**Supplementary Information**

**Rules of amino acid convergence: Not how many, but who in avian vocal learning clades**

Lee et al.

**Table of contents**

[**Supplementary Notes: Methods** 3](#_Toc54879065)

[**Supplementary Note 1: Data description and CSAV analysis** 3](#_Toc54879066)

[**Supplementary Note 2: Control species set designs** 5](#_Toc54879067)

[**Supplementary Note 3: Correlations between identical and different amino acid convergences** 5](#_Toc54879068)

[**Supplementary Note 4: Phylogenetic features related to the number of molecular convergences** 7](#_Toc54879069)

[**Supplementary Note 5: Convergent single codon and single nucleotide variants** 8](#_Toc54879070)

[**Supplementary Note 6: Correlation tests among molecular convergences and phylogenetic features** 8](#_Toc54879071)

[**Supplementary Note 7: PCA and ML tree analyses with rifleman** 9](#_Toc54879072)

[**Supplementary Note 8: Gene ontology functional annotations and gene network analyses** 9](#_Toc54879073)

[**Supplementary Note 9: Fixed differences of AVL-CSAVs within populations of zebra finch and chicken** 10](#_Toc54879074)

[**Supplementary Note 10: Positive selection calculations** 11](#_Toc54879075)

[**Supplementary Note 11: Singing regulated and specialized expressed genes in song learning nuclei** 12](#_Toc54879076)

[**Supplementary Note 12: Human specific amino acid substitutions** 14](#_Toc54879077)

[**Supplementary Note 14: Institutional Review for animal cares and experiments** 15](#_Toc54879078)

[**Supplementary Tables** 16](#_Toc54879079)

[**Supplementary Figures** 19](#_Toc54879080)

[**Supplementary Data** 37](#_Toc54879081)

[**Supplementary References** 38](#_Toc54879082)

# **Supplementary Notes: Methods**

# **Supplementary Note 1: Data description and CSAV analysis**

In our preliminary studies, the Avian Phylogenomics Project (now the Bird10K project) defined 8,295 singleton orthologous gene sets across 48 avian species, and constructed the phylogenetic avian family tree consisting of 34 orders1-3 (**Fig. 1a, Supplementary Fig. 1**). This 1:1 orthologous gene set was identified by reciprocal best blast hits and synteny, using two species as a reference: chicken and zebra finch3. They were then aligned across all species using SATé+MAFFT and SATé+Prank3, for both nucleotide and amino acid sequences. Alignment frameshift errors were corrected when translating into amino acid sequence alignments. In our previous analyses of convergent amino acid substitutions2, we used Gblocks4 to remove poorly scored alignments with sequence divergences and columns with gaps in at least one species included. However, here we found that this was too aggressive, removing 65% of the whole regions of aligned sequences. For example, vocal learner-specific amino acid substitutions of *DRD1B* was excluded because of gaps in one of outgroup species (Lizard) (**Supplementary Fig. 2**). Therefore, we used whole regions of alignments without the trimming step in the current study (**Supplementary Table 1**). The peptide and codon-wise nucleotide sequence alignments are accessible at the following link (https://github.com/chulbioinfo/CSAVanalysis)

We initially developed an algorithm to find convergent amino acid substitutions specific to a group of species, called Target-specific Amino Acid Substitution (TAAS) analysis2. It also detected insertion/deletions specific to a group of species. In this study, we changed the analysis name intuitively from TAAS to convergent single amino acid variants (CSAV) analysis (**Fig. 1b**). The CSAV analysis focuses on identifying amino acid sites with convergent variants specific to multi-species from polyphyletic lineages (**Supplementary Fig. 3**), while TAAS analysis can identify both convergent and apomorphic amino acid variants from polyphyletic clades and a monophyletic clades, respectively. We considered convergent variants as consisting of either identical (convergent and parallel) and different (divergent) amino acid substitutions at the same site among independent lineages; these amino acid substitutions share mutually exclusive variants between a target group of species relative to all other species tested (**Supplementary Fig. 3**).

We improved the CSAV analysis with 2 major updates: determined statistical significance of the CSAVs, and identify identical and different convergent substitutions based on entropy scores across sites in the alignment. To estimate the statistical differences of CSAVs, we employed Fisher’s exact test. We generated 2x2 contingency tables composed of two factors such as phenotypic information (vocal learner versus vocal non-learner) and genetic information (amino acids observed in vocal learners versus vocal non-learners) of each species in each homologous site. By using these 2x2 contingency tables, we performed Fisher’s exact test with the alternative hypothesis that the odds ratio is greater than 1. This statistical test was performed k times on each amino acid position in the targeted gene (k = length of multiple sequence alignment of a gene). Therefore, we adjusted for multiple testing by Bonferroni correction on each gene (**Supplementary Data 1**).

To classify CSAV into identical and different variants, we calculated Shannon’s entropy score (H)5 as:

Hj = - (1)

where Pij is the observed frequency of amino acid i at site j.

Identical convergent amino acid variants (iCSAV: Types 1 and 2) had entropy scores = 0; different convergent amino acid variants (dCSAV: Types 3 and 4) had entropy scores > 0 (**Supplementary Data 1**). The source code and peptide sequence alignments for CSAV analysis are accessible at the following link (https://github.com/chulbioinfo/CSAVanalysis). Examples of vocal learner specific convergent amino acid substitutions were summarized and visualized by using WebLogo (v2.8.2)6,7 (**Fig. 1c**).

# **Supplementary Note 2: Control species set designs**

Considering that we have 6 vocal learning species we calculated all 6 species combinations of 47 birds in the avian family tree excluding Rifleman, which was 10,737,573 combinations (**Supplementary Table 2a**). Of these, 8,281 combinations of 6 species originated from 3 independent lineages like 3 vocal learning clades (songbirds, parrots, hummingbirds). From these combinations, we designed 2 main types of control sets (**Supplementary Table 2a**): Random control sets consisting of 1,000 random species combinations from the 8,281 set of 6 species with 3 independent origins; and core control sets consisting of 59 possible convergent combinations of species that have a similar phylogenetic history to vocal learners, but contained 6 species originated from 2 clades out of 3 vocal learning clades and 1 vocal non-learning clade. Additionally, we include 2 control sets with target species sharing other convergent traits among birds: 4 origins for 6 species of birds of prey and 4 origins for 15 species of water birds in the tree (**Supplementary Fig. 1**).

# **Supplementary Note 3: Correlations between identical and different amino acid convergences**

Castoe et al8 was the first to report that convergent (our identical) substitutions are proportional to divergent (our different) substitutions, in pairs of species, based on 34 squamate reptile species plus 6 tetrapod species, in pairs of species analyses. Thomas et al9 uncovered that in pairs of species in 9 species of mammals, the number of convergent substitutions was also correlated to the number of divergent substitutions, and echolocating mammals (bat and dolphin) did not exceed the expected proportion of convergent substitutions. Here we replaced the ambiguous terminology of *convergent and divergent convergent substitutions* with *identical convergent (iCSAVs) and different convergent (dCSAV) single amino acid variants* (**Supplementary Fig. 3**), and developed a CSAV script that identifies whether iCSAVs are proportional to dCSAVs beyond pairs in multiple combinations of species (<https://github.com/chulbioinfo/CSAVanalysis>).

To check statistical significances of correlations between iCSAV and dCSAV, we calculated Spearman rank correlation coefficient as:

where and are each rank of x and y, respectively; and and correspond to the means of rank(x) and rank(y), respectively. By using ‘cor.test’ function with the option method = “spearman” in R package (ver. 3.5.1), we tested correlations between iCSAVs and dCSAVs in the multiple combinations of species (e.g. avian vocal learners, 1,000 random control sets, and 61 core control sets). After then, we performed linear regression analysis for modeling the relationship between iCSAVs and dCSAVs based on ‘lm’ function, and visualized it with ‘plot’, ‘points’, and ‘abline’ function in R package (ver. 3.5.1)10. We also performed Bonferroni Outlier Test to check whether the number of convergent substitutions of vocal learners or other species combinations is an outlier, as determined by residuals from regression model with the ‘outlierTest’ function in R package (ver. 3.5.1) 10,11; option for limitation of the max number of outliers as 3: ‘n.max=3’. The source code and dataset to perform above analyses are accessible at the following link (https://github.com/chulbioinfo/CSAVanalysis).

# **Supplementary Note 4: Phylogenetic features related to the number of molecular convergences**

We performed multiple clade-wise comparisons of at least 3 polyphyletic clades to find relationships between convergent variants and various phylogenetic features. Using the branch lengths of the avian total evidence phylogenetic tree from Jarvis et al1 (**Fig. 1a, Supplementary Fig. 1**), we calculated four types of phylogenetic branch measures for convergent groups of species: product of origin branch lengths (POB), product of terminal branch lengths (PTB), distance between terminal branches (DTB), and distance between terminal nodes (DTN) (**Fig. 2a-d**). POB was calculated by multiplying lengths of most recent common ancestral (MRCA, origin) branches of each target clade and PTB as branch lengths of terminal taxa. DTB was calculated as a summation adding lengths of all branches between the MRCA node of the 47 birds and each terminal taxon, whereas the DTN was calculated as the summation between the MRCA node and the most recent ancestral nodes of each terminal taxon (**Fig. 2a-d**). The source code to calculate each phylogenetic feature is accessible at the following link (Https://github.com/chulbioinfo/CSAVanalysis).

We tested whether the convergent variants (CSAV and the underlying iCSAV and dCSAV subsets) are proportional to 4 types of phylogenetic features of avian vocal learning set and control sets or not by applying Spearman rank sum test and linear regression analysis (Note 3). We also performed Bonferroni Outlier Test to check whether the number of convergent substitutions of vocal learners is an outlier detected by residuals from regression model or not with ‘outlierTest’ function in R package (ver. 3.5.1) 10,11; with the option for limitation of the max number of outliers as 3: ‘n.max=3’. The source code to perform correlation tests, regression analyses, and outlier test for 3 types of amino acid convergences and 4 phylogenetic features is accessible at the following link (Https://github.com/chulbioinfo/CSAVanalysis).

# **Supplementary Note 5: Convergent single codon and single nucleotide variants**

In order to check fundamental variants of amino acid convergences, we modified the CSAV algorithm (**Supplementary Note 2**) to detect convergent single codon variants (CSCV) and convergent single nucleotide variants (CSNV). Instead of searching for amino acid variants (among 20 of them) in the protein-coding sequence alignments, the CSCV algorithm was changed to search codon and the CSNV for single nucleotide variants mutually exclusive between two group of species in the codon-wise nucleotide sequence alignments. After identifying 3 types (all, identical, and different) of convergent variants at three levels (amino acids, codons, and nucleotides) based on CSAV, CSCV, and CSNV analyses, we manually checked overlaps among those variants. We counted the number of variants by considering all possible relations among CSAVs and CSNVs in CSCVs, and drew Vann diagrams of each type of species sets (1 avian vocal learning set, 1,000 random control sets, and 61 core control sets). The intuitive examples of CSAVs arising from CSNVs at a single site and nonsynonymous complex multiple nucleotide variants (CMNV) were selected within convergent sites with identical CSAVs specific to vocal learners.

# **Supplementary Note 6: Correlation tests among molecular convergences and phylogenetic features**

We performed correlation tests, regression analyses, and outlier tests for the number of 3 types of convergent variants at 3 levels and 4 phylogenetic features and visualized their relationships in random control sets and in core control sets. Detail of methods are described in **Supplementary Note 3**. The source code and the data matrix are accessible at the following link (https://github.com/chulbioinfo/CSAVanalysis).

# **Supplementary Note 7: PCA and ML tree analyses with rifleman**

With the CSAV sites found in vocal learners, Rifleman was added and principle component analysis (PCA) was performed using the method as implemented in JalView12. Focusing on the AVL-CSAV and AVL-iCSAV sites, pairwise scores between bird species was computed by summing the substitution scores from BLOSUM62. Then, we performed spectral decomposition of the score matrix to obtain principal component (PC) vector and eigenvalue of the respective vectors. Sorting the PCs in the descending order of eigen values, we defined the first two vectors as PC1 and PC2. The PCA biplot was computed using these two vectors. For the maximum likelihood (ML) tree, we constructed it using MEGA13, and selected the JTT model, on the part of the amino acid sequence alignment of all AVL-CSAV sites or AVL-iCSAV sites.

# **Supplementary Note 8: Gene ontology functional annotations and gene network analyses**

To investigate if there were enriched functions of genes with convergent amino acid substitutions in the vocal learning set and control sets, we summarized 9,513 lists of genes with convergent variants considering combinations of 3 types (all, identical, and different variants) at 3 levels (amino acid, codon, and nucleotide levels) specific to each set (n=1,057). We conducted Gene Ontology (GO) analysis by using g:Profiler (v 0.3.5.)14 with the default option. and ClueGO (ver. 2.3.3.)15 in Cytoscape16 with the following options: GO BiologicalProcess-GOA (released in 08.04.2016); all of GO tree interval; all of GO Term/Pathway selection; multiple testing correction by Bonferroni (adjusted p-value < 0.05); and default options of others. After then, we tested whether the number of genes is correlated with the number of significant GO terms and the significances of GO terms, by applying regression analyses using ‘lm’ function. We visualized the results with ‘plot’, ‘points’, and ‘abline’ functions in the R package (ver. 3.5.1)10.

After then, focusing on 2 lists enriched for learning process: AVL-iCSAV gene list and a Ctrl-dCSCV set (different codon convergences specific to Dalmatian pelican, little egret, houbara bustard, red-crested turaco, white-throated tinamou, and ostrich), we searched cellular positions of convergent genes related to learning and analyzed protein-protein interactions among convergent genes by using CluePedia ver. 1.3.3.17 in Cytoscape16, selecting the following databases: STRING-ACTIONS\_v10.0 (released in 07.05.2015); activation v10.0; binding v10.0; catalysis v10.0; expression v10.0; inhibition v10.0; ptmod v10.0, and reaction v10.0. Sequences of the convergent variants of 2 gene lists associated with learning were summarized and visualized by WebLogo (v2.8.2)6,7.

# **Supplementary Note 9: Fixed differences of AVL-CSAVs within populations of zebra finch and chicken**

CSAV analysis was performed with the assumption that a haploid sequences identified are representative of the species. However, variation is also prevalent within a species. More than 20 million (20,739,045) and 1.6 million (1,661,545) variants have been reported in chicken (n=9,586) and zebra finch (n=1,257), respectively, according to Ensembl database release 8418,19. Hence, we performed additional analysis to check if the AVL-CSAV sequences we identified not due to within species variation. Local alignment was conducted for the CDS sequences containing AVL-CSAVs using BLAST (ver. 2.8.1)20 to find the position of CSAV on the chromosome sequence of chicken (Galgal4) and zebra finch (taeGut3.2.4) according to Ensembl database release 8418. Fixation of sequence in a species was assessed by comparing the chromosomic position of all AVL-CSAVs with the polymorphism data of chicken and zebra finch obtained from Ensembl dbSNP build 145 and 139 of chicken and zebra finch, respectively21. AVL-CSAVs overlapping with polymorphism was considered polymorphic.

We also performed additional fixation analyses on several genes amplified by PCR from red blood cells in blood of zebra finch (n = 3 males and 3 females) and chicken (n = 3 males and 3 females). The *D1B* gene was cloned from genomic DNA by using zebra finch specific primers (forward 5’-GCC CTG CGT CAG TGA GAC CA-3’ and reverse 5’-CCG CCA GCC CCC TGT ATG AC-3’) and white-leghorn chicken specific primers (forward 5’-CAG ATC TCC CCC GAC CCC GA-3’ and reverse 5’-GGC AAC AAT GCC GCC TGG AG-3’). The PCR reaction was conducted a total volume of 20 ul containing 100 ng genomic DNA, 10x PCR buffer, 0.4 μl dNTP (10 mM each), 10 pmol of each primer, and 0.5 U Taq polymerase (BioFACT) in the following thermocycling conditions: 2 min at 95°C, followed by 35 cycles of 20 s at 95°C, 40 s at 60°C, 2 min at 72°C, and, finally, 5 min at 72°C. The PCR products were cloned into the pGEM-T easy vector (Promega) and sequenced using an ABI Prism 3730 XL DNA Analyzer (Thermo Fisher–Applied Bio- systems).

# **Supplementary Note 10: Positive selection calculations**

The *dN* (the rate of non-synonymous substitution), *dS* (the rate of synonymous substitution) and ω = *dN/dS* were estimated along each branch of the phylogenetic tree and across sites by using the branch-site model A, implemented in codeml within PAML ver. 4.622 with F3X4 codon frequencies. We assumed the vocal learning trait in birds was originated from the most recent common ancestral branches of each vocal learning clade (**Fig. S10**). Log likelihood ratio test (LRT, *D* value) was performed to compare the null hypothesis with a fixed ω (model 2) and an alternative hypothesis with an estimated ω (model 2). Orthologs with ω2 Foreground > 1 and number of accelerated sites (BEB > 0.5) > 0 were retained (branches tested for positive selection are referred to as “foreground” branches and all other are referred to as “background” branches).

Out of 8,295 orthologous gene sets of 47 birds excluding Rifleman, we focused on 2 gene lists with convergent variants (CSAVs) specific to avian vocal learners and the closest control set to determine adaptive evolution of those genotypes. The data set of coding sequences of each gene list, including alignment gaps in species, was analyzed with a codeml option (cleandata = 0) and robust cutoff of adjusted p-value (<0.05; FDR). False discovery rates were calculated in R (ver.3.0.1) (**Supplementary Data 2**).

# **Supplementary Note 11: Singing regulated and specialized expressed genes in song learning nuclei**

We obtained and analyzed 8 gene expression profiles that overlapped with those among the 8,295 orthologous gene set:

1) SRG\_2014: A data set of 1,108 singing-regulated genes in zebra finch by using microarray approaches from Whiteney et al.23.

2) DEG\_2014: A data set of 1,849 differentially expressed genes between song nuclei (RA, HVC, LMAN, and Area X) from Whiteney et al.23 and Pfenning et al.23, were we selected those that had expression in one nucleus different from all others (NUC VS other NUCs) .

3) DEG\_2019: A data set of 1,148 differentially expressed genes between a song nucleus relative to its surrounding brain region (NUC VS SUR) that were obtained using the micro-dissected method (Gedman et al in preparation).

4) DEG\_2020: A data set of differentially expressed genes obtained by the laser capture microscope (LCM) (Gedman et al in preparation) in 5 different comparisons: (a) 2,065 differentially expressed genes among four song nuclei (RA, HVC, LMAN, and Area X) relative to the surrounding brain regions (NUC VS SUR), (b) 4,148 differentially expressed genes between a song nuclei relative to another song nuclei (NUC VS NUC), (c) 3,308 differentially expressed genes between a surrounding region of a song nucleus relative to another surrounding region (SUR VS SUR), (d) 1,942 differentially expressed genes among a song nucleus relative to the other song nuclei (NUC VS other NUCs), and (e) 1,388 differentially expressed genes among a surrounding region of a song nucleus relative to the other surrounding regions (SUR VS other SURs).

In brief, for specialized gene sets 3 and 4, tissue samples were collected from 4 adult male zebra finches that were kept in the dark for at least 2 hours to limit singing behavior and movement to ensure no immediate early gene activity in the song system or surrounding brain regions, respectively. Each brain was extracted, bisected along the midline, and frozen in TissueTek block mold on dry ice, in <2-5 minutes to ensure high RNA integrity. For microdissected samples, brain regions were visualized under a brightfield dissecting microscope with small scissors and forceps. For LCM, one hemisphere/bird was sectioned on a cryostat at 12uM and mounted on PEN membrane slides. Sections on the slides were dehydrated visualized under an LCM microscope, and specific song nuclei and their adjacent non-vocal motor control regions laser dissected. For both microdissected and LCM samples, RNA was isolated from each sample using the Picopure RNA Isolation kit, and stored at -80oC until all samples were collected. Samples were then randomized across batches to minimize batch effects, and cDNA was generated using the UltraLow-input RNAseq kit from Clonetech. Each library was prepped and indexed for sequencing using the NEB Next-flex library prep kit. Sequencing was conducted on the Nextseq 500 system from Illumina.

Quality of all raw sequence reads were verified using fastqc, trimming off low-quality (<30) and adapter sequences using fastq-mcf. Reads were mapped using STAR (v=2.7.2b) and counted using featureCounts (v=2.0.0). Final gene x sample matrix was used as input for DESeq2 for differential expression analysis. Each nucleus-surround pair had a linear model with one variable (~ spec) where “spec” was either “center” (vocal motor nucleus) or “surr” (non-vocal motor surround). Genes were considered differentially expressed (increased or decreased in song nuclei versus surround) if they passed multiple test corrections (q < 0.05) (**Supplementary Data 2**).

# **Supplementary Note 12: Human specific amino acid substitutions**

To test if genes with convergent avian vocal learner substitutions also have human specific substitutions different from vocal non-learning primates, we downloaded CDS sequences of 18,565 orthologous genes with of 24 primates from Ensembl (release 94)24, and performed multiple sequence alignments of each orthologous gene set by using PRANK 25 with ‘-codon’ and ‘-F’ for the codon-wise recognition and trusting the inference of insertions 26, respectively **(Supplementary Table 3**). We modified CSAV analysis into single amino acid variants (SAV) analysis to detect species-specific amino acid variants (<https://github.com/chulbioinfo/CSAVanalysis>). For 18,565 orthologous genes, we applied it to find human-specific sites with mutually exclusive amino acid substitutions between human and the other 23 primates. We then tested if any fell within the 141 genes with convergent avian vocal learner substitutions (**Supplementary Data 3**).

To validate the conservation of the famous 2 human-specific substitutions in *FOXP2* gene, we checked the p.Thr303Asn and p.Asn325Ser amino acid substitutions are overlapped with human variants in accessible data sources in UCSC genome browser (common dbSNPs: 153, 151, 150, 147, 146, 144, 142, and 141, All SNPs: 142 and 141, Flagged SNPs: 151, 150, 147, 146, 144, 142, and 141, Multi. SNPs: 151, 150, 147, 146, 144, 142, and 141, Platinum genome variants, Reported variants by RepeatMasker, Microsattellite, Segmental Dups, simple repeats, and Genomic intervals masked by WindowMasker + SDust).

**Supplementary Note 13: Functional domains**

To test whether the convergent variants of avian vocal learners and human-specific amino acid substitutions are localized on conserved functional domains, we performed NCBI conserved domain search27 with human peptide sequences of key candidate genes. We used following options: Data source as ‘cdd v3.16’ and datatype as ‘hitsConcise Results’. After checking locations of functional domains, we compared the locations of AVL-CSAV sites and Human-SAV sites in human peptide sequences of each gene. The location information of functional domain, AVL-CSAV sites, and Human-SAV sites was visualized by using IBS28.

# **Supplementary Note 14: Institutional Review for animal cares and experiments**

The care and experimental use of animals (zebra finch or chicks) were approved by the Institute of Laboratory Animal Resources, Seoul National University (SNU-150827-1) and the Rockefeller University IACUC. The experimental animals were maintained according to a standard management program at the University Animal Farm, Seoul National University in or the Rockefeller University. The procedures for animal management adhered to the standard operating protocols of the laboratory at Seoul National University, Korea or the at the Rockefeller University.

# **Supplementary Tables**

**Supplementary Table 1. Summary of avian multiple sequence alignments.** The multiple peptide and coding sequence alignments are accessible at the following link (Https://github.com/chulbioinfo/CSAVanalysis).

|  |  |
| --- | --- |
|  | **Avian multiple sequence alignments** |
| Total # of species in alignments | 48 birds, 3 reptiles, 1 primate  (5 outgroup species, rifleman, American alligator, Green lizard, Green turtle, and human) |
| Orthologous define method | RBH approach with 2 representative species (Chicken and Zebra finch) |
| Total # of orthologous gene sets | 8,295 |
| Total amino acid lengths of peptide alignments (aa) | 4,519,041 |
| Total nucleotide lengths of codon alignments (bp) | 13,557,123 |
| Alignment programs | SATé+MAFFT, SATé+Prank |

**Supplementary Table 2. Summary of 6 species combinations in 47 birds excluding rifleman (*Acanthisitta chloris*).** (a) Summary of number of origins of 10,737,573 control sets with 6 species combinations. 8,281 sets with 6 species originated from 3 independent lineages like avian vocal learners are marked as bold. (b) Summary of random and core control sets which were analyzed in this study.

**(a)**

|  |  |
| --- | --- |
| **Number of origins (independent lineages)** | **Number of sets** |
| 1 | 3 |
| 2 | 165 |
| **3** | **8,281** |
| 4 | 199,976 |
| 5 | 1,795,696 |
| 6 | 8,733,452 |
| Total | 10,737,573 |

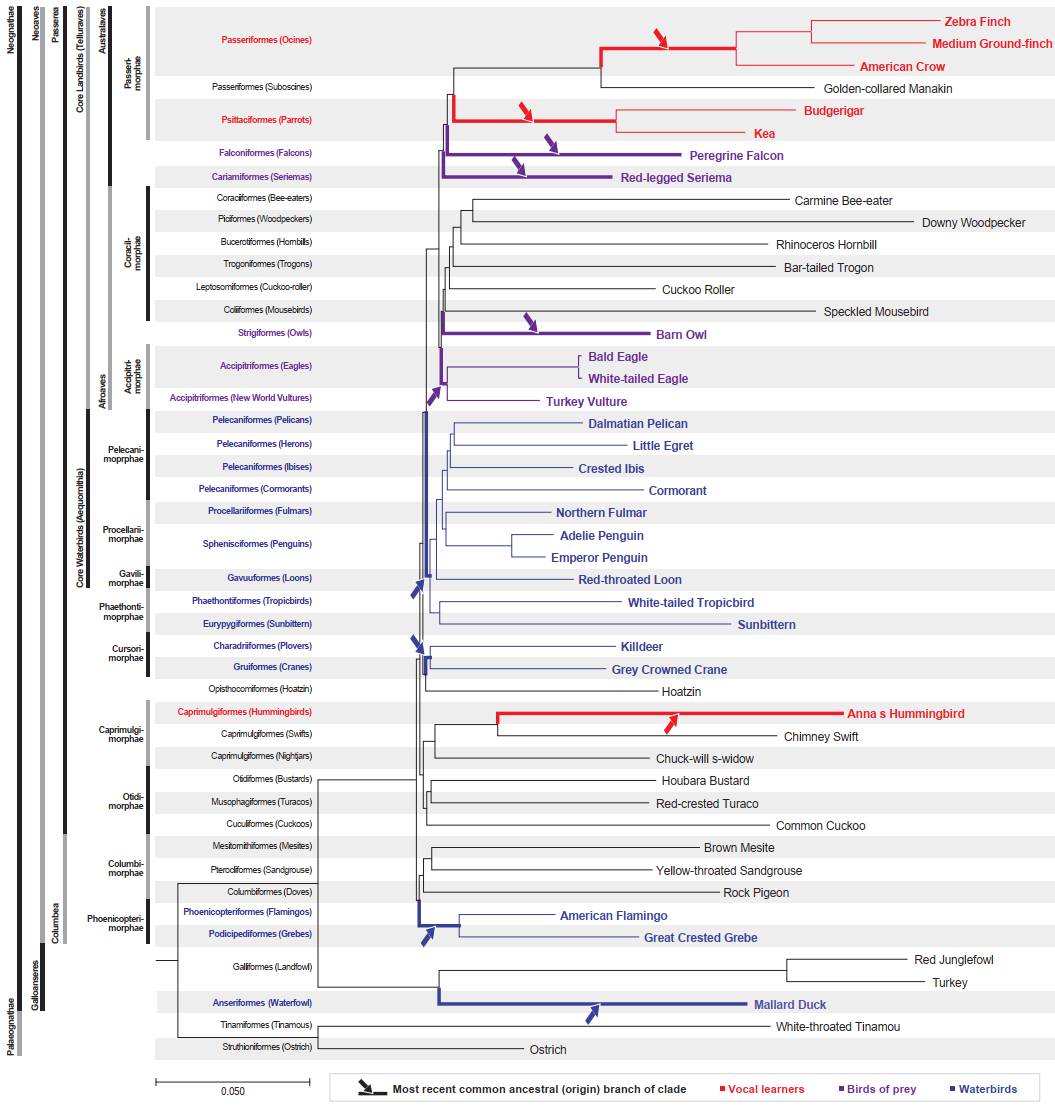
**(b)**

|  |  |  |
| --- | --- | --- |
| **Type of control sets** | **Number of sets** | **Description** |
| Random  control sets | 1,000 | Randomly selected within 8,281 sets with 3 origins |
| Core  control sets | 61 | 59 sets consisted of 6 species from 2 vocal learning clades and 1 vocal non-learning clade, 2 sets included 6 birds of prey and 15 water birds |

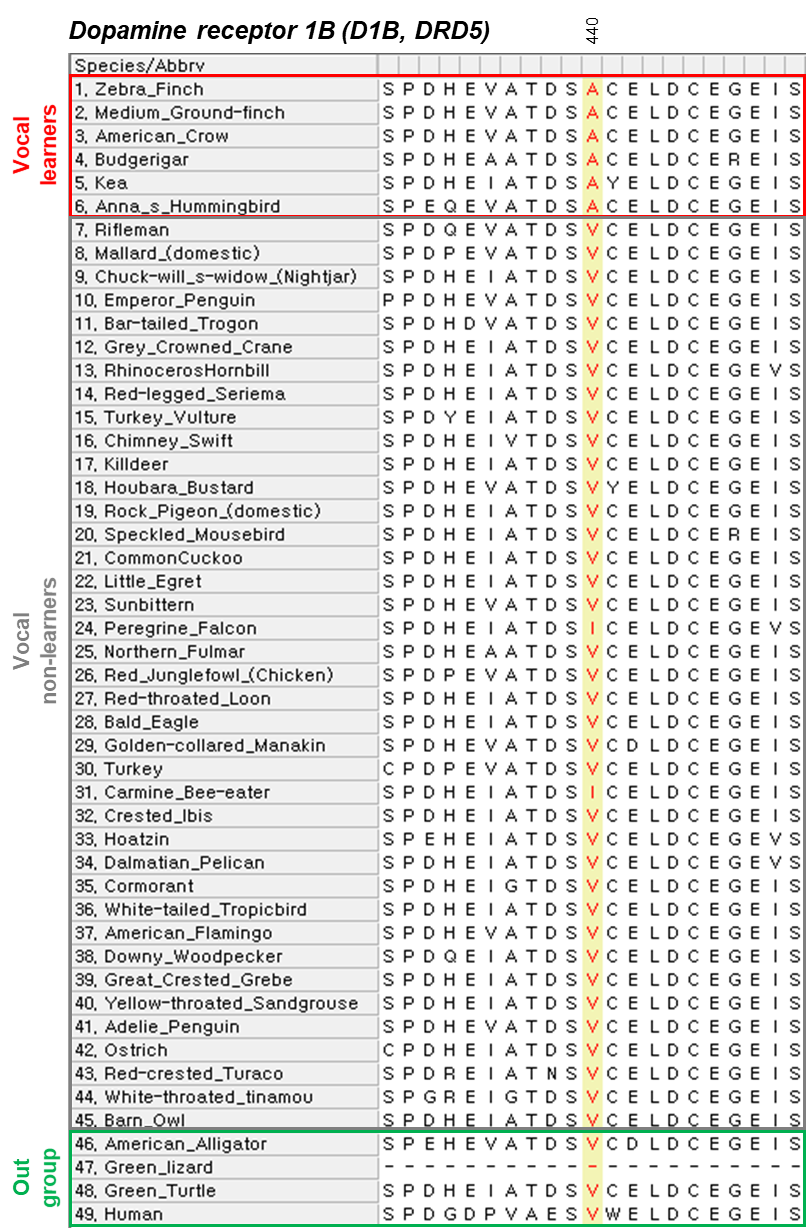
**Supplementary Table 3. Summary of primate multiple sequence alignments.** The multiple peptide and coding sequence alignments are accessible at the following link (Https://github.com/chulbioinfo/CSAVanalysis).

|  |  |
| --- | --- |
|  | **Avian multiple sequence alignments** |
| Total # of species in alignments | 24 primates  (1 human, 23 non-human primates) |
| Total # of orthologous gene sets | 18,565 |
| Total amino acid lengths of peptide alignments (aa) | 13,692,899‬ |
| Total nucleotide lengths of codon alignments (bp) | 41,078,697 |
| Orthology define | Ensembl compara (release 94)24 with 1 representative species (Human) |
| Alignment programs | PRANK |

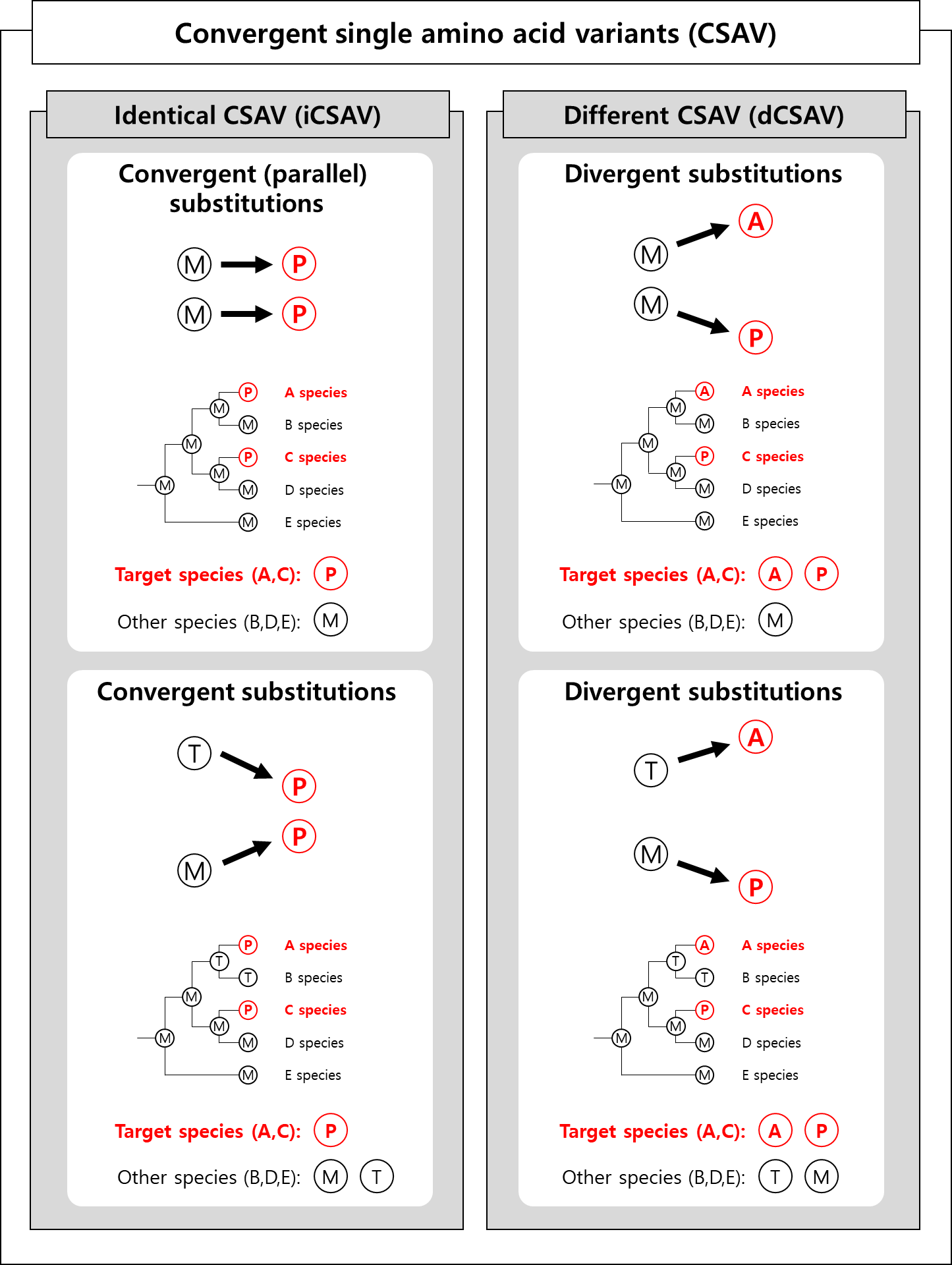
# **Supplementary Figures**



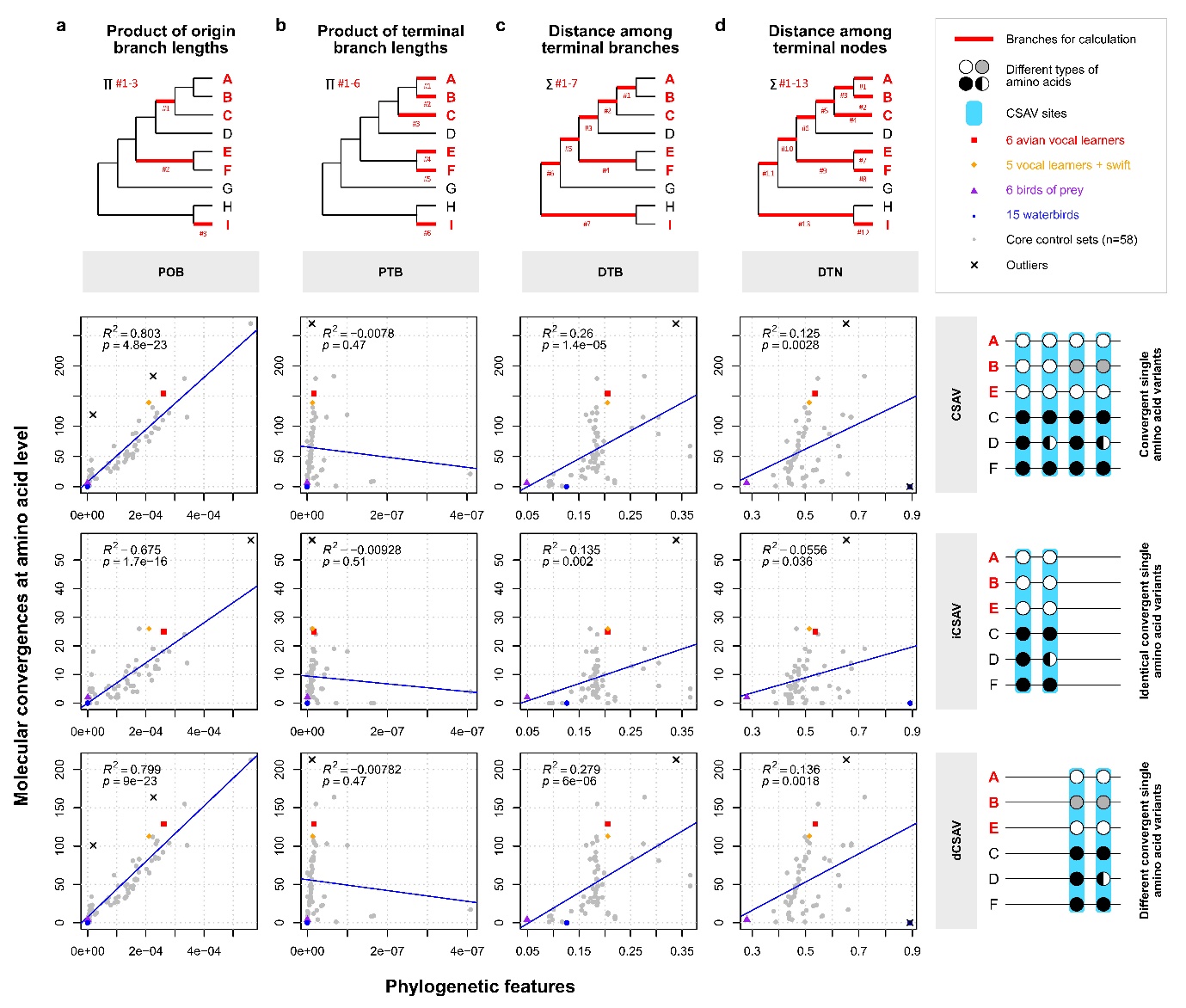
Supplementary Figure 1. Estimated branch lengths of avian lineages. Numbers on branches = number of substitutions per site, as generated by ExaML analyses. Red characters indicate avian vocal learners, and red lines indicate branches of avian vocal learners. Scale of branch length is number of subsitutions per site. Modified from Jarvis et al.1.



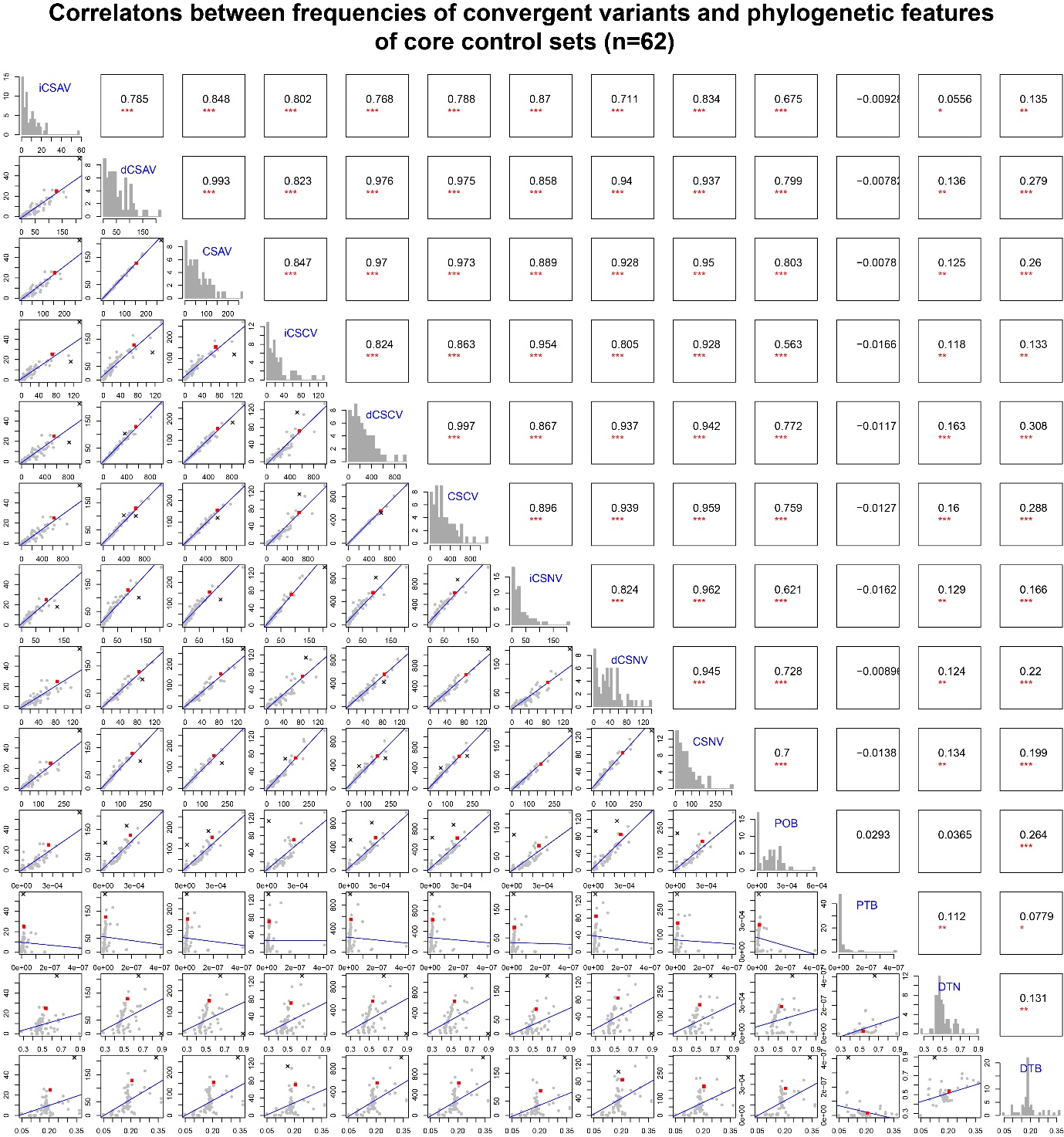
**Supplementary Figure 2. Example of a trimmed region with a low alignment score caused by a regional deletion in an outgroup species.** Note the outgroup lizard has missing sequence, which would have caused the entire sequence be removed for all species using G-blocks, and thus the convergent site in vocal learning birds (yellow) would have not been discovered.



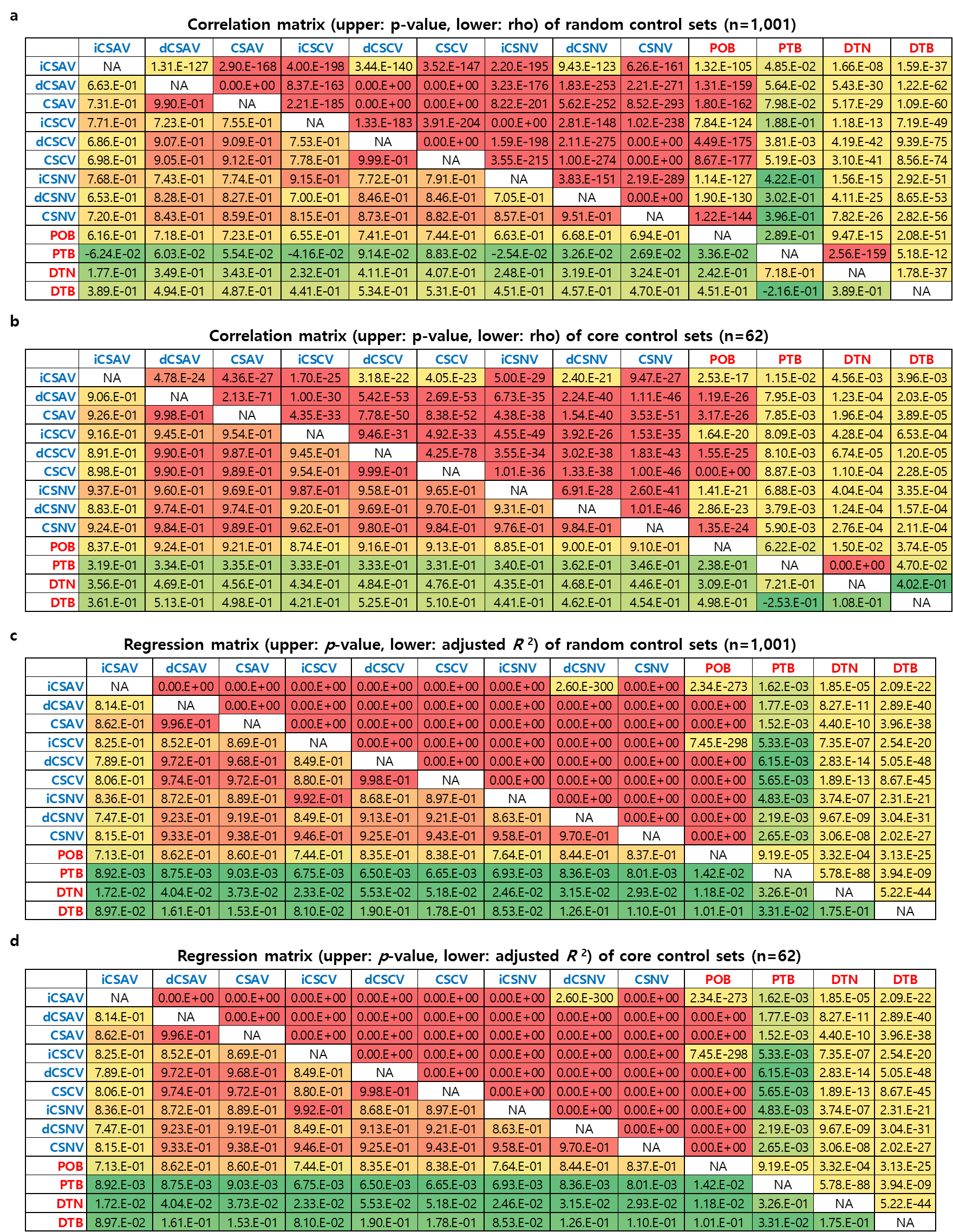
**Supplementary Figure 3. Evolutionary histories generating convergent single amino acid variants.** Convergent substitutions consist of identical and different convergent substitutions. As a convergent feature of polyphyletic species from independent lineages, both types of substitutions of target species have the mutual exclusiveness of amino acid information between target species (A and C species) and the other species (B, D, and E species) at the terminal nodes. Based on considerations on amino acid information of ancestral nodes, identical convergences are classified as 2 types by evolutionary histories: convergent parallel substitutions from same amino acids (M) at ancestral state to same amino acids (P) at their daughter taxa, and convergent substitutions from different amino acids (M, T) at ancestral state to same amino acids (P) at their daughter taxa. On the other hands, different convergences are also classified as 2 types by evolutionary histories but called as divergent substitutions: from same amino acids (M) at ancestral state to different amino acids (A, P) at their daughter taxa, and from different amino acids (M, T) at ancestral state to different amino acids (A, P) at their daughter taxa. Target and the other species are marked as red and black colors, respectively.

****

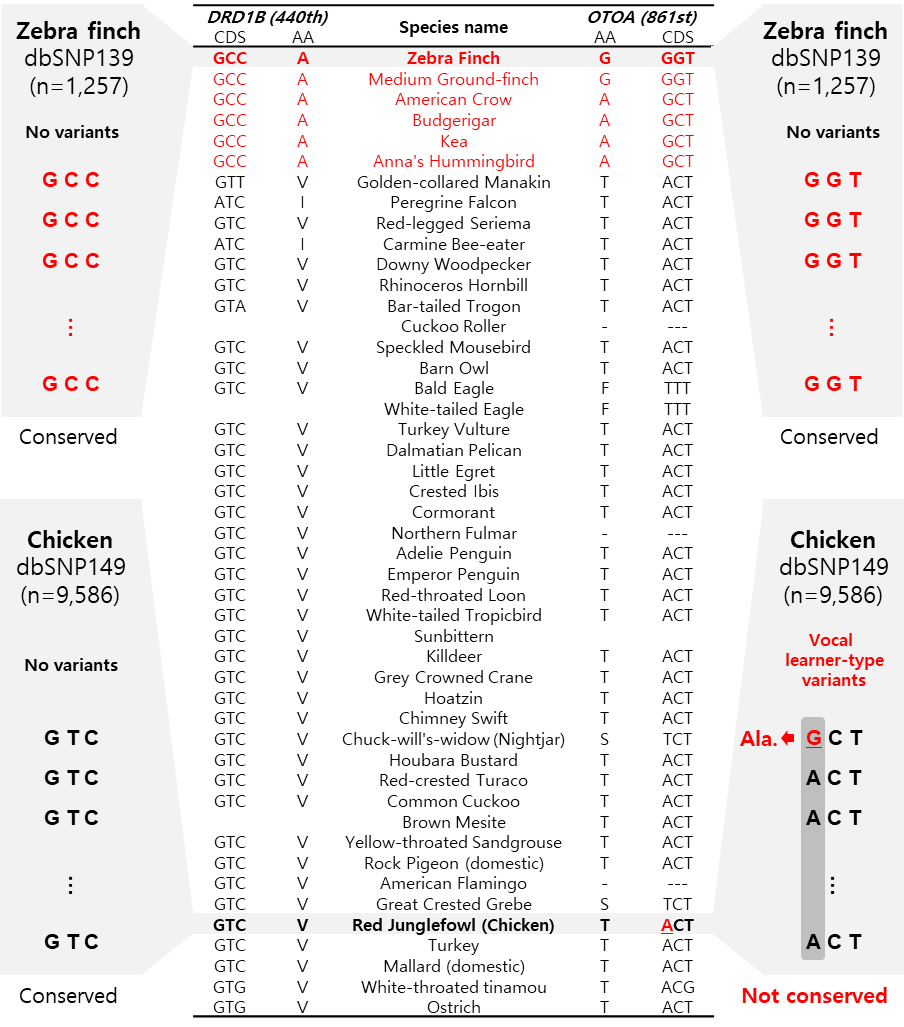
**Supplementary Figure 4. Amino acid convergence amount is correlated to the product of origin branch lengths.** Shown are regression analyses of amino acid convergence in the set of vocal learning birds and 61 core control sets of avian species with four phylogenetic tree features: (a) product of origin branch lengths; (b) product of terminal branch lengths; (c) distance among terminal branches; and (d) distance among terminal nodes. Top row, example type of tree branches used (red lines) associated with species clades that have a convergent trait (red text) with example calculations listed below each tree. Bottom three rows, correlation plots with all four types of convergent amino acids (CSAV) and the identical CSAV (Types 1+2) and different (Types 3+4) separated out. Left, legend key for control species combinations and species with known convergent traits, and the four types of CSAV (based on pattern in **Fig. 1b**). Statistics calculated as a linear regression, with adjusted *R*2 and *p* value using the 'lm’ function in the R package (ver. 3.5.1).



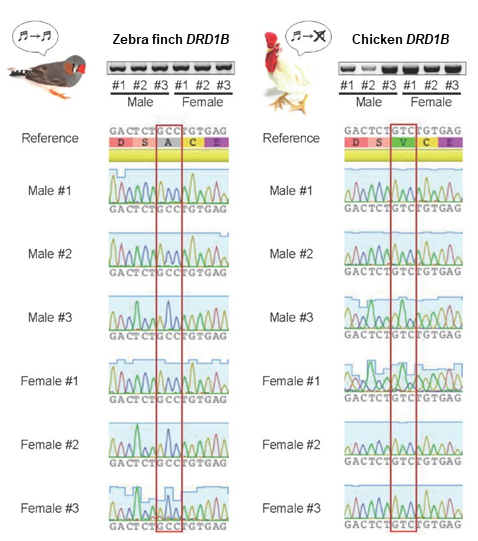
**Supplementary Figure 5. Significantly strong correlations among convergent variants at multiple levels and the product of origin branch lengths (POB) of random control sets.** *p* values and Adjusted squared R of correlations are visualized at upper diagonal matrix (*p*<0.05\*, *p*<0.01\*\*, and *p*<0.001\*\*\*). Histograms of frequencies of each convergent variant and values of each phylogenetic feature are visualized at diagonal matrix. Scatter plots between frequencies of convergent variant and values of phylogenetic features are visualized at lower diagonal matrix. Grey and red spots indicate control sets and set of avian vocal learners, respectively. Blue lines and black ‘X’ marks indicate regression lines and outliers, respectively. POB = product of origin branch lengths, PTB = product of terminal branch lengths, DTB = distance between terminal branches, DTN = distance between terminal nodes, CSAV = convergent single amino acid variants, iCSAV = identical CSAV, dCSAV = different CSAV, CSCV = convergent single codon variants, iCSCV = identical CSCV, dCSCV = different CSCV, CSNV = convergent single nucleotide variants, iCSNV = identical CSNV, dCSNV = different CSNV.



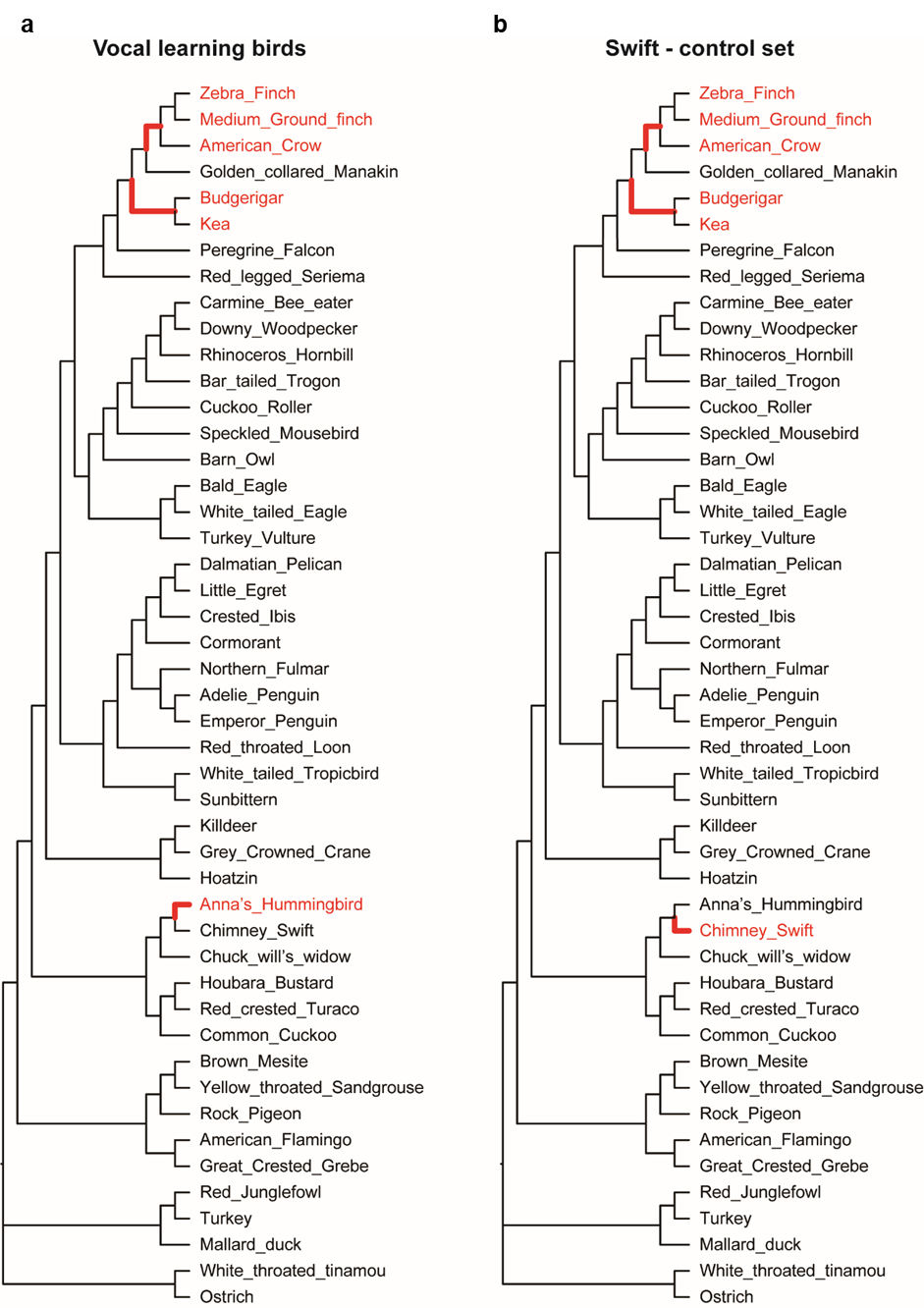
**Supplementary Figure 6. Significantly strong correlations among convergent variants at multiple levels and the product of origin branch lengths (POB) of random and core control sets.** (a, b) *rho* and *p* values of Spearman’s rank correlation tests for random and core control sets. (c, d) Adjusted *R2* and *p* values of linear regressions for random and core control sets. CSAV = convergent single amino acid variants, iCSAV = identical CSAV, dCSAV = different CSAV, CSCV = convergent single codon variants, iCSCV = identical CSCV, dCSCV = different CSCV, CSNV = convergent single nucleotide variants, iCSNV = identical CSNV, dCSNV = different CSNV, POB = product of MRCA (=origin) branch lengths, PTB = product of terminal branch lengths, DTB = distance between terminal branches, DTN = distance between terminal nodes.

****

**Supplementary Figure 7. Examples of fixed and unfixed differences wthin each population.** The central table indicate convergent single amino acid variants of vocal learning birds in *DRD1B* and *OTOA*. Numbers in parentheses indicate positions in peptide alignments of each gene. Bold characters in the species name column indicate representative species of vocal learners and non-learners which are marked as red and black, respectively. AA and CDS column show amino acids and codons of each species at the AVL-CSAV sites of each gene. Blank and ‘- (gap)‘ indicate absence of orthologous gene in the species’ genome and deletions in the species. Under bar at the first site of the AVL-CSAV site in *OTOA* gene of chicken indicates a nonsynonymous SNP in chicken population (dbSNP149, number of samples = 9,586). Except for the case of *OTOA* gene of chicken, all of AVL-CSAV sites are conserved within zebra finch population (dbSNP139, number of samples = 1,257) and chicken population without any nonsynonymous substitutions.

****

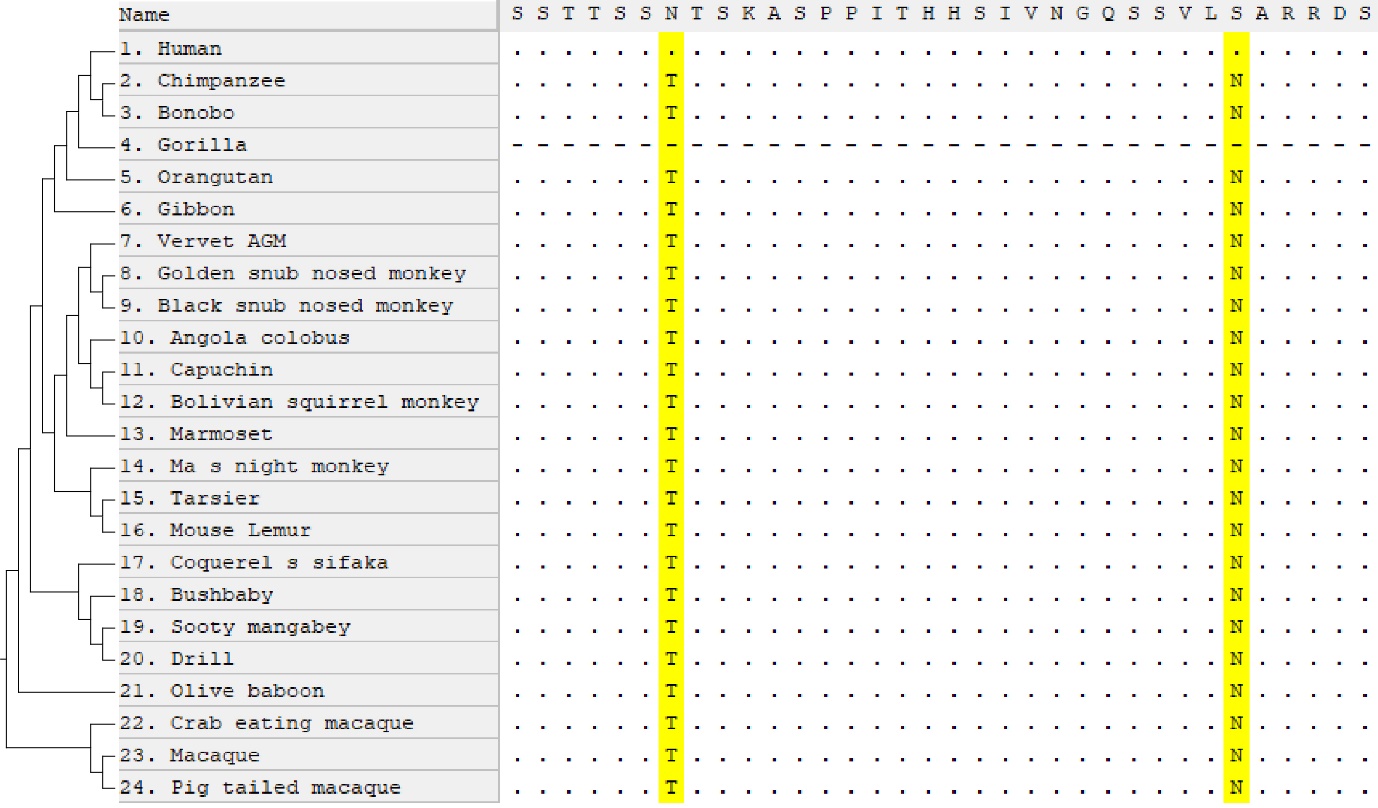
**Supplementary Figure 8. Fixed differences of the avian vocal learner-specific identically convergent amino acid variant (AVL-iCSAV) in *DRD1B.*** Shown are sequences determined from PCR reactions from individual animals. All of 3 male and 3 female samples of zebra finch and chicken showed fixation of the vocal learner-type codon (GCC) and vocal non-learner-type codon (GTC) at the AVL-iCSAV site in *DRD1B*, respectively.

****

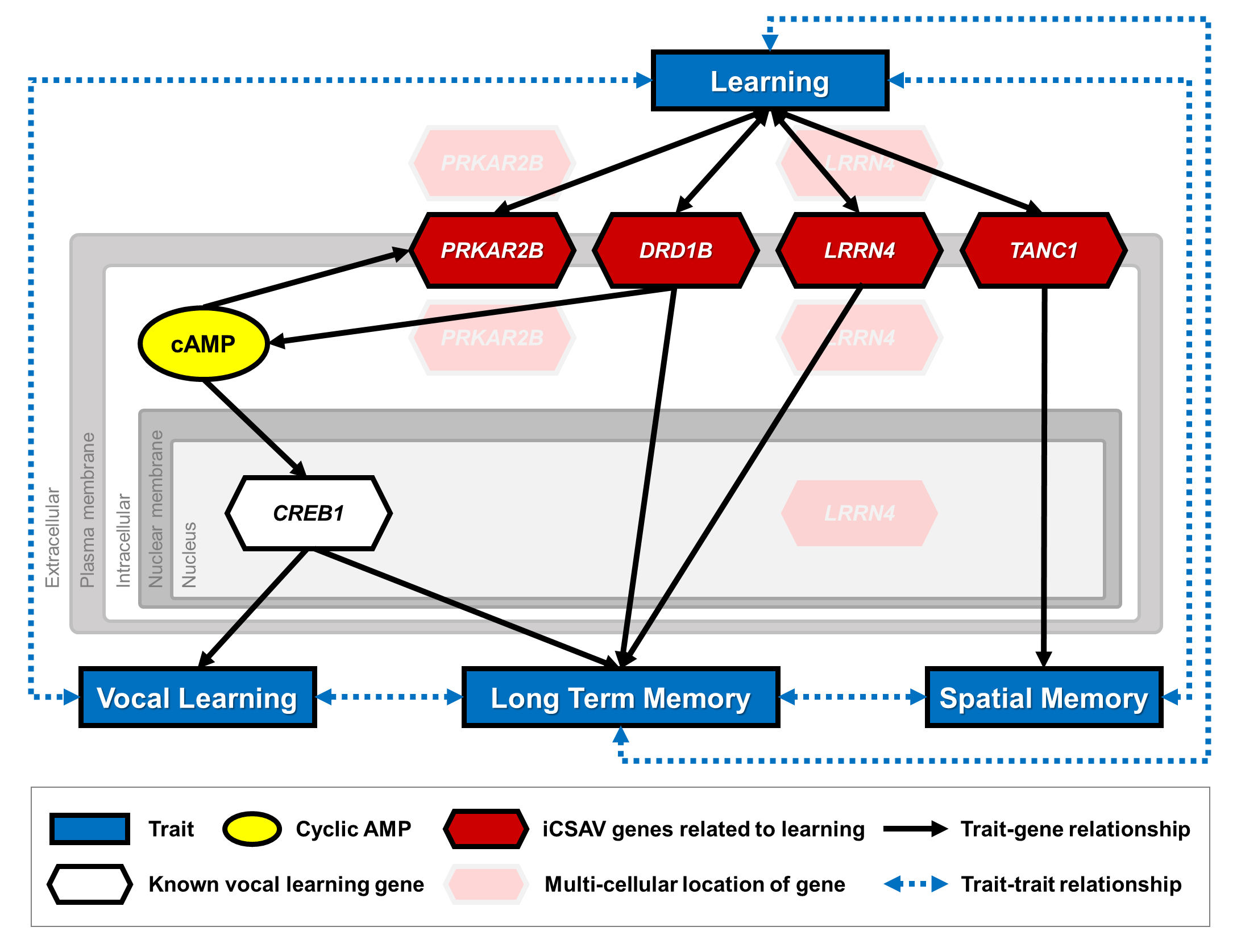
**Supplementary Figure 9. Evolutionary models of positive selection on avian vocal learner set and their closest relative set (Swift).** (**a**) parsimonious hypothesis to get independent gains of vocal learning ability for each vocal learning clade. (**b**) parsimonious hypothesis for positive selection on species of the closest control set like vocal learners’ set. Red characters indicate target species of each set. Bold branches indicate the most recent common ancestral (=origin) branches of each clade of target species which are assumed as foreground branches under positive selection.

****

**Supplementary Figure 10. Different concepts to define differentially expressed genes in song nuclei.** ‘NUC VS NUC’ comparison: gene supported by a significantly differential expression between a song nucleus relative to at least one of other song nuclei. ‘NUC VS other NUCs’ comparison: gene supported by 3 significantly differential expression among a song nucleus relative to the other song nuclei.

****

**Supplementary Figure 11. Human-specific amino acid substitutions in *FOXP2* in 24 primate alignments.** Peptide alignment of *FOXP2* gene of 24 primates by using prank alignment program 25. The cladogram of primates from Ensembl (release 94) 24. Yellow highlighted sites indicate the famous language-related human-specific substitutions in *FOXP2* gene: p.Thr303Asn and p.Asn325Ser. Peptide sequences and the tree are visualized with MEGA 29.

****

**Supplementary Figure 12. Cyclic AMP based vocal learning pathway.** Red hexagons indicate learning genes with identically convergent amino acid variants iCSAV specific avian vocal learners. Transparent red hexagons indicate muti-cellular location of the genes. White hexagon indicates vocal learning gene, *CREB1*. Yellow circle indicates cyclic AMP (cAMP). Grey-scale rectangular indicates cellular positions of the genes and cAMP. Blue rectangular indicates traits related to learning genes. Black arrows and dashed blue arrows indicate trait-gene relationships and trait-trait relationships, respectively.

# **Supplementary Data**

**Supplementary Data 1. Convergent single amino acid variants of avian vocal learners mutually exclusive to avian vocal non-learners.**

**Supplementary Data 2. Gene profiles of convergent variants, positive selection, and differential gene expression on song nuclei and these surrounding regions.**

**Supplementary Data 3. Single amino acid variants of human compared to non-human primates in AVL-CSAV genes.**

# **Supplementary References**

1 Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320-1331 (2014).

2 Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311-1320 (2014).

3 Jarvis, E. D. *et al.* Phylogenomic analyses data of the avian phylogenomics project. *GigaScience* **4**, doi:10.1186/s13742-014-0038-1 (2015).

4 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* **17**, 540-552 (2000).

5 Reza, F. M. *An introduction to information theory*. (Courier Corporation, 1961).

6 Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18**, 6097-6100 (1990).

7 Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome research* **14**, 1188-1190 (2004).

8 Castoe, T. A. *et al.* Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences* **106**, 8986-8991 (2009).

9 Thomas, G. W. & Hahn, M. W. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. *Molecular biology and evolution* **32**, 1232-1236 (2015).

10 Team, R. C. R: A language and environment for statistical computing. (2013).

11 Fox, J. *et al.* Package ‘car’. *Vienna: R Foundation for Statistical Computing* (2012).

12 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191 (2009).

13 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution* **35**, 1547-1549 (2018).

14 Raudvere, U. *et al.* g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research* **47**, W191-W198 (2019).

15 Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091-1093 (2009).

16 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).

17 Bindea, G., Galon, J. & Mlecnik, B. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics*, btt019 (2013).

18 Yates, A. *et al.* Ensembl 2016. *Nucleic acids research* **44**, D710-D716 (2016).

19 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic acids research* **46**, D754-D761 (2018).

20 Madden, T. in *The NCBI Handbook [Internet]. 2nd edition* (National Center for Biotechnology Information (US), 2013).

21 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308-311 (2001).

22 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586-1591 (2007).

23 Pfenning, A. R. *et al.* Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**, 1256846, doi:10.1126/science.1256846 (2014).

24 Cunningham, F. *et al.* Ensembl 2019. *Nucleic acids research* **47**, D745-D751 (2019).

25 Löytynoja, A. in *Multiple sequence alignment methods* 155-170 (Springer, 2014).

26 Löytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632-1635 (2008).

27 Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic acids research* **39**, D225-D229 (2010).

28 Liu, W. *et al.* IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* **31**, 3359-3361 (2015).

29 Kumar, S., Tamura, K. & Nei, M. MEGA: molecular evolutionary genetics analysis software for microcomputers. *Bioinformatics* **10**, 189-191 (1994).