www.sciencemag.org/content/364/6435/74/suppl/DC1

Supplementary Materials for

Convergent regulatory evolution and loss of flight in paleognathous birds

Timothy B. Sackton*, Phil Grayson, Alison Cloutier, Zhirui Hu, Jun S. Liu, Nicole E. Wheeler,
Paul P. Gardner, Julia A. Clarke, Allan J. Baker, Michele Clamp, Scott V. Edwards*

*Corresponding author. Email: tsackton@g.harvard.edu (T.B.S.); sedwards@fas.harvard.edu (S.V.E.)

Published 5 April 2019, *Science* **364**, 74 (2019)
DOI: 10.1126/science.aat7244

This PDF file includes:

Figs. S1 to S15
Tables S1 to S8
Captions for data S1 to S5
References

Other supplementary material for this manuscript includes:

Data S1 to S5 (Excel format)

Materials and Methods

Sequencing and assembly of new genomes

We describe in detail the sequencing and assembly of 10 new paleognath genomes, analyzed in conjunction with 3 existing paleognath assemblies and the newly sequenced little bush moa (described in detail in a companion manuscript (*24*)). Specimens sequenced and related information for all 14 new and existing paleognath genomes are presented in Table S1.

Samples

Tissue samples were obtained for 10 species of paleognaths. Kiwi samples (*Apteryx haastii*, *Apteryx owenii*, *Apteryx rowi*) were obtained from samples available at the Royal Ontario Museum. DNA was extracted from blood from a single male individual for each species. Southern cassowary (*Casuarius casuarius*) DNA was obtained from a liver sample of a female individual provided by the Australian National Wildlife Collection. Thicket tinamou (*Crypturellus cinnamomeus*) DNA was extracted from tissue from a single male collected by Douglas Causey in Guanacaste, Costa Rica and accessioned in the collections of the Museum of Comparative Zoology, Harvard University. Emu (*Dromaius novaehollandiae*) DNA was obtained from a muscle sample from a single farm-raised male individual purchased from Songline Emu Farm in Gill, Massachusetts, and collected by Dan Janes. Elegant crested tinamou (*Eudromia elegans*) DNA was obtained from blood from a captive bred male in the Toronto Zoo, Ontario, Canada, collected by Graham Crawshaw. Chilean tinamou (*Nothoprocta perdicaria*) DNA was obtained from a blood sample collected by Kimberly Cheng, from a single captive bred male individual from a research flock housed in Chilliwack, British Columbia. Greater rhea (*Rhea americana*) DNA was obtained from a blood sample from a single male individual originally collected by Mark Peck of the Royal Ontario Museum. Lesser rhea (*Rhea pennata*) DNA was obtained from a liver biopsy provided by the Smithsonian National Zoo, from a single male originally born at Seaworld, Orlando, Florida. All DNA samples were imported to MCZ for sequencing under permits issued to the Museum of Comparative Zoology (USDA and CITES) and the Royal Ontario Museum (CITES). All experimental procedures (including for specimen collection) were approved by the relevant animal care committees at Harvard University or the Royal Ontario Museum.

DNA sequencing

DNA extraction was carried out following standard protocols using either the DNeasy blood & tissue kit (Qiagen) or the E.Z.N.A Tissue DNA Kit (Omega). Genome library preparation was carried out as previously described (*61*) to create libraries compatible with the ALLPATH-LG assembly method (*62*). Briefly, this algorithm requires at least one overlapping fragment library (i.e., a short insert library sequenced with paired-end reads long enough to overlap one another) alongside a jumping or mate-pair library. Short fragment libraries with an insert size of 220 bp were generated using the PrepX ILM 32i DNA library Kit (Wafergen Biosystems) and 3 kb jumping libraries were generated using the Nextera Mate Pair Sample Preparation Kit (Illumina). Both libraries were sequenced primarily or exclusively on the Illumina HiSeq 2500 using the V4 high output kit, generating 2x125 bp reads. To test library quantification and multiplexing accuracy, some samples were also sequenced on the Illumina HiSeq 2500 under rapid run mode, which generated 2x150 bp reads. Reads were demultiplexed using standard options with bcl2fastq and allowing no mismatches in the index read, except for one run (run id: C57WLANXX) where a high number of unidentified reads led us to reprocess allowing a Hamming distance of 3 in the index read, and a further unpooled run (run id: C57WPANXX) where all unidentified reads not mapping to PhiX were retained. All reads for each species were used for assembly, described below. Coverage per genome across all libraries was typically between 40-60x. Raw sequencing data is available from the NCBI SRA under BioProject PRJNA433110.

Assembly

We assembled each genome using the recommended ALLPATHS-LG approach, which we have previously described in detail in the context of avian-specific assemblies (*61*). Prior to assembly, we first trimmed adaptor sequence from both fragment and jumping libraries using Trimmomatic version 0.32 (*63*) with the ILLUMINACLIP option (ILLUMINACLIP:adapters.fa:2:30:10:1:true). The adapter sequence file consisted of both Apollo library preparation paired-end adapters and Nextera adapters. After adapter trimming, we verified sequence quality with FastQC (version 10.1), with the -k 5 option. All sequences used for assembly passed basic quality control checks. We assembled all libraries for each species with ALLPATHS-LG version 50191 (*62*), with HALPOIDIFY=TRUE and PLOIDY=2 options, but otherwise using the default parameters. To assess quality, we used BUSCO version 3.0.2 to infer the presence of highly conserved genes across our assemblies, which provides a reliable measure of assembly completeness (*64*). We ran BUSCO using the Odb9 Vertebrate set with chicken AUGUSTUS models and otherwise default parameters. Final assembly contiguity statistics (N50) as reported by ALLPATHS, and completeness statistics estimated using BUSCO, are presented in Table S2.

Moa genome sequencing

We assembled a reference-based nuclear genome assembly for the extinct little bush moa (*Anomalopteryx didiformis*) as described (*24*). In brief, DNA was extracted from the interior of a single toe bone following microblasting to remove the outer bone surface and grinding of the remaining material into fine powder, with subsequent enzymatic digestion (*65*) followed by purification using commercially available silica spin columns (DNeasy Blood & Tissue Kit; Qiagen, Germantown MD, USA). Sample processing occurred in a dedicated ancient DNA (aDNA) workspace in the Royal Ontario Museum (Toronto, Canada) and adhered to established best practices to minimize contamination risk for aDNA samples (*66*, *67*). Five sequencing libraries were constructed from the recovered DNA: 2 Illumina TruSeq DNA v3 libraries were constructed from DNA size selected in the 200-400 bp range that was subsequently sheared to a 200 bp insert size, an additional 2 libraries were generated using the Illumina Nextera XT Sample Preparation Kit with DNA below 500 bp in length, and a final Nextera library was built from input DNA in the 500bp-2Kb size range sheared to <700bp. All libraries were sequenced in paired-end mode (2x101 bp reads) on a HiSeq 2500 platform with Illumina v3 chemistry, yielding 143.4 Gbp of raw sequence data in total. Following pre-processing with Trimmomatic v. 0.32 (*63*) and options ILLUMINACLIP:[adapter_file]:2:30:10:1:true:SLIDINGWINDOW:10:13 MINLEN:25, reads were mapped to the draft emu genome reported here with Stampy v. 1.0.28 and a user-defined substitution parameter of 0.0839 that was estimated from an initial mapping to the emu reference using default Stampy parameters. Mapped reads were post-processed using Picard Tools v. 2.6.0 to mark and remove duplicates, Samtools v. 1.3.1 mpileup (*68*) was used to generate variant call format (VCF) output with MAPQ ≥ 30 and BASEQ ≥ 20, and a moa consensus sequence was generated with BCFTools v. 1.2. Reads were re-mapped to this initial moa genome assembly with Bowtie2 v. 2.2.9 (*69*) or improved mapping of short and/or more divergent reads, with subsequent post-processing steps as described above.

We additionally generated an error corrected version of the moa genome assembly (moa v2) using mapDamage2 v. 2.0.7 (*70*), which employs a Bayesian statistical model to recalibrate base quality scores to reflect probabilities of nucleotide misincorporations from post-mortem damage characteristic of ancient DNA. We first trimmed reads with Trimmomatic as described above, and then merged overlapping read pairs with PEAR v. 0.9.7 (*71*) before mapping reads to the moa reference described above with Bowtie2 v. 2.2.9 (*69*). We used Samtools to filter reads with MAPQ < 30, and removed duplicates with Picard as described above before passing the resulting BAM file to mapDamage with default parameters and the --rescale option invoked. Samtools v. 0.1.11 was used to generate files in 'pileup' format for each scaffold with BASEQ ≥ 20 specified for recalibrated base quality scores. The error corrected moa consensus sequence was called from this pileup output using a custom Perl script to mask bases with 'N's for positions with no coverage at the specified BASEQ cutoff. The original (prior to error-correction) version of the moa genome assembly was used for all phylogenetic analysis; the error-corrected version was used for all molecular evolutionary analysis of protein-coding and non-coding regions.

Section 1: Key summary

Data inputs:

Raw sequencing reads (fastq files): available at NCBI (BioProject PRJNA433110)

Methods and code:

Trim sequence reads (TRIMMOMATIC); assembly genomes (ALLPATHS); check completeness (BUSCO)

Github link: https://github.com/tsackton/ratite-genomics/tree/master/01_assembly

Data outputs:

Assembled genomes (fasta files): available at NCBI (BioProject PRJNA433110, PRJNA433423)

Annotating protein-coding genes

In order to annotate our newly assembled genomes, we generated new RNA-seq data from multiple tissues of two species. These data, along with published RNA-seq data from additional species, were used as input to MAKER to produce annotations (*72*).

RNA-seq data generation

We made RNA-seq libraries from flash frozen tissue stored in RNAlater at -80 degrees Celsius for two female emus (MCZ Cryo ID 6601 and 6608) and two Chilean tinamous (one male – ROM collection ID AJB6179, and one female – ROM collection ID AJB6180). For each species, we processed brain, liver, and gonad tissues from each individual for library preparation and sequencing. Following homogenization with a Tissumizer (Tekmar) RNA was extracted from each specimen using the RNeasy mini kit (Qiagen). Total RNA samples were processed with a PrepX PolyA kit (Wafergen Biosystems) on the Apollo 324 (IntegenX) to enrich for polyadenylated transcripts. Stranded RNA-seq libraries were generated from these samples using the PrepX mRNA 48 kit (Wafergen Biosystems) on the Apollo 324. Following PCR and bead cleanup with AmpureXP beads (Beckman Coulter), libraries were visualized on a 2200 TapeStation (Agilent) and quantified using a Library Quantification Kit (KAPA Biosystems) for multiplexing. Eleven libraries were sequenced on a NextSeq High 150 flowcell (Illumina) as 75 bp paired-end reads (the tube labeled Tinamou Female Liver sample failed during library preparation and was not included in the sequencing run). All raw FASTQ files generated are available from the NCBI SRA under BioProject PRJNA433114.

In addition to this newly sequenced data, we used previously published RNA-seq data (SRA accession ERR522068) from an additional species of kiwi, the North Island brown kiwi (*Apteryx mantelli*) to improve genome annotation (*26*). We also included previously published RNA-seq data from emu embryonic brain tissue (*73*) in our Trinity and Tophat runs to improve transcriptome coverage.

RNA-seq data analysis

Prior to the analysis described below, all RNA-seq data were trimmed using Trimmomatic version 0.32 to remove adaptor sequences and low-quality bases. We used the following Trimmomatic options: ILLUMINACLIP:adapters.fa:2:30:10:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:10 MINLEN:25.

Preliminary analysis for MAKER annotations

To prepare RNA-seq data for MAKER, we used two different approaches. First, we generated Trinity (*74*) de novo assemblies from all three species (North Island brown kiwi, emu, Chilean tinamou), and used those as EST evidence in MAKER. To produce de novo assemblies, we ran Trinity (version 2.0.6) with in-silicio normalization, min_kmer_cov = 1, and group_pairs_distance = 800 on the full set of reads from each species. Second, we mapped reads using TopHat to related species and used the junction files as additional splice-site evidence in MAKER. To map reads, we generated bowtie indexes for each genome using default options in bowtie-build, and mapped RNA-seq reads with TopHat version 2.013 (*75*) and the following options: -read-gap-length 3 --read-edit-dist 9 -i 20 --b2-very-sensitive --no-coverage-search. For our libraries we used --library-type fr-secondstrand, and for the public (unstranded) kiwi libraries we used --library-type fr-unstranded. We then post-processed the junctions.bed files produced by TopHat to retain only junctions supported by at least 5 reads, and converted to GFF files for MAKER.

MAKER annotation

To annotate our newly sequenced genomes, we used the program MAKER, version 2.31.8. We ran MAKER initially on all 10 species, using chicken-trained versions of the gene predictors Augustus and SNAP, and using as evidence the RNA-seq Trinity assemblies data described above (as either same or alternate species EST evidence) and complete proteomes from 10 high quality or closely related vertebrate species (human, mouse, zebrafish, zebra finch, chicken, turkey, collared flycatcher, ostrich, white-throated tinamou, green anole). After this initial run, we extracted high-quality models from the emu and the Chilean tinamou genomes using the maker2zff script included with MAKER2, and used those to train SNAP and Augustus (version 3.1).

We then reran MAKER with updated RNA-seq data and the trained versions of Augustus and SNAP. For this second run, we used the emu-trained versions of the gene predictors for other ratite genomes (kiwi, rheas, and cassowary) and the Chilean tinamou-trained versions of the gene predictors for the other tinamou genomes (thicket tinamou, elegant crested tinamou). For RNA-seq evidence, we included two additional data sources: 1) kiwi Trinity assemblies using the published RNA-seq described above and 2) TopHat junction evidence for both same-species and cross-species mappings. For all three kiwi, we used the North Island Brown kiwi Trinity assemblies and TopHat junctions as same-species evidence, and the emu data as cross-species evidence. For emu, we use the emu data as same species evidence, and the kiwi data as cross-species evidence. For cassowary, we use both the kiwi and emu data as cross-species evidence. For rheas, we used the emu and kiwi Trinity assemblies as cross-species EST data, but only the emu junction data. Finally, for tinamous we use the tinamou data as either same-species or cross-species evidence as appropriate. We acknowledge these choices are somewhat arbitrary, although informed by the phylogenetic relationships of the species. Unfortunately, the computational costs of MAKER are too high to explore and optimize these choices.

Final MAKER models were extracted from the GFF files produced and form the basis of our subsequent analysis, using the cleanup_maker.sh script to merge fasta files, merge GFF output, and update default MAKER ids to species-specific keys.

Quality control

Subsequent to our MAKER runs, we discovered that one of our samples (the second emu female liver sample; BioSample SAMN08476475) suffered from a sample labeling problem and is actually a Chilean tinamou sample. Although this was treated as same-species (instead of cross-species) data for our MAKER analysis, the low mapping probabilities of tinamou reads against the emu genome mean that this has little, if any, effect on our gene models.

We used several approaches to assess the quality of the final MAKER annotations. First, we examined MAKER-derived annotation edit distance (AED) scores, which reflect the distance between evidence and produced annotations (lower is better). The distribution of AED scores (Figure S1) reflects that expected from generally high quality annotations.

As an additional check on annotation quality and completeness, we ran BUSCO (3.0.2) in protein mode to assess the completeness of the predicted proteomes for each species. We also compared each protein using hmmscan (HMMER 3.1b; default options) (*76*) against a set of HMMs built from vertebrate eggNOG (*77*) models (version 3.0; hmms built using HMMER 3.1b with default options) to assess the fraction of predicted proteins that can be assigned to an existing homologous group. Finally, we used blastp to search all predicted chicken proteins (NCBI reference build 102) against each predicted protein individually, using an E-value cutoff of 1e-03, to identify the fraction of chicken genes potentially missing from our MAKER annotations. All QC metrics suggest that our annotations are largely complete (Table S3).

Section 2: Key summary

Data inputs:

Raw RNA sequencing reads (fastq files): BioProject PRJNA433110
Assembled genomes
Annotated proteins from NCBI

Methods and code:

Prepare RNA-seq data (TRIMMOMATIC), map to genome (TopHat), assemble transcriptomes (Trinity); make preliminary annotations (MAKER); train gene predictors with preliminary MAKER models (Augustus, SNAP); produce final annotations (MAKER)

Github link: https://github.com/tsackton/ratite-genomics/tree/master/02_annotation

Data outputs:

Annotated genomes (GFF): Dryad

Protein and transcript sequences (fasta): Dryad

Homology inference and alignment of protein-coding genes

Identifying homologous groups

We used the program OMA (version 1.0) (*78*) to infer patterns of homology among protein-coding genes across sequenced birds and reptile genomes, including our new paleognath annotations. We selected 30 existing bird annotations from NCBI (based on quality of genomes) and 3 outgroups to include in our analysis, in addition to our 10 new paleognath gene sets produced as described above, for a total of 43 species. These include the following existing avian species: *Anas platyrhynchos, Aptenodytes forsteri, Aquila chrysaetos canadensis, Calypte anna, Chaetura pelagica, Charadrius vociferus, Columba livia, Corvus brachyrhynchos, Cuculus canorus, Egretta garzetta, Falco peregrinus, Ficedula albicollis, Gallus gallus. Geospiza fortis, Haliaeetus leucocephalus, Meleagris gallopavo, Melopsittacus undulatus, Nipponia nippon, Picoides pubescens, Pseudopodoces humilis, Pygoscelis adeliae, Serinus canaria. Struthio camelus australis, Taeniopygia gutta, Tinamus guttatus. Balearica regulorum, Fulmarus glacialis, Leptosomus discolor,* and *Mesitornis unicolor*, plus the non-avian reptile outgroups *Alligator mississippiensis, Anolis carolinensis*, and *Chrysemys picta*. A full table including NCBI annotation versions, common names, and other information is available at Github (oma_species_list.csv). For each gene set, we selected the longest transcript to represent that protein in our homology search using custom Python scripts (gff_parsing subdirectory from the code link, below).

We then ran OMA using mostly default options (see code on Github for parameter files used). Once OMA had completed, we checked and improved the annotated homology groups using HMMs. We started by building alignments for each homologous group defined by OMA using MAFFT v7.221 (*79*) with the --globalpair and --maxiterate 1000 options. We then built HMMs for each protein alignment using HMMER 3.1b hmmbuild with default options, and searched each hmm against the full set of input proteins to OMA, in order to 1) verify that proteins assigned to an homologous group are recovered by searching with the HMM built from that group, and 2) assign unassigned proteins as best as possible. Subsequent to the HMM search, we improved the OMA output with a custom Python script which uses a graph-based algorithm to add gene models not assigned to any OMA group to best hit groups conditional on a high quality match. After this step, we were left with a final set of 45,367 homologous groups (HOGs), which we use in the following analyses.

Aligning homologous groups

The 45,367 total HOGs were filtered to retain 16,151 groups with sequence from at least four taxa. Protein sequences for these HOGs were aligned with default options in MAFFT v.

7.245, following which three rounds of alignment filtering were used to remove poorly aligning sequences. In the first round, custom Perl scripts were used to remove entire alignment columns if: more than 30% of sequences had a gap at that position, or fewer than ten sequences total had a non-gap character, or there was not at least one sequence with a non-gap character from at least two of the three major taxonomic groups (paleognaths, neognaths, and non-avian outgroups). In the second round of filtering, poorly aligning regions of individual sequences were masked in the output alignments from round 1 using the sliding-window amino acid similarity approach with default parameters employed by the Avian Phylogenomics Project (*25*) (scripts accessed from ftp://climb.genomics.cn/pub/10.5524.101001_10200/10/041/ on Sept. 30, 2015). A third round of sequence filtering then removed columns with low sequence representation using the same criteria as outlined for round 1.

Following filtering, we flagged sequences to retain within each alignment by requiring that all of the following apply: the original (unfiltered) sequence length was at least 50% of the average input sequence length of all unfiltered sequences for that locus, the filtered length of an individual sequence was at least 50% of its original pre-filtering length, and there was less than 1 gap per bp aligned sequence in the individual filtered sequence. Failure to meet any of these criteria resulted in the entire sequence being excluded from the HOG. After removal of individual sequences, entire HOGs were retained for downstream analyses only if they contained a maximum of three sequences for any given taxon and the total number of sequences did not exceed 1.5X of the total number of alignment taxa. We also required the presence of at least 50% of all avian species in the alignment and at least 50% of all paleognath species. These criteria resulted in 11,274 HOGs retained for further analysis, of which four were excluded due to failures to complete alignments or analysis runs in a reasonable time. The retained 11,270 HOGs map to 11,930 chicken NCBI gene IDs (annotation release 102), which is 69.5% of all chicken gene models, or 84.1% of all chicken gene models that are not members of moderate or larger (>2 paralogs in chicken) multigene families.

Nucleotide alignments were compiled for avian CDS sequences retained in each HOG following filtering as described above (and omitting non-avian outgroups). For publicly available genomes, the GenBank transcript associated with each protein was accessed and sequence ends were padded with Ns if necessary so that all transcripts began in phase 0 and contained a multiple of 3 bases. For paleognath draft genomes, the gffread utility from Cufflinks v. 2.2.1 (*80*) was used to output CDS sequences from MAKER genome annotations. CDS sequences for each HOG were aligned with the codon model in Prank v. 150803 (*81*). Poorly aligning regions of individual sequences were masked using the filter_alignment_fasta_v1.3B.pl script released by the Avian Phylogenomics Project accessed from ftp://climb.genomics.cn/pub/10.5524.101001_10200/10/041/ on Sept. 30, 2015), following which alignment columns for codons containing only gaps/Ns were removed with custom Perl scripts that maintained the alignment reading frame.

Guide trees for each HOG were built from the filtered CDS alignments with RAxML v.8.1.4 (*82*) using 200 rapid bootstrap replicates followed by a thorough maximum likelihood tree search. All loci were run with a GTR+GAMMA substitution model, and with 3rd codon positions specified as a separate alignment partition to codon positions 1+2.

In addition to the data set described above, which included taxa from the whole-genome alignment, we also compiled an extended data set adding sequence from the little bush moa (*24*) and from four cormorant species (*34*): *Nannopterum auritus* [Double-Crested Cormorant], *Nannopterum brasilianus* [Neotropic Cormorant], *Nannopterum harrisi* [Galápagos Flightless Cormorant], and *Urile pelagicus* [Pelagic Cormorant].

Assignment of cormorant protein-coding genes to the 11,270 HOGs comprising the final data set was determined by first aligning each HOG nucleotide sequence in translated protein space using PRANK, then producing a new profile Hidden Markov Model for each of the 11,270 final HOG alignments using HMMER. Each cormorant protein was then searched against all HOG HMMs using the hmmsearch program in HMMER. To match proteins to HOGs, we kept the best hit only for each cormorant protein among all HOG HMM matches.

For moa, we used whole-scaffold alignments of the mapDamage corrected moa assembly with the emu reference genome to map emu coordinates to their corresponding moa sequence. Coordinates for the 120,043 emu exons in 10,945 transcripts included in the final HOG CDS alignments were individually mapped to moa using a custom Perl script, and each moa-emu exon nucleotide alignment was extracted from the full scaffold alignment.

A second Perl script was used to construct moa CDS sequences while attempting to correct errors introduced through the gene-unaware mapping used to generate the initial reference-based moa genome assembly (e.g. moa sequence could contain apparent frameshifts resulting from indel placement and insertion of ‘N’s for unmapped regions). For exon mappings with no moa bases (‘N’s only), a string of ‘N’s equivalent to the emu reference exon length was used. For exon mappings with ‘called’ moa bases, moa sequence was retained if it differed from the emu reference length by a multiple of 3, contained no internal stop codons, and had ≥ 50% amino acid identity to the emu reference sequence; otherwise, the moa sequence was replaced by a string of ‘N’s equivalent to the emu reference length. When moa exon mappings differed by a non-triplet length from the emu reference, the moa/emu exon nucleotide sequences were realigned with MAFFT v. 7.245 using the ‘ginsi’ option. Alignment columns that contained a gap character in emu and ‘N’ in moa were removed, whereas ‘N’s were inserted in moa for alignment positions that were a gap in moa and nongap character in emu. Following this refinement, moa sequence was retained for exons that differed from emu by a multiple of 3, contained no internal stop codons, and had ≥ 50% amino acid identity with emu; otherwise, moa sequence was replaced by a string of ‘N’s equivalent to the emu reference length. Moa sequence for all processed exon mappings were concatenated into transcripts and moa transcripts with no internal stop codons that spanned ≥ 30% of the emu reference length or were at least 100 amino acids long were retained.

Amino acid sequences for moa and cormorants were appended to FASTAs containing unaligned sequences for all 11,270 HOGs in the final data set, aligned with MAFFT, and filtered to identify sequences to retain as described for the original data set above (retaining all sequences from the original data set, but excluding poorly aligning moa and cormorant sequences as well as instances where one of these new species had > 3 retained sequences in a HOG). Nucleotide sequences for retained moa and cormorant loci were added to the unaligned CDS fastas from the original data set, aligned with PRANK, filtered to remove poorly aligning regions, and guide trees were built with RAxML as described for the original data set above.

Section 3: Key summary

Data inputs:

Newly annotated proteins (section 2)
Publicly available protein sequences (NCBI)

Methods and code:

Infer homologous groups (OMA); update using Hidden Markov Models (HMMER); align and filter homologous groups (MAFFT); make gene trees (RAXML)

Github link: https://github.com/tsackton/ratite-genomics/tree/master/03_homology

Data outputs:

Homology matrix and associated information: Dryad, Github
Protein and transcript alignments (fasta): Dryad
Gene trees for each homologous group: Dryad

Whole genome alignment and CNEE identification

Producing the whole genome alignment

To produce a whole genome alignment of birds and non-avian reptile outgroups, we used the progressiveCactus software (*83*), which uses a progressive algorithm and a guide tree to produce a whole genome alignment and genome history reconstruction, including inferred ancestral sequences. We selected a total of 42 sequenced genomes to align: 35 bird species (our 10 new paleognath genomes, 2 existing paleognath genomes, and 23 neognaths), and 7 non-avian reptilian outgroups. A full list of species, genome assembly versions, and references are available at Github (alignment_species_list.tsv). In general we aimed for reasonably complete phylogenetic coverage while avoiding genomes that are of sufficiently low quality to be excessively gappy.

To gain the benefits of the progressive alignment algorithm in progressiveCactus, we generated a partially resolved guide tree of the 42 species in our alignment, based on previously published analyses (*20*, *25*, *84*). To estimate neutral branch length for this partially resolved tree, we combined estimates from a published tree based on 4-fold degenerate sites in some birds and non-avian reptiles (*84*) and a preliminary UCE tree generated from our genome data, based on extracting UCEs from unannotated genome assemblies by BLASTN and then tree building as described above.

Using this guide tree, we ran a modified version of progressiveCactus that replaces the default meta-scheduler (jobTree) with a version that can use the SLURM workload manager operating on the Harvard Odyssey cluster (https://github.com/harvardinformatics/jobTree). We ran progressiveCactus largely with default options, but given uncertainties in our branch length estimates we used slightly larger than default branch length cutoffs for each level of lastz optimization. The configuration file we used is available on Github.

To aid in visualization and analysis of our whole genome alignment, we also produced a UCSC Genome Browser track hub containing our progressiveCactus alignment, as well as gene and conserved element annotations (see below). We used the hal2assemblyHub.py script distributed as part of halTools/progressiveCactus to generate the input files for our assembly hub.

Annotating conserved non-exonic elements

To identify conserved non-exonic elements (CNEEs), we started from the whole genome alignment described above, and used the phast package to identify conserved elements (*38*, *85*). First, we extracted fourfold degenerate sites from our whole genome alignment and use them to estimate a neutral model with phyloFit for three trees which differ in the placement of rheas (tree

1: rheas are sister to tinamous; tree 2: Mitchell et al tree; tree 3: rheas are sister to emu/cassowary/kiwi clade). We then corrected the estimated neutral models for base composition statistics using the phast program modFreqs, and named all ancestral nodes with tree_doctor. Next, we estimated rho (expected substitution rate of conserved elements relative to neutrality) using phastCons with the –estimate_rho option, run separately on non-overlapping 1 MB chunks of the input alignment. Conserved models for each chunk were combined with phyloBoot and then used for initial conserved element prediction in phastCons. We investigated several possible parameter values for both the –target_coverage and –expected_length options in phastCons, but determined that these made little difference to predicted conserved elements. We then ran our final phastCons run on the whole genome to estimate conserved elements, using the Mitchell et al. tree (tree2), although >99% of elements are identical when run using alternate topologies (tree1 or tree3). After estimating conserved elements, we first merged elements within 5 bp of each other into single conserved elements, and then extracted a final set of 284,001 CNEEs representing non-exonic conserved elements at least 50 bp long.

Section 4: Key summary

Data inputs:

Genome sequence (NCBI)

Methods and code:

Align whole genomes (progressiveCactus); annotate conserved non-coding regions (PHAST)

Github link: https://github.com/tsackton/ratite-genomics/tree/master/04_wga

Data outputs:

Whole genome alignment (hal): UCSC Track Hub
Conserved element predictions (bed): Dryad, Github
Conserved element annotations (bed, text): Dryad, Github

Phylogenomic inference of paleognath relationships

Data set compilation

Conserved non-exonic elements (CNEEs)

A candidate set of 811,696 intergenic CNEEs was constructed by filtering 1,949,832 conserved elements (CEs) called from the whole-genome alignment (see below) to remove elements that overlapped annotated exons, genes, or CDS features in the galGal4 chicken genome release (NCBI annotation version 102), using BEDTools v. 2.26.0 (*86*). To avoid biasing phylogenetic reconstruction, we considered only those CNEEs that were consistently identified irrespective of the placement of rheas in alternative guide trees used when calling conserved elements. The candidate set was further filtered to retain 16,852 CNEEs of minimum 250 bp in length in chicken, and HAL Tools v.2.1 (*87*) was used to lift over chicken reference coordinates for candidate CNEEs to each paleognath species included in the whole-genome alignment (N= 12 species). To avoid including paralogous regions, output from halLiftover was parsed to retain CNEEs where the chicken reference corresponded to a unique region in each target species, and CNEEs with no missing taxa and at least as many variable sites as there were sequences, following alignment as described below, were retained (N= 14,528 loci). BEDTools was used to remove CNEEs with any overlap to the set of ultraconserved elements (UCEs) used for phylogenetic inference, leaving a data set of 12,676 CNEEs (note that overlap with the intronic data set is already addressed by considering only intergenic CNEEs).

Sequence for the North Island brown kiwi (*Apteryx mantelli*), which was not included in the whole-genome alignment, was identified with blastn searches using sequence from each of the three kiwi species present in the whole-genome alignment as queries. Blastn searches were run with NCBI's 'somewhat similar' parameters (evalue 1e-10, perc_identity 10, penalty -3, reward 2, gapopen 5, gapextend 2, word size 11), followed by strict criteria for sequence inclusion that required consistent best hits across query species, with a single HSP covering at least 50% of the query sequence and minimum 80% sequence identity.

Sequence for the extinct little bush moa (*Anomalopteryx didiformis*) was added from a reference-based genome assembly built from mapping moa reads to the draft emu genome included in the whole-genome alignment. Emu coordinates determined from the liftover approach outlined above were used to retrieve the corresponding mapped region in moa, and moa sequence was included if it covered at least 30% of the emu reference or was at least 200 bp in length, excluding Ns.

An additional data set was compiled to test the robustness of inferred paleognath relationships using an alternative outgroup and more comprehensive taxon sampling across birds. The strategy outlined above was used to lift over chicken coordinates for the 12,676 CNEE loci used for paleognath phylogenetic inference to all neognath birds included in the whole-genome alignment (N= 23) and to the alternative outgroup American alligator (*Alligator mississippiensis*). Loci with no missing taxa and a minimum unaligned sequence length of 250 bp per species were retained (N= 6,931 loci, with 38 taxa total).

Introns

Bedtools was used to output coordinates for introns that did not overlap any annotated exon feature in the galGal4 genome annotation. Chicken coordinates for both these 'nonoverlapping' introns, as well as for all coding exons, were lifted over to each paleognath species in the whole-genome alignment as described for CNEEs above. Liftover output was parsed to retain unique regions in the target species (e.g. omitting reference chicken coordinates that correspond to multiple regions in the target genome), and expected exon/intron boundaries in the target sequence were refined by padding exon coordinates to be flush with the reference chicken sequence. Intron liftovers falling within the expected target region between adjacent exon liftovers were omitted if greater than 100 kb in length, or if greater than 10 kb and more than 50% longer than the chicken reference. Candidate introns were required to have at least 100 bp of sequence for all paleognath species in the whole-genome alignment, and introns from one CDS per alternatively spliced transcript in chicken were chosen based on the longest combined intron length in ostrich to generate a set of 33,066 introns. Sequence for the North Island brown kiwi and little bush moa were added to each of these loci as described for CNEEs above. Candidate loci were aligned with MAFFT v. 7.245 using default parameters, and one intron per gene was chosen, requiring a minimum average pairwise sequence identity of 70% and less than 0.5 gaps per bp of aligned sequence, and then choosing based on the fewest number of missing taxa and longest average unaligned input sequence length across taxa. The final data set consists of 5,016 introns, each of which originates from a different gene.

Ultraconserved elements (UCEs)

We used the data set of 3,679 UCEs compiled by the Avian Phylogenomics Project (accessed from http://dx.doi.org/10.5524/101041). To generate the most complete data matrix possible from UCE reference sequences for individual taxa, which often do not span the entire locus, we used both the ostrich (*Struthio camelus*) and white-throated tinamou (*Tinamus guttatus*) included in the Avian Phylogenomics Project data in addition to chicken as reference taxa to lift over to each other paleognath species included in the whole-genome alignment. The resulting liftover data were filtered to retain non-duplicated regions in the target genome, and liftovers that were consistent across reference species were tiled to obtain the longest total target sequence. We allowed a maximum of one missing paleognath species per locus from taxa included in the whole-genome alignment to produce a final data set of 3,158 UCE loci. Blastn searches were used to add sequence from the North Island brown kiwi, and coordinate mapping from the emu reference to add sequence for little bush moa, as described for CNEEs above.

Sequence alignment and alignment trimming/filtering

Sequences for individual loci were aligned with MAFFT v. 7.245 using default options for global iterative alignment for CNEEs (option 'ginsi') and local pairwise alignment for introns and UCEs (option 'linsi'). Two additional data sets were generated from these MAFFT alignments, except in the case of CNEEs with additional sampling across neognaths. trimAl v. 1.2rev59 was used for column-based filtering with the 'automated1' option to heuristically choose trimming parameters based on input alignment characteristics for each locus (*88*). Additionally, alignments with full matrix occupancy (no 'missing data') were generated with custom Perl scripts that filtered the original MAFFT alignments to retain loci with no missing taxa and to exclude alignment columns containing gaps, undetermined bases (Ns), or ambiguity characters and to omit loci with post-filtering total alignment length below 200 bp.

Gene tree inference

RAxML v. 8.1.5 was used to infer the highest scoring maximum likelihood tree from unpartitioned alignments for each locus with a GTR+GAMMA substitution model and 20 independent tree searches beginning from random starting tree topologies. RAxML was also used to infer topologies for 500 bootstrap replicates for each locus, again using unpartitioned alignments and a GTR+GAMMA substitution model.

Species tree inference

Both coalescent-based MP-EST inference from gene tree topologies and maximum likelihood inference from concatenated alignments with ExaML were used to infer topologies for each data set (e.g. original, trimAl trimmed, and filtered MAFFT alignments). Analyses were run for each marker type (e.g. CNEEs, introns, UCEs) as well as the total evidence nucleotide tree (TENT) that combined loci across the three marker types. Due to computational considerations, the CNEE data set with expanded taxon sampling was analyzed with MP-EST only.

For each data set, MP-EST v.1.5 (*30*) was used to infer the species tree topology from the maximum likelihood RAxML gene trees, and node support was estimated using MP-EST with RAxML gene tree bootstrap replicates. For both species tree and bootstrap topology searches, MP-EST was run in triplicate beginning from a different random number seed each time, and with ten independent tree searches within each of the triplicate runs. Additionally, replicate MP-

EST bootstrap data sets were independently generated by randomly assigning RAxML bootstraps from each gene.

ExaML v. 3.0.16 (*89*) was run with a GTR+GAMMA substitution model on fully partitioned, concatenated alignments for each data set (e.g. each locus as a separate partition). The topology was inferred from 21 full maximum likelihood tree searches, 20 of which began with complete random starting trees, and one additional search beginning with the random stepwise addition order parsimony tree (note that with the relatively few taxa present in our data sets, identical parsimony starting trees were always produced from different random number seeds). Additionally, at least 50 ExaML bootstrap replicates were run for each data set, in each case using a GTR+GAMMA model and fully partitioned alignment. Convergence of bootstrap replicates was assessed according to the bootstopping analysis with majority-rule consensus tree criterion in RAxML (option -I autoMRE), and additional bootstrap replicates were added as necessary until convergence was reached, or to a maximum of 250 bootstrap replicates.

For both MP-EST and ExaML, bootstrap support values were plotted on the inferred species tree topology with RAxML, and trees were outgroup-rooted with chicken, or with American alligator for the CNEE data set with increased taxon sampling, using ETE v. 3 (*90*). Branch lengths for MP-EST species trees, which are output in coalescent units, were re-estimated in units of substitutions per site by constraining fully partitioned, concatenated alignments to the MP-EST species tree topology with ExaML using option -f E.

Tests of ratite monophyly versus paraphyly

Likelihood-ratio statistics were used to test support for the paraphyletic ratite clade recovered in all analyses against an alternative hypothesis of ratite monophyly where the tinamous are placed as the sister group to the ratites. For both MP-EST and ExaML, the test statistic was calculated by taking the difference between the log likelihood score from the unconstrained tree search of the total evidence nucleotide data set (N= 20,850 loci, using RAxML maximum likelihood gene trees as input to MP-EST and the fully partitioned alignment for ExaML) against the log likelihood when the tree search was constrained to a user defined tree that pruned the tinamou clade from each respective species tree and regrafted it as the sister to the ratites. For MP-EST, this test statistic lnL(species tree) – lnL(ratite monophyly constraint) was compared against a null distribution where likelihood ratios were computed from each of the 500 MP-EST bootstrap replicates for the TENT data set. Unlike MP-EST, not all ExaML bootstrap replicates recovered identical topologies for the placement of rheas. Therefore, ratite monophyly constraint trees were allowed to differ across bootstrap replicates for ExaML and were constructed by pruning the tinamou clade from each unconstrained bootstrap search and placing it as the sister to the ratites (therefore allowing the interrelationships among ratites to vary across ExaML bootstrap replicates). Additionally, due to computational constraints, it was only possible to run a limited number of bootstrap replicates for ExaML (N= 83).

Investigation of gene tree heterogeneity and the anomaly zone

Both best maximum likelihood gene trees inferred by RAxML and majority-rule extended consensus gene trees generated from RAxML bootstrap replicates were used in tests of gene tree heterogeneity. Since these tests require identical taxon sampling across loci, North Island brown kiwi, which was missing from many loci by virtue of the fact that it was added by blastn searches after data set compilation for species included in the whole-genome alignment, was pruned from each gene tree and loci with any additional missing taxa were omitted from the data set (retaining

20,491 of 20,850 loci in the total evidence data set, or 98.3% of loci). HashRF v. 6.0.0 was used to calculate pairwise Robinson-Foulds distances between all rooted gene trees, and a custom Perl script was used to parse these values into sets of loci with identical gene tree topology. For gene tree topologies occurring at higher frequency than the inferred MP-EST species tree topology, we used ETE to parse bootstrap support for clades that conflicted with the species tree and PAUP v. 4.0a151 (*91*) to calculate the number of substitutions occurring on conflicting branches under a parsimony criterion. Support for recovered gene tree topologies was further assessed by comparing the gene tree likelihood for each locus to that obtained when the topology was constrained to match the species tree with RAxML. Equation 4 from Degnan and Rosenberg (*32*) was used to calculate the function $a(x)$ for each internal branch x in the MP-EST TENT species tree, where $a(x)$ represents the boundary of the anomaly zone. These values were compared to MP-EST coalescent branch lengths for each descendant branch y in the MP-EST TENT tree, with $y < a(x)$ consistent with branch lengths expected to produce anomalous gene trees.

Comparing gain of flight and loss of flight scenarios

Numerous previous authors (*18*, *20*) have made qualitative arguments favoring multiple independent losses of flight in ratites (as opposed to gain of flight in tinamous independently of neognaths) based on the presumed difficulty in re-evolving a complex morphological phenotype. However, a strict parsimony reconstruction with equal rates of flight loss and flight gain would support a most parsimonious reconstruction in which flight is gained once in neognaths and once in tinamous independently. To explore this, we took two approaches. First, we conducted a maximum parsimony ancestral reconstruction in Mesquite v. 3.51 (*92*) in which we varied the cost of gain and loss of flight. Second, we used a maximum likelihood model (implemented in the corHMM R package v. 1.2.2 (*93*)) to estimate the probability that the key node ancestral to tinamous and moa is volant (implying independent losses of flight across ratites) under a variety of rates of gain and loss. We used the "rayDisc" function under the asymmetric rates regime (model="ARD"), starting with an ultrametric version of the MP-EST tree in Fig. 1. This tree was ultrametricized using the "chronopl" function in the ape R package (*94*) with a lambda =0.1. An ultrametricized tree was used because this provides a more realistic setting in which to examine character evolution than a tree with unequal branch lengths leading to the present.

Section 5: Key summary

Data inputs:

Combined datasets described in the first part of this section, consisting of UCEs, CNEEs, and introns.

Methods and code:

Align input data (MAFFT); build trees (ExaML, RAXML, MP-EST)
Github link: https://github.com/tsackton/ratite-genomics/tree/master/05_phylogenomics

Data outputs:

Aligned markers: Dryad
Phylogenetic trees: Dryad
Estimated species tree: Github

Analysis of protein-coding genes

Evolutionary models

To generate data for both our analysis of rates of protein evolution along each branch, and to estimate convergent and divergent amino acid substitutions, we ran a series of PAML (*95*) and HyPhy (*96–98*) models on each protein-coding alignment (described above), using PAML v 4.8 and HyPhy v2.2.1. For each alignment, we ran all models with both the species tree topology and the specific gene tree topology where possible; models with at least one gene duplication were run with only the gene tree topology, although we exclude these from most results. In total, we ran three PAML models and two HyPhy models. With PAML, we ran model M0, which simply fits a single ω rate to each alignment, and also extracted the maximum likelihood ancestral state reconstruction from this model run, which we use to compute convergent and divergent amino acid substitutions. We also ran the free-ratio branch model, which allows a separate ω value for each branch. Finally, we ran aaml to estimate amino acid branch lengths on fixed topologies, which we use as the input to the branch tests, below. With HyPhy, we ran the aBS-REL model (*98*), which detects lineages experiencing positive selection in a subset of codons, and the RELAX model, which detects relaxed selection on a pre-specified clade (*99*). We ran RELAX with two sets of target lineages, first with ratites and second with vocal learners. Both runs we then parse with custom code and analysis in R. Prior to running these models, we filtered alignments to remove codons which contain gaps in greater than 80% of species, with the exception of a few instances where this removed too much sequence and led to program crashes, in which case we lowered the gap threshold to 50%.

All models were run on the ‘default’ protein dataset described above, which includes 43 species of birds and outgroups, but does not include the moa or the cormorants. We additionally ran the BS-REL tests for lineage-specific positive selection, and the aaml models to estimate amino acid divergence for RERconverge (*100–102*) on the extended dataset including both moa and the cormorant species from (*34*).

Analysis of amino acid convergence

Using the inferred ancestral states at each position in each alignment from the PAML M0 models, we define convergent amino acid substitutions between pairs of lineages as those that occur at the same position and have the same amino acid in the descendant sequences. This includes both 'parallel' changes (e.g., same starting amino acid and same ending amino acid) and 'convergent' changes (e.g., different starting amino acid, same ending amino acid). All other amino acid substitutions that occur at the same position on a pair of branches are defined as divergent. We compute numbers of convergent and divergent substitutions for all branch pairs for each alignment using a custom Python script.

To test whether ratites have more convergent substitutions than expected, we classified each pair of branches as either ratite-ratite or other, and then summed all substitutions (convergent or divergent) across all alignments. Some branch pairs do not appear in all alignments, because each alignment is allowed to have a small number of missing species. After summing substitution types, we fit a linear model that included main effects of divergent substitution number, evolutionary distance (calculated as the distance between nodes on the neutral tree calculated from four-fold degenerate sites), their interaction, and a term for if the branch pair was a ratite-ratite pair using R, to test for a significant effect of ratite-ratite branch pairs on the number of convergent substitutions. A significant effect of the ratite-ratite term would indicate that convergent substitutions are more likely between ratite-ratite pairs than other taxa, correcting for expected number of convergent substitutions based on evolutionary distance and divergent substitution number. For this analysis, we only use the 'default' (excluding moa and cormorant) protein alignments.

Analysis of branch rate shifts and RELAX results

We used the RERconverge package (*100*) to test for convergent rate shifts along target lineages. We first extracted the amino acid tree for each alignment from the PAML aaml runs, and then normalized each branch rate using RERconverge functions with the transform = "sqrt", weighted = TRUE, and scale = TRUE options. In order to correct for biases in the RERconverge results due to an unbalanced tree topology (the large rhea/emu/kiwi/cassowary clade), we repeated the RERconverge analysis after downsampling target lineages to retain only three ratites (moa, ostrich, and one of the rhea/kiwi/emu/cassowary clade). This follows the approach described below to estimate convergence in non-coding regions. We ran RERconverge on both the 'default' (including moa) and 'reduced' (excluding moa) datasets, but did not run the downsampling analysis on the reduced dataset as we would be left with only two independent lineages to test in this case, leading to low power.

To verify these results, we also estimated the K parameter (a measure of intensity of selection) and the P-value from HyPhy's RELAX model, with ratites as the target lineage. We extracted K and P-values, normalized K values by subtracting from 1, and then rank-transformed both normalized K and normalized amino acid branch rate for comparison. Due to computational time to complete this analysis, we only analyzed the 'reduced' dataset for this comparison.

Analysis of BS-REL results

We used aBS-REL to identify, for each alignment, the number of ratite lineages with evidence for a class of codons with $\omega > 1$ and the number of non-ratite lineages with evidence for a class of codons with $\omega > 1$. We refer to these as "target selected" and "non-target selected", respectively. From these data, we can compute three quantities: the number of alignments specifically selected in ratites (target selected > 0, non-target selected == 0), the number of alignments convergently selected in ratites (target selected > 1), and the intersection (convergently selected). We then extract galGal4 gene symbols for each HOG in these three groups, and test for functional enrichment using the clusterProfiler package (*103*) and the Bioconductor galGal annotation package. We also tested for functional enrichments among genes with a bias towards selection in ratites, defined as those where > 50% of selected lineages are ratites.

To test for excess convergence in ratite lineages detected with BS-REL, we first computed the proportion of lineages selected for each alignment, and treated that as the probability that any single lineage would be the target of selection. We then simulate *n* random draws from a binomial (where *n* is the total number of target lineages for that alignment) 10000 times, using the probability of selection as the probability of success, and count the number of times we observe 2 or more target lineages under selection (and no non-target lineages under selection). We then normalize this as a proportion of the total number of lineage-specific selection we observe, and calculate an empirical P-value. This permutation procedure thus tests the null hypothesis that evidence for positive selection is independent across lineages – observing selection in one ratite lineage does not increase the probability that we will observe selection in another ratite lineage. Rejecting this null is thus consistent with the hypothesis that genes selected in one ratite lineage are likely to be experience positive selection in multiple lineages, potentially due to shared selection pressure driven by convergent phenotypes.

Pseudogenization, loss, or functional degradation of proteins in ratites

To screen for proteins that may have specifically lost function in multiple ratite lineages, we used two approaches. First, we took all annotated proteins in chicken and ran blastp to screen for presence in all ratites and other bird genomes. We consider a protein present if the protein has a hit with at least 50% identity and where the target is at least 70% of the length of the query. As an alternative approach, we used deltaBS (*33*), a profile Hidden Markov model method, as follows. All protein coding genes and isoforms were compared to vertebrate protein family hidden Markov models from the eggNOG database (*77*), using hmmsearch (HMMER3.0 package) to call orthologs. hmmsearch results were filtered to exclude model hits with E-values <0.0001. Next, for each species, the top protein hit to each model was taken, and top hits to each model were treated as orthologous groups. Orthologous groups with representatives from fewer than 30 species were excluded from analysis, restricting the analysis to proteins found in most birds in the study. Of the remaining orthologous groups, those showing the highest 5% of variance in scores were excluded, to reduce the likelihood that mis-called orthologs were being included in the study. Once ortholog relationships had been established, scores for each protein in each orthologous group were compared using a phylogenetic generalized least squares approach in R (pgls, caper package (*104*)) to identify any differences in score that could not be explained by phylogeny. P-values and coefficients from each analysis were recorded, and p-values were corrected using the Benjamini-Hochberg approach. To verify consistency with alternate datasets, we repeated this analysis on the 'default' (including moa) and the 'extended' (including cormorant) datasets. Because of low quality gene models and sensitivity of this method to premature stop codons in poorly predicting proteins, we filtered to remove all protein alignments with either an X (undetermined amino acid) or a premature stop for both the moa and cormorant datasets, leaving 2,812 genes for the moa analysis and 5,341 genes for the cormorant analysis. In both cases our results are consistent across input datasets, however.

Section 6: Key summary

Data inputs:

Aligned protein coding genes (section 3)

Methods and code:

Estimate evolutionary parameters (PAML, HyPhy); statistical analysis (R); test for gene loss or function-altering substitutions (blastp; deltaDBS)

Github link:
https://github.com/tsackton/ratite-genomics/tree/master/06_protein_coding_analysis

Data outputs:

Aligned protein coding genes as trimmed for PAML/HyPhy: Dryad
Raw and parsed PAML and HyPhy outputs: Dryad, Github

Analysis of conserved non-exonic elements

Identifying changes in conservation in CNEEs

Data preparation

As described for protein-coding genes in Section 3, we analyzed evolutionary patterns of CNEEs for both a data set limited to taxa included in the whole-genome alignment, as well as an expanded data set that included an additional four cormorant species and the little bush moa. We first extracted sequence for each CNEE from all species included in the whole-genome alignment using HAL Tools v.2.1 liftover to generate psl files mapping chicken reference coordinates to each target species. Custom Perl scripts were used to produce FASTA format files from halLiftover output, with sequence for individual CNEEs omitted when the target region was duplicated in a given taxon or was >2X the reference chicken length. Individual CNEEs were aligned using MAFFT v. 7.245 with the 'ginsi' option, following which all CNEEs were concatenated into a single FASTA with an accompanying partition file defining the start and end positions of individual loci.

To produce an 'extended' alignment including the flightless cormorant and related species, we produced a secondary whole genome alignment of chicken (galGal4) and four cormorant species (*Nannopterum auritus* [Double-Crested Cormorant], *Nannopterum brasilianus* [Neotropic Cormorant], *Nannopterum harrisi* [Galápagos Flightless Cormorant], and *Urile pelagicus* [Pelagic Cormorant]), using progressiveCactus as described above. Using this secondary alignment, we lifted CNEE coordinates over to cormorants (from chicken) and extracted FASTA files for each element as described above.

For the little bush moa, we mapped emu reference coordinates for individual CNEEs to their counterparts in the mapDamage corrected moa genome assembly using moa-emu whole-scaffold alignments that were generated during the reference-based approach to moa genome assembly. We omitted moa sequence for an individual CNEE when its length was >2X the reference emu length, or when >30% of the moa-emu alignment for this region consisted of moa insertions relative to the emu. Sequences for cormorants and moa were added to unaligned FASTAs of individual CNEE loci for all other taxa, and *de novo* alignment and subsequent concatenation of all loci was performed as described above. These CNEEs were realigned as described above, except that reference gaps were not removed. Finally, we produced a 'reduced' CNEE alignment set, in which the moa sequences were removed from the original set and then each CNEE was realigned as described. Overall, then, we report results from three different alignment sets: 'default', containing moa but not flightless cormorant and related sister taxa; 'reduced', excluding moa; and 'extended', including both moa and flightless cormorant.

PhyloAcc analysis

Our main analysis of rate variation in CNEEs employed a new Bayesian method, PhyloAcc, to identify CNEEs that experience changes in conservation state across the avian phylogeny (*44*). In brief, PhyloAcc uses a latent conservation state model, in which each branch can belong to one of four states which define the relative substitution rate for that CNEE on that branch (compared to neutrality, defined by the neutral model produced with phyloFit, described above). The states (Z) are: -1 (missing), 0 (neutral), 1 (conserved), or 2 (accelerated, which is parameterized as 'different from conserved' and so does not have to be strictly greater than the neutral rate). While the neutral state is defined as having a relative rate of 1, the conserved and accelerated rates are estimated from the data for each CNEE. We ran PhyloAcc on all three alignment sets described above (reduced, default, extended), where in the extended dataset including cormorants we added flightless cormorant to the ‘target’ flightless lineages, and in the default dataset we include the damage-corrected moa sequence, but no cormorants. For each alignment set, we use all four combinations of two parameters: one that controls the transition probability from neutral to conserved (and thus how deeply on the tree CNEEs are likely to arise), and one that controls whether gaps are treated as missing data or evidence for acceleration. Configuration files for PhyloAcc for each version are available from Dryad. In the main text, we present the version where gaps are treated as missing data, and gains are more likely, as our main result but our conclusions are qualitatively identical under all scenarios and consistent across different datasets as well (Fig. S12-13).

We used PhyloAcc to compute the maximum a posteriori (MAP) Z matrix (matrix of latent states), two Bayes factors to test for ratite acceleration and ratite specificity, and the posterior probability of each Z state for each branch in each alignment, as described previously (*44*). We define Bayes Factor 1 (BF1) as the Bayes factor comparing a null model (no acceleration allowed on any branch) to the ratite model (acceleration allowed only on ratite branches). We define Bayes Factor 2 (BF2) as the Bayes factor comparing the ratite model to the full model (acceleration allowed on any branch). BF1 identifies elements accelerated in ratites irrespective of the pattern in the rest of the clade, whereas BF2 identifies elements with acceleration specific to ratites. We define ratite-accelerated elements as those with BF1 >= 10, and BF2 >= 1, and additionally require that the estimated posterior number of losses in internal paleognath branches + tinamou branches be less than 1. To identify convergent acceleration, we define the number of losses from the model, based on the posterior probability matrix, and consider an element as convergently accelerated if the posterior expected number of independent losses is >= 1.8.

To test for an excess of convergence in ratites in a consistent way, we focus on three-way convergences. We randomly sample trios of ratites from different clades (focusing on the parsimonious case with flight lost once in ostrich, once in moa, and once in emu/cassowary/kiwi/rhea), and compute the number of times all three are in non-conserved states (and all other non-ratite lineages are in the conserved state, using a posterior probability cutoff of >= 0.90 to define acceleration). We then repeat this procedure for 10000 random subsets of non-ratite lineages (excluding sister taxa, and removing duplicate tip sets, so the actual number of permutations is much lower) to estimate the rate of convergence in random lineages not expected to show evidence for phenotypic convergence. While this procedure does not explicitly account for power differences to detect acceleration among lineages, the observed number of accelerated elements we see in ratites is substantially greater than would be predicted by chance based on rates of acceleration in individual lineages, and the proportion of accelerated elements among ratites is also much greater than the proportion among random neognaths trios. Together these observations suggest that power differences do not play a role in our observations.

PhyloAcc is available from https://github.com/xyz111131/PhyloAcc and is described in more detail in a companion manuscript (*44*).

GO enrichment

To test for GO terms and other functional elements enriched in association with ratite-accelerated or convergently accelerated CNEEs, we used a permutation approach to account for the fact that CNEEs themselves are biased towards particular gene functions. For these permutations, we focused on the default (moa included) dataset, and the ‘gain’ PhyloAcc parameter set. First, we define ratite-accelerated and convergently accelerated CNEEs as above, but additionally require that no neognath species has strong evidence for acceleration (posterior probability of acceleration > 0.90). We then, for each analysis set, randomly sample N CNEEs (where N = number of accelerated or convergently accelerated CNEEs) 5000 times (although due to job failures the actual number of permutations is somewhat lower), and calculate P-values and enrichment scores for each GO term (separately for molecular function and biological process ontologies). We then use these permutation empirical distributions to compute P-values for each real set, resulting in tests for enrichment corrected for the background distribution of functional terms associated with CNEEs.

Gene enrichment

To identify genes with an excess of ratite-accelerated or convergently accelerated CNEEs nearby, we again used a permutation approach. We started by assigning each CNEE to the gene represented by the nearest transcription start site. Then, we randomly permuted the vector of significantly accelerated or convergently accelerated CNEEs 10,000 times, so each permutation is effectively a random sample without replacement of N CNEEs, where N is the number of significant elements. This permutation preserves the CNEE-gene associations but randomly reassigns which CNEEs are significant. For each permutation, we computed the number of significant CNEEs assigned to each gene. We then used these as an empirical null distribution to compute a P-value for each gene reflecting the probability that we would observe X or more significant CNEEs associated with that gene among 10,000 random permutations.

Spatial enrichment

Because CNEE-gene assignments are difficult and CNEEs may not always regulate the nearest gene, we also looked for regions of the genome enriched for accelerated or convergently accelerated CNEEs, irrespective of CNEE-gene annotations. To do this, we divided the genome into 1 Mb sliding windows (100 kb slide), and for each window computed the probability of observing X significant CNEEs (accelerated or convergently accelerated) based on the binomial distribution where the number of trials is the number of CNEEs in the window and the probability of success is the proportion of all CNEEs that are significant.

phyloP analysis

To check the robustness of our results using the new Bayesian method, we also screened for accelerated CNEEs using phyloP (*105*). Here, we tested each ratite clade, along with tinamous, independently and estimated acceleration on that clade relative to the tree as a whole. We consider a clade to be ratite-accelerated by phyloP if we detect a significant acceleration (FDR 5%) in the ratite clade, but not in the tinamou clade. Because we used the whole genome alignment for this analysis, we only used the 'reduced' dataset (excluding moa) for these tests.

Section 7: Key summary

Data inputs:

CNEEs; moa and cormorant sequences

Methods and code:

Produce CNEE alignments (haltools, MAFFT); analysis changes in conservation across tree (phyloP, PhyloAcc); statistical analysis (R)

Github link:

https://github.com/tsackton/ratite-genomics/tree/master/ 07_cnee_analysis

Data outputs:

CNEE alignments: Dryad

PhyloAcc outputs: Dryad, Github

ATAC-seq sample preparation, sequencing, and analysis

ATAC-seq library preparation

ATAC-seq library preparation was carried out according to published protocols (*55*) with minor modifications. Three biological replicates were generated for each of the following tissues from chicken: e4.5 forelimb, e4.5 hindlimb, e9 flight muscle, e9 superior sternum, e9 inferior sternum, e10 full sternum, e10 keel and e10 flight muscle. These tissues were chosen based on their association with the flightless phenotype in ratites, comprised of reduced or vestigial forelimbs, more robust hindlimbs, reductions in flight (pectoral) musculature, and total loss of the sternal keel.

Single cell suspensions were generated as follows: tissue was dissected from the chicken embryo in cold PBS. For forelimb, hindlimb and flight muscle, tissue was immediately transferred to 1x Trypsin in EDTA Solution (Sigma) for 10-15 minutes at room temperature. Following trypsinization, the tissue was transferred to a neutralizing culture media (DMEM (Gibco) with 10% FBS (Gibco) and 1% Pen Strep (Gibco)) and pipette-mixed until homogenized. Homogenate was filtered through 35µm nylon mesh filters and filtrate was taken immediately to cell counting. For sternum and keel, dissected tissue was immediately transferred to collagenase 2 (Worthington Biochemical) in DMEM and Pen Strep (5 mg/ml collagenase 2 in DMEM with 1% Pen Strep) and incubated at 37 degrees Celsius for 30 minutes. Following the first incubation, the solution was mixed by pipette and placed in a shaking incubator for an additional 30-45 minutes at 37 degrees C. The solution was then pipette-mixed until homogenized and spun in a bench-top centrifuge for 5 minutes at 4 degrees C. The supernatant was discarded and the pellet was resuspended in the DMEM neutralizing culture media and filtered as above before proceeding to cell counting.

Cells were counted on an inverted light microscope using a hemocytometer. Cells to be counted were diluted in Trypan Blue (Sigma) to distinguish living cells from dead. Between 25,000 and 50,000 live cells were isolated and utilized for each library. Nucleus isolation, transposition, and reaction cleanup were carried out as described (*55*). Transposition utilized the Tn5 transposase and buffer from the Nextera DNA Library Preparation Kit (Illumina). Following transposition reaction cleanup, all libraries were PCR amplified for 11 cycles. 45 ul of product was taken through bead purification (18 ul or 0.4x to remove large fragments, followed by 63 ul or 1.4x to remove small fragments) with AmpureXP beads (Beckman Coulter). Libraries were visualized on a HS DNA Bioanalyzer Chip (Agilent) and quantified using a Library Quantification Kit (KAPA Biosystems) for multiplexing. ATAC-seq libraries were sequenced as 75 bp paired-end reads on a NextSeq High 150 flowcell (Illumina).

ATAC-seq Analysis

Raw ATAC-seq reads were trimmed using NGmerge (https://github.com/harvardinformatics/NGmerge) and mapped to the galGal4 genome using Bowtie2 (with the –X 2000 option). MACS2 was utilized for peak calling using the following pipeline to identify consistent peaks between biological replicates for a given tissue: (1) each individual library (n=24) was passed through MACS2 with a relaxed significance threshold (p-value < 0.05); (2) the biological replicates (n=3) for each tissue (n=8) were pooled together and passed through MACS2 with a stringent significance threshold (q-value < 0.05); (3) peak boundaries were defined by the peaks called for the pooled dataset; (4) bedtools intersect and bedtools annotate were utilized to identify pooled peaks (from step 2) that overlapped with peaks called individually for the three biological replicates in step 1; (5) a peak was only considered significant for a given tissue if it was called in the pooled dataset (q-value < 0.05), overlapped peaks in all three biological replicates, and also possessed a stringent significance value (q-value < 0.05) across all three individual biological replicates (as identified in step 1). This pipeline generated a single bed file containing consistent peak calls for each of the eight tissues we analyzed. Enrichment scores were calculated for CNEEs using the Genomic Association Test (GAT) sampling 10,000 times with the genome as background (*106*).

Section 8: Key summary

Data inputs:

ATAC-seq sequencing files (fasta): NCBI
Genomes, annotations: NCBI, Dryad

Methods and code:

Align reads (bowtie2), call peaks (MACS2), statistical analysis (R)
Github link: https://github.com/tsackton/ratite-genomics/tree/master/ 08_atacseq

Data outputs:

Peak calls: Dryad
Analyzed results: Dryad, Github

Enhancer candidate identification and functional testing

Identifying enhancer candidates

Following ATAC-seq quality control and association tests, the 284,001 CNEEs identified in our study were utilized as the input for a bedtools annotate run with all consistent peaks from our ATAC-seq libraries, alongside the ChIP-seq peaks from Seki et al (*39*). Candidate elements for functional testing were defined as chicken forelimb ATAC-seq peaks that overlapped a convergent ratite accelerated region (RAR) and at least one H3K27ac or H3K4me1 ChIP-seq peak from any stage. The 54 RARs identified at this step were further filtered into forelimb specific (found only in our forelimb ATAC-seq peaks) and multi-tissue (found in five or more of our ATAC-seq libraries). The multi-tissue peak candidates were then filtered a final time to obtain those ATAC-seq peaks that were found in at least five of the ten H3K27ac or H3K4me1 peak sets and none of the five H3K27me3 peak sets, the latter of which have been shown to be unable to drive reporter expression (*107*). Final candidates were selected based on visual inspection of their alignment and overlap with both ChIP- and ATAC-seq peaks in the genome browser and represented both forelimb-specific and multi-tissue ATAC-seq peaks. For selected enhancer candidates we also generated a list of potential binding sites using TRANFAC run with vertebrate transcription factor binding site models but otherwise default setting.

Screening enhancer activity in candidate CNEEs

Candidate RAR-associated ATAC-seq regions were screened as described below until we identified a chicken element that was able to drive consistent GFP expression in the developing chicken limb 24 hours after electroporation. We focused on genomic regions encompassed by forelimb ATAC-seq peaks since the enhancer screen would be carried out in the chicken forelimb. The chicken was chosen for these experiments because we sought to functionally test the sequence divergence identified between volant and flightless birds in an avian background, and the forelimb was chosen due to its important role in avian flight and its accessibility for electroporation. In total, we carried out preliminary screens on five putative limb enhancers to identify one to characterize fully as discussed below. Three of these initial candidates had forelimb specific ATAC-seq peaks, and two were taken from the class with multi-tissue ATAC-seq peaks. One of the two multi-tissue candidates was able to drive consistent GFP in the developing forelimb; we focused further characterizing this candidate. The five regions included the following CNEEs: mCE1623056 near TBX5, mCE1084527 near TMEM26, mCE1747138 near BMP7, mCE225714 near SHOX, and mCE967994 near TEAD1. Only the genomic region under the ATAC-seq peak including mCE967994 displayed strong, consistent GFP in chicken forelimbs during initial screening, and this element was selected for further study as described below. Additional candidates, while they did not drive GFP expression under default conditions, where not extensively optimized to increase the chance of observing activity, making it challenging to interpret the lack of signal in these experiments.

The enhancer screen vector was constructed by Hajime Ogino and generously provided by Mikiko Tanaka (*108*, *109*). This plasmid contains a multiple cloning site upstream of the β-actin basal promoter and EGFP (see Figure 4D, main text). To standardize the insert site and orientation of candidate enhancers within the vector, Gibson assembly was utilized (*110*). First, a stock of SmaI (blunt-end cut) enhancer screen plasmid was produced. All candidates were PCR amplified from genomic DNA using primers designed to also include tails with sequence complementary to each side of the SmaI cut-site. Following PCR amplification with high fidelity TaKaRa Long-Amp DNA Polymerase, products of the desired size were isolated using gel electrophoresis, gel extraction, and gel cleanup (Qiagen QIAquick). Cleaned PCR products were utilized in a Gibson assembly reaction using a 4:1 insert:plasmid ratio and allowed to ligate for 30-60 minutes. Competent E. Coli DH5α cells were transformed with the Gibson product, plated on selective media and allowed to grow overnight at 37 degrees Celsius. The following morning, colonies were picked to grow in 3 ml of selective media and colony PCR was performed on each one to screen for inserts of the correct size. Sequence identity and orientation was confirmed using Sanger sequencing in the T3 direction (through the multiple cloning site and into the product) following Qiagen Plasmid Mini- or Maxiprep.

Enhancer activity of mCE967994 sequence from flightless and volant species

The approximately 1.68 kb chicken genomic region (galGal4, NC_006092.3:7296321-7297998) surrounding the forelimb ATAC-seq peak containing RAR mCE967994 was amplified via PCR from chicken with the following forward (F) and reverse (R) primers:

F: gtggcggccgctctagactagtggatccccGTGATGCTAAATACAGATGGGTGT
R: gcagccgccgcctcgccatacctgcagcccTTAAGTAGAATTCCAGAAGCTGTCT

Homologous genomic regions were amplified from elegant crested tinamou (eudEle v1, scaffold_167:491258-493043) with the following forward (F) and reverse (R) primers:
F: gtggcggccgctctagactagtggatccccATGATATTAAATAAGGATTAGTGTTGCAGGT
R:gcagccgccgcctcgccatacctgcagcccGATTTTTTTTTTTTTTTAAAGCTAAATTCCAGCA

Homologous genomic regions were amplified from the greater rhea (rheAme v1, scaffold_17:331056-332770) with the following forward (F) and reverse (R) primers:
F: gtggcggccgctctagactagtggatccccAATGCTATTAAGTAAGGATTAGAGT
R: gcagccgccgcctcgccatacctgcagcccTTAAAGTAGAATTCCAGCAGTGT

Lowercase letters in the primers are complementary to the SmaI plasmid cut-site, while uppercase letter are complementary to the genomic DNA.

Forelimb field electroporation was carried out as previously described on hour 50-52 (~HH14) chicken embryos (*111*), with modifications described below. The DNA solution used for the electroporation contained the following working concentrations: 6 μg/μL of enhancer screen construct, <1 μg/μL of pCAGGS-H2B-RFP (for electroporation control), and 1% Fast Green. The Nepa Gene Super Electroporator Type II (NEPA21) was used with Nepa Gene CUY613P1 electrodes. The poring voltage was set to 50.0 V with pulse length 5.0 ms, pulse interval 10.0 ms, 3 pulses, + polarity. The transfer voltage was set to 5.0 V with pulse length 10.0 ms, pulse interval 10.0 ms, 5 pulses, and +/- polarity. The following day, ~20-24 hours after electroporation (HH19-20), embryos were examined and photographed at a total magnification of 5.8X in bright field, dsRED, and GFP LP using a Leica M165 FC Stereoscope with a Sola II LED Light Engine Illuminator. Embryos were scored as displaying strong GFP expression, partial GFP expression, or no GFP expression by eye during microscopy and again upon review of the images at a later date.

Section 9: Key summary

Data / laboratory inputs:
Enhancer screen electroporation vectors
Experimental materials

Methods and code:
Electroporate vectors; screen for control (RFP) and GFP signal

Data outputs:
Images: Dryad

Image Credits

All images used in Figure 1 and Figure 4 are either in the public domain or available with a Creative Commons license. Public domain images: southern cassowary, little spotted kiwi, North Island brown kiwi, greater rhea, Chilean tinamou, Thicket tinamou, little bush moa, ostrich, and continent shapefiles. The emu image [CC-BY-SA] is attributed to GFDL via Wikimedia Commons, the red jungle fowl/chicken image [CC-BY-SA] is attributed to Subramanya C K via Wikimedia Commons, and the tinamou image in Figure 4 [CC-BY-SA] is attributed to Dominic Sherony via Wikimedia Commons.

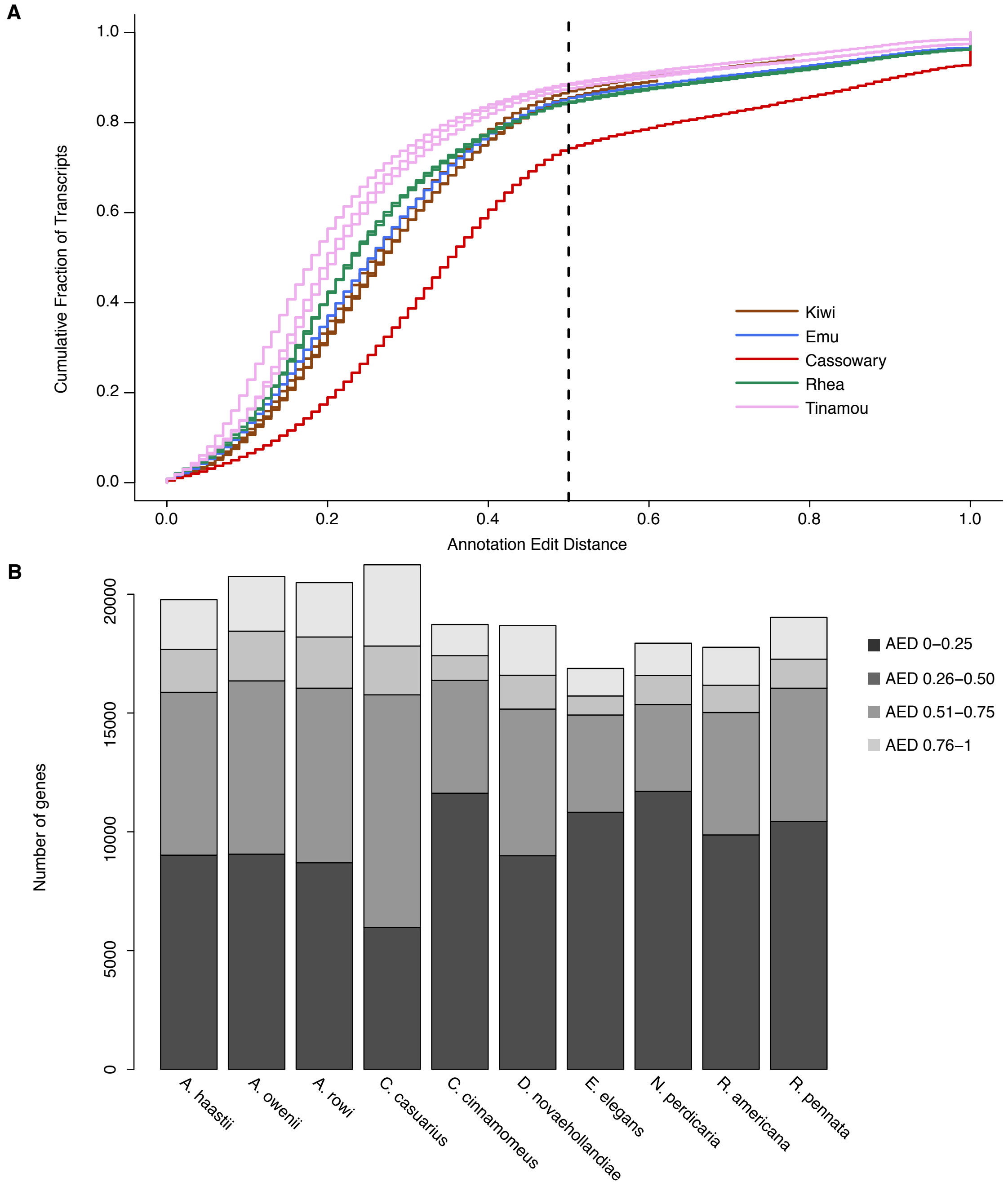


Fig. S1. Genome annotations are of high quality in all newly sequenced species.
A) Distribution of annotation edit distances from MAKER annotations across our genomes, which generally show acceptable to high quality results. B) Number of genes annotated across each species, in bins based on annotation edit distance. Lower numbers indicate better agreement with evidence (RNA-seq and cross species protein mapping). In all cases, a substantial majority of genes are annotated with AED below 0.5.

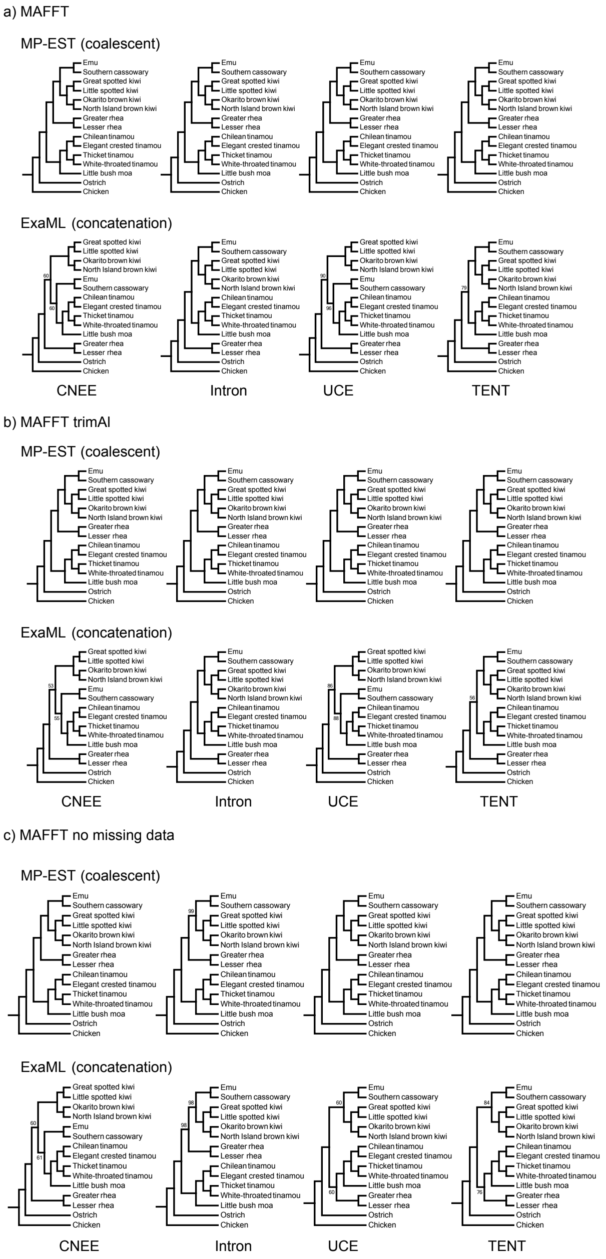


Fig. S2. Inference of paleognath relationships from MP-EST coalescent based and ExaML concatenation analyses.

Topologies are given for separate analyses of conserved non-exonic elements (CNEEs), introns, and ultraconserved elements (UCEs), as well as for the total evidence nucleotide tree (TENT) combining loci from all three marker types. Only bootstrap support values < 100% are drawn. Topologies are shown for (A) MAFFT sequence alignments with no additional trimming or filtering, (B) alignments trimmed with the automated heuristic column-based filtering of trimAl, and (C) alignments requiring a full data matrix with no missing taxa per locus and with columns containing gaps or undetermined bases removed.

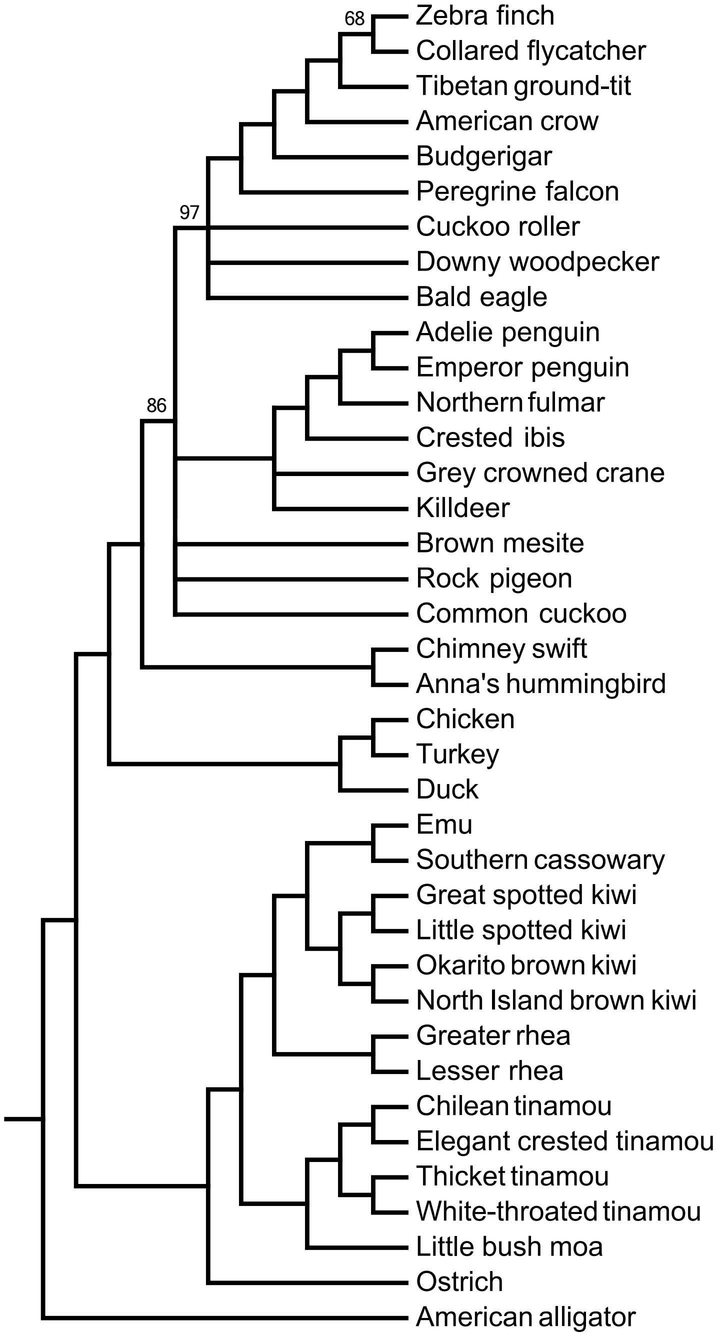
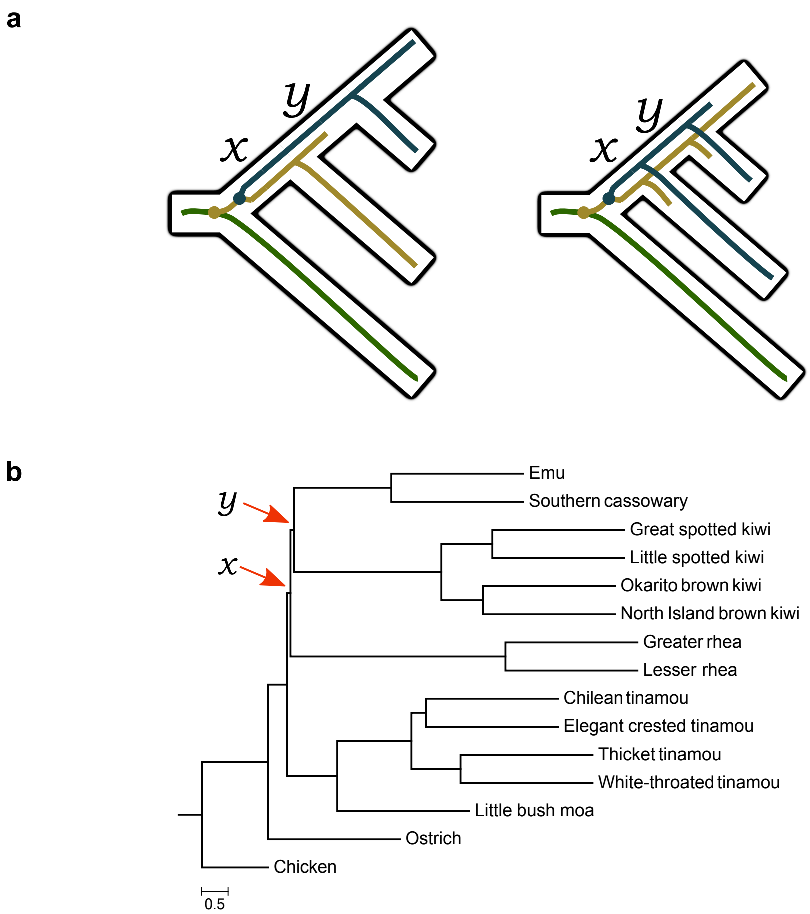


Fig. S3. MP-EST species tree for 6,931 CNEEs.

The recovered topology for paleognaths is robust to the choice of an alternative outgroup (American alligator), and the inclusion of additional avian ingroup taxa. Branches with support < 50% are collapsed, and only bootstrap support values < 100% are drawn.



c

Clade 1	Clade 2	*x*	*a(x)*	*y*	AGTs expected?
All Kiwi	Spotted kiwi	2.792	-0.389	0.965	No
All Kiwi	Brown kiwi	2.792	-0.389	0.788	No
Emu/Cassowary + Kiwi	Emu/Cassowary	0.066	0.519	1.844	No
Emu/Cassowary + Kiwi	Kiwi	0.066	0.519	2.792	No
Emu/Cassowary + Kiwi + Rheas	Rheas	0.063	0.543	4.075	No
Emu/Cassowary + Kiwi + Rheas	Emu/Cassowary + Kiwi	0.063	0.543	0.066	Yes
All Tinamous	Thicket & White-throated tinamous	1.406	-0.331	0.930	No
All Tinamous	Chilean & Elegant crested tinamous	1.406	-0.331	0.276	No
Moa + Tinamous	Tinamous	0.948	-0.272	1.406	No
Non-ostrich palaeognaths	Emu/Cassowary + Kiwi + Rheas	0.366	-0.080	0.063	No
Non-ostrich palaeognaths	Moa + Tinamous	0.366	-0.080	0.948	No

Fig. S4. Evidence for incomplete lineage sorting.

(A) Incomplete lineage sorting (ILS) can produce anomalous gene trees (AGTs) across pairs of short internal branches, labelled x and y. Colored lines represent different alleles, and circles indicate mutational events. (B) MP-EST species tree for the 20,850 locus total evidence data set, with internal branch lengths in coalescent units from MP-EST analysis of maximum likelihood RAxML gene trees. Terminal branch lengths are uninformative and are drawn as a constant value across all taxa. (C) Branch lengths in coalescent units for all pairs of branches x and y across the species tree, with the boundary of the anomaly zone a (x) calculated following Equation 4 from (*32*). AGTs are expected when y < a (x).

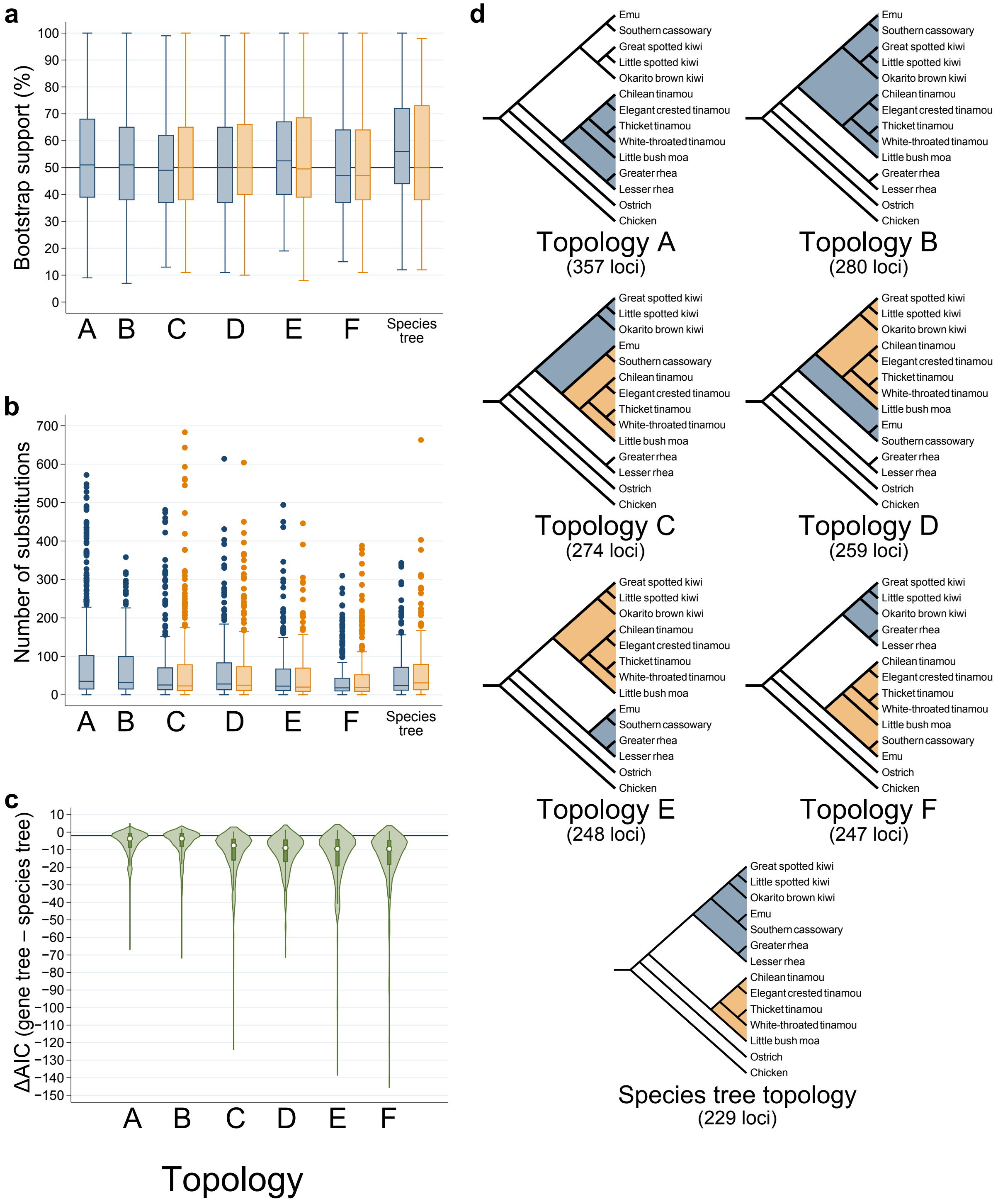


Fig. S5. Gene trees that conflict with the MP-EST species tree have strong support. Boxplots of bootstrap support (A) and the number of substitutions under a parsimony criterion (B) for clades that conflict with the species tree topology. A reference line of 50% bootstrap

support is drawn in (A). C) Violin plots of the difference in Akaike information criteria (AIC) for the recovered gene tree topology for each locus relative to the AIC when the sequence alignment is constrained to the species tree topology. A reference line is drawn at ΔAIC= -2, indicating substantial support in favor of the gene tree topology over that of the species tree for a given locus. D) Diagrams of the six gene tree topologies that occur at highest frequency (topologies A–F), and of the species tree topology. Clades considered in parts A and B are shaded and marked with star symbols.

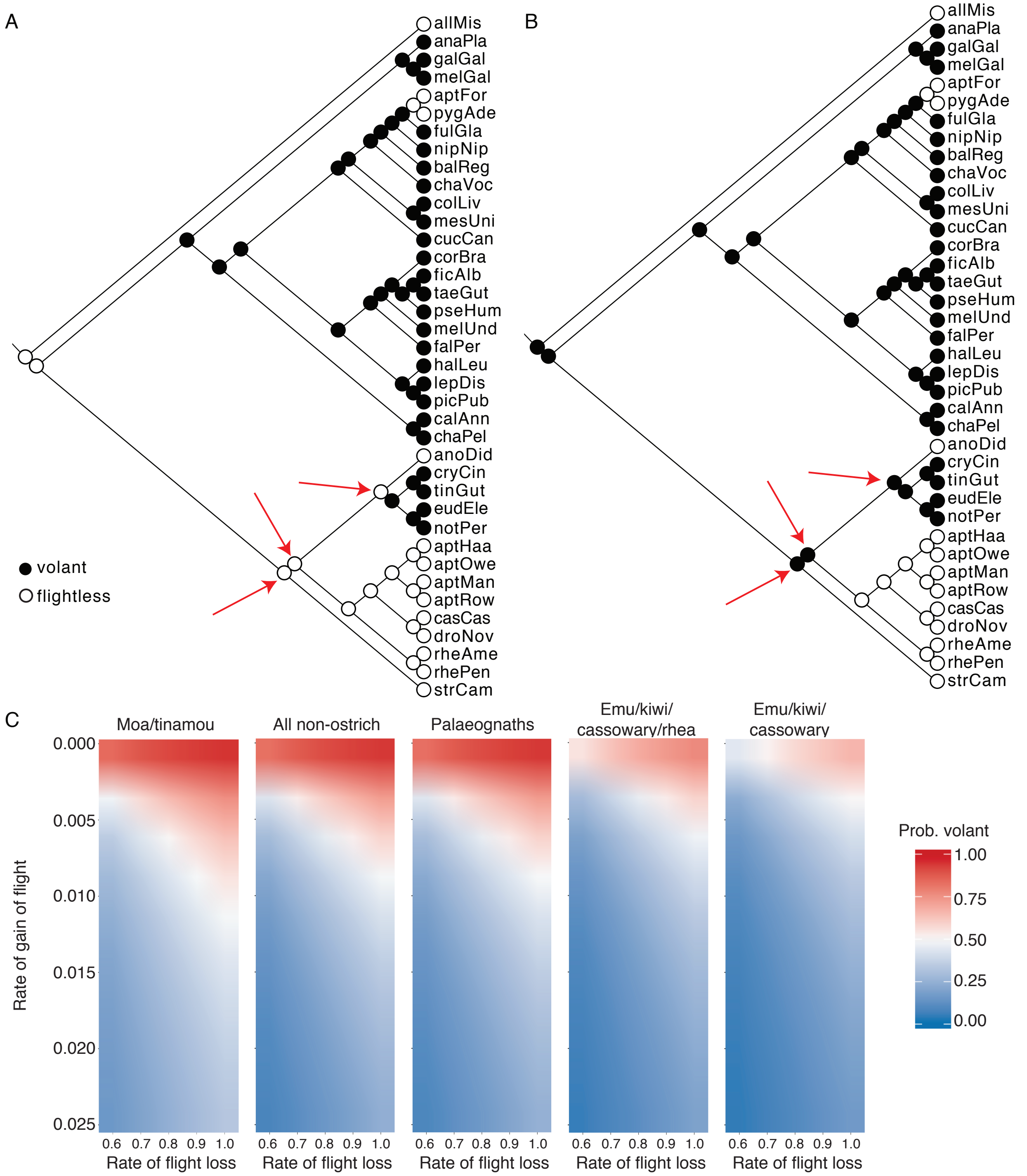


Fig. S6. Conditions supporting multiple independent losses of flight in ratites.
We used both parsimony and maximum likelihood methods to explore under what conditions multiple independent losses of flight in ratites is supported over a single gain of flight in tinamous. (A, B) Parsimony reconstructions from Mesquite under equal costs (A) and unequal costs (B, where cost of gain of flight (in parsimony steps) is three times or greater the cost of loss of flight). Key internal nodes in paleognaths are highlighted with red arrows. Multiple independent losses of flight is not supported in the equal cost reconstruction, but becomes much

more likely once the cost of gain in steps is at least three times the cost of loss. (C) Probability that key internal nodes in paleognaths are volant under a range of loss and gain rates, fit using the maximum likelihood methods implemented in corHMM. Descendant taxa of each node are labeled along the top of each panel. As long as rates of gain are relatively low compared to rates of loss, multiple independent losses of flight become well supported.

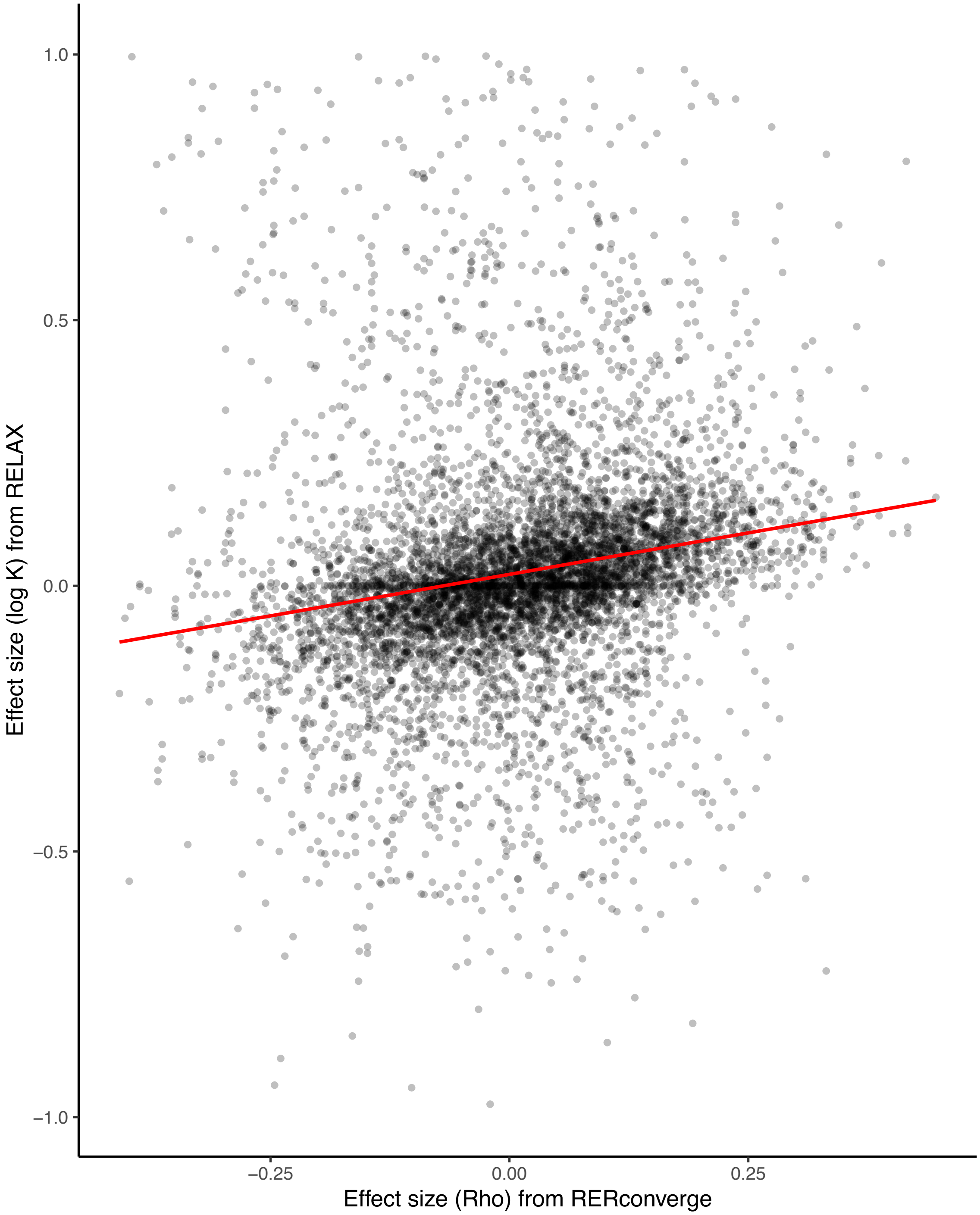


Fig. S7. Effect size estimates from RELAX and RERconverge are highly correlated. To validate the RERconverge results, we compared the effect size estimated by RERconverge (rho, the correlation coefficient of relative rate compared to the binary trait tree) to that estimated by RELAX (K, the change in constraint associated with target lineages). The red line represents the best fit line; effect sizes are highly correlated (rho = 0.32, P < 2.2e-16).

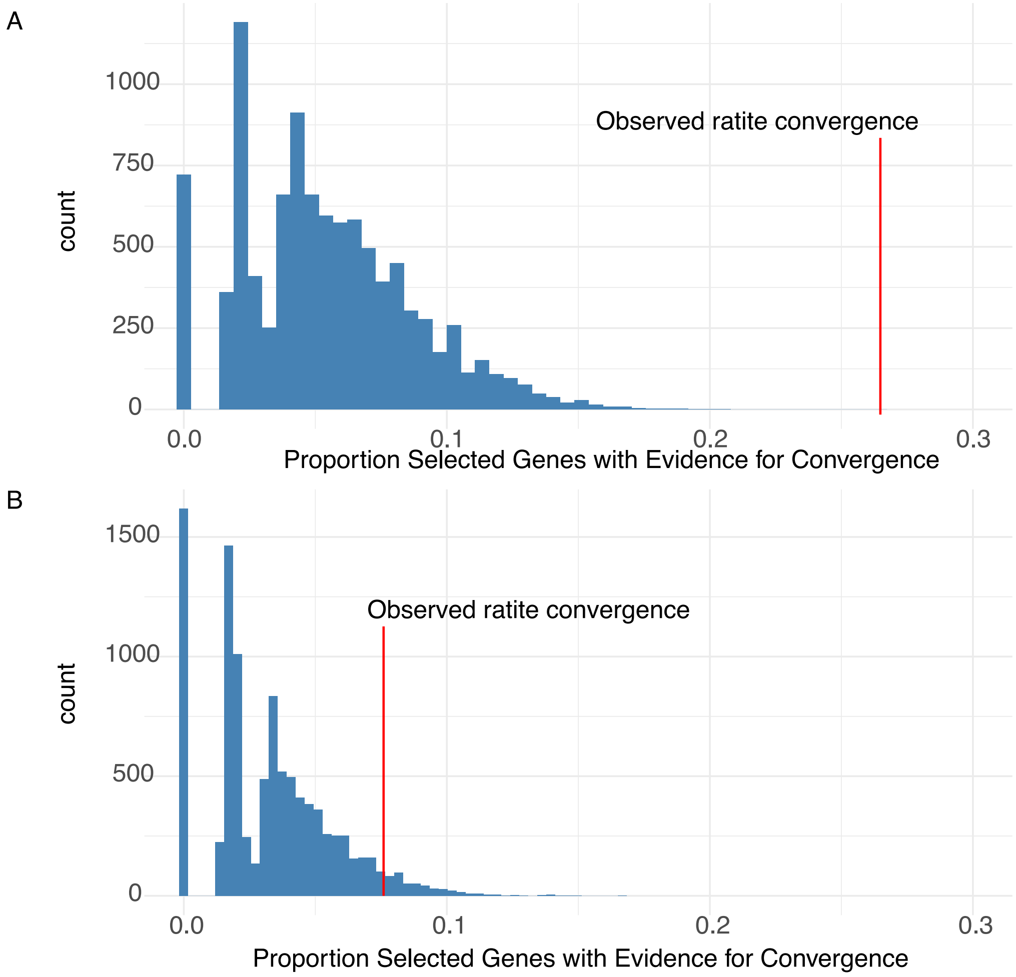


Fig. S8. Proportion of convergently positively selected genes

To determine if there were more genes with evidence for positive selection in multiple ratite lineages than we would expect under a model of independent selection (no convergence), we used a permutation test (see methods; light blue histogram). Red lines indicate the observed degree of convergent selection in protein-coding genes for the 'default dataset (including moa; panel A; $P < 0.001$) and the 'reduced' dataset (excluding moa; panel B, $P = 0.0475$).

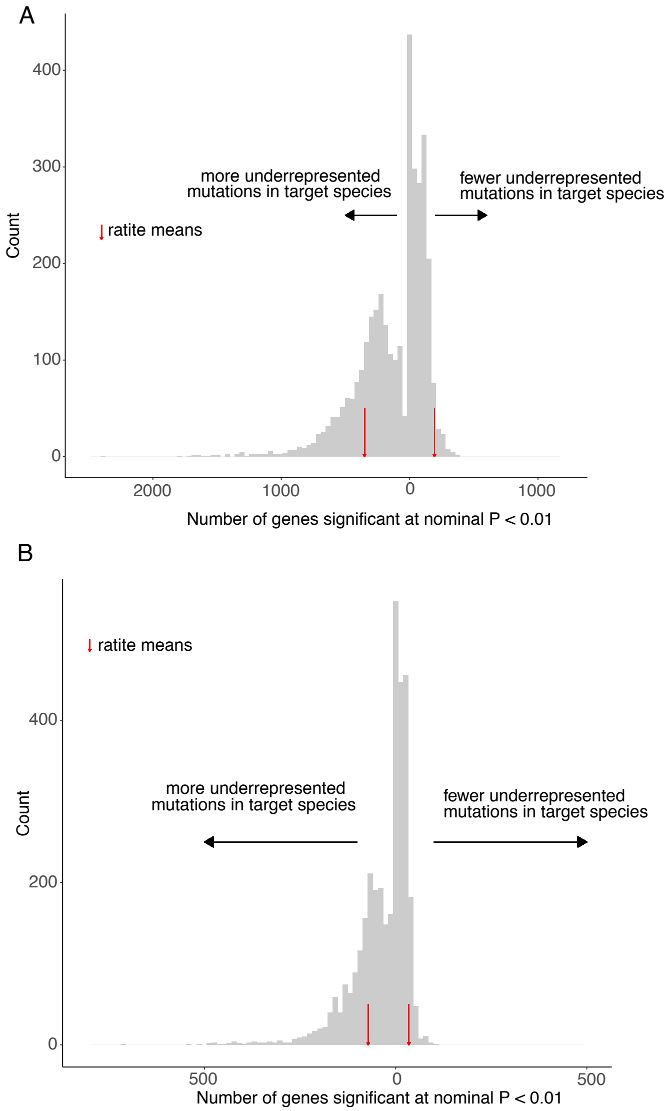


Fig. S9. Ratites are not enriched for convergent, putatively function-altering substitutions in protein-coding genes.
We sampled random trios of non-sister birds species (excluding trios that are entirely vocal learners or entirely ratites) and computed the number of proteins with significantly high or low DeltaBS scores for each trio compared to all other birds, as described in methods. Counts of proteins significant at a nominal P = 0.01 for both positive (amino acid sequence in trio is a better match to consensus than outside trio) and negative (amino acid sequence in trio is more diverged from the consensus than outside trio) DeltaBS scores are plotting for these random samples. Mean counts for ratite trios (also excluding sister lineages) are plotted as vertical lines. (A) Reduced dataset (excluding moa). (B) Original dataset (including moa).

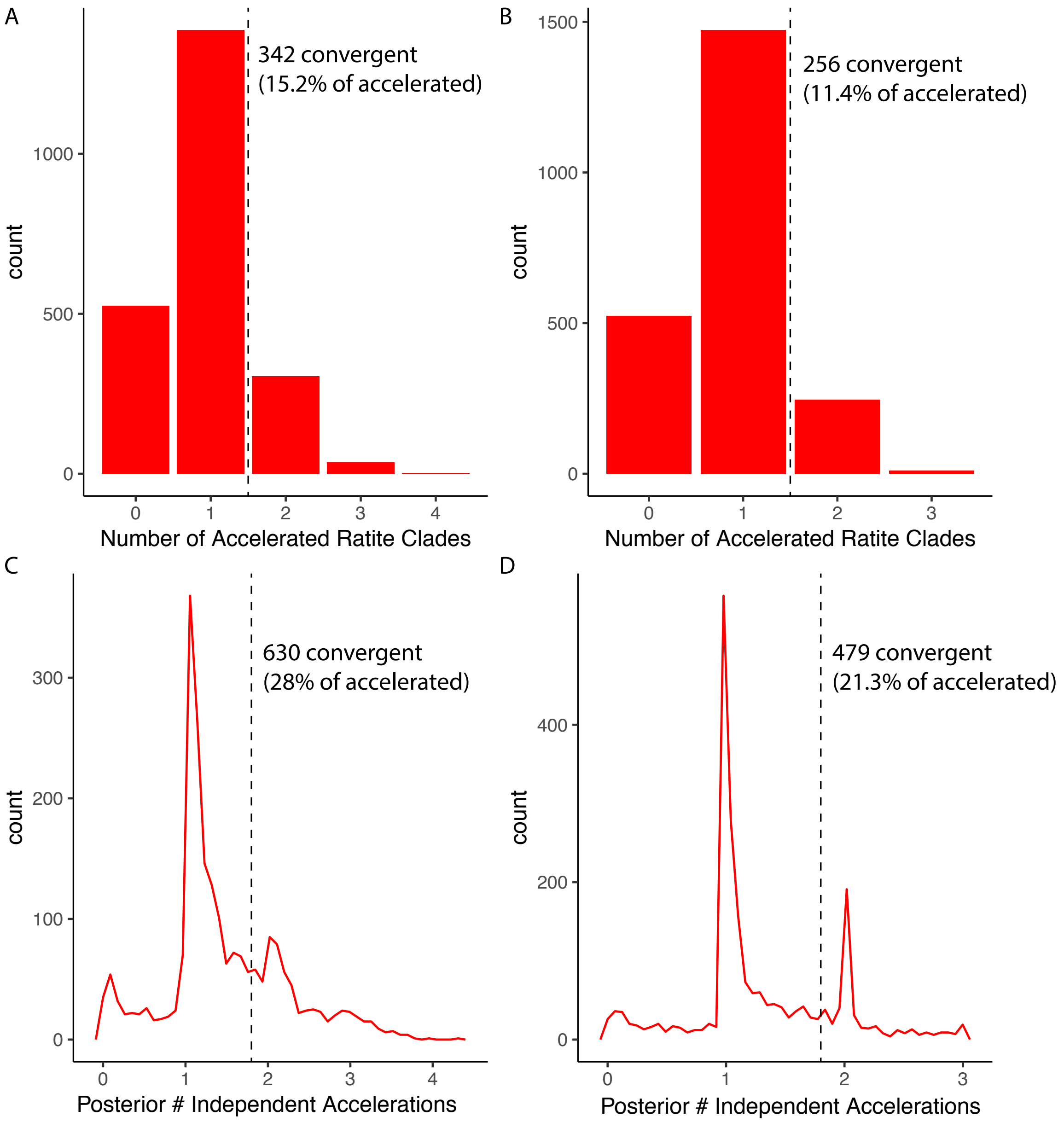


Fig. S10. Number of convergently accelerated elements under different scenarios.
Distribution of the number of independent accelerations for ratite-accelerated CNEEs. Black lines show cut-off for defining an element as convergently accelerated. In all plots the 0 bar represents elements with evidence for acceleration across ratites but for which we cannot identify with confidence the specific clades; in all cases we also exclude CNEEs with evidence for acceleration (posterior prob > 0.90) in any non-ratite lineage. (A) Discrete distribution where an element is considered accelerated in a clade (ostrich, moa, emu/cassowary, kiwi, or rhea) if the posterior probability of loss in that clade is greater than 0.90. (B) Discrete distribution as in (A), but using a parsimony restricted set of clades such that at most a single independent acceleration is allowed among emu/kiwi/cassowary/rhea. (C) Posterior estimated number of independent accelerations, allowing at most one acceleration per clade. (D) As (C), except restricting ‘clade’ to the three parsimony-consistent losses of flight (moa, ostrich, and emu/kiwi/cassowary/rhea).

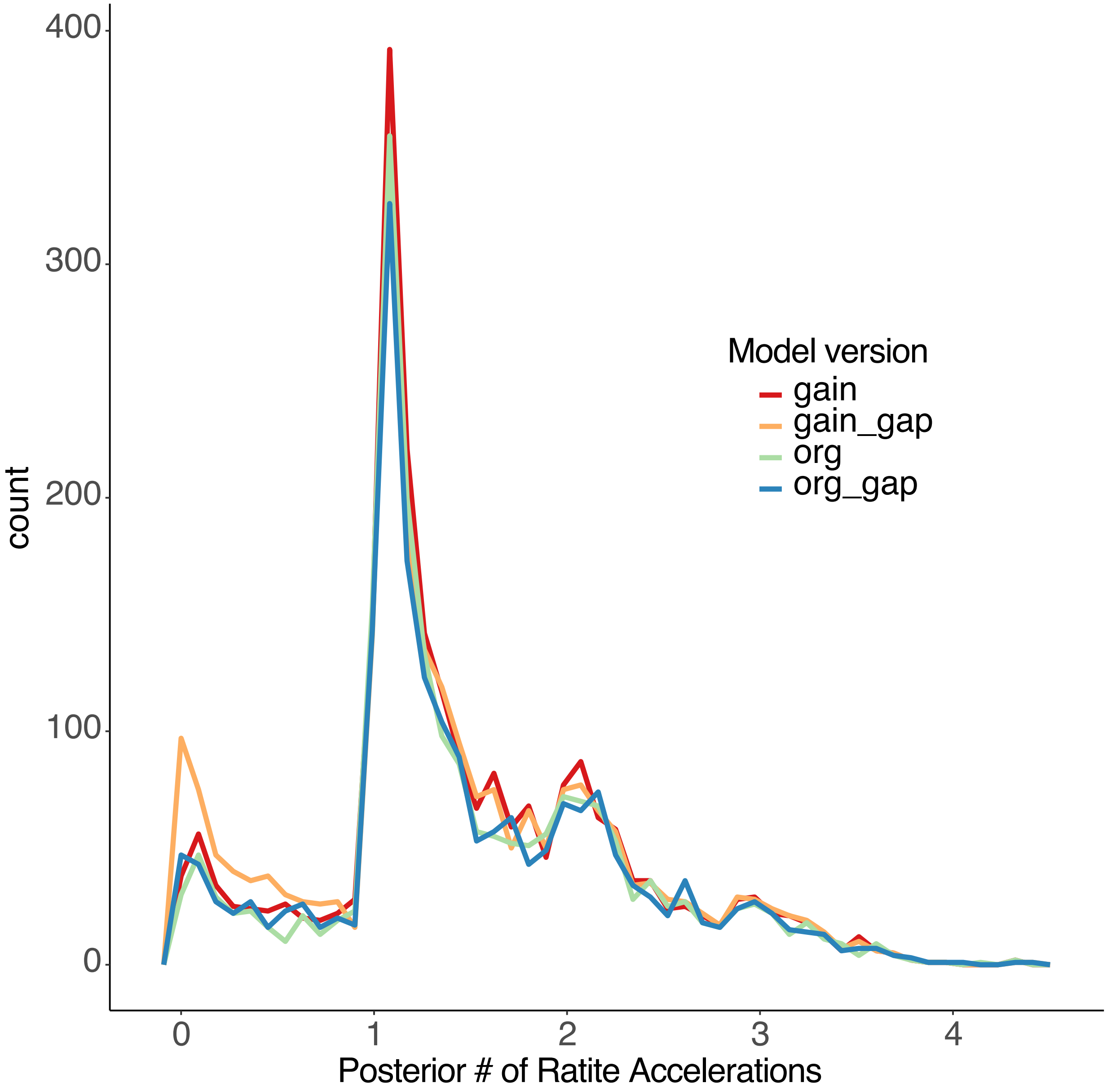


Fig. S11. PhyloAcc results are consistent across parameter sets.

To verify that our conclusions are not dependent on modeling assumptions in PhyloAcc, we repeated our analysis on four different parameter sets: gain [prior(gain) ~ Beta(9,1), treat gaps as unknown], gain_gap [prior(gain) ~ Beta(9,1), penalize gap branches in conserved state], gap [prior(gain) ~ Beta(3,1), penalize gaps branches in conserved state], and orig [prior(gain) ~ Beta(3,1), treat gaps as unknown]. The "gain" versions have a higher rate of transition from neutral to conserved, and the "gap" versions assume that gaps are unlikely to occur on conserved branches. In all cases, we observe nearly identical number of accelerated elements with very similar distributions for the number of convergent accelerations.

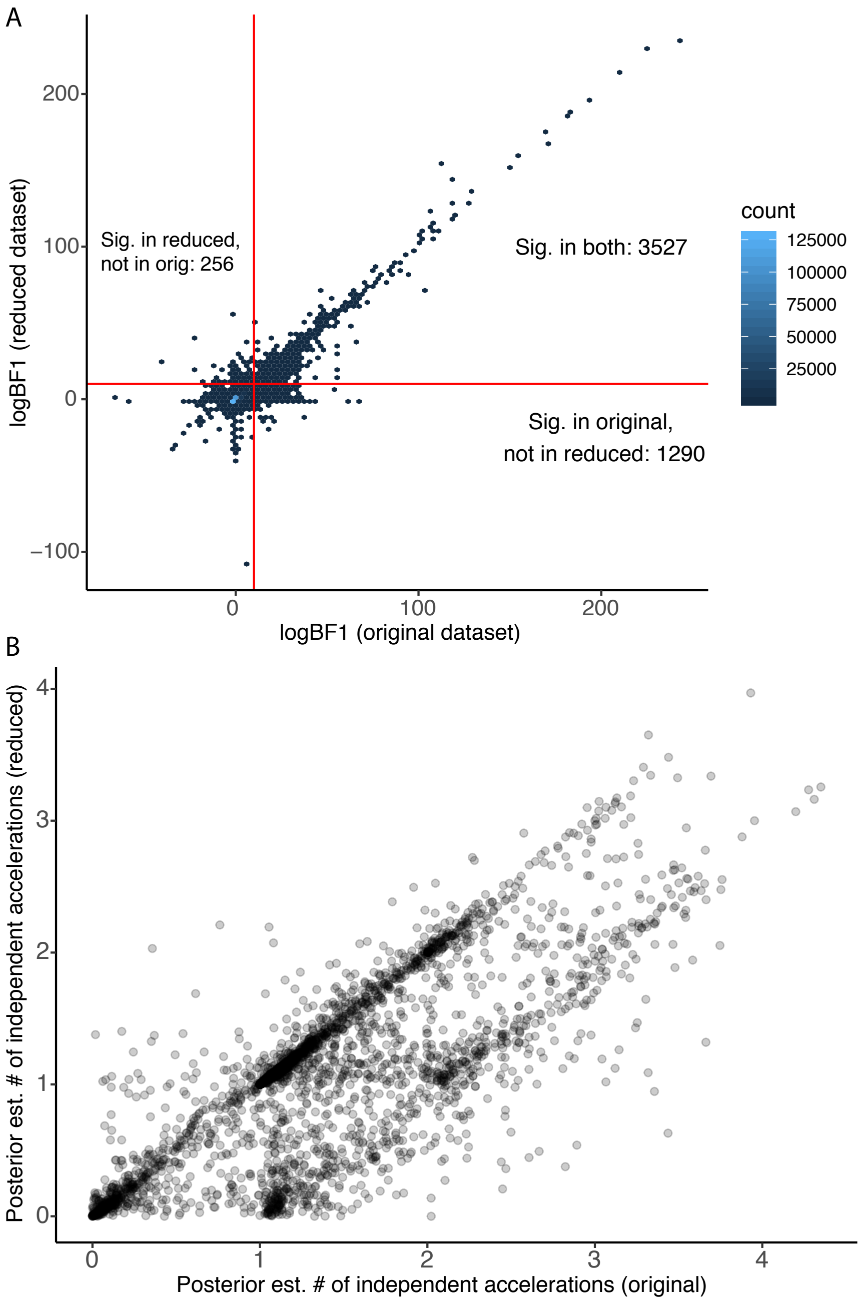


Fig. S12. Ratite accelerated elements are largely consistent between datasets.

To verify that our results are not sensitive to taxon choice, we repeated our analysis (using the “gain” parameter set, described above) with a reduced dataset that excludes the moa genome assembly. (A) The relative support for ratite acceleration (BF1) is highly correlated across datasets (each point represents a 3x3 bin colored according to the number of elements in that bin). Red lines indicate BF1 = 10 for each dataset; points in the upper right quadrant and therefor significant in both datasets. Counts of points in each of the three quadrants with BF1 > 10 in at least one dataset are labeled. (B) Posterior estimates of the number of independent losses in ratites are also highly correlated, with an expected shift in the original dataset which includes one extra ratite species.

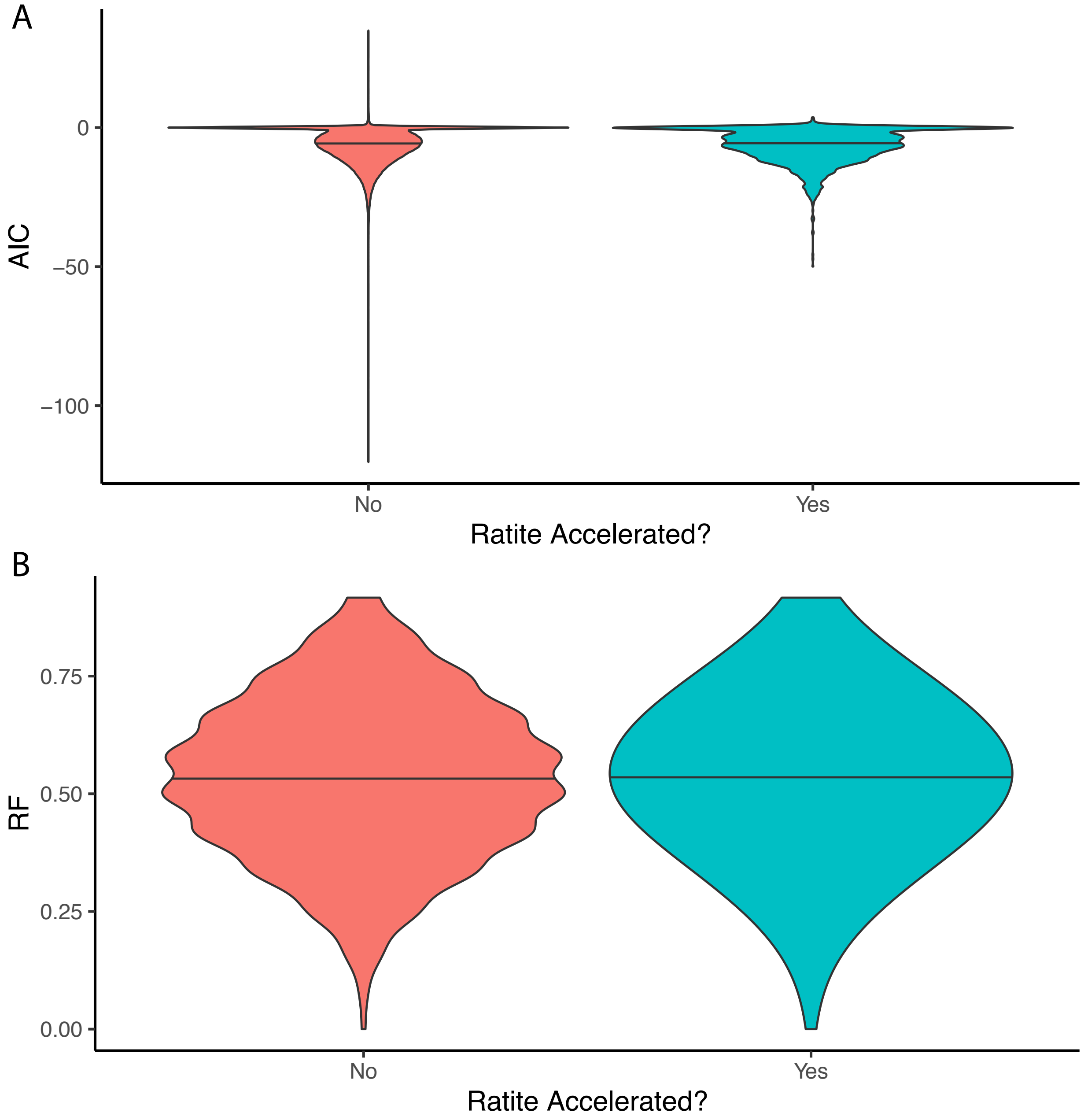


Fig. S13. No evidence for increased discordance in ratite-accelerated regions.
Because gene tree discordance can produce apparent increases in substitution rate on a species tree, we tested whether ratite-accelerated regions are more likely to have discordant gene trees, as would be predicted if discordance contributed to the probability of detecting acceleration. Using either difference in AIC between gene and species tree (panel A) or normalized Robinson-Folds distance (B), we find no difference between ratite-accelerated regions (defined as elements with BF1 > 10, BF2 > 1, and no evidence of tinamou acceleration) and other CNEEs.

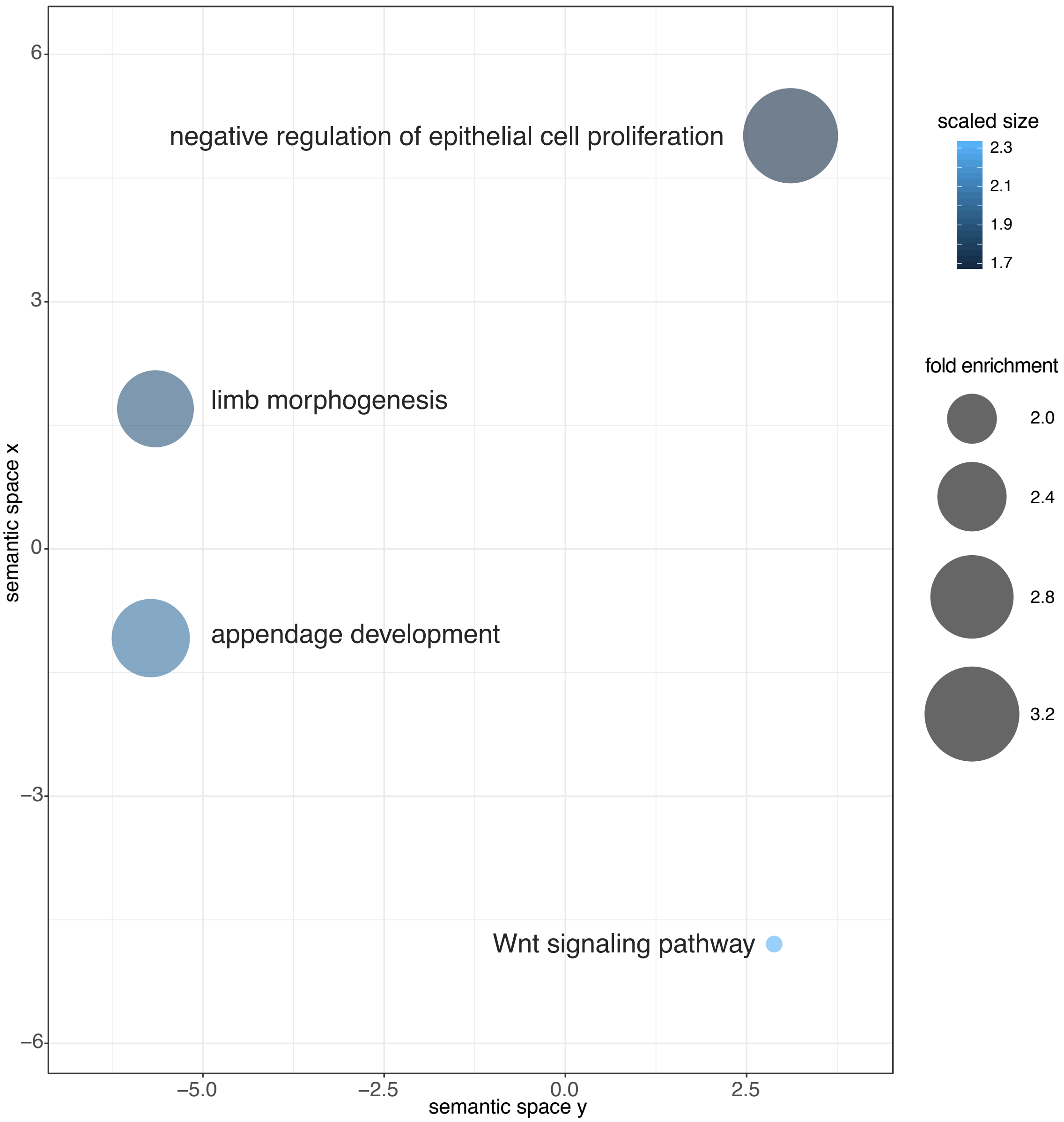


Fig. S14. Semantic clustering of GO terms enriched among genes near convergently accelerated elements.

To reduce redundancy among GO terms enriched among genes near convergently accelerated elements, we used REVIGO (*112*) to group biological process GO terms using default parameters (we do not cluster molecular function terms or terms enriched near accelerated elements regardless of convergent status, as there are only a few terms in each of these cases). Non-redundant terms are plotted in arbitrary semantic space, where terms with more similar constituent of genes appear closer together. Terms are colored based on set size (terms with more genes are lighter blue) and sized based on enrichment (proportion of genes in target set with term divided by proportion of genes in background with term).

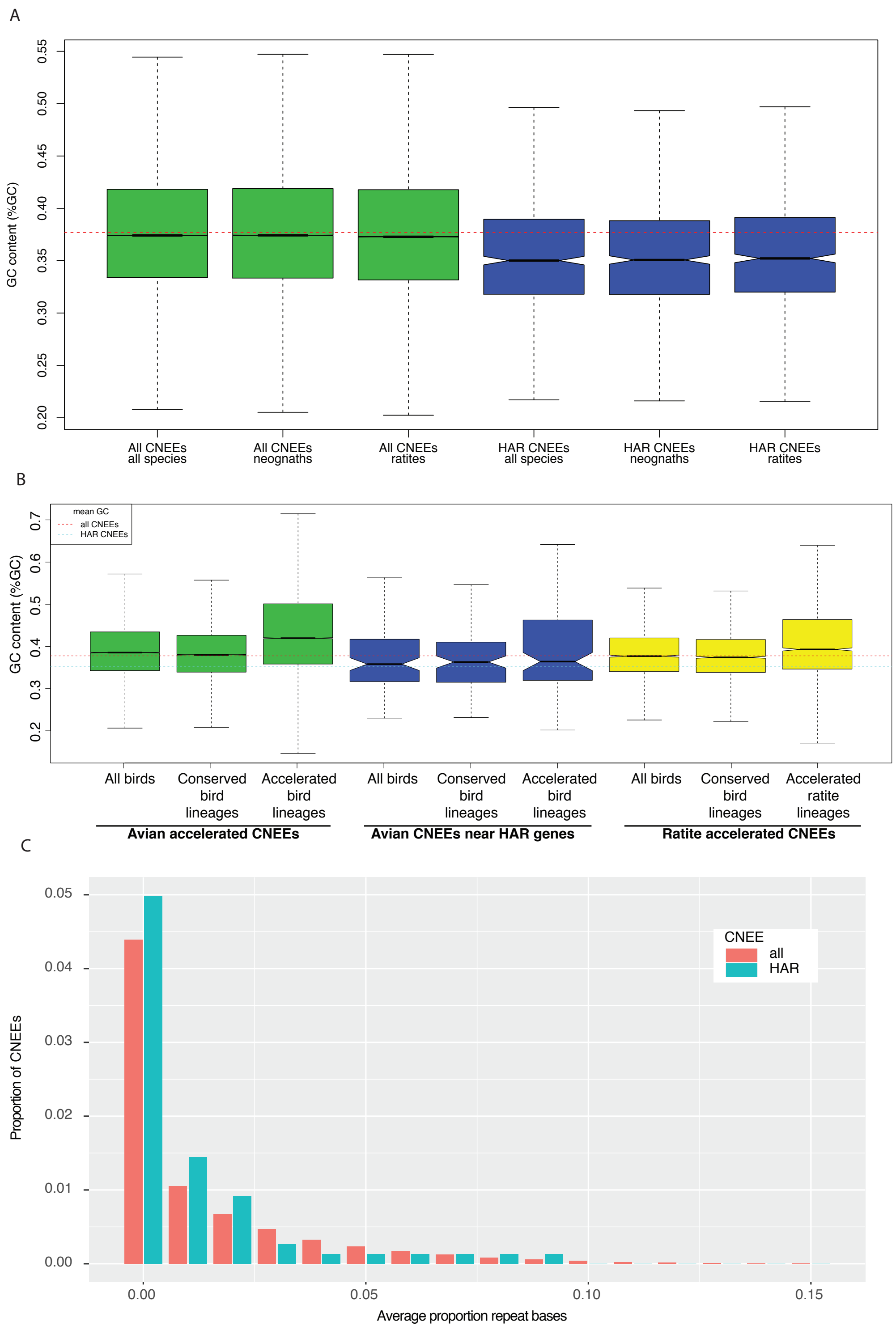


Fig. S15. Conserved elements near genes associated with HARs are not atypical.
We identified a subset of genes enriched near ratite-accelerated regions that are also enriched near human-accelerated regions (HARs) from (51). A) CNEEs in birds near genes associated with HARs have a lower GC content than the entire set of avian CNEEs. Because GC content is a potential indicator of gene conversion, this suggests that avian CNEEs near genes enriched for

HARs are not being detected because of higher levels of gene conversion. B) Accelerated CNEEs exhibit higher GC content than the total set of avian CNEEs, both across all birds (green, n=32918) or within ratites (yellow, n= 2639). CNEEs where Bayes Factor 1 (comparing the full model to the null model with no acceleration) is > 10 are considered accelerated. The possibility that accelerations of CNEEs are driven by gene conversion, which would be predicted to increase GC content, is accommodated by PhyloAcc, the Bayesian acceleration model (44). However, the avian CNEEs near genes enriched for HARs (n=111, blue), including those undergoing acceleration within birds, do not exhibit a marked increase in GC content relative to all CNEEs, indicating again that GC biased gene conversion is unlikely to be driving the coincidence of avian and human genes exhibiting high numbers of accelerated elements this and other studies (51-53). C) The abundance of simple sequence repeats is no higher for avian CNEEs near genes enriched for HARs than for the entire set of CNEEs, indicating that simple sequence repeats is unlikely to be driving the coincidence of avian and human genes exhibiting high numbers of accelerated elements in this and other studies (51-53). For each CNEE, we searched for simple repeats (i.e. (CA)n or (TG)n with n>=4) in all species and computed the repeat content as the average length of repeats across all species divided by the total length of each element. We then compared repeat content in the 762 CNEEs associated with HARs (around DACH1, NF1B or NPAS3) with all CNEEs. The x-axis shows repeat content or average proportion of repeats; each bar is the proportion of CNEEs with repeat content above the number on x-axis. Fewer than 5% of CNEEs across any species harbor simple sequence repeats. For each cutoff of repeat content, there is no significant difference between the proportion of HAR-associated CNEEs and other CNEEs.

Table S1.
Source information for Paleognath genomes included in this study

Code	Scientific name	Common name	Accession	Institution[1]	Collection locality	Sex	Reference
anoDid	*Anomalopteryx didiformis*	Little bush moa	TW95	ROM	unprovenanced	unknown	This study, (24)
aptHaa	*Apteryx haastii*	Great spotted kiwi	R29945	ROM	Hurunui, Canterbury, New Zealand	Male	This study
aptMan	*Apteryx mantelli*	North Island brown kiwi	multiple[2]	multiple[2]	see ref.[2]	Female	(26)
aptOwe	*Apteryx owenii*	Little spotted kiwi	O 20599	ROM	Te Mimiorakopa, Kapiti Island, New Zealand	Male	This study
aptRow	*Apteryx rowi*	Okarito brown kiwi	R 43461	ROM	Okarito, Westland, New Zealand	Male	This study
casCas	*Casuarius casuarius*	Southern cassowary	B32419	ANWC	Babinda, Queensland, Australia	Female	This study
cryCin	*Crypturellus cinnamomeus*	Thicket tinamou	337707	MCZ	Cafetal stream, Guanacaste, Costa Rica	Male	This study
droNov	*Dromaius novaehollandiae*	Emu	Cryogenic 6597	MCZ	Songline Emu Farm, Gill, Massachusetts[3]	Male	This study
eudEle	*Eudromia elegans*	Elegant crested tinamou	44215	ROM	Toronto Zoo, Ontario, Canada	Male	This study
notPer	*Nothoprocta perdicaria*	Chilean tinamou	Ctin4	ROM	Chilliwack, British Columbia, Canada[3]	Male	This study
tinGut	*Tinamus guttatus*	White-throated tinamou	B-42614	LSU	Loreto Department, Peru	Female	(25)
rheAme	*Rhea americana*	Greater rhea	MKP 2758	ROM	Viedma, Rio Negro, Argentina	Male	This study
rhePen	*Rhea pennata*	Lesser rhea	206185 (NZPBD100-12)	SNZ	Sea World, Orlando, Florida[4]	Male	This study
strCam	*Struthio camelus*	Ostrich	202443	SDZ	Botswana, Africa	Female	(25)

[1] ANWC: Australian National Wildlife Collection
MCZ: Harvard University Museum of Comparative Zoology
ROM: Royal Ontario Museum
SNZ: Smithsonian National Zoo
SDZ: San Diego Zoo

[2] Assembly produced from multiple individuals, see reference for details

[3] Farmed research animals

[4] Captive born individual

Table S2.

Sequencing and assembly information for new genomes

Scientific name	Total Frag. Reads	Total MP Reads	Scaffold N50	Contig N50	BUSCO score[1]
Apteryx haastii	294,924,492	277,444,152	1360	105.6	C:97.1, F.2.0, M:0.9
Apteryx owenii	447,749,042	451,574,758	1617	128.4	C:97.6, F:1.6, M:0.8
Apteryx rowi	357,835,928	436,303,038	1665	120.3	C:98.1, F:1.3, M:0.6
Casuarius casuarius	419,359,436	459,568,992	3700	133.1	C:97.8, F:1.3, M:0.9
Crypturellus cinnamomeus	269,962,512	322,211,994	2425	50.3	C:97.2, F:1.4, M:1.4
Dromaius novaehollandiae	501,233,652	397,997,402	3322	138.8	C:97.7, F:1.4, M:0.9
Eudromia elegans	342,759,376	329,360,988	3283	98.8	C: 98.1, F: 1.1, M: 0.8
Nothoprocta perdicaria	429,304,722	336,279,414	3348	75.8	C: 97.4, F: 1.8, M: 0.8
Rhea americana	390,468,074	375,784,406	4082	68.7	C: 97.6, F: 1.2, M: 1.2
Rhea pennata	312,785,672	331,128,486	3846	55.9	C: 96.3, F: 1.5, M: 2.2

[1] BUSCO scores are percentage of complete (C), fragmented (F), and missing (M) vertebrate BUSCO models in each genome.

Table S3.

Predicted genes in each genome

Scientific name	# Gene models	% ortho (OMA)	% chicken blastp[1]	% eggNOG hit[2]	BUSCO score[3]
Apteryx haastii	19,973	72.4%	98.80%	86.6%	C: 94.5, F: 4.3, M:1.2
Apteryx owenii	20,969	70.9%	99.07%	86.1%	C: 94.6, F: 3.9, M: 1.5
Apteryx rowi	20,710	71.2%	99.00%	85.7%	C: 95.1, F: 3.6, M: 1.3
Casuarius casuarius	21,342	72.4%	99.02%	87.1%	C: 93.6, F: 5.2, M: 1.2
Crypturellus cinnamomeus	18,883	83.7%	98.93%	95.7%	C: 95.3, F: 3.8, M: 0.9
Dromaius novaehollandiae	18,858	77.0%	98.90%	89.2%	C: 96.1, F: 3.1, M: 0.8
Eudromia elegans	17,021	84.9%	98.77%	95.6%	C: 95.4, F: 3.7, M 0.9
Nothoprocta perdicaria	18,151	78.5%	98.91%	88.8%	C: 96.1, F: 2.9, M: 1.0
Rhea americana	17,916	81.6%	98.61%	94.3%	C: 94.0, F: 4.9, M: 1.1
Rhea pennata	19,203	80.5%	98.88%	94.1%	C: 93.8, F: 5.4, M: 0.8

[1] Percent of chicken proteins with a blast hit to target species predicted proteome.

[2] Percent of predicted proteins in target species with a hit to a vertebrate eggNOG model.

[3] BUSCO scores are percentage of complete (C), fragmented (F), and missing (M) vertebrate BUSCO models in each proteome.

Table S4.

Phylogenomic data sets used for MP-EST coalescent and ExaML concatenation species tree inference.

Marker type	Data set	Num. loci	Aligned sequence length (bp)
CNEEs	MAFFT	12,676	4,794,620
	MAFFT trimAl	12,676	4,667,306
	MAFFT no missing data	11,125	3,851,232
	MAFFT allspecies (American alligator outgroup)	6,931	2,937,637
Introns	MAFFT	5,016	27,890,802
	MAFFT trimAl	5,016	23,346,653
	MAFFT no missing data	2,117	2,333,861
UCEs	MAFFT	3,158	8,498,759
	MAFFT trimAl	3,158	7,361,135
	MAFFT no missing data	1,837	2,409,470
TENT	MAFFT	20,850	41,184,181
	MAFFT trimAl	20,850	35,375,094
	MAFFT no missing data	15,079	8,594,563

Table S5. Model fits with number of convergent substitutions as response variable

Data Filtering	Alignments	Model Terms	Estimate[1]	P-value
Default	6,337	Divergent Substitutions	0.7885	< 2e-16
(1:1 orthologs only,		Evolutionary Distance	-766.97	0.00146
< 3 missing taxa, species tree)		Divergent:Distance interaction	-1.01	< 2e-16
		Is Ratite?	**46.39**	**0.60375**
		Overall	*0.9767*	*< 2e-16*
Default	6,337	Divergent Substitutions	0.1113	6.43e-11
(1:1 orthologs only,		Log (Evolutionary Distance)	-126.8	0.0284
< 3 missing taxa, species tree)		Divergent:Distance interaction	-0.3035	< 2e-16
		Is Ratite?	**68.82**	**0.4818**
		Overall	*0.9778*	*< 2e-16*
Default	6,337	Divergent Substitutions	1.4676	< 2e-16
(1:1 orthologs only,		Exp (Evolutionary Distance)	-580.28	0.00172
< 3 missing taxa, species tree)		Divergent:Distance interaction	-0.724	< 2e-16
		Is Ratite?	**52.54**	**0.55808**
		Overall	*0.9757*	*< 2e-16*
Gene Tree	6,342	Divergent Substitutions	0.6981	< 2e-16
(1:1 orthologs only,		Evolutionary Distance	-620.6	3.9e-07
< 3 missing taxa, gene tree)		Divergent:Distance interaction	-0.8271	< 2e-16
		Is Ratite?	**-5.018**	**0.9054**
		Overall	*0.9809*	*< 2e-16*
Gene Tree	6,342	Divergent Substitutions	0.1171	< 2e-16
(1:1 orthologs only,		Log (Evolutionary Distance)	-752.2	0.00266
< 3 missing taxa, gene tree)		Divergent:Distance interaction	-0.2672	< 2e-16
		Is Ratite?	**31.23**	**0.47443**
		Overall	*0.982*	*< 2e-16*
Gene Tree	6,342	Divergent Substitutions	1.241	< 2e-16
(1:1 orthologs only,		Exp (Evolutionary Distance)	-501.7	1.79e-07
< 3 missing taxa, gene tree)		Divergent:Distance interaction	-0.5837	< 2e-16
		Is Ratite?	**-5.576**	**0.896**
		Overall	*0.9798*	*< 2e-16*
Strict	3,946	Divergent Substitutions	0.7204	< 2e-16
(1:1 orthologs only,		Evolutionary Distance	-380.84	1.42e-06
no missing taxa, gene tree)		Divergent:Distance interaction	-0.852	< 2e-16
		Is Ratite?	**-3.15**	**0.9080**
		Overall	*0.9804*	*< 2e-16*
Strict	3,946	Divergent Substitutions	0.1232	< 2e-16
(1:1 orthologs only,		Log (Evolutionary Distance)	-45.81	0.00544
no missing taxa, gene tree)		Divergent:Distance interaction	-0.2741	< 2e-16
		Is Ratite?	**19.7**	**0.49144**
		Overall	*0.9811*	*< 2e-16*
Strict	3,946	Divergent Substitutions	1.281	< 2e-16
(1:1 orthologs only,		Exp (Evolutionary Distance)	-307.56	6.28e-07
no missing taxa, gene tree)		Divergent:Distance interaction	-0.6025	< 2e-16
		Is Ratite?	**-3.454**	**0.9**
		Overall	*0.9794*	*< 2e-16*

[1] Estimate is model coefficient for model terms, and the adjusted R-squared for the entire model for the overall term.

Table S6. RERconverge results under different datasets.

We computed the number of nominally significant genes (P < 0.01) and the number of genes with false discovery rate < 0.01 or < 0.10 for the two full datasets (‘default’, including moa; ‘reduced’ excluding moa), and also for subsets of the ‘default’ dataset that include non-sister ratite trios. For comparison, we also include results for convergent protein evolution in vocal learners, which has been previously reported (*113*). While the full datasets have some weak evidence of convergent shifts in rates of protein evolution in ratites, this seems largely to be driven by the unbalanced taxon sampling as the signal is strongly attenuated in downsampled datasets with more balanced sampling.

Dataset	*Targets*	*P < 0.01 up*	*P < 0.01 down*	*FDR 1% up*	*FDR 1% down*	*FDR 10% up*	*FDR 10% down*
Default	Ratites	214 (2.8%)	238 (3.1%)	**0**	**0**	123	151
Default	Vocal learners	387 (5.1%)	246 (3.2%)	40	6	335	187
Reduced	Ratites	225 (2.6%)	218 (2.5%)	**0**	**0**	61	65
Reduced	Vocal learners	439 (5.0%)	300 (3.5%)	48	7	380	233
Subset	Moa, ostrich, emu	25 (0.33%)	70 (0.92%)	**0**	**0**	**0**	**0**
Subset	Moa, ostrich, kiwi	45 (0.59%)	78 (1.0%)	**0**	**0**	**0**	**0**
Subset	Moa, ostrich, kiwi	45 (0.59%)	72 (0.95%)	**0**	**0**	**0**	**0**
Subset	Moa, ostrich, kiwi	39 (0.51%)	72 (0.95%)	**0**	**0**	**0**	**0**
Subset	Moa, ostrich, cassowary	31 (0.41%)	63 (0.83%)	**0**	**0**	**0**	**0**
Subset	Moa, ostrich, rhea	34 (0.45%)	93 (1.2%)	**0**	**0**	**0**	**0**
Subset	Moa, ostrich, rhea	40 (0.53%)	102 (1.3%)	**0**	**0**	**0**	**0**

Table S7. Counts of clades with positively selected genes in BS-REL analysis

Clade(s)	*Selected count, reduced*	*Selected count, default*
Emu	99	40
Kiwi	95	46
Ostrich	32	10
Rhea	99	62
Moa	N/A	17
Emu-Kiwi	8	4
Rhea-Emu	9	10
Rhea-Kiwi	7	6
Moa-Emu	N/A	14
Moa-Kiwi	N/A	6
Moa-Ostrich	N/A	2
Moa-Rhea	N/A	1
Ostrich-Emu	0	2
Ostrich-Kiwi	1	1
Ostrich-Rhea	1	4
Moa-Emu-Kiwi	N/A	2
Moa-Rhea-Emu	N/A	3
Moa-Rhea-Kiwi	N/A	1
Ostrich-Rhea-Emu	0	1
Ostrich-Rhea-Kiwi	0	1
Rhea-Emu-Kiwi	1	3
Moa-Rhea-Emu-Kiwi	N/A	2

Table S8. Counts of clades with convergently accelerated CNEEs in PhyloAcc analysis

Species are defined as accelerated Bayes Factor 1 > 10 and Bayes Factor 2 > 1 and there is no evidence for acceleration in tinamous (posterior estimated losses < 1 among all tinamou lineages). Lineages are defined as accelerated if posterior probability of acceleration in that lineage is at least 0.80

Clade(s)	*Accelerated count, reduced*	*Accelerated count, default*
None	402	448
Emu	233	216
Kiwi	164	145
Ostrich	243	218
Rhea	495	455
Moa	N/A	383
Emu-Kiwi	19	15
Rhea-Emu	54	47
Rhea-Kiwi	51	49
Moa-Emu	N/A	38
Moa-Kiwi	N/A	39
Moa-Ostrich	N/A	74
Moa-Rhea	N/A	49
Ostrich-Emu	24	23
Ostrich-Kiwi	14	12
Ostrich-Rhea	55	48
Rhea-Emu-Kiwi	6	8
Moa-Emu-Kiwi	N/A	13
Moa-Rhea-Emu	N/A	9
Moa-Rhea-Kiwi	N/A	9
Ostrich-Emu-Kiwi	4	3
Ostrich-Rhea-Emu	10	9
Ostrich-Rhea-Kiwi	8	8
Moa-Ostrich-Kiwi	N/A	5
Moa-Ostrich-Emu	N/A	8
Moa-Ostrich-Rhea	N/A	21
Ostrich-Moa-Rhea-Kiwi	N/A	2
Ostrich-Rhea-Emu-Kiwi	1	1

Additional Data 1 (xls)

List of GO terms associated with positively selected genes based on BS-REL. The “spec” set represents genes only selected in ratites; the “enrich” set represents genes biased towards selection in ratites. All results calculated with the R package clusterProfiler.

Additional Data 2 (xls)

List of GO terms associated with ratite-accelerated and convergently accelerated CNEEs. The “crar” set represents convergently accelerated CNEEs, the “rar” set represents accelerated CNEEs (regardless of convergence), and the “crar_dollo” set represents convergently accelerated CNEEs conditional on acceleration in at least two independent losses of flight assuming the conservative three-loss scenario. See methods for details.

Additional Data 3 (xls)

List of the 54 candidate enhancers representing the intersection between ATAC-seq, CHiP-seq, and PhyloAcc results.

Additional Data 4 (xls)

List of all genes with at least one accelerated CNEE nearby. The ‘crar’ columns give the count of convergently accelerated CNEEs near each gene, and the empirical (permutation) q-value of enrichment. The ‘rar’ columns give the same but for accelerated CNEEs regardless of convergence. Genes with NA in the ‘crar’ columns have accelerated CNEEs but no convergently accelerated CNEEs nearby.

Additional Data 5 (xls)

List of regions of the genome (galGal4) enriched for either accelerated CNEEs (RAR enrich column) or convergently accelerated CNEEs (CRAR enrich column).

References

1. J. B. Losos, Convergence, adaptation, and constraint. *Evolution* **65**, 1827–1840 (2011). doi:10.1111/j.1558-5646.2011.01289.x Medline

2. A. Martin, V. Orgogozo, The Loci of repeated evolution: A catalog of genetic hotspots of phenotypic variation. *Evolution* **67**, 1235–1250 (2013). Medline

3. D. L. Stern, The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013). doi:10.1038/nrg3483 Medline

4. J. F. Storz, Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet.* **17**, 239–250 (2016). doi:10.1038/nrg.2016.11 Medline

5. Y. F. Chan, M. E. Marks, F. C. Jones, G. Villarreal Jr., M. D. Shapiro, S. D. Brady, A. M. Southwick, D. M. Absher, J. Grimwood, J. Schmutz, R. M. Myers, D. Petrov, B. Jónsson, D. Schluter, M. A. Bell, D. M. Kingsley, Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**, 302–305 (2010). doi:10.1126/science.1182213 Medline

6. Y. Zhen, M. L. Aardema, E. M. Medina, M. Schumer, P. Andolfatto, Parallel molecular evolution in an herbivore community. *Science* **337**, 1634–1637 (2012). doi:10.1126/science.1226630 Medline

7. K. L. Cooper, K. E. Sears, A. Uygur, J. Maier, K.-S. Baczkowski, M. Brosnahan, D. Antczak, J. A. Skidmore, C. J. Tabin, Patterning and post-patterning modes of evolutionary digit loss in mammals. *Nature* **511**, 41–45 (2014). doi:10.1038/nature13496 Medline

8. S. Yeaman, K. A. Hodgins, K. E. Lotterhos, H. Suren, S. Nadeau, J. C. Degner, K. A. Nurkowski, P. Smets, T. Wang, L. K. Gray, K. J. Liepe, A. Hamann, J. A. Holliday, M. C. Whitlock, L. H. Rieseberg, S. N. Aitken, Convergent local adaptation to climate in distantly related conifers. *Science* **353**, 1431–1433 (2016). doi:10.1126/science.aaf7812 Medline

9. R. Partha, B. K. Chauhan, Z. Ferreira, J. D. Robinson, K. Lathrop, K. K. Nischal, M. Chikina, N. L. Clark, Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife* **6**, e25884 (2017). doi:10.7554/eLife.25884 Medline

10. S. Xu, Z. He, Z. Guo, Z. Zhang, G. J. Wyckoff, A. Greenberg, C. I. Wu, S. Shi, Genome-Wide Convergence during Evolution of Mangroves from Woody Plants. *Mol. Biol. Evol.* **34**, 1008–1015 (2017). Medline

11. M. J. Berger, A. M. Wenger, H. Guturu, G. Bejerano, Independent erosion of conserved transcription factor binding sites points to shared hindlimb, vision and external testes loss in different mammals. *Nucleic Acids Res.* **46**, 9299–9308 (2018). doi:10.1093/nar/gky741 Medline

12. W. K. Meyer, J. Jamison, R. Richter, S. E. Woods, R. Partha, A. Kowalczyk, C. Kronk, M. Chikina, R. K. Bonde, D. E. Crocker, J. Gaspard, J. M. Lanyon, J. Marsillach, C. E. Furlong, N. L. Clark, Ancient convergent losses of *Paraoxonase 1* yield potential risks for modern marine mammals. *Science* **361**, 591–594 (2018). doi:10.1126/science.aap7714 Medline

13. V. Sharma, N. Hecker, J. G. Roscito, L. Foerster, B. E. Langer, M. Hiller, A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat. Commun.* **9**, 1215 (2018). doi:10.1038/s41467-018-03667-1 Medline

14. P. Houde, Ostrich ancestors found in the Northern Hemisphere suggest new hypothesis of ratite origins. *Nature* **324**, 563–565 (1986). doi:10.1038/324563a0 Medline

15. S. R. B. Bickley, M. P. O. Logan, Regulatory modulation of the T-box gene *Tbx5* links development, evolution, and adaptation of the sternum. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17917–17922 (2014). doi:10.1073/pnas.1409913111 Medline

16. S. J. Nesbitt, J. A. Clarke, The Anatomy and Taxonomy of the Exquisitely Preserved Green River Formation (Early Eocene) Lithornithids (Aves) and the Relationships of Lithornithidae. *Bull. Am. Mus. Nat. Hist.* **406**, 1–91 (2016). doi:10.1206/0003-0090-406.1.1

17. J. Cracraft, Phylogeny and evolution of the ratite birds. *Ibis* **116**, 494–521 (1974). doi:10.1111/j.1474-919X.1974.tb07648.x

18. J. Harshman, E. L. Braun, M. J. Braun, C. J. Huddleston, R. C. K. Bowie, J. L. Chojnowski, S. J. Hackett, K.-L. Han, R. T. Kimball, B. D. Marks, K. J. Miglia, W. S. Moore, S. Reddy, F. H. Sheldon, D. W. Steadman, S. J. Steppan, C. C. Witt, T. Yuri, Phylogenomic evidence for multiple losses of flight in ratite birds. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 13462–13467 (2008). doi:10.1073/pnas.0803242105 Medline

19. A. J. Baker, O. Haddrath, J. D. McPherson, A. Cloutier, Genomic support for a moa-tinamou clade and adaptive morphological convergence in flightless ratites. *Mol. Biol. Evol.* **31**, 1686–1696 (2014). doi:10.1093/molbev/msu153 Medline

20. K. J. Mitchell, B. Llamas, J. Soubrier, N. J. Rawlence, T. H. Worthy, J. Wood, M. S. Y. Lee, A. Cooper, Ancient DNA reveals elephant birds and kiwi are sister taxa and clarifies ratite bird evolution. *Science* **344**, 898–900 (2014). doi:10.1126/science.1251981 Medline

21. A. Grealy, M. Phillips, G. Miller, M. T. P. Gilbert, J.-M. Rouillard, D. Lambert, M. Bunce, J. Haile, Eggshell palaeogenomics: Palaeognath evolutionary history revealed through ancient nuclear and mitochondrial DNA from Madagascan elephant bird (Aepyornis sp.) eggshell. *Mol. Phylogenet. Evol.* **109**, 151–163 (2017). doi:10.1016/j.ympev.2017.01.005 Medline

22. T. Yonezawa, T. Segawa, H. Mori, P. F. Campos, Y. Hongoh, H. Endo, A. Akiyoshi, N. Kohno, S. Nishida, J. Wu, H. Jin, J. Adachi, H. Kishino, K. Kurokawa, Y. Nogi, H. Tanabe, H. Mukoyama, K. Yoshida, A. Rasoamiaramanana, S. Yamagishi, Y. Hayashi, A. Yoshida, H. Koike, F. Akishinonomiya, E. Willerslev, M. Hasegawa, Phylogenomics and Morphology of Extinct Paleognaths Reveal the Origin and Evolution of the Ratites. *Curr. Biol.* **27**, 68–77 (2017). doi:10.1016/j.cub.2016.10.029 Medline

23. See supplementary materials.

24. A. Cloutier *et al.*, First nuclear genome assembly of an extinct moa species, the little bush moa (Anomalopteryx didiformis). bioRxiv 262816 [preprint]. 11 February 2018.

25. E. D. Jarvis, S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, J. Li, F. Zhang, H. Li, L. Zhou, N. Narula, L. Liu, G. Ganapathy, B. Boussau, M. S. Bayzid, V. Zavidovych, S. Subramanian, T. Gabaldón, S. Capella-Gutiérrez, J. Huerta-Cepas, B. Rekepalli, K. Munch, M. Schierup, B. Lindow, W. C. Warren, D. Ray, R. E. Green, M. W. Bruford, X. Zhan, A. Dixon, S. Li, N. Li, Y. Huang, E. P. Derryberry, M. F. Bertelsen, F. H. Sheldon, R. T. Brumfield, C. V. Mello, P. V. Lovell, M. Wirthlin, M. P. C. Schneider, F. Prosdocimi, J. A. Samaniego, A. M. Vargas Velazquez, A. Alfaro-Núñez, P. F. Campos, B. Petersen, T. Sicheritz-Ponten, A. Pas, T. Bailey, P. Scofield, M. Bunce, D. M. Lambert, Q. Zhou, P. Perelman, A. C. Driskell, B. Shapiro, Z. Xiong, Y. Zeng, S. Liu, Z. Li, B. Liu, K. Wu, J. Xiao, X. Yinqi, Q. Zheng, Y. Zhang, H. Yang, J. Wang, L. Smeds, F. E. Rheindt, M. Braun, J. Fjeldsa, L. Orlando, F. K. Barker, K. A. Jønsson, W. Johnson, K.-P. Koepfli, S. O'Brien, D. Haussler, O. A. Ryder, C. Rahbek, E. Willerslev, G. R. Graves, T. C. Glenn, J. McCormack, D. Burt, H. Ellegren, P. Alström, S. V. Edwards, A. Stamatakis, D. P. Mindell, J. Cracraft, E. L. Braun, T. Warnow, W. Jun, M. T. P. Gilbert, G. Zhang, Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014). doi:10.1126/science.1253451 Medline

26. D. Le Duc, G. Renaud, A. Krishnan, M. S. Almén, L. Huynen, S. J. Prohaska, M. Ongyerth, B. D. Bitarello, H. B. Schiöth, M. Hofreiter, P. F. Stadler, K. Prüfer, D. Lambert, J. Kelso, T. Schöneberg, Kiwi genome provides insights into evolution of a nocturnal lifestyle. *Genome Biol.* **16**, 147 (2015). doi:10.1186/s13059-015-0711-4 Medline

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

28. S. V. Edwards, A. Cloutier, A. J. Baker, Conserved Nonexonic Elements: A Novel Class of Marker for Phylogenomics. *Syst. Biol.* **66**, 1028–1044 (2017). doi:10.1093/sysbio/syx058 Medline

29. M. J. Phillips, G. C. Gibb, E. A. Crimp, D. Penny, Tinamous and moa flock together: Mitochondrial genome sequence analysis reveals independent losses of flight among ratites. *Syst. Biol.* **59**, 90–107 (2010). doi:10.1093/sysbio/syp079 Medline

30. L. Liu, L. Yu, S. V. Edwards, A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* **10**, 302 (2010). doi:10.1186/1471-2148-10-302 Medline

31. A. Cloutier *et al.*, Whole-genome analyses resolve the phylogeny of flightless birds (Palaeognathae) in the presence of an empirical anomaly zone. bioRxiv 262949 [preprint]. 9 February 2018.

32. J. H. Degnan, N. A. Rosenberg, *Public Libr. Sci. Genet.* **2**, 762–768 (2006).

33. N. E. Wheeler, L. Barquist, R. A. Kingsley, P. P. Gardner, A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. *Bioinformatics* **32**, 3566–3574 (2016). Medline

34. A. Burga, W. Wang, E. Ben-David, P. C. Wolf, A. M. Ramey, C. Verdugo, K. Lyons, P. G. Parker, L. Kruglyak, A genetic signature of the evolution of loss of flight in the Galapagos cormorant. *Science* **356**, eaal3345 (2017). doi:10.1126/science.aal3345

35. M. Kolanczyk, N. Kossler, J. Kühnisch, L. Lavitas, S. Stricker, U. Wilkening, I. Manjubala, P. Fratzl, R. Spörle, B. G. Herrmann, L. F. Parada, U. Kornak, S. Mundlos, Multiple roles for neurofibromin in skeletal development and growth. *Hum. Mol. Genet.* **16**, 874–886 (2007). doi:10.1093/hmg/ddm032 Medline

36. N. Kossler, S. Stricker, C. Rödelsperger, P. N. Robinson, J. Kim, C. Dietrich, M. Osswald, J. Kühnisch, D. A. Stevenson, T. Braun, S. Mundlos, M. Kolanczyk, Neurofibromin (Nf1) is required for skeletal muscle development. *Hum. Mol. Genet.* **20**, 2697–2709 (2011). doi:10.1093/hmg/ddr149 Medline

37. S. B. Carroll, B. Prud'homme, N. Gompel, Regulating evolution: How gene switches make life. *Sci. Am.* **298**, 60–67 (May 2008). doi:10.1038/scientificamerican0508-60 Medline

38. A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, D. Haussler, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005). doi:10.1101/gr.3715005 Medline

39. R. Seki, C. Li, Q. Fang, S. Hayashi, S. Egawa, J. Hu, L. Xu, H. Pan, M. Kondo, T. Sato, H. Matsubara, N. Kamiyama, K. Kitajima, D. Saito, Y. Liu, M. T. P. Gilbert, Q. Zhou, X. Xu, T. Shiroishi, N. Irie, K. Tamura, G. Zhang, Functional roles of Aves class-specific cis-regulatory elements on macroevolution of bird-specific features. *Nat. Commun.* **8**, 14229 (2017). doi:10.1038/ncomms14229 Medline

40. L. A. Pennacchio, N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K. D. Lewis, I. Plajzer-Frick, J. Akiyama, S. De Val, V. Afzal, B. L. Black, O. Couronne, M. B. Eisen, A. Visel, E. M. Rubin, In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006). doi:10.1038/nature05295 Medline

41. A. Visel, S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis, A. Holt, I. Plajzer-Frick, V. Afzal, E. M. Rubin, L. A. Pennacchio, Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* **40**, 158–160 (2008). doi:10.1038/ng.2007.55 Medline

42. J. A. Capra, G. D. Erwin, G. McKinsey, J. L. R. Rubenstein, K. S. Pollard, Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. B* **368**, 20130025 (2013). doi:10.1098/rstb.2013.0025 Medline

43. D. E. Dickel, A. R. Ypsilanti, R. Pla, Y. Zhu, I. Barozzi, B. J. Mannion, Y. S. Khin, Y. Fukuda-Yuzawa, I. Plajzer-Frick, C. S. Pickle, E. A. Lee, A. N. Harrington, Q. T. Pham, T. H. Garvin, M. Kato, M. Osterwalder, J. A. Akiyama, V. Afzal, J. L. R. Rubenstein, L. A. Pennacchio, A. Visel, Ultraconserved Enhancers Are Required for Normal Development. *Cell* **172**, 491–499.e15 (2018). doi:10.1016/j.cell.2017.12.017 Medline

44. Z. Hu, T. B. Sackton, S. V. Edwards, J. S. Liu, Bayesian detection of convergent rate changes of conserved noncoding elements on phylogenetic trees. *Mol. Biol. Evol.* msz049 (2019). doi:10.1093/molbev/msz049

45. F. K. Mendes, M. W. Hahn, Gene Tree Discordance Causes Apparent Substitution Rate Variation. *Syst. Biol.* **65**, 711–721 (2016). doi:10.1093/sysbio/syw018 Medline

46. C. Rallis, B. G. Bruneau, J. Del Buono, C. E. Seidman, J. G. Seidman, S. Nissim, C. J. Tabin, M. P. Logan, Tbx5 is required for forelimb bud formation and continued outgrowth. *Development* **130**, 2741–2751 (2003). doi:10.1242/dev.00473 Medline

47. D. G. Ahn, M. J. Kourakis, L. A. Rohde, L. M. Silver, R. K. Ho, T-box gene tbx5 is essential for formation of the pectoral limb bud. *Nature* **417**, 754–758 (2002). doi:10.1038/nature00814 Medline

48. D. Li, R. Sakuma, N. A. Vakili, R. Mo, V. Puviindran, S. Deimling, X. Zhang, S. Hopyan, C. C. Hui, Formation of proximal and anterior limb skeleton requires early function of Irx3 and Irx5 and is negatively regulated by Shh signaling. *Dev. Cell* **29**, 233–240 (2014). doi:10.1016/j.devcel.2014.03.001 Medline

49. Y. Kida, Y. Maeda, T. Shiraishi, T. Suzuki, T. Ogura, Chick Dach1 interacts with the Smad complex and Sin3a to control AER formation and limb development along the proximodistal axis. *Development* **131**, 4179–4187 (2004). doi:10.1242/dev.01252 Medline

50. H. Peters, A. Neubüser, K. Kratochwil, R. Balling, Pax9-deficient mice lack pharyngeal pouch derivatives and teeth and exhibit craniofacial and limb abnormalities. *Genes Dev.* **12**, 2735–2747 (1998). doi:10.1101/gad.12.17.2735 Medline

51. K. S. Pollard, S. R. Salama, B. King, A. D. Kern, T. Dreszer, S. Katzman, A. Siepel, J. S. Pedersen, G. Bejerano, R. Baertsch, K. R. Rosenbloom, J. Kent, D. Haussler, Forces shaping the fastest evolving regions in the human genome. *PLOS Genet.* **2**, e168 (2006). doi:10.1371/journal.pgen.0020168 Medline

52. B. M. Booker, T. Friedrich, M. K. Mason, J. E. VanderMeer, J. Zhao, W. L. Eckalbar, M. Logan, N. Illing, K. S. Pollard, N. Ahituv, Bat Accelerated Regions Identify a Bat Forelimb Specific Enhancer in the HoxD Locus. *PLOS Genet.* **12**, e1005738 (2016). doi:10.1371/journal.pgen.1005738 Medline

53. D. Kostka, A. K. Holloway, K. S. Pollard, Developmental Loci Harbor Clusters of Accelerated Regions That Evolved Independently in Ape Lineages. *Mol. Biol. Evol.* **35**, 2034–2045 (2018). doi:10.1093/molbev/msy109 Medline

54. K. T. Xie, G. Wang, A. C. Thompson, J. I. Wucherpfennig, T. E. Reimchen, A. D. C. MacColl, D. Schluter, M. A. Bell, K. M. Vasquez, D. M. Kingsley, DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science* **363**, 81–84 (2019). doi:10.1126/science.aan1425 Medline

55. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013). doi:10.1038/nmeth.2688 Medline

56. A. R. Gehrke, I. Schneider, E. de la Calle-Mustienes, J. J. Tena, C. Gomez-Marin, M. Chandran, T. Nakamura, I. Braasch, J. H. Postlethwait, J. L. Gómez-Skarmeta, N. H. Shubin, Deep conservation of wrist and digit enhancers in fish. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 803–808 (2015). doi:10.1073/pnas.1420208112 Medline

57. H. Zhang, C.-Y. Liu, Z.-Y. Zha, B. Zhao, J. Yao, S. Zhao, Y. Xiong, Q.-Y. Lei, K.-L. Guan, TEAD transcription factors mediate the function of TAZ in cell growth and epithelial-mesenchymal transition. *J. Biol. Chem.* **284**, 13355–13362 (2009). doi:10.1074/jbc.M900843200 Medline

58. B. Zhao, X. Ye, J. Yu, L. Li, W. Li, S. Li, J. Yu, J. D. Lin, C.-Y. Wang, A. M. Chinnaiyan, Z.-C. Lai, K.-L. Guan, TEAD mediates YAP-dependent gene induction and growth control. *Genes Dev.* **22**, 1962–1971 (2008). doi:10.1101/gad.1664408 Medline

59. A. Sawada, H. Kiyonari, K. Ukita, N. Nishioka, Y. Imuta, H. Sasaki, Redundant roles of Tead1 and Tead2 in notochord development and the regulation of cell proliferation and survival. *Mol. Cell. Biol.* **28**, 3177–3189 (2008). doi:10.1128/MCB.01759-07 Medline

60. S. Lamichhaney *et al*., Integrating natural history-derived phenomics with comparative genomics to study the genetic architecture of convergent evolution. bioRxiv 574756 [preprint]. 12 March 2019.

61. P. Grayson, S. Y. W. Sin, T. B. Sackton, S. V. Edwards, in *Avian and Reptilian Developmental Biology* (Humana, 2017); https://link.springer.com/protocol/10.1007/978-1-4939-7216-6_2), *Methods in Molecular Biology*, pp. 11–46.

62. J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, D. B. Jaffe, ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res.* **18**, 810–820 (2008). doi:10.1101/gr.7337908 Medline

63. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014). doi:10.1093/bioinformatics/btu170 Medline

64. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015). doi:10.1093/bioinformatics/btv351 Medline

65. E. Hagelberg, in *Ancient DNA* (Springer, 1994), pp. 195–204; https://link.springer.com/chapter/10.1007/978-1-4612-4318-2_13.

66. A. Cooper, H. N. Poinar, Ancient DNA: Do it right or not at all. *Science* **289**, 1139 (2000). doi:10.1126/science.289.5482.1139b Medline

67. M. Knapp, A. C. Clarke, K. A. Horsburgh, E. A. Matisoo-Smith, Setting the stage - building and working in an ancient DNA laboratory. *Ann. Anat.* **194**, 3–6 (2012). doi:10.1016/j.aanat.2011.03.008 Medline

68. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline

69. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi:10.1038/nmeth.1923 Medline

70. H. Jónsson, A. Ginolhac, M. Schubert, P. L. F. Johnson, L. Orlando, mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013). doi:10.1093/bioinformatics/btt193 Medline

71. J. Zhang, K. Kobert, T. Flouri, A. Stamatakis, PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014). doi:10.1093/bioinformatics/btt593 Medline

72. C. Holt, M. Yandell, MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011). doi:10.1186/1471-2105-12-491 Medline

73. B. Vicoso, V. B. Kaiser, D. Bachtrog, Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6453–6458 (2013). doi:10.1073/pnas.1217027110 Medline

74. B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013). doi:10.1038/nprot.2013.084 Medline

75. D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg, TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013). doi:10.1186/gb-2013-14-4-r36 Medline

76. S. R. Eddy, Accelerated Profile HMM Searches. *PLOS Comput. Biol.* **7**, e1002195 (2011). doi:10.1371/journal.pcbi.1002195 Medline

77. S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. von Mering, P. Bork, eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40**, D284–D289 (2012). doi:10.1093/nar/gkr1060 Medline

78. A. M. Altenhoff, M. Gil, G. H. Gonnet, C. Dessimoz, Inferring hierarchical orthologous groups from orthologous gene pairs. *PLOS ONE* **8**, e53786 (2013). doi:10.1371/journal.pone.0053786 Medline

79. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi:10.1093/molbev/mst010 Medline

80. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010). doi:10.1038/nbt.1621 Medline

81. A. Löytynoja, Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* **1079**, 155–170 (2014). doi:10.1007/978-1-62703-646-7_10 Medline

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

83. B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, D. Haussler, Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011). doi:10.1101/gr.123356.111 Medline

84. R. E. Green, E. L. Braun, J. Armstrong, D. Earl, N. Nguyen, G. Hickey, M. W. Vandewege, J. A. St. John, S. Capella-Gutiérrez, T. A. Castoe, C. Kern, M. K. Fujita, J. C. Opazo, J. Jurka, K. K. Kojima, J. Caballero, R. M. Hubley, A. F. Smit, R. N. Platt, C. A. Lavoie, M. P. Ramakodi, J. W. Finger Jr., A. Suh, S. R. Isberg, L. Miles, A. Y. Chong, W. Jaratlerdsiri, J. Gongora, C. Moran, A. Iriarte, J. McCormack, S. C. Burgess, S. V. Edwards, E. Lyons, C. Williams, M. Breen, J. T. Howard, C. R. Gresham, D. G. Peterson, J. Schmitz, D. D. Pollock, D. Haussler, E. W. Triplett, G. Zhang, N. Irie, E. D. Jarvis, C. A. Brochu, C. J. Schmidt, F. M. McCarthy, B. C. Faircloth, F. G. Hoffmann, T. C. Glenn, T. Gabaldón, B. Paten, D. A. Ray, Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* **346**, 1254449 (2014). doi:10.1126/science.1254449 Medline

85. M. J. Hubisz, K. S. Pollard, A. Siepel, PHAST and RPHAST: Phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011). doi:10.1093/bib/bbq072 Medline

86. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi:10.1093/bioinformatics/btq033 Medline

87. G. Hickey, B. Paten, D. Earl, D. Zerbino, D. Haussler, HAL: A hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013). doi:10.1093/bioinformatics/btt128 Medline

88. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009). doi:10.1093/bioinformatics/btp348 Medline

89. A. M. Kozlov, A. J. Aberer, A. Stamatakis, ExaML version 3: A tool for phylogenomic analyses on supercomputers. *Bioinformatics* **31**, 2577–2579 (2015). doi:10.1093/bioinformatics/btv184 Medline

90. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016). doi:10.1093/molbev/msw046 Medline

91. D. L. Swofford, *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) 4.0.b5* (2001).

92. W. P. Maddison, D. R. Maddison, *Mesquite: A Modular System for Evolutionary Analysis* (2018); www.mesquiteproject.org.

93. J. M. Beaulieu, B. C. O'Meara, Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction. *Syst. Biol.* **65**, 583–601 (2016). doi:10.1093/sysbio/syw022 Medline

94. E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2018). doi:10.1093/bioinformatics/bty633

95. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007). doi:10.1093/molbev/msm088 Medline

96. S. L. K. Pond, S. D. W. Frost, S. V. Muse, HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679 (2005). doi:10.1093/bioinformatics/bti079 Medline

97. S. L. Kosakovsky Pond, B. Murrell, M. Fourment, S. D. W. Frost, W. Delport, K. Scheffler, A random effects branch-site model for detecting episodic diversifying selection. *Mol. Biol. Evol.* **28**, 3033–3043 (2011). doi:10.1093/molbev/msr125 Medline

98. M. D. Smith, J. O. Wertheim, S. Weaver, B. Murrell, K. Scheffler, S. L. Kosakovsky Pond, Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32**, 1342–1353 (2015). doi:10.1093/molbev/msv022 Medline

99. J. O. Wertheim, B. Murrell, M. D. Smith, S. L. Kosakovsky Pond, K. Scheffler, RELAX: Detecting relaxed selection in a phylogenetic framework. *Mol. Biol. Evol.* **32**, 820–832 (2015). doi:10.1093/molbev/msu400 Medline

100. A. Kowalczyk *et al.*, RERconverge: an R package for associating evolutionary rates with convergent traits. bioRxiv 451138 [preprint]. 23 October 2018.

101. R. Partha, A. Kowalczyk, N. L. Clark, M. Chikina, Robust methods for detecting convergent shifts in evolutionary rates. bioRxiv 457309 [preprint]. 30 October 2018.

102. M. Chikina, J. D. Robinson, N. L. Clark, Hundreds of Genes Experienced Convergent Shifts in Selective Pressure in Marine Mammals. *Mol. Biol. Evol.* **33**, 2182–2192 (2016). doi:10.1093/molbev/msw112 Medline

103. G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omics J. Integr. Biol.* **16**, 284–287 (2012). doi:10.1089/omi.2011.0118

104. D. Orme, *The Caper Package: Comparative Analysis of Phylogenetics and Evolution in R* (2018); https://cran.r-project.org/web/packages/caper/vignettes/caper.pdf.

105. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010). doi:10.1101/gr.097857.109 Medline

106. A. Heger, C. Webber, M. Goodson, C. P. Ponting, G. Lunter, GAT: A simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**, 2046–2048 (2013). doi:10.1093/bioinformatics/btt343 Medline

107. E. Calo, J. Wysocka, Modification of enhancer chromatin: What, how, and why? *Mol. Cell* **49**, 825–837 (2013). doi:10.1016/j.molcel.2013.01.038 Medline

108. H. Ochi, T. Tamai, H. Nagano, A. Kawaguchi, N. Sudou, H. Ogino, Evolution of a tissue-specific silencer underlies divergence in the expression of pax2 and pax8 paralogues. *Nat. Commun.* **3**, 848 (2012). doi:10.1038/ncomms1851 Medline

109. K. Onimaru, S. Kuraku, W. Takagi, S. Hyodo, J. Sharpe, M. Tanaka, A shift in anterior-posterior positional information underlies the fin-to-limb evolution. *eLife* **4**, e07048 (2015). doi:10.7554/eLife.07048

110. D. G. Gibson, L. Young, R.-Y. Chuang, J. C. Venter, C. A. Hutchison 3rd, H. O. Smith, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009). doi:10.1038/nmeth.1318 Medline

111. T. Suzuki, T. Ogura, Congenic method in the chick limb buds by electroporation. *Dev. Growth Differ.* **50**, 459–465 (2008). doi:10.1111/j.1440-169X.2008.01054.x Medline

112. F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLOS ONE* **6**, e21800 (2011). doi:10.1371/journal.pone.0021800 Medline

113. G. Zhang, C. Li, Q. Li, B. Li, D. M. Larkin, C. Lee, J. F. Storz, A. Antunes, M. J. Greenwold, R. W. Meredith, A. Ödeen, J. Cui, Q. Zhou, L. Xu, H. Pan, Z. Wang, L. Jin, P. Zhang, H. Hu, W. Yang, J. Hu, J. Xiao, Z. Yang, Y. Liu, Q. Xie, H. Yu, J. Lian, P. Wen, F. Zhang, H. Li, Y. Zeng, Z. Xiong, S. Liu, L. Zhou, Z. Huang, N. An, J. Wang, Q. Zheng, Y. Xiong, G. Wang, B. Wang, J. Wang, Y. Fan, R. R. da Fonseca, A. Alfaro-Núñez, M. Schubert, L. Orlando, T. Mourier, J. T. Howard, G. Ganapathy, A. Pfenning, O. Whitney, M. V. Rivas, E. Hara, J. Smith, M. Farré, J. Narayan, G. Slavov, M. N. Romanov, R. Borges, J. P. Machado, I. Khan, M. S. Springer, J. Gatesy, F. G. Hoffmann, J. C. Opazo, O. Håstad, R. H. Sawyer, H. Kim, K.-W. Kim, H. J. Kim, S. Cho, N. Li, Y. Huang, M. W. Bruford, X. Zhan, A. Dixon, M. F. Bertelsen, E. Derryberry, W. Warren, R. K. Wilson, S. Li, D. A. Ray, R. E. Green, S. J. O'Brien, D. Griffin, W. E. Johnson, D. Haussler, O. A. Ryder, E. Willerslev, G. R. Graves, P. Alström, J. Fjeldså, D. P. Mindell, S. V. Edwards, E. L. Braun, C. Rahbek, D. W. Burt, P. Houde, Y. Zhang, H. Yang, J. Wang, Avian Genome Consortium, E. D. Jarvis, M. T. P. Gilbert, J. Wang, Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014). doi:10.1126/science.1251385 Medline