Title
Diversity, disparity, and exploitation in the ray-finned fishes

Author
Chang, Jonathan

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Diversity, disparity, and exploitation in the ray-finned fishes

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

by

Jonathan Chang

2017

ABSTRACT OF THE DISSERTATION

Diversity, disparity, and exploitation in the ray-finned fishes

by

Jonathan Chang
Doctor of Philosophy in Biology
University of California, Los Angeles, 2017
Professor Michael Edward Alfaro, Chair

Understanding the process that underlie the disparity in species richness across different taxonomic groups is a fundamental question in evolutionary biology. Several difficulties hinder deeper investigation into this field, namely the lack of high quality phylogenetic and phenotypic data to appropriately test competing hypotheses. I use ray-finned fish (class Actinopterygii), which comprise over half of all vertebrate diversity with 30,000 species in 500 families, as a study system to understand the processes that generate biological diversity. In chapter one, I combine previously-published molecular sequence data to generate a new phylogeny of ray-finned fish containing over 11,000 species and time-calibrate it using over 130 fossils. In chapter two, I develop a new method to collect large amounts of morphological data using crowdsourcing. In chapter three, I develop a new method to estimate completely sampled phylogenies using taxonomic information and birth-death-sampling estimators. In chapter four, I present an accessible web resource to distribute phylogenetic data about actinopterygian fishes. In chapter five, I estimate the distribution of exploitation on the fish tree of life, and test whether certain lineages are disproportionately exploited, and whether certain life history or ecological characteristics predispose species to fishing pressure.

The dissertation of Jonathan Chang is approved.

Kaustuv Roy

Van Maurice Savage

Thomas Bates Smith

Michael Edward Alfaro, Committee Chair

University of California, Los Angeles

2017

Phylogenies cloud the true genealogical process.

To my grandparents—their bright spirits would pierce any haze.

Yi-Ping Chang	1922—	Gansu
Kwun-Rwer Chang Li	1928–2004	
Jon-Chun Jair Wu	1926–2003	Chengdu, Sichuan
Quen-Fung Jair	1914–1992	Guangping, Hebei

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

To my advisor, Michael Alfaro, thank you for for your stellar mentorship over the years. I first stepped into Mike's office on January 26, 2009 to work in his lab as an undergrad, and I'm pleased to report that nearly a decade later, Mike has made science just as fun and exciting as that first day. His unwavering and selfless support, ranging from my rollercoaster admissions process to countless last-minute deadlines, has helped me grow into the scientist that I am today. This dissertation would not exist without his guidance and I am sincerely and forever grateful.

Thank you to my committee members, Van Savage, Tom Smith, and Kaustuv Roy. You were an inspiration and guiding light for me, and have been incredibly helpful for navigating my future career. To my reading committee members, Lawren Sack and John Novembre: thank you for supporting and believing in me when few others did.

To all the Alfaro lab postdocs that I've loved and lost over the years— even though your time in the lab was brief, the influence you had was lasting. Francesco Santini, you taught me how to take fish, science, and fish science seriously; I'll always look forward to our conference time together. Graham Slater, I'll never look at arctoids and whales in quite the same way; thank you for your insight into comparative methods and teaching me how papers are written. Erik Gjesfjeld, you brought into the lab a fantastic new perspective on diversification; thank you for being my friend and dedicated science champion. Sharlene Santana and Bruno Frédérich, thank you for showing the lab how to be awesome mentors and always being there when we had science emergencies.

To my collaborators and coauthors— thank you for your endless patience while I figured out how to do this whole science thing. Brant Faircloth got me started early on the awesomeness of phylogenomics and helped me win my first grant ever. Dan Rabosky has been an extremely patient and generous collaborator and answered all the silly questions I've asked about comparative methods. Matt McGee kept me ichthyologically grounded when I threatened to float off into phylospace. Brian Sidlauskas, Luke Harmon, and Stephen Smith all pulled through for me when times were tough and I needed a helping

hand.

To my instructional mentors and colleagues—Deb Pires, Tonya Kane, and all the teaching assistants and writing consultants I've worked with—thank you for reminding me whenever I lost my way why exactly we do this, and the indispensable role of the public research university.

To the Alfaro labbies, past and present—your solidarity both in and out of the lab, during late night Alfaro email storms, Alfaro bowling contests, surprise Alfaro birthday parties, Alfaro game nights, and Alfaro-forgot-to-pay-the-bills blackouts, has been so wonderful over the past decade. Mai Nguyen and Tina Marcroft, thanks for helping me get my footing and showing that science can be a source of joy in this world. Janet Buckner and Princess Gilbert, thanks for helping me become a competent and fearless graduate student. Mericien Venzon, Andrew Noonan, Jimmy Zheng, Zack Herbst, Binal Patel, David Černý, and Chris Rice, thanks for assisting me on various crazy projects, and always surprising me with the incredible amount of stupendous work you were willing to do. Mark Phuong and Tyler McCraney, thanks for being a bright light during the darkest middle part of graduate school, and for staying late to work with me when everyone else just peaced out. And to Mark Juhn and Elizabeth Karan, thanks for reminding me all over again just what a joy and privilege it is to work in science.

To my friends—Liz Carlen, thank you for being my constant voice of reason and steady supporter; I'm certain we'll one day get to do field work together. Bernard Kim, I'll always be happy to nerd the @#$%& out with you, even when it's terribly embarrassing. Jacqueline Robinson, thanks for showing us how to be composed, even during all of our interminable meltdowns. Grace John, thanks for sticking with me through the spike-lined pits, snake-filled pits, and regular ol' bottomless pits of graduate school. Without your support, it really would have been the pits. (How pitiful...)

And to my family—my parents and especially my brother, Henry. Thank you for all of your love and care during all these years, and feeding and watering me even when I didn't really want to be fed or watered. Thank you for supporting me from day one.

This dissertation would not be possible without the financial support of the Encyclopedia of Life David M. Rubenstein Fellowship, the UCLA George A. Bartholomew Fellowship for Field Biology, a National Science Foundation Doctoral Dissertation Improvement Grant, and the University of Washington Stephen and Ruth Wainwright Fellowship. Travel funding to disseminate the products of this research was supported by the Society of Systematic Biologists, the National Evolutionary Synthesis Center, the Society for the Study of Evolution, and several UCLA Research and Travel Grants.

Chapter 1 is adapted from portions from a manuscript in review: Daniel L. Rabosky,[1] Jonathan Chang,[1] Pascal O. Title,[1] Peter F. Cowman, Lauren Sallan, Matt Friedman, Kristin Kaschner, Cristina Garilao, Thomas J. Near, Michael E. Alfaro,[1] A global tropical depression in speciation rate for marine fishes. DLR, JC, POT, and MEA conceived and designed experiments, developed analysis tools, conducted analyses, and wrote the initial paper. LC and MF contributed fossil calibrations. PFC and TJN contributed taxonomic information and new molecular sequences. KK and CG contributed spatial informatics. All authors contributed to and edited the final paper.

Chapter 2 is a reprint of: Jonathan Chang, Michael E. Alfaro (2016) Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data. *Methods in Ecology and Evolution* 7(4):472-482 doi:10.1111/2041-210x.12508. JC performed the experiments, analyzed the data, and built analysis tools. JC and MEA conceived and designed the experiments and wrote the paper.

Chapter 3 is in preparation for submission: Jonathan Chang, Michael E. Alfaro, TACT: Taxonomic Addition for Complete Trees using birth-death-sampling estimators. JC drafted the manuscript, developed the methods, and wrote the software. MEA planned the work and contributed to the final manuscript.

Chapter 4 is in preparation for submission: Jonathan Chang, Michael E. Alfaro, An online resource for the ray-finned fish tree of life. JC drafted the manuscript, developed the methods, and wrote the software. All authors planned the work and contributed to the

[1]These authors contributed equally

final manuscript.

Chapter 5 is a version of a manuscript in preparation: Jonathan Chang, Kaustuv Roy, Julia K. Baum, Michael E. Alfaro, Devouring the fish tree of life: the phylogenetic distribution of human exploitation. JC analyzed the data and built analysis tools, and wrote the paper. All authors helped design and conceptualize the study and contributed to the final manuscript.

EPIGRAPH

"Slayer." Dareon appeared beside him, oblivious to Sam's pain. "A sweet night, for once. Look, the stars are coming out. We might even get a bit of moon. Might be the worst is done."

"No." Sam wiped his nose, and pointed south with a fat finger, toward the gathering darkness. "There," he said. No sooner had he spoken than lightning flashed, sudden and silent and blinding bright. The distant clouds glowed for half a heartbeat, mountains heaped on mountains, purple and red and yellow, taller than the world. "The worst isn't done. The worst is just beginning, and there are no happy endings."

"Gods be good," said Dareon, laughing. "Slayer, you are *such* a craven."

— *A Feast for Crows*, Chapter 15, Samwell II.

Biographical Sketch: **Jonathan Chang**

(a) Professional Preparation

Institution	Major/Area	Degree/Year
University of California, Los Angeles	Ecology & Evolutionary Biology	B.S., 2011

(b) Appointments

2016—: **Writing Consultant**, Graduate Writing Center, UCLA

2015—: **TA Consultant**, Office of Instructional Development, UCLA

2011—: **PhD Candidate**, Department of Ecology and Evolutionary Biology, UCLA

2011—: **Teaching Assistant**, Department of Ecology and Evolutionary Biology, UCLA

(c) Awards

1. UCLA A. M. Schechtman Award for distinguished teaching (2017)
2. Society for Integrative and Comparative Biology, David and Marvalee Wake Award for best student presentation (2016)

(d) Grants and Fellowships

1. UCLA. George A. Bartholomew Fellowship and Research Award. (2017) $9,000
2. National Science Foundation. Doctoral Dissertation Improvement Grant (Co-PI). Testing macroevolutionary predictions of diversity and disparity in the ray-finned fishes. (2016) $20,020 (DEB-1601830)
3. Encyclopedia of Life. David M. Rubenstein Fellowship (PI). Using massively crowdsourced data to examine morphological impacts of extinction risk in ray-finned fishes. (2013) $52,280 (EOL-33066-13)
4. UCLA. Whitcome Summer Undergraduate Research Fellowship. Phylogenomic approaches to resolving evolutionary relationships among ray-finned fishes. (2010) $3,000

(e) Publications

1. DL Rabosky, JS Mitchell, **J Chang** (2017). Is BAMM flawed? Theoretical and practical concerns in the analysis of multi-rate diversification models. Systematic Biology 66(4):477–498 doi:10.1093/sysbio/syx037
2. E Gjesfjeld, **J Chang**, D Silvestro, C Kelty, ME Alfaro (2016). Competition and extinction explain the evolution of diversity in American automobiles. Palgrave Communications 2:16019 doi:10.1057/palcomms.2016.19.
3. **J Chang**, ME Alfaro (2015). Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data. Methods in Ecology and Evolution 7:472–482 doi:10.1111/2041-210X.12508

4. PS Gilbert, **J Chang**, C Pan, EM Sobel, JS Sinsheimer, BC Faircloth, ME Alfaro (2015). Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. Molecular Phylogenetics and Evolution 92:140-146 doi:10.1016/j.ympev.2015.05.027

5. DL Rabosky, F Santini, J Eastman, SA Smith, B Sidlauskas, **J Chang**, ME Alfaro (2013). Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nature Communications 4:1958 doi:10.1038/ncomms2958

6. BC Faircloth, **J Chang**, ME Alfaro (2012). TAPIR enables high-throughput estimation and comparison of phylogenetic informativeness using locus-specific substitution models. arXiv preprint 1202.1215

(f) Synergistic Activities

Mentoring: Mentored two undergraduate students and one high school student, who presented posters at the UCLA Annual Biology Research Symposium in 2014 and 2015, one of which won the *Best Undergraduate Student Poster* award. Worked at the Graduate Writing Center (2016–2017) to help graduate students professionalize their academic writing.

Professional service: Served on Society for Integrative and Comparative Biology, Student Postdoctoral Affairs Committee 2017–2020.

University service: Served as departmental faculty-student liasion (2016–2016) and on departmental seminar committee (2014–2015).

Community service: (i) Exploring Your Universe (2013–2016): developed and demonstrated interactive science activities for the public. (ii) Los Angeles County Science Fair, Animal Physiology Chair (2014–2015): coordinated and judged middle and high school student projects.

Other service: Maintainer of the Homebrew Science package manager for macOS and Linux.

Teaching: (i) UCLA/La Kretz Workshop in Conservation Genomics (2013–2015): created and presented workshop materials that taught attendees the basics of comparative methods in phylogenetics using R. (ii) Developed and presented multiple workshops for new teaching assistants, including how to teach scientific writing. (iii) Developed course modules for the TA training course in the life sciences departments, including topics on diversity and inclusion, writing a teaching philosophy, and active learning techniques.

CHAPTER 1

The complete ray-finned fish tree of life using multilocus molecular data, taxonomy, and birth-death models

1.1 Summary

Ray-finned fishes (Actinopterygii) represent nearly half of all known vertebrate diversity, yet their evolutionary relationships and the timing of their diversification remain poorly understood. Three recent manuscripts published by several groups (Rabosky *et al.* 2013, Near *et al.* 2013, Betancur-R *et al.* 2013) have attempted to resolve this controversy with large multilocus studies of sequenced nuclear and mitochondrial datasets. Here we present a new multilocus phylogeny combining these and other datasets, representing all known orders and most families of ray-finned fishes. This new time-calibrated phylogeny resolves, to the species level, the relationships among nearly a third (c. 11,000 spp.) of all extant actinopterygian diversity. We time-calibrate this phylogeny using a large fossil dataset of 139 calibration points. We present a method to add unsampled species to a backbone phylogeny using taxonomic constraints and constant-rate birth-death-incomplete sampling estimators, and apply this method to our inferred molecular phylogeny to generate a distribution of the complete ray-finned fish tree of life. We show that this method to generate complete phylogenies using a combination of molecular data and taxonomic placements improves estimates of diversification rates compared to an incompletely sampled phylogeny of only molecular data. We also build a website, fishtreeoflife.org, to disseminate our final phylogeny and taxonomy. Our completed tree inference and web product will be useful for downstream comparative analyses at all levels of evolutionary study.

1.2 Matrix Assembly

To build a multilocus phylogeny, we first generated a multiple sequence alignment (MSA) using PHLAWD. We then used nucleotide BLAST to identify and filter out sequences that were likely to be misidentified or contaminated.

1.2.1 Baited sequence alignment with PHLAWD

PHLAWD uses a baited approach where sequences for a clade of interest are compared to NCBI GenBank sequences and used to download homologous gene regions. We acquired bait sequences for 24 genes from several sources: the "ETOL" set, from the Euteleost Tree of Life project (Betancur-R *et al.* 2013), the "Rabosky" set (Rabosky *et al.* 2013), and the "Near" set (Near *et al.* 2013). A full accounting of baited gene sources is available in Table 1.1.

All PHLAWD analyses used a modified version of the original software. The original version of PHLAWD (github.com/blackrim/phlawd) entered maintenance mode in 2012, and was subsequently modified by Cody Hinchliff (github.com/chinchliff/phlawd) to fix a number of bugs and speed up analyses. Our modified version (github.com/jonchang/phlawd) fixes other bugs and supports including daily updates in addition to the bimonthly GenBank releases.

Our modified version of PHLAWD then assesses these homologous sequences for saturation, and if saturated, broken up into sub-matrices aligned with MAFFT (Katoh and Standley 2013) corresponding to a user taxonomy or guide tree. We conducted a PHLAWD-mediated GenBank search for each gene with the parameters MAD (median average deviation) = 0.01, coverage = 0.2, and identity = 0.2 for NCBI taxon Actinopterygii. Using the NCBI taxonomy, these sub-matrices were then aligned together using profile alignment as provided in MUSCLE (Edgar 2004). We used GNU Parallel (Tange 2011) to parallelize this search, as the built-in parallelization in PHLAWD can occasionally stall using high numbers of threads.

To further increase the genetic coverage of our dataset, we downloaded the full Barcode

of Life (BOLD) database sequences (Ratnasingham and Hebert 2007) and extracted the longest cytochrome oxidase subunit 1 (*coi*) gene for each species in Actinopterygii. We also downloaded full mitochondrial chromosomes for each actinopterygian species and extracted the *nd2* and *nd4* genes (Table 1.1, "mt-genome"). This preliminary alignment included 15,606 species.

	ETOL	Rabosky	Near	BOLD*	mt-genome*
12s		✓			
16s	✓	✓			
4c4		✓			
coi				✓	
cytb		✓			
enc1	✓	✓	✓		
ficd					
glyt	✓	✓	✓		
hoxc6a	✓				
kiaa1239	✓				
myh6	✓	✓	✓		
nd2					✓
nd4					✓
panx2	✓				
plagl2	✓	✓	✓		
ptr	✓		✓		
rag1	✓	✓	✓		
rag2	✓				
rhodopsin	✓	✓			
ripk4	✓				
sh3px3	✓		✓		
sidkey	✓				
sreb2	✓	✓	✓		
svep1	✓				
tbr1	✓	✓	✓		
vcpip	✓				
zic1	✓	✓	✓		

Table 1.1: Gene sources for PHLAWD analyses. Sources marked (*) were not used for baited PHLAWD searches and instead included directly into the character matrix. A full distribution of the final alignment by species is shown in Figure 1.1.

1.2.2 Alignment error-correction

To filter out misidentified sequences, we ran a local nucleotide BLAST search (Camacho *et al.* 2009) on our combined PHLAWD and mitochondrial sequences. Using the closest non-self BLAST match, we ensured that no PHLAWD sequences matched with a high identity to a species outside of the original species family, and checked for contamination by excluding sequences that aligned with high identity to a non-actionpterygian such as *Homo*.

For example, the *enc1* sequence for *Amia calva* (Accession EF032974.1), in family Amiidae, matches with 99.87% identity to *Lepomis cyanellus* (Accession KF139483.1), in family Centrarchidae, despite there being other *enc1* closer for this species. This specific sequence was therefore excluded from the final analysis (Table 1.3).

We used previously described sequencing protocols (Near *et al.* 2013) to generate new multilocus data for 442 species. These were directly added and aligned to the character matrix. Alignments were then quality checked by eye to ensure that coding genes were in frame. The distribution of genes on the final matrix is shown in Figure 1.1.

1.3 Taxonomic reconciliation

We wrote a custom web scraper in Python to download all accepted scientific names, synonyms, and taxonomy for Actinopterygii fishes from FishBase (Froese and Pauly 2014). We then loaded all aligned PHLAWD sequences into an SQLite database to record all taxonomic changes in a consistent format.

We then used a custom Python script to attempt to reconcile the GenBank species names against our known FishBase taxonomy. Species names were matched using the following algorithms, in order:

1. Exact scientific name
2. Exact valid synonym
3. Exact common name

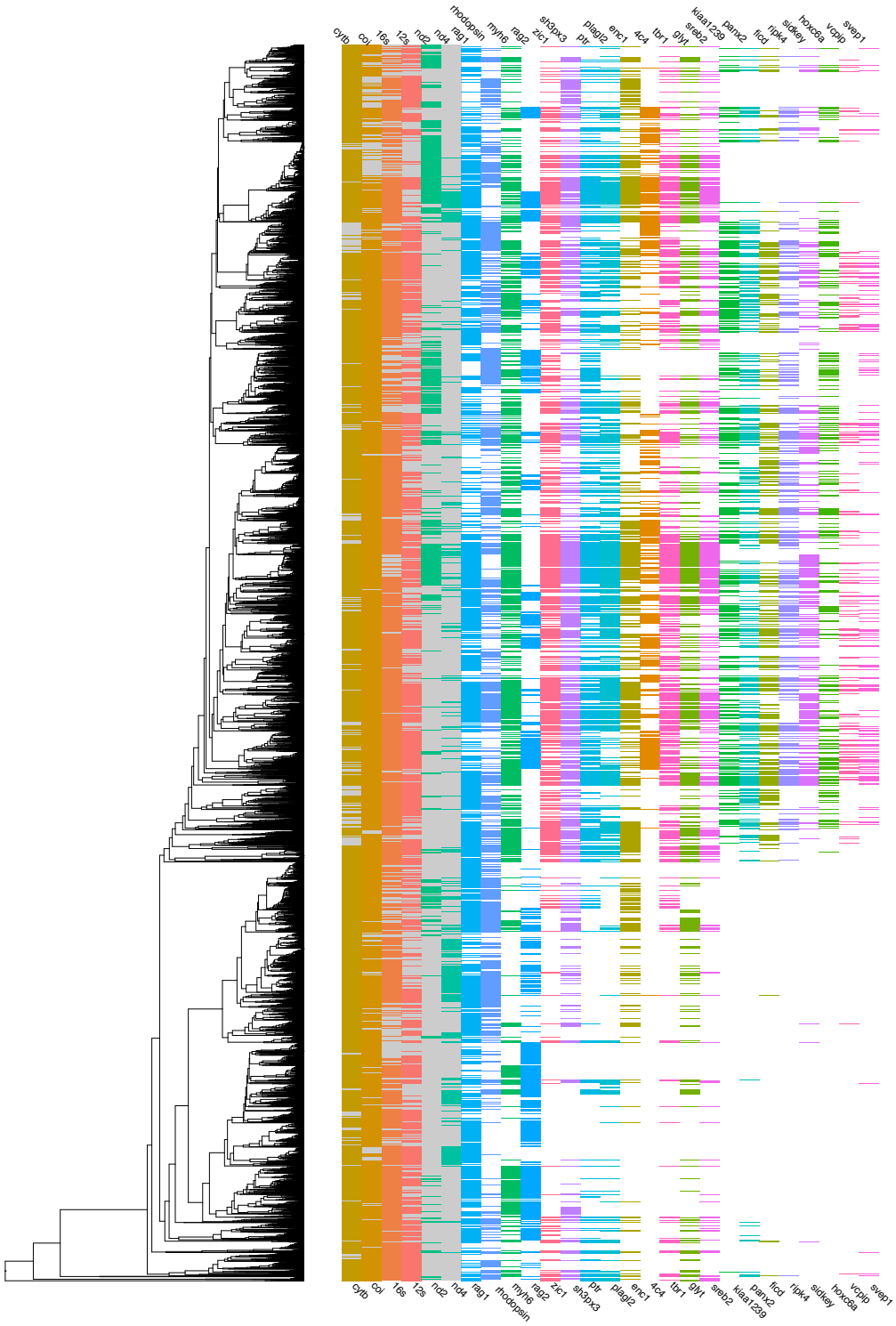

Figure 1.1: Molecular character completeness by species and locus. Each cell in the coverage matrix (**right**) corresponds to the presence of molecular data for that species and locus combination, arranged by phylogenetic position (**left**) so that groups of related species can be drawn as contiguous blocks of color. Loci are organized into blocks of mitochondrial (gray background) and nuclear loci, and secondarily ordered by coverage within major locus type.

Matching method	Count
Exact scientific name	11,368
Exact synonym	623
Manual taxonomic corrections	131
Unmatched-but-unambiguous	84
Exact scientific name, no subspecies	69
Fuzzy scientific name	61
Fuzzy synonym	12
Exact synonym, no subspecies	9

Table 1.2: Taxonomic reconciliation by match type

4. Exact scientific name without subspecies epithet

5. Exact valid synonym, without subspecies epithet

6. Apply manual taxonomic corrections

7. Fuzzy match against scientific names based on the gestalt pattern matching algorithm (Ratcliff and Metzener 1988)

8. Fuzzy match against valid synonyms based on the gestalt pattern matching algorithm

9. Adding unambiguous-but-unmatched species with more than 2 genes, as these are likely to be new species that had not yet been included in FishBase

After these automated mechanisms, we examined matches by hand and manually corrected any mis-assignations, then checked for sequences that were identical, yet were mapped to different species. Our taxonomic reconciliation process matched 46 of 46 orders of fish (100%), 454 of 480 families (94.6%), and 3,368 of 4,853 genera (69.4%), as measured against FishBase. The method for how these matches were accomplished is available in Table 1.2.

1.4 Rogue search with RogueNaRok

To eliminate rogue taxa, which reduce the bootstrap support of phylogenies due to their unstable position, we conducted a RogueNaRok analysis (Aberer *et al.* 2013) and searched for sets of up to 3 species that could be dropped to improve bootstrap support on an unconstrained phylogenetic analysis. RogueNaRok iteratively removes taxa and estimates their impact on bootstrap support; this impact is dependent on the identity of all other taxa removed before it. We therefore excluded all taxa or sets of taxa up to the point where dropping any subsequent taxa would fail to improve bootstrap support by more than 1. A total of 645 species were removed in this manner, with 152 and 102 species removed as part of a 2-species and 3-species set, respectively.

1.5 Tree search with RAxML

We conducted an initial tree search using RAxML v8.1.17 (Stamatakis 2014) using the fast ML search convergence criterion for large trees (option `-D`) and the SEV-based implementation for gap columns (option `-U`, Izquierdo-Carrasco *et al.* 2011). The analysis took approximately 4 days of wall-clock time on a 24-core Intel Xeon E5-2690V3 x2 compute machine.

We then generated individual family-level phylogenies by extracting the subtree descended from the most recent common ancestor of all species in each family, and automatically marked descendent taxa that were from outside the focal family. We then assessed the quality of the phylogeny on a family-by-family basis, and marked any taxa that exhibited rogue behavior (Table 1.3).

We then removed tips that had extremely long branches, as these potentially indicated areas of poor sequence quality or alignment. Using the final filtered dataset, which contained 11,644 tips, we reran a maximum likelihood analysis in RAxML and computed node support values using the SH-like statistic, as it is conservative at estimating support values like standard bootstrapping but runs much faster (Anisimova and Gascuel 2006,

Anisimova *et al.* 2011).

gene	n	gene	n
12s	27	*ptr*	13
16s	84	*rag1*	27
4c4	17	*rag2*	6
coi	175	*rhodopsin*	20
cytb	70	*ripk4*	8
enc1	9	*sh3px3*	8
ficd	14	*sidkey*	8
glyt	1	*sreb2*	4
hoxc6a	7	*svep1*	6
kiaa1239	5	*tbr1*	11
myh6	15	*vcpip*	8
panx2	11	*zic1*	14
plagl2	9	**total sequences**	577

Table 1.3: Gene sequences excluded due to rogue behavior or high identity BLAST matches outside of their species' assigned family.

1.6 Fossil calibrations

We devised an extensive list of fossil-based minima for divergences in actinopterygian phylogeny. Many of these derived from past molecular clock analyses, but others are new to this study. Extinct taxa, along with relevant phylogenetic and age justifications, are supplied in Table 1.4. We applied these fossils as node-based calibrations, with upper age bound specified by a modified implementation of the Whole Tree Extension of the Hedman Algorithm (WHETA, Hedman 2010, Lloyd 2016). This approach yields probabilistic maximum age constraints on given nodes based on: a minimum age specified by the oldest fossil descended from that node; the stratigraphically consistent sequence of older fossil outgroups to that node; and a hard maximum age defined by the investigator.

Concatenated outgroup-age sequences were submitted to the Hedman (2010) algorithm, with a hard upper age constraint of 430 Ma. This choice of maximum age is unlikely to bias our estimates substantially, as we only applied this method for nodes within

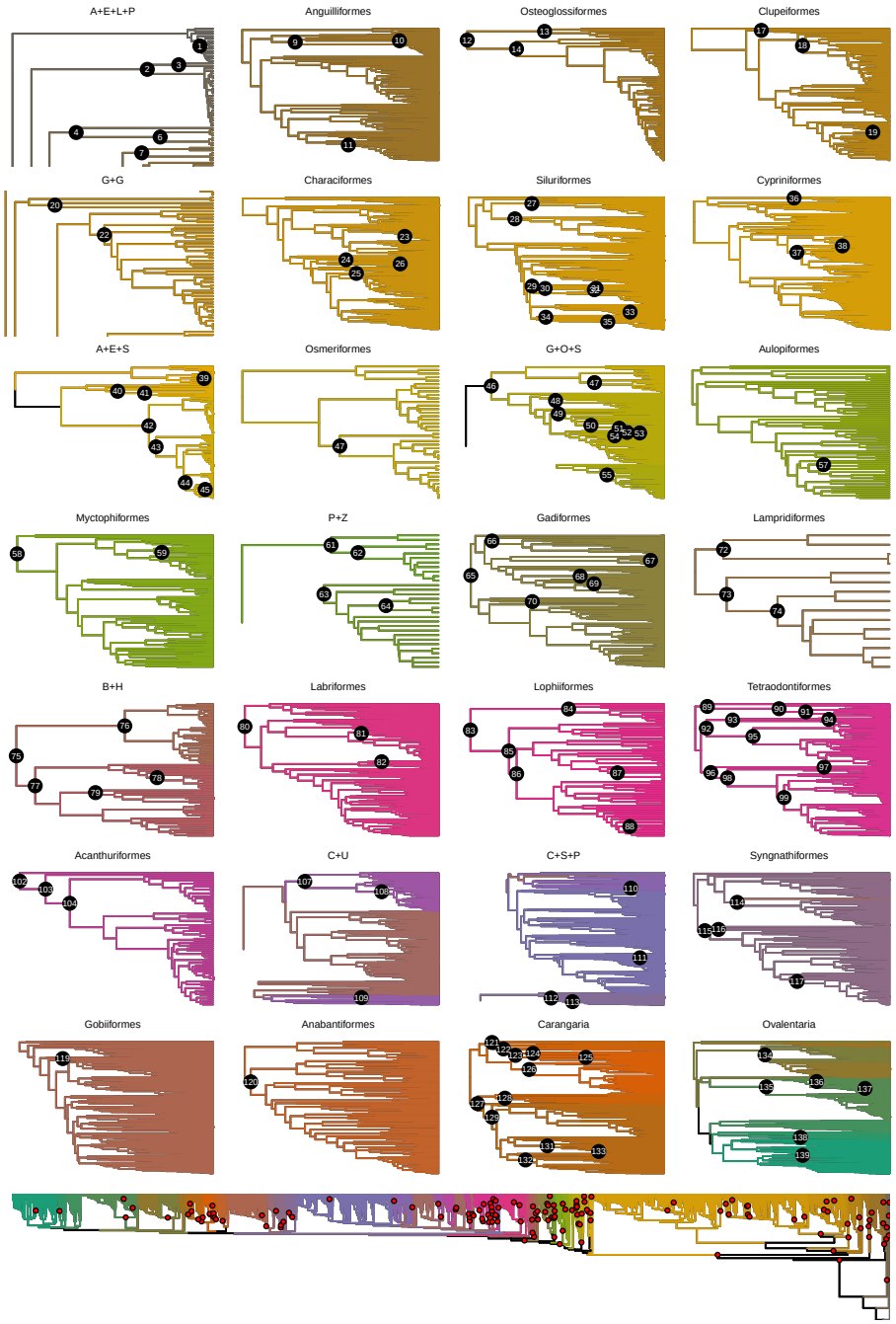

Figure 1.2: Phylogenetic placement of fossil calibrations in major fish lineages. Major lineages are broken into subclades (**top**) to visualize fossil calibrations and are colored by taxonomic order. Numbered nodes correspond to Table 1.4: Fossil Calibrations. The same calibrations are red circles in the full phylogeny (**bottom**). Abbrevations: A+E+S: Argentiniformes, Esociformes, Salmoniformes; G+O+S: Osmeriformes, Galaxiiformes, Stomiatiformes; A+E+L+P: Acipenseriformes, Elopiformes, Lepisosteiformes, Polypteriformes; P+Z: Percopsiformes, Zeiformes; G+G: Gonorynchiformes, Gymnotiformes; C+U: Chaetodontiformes, Uranoscopiformes; C+S+P: Centrarchiformes, Scombriformes, Perciformes; B+H: Beryciformes, Holocentriformes

the actinopteran crown where times of origin are generally accepted to be substantially younger than this Silurian bound. In practice, the credible intervals estimated by the algorithm are relatively insensitive to the choice of the hard maximum age constraint.

Table 1.4: Fossil calibrations used in the new phylogeny

ID	Fossil taxon	Minimum	Maximum
1	*Polypterus faraou*	7	n/a
2	*Protopsephurus luii*	120.8	233.77
3	*Polyodon tuberculata*	63.1	177.68
4	*Watsonulus eugnathoides*	251.2	n/a
5	*Anaethalion zapporum*	151.2	192.78
6	*Arratiaelops vectensis*	126	157.95
7	*Atractosteus falipoui*	93.9	145.37
8	*Baugeichthys caeruleus*	129.4	173.63
9	*Anguilla ignota*	47	120.60
10	*Serrivomer sp.*	12.62	87.40
11	*Echelus branchialis*	53.7	148.62
12	*Paralycoptera wui*	107	159.70
13	*Joffrichthys symmetropterus*	58.551	138.12
14	*Palaeonotopterus greenwoodi*	93.9	142.68
15	*Leptolepides haerteisi*	150.94	177.47
16	*Tischlingerichthys viohli*	150.94	167.14
17	*Trollichthys bolcensis*	49	124.51
18	*Eoengraulis fasoloi*	49	150.44
19	*Dorosoma petenense*	1.8	117.52
20	*Rubiesichthys gregalis*	126.3	155.78
21	*Characiformes*indet.	93.9	143.01
22	*Humboldtichthys kirschbaumi*	7.246	117.69
23	*Megapiranha paranensis*	6	88.33
24	*Lignobrycon ligniticus*	24.5	91.06
25	*Megacheirodon unicus*	24.5	119.07
26	*Salminus noriegai*	7.246	63.10
27	*Corydoras revelatus*	39.5	99.30
28	*Taubateia paraiba*	24.5	75.23
29	*Cetopangasius chaetobranchus*	5.333	94.60
30	*Astephus sp.*	59.36	123.17
31	*Ameiurus pectinatus*	33.97	98.58
32	*Pylodictis olivaris*	16.3	73.03
33	*Brachyplatystoma promagdlaena*	12.8	74.60
34	*Chrysichthys mahengeensis*	45	100.11
35	*Synodontis sp.*	28.1	77.18
36	*Amyzon aggregatum*	48.88	121.74
37	*Cyprinus maomingensis*	40.14	97.506
38	*Huashancyprinus robustispinus*	23.03	73.13

Table 1.4: Fossil calibrations used in the new phylogeny

ID	Fossil taxon	Minimum	Maximum
39	*Macropinna sp*	7.246	107.96
40	*Estesesox foxi*	76.4	139.75
41	*Esox kronneri*	51.57	114.07
42	*Eosalmo driftwoodensis*	51.43	113.95
43	*Hucho sp.*	15.4	85.95
44	*Oncorhynchus ('Smilodonichthys') rastrosus*	8.2	60.97
45	*Oncorhynchus keta*	4.8	40.55
46	*Paravinciguerria praecursor*	93.9	142.01
47	*Speirsaenigma lindoei*	56.83	118.46
48	*Sigmops sp.*	12.62	88.65
49	*Polypnoides laevis*	41.3	116.39
50	*Argyropelecus sp.*	32.02	91.37
51	*Argyropelecus logearti*	12.62	66.09
52	*Chauliodus testa*	7.246	87.64
53	*Chauliodus sloani*	2.588	59.91
54	*Stomias affinis*	5.33	60.59
55	*Galaxias effusus*	23	133.81
56	*Apateodus glyphodus*	103.13	159.40
57	*Alepisaurus 'ferox'*	15.97	132.84
58	*Eomyctophum koraense*	32.02	116.47
59	*Bolinichthys sp.*	5.33	87.37
60	*Homonotichthys dorsalis*	93.6	125.92
61	*Massamorichthys wilsoni*	63.1	94.31
62	*Trichophanes foliarum*	33.07	78.36
63	*Cretzeus rinaldii*	69.71	108.29
64	*Zenopsis clarus*, *Zenopsis tyleri*, and *Zenopsis hoernesi*	32.02	90.18
65	*Rhinocephalus planiceps*	53.7	92.87
66	*Nezumia lindsayi*	41.3	77.48
67	*Merluccius cf. merluccius*	5.333	43.59
68	*Gaidropsarus pilleri*	13.53	59.80
69	*Gadiculus cf. jonas*	5.333	43.59
70	*Bregmaceros filamentosus*	41.3	77.48
71	*Aipichthys velifer*	98	142.86
72	*Turkmene finitimus*	54.17	119.25
73	*Eolophotes lenis*	41.3	96.16
74	*Trachipterus mauritanicus*	5.333	70.37
75	*Stichocentrus liratus*	98	126.91
76	*Berybolcensis leptacanthus*	49	106.58
77	*Hoplopteryx lewesensis*	93.6	114.14
78	*Gephyroberyx robustus*	32.02	97.35
79	*Miobarbourisia aomori*	9.83	96.05
80	*Phyllophyarngodon longipinnis*	49	79.89
81	*Calotomus priesli*	13.53	48.13
82	*Tautoga sp.*	15	62.77

Table 1.4: Fossil calibrations used in the new phylogeny

ID	Fossil taxon	Minimum	Maximum
83	*Caruso brachysomus*	49	79.89
84	*Eosladenia caucasica*	38	66.25
85	*Tarkus squirei*	49	68.45
86	*Eophryne barbutii*	49	60.82
87	*Oneiroides sp.*	7.42	50.56
88	*Antennarius monodi*	5.333	49.96
89	*Ctenoplectus williamsi*	53.7	80.82
90	*Eolactoria sorbinii*	49	69.59
91	*Oligolactoria bubiki*	30.28	57.88
92	*Eospinus daniltshenkoi*	54.17	81.25
93	*Protacanthodes nimesensis*	49	69.59
94	*Carpathospinosus propheticus*	26.93	57.59
95	*Oligobalistes robustus*	32.02	66.67
96	*Balkaria histiopterygia*	55.8	94.38
97	*Austromola angerhoferi*	21.12	75.55
98	*Heptadiodon echinus*	49	79.89
99	*Archaeotetraodon winterbottomi*	32.02	65.48
100	*Eoscatophagus frontalis*	49	62.72
101	*Siganopygaeus rarus*	54.17	70.72
102	*Luvarus necopinatus*	54.17	81.25
103	*Eozanclus brevirostris*	49	69.59
104	*Proacanthurus tenuis*	49	61.58
105	*Malacanthus carosii*	13.53	39.37
106	*Lopholatilus chamaeleonticeps*	13.82	51.73
107	*Chaetodontidae indet.* (Tholichthys larval stage)	29.62	66.09
108	*Chaetodon ficheuri*	5.333	50.81
109	*Astroscopus countermani*	7.246	38.13
110	*Archoplites clarki*	15.4	52.33
111	*Gasterosteus cf. wheatlandi*	13.1	39.39
112	*Argestichthys vysotzkyi*	54.17	94.42
113	*Eocoelopoma portentosum*	54.17	80.88
114	*Eochampsodon elongatus*	38	92.15
115	*Gasterorhamphosus zuppichinii*	69.71	109.38
116	*Gerpegezhus paviai*	55.8	94.38
117	*Hippocampus samarticus*	49	79.89
118	*Carlomonnius quasigobius*	49	93.62
119	*Lepidocottus aries*	23.03	74.76
120	*Anchichanna kuldanensis*	41.3	78.65
121	*Eolates gracilis*	49	79.89
122	*Mene purdyi*	55.2	94.59
123	*Ductor vestenae*	49	68.45
124	*Oligoremora rhenana*	29.62	56.87
125	*Scomberoides spinosus*	19.3	55.47
126	*Eastmanalepes primaevus*	49	79.89

Table 1.4: Fossil calibrations used in the new phylogeny

ID	Fossil taxon	Minimum	Maximum
127	*Heteronectes chaneti*	49	79.89
128	*Sphyraena bolcensis*	49	68.45
129	*Eobothus minimus*	49	68.45
130	*Oligopleuronectes germanicus*	29.62	50.05
131	*Oligobothus pristinus*	29.62	50.05
132	*Eubuglossus eocenicus*	41.2	59.10
133	*Bothus sp.*	11.056	41.25
134	*Palaeopomacentrus orphae*	49	68.45
135	*Mahengechromis spp*	45	67.73
136	*Gymnogeophagus eocenicus*	39.5	57.87
137	*Nandopsis woodringi*	3.6	45.85
138	*Ramphexocoetus volans*	49	79.89
139	*Francolebias aymardi*	28.1	64.70

The final time-calibrated phylogeny is shown in Figure 1.1, and a breakdown of where fossil calibrations are placed on the phylogeny are Figure 1.2.

1.7 Placing unsampled species

We compared the taxonomic classification across Fishbase (Froese and Pauly 2014), the Catalog of Fishes (Eschmeyer *et al.* 2017), and the Euteleost Tree of Life project (Betancur-R *et al.* 2013). Based on these taxonomic authors, we built a new classification scheme and explored shallower phylogenetic groups where non-monophyly was found in our phylogeny. This combined "Phylogenetic Fish Classification" was then used for the purpose of taxonomic back-filling of taxa without molecular data or those that were removed during the curation stage. Using the time-calibrated phylogeny as a backbone, we generated a distribution of trees where missing taxa were placed according to our PFC taxonomy.

For each of the unsampled species of ray-finned fish, we assigned the most restrictive taxonomic rank (e.g., genus, family, order) that was recovered as monophyletic in our maximum likelihood phylogeny. We computed rank-specific estimates of the speciation and extinction rate under a constant rate model, conditioned on the sampling fraction

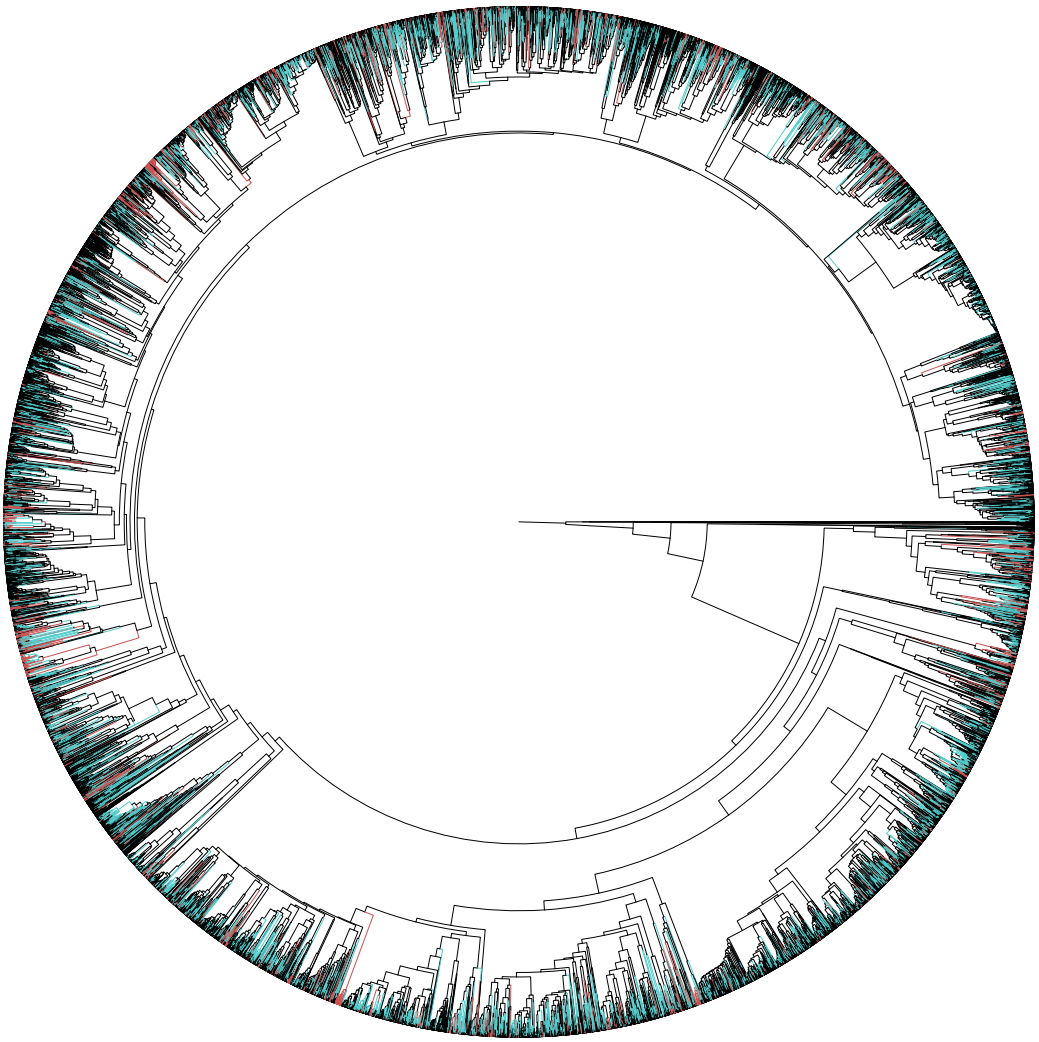

Figure 1.3: One realization of the all-taxon assembled (ATA) phylogeny. Black edges indicate lineages that were inferred using genetic data, blue indicate single species that were placed taxonomically, and red indicates entire clades that were placed taxonomically.

(Stadler 2009), and used these rates to generate waiting times for unsampled species. However, if the taxonomic node had fewer than 3 tips, or if the probability of sampling the crown age of that node given the number of sampled taxa (Sanderson 1996) was less than 0.8, we searched all the ancestors of that taxonomic node that fulfilled the previous criteria. The generated waiting times were bounded between the crown age of that clade and the present time ($t = 0$). However, if the crown capture probability was less than 0.8, the maximum generated age was extended to the stem age of the taxonomic node. If placement was impossible due to monophyletic constraints (see below), the waiting time was then bounded between the stem age and the crown age of that taxonomic node.

These waiting times were used to randomly attach unsampled species to an existing branch within their assigned taxonomic rank, as long as these new species did not break the monophyly of nodes that were recovered as monophyletic and assigned a taxonomic rank, and constrained to not produce negative branch lengths due to a child node being added that was older than a parent node. If all of the child branches of a taxonomic node belonged to a monophyletic node, or if the crown capture probability was less than 0.8, the new species was instead assigned to the stem of that clade.

This procedure is similar to stochastic polytomy resolution as implemented in PASTIS (Thomas *et al.* 2013), but permits construction of extremely large phylogenies using all molecular data in a single analysis, rather than a two-stage process that begins with a reduced backbone dataset followed by separate tree searches for each crown lineage that jointly estimate the placement of species with and without molecular data. Additionally, our procedure produces a local estimate of diversification rate at every taxonomic rank, rather than computing a single rate at the rank at which the crown lineages will be grafted onto the backbone phylogeny. This permits a more accurate placement of unsampled taxa as diversification rate heterogeneity below the order or family level might significantly bias the inferred waiting times.

These functions were all implemented in a custom Python script based on the code from the R packages TreePar and SimTree (Stadler 2009, 2011a,b). This procedure was repeated 100 times to generate a distribution of fully-sampled ray-finned fish phylogenies,

which we term as the all-taxon assembled (ATA) trees (Figure 1.3). Our distribution of ATA phylogenetic trees of ray-finned fishes contained 31,526 species.

1.8 Estimating diversification rates

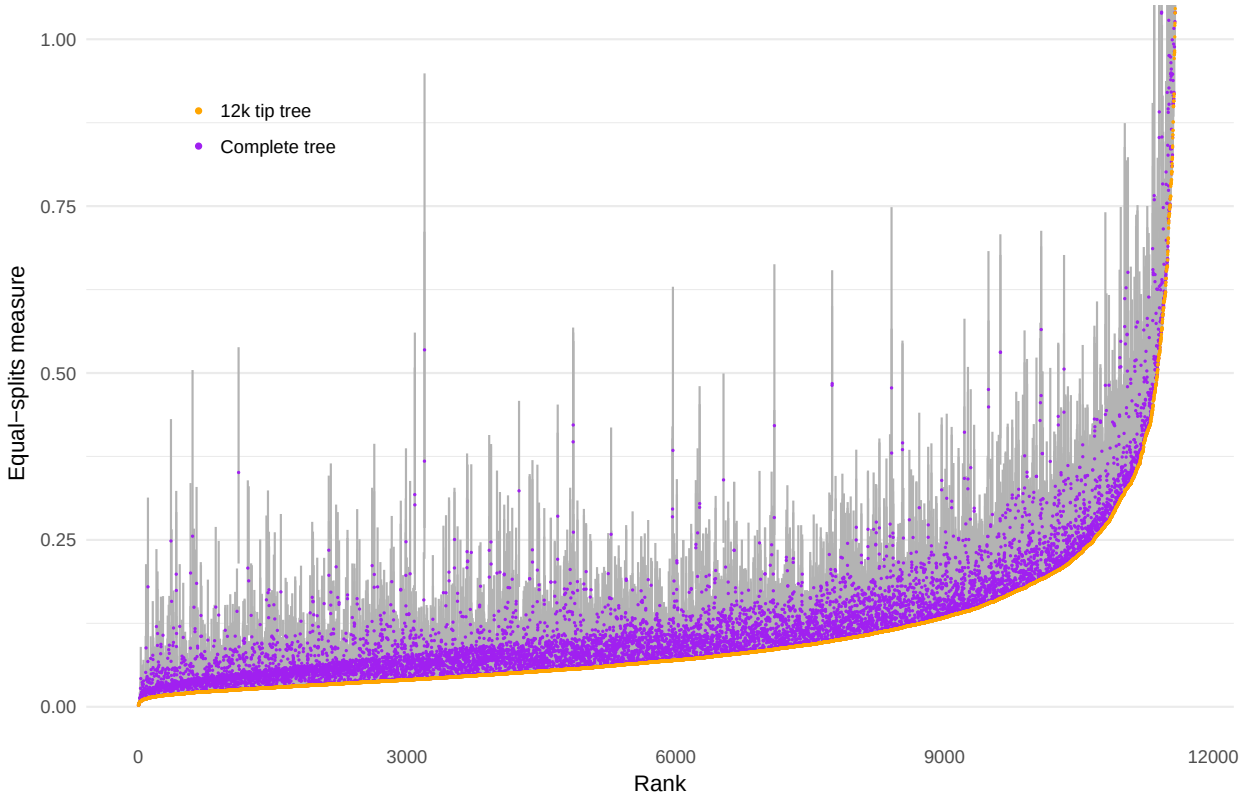

Figure 1.4: Estimates of the equal-splits (DR) rates measure compared between the molecular phylogeny (orange) and complete ATA phylogeny (purple). The points for the complete phylogeny represent the median of all DR calculations conducted on the 100 ATA phylogenies; the grey bars indicate the interquartile range of DR rate estimates.

We estimated speciation rates across the ATA phylogenies using DR (Jetz *et al.* 2012, Equation 1.1), a summary statistic that infers recent speciation rates for all tips in the phylogeny without requiring a formal parametric inference model:

$$DR = \left(\sum_{j_1}^{N_1} l_j \frac{1}{2^{j-1}} \right)^{-1} \tag{1.1}$$

where, for any given species, N is the number of branches from root to tip, j is the depth

of the branch, and l_j is the length of the branch j. DR can be intuitively described as the "splitting rate" of a tip, with the contribution of splits farther in the past decaying exponentially.

Taxonomically-placed species that lack genetic data may bias inference in certain scenarios, particularly when considering hypotheses of trait evolution (Rabosky 2016). However, when estimating the DR statistic, bias is actually reduced as DR requires an accurate estimate of the number of nodes from the tips to the roots. Adding unsampled taxa increases the DR statistic for approximately two-thirds of species with molecular information, improving the estimates of speciation rate for large, incompletely sampled phylogenies (Figure 1.4). Furthermore, uncertainty in the placement algorithm, expressed as the paraphyly of the group of interest, will cause the species that the algorithm is trying to place to be assigned to the next highest monophyletic rank available. In practice, this tends to place species in more inclusive (and therefore older) groups, having the ultimate effect of diluting any signal of atypically-fast speciation rates. The placement of these unsampled taxa are therefore conservative for the purposes of diversification analyses.

1.9 Results and Discussion

In this study, we have improved on previously-published phylogenies of ray-finned fishes by nearly doubling the previous extent of taxon sampling (Rabosky *et al.* 2013). Furthermore, we have leveraged taxonomic information to generate a complete distribution of taxonomically informed species placements, for a complete fish tree of life similar to efforts in the birds (Jetz *et al.* 2012). Our improved phylogeny incorporates more fossil calibrations and more sequences; these new sequences represent a significant advance on the state-of-the-art as we require no monophyletic calibrations and let the data fully inform the branching relationships in our phylogeny. In subsequent chapters, I will discuss the taxonomically informed species placement algorithm in detail, and present new work showcasing our fish tree of life via a web portal.

1.10 References

Andre J. Aberer, Denis Krompass, and Alexandros Stamatakis. Pruning rogue taxa improves phylogenetic accuracy: An efficient algorithm and webservice. *Systematic Biology*, 62(1):162–166, 2013.

Maria Anisimova and Olivier Gascuel. Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology*, 55(4):539–552, aug 2006.

M. Anisimova, M. Gil, J.-F. Dufayard, C. Dessimoz, and O. Gascuel. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Systematic Biology*, 60(5):685–699, oct 2011.

Ricardo Betancur-R, Richard E. Broughton, Edward O. Wiley, Kent Carpenter, J. Andrés López, Chenhong Li, Nancy I. Holcroft, Dahiana Arcila, Millicent Sanciangco, James C. Cureton, Feifei Zhang, Thaddaeus Buser, Matthew A. Campbell, Jesus A. Ballesteros, Adela Roa-Varon, Stuart Willis, W. Calvin Borden, Thaine Rowley, Paulette C. Reneau, Daniel J. Hough, Guoqing Lu, Terry Grande, Gloria Arratia, and Guillermo Ortí. The Tree of Life and a New Classification of Bony Fishes. *PLoS Currents*, 2013.

Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

Robert C. Edgar. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, jan 2004.

W. N. Eschmeyer, R. Fricke, and R. van der Laan. Catalog of Fishes: genera, species, references, 2017.

R Froese and Daniel Pauly. FishBase, 2014.

Matthew M. Hedman. Constraints on clade ages from fossil outgroups. *Paleobiology*, 36(1):16–31, 2010.

Fernando Izquierdo-Carrasco, Stephen A Smith, and Alexandros Stamatakis. Algorithms, data structures, and numerics for likelihood-based phylogenetic inference of huge trees. *BMC Bioinformatics*, 12(1):470, 2011.

W Jetz, G.H. H Thomas, J.B. B Joy, K. Hartmann, and A.O. O Mooers. The global diversity of birds in space and tim. *Nature*, 491(7424):1–5, 2012.

Kazutaka Katoh and Daron M. Standley. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30:772–780, 2013.

Graeme T. Lloyd. Estimating morphological diversity and tempo with discrete character-taxon matrices : implementation , challenges , progress , and future directions. *Biological Journal of the Linnean Society*, (1998), 2016.

Thomas J. Near, Alex Dornburg, Ron I Eytan, Benjamin P Keck, W leo Smith, Kristen L Kuhn, Jon A Moore, Samantha A. Price, Frank T Burbrink, Matt Friedman, and Peter C. Wainwright. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. *Proceedings of the National Academy of Sciences*, 110:12738–12743, 2013.

Daniel L. Rabosky, Francesco Santini, Jonathan Eastman, Stephen A Smith, Brian Sidlauskas, Jonathan Chang, and Michael E. Alfaro. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nature Communications*, 4:1958, jun 2013.

Daniel L Rabosky. No substitute for real data: a cautionary note on the use of phylogenies from birth–death polytomy resolvers for downstream comparative analyses. *Evolution*, 69(12):3207–3216, 2016.

John W Ratcliff and David Metzener. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13:46–72, 1988.

S Ratnasingham and PDN Hebert. Barcoding BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, 7:355–364, 2007.

M. J. Sanderson. How Many Taxa Must Be Sampled to Identify the Root Node of a Large Clade? *Systematic Biology*, 45(2):168–173, jun 1996.

Tanja Stadler. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261(1):58–66, 2009.

Tanja Stadler. Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences*, 108(15):6187–6192, 2011.

Tanja Stadler. Simulating trees with a fixed number of extant species. *Systematic Biology*, 60(5):676–684, 2011.

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

Ole Tange. GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine*, 36(1):42–47, feb 2011.

Gavin H Thomas, Klaas Hartmann, Walter Jetz, Jeffrey B Joy, Aki Mimoto, and Arne O Mooers. Pastis: an r package to facilitate phylogenetic assembly with soft taxonomic inferences. *Methods in Ecology and Evolution*, 4(11):1011–1017, 2013.

Methods in Ecology and Evolution

Methods in Ecology and Evolution 2016, **7,** 472–482 doi: 10.1111/2041-210X.12508

Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data

Jonathan Chang[1]* and Michael E. Alfaro[1]

[1]*Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA, USA*

Summary

1. Advances in genomics and informatics have enabled the production of large phylogenetic trees. However, the ability to collect large phenotypic data sets has not kept pace.

2. Here, we present a method to quickly and accurately gather morphometric data using crowdsourced image-based landmarking.

3. We find that crowdsourced workers perform similarly to experienced morphologists on the same digitization tasks. We also demonstrate the speed and accuracy of our method on seven families of ray-finned fishes (Actinopterygii).

4. Crowdsourcing will enable the collection of morphological data across vast radiations of organisms and can facilitate richer inference on the macroevolutionary processes that shape phenotypic diversity across the tree of life.

Key-words: Actinopterygii, comparative methods, large-scale annotation, macroevolution, Mechanical Turk

Introduction

Integrating phenotypic data, such as anatomy, behaviour, physiology and other traits, with phylogenies is a powerful strategy for investigating the patterns of biological evolution. Recent advances in next-generation sequencing (Meyer, Stenzel & Hofreiter 2008; Shendure & Ji 2008) and sequence capture technologies (Faircloth *et al*. 2012; Lemmon, Emme & Lemmon 2012) have made phylogenetic inference of large radiations of organisms possible (McCormack *et al*. 2012, 2013; Faircloth *et al*. 2013, 2015). However, similar breakthroughs for generating new phenotypic data sets have been comparatively uncommon, likely due to the high expense and effort required (reviewed in Burleigh *et al*. 2013).

Creating these large phenotypic data sets has generally required an extended dedicated effort of measuring and describing morphological or behavioural traits that are then coded into a comprehensive data matrix. One such example is the Phenoscaping project (http://kb.phenoscape.org; Deans *et al*. 2015), and related efforts in the Vertebrate Taxonomy Ontogeny (Midford *et al*. 2013) and Hymenoptera Anatomy Ontology (Yoder *et al*. 2010), which require large amounts of researcher effort to collate. Other approaches include using machine learning (Dececchi *et al*. 2015), machine vision (Corney *et al*. 2012a, b) or natural language processing (Cui 2012) to identify or infer phenotypes. These statistical techniques function ideally with either a large training data set (e.g., a predefined ontogeny data base) or a complex model (Brill 2003; Halevy, Norvig & Pereira 2009; Hastie, Tibshirani & Friedman 2009), both of which also require intensive researcher effort to build and validate. Finally, methods such as high-throughput infrared imaging, mass spectrometry and chromatography have been successfully used in plant physiology (Furbank & Tester 2011) and microbiology (Skelly *et al*. 2013), but these methods may not be applicable for zoological researchers. These approaches all share a similar goal of collecting large comparative data sets, but also require large investments in researcher effort. This bottleneck in researcher availability has limited the scope of work in comparative biology.

Although it is now possible to build phylogenetic trees with thousands of tips, and phenotypic data sets have similarly been growing larger and larger, studies at this scale tend to be limited to a few broad types of traits, including geographic occurrences (Jetz *et al*. 2012), one or two continuous characters (Harmon *et al*. 2010; Rabosky *et al*. 2013), a single discrete character (Goldberg *et al*. 2010; Aliscioni *et al*. 2012; Price *et al*. 2012), or some combination of these (Pyron & Burbrink 2014; Zanne *et al*. 2014). Most morphological evolutionary studies are constrained by a fundamental trade-off in effort. Although the collection of detailed phenotypic measurements is often required to fully analyse complex form–function or ecology–phenotype relationships (Schluter 2000; Alfaro, Bolnick & Wainwright 2004 2005; Wainwright *et al*. 2005; Collar & Wainwright 2006; Price *et al*. 2010; Frédérich *et al*. 2013), rich methods of data collection such as computed tomography (CT) scanning are time intensive and do not permit easy scaling to hundreds or thousands of species. Analysis of more complex traits at this scale has the potential to greatly enrich our understanding of macroevolutionary processes, by permitting more refined hypothesis testing.

*Correspondence author: E-mail: jonathan.chang@ucla.edu

Here, we present a method and toolkit to efficiently collect two-dimensional geometric morphometric phenotypic data at a high-throughput 'phenomic' scale. We developed a novel web browser-based image landmarking application and use Amazon Mechanical Turk (https://www.mturk.com) to distribute digitization tasks to remote workers (hereafter *turkers*) over the Internet, who are paid for their contributions. We evaluate the accuracy and precision of turkers by assigning identical image sets and digitization protocols to users who are experienced with fish morphology (hereafter *experts*), and compare the inter- and intra-observer differences between turkers and experts. To illustrate the efficiency of this approach, we construct a phylogenetic analysis pipeline to download photographs and phylogenies of seven actinopterygiian families from the web, collect Mechanical Turk shape results, analyse the body shape evolution using BAMM (Rabosky 2014) and compare the time required for this workflow to traditional approaches. Although we focus on collecting two-dimensional geometric morphometric data, we address the challenges that will be common to all studies that crowdsource phenotypic data. We also discuss the role that crowdsourcing is best suited in large-scale morphological analyses, and suggest ways to integrate crowdsourced data as part of larger initiatives to digitize biodiversity.

Materials and methods

AMAZON MECHANICAL TURK

Amazon Mechanical Turk ('MTurk') is a web-based service where Requesters can request work, known as Human Intelligence Tasks ('HITs') to be performed by Workers. Workers submit the tasks over the Internet, where Requesters review the completed work, and, if they are satisfied with the results, accept the work and pay the Worker (for a detailed overview, see Mason & Suri 2012). We use MTurk as a platform to distribute our geometric morphometric tasks and financially compensate the worker accordingly. Scientific collection of data over MTurk and similar services has generally been limited to the fields of psychology and computer science, and there have been few attempts to crowdsource biological trait data (Burleigh *et al.* 2013).

WEB-BASED GEOMETRIC MORPHOMETRICS

We developed an geometric morphometric digitization application that runs completely on the user's local web browser, using the HTML5 Canvas interface. This simplifies the infrastructure challenge of needing to serve many crowdsourced workers simultaneously, since workers will not need to download desktop software such as tpsDig (http://life.bio.sunysb.edu/ee/rohlf/software.html) before generating data. The web application is configured with a JavaScript Object Notation (JSON) file that describes the landmarks necessary to complete an image digitization task (Fig. S1). Point landmarks, semilandmark curves and linear measurements are all supported. The software is available at https://github.com/jonchang/eol-mturk-landmark.

Although digitizing and landmarking a single image (microtasks *sensu* Good & Su 2013) is effective for high-throughput work on MTurk, it is unsuitable for conducting controlled experiments. To solve this issue, we also created a server-side application backend that automatically distributes tasks according to a configurable set of images and experimental protocol. This application mimics an official Amazon Mechanical Turk interface endpoint, to facilitate drop-in replacement for an existing MTurk workflow. External non-MTurk workers can also participate in the same experiment, ensuring consistent comparisons across separate groups. The software is available at https://github.com/jonchang/fake-mechanical-turk.

RELIABILITY ANALYSIS

Collecting landmark-based geometric morphometric data at a broad scale permits detailed analysis of different sources of error, such as among- and within-observer variation (Von Cramon-Taubadel *et al.* 2007). To assess whether the quality of data gathered by workers recruited through Amazon Mechanical Turk was significantly different than traditionally collected data, we asked turkers (n = 21) and experts (n = 8) to landmark a set of five fish images, five times each. Turkers were compensated $25 for the entire task. All participants used the same protocol (Appendix S2) and same software to digitize the same set of fishes (Tables S1 and S2). The landmarks were carefully selected based on previously published literature concerning fish shape (Fig. S2; Fink & Zelditch 1995; Cavalcanti, Monteiro & Lopes 1999; Rüber & Adams 2001; Klingenberg, Barluenga & Meyer 2003; Chakrabarty 2005; Frédérich *et al.* 2008; Claverie & Wainwright 2014; Thacker 2014). We also ensured that the chosen landmarks included morphological features that were relatively straightforward to digitize (e.g. the position of the eye) and features that were likely to be more challenging to digitize (e.g. the most anterior and most dorsal points of the preopercle), in order to test for turker and expert differences over a spectrum of difficulties. We report the interobserver reliability for turkers and experts by computing the ratio of the among-individual and the sum of the among-individual and measurement error variance components in a repeated measures nested MANOVA (Palmer & Strobeck 1986; Zelditch, Swiderski & Sheets 2012). To test whether workers were consistently measuring the same shape, we examined the per-worker consistency, as estimated by the morphological disparity (Procrustes variance; Zelditch, Swiderski & Sheets 2012) of each worker's measured shapes. We then summarized the consistency within groups and compared the median consistency of turkers and experts. To determine whether turkers improved with experience, we excluded the first three images that turkers worked on, and calculated the distance between their mean shape and the mean shape of experts. We then repeated this, but without excluding the first three images that turkers digitized. To determine whether turkers worked faster with experience, we compared the time it took turkers to complete their first image compared to their fifth image.

To assess the differences between turker and experts on a per-landmark basis, we first compared for each landmark the median position of all turkers to the median position of all experts. We assumed that the expert median was the true position of that landmark, and calculated the absolute Euclidian distance in pixels. Larger distances would indicate low turker accuracy, while smaller distances would indicate high turker accuracy. Because the specimens digitized in this study varied in size, we also report turker accuracy as both distance in millimetres and as a fraction of the specimen's total length (TL). We then examined the variance in turker landmarks. For each landmark, we rotated the cloud of points to maximize variance in one dimension, and calculated the log-ratio of median absolute deviations (MAD) between turkers and experts. This rotation is a conservative approach for assessing the difference in variance between these two groups, because it maximizes any apparent differences in landmark position. A positive log-ratio indicated that experts had lower variance than turkers, while a negative log-ratio indicated that turkers had lower variance. For all subsequent

analysis, we excluded landmarks where turkers performed especially poorly, where either the accuracy or precision components for a given landmark exceeded 1·5 times the interquartile range of that component.

To determine whether turkers and experts were statistically distinguishable, we performed a nonparametric MANOVA using the randomized residual permutation procedure (RRPP) with 1000 iterations (Collyer, Sekora & Adams 2015). The RRPP method reduces the effect of the 'curse of dimensionality' ($P >> n$, where the number of predictors greatly exceeds the number of observations), a common problem in geometric morphometrics, and has been shown to have increased statistical power compared to a method where the raw data are randomized instead (Anderson & Braak 2003). We test for a difference between mean turker and expert shapes against a null model of no difference between turker and expert changes, taking into account species-specific differences. A difference between models was considered significant if the P-value was less than $\alpha = 0{\cdot}05$.

As a separate test, we use linear discriminant analysis (LDA, Ripley 1996), a statistical classification algorithm that finds features to differentiate between different classes of data, in this case turkers and experts. We assessed the accuracy of the LDA classification using 10-fold cross validation (CV), which splits our data into 10 equally sized groups, using nine for training and one for validation (Kohavi 1995; Hastie, Tibshirani & Friedman 2009). An acceptable misclassification rate varies depends on application, but here we use a 25% misprediction rate as a standard for sufficient accuracy. This is a highly forgiving standard, since a 50% misprediction rate is no better than a coin flip, and a 25% misprediction rate would still erroneously classify one in four turkers as experts or vice versa. We also use quadratic discriminant analysis (QDA), which relaxes some of the assumptions of LDA, and similarly report the QDA misclassification rate.

We calculated the per-individual median shape for each species used, as well as the consensus turker and morphologist shapes, and projected these shapes into Procrustes space, to visualize the orthogonalized differences in median shape among and between the types of digitizers.

EXAMPLE: A PHENOMIC PIPELINE FOR COMPARATIVE PHYLOGENETIC ANALYSIS

A common strategy in fish comparative studies is to examine evolutionary dynamics within a single family (Ferry-Graham *et al.* 2001; Alfaro, Bolnick & Wainwright 2005; Alfaro, Santini & Brock 2007; Rocha *et al.* 2008; Hernandez, Gibb & Ferry-Graham 2009; Dornburg *et al.* 2011; Frédérich *et al.* 2013; Santini, Sorenson & Alfaro 2013; Sorenson *et al.* 2013; Claverie & Wainwright 2014; Thacker 2014), potentially due to the extensive amount of time necessary to collect data. To demonstrate the utility of obtaining comparative data using our method, we use previously published phylogenies for seven fish families: Acanthuridae (Sorenson *et al.* 2013), Balistoidae, Tetraodontidae (Santini, Sorenson & Alfaro 2013), Apogonidae, Chaetodontidae, Labridae (Cowman & Bellwood 2011; Choat *et al.* 2012), and Pomacentridae (Frédérich *et al.* 2013). We match species in these phylogenies to left-lateral images from the Encyclopedia of Life (http://eol.org/) using their application programming interface (Table S5; Parr *et al.* 2014). Crowdsourced workers placed landmarks describing body shape variation following a standard protocol (Appendix S2) and were compensated \$0·15 per completed image.

To test whether our method could be faster than a single expert digitizing a data set, we extrapolated the time it would take for a single expert to measure all images at 1× replication, based on the average time an expert took to digitize a single image. We compared this predicted measurement time to the total time required for turkers to complete all digitization tasks at 5× replication, from initial upload to final submission. If the turkers in aggregate annotated images more quickly than a single expert would have, this suggests that the parallelization afforded by crowdsourcing is effective at reducing the total time required for data collection.

The Cartesian position of turker-collected landmarks was used in a generalized Procrustes analyses (Gower 1975; Rohlf & Slice 1990), which centres, scales and rotates landmark configurations to minimize the least-squares distance between shapes. We then determined the major components of shape variation using a Procrustes-aligned principal components analysis (PCA) (Mardia, Kent & Bibby 1979; Bookstein 1991) with the R package *geomorph* (Adams & Otarola-Castillo 2013), and retain the principal component axes whose eigenvalues exceeded the corresponding random broken-stick component (Jackson 1993; Legendre & Legendre 1998) for all subsequent analyses.

To illustrate the potential of how crowdsourcing could be integrated into an pipeline that could allow rapid collection and analysis of phenotypic data, we used Bayesian Analysis of Macroevolutionary Mixtures (BAMM; Rabosky 2014) to estimate rates of body shape evolution for all seven families. BAMM estimates the location of rate shifts in character evolution using a transdimensional (reversible jump) Markov Chain Monte Carlo method that samples a variety of models of trait evolution. Any missing trait data is treated as a latent variable in the analysis. We assessed convergence and mixing using Tracer (Rambaut *et al.* 2014). We also repeated each analysis and simulated under the prior (without data) to exclude rate heterogeneity that occurred solely due to stochastic processes. We use a Bayes Factor criterion of $BF > 5$ to enumerate the set of credible shifts (Shi & Rabosky 2015) and visualized them using BAMMtools (Rabosky *et al.* 2014).

Results

RELIABILITY ANALYSIS

For nearly 90% of the points measured, turkers differed from the expert consensus by less than 30 pixels, with half of all landmarks having less than 3 pixels of difference (10 px = 0·68–4·2 mm, 1·3–1·5% TL, Figs 1 and S3, Table S1). The most accurate and precise points are those that are related to the position of the eye (landmarks E1 and E2). The least accurate are those in the opercular series (O1–O5), particularly the ones related to the preopercle (O1–O3) likely because in certain groups (e.g. Tetraodontidae) the preopercle is difficult to visualize from external morphology alone. Experts were generally more precise than turkers; however, there were some landmarks where the turkers converged on very similar locations. Based on these results, we exclude in subsequent analyses the landmarks relating to the distal margins of all fins (A3, A4, P3, P4, D3, D4), the preopercle bones (O1–O3), the dorsal fin for triggerfishes (D1, D2) and the opercular opening for pufferfishes (O4–O5), due to low turker accuracy.

The interobserver reliability of turkers and experts as measured by the ratio of among-individual and sum of the among-individual and measurement error ANOVA components was 96·4% and 90·9%, respectively. Although there is no current standard for acceptable levels of measurement reliability (Von Cramon-Taubadel *et al.* 2007), these percentages are not low enough to suggest weaknesses in the measurement protocol.

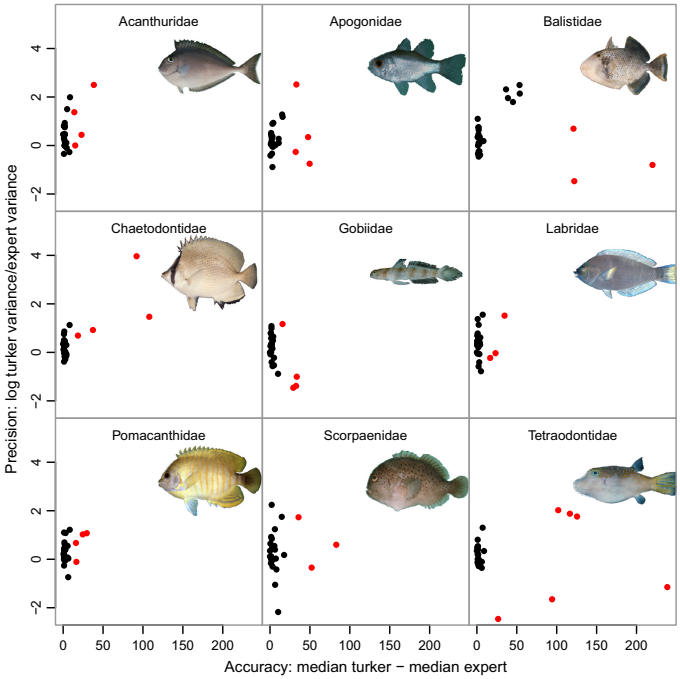

Fig. 1. Per-family breakdown of accuracy vs. precision for each landmark. Accuracy is represented as the difference between the median turker location for that landmark and the median expert location, with the expert location assumed to be the true location. Precision is represented as the log-ratio of median absolute deviations between turkers and experts. More positive numbers indicate better expert precision, whereas more negative numbers indicate better turker precision. Points highlighted in red are those determined to be outliers (1·5 × IQR). A labelled version of this figure is available as Fig. S3. Photo credit J.E. Randall (used with permission under a CC-BY-NC 3.0 licence).

Table 1. Misprediction rate of linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) with 10-fold cross validation for each fish image. The discriminant model for each family was unable to meet the standard of one in four misclassifications, and in some cases, the more flexible QDA method performed worse than the LDA model

Family	LDA	QDA
Acanthuridae	0·504	0·428
Apogonidae	0·450	0·472
Balistidae	0·444	0·411
Chaetodontidae	0·400	0·422
Gobiidae	0·481	0·462
Labridae	0·389	0·389
Pomacanthidae	0·462	0·431
Scorpaenidae	0·504	0·472
Tetraodontidae	0·455	0·460

Turkers were less consistent than the average expert (Table S3); however, the overall difference in consistency between turkers and experts was generally quite small. We did not find evidence that turkers improved over time. Excluding the first three images did not markedly change turkers' performance compared to experts (Table S4). Turkers took extra time to complete their first task, with a median completion time of 8·93 min, compared to 2·43 min on their fifth task.

The nonparametric MANOVA with RRPP failed to detect a significant difference between turker and expert shapes (P = 0·394, Z = 1·0067363, F = 0·9938314). Similarly, both linear and quadratic discriminant analysis with 10-fold cross validation (Table 1) were unable to reliably distinguish between these two groups, for any given family. Although for some images the classifier showed slight improvement beyond a 50% coin flip, in all cases our model fell short based on a one in four (25%) acceptable misclassification rate. We conclude that, for any given sample of landmarks, it is challenging to statistically distinguish between expert-provided and turker-provided landmark configurations.

We projected turker and expert shape configurations into morphospace (Figs 2 and S4). Although the overall space occupied by each family's shape configurations varies, the aggregated median turker and expert shapes are not qualitatively different. The only exception is the triggerfishes (Balistidae), likely due to turker confusion over the exact location of dorsal fin due to their reduced anterior dorsal fin.

PHENOMIC PIPELINE FOR COMPARATIVE PHYLOGENETIC ANALYSIS

We were able to match 147 of 950 species to images in EOL's data base (Acanthuridae: 8/45, Apogonidae: 19/86, Balistoidae: 23/86, Chaetodontidae: 12/103, Labridae: 31/316, Pomacentridae: 30/208, Tetraodontidae: 24/106). Due to the low number of images matched for acanthurids, apogonids and chaetodontids, we focused on the other four families with better taxon sampling for the comparative BAMM analysis.

At 5× replication, 19789s (*c.* 5·5 h) elapsed between initial upload of the task to Amazon Mechanical Turk and submission of the last task by a turker (Fig. 3). We estimate that a

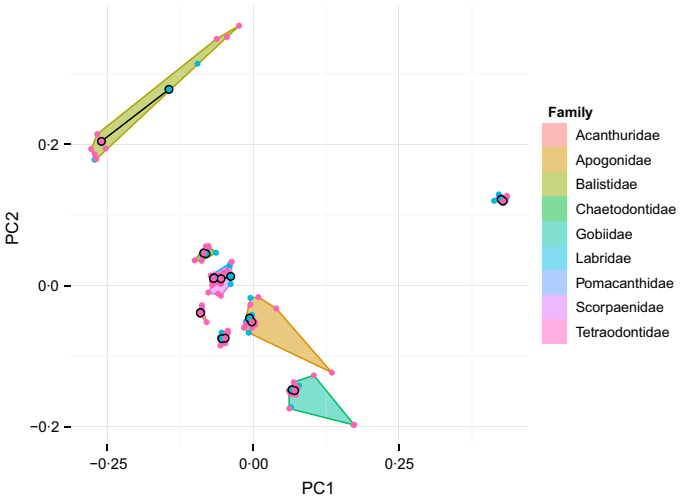

Fig. 2. Morphospace projection for each observer's mean shape. Blue points indicate experts, while red points indicate turkers. The mean shape for all turkers and experts for a given family is the point outlined in black for each family, and connected with a black line to help emphasize the difference between turker and expert mean shapes. The convex hull for each family is drawn to show the amount of among-observer shape variation.

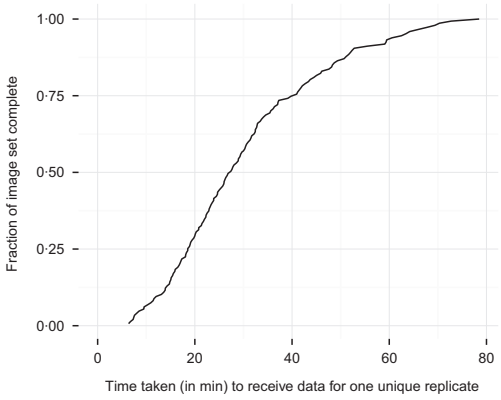

Fig. 3. Line plot showing time to receive results for any given image (*x* axis) and the total fraction of the data set received (*y* axis). Landmarks were first received 8 min after creation of the Amazon MTurk task, and at least one replicate was received for every image at the 80 min mark.

single expert would need 25151·7s (*c.* 6·99 h) to complete all images at 1× replication, extrapolated from a median expert time per image of 171·1 s (*c.* 2·85 min). Our projected expert would need 125758·5s (*c.* 1·46 days) if they had to work at 5× replication.

Using the broken-stick method of determining a PCA stopping point, we analysed PC 1 through PC 5. We project per-species consensus shapes into Procrustes space (Figs 4, S5 and S6). The BAMMtools analysis uncovered heterogeneity in the rate of body shape evolution in each family (Figs 5 and S7). Significant shifts in the rate of shape evolution were detected within two families: Labridae and Pomacentridae. Two significant shifts in shape evolution rate occur in the wrasses (Labridae). The first rate shift occurs deep in the tree, corresponding to the lineage containing the labrine, scarine and cheiline tribes. The other shift is nested within that group, in *Sparisoma*. One shift in the rate of shape evolution occurs in the damselfishes (Pomacentridae) in the genus *Amphiprion*.

Discussion

We have shown that crowdsourcing through Amazon Mechanical Turk is a tractable approach for generating reliable trait data at an unprecedented scale. Using this framework, it is possible to distribute thousands of images to workers, collect the data and send it to a comparative analysis pipeline. We have also demonstrated that it is possible to identify the set of geometric morphometric landmarks that can be reliably captured by nonspecialists. We found that for certain landmarks there was significant between- and within-group disagreement. Points belonging to the opercular series and those locating the distal margin of the dorsal and anal fins were particularly challenging for turkers, compared to the experts. Based on these results, nonspecialist turkers are unlikely to replace experts for all morphometric tasks. However, by digitizing less than 5% of our data set with experts, we were able to identify groups of landmarks that exhibited extremely poor performance and excluded these. Furthermore, we were able to obtain biologically significant results from a data set collected entirely by turkers. By combining expert knowledge with the sheer scale of the Amazon Mechanical Turk workforce, it is possible to collect and assess large quantities of morphometric data, with an order of magnitude improvement in throughput over traditional approaches.

RELIABILITY OF CROWDSOURCED WORKERS

One advantage of the crowdsourced method we develop here is that interobserver error can be readily assessed. Traditional geometric morphometric studies often rely on a single observer for practical reasons, as the pool of trained geometric morpho-

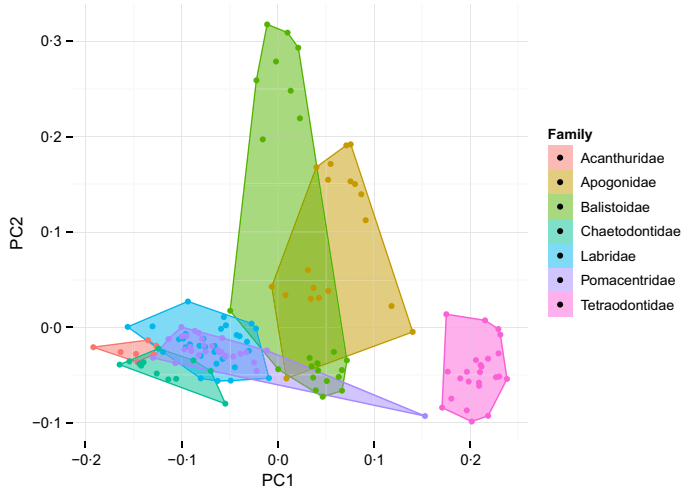

Fig. 4. Morphospace for seven families of ray-finned fishes. Each point indicates a separate species; families are separated by colours. The convex hull for each family is drawn to show area of morphospace occupied by each family. The other PC axes are shown in Figs S5 and S6.

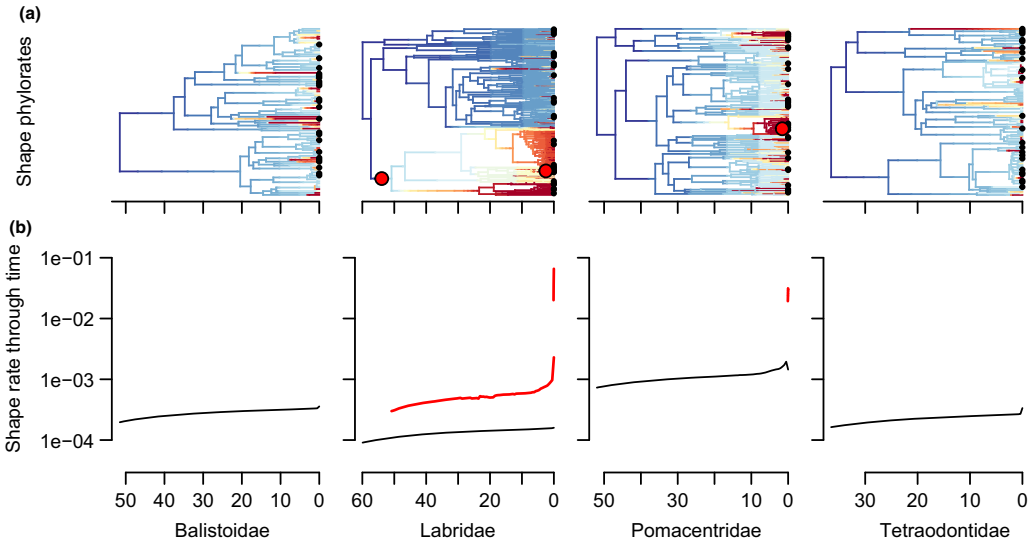

Fig. 5. Rates of shape evolution for PC1 across four families of fishes. (a) Phylorate plots colour branch lengths by rates of shape evolution, where warmer colours indicate faster rates of evolution. Significant rate shift events ($P > 0{\cdot}95$) are indicated on the phylorate plot as a red circle on the corresponding branch. Black circles at the tips indicate the species that had shape data collected. (b) Median log rates of shape evolution through time, where black lines indicate the background rate and red lines indicate the rate of phenotypic evolution in a clade experiencing a significant shift in rate, corresponding to red circles in (a). The other three families are available in Fig. S7.

metricians is limited, to ensure accurate comparisons of the same landmark across specimens, and to avoid individually driven systematic biases in data collection. Although this common practice may reduce bias, it also precludes meaningful assessment of differences among observers. Our results show that interobserver variance can be substantial for some landmarks even among expert digitizers. Therefore, explicitly accounting for interobserver error is critical to determine the efficacy of each individual landmark and the replicability of the study as a whole. Interobserver error signals which landmarks can be relied on and which merit further consideration, as we have done in this analysis. The quantification of interobserver error is a strict requirement of our workflow, as it would otherwise be impossible to arrive at a single consensus shape across several turkers working independently. This requirement ensures that interobserver error is not ignored or bypassed due to the difficulty of assessing it.

In our analysis, we assessed the quality of a variety of landmarks between turkers and experts. Unsurprisingly, turkers performed exceptionally poorly for several landmarks requir-

ing knowledge of fish anatomy. For example, the landmarks that describe the shape of the fish's caudal fin asked workers to mark the distal tip of the first principal fin ray. Even when turkers are armed with a definition and a comparison between procurrent and principal fin rays, the experts' experience and training allowed them to substantially outperform turkers in identifying this point. Furthermore, experts generally had lower disagreement in their landmark placement when compared to turkers, even for landmarks that turkers found especially difficult. These differences between experts and MTurk workers have also been observed in image categorization tasks (Deng *et al*. 2009; Van Horn *et al*. 2015). However, it is possible that an improved training protocol could result in better collection of these difficult landmarks. Turkers have been found to perform well in extremely detailed video annotation tasks (Vondrick, Patterson & Ramanan 2013), provided that researchers conduct pretask training and post-task validation. Implementing these pretask requirements would be a straightforward avenue to improve accuracy for future work.

THE ROLE OF CROWDSOURCED PHENOTYPIC DATA COLLECTION IN MODERN COMPARATIVE STUDIES

The traditional way of collecting phenotypic data involves enormous researcher effort and significant morphological expertise. For example, Brusatte *et al*. (2014) used a 853 character discrete character matrix for 150 taxa to estimate the rate of morphological evolution in the transition from theropod dinosaurs to modern birds. These data were collected over the course of 20 years as part of the Therapod Working Group (Brusatte *et al*. 2014). O'Leary *et al*. (2013) combined the work of MorphoBank contributors (O'Leary & Kaufman 2011) with literature review to generate 4541 characters for 86 species. Rabosky *et al*. (2013) examined 7822 species of ray-finned fish and used a single quantitative measure (body size) collected from FishBase (Froese & Pauly 2014), whose data are contributed from the scientific literature by experts. All of these studies share the same requirement for intensive researcher effort, but the data collected are generally either broad (many species) or deep (many characters). In this study, we collected a phenotypically rich data set across great taxonomic breadth. This approach can easily be scaled to permit unprecedented, massive comparative analyses on new, phenotypically rich data sets.

This method does not threaten to replace experienced morphologists. Although certain conspicuous landmarks can be rapidly collected by turkers, other types of analyses will require landmarks that can only be identified by experts and thus cannot use the high-throughput method presented here. Although this can likely be alleviated by implementing more sophisticated training regimes, the implicit anatomical knowledge that morphologists have must be made explicit in the form of a written protocol for turkers to follow. The cost of developing a clearer and simpler protocol that still captures the essence of the morphological characters of interest must be weighed against the benefit of higher throughput from turker data collection, and for many such analyses, this trade-off is impractical. However, for such analyses where crowdsourcing is a viable alternative, our approach allows experts to move beyond data collection and into a role of developing training materials for nonspecialists and validating the data collected by crowdsourced workers.

Approaches involving statistical techniques like machine vision and natural language processing have yet to make significant headway in automatically collecting morphological data. Although methods to automatically measure leaves exist (Corney *et al*. 2012a, b), these require 2D specimens to eliminate parallax error, as well as high-contrast mounting paper backgrounds for effective automatic outline detection. More sophisticated methods for lower-quality images or organisms with more 3D structure have yet to be developed. Natural language processing of the scientific literature could potentially be used for automatic extraction of morphological characters using DeepDive (Peters *et al*. 2014; Shin *et al*. 2015), but it may require impractically large corpus sizes (Brill 2003; Halevy, Norvig & Pereira 2009). Instead of using any one method exclusively, crowdsourcing can augment and enhance these statistical techniques. For example, the algorithm in Corney *et al*. (2012a) occasionally captures non-leaf objects and systematically underestimates leaf sizes. MTurk workers could improve this method by confirming the presence of a leaf in the image segment and measure the leaf size to ground truth the algorithm's results.

A third alternative to using expert morphologists and crowdsourced workers is to collect data through citizen science. Citizen scientists are enthusiasts that volunteer to collect data or contribute annotations to a scientific endeavour. They can specialize in a particular field, such as birds, plants or fungi. Compared to Amazon Mechanical Turk workers, citizen scientists are typically unpaid, but can produce higher-quality work due to their expertise. For example, a study comparing citizen scientists and MTurk workers showed that for an image segmentation task, MTurk workers had higher throughput and comparable accuracy to citizen scientists, but MTurk workers performed poorly when asked to identify birds to the species level (Van Horn *et al*. 2015). Volunteer citizen scientists can be inexpensive to use, but the pool of available MTurk workers is likely much larger. This larger participant pool means that tasks can be completed much faster due to the ability of multiple individuals to work in parallel; the financial motivation additionally ensures that higher-paying tasks are completed more quickly (Ipeirotis 2010; Mason & Suri 2012). Balancing the desired speed and quality of results, and the cost of data collection will be an important consideration for any future study using crowdsourcing.

SUITABILITY FOR OTHER SYSTEMS

Our novel pipeline to download images, upload them to Amazon MTurk and process them using BAMM and BAMMtools showcases the ability to rapidly collect phenotypic data. Most of the time taken to collect these data were spent on waiting for worker results; however, a majority of the data had already been collected at the 1-h mark. An online methodology could

conceivably improve on this analysis time, by iteratively refining its results as new data streamed in from Amazon's servers.

Although there are limitations in the type and accuracy of data that can be collected through MTurk crowdsourcing, even a simplified protocol can produce meaningful biological results that are concordant with previous hypotheses in these groups. Despite our low sampling fraction, we detected a significant shift in the rate of body shape evolution in Labridae, restricted to the wrasse tribes Labrini, Cheilini and Scarini. The scarines and cheilines are mostly reef associated (Froese & Pauly 2014), which has been proposed as an environment that drives diversification rate changes in marine teleosts (Alfaro, Santini & Brock 2007; Cowman & Bellwood 2011; Price *et al*. 2011). These results suggest that evolution of body form may also be influenced by environmental association (Claverie & Wainwright 2014). Although the example we present here was necessarily limited, extending this technique to generate new phenotypic data sets for existing large phylogenetic trees such as fishes (Rabosky *et al*. 2013), birds (Jetz *et al*. 2012), mammals (Bininda-Emonds *et al*. 2007) and angiosperms (Zanne *et al*. 2014) would be straightforward, especially for taxa where image data are already aggregated in a data base such as FishBase (Froese & Pauly 2014) or the Encyclopedia of Life (Parr *et al*. 2014).

FUTURE CHALLENGES FOR GENERATING MASSIVE PHENOTYPIC DATA SETS

Our approach hits a 'sweet spot' on the three axes of expertise, effort and computational complexity. We use researcher expertise to identify a comparative hypothesis, and design a data collection protocol to specifically test this hypothesis. Amazon Mechanical Turk supplies a large source of worker effort that collects data according to protocol. Finally, computational statistical techniques validate the accuracy of our data and identify outliers and other errors in data collection. Researchers do not have to spend time digitizing collections, workers need not generate biological hypotheses, and biologists will not have to solve open questions in the fields of machine vision and natural language processing in order to answer questions in comparative biology. The task of phenomic-scale data collection is split up and efficiently allocated according to the strengths of each role, without overly relying on any single role to carry out the entire task.

Although we have shown that crowdsourcing can increase these speed of data collection, we are still dependent on high-quality image data sets, as evidenced by our low sampling fraction for three of the seven families analysed. The problem of difficult-to-retrieve *dark data* is well known (Heidorn 2008), but without either physical access to the collections or an image of the specimen, morphological data are impossible to acquire. The need to collect, identify, photograph and publish specimen images remains as another obstacle to high-throughput phenotyping. Efforts are underway to digitize more biodiversity resources, such as the National Science Foundation's iDigBio initiative (https://www.idigbio.org) in the U.S. and the Natural History Museum's iCollections project (http://www.nhm.ac.uk/our-science/our-work/digital-museum/digital-collections-programme.html) in the U.K. Whole-drawer imaging of insect collections and scanning of herbarium pressings are already well underway, but one future direction would be to expand this to other avenues: skeletal imaging with radiographs, 3D morphometrics using laser or CT scanning, of both fossils and extant organisms. Much work and engineering expertise will be required to extend our framework into the physical world to further streamline data collection, but these efforts will likely result in a huge increase in the quality and quantity of phenotypic data.

Our work fills the niche of gathering phenotypic data across large radiations, which has been a challenging open research question (Burleigh *et al*. 2013). Even seemingly obvious phenotypes, such as the woodiness of plant species, are incomplete and sampled in a biased manner (FitzJohn *et al*. 2014), potentially misleading inference on a global scale. This method unlocks the potential of high-throughput data collection and shifts the data bottleneck for morphological research onto acquiring suitable images for quantification, and developing higher-quality worker training regimens to enable collection of more sophisticated data. The burden is now on experienced taxonomists and morphologists to create protocols that are simple enough to be understood by MTurk workers, but comprehensive enough to test hypotheses of interest across the tree of life.

Our results suggest that, where possible, crowdsourcing should be an integral part of any large-scale morphological analysis. Crowdsourcing can play a key role in unlocking the 'dark data' present in biodiversity collections by providing a high-throughput way to extract the phenotypic data present in specimens. Furthermore, coordinating efforts from digitizing museum collections, natural language processing and machine vision software, citizen scientists, expert morphologists and taxonomists, and crowdsourced Mechanical Turk workers would result in an extremely powerful pipeline that could generate a 'phenoscape' across the tree of life.

Acknowledgements

We thank P. Chakrabarty and G. Thomas for helpful comments on the manuscript, as well as T. Marcroft, B. Frederich, V. Liu, R. Aguilar, R. Ellingson, F. Pickens, C. LaRochelle, and the 67 Amazon Mechanical Turk workers that contributed their time and effort. We also thank D. Rabosky, B. Sidlauskas, M. McGee, A. Summers and M. Burns for insightful discussions about fish morphology and digitization protocols. M. Venzon and T. Claverie provided unpublished figures that assisted this study. K. Staab and T. Kane allowed 156 undergraduate students to beta test the methods. This work was supported by an Encyclopedia of Life David M. Rubenstein Fellowship (EOL-33066-13), a Stephen and Ruth Wainwright Fellowship, and a UCLA Research and Conference Award to JC. Travel support to present this research was provided by the Society for Study of Evolution.

Data accessibility

Data collected for this paper have been archived on Dryad http://dx.doi.org/10.5061/dryad.gh4k7 (Chang & Alfaro 2015). Source code is available on GitHub for the web interface (https://github.com/jonchang/eol-mturk-landmark), repeatability experiment (https://github.com/jonchang/fake-mechanical-turk) and this manuscript (https://github.com/jonchang/fish.reliability).

Author contributions

JC MEA conceived and designed the experiments. JC performed the experiments. JC analysed the data. JC contributed reagents/materials/analysis tools. JC MEA wrote the paper.

References

Adams, D. & Otarola-Castillo, E. (2013). Geomorph: an R package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and Evolution*, **4**, 393–399.

Alfaro, M.E., Bolnick, D.I. & Wainwright, P.C. (2004). Evolutionary dynamics of complex biomechanical systems: an example using the four-bar mechanism. *Evolution*, **58**, 495–503.

Alfaro, M.E., Bolnick, D.I. & Wainwright, P.C. (2005). Evolutionary consequences of many-to-one mapping of jaw morphology to mechanics in labrid fishes. *The American Naturalist*, **165**, E140–E154.

Alfaro, M.E., Santini, F. & Brock, C.D. (2007). Do reefs drive diversification in marine teleosts? Evidence from the pufferfish and their allies (order tetraodontiformes). *Evolution*, **61**, 2104–2126.

Aliscioni, S., Bell, H.L., Besnard, G., Christin, P.A., Columbus, J.T., Duvall, M.R. *et al*. (2012). New grass phylogeny resolves deep evolutionary relationships and discovers C 4 origins. *New Phytologist*, **193**, 304–312.

Anderson, M. & Braak, C.T. (2003). Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation*, **73**, 85–113.

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R. *et al*. (2007). The delayed rise of present-day mammals. *Nature*, **446**, 507–512.

Bookstein, F.L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press, Cambridge.

Brill, E. (2003). Processing Natural Language without Natural Language Processing. *Computational Linguistics and Intelligent Text Processing*, **2588**, 360–369.

Brusatte, S.L., Lloyd, G.T., Wang, S.C. & Norell, M.A. (2014). Gradual Assembly of Avian Body Plan Culminated in Rapid Rates of Evolution across the Dinosaur-Bird Transition. *Current Biology*, **24**, 1–7.

Burleigh, J.G., Alphonse, K., Alverson, A.J., Bik, H.M., Blank, C., Cirranello, A.L. *et al*. (2013). Next-generation phenomics for the Tree of Life. *PLoS Currents Tree of Life*.

Cavalcanti, M.J., Monteiro, L.R. & Lopes, P.R.D. (1999). Landmark-based morphometric analysis in selected species of serranid fishes (Perciformes: Teleostei). *Zoological Studies*, **38**, 287–294.

Chakrabarty, P. (2005). Testing Conjectures about Morphological Diversity in Cichlids of Lakes Malawi and Tanganyika. *Copeia*, **2005**, 359–373.

Chang, J. & Alfaro, M.E. (2015) Data from: Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data. *Dryad Digital Repository*, http://dx.doi.org/10.5061/dryad.gh4k7.

Choat, J.H., Klanten, O.S., Van Herwerden, L., Robertson, D.R. & Clements, K.D. (2012). Patterns and processes in the evolutionary history of parrotfishes (Family Labridae). *Biological Journal of the Linnean Society*, **107**, 529–557.

Claverie, T. & Wainwright, P.C. (2014). A morphospace for reef fishes: elongation is the dominant axis of body shape evolution. *PLoS ONE*, **9**, e112732.

Collar, D.C. & Wainwright, P.C. (2006). Discordance between morphological and mechanical diversity in the feeding mechanism of centrarchid fishes. *Evolution*, **60**, 2575–2584.

Collyer, M.L., Sekora, D.J. & Adams, D.C. (2015). A method for analysis of phenotypic change for phenotypes described by high-dimensional data. *Heredity*, **115**, 1–9.

Corney, D.P.A., Clark, J.Y., Lilian Tang, H. & Wilkin, P. (2012a). Automatic extraction of leaf characters from herbarium specimens. *Taxon*, **61**, 231–244.

Corney, D.P.A., Tang, H.L., Clark, J.Y., Hu, Y. & Jin, J. (2012b). Automating digital leaf measurement: The tooth, the whole tooth, and nothing but the tooth. *PLoS ONE*, **7**, 1–10.

Cowman, P.F. & Bellwood, D.R. (2011). Coral reefs as drivers of cladogenesis: Expanding coral reefs, cryptic extinction events, and the development of biodiversity hotspots. *Journal of Evolutionary Biology*, **24**, 2543–2562.

Cui, H. (2012). CharaParser for fine-grained semantic annotation of organism morphological descriptions. *Journal of the American Society for Information Science and Technology*, **63**, 738–754.

Deans, A.R., Lewis, S.E., Huala, E., Anzaldo, S.S., Ashburner, M., Balhoff, J.P. *et al*. (2015). Finding our way through phenotypes. *PLoS Biology*, **13**, e1002033.

Dececchi, T.A., Balhoff, J.P., Lapp, H. & Mabee, P.M. (2015). : Using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. *Systematic Biology*, **64**, 936-952.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009) ImageNet: a large-scale hierarchical image database. Proc. CVPR, 248–255.

Dornburg, A., Sidlauskas, B., Santini, F., Sorenson, L., Near, T.J. & Alfaro, M.E. (2011). The influence of an innovative locomotor strategy on the phenotypic diversification of triggerfish (family: balistidae). *Evolution*, **65**, 1912–1926.

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. *Systematic Biology*, **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.

Faircloth, B.C., Branstetter, M.G., White, N.D. & Brady, S.G. (2015). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources*, **15**, 489–501.

Ferry-Graham, L.A., Wainwright, P.C., Darrin Hulsey, C. & Bellwood, D.R. (2001). Evolution and mechanics of long jaws in butterflyfishes (Family Chaetodontidae). *Journal of Morphology*, **248**, 120–143.

Fink, W.L. & Zelditch, M.L. (1995). Phylogenetic Analysis of Ontogenic Shape Transformations - a Reassessment of the Piranha Genus Pygocentrus (Teleostei). *Systematic Biology*, **44**, 343–360.

FitzJohn, R.G., Pennell, M.W., Zanne, A.E., Stevens, P.F., Tank, D.C. & Cornwell, W.K. (2014). How much of the world is woody? *Journal of Ecology,* **102,** 1266–1272.

Frédérich, B., Adriaens, D. & Vandewalle, P. (2008). Ontogenetic shape changes in Pomacentridae (Teleostei, Perciformes) and their relationships with feeding strategies: A geometric morphometric approach. *Biological Journal of the Linnean Society*, **95**, 92–105.

Frédérich, B., Sorenson, L., Santini, F., Slater, G.J. & Alfaro, M.E. (2013). Iterative ecological radiation and convergence during the evolutionary history of damselfishes (Pomacentridae). *The American Naturalist*, **181**, 94–113.

Froese, R. & Pauly, D. (2014). *FishBase*. URL: http://www.fishbase.org.

Furbank, R.T. & Tester, M. (2011). Phenomics - technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, **16**, 635–644.

Goldberg, E.E., Kohn, J.R., Lande, R., Robertson, K.A., Smith, S.A. & Igić, B. (2010). Species selection maintains self-incompatibility. *Science*, **330**, 493–495.

Good, B.M. & Su, A.I. (2013). Crowdsourcing for bioinformatics. *Bioinformatics*, **29**, 1925–1933.

Gower, J.C. (1975). Generalized procrustes analysis. *Psychometrika*, **40**, 33–51.

Halevy, A., Norvig, P. & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, **24**, 8-12.

Harmon, L.J., Losos, J.B., Jonathan Davies, T., Gillespie, R.G., Gittleman, J.L., Bryan Jennings, W. *et al*. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, **64**, 2385–2396.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd edn. Springer, New York.

Heidorn, P.B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, **57**, 280–299.

Hernandez, L.P., Gibb, A.C. & Ferry-Graham, L.A. (2009). Trophic apparatus in cyprinodontiform fishes: Functional specializations for picking and scraping behaviors. *Journal of Morphology*, **270**, 645–661.

Ipeirotis, P.G. (2010). Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, **17**, 16.

Jackson, D.A. (1993). Stopping rules in principal components analysis : A comparison of heuristical and statistical approaches. *Ecology*, **74**, 2204–2214.

Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K. & Mooers, A.O. (2012). The global diversity of birds in space and tim. *Nature*, **491**, 1–5.

Klingenberg, C.P., Barluenga, M. & Meyer, A. (2003). Body shape variation in cichlid fishes of the Amphilophus citrinellus species complex. *Biological Journal of the Linnean Society*, **80**, 397–408.

Kohavi, R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint conference on artificial intelligence*, pp. 1137–1143.

Legendre, P. & Legendre, L. (1998). *Numerical Ecology*, 2nd edn. Elsevier, Amsterdam.

Lemmon, A.R., Emme, S.A. & Lemmon, E.M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, **61**, 727–744.

Mardia, K.V., Kent, J.T. & Bibby, J. (1979) *Multivariate Analysis*, 1st edn. Academic Press, London.

Mason, W. & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, **44**, 1–23.

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 Research*, **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 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing. *PLoS ONE*, **8**, e54848.

Meyer, M., Stenzel, U. & Hofreiter, M. (2008). Parallel tagged sequencing on the 454 platform. *Nature Protocols*, **3**, 267–278.

Midford, P.E., Dececchi, T.A., Balhoff, J.P., Dahdul, W.M., Ibrahim, N., Lapp, H. *et al*. (2013). The vertebrate taxonomy ontology: a framework for reasoning across model organism and species phenotypes. *Journal of Biomedical Semantics*, **4**, 34.

O'Leary, M.A. & Kaufman, S. (2011). MorphoBank: Phylophenomics in the 'cloud'. *Cladistics*, **27**, 529–537.

O'Leary, M.A., Bloch, J.I., Flynn, J.J., Gaudin, T.J., Giallombardo, A., Giannini, N.P. *et al*. (2013). The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science*, **339**, 662–667.

Palmer, A.R. & Strobeck, C. (1986). Fluctuating asymmetry: measurement, analysis, patterns. *Annual Review of Ecology and Sustematics*, **17**, 391–421.

Parr, C.S., Wilson, N., Leary, P., Schulz, K.S., Lans, K., Walley, L. *et al*. (2014) The encyclopedia of life v2: providing global access to knowledge about life on earth. *Biodiversity Data Journal*, **2**, e1079.

Peters, S.E., Zhang, C., Livny, M. & Christopher, R. (2014) A machine-compiled macroevolutionary history of Phanerozoic life. arXiv:1406.2963 [cs.DB].

Price, S.A., Wainwright, P.C., Bellwood, D.R., Kazancioglu, E., Collar, D.C. & Near, T.J. (2010). Functional innovations and morphological diversification in parrotfish. *Evolution*, **64**, 3057–3068.

Price, S.A., Holzman, R., Near, T.J. & Wainwright, P.C. (2011). Coral reefs promote the evolution of morphological diversity and ecological novelty in labrid fishes. *Ecology Letters*, **14**, 462–469.

Price, S.A., Hopkins, S.S.B., Smith, K.K. & Roth, V.L. (2012). Tempo of trophic evolution and its impact on mammalian diversification. *Proceedings of the National Academy of Sciences USA*, **109**, 7008–7012.

Pyron, R.A. & Burbrink, F.T. (2014). Early origin of viviparity and multiple reversions to oviparity in squamate reptiles. *Ecology Letters*, **17**, 13–21.

Rabosky, D.L. (2014). Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE*, **9**, e89543.

Rabosky, D.L., Santini, F., Eastman, J.M., Smith, S.A., Sidlauskas, B., Chang, J. & Alfaro, M.E. (2013). Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nature Communications*, **4**, 1958.

Rabosky, D.L., Grundler, M., Anderson, C., Title, P., Shi, J.J., Brown, J.W., Huang, H. & Larson, J.G. (2014) BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution*, **5**, 701–707.

Rambaut, A., Suchard, M.A., Xie, D. & Drummond, A.J. (2014). Tracer 1.6. URL: http://beast.bio.ed.ac.uk/tracer

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks, 1st edn*. Cambridge University Press, Cambridge.

Rocha, L.A., Lindeman, K.C., Rocha, C.R. & Lessios, H.A. (2008). Historical biogeography and speciation in the reef fish genus Haemulon (Teleostei: Haemulidae). *Molecular Phylogenetics and Evolution*, **48**, 918–928.

Rohlf, F. & Slice, D. (1990). Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Biology*, **39**, 40–59.

Rüber, L. & Adams, D.C. (2001). Evolutionary convergence of body shape and trophic morphology in cichlids from Lake Tanganyika. *Journal of Evolutionary Biology*, **14**, 325–332.

Santini, F., Sorenson, L. & Alfaro, M.E. (2013). A new multi-locus timescale reveals the evolutionary basis of diversity patterns in triggerfishes and filefishes (Balistidae, Monacanthidae; Tetraodontiformes). *Molecular Phylogenetics and Evolution*, **69**, 165–176.

Schluter, D. (2000). *The Ecology of Adaptive Radiations*. Oxford University Press, Oxford, UK.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Shi, J.J. & Rabosky, D.L. (2015). Speciation dynamics during the global radiation of extant bats. *Evolution*, **69**, 1528–1545.

Shin, J., Wang, F., De Sa, C., Zhang, C., Wu, S. & Ré, C. (2015). Incremental Knowledge Base Construction Using DeepDive. *Proceedings of the VLDB Endowment*, **8**, 1310–1321.

Skelly, D.A., Merrihew, G.E., Riffle, M., Connelly, C.F., Kerr, E.O., Johansson, M. *et al*. (2013). Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Research*, **23**, 1496–1504.

Sorenson, L., Santini, F., Carnevale, G. & Alfaro, M.E. (2013). A multi-locus timetree of surgeonfishes (Acanthuridae, Percomorpha), with revised family taxonomy. *Molecular Phylogenetics and Evolution*, **68**, 150–160.

Thacker, C.E. (2014). Species and shape diversification are inversely correlated among gobies and cardinalfishes (Teleostei: Gobiiformes). *Organisms Diversity & Evolution*, **14**, 419–436.

Van Horn, G., Branson, S., Farrell, R., Barry, J. & Tech, C. (2015) Building a bird recognition app and large scale dataset with citizen scientists : The fine print in fine-grained dataset collection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604.

Von Cramon-Taubadel, N., Frazier, B.C. & Lahr, M.M. (2007). The problem of assessing landmark error in geometric morphometrics: Theory, methods, and modifications. *American Journal of Physical Anthropology*, **134**, 24–35.

Vondrick, C., Patterson, D. & Ramanan, D. (2013). Efficiently scaling up crowdsourced video annotation: A set of best practices for high quality, economical video labeling. *International Journal of Computer Vision*, **101**, 184–204.

Wainwright, P.C., Alfaro, M.E., Bolnick, D.I. & Hulsey, C.D. (2005). Many-to-one mapping of form to function: a general principle in organismal design? *Integrative and Comparative Biology*, **45,** 256–262.

Yoder, M.J., Mikó, I., Seltmann, K.C., Bertone, M.A. & Deans, A.R. (2010). A gross anatomy ontology for hymenoptera. *PLoS ONE*, **5**, e15991.

Zanne, A.E., Tank, D.C., Cornwell, W.K., Eastman, J.M., Smith, S.A FitzJohn, R.G. *et al*. (2014). Three keys to the radiation of angiosperms into freezing environments. *Nature*, **506**, 89–92.

Zelditch, M.L., Swiderski, D. & Sheets, H.D. (2012). *Geometric Morphometrics for Biologists: A Primer*, 2nd edn. Academic Press, San Diego.

Received 16 July 2015; accepted 25 October 2015
Handling Editor: Robert Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Supplementary material.

Table S1. Images digitized by turkers and experts to compare their performance.

Table S2. Online URLs of images from Table S1.

Table S3. Five number summaries of turker and expert consistency.

Table S4. Comparison of the Procrustes distance between the mean turker shape and the mean expert shape, for a full dataset, and a dataset excluding the first three images that turkers worked on.

Table S5. Families, species names, and URLs of the images hosted on Encyclopedia of Life for the section 'Example: a phenomic pipeline for comparative phylogenetic analysis'.

Figure S1. A screenshot of the web app that turkers used to digitize images.

Figure S2. Description of landmarks used to digitize fish body shape.

Figure S3. Version of Figure 1 where points are annotated with the landmark label.

Figure S4. Morphospace projection of PC3 and PC4 for each observer's mean shape.

Figure S5. Morphospace of PC3 and PC4 for seven families of ray-finned fishes.

Figure S6. Morphospace of PC5 and PC6 for seven families of ray-finned fishes.

Figure S7. Rates of shape evolution for PC1 across three families of fishes.

Appendix S2. Landmarking protocol.

Appendix S3. CSV file used to generate Table S5.

Supplementary material

Below is an example JSON file that demonstrates the utility of our web app:

```
{
    "C": {
        "kind": "point",
        "help": "Click the center of the eye."
    },
    "D": {
        "kind": "line",
        "help": "Click and drag from the left edge of the eye
                 to the right edge of the eye."
    },
    "O": {
        "kind": "curve",
        "help": "Click and drag over the outline of the eye,
                 starting from the leftmost point of the eye."
    }
}
```

Each digitization task to be completed is given a short abbreviation to aid task identification ("C" for center, 'D" for diameter, and "O" for outline), and the type of task, "point", "line" or "curve", for homologous landmarks, linear measurements, and sliding semilandmarks can be specified. There is also an optional short help snippet displayed inline, which serve as a brief reminder for each landmark and complements a larger and more detailed protocol document that workers are required to read before beginning work.

Table S1: Images digitized by turkers and experts to compare their performance. TL = total length of the specimen, in cm. PX = The total length of the specimen, in number of image pixels.

Family	Species	Author	Rights	TL	px
Acanthuridae	Naso annulatus	John E Randall	cc-by-nc 3.0	19.8	704
Apogonidae	Nectamia ignitops	John E Randall	cc-by-nc 3.0	9.4	707
Balistidae	Pseudobalistes flavimarginatus	John E Randall	cc-by-nc 3.0	22.9	669
Chaetodontidae	Chaetodon citrinellus	John E Randall	cc-by-nc 3.0	11.4	706
Gobiidae	Amblyeleotris neglecta	John E Randall	cc-by-nc 3.0	7.5	708
Labridae	Anampses cuvier	John E Randall	cc-by-nc 3.0	31.0	738
Pomacanthidae	Centropyge eibli	John E Randall	cc-by-nc 3.0	7.1	750
Scorpaenidae	Caracanthus maculatus	John E Randall	cc-by-nc 3.0	4.7	696
Tetraodontidae	Canthigaster epilampra	John E Randall	cc-by-nc 3.0	8.4	689

Table S2: Online URLs of images from Supplemental Table S1

Family	URL
Acanthuridae	http://eol.org/data_objects/21028048
Apogonidae	http://www.fishbase.org/Photos/PicturesSummary.php?ID=63749&what=species
Balistidae	http://eol.org/data_objects/21028158
Chaetodontidae	http://eol.org/data_objects/21022257

Family	URL
Gobiidae	http://eol.org/data_objects/30887808
Labridae	http://eol.org/data_objects/21016334
Pomacanthidae	http://www.fishbase.org/photos/PicturesSummary.php?resultPage=5&ID=10870&what=sp
Scorpaenidae	http://eol.org/data_objects/21043145
Tetraodontidae	http://eol.org/data_objects/21042893

Landmarks used

The landmarks used are shown in Supplemental Figure S2. The image is redrawn from Chakrabarty (2005), which was itself redrawn from Nelson (1994).

Testing turker vs expert consistency

Table S3: Five number summaries of turker and expert consistency. The summary statistics are multiplied by 1,000,000 to facilitate comparisons.

family	role	minimum	1st quartile	median	3rd quartile	maximum
Acanthuridae	expert	63	78	107	116	488
Acanthuridae	turker	85	96	141	1518	8340
Apogonidae	expert	411	780	5165	9540	9900
Apogonidae	turker	124	337	1208	6524	15131
Balistidae	expert	110	187	263	318	372
Balistidae	turker	89	121	410	5270	6713
Chaetodontidae	expert	36	87	143	313	550
Chaetodontidae	turker	92	115	238	404	8397
Gobiidae	expert	155	434	458	524	581
Gobiidae	turker	87	186	324	2883	25698
Labridae	expert	23	82	104	124	192
Labridae	turker	83	87	134	589	6236
Pomacanthidae	expert	53	61	78	119	148
Pomacanthidae	turker	67	99	146	197	31122
Scorpaenidae	expert	252	264	277	5470	10663
Scorpaenidae	turker	142	233	274	350	3275
Tetraodontidae	expert	184	308	483	723	912
Tetraodontidae	turker	86	131	152	320	431

Do turkers improve with experience?

Table S4: Comparison of the Procrustes distance between the mean turker shape and the mean expert shape, for a full dataset, and a dataset excluding the first three images that turkers worked on. The ratio is computed by dividing the full dataset's distance by the reduced dataset's distance, in order to compare the relative distance change among the different images digitized.

Family	Procrustes distance: full dataset	Reduced dataset	Ratio
Acanthuridae	0.02459	0.02581	0.95287
Apogonoidae	0.04509	0.04857	0.92834

Family	Procrustes distance: full dataset	Reduced dataset	Ratio
Balistidae	0.05344	0.05664	0.94348
Chaetodontidae	0.01415	0.01488	0.95059
Gobiidae	0.05789	0.06009	0.96339
Labridae	0.01842	0.01832	1.00584
Pomacanthidae	0.01239	0.01226	1.01019
Scorpaenidae	0.02529	0.02529	0.99999
Tetraodontidae	0.02974	0.02974	1.00004

R information

Information on the versions of R packages used to analyze this data.

```
## Session info ------------------------------------------------------------------------

##  setting  value
##  version  R version 3.2.2 (2015-08-14)
##  system   x86_64, linux-gnu
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/Los_Angeles
##  date     2015-09-22

## Packages ----------------------------------------------------------------------------

##  package           * version    date       source
##  animation           2.4        2015-08-16 CRAN (R 3.2.2)
##  ape               * 3.3        2015-05-29 CRAN (R 3.2.2)
##  assertthat          0.1        2013-12-06 CRAN (R 3.2.2)
##  BAMMtools         * 2.0.5      2015-09-21 Github (jonchang/BAMMtools@2e402c9)
##  bibtex              0.4.0      2014-12-31 CRAN (R 3.2.2)
##  bitops              1.0-6      2013-08-17 CRAN (R 3.2.2)
##  class               7.3-13     2015-06-29 CRAN (R 3.2.2)
##  cluster             2.0.3      2015-07-21 CRAN (R 3.2.2)
##  clusterGeneration   1.3.4      2015-02-18 CRAN (R 3.2.2)
##  coda                0.17-1     2015-03-03 CRAN (R 3.2.2)
##  codetools           0.2-14     2015-07-15 CRAN (R 3.2.2)
##  colorspace          1.2-6      2015-03-11 CRAN (R 3.2.2)
##  DBI                 0.3.1      2014-09-24 CRAN (R 3.2.2)
##  deSolve             1.12       2015-07-06 CRAN (R 3.2.2)
##  devtools            1.9.1      2015-09-11 CRAN (R 3.2.2)
##  digest              0.6.8      2014-12-31 CRAN (R 3.2.2)
##  directlabels      * 2013.6.15  2013-07-23 CRAN (R 3.2.2)
##  dplyr             * 0.4.3      2015-09-01 CRAN (R 3.2.2)
##  evaluate            0.8        2015-09-18 CRAN (R 3.2.2)
##  expm                0.99-1.1   2014-02-12 CRAN (R 3.2.2)
##  fish.reliability  * 0.0.0.9000 2015-09-17 local
##  formatR             1.2.1      2015-09-18 CRAN (R 3.2.2)
##  geiger              2.0.6      2015-09-07 CRAN (R 3.2.2)
##  geomorph          * 2.1.6      2015-04-02 Github (jonchang/geomorph@82ced7c)
##  ggplot2           * 1.0.1      2015-03-17 CRAN (R 3.2.2)
```

```
##  gtable              0.1.2      2012-12-05 CRAN (R 3.2.2)
##  htmltools           0.2.6      2014-09-08 CRAN (R 3.2.2)
##  httr                1.0.0      2015-06-25 CRAN (R 3.2.2)
##  igraph              1.0.1      2015-06-26 CRAN (R 3.2.2)
##  ipred             * 0.9-5      2015-07-28 CRAN (R 3.2.2)
##  jpeg                0.1-8      2014-01-23 CRAN (R 3.2.2)
##  kfigr             * 1.2        2015-07-15 CRAN (R 3.2.2)
##  knitcitations     * 1.0.6      2015-05-26 CRAN (R 3.2.2)
##  knitr             * 1.11       2015-08-14 CRAN (R 3.2.2)
##  lattice           * 0.20-33    2015-07-14 CRAN (R 3.2.2)
##  lava                1.4.1      2015-06-22 CRAN (R 3.2.2)
##  lazyeval            0.1.10     2015-01-02 CRAN (R 3.2.2)
##  lubridate         * 1.3.3      2013-12-31 CRAN (R 3.2.2)
##  magrittr          * 1.5        2014-11-22 CRAN (R 3.2.2)
##  maps                2.3-11     2015-08-03 CRAN (R 3.2.2)
##  MASS              * 7.3-43     2015-07-16 CRAN (R 3.2.2)
##  Matrix              1.2-2      2015-07-08 CRAN (R 3.2.2)
##  memoise             0.2.1      2014-04-22 CRAN (R 3.2.2)
##  mgcv                1.8-7      2015-07-23 CRAN (R 3.2.2)
##  mnormt              1.5-3      2015-05-25 CRAN (R 3.2.2)
##  msm                 1.5        2015-01-06 CRAN (R 3.2.2)
##  munsell             0.4.2      2013-07-11 CRAN (R 3.2.2)
##  mvtnorm             1.0-3      2015-07-22 CRAN (R 3.2.2)
##  nlme                3.1-121    2015-06-29 CRAN (R 3.2.2)
##  nnet                7.3-10     2015-06-29 CRAN (R 3.2.2)
##  nnls                1.4        2012-03-19 CRAN (R 3.2.2)
##  numDeriv            2014.2-1   2015-05-04 CRAN (R 3.2.2)
##  permute           * 0.8-4      2015-05-19 CRAN (R 3.2.2)
##  phangorn            1.99.14    2015-07-09 CRAN (R 3.2.2)
##  phytools            0.4-60     2015-07-10 CRAN (R 3.2.1)
##  plotrix             3.5-12     2015-05-16 CRAN (R 3.2.2)
##  plyr                1.8.3      2015-06-12 CRAN (R 3.2.1)
##  prodlim             1.5.1      2014-12-10 CRAN (R 3.2.2)
##  proto               0.3-10     2012-12-22 CRAN (R 3.2.2)
##  quadprog          * 1.5-5      2013-04-17 CRAN (R 3.2.2)
##  R6                  2.1.1      2015-08-19 CRAN (R 3.2.2)
##  Rcpp                0.12.1     2015-09-10 CRAN (R 3.2.2)
##  RCurl               1.95-4.7   2015-06-30 CRAN (R 3.2.2)
##  readr             * 0.1.1      2015-05-29 CRAN (R 3.2.2)
##  RefManageR          0.8.63     2015-06-09 CRAN (R 3.2.2)
##  reshape2          * 1.4.1      2014-12-06 CRAN (R 3.2.2)
##  rgl               * 0.95.1337  2015-09-19 CRAN (R 3.2.2)
##  rjson             * 0.2.15     2014-11-03 CRAN (R 3.2.2)
##  RJSONIO             1.3-0      2014-07-28 CRAN (R 3.2.2)
##  rmarkdown           0.8        2015-08-30 CRAN (R 3.2.2)
##  rpart               4.1-10     2015-06-29 CRAN (R 3.2.2)
##  scales            * 0.3.0      2015-08-25 CRAN (R 3.2.2)
##  scatterplot3d       0.3-36     2015-07-30 CRAN (R 3.2.2)
##  stringi             0.5-5      2015-06-29 CRAN (R 3.2.2)
##  stringr           * 1.0.0      2015-04-30 CRAN (R 3.2.2)
##  subplex             1.1-6      2015-07-11 CRAN (R 3.2.2)
##  survival            2.38-3     2015-07-02 CRAN (R 3.2.2)
##  tidyr             * 0.3.1      2015-09-10 CRAN (R 3.2.2)
##  vegan             * 2.3-0      2015-05-26 CRAN (R 3.2.2)
```

```
##  XML                 3.98-1.3   2015-06-30 CRAN (R 3.2.2)
##  yaml                2.1.13     2014-06-12 CRAN (R 3.2.2)
```

Table S5: Families, species names, and URLs of the images hosted on Encyclopedia of Life for the section "Example: a phenomic pipeline for comparative phylogenetic analysis".

family	species	url
Acanthuridae	Acanthurus bahianus	http://media.eol.org/content/2013/12/10/00/96795_o
Acanthuridae	Acanthurus chirurgus	http://media.eol.org/content/2012/01/28/04/93256_o
Acanthuridae	Acanthurus coeruleus	http://media.eol.org/content/2009/05/19/10/17601_o
Acanthuridae	Ctenochaetus truncatus	http://media.eol.org/content/2013/04/25/06/63463_o
Acanthuridae	Ctenochaetus truncatus	http://media.eol.org/content/2013/04/25/06/63463_o
Acanthuridae	Naso elegans	http://media.eol.org/content/2013/04/25/03/31014_o
Acanthuridae	Naso minor	http://media.eol.org/content/2013/03/12/02/96923_o
Acanthuridae	Prionurus scalprum	http://media.eol.org/content/2013/03/12/02/49687_o
Acanthuridae	Zebrasoma velifer	http://media.eol.org/content/2012/12/08/06/09619_o
Apogonidae	Apogon amboinensis	http://media.eol.org/content/2013/03/12/02/16100_o
Apogonidae	Apogon aurolineatus	http://media.eol.org/content/2012/01/28/03/22795_o
Apogonidae	Apogon carinatus	http://media.eol.org/content/2013/03/12/02/02019_o
Apogonidae	Apogon cathetogramma	http://media.eol.org/content/2009/05/19/11/66920_o
Apogonidae	Apogon ellioti	http://media.eol.org/content/2012/01/28/02/17488_o
Apogonidae	Apogon erythrinus	http://media.eol.org/content/2013/03/12/03/67429_o
Apogonidae	Apogon fuscus	http://media.eol.org/content/2012/12/08/06/28366_o
Apogonidae	Apogon lineatus	http://media.eol.org/content/2013/04/25/04/86868_o
Apogonidae	Apogon maculatus	http://media.eol.org/content/2013/05/12/08/86018_o
Apogonidae	Apogon niger	http://media.eol.org/content/2012/12/08/06/91453_o
Apogonidae	Apogon semilineatus	http://media.eol.org/content/2013/03/12/03/55521_o
Apogonidae	Astrapogon puncticulatus	http://media.eol.org/content/2012/01/28/03/87521_o
Apogonidae	Cheilodipterus isostigmus	http://media.eol.org/content/2014/03/21/04/91392_o
Apogonidae	Eleotris acanthopoma	http://media.eol.org/content/2013/03/12/04/35642_o
Apogonidae	Fowleria isostigma	http://media.eol.org/content/2012/12/08/06/93533_o
Apogonidae	Gillichthys mirabilis	http://media.eol.org/content/2009/05/21/16/74503_o
Apogonidae	Glossamia aprion	http://media.eol.org/content/2013/04/25/02/81035_o
Apogonidae	Glossamia aprion	http://media.eol.org/content/2013/04/25/02/81035_o
Apogonidae	Zoramia fragilis	http://media.eol.org/content/2009/05/19/11/21625_o
Apogonidae	Zoramia leptacantha	http://media.eol.org/content/2009/05/19/11/27556_o
Balistoidae	Abalistes stellatus	http://media.eol.org/content/2012/12/08/06/90150_o
Balistoidae	Aluterus heudelotii	http://media.eol.org/content/2013/11/25/09/53533_o
Balistoidae	Aluterus schoepfii	http://media.eol.org/content/2013/04/05/10/38088_o
Balistoidae	Aluterus scriptus	http://media.eol.org/content/2012/01/28/03/80122_o
Balistoidae	Balistes capriscus	http://media.eol.org/content/2012/01/28/03/21930_o
Balistoidae	Balistes punctatus	http://media.eol.org/content/2012/01/28/03/92305_o
Balistoidae	Balistes vetula	http://media.eol.org/content/2013/04/25/06/10464_o
Balistoidae	Cantherhines dumerilii	http://media.eol.org/content/2013/04/25/03/94197_o
Balistoidae	Cantherhines pullus	http://media.eol.org/content/2013/04/05/10/14161_o
Balistoidae	Canthidermis sufflamen	http://media.eol.org/content/2009/05/19/11/00557_o
Balistoidae	Chaetodermis penicilligerus	http://media.eol.org/content/2013/03/12/03/29267_o
Balistoidae	Monacanthus ciliatus	http://media.eol.org/content/2009/11/17/08/05018_o
Balistoidae	Monacanthus tuckeri	http://media.eol.org/content/2012/01/28/03/99759_o
Balistoidae	Paramonacanthus sulcatus	http://media.eol.org/content/2013/03/12/03/29915_o
Balistoidae	Pseudomonacanthus macrurus	http://media.eol.org/content/2012/01/28/01/71658_o
Balistoidae	Stephanolepis auratus	http://media.eol.org/content/2013/04/25/02/89115_o
Balistoidae	Stephanolepis hispidus	http://media.eol.org/content/2012/01/28/04/00266_o
Balistoidae	Sufflamen albicaudatum	http://media.eol.org/content/2009/05/19/11/74954_o
Balistoidae	Sufflamen chrysopterum	http://media.eol.org/content/2013/04/25/02/44206_o

family	species	url
Balistoidae	Sufflamen fraenatum	http://media.eol.org/content/2009/05/19/11/52331_o
Balistoidae	Thamnaconus tessellatus	http://media.eol.org/content/2012/12/08/06/25448_o
Balistoidae	Xanthichthys lineopunctatus	http://media.eol.org/content/2013/04/25/04/74369_o
Balistoidae	Xanthichthys ringens	http://media.eol.org/content/2011/02/12/06/34197_o
Chaetodontidae	Chaetodon burgessi	http://media.eol.org/content/2011/12/13/10/78329_o
Chaetodontidae	Chaetodon capistratus	http://media.eol.org/content/2012/01/28/00/41497_o
Chaetodontidae	Chaetodon interruptus	http://media.eol.org/content/2013/04/25/03/86257_o
Chaetodontidae	Chaetodon quadrimaculatus	http://media.eol.org/content/2013/04/25/05/89919_o
Chaetodontidae	Chaetodon robustus	http://media.eol.org/content/2012/01/28/03/97570_o
Chaetodontidae	Chaetodon sedentarius	http://media.eol.org/content/2011/02/12/04/03772_o
Chaetodontidae	Chaetodon striatus	http://media.eol.org/content/2012/01/28/03/59955_o
Chaetodontidae	Chaetodon zanzibarensis	http://media.eol.org/content/2013/04/25/06/44731_o
Chaetodontidae	Hemitaurichthys thompsoni	http://media.eol.org/content/2012/09/02/10/71444_o
Chaetodontidae	Heniochus singularis	http://media.eol.org/content/2009/05/19/13/15675_o
Chaetodontidae	Pomacanthus arcuatus	http://media.eol.org/content/2011/10/06/08/27615_o
Chaetodontidae	Prognathodes aculeatus	http://media.eol.org/content/2011/02/12/04/19347_o
Labridae	Bodianus oxycephalus	http://media.eol.org/content/2013/03/12/02/23019_o
Labridae	Bodianus pulchellus	http://media.eol.org/content/2011/02/12/04/09554_o
Labridae	Bodianus scrofa	http://media.eol.org/content/2013/04/25/05/26099_o
Labridae	Bodianus tanyokidus	http://media.eol.org/content/2013/03/12/02/40853_o
Labridae	Chlorurus atrilunula	http://media.eol.org/content/2013/04/25/02/79996_o
Labridae	Chlorurus oedema	http://media.eol.org/content/2013/03/12/02/72014_o
Labridae	Ctenolabrus rupestris	http://media.eol.org/content/2009/09/03/04/69448_o
Labridae	Halichoeres bivittatus	http://media.eol.org/content/2014/08/22/13/06369_o
Labridae	Halichoeres dispilus	http://media.eol.org/content/2009/05/19/16/96697_o
Labridae	Halichoeres dispilus	http://media.eol.org/content/2009/05/19/16/96697_o
Labridae	Halichoeres radiatus	http://media.eol.org/content/2013/04/05/11/82296_o
Labridae	Iniistius aneitensis	http://media.eol.org/content/2009/05/19/16/84671_o
Labridae	Labrus bergylta	http://media.eol.org/content/2013/04/05/10/97708_o
Labridae	Macropharyngodon bipartitus	http://media.eol.org/content/2009/05/19/16/85624_o
Labridae	Notolabrus gymnogenis	http://media.eol.org/content/2013/04/25/03/13097_o
Labridae	Oxycheilinus digramma	http://media.eol.org/content/2012/12/08/06/35745_o
Labridae	Pseudolabrus eoethinus	http://media.eol.org/content/2013/03/12/02/00626_o
Labridae	Scarus guacamaia	http://media.eol.org/content/2013/04/05/10/15436_o
Labridae	Scarus hoefleri	http://media.eol.org/content/2013/04/05/11/45097_o
Labridae	Sparisoma amplum	http://media.eol.org/content/2013/04/25/06/88651_o
Labridae	Sparisoma aurofrenatum	http://media.eol.org/content/2011/02/12/04/39538_o
Labridae	Sparisoma chrysopterum	http://media.eol.org/content/2013/04/05/10/24098_o
Labridae	Sparisoma cretense	http://media.eol.org/content/2012/01/28/03/15809_o
Labridae	Sparisoma rubripinne	http://media.eol.org/content/2013/04/05/10/68601_o
Labridae	Symphodus melops	http://media.eol.org/content/2010/03/24/05/11347_o
Labridae	Symphodus ocellatus	http://media.eol.org/content/2010/03/24/05/66624_o
Labridae	Tautoga onitis	http://media.eol.org/content/2013/04/05/10/97919_o
Labridae	Tautogolabrus adspersus	http://media.eol.org/content/2013/04/05/11/47611_o
Labridae	Tautogolabrus adspersus	http://media.eol.org/content/2013/04/05/10/66839_o
Labridae	Thalassoma lucasanum	http://media.eol.org/content/2009/05/19/17/72515_o
Labridae	Thalassoma noronhanum	http://media.eol.org/content/2013/05/12/08/26587_o
Labridae	Thalassoma noronhanum	http://media.eol.org/content/2013/05/12/08/26587_o
Labridae	Xyrichtys novacula	http://media.eol.org/content/2013/04/05/10/09513_o
Labridae	Xyrichtys splendens	http://media.eol.org/content/2013/05/12/08/03204_o
Pomacentridae	Abudefduf taurus	http://media.eol.org/content/2012/01/28/03/60668_o
Pomacentridae	Amblyglyphidodon orbicularis	http://media.eol.org/content/2009/05/19/19/82131_o

family	species	url
Pomacentridae	Amblyglyphidodon orbicularis	http://media.eol.org/content/2009/05/19/19/82131_o
Pomacentridae	Amblyglyphidodon ternatensis	http://media.eol.org/content/2012/01/28/00/44746_o
Pomacentridae	Amphiprion akallopisos	http://media.eol.org/content/2013/04/25/02/61114_o
Pomacentridae	Amphiprion chagosensis	http://media.eol.org/content/2013/04/25/02/86585_o
Pomacentridae	Amphiprion latifasciatus	http://media.eol.org/content/2013/04/25/03/12923_o
Pomacentridae	Amphiprion sebae	http://media.eol.org/content/2013/04/05/15/67914_o
Pomacentridae	Chromis atrilobata	http://media.eol.org/content/2009/05/19/20/99377_o
Pomacentridae	Chromis bami	http://media.eol.org/content/2012/01/28/00/03625_o
Pomacentridae	Chromis cyanea	http://media.eol.org/content/2011/02/12/04/46652_o
Pomacentridae	Chromis enchrysura	http://media.eol.org/content/2009/11/17/09/46367_o
Pomacentridae	Chromis multilineata	http://media.eol.org/content/2011/02/12/04/18810_o
Pomacentridae	Chromis ovatiformis	http://media.eol.org/content/2009/05/19/20/92849_o
Pomacentridae	Chromis punctipinnis	http://media.eol.org/content/2011/10/14/18/41733_o
Pomacentridae	Chrysiptera brownriggii	http://media.eol.org/content/2013/03/12/03/67671_o
Pomacentridae	Chrysiptera starcki	http://media.eol.org/content/2013/03/12/02/62703_o
Pomacentridae	Dischistodus pseudochrysopoecilus	http://media.eol.org/content/2012/01/28/02/21322_o
Pomacentridae	Hypsypops rubicundus	http://media.eol.org/content/2011/10/14/18/76684_o
Pomacentridae	Microspathodon chrysurus	http://media.eol.org/content/2011/02/12/04/04869_o
Pomacentridae	Pomacentrus alleni	http://media.eol.org/content/2013/04/25/02/35835_o
Pomacentridae	Pomacentrus caeruleopunctatus	http://media.eol.org/content/2013/04/25/06/57664_o
Pomacentridae	Pomacentrus callainus	http://media.eol.org/content/2009/05/19/20/92808_o
Pomacentridae	Pomacentrus coelestis	http://media.eol.org/content/2012/09/02/10/24168_o
Pomacentridae	Stegastes adustus	http://media.eol.org/content/2013/05/12/08/29190_o
Pomacentridae	Stegastes albifasciatus	http://media.eol.org/content/2012/01/28/01/89740_o
Pomacentridae	Stegastes diencaeus	http://media.eol.org/content/2013/05/12/08/44790_o
Pomacentridae	Stegastes lividus	http://media.eol.org/content/2013/03/12/03/33534_o
Pomacentridae	Stegastes partitus	http://media.eol.org/content/2013/05/12/08/52466_o
Pomacentridae	Stegastes variabilis	http://media.eol.org/content/2011/02/12/05/65583_o
Pomacentridae	Teixeirichthys jordani	http://media.eol.org/content/2013/03/12/02/21560_o
Tetraodontidae	Arothron firmamentum	http://media.eol.org/content/2013/03/12/02/32649_o
Tetraodontidae	Arothron firmamentum	http://media.eol.org/content/2013/03/12/02/32649_o
Tetraodontidae	Canthigaster papua	http://media.eol.org/content/2011/10/14/18/41134_o
Tetraodontidae	Canthigaster rostrata	http://media.eol.org/content/2012/01/28/04/24086_o
Tetraodontidae	Chelonodon pleurospilus	http://media.eol.org/content/2013/04/25/05/38964_o
Tetraodontidae	Colomesus asellus	http://media.eol.org/content/2013/09/03/14/84092_o
Tetraodontidae	Colomesus psittacus	http://media.eol.org/content/2013/04/25/05/37704_o
Tetraodontidae	Lagocephalus laevigatus	http://media.eol.org/content/2011/02/12/05/11763_o
Tetraodontidae	Lagocephalus laevigatus	http://media.eol.org/content/2011/02/12/05/11763_o
Tetraodontidae	Lagocephalus suezensis	http://media.eol.org/content/2009/05/19/23/43115_o
Tetraodontidae	Lagocephalus wheeleri	http://media.eol.org/content/2013/03/12/03/36662_o
Tetraodontidae	Sphoeroides annulatus	http://media.eol.org/content/2009/05/19/23/51988_o
Tetraodontidae	Sphoeroides dorsalis	http://media.eol.org/content/2009/11/17/11/88894_o
Tetraodontidae	Sphoeroides pachygaster	http://media.eol.org/content/2013/03/12/03/17803_o
Tetraodontidae	Sphoeroides parvus	http://media.eol.org/content/2009/11/17/11/36111_o
Tetraodontidae	Takifugu niphobles	http://media.eol.org/content/2012/01/27/23/77163_o
Tetraodontidae	Takifugu oblongus	http://media.eol.org/content/2013/03/12/02/31339_o
Tetraodontidae	Takifugu ocellatus	http://media.eol.org/content/2012/01/27/23/63374_o
Tetraodontidae	Takifugu poecilonotus	http://media.eol.org/content/2013/04/25/05/06640_o
Tetraodontidae	Takifugu porphyreus	http://media.eol.org/content/2012/01/27/23/07838_o
Tetraodontidae	Takifugu rubripes	http://media.eol.org/content/2012/01/27/21/32161_o
Tetraodontidae	Takifugu vermicularis	http://media.eol.org/content/2013/03/12/03/96506_o
Tetraodontidae	Takifugu xanthopterus	http://media.eol.org/content/2012/01/27/23/61990_o

family	species	url
Tetraodontidae	Tetractenos hamiltoni	http://media.eol.org/content/2013/04/25/03/88631_o
Tetraodontidae	Torquigener hypselogeneion	http://media.eol.org/content/2013/03/12/03/43051_o
Tetraodontidae	Torquigener hypselogeneion	http://media.eol.org/content/2013/03/12/03/43051_o
Tetraodontidae	Tylerius spinosissimus	http://media.eol.org/content/2013/03/12/02/07164_o

References

Chakrabarty, P. (2005). Testing Conjectures about Morphological Diversity in Cichlids of Lakes Malawi and Tanganyika. *Copeia*, **2005**, 359–373.

Nelson, J.S. (1994). *Fishes of the World*, 3rd ed.n. John Wiley; Sons, Inc., New York.

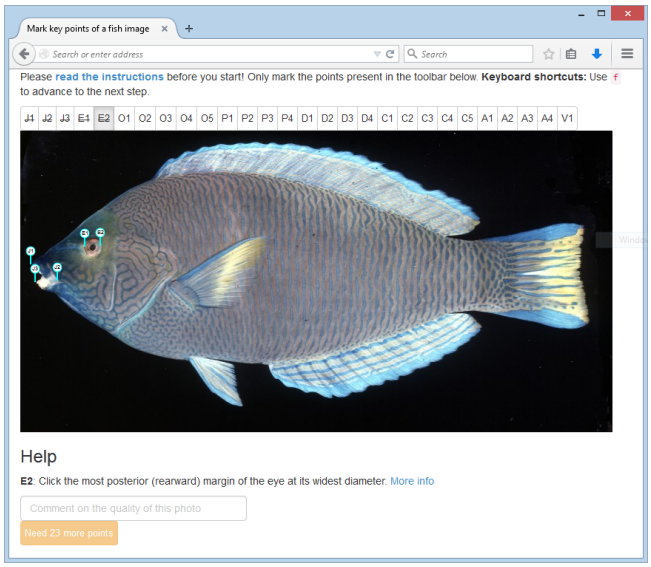

Figure S1: A screenshot of the web app that turkers used to digitize images. A live demonstration is available at https://jonchang.github.io/eol-mturk-landmark/

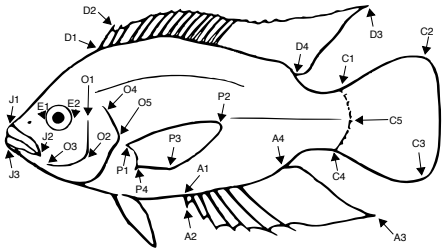

Figure S2: Description of landmarks used to digitize fish body shape. (J1) rostral tip of premaxilla (J2) ventral tip of premaxilla (J3) rostral tip of dentary (E1) anterior margin of midline through eye (E2) posterior margin of midline through eye (O1) dorsal end of preopercle (O2) ventral elbow of preopercle (O3) anterior end of preopercle (O4) dorsal end of opercle (O5) posterior end of opercle (D1) anterior insertion of dorsal fin (D2) distal tip of the anterior dorsal fin ray (D3) distal tip of the posterior dorsal fin ray (D4) posterior insertion of dorsal fin (P1) dorsal insertion of pectoral fin (P2) distal tip of the dorsal pectoral fin ray (P3) distal tip of the ventral pectoral fin ray (P4) ventral insertion of anal fin (A1) anterior insertion of anal fin (A2) distal tip of the anterior anal fin ray (A3) distal tip of the posterior anal fin ray (A4) posterior insertion of anal fin (C1) dorsal insertion of the caudal fin (C2) distal tip of the dorsal caudal fin ray (C3) distal tip of the the ventral caudal fin ray (C4) ventral insertion of the caudal fin (C5) midpoint of the caudal margin of the caudal peduncle.

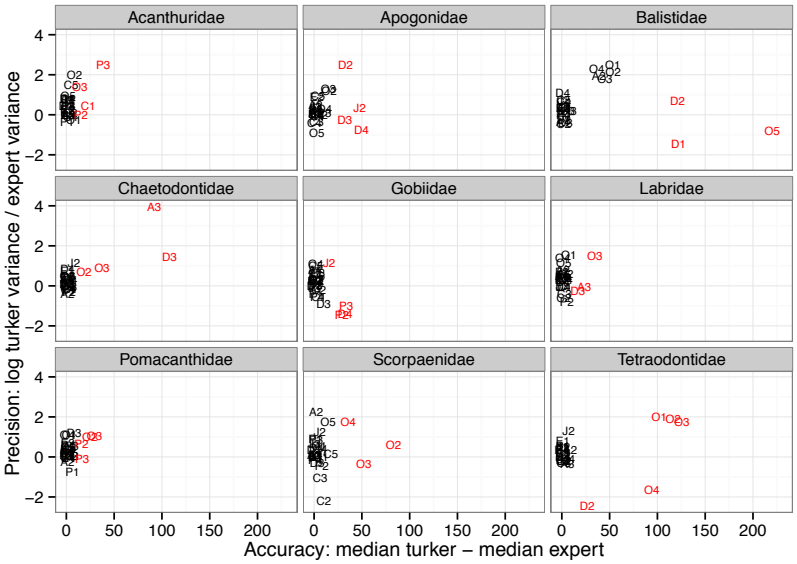

Figure S3: Version of Figure 1 where points are annotated with the landmark label.

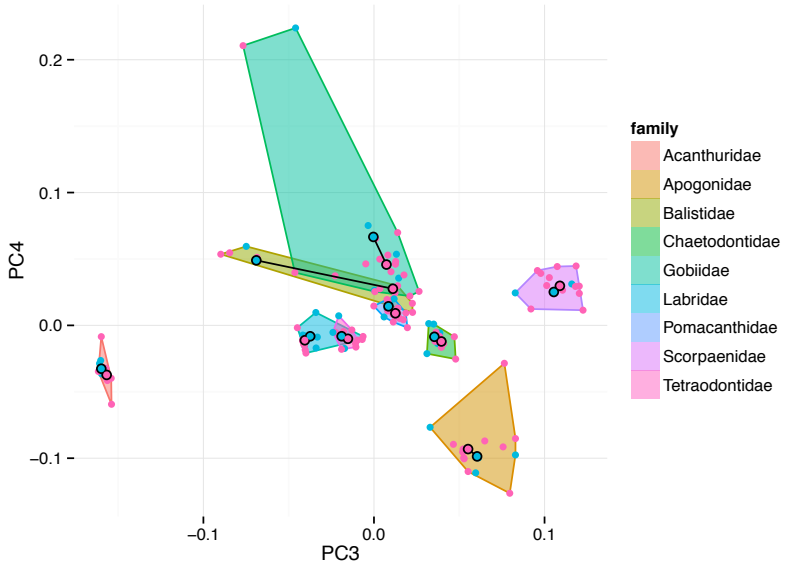

Figure S4: Morphospace projection of PC3 and PC4 for each observer's mean shape. Blue points indicate experts, while red points indicate turkers. The mean shape for all turkers and experts for a given family is the point outlined in black for each family, and connected with a black line to help emphasize the difference between turker and expert mean shapes. The convex hull for each family is drawn to show the amount of among-observer shape variation. PC 1 and 2 are shown in Fig 2.

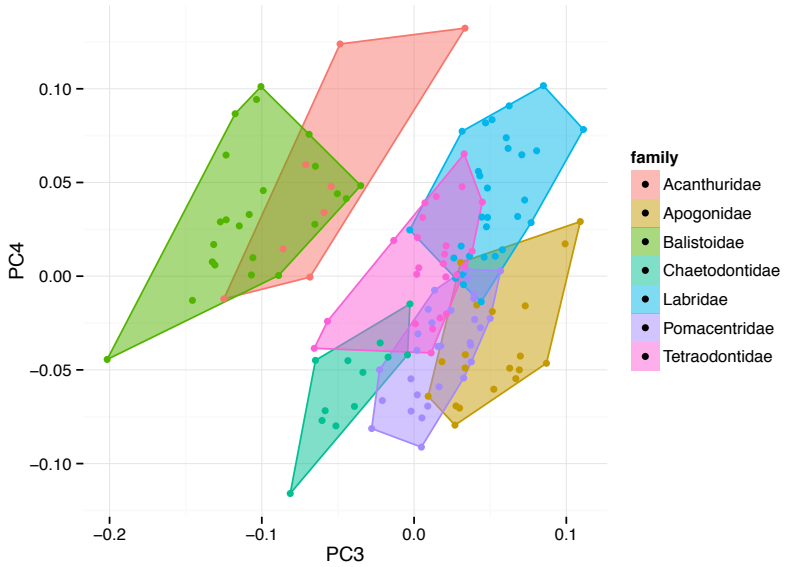

Figure S5: Morphospace of PC3 and PC4 for seven families of ray-finned fishes. Each point indicates a separate species; families are separated by colors. The convex hull for each family is drawn to show area of morphospace occupied by each family. The other PC axes are shown in Figs 4 and S6.

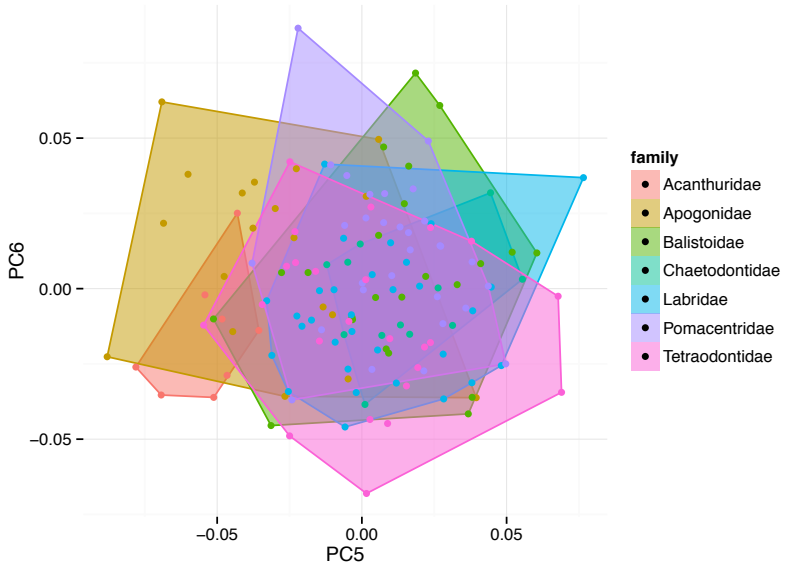

Figure S6: Morphospace of PC5 and PC6 for seven families of ray-finned fishes. Each point indicates a separate species; families are separated by colors. The convex hull for each family is drawn to show area of morphospace occupied by each family. The other PC axes are shown in Figs 4 and S5.

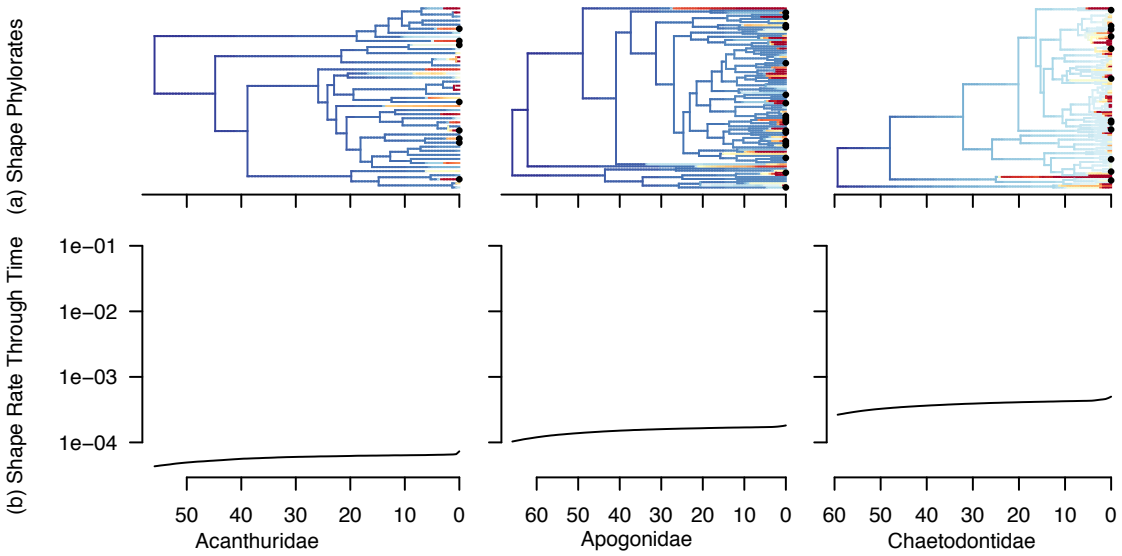

Figure S7: Rates of shape evolution for PC1 across three families of fishes. (a) Phylorate plots color branch lengths by rates of shape evolution, where warmer colors indicate faster rates of evolution. No significant rate shift events were detected within these families. (b) Median log rates of shape evolution through time. Analysis for the other four families are available in the main text, Figure 5.

CHAPTER 3

TACT: Taxonomic Addition for Complete Trees using birth-death-sampling estimators

3.1 Abstract

Phylogenies are critical components for analyses in ecology and evolutionary biology. However, many phylogenetic trees are incompletely sampled due to limitations in data and specimen availability. Generating complete phylogenies in the face of incomplete sampling generally requires attaching missing (unsampled) taxa to a backbone, the placement of which is estimated by imputation from the sampled data. Current methods for placing unsampled taxa onto a backbone generally assume a constant-rate branching process, a model that is unrealistic in the face of widespread rate heterogeneity in empirical systems. Here we present a method, TACT: Taxonomic Addition for Complete Trees, that uses birth-death-sampling estimators at nodes across an ultrametric backbone phylogeny to estimate branching times for unsampled taxa, then uses taxonomic information to compatibly place these new taxa onto a backbone phylogeny. Distributions of these completely sampled trees can greatly improve inference in diversification analyses, decreasing these methods' susceptibility to incorrect inference due to uneven sampling or rate heterogeneity across the tree of life.

3.2 Introduction

Phylogenetic trees, which describe the historical relationships between organisms, have become indispensable tools for answering questions in ecology and evolutionary biology,

ranging from systematics to biogeography and conservation. In macroevolutionary studies particularly, phylogenies have been used to understand historical patterns of diversification and fit various models that could generate the observed pattern of diversity. Studies of this nature, however, must be cautious when using incompletely sampled phylogenies, due to the potential of misleading results (Pybus and Harvey 2000).

Despite the advantages of using completely sampled phylogenies for evolutionary inference, these are still rare in empirical studies, save for young radiations of some model organisms, such as swordtails and Darwin's finches (Kang *et al.* 2013, Lamichhaney *et al.* 2015). In contrast, major groups of organisms that span deep time, such as mammals, birds, fishes, and plants, lack complete resolution (Bininda-Emonds *et al.* 2007, Jetz *et al.* 2012, Rabosky *et al.* 2013, Zanne *et al.* 2014). Even in smaller groups of organisms where complete sampling is more feasible, many obstacles stand in the way of completely sampled molecular phylogeny, ranging from the poor accessibility of tropical specimens (Reddy 2014) to recent extinction (but see e.g., Cooper *et al.* 2001).

Assuming a uniform global sampling fraction when sampling is non-random can severely bias estimates of diversification rates even in small ($n < 100$) phylogenies (Cusimano and Renner 2010, Brock *et al.* 2011, Höhna *et al.* 2011). Due to this issue, researchers have generally turned to three methods to address the lack of complete species-level phylogenies: modifying the likelihood function to condition on the degree of sampling, using terminally unresolved trees, or adding unsampled lineages with stochastic polytomy resolvers.

To account for unsampled lineages, several comparative methods (FitzJohn *et al.* 2009, Rabosky 2014) modify their likelihood function to account for the degree of incomplete sampling. The reconstructed birth-death process (Nee *et al.* 1994) uses a likelihood function that assumes complete sampling for extant taxa. It is possible to modify the extinction parameter to not just represent the probability of a lineage going extinct, but the probability that any lineage is not sampled at the present (Fig. 3.1).

This technique, however, requires that the downstream comparative method both (a)

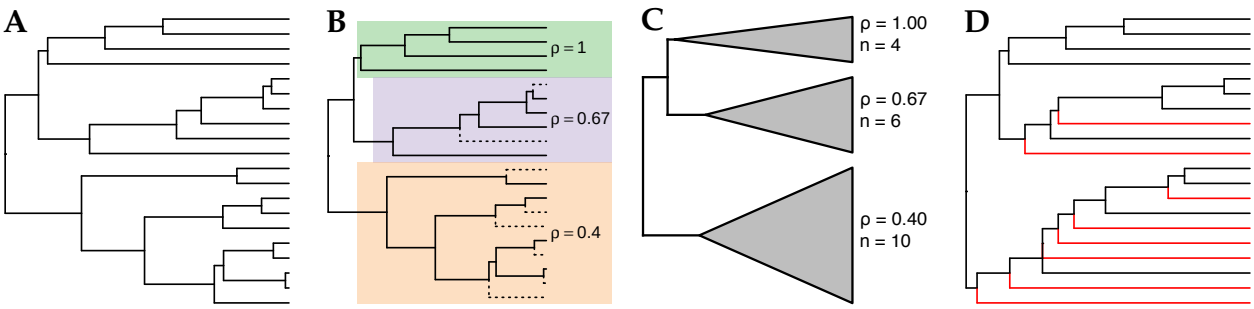

Figure 3.1: Comparison of different methods to accommodate incompletely sampled phylogenies. **(A)** The reconstructed phylogeny, which is unobserved by the researcher. **(B)** Incompletely sampled phylogeny, with unsampled lineages represented as dotted lines. Clades, which are highlighted in different colors, can be defined by the researcher and assigned different sampling fractions ρ, as used in BAMM. However, other methods, including BiSSE, must instead specify a global sampling fraction for the phylogeny, in this case $\rho = 0.6$. **(C)** Incompletely sampled phylogeny, with clades collapsed down to terminal exemplar lineages, represented as triangles, and with researcher assigned sampling fractions ρ and total richness n, as used in methods such as MEDUSA. Note that MEDUSA can use the fully-resolved version of the top "green" clade, since it is completely sampled; however, other methods will generally require all terminal taxa to be resolved to the same rank, such as the genus level, even if some genera are fully-resolved. **(D)** One realization of a stochastic polytomy resolution approach, which is used in methods such as PASTIS and TACT. Stochastically placed tips are highlighted in red.

implements this analytic correction, and (b) permits this correction to be applied non-uniformly across the phylogeny, which is an assumption of the original method (Pybus and Harvey 2000, Maddison *et al.* 2007). Methods such as BiSSE (Maddison *et al.* 2007), BAMM (Rabosky 2014), and RPANDA (Morlon *et al.* 2016) do implement this correction, but only BAMM allows the researcher to specify different sampling fraction across portions of a phylogeny. Alternative techniques add an extra parameter to account for the degree of overdispersed sampling (α in Brock *et al.* 2011), or replace the assumed uniform probability for sampling a lineage with one that maximizes or minimizes the edge lengths in the sampled phylogeny (f_{DS} and f_{CS} in Höhna *et al.* 2011). However, determining the precise α value or which sampling scheme to use can be problematic (Cusimano *et al.* 2012).

In other situations where species-level resolution is not possible, researchers can also use comparative methods that analyze phylogenies with terminal exemplar taxa, where a single tip may represent an entire genus or family, with sampling fractions and estimated richness assigned to the tips (Figure 3.1; e.g., MEDUSA, Alfaro *et al.* 2009, Pennell *et al.* 2014).

MEDUSA-like methods that rely on the mathematics of Magallon and Sanderson (2001) may have more power to detect diversification rate shifts in the face of rampant incomplete lineage sampling ($\rho < 0.5$)[1].

Incomplete sampling corrections that use the extinction rate to represent the probability of not observing a lineage for any reason have difficulty in young, species rich clades where the extinction rate is near zero and therefore a birth-death diversification model is inappropriate. However, comparative methods that use terminal exemplar taxa may in some cases be to conservative, as it becomes impossible to estimate parameters or rate shifts below the level of exemplar representation. Using phylogenies terminally unresolved tips may therefore discard data that would otherwise contribute to diversification analyses.

A final way to deal with incomplete sampling is to use a stochastic polytomy resolver to place missing (unsampled) species onto a sampled backbone phylogeny (Kuhn *et al.* 2011, Thomas *et al.* 2013). These have been used to successfully generate distributions of complete phylogenies for birds (Jetz *et al.* 2012), or to place recently-extinct or hypothesized species onto otherwise complete phylogenies (Revell *et al.* 2015). This class of methods lie upstream of the former two comparative methods, and are therefore fully agnostic with respect to the eventual comparative method used. Due to the advantages of this technique, we have developed a new method, TACT: Taxonomic Addition for Complete Trees, that uses taxonomic information combined with a time-calibrated backbone phylogeny to compatibly place unsampled lineages using a birth-death-incomplete sampling estimator. We describe the method and compare it to other stochastic polytomy resolvers.

[1]Dan Rabosky, "Hence, if you specify incomplete sampling fractions in BAMM, I would be surprised if you get the same results [as MEDUSA], at least if your tree is fairly incomplete (< 50% taxa sampled)", https://github.com/macroevolution/bamm/issues/135#issuecomment-106432496

3.3 The TACT Method

3.3.1 Method description

TACT requires a time-calibrated phylogeny; these can be estimated in a number of software programs, including BEAST (Drummond *et al.* 2012), MCMCtree (Rannala and Yang 2007), r8s (Sanderson 2003), and treePL (Smith and O'Meara 2012). The researcher must also supply a taxonomic tree where each tip in their clade of interest is represented, and polytomies at higher ranks (e.g., genus, family) represent monophyletic constraints that will be tested against the backbone phylogeny and potentially resolved in the complete tree. These taxonomy trees could be downloaded from the Open Tree of Life project (Hinchliff *et al.* 2015) or generated using the `as.phylo.formula` function from the R package ape (Paradis *et al.* 2004). For convenience, we also supply a command that will generate a taxonomy tree from a comma-separated values (CSV) file, `tact_build_taxonomic_tree`, where each column represents a taxonomic rank, from most inclusive to least inclusive, with the last column as the species name, and each row represents a separate species. This file must also be sorted alphabetically.

Once the researcher has both a time-calibrated phylogeny and a taxonomy tree, the `tact_add_taxa` command will start placing unsampled species to generate a realization of the complete phylogeny. For each defined taxonomy rank that was recovered as monophyletic in the backbone phylogeny, TACT performs a maximum likelihood estimate of the birth and death rates under the birth-death-sampling equation (Stadler 2009). These estimated rates are then used to parameterize a birth-death model to generate new waiting times for the unsampled species at this taxonomic rank.

However, if the probability of sampling the crown age of that node given the number of sampled taxa (Sanderson 1996) is below a user-defined threshold (option `--min-ccp`, default $= 0.8$), we instead walk up the ancestor chain to identify a valid taxonomic node that does fulfill that criteria.

The generated waiting times are bounded between the crown age of that clade and

the present time ($t = 0$). However, if the crown capture probability was less than 0.8, the maximum permitted age is extended to the stem age of the taxonomic node. These waiting times are used to randomly attach unsampled species to an existing branch within their assigned taxonomic rank, as long as these new species did not break the monophyly of nodes that were recovered as monophyletic and assigned a taxonomic rank, and constrained to not produce negative branch lengths due to a child node being added that was older than a parent node.

In the special case where all of the child branches of a taxonomic node belong to a monophyletic node, or if the crown capture probability of the entire clade is less than 0.8, the new species is instead assigned to the stem of that clade, and the waiting time will be bounded between the stem age and the crown age of that taxonomic node.

3.3.2 Implementation

TACT is implemented as a Python 2.7 package. Certain likelihood functions are based on code from the R packages TreePar and SimTree (Stadler 2009, 2011a,b). TACT depends on the Python packages SciPy (Oliphant 2007) and DendroPy (Sukumaran and Holder 2010), and uses the truncated-Newton bound-constrainted optimizer (Nash 1982) to perform the maximum-likelihood estimate of the speciation and extinction parameters.

3.4 Conclusion

Our method is similar to stochastic polytomy resolution as implemented in PASTIS (Thomas *et al.* 2013), but can instead use all available molecular data to construct the backbone phylogeny in a single analysis, rather than a two-stage process that begins with a reduced backbone dataset followed by separate tree searches for each crown lineage that jointly estimate the placement of species with and without molecular data (Jetz *et al.* 2012). Additionally, the MrBayes software (Ronquist *et al.* 2012) that PASTIS relies on for inference assumes a homogenous birth-death process across the entire phylogenetic tree, an unreal-

istic assumption given the common observation of diversification rates that vary among lineages and through time (e.g., Foote *et al.* 1999, Magallon and Sanderson 2001). TACT produces a local estimate of diversification rate at every taxonomic rank, permitting a more accurate placement of unsampled taxa, as diversification rate heterogeneity might significantly bias the inferred waiting times.

Although TACT uses the likelihood equations from CorSiM (Cusimano *et al.* 2012), it still represents a significant improvement on the functionality of that method. For example, CorSiM only generates estimated waiting times based on the sampled richness and waiting times; it does not place these splits onto a backbone phylogeny. It also does not easily permit diversification rate heterogeneity, and the researcher must instead manually separate clades that have may have different birth-death rates and estimate new waiting times on these hand-split lineages. With respect to CorSiM, TACT is much more flexible, as it will automatically split lineages to accommodate potential rate heterogeneity, and will also attempt to place unsampled lineages into a complete phylogeny subject to taxonomic constraints.

TACT easily accommodates among-lineage diversification rate heterogeneity, but we note that, in some instances, temporal diversification rate heterogeneity may also impact comparative inference. More complex, variable-rate diversification models, such as the ones implemented in GEIGER (Pennell *et al.* 2014), may be more appropriate in these situations. However, as the constant-rate process is generally assumed to be the null model in most comparative methods, we expect that the false positive rate for TACT-generated phylogenies to be low. Our conservative placement of species, where we assign species belonging to non-monophyletic ranks to the next highest monophyletic rank, will also tend to decrease the probability of Type I errors.

We have a presented a new method, TACT, that generates distributions of completely-resolved species-level phylogenies that can be used for any downstream comparative method. Reducing the deleterious impacts of nonrandom species sampling can greatly improve diversification analyses (Cusimano and Renner 2010, Brock *et al.* 2011, Höhna *et al.* 2011), and we see TACT playing an important role in comparative analyses where

phylogenies are incompletely sampled.

3.5 Acknowledgements

We thank Dan Rabosky for providing feedback on the methods, software, and manuscript prior to publication. This work was supported by an Encyclopedia of Life Rubenstein Fellowship (EOL-33066-13) and an NSF Doctoral Dissertation Improvement Grant (DEB-1601830) to JC. Travel support to disseminate this research was provided to JC by UCLA and the Society of Systematic Biologists.

3.6 Data Accessibility

All code is available on GitHub, https://github.com/jonchang/tact.

3.7 References

Michael E. Alfaro, Francesco Santini, Chad Brock, Hugo Alamillo, Alex Dornburg, Daniel L. Rabosky, Giorgio Carnevale, and Luke J Harmon. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences*, 106:13410–13414, 2009.

Olaf R P Bininda-Emonds, Marcel Cardillo, Kate E Jones, Ross D E MacPhee, Robin M D Beck, Richard Grenyer, Samantha A Price, Rutger A Vos, John L Gittleman, and Andy Purvis. The delayed rise of present-day mammals. *Nature*, 446(7135):507–512, 2007.

Chad D Brock, Luke J Harmon, and Michael E Alfaro. Testing for temporal variation in diversification rates when sampling is incomplete and nonrandom. *Systematic Biology*, 60(4):410–419, 2011.

Alan Cooper, Carles Lalueza-Fox, Simon Anderson, Andrew Rambaut, Jeremy Austin, and Ryk Ward. Complete mitochondrial genome sequences of two extinct moas clarify

ratite evolution. *Nature*, 409(6821):704–707, 2001.

Natalie Cusimano and Susanne S Renner. Slowdowns in diversification rates from real phylogenies may not be real. *Systematic Biology*, 59(4):458–464, 2010.

Natalie Cusimano, Tanja Stadler, and Susanne S. Renner. A new method for handling missing species in diversification analysis applicable to randomly or nonrandomly sampled phylogenies. *Systematic Biology*, 61(5):785–792, 2012.

Alexei J. Drummond, Marc A. Suchard, Dong Xie, and Andrew Rambaut. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29:1969–1973, 2012.

Richard G. FitzJohn, Wayne P. Maddison, and Sarah P. Otto. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology*, 58:595–611, 2009.

Mike Foote, John P Hunter, Christine M Janis, and J John Sepkoski. Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals. *Science*, 283(5406):1310–1314, 1999.

Cody E. Hinchliff, Stephen A. Smith, James F. Allman, J. Gordon Burleigh, Ruchi Chaudhary, Lyndon M. Coghill, Keith A. Crandall, Jiabin Deng, Bryan T. Drew, Romina Gazis, Karl Gude, David S. Hibbett, Laura A. Katz, H. Dail Laughinghouse, Emily Jane McTavish, Peter E. Midford, Christopher L. Owen, Richard H. Ree, Jonathan A. Rees, Douglas E. Soltis, Tiffani Williams, and Karen A. Cranston. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, 112(41):201423041, 2015.

Sebastian Höhna, Tanja Stadler, Fredrik Ronquist, and Tom Britton. Inferring speciation and extinction rates under different sampling schemes. *Molecular Biology and Evolution*, 28(9):2577–2589, 2011.

W Jetz, G.H. H Thomas, J.B. B Joy, K. Hartmann, and A.O. O Mooers. The global diversity of birds in space and time. *Nature*, 491(7424):1–5, 2012.

Ji Hyoun Kang, Manfred Schartl, Ronald B. Walter, and Axel Meyer. Comprehensive phylogenetic analysis of all species of swordtails and platies (Pisces: Genus Xiphophorus) uncovers a hybrid origin of a swordtail fish, *Xiphophorus monticolus*, and demonstrates that the sexually selected sword originated in the ancestral lineage of the genus, but was lost again secondarily. *BMC Evolutionary Biology*, 13(1):25, Jan 2013.

Tyler S Kuhn, Arne Ø Mooers, and Gavin H Thomas. A simple polytomy resolver for dated phylogenies. *Methods in Ecology and Evolution*, 2(5):427–436, 2011.

Sangeet Lamichhaney, Jonas Berglund, Markus Sällman Almén, Khurram Maqbool, Manfred Grabherr, Alvaro Martinez-Barrio, Marta Promerová, Carl-Johan Rubin, Chao Wang, Neda Zamani, et al. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, 518(7539):371–375, 2015.

Wayne P Maddison, Peter E Midford, and Sarah P Otto. Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 56:701–710, 2007.

Susana Magallon and Michael J. Sanderson. Absolute diversification rates in angiosperm clades. *Evolution*, 55(9):1762–1780, 2001.

Hélène Morlon, Eric Lewitus, Fabien L Condamine, Marc Manceau, Julien Clavel, and Jonathan Drury. RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. *Methods in Ecology and Evolution*, 7(5):589–597, 2016.

Stephen G Nash. *Truncated-Newton Methods*. PhD thesis, Stanford University, 1982.

Sean Nee, Robert M May, and H Harvey PAUL. The reconstructed evolutionary process. *Phil. Trans. R. Soc. Lond. B*, 344:305–311, 1994.

Travis E Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3), 2007.

Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

Matthew W. Pennell, Luke J. Harmon, and Josef C. Uyeda. Is there room for punctuated equilibrium in macroevolution? *Trends in Ecology and Evolution*, 29(1):23–32, 2014.

Oliver G Pybus and Paul H Harvey. Testing macro–evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London B: Biological Sciences*, 267(1459):2267–2272, 2000.

Daniel L. Rabosky, Francesco Santini, Jonathan Eastman, Stephen A Smith, Brian Sidlauskas, Jonathan Chang, and Michael E. Alfaro. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nature Communications*, 4:1958, jun 2013.

Daniel L. Rabosky. Automatic Detection of Key Innovations, Rate Shifts, and Diversity-Dependence on Phylogenetic Trees. *PLoS ONE*, 9(2):e89543, feb 2014.

Bruce Rannala and Ziheng Yang. Inferring speciation times under an episodic molecular clock. *Systematic Biology*, 56(3):453–466, 2007.

Sushma Reddy. Whats missing from avian global diversification analyses? *Molecular phylogenetics and evolution*, 77:159–165, 2014.

Liam J. Revell, D. Luke Mahler, R. Graham Reynolds, and Graham J. Slater. Placing cryptic, recently extinct, or hypothesized taxa into an ultrametric phylogeny using continuous character data: A case study with the lizard anolis roosevelti. *Evolution*, 69(4):1027–1035, 2015.

Fredrik Ronquist, Maxim Teslenko, Paul Van Der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542, 2012.

M. J. Sanderson. How Many Taxa Must Be Sampled to Identify the Root Node of a Large Clade? *Systematic Biology*, 45(2):168–173, jun 1996.

Michael J Sanderson. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–302, 2003.

Stephen A. Smith and Brian C. O'Meara. TreePL: Divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics*, 28(20):2689–2690, 2012.

Tanja Stadler. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261(1):58–66, 2009.

Tanja Stadler. Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences*, 108(15):6187–6192, 2011.

Tanja Stadler. Simulating trees with a fixed number of extant species. *Systematic Biology*, 60(5):676–684, 2011.

Jeet Sukumaran and Mark T Holder. DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571, 2010.

Gavin H Thomas, Klaas Hartmann, Walter Jetz, Jeffrey B Joy, Aki Mimoto, and Arne O Mooers. PASTIS: an R package to facilitate phylogenetic assembly with soft taxonomic inferences. *Methods in Ecology and Evolution*, 4(11):1011–1017, 2013.

Amy E Zanne, David C Tank, William K Cornwell, Jonathan M Eastman, Stephen a Smith, Richard G FitzJohn, Daniel J McGlinn, Brian C O'Meara, Angela T Moles, Peter B Reich, Dana L Royer, Douglas E Soltis, Peter F Stevens, Mark Westoby, Ian J Wright, Lonnie Aarssen, Robert I Bertin, Andre Calaminus, Rafaël Govaerts, Frank Hemmings, Michelle R Leishman, Jacek Oleksyn, Pamela S Soltis, Nathan G Swenson, Laura Warman, and Jeremy M Beaulieu. Three keys to the radiation of angiosperms into freezing environments. *Nature*, 506(7486):89–92, 2014.

CHAPTER 4

An online resource for the ray-finned fish tree of life

4.1 Abstract

Using phylogenetic trees has been critical for comparative researchers investigating problems in ecology, evolution, and biodiversity. Yet despite the increase in the number of phylogenies published, phylogenetic inference itself remains a specialized skill requiring expert knowledge to correctly perform, potentially limiting the pool of phylogenetic information available. Research requiring phylogenetic data is therefore challenges in obtaining such data, potentially slowing down progress in evolutionary biology. To resolve this problem, here we present a web resource for a recent phylogeny, that provides convenient access to our sequences, phylogenies, and fossil calibrations. These data are already vetted for quality, and as they are also available in pre-subsetted varieties (by e.g., family), they will facilitate phylogenetic reuse and increase access to phylogenetic datasets. We demonstrate some example use cases and conclude by advocating for similar approaches in other taxonomic groups.

4.2 Introduction

Phylogenies are now commonplace in analyses in evolutionary biology, and are used for myriad purposes, including studies of classification, diversification, trait evolution, and community composition. The light that phylogenetic research can shine on open questions in biology is clouded by the fact that inferring phylogenies is quite challenging and fraught with peril for non-specialist researchers. One way of avoiding these pitfalls is re-

using existing phylogenies, which can make phylogenetic knowledge accessible without requiring researchers to collaborate with phylogenetic experts or learn phylogenetics themselves. However, surveys of the biological literature have estimated that 60-95% of previously-published phylogenetic datasets are no longer accessible (Stoltzfus *et al.* 2012, Drew *et al.* 2013, Magee *et al.* 2014, McTavish *et al.* 2017), pointing to a disturbing failure of the scientific community to share data and potentially creating a major barrier to new comparative analyses.

One alternative solution is a "tree of life" approach, to centralize research effort in order to create a standard phylogenetic dataset that anyone can subset and reuse (McTavish *et al.* 2017). These broad phylogenies, in diverse groups such as mammals, birds, fishes, squamate reptiles, and plants (Bininda-Emonds *et al.* 2007, Jetz *et al.* 2012, Rabosky *et al.* 2013, Pyron *et al.* 2013, Zanne *et al.* 2014), represent the best target for phylogenetic re-use, as their diverse sampling means it is likely to cover the species that a typical taxon-focused researcher would be interested in. However, even with the release of tools such as the Open Tree of Life (Hinchliff *et al.* 2015), it is still not easy to reuse or subset these megaphylogenies, nor is it straightforward to integrate them with other data sources without substantial programming expertise.

Progress in evolutionary biology has therefore been hindered by the difficulty of accessing, reusing, and remixing phylogenetic data. Reuse is hamstrung by three major problems: vetting and curation (how to ensure that high-quality data and methods generated a phylogeny?), removing existing data (how to only use a portion of the megaphylogeny?), and adding new data (how to add new data to an existing megaphylogeny or a portion thereof?).[1]

We briefly survey existing approaches to making phylogenetic knowledge accessible in general, and efforts to do so specifically in the ray-finned fishes (Actinopterygii), the most diverse group of vertebrates with approximately 33,000 species.

[1]We do not address a potential fourth issue: community resistance to re-used trees (e.g., the common belief that, if you need a phylogeny, you should build it yourself to ensure its correctness).

4.2.1 General efforts

The 10kTrees project (10ktrees.nunn-lab.org; Arnold *et al.* 2010) permits researchers to download phylogenies for mammals, namely primates, even-toed ungulates, odd-toed ungulates, and carnivorans. Within these groups, the phylogram and chronogram are available to download, and taxonomic subsets of these phylogenies can also be custom generated and downloaded. In addition, the full multiple sequence alignment can also be downloaded. However, the fossil calibration information is not available except as text.

The BirdTree.org website (Jetz *et al.* 2012) similarly permits taxonomic subsets of the chronogram to be downloaded, as well as the full chronogram and multiple sequence alignment. Fossil calibration information cannot be downloaded.

The DateLife and Phylomatic projects (datelife.org; phylodiversity.net; Webb and Donoghue 2005, Stoltzfus *et al.* 2013) permit researchers to download taxonomic subsets from many published time-calibrated phylogenies. However, related data pertaining to these phylogenies, such as fossil calibrations and sequence alignments, cannot be downloaded from this service.

The TimeTree website (Hedges *et al.* 2006) permits researchers to interactively download phylogenies of taxonomic subsets using data from many different published phylogenies. However, machine reuse and synthesis is explicitly forbidden by the website, and a full data download is not available.

The Open Tree of Life project (opentreeoflife.org; Hinchliff *et al.* 2015) has an interactive interface to browse and download subsets of their synthetic phylogeny, possibly including polytomies at nodes where precise phylogenetic data are not available. The source phylogenies for their synthesis can all be downloaded, including expert curation information and taxa mapping. However, Open Tree phylogenies are all cladograms as they do not incorporate information about the timing of splitting events on a tree.

4.2.2 Efforts in ray-finned fish

The Euteleost Tree of Life (ETOL) phylogeny (Betancur-R *et al.* 2013) distributed in machine-readable formats a multiple sequence alignment and a phylogram. The fossil calibrations and chronogram were only present in the manuscript as text and figures, but we note that a chronogram was later made available for an update of the ETOL phylogeny (Betancur-R *et al.* 2017). A website for the ETOL project was originally published at fishtree.org, but that site is not operational as of 2017; deepfin.org now appears to host links to various iterations of the ETOL classification.

The Rabosky phylogeny (Rabosky *et al.* 2013) distributed in machine-readable formats a chronogram and a table of GenBank accession numbers. Fossil calibrations were present as a table in the text.

The Near phylogeny (Near *et al.* 2013) distributed in machine readable format a multiple sequence alignment, but the fossil calibrations and chronogram were only present in the manuscript as text and figures.

4.2.3 Our approach

Here we describe a new community resource, fishtreeoflife.org. This website provides our most recent phylogeny (Rabosky *et al.* in review) for the ray-finned fishes (class Actinopterygii), the most species-rich group of vertebrates representing over half of their diversity with approximately 33,000 species. We also include fossil information used to time-calibrate this phylogeny, and organize these data taxonomically using a new taxonomy derived from this new phylogeny and others (Rabosky *et al.* 2013, Betancur-R *et al.* 2013, Near *et al.* 2013). We finally demonstrate a few use cases for comparative biologists, and suggest that this pattern of providing resources be used as a template for the other branches on the tree of life.

4.3 Description

Our website aims to serve as a portal for comparative ichthyological research. Similar to Dryad, the full datasets used to generate our phylogeny are available for download, including the multiple sequence alignment, the phylogram from RAxML (Stamatakis 2014), the time-calibrated phylogeny from treePL (Smith and O'Meara 2012), and the fossil calibrations used to calibrate that phylogeny.

We also have tools for researchers to explore different subsets of the fish tree of life, browsing by taxonomic family and browsing by fossil calibration.

4.3.1 Browsing taxonomic subsets

We expect most researchers to approach our online resource from a taxon-specific perspective. We therefore have created a page for each family we included in our taxonomy. On each family page, researchers can see the full list of species included in the family, whether that species is placed on the phylogeny with molecular or merely taxonomic data, the fossil calibrations associated with that family, and downloads related to that family. These downloads include both the phylogram, inferred via RAxML (Stamatakis 2014), and the chronogram, time-calibrated by treePL (Smith and O'Meara 2012).

Researchers can directly use these phylogenies in e.g., R using the APE package (Paradis *et al.* 2004). The following example downloads the tree for the Acanthuridae (surgeonfishes) and generates a lineage-through-time plot:

```
library(ape)
url <- "https://fishtreeoflife.org/downloads/family/Acanthuridae.tre"
tree <- read.tree(url)
ltt.plot(tree)
```

Aligned sequence data are also available to download. This permits a researcher who has collected their own genetic data to simply use profile alignment from e.g., MAFFT

(Katoh and Toh 2008) to incorporate their new data into our existing, validated multiple sequence alignment. This increases the speed at which researchers can, for example, infer a new phylogenetic tree for a taxon of interest.

4.3.2 Browsing fossil calibrations

We have also included a fossil section to our Fish Tree of Life website. The fossils page lists all 139 fossil calibrations used in our analysis, as well as the phylogenetic placement of those fossils on the phylogeny.

Each fossil has its own page associated with it, which includes the exact fossil taxon being used to calibrate the group (e.g., crown Acanthuridae), as well as the minimum age, authority for fossil placement, and fossil locality. We also incorporate the maximum 95% estimated age through the Hedman fossil outgroup process (Hedman 2010), and list the fossil outgroup sequence used to calculate that maximum estimated age. These ages were used in the treePL (Smith and O'Meara 2012) dating analysis to provide upper bounds for calibrations.

4.3.3 Downloading data

Downloads of our entire dataset are available through the Downloads section of the website. Information on individual pages, such as the family-level taxonomy pages, can also be downloaded in a machine-readable Javascript Object Notation (JSON) format. Phylogenetic and sequence data are also provided, in the standard Newick and Phylip formats, for each of these subsets as well. We anticipate that these pre-subsetted data will lower the barrier of entry for comparative researchers to begin using phylogenetic and molecular data without tedious preparation and data cleaning and integration steps.

4.3.4 Contributing data

Researchers from other lab groups can easily contribute additional phylogenies, sequence matrices, and fossil calibrations and add them to our dataset. As our web resource is developed using Github, and all the associated data are stored on Github, users simply need to create a "pull request" that adds their own data into the repository. The merged pull request will automatically build the Fish Tree of Life website using our continuous integration infrastructure. Based on the data provided, the website will automatically include links to additional, user-contributed datasets or subsets of datasets on the appropriate taxon or fossil page.

4.4 Conclusion

We have presented a comprehensive web resource for comparative ichthyologists, and researchers generally interested in macroevolutionary questions. Our resource has numerous facilities to permit researchers to easily use manageable subsets of an otherwise dauntingly large dataset. We believe that this is a key step forward to making phylogenetic data available to comparative researchers, and will help to close the gap between researchers skilled at generating phylogenies, and researchers interested in answering other empirical or theoretical questions that may not necessarily have an affinity for phylogenetic inference.

Our website can be accessed at `https://fishtreeoflife.org`. The source code is available on on GitHub, `https://github.com/jonchang/fishtreeoflife.org`.

4.5 Acknowledgements

This work was supported by an Encyclopedia of Life Rubenstein Fellowship (EOL-33066-13) and an NSF Doctoral Dissertation Improvement Grant (DEB-1601830) to JC. Travel support to disseminate this research was provided to JC by UCLA and the Society of Systematic Biologists.

4.6 References

Christian Arnold, Luke J. Matthews, and Charles L. Nunn. The 10ktrees website: A new online resource for primate phylogeny. *Evolutionary Anthropology: Issues, News, and Reviews*, 19(3):114–118, 2010.

Ricardo Betancur-R, Richard E. Broughton, Edward O. Wiley, Kent Carpenter, J. Andrés López, Chenhong Li, Nancy I. Holcroft, Dahiana Arcila, Millicent Sanciangco, James C. Cureton, Feifei Zhang, Thaddaeus Buser, Matthew A. Campbell, Jesus A. Ballesteros, Adela Roa-Varon, Stuart Willis, W. Calvin Borden, Thaine Rowley, Paulette C. Reneau, Daniel J. Hough, Guoqing Lu, Terry Grande, Gloria Arratia, and Guillermo Ortí. The Tree of Life and a New Classification of Bony Fishes. *PLoS Currents*, 2013.

Ricardo Betancur-R, Edward O Wiley, Gloria Arratia, Arturo Acero, Nicolas Bailly, Masaki Miya, Guillaume Lecointre, and Guillermo Ortí. Phylogenetic classification of bony fishes. *BMC evolutionary biology*, 17(1):162, 2017.

Olaf R P Bininda-Emonds, Marcel Cardillo, Kate E Jones, Ross D E MacPhee, Robin M D Beck, Richard Grenyer, Samantha A Price, Rutger A Vos, John L Gittleman, and Andy Purvis. The delayed rise of present-day mammals. *Nature*, 446(7135):507–512, 2007.

Bryan T Drew, Romina Gazis, Patricia Cabezas, Kristen S Swithers, Jiabin Deng, Roseana Rodriguez, Laura A Katz, Keith A Crandall, David S Hibbett, and Douglas E Soltis. Lost branches on the tree of life. *PLoS Biology*, 11(9):e1001636, 2013.

S Blair Hedges, Joel Dudley, and Sudhir Kumar. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23):2971–2972, 2006.

Matthew M. Hedman. Constraints on clade ages from fossil outgroups. *Paleobiology*, 36(1):16–31, 2010.

Cody E. Hinchliff, Stephen A. Smith, James F. Allman, J. Gordon Burleigh, Ruchi Chaudhary, Lyndon M. Coghill, Keith A. Crandall, Jiabin Deng, Bryan T. Drew, Romina Gazis,

Karl Gude, David S. Hibbett, Laura A. Katz, H. Dail Laughinghouse, Emily Jane McTavish, Peter E. Midford, Christopher L. Owen, Richard H. Ree, Jonathan A. Rees, Douglas E. Soltis, Tiffani Williams, and Karen A. Cranston. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, 112(41):201423041, 2015.

W Jetz, G.H. H Thomas, J.B. B Joy, K. Hartmann, and A.O. O Mooers. The global diversity of birds in space and time. *Nature*, 491(7424):1–5, 2012.

Kazutaka Katoh and Hiroyuki Toh. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, 9(4):286–98, jul 2008.

Andrew F Magee, Michael R May, and Brian R Moore. The dawn of open access to phylogenetic data. *PLoS One*, 9(10):e110268, 2014.

Emily Jane McTavish, Bryan T. Drew, Ben Redelings, and Karen A. Cranston. How and why to build a unified tree of life. *BioEssays*, 39(11):1700114–n/a, 2017. 1700114.

Thomas J. Near, Alex Dornburg, Ron I Eytan, Benjamin P Keck, W leo Smith, Kristen L Kuhn, Jon A Moore, Samantha A. Price, Frank T Burbrink, Matt Friedman, and Peter C. Wainwright. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. *Proceedings of the National Academy of Sciences*, 110:12738–12743, 2013.

Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

R Alexander Pyron, Frank T Burbrink, and John J Wiens. A phylogeny and revised classification of squamata, including 4161 species of lizards and snakes. *BMC evolutionary biology*, 13(1):93, 2013.

Daniel L. Rabosky, Francesco Santini, Jonathan Eastman, Stephen A Smith, Brian Sidlauskas, Jonathan Chang, and Michael E. Alfaro. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nature Communications*, 4:1958, jun 2013.

Daniel L Rabosky, Jonathan Chang, Pascal O Title, Peter F Cowman, Lauren Sallan, Matt Friedman, Kristin Kaschner, Cristina Garilao, Thomas J Near, and Michael E Alfaro. in review.

Stephen A. Smith and Brian C. O'Meara. TreePL: Divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics*, 28(20):2689–2690, 2012.

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

Arlin Stoltzfus, Brian O'meara, Jamie Whitacre, Ross Mounce, Emily L Gillespie, Sudhir Kumar, Dan F Rosauer, and Rutger A Vos. Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. *BMC Research Notes*, 5(1):574, 2012.

Arlin Stoltzfus, Hilmar Lapp, Naim Matasci, Helena Deus, Brian Sidlauskas, Christian M. Zmasek, Gaurav Vaidya, Enrico Pontelli, Karen Cranston, Rutger Vos, Campbell O. Webb, Luke J. Harmon, Megan Pirrung, Brian O'Meara, Matthew W. Pennell, Siavash Mirarab, Michael S. Rosenberg, James P. Balhoff, Holly M. Bik, Tracy A. Heath, Peter E. Midford, Joseph W. Brown, Emily Jane McTavish, Jeet Sukumaran, Mark Westneat, Michael E. Alfaro, Aaron Steele, and Greg Jordan. Phylotastic! making tree-of-life knowledge accessible, reusable and convenient. *BMC Bioinformatics*, 14(1):158, May 2013.

Campbell O. Webb and Michael J. Donoghue. Phylomatic: tree assembly for applied phylogenetics. *Molecular Ecology Notes*, 5(1):181–183, 2005.

Amy E Zanne, David C Tank, William K Cornwell, Jonathan M Eastman, Stephen a Smith, Richard G FitzJohn, Daniel J McGlinn, Brian C O'Meara, Angela T Moles, Peter B Reich, Dana L Royer, Douglas E Soltis, Peter F Stevens, Mark Westoby, Ian J Wright, Lonnie Aarssen, Robert I Bertin, Andre Calaminus, Rafaël Govaerts, Frank Hemmings, Michelle R Leishman, Jacek Oleksyn, Pamela S Soltis, Nathan G Swenson, Laura Warman, and Jeremy M Beaulieu. Three keys to the radiation of angiosperms into freezing environments. *Nature*, 506(7486):89–92, 2014.

CHAPTER 5

Devouring the fish tree of life: the phylogenetic distribution of human exploitation

5.1 Abstract

Humans intensively harvest fishes, and although size-selective exploitation is known to cause large changes in exploited species' phenotypes, the macroevolutionary implications of this pervasive harvest remain unexplored. "Anthropogenic filtering", where human consumers preferentially exploit fishes with specific phenotypes, ecologies, or habitats, could pose a heightened risk to fish diversity at the macroevolutionary scale by exploiting fishes that are particularly vulnerable or provide unique functions to their ecosystems. We test this hypothesis with respect to three axes of fish biodiversity: 1) phylogenetic diversity, 2) phenotypic diversity measured through body size, and 3) ecological diversity measured through habitat. Consistent with the anthropogenic filter hypothesis, we find that fished species are more closely related to each other than expected. We also show that exploited species tend to be larger than their unexploited sister lineages, and that exploited species are overrepresented in reef-associated systems. Our results suggest that human exploitation of fishes is likely to be more disruptive to ecosystem function due to size-selective harvesting at the macroevolutionary level, and exerts heightened pressure on fish biodiversity, both in productive reef-associated environments, and overall due to its uneven phylogenetic distribution. Our results have broad implications for marine conservation efforts to mitigate these potentially negative effects of anthropogenic exploitation.

5.2 Main text

Human harvesting of fish species dates back to some of the earliest archeological records (Jerardino *et al.* 1992), but over time these subsistence harvests have given way to industrial fishing operations, which have had immense impacts on global fish populations (Worm *et al.* 2009). This harvesting has had increasingly well-documented effects on the ecology and population biology of individual species and assemblages (Jorgensen 1990, Law 2000, Fenberg and Roy 2008). For example, size-selective harvesting acts as a powerful selective force affecting not only body sizes of harvested species, but also many aspects of their life histories, such as age at maturity and fecundity (Jorgensen 1990, Law 2000, Heino *et al.* 2015). Because body size plays an important role in macroevolutionary dynamics (Jablonski 1996, Rabosky *et al.* 2013), these fisheries-induced declines in body size have significantly altered their natural evolutionary trajectories (Fenberg and Roy 2008).

These species-level assessments, especially in the wake of collapsing fisheries, have raised awareness and informed conservation priorities centered on species- or stock-specific management. Commercial exploitation may increase the vulnerability of exploited species to extinction through several mechanisms, including reduced population size, restriction of geographic range, and habitat or ecosystem alteration. Exploitation may also render species functionally extinct via the same mechanisms (Anderson *et al.* 2011, Galetti *et al.* 2013). However, we are unable to quantify the phylogenetic extent of exploitation or the threat to aspects of biodiversity, due to our lack of a broader macroevolutionary perspective that assesses harvesting on the fish tree of life. As larger lineages tend to play a more critical role in ecosystem function than smaller lineages (Solan 2004, Séguin *et al.* 2014), concentrated exploitation of fish lineages with shared phenotypic characters and/or habitat would alter their evolutionary trajectory and impair ecosystem function and productivity. A phylogenetic perspective also permits a broader point of view through deep time, as comparative biologists and paleontologists have often invoked "species selection" to explain why certain lineages seem to flourish and others fail to thrive (Jablonski 2008). Therefore, from a macroevolutionary perspective, humans may be acting as agents of

species selection (Carroll *et al.* 2014). Though the extent of the potential threat of human-mediated species selection to fish biodiversity is unknown, we hypothesize that human exploitation may act as such a filter on biodiversity at macroevolutionary scales, the "anthropogenic filter". Here we test this hypothesis with respect to three important aspects of fish biodiversity: phylogenetic, phenotypic, and ecological diversity.

Using a time-calibrated phylogeny of ray-finned fishes consisting of 11638 species (38.79% of ray-finned fish diversity and 65.21% of exploited species), we test the prediction of phylogenetic clustering by assessing the degree of relatedness associated with exploited species (Figure 5.1). If exploited species tend to be related to each other, this suggest that some shared evolutionary characteristic, such as a specific phenotype or ecology, predisposes lineages to experience exploitation, therefore enhancing the risk to ecosystem functioning (Purvis 2000). Exploited species were significantly phylogenetically clustered at both shallow and deep scales ($p_{mpd} = 0.025$, $p_{mntd} = 0.073$, Table 5.1). This result suggests that there is an intrinsically greater threat to fish biodiversity than would be expected if exploitation were randomly distributed across the phylogeny, and that the species-level effects of exploitation on life history traits, such as body size and reproductive age, are phylogenetically distributed in such a way to amplify their threat to biodiversity.

Exploitation type	Number of taxa	MPD	MNTD	%E(PD)
exploited	3106	0.001***	0.001***	10.0%***
unexploited	8388	0.088.	0.136	5.6%***
highly commercial	196	0.28	0.313	24.6%***
commercial	1505	0.025*	0.049*	12.6%***
minor commercial	1162	0.001***	0.001***	8.4%***
subsistence fisheries	243	0.472	0.505	13.1%***
of no interest	8166	0.025*	0.001***	5.5%***
of potential interest	35	0.313	0.313	-2.7%

Table 5.1: Statistics on the distribution of exploitation across the phylogeny. The 'exploited' and 'unexploited' categories are aggregations of the other categories. Exploited = commercial, highly commercial, minor commercial, subsistence fisheries. Unexploited = of no interest, of potential interest. Significance codes: $p < 0.001$: ***; $p < 0.01$: **; $p < 0.05$: *; $p < 0.1$: .

This pattern of clustering could be due to shared phenotypes or shared ecologies,

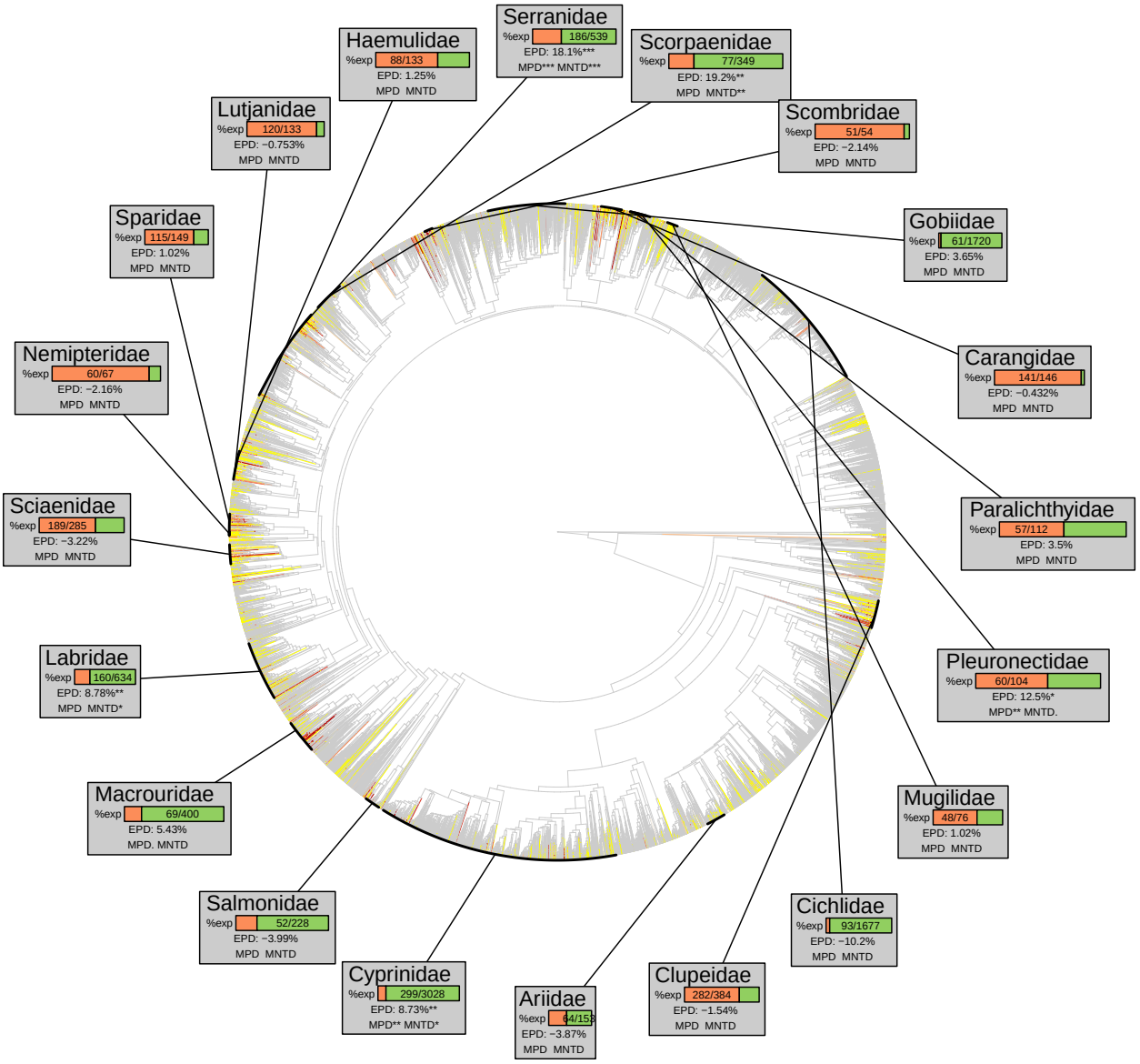

Figure 5.1: Phylogeny of ray-finned fishes, with species tips colored by mass caught between 2004–2014, or gray if that species was not commercially fished. The top 20 families by number of exploited species are highlighted as arcs drawn across the phylogeny; each box contains details on the fraction of exploited species richness within that family, the percent difference in expected phylogenetic diversity (dE(PD)) and its significance, and the significance of the mean nearest taxon distance (MNTD) and mean phylogenetic distance (MPD) for that family. Significance codes: $p < 0.001$: ***; $p < 0.01$: **; $p < 0.05$: *; $p < 0.1$:

or merely due to phylogenetic relatedness. Heritable factors that promote risk, such as body size, could explain the observed clustering pattern. Therefore, we test whether exploited species tend to be larger than unexploited species while also correcting for phylogenetic non-independence. As many large-bodied fish species are the basis of major commercial industries, we expect that larger-sized lineages are similarly preferentially exploited at the phylogenetic scale due to a species selection effect. Although exploited fish exist in a range of body sizes, we find that after accounting for shared ancestry using a phylogenetic generalized linear mixed model, lineages which have tended to evolve larger size are preferentially fished as well ($p < 0.001$). This effect is not merely driven by a few large clades; a more conservative sister lineage test similarly found that 701 out of 963 fully-exploited clades had average body sizes greater than their unexploited sister clades (Figure 5.2). Both of these analyses are consistent with the prediction that lineages that have evolved larger sizes tend to be more exploited. The strong signal of species-level size-selective harvesting suggests the possibility of fisheries-induced changes in species' phenotypes.

We also consider whether accessiblity to fishing grounds and economic productivity also plays a role in predicting exploitation. Commercially-exploited species should generally be present in shallower reef-associated environments, and less often in deep water environments where it is more challenging and unprofitable to fish. We therefore test whether habitat can predict exploitation and find that fishes in reef-associated environments are significantly and disproportionately overrepresented, being 2.10 times more likely to be exploited than the average fish ($p = 0.02$). In contrast, fishes in deeper waters, such as bathypelagic species, which live and feed below 300m, were 4.14 times less likely to be exploited ($p < 0.001$).

The disproportionate impact of fishing on these productive reef-associated habitats, combined with our previous pattern of size-selective harvesting, potentially has deleterious effects on ecosystem function. Large fish tend to play an important ecological role as top predators that regulate levels of smaller prey fish (Pauly *et al.* 1998, Jackson *et al.* 2001, Essington *et al.* 2006), and any reduction or extirpation of their populations will have a

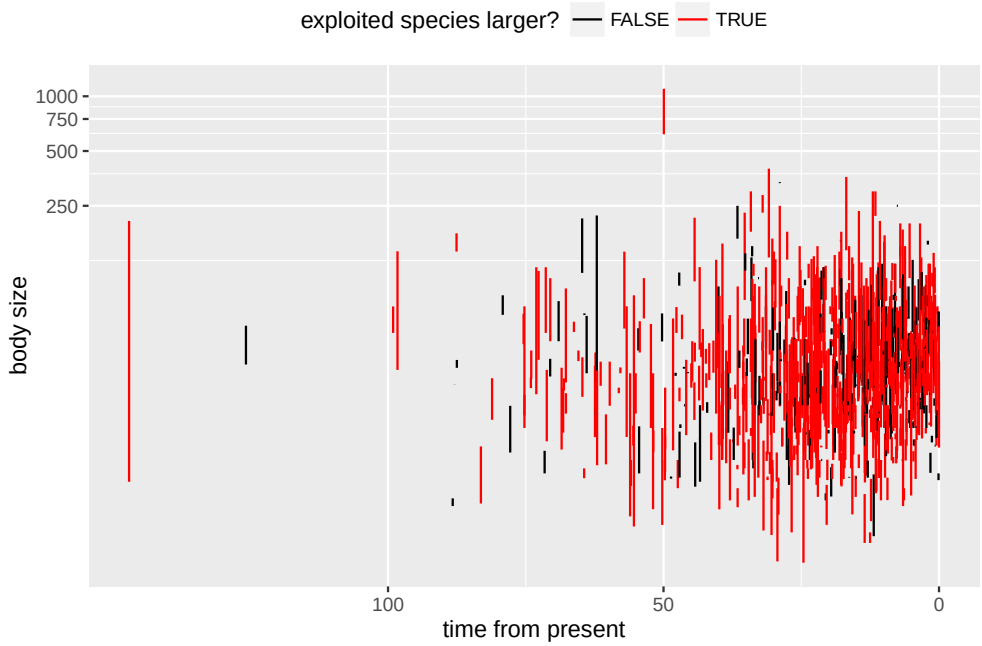

Figure 5.2: Sister lineage comparison of body size in exploited clades to unexploited clades. Each line represents the difference in body size between the left and right clade of a sister lineage comparison. For comparisons where the exploited clade is larger, the line is colored red; otherwise the line is colored black.

major impact on the stability of these ecosystems. Large-bodied fish also tend to occupy roles that have less functional redundancy than smaller-bodied species due to their lower abundances and higher trophic level (Bellwood *et al.* 2003, Séguin *et al.* 2014). Supposing that the reduction of diversity of large-bodied guilds of fish is maintained through time, the long-term productivity of habitats where fish are disproportionately harvested such as reefs would precipitously decline due to fewer predators and smaller individuals caused by species-level size-selective harvesting.

The phylogenetic structure of exploited species suggests that multiple processes are contributing to these observed patterns. In particular, the clustering of exploitation pressure close to the tips of the tree indicate exploited species have traits that tend to co-occur with recent diversification events. The observed link between speciation and rates of body size evolution in fishes (Rabosky *et al.* 2013, Heim and Knope 2015a) suggests that, if a single process generates both species richness and phenotypic disparity, filtering out fish species that have evolved morphological novelty may also reduce the rate at which fish species originate. We found a significant difference in affected phylogenetic diversity affected by fishing compared to a null model, ($p_{pd} = 0.001$). These results corroborate the idea that the anthropogenic filter could be reducing the density of lineages that are particularly exceptional from an evolutionary perspective.

The initial impacts anthropogenic filter and its downstream ecological and evolutionary consequences on ray-finned fishes is potentially alarming. Although there are no known recent extinctions of marine fishes, the threat to large-bodied fishes in the context of an anthropogenically-induced mass extinction in the marine realm cannot be ignored (Payne *et al.* 2016). Furthermore, the paleontological record suggests that the tendency for lineages evolve larger body size, termed Cope's rule, is often observed to co-occur with the tendency to evolve specialization and experience increased rates of extinction (Hallam 1975, Van Valkenburgh *et al.* 2004, Heim and Knope 2015b). Compounding these macroevolutionary risk factors with commercial harvesting that clusters on specific large-bodied clades could lead to an "anthropogenic filter" effect of these ecologically important and evolutionarily distinctive lineages. A new perspective on anthropogenic exploitation in light of our

results suggests that the current regime of exploitation is extremely deleterious, both in the short-term, by potentially reducing ecosystem function, and in the long-term, by robbing the fish tree of life of its evolutionary novelty or altering its evolutionary pressures through divergent selection. The pervasiveness of exploitation is exacerbated by the threat of the shifting baseline, as measuring the effect of anthropogenically-induced changes on a macroevolutionary timescale can be extremely challenging unless historical data exists (Simenstad *et al.* 1978, Dayton *et al.* 1998). The effect of the "anthropogenic filter" suggests that a redoubling of effort in fishery conservation efforts are warranted, due to the combined impact that clustered, size-selective, habitat- and ecology-specific harvesting will have on compromising ecosystem function and altering macroevolutionary trajectories.

5.3 Acknowledgements

We thank Jon Eastman for assisting with initial analyses. Code is hosted on Github (`https://github.com/jonchang/fisheries-exploitation`).

5.4 Methods

5.4.1 Data collected

We used a previously published phylogeny of ray-finned fishes for all the analyses in this study (Rabosky *et al.* 2013). We dropped species whose placement in the original phylogeny were not consistent with previously-published literature (see Supplemental Information). We collected exploitation data from FishBase (Froese and Pauly 2014), the United Nations Food and Agriculture Organization (FAO) Fisheries Report (FAO 2012), and the RAM Legacy Stock Assessment Database verison 3.0 (Ricard *et al.* 2012). If a species was recorded with any landings from 2003–2013 in the FAO database but was not in the FishBase dataset, it was coded as "commercial". We binned together exploited species ("highly commercial", "commercial", "minor commercial", and "subsistence fisheries").

Family	Exploited	Richness	% exploited	EPD	MPD	MNTD
Cyprinidae	299	3028	9.9%	8.7%**	0.003**	0.012*
Clupeidae	282	384	73.4%	-1.5%	0.927	0.757
Sciaenidae	189	285	66.3%	-3.2%	0.401	0.739
Serranidae	186	539	34.5%	18.1%***	0.001***	0.001***
Labridae	160	634	25.2%	8.8%**	0.667	0.03*
Carangidae	141	146	96.6%	-0.4%	0.8555	0.618
Lutjanidae	120	133	90.2%	-0.8%	0.981	0.7635
Sparidae	115	149	77.2%	1.0%	0.742	0.314
Cichlidae	93	1677	5.5%	-10.2%	0.998	0.759
Haemulidae	88	133	66.2%	1.3%	0.365	0.361
Scorpaenidae	77	349	22.1%	19.2%**	0.119	0.004**
Macrouridae	69	400	17.2%	5.4%	0.064.	0.63
Ariidae	64	153	41.8%	-3.9%	1	0.437
Gobiidae	61	1720	3.5%	3.6%	0.517	0.215
Nemipteridae	60	67	89.6%	-2.2%	0.6335	0.8755
Pleuronectidae	60	104	57.7%	12.5%*	0.004**	0.092.
Paralichthyidae	57	112	50.9%	3.5%	0.876	0.217
Salmonidae	52	228	22.8%	-4.0%	0.253	0.551
Scombridae	51	54	94.4%	-2.1%	0.617	0.593
Mugilidae	48	76	63.2%	1.0%	0.218	0.317

Table 5.2: Statistics for the top 20 families by number of exploited species.

5.4.2 Distribution of exploitation

To determine the level of phylogenetic clustering of exploitation risk among fish species, we computed the mean pairwise distance (MPD) and mean nearest taxon distance (MNTD) statistics (Webb *et al.* 2002). MPD is thought to be more sensitive to clustering towards the tips, while MNTD reveals clustering deeper in the tree. Statistical significance of phylogenetic clustering was determined by calculating standardized effect size (SES), which compares the the empirical statistics to a null distribution of statistics generated by randomizing tip labels 1,000 times. SES values less than 0.05 were interpreted as significant clustering.

We also calculated the phylogenetic diversity metric (PD Faith 1992, Webb *et al.* 2008) for all exploited fish species, and computed the SES value to assess significance. To quantify the potential increased loss of phylogenetic diversity compared to a model where

Family	Exploited	Richness	% exploited
Carangidae	141	146	96.6%
Lutjanidae	120	133	90.2%
Nemipteridae	60	67	89.6%
Scombridae	51	54	94.4%
Lethrinidae	36	38	94.7%
Gadidae	20	24	83.3%
Centropomidae	12	12	100.0%
Istiophoridae	11	12	91.7%
Trachinidae	8	9	88.9%
Triacanthidae	7	7	100.0%
Elopidae	6	7	85.7%
Fistulariidae	4	4	100.0%
Glaucosomatidae	4	4	100.0%
Polyprionidae	4	4	100.0%
Drepaneidae	3	3	100.0%
Psettodidae	3	3	100.0%
Anoplopomatidae	2	2	100.0%
Chirocentridae	2	2	100.0%
Coryphaenidae	2	2	100.0%
Dinopercidae	2	2	100.0%
Lateolabracidae	2	2	100.0%
Lobotidae	2	2	100.0%
Megalopidae	2	2	100.0%
Polyodontidae	2	2	100.0%
+16 monotypics	1	1	100.0%

Table 5.3: Statistics for families with more than 80% exploitation.

exploitation pressures are randomly distributed among tips, we calculated the percentage difference in expected PD (EPD Parhar and Mooers 2011).

To determine whether tree-wide patterns of clustered exploitation also applied within certain families, we split up the phylogeny by taxonomic family, and repeated the MPD, MNTD, and PD analyses on the family-level trees. We report details on the 20 families that have the most exploited species. All calculations were performed using custom routines written in R based on the packages ape (Paradis *et al.* 2004), picante (Kembel *et al.* 2010), and spacodiR (Eastman *et al.* 2011).

5.4.3 Relationship of exploitation to phenotype and ecology

To test how body size relates to exploitation, we gathered standard and total lengths of adult males from FishBase using the rfishbase package (Boettiger *et al.* 2012). For species that did not have a total length measurement, we used a regression analysis to convert standard length to total length, based on published conversion tables (Echeverria and Lenarz 1984, Gaygusuz *et al.* 2006). We corrected for relatedness using both a phylogenetic generalized least squares analysis, and a phylogenetic logistic regression (Ives and Garland 2010, Tung Ho and Ané 2014), in order to test whether exploited species tended to be larger than unexploited species. We also perform a sister-taxa comparison, and compare the average body size of fully-exploited clades and unexploited sister clades. The sister taxon approach is more conservative because unlike PGLS, it does not rely on correcting for shared ancestry using a Brownian correlation matrix, as sister taxa are by definition of equal age.

We also performed an integrated analysis using generalized linear mixed models (GLMM) as implemented in the MCMCglmm package (Hadfield 2010). We fit four logistic models with exploitation as a binary response variable and the phylogenetic covariance as a random effect, with the fixed effect ranging from a full model that included log-transformed body size and habitat type, to a null intercept only model. To evaluate model performance we used the deviance information criterion (DIC), where smaller values indicated better model fit. We also assessed convergence using the coda package (Plummer *et al.* 2006). The full model and model using body-size only predictors fit far better than the habitat-only and null model according to DIC, therefore, all reported GLMM results are from those analyses only.

5.5 References

Philip S L Anderson, Matt Friedman, Martin D Brazeau, and Emily J Rayfield. Initial radiation of jaws demonstrated stability despite faunal and environmental change.

Nature, 476(7359):206–209, 2011.

David R. Bellwood, Andrew S. Hoey, and J. Howard Choat. Limited functional redundancy in high diversity systems: Resilience and ecosystem function on coral reefs. *Ecology Letters*, 6:281–285, 2003.

Carl Boettiger, Duncan Temple Lang, and Peter Wainwright. rfishbase: exploring, manipulating and visualizing FishBase data from r. *Journal of Fish Biology*, nov 2012.

Scott P Carroll, Peter Søgaard Jørgensen, Michael T Kinnison, Carl T Bergstrom, R Ford Denison, Peter Gluckman, Thomas B Smith, Sharon Y Strauss, and Bruce E Tabashnik. Applying evolutionary biology to address global challenges. *Science*, 346(6207):1245993, 2014.

Paul K. Dayton, Mia J. Tegner, Peter B. Edwards, and Kristin L. Riser. Sliding baselines, ghosts, and reduced expectations in kelp forest communities, 1998.

Jonathan M Eastman, C E Timothy Paine, and Olivier J Hardy. spacodiR: structuring of phylogenetic diversity in ecological communities. *Bioinformatics*, 27(17):2437–8, sep 2011.

Tina Echeverria and William H Lenarz. Conversions between total, fork, and standard lengths in 35 species of Sebastes from California. *Fishery Bulletin*, 8(1):249–251, 1984.

Timothy E Essington, Anne H Beaudreau, and John Wiedenmann. Fishing through marine food webs. *Proceedings of the National Academy of Sciences*, 103(9):3171–5, feb 2006.

Daniel P Faith. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1):1–10, jan 1992.

FAO. The state of world fisheries and aquaculture. Technical report, 2012.

Phillip B. Fenberg and Kaustuv Roy. Ecological and evolutionary consequences of size-selective harvesting: how much do we know? *Molecular Ecology*, 17(1):209–20, jan 2008.

R Froese and Daniel Pauly. FishBase, 2014.

Mauro Galetti, Roger Guevara, Marina C. Côrtes, Rodrigo Fadini, 4 Sandro Von Matter, 5 Abraão B. Leite, 1 Fábio Labecca, 1 Thiago Ribeiro, 1 Carolina S. Carvalho, 7 Rosane G. Collevatti, 5 Mathias M. Pires, 6 Paulo R. Guimarães Jr., 6 Pedro H. Brancalion, and 1 Pedro Jordano8 Milton C. Ribeiro. Functional Extinction of Birds Drives Rapid Evolutionary Changes in Seed Size. *Science*, 340(May):1086–1091, 2013.

Özcan Gaygusuz, Çi Gürsoy, and M Özulug. Conversions of total, fork, and standard length measurements based on 42 marine and freshwater fish species (from Turkish waters). *Turkish Journal of Fisheries and Aquatic Sciences*, 84:79–84, 2006.

Jarrod D Hadfield. Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22, 2010.

A. Hallam. Evolutionary size increase and longevity in Jurassic bivalves and ammonites. *Nature*, 258(5535):493–496, dec 1975.

Noel A. Heim and Matthew Knope. Cope's rule in the evolution of marine animals. *Science*, 347(6224):867–870, 2015.

Noel A. Heim and Matthew Knope. Cope's rule in the evolution of marine animals. *Science*, 347(6224):867–870, 2015.

Mikko Heino, Beatriz Díaz Pauli, and Ulf Dieckmann. Fisheries-Induced Evolution. *Annual Review of Ecology, Evolution, and Systematics*, 46(1):annurev–ecolsys–112414–054339, 2015.

Anthony R. Ives and Theodore Garland. Phylogenetic logistic regression for binary dependent variables. *Systematic Biology*, 59(1):9–26, 2010.

David Jablonski. Body size and macroevolution. In David Jablonski, Douglas H. Erwin, and Jere H. Lipps, editors, *Evolutionary Paleobiology*, pages 256–289. University of Chicago Press, Chicago, 1 edition, 1996.

David Jablonski. Species Selection: Theory and Data. *Annual Review of Ecology, Evolution, and Systematics*, 39(1):501–524, dec 2008.

Jeremy B C Jackson, M X Kirby, W H Berger, K a Bjorndal, L W Botsford, B J Bourque, R H Bradbury, R Cooke, J Erlandson, J a Estes, T P Hughes, S Kidwell, C B Lange, H S Lenihan, J M Pandolfi, C H Peterson, R S Steneck, M J Tegner, and R R Warner. Historical overfishing and the recent collapse of coastal ecosystems. *Science*, 293(5530):629–37, jul 2001.

Antonieta Jerardino, Juan Carlos Castilla, José Miguel Ramírez, Nuriluz Hermosilla, Source Latin, American Antiquity, and No Mar. Early Coastal Subsistence Patterns in Central Chile : A Systematic Study of the Marine- Invertebrate Fauna from the Site of Curaumilla-1. *Latin American Antiquity*, 3(1):43–62, 1992.

T. Jorgensen. Long-term changes in age at sexual maturity of Northeast Arctic cod (Gadus morhua L.). *ICES Journal of Marine Science*, 46(3):235–248, jan 1990.

Steven W Kembel, Peter D Cowan, Matthew R Helmus, William K. Cornwell, Hélène Morlon, David D Ackerly, Simon P Blomberg, and Campbell O Webb. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, 26(11):1463–4, jun 2010.

R. Law. Fishing, selection, and phenotypic evolution. *ICES Journal of Marine Science*, 57(3):659–668, jun 2000.

Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

Rakesh K. Parhar and Arne Ø. Mooers. Phylogenetically clustered extinction risks do not substantially prune the Tree of Life. *PloS one*, 6(8):e23528, jan 2011.

Daniel Pauly, A W Trites, E Capuli, and V Christensen. Diet composition and trophic levels of marine mammals. *ICES Journal of Marine Science*, 55:467–481, 1998.

Jonathan L Payne, Andrew M Bush, Noel A Heim, Matthew L Knope, and Douglas J McCauley. Ecological selectivity of the emerrging mass extinction in the oceans. *Science*, 353:1284–1286, 2016.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

Andy Purvis. Nonrandom Extinction and the Loss of Evolutionary History. *Science*, 288(5464):328–330, apr 2000.

Daniel L. Rabosky, Francesco Santini, Jonathan Eastman, Stephen A Smith, Brian Sidlauskas, Jonathan Chang, and Michael E. Alfaro. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nature Communications*, 4:1958, jun 2013.

Daniel Ricard, Cóilín Minto, Olaf P. Jensen, and Julia K. Baum. Examining the knowledge base and status of commercially exploited marine species with the RAM Legacy Stock Assessment Database. *Fish and Fisheries*, 13(4):380–398, 2012.

Annie Séguin, Éric Harvey, Philippe Archambault, Christian Nozais, and Dominique Gravel. Body size as a predictor of species loss effect on ecosystem functioning. *Scientific Reports*, 4:4616, apr 2014.

Charles A. Simenstad, James A. Estes, and Karl W. Kenyon. Aleuts, Sea Otters, and Alternate Stable-State Communities. *Science*, 200(4340):403–411, apr 1978.

Martin Solan. Extinction and Ecosystem Function in the Marine Benthos. *Science*, 306(5699):1177–1180, nov 2004.

Lam Si Tung Ho and Cécile Ané. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Systematic Biology*, 63(3):397–408, 2014.

Blaire Van Valkenburgh, Xiaoming Wang, and John Damuth. Cope's rule, hypercarnivory, and extinction in North American canids. *Science*, 306(5693):101–4, 2004.

Campbell O Webb, David D Ackerly, Mark A McPeek, and Michael J Donoghue. Phylogenies and Community Ecology. *Annual Review of Ecology and Sustematics*, 33:475–505, 2002.

Campbell O Webb, David D Ackerly, and Steven W Kembel. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, 24(18):2098–100, sep 2008.

Boris Worm, Ray Hilborn, Julia K Baum, Trevor a Branch, Jeremy S Collie, Christopher Costello, Michael J Fogarty, Elizabeth a Fulton, Jeffrey a Hutchings, Simon Jennings, Olaf P Jensen, Heike K Lotze, Pamela M Mace, Tim R McClanahan, Cóilín Minto, Stephen R Palumbi, Ana M Parma, Daniel Ricard, Andrew a Rosenberg, Reg Watson, and Dirk Zeller. Rebuilding global fisheries. *Science*, 325(5940):578–85, jul 2009.