are not limited to the specific clades tested here, our results suggest that UCEs are likely to be an effective means for resolving relationships within percomorphs across a range of time scales.

Acknowledgments

This work was supported by the National Institute of Health-Genomic Analysis Training Program (T32HG002536 to P.S.G.); the U.S. Department of Education Graduate Assistance in Areas of National Need (to P.S.G); the Whitcome Research Fellowship (to J.C.); and the National Science Foundation (DMS 1264153 to J.S.S., DEB 6701648 and DEB 6861953 to M.E.A.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2015.05.027.

References

Abascal, F., Zardoya, R., Telford, M.J., 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucl. Acids Res. 38, W7–W13.
Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., Haussler, D., 2004. Ultraconserved elements in the human genome. Science 304, 1321–1325.
Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2005. GenBank. Nucl. Acids Res. 33, D34–D38.
Betancur-R, R., Broughton, R.E., Wiley, E.O., Carpenter, K., Lopez, J.A., Li, C., Holcroft, N.I., Arcila, D., Sanciangco, M., Cureton Ii, J.C., Zhang, F., Buser, T., Campbell, M.A., Ballesteros, J.A., Roa-Varon, A., Willis, S., Borden, W.C., Rowley, T., Reneau, P.C., Hough, D.J., Lu, G., Grande, T., Arratia, G., Orti, G., 2013. The tree of life and a new classification of bony fishes. PLoS Curr. 5. http://dx.doi.org/10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288.
Broughton, R.E., Betancur-R, R., Li, C., Arratia, G., Ortí, G., 2013. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. PLoS Curr. 5. http://dx.doi.org/10.1371/currents.tol.2ca8041495ffafd0c92756e75247483e.
Chen, W.-J., Santini, F., Carnevale, G., Chen, J.-N., Liu, S.-H., Lavoué, S., Mayden, R.L., 2014. New insights on early evolution of spiny-rayed fishes (Teleostei: Acanthomorpha). Front. Mar. Sci. 1. http://dx.doi.org/10.3389/fmars.2014.00053.
Crawford, N.G., Faircloth, B.C., 2014. Cloudforest: Code to Calculate Species Trees from Large Genomic Datasets. doi: http://dx.doi.org/10.5281/zenodo.12259.
Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K., Glenn, T.C., 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of Archosaurs. Biol. Lett. 8, 783–786.
Derti, A., Roth, F.P., Church, G.M., Wu, C.T., 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat. Genet. 38, 1216–1220.
Dornburg, A., Townsend, J.P., Friedman, M., Near, T.J., 2014. Phylogenetic informativeness reconciles ray-finned fish molecular divergence times. BMC Evol. Bio. 14, 169. http://dx.doi.org/10.1186/s12862-014-0169-0.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32, 1792–1797.
Faircloth, B.C., Branstetter, M.G., White, N.D., Brady, S.G., 2014. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. Res. 15, 489–501. http://dx.doi.org/10.1111/1755-0998.12328.
Faircloth, B.C., Chang, J., Alfaro, M.E., 2012a. TAPIR Enables High-throughput Estimation and Comparison of Phylogenetic Informativeness using Locus-specific Substitution Models. arXiv preprint arXiv:12021215 2012, p. 1215.
Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012b. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726.
Faircloth, B.C., Sorenson, L., Santini, F., Alfaro, M.E., 2013. A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). PLoS One 8, e65923.
Harrell Jr., F.E., Dupont, M.C., 2014. R Package Hmisc. R Foundation for Statistical Computing, Vienna, Austria.
Harris, R.S., 2007. Improved Pairwise Alignment of Genomic DNA. Computer Science and Engineering. The Pennsylvania State University, PA, USA.
Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S.C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., Birney, E., 2007. Ensembl 2007. Nucl. Acids Res. 35, D610–617.
Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17, 754–755.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, David., 2002. The human genome browser at UCSC. Genome Res. 12, 996–1006.
Klopfstein, S., Kropf, C., Quicke, D.L., 2010. An evaluation of phylogenetic informativeness profiles and the molecular phylogeny of diplazontinae (Hymenoptera, Ichneumonidae). Syst. Biol. 59, 226–241.
Li, B., Dettai, A., Cruaud, C., Couloux, A., Desoutter-Meniger, M., Lecointre, G., 2009. RNF213, a new nuclear marker for acanthomorph phylogeny. Mol. Phylog. Evol. 50, 345–363.
Li, C., Lu, G., Orti, G., 2008. Optimal data partitioning and a test case for ray-finned fishes (Actinopterygii) based on ten nuclear loci. Syst. Biol. 57, 519–539.
Li, C., Orti, G., Zhang, G., Lu, G., 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol. Biol. 7, 44.
Lopez-Giraldez, F., Moeller, A.H., Townsend, J.P., 2013. Evaluating phylogenetic informativeness as a predictor of phylogenetic signal for metazoan, fungal, and mammalian phylogenomic data sets. Biomed. Res. Int. 2013, 621604. http://dx.doi.org/10.1155/2013/621604.
Lopez-Giraldez, F., Townsend, J.P., 2011. PhyDesign: an online application for profiling phylogenetic informativeness. BMC Evol. Biol. 11, 152.
McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22, 746–754.
McCormack, J.E., Harvey, M.G., Faircloth, B.C., Crawford, N.G., Glenn, T.C., Brumfield, R.T., 2013. A phylogeny of birds based on over 1500 loci collected by target enrichment and high-throughput sequencing. PLoS One 8, e54848.

Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D., Kosakovsky Pond, S.L., Nekrutenko, A., Giardine, B., Harris, R.S., Tyekucheva, S., Diekhans, M., Pringle, T.H., Murphy, W.J., Lesk, A., Weinstock, G.M., Lindblad-Toh, K., Gibbs, R.A., Lander, E.S., Siepel, A., Haussler, D., Kent, W.J., 2007. 28-Way vertebrate alignment and conservation track in the UCSC genome browser. Genome Res. 17, 1797–1808.

Near, T.J., Dornburg, A., Eytan, R.I., Keck, B.P., Smith, W.L., Kuhn, K.L., Moore, J.A., Price, S.A., Burbrink, F.T., Friedman, M., Wainwright, P.C., 2013. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proc. Natl. Acad. Sci. USA 110, 12738–12743.

Near, T.J., Dornburg, A., Kuhn, K.L., Eastman, J.T., Pennington, J.N., Patarnello, T., Zane, L., Fernández, D.A., Jones, C.D., 2012. Ancient climate change, antifreeze, and the evolutionary diversification of Antarctic fishes. Proc. Natl. Acad. Sci. 109, 3434–3439.

Pond, S.L., Frost, S.D., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679.

Rambaut, A., Suchard, M.A., Xie, D., Drummond, A.J., 2014. Tracer v.16- MCMC Trace Analysis Tool. <http://beast.bio.ed.ac.uk/Tracer>.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542.

Santini, F., Harmon, L.J., Carnevale, G., Alfaro, M.E., 2009. Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. BMC Evol. Biol. 9, 194.

Santini, F., Nguyen, M.T.T., Sorenson, L., Waltzek, T.B., Lynch Alfaro, J.W., Eastman, J.M., Alfaro, M.E., 2013. Do habitat shifts drive diversification in teleost fishes? An example from the pufferfishes (Tetraodontidae). J. Evol. Biol. 26, 1003–1018.

Schoch, C.L., Sung, G.H., Lopez-Giraldez, F., Townsend, J.P., Miadlikowska, J., Hofstetter, V., Robbertse, B., Matheny, P.B., Kauff, F., Wang, Z., Gueidan, C., Andrie, R.M., Trippe, K., Ciufetti, L.M., Wynns, A., Fraker, E., Hodkinson, B.P., Bonito, G., Groenewald, J.Z., Arzanlou, M., de Hoog, G.S., Crous, P.W., Hewitt, D., Pfister, D.H., Peterson, K., Gryzenhout, M., Wingfield, M.J., Aptroot, A., Suh, S.O., Blackwell, M., Hillis, D.M., Griffith, G.W., Castlebury, L.A., Rossman, A.Y., Lumbsch, H.T., Lucking, R., Budel, B., Rauhut, A., Diederich, P., Ertz, D., Geiser, D.M., Hosaka, K., Inderbitzin, P., Kohlmeyer, J., Volkmann-Kohlmeyer, B., Mostert, L., O'Donnell, K., Sipman, H., Rogers, J.D., Shoemaker, R.A., Sugiyama, J., Summerbell, R.C., Untereiner, W., Johnston, P.R., Stenroos, S., Zuccaro, A., Dyer, P.S., Crittenden, P.D., Cole, M.S., Hansen, K., Trappe, J.M., Yahr, R., Lutzoni, F., Spatafora, J.W., 2009. The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. Syst. Biol. 58, 224–239.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., Haussler, D., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050.

Simons, C., Pheasant, M., Makunin, I.V., Mattick, J.S., 2006. Transposon-free regions in mammalian genomes. Genome Res. 16, 164–172.

Smith, B.T., Harvey, M.G., Faircloth, B.C., Glenn, T.C., Brumfield, R.T., 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol. 63, 83–95.

Smith, W.L., Craig, M.T., Quattro, J.M., 2007. Casting the Percomorph Net Widely: the importance of broad taxonomic sampling in the search for the placement of Serranid and Percid fishes. Copeia 1, 35–55.

Stephen, S., Pheasant, M., Makunin, I.V., Mattick, J.S., 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol. Biol. Evol. 25, 402–408.

Sun, K., Meiklejohn, K.A., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T., 2014. The evolution of peafowl and other taxa with ocelli (eyespots): a phylogenomic approach. Proc. Roy. Soc. B 281. http://dx.doi.org/10.1098/rspb.2014.0823.

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.

Townsend, J.P., 2007. Profiling phylogenetic informativeness. Syst. Biol. 56, 222–231.

Townsend, J.P., Leuenberger, C., 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Syst. Biol. 60, 358–365.

Townsend, J.P., Lopez-Giraldez, F., Friedman, R., 2008. The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree. J. Mol. Evol. 67, 437–447.

Wainwright, P.C., Smith, W.L., Price, S.A., Tang, K.L., Sparks, J.S., Ferry, L.A., Kuhn, K.L., Eytan, R.I., Near, T.J., 2012. The evolution of pharyngognathy: a phylogenetic and functional appraisal of the pharyngeal jaw key innovation in labroid fishes and beyond. Syst. Biol. 61, 1001–1027.

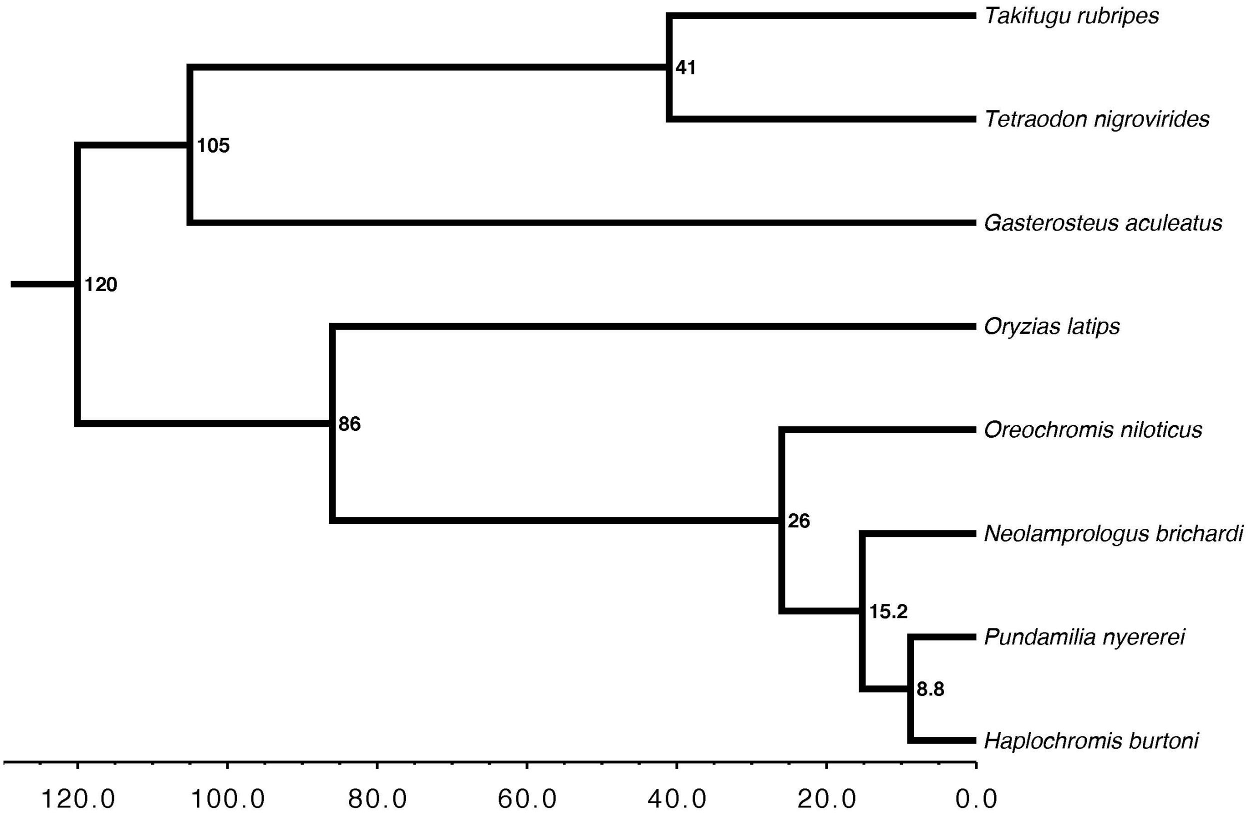
Warnes, G.R., Bolker, B., Lumley T., 2014. Gtools: Various R Programming Tools. R Package Version 3.4.1. <http://CRAN.R-project.org/package=gtools>.

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer, New York.

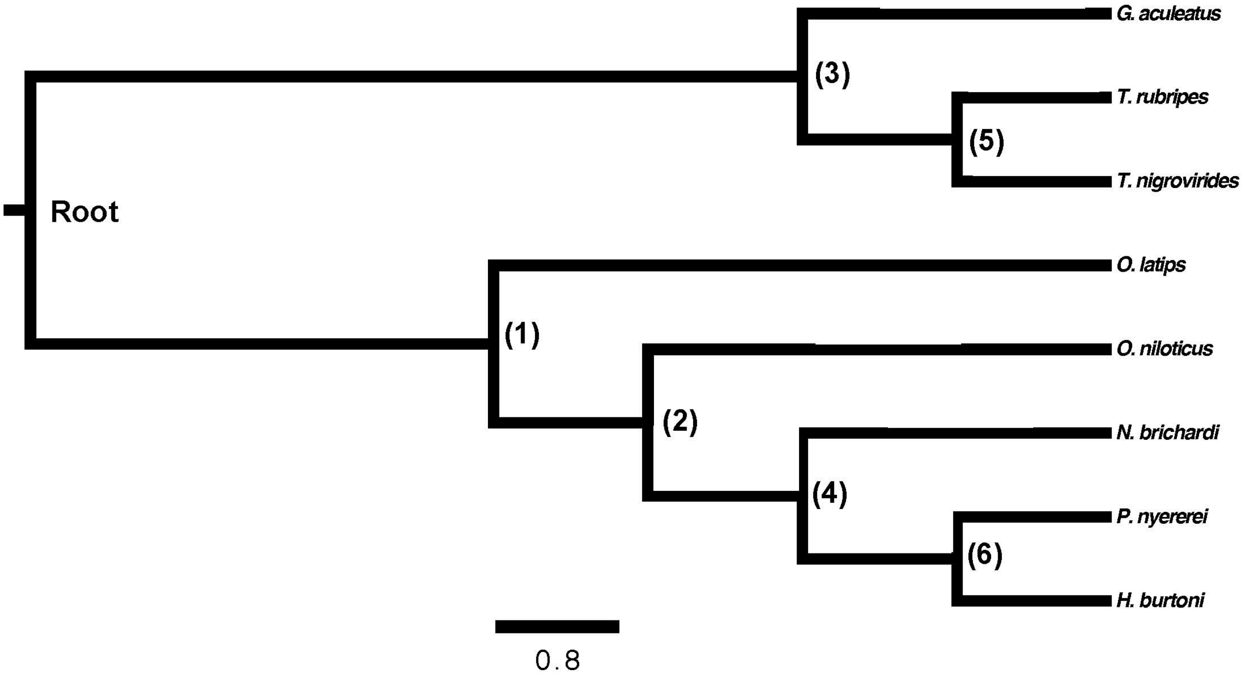
Wickham, H., 2011. The split-apply-combine strategy for data analysis. J. Stat. Softw. 40, 1–29.

Appendix

In this appendix to chapter 2, I provide the phylogenies used for calculating PI. As shown in the first two figures, the topology of each phylogeny is identical. I also show the average and maximum PI for each UCEs core versus that same UCE's flanking regions. The core and flanking regions have similar maximum PI but the flanking region has a 10 fold higher average PI than the core. Finally, I provide information regarding the traditional genes used to infer the phylogeny of the fishes that the UCEs were compared to.

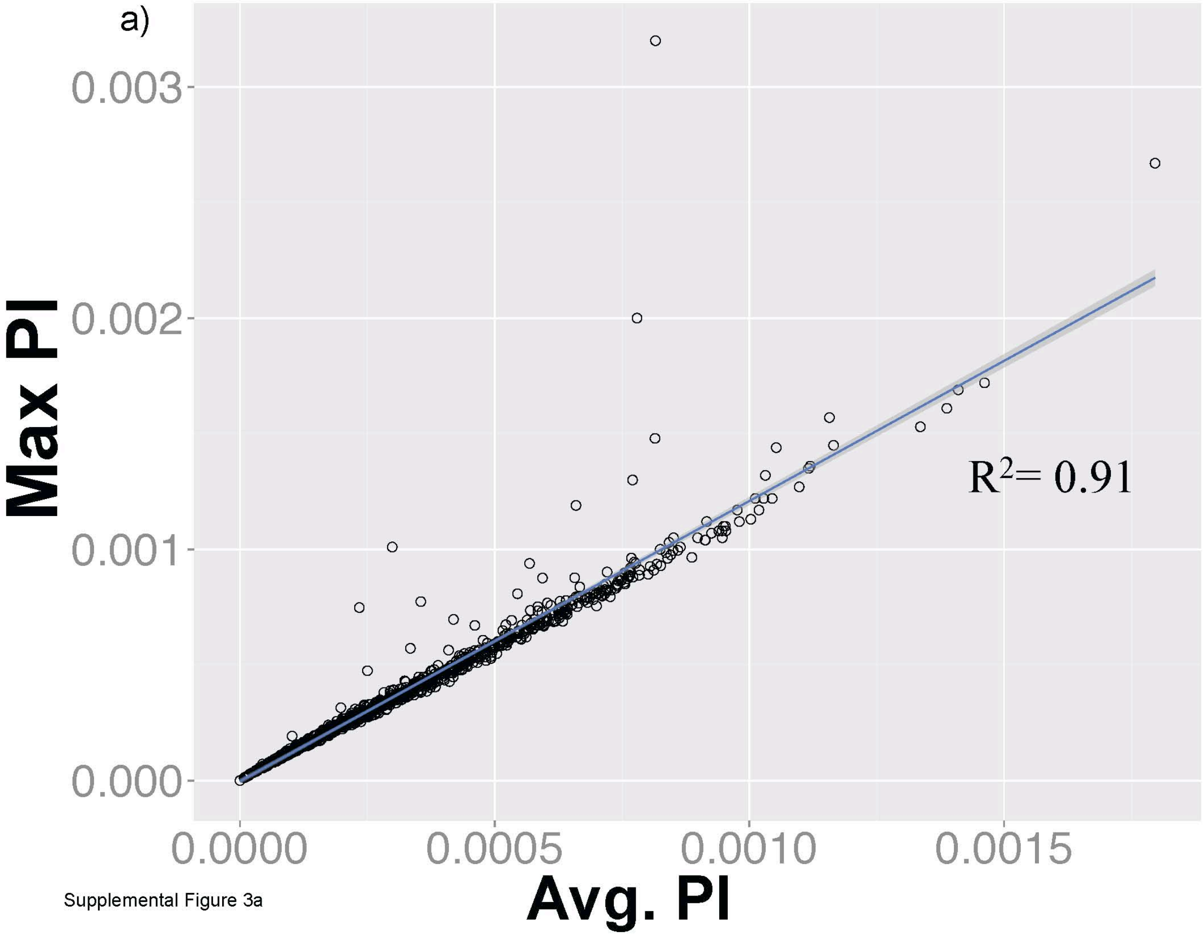


Supplementary Fig. S1.
In silico time calibrated phylogeny used in TAPIR analysis. Time axis is in millions of years before the present.

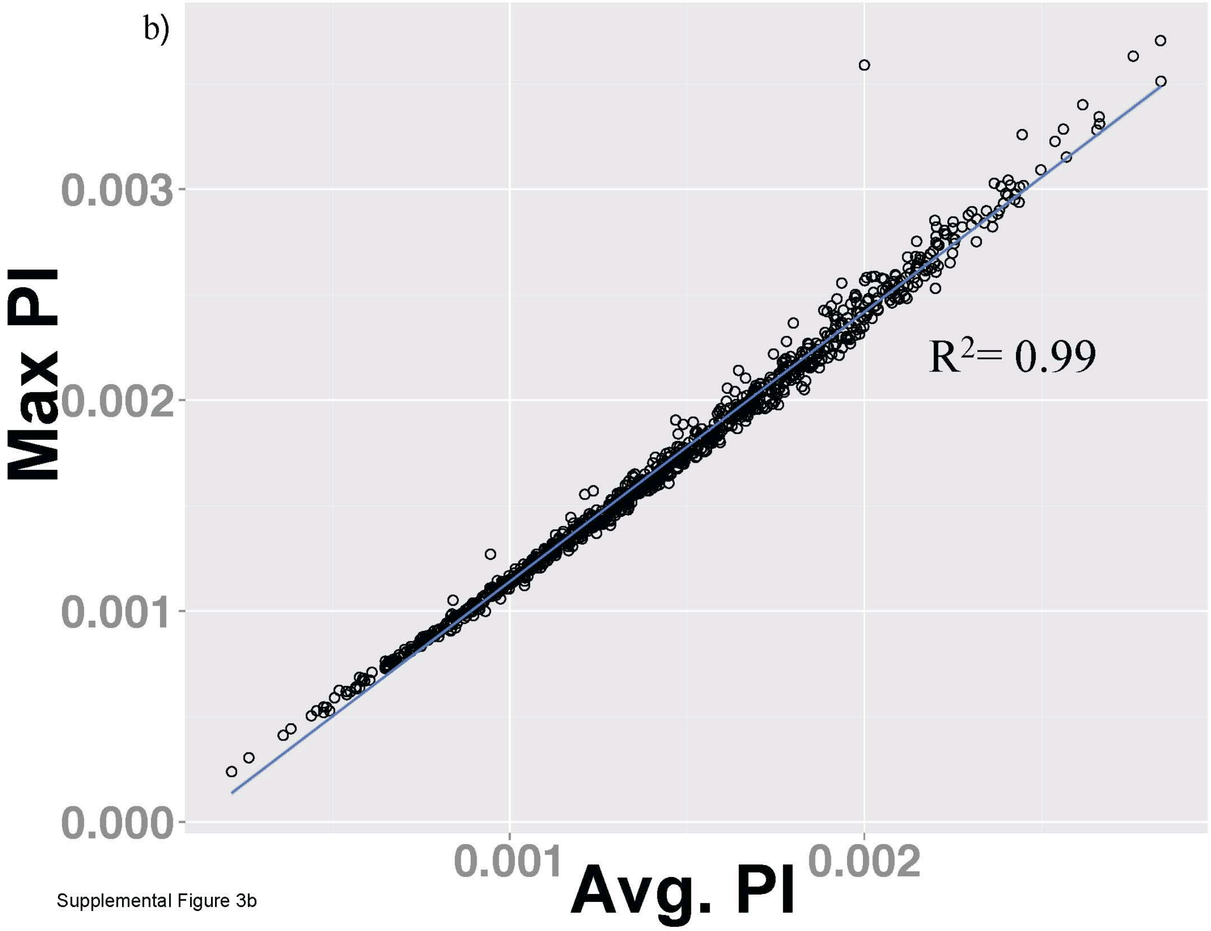


Supplemental Figure 2

Supplementary Fig. S2.
Phylogenetic reconstruction of the eight percomorph species used in our analysis based on ten protein coding genes from Li et al. (2007) and 988 UCEs. The posterior probabilities for all internal nodes were near 1 (see Section 2.6 and Supplemental Table 2 for details).



Supplemental Figure 3a



Supplementary Fig. S3
Linear regression of maximum PI per nucleotide against average PI per nucleotide for the core and flanking regions. (a) Core Regions: adjusted R^2 = 0.91; slope = 1.212; and Y-intercept = 3.225 × 10^{-6}. (b) Flanking regions: adjusted R^2 = 0.99; slope = 1.277; and Y-intercept = −1.364 × 10^{-4}

Supplemental Table S1. Accession data for individual gene information (all from Li et al. 2007) downloaded from the ENSEMBL Genome Browser (Hubbard et al., 2007), the UCSC genome browser (Kent et al., 2002) and GenBank (Benson et. al., 2005).

LOCUS	**Accession number for template *Oryzias latipes***
zic1	EF032914.1
myh6	EF032927.1
RYR3	EF032940.1
Ptr	EF032953.1
tbr1	EF032966.1
ENC1	EF032979.1
Glyt	EF032992.1
SH3PX3	EF033005.1
plag12	EF033018.1
sreb2	EF033031.1

Supplemental Table S2. Branch lengths and the corresponding 95% Bayesian credible intervals (BCI) based on phylogenetic reconstruction using the ten protein coding genes (Li et al., 2007) or Ultra Conserved Elements (UCEs) and the corresponding flanking regions. See Materials and Methods and Supplemental Figure 2 for details.

Branch	**Protein-Coding Average**	**Lower 95% BCI**	**Upper 95% BCI**	**UCE Average**	**Lower 95% BCI**	**Upper 95% BCI**
(Root,3)	0.0268	0.0158	0.0389	0.0198	0.0191	0.0205
(Root,1)	0.0268	0.0158	0.389	0.0198	0.0191	0.0205
(1,2)	0.0648	0.0531	0.0767	0.0528	0.0521	0.0536
(2,4)	0.0083	0.0049	0.0117	0.0046	0.0044	0.0048
(4,6)	0.0037	0.002	0.0056	0.0024	0.0023	0.0025
(3,5)	0.1736	0.1451	0.2054	0.1202	0.119	0.1214
(3, *G. aculeatus*)	0.0837	0.0678	0.0984	0.1041	0.1031	0.1051
(5, *T. rubripes*)	0.1305	0.1084	0.1547	0.0539	0.0531	0.0546
(5, *T. nigrovirides*)	0.1435	0.1205	0.1681	0.0739	0.073	0.747
(1, *O. latipes*)	0.1334	0.1154	0.1517	0.167	0.1656	0.1683
(2, *O. niloticus*)	0.0069	0.0037	0.0101	0.0055	0.0053	0.0057
(4, *N. brichardi*)	0.0048	0.0028	0.0069	0.0046	0.0028	0.0069
(6, *P. nyererei*)	0.0029	0.0015	0.0045	0.0021	0.002	0.022
(6, *H. burtoni*)	0.0013	0.0004	.0024	0.0019	0.0018	0.0019

Chapter 3

Filtering nucleotide sites from ultraconserved elements by phylogenetic signal to noise ratio improves the precision of the avian phylogeny.

PRINCESS S. GILBERT, JING WU, MARGARET W. SIMON, JANET S. SINSHEIMER, MICHAEL E. ALFARO

ABSTRACT

Despite genome scale analyses, high-level relationships among Neoaves birds remain contentious. The placements of the Neoaves superorders are notoriously difficult to resolve because they involve deep splits followed by short internodes. We present a novel and easy to implement site filtering approach based on signal to noise ratios. Using our approach, we investigate whether filtering UCE loci on their phylogenetic signal to noise ratio helps to resolve key nodes in the Neoaves tree of life. We find that filtering UCE data allows us to recover relationships that are not recovered from UCE data without filtering. These relationships include the Columbea + Passerea sister relationship and the Phaethontimorphae + Aequornithia sister relationship. We also find increased statistical support for more recent nodes (i.e. the Pelecanidae + Ardeidae sister relationship, the Eucavitaves clade, and the Otidiformes + Musophagiformes sister relationship). However it is also possible to reduce support for well-established clades and we believe this is the effect of removing too many sites with moderate signal to noise ratios from the UCE datasets. Nonetheless, our results suggest that using our filtering approach as a part of the phylogenomic pipeline can result in the recovery of difficult to resolve splits.

Key words: [Ultraconserved elements; Neoaves; phylogenetic signal; Non-coding DNA]

INTRODUCTION

Phylogenetic reconstruction has greatly benefitted from the recent increase in genome-wide sequence data available on many taxa. The expectation was that with these data, all phylogenetic relationships not subjected to incomplete lineage sorting (ILS) or horizontal gene transfers could be resolved. Yet there are still numerous phylogenetic relationships that are not certain, reminding us that more data can mean more noise not just more signal. In phylogenies with short internodes, there is little opportunity to observe molecular changes on internode branches that would lead to correct resolution. There is also a greater chance of finding misleading change on the subtending branches.

In order to enrich their data in signal and reduce noise, researchers conducting phylogenomic studies have explored ways partition data that only incorporate rates optimal to resolve the phylogenetic relationship in question (Philippe et al. 2011). Assessing markers by their phylogenetic informativeness (PI) is one means of selecting sites across a dataset that are appropriate to resolve a specific phylogenetic question. It has the potential to detect which sites will be able to resolve a short internode followed by long branches (Dornburg et al. 2014; Dornburg et al. 2016; Gilbert et al. 2015; Prum et al. 2015; Townsend et al. 2007; Townsend et al. 2010; Townsend et al. 2011). PI tracks the power of a marker or site to resolve a hypothetical, un-rooted, 4-taxon polytomy (Townsend 2007). However, resolution of such a polytomy can be achieved correctly or incorrectly. Thus, focus instead is needed on the ratio of phylogenetic signal to noise. Townsend et al. (2012) developed the measures of phylogenetic signal and noise based, again, on the phylogenetic quartet (Bandelt & Dress 1986) and their model applies estimates of nucleotide composition and the evolutionary rates of characters to approximate the probability of phylogenetic signal and noise due to convergent or parallel evolution (Bandelt & Dress 1986; Townsend 2012). Although still infrequently applied, ranking phylogenetic markers or removing sites with low signal to noise ratios, especially when analyzing unresolved nodes (polytomies), has been successful (Chen et al. 2015).

Ultraconserved elements (UCEs) are small fragments of DNA that are very similar (greater than 80% identical sequence) across distantly related taxa (Bejerano et al. 2004, Siepel et al. 2005). UCEs have quickly gained popularity in phylogenomics because (1) of the computational ease with which they can be designed for non-model organisms, (2) hundreds or thousands of UCEs can quickly be sequenced using high-throughput technology (targeted enrichment or capture array) and (3) nucleotide variation predominantly found in the UCE flanking regions, carries micro and macro evolutionary signal (Faircloth et al. 2012). Phylogenomic studies using UCEs have improved our understanding of many animal relationships, notably, ray-finned fishes (Faircloth et al. 2013), non-avian reptiles (Crawford et al. 2012), birds (Sun et al. 2014; McCormack et al. 2013; Jarvis et al. 2014), mammals (McCormack et al. 2012), and arthropods (Faircloth et al. 2014).

Some of the deepest branches within Neoaves are poorly resolved (Claramunt et al. 2015, Jarvis et al. 2014; Jetz et al. 2012; Prum et al. 2015; Thomas 2015). Neoaves include all the bird species except for the flightless 'ratite' birds and tinamous (Palaeognathae) and the chickens, turkeys, pheasants, megapodes, ducks, geese and swans (Galloanseres). Although debated (Brown et al. 2007; Cracraft et al. 2015; Ericson et al. 2006; Mitchell et al. 2015) it is believed that nearly all neoavian orders evolved between 50-70MYA (Jarvis et al. 2014). Considerable incomplete lineage sorting (Feducia, 1995; Poe & Chubb, 2004), measured most recently via indels and transposable elements (Suh et al. 2015), were cited as possibly affecting the inference of the deepest branches of Neoaves and Afroaves (Jarvis et al. 2014). Jarvis et al. (2014) also found that phylogenies of 48 bird species constructed using UCEs exhibited lower resolution on deep branches in Neoaves than the phylogenies constructed using both gene and UCE data. This lower resolution is explained not only by the reduction in data but also the lower rate of evolution of the UCEs relative to genes (Jarvis et al. 2014).

Here, we reconstruct the UCE phylogeny of the same 48 bird species used by Jarvis et al. (2014). We apply Townend's model to get phylogenetic signal to noise ratio estimates (Townsend et al., 2012) in order to select the specific sites that are likely to resolve the deepest branches of Neoaves (nodes occurring between 60-62MYA) and reconstruct the phylogeny using only these sites. Again using Townend's signal to noise ratios, we also select the avian UCE nucleotide sites that should, in principle, be optimized for resolving speciation events separated by longer internodes and which have occurred more recently in time (nodes occurring between 27-64MYA, Figure 1). We then reconstruct the phylogeny of the Neoaves using these filtered sites and compare the phylogeny to one based on the unfiltered UCE sites and the total evidence based maximum likelihood phylogeny of Jarvis (ExaML-TENT), a reconstruction based on UCEs, exonic and intronic regions. Our implementation of the signal to noise ratio for filtering sites is generally applicable and simple to implement so that it can be used with any large genomic data set, not just UCEs, to improve phylogenetic resolution.

MATERIALS AND METHODS

2.1 Rationale for Neoaves Nodes Chosen for Signal to Ratio Calculations

We chose two general depths in the phylogeny, one representing a series of deep divergences followed by long branches, the second representing a more recent rapid radiation with longer internodes between branching events. The deepest branching nodes of Neoaves occurred between 60-70MYA (Figure 1). Thus this region provides an excellent test case for resolving a deep branching (62MYA) with short internodes (5 million years) problem (Figure 1, red). These nodes also exhibited low bootstrap support values in the UCE species phylogeny published in Jarvis et al. (2014). Using their publically available files (Aberer et al., 2014) we recreated and annotated the phylogram (Figure 1) and cladograms (Figs. 2,3,5 &7) to reflect the results of Jarvis et al. (2014).

The second problem focuses on the rest of the Neoaves orders and therefore the majority of all remaining extant bird lineages, evolved by 27MYA (Figure 1, blue). For example, major groups like all passerine birds, bee-eaters, woodpeckers, hummingbirds, swifts, flamingos and grebes (Figure 1, blue). Besides being of intrinsic interest in understanding bird systematics, filtering based on this time period provides a second test case for a shallower reconstruction problem: resolving recently evolved clades (27MYA) with moderate internode lengths (75 MYA) (Figure 1, blue).

2.2 Phylogenetic signal to noise analysis

Townsend et al. (2012) developed a model that estimates the probability of phylogenetic signal, probability of phylogenetic noise (due to convergent or parallel evolution) and the probability of a true polytomy for a given locus at a given node. These estimates incorporate the date of the node and the length of the subtending branches following that node. Thus, the model relies on evolutionary rates and estimates of node age and internode length. The evolutionary rate is simply the substitution rate of a character. The character state space is based on the percentage of each nucleotide type and the transition - transversion rates (rTA, rTG, rCA, rCG). The time components of the calculation are defined by the time at which the nodes of interest occur and the length of the descendant branches from that point.

For the deep (60-62MYA) and shallow (27-64MYA) branching questions we calculated the probability of signal (C), probability of noise (N) and the probability of a true polytomy (P) for each site in each UCE. Sequence alignment data were downloaded from Aberer et al. (2014). We used Mathematica versions 10.2-10.4 (Wolfram Research, Inc., 2016) and modified computer code from Townsend et al. (2012) and Phydesign (Lopez-Giraldez and Townsend, 2011) to calculate these measures. To do so required calculating the transition -transversion rates, the percentage of each nucleotide type and the substitution per site rate of each nucleotide in each UCE, which we did using TAPIR (Faircloth et al. 2012*a*). TAPIR creates a separate JSON file (Ooms 2014) for each UCE. We processed each JSON file to isolate the required inputs with a computational pipeline, written in the statistical computing language R, to remove all information except the transition -transversion rates, the percentage of each nucleotide type and the substitution per site rate of each nucleotide (all scripts to be made available on Dryad). We then used these three pieces of information along with the node age and internode length to calculate the probability of C, N and P.

For our analysis we customized the Mathematica notebook in order to calculate C, N, and P for each site across unfiltered UCEs. Sites with a zero rate of change lead to P = 1 and therefore were excluded from the calculation of phylogenetic C, N, and P. Sites with higher than 0.2 substitutions per site were also excluded to eliminate artificially high estimates that resulted from indels introduced in UCE sequence alignment in regions of high uncertainty (Supplemental and Appendix Figs., Philippe and Roure 2011). Signal to noise (SN) can be defined in several ways (Townsend et al., 2012). We used SN = C/(C+N) which is equivalent to SN = C/(1-P). We looked for sites that sufficiently shifted the distribution of SN towards the maximum (Supplemental Figures 1-4) and these fell within the top 20% of the distribution. Calculating the SN ratio probabilities for each nucleotide within the UCE dataset allowed us to isolate 768,612 unique sites from 3,603 UCEs (Supplemental Figs. 1-4) within the top 20% of all non-zero signal probabilities for the each of the two time periods we analyzed (60-62MYA & 27-64MYA). The resulting datasets were used to create the phylogenetic reconstructions described in section 2.4.

In order to test the effect of filtering the UCE data we looked at two comparisons: The unfiltered UCE phylogeny compared to each of the filtered phylogenies (deep and shallow) and the unfiltered UCE phylogeny compared to the UCE and ExaML-TENT phylogenies presented in Jarvis et al. (2014). For parts of the phylogenies that remained the same we compared bootstrap support values at the recovered node. For parts of the phylogenies that were different we sought confirmation from independent studies before accepting the clade as valid.

2.4 Phylogenetic Reconstruction

We chose to concatenate the UCEs because it was computationally far less intensive. We used a general time reversible model of evolution with gamma distributed rate variation among sites to compute 20 distinct maximum likelihood topologies starting from 20 distinct randomized maximum parsimony starting topologies (scripts available from DRYAD.org.). We parallelized the computations with 24 threads of execution spread over 12 processing cores in RAxML (Stamatakis 2014). We computed 100 bootstrap alignment replicates under the GTRGAMMA model for the unfiltered data and 200 bootstraps for the filtered data. We then reconciled the best phylogeny (highest GRTGAMMA likelihood score) with the bootstrap replicates. Results were visualized using customized R Code (R Core Team, 2016). We used exact test of proportions to compare bootstrap support for equivalent nodes in the different phylogenies and use p-values to determine significance at a significance level of alpha = 0.05. For those nodes observed in all three data sets, deep filtered, shallow filtered or unfiltered, we also scored the nodes as more or less confidently observed in the filtered versus unfiltered data. We used a binomial model with success probability for a single node to be observed with more confidence as 0.5 to test whether there was a significant increase in support overall. In order to reduce the possible reasons for topological differences and bootstrap support differences for nodes, we compared our filtered UCE phylogenies to the unfiltered UCE phylogeny we reconstructed (described in section 2.4) and not the UCE species phylogeny available from Jarvis et al. (2014).

2.5 Heatmap Analysis

Filtered data can improve confidence by increasing the average signal to noise ratio but if the data are too aggressively filtered there can be a loss of confidence.   In order to examine this balance in our case, we created a heatmap to summarize the results for 33 nodes observed in all three UCE phylogenies. We arranged these nodes by their relative age based on Jarvis et al. (2014)'s analyses.

RESULTS

As expected, because we used concatenated UCEs and Jarvis et al. (2014) created a phylogeny for each UCE and then constructed a species phylogeny from the collection, we find differences between our unfiltered UCE phylogeny and Jarvis' UCE phylogeny. These differences are shown in figure 2. Because these differences are due not to the data or filtering of the data but to the phylogenetic methods, we compare phylogenies reconstructed from filtered UCE data only to the phylogeny we constructed using the concatenated unfiltered UCE data.

3.1 Unfiltered UCE phylogeny vs. ExaML-TENT (Jarvis et al., 2014) phylogeny

There are a number of differences between our unfiltered UCE phylogeny and the ExaML-TENT phylogeny of Jarvis et al. (2014) that used both exon data and UCE data to reconstruct the phylogeny (Figure 3). These differences likely occurred because of difference in the data (the ExaML-TENT was based on introns, exons, and UCE datasets) and because of the differences in the assumptions of reconstruction computations, for example, our UCE concatenation in RaxML vs. gene phylogeny/species phylogeny analysis in ExaML. Our unfiltered UCE phylogenetic reconstruction (Figure 3, left phylogeny, Node H) placed the Coliiformes as the outgroup to a clade containing the Cavitaves, Strigiformes and Accipitrimorphae (left phylogeny, Node HH) while the ExaML-TENT analysis from Jarvis et al. (2014) placed the speckled mousebird as sister to the clade Cavitaves (right phylogeny, Node F).

Other differences between these two phylogenies were in the placement of the Caprimulgimorphae clade (Figure 3, Node V, highlighted in brown) and the Phaethontimorphae clade (Node P, highlighted in light blue). In the Jarvis et al. (2014) ExaML-TENT phylogeny Caprimulgimorphae (right phylogeny, Node V) was placed sister to the Otidimorphae (right phylogeny, Node X), highlighted in orange (right phylogeny, Node Y, 91% BS) and

Phaethontimorphae (right phylogeny, Node P) was placed sister to the core waterbirds (right phylogeny, Node O), Aequornithia (right phylogeny, Node Q, 70% BS). However, our analysis placed Caprimulgimorphae (left phylogeny, Node V) sister to Phaethontimorphae (left phylogeny, Node P) resulting in Node JJ (left phylogeny, 42% BS) and the core landbirds (left phylogeny, Node I) sister to the core waterbirds (Figure 3, left phylogeny, Node O) resulting in Node II (left phylogeny, 100% BS).

Our unfiltered UCE phylogeny did not recover the Jarvis et al. (2014) highly supported Columbea clade (Figure 3, right phylogeny, Node DD, 100% BS). Instead it placed Columbimorphae sister to all Passerea (left phylogeny, Node NN, 57% BS) and Columbimorphae + Passerea sister to Phoenicopterimorphae (left phylogeny, Node OO, 73% BS). This result places Phoenicopterimorphae (Node CC) instead of Columbea (Node DD) as the sister to all the remaining Neoaves (Node OO) and is the same topology as that found in the UCE species phylogeny from Jarvis et al. (2014)(See Figure 2, Node OO).

There were nodes that had the same topology in the two phylogenies for which we observed changes in bootstrap confidence. In Telluraves (Figure 3, left phylogeny, Node I) and within Afroaves (left phylogeny, Node H) we observed a slight but significant decrease in support for the Coraciiformes + Piciformes sister relationship (Node AAA, 96% BS, left phylogeny vs. 100% BS, right phylogeny, $p = 0.0119$). We also observed a decrease in node support for Eucavitaves but this decrease is not statistically significant (Node D, 66% BS, left phylogeny vs. 72% BS, right phylogeny, $p = 0.2886$). Within the core waterbirds (Node O) we observed a significant decrease in support for the Dalmatian pelican + little egret sister relationship (Node J, 90% BS, left phylogeny vs. 100% BS, right phylogeny, $p < 0.0005$). We see no virtually change in support for the hoatzin + Cursorimorphae sister relationship (Node T, 90% BS, left phylogeny vs. 91% BS, right phylogeny, p

= 0.8341). Support for the monophyletic clade Otidimorphae slightly decreased (Node X, 93% BS, left phylogeny vs. 100% BS, right phylogeny, p = 0.0004). Support for Columbimorphae remained the essentially the same (Node BB, 99% BS, left phylogeny vs. 100% BS, right phylogeny, p = 0.3333).

3.2 Comparison of the shallow filtered UCE phylogeny vs. the unfiltered UCE phylogeny

For the phylogeny constructed with sites having the highest signal to noise ratio for species divergences occurring between 27-64MYA (shallow filtered), Coliiformes was the sister lineage of Strigiformes (Figure 4, Node SS, left phylogeny). This placement differed from the Coliiformes placement as the sister to all remaining Afroaves based on the unfiltered UCE phylogeny (Figure 4, Node H, right phylogeny). We note however that in both cases the bootstrap support was low (Node SS, 51% BS, left phylogeny vs. Node H, 55% BS, right phylogeny).

In general, for the portions of the two phylogenies that had the same topology, we observed increased support due to filtering sites on signal to noise ratios. Specifically, for the placement of Caprimulgimorphae (Figure 4, Node V, brown branches) as sister to Phaethontimorphae (Node P, light blue branches) we observed significantly increased support (Node JJ, 58% BS, left phylogeny vs. 42% BS, right phylogeny, p=0.0101). We also observed significantly higher support for the entire Otidimorphae clade (Node X, orange branches; 100% BS, left phylogeny vs. 93% BS, right phylogeny, p =0.0004) as well as for the sister placement of Otidiformes to Musophagiformes (Node W, orange branches; 92% BS, left phylogeny vs. 77% BS, right phylogeny, p = 0.0005).

We also observed an increase in support for major Passerea clades. Specifically, the Otidimorphae + (Cursorimorphae+ (Caprimulgimorphae+ Phaethontimorphae)+ Aequornithia +Telluraves)(Figure 4, Node MM, 86% BS, left phylogeny vs. 56% BS, right phylogeny, p = 0.0001), Cursorimorphae + (Caprimulgimorphae+ Phaethontimorphae)+ Aequornithia + Telluraves)(Figure

4, Node LL, 87% BS, left phylogeny vs. 54% BS, right phylogeny, p < 0.00005) and (Caprimulgimorphae+ Phaethontimorphae)+ Aequornithia + Telluraves (Node KK, 65% BS, left phylogeny vs. 42% BS, right phylogeny, p = 0.0002).

The most important difference between the shallow filtered UCE phylogeny and the unfiltered UCE phylogeny lies in the relationship between Columbimorphae and Phoenicopterimorphae (Nodes BB and CC, both highlighted in purple). The shallow filtered UCE sites recover a sister relationship between Columbimorphae and Phoenicopterimorphae (i.e. the Columbea clade, Node DD) which fails to be recovered in our unfiltered UCE phylogeny or the unfiltered UCE species phylogeny from Jarvis et al (2014).

The bootstrap support for and within the Telluraves clade remains very strong (Figure 4, Node I, 100% BS). We do however see a significant decrease in support for the Cursorimorphae clade (Node S, 44% BS left phylogeny vs. 99% BS right phylogeny, p < 0.00005) as well as the Cursorimorphae + hoatzin sister relationship (Node T, 71% BS, left phylogeny vs. 90% BS right phylogeny, p = 0.0001). The decreased support for Columbimorphae is not significant (Node BB, 96% BS, left phylogeny vs. 99% BS, right phylogeny, p = 0.0979). Within Neoaves the Telluraves + Aequornithia sister relationship support decreased dramatically (Node II, 36% BS, left phylogeny vs. 100% BS, right phylogeny, p < 0.00005).

3.3 Comparison of shallow filtered UCE phylogeny vs. ExaML-TENT phylogeny

When comparing the shallow filtered phylogeny to the ExaML-TENT phylogeny, the placement of Strigiformes (Figure 5, Node SS or G), Phaethontimorphae (Node P, light blue branches), Caprimulgimorphae (Node V, brown branches), Otidimorphae (Node X, orange branches), and Aequornithia (Node O, navy blue branches) differed. The shallow filtered phylogeny placed Strigiformes sister to the Coliiformes, although bootstrap support was low (Node

SS, 51% BS, left phylogeny) and this broke up the Coraciimorphae (Node TT, 50% BS, left phylogeny), found to be monophyletic in ExaML-TENT phylogeny (Node F, right phylogeny).

Phaethontimorphae (Figure 5, Node P, light blue) was placed sister to Caprimulgimorphae (Node V, brown), (Node JJ, 58% BS, left phylogeny) and Otidimorphae (Node X, orange) was placed sister to all remaining, extant Passerea (Node MM; 86% BS, left phylogeny). In the ExaML-TENT phylogeny, Strigiformes is placed sister to Coraciimorphae (Node G, 84% BS, right phylogeny), Phaethontimorphae (Node P, light blue) is placed sister to the core water birds, Aequornithia (Node Q, 70% BS, right phylogeny), and Caprimulgimorphae (Node V, brown) is placed sister Otidimorphae (Node Y, 91% BS, right phylogeny). Caprimulgimorphae + Otidimorphae (Node Y) are placed sister to all remaining Passerea (Node Z, 91% BS, right phylogeny). The ExaML-TENT phylogeny and the shallow filtered phylogeny both recovered a monophyletic Columbea clade albeit with decreased support for the shallow filtered phylogeny (Node DD, 55% BS, left phylogeny and 100% BS, right phylogeny, purple branches, $p < 0.00005$).

Within the core landbirds, Telluraves (Figure 5, Node I), we observed increased support for the Eucavitaves clade in the shallow filtered phylogeny (Node D, 100% BS left phylogeny vs. 72% BS right phylogeny, $p < 0.00005$). We also found support for the monophyly of members of the Otidimorphae clade (Node X, orange branches). Specifically we, too, found Otidiformes to be closely related to Musophagiformes as was found in the ExaML-TENT phylogeny but the shallow filtered dataset yielded increased bootstrap support (Node W, 92%BS, left phylogeny vs. 55%BS, right phylogeny, $p < 0.00005$).

In figure 5 we observe a large and statistically significant decrease in support for Cursorimorphae (Figure 5, Node S, 44% BS, left phylogeny vs. 96% BS, right phylogeny, $p < 0.00005$) and a decrease in support for the Cursorimorphae + hoatzin sister relationship (Node T, 71% BS, left phylogeny vs. 91% BS, right phylogeny, p< 0.00005). We observed a slight but significant decrease in support for Columbimorphae (Node BB, 96% BS, left phylogeny vs. 100% BS, right phylogeny, p =0.0073) and we saw a substantial decrease in support for Columbea (Node DD, 55% BS, left phylogeny vs. 100% BS, right phylogeny, p < 0.00005). None of the Neoaves backbone nodes overlapped so support values could not be compared.

3.4 Comparison of the deep filtered UCE phylogeny vs. unfiltered UCE phylogeny

We now turn our attention to the results of filtering designed to improve resolution of deeper branched polytomies. The deep filtered phylogeny's topology differed unfiltered phylogeny's topology from the within the Afroaves clade (Figure 6, Node H,). As mentioned in an earlier comparison (see section 3.2 and Figure 3), our unfiltered UCE phylogenetic reconstruction (Figure 6, Node H, right phylogeny) placed Coliiformes as sister to the clade containing Cavitaves, Strigiformes and Accipitrimorphae (the eagles and vultures). In contrast, in the deep filtered phylogeny, the Accipitrimorphae (Node FF, left phylogeny) were placed sister to a clade containing Coliiformes (Node UU), Strigiformes (Node GG) and Cavitaves (Node E). Although the placement of Coliiformes differed between the two topologies, support for Afroaves increased in the deep filtered phylogeny (Node H, 100% BS, left phylogeny vs. 55% BS, right phylogeny). Also, Strigiformes (Node GG) was placed sister to Cavitaves (Node E), which forces the Coraciimorphae clade (Node UU, 56% BS, left phylogeny) to be paraphyletic and differs from the ExaML-TENT phylogeny.

Topologies also differed between the two phylogenies on the placement of Caprimulgimorphae (i.e. hummingbirds, swifts and nightjars). The deep filtered phylogeny placed Caprimulgimorphae (Figure 6, Node V, brown branches) sister to all Telluraves (Node I, green branches) with strong bootstrap support (Node YY, 100%BS, left phylogeny). In the unfiltered

UCE phylogeny, Caprimulgimorphae (Node V, right phylogeny) was placed sister to Phaethontimorphae (Node P, light blue branches, the tropicbirds and sunbittern) with low support (Node JJ, 42% BS, left phylogeny).

Additionally, the deep filtered phylogeny placed Aequornithia (Figure 6, Node O, navy blue branches) sister to Phaethontimorphae (Node P, light blue branches) with 53% BS (Node Q, left phylogeny). This result contrasts with Aequornithia's (Node O) placement as sister to the core landbirds (Node I, Telluraves, green branches) in the unfiltered UCE phylogeny (Node II, 100% BS, right phylogeny).

In the deep filtered phylogeny, Opisthocomiformes was placed sister to Gruiformes with 45% BS (Figure 6, Node WW, left phylogeny). In the unfiltered phylogeny, Opisthocomiformes is placed sister to Cursorimorphae (Node T, 90% BS, right phylogeny), a clade that includes Gruiformes and Charadriiformes (Node S, 99% BS, right phylogeny).

The deep filtered phylogeny placed the Columbiformes sister to Phoenicopterimorphae (Figure 6, Node VV, 77% BS, left phylogeny), which includes Phoenicopteriformes and Podicipediformes. This placement of Columbiformes contradicts the placement of Columbiformes in the phylogeny reconstructed using unfiltered UCE sites. In the unfiltered UCE phylogeny Columbiformes is placed sister to Mesitornithiformes and Pterocliformes (Node AA), a clade referred to as Columbimorphae (Node BB, 99% BS, right phylogeny). This difference between the deep filtered phylogeny and the unfiltered phylogeny is interesting because the relationship among the Columbimorphae and Phoenicopterimorphae results in the recovery of the monophyletic Columbea clade (Node DD, 73%BS, left phylogeny, purple branches).

For portions of the two phylogenies with the same topology, we found essentially no change in the support for the placement of the Coraciiformes and Piciformes (Figure 6, Node AAA 99%

BS, left phylogeny vs. 96% BS, right phylogeny, p=0.0979). We found dramatically increased support for the placement of Trogoniformes within the clade Eucavitaves (Node D, 100% BS, left phylogeny vs. 66% BS, right phylogeny, p < 0.00005). We found 100% BS support for the Pelicanidae + Ardeidae sister relationship both from the Order Pelecaniformes (Node J, left phylogeny) while this relationship was only recovered with 90% BS in the unfiltered UCE dataset (Node J, right phylogeny, p< 0.00005). We observed slightly but significantly increased support for the entire Otidimorphae clade (Node X, 99% BS, left phylogeny vs. 93% BS, right phylogeny, orange branches, p =0.0074) and a greater increase in support for the sister placement of Otidiformes to Musophagiformes (Node W, 92% BS, left phylogeny vs. 77% BS, right phylogeny, orange branches, p <0.00005).

Within Afroaves we observed a dramatic decrease in support for the Cavitaves + Strigiformes sister relationship (Figure 6, Node GG 42% BS, left phylogeny vs. 100 % BS, right phylogeny, p <0.00005). Within the core waterbirds (Node O) we observed a slight decrease in support for the Procellariimorphae clade (Node M, 93% BS, left phylogeny vs. 100% BS, right phylogeny, p=0.0062). We observed a big increase in support for the Passerea backbone node splitting Otidimorphae from all remaining extant Passerea (Node MM, 82% BS, left phylogeny vs. 56% BS, right phylogeny, p<0.00005), and non-significant increases in the support for Node LL (61% BS, left phylogeny vs. 54% BS, right phylogeny, p=0.2645) and a non-significant decrease in support for the Passerea+Columbea node (Figure 6, Node EE, 98% BS, left phylogeny vs. 100% BS, right phylogeny, p = 0.3052).

3.5 Comparison of the deep filtered UCE phylogeny vs. ExaML-TENT phylogeny

Four topological changes were observed between the deep filtered phylogeny and the ExaML-TENT phylogeny (Figure 7). Within Telluraves (Node I, green branches), we found support for the paraphyly of Coraciimorphae (Node UU, 56% BS, left phylogeny) and for the inclusion of Strigiformes (Node GG, 42% BS, left phylogeny), an order not included within Coraciimorphae by Jarvis et al. (2014) (Node G, 84% BS right phylogeny). We did not recover the sister relationship of Caprimulgimorphae (Node V, left phylogeny, brown branches) to Otidimorphae (Node X, left phylogeny, orange branches) as was found in ExaML-TENT phylogeny (Node Y, 91% BS, right phylogeny). Additionally, in deep filtered phylogeny the placement of Opisthocomiformes (Node WW, 45% BS) split the sister relationship of the Gruiformes and Charadriiformes (Node XX, 43% BS, left phylogeny), a clade which was highly supported in the ExaML-TENT phylogeny.

As mentioned previously, the most important similarity between the optimized phylogeny (Figure 7, left phylogeny) and the ExaML-TENT phylogeny lies in the relationship among the Columbimorphae and Phoenicopterimorphae (Node DD, purple branches). Even though the placement of the Columbiformes (Node VV, 77% BS left phylogeny) as sister to the aquatic Phoenicopterimorphae (Node CC, purple branches) is unique to this study, the monophyly of Columbea (Node DD) is recovered in the deep filtered phylogeny as in the ExaML-TENT phylogeny.

Australaves in both phylogenies has identical node support values (Figure 7, Node C, 100% BS, both phylogenies) and we see no essentially change in support for the Coraciiformes + Piciformes sister relationship (Node AAA, 99% BS, left phylogeny vs. 100% BS, right phylogeny, $p=0.4987$). The Coraciimorphae+ Strigiformes clade has decreased support (Node UU, 56% BS left phylogeny vs. Node G, 84% BS, right phylogeny, $p<0.00005$). Within Aequornithia , the water birds (Node O) we see a slight but significant decrease in support for Procellariimorphae (Node M, 93% BS, left phylogeny vs. 100% BS, right phylogeny, $p = 0.0001$). We also see a decrease in the Aequornithia + Phaethontimorphae sister relationship (Node Q, 53% BS, left phylogeny vs. 70%

BS, right phylogeny, p=0.0007). We see no essentially change in support for Otidimorphae (Node X, 99% BS, left phylogeny vs. 100% BS, right phylogeny, p=0.4987) but a large and significant decrease in support for Columbea (Node DD, 73% BS, left phylogeny vs. 100% BS, right phylogeny, p<0.00005). Along the backbone of Neoaves we see essentially no change in support for the Columbea + Passerea sister relationship (Node EE, 98% BS, left phylogeny vs. 100% BS, right phylogeny, p=0.1231). The remaining neoavian backbone nodes are different between the two phylogenies.

3.6 Overall comparison of consistent clades from unfiltered and filtered phylogenies.

Of the 33 nodes consistently observed in all three UCE phylogenies, 22 of these nodes had 100% bootstrap support (Figure 8). Of these 22, eight nodes had estimated ages of less 39 MYA. There were 11 nodes in which one or more of the phylogenies had less than 100% BS support. Seven of these nodes had estimated ages of greater than 62 MYA.

Comparing the shallow filtered phylogeny (27-64MYA) to the unfiltered phylogeny, we found eight nodes with higher support in the shallow filtered phylogeny than in the unfiltered UCE phylogeny and only one node that had higher support in the unfiltered phylogeny than in the shallow filtered phylogeny (Figure 8). These results suggest a significant improvement in support (p=0.0195). There were eight nodes that had higher support in the deep phylogeny (60-62MYA) than in the unfiltered phylogeny. We found three nodes that had higher support in the unfiltered phylogeny than in the deep filtered phylogeny (Figure 8). Although not significant, there is a trend towards improved confidence with deep filtering (p=0.1133).

4 DISCUSSION

Here we present a novel pipeline that can be used on genomic datasets to increase the ability of those data to resolve phylogenetically difficult problems. We use this pipeline to find sites in the UCEs that are most appropriate for answering specific questions in neoavian evolution (Figure 1). The improvements afforded by our pipeline fell into two categories: Increased bootstrap support for clades that were already supported and support for clades that had not been supported before.

We found increased bootstrap support for many clades after filtering our UCE data spanning both deep (Figures 4-7, Nodes W, H, X, LL, MM, all greater than 62MYA), and more shallow time spans (Figures 4-7, Node AAA, 41 MYA, D and J 55 MYA). Because the resulting clades were recovered with high support in both filtered phylogenies we believe that by optimizing the UCE data for species divergences occurring either between 60-62 MYA or 27-64 MYA we were able to remove sites that carried higher amounts phylogenetic noise. Encouragingly, these clades were also found in the much larger and more exhaustive total evidence based phylogeny (ExaML, Figures 1, 3, 5, and 7).

There were singular changes that were not supported by both filtered datasets but which were biologically compelling. The increased support for the Phaethontimorphae + Aequornithia sister relationship in our deep filtered phylogeny (Figures 6 and 7, Node Q) was not observed in the shallow filtered or unfiltered phylogenies. Phaethontimorphae include the tropicbirds and sunbittern while Aequornithia includes the majority of all neoavian waterbirds and together these clades share similar aquatic behaviors and habitats. Additionally, this relationship was also found in the total evidence based phylogeny from Jarvis et al. 2014 and Prum et al. (2015). We suspect deep filtering the UCE sites "turned down the noise" that essentially resolved this relationship incorrectly due to homoplasy or convergence.

Another change that was found in the deep filtered phylogeny but not the shallow filtered, unfiltered or the ExaML phylogenies was the strong placement of Caprimulgimorphae (the hummingbirds, swifts and nightjars) as sister to the all core landbirds, Telluraves (Node YY, Figures 6 and 7). This relationship is intriguing, however, as with all our findings, we acknowledge that this outcome is sensitive to the dates for which we selected the highest signal. This increased support for the Caprimulgimorphae + Telluraves sister relationship contradicts the Aequornithia + Telluraves sister relationship (Node II) recovered in shallow filtered phylogeny (Figure 4 and Figure 5). Contradictions between our two filtered phylogenies are to be expected as the underlying datasets these topologies are built upon are targeting different time periods. Thus these topological discrepancies highlight the importance of accurate species divergence estimations, as these estimations heavily impact the subsequently optimized dataset and the resulting phylogenetic reconstructions.

Filtering UCEs at the base pair level for their signal to noise ratio is novel and can improve signal resolution. However it is unclear how support for nodes outside of the targeted filtered range should be interpreted. Filtering depends on the number of sites that might possibly resolve a toplogy at a given period of time. Some arrangements might be intractable even with large amounts of data if there are too many rapid divergences. Likewise clades with especially patchy fossil records (like birds) reduce the accuracy of time-calibrated phylogenies and thus the effectiveness of filtering. Bootstrap support depends both on the strength of the relationships (e.g. a long central branch) and the amount of data used to infer the relationships so it is possible to filter too aggressively and reduce bootstrap confidence in a correctly inferred clade

Being able to resolve divergences at multiple depths across a phylogeny is a strength of UCEs but they, as well as other genome-wide markers, are not immune to the lack of resolution for certain nodes. Little is known about filtering non-exonic phylogenomic datasets such as a collection of UCEs to decrease the effects of systematic bias during phylogeny reconstruction. Thus phylogenetics requires more sensitive methods and better study designs allowing careful selection of the most appropriate data to resolve these nodes.

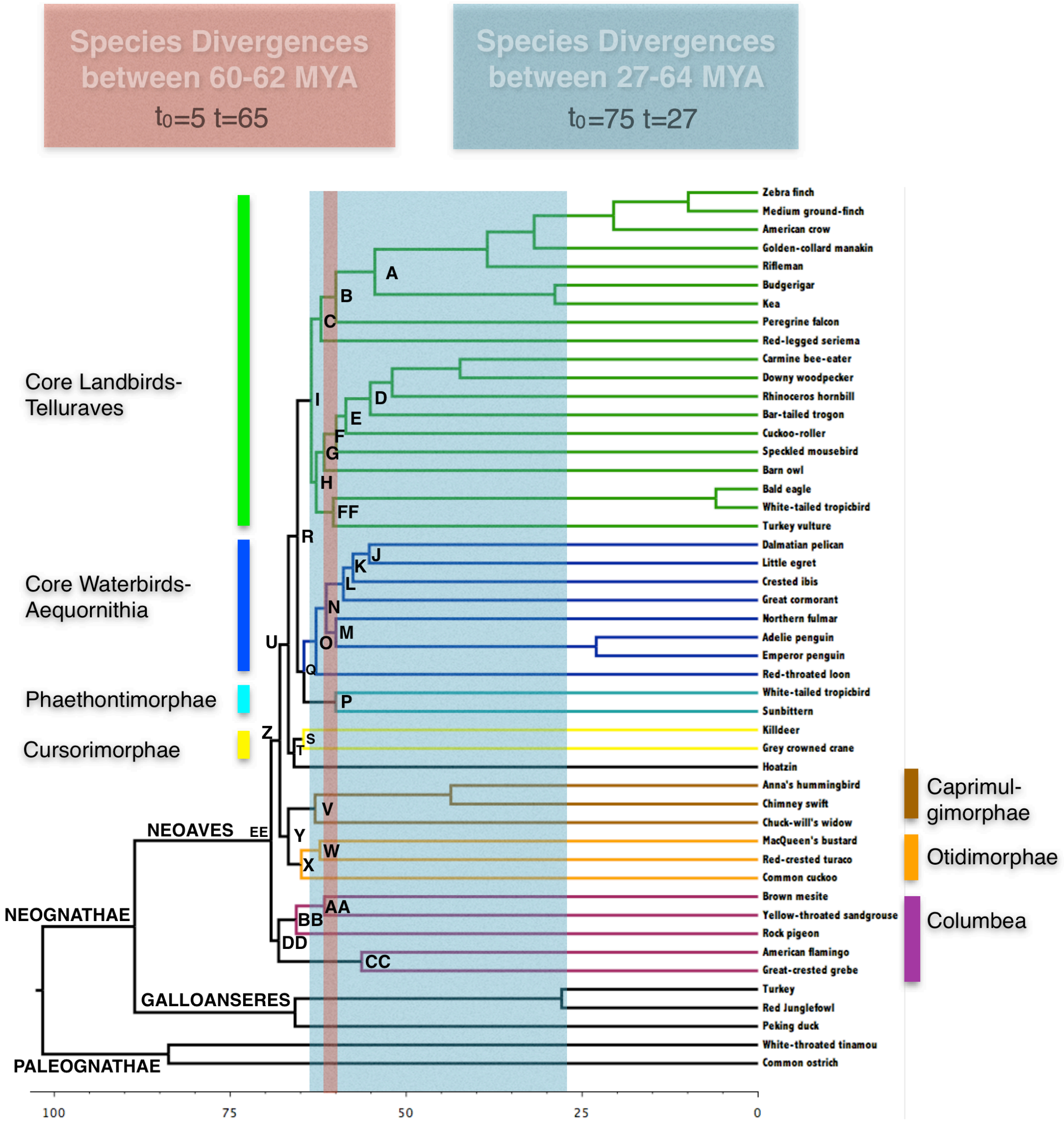
We believe that we have shown that by implementing our pipeline and partitioning data on the signal to noise ratio (Townsend et al. 2012), it is possible to improve bootstrap support and recover relationships that otherwise would require total-evidence based datasets. Independent and genomically exhaustive bodies of evidence also supported these recovered relationships. But we also have demonstrated that incongruent topologies can be found when datasets, composed of sites selected for different target eras, are used to answer the same phylogenetic questions. As with exonic data, a non-trivial number of UCE sites have rates that are too fast or too slow to resolve certain nodes. Thus using UCE sites optimized for certain epochs can improve certain bootstrap support values in a phylogenomic reconstruction and lead to findings with higher confidence.

For future studies, we recommend investigating the level of partitioning required to yield high supported, fully resolved nodes along every time span of given phylogeny. A comparison study of filtered UCE data at each important neoavian node would provide the microscale analysis that might be required to fully resolve evolutionary patterns within Neoaves, especially along those backbone nodes that have undergone a rapid radiation. The two partitions presented here are limited in the resolution they can provide and the conclusions that can be made. A more exhaustive sampling of nodes across Neoaves would be helpful. And it may be that the deep filtered dataset was too narrow a time span to yield statistically strong improvements. Likewise the shallow filtered dataset may have been too wide. However we do not dispute the importance of thoughtful data filtering. Filtering, and therefore, optimizing non-exonic phylogenomic datasets on the signal to noise ratio of a given node requiring resolution maximizes the chances of recovering relationships that would otherwise require more data. We find that our study provides evidence that filtering genomic datasets can result in the recovery of clades otherwise drowned-out by noise, thus filtering on the signal to noise ratio is a worthwhile step in the phylogenomic pipeline or as part of a sensitivity analyses.

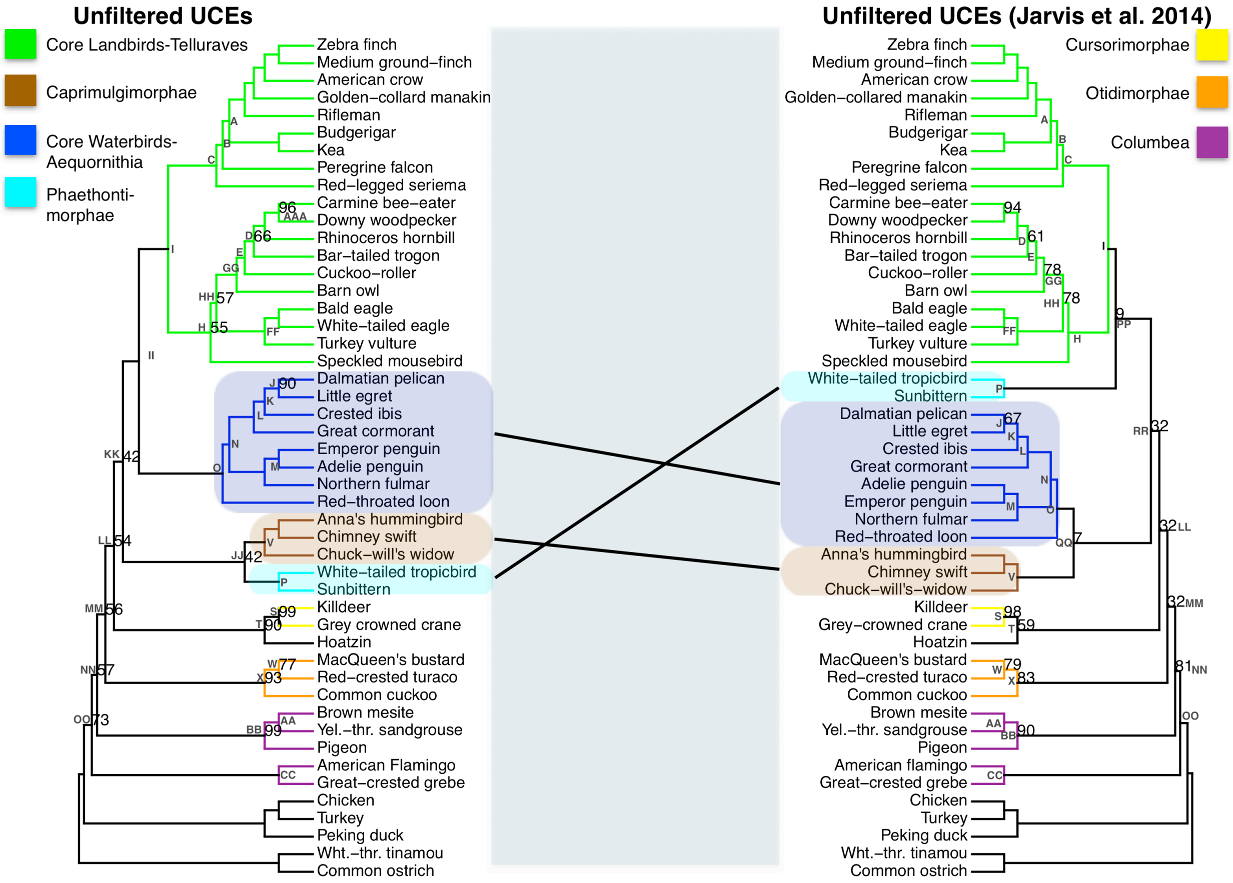
5. ACKNOWLEDGEMENTS

The authors would like to thank Ms. Tina Marcroft, Ms. Janet Buckner, Mr. David Cerny, and Dr. David A. Gold. JSS is funded by NIH grant GM053275 and NSF grant DMS 1264153.

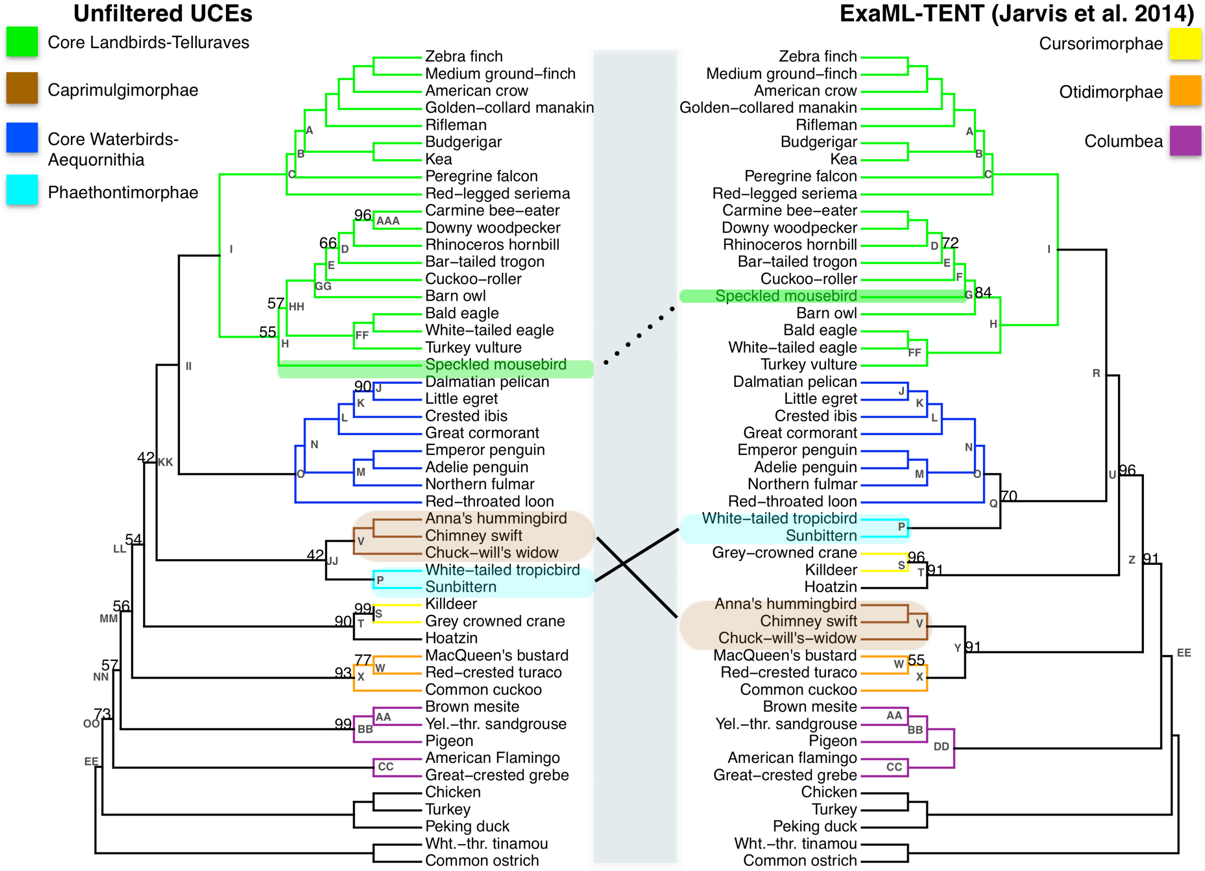
Figures



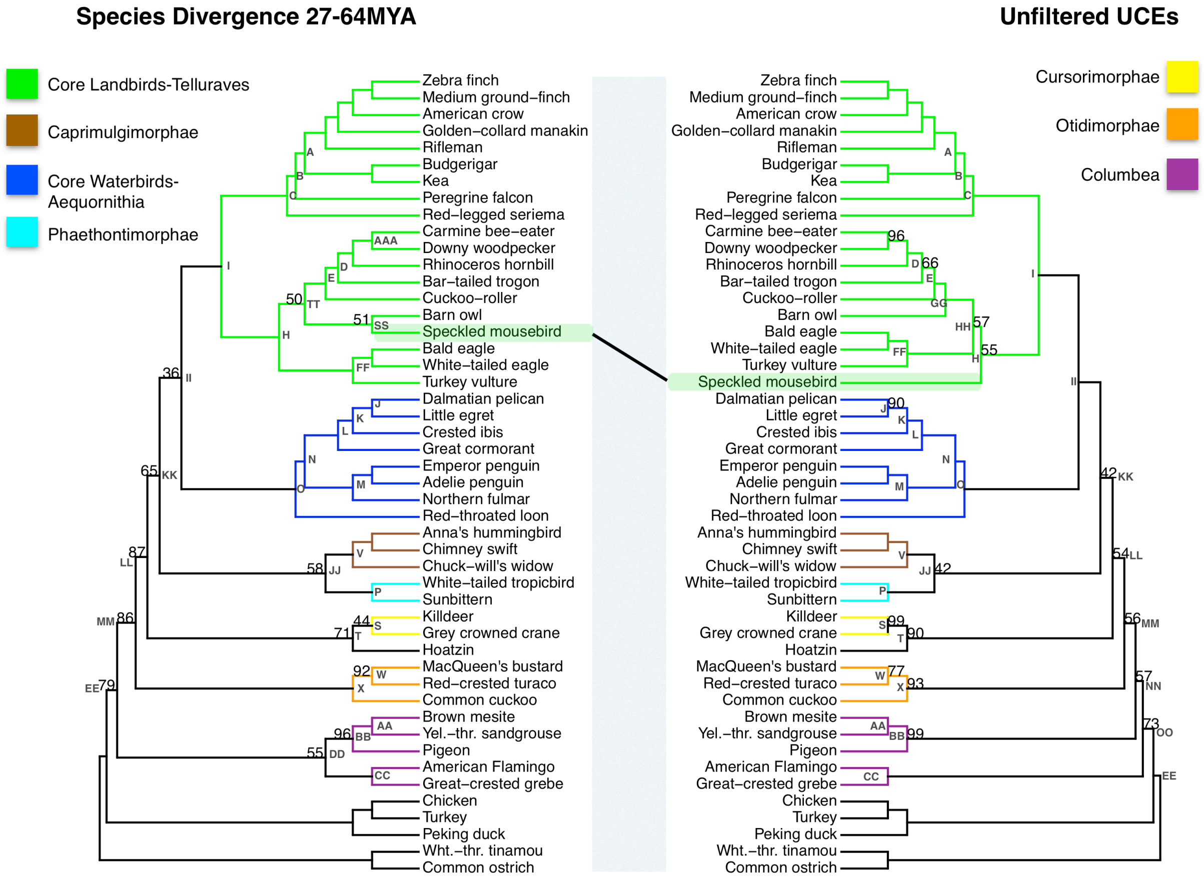
3-1. Regions of the avian phylogeny for which phylogenetic signal, noise and polytomy probabilities were calculated. The red and blue colors denote 60-62 MYA (deep) and 27-64 (shallow) MYA respectively and highlight the avian species divergences occurring these periods. The area between colored bars denotes internode length plus the average subtending branch length of each partition. The time-calibrated phylogeny is from Jarvis et al. (2014).



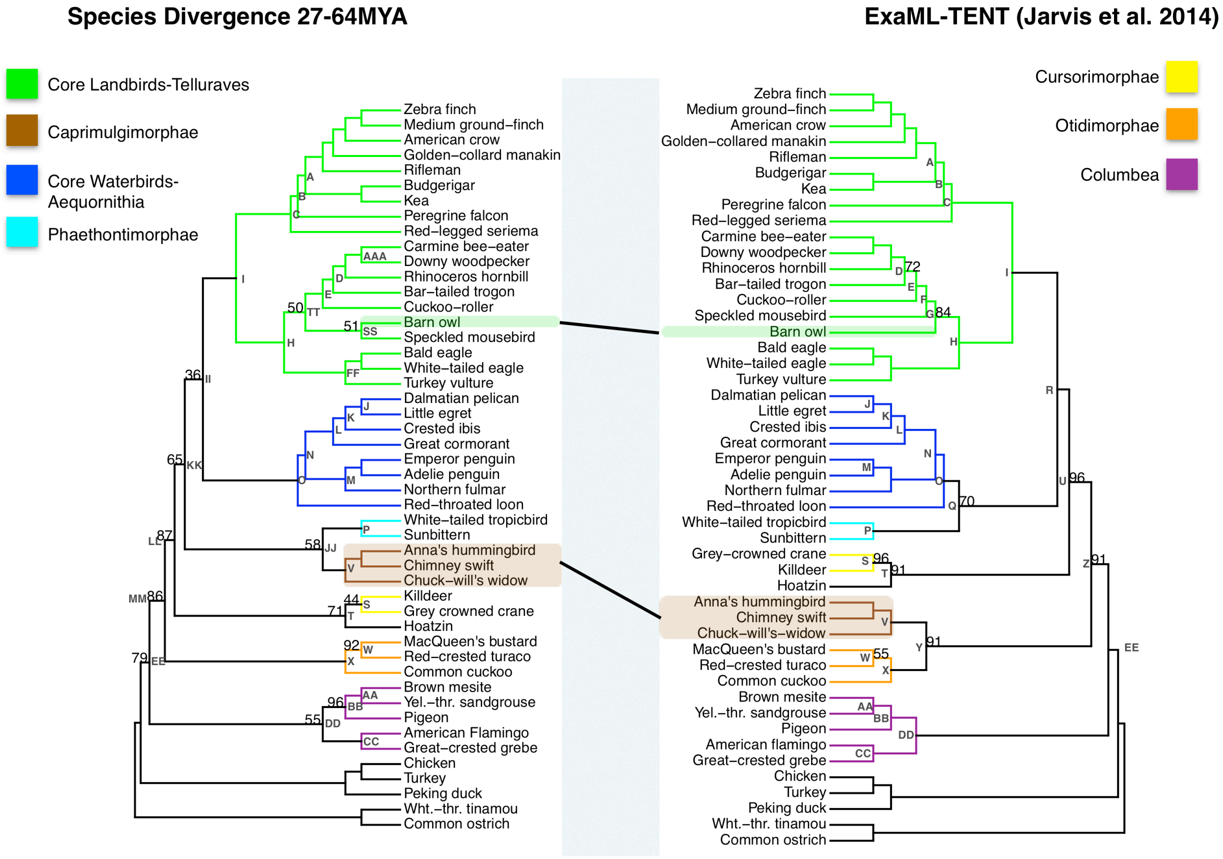
3-2. The phylogenetic reconstruction using unfiltered UCEs (left) and the UCE species tree of the 48 bird species from Jarvis et al. (2014) (right). Bootstrap support values less than 100% are shown for each internal node.



3-3. The phylogeny using unfiltered UCEs (left) and the ExaML-TENT phylogeny from Jarvis et al. (2014) (right). Bootstrap support values less than 100% are shown for each internal node.



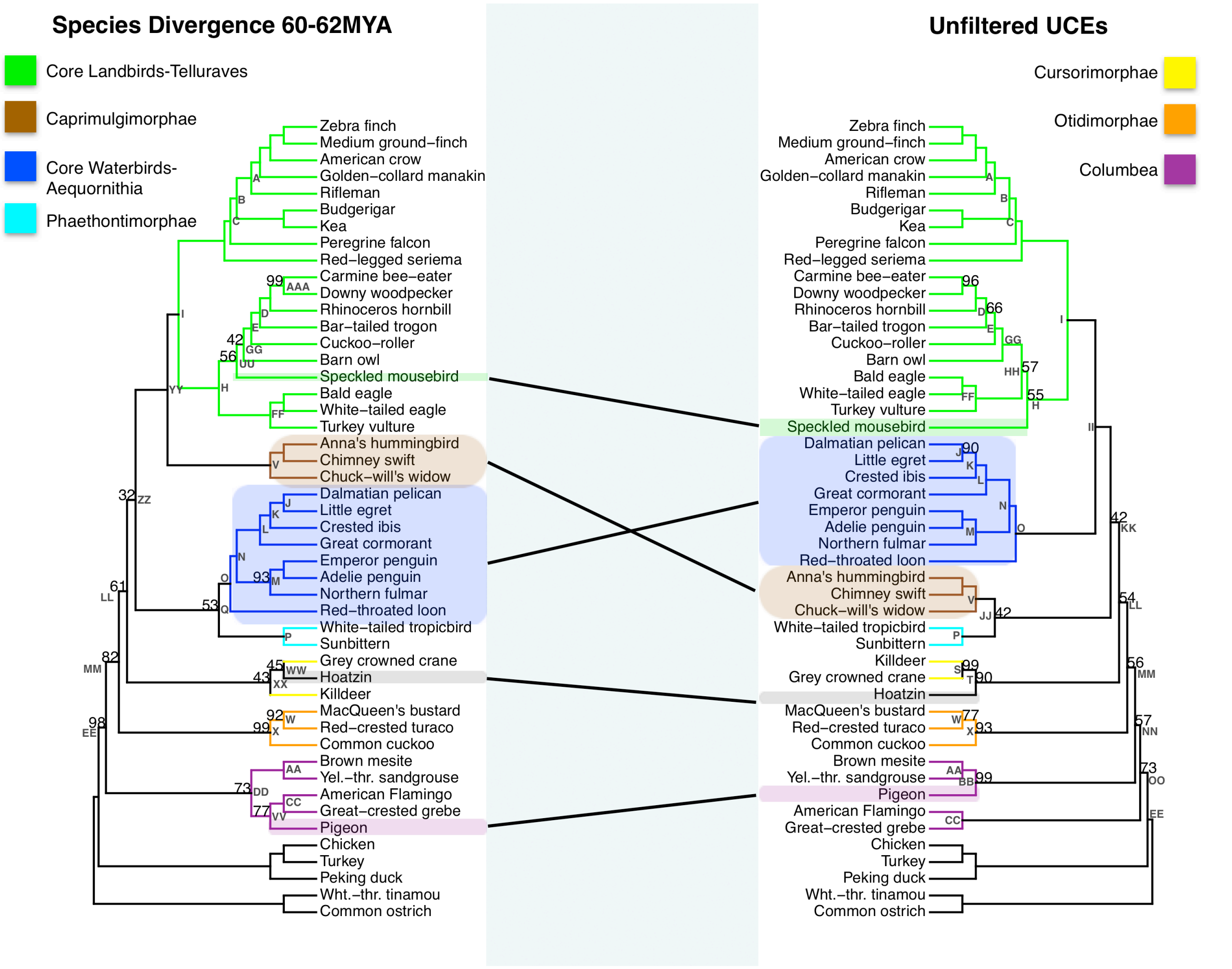
3-4. The phylogenetic reconstruction based on UCE nucleotide positions which had phylogenetic signal within in the top 20th percent of the UCE's adjusted phylogenetic signal score for species divergences between 27-64MYA (shallow filtered, left) and the phylogenetic reconstruction using all nucleotide positions (unfiltered, right). Bootstrap support values less than 100% are shown for each internal node.



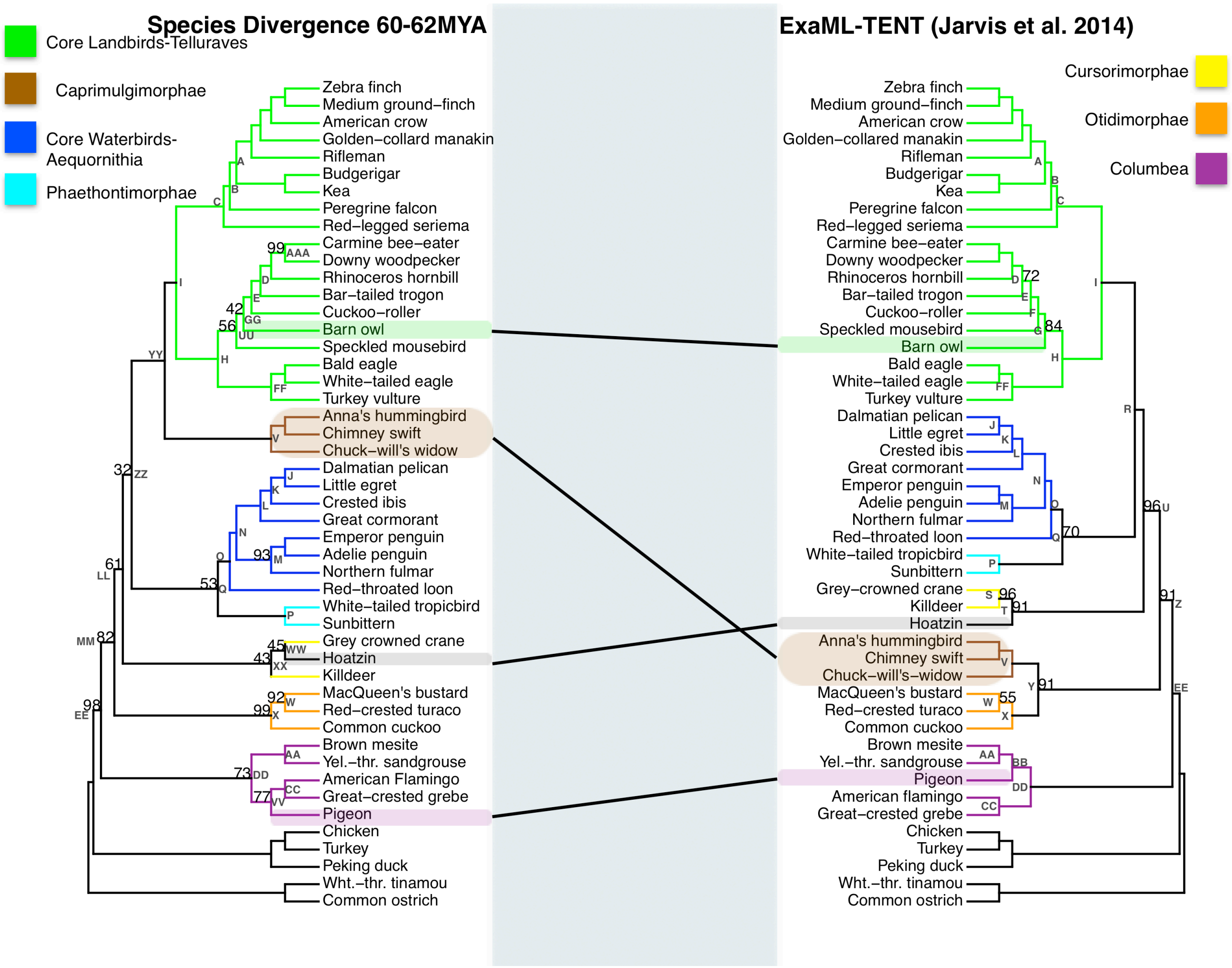
Species Divergence 27-64MYA

ExaML-TENT (Jarvis et al. 2014)

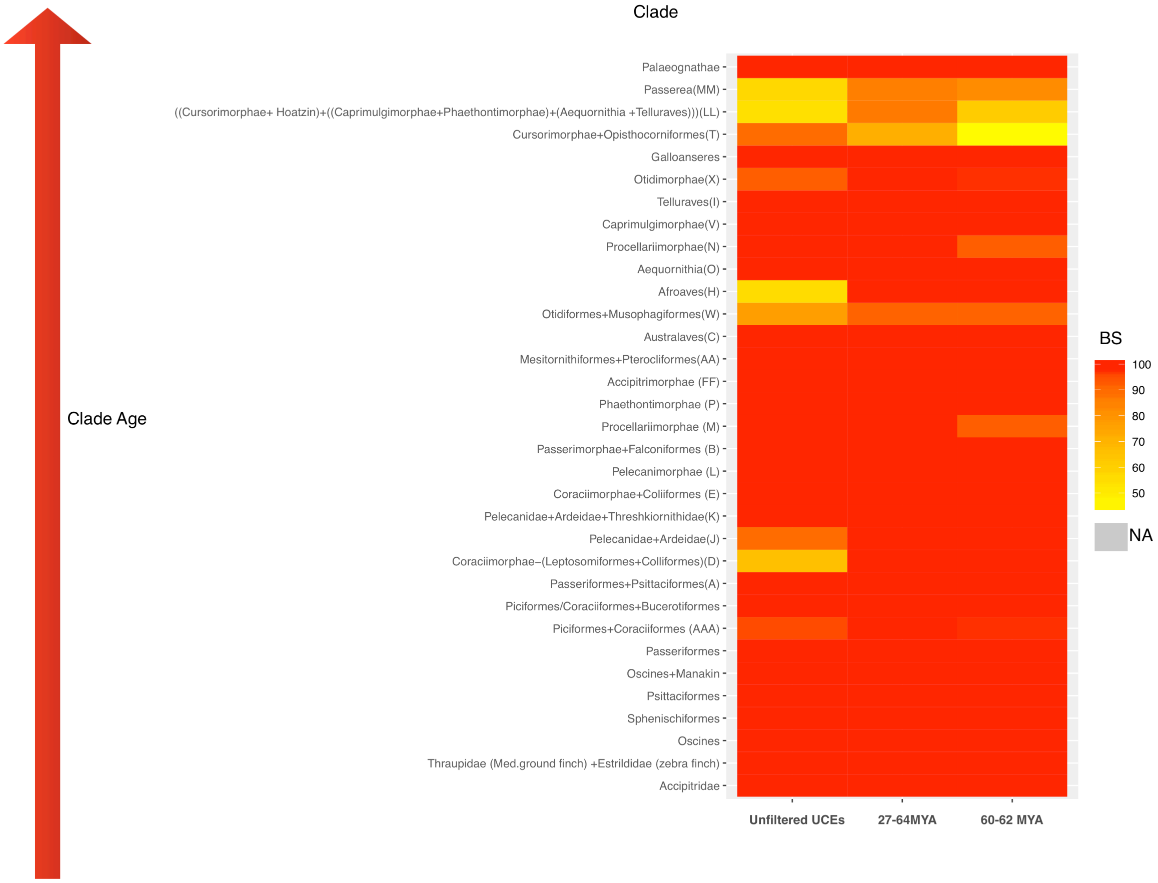
3-5. The shallow filtered UCE phylogenetic reconstruction (left) and the ExaML-TENT phylogenetic reconstruction (right). Bootstrap support values less than 100% are shown for each internal node.



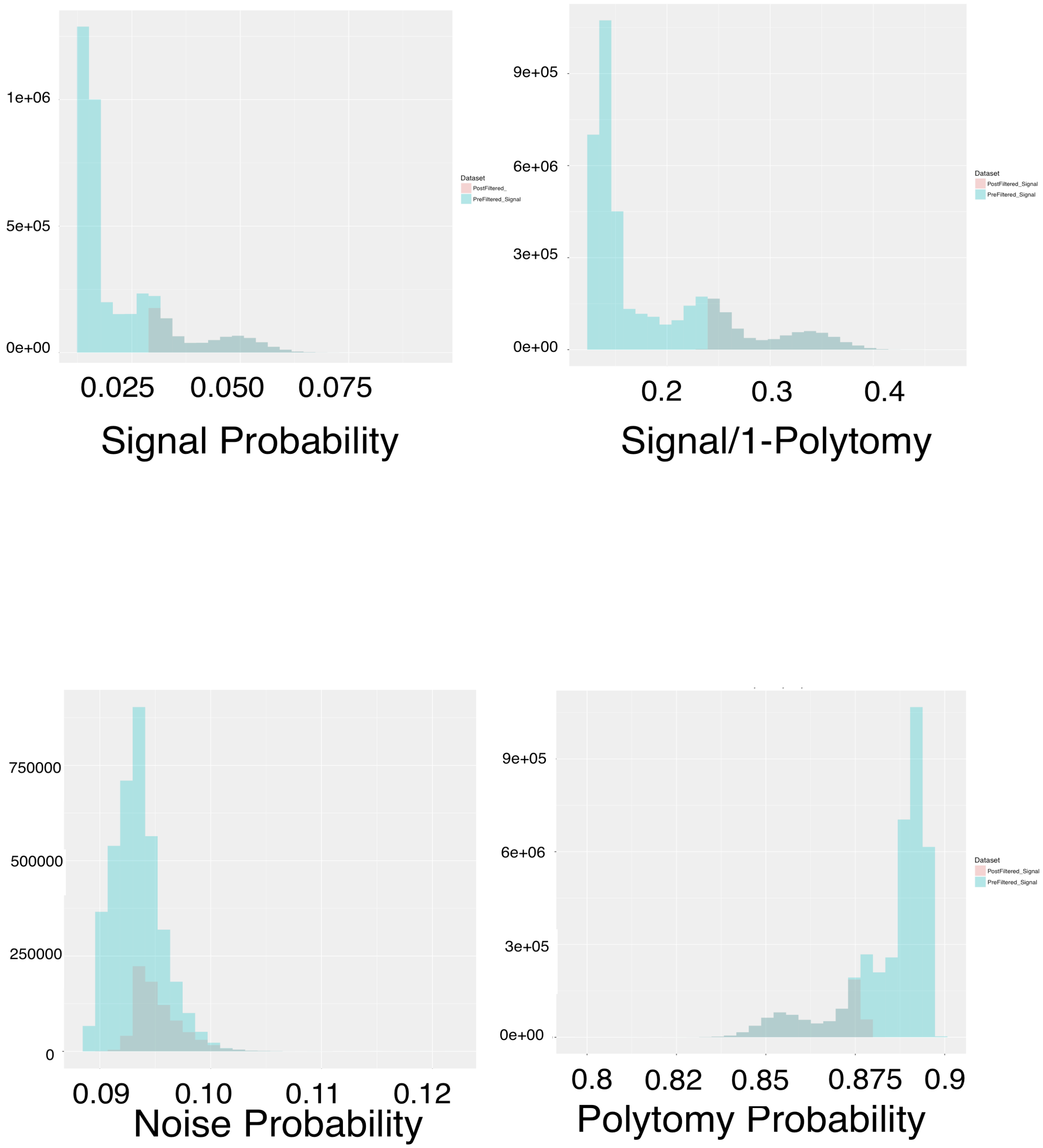
3-6. The phylogenetic reconstruction based on nucleotide UCE positions which had phylogenetic signal within in the top 20th percent of the UCE's adjusted phylogenetic signal score for species divergences between 60-62MYA (deep filtered, left) and the phylogenetic reconstruction using all nucleotide positions (unfiltered, right). Bootstrap support values less than 100% are shown for each internal node.



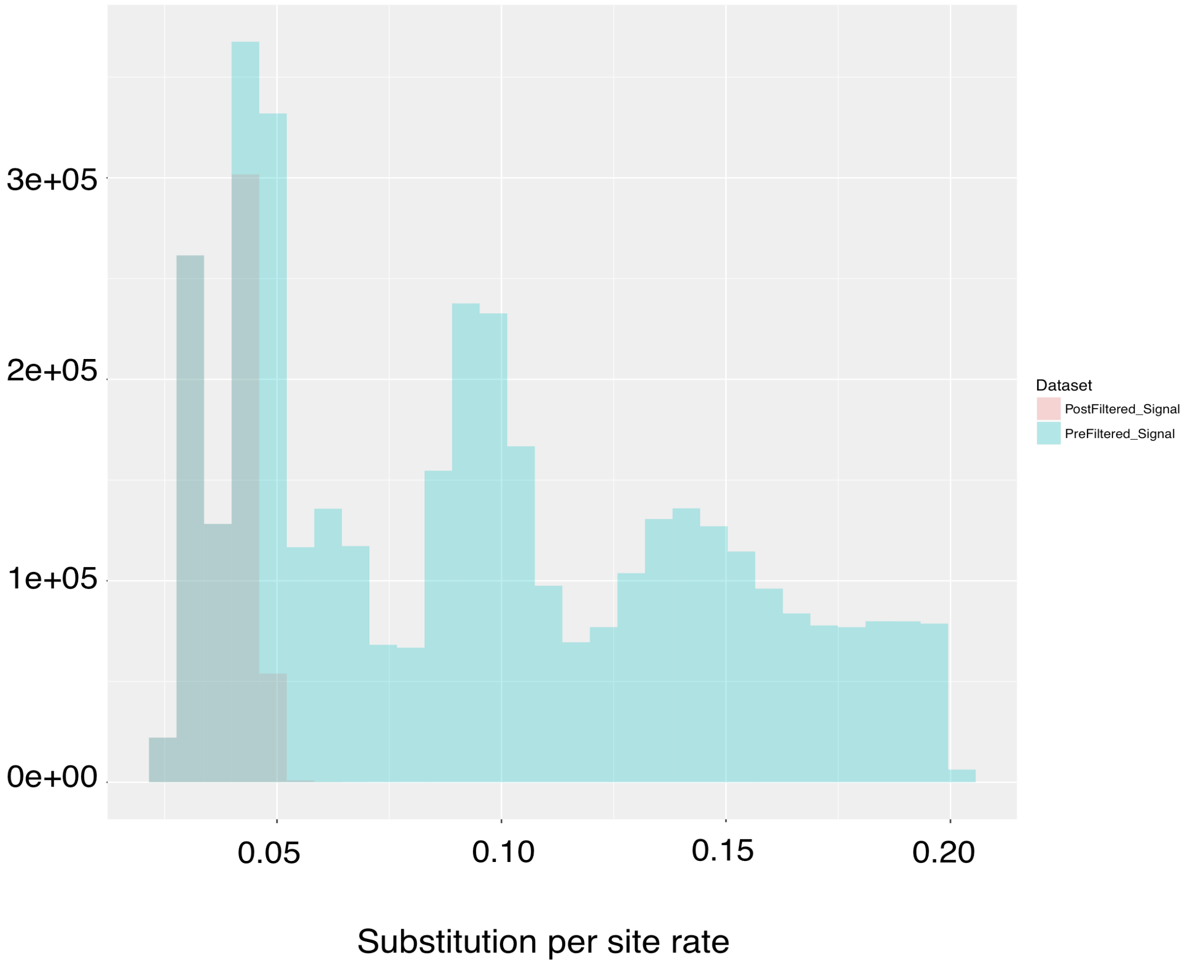
3-7. The deep filtered phylogenetic reconstruction (left) and the ExaML-TENT phylogenetic reconstruction. Bootstrap support values less than 100% are shown for each internal node.



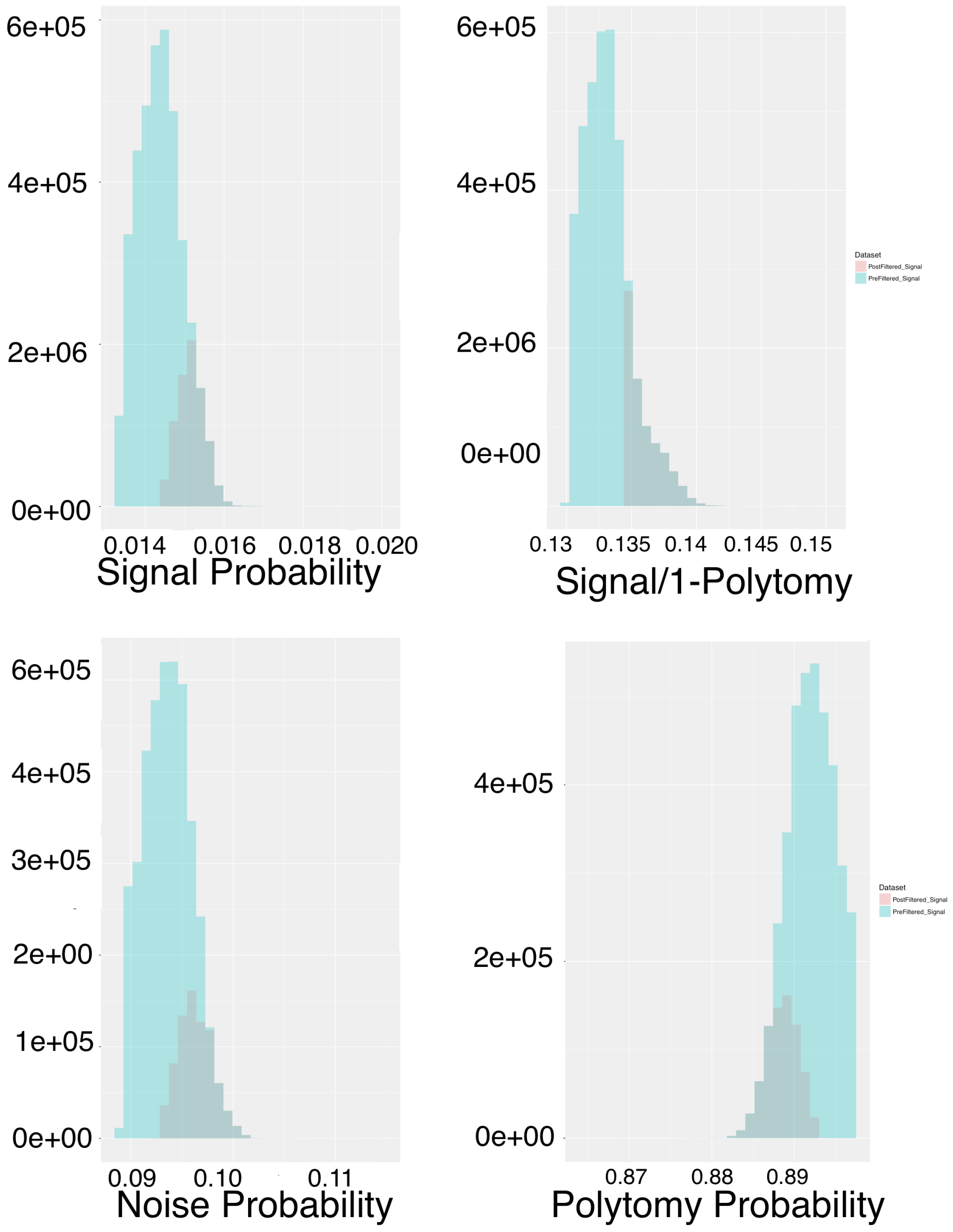
3-8. Clade support results found consistently in the unfiltered UCE phylogeny, the shallow filtered phylogeny, and deep filtered phylogeny. Each column represents a dataset; each row represents a clade that was recovered all three phylogenies.



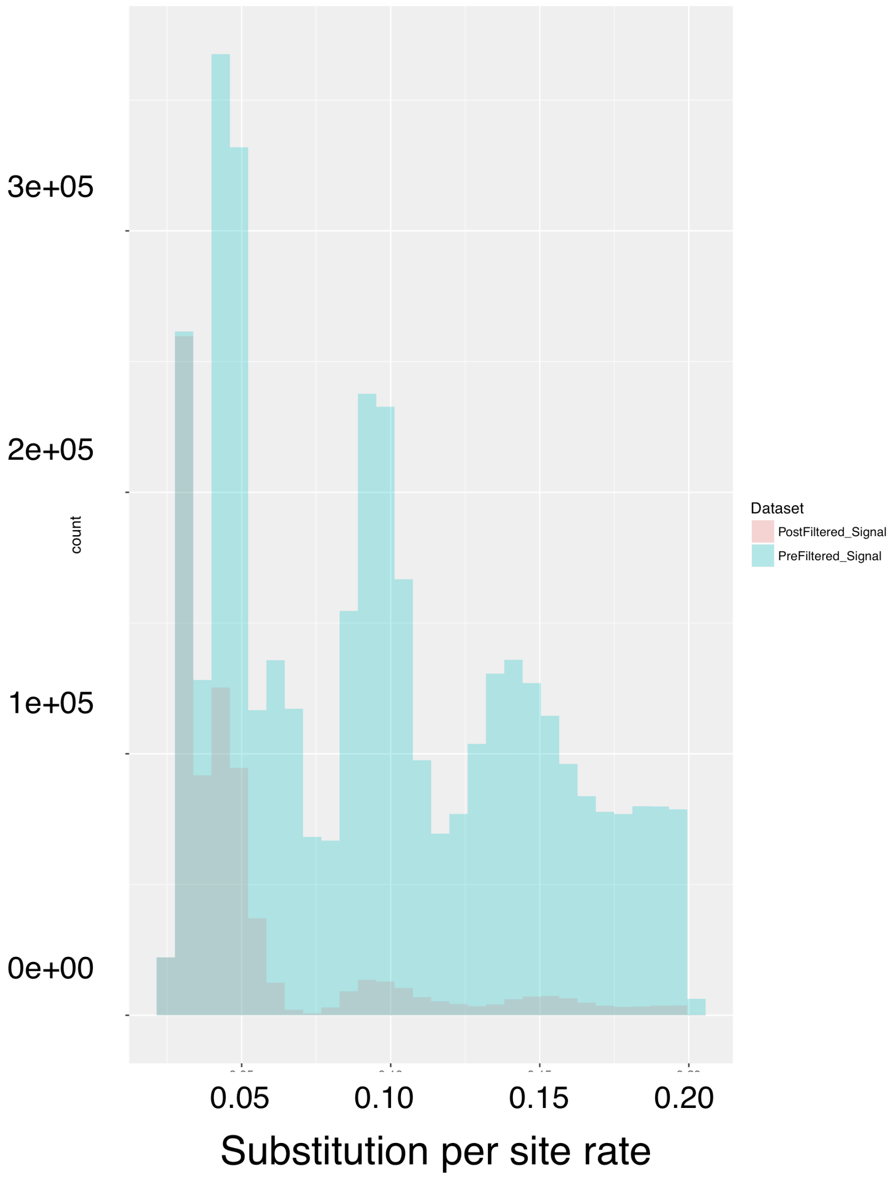
3-S1. Noise, polytomy, signal probabilities and adjusted signal (signal to noise ratio) for each UCE nucleotide before and after filtering for sites with the highest adjusted phylogenetic signal during species divergences 27-64MYA.



3-S2. Substitution rate per site for each UCE nucleotide before and after filtering for sites with the highest adjusted phylogenetic signal during species divergences 27-64MYA.

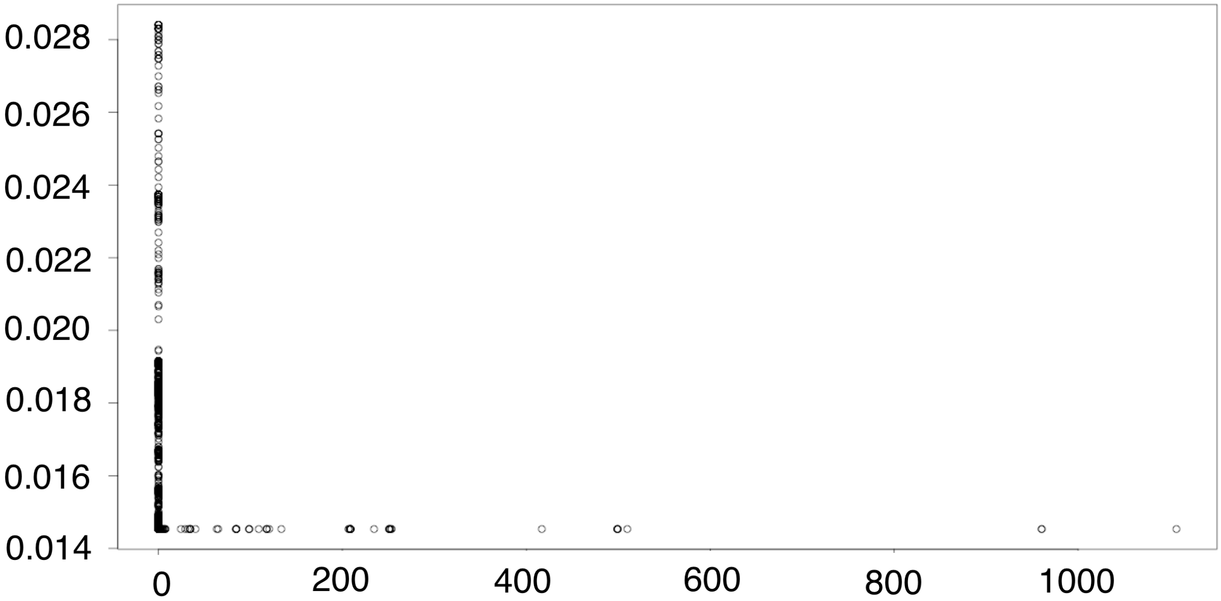


3-S3. Noise, polytomy, signal probabilities and adjusted signal (signal to noise ratio) for each UCE nucleotide before and after filtering for sites with the highest signal during species divergences 60-62MYA.

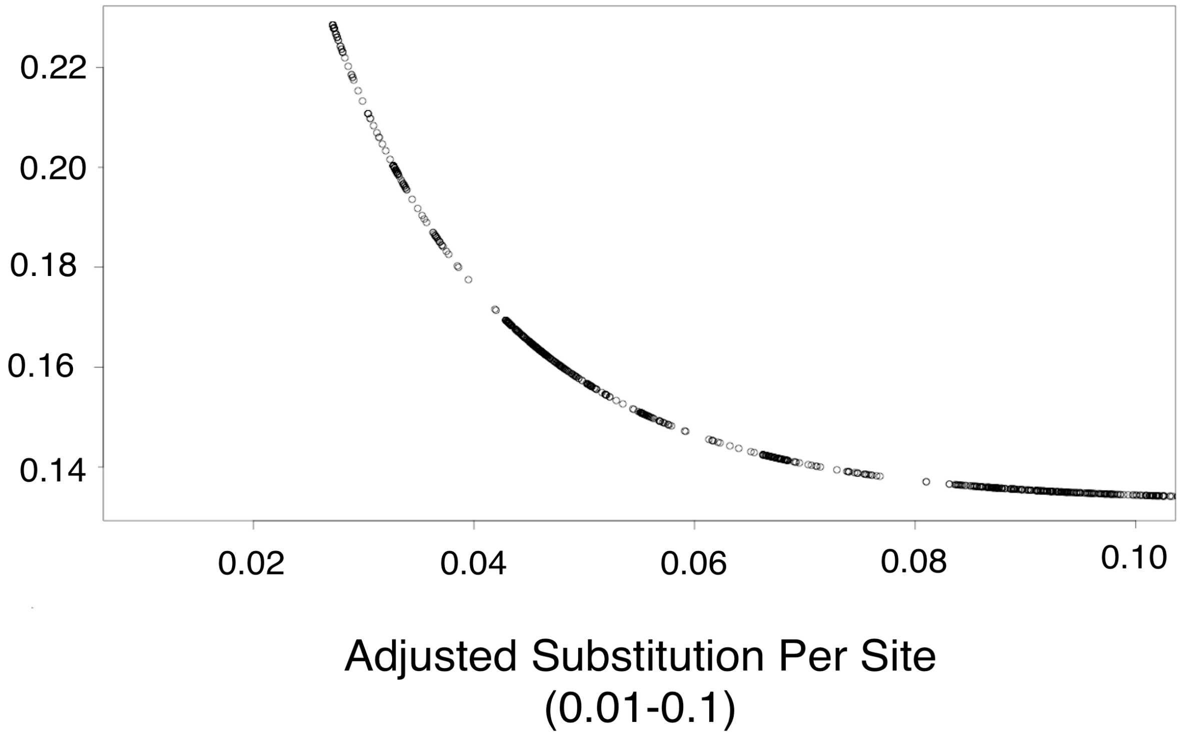


3-S4. Substitution rate per site for each UCE nucleotide before and after filtering for sites with the highest adjusted phylogenetic signal during species divergences 60-62MYA.

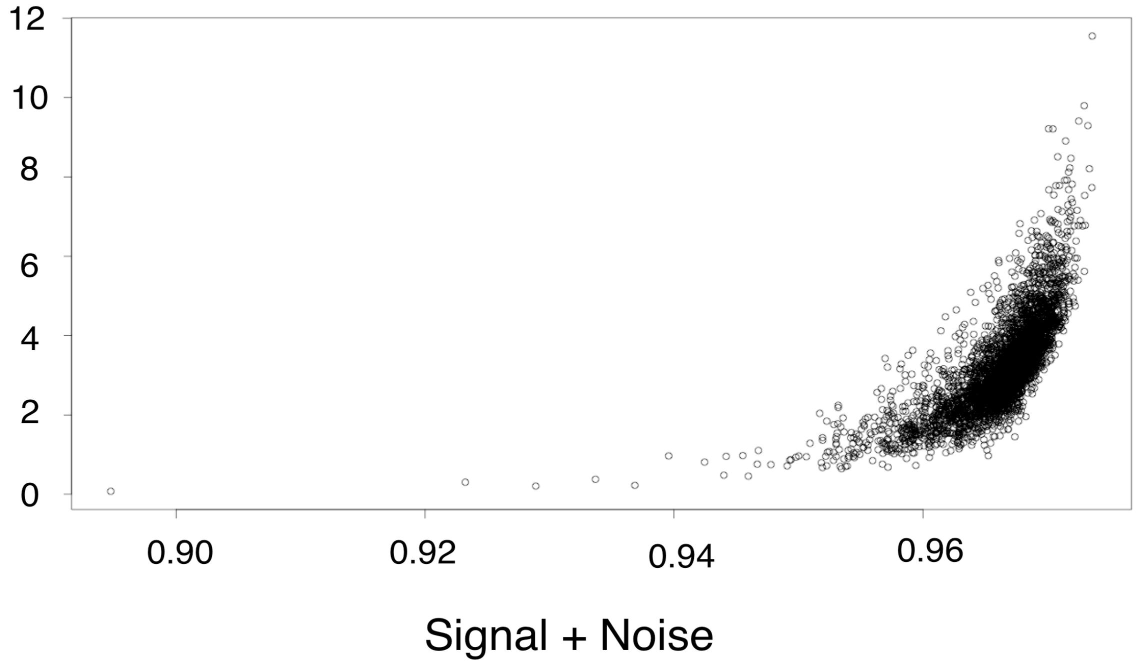
Appendix



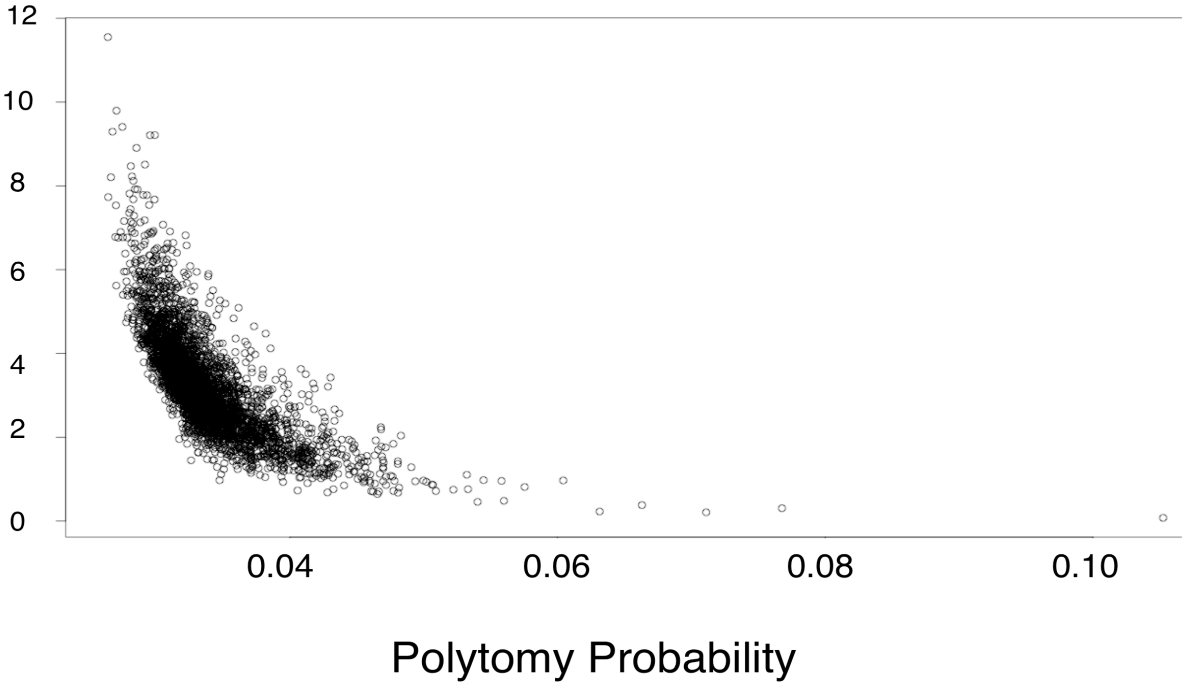
3-A1 Appendix Figure 1. Signal vs. Lambda. The x-axis is the substitution per site rate of each nucleotide in one UCE. The y-axis is the phylogenetic signal at that nucleotide. Here we see no relationship between the x and y axis.



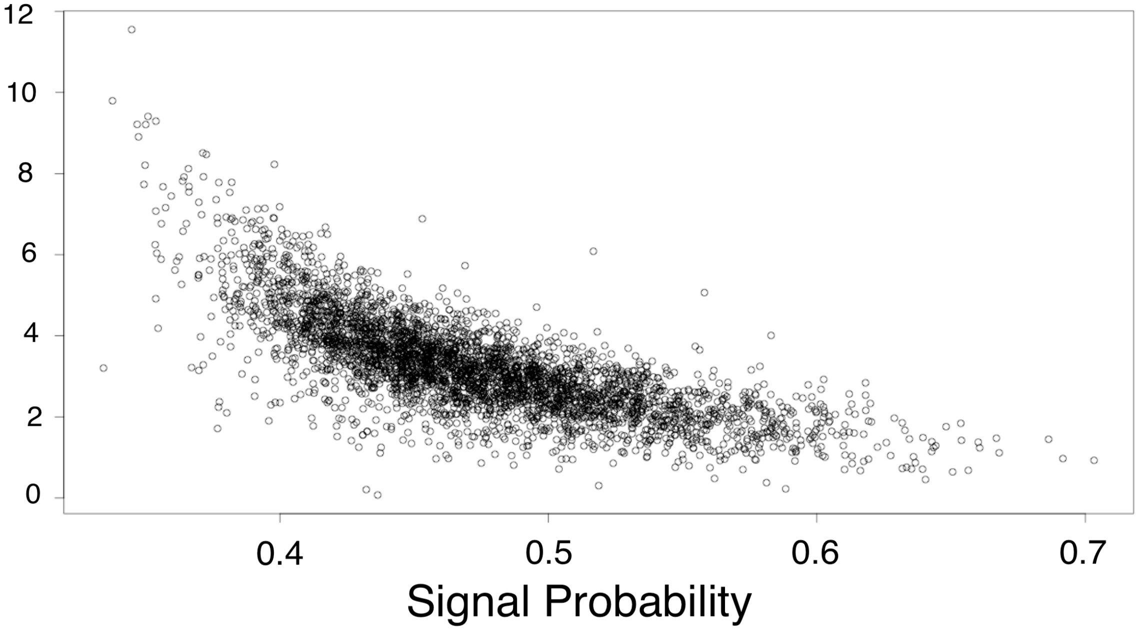
3-A2 Appendix Figure 2 Signal vs. Adjusted Lambda. The x-axis is nucleotide sites in a given UCE that fall between 0.01 and 0.1 substitutions per site rate. For these sites we observe a correlation between phylogenetic signal and substitution per site rate for these nucleotides. Thus we only kept UCEs sites whose substitution rate fell between 0.01 and 0.1 substitutions per site.



3-A3 Appendix Figure 3 Phylogenetic Informativeness vs Signal+Noise. The x-axis is the sum of phylogenetic signal and noise for a given UCE. The y-axis is the phylogenetic informativeness of that UCE for the same time span. Here we see a positive correlation between the two variables.



3-A4 Appendix Figure 4 Phylogenetic Informativeness vs. Polytomy Probability. The y-axis is the probability that a given UCE will be unable to resolve a polytomy. The y-axis is the phylogenetic informativeness of that UCE during the same time span. Here we see a negative correlation.



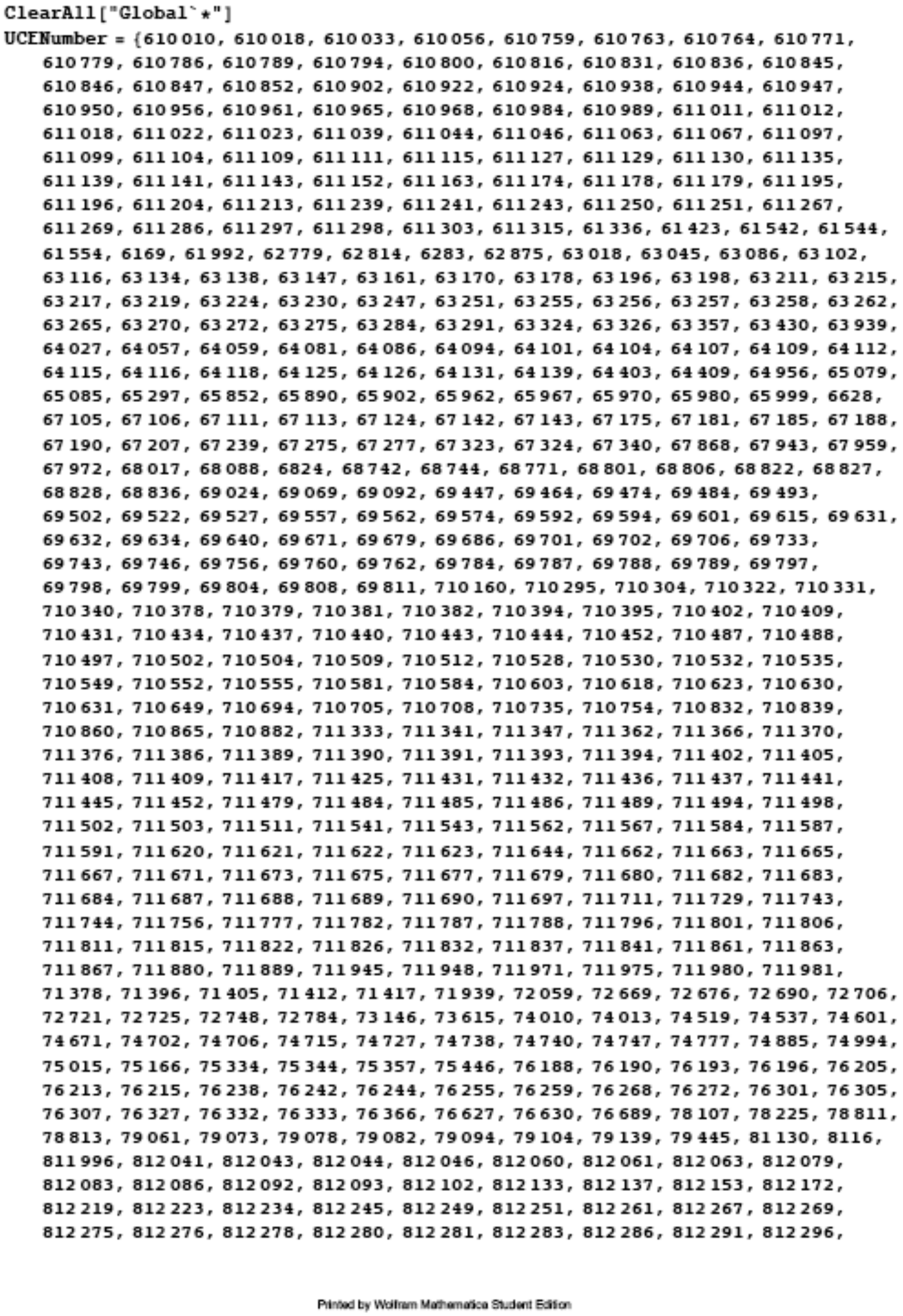
3-A5 Appendix Figure 5 Phylogenetic Informativeness vs. Signal Probability. The x-axis is probability that a given UCE has phylogenetic signal for a given time span. The y-axis is the phylogenetic informativeness of that same UCE during the same time span. Here we see a negative correlation indicating that phylogenetic informativeness and signal probability are not positively correlated.

```
#Hi! This script takes a JSON file and parses it and prints the rate, base pair switches, and base pair frequencies to an individual file.


#The following section runs a function reads in each file in the directory of your choice. This folder cannot include anything but the rate files when you start.
#setwd("~/Documents/Research/Aves/SignalNoise/Analyses/SignalNoise_TESTFOLDER/R_Terminal_Test")
#Set for 988 UCES
ListOfJSONFiles<-list.files(full.names=TRUE)

ReadJSONFile<-function(x){
  require(jsonlite)
  FileToProcess<-fromJSON(x)
  return (FileToProcess)
}
InterpretedJSONFiles3<-sapply(ListOfJSONFiles, FUN=ReadJSONFile) #use sapply instead of lapply so that each entry will have the UCE name
#lapply(InterpredJSONFiles,FUN=SubstitutionMatrixPull)


#SubstitutionMatrixFileCreation<-function(ListOfJSONFiles){
 # JSON.i<-ListOfJSONFiles
  #JSON.i<-"./1043_oryLat2.nex.rates"
  #UniqueJSONFilename<-paste(JSON.i,".switches.txt",sep="")
  #NewFiles<-file.create(UniqueJSONFilename)
  #return(NewFiles)
#}
#Create the new empty files
#lapply(ListOfJSONFiles,FUN=SubstitutionMatrixFileCreation)



#findMatrix<-function (FileList) {
 # Matrix.i<-x[[1]]$subs_matrix
#  rCA<-cat("rCA=",Matrix.i$AC,"\n")
#  rTG<-cat("rTG=",Matrix.i$GT,"\n")
#  rAG<-cat("rAG=",Matrix.i$AG,"\n")
#  rCG<-cat("rCG=",Matrix.i$CG,"\n")
#  rAT<-cat("rAT=",Matrix.i$AT,"\n")
 # rCT<-cat("rCT=",Matrix.i$CT,"\n")
#}
#testpsg<-findMatrix(FileList=InterpretedJSONFiles3)
#anothertestpsg<-lapply(InterpretedJSONFiles3,FUN=findMatrix)
```

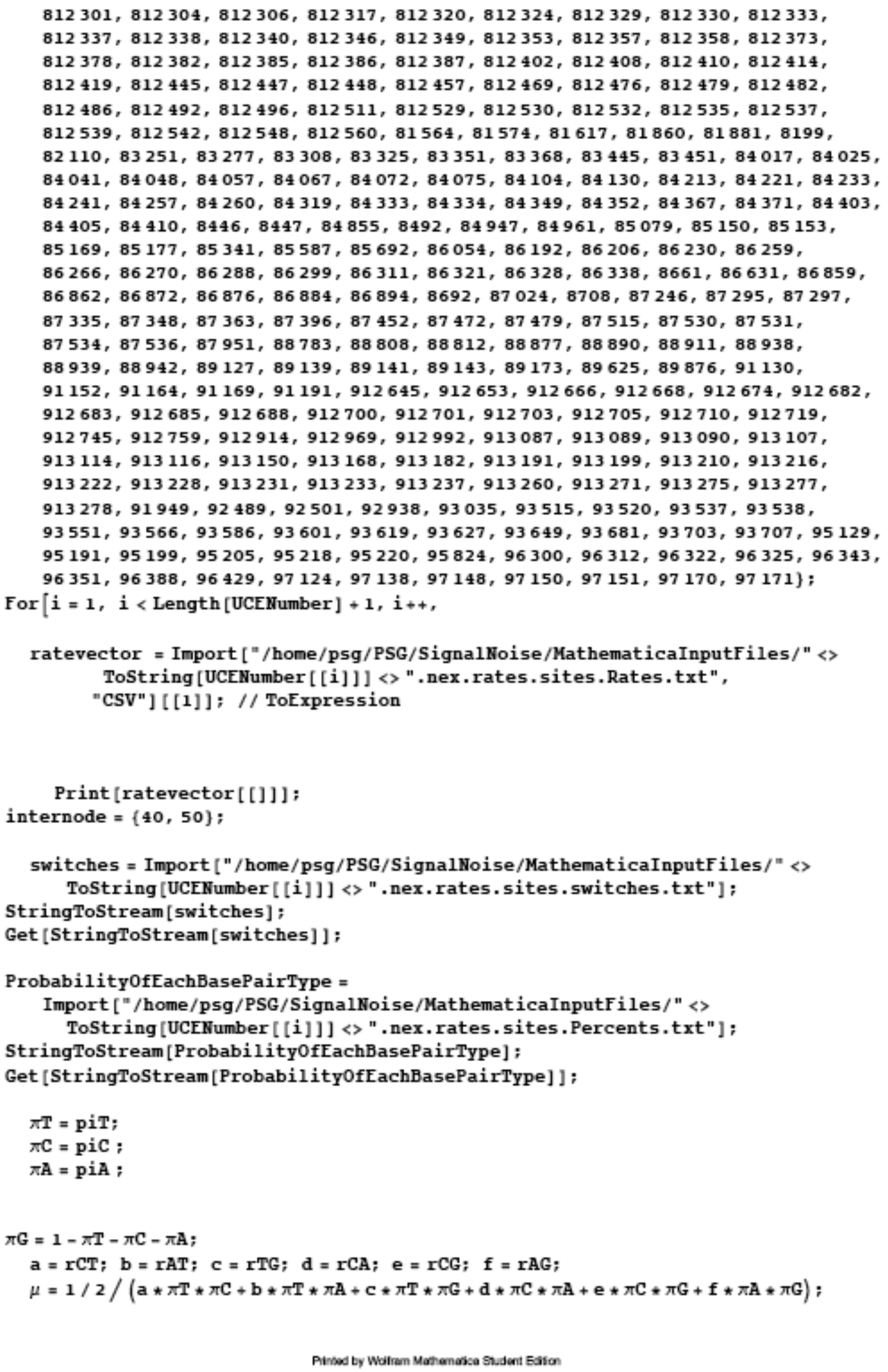
```
################################################################
#######  Substitution Matrix
#########################################################
for (i in 1:988) {
  CurrentFileName <-names(InterpretedJSONFiles3)
  EachEntryName<-
names(InterpretedJSONFiles3[names(InterpretedJSONFiles3)==CurrentFileName][])
  #CurrentFileName.i<-EachEntryName #Gets mad here
  #File2BCreated<-paste(CurrentFileName.i,".switches.txt", sep="")
  File2BCreated<-paste(EachEntryName,".switches.txt", sep="")
  file.create(File2BCreated)
}

for (i in 1:988){
  EachEntry<-InterpretedJSONFiles3[names(InterpretedJSONFiles3)==CurrentFileName][]
  #Matrix.i<-EachEntry[[1]]$subs_matrix
  #file.append(file=File2BCreated[i], append=T)
  sink(file=File2BCreated[i])
  rCA<-cat("rCA=",EachEntry[[i]]$subs_matrix$AC,"\n")
  rTG<-cat("rTG=",EachEntry[[i]]$subs_matrix$GT,"\n")
  rAG<-cat("rAG=",EachEntry[[i]]$subs_matrix$AG,"\n")
  rCG<-cat("rCG=",EachEntry[[i]]$subs_matrix$CG,"\n")
  rAT<-cat("rAT=",EachEntry[[i]]$subs_matrix$AT,"\n")
  rCT<-cat("rCT=",EachEntry[[i]]$subs_matrix$CT,"\n")
  sink()
}
###########################################################
Percents
################################################################
##########
for (i in 1:988) {
  CurrentFileName <-names(InterpretedJSONFiles3)
  EachEntryName<-
names(InterpretedJSONFiles3[names(InterpretedJSONFiles3)==CurrentFileName][])
  #CurrentFileName.i<-EachEntryName #Gets mad here
  #File2BCreated<-paste(CurrentFileName.i,".switches.txt", sep="")
  File2BCreated_Percents<-paste(EachEntryName,".Percents.txt", sep="")
  file.create(File2BCreated_Percents)
}

for (i in 1:988){
  EachEntry<-InterpretedJSONFiles3[names(InterpretedJSONFiles3)==CurrentFileName][]
  sink(file=File2BCreated_Percents[i])
  PiSymbolA<-cat("piA=",EachEntry[[i]]$freqs$A,"\n")
  PiSymbolC<-cat("piC=",EachEntry[[i]]$freqs$C,"\n")
  PiSymbolT<-cat("piT=",EachEntry[[i]]$freqs$T,"\n")
  sink()
}
################################################## Rates
#############################################################
#######################
for (i in 1:988) {
  CurrentFileName <-names(InterpretedJSONFiles3)
  EachEntryName<-
names(InterpretedJSONFiles3[names(InterpretedJSONFiles3)==CurrentFileName][])
  #CurrentFileName.i<-EachEntryName #Gets mad here
  #File2BCreated<-paste(CurrentFileName.i,".switches.txt", sep="")
  File2BCreated_Rates<-paste(EachEntryName,".Rates.txt", sep="")
  file.create(File2BCreated_Rates)
}

for (i in 1:988){
  EachEntry<-InterpretedJSONFiles3[names(InterpretedJSONFiles3)==CurrentFileName][]
  sink(file=File2BCreated_Rates[i])
  cat(EachEntry[[i]]$rates$rate,sep=",")
  sink()
}
```

3-A6 Appendix Figure 6 UCE pre-processing R code. Code can also be found at

https://github.com/PrincessG?tab=repositories

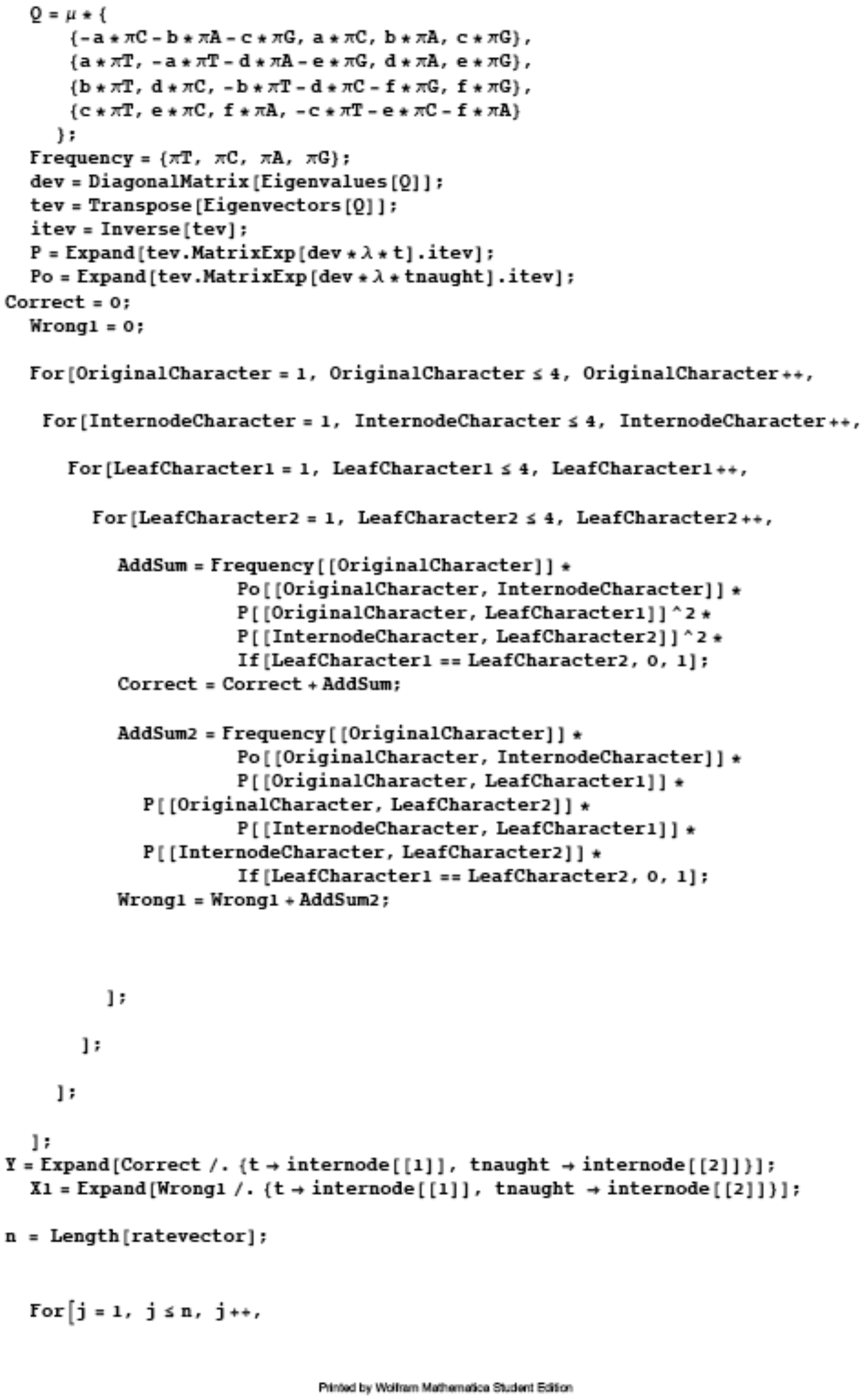


```
ClearAll["Global`*"]
UCENumber = {610 010, 610 018, 610 033, 610 056, 610 759, 610 763, 610 764, 610 771,
    610 779, 610 786, 610 789, 610 794, 610 800, 610 816, 610 831, 610 836, 610 845,
    610 846, 610 847, 610 852, 610 902, 610 922, 610 924, 610 938, 610 944, 610 947,
    610 950, 610 956, 610 961, 610 965, 610 968, 610 984, 610 989, 611 011, 611 012,
    611 018, 611 022, 611 023, 611 039, 611 044, 611 046, 611 063, 611 067, 611 097,
    611 099, 611 104, 611 109, 611 111, 611 115, 611 127, 611 129, 611 130, 611 135,
    611 139, 611 141, 611 143, 611 152, 611 163, 611 174, 611 178, 611 179, 611 195,
    611 196, 611 204, 611 213, 611 239, 611 241, 611 243, 611 250, 611 251, 611 267,
    611 269, 611 286, 611 297, 611 298, 611 303, 611 315, 61 336, 61 423, 61 542, 61 544,
    61 554, 6169, 61 992, 62 779, 62 814, 6283, 62 875, 63 018, 63 045, 63 086, 63 102,
    63 116, 63 134, 63 138, 63 147, 63 161, 63 170, 63 178, 63 196, 63 198, 63 211, 63 215,
    63 217, 63 219, 63 224, 63 230, 63 247, 63 251, 63 255, 63 256, 63 257, 63 258, 63 262,
    63 265, 63 270, 63 272, 63 275, 63 284, 63 291, 63 324, 63 326, 63 357, 63 430, 63 939,
    64 027, 64 057, 64 059, 64 081, 64 086, 64 094, 64 101, 64 104, 64 107, 64 109, 64 112,
    64 115, 64 116, 64 118, 64 125, 64 126, 64 131, 64 139, 64 403, 64 409, 64 956, 65 079,
    65 085, 65 297, 65 852, 65 890, 65 902, 65 962, 65 967, 65 970, 65 980, 65 999, 6628,
    67 105, 67 106, 67 111, 67 113, 67 124, 67 142, 67 143, 67 175, 67 181, 67 185, 67 188,
    67 190, 67 207, 67 239, 67 275, 67 277, 67 323, 67 324, 67 340, 67 868, 67 943, 67 959,
    67 972, 68 017, 68 088, 6824, 68 742, 68 744, 68 771, 68 801, 68 806, 68 822, 68 827,
    68 828, 68 836, 69 024, 69 069, 69 092, 69 447, 69 464, 69 474, 69 484, 69 493,
    69 502, 69 522, 69 527, 69 557, 69 562, 69 574, 69 592, 69 594, 69 601, 69 615, 69 631,
    69 632, 69 634, 69 640, 69 671, 69 679, 69 686, 69 701, 69 702, 69 706, 69 733,
    69 743, 69 746, 69 756, 69 760, 69 762, 69 784, 69 787, 69 788, 69 789, 69 797,
    69 798, 69 799, 69 804, 69 808, 69 811, 710 160, 710 295, 710 304, 710 322, 710 331,
    710 340, 710 378, 710 379, 710 381, 710 382, 710 394, 710 395, 710 402, 710 409,
    710 431, 710 434, 710 437, 710 440, 710 443, 710 444, 710 452, 710 487, 710 488,
    710 497, 710 502, 710 504, 710 509, 710 512, 710 528, 710 530, 710 532, 710 535,
    710 549, 710 552, 710 555, 710 581, 710 584, 710 603, 710 618, 710 623, 710 630,
    710 631, 710 649, 710 694, 710 705, 710 708, 710 735, 710 754, 710 832, 710 839,
    710 860, 710 865, 710 882, 711 333, 711 341, 711 347, 711 362, 711 366, 711 370,
    711 376, 711 386, 711 389, 711 390, 711 391, 711 393, 711 394, 711 402, 711 405,
    711 408, 711 409, 711 417, 711 425, 711 431, 711 432, 711 436, 711 437, 711 441,
    711 445, 711 452, 711 479, 711 484, 711 485, 711 486, 711 489, 711 494, 711 498,
    711 502, 711 503, 711 511, 711 541, 711 543, 711 562, 711 567, 711 584, 711 587,
    711 591, 711 620, 711 621, 711 622, 711 623, 711 644, 711 662, 711 663, 711 665,
    711 667, 711 671, 711 673, 711 675, 711 677, 711 679, 711 680, 711 682, 711 683,
    711 684, 711 687, 711 688, 711 689, 711 690, 711 697, 711 711, 711 729, 711 743,
    711 744, 711 756, 711 777, 711 782, 711 787, 711 788, 711 796, 711 801, 711 806,
    711 811, 711 815, 711 822, 711 826, 711 832, 711 837, 711 841, 711 861, 711 863,
    711 867, 711 880, 711 889, 711 945, 711 948, 711 971, 711 975, 711 980, 711 981,
    71 378, 71 396, 71 405, 71 412, 71 417, 71 939, 72 059, 72 669, 72 676, 72 690, 72 706,
    72 721, 72 725, 72 748, 72 784, 73 146, 73 615, 74 010, 74 013, 74 519, 74 537, 74 601,
    74 671, 74 702, 74 706, 74 715, 74 727, 74 738, 74 740, 74 747, 74 777, 74 885, 74 994,
    75 015, 75 166, 75 334, 75 344, 75 357, 75 446, 76 188, 76 190, 76 193, 76 196, 76 205,
    76 213, 76 215, 76 238, 76 242, 76 244, 76 255, 76 259, 76 268, 76 272, 76 301, 76 305,
    76 307, 76 327, 76 332, 76 333, 76 366, 76 627, 76 630, 76 689, 78 107, 78 225, 78 811,
    78 813, 79 061, 79 073, 79 078, 79 082, 79 094, 79 104, 79 139, 79 445, 81 130, 8116,
    811 996, 812 041, 812 043, 812 044, 812 046, 812 060, 812 061, 812 063, 812 079,
    812 083, 812 086, 812 092, 812 093, 812 102, 812 133, 812 137, 812 153, 812 172,
    812 219, 812 223, 812 234, 812 245, 812 249, 812 251, 812 261, 812 267, 812 269,
    812 275, 812 276, 812 278, 812 280, 812 281, 812 283, 812 286, 812 291, 812 296,
```



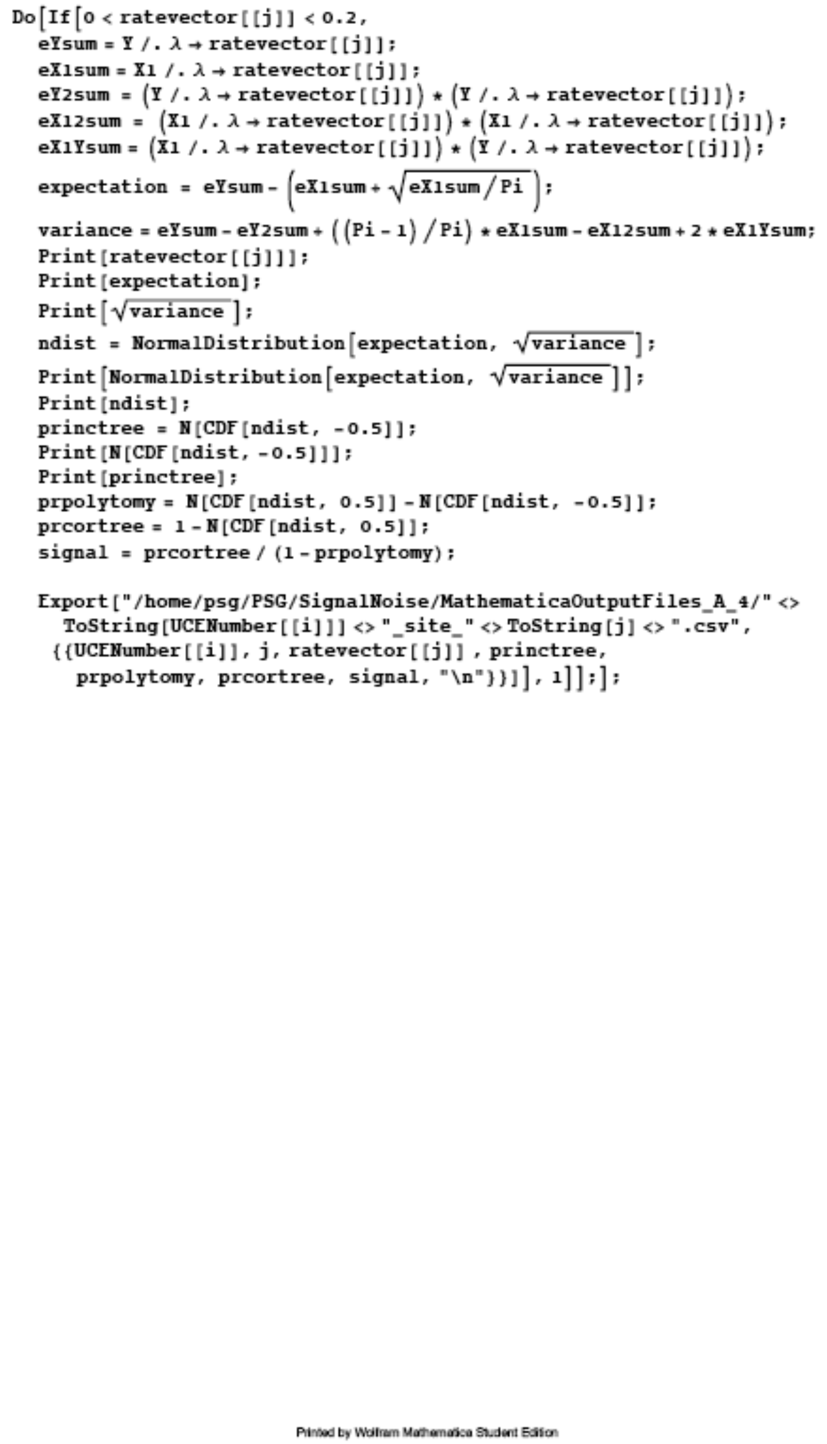
```
    812 301, 812 304, 812 306, 812 317, 812 320, 812 324, 812 329, 812 330, 812 333,
    812 337, 812 338, 812 340, 812 346, 812 349, 812 353, 812 357, 812 358, 812 373,
    812 378, 812 382, 812 385, 812 386, 812 387, 812 402, 812 408, 812 410, 812 414,
    812 419, 812 445, 812 447, 812 448, 812 457, 812 469, 812 476, 812 479, 812 482,
    812 486, 812 492, 812 496, 812 511, 812 529, 812 530, 812 532, 812 535, 812 537,
    812 539, 812 542, 812 548, 812 560, 81 564, 81 574, 81 617, 81 860, 81 881, 8199,
    82 110, 83 251, 83 277, 83 308, 83 325, 83 351, 83 368, 83 445, 83 451, 84 017, 84 025,
    84 041, 84 048, 84 057, 84 067, 84 072, 84 075, 84 104, 84 130, 84 213, 84 221, 84 233,
    84 241, 84 257, 84 260, 84 319, 84 333, 84 334, 84 349, 84 352, 84 367, 84 371, 84 403,
    84 405, 84 410, 8446, 8447, 84 855, 8492, 84 947, 84 961, 85 079, 85 150, 85 153,
    85 169, 85 177, 85 341, 85 587, 85 692, 86 054, 86 192, 86 206, 86 230, 86 259,
    86 266, 86 270, 86 288, 86 299, 86 311, 86 321, 86 328, 86 338, 8661, 86 631, 86 859,
    86 862, 86 872, 86 876, 86 884, 86 894, 8692, 87 024, 8708, 87 246, 87 295, 87 297,
    87 335, 87 348, 87 363, 87 396, 87 452, 87 472, 87 479, 87 515, 87 530, 87 531,
    87 534, 87 536, 87 951, 88 783, 88 808, 88 812, 88 877, 88 890, 88 911, 88 938,
    88 939, 88 942, 89 127, 89 139, 89 141, 89 143, 89 173, 89 625, 89 876, 91 130,
    91 152, 91 164, 91 169, 91 191, 912 645, 912 653, 912 666, 912 668, 912 674, 912 682,
    912 683, 912 685, 912 688, 912 700, 912 701, 912 703, 912 705, 912 710, 912 719,
    912 745, 912 759, 912 914, 912 969, 912 992, 913 087, 913 089, 913 090, 913 107,
    913 114, 913 116, 913 150, 913 168, 913 182, 913 191, 913 199, 913 210, 913 216,
    913 222, 913 228, 913 231, 913 233, 913 237, 913 260, 913 271, 913 275, 913 277,
    913 278, 91 949, 92 489, 92 501, 92 938, 93 035, 93 515, 93 520, 93 537, 93 538,
    93 551, 93 566, 93 586, 93 601, 93 619, 93 627, 93 649, 93 681, 93 703, 93 707, 95 129,
    95 191, 95 199, 95 205, 95 218, 95 220, 95 824, 96 300, 96 312, 96 322, 96 325, 96 343,
    96 351, 96 388, 96 429, 97 124, 97 138, 97 148, 97 150, 97 151, 97 170, 97 171};
For[i = 1, i < Length[UCENumber] + 1, i++,

  ratevector = Import["/home/psg/PSG/SignalNoise/MathematicaInputFiles/" <>
        ToString[UCENumber[[i]]] <> ".nex.rates.sites.Rates.txt",
       "CSV"][[1]]; // ToExpression



    Print[ratevector[[]]];
internode = {40, 50};

  switches = Import["/home/psg/PSG/SignalNoise/MathematicaInputFiles/" <>
     ToString[UCENumber[[i]]] <> ".nex.rates.sites.switches.txt"];
StringToStream[switches];
Get[StringToStream[switches]];

ProbabilityOfEachBasePairType =
   Import["/home/psg/PSG/SignalNoise/MathematicaInputFiles/" <>
     ToString[UCENumber[[i]]] <> ".nex.rates.sites.Percents.txt"];
StringToStream[ProbabilityOfEachBasePairType];
Get[StringToStream[ProbabilityOfEachBasePairType]];

  πT = piT;
  πC = piC ;
  πA = piA ;


πG = 1 - πT - πC - πA;
  a = rCT; b = rAT; c = rTG; d = rCA; e = rCG; f = rAG;
  μ = 1 / 2 / (a * πT * πC + b * πT * πA + c * πT * πG + d * πC * πA + e * πC * πG + f * πA * πG);
```

Printed by Wolfram Mathematica Student Edition



```
  Q = μ * {
      {-a * πC - b * πA - c * πG, a * πC, b * πA, c * πG},
      {a * πT, -a * πT - d * πA - e * πG, d * πA, e * πG},
      {b * πT, d * πC, -b * πT - d * πC - f * πG, f * πG},
      {c * πT, e * πC, f * πA, -c * πT - e * πC - f * πA}
     };
  Frequency = {πT, πC, πA, πG};
  dev = DiagonalMatrix[Eigenvalues[Q]];
  tev = Transpose[Eigenvectors[Q]];
  itev = Inverse[tev];
  P = Expand[tev.MatrixExp[dev * λ * t].itev];
  Po = Expand[tev.MatrixExp[dev * λ * tnaught].itev];
Correct = 0;
  Wrong1 = 0;

  For[OriginalCharacter = 1, OriginalCharacter ≤ 4, OriginalCharacter ++,

   For[InternodeCharacter = 1, InternodeCharacter ≤ 4, InternodeCharacter ++,

     For[LeafCharacter1 = 1, LeafCharacter1 ≤ 4, LeafCharacter1 ++,

       For[LeafCharacter2 = 1, LeafCharacter2 ≤ 4, LeafCharacter2 ++,

         AddSum = Frequency[[OriginalCharacter]] *
                   Po[[OriginalCharacter, InternodeCharacter]] *
                   P[[OriginalCharacter, LeafCharacter1]] ^ 2 *
                   P[[InternodeCharacter, LeafCharacter2]] ^ 2 *
                   If[LeafCharacter1 == LeafCharacter2, 0, 1];
         Correct = Correct + AddSum;

         AddSum2 = Frequency[[OriginalCharacter]] *
                   Po[[OriginalCharacter, InternodeCharacter]] *
                   P[[OriginalCharacter, LeafCharacter1]] *
           P[[OriginalCharacter, LeafCharacter2]] *
                   P[[InternodeCharacter, LeafCharacter1]] *
           P[[InternodeCharacter, LeafCharacter2]] *
                   If[LeafCharacter1 == LeafCharacter2, 0, 1];
         Wrong1 = Wrong1 + AddSum2;



        ];

      ];

    ];

  ];
Y = Expand[Correct /. {t → internode[[1]], tnaught → internode[[2]]}];
  X1 = Expand[Wrong1 /. {t → internode[[1]], tnaught → internode[[2]]}];

n = Length[ratevector];


  For[j = 1, j ≤ n, j ++,
```

Printed by Wolfram Mathematica Student Edition



```
Do[If[0 < ratevector[[j]] < 0.2,
  eYsum = Y /. λ → ratevector[[j]];
  eX1sum = X1 /. λ → ratevector[[j]];
  eY2sum = (Y /. λ → ratevector[[j]]) * (Y /. λ → ratevector[[j]]);
  eX12sum =  (X1 /. λ → ratevector[[j]]) * (X1 /. λ → ratevector[[j]]);
  eX1Ysum = (X1 /. λ → ratevector[[j]]) * (Y /. λ → ratevector[[j]]);
  expectation = eYsum - (eX1sum + √eX1sum / Pi );
  variance = eYsum - eY2sum + ((Pi - 1) / Pi) * eX1sum - eX12sum + 2 * eX1Ysum;
  Print[ratevector[[j]]];
  Print[expectation];
  Print[√variance ];
  ndist = NormalDistribution[expectation, √variance ];
  Print[NormalDistribution[expectation, √variance ]];
  Print[ndist];
  princtree = N[CDF[ndist, -0.5]];
  Print[N[CDF[ndist, -0.5]]];
  Print[princtree];
  prpolytomy = N[CDF[ndist, 0.5]] - N[CDF[ndist, -0.5]];
  prcortree = 1 - N[CDF[ndist, 0.5]];
  signal = prcortree / (1 - prpolytomy);

  Export["/home/psg/PSG/SignalNoise/MathematicaOutputFiles_A_4/" <>
    ToString[UCENumber[[i]]] <> "_site_" <> ToString[j] <> ".csv",
   {{UCENumber[[i]], j, ratevector[[j]] , princtree,
     prpolytomy, prcortree, signal, "\n"}}]], 1]];];
```

Printed by Wolfram Mathematica Student Edition

3-A7 Appendix Figure 7 Mathematica Signal to Noise Calculation code.

```
#Load tree files
require(phytools)
packageVersion("phytools")
require(ape)
Moderate_tree<-
read.tree("~/Documents/Research/Aves/Figures/QuestionD_Trees/RAxML_bipartitions.Run.7.1
2.2016.tree")
plotTree(Moderate_tree)
Total_UCE_tree<-
read.tree("~/Documents/Research/Aves/Figures/Total_Trees/RAxML_bipartitions.Reconciled_T
ree.10.2.16.tree")
plotTree(Total_UCE_tree)
Extreme_tree<-
read.tree('~/Documents/Research/Aves/Figures/QuestionE_Trees/RAxML_bipartitions.Questio
nE_Reconciled_Tree.10.6.16.tree')
plotTree(Extreme_tree)
Jarvis_TENT<-
read.tree("~/Documents/Research/Aves/Newick_tree_files/TENT.ExaML.ShortenedNames.PSG
.tre")
plotTree(Jarvis_TENT)
Jarvis_UCE<-
read.tree("~/Documents/Research/Aves/Newick_tree_files/UCE.RAxML.unpartitioned_Shortene
dNames.PSG.tre")
plotTree(Jarvis_UCE)
Prum<-read.tree("~/Documents/Research/Aves/Newick_tree_files/Holocentrus-
Prum_et_al_2015-
a03a2b5/Trees/Concatenated/ExaBayes/ExaBayes_ConsensusExtendedMajorityRuleNewick_259l
ocus.tre")
#

str(Extreme_tree)
# "acanthisitta_chloris","Rifleman"
# "anas_platyrhynchos_domestica","Peking duck"
# "apaloderma_vittatum","Bar-tailed trogon"
# "aptenodytes_forsteri","Emperor penguin"
# "balearica_regulorum_gibbericeps","Grey crowned crane"
# "buceros_rhinoceros_silvestris","Rhinoceros hornbill"
# "calypte_anna","Anna's hummingbird"
# "caprimugus_carolinensis","Chuck-will's widow"
# "cariama_cristata","Red-legged seriema"
# "cathartes_aura","Turkey vulture"
# "chaetura_pelagica","Chimney swift"
# "charadrius_vociferus","Killdeer"
# "chlamydotis_undulata","MacQueen's bustard"
# "colius_striatus","Speckled mousebird"
# "columba_livia","Pigeon"
# "corvus_brachyrhynchos","American crow"
# "cuculus_canorus","Common cuckoo"
# "egretta_garzetta","Little egret"
# "eurypyga_helias","Sunbittern"
# "falco_peregrinus","Peregrine falcon"
# "fulmarus_glacialis","Northern fulmar"
# "galga","Chicken"
# "gavia_stellata","Red-throated loon"
# "geospiza_fortis","Medium ground-finch"
# "haliaeetus_albicilla","White-tailed eagle"
# "haliaeetus_leucocephalus","Bald eagle"
# "leptosomus_discolor","Cuckoo-roller"
# "manacus_vitellinus","Golden-collard manakin"
# "meleagris_gallopavo","Turkey"
# "melopsittacus_undulatus","Budgerigar"
# "merops_nubicus","Carmine bee-eater"
# "mesitornis_unicolor","Brown mesite"
# "nestor_notabilis","Kea"
# "nipponia_nippon","Crested ibis"
# "ophisthocomus_hoazin","Hoatzin"
# "pelecanus_crispus","Dalmatian pelican"
# "phaethon_lepturus","White-tailed tropicbird"
# "phalacrocorax_carbo","Great cormorant"
# "phoenicopterus_ruber","American Flamingo"
# "picoides_pubescens","Downy woodpecker"
# "podiceps_cristatus","Great-crested grebe"
# "pterocles_guturalis","Yel.-thr. sandgrouse"
# "pygoscelis_adeliae","Adelie penguin"
# "struthio_camelus","Common ostrich"
# "taeniopygia_guttata","Zebra finch"
# "tauraco_erythrolophus","Red-crested turaco"
# "tinamus_major","Wht.-thr. tinamou"
# "tyto_alba","Barn owl"


mgsub <- function(pattern, replacement, x, ...) {
 if (length(pattern)!=length(replacement)) {
  stop("pattern and replacement do not have the same length.")
 }
 result <- x
 for (i in 1:length(pattern)) {
  result <- gsub(pattern[i], replacement[i], result, ...)
 }
 result
}
```

```
Fixed_Tip_Labels<-
mgsub(c("acanthisitta_chloris","anas_platyrhynchos_domestica","apaloderma_vittatum","aptenodyt
es_forsteri","balearica_regulorum_gibbericeps","buceros_rhinoceros_silvestris","calypte_anna","cap
rimugus_carolinensis","cariama_cristata","cathartes_aura","chaetura_pelagica","charadrius_vociferus
","chlamydotis_undulata","colius_striatus","columba_livia","corvus_brachyrhynchos","cuculus_can
orus","egretta_garzetta","eurypyga_helias","falco_peregrinus","fulmarus_glacialis","galga","gavia_ste
llata","geospiza_fortis","haliaeetus_albicilla","haliaeetus_leucocephalus","leptosomus_discolor","ma
nacus_vitellinus","meleagris_gallopavo","melopsittacus_undulatus","merops_nubicus","mesitornis_
unicolor","nestor_notabilis","nipponia_nippon","ophisthocomus_hoazin","pelecanus_crispus","pha
ethon_lepturus","phalacrocorax_carbo","phoenicopterus_ruber","picoides_pubescens","podiceps_c
ristatus","pterocles_guturalis","pygoscelis_adeliae","struthio_camelus","taeniopygia_guttata","taurac
o_erythrolophus","tinamus_major","tyto_alba"),c("Rifleman","Peking duck","Bar-tailed
trogon","Emperor penguin","Grey crowned crane","Rhinoceros hornbill","Anna's
hummingbird","Chuck-will's widow","Red-legged seriema","Turkey vulture","Chimney
swift","Killdeer","MacQueen's bustard","Speckled mousebird","Pigeon","American
crow","Common cuckoo","Little egret","Sunbittern","Peregrine falcon","Northern
fulmar","Chicken","Red-throated loon","Medium ground-finch","White-tailed eagle","Bald
eagle","Cuckoo-roller","Golden-collard manakin","Turkey","Budgerigar","Carmine bee-
eater","Brown mesite","Kea","Crested ibis","Hoatzin","Dalmatian pelican","White-tailed
tropicbird","Great cormorant","American Flamingo","Downy woodpecker","Great-crested
grebe","Yel.-thr. sandgrouse","Adelie penguin","Common ostrich","Zebra finch","Red-crested
turaco","Wht.-thr. tinamou","Barn owl"),Extreme_tree$tip.label)

Extreme_tree$tip.label<-Fixed_Tip_Labels
plotTree(Extreme_tree)

Fixed_Total_Tip_Labels<-
mgsub(c("acanthisitta_chloris","anas_platyrhynchos_domestica","apaloderma_vittatum","aptenodyt
es_forsteri","balearica_regulorum_gibbericeps","buceros_rhinoceros_silvestris","calypte_anna","cap
rimugus_carolinensis","cariama_cristata","cathartes_aura","chaetura_pelagica","charadrius_vociferus
","chlamydotis_undulata","colius_striatus","columba_livia","corvus_brachyrhynchos","cuculus_can
orus","egretta_garzetta","eurypyga_helias","falco_peregrinus","fulmarus_glacialis","galga","gavia_ste
llata","geospiza_fortis","haliaeetus_albicilla","haliaeetus_leucocephalus","leptosomus_discolor","ma
nacus_vitellinus","meleagris_gallopavo","melopsittacus_undulatus","merops_nubicus","mesitornis_
unicolor","nestor_notabilis","nipponia_nippon","ophisthocomus_hoazin","pelecanus_crispus","pha
ethon_lepturus","phalacrocorax_carbo","phoenicopterus_ruber","picoides_pubescens","podiceps_c
ristatus","pterocles_guturalis","pygoscelis_adeliae","struthio_camelus","taeniopygia_guttata","taurac
o_erythrolophus","tinamus_major","tyto_alba"),c("Rifleman","Peking duck","Bar-tailed
trogon","Emperor penguin","Grey crowned crane","Rhinoceros hornbill","Anna's
hummingbird","Chuck-will's widow","Red-legged seriema","Turkey vulture","Chimney
swift","Killdeer","MacQueen's bustard","Speckled mousebird","Pigeon","American
crow","Common cuckoo","Little egret","Sunbittern","Peregrine falcon","Northern
fulmar","Chicken","Red-throated loon","Medium ground-finch","White-tailed eagle","Bald
eagle","Cuckoo-roller","Golden-collard manakin","Turkey","Budgerigar","Carmine bee-
eater","Brown mesite","Kea","Crested ibis","Hoatzin","Dalmatian pelican","White-tailed
tropicbird","Great cormorant","American Flamingo","Downy woodpecker","Great-crested grebe","Yel.-thr. sandgrouse","Adelie penguin","Common ostrich","Zebra finch","Red-crested turaco","Wht.-thr. tinamou","Barn owl"),Total_UCE_tree$tip.label)

Total_UCE_tree$tip.label<-Fixed_Total_Tip_Labels
plotTree(Total_UCE_tree)


Fixed_Moderate_Tip_Labels<-mgsub(c("acanthisitta_chloris","anas_platyrhynchos_domestica","apaloderma_vittatum","aptenodytes_forsteri","balearica_regulorum_gibbericeps","buceros_rhinoceros_silvestris","calypte_anna","caprimugus_carolinensis","cariama_cristata","cathartes_aura","chaetura_pelagica","charadrius_vociferus","chlamydotis_undulata","colius_striatus","columba_livia","corvus_brachyrhynchos","cuculus_canorus","egretta_garzetta","eurypyga_helias","falco_peregrinus","fulmarus_glacialis","galga","gavia_stellata","geospiza_fortis","haliaeetus_albicilla","haliaeetus_leucocephalus","leptosomus_discolor","manacus_vitellinus","meleagris_gallopavo","melopsittacus_undulatus","merops_nubicus","mesitornis_unicolor","nestor_notabilis","nipponia_nippon","ophisthocomus_hoazin","pelecanus_crispus","phaethon_lepturus","phalacrocorax_carbo","phoenicopterus_ruber","picoides_pubescens","podiceps_cristatus","pterocles_guturalis","pygoscelis_adeliae","struthio_camelus","taeniopygia_guttata","tauraco_erythrolophus","tinamus_major","tyto_alba"),c("Rifleman","Peking duck","Bar-tailed trogon","Emperor penguin","Grey crowned crane","Rhinoceros hornbill","Anna's hummingbird","Chuck-will's widow","Red-legged seriema","Turkey vulture","Chimney swift","Killdeer","MacQueen's bustard","Speckled mousebird","Pigeon","American crow","Common cuckoo","Little egret","Sunbittern","Peregrine falcon","Northern fulmar","Chicken","Red-throated loon","Medium ground-finch","White-tailed eagle","Bald eagle","Cuckoo-roller","Golden-collard manakin","Turkey","Budgerigar","Carmine bee-eater","Brown mesite","Kea","Crested ibis","Hoatzin","Dalmatian pelican","White-tailed tropicbird","Great cormorant","American Flamingo","Downy woodpecker","Great-crested grebe","Yel.-thr. sandgrouse","Adelie penguin","Common ostrich","Zebra finch","Red-crested turaco","Wht.-thr. tinamou","Barn owl"),Moderate_tree$tip.label)

Moderate_tree$tip.label<-Fixed_Moderate_Tip_Labels
plotTree(Moderate_tree)

plotTree(Jarvis_TENT) #File already has names changed within it.

plotTree(Jarvis_UCE) #File already has names changed within it.

write.csv(x=Prum$tip.label,file="~/Desktop/PrumTips.csv")

#############################################################
#############################################################
#############################################################
#############################################################
#############################################################
#############################################################
```

```
rr.82_edge<-reroot(Extreme_tree, 82)
plotTree(rr.82_edge)

#
plotTree(Total_UCE_tree,node.numbers=T)
Total.rr.80<-reroot(Total_UCE_tree,80)

#
plotTree(Moderate_tree,node.numbers=T)
Moderate.rr.57<-reroot(Moderate_tree,57)

#
plotTree(Jarvis_TENT)
TENT.rr.76<-reroot(Jarvis_TENT,76)
#
plotTree(Jarvis_UCE)
UCE_Jarvis.rr.51<-reroot(Jarvis_UCE,51)
############################################################
############################################################
ultra82_p.05<-compute.brlen(rr.82_edge,power = 0.5)
plotTree(ultra82_p.05)
write.tree(phy = ultra82_p.05,file = "~/Documents/Research/Aves/Figures/roundtrip.tree")
#Roundtripping because there is a bug

Ultra_Total_p.05<-compute.brlen(Total.rr.80,power=0.5)
plotTree(Ultra_Total_p.05)
write.tree(phy = Ultra_Total_p.05,file =
"~/Documents/Research/Aves/Figures/Total_Ultra.tree")

Ultra_Moderate_p.05<-compute.brlen(Moderate.rr.57,power=0.5)
plotTree(Ultra_Moderate_p.05)
write.tree(phy=Ultra_Moderate_p.05,file="~/Documents/Research/Aves/Figures/Moderate_Ultra
.tree")

Ultra_TENT_p.05<-compute.brlen(TENT.rr.76,power=0.5)
plotTree(Ultra_TENT_p.05)
write.tree(phy=Ultra_TENT_p.05,file="~/Documents/Research/Aves/Figures/Jarvis_TENT_Ult
ra.tree")

Ultra_UCE.Jarvis_p.05<-compute.brlen(UCE_Jarvis.rr.51,power=0.5)
plotTree(Ultra_UCE.Jarvis_p.05)
write.tree(phy=Ultra_UCE.Jarvis_p.05,file="~/Documents/Research/Aves/Figures/Jarvis_UCE_
Ultra.tree")
```

```
##############################################################
##############################################################
Roundtree<-read.tree(file = "~/Documents/Research/Aves/Figures/roundtrip.tree") #Extreme
Tree
plotTree(Roundtree)

rt.49.rt<-rotateNodes(Roundtree,49)
rt.50<-rotateNodes(rt.49.rt,50)
rt.51<-rotateNodes(rt.50,51)
rt.all<-rotateNodes(rt.51,"all")
rt.51.rt<-rotateNodes(rt.all,51)
rt.50.rt.rt<-rotateNodes(rt.51.rt,50)
rt.92<-rotateNodes(rt.50.rt.rt,92)
rt.91<-rotateNodes(rt.92,91)
rt.90<-rotateNodes(rt.91,90)
rt.87<-rotateNodes(rt.90,87)
rt.86<-rotateNodes(rt.87,86)
rt.80<-rotateNodes(rt.86,80)
rt.82<-rotateNodes(rt.80,82)
rt.76<-rotateNodes(rt.82,76)
rt.75<-rotateNodes(rt.76,75)
rt.56<-rotateNodes(rt.75,56)
rt.57<-rotateNodes(rt.56,57)
rt.60<-rotateNodes(rt.57,60)
rt.61<-rotateNodes(rt.60,61)
rt.62<-rotateNodes(rt.61,62)
rt.63<-rotateNodes(rt.62,63)
rt.68<-rotateNodes(rt.63,68)
rt.70<-rotateNodes(rt.68,70)
rt.71<-rotateNodes(rt.70,71)
rt.72<-rotateNodes(rt.71,72)
rt.66<-rotateNodes(rt.72,66)
rt.89<-rotateNodes(rt.66,89)
rt.85<-rotateNodes(rt.89,85)
rt.95<-rotateNodes(rt.85,95)
rt.83<-rotateNodes(rt.95,83)
rt.89<-rotateNodes(rt.83,89)
rt.91<-rotateNodes(rt.89,91)
rt.69<-rotateNodes(rt.91,69)
rt.59<-rotateNodes(rt.69,59)
rt.49<-rotateNodes(rt.59,49)
plotTree(rt.49)
write.tree(phy = rt.49,file =
"~/Documents/Research/Aves/Figures/QuestionE_Trees/RotatedNodes_Extreme.tree")
```

```
Total_Ultra<-read.tree(file = "~/Documents/Research/Aves/Figures/Total_Ultra.tree")
plotTree(Total_Ultra)
T_rt.49<-rotateNodes(Total_Ultra,49)
T_rt.50<-rotateNodes(T_rt.49,50)
T_rt.51<-rotateNodes(T_rt.50,51)
T_rt.50<-rotateNodes(T_rt.51,50)
T_rt.92<-rotateNodes(T_rt.50,92)
T_rt.91<-rotateNodes(T_rt.92,91)
T_rt.90<-rotateNodes(T_rt.91,90)
T_rt.87<-rotateNodes(T_rt.90,87)
T_rt.86<-rotateNodes(T_rt.87,86)
T_rt.80<-rotateNodes(T_rt.86,80)
T_rt.82<-rotateNodes(T_rt.80,82)
T_rt.76<-rotateNodes(T_rt.82,76)
T_rt.75<-rotateNodes(T_rt.76,75)
T_rt.56<-rotateNodes(T_rt.75,56)
T_rt.57<-rotateNodes(T_rt.56,57)
T_rt.60<-rotateNodes(T_rt.57,60)
T_rt.61<-rotateNodes(T_rt.60,61)
T_rt.62<-rotateNodes(T_rt.61,62)
T_rt.63<-rotateNodes(T_rt.62,63)
T_rt.68<-rotateNodes(T_rt.63,68)
T_rt.70<-rotateNodes(T_rt.68,70)
T_rt.71<-rotateNodes(T_rt.70,71)
T_rt.72<-rotateNodes(T_rt.71,72)
T_rt.73<-rotateNodes(T_rt.72,73)
T_rt.66<-rotateNodes(T_rt.73,66)
T_rt.49<-rotateNodes(T_rt.66,49)
T_rt.52<-rotateNodes(T_rt.49,52)
T_rt.53<-rotateNodes(T_rt.52,53)
T_rt.54<-rotateNodes(T_rt.53,54)
T_rt.55<-rotateNodes(T_rt.54,55)
T_rt.58<-rotateNodes(T_rt.55,58)
T_rt.57<-rotateNodes(T_rt.58,57)
T_rt.69<-rotateNodes(T_rt.57,69)
T_rt.72.rt<-rotateNodes(T_rt.69,72)
T_rt.73.rt<-rotateNodes(T_rt.72.rt,73)
T_rt.59<-rotateNodes(T_rt.73.rt,59)
T_rt.62.rt<-rotateNodes(T_rt.59,62)
T_rt.63.rt<-rotateNodes(T_rt.62.rt,63)
T_rt.87.rt<-rotateNodes(T_rt.63.rt,87)
T_rt.67.rt<-rotateNodes(T_rt.87.rt,67)
T_rt.76<-rotateNodes(T_rt.67.rt,76)
T_rt.80<-rotateNodes(T_rt.76,80)
T_rt.83<-rotateNodes(T_rt.80,83)
T_rt.85<-rotateNodes(T_rt.83,85)
```

```
T_50<-rotateNodes(T_rt.85,50)
T_49.rt<-rotateNodes(T_50,49)
T_93.rt<-rotateNodes(T_49.rt,93)
T_92.rt<-rotateNodes(T_93.rt,92)
T_77.rt<-rotateNodes(T_92.rt,77)
T_67.rt<-rotateNodes(T_77.rt,67)
write.tree(phy = T_67.rt,file = "~/Documents/Research/Aves/Figures/RotatedNodes_Total.tree")


Roundtrip_Moderate_tree_rooted_ultra<-
read.tree(file="~/Documents/Research/Aves/Figures/Moderate_Ultra.tree")
plotTree(Roundtrip_Moderate_tree_rooted_ultra)
nodelabels()
M_rt.49<-rotateNodes(Roundtrip_Moderate_tree_rooted_ultra,49)
m_rt.50<-rotateNodes(M_rt.49,50)
M.rt.51<-rotateNodes(m_rt.50,51)
M.rt.52<-rotateNodes(M.rt.51,52)
M.rt.61<-rotateNodes(M.rt.52,61)
M.rt.62<-rotateNodes(M.rt.61,62)
M.rt.81<-rotateNodes(M.rt.61,81)
M.rt.86<-rotateNodes(M.rt.81,86)
M.rt.82<-rotateNodes(M.rt.86,82)
M.rt.62<-rotateNodes(M.rt.82,62)
M.rt.52<-rotateNodes(M.rt.62,52)
M.rt.64<-rotateNodes(M.rt.52,64)
M.rt.65<-rotateNodes(M.rt.64,65)
M.rt.66<-rotateNodes(M.rt.65,66)
M.rt.68<-rotateNodes(M.rt.66,68)
M.rt.55<-rotateNodes(M.rt.68,55)
M.rt.94<-rotateNodes(M.rt.55,94)
M.rt.53<-rotateNodes(M.rt.94,53)
M.rt.77<-rotateNodes(M.rt.53,77)
M.rt.79<-rotateNodes(M.rt.77,79)
M.rt.72<-rotateNodes(M.rt.79,72)
M.rt.85<-rotateNodes(M.rt.72,85)
M.rt.58<-rotateNodes(M.rt.85,58)
M.rt.89<-rotateNodes(M.rt.58,89)
M.rt.91<-rotateNodes(M.rt.89,91)
M.rt.92<-rotateNodes(M.rt.91,92)

Roundtrip_TENT_tree_rooted_ultra<-
read.tree(file="~/Documents/Research/Aves/Figures/Jarvis_TENT_Ultra.tree") #Jarvis ExaML
plotTree(Roundtrip_TENT_tree_rooted_ultra)
nodelabels()
J.rt.49<-rotateNodes(Roundtrip_TENT_tree_rooted_ultra,49)
J.rt.50<-rotateNodes(J.rt.49,50)
```

```
J.rt.51<-rotateNodes(J.rt.50,51)
J.rt.52<-rotateNodes(J.rt.51,52)
J.rt.53<-rotateNodes(J.rt.52,53)
J.rt.54<-rotateNodes(J.rt.53,54)
J.rt.55<-rotateNodes(J.rt.54,55)
J.rt.56<-rotateNodes(J.rt.55,56)
J.rt.57<-rotateNodes(J.rt.56,57)
J.rt.60<-rotateNodes(J.rt.57,60)
J.rt.61<-rotateNodes(J.rt.60,61)
J.rt.62<-rotateNodes(J.rt.61,62)
J.rt.68<-rotateNodes(J.rt.62,68)
J.rt.70<-rotateNodes(J.rt.68,70)
J.rt.76<-rotateNodes(J.rt.70,76)
J.rt.78<-rotateNodes(J.rt.76,78)
J.rt.80<-rotateNodes(J.rt.78,80)
J.rt.94<-rotateNodes(J.rt.80,94)
J.rt.74<-rotateNodes(J.rt.94,74)
J.rt.85<-rotateNodes(J.rt.74,85)
J.rt.95<-rotateNodes(J.rt.85,95)
J.rt.90<-rotateNodes(J.rt.95,90)
J.rt.87<-rotateNodes(J.rt.90,87)

Roundtrip_UCE.Jarvis_rooted_ultra<-
read.tree(file="~/Documents/Research/Aves/Figures/Jarvis_UCE_Ultra.tree")
plotTree(Roundtrip_UCE.Jarvis_rooted_ultra)
nodelabels()
UCE.j.rt.51<-rotateNodes(Roundtrip_UCE.Jarvis_rooted_ultra,51)
UCE.j.rt.94<-rotateNodes(UCE.j.rt.51,94)
UCE.j.rt.95<-rotateNodes(UCE.j.rt.94,95)
UCE.j.rt.59<-rotateNodes(UCE.j.rt.95,59)
UCE.j.rt.63<-rotateNodes(UCE.j.rt.59,63)
UCE.j.rt.91<-rotateNodes(UCE.j.rt.63,91)
UCE.j.rt.67<-rotateNodes(UCE.j.rt.91,67)
UCE.j.rt.76<-rotateNodes(UCE.j.rt.67,76)
UCE.j.rt.78<-rotateNodes(UCE.j.rt.76,78)
UCE.j.rt.82<-rotateNodes(UCE.j.rt.78,82)
UCE.j.rt.74<-rotateNodes(UCE.j.rt.82,74)
UCE.j.rt.69<-rotateNodes(UCE.j.rt.74,69)
UCE.j.rt.50<-rotateNodes(UCE.j.rt.69,50)
UCE.j.rt.53<-rotateNodes(UCE.j.rt.50,53)
UCE.j.rt.75<-rotateNodes(UCE.j.rt.53,75)
UCE.j.rt.83<-rotateNodes(UCE.j.rt.75,83)
UCE.j.rt.68<-rotateNodes(UCE.j.rt.83,68)
UCE.j.rt.71<-rotateNodes(UCE.j.rt.68,71)
######################################################################
######################################################################
```

```
############################################################
############################################################

#Coloring clades for the extreme tree
cols_Extreme<-c("green","sienna4","blue2","cyan","darkorange","darkmagenta","yellow","black");
names(cols_Extreme)<-c(1,2,3,4,5,6,7,8)
blacktips<-paintSubTree(rt.49,node=49,state="8",stem=FALSE)
greentips<-paintSubTree(blacktips,node=56,state="1",stem=FALSE)
sienna4tips<-paintSubTree(greentips,node=74,state="2",stem=FALSE)
bluetips<-paintSubTree(sienna4tips,node=78, state="3",stem=FALSE)
orangetips<-paintSubTree(bluetips,node=87,state="5",stem=FALSE)
purpletips<-paintSubTree(orangetips,node=89,state="6",stem=FALSE)
yellowtips.1<-paintSubTree(purpletips,node=16,state="7",stem=TRUE)
yellowtips.2<-paintSubTree(yellowtips.1,node=14,state="7",stem=TRUE)
cyantips<-paintSubTree(yellowtips.2,node=77,state="4",stem=FALSE)
ExtremeTree<-cyantips
plotSimmap(ExtremeTree,cols_Extreme,lwd=4,pts=F)


#Coloring clades for the total tree
cols_Total<-c("green","sienna4","blue2","cyan","darkorange","darkmagenta","yellow","black");
names(cols_Total)<-c(1,2,3,4,5,6,7,8)
blacktips.t<-paintSubTree(T_67.rt,node=49,state="8",stem=FALSE)
greentips.t<-paintSubTree(blacktips.t,node=57,state="1",stem=FALSE)
sienna4tips.t<-paintSubTree(greentips.t,node=83,state="2",stem=FALSE)
bluetips.t<-paintSubTree(sienna4tips.t,node=75, state="3",stem=FALSE)
orangetips.t<-paintSubTree(bluetips.t,node=88,state="5",stem=FALSE)
purpletips.t.1<-paintSubTree(orangetips.t,node=90,state="6",stem=FALSE)
purpletips.t.2<-paintSubTree(purpletips.t.1,node=92,state="6",stem=FALSE)
yellowtips.t.1<-paintSubTree(purpletips.t.2,node=16,state="7",stem=TRUE)
yellowtips.t.2<-paintSubTree(yellowtips.t.1,node=15,state="7",stem=TRUE)
cyantips.t<-paintSubTree(yellowtips.t.2,node=85,state="4",stem=FALSE)
TotalTree<-cyantips.t
plotSimmap(TotalTree,cols_Total,lwd=4)


#Coloring clades for the moderate tree
cols_moderate<-c("green","sienna4","blue2","cyan","darkorange","darkmagenta","yellow","black");
names(cols_moderate)<-c(1,2,3,4,5,6,7,8)
blacktips.m<-paintSubTree(M.rt.92,node=49,state="8",stem=FALSE)
greentips.m<-paintSubTree(blacktips.m,node=62,state="1",stem=FALSE)
sienna4tips.m<-paintSubTree(greentips.m,node=58,state="2",stem=FALSE)
bluetips.m<-paintSubTree(sienna4tips.m,node=80,state="3",stem=FALSE)
orangetips.m<-paintSubTree(bluetips.m,node=53,state="5",stem=FALSE)
purpletips.m.1<-paintSubTree(orangetips.m,node=92,state="6",stem=FALSE)
purpletips.m.2<-paintSubTree(purpletips.m.1,node=90,state="6",stem=FALSE)
```

```
yellowtips.m<-paintSubTree(purpletips.m.2,node=88,state="7",stem=FALSE)
cyantips.m<-paintSubTree(yellowtips.m,node=60,state="4",stem=FALSE)
ModerateTree<-cyantips.m
plotSimmap(ModerateTree,cols_moderate,lwd=4)

#Coloring clades for the TENT tree
cols_Jarvis<-c("green","sienna4","blue2","cyan","darkorange","darkmagenta","yellow","black")
names(cols_Jarvis)<-c(1,2,3,4,5,6,7,8)
blacktips.j<-paintSubTree(J.rt.87,node=49,state="8",stem=FALSE)
greentips.j<-paintSubTree(blacktips.j,node=55,state="1",stem=FALSE)
siennatips.j<-paintSubTree(greentips.j,node=87,state="2",stem=FALSE)
bluetips.j<-paintSubTree(siennatips.j,node=75,state="3",stem=FALSE)
orangetips.j<-paintSubTree(bluetips.j,node=85,state="5",stem=FALSE)
purpletips.j<-paintSubTree(orangetips.j,node=89,state="6",stem=FALSE)
yellowtips.j<-paintSubTree(purpletips.j,node=83,state="7",stem=FALSE)
cyantips.j<-paintSubTree(yellowtips.j,node=74,state="4",stem=FALSE)
Jarvis_ExaML_Tree<-cyantips.j
plotSimmap(Jarvis_ExaML_Tree,cols_Jarvis,lwd=4)

#Coloring clades for the Jarvis UCE tree
cols_UCE.j<-c("green","sienna4","blue2","cyan","darkorange","darkmagenta","yellow","black")
names(cols_UCE.j)<-c(1,2,3,4,5,6,7,8)
blacktips.uce.j<-paintSubTree(UCE.j.rt.71,node=49,state="8", stem=FALSE)
greentips.uce.j<-paintSubTree(blacktips.uce.j,node=66,state="1",stem=FALSE)
siennatips.uce.j<-paintSubTree(greentips.uce.j,node=85,state="2",stem=FALSE)
bluetips.uce.j<-paintSubTree(siennatips.uce.j,node=87,state="3",stem=FALSE)
purpletips.uce.j.1<-paintSubTree(bluetips.uce.j,node=55,state="6",stem=FALSE)
purpletips.uce.j.2<-paintSubTree(purpletips.uce.j.1,node=53,state="6",stem=FALSE)
yellowtips.uce.j<-paintSubTree(purpletips.uce.j.2,node=62,state="7",stem=FALSE)
cyantips.uce.j<-paintSubTree(yellowtips.uce.j,node=65,state="4",stem=FALSE)
orangetips.uce.j<-paintSubTree(cyantips.uce.j,node=58,state="5",stem=FALSE)
Jarvis.UCE.Tree<-orangetips.uce.j
##########################################################
##########################################################


#Extreme vs Total
par(mfrow=c(1,2))
   #Left Panel
   #par(mai=c(0.2,5,0,5))
#write.csv(x=ExtremeTree$node.label,file="~/Desktop/NewBoot.txt")
#Find and replace "100" with "", then create a new variable
NewBoot2<-
c("","","98","82","61","32","","","","","","56","42","","","","99","","","","","","","","","","","53","",""
,"","","","","93","","43","45","99","92","73","","77","","","","")
ExtremeTree$node.label<-NewBoot2
```

```
plotSimmap(ExtremeTree,cols_Extreme,lwd=2,fsize=.70)
nodelabels(ExtremeTree$node.label,adj=c(1,0),bg="white", frame="none",cex=0.7)
    #Right Panel
    #par(mai=c(0.4,0,0,0.2))
#write.csv(x=TotalTree$node.label,file="~/Desktop/Total.bootstrap.csv")
#Find and replace "100" with "", then create a new variable
Total_bootstrap<-
c("","","73","57","56","54","42","","","55","57","","","66","","96","","","","","","","","","","","","",""
,"","","","90","42","","","","90","99","93","77","99","","","","","")
TotalTree$node.label<-Total_bootstrap
plotSimmap(TotalTree,cols_Total,lwd=2,direction="leftwards",fsize = .70)
nodelabels(TotalTree$node.label,adj=c(0,0), bg = "white",frame="none",cex=0.7)
dev.off()

#Extreme vs Jarvis ExaML
par(mfrow=c(1,2),tcl=-0.5,omi=c(0,0,0,0),mai=c(0,0,0,0))
  #LeftPanel
NewBoot2<-
c("","","98","82","61","32","","","","","","56","42","","","","99","","","","","","","","","","","53","",""
,"","","","","93","","43","45","99","92","73","","77","","","","")
ExtremeTree$node.label<-NewBoot2
plotSimmap(ExtremeTree,cols_Extreme,lwd=2,fsize=.70)
nodelabels(ExtremeTree$node.label,adj=c(1,0),bg="white", frame="none",cex=0.7)
  #RightPanel
write.csv(x=Jarvis_ExaML_Tree$node.label,file="~/Desktop/ExaML.bootstrap.csv")
ExaML_Boot<-
c("","","","91","96","","","","","","","","","","","","","","84","","","72","","","70","","","","","","","","
","91","96","91","","55","","","","","","","","")
Jarvis_ExaML_Tree$node.label<-ExaML_Boot
plotSimmap(Jarvis_ExaML_Tree,cols_Jarvis,lwd=2,direction="leftwards",fsize=.70)
nodelabels(Jarvis_ExaML_Tree$node.label,adj=c(0,0), bg = "white",frame="none",cex=0.7)

#Extreme vs Jarvis UCE
par(mfrow=c(1,2),tcl=-0.5,omi=c(0,0,0,0),mai=c(0,0,0,0))
#LeftPanel
NewBoot2<-
c("","","98","82","61","32","","","","","","56","42","","","","99","","","","","","","","","","","53","",""
,"","","","","93","","43","45","99","92","73","","77","","","","")
ExtremeTree$node.label<-NewBoot2
plotSimmap(ExtremeTree,cols_Extreme,lwd=2,fsize=.70)
nodelabels(ExtremeTree$node.label,adj=c(1,0),bg="white", frame="none",cex=0.7)
#RightPanel
write.csv(x=Jarvis.UCE.Tree$node.label,file="~/Desktop/JArvisUCE.bootstrap.csv")
UCE.j_Boot<-
c("","","","","","81","90","","32","83","79","32","59","98","32","9","","","","78","78","","61","","94"
,"","","","","","","","","","","7","","","","","","","","","67","","")
```

```
Jarvis.UCE.Tree$node.label<-UCE.j_Boot
plotSimmap(Jarvis.UCE.Tree,cols_,lwd=2,direction="leftwards",fsize=.70)
nodelabels(Jarvis.UCE.Tree$node.label,adj=c(0,0), bg = "white",frame="none",cex=0.7)




###########################################################
#####      Moderate vs Total
par(mfrow=c(1,2),tcl=-0.5,omi=c(0,0,0,0),mai=c(0,0,0,0))

#Left Panel
plotTree(ModerateTree)
ModerateVTotalTree<-rotateNodes(ModerateTree,57)
plotTree(ModerateVTotalTree)
cols_moderate_V_Total<-
c("green","sienna4","blue2","cyan","darkorange","darkmagenta","yellow","black");
names(cols_moderate_V_Total)<-c(1,2,3,4,5,6,7,8)
blacktips.m.v<-paintSubTree(ModerateVTotalTree,node=49,state="8",stem=FALSE)
greentips.m.v<-paintSubTree(blacktips.m.v,node=62,state="1",stem=FALSE)
sienna4tips.m.v<-paintSubTree(greentips.m.v,node=58,state="2",stem=FALSE)
bluetips.m.v<-paintSubTree(sienna4tips.m.v,node=80,state="3",stem=FALSE)
orangetips.m.v<-paintSubTree(bluetips.m.v,node=53,state="5",stem=FALSE)
purpletips.m.1.v<-paintSubTree(orangetips.m.v,node=92,state="6",stem=FALSE)
purpletips.m.2.v<-paintSubTree(purpletips.m.1.v,node=90,state="6",stem=FALSE)
yellowtips.m.v<-paintSubTree(purpletips.m.2.v,node=88,state="7",stem=FALSE)
cyantips.m.v<-paintSubTree(yellowtips.m.v,node=60,state="4",stem=FALSE)
ModerateTree.v<-cyantips.m.v
plotSimmap(ModerateTree.v,cols_moderate_V_Total,lwd=2,fsize=.70)
write.csv(x=ModerateTree$node.label,file="~/Desktop/Moderate.boostrap.csv")
ModeBoots_<-
c("","","79","86","","92","87","65","58","","","","36","","","","","","","","","","","","","50","51","",""
,"","","","","","","","","","71","44","55","96","","","","","")
ModerateVTotalTree$node.label<-ModeBoots_
nodelabels(ModerateTree$node.label,adj=c(1.1,0),bg="white", frame="none",cex=0.70)
#Right Panel
#TotalTree$node.label<-Total_bootstrap
Total_bootstrap<-
c("","","73","57","56","54","42","","","55","57","","","66","","96","","","","","","","","","","","","",""
,"","","","90","42","","","","90","99","93","77","99","","","","","")
TotalTree$node.label<-Total_bootstrap
plotSimmap(TotalTree,cols_Total,lwd=2,direction="leftwards",fsize = .70)
nodelabels(TotalTree$node.label,adj=c(0,0), bg = "white",frame="none",cex=0.7)

dev.off()
```

```
############################################################
#######   Moderate Vs ExaML Tree
par(mfrow=c(1,2),tcl=-0.5,omi=c(0,0,0,0),mai=c(0,0,0,0))
#Left Panel
ModeBoots_<-
c("","","79","86","","92","87","65","58","","","","36","","","","","","","","","","","","","50","51","",""
,"","","","","","","","","","71","44","55","96","","","","")
ModerateTree$node.label<-ModeBoots_
plotSimmap(ModerateTree,cols_moderate,lwd=2,mar = c(1,1,1,1),fsize=.70)
nodelabels(ModerateTree$node.label,adj=c(1.1,0),bg="white", frame="none",cex=0.70)
#Right Panel
ExaML_Boot<-
c("","","","91","96","","","","","","","","","","","","","","84","","","72","","","70","","","","","","","","
","91","96","91","","55","","","","","","","","","")
Jarvis_ExaML_Tree$node.label<-ExaML_Boot
plotSimmap(Jarvis_ExaML_Tree,cols_Jarvis,lwd=2,direction="leftwards",fsize=.70)
nodelabels(Jarvis_ExaML_Tree$node.label,adj=c(0,0), bg = "white",frame="none",cex=0.7)
dev.off()
############################################################
#######   Moderate Vs Jarvis UCE
par(mfrow=c(1,2),tcl=-0.5,omi=c(0,0,0,0),mai=c(0,0,0,0))
#Left Panel
ModeBoots_<-
c("","","79","86","","92","87","65","58","","","","36","","","","","","","","","","","","","50","51","",""
,"","","","","","","","","","71","44","55","96","","","","")
ModerateTree$node.label<-ModeBoots_
plotSimmap(ModerateTree,cols_moderate,lwd=2,mar = c(1,1,1,1),fsize=.70)
nodelabels(ModerateTree$node.label,adj=c(1.1,0),bg="white", frame="none",cex=0.70)
#RightPanel
write.csv(x=Jarvis.UCE.Tree$node.label,file="~/Desktop/JArvisUCE.bootstrap.csv")
UCE.j_Boot<-
c("","","","","","81","90","","32","83","79","32","59","98","32","9","","","","78","78","","61","","94"
,"","","","","","","","","","","7","","","","","","","","","67","","")
Jarvis.UCE.Tree$node.label<-UCE.j_Boot
plotSimmap(Jarvis.UCE.Tree,cols_Jarvis,lwd=2,direction="leftwards",fsize=.70)
nodelabels(Jarvis.UCE.Tree$node.label,adj=c(0,0), bg = "white",frame="none",cex=0.7)

############################################################
#######   Total Vs Jarvis UCE Tree
par(mfrow=c(1,2),tcl=-0.5,omi=c(0,0,0,0),mai=c(0,0,0,0))
#Left Panel
Total_bootstrap<-
c("","","73","57","56","54","42","","","55","57","","","66","","96","","","","","","","","","","","","",""
,"","","","90","42","","","","90","99","93","77","99","","","","","")
TotalTree$node.label<-Total_bootstrap
plotSimmap(TotalTree,cols_Total,lwd=2,direction=,fsize = .70)
```

```
nodelabels(TotalTree$node.label,adj=c(0,0), bg = "white",frame="none",cex=0.7)

#Right Panel
UCE.j_Boot<-
c("","","","","","81","90","","32","83","79","32","59","98","32","9","","","","78","78","","61","","94"
,"","","","","","","","","","","7","","","","","","","","","67","","")
Jarvis.UCE.Tree$node.label<-UCE.j_Boot
plotSimmap(Jarvis.UCE.Tree,cols_Jarvis,lwd=2,direction="leftwards",fsize=.70)
nodelabels(Jarvis.UCE.Tree$node.label,adj=c(0,0), bg = "white",frame="none",cex=0.7)
dev.off()
##########################################################
#######   Total Vs ExaML Tree
par(mfrow=c(1,2),tcl=-0.5,omi=c(0,0,0,0),mai=c(0,0,0,0))
#Left Panel
Total_bootstrap<-
c("","","73","57","56","54","42","","","55","57","","","66","","96","","","","","","","","","","","","",""
,"","","","90","42","","","","90","99","93","77","99","","","","","")
TotalTree$node.label<-Total_bootstrap
plotSimmap(TotalTree,cols_Total,lwd=2,direction="leftwards",fsize = .70)
nodelabels(TotalTree$node.label,adj=c(0,0), bg = "white",frame="none",cex=0.7)

#Right Panel
plotSimmap(Jarvis_ExaML_Tree,cols_Jarvis,lwd=2,direction="leftwards",fsize=.70)
nodelabels(Jarvis_ExaML_Tree$node.label,adj=c(0,0), bg = "white",frame="none",cex=0.7)
dev.off()
```

3-A8 Appendix Figure 8 Phylogeny Comparison Image production R code. Code can also be found online at https://github.com/PrincessG?tab=repositories

REFERENCES

[database] Aberer, A., J., Faircloth, B., C., Ho, S., Y., Houde, P., Jarvis, E., D., Li, C., Mirarab, S., Nabholz, B., Howard, J., T., Suh, A., Weber, C., C., da Fonseca, R., R., Alfaro-Nunez, A., Narula, N., Liu, L., Burt, D., W., Ellegren, H., Edwards, S., V., Stamatakis, A., Mindell, D., P., Cracraft, J., Braun, E., L., Warnow, T., Wang, J., Gilbert, M., P., Zhang, G., 2014. Phylogenomic analyses data of the avian phylogenomics project. GigaScience Database. doi: http://dx.doi.org/10.5524/101041

Bandelt H., Dress A., 1986.Reconstructing the shape of a tree from observed dissimilarity data. Adv. Appl. Math. 7, 309–343.

Bejerano, G., Pheasant M., Makunin I., Stephen S., Kent W. J., Mattick J.S., Haussler D., 2004. Ultraconserved Elements in the Human Genome. Science 5675, 1321-1325.

Brown J.W., Payne R. B., Mindell D. P., 2007. Nuclear DNA does not reconcile 'rocks' and 'clocks' in Neoaves: a comment on Ericson *et al*. Biol. Lett., 3 257-260
doi: 10.1098/rsbl.2006.0611

Chen M., Liang D., Zhang P., 2015. Selecting Question-Specific Genes to Reduce Incongruence in Phylogenomics: A Case Study of Jawed Vertebrate Backbone Phylogeny. Syst. Biol. 64, 1104-1120. doi: 10.1093/sysbio/syv059

Claramunt S., Cracraft J., 2015A new time tree reveals Earth history's imprint on the evolution of modern birds. Sci. Adv. **1**, e1501005.

Cracraft, J. Cracraft J., Houde P., Ho S.Y., Mindell D.P., Fjeldså J., Lindow B., Edwards S.V., Rahbek C., Mirarab S., Warnow T., Gilbert M.T., Zhang G., Braun E.L., Jarvis E.D. 2015. Response to Comment on "Whole-genome analyses resolve early branches in the tree of life of modern birds". Science 349, 1460
doi:10.1126/science.aab1578.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn T.C**.**, 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8,783-786. pmid: 22593086 doi:10.1098/rsbl.2012.0331.

Dornburg, A., Fisk, J. N., Tamagnan, J., Townsend J. P., 2016. PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. BMC Evol. Biol. 16, 262-269. doi: 10.1186/s12862-016-0837-3

Dornburg A., Townsend J.P., Friedman M., Near T.J., 2014. Phylogenetic informativeness reconciles ray-finned fish molecular divergence times. BMC Evol. Biol. 14, 169-183. URL http://www.biomedcentral.com/1471-2148/14/169

Ericson G.P., Anderson C.L., Britton T., Elzanowski A., Johansson U. S., Källersjö M., Ohlson J.I.,Parsons, T. J., Zuccon D., Mayr G., 2006 Diversification of Neoaves: integration of molecular sequence data an fossils Biol Lett. 2, 543-547
doi: 10.1098/rsbl.2006.0523.

Faircloth, B.C., Branstetter, M.G., White, N.D., Brady, S.G., 2014. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. Res. 15, 489–501. http:// dx.doi.org/10.1111/1755-0998.12328.

Faircloth, B.C., Chang, J., Alfaro, M.E., 2012. TAPIR Enables High-throughput Estimation and Comparison of Phylogenetic Informativeness using Locus specific Substitution Models. arXiv preprint arXiv:12021215 2012, p. 1215.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C., 2012. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. Syst Biol 61, 717-726. pmid: 22232343doi:10.1093/sysbio/sys004.

Faircloth B.C., Sorenson L., Santini F., Alfaro M.E**.,** 2013. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). PLoS ONE 8, e65923. doi:10.1371/journal.pone.0065923.

Feduccia A.,1995. Explosive evolution in tertiary birds and mammals. Science 267, 637-638.

Gilbert P.S., Chang J., Pan C., Sobel E.M., Sinsheimer J.S., Faircloth B.C., Alfaro M.E., 2015. Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. Mol. Phylogen. Evol. 92,140–146.

Green R.E., Braun E.L., Armstrong J., Earl D., Nguyen N., Hickey G., Vandewege M.W., St John J.A., Capella-Gutiérrez S., Castoe T.A., Kern C., Fujita M.K., Opazo J.C., Jurka J., Kojima K.K., Caballero J., Hubley R.M., Smit A.F., Platt R.N., Lavoie C.A., Ramakodi M.P., Finger J.W. Jr., Suh A., Isberg S.R., Miles L., Chong A.Y., Jaratlerdsiri W., Gongora J., Moran C., Iriarte A., McCormack J., Burgess S.C., Edwards S.V., Lyons E., Williams C., Breen M., Howard J.T., Gresham C.R., Peterson D.G., Schmitz J., Pollock D.D., Haussler D., Triplett E.W., Zhang G., Irie N., Jarvis E.D., Brochu C.A., Schmidt C.J., McCarthy F.M., Faircloth B.C., Hoffmann F.G., Glenn T.C., Gabaldón T., Paten B., Ray D.A., 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science 346,1335-1346.

Hackett, S.J., Kimball, R.T., Reddy, S., Bowie, R.C., Braun, E.L., Braun, M.J., Chojnowski, J.L., Cox, W.A., Han, K.L., Harshman, J. and Huddleston, C.J., 2008. A phylogenomic study of birds reveals their evolutionary history. Science 320,1763-1768.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.l.C., Prosdocimi F., Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jnsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alstrm P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G., 2014.Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346,1320-1331.

Jetz W., Thomas G.H., Joy J.B., Hartmann K., Mooers A.O., 2012. The global diversity of birds in space and time. Nature 491, 444-448.

Lopez-Giraldez, F., Townsend, J.P., 2011. PhyDesign: an online application for profiling phylogenetic informativeness. BMC Evol. Biol. 11, 152.

McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn T.C**.,** 2012. Ultraconserved Elements Are Novel Phylogenomic Markers that Resolve Placental Mammal Phylogeny when Combined with Species Tree Analysis. Genome Res. 22, 746–754. pmid: 22207614 doi: 10.1101/gr.125864.111.

McCormack J.E., Harvey M.G., Faircloth B.C., Crawford N.G., Glenn T.C., Brumfield R.T**.,** 2013. A Phylogeny of Birds Based on Over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing. PLoS ONE 8, e54848. pmid: 23382987 doi:10.1371/journal.pone.0054848.

Mitchell, K. J., Cooper, A., Phillips, M. J. 2015 *Science* **349,** 1460
doi: 10.1126/science.aab1062

Ooms J., 2014. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL http://arxiv.org/abs/1403.2805.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G. and Baurain, D., 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol, 9, e1000602. doi:10.1371/journal.pbio.
1000602

Philippe H., Roure B., 2011. Difficult phylogenetic questions: more data, maybe; better methods, certainly. BMC Biology 9, 91-95 http://www.biomedcentral.com/1741-7007/9/91

Poe S., Chubb A.L, 2004. Birds in a bush: Five genes indicate explosive evolution of avian orders.Evolution, 58, 404-415
doi:10.1111/j.0014-3820.2004.tb01655.x

Pond, S.L., Frost, S.D., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679.

Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M., Lemmon, A.R., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526, 569-573.

R Core Team, 2016 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., Haussler, D., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050.

Stamatakis A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30,1312-1313. doi: 10.1093/bioinformatics/btu033

Suh A., Smeds, L., Ellegren H., 2015. The Dynamics of Incomplete Lineage Sorting across the Ancient Adaptive Radiation of Neoavian Birds. PLoS Biol 13, e1002224. doi:10.1371/journal.pbio.1002224

Sun K., Meiklejohn K.A., Faircloth B.C., Glenn T.C., Braun E.B., Kimball R.T**.,** 2014. The evolution of peafowl and other taxa with ocelli (eyespots): A phylogenomic approach. Proc R Soc Lond B Biol Sci 281, 20140823. doi:10.1098/rspb.2014.0823.

Thomas G.H., 2015. Evolution: An avian explosion. Nature, 526,516-517. doi:10.1038/nature15638.

Townsend, J.P., 2007. Profiling phylogenetic informativeness. Syst. Biol. 56, 222–231.

Townsend, J.P. and Lopez-Giraldez, F., 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Systematic biology, 59, 446-457 doi: http://dx.doi.org/doi.org/10.1093/sysbio/syq025

Townsend, J.P. and Leuenberger, C., 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Systematic biology, 60,358-365.

Townsend, J.P., Su Z., Tekle Y.I., 2012. Phylogenetic Signal and Noise: Predicting the Power of a Data Set to Resolve Phylogeny. Syst. Biol. 61, 835-849.

Wolfram Research, Inc., 2016. Mathematica, Version 10.4, Champaign, IL.

Yu G., Smith D., Zhu H., Guan Y., Lam T.T.Y., 2016.
ggtree: an R package for visualization and annotation of phylogenetic tree with different types of meta-data. revised.

Chapter 4

Conclusion

Conclusions

The advent of high-throughput sequencing has revolutionized systematics. Because of these advances we can now approach questions in systematics and phylogenetics with methodologies that incorporate information found across the entire genome and not just individual gene regions. A phylogenomic approach consists of using large-scale or multiple-loci sequence data obtained from intensive sequencing efforts or available genomic databases to reconstruct a reliable evolutionary history of organisms (Chen et al. 2004, Chen et al. 2010; Delsuc et al. 2005). But even though the approach is now different, phylogeneticists are still investigating questions that pre-date the genomic revolution. Which approach is more informative, increased character sampling or taxon sampling (Rosenberg & Kumar, 2001; Rosenberg & Kumar, 2003; Zwickl and Hillis 2002)? What is the most accurate way to differentiate rapid radiations as evidenced by historical polytomies (Poe & Chubb 2004; Rokas et al. 2005)? And most pertinent to my dissertation, how does one use genome-scale sequence data to build robust and accurate phylogenies (Rosenberg & Kumar 2003;Philippe et al. 2005; Jeffroy et al. 2006; Delsuc et al.2005)? This final question inspired my dissertation and as a result my dissertation is one of the first to empirically assess the phylogenetic utility of ultraconserved elements.

I developed a diagnostic framework for finding UCEs with the power to resolve specific phylogenetic questions. This framework consists of identifying phylogenetically informative profile of each individual UCE, and identifying the phylogenetically informative profile of each site within each UCE, identifying signal, noise and polytomy probabilities of each UCE or each site within each UCE, and identifying the signal to noise ratio for the each site across each UCE. In Gilbert et al. (2015) ultraconserved elements were compared to protein-coding genes identified in percomorph fishes. We found that collectively UCEs and their flanking regions had phylogenetically informativeness measures that surpassed these protein coding genes. We also found that the more informativeness was found in the flanking region of UCEs than their cores. In Gilbert et al. (in prep) I found that filtering sites within UCEs based on signal to noise ratios for specific node ages and internode lengths affected our ability to resolve a various nodes across the neoavian phylogeny. For example, we found increased support for the Eucavitaves clade, and the Columbea +Passerea sister relationship after filtering for sites with the highest signal to noise ratio for nodes occurring between 60-62MYA and 27-64MYA.

The ultimate goal of phylogenetics is to arrive at the true species tree for a given taxonomic clade. This goal can be arrived at through the collective evidence of individual gene trees with their own varying evolutionary histories (Doyle et al. 1992,1997; Slowinski & Page, 1999; and Avise, 2000). However, complications to this methodology fall into four main categories: gene duplication, horizontal gene transfer, deep coalescence or incomplete lineage sorting, and branch length heterogeneity (Maddison et al. 1997; Edwards et al. 2009). Undetected gene duplication can lead to spurious conclusions, but if utilized (as in many ray-finned fish studies) questions relating to large-scale genomic change and biodiversity can be addressed (Kocker et al. 2004, Cossins et al. 2005, Volff et al 2005). Incomplete lineage sorting refers to the situation where genes within individual species have not fully coalesced resulting in gene trees of conflicting topologies for the same species. Branch length heterogeneity describes differences in branch lengths for genes that result in identical topologies. This is closely related to heterotachy; the heterogeneity of character change rates over time (Kolaczkowski et al. 2004, 2008).

A potential solution to the problems introduced above is to have a high character to taxon sampling ratio and this can result in phylogenies with high support. Rosenberg and Kumar (2001) conducted a simulation study that showed that incomplete taxon sampling had much smaller effects on the accuracy of a phylogeny than the number of sites and or substitution rate. Poor character sampling (in the presence of a poor phylogenetic signal) was much more likely to lead to spurious results than poor taxon sampling. Weins et al. (2006) also used simulations to determine that incompletely sampled taxa could still be accurately placed in phylogenies provided a large amount of characters were sampled. Gilbert et al (2015) found evidence in support of these conclusions also, but more work remains.

Some have found that increasing character sampling also increased the risk of long branch attraction (Rosenberg and Kumar 2001, 2003; Weins et al. 2006) and high character sampling does not seem to be enough in the most difficult of phylogenetic situations. Branching events separated by short internodes that are then followed by long branches elude phylogeneticists in their quest to curate the tree of life. In these situations a combination of carefully curated taxa and a large amount of sequence data is required to resolve nodes and recover a reasonable amount of statistical confidence. The work completed Chapter 3 of this dissertation is to that end but further analyses would include an exhaustive node and internode length sampling across the neoaves phylogeny to identify UCE positions appropriate for each phylogeny region. And it would be informative to explore the advantages and disadvantages of UCE concatenation vs. gene tree/species tree analyses in context of signal to noise ratio filtering.

REFERENCES

Avise, J. C. 2000. *Phylogeography: The history and formation of species.* Harvard Univ Pr.

Chen, W.-J., Mayden R.L., 2010. A phylogenomic perspective on the new era of Icthyology. Bioscience 60, 421-432 doi: 10.1525/bio.2010.60.6.6

Chen, W.-J., Ortí, G., & Meyer, A. (2004). Novel evolutionary relationship among four fish model systems. *Trends Genet*, *20*, 424 - 431.

Cossins, A. R., & Crawford, D. L. (2005). Fish as models for environmental genomics. *Nat. Rev. Genet.*, *6*, 324-333.

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet, 6, 361-75.

Doyle, J. J. (1992). Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst. Bot., 17, 144-163.

Doyle, J. J., Doyle, J. L., Ballenger, J. A., Dickson, E. E., Kajita, T., & Ohashi, H. (1997). A phylogeny of the chloroplast gene rbcl in the leguminosae: Taxonomic correlations and insights into the evolution of

nodulation. Am J Bot, 84(4), 541.

Gilbert P.S., Chang J., Pan C., Sobel E., Sinsheimer, J.S., Alfaro M.E. (2015). Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. Mol., Phy. Evol. 92, 140-146.

Gilbert P.S., Wu J., Simon M.A, Sinsheimer J.S., Alfaro M.E., (*In prep*). Filtering nucleotide sites from ultraconserved elements by phylogenetic signal to noise improves the precision of the avian phylogeny.

Jeffroy, O., Brinkmann, H., Delsuc, F., & Philippe, H. (2006). Phylogenomics: The beginning of incongruence?.TRENDS in Genet., 22, 225-231.

Kocher, T. D. (2004). Adaptive evolution and explosive speciation: The cichlid fish model. Nat. Rev. Genet., 5, 288-98.

Kolaczkowski, B., & Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature, 431, 980-984.

Kolaczkowski, B., & Thornton, J. W. (2008). A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol. Biol. Evol., 25, 1054.

Maddison, W. P. (1997). Gene trees in species trees. Syst. Biol., 46, 523-536.

Poe, S., & Chubb, A. L. (2004). Birds in a bush: Five genes indicate explosive evolution of avian orders. Evolution, 58, 404-415.

Philippe, H., Delsuc, F., Brinkmann, H., & Lartillot, N. (2005). Phylogenomics. Annu. Rev. Ecol. Evol. Syst., 36, 541-562.

Rokas A. and Carroll S. B. (2008). Frequent and Widespread Parallel Evolution of Protein Sequences, Mol. Biol. Evol., 25, 1943-1953. doi:10.1093/molbev/msn143

Rosenberg, M. S., & Kumar, S. (2001). Incomplete taxon sampling is not a problem for phylogenetic inference. PNAS, 98, 10751-10756.

Rosenberg, M. S., & Kumar, S. (2003). Taxon sampling, bioinformatics, and phylogenomics. Syst Biol, 52, 119-24.

Slowinski, J. B., & Page, R. D. M. (1999). How should species phylogenies be inferred from sequence data?. Syst. Biol., 48, 814.

Volff, J. N. (2004). Genome evolution and biodiversity in teleost fish. Heredity, 94(3), 280-294.

Wiens, J. J., Graham, C. H., Moen, D. S., Smith, S. A., & Reeder, T. W. (2006). Evolutionary and ecological causes of the latitudinal diversity gradient in hylid frogs: Treefrog trees unearth the roots of high tropical diversity. *Am. Nat.*, *168* 579-596.

Zwickl, D. J., & Hillis, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. Syst Biol, 51, 588-98.