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

Chul Lee
Seoul National University
Seoae Cho
C&K Genomics
Kyu-Won Kim
Kongju National University
DongAhn Yoo
Seoul National University
Jae Yong Han
Seoul National University https://orcid.org/0000-0003-3413-3277
Hong Jo Lee
Seoul National University
Gregory Gedman
Rockefeller University
Jean-Nicholas Audet
Rockefeller University
Erina Hara
Duke University Medical Center
Miriam Rivas
Integrated Laboratory Systems, Inc.
Osceola Whitney
The City College of New York
Andreas Pfenning
Carnegie Mellon University https://orcid.org/0000-0002-3447-9801
Heebal Kim
Seoul National University https://orcid.org/0000-0003-3064-1303
Erich Jarvis (ejarvis@rockefeller.edu)
Rockefeller University https://orcid.org/0000-0001-8931-5049

Article

Keywords: Convergent evolution, Single amino acid variant (SAV), Single codon variant (SCV), Single nucleotide variants (SNV), Product of original branch lengths (POB), Positive selection, Differential gene expression, Dopamine receptor D1B (DRD5), PRKAR2B.

Posted Date: November 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-100389/v1

Abstract

Single amino acid variants (SAVs) may provide clues to understanding evolution of traits. A complex trait that has evolved convergently among species is vocal learning, the rare ability to imitate sounds heard and an important component of spoken-language. Here we assessed whether convergent vocal learning bird species have convergent SAVs (CSAVs) that could be associated with their specialized trait. We analyzed avian genomes and identified CSAVs in vocal learners, but also in most species combinations tested. The number of CSAVs among species was proportional to the product of the most recent common ancestor (MRCA; origin) branch lengths of the species in question, and vocal learning birds did not exceed the overall proportion in most test. However, genes with identical CSAVs (iCSAVs) in vocal learning species were uniquely enriched in 'learning' functions, and a subset of iCSAV genes were under positive selection and had enriched specialized regulation in vocal learning and their adjacent brain subdivisions. Several top candidate genes converge on the cAMP signaling pathway, including *DRD1B* and *PRKAR2B*. Our findings suggest a complex mechanism of amino acid convergences and specialized gene regulation upon which selection acts for specialized convergent traits.

Introduction

Single amino acid variants (SAV) are one of the potential drivers of evolution for various traits. For example, the Forkhead box P2 (*FOXP2*) transcription factor has two well-known human-specific SAVs which might have been positively selected for learning behavior related to language[1,2]. Mutant mice humanized for the two SAVs of *FOXP2* showed more advanced learning abilities[3] and alterations of cortico-basal ganglia circuits[4,5,6], which play critical roles in spoken-language[7]; and mice containing a heterozygous missense mutation that causes speech syllable apraxia in humans also showed syllable sequencing deficits[8,9].

A crucial component of spoken-language is vocal learning, the ability to produce vocalizations through imitation, and is a convergent trait observed in only a few animals, including songbirds, parrots, and hummingbirds among birds, and bats, dolphins/whales, seals, elephants, and humans among mammals[7,10–13]. Both vocal learners and vocal non-learners share an auditory pathway that controls auditory learning, while only the vocal learning birds and humans have been found to share a specialized convergent forebrain pathway that controls vocal learning[11–14]. Supporting the hypothesis of independent origins of vocal learning, the recent genome-scale phylogenetic tree reported by the Avian Phylogenomics Project showed that the three avian vocal-learner lineages are indeed not monophyletic[15]. Even though songbirds and parrots are relatively closely related, the closest lineage to songbirds[15], sub-oscines, is a vocal non-learning lineage

In the first genome-scale analyses for vocal learning in the avian lineage, genes with positively selected changes in zebra finch (a songbird) compared to chicken were identified[16]. Some of the positively selected genes were in ion channels, which are known to control neurological function, behavior and disease[16]. However, the comparison was made narrowly between only one vocal learner (zebra finch) with one vocal non-learner (chicken), which is a very distant[15] relative, like a marsupial is to a placental mammal.

The big bang of draft genome sequences of the Avian Phylogenomics Project, consisting of 48 avian species that represent 34 orders of birds[17], provided an unprecedented opportunity to investigate genetic features specific to polyphyletic vocal learning clades. These studies found convergent brain gene expression specializations in vocal-learning birds and human[11,12,14]. We also found mutually exclusive convergent amino acid substitutions unique to vocal learners, using a novel method (Target-specific Amino Acid Substitutions [TAAS] analysis)[17]. However, the study overlooked several viewpoints reported around that time for molecular convergences[18,19]; it did not separately test for identical (i.e. true convergence) versus different (divergent) convergent amino acid substitutions; it did not test for preponderance of convergences and divergences over random control sets of species; and it did not test for possible influences of close phylogenetic relationships.

Here, we investigated basic rules of molecular convergent evolution and their functions in various combinations of avian species, including vocal learners. We improved methods for comparative genomics of convergent and divergent amino acid

variants among species, and tested whether vocal learning birds have more molecular convergences or divergences than control sets. We discovered phylogenetic features associated with the number of convergent and divergent substitutions among species beyond those of previous studies[19,20], and the underlying nucleotide variant changes associated with these amino acid substitutions. We found a preponderance of higher changes in vocal learners when considering the most recent common ancestor branch lengths, and we did found an enrichment in learning functions, positive selection, and specialized gene expression in vocal learning brain regions and the subdivisions they reside for a subset of genes with amino acid convergences of vocal learning birds.

Results

Convergent amino acid variants specific to avian vocal learners. Among the genomes of 48 avian species spanning most orders[17], we compared 6 species from the three vocal learning orders or suborders (songbirds: zebra finch, medium ground finch, and American crow; parrots: budgerigar, and kea; and hummingbirds: Anna's hummingbird) with 41 vocal non-learning birds (Fig. 1a, **Fig. S1**)[15]. Rifleman, a close relative of songbirds and suboscines, was initially excluded because of the uncertainty of its vocal learning ability, although assumed to be a vocal non-learner[15]. These were short-read genome assemblies and do not have all repetitive and GC-rich regions assembled, but do have most protein coding genes assembled[21]. We performed CSAV analysis for 4,519,041 homologous amino acid sites in an alignment of 8,295 orthologous genes in birds (**Supplementary Table 1**; methods in **Supplementary Note 1**). Out of these homologous sites, 154 sites (0.0034%) detected in 141 genes (1.7%) contained convergent single amino acid variants specific to avian vocal learners (AVL-CSAV; **Supplementary Data 1**). These amino acid differences were significantly supported (adjusted $p < 0.05$) by Fisher's exact test[22] (**Supplementary Data 1**). We classified AVL-CSAVs into four types, two identical and two different types by applying a Shannon's entropy test[23] (Fig. 1b; **Supplementary Note 1** and **Supplementary Data 1**); the terminology, 'identical' and 'different' were used, instead of 'convergent' and 'divergent' or 'parallel'[20,24], as the term 'convergent' can be ambiguously interpreted as pertaining to only identical CSAVs or both identical and divergent CSAVs. Out of 154 AVL-CSAV sites, 25 sites (16%) were identical CSAVs (AVL-iCSAV) and 129 sites (84%) were different CSAVs (AVL-dCSAV) among vocal learners relative to vocal non-learners (**Supplementary Data 1**). For example, the 253rd site of *B3GNT2* was a Type 1 (AVL-iCSAV) site with asparagine (N) observed in all avian vocal learning species and histidine (H) in all vocal non-learning species; while the 217th site of *SMRC8* was a Type 3 (AVL-dCSAV) site with glutamine, valine, and leucine (Q, V, and L) observed in avian vocal learners and isoleucine (I) in all vocal non-learners (Fig. 1c).

Avian vocal learners did not show a preponderance of identical convergent single amino acid variants. We next tested whether avian vocal learners have a higher frequency of convergent substitutions relative to control sets of species. Considering the polyphyletic relationship of the 6 vocal learning species examined, we designed 2 types of clade-specific control sets: 1) Random controls consisting of 1,000 different species combinations with 6 randomly selected target species from 3 independent lineages without considering any traits; 2) Core controls consisting of all possible 59 sets given the phylogeny with 6 target species having at least 2 vocal learning clades and 1 non-learning clade, and 2 sets with different numbers of target species with other convergent traits (6 birds of prey from 4 clades and 15 waterbirds from 4 clades: **Supplementary Note 2; Supplementary Table 2**). We conducted the CSAV analysis for this total of 1,056 random and core control species combinations, and identified convergent identical (Ctrl-iCSAV) and different (Ctrl-dCSAV) amino acid substitutions in all control sets.

As an extension to the previous studies on convergent evolution in reptile and mammalian lineages[18–20] that tested pair-wise combinations of two species (**Supplementary Note 3**), we found strong correlations between iCSAVs and dCSAVs tested in higher dimensional combinations of species (Fig. 1d,e). Although higher than the expectation (20.48 and 23.04 respectively) according to the regression with random and core control sets, the number of convergent substitutions in vocal learning birds was not an outlier (adjusted $p > 0.05$, Bonferroni Outlier Test [25]) from the trend observed in the control sets (Fig. 1d,e). Several outliers did exist among the control sets, with the highest residual being 32.46 in one of the random control sets (4 Passeriformes, budgerigar, and falcon), and 17.61 in a core control set (3 songbirds, Anna's hummingbird, and 2 land fowls;

Fig. 1d,e). These species combinations of 2 control sets with the highest residuals, however, do not share any known convergent traits as far as we are aware. In the groups of species with other known convergent traits (birds of prey and waterbirds), both iCSAVs and dCSAVs were less frequently observed (Fig. 1e). These findings support that identical convergent single amino acid substitutions are widespread, and their numbers vary in different species combinations that does not appear to readily correlate with known convergent traits.

Amino acid convergences are associated with product of origin branch lengths. We sought a measure of molecular convergence that controls for phylogenetic relationships. According to previous studies on mammalian or drosophila genomes[24], and vertebrate mitochondrial genomes[19], fewer convergent substitutions are expected with the greater phylogenetic tree branch distance. However, the correlations found in these studies showed high variations, which makes it difficult to identify the outliers. Here, we took into consideration additional phylogenetic features, including the relationship between the convergent variants versus the most recent common ancestor [MRCA] branch of each clade (origin branches), the terminal branches, and the nodes of the tree (Fig. 2, **top row**; **Supplementary Note 4; Supplementary Fig. 1**). We observed strong and significant correlations between CSAVs and the product of MRCA branch length lengths (POB) for both random control (Fig. 1f **and** Fig. 2) and core control sets (Fig. 1g and **Supplementary Fig. 4**). The correlation was also observed for both iCSAV and dCSAV (Fig. 2a). Much weaker correlations of CSAVs were observed with the product of terminal branches (PTB), distances among terminal branches (DTB) and terminal nodes (DTN; Fig. 2b-d). In contrast to the iCSAV versus dCSAV correlation analyses, in the POB phylogenetic analyses the number of CSAVs in avian vocal learners was a significant outlier relative to random control species sets (adjusted $p < 1.41e\text{-}08$; Fig. 2a). The outlier position of CSAVs was driven mostly by the dCSAVs (adjusted $p < 1.64e\text{-}07$), but also combined trend for the iCSAVs in vocal learners (Fig. 2a). Two other species combinations were outliers, but different from the iCSAV versus dCSAV analyses: (4 Passeriformes, chicken, and duck) and (3 songbirds, Anna's hummingbird, carmine bee-eater, and downy woodpecker). However, the avian vocal learners were not a significant outlier relative to core control sets only (**Supplementary Fig. 4**). These findings suggest that POB value can largely explain convergent variants at the amino acid level, where the longer their ancestral branch lengths the greater frequencies of convergence amino acid variants, and that vocal learning birds are somewhat different in this regard relative to other species combinations.

Convergent amino acid substitutions can arise from complex nucleotide variants at multiple sites. To investigate what types of codon and nucleotide variants can cause CSAVs, we modified the CSAV method to detect convergent single codon variants (CSCVs) made up of 3-nucleotides and convergent single nucleotide variants (CSNVs) in those codons (Fig. 3a,b; **Supplementary Note 5**). Similar to the CSAV analysis, the codon and nucleotide variants were classified into 2 identical types and 2 different types (Fig. 3a,b). A CSAV can theoretically originate from identical single nucleotide variants at a homologous codon position (CSNVs) or complex multiple nucleotide variants (CMNV) at different codon positions (Fig. 3c). However, some CSNVs can also give rise to no change in the amino acid, namely synonymous substitutions (Fig. 3d). We checked for overlaps among CSAVs, CSCVs, and CSNVs to trace the source of the convergent amino acid substitutions at the codon and nucleotide levels.

Analyzing 4,519,041 homologous codons and 13,557,123 homologous nucleotides of the 8,295 singleton orthologous genes in birds, we found 626 CSCV sites specific to avian vocal learners (AVL-CSCVs; Fig. 3e). Among them, 56 (15.7%) showed nonsynonymous CSNVs and 98 (8.9%) nonsynonymous CMNVs, resulting in 154 CSAVs among vocal learners (Fig. 3e). The remaining CSCVs consisted of 113 (18.1%) synonymous CSNVs and 359 (57.3%) synonymous CMNVs (Fig. 3e). An example of a nonsynonymous CSAV explained by a CSNV is in the 253rd codon site of *B3GNT2*, where all vocal learners had the same convergent nucleotide (A), codon (AAT), and amino acid (N, Asparagine) sequence relative to all vocal non-learners (e.g. C; CAT or CAC; and H, Histidine; Fig. 3f). An example of a nonsynonymous CSAVs explained by a CMNV is the 482nd site of *LRRN4*, where all of vocal learners showed identical amino acid convergence (AVL-iCSAV) to Histidine (H), while their codons consist of CAC or CAT for vocal learners and TTT, TAT, TAC, CGC, and TCG for vocal non-learners (Fig. 3f).

In the random and core control species sets, although we found different total numbers of CSCVs and their corresponding CSAVs and CSNVs, their relative proportions (%) were similar to each other and to that of vocal learners (Fig. 3e); about 1/3 of

CSAVs of random and core control sets originated from CSNVs at each homologous nucleotide site, while 2/3 originated from multiple nucleotide changes at different nucleotide sites (Fig. 3e). These findings suggest that amino acid convergences originate not only from identical single nucleotide substitutions at each homologous site but also from complex nucleotide substitutions at multiple sites within a codon.

Convergent variants at all levels are best explained by the product of MRCA branch lengths. Next, we performed correlation tests between nine types of sequence variants (three types of convergences [all, identical, and different] at three levels [amino acid, codon, and nucleotide]) and four types of phylogenetic features (POB, PTB, DTB, and DTN; **Supplementary Note 6**). As expected, all nine types of convergent variants (CSAVs, iCSAVs, dCSAVs, CSCVs, iCSCVs, dCSCVs, CSNVs, iCSNVs, and dCSNVs) were highly correlated with each other, in both random control (Fig. 4) or core control species sets (**Supplementary Fig. 5**). For the phylogenetic features, the POB showed the strongest correlation with all variant types, where the others (DTB, DTN, and PTB) were either weaker or not correlated at all (Fig. 4). Correlations were overall weaker in the core control sets of species (Fig. 4 **vs Supplementary Figs. 5,6**), presumably due to a smaller number of species combinations than the random control sets. Unlike the CSAVs, the numbers of vocal learner-specific variants did not exceed that of control sets for the other variant types (Fig. 4).

Post hoc analyses of vocal learner substitutions in Rifleman. Rifleman and more broadly the New Zealand Wrens, a close relative of vocal learning songbirds, have been assumed to be a vocal non-learner[15]. Although rifleman was excluded from the initial CSAV search, we can ask whether its sequences match those of vocal non-learners as assumed. We applied principal component analysis (PCA) and phylogenetic analysis for the 154 AVL-CSAV sites and the subset of 25 AVL-iCSAV sites (**Supplementary Note 7**). PC1 and PC2 accounted for 53% and 66% of the total variances of the AVL-CSAV sites and AVL-iCSAV sites, respectively (Fig. 5a,b). The vocal learning birds clustered away from the vocal non-learning group as expected (Fig. 5a,b). For the AVL-CSAV sites, rifleman clustered with the vocal non-learners (Fig. 5a). For the AVL-iCSAV subset, rifleman was separate from the two groups, but was still closer to vocal non-learners (Fig. 5b). Phylogenetic analysis of these AVL-CSAVs was consistent with the PCA results, where instead of branching with its closest relatives, the songbirds, rifleman was on a branch outside and next to the vocal learners (Fig. 5c,d). These results support the assumption that rifleman is a vocal non-learner. The results also suggest that some of the convergent sites and genes could be associated with their convergent trait.

Functions of convergent genes. To investigate the biological functions of genes with convergent variants in vocal learners and in control sets, gene ontology (GO) analysis was performed on 9,513 (1,057 species sets * 9 variants types) gene lists with 1 or more convergent variants (**Supplementary Note 8**). Among them, at least one significant (adjusted $p < 0.05$) GO term was found for 1,038 gene lists (10.9%). We further found a positive correlation between the number of significant GO terms and the number of genes with convergent variants in each list (Fig. 6a); however, we found negative correlation between the average adjusted p value of significant GO terms and the number of genes with convergent variants (Fig. 6b).

In vocal learning birds, we did not find any GO enrichment for the total AVL-CSAV gene list. However, the AVL-iCSAVs gene list subset was significantly enriched for 'learning' (GO:0007612, adjusted $p = 0.042$). Four genes were responsible for this enrichment (*DRD1B* [also known as *DRD5*], *LRRN4*, *PRKAR2B*, and *TANC1*; Fig. 6c). The identical amino acid convergences of *DRD1B*, *PRKAR2B*, and *TANC1* also showed identical codon convergences (AVL-iCSCVs) in all vocal learners, while that of *LRRN4* showed different synonymous codon changes (AVL-dCSCV; Fig. 6d). Of the 1,056 control species combinations, only one control gene list (a Ctrl-dCSCV gene list) showed significant enrichment for 'associative learning' (GO:0007612, adjusted $p = 0.035$); the associated set of species (Fig. 6e) did not include any vocal learners, but a convergent variant in *LRRN4* contributed to this functional enrichment (Fig. 6f). The findings indicate that convergent genes in vocal learners do function in the brain and for learning.

Convergent amino acid sites are fixed and positively selected. We checked for additional evidence of whether the CSAVs in vocal learners are reliable, by checking for sequence assembly artifacts and SNPs within species. We used the dbSNP database of a representative vocal learner (zebra finch; n = 1,257 samples; build 139) and a vocal non-learner (chicken; n =

9,586 sample, build 145; **Supplementary Note 9**). At the 154 AVL-CSAV sites, zebra finches showed complete fixation without any nonsynonymous polymorphisms. However, one missense SNP was found in the chicken *OTOA* gene (c.2581A > G, p.Thr861Ala) resulting in an amino acid change identical to that of vocal learners (**Supplementary Fig. 7**). We also validated fixation of the convergent substitution in *DRD1B* by PCR of genomic DNA and sequencing of 3 male and 3 female zebra finches and chickens (**Supplementary Fig. 8**). These findings indicate that the vast majority (99.4%) of the convergent amino acids we identified in vocal learners are the result of true species-specific variants.

In addition, to consider positive selection on the convergent sites, we performed dN/dS analysis with the branch-site model for CSAV genes in the avian vocal learning species and their closest control set (songbirds, parrots, and swifts; **Supplementary Fig. 9**). We found that under half of the iCSAV sites showed signs of positive selection (Likelihood ratio value (D) > 0, ω_2.foreground > 1) in the vocal learning birds (10 of 25 genes, 40%) and the closest control set (12 of 26 genes, 46%), with 12% (3) and 23% (6) being statistically significant, respectively (adjusted $p < 0.05$, posterior probability > 0.5). These findings suggest that a subset of iCSAV genes in different species combinations have been positively selected whether the species share a convergent trait or not, and it does not seem to need a greater number of positively selected sites for the avian vocal learning ability.

Genes with convergent identical changes are specialized in vocal learning and the associated brain subdivisions. We next tested if the AVL-CSAV genes are expressed in vocal learning brain circuits. We analyzed 8 brain transcriptome data sets, which include genes that show singing-regulated expression (increased or decreased) in song learning nuclei of songbirds[26] (Area X, HVC, LMAN, and RA), differential expression (increased or decreased) in song nuclei compared to their surrounding non-vocal motor brain regions (NUC vs SUR), one song nucleus compared among the other song nuclei (NUC vs other NUCs), and the surrounding regions of each song nucleus compared to the other surrounding regions (SUR vs other SURs), from independent experimental data sets (DEG_2014: microarray method in 2014[14,26], DEG_2019: micro-dissected RNA sequencing in 2019[27], and DEG_2020: laser capture microscope with RNA sequencing in 2020[27]; **Supplementary Note 11, Supplementary Fig. 10**, and **Supplementary Data 2**). Relative to the average of all genes (8,295 avian orthologous genes) measured, we found no enrichment of AVL-CSAV genes (0–13%) among the singing regulated genes or song nuclei specialized genes relative to the surrounding brain regions, whether positively selected or not (Fig. 7a). However, in two independent transcriptome experiments, we found 60 to 100% of the positively selected AVL-iCSAV (AVL-PS-iCSAV) genes were enriched among the differentially expressed genes in one song nucleus relative to the others, and some of those were also enriched to a lesser degree in the adjacent surrounding brain subdivision relative to the others (Fig. 7a). These enrichments were not found for the AVL-dCSAV genes, not for any positively selected gene set in the closest related control set (Fig. 7a), nor for all control sets even before positive selection analyses (not shown). Out of 4 song nuclei, Area X involved in song learning showed the highest number of differentially regulated genes out of singleton orthologous genes of birds or genes with amino acid convergences specific to avian vocal learners in comparisons among a song nucleus and the other song nuclei (NUC vs other NUCs, DEG_2020; **Supplementary Data 2**, Fig. 7b).

Out of 25 AVL-iCSAV genes, a total of 8 genes (32%) had positively selected sites in vocal learners and differential expression specific to a song nucleus and surrounding brain subdivision: *B3GNT2, DRD1B, FNDC1, HMGXB3, MTFR1, PIK3R4, PRKAR2B, and SMPD3* (Table 1). These include two genes, *DRD1B* and *PRKAR2B*, revealed in the GO analyses for learning functions (Fig. 6c, d). Further *DRD1B* has specialized up-regulation specific to adult Area X compared to its surrounding striatum (Table 1 and Fig. 7c)[28–30]

Table 1. Avian vocal learning-related candidate genes with iCSAV under positive selection supported by differential expression on song nuclei and surrounding regions. NUC vs SUR: song nucleus compared to its surrounding non-vocal motor brain regions. NUC vs other NUCs: a song nucleus compared to other song nuclei. SUR vs other SURs: a surrounding region of song nuclei compared to another song nucleus.

Gene Symbol	Ensembl Gene ID	CSAV position	AA of vocal learners	AA of vocal non-learners	Likelihood ratio value (D)	Adjusted p value (FDR)	dN/dS (ω_2) of foreground branch	Posterior probability (B.E.B.)	DEG2014 (NUC vs other NUCs)	DEG2019 (NUC vs SUR)	DEG2020 (NUC vs SUR)	DEG2020 (NUC vs other NUCs)	DEG2020 (SUR vs other SURs)
B3GNT2	ENSTGUG00000002218	253	N	H	1.14	2.75.E-01	3.3	0.999**	AreaX Up	RA Down		AreaX Up	AreaX Up
DRD1B (DRD5)	ENSTGUG00000009908	440	A	IV	1.22	2.74.E-01	3.7	0.5	AreaX Up	AreaX Up	LMAN Down	AreaX Up & LMAN Down	AreaX Up
FNDC1	ENSTGUG00000011376	1175	S	G	4.31	6.37.E-02	6.7	0.981*		RA Up		RA Up	
HMGXB3	ENSTGUG00000000975	325	D	-E	0.99	2.97.E-01	2.3	0.995**	AreaX Up			AreaX Up	
MTFR1	ENSTGUG00000011257	104	T	-AGP	3.24	1.01.E-01	10.1	0.521			LMAN Up	AreaX Down	
PIK3R4	ENSTGUG00000004027	673	C	R	4.87	4.93.E-02*	10.4	0.997**	AreaX Up			AreaX Up	
PRKAR2B	ENSTGUG00000003068	98	V	-I	25.44	3.57.E-06**	295.8	0.999**	Ra Down			AreaX Up & RA Down	AreaX Up
SMPD3	ENSTGUG00000006964	311	C	-Y	6.00	3.37.E-02*	14.3	0.994**	AreaX Up			AreaX Up & RA Down	AreaX Up & RA Down

Convergent evolution between vocal learning birds and human in the same genes. Lastly, we investigated amino acid convergences between vocal learning birds and humans. We performed SAV analysis for 18,565 singleton orthologous genes in primates, and discovered 11,317 genes (61.0%) contained 126,087 human-specific amino acid substitutions relative to other primates (Human-SAV; **Fig. 8; Supplementary Note 12**), including the well-known p.Thr303Asn and p.Asn325Ser amino acid substitutions in *FOXP2* specific to humans (**Supplementary Fig. 11**). Since this is over 60% the coding genes in humans, it is unlikely that the majority of these substitutions are associated with differences between humans and other primates. Narrowing them down to our candidate gene list to those found in avian vocal learning species, out of 141 AVL-CSAV genes, 125 singleton orthologous genes were identified in human (**Supplementary Data 3**). Among the 125 genes, 85.6% (107 genes; well over 61%) contained 396 Human-SAV sites compared to other non-human primates (**Fig. 8**). For random and core control sets of birds, on average 64.8% and 77.9% included Human-SAVs, respectively (**Fig. 8**). Despite this lower average % in control sets, several random and core control species sets still showed higher percentages with Human-SAV (**Fig. 8**). These findings are consistent with analyses among birds, that although vocal learning birds and humans have convergences in the same genes, the total number is not higher than some control sets that may not have trait convergences with humans.

We scanned for overlaps of the amino acid substitutions in vocal learning birds and in humans. Out of 107 bird-primate orthologous genes with 120 AVL-CSAV sites among birds and 396 Human-SAV sites among primates, only *SETD4* had an overlapping site, in the 103rd position of the primate and avian peptide alignment (**Fig. 9a**). However, the amino acid Arginine (R) in one avian vocal learning bird (Budgerigar) was identical to that of non-human primates, as opposed to Glycine (G) in human; vocal non-learning birds had a Glutamine (Q; **Fig. 9a**). In contrast to this result, we found that of the 107 genes with avian vocal learner convergences and with human-specific substitutions among primates, 79 showed 111 conserved protein domains containing at least one of AVL-CSAV and Human-SAV site, respectively (**Supplementary Note 13**). This included four genes with identical convergent substitutions in vocal learning birds, enrichment for learning in the GO analyses, with positive

selection and differential expression between song learning nuclei (**Figs. 9;** *DRD1B*, *LRRN4*, *PRKARB*, and *TANC1*). *DRD1B* showed one AVL-iCSAV site and one Human-SAV site located in the c-terminal intracellular region that interact with *DRD2*[31], a gene also associated with a spoken language disorder[32] (**Fig. 9b**). These convergent variants specific to vocal learning birds and humans, but at different sites in the signaling domain of the *DRD1B* receptor along with its specialized expression serve as candidates for changing dopamine receptor function in human speech and songbird vocal learning circuits[28,33]. These findings indicate that in more distantly related species, convergence in genes associated with convergent traits may occur with lineage-specific convergences occurring in different amino acid sites, but in the same protein domain.

Discussion

As the primary structure of proteins, amino acid substitutions can contribute to various traits including human language[34–36]. Our findings give us new insights into convergent evolution of amino acid substitutions, and possible influence on convergent traits. We discovered correlations between the frequency of amino acid convergence with the product of ancestral branch lengths. These amino acid convergences originate from identical but also complex nucleotide substitutions in each codon across species. Our finding that around 60% of convergent amino acid variants (e.g. codons) originate from more complex patterns of convergent nucleotide variants amongst species suggest that identical protein coding convergence due to different nucleotide convergence is more common. Remarkably, although vocal learners did not have a higher preponderance of identical amino acid convergences above background levels, we find that a subset of the sites and the associated genes have been positively selected upon and have specialized expression between different brain subdivisions. To explain our findings, we propose a hypothesis of selection on a background of convergent substitutions for convergent traits.

Our phylogenetic analyses suggest that the background level and rate of convergent substitutions is a function of the product of substitutions rates along the MRCA branches of each clade. Only in the MRCA analyses did we find correlations between the phylogenetic feature, POB, and convergent substitutions in species from multiple independent lineages, where other analyses and studies have attempted and failed to find[24,37–39]. Our positive selection, functional association, and gene expression analyses suggest that selection occurs on some of these convergent substitutions to contribute to evolving novel, convergent traits, in our case vocal learning. Similarly, we predict that some of the other convergent substitutions we found in birds of prey or waterbirds will be associated functionally with those traits. According to this hypothesis, it is not about how many genes show convergence, but which specific genes (e.g. who) show convergence, as the most important factor to consider.

Our findings of an association between convergent identical amino acid changes in vocal learners and specialized expression specific to a vocal learning nucleus and the associated surrounding brain subdivision was both intriguing and perplexing to us. If anything, we were testing a more logical outcome of amino acid convergence in genes that show singing-regulated gene expression or specialized expression in vocal learning brain regions relative to the surrounding brain subdivisions. But the unexpected relationship with vocal learning and brain subdivision specialization we believe is real, as we replicated multiple times, and there is 100% overlap of the most significantly selected genes in vocal learners and brain subdivision gene expression specificity. These findings suggest that there is selection of protein coding sequence changes in vocal learners for a set of genes that have brain region specific expression, particularly in the striatum. Further, one of the striatum-specific genes, *DRD1B*, also had specialized up-regulation in Area X of the striatum, suggesting further regulatory genomic region changes. Often coding and regulatory genomic sources of trait evolution are pitted against in each other as alternatives[40], but our findings suggest that they could synergistically influence evolution of each other. Studies in our group are underway to find the regulatory regions of these genes, and to determine what non-coding sequence changes are the cause of their specialized regulation. Based on convergent variants, positive selection, and differential gene expression in song nuclei and the surrounding regions, we suggest 8 key candidate genes for associations with the vocal learning ability in birds (Table 1).

When searching for convergent substitutions among species, we believe our approach of multi-wise comparisons and the product of the MRCA branch lengths (POB) maybe more informative than past approaches. Previous studies found correlations between convergent identical and different substitutions (previously called convergent and divergent

substitutions) between pairs of species among reptiles[18] or mammals[20]. We further find that such a relationship exists in higher dimensional combinations of species, but this type of analyses does not control for species relationships. Several other studies found that the rate or number of convergent identical and different substitutions decreases with increasing genetic distance between two lineages[19,24]. Our findings with the product of the MRCA branch lengths in convergent polyphyletic clades suggest that the deeper in time their common ancestor, the more likely to find higher proportions of detectable convergences at the amino acid, codon, and nucleotide levels. These analyses provided a new null hypothesis of convergent evolution according to phylogeny.

The biological function of genes with amino acid convergences specific to avian vocal learners gave us new insights into the potential molecular mechanisms of vocal learning. The four convergent learning-related genes with AVL-iCSAV sites includes the *DRD1B* dopamine receptor associated with learning[41], and *LRRN4* that affects long lasting memory[29,42], fundamental traits of vocal learning[43]. *TANC1* regulates dendritic spines and spatial memory[44]. At the mechanistic level, *DRD1B*, through its G-protein, regulates activity of adenylyl cyclase's synthesis of cAMP in the cell membrane[30,45]; *PRKAR2B*, or Protein Kinase cAMP-dependent Type II Regulatory Subunit Beta, is an enzyme that activates cAMP-dependent protein kinase (PKA) inside the cell[46]. Additionally, one of the most well-known genes that PKA inhibits is involved in learning, including vocal learning[47], namely the cAMP response element binding protein (*CREB1)*, a transcription factor responsive to cAMP signaling via PKA, which regulates genes that converts short-term memories into long-term memories[48]. The combined findings suggest that some genes with convergent identical amino acid changes may have a nexus at targeting the cAMP signaling pathway associated with the vocal learning ability (**Supplementary Fig. 12**).

Although our study illuminated novel findings, it spurs on ideas for future studies. Vocal learning species could share other convergent traits besides vocal learning[7,13,49,50], and the identified genes could be associated with these other traits. The basic rules of convergent evolution we discovered in protein coding regions leave open the possibility that similar or different rules apply to non-coding regions. The greater association of rifleman CSAVs with vocal non-learning species could be further tested with brain and behavior studies, to see if indeed they do not have a vocal learning forebrain circuit or advanced vocal learning behavior. We identified new candidate genes and specific nucleotide substitutions that can be genetically manipulated when the technology is more advanced[3,51] to test possible causal roles in the evolution and function of vocal learning. It will be useful to determine if the convergent rules we identified here are specific to birds, or are more widespread across life forms.

Declarations

Author contributions

CL, SC, KK, HK, and EDJ designed the study. CL and EDJ collected all of raw data for this study. CL, KK, HK, and EDJ developed algorithms and programs to identify convergent variants. CL, HK, and EDJ designed random and core control sets by considering phylogenetic relationships among vocal learning birds, and developed algorithms and programs to calculate phylogenetic features. CL performed most of analyses. CL and DY performed PCA and phylogenetic analysis to estimate the vocal learning ability of Rifleman, and conducted fixed difference analyses for public SNP resources of zebra finch and chicken. JH and HL validated fixed differences of convergent amino acid variants in zebra finch and chicken. GG, JA, EH, MR, OW, ARP, and EDJ provided DEG profiles based their independent studies with DEG analyses on song nuclei and those surrounding regions. CL and EDJ wrote the draft paper and all of authors reviewed it. HK and EDJ supervised this study.

Acknowledgement

This work was funded by 'Population genomics of Korean long-tailed fowl' the Program for Agriculture Science and Technology Development (Project No. PJ0133402) of the Rural Development Administration (RDA). This study was supported by HHMI, USA.

References

1. Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. **418**, 869 (2002).
2. Scharff, C. & Petri, J. Evo-devo, deep homology and FoxP2: implications for the evolution of speech and language. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **366**, 2124–2140, doi:10.1098/rstb.2011.0001 (2011).
3. Schreiweis, C. *et al.* Humanized Foxp2 accelerates learning by enhancing transitions from declarative to procedural performance. *Proceedings of the National Academy of Sciences* **111**, 14253–14258 (2014).
4. Enard, W. *et al.* A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell* **137**, 961–971 (2009).
5. Reimers-Kipping, S., Hevers, W., Pääbo, S. & Enard, W. Humanized Foxp2 specifically affects cortico-basal ganglia circuits. *Neuroscience* **175**, 75–84 (2011).
6. Enard, W. FOXP2 and the role of cortico-basal ganglia circuits in speech and language evolution. *Current opinion in neurobiology* **21**, 415–424 (2011).
7. Jarvis, E. D. Evolution of vocal learning and spoken language. *Science* **366**, 50–54 (2019).
8. Chabout, J. *et al.* A Foxp2 mutation implicated in human speech deficits alters sequencing of ultrasonic vocalizations in adult male mice. *Frontiers in behavioral neuroscience* **10**, 197 (2016).
9. Castellucci, G. A., McGinley, M. J. & McCormick, D. A. Knockout of Foxp2 disrupts vocal development in mice. *Scientific reports* **6**, 23305 (2016).
10. Nottebohm, F. The origins of vocal learning. *American Naturalist*, 116–140 (1972).
11. Jarvis, E. D. Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Sciences* **1016**, 749–777 (2004).
12. Petkov, C. I. & Jarvis, E. Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Frontiers in evolutionary neuroscience* **4**, 12 (2012).
13. Nowicki, S. & Searcy, W. A. The evolution of vocal learning. *Current Opinion in Neurobiology* **28**, 48–53, doi:https://doi.org/10.1016/j.conb.2014.06.007 (2014).
14. Pfenning, A. R. *et al.* Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**, 1256846 (2014).
15. Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).
16. Warren, W. C. *et al.* The genome of a songbird. *Nature* **464**, 757–762 (2010).
17. Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
18. 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).
19. Goldstein, R. A., Pollard, S. T., Shah, S. D. & Pollock, D. D. Nonadaptive amino acid convergence rates decrease over time. *Molecular biology and evolution* **32**, 1373–1381 (2015).
20. 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).
21. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv* (2020).
22. Fisher, R. A. *The design of experiments*. (Oliver And Boyd; Edinburgh; London, 1937).
23. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *Journal of molecular biology* **188**, 415–431 (1986).
24. Zou, Z. & Zhang, J. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Molecular biology and evolution* **32**, 2085–2096 (2015).

25. Fox, J. *et al.* Package 'car'. *Vienna: R Foundation for Statistical Computing* (2012).
26. Whitney, O. *et al.* Core and region-enriched networks of behaviorally regulated genes and the singing genome. *Science* **346**, 1256780 (2014).
27. al., G. e. in preparation. (2020).
28. Kubikova, L., Wada, K. & Jarvis, E. D. Dopamine receptors in a songbird brain. *Journal of Comparative Neurology* **518**, 741–769 (2010).
29. da Silva, W. C., Köhler, C. C., Radiske, A. & Cammarota, M. D1/D5 dopamine receptors modulate spatial memory formation. *Neurobiology of learning and memory* **97**, 271–275 (2012).
30. Rangel-Barajas, C., Coronel, I. & Florán, B. Dopamine receptors and neurodegeneration. *Aging and disease* **6**, 349 (2015).
31. O'Dowd, B. F., Nguyen, T., Ji, X. & George, S. R. D5 dopamine receptor carboxyl tail involved in D5-D2 heteromer formation. *Biochemical and biophysical research communications* **431**, 586–589, doi:10.1016/j.bbrc.2012.12.139 (2013).
32. Eicher, J. D. *et al.* Associations of prenatal nicotine exposure and the dopamine related genes ANKK1 and DRD2 to verbal language. *PloS one* **8**, e63762, doi:10.1371/journal.pone.0063762 (2013).
33. Simonyan, K., Horwitz, B. & Jarvis, E. D. Dopamine regulation of human speech and bird song: A critical review. *Brain and Language* **122**, 142–150, doi:https://doi.org/10.1016/j.bandl.2011.12.009 (2012).
34. Lai, C. S., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F. & Monaco, A. P. J. N. A forkhead-domain gene is mutated in a severe speech and language disorder. **413**, 519 (2001).
35. Enard, W. J. C. o. i. n. FOXP2 and the role of cortico-basal ganglia circuits in speech and language evolution. **21**, 415–424 (2011).
36. Berwick, R. C., Friederici, A. D., Chomsky, N. & Bolhuis, J. J. Evolution, brain, and the nature of language. *Trends in cognitive sciences* **17**, 89–98 (2013).
37. Speed, M. P. & Arbuckle, K. Quantification provides a conceptual basis for convergent evolution. *Biological Reviews* **92**, 815–829 (2017).
38. Storz, J. F. Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews Genetics* **17**, 239, doi:10.1038/nrg.2016.11 (2016).
39. Rittschof, C. C. & Robinson, G. E. in *Current Topics in Developmental Biology* Vol. 119 (ed Virginie Orgogozo) 157–204 (Academic Press, 2016).
40. Carroll, S. B. Evolution at two levels: on genes and form. *PLoS Biol* **3**, e245, doi:10.1371/journal.pbio.0030245 (2005).
41. Wong, P. C., Morgan-Short, K., Ettlinger, M. & Zheng, J. J. c. Linking neurogenetics and individual differences in language learning: The dopamine hypothesis. **48**, 1091–1102 (2012).
42. Bando, T. *et al.* Neuronal leucine-rich repeat protein 4 functions in hippocampus-dependent long-lasting memory. *Molecular and cellular biology* **25**, 4166–4175 (2005).
43. Gobes, S. M. H. & Bolhuis, J. J. Birdsong Memory: A Neural Dissociation between Song Recognition and Production. *Current Biology* **17**, 789–793, doi:https://doi.org/10.1016/j.cub.2007.03.059 (2007).
44. Han, S. *et al.* Regulation of dendritic spines, spatial memory, and embryonic development by the TANC family of PSD-95-interacting proteins. *Journal of Neuroscience* **30**, 15102–15112 (2010).
45. Sunahara, R. K. *et al.* Cloning of the gene for a human dopamine D5 receptor with higher affinity for dopamine than D1. *Nature* **350**, 614–619, doi:10.1038/350614a0 (1991).
46. Solberg, R. *et al.* Mapping of the regulatory subunits RI beta and RII beta of cAMP-dependent protein kinase genes on human chromosome 7. *Genomics* **14**, 63–69 (1992).
47. Abe, K., Matsui, S. & Watanabe, D. Transgenic songbirds with suppressed or enhanced activity of CREB transcription factor. *Proceedings of the National Academy of Sciences* **112**, 7599–7604 (2015).
48. Kandel, E. R. The molecular biology of memory: cAMP, PKA, CRE, CREB-1, CREB-2, and CPEB. *Molecular brain* **5**, 14 (2012).
49. Naguib, M. & Riebel, K. in *Biocommunication of animals* 233–247 (Springer, 2014).

50. Mason, N. A. *et al.* Song evolution, speciation, and vocal learning in passerine birds. *Evolution* **71**, 786–796, doi:10.1111/evo.13159 (2017).
51. Liu, W.-c. *et al.* Human mutant huntingtin disrupts vocal learning in transgenic songbirds. *Nature neuroscience* **18**, 1617 (2015).
52. Team, R. C. R: A language and environment for statistical computing. (2013).
53. Frankish, A. *et al.* Ensembl 2018. *Nucleic Acids Research* **46**, D754-D761, doi:10.1093/nar/gkx1098 %J Nucleic Acids Research (2017).

Unsectioned Paragraphs

Authors

Figures

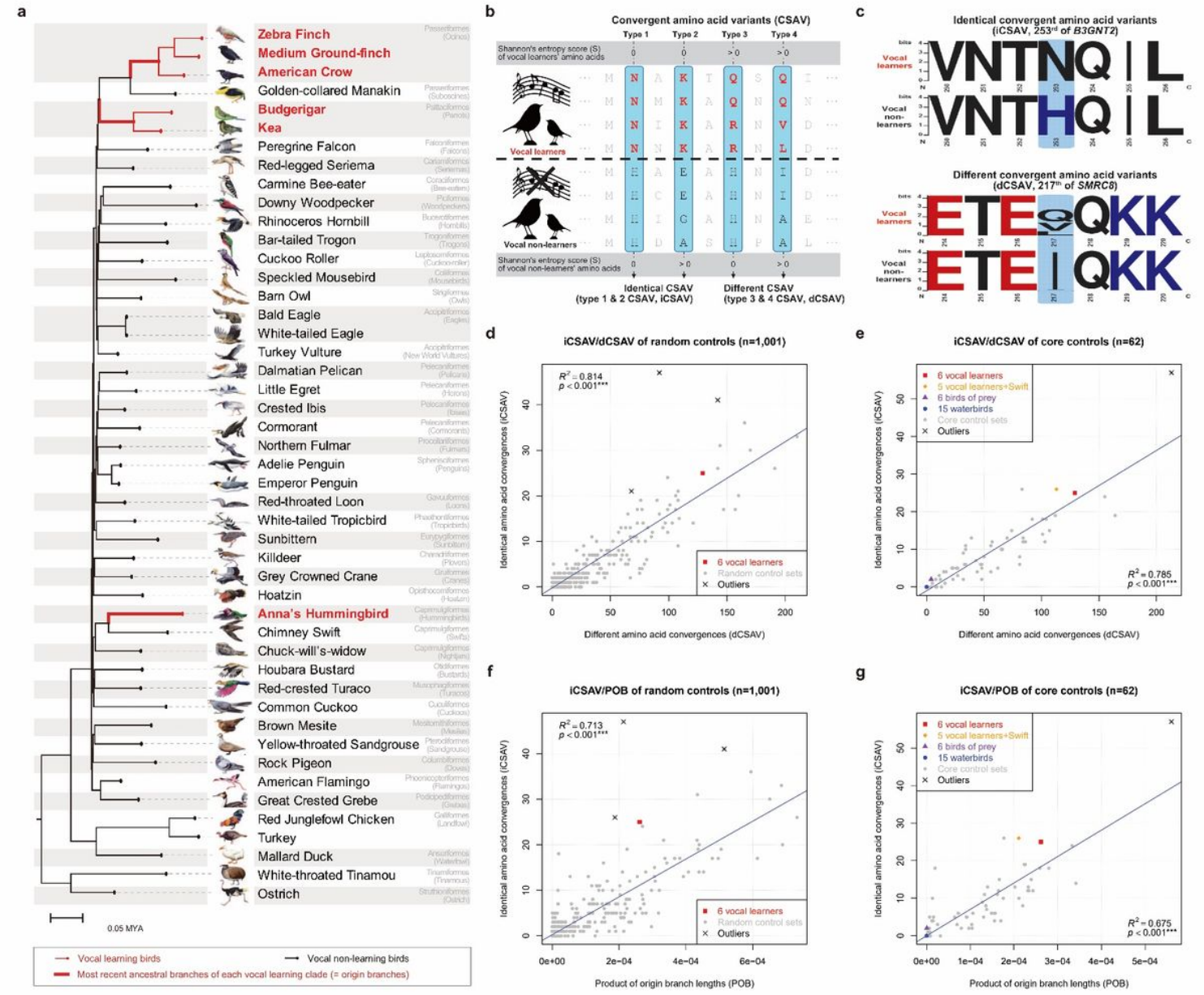

Figure 1

Convergent amino acid substitutions in vocal learning birds compared to controls. (a) Avian family tree and genomes analyzed. The branch lengths of the 48 birds is estimated from the RAXML tree of Jarvis, et al.15 (Supplementary Fig. 1). Red, avian vocal learning lineages. MRCA (origin branch) of each vocal learning clade is indicated as a bold red line. (b) Illustration of the four types of convergent single amino acid variants (CSAV, sky blue-colored boxes) and their expected Shannon entropy scores in the convergent versus control groups of species. (c) Example cases of an identical CSAV (iCSAV, type 1) site in B3GNT2 and a different CSAV (dCSAV, type 3) site in SMRC8. (d, e) Correlation plots between iCSAV (Types 1+2, y-axis) and dCSAV (Types 3+4, x-axis) of random and core control species sets, respectively. (f, g) Correlation plots between iCSAV (Types 1+2) and product of origin branch lengths (POB) of random and core control sets, respectively. Different combinations of species with known convergent traits are indicated by colored symbols. Statistics calculated as a linear regression, with adjusted R2 and p values using the 'lm' function in the R package (ver. 3.5.1)52.

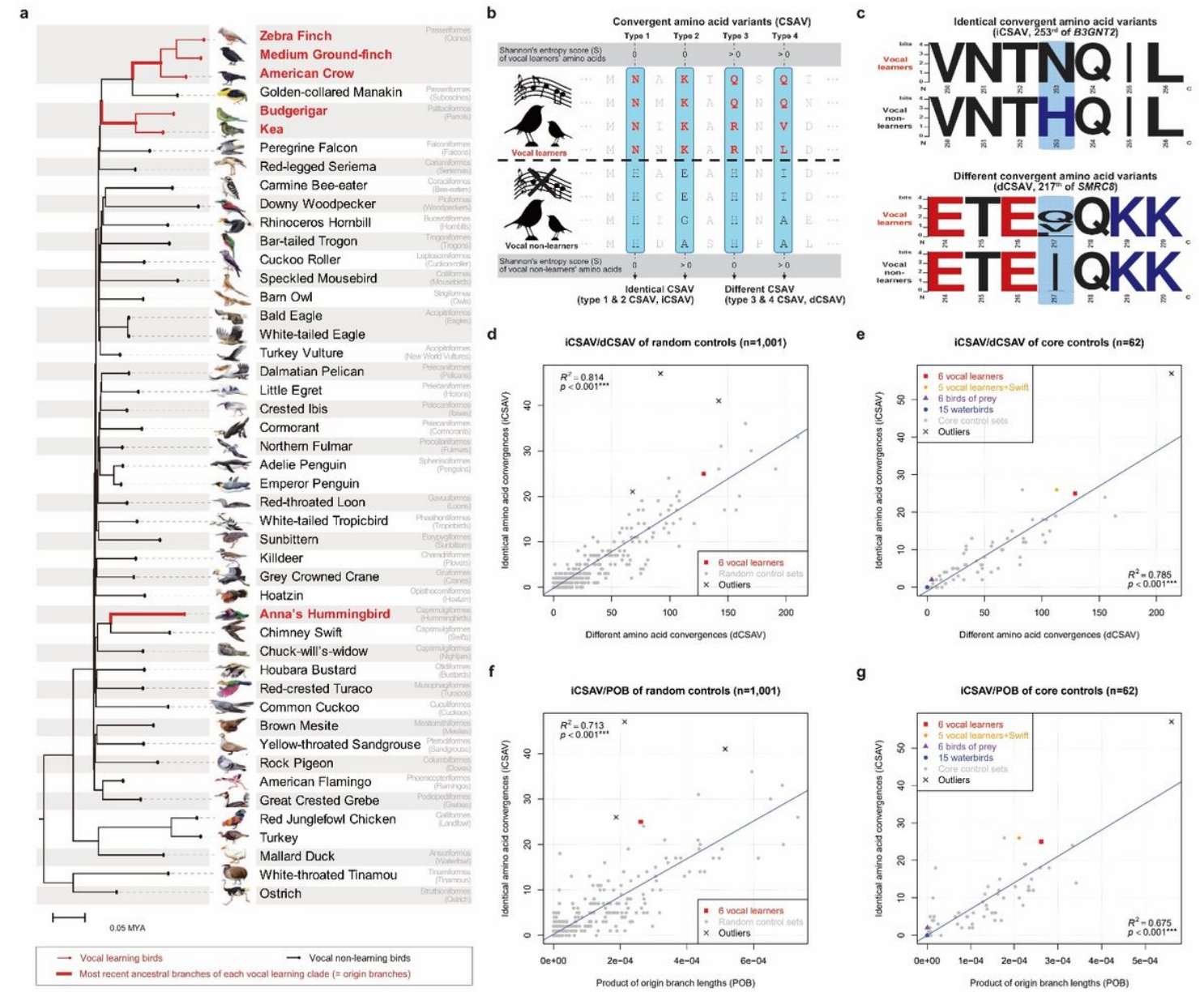

Figure 1

Convergent amino acid substitutions in vocal learning birds compared to controls. (a) Avian family tree and genomes analyzed. The branch lengths of the 48 birds is estimated from the RAXML tree of Jarvis, et al.15 (Supplementary Fig. 1). Red, avian vocal learning lineages. MRCA (origin branch) of each vocal learning clade is indicated as a bold red line. (b) Illustration of the four types of convergent single amino acid variants (CSAV, sky blue-colored boxes) and their expected Shannon entropy

scores in the convergent versus control groups of species. (c) Example cases of an identical CSAV (iCSAV, type 1) site in B3GNT2 and a different CSAV (dCSAV, type 3) site in SMRC8. (d, e) Correlation plots between iCSAV (Types 1+2, y-axis) and dCSAV (Types 3+4, x-axis) of random and core control species sets, respectively. (f, g) Correlation plots between iCSAV (Types 1+2) and product of origin branch lengths (POB) of random and core control sets, respectively. Different combinations of species with known convergent traits are indicated by colored symbols. Statistics calculated as a linear regression, with adjusted R2 and p values using the 'lm' function in the R package (ver. 3.5.1)52.

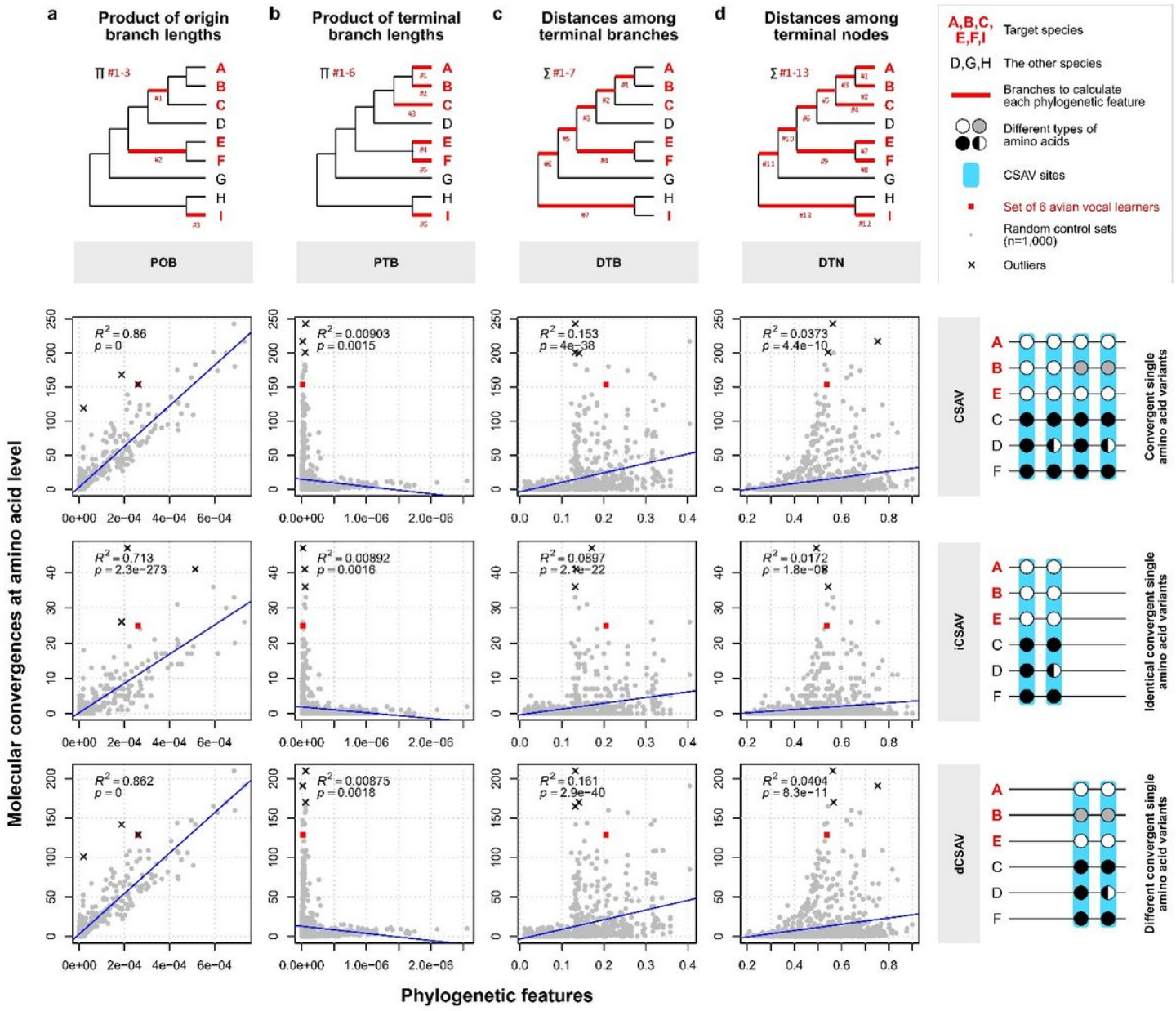

Figure 2

Amino acid convergence amount is correlated to the product of origin branch lengths. Shown are regression analyses of amino acid convergence for (a) total CSAV (Types 1-4), (b) iCSAV (Types 1+2), and (c) dCSAV (Types 3+4) in the vocal learning set and 1,000 random control sets of avian species with four phylogenetic tree features (top row): product of origin branch lengths; product of terminal branch lengths; distance among terminal branches; and distance among terminal nodes. In the example type of tree branches, red lines show the branches used for the calculations and red text the species clades that have a convergent trait. Right, 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 R2 and p value using the 'lm' function in the R package (ver. 3.5.1)52.

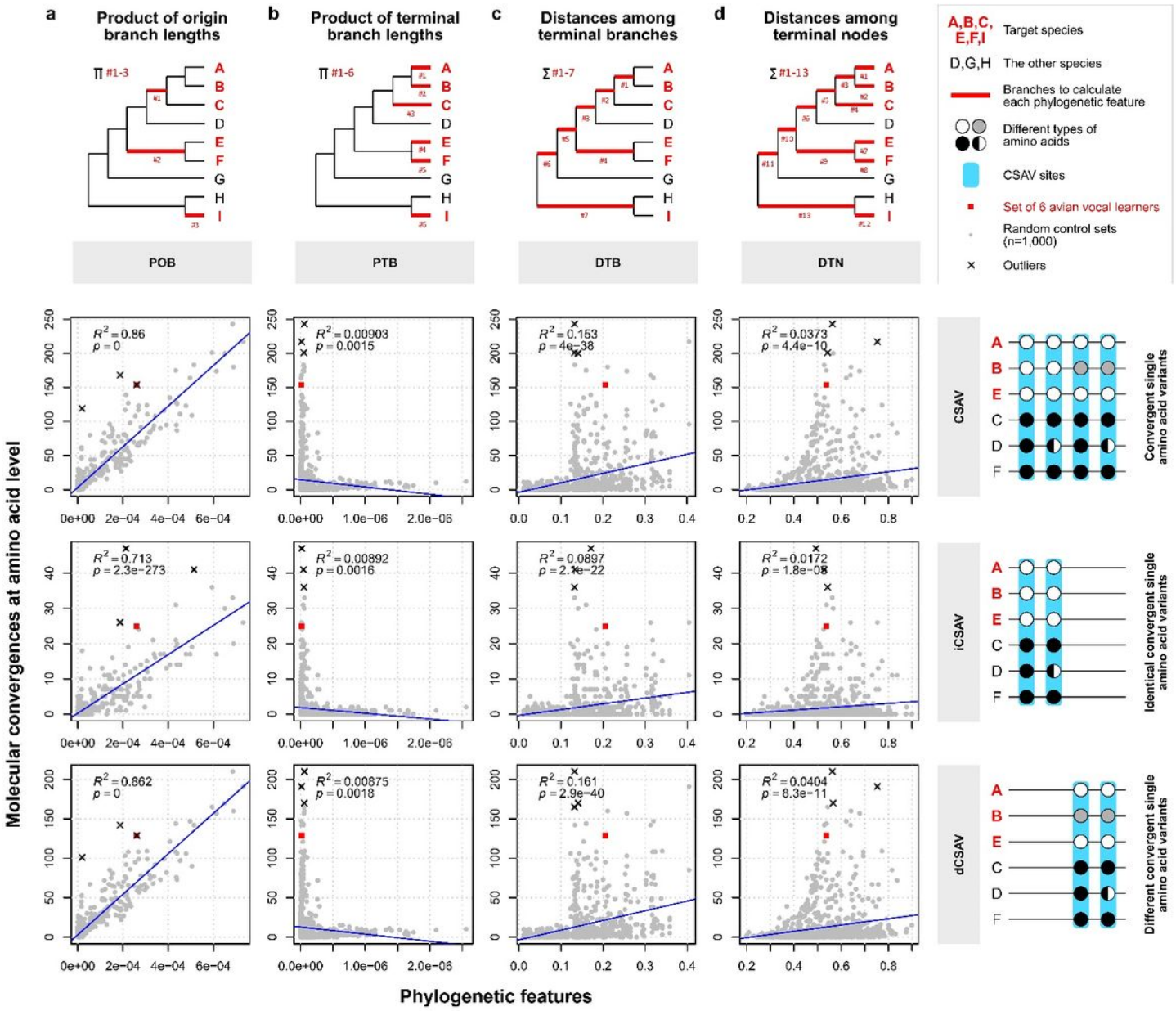

Figure 2

Amino acid convergence amount is correlated to the product of origin branch lengths. Shown are regression analyses of amino acid convergence for (a) total CSAV (Types 1-4), (b) iCSAV (Types 1+2), and (c) dCSAV (Types 3+4) in the vocal learning set and 1,000 random control sets of avian species with four phylogenetic tree features (top row): product of origin branch lengths; product of terminal branch lengths; distance among terminal branches; and distance among terminal nodes. In the example type of tree branches, red lines show the branches used for the calculations and red text the species clades that have a convergent trait. Right, 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 R2 and p value using the 'lm' function in the R package (ver. 3.5.1)52.

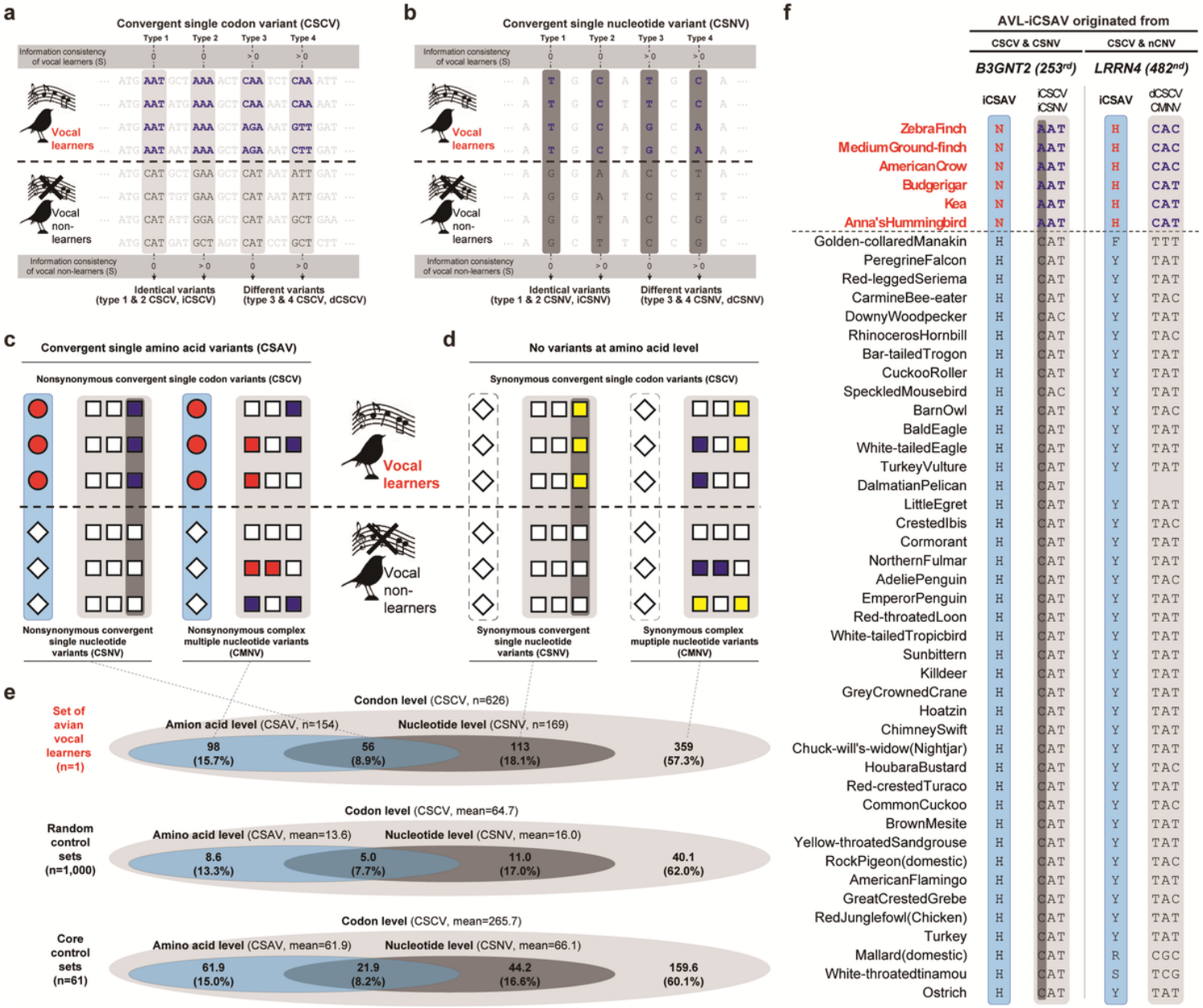

Figure 3

Nucleotide substitution source of convergent amino acid substitutions. (a) Concept of four types of convergent single codon variants (CSCV). (b) Concept of four types of convergent single nucleotide variants (CSNV). Explanation of Shannon's entropy is similar as in Fig. 1b 23. (c) Concept of convergent single amino acid variants (CASV) explained by CSCV. Left case, CSAVs caused by CSCVs with CSNV at a homologous nucleotide site. Right case, CSAVs caused by CSCVs with multiple nucleotide variants at different sites (Non-SNVs). (d) Concept of CSCV explained by synonymous substitutions between species, which are those do not cause amino acid changes. Left case, synonymous CSCVs with CSNV at a homologous nucleotide site. Right case, CSCVs with multiple nucleotide variants at different sites (Non-CSNVs). (e) Venn diagrams of the different subsets of CSCV caused by the four types of nucleotide substitutions outlined in (c) and (d), in avian vocal learners, random control sets of species, and the core control set. (f) Examples of identical amino acid convergences among vocal learners (Type 1 AVL-iCSAVs) originating from identical CSNVs at the same site (in B3GNT2) or non-CSNV complex nucleotide variants at multiple sites (in LRRN4). Red text, avian vocal learners. Sky blue boxes, sites with CSAVs; dark grey box, CSNV; light grey boxes, CSCVs.

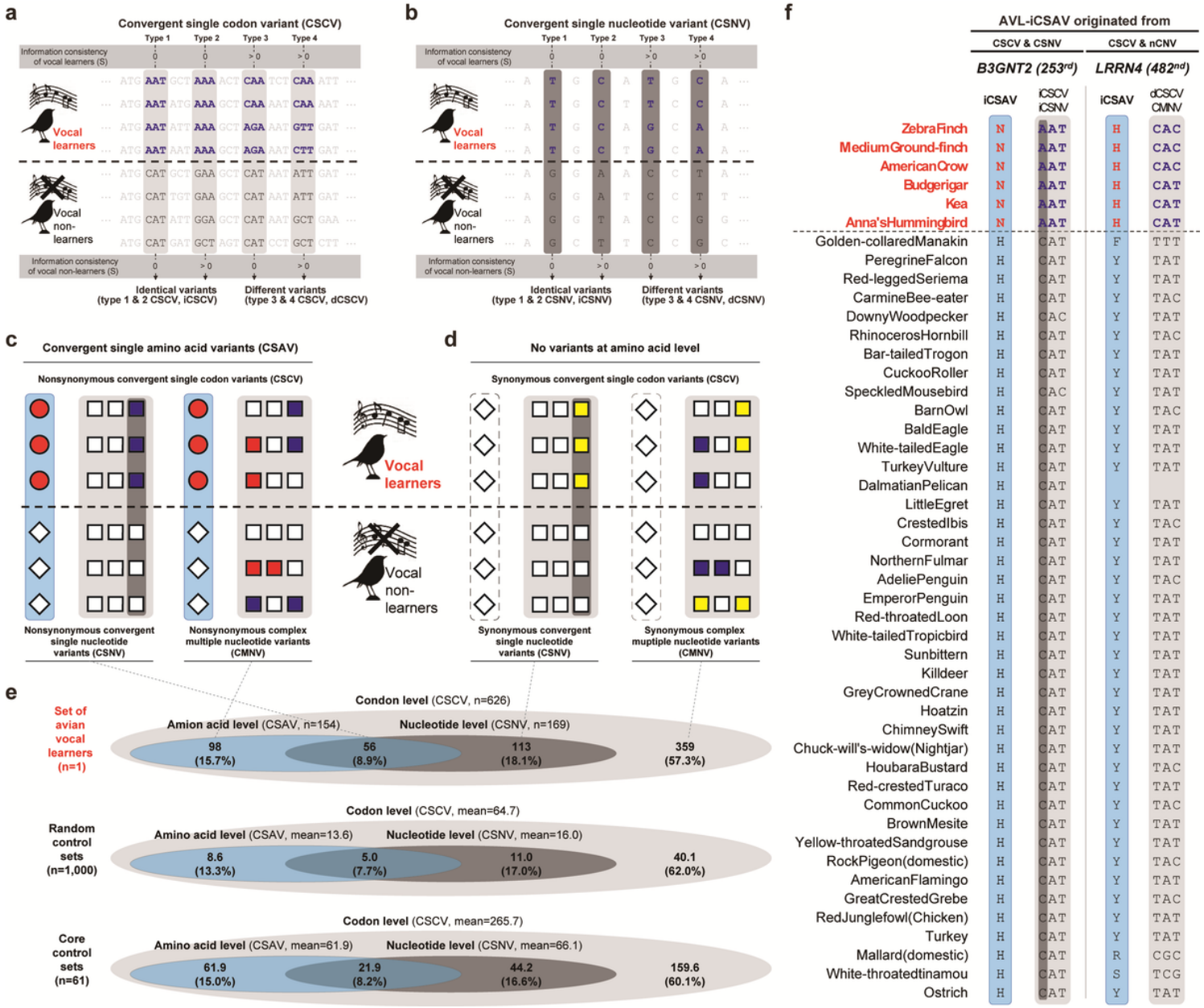

Figure 3

Nucleotide substitution source of convergent amino acid substitutions. (a) Concept of four types of convergent single codon variants (CSCV). (b) Concept of four types of convergent single nucleotide variants (CSNV). Explanation of Shannon's entropy is similar as in Fig. 1b 23. (c) Concept of convergent single amino acid variants (CASV) explained by CSCV. Left case, CSAVs caused by CSCVs with CSNV at a homologous nucleotide site. Right case, CSAVs caused by CSCVs with multiple nucleotide variants at different sites (Non-SNVs). (d) Concept of CSCV explained by synonymous substitutions between species, which are those do not cause amino acid changes. Left case, synonymous CSCVs with CSNV at a homologous nucleotide site. Right case, CSCVs with multiple nucleotide variants at different sites (Non-CSNVs). (e) Venn diagrams of the different subsets of CSCV caused by the four types of nucleotide substitutions outlined in (c) and (d), in avian vocal learners, random control sets of species, and the core control set. (f) Examples of identical amino acid convergences among vocal learners (Type 1 AVL-iCSAVs) originating from identical CSNVs at the same site (in B3GNT2) or non-CSNV complex nucleotide variants at multiple sites (in LRRN4). Red text, avian vocal learners. Sky blue boxes, sites with CSAVs; dark grey box, CSNV; light grey boxes, CSCVs.

Figure 4

Correlations among convergent nucleotide variants, amino acid variants, and phylogenetic features of random control sets. p values and Adjusted R2 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 in lower diagonal matrices. 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. Correlations of core control sets are shown in Supplementary Figure 4.

Figure 4

Correlations among convergent nucleotide variants, amino acid variants, and phylogenetic features of random control sets. p values and Adjusted R2 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 in lower diagonal matrices. 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. Correlations of core control sets are shown in Supplementary Figure 4.

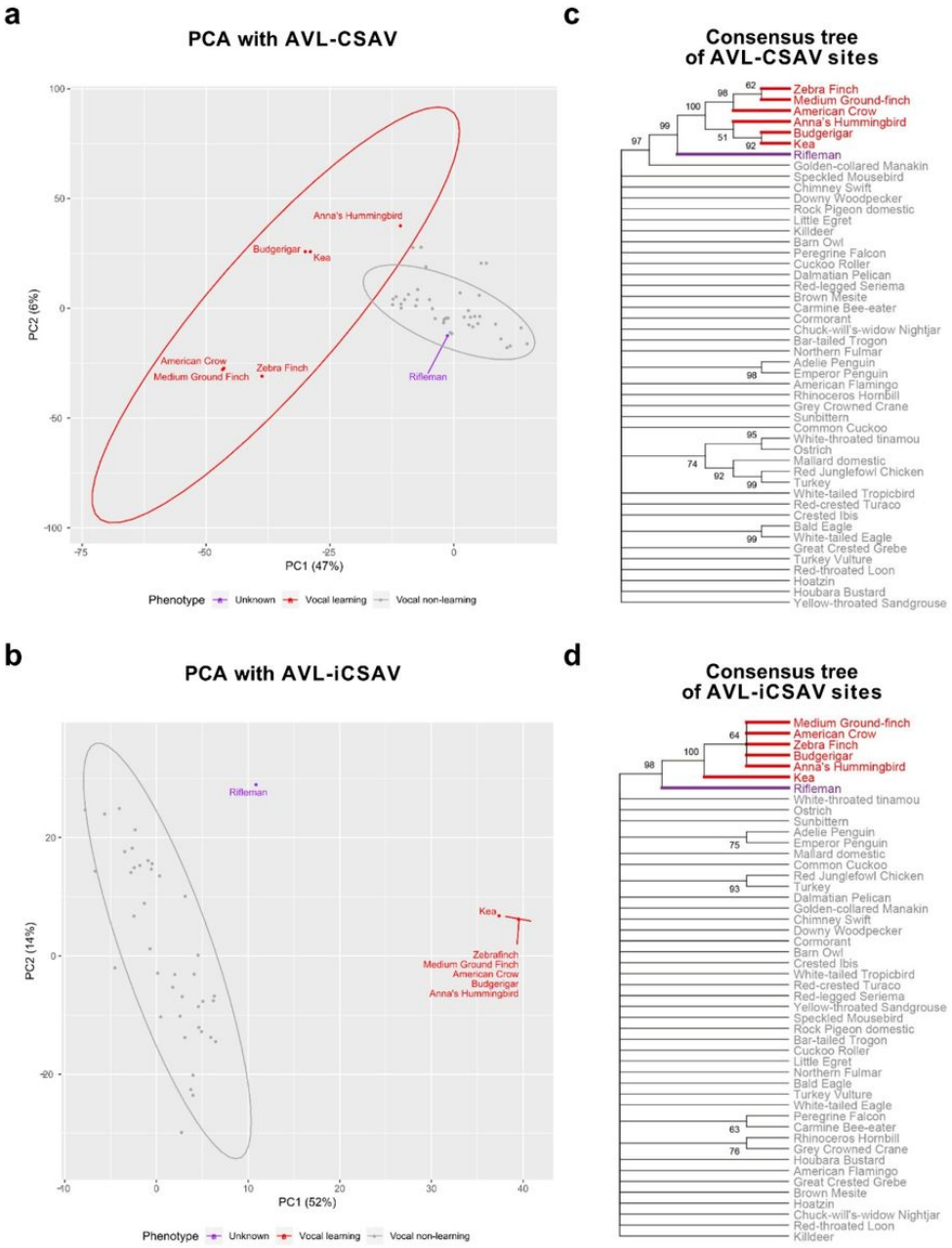

Figure 5

Rifleman amino acid profile similar to vocal non-learners. (a) Principle component analysis (PCA) of all four types of AVL-CSAV sites. (b) PCA of identical (Type 1+2) AVL-iCSAV sites. (c) Consensus tree based on AVL-CSAV sites. (d) Consensus tree based on AVL-iCSAV sites. Red, avian vocal learners; Grey, avian vocal non-learners; Purple, rifleman.

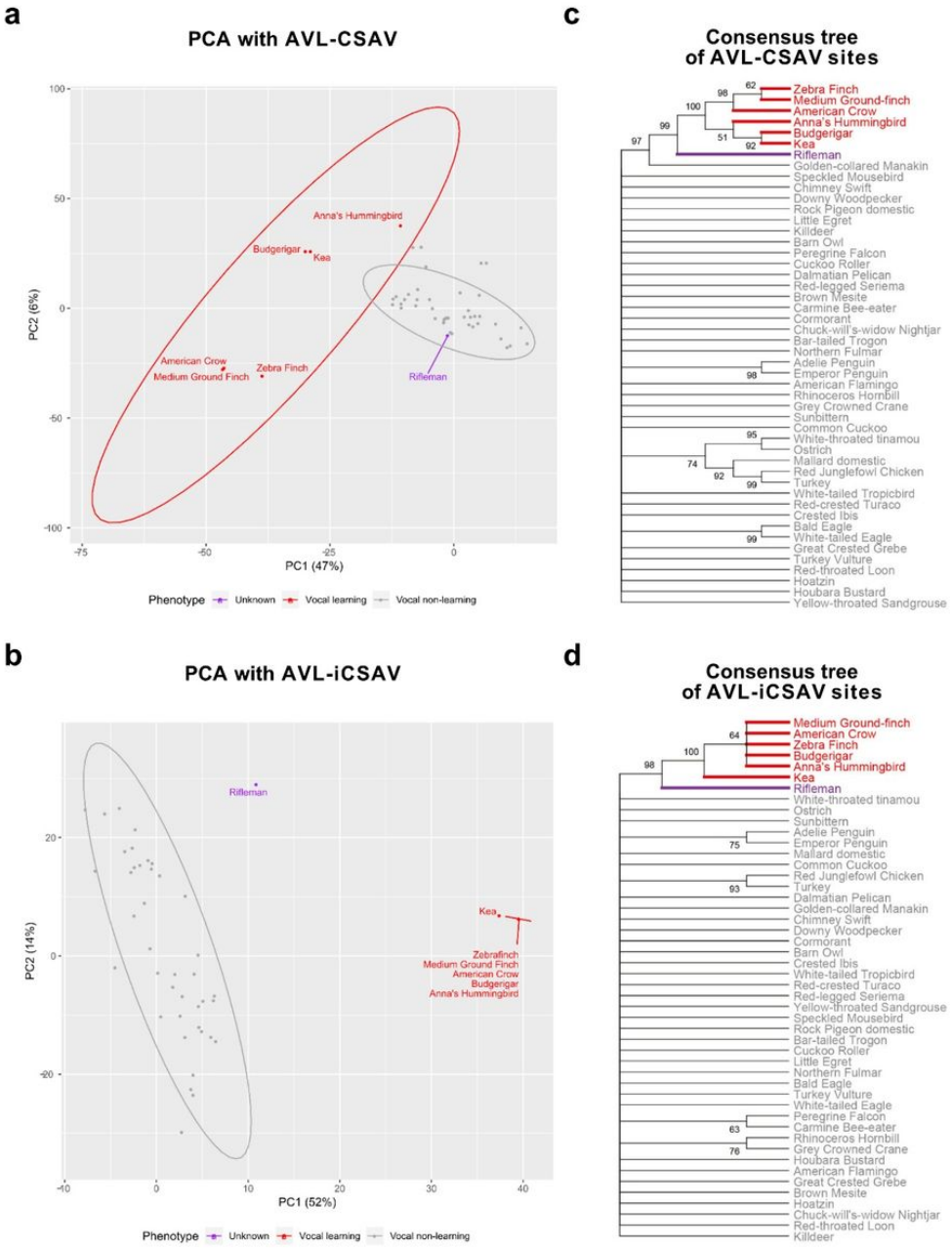

Figure 5

Rifleman amino acid profile similar to vocal non-learners. (a) Principle component analysis (PCA) of all four types of AVL-CSAV sites. (b) PCA of identical (Type 1+2) AVL-iCSAV sites. (c) Consensus tree based on AVL-CSAV sites. (d) Consensus tree based on AVL-iCSAV sites. Red, avian vocal learners; Grey, avian vocal non-learners; Purple, rifleman.

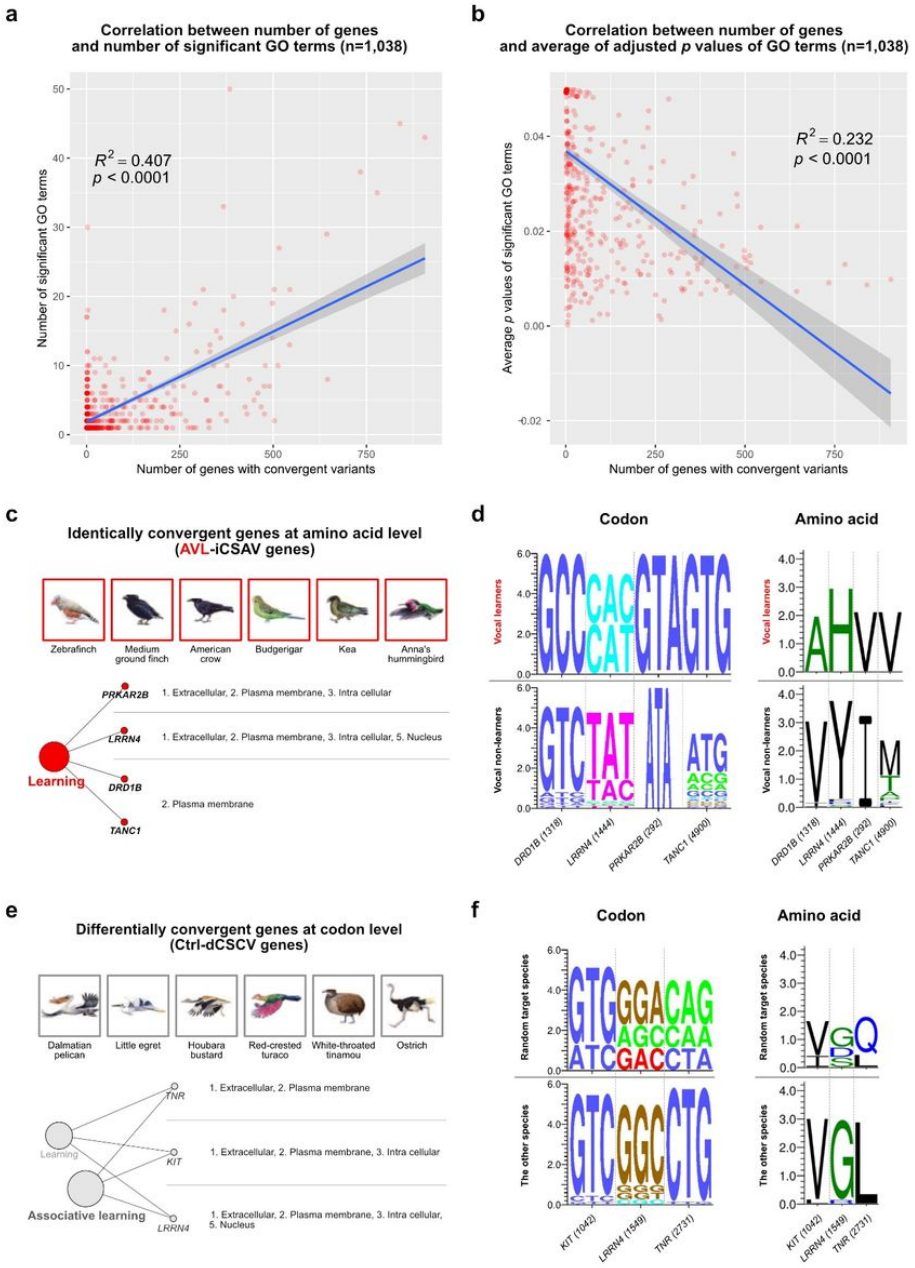

Figure 6

Functional ontology of genes with convergent substitutions. (a) Correlation plot between the number of significantly enriched GO terms and the number of genes with convergent variant in each species data set (n=1,038) with 9 types (CSAV, iCSAV, dCSAV, CSCV, iCSCV, dCSCV, CSNV, iCSNV, and dCSNV). (b) Correlation plot between averages of p-values of significant GO terms and the number of genes with convergent sites of each species data set. Regression lines are shown with confidence interval (>0.95). Adjusted R2 and p value were calculated by 'lm' function in the R package (ver. 3.5.1)52. (c) Gene ontology analysis for learning associated genes with AVL-iCSAV (adj. p < 0.05). (d) Codon and amino acid logos of learning associated genes with AVL-iCSAV. (e) Gene ontology analysis in the only other species combination with convergent learning associated genes, a random Ctrl-dCSCV set (adj. p < 0.05). (f) Codon and amino acid logos of the learning associated genes in the random Ctrl-dCSCV set.

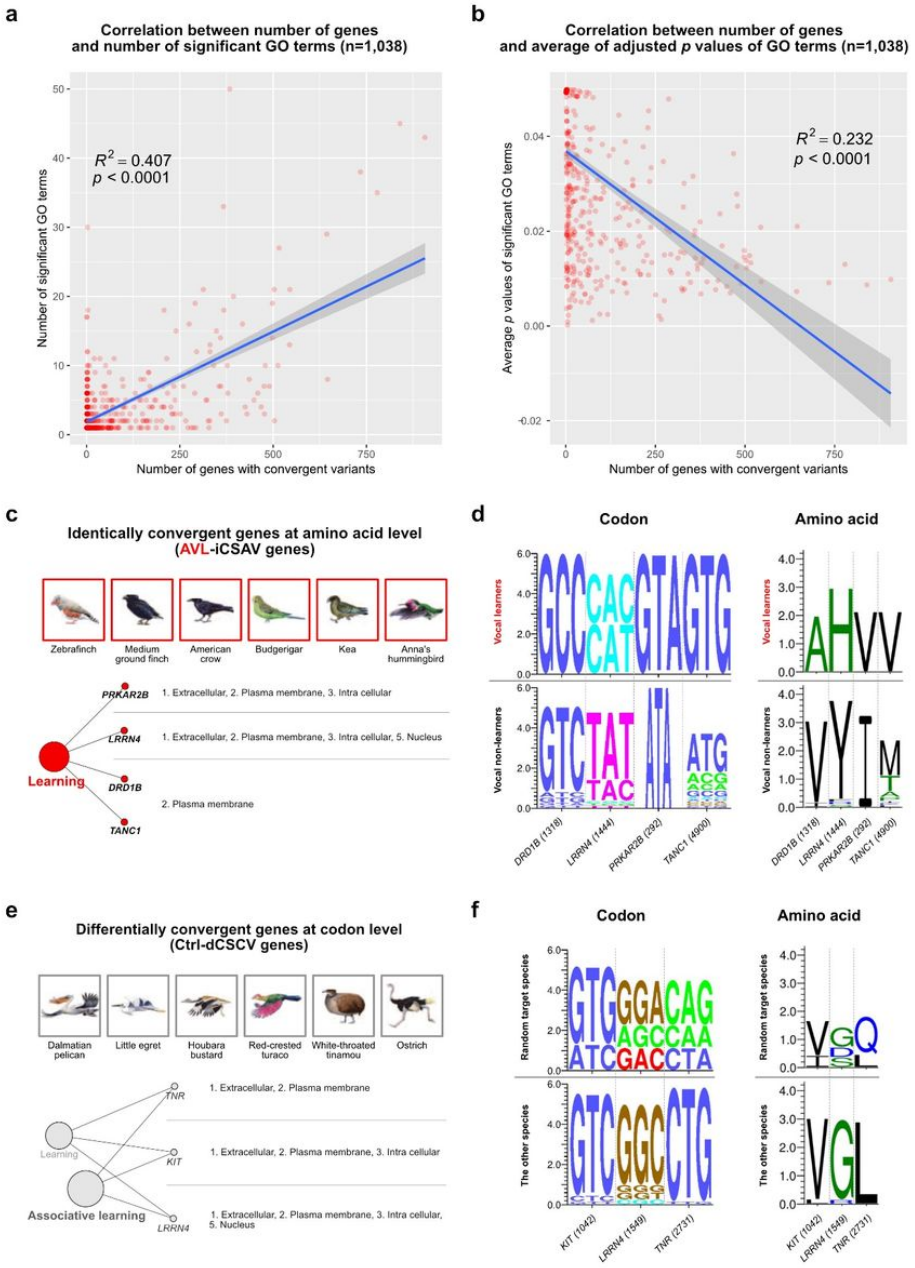

Figure 6

Functional ontology of genes with convergent substitutions. (a) Correlation plot between the number of significantly enriched GO terms and the number of genes with convergent variant in each species data set (n=1,038) with 9 types (CSAV, iCSAV, dCSAV, CSCV, iCSCV, dCSCV, CSNV, iCSNV, and dCSNV). (b) Correlation plot between averages of p-values of significant GO terms and the number of genes with convergent sites of each species data set. Regression lines are shown with confidence interval (>0.95). Adjusted R2 and p value were calculated by 'lm' function in the R package (ver. 3.5.1)52. (c) Gene ontology analysis for learning associated genes with AVL-iCSAV (adj. p < 0.05). (d) Codon and amino acid logos of learning associated genes with AVL-iCSAV. (e) Gene ontology analysis in the only other species combination with convergent learning associated genes, a random Ctrl-dCSCV set (adj. p < 0.05). (f) Codon and amino acid logos of the learning associated genes in the random Ctrl-dCSCV set.

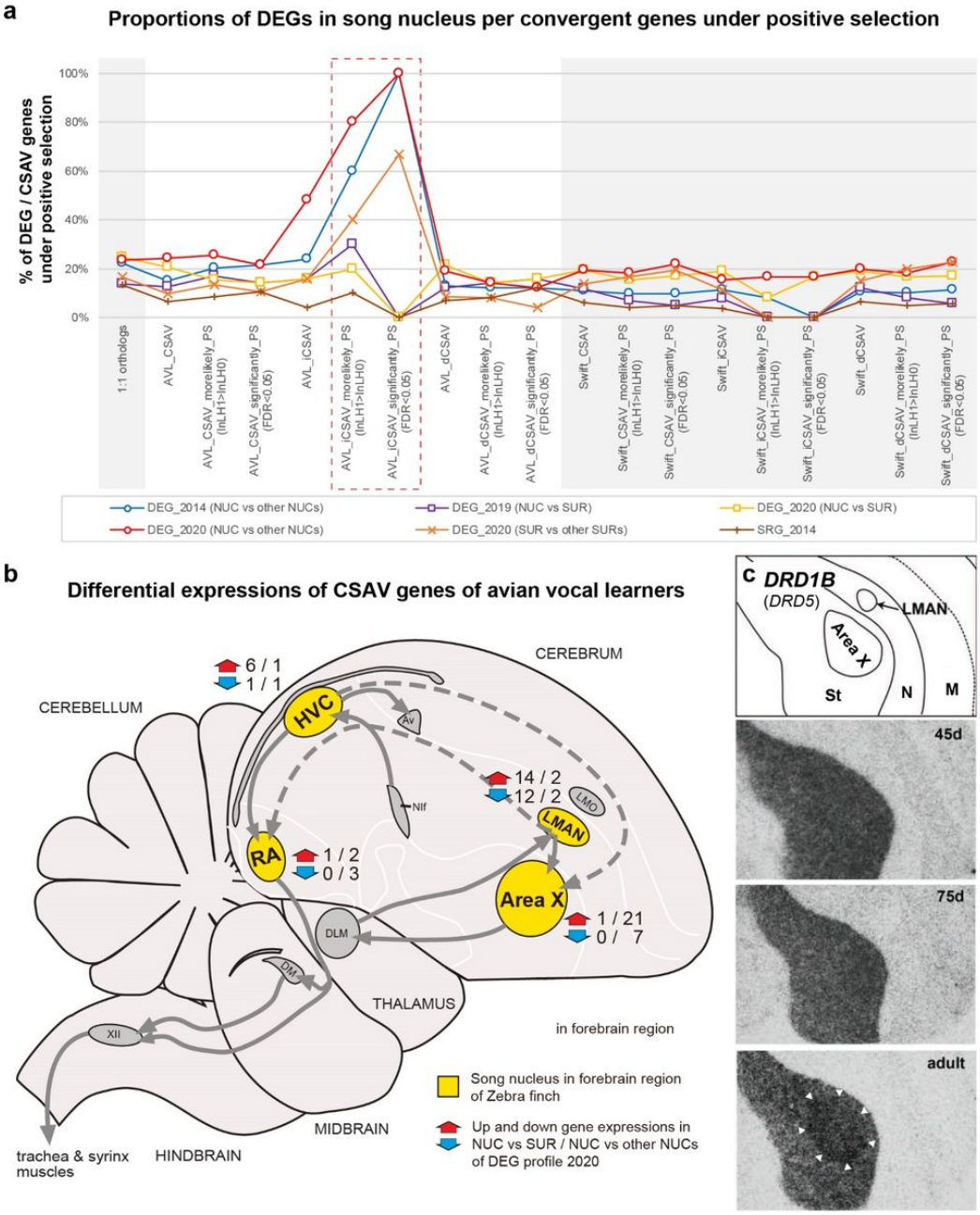

Figure 7

Sequence convergence, positive selection, and specialized gene expression. (a) Proportions of singing regulated genes (SRG) or differentially expressed genes (DEG) in song learning nuclei or adjacent brain subdivisions of the zebra finch brain (y-axis) that have convergent amino acid coding sequences and have been positively selected, in vocal learners and the closest control set of species (x-axis). DEGs collected from three independent sources based on microarray14,25, micro-dissected RNA sequencing27, and laser capture microscope RNA sequencing27 data sets from 2014, 2019, and 2020 analyses, respectively (Supplementary Note 11). NUC vs SUR: song nucleus compared to its surrounding non-vocal motor brain regions. NUC vs NUC: a song nucleus compared to another song nucleus. SUR vs SUR: a surrounding region of a song nucleus compared to another song nucleus. (b) Songbird brain diagram showing the song learning system. Yellow, forebrain song learning brain regions with SRG and DEGs measured. Grey, other song learning nuclei. Grey arrows, connections between the song nuclei. Red-up arrow and blue-down arrow indicates number of AVL-iCSAV genes supported by differential expressions in 'NUC vs other NUCs' and 'SUR vs other SURs' comparisons / in the DEG_2020 data source. (c) DRD1B mRNA expression pattern in zebra finch Area X and surrounding striatum at 3 different development time points, with specialized expression (white arrows) appearing by adulthood. Image used with permission from Kubikova et al.28.

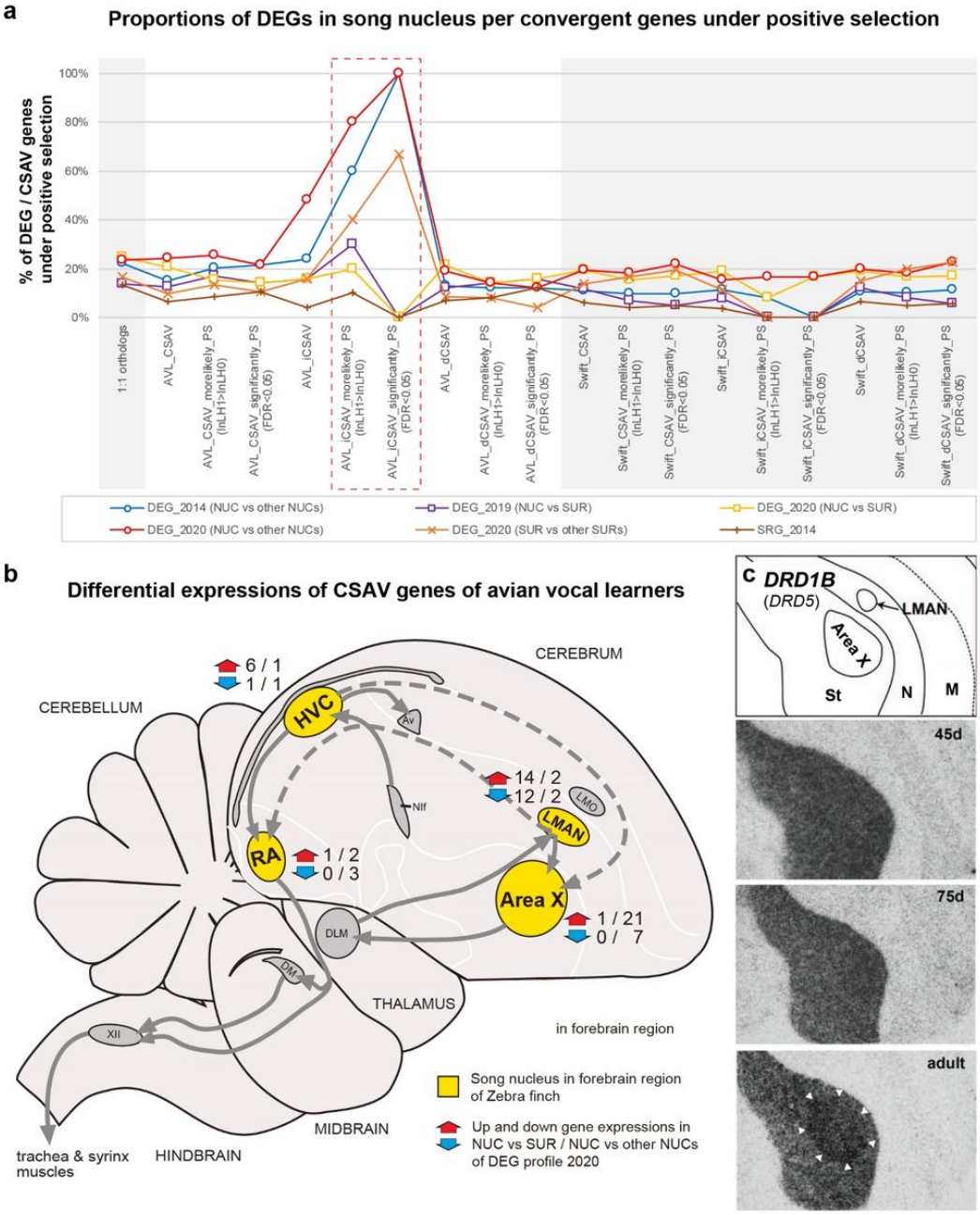

Figure 7

Sequence convergence, positive selection, and specialized gene expression. (a) Proportions of singing regulated genes (SRG) or differentially expressed genes (DEG) in song learning nuclei or adjacent brain subdivisions of the zebra finch brain (y-axis) that have convergent amino acid coding sequences and have been positively selected, in vocal learners and the closest control set of species (x-axis). DEGs collected from three independent sources based on microarray14,25, micro-dissected RNA sequencing27, and laser capture microscope RNA sequencing27 data sets from 2014, 2019, and 2020 analyses, respectively (Supplementary Note 11). NUC vs SUR: song nucleus compared to its surrounding non-vocal motor brain regions. NUC vs NUC: a song nucleus compared to another song nucleus. SUR vs SUR: a surrounding region of a song nucleus compared to another song nucleus. (b) Songbird brain diagram showing the song learning system. Yellow, forebrain song learning brain regions with SRG and DEGs measured. Grey, other song learning nuclei. Grey arrows, connections between the song nuclei. Red-up arrow and blue-down arrow indicates number of AVL-iCSAV genes supported by differential expressions in 'NUC vs other NUCs' and 'SUR vs other SURs' comparisons / in the DEG_2020 data source. (c) DRD1B mRNA expression pattern in zebra finch Area X and surrounding striatum at 3 different development time points, with specialized expression (white arrows) appearing by adulthood. Image used with permission from Kubikova et al.28.

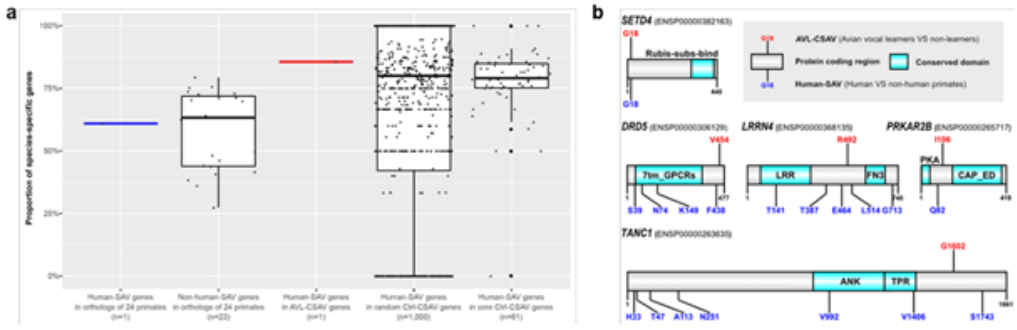

Figure 8

Proportion of genes with human specific amino acid substitutions. Human-SAV, is % of genes with human specific amino acid substitutions compared to 24 non-human primates (n=18,852 total genes analyzed). Non-human primate SAV, is % of genes with non-human primate specific amino acid substitutions in each of 23 non-human primates compared to all other primates including human. Human SAV in AVL-CSAV, is % of genes with avian vocal learner-specific substitutions that overlap with genes containing human-specific substitutions (n=107 out of 124 orthologs identified between vocal learning birds and humans). Human-SAV in random Ctrl-CSAV, same type of results but for 1000 random control sets of avian species. Human-SAV in core Ctrl-CSAV, same type of results but for 61 core control sets of avian species.

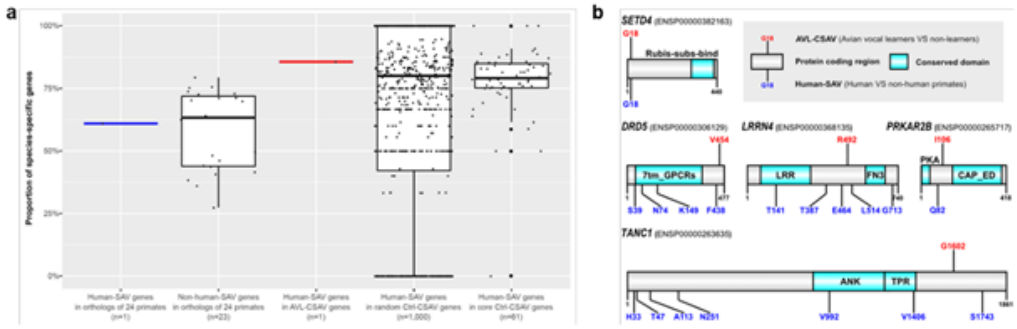

Figure 8

Proportion of genes with human specific amino acid substitutions. Human-SAV, is % of genes with human specific amino acid substitutions compared to 24 non-human primates (n=18,852 total genes analyzed). Non-human primate SAV, is % of genes with non-human primate specific amino acid substitutions in each of 23 non-human primates compared to all other primates including human. Human SAV in AVL-CSAV, is % of genes with avian vocal learner-specific substitutions that overlap with genes containing human-specific substitutions (n=107 out of 124 orthologs identified between vocal learning birds and humans). Human-SAV in random Ctrl-CSAV, same type of results but for 1000 random control sets of avian species. Human-SAV in core Ctrl-CSAV, same type of results but for 61 core control sets of avian species.

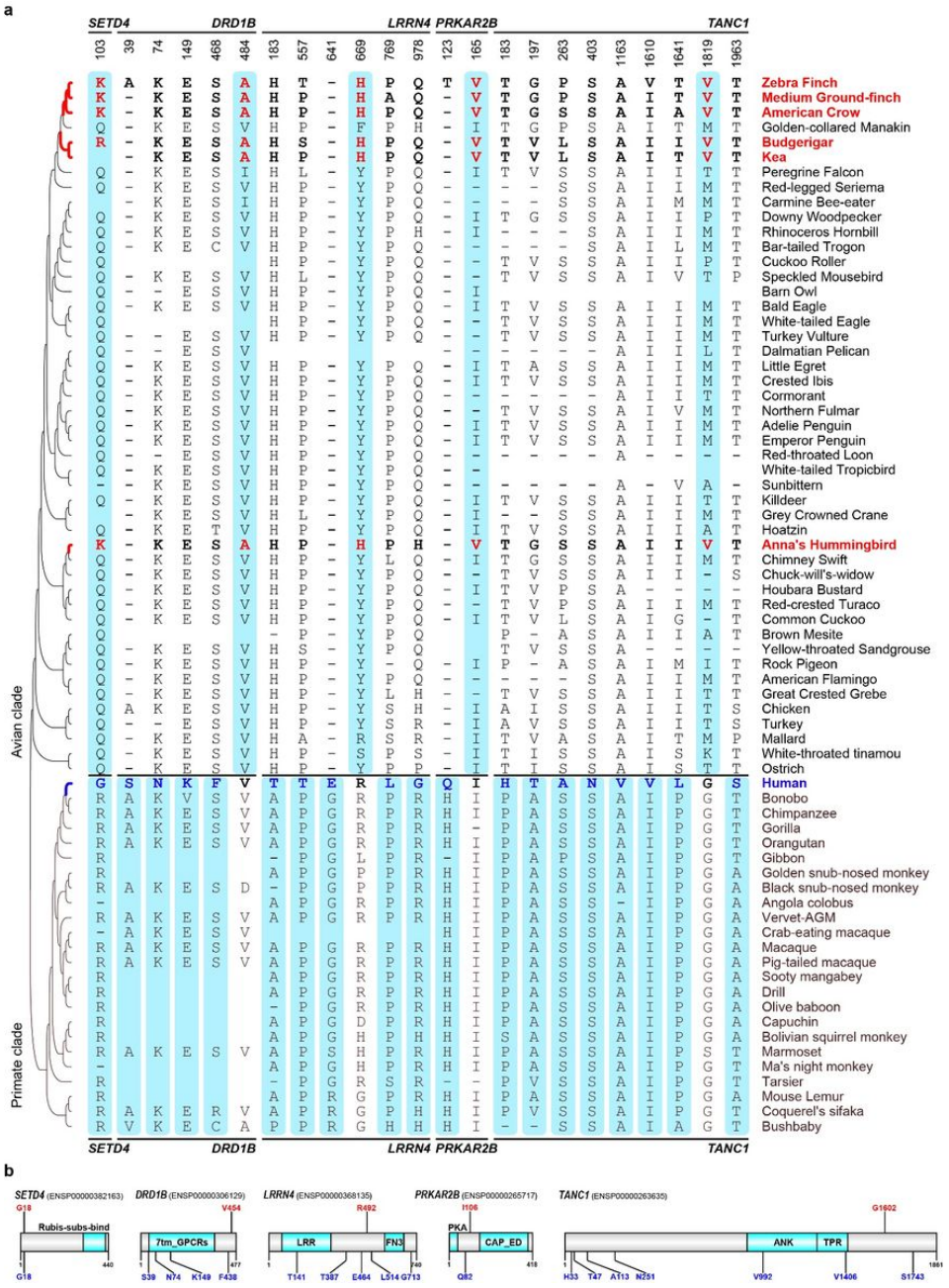

Figure 9

Candidate genes with avian vocal learner and human specific substitutions. (a) Red texts, vocal learning bird-specific amino acid substitutions compared to vocal non-learning birds. Blue texts, human-specific amino acid substitutions compared to non-human primates. Amino acid of avian vocal learners and human marked as red and blue bold letters. Cladogram of birds and primates from the Jarvis et al tree15 and Ensembl tree53, respectively. (b) Amino acid positions of avian vocal learner-specific amino acid substitutions in zebra finch and human-specific amino acid substitutions in human in the same genes. Rubis-subs-bind: Rubisco LSMT substrate-binding. 7tm_GPCRs: seven-transmembrane G protein-coupled receptor superfamily. LRR: Leucine-rich repeat (LRR) protein. FN3: Fibronectin type 3 domain. PKA: Dimerization/Docking domain of the Type II beta Regulatory subunit of cAMP-dependent protein kinase. CAP_ED: effector domain of the CAP family of transcription factors. ANK: ankyrin repeats, and TPR: Tetratricopeptide (TPR) repeat.

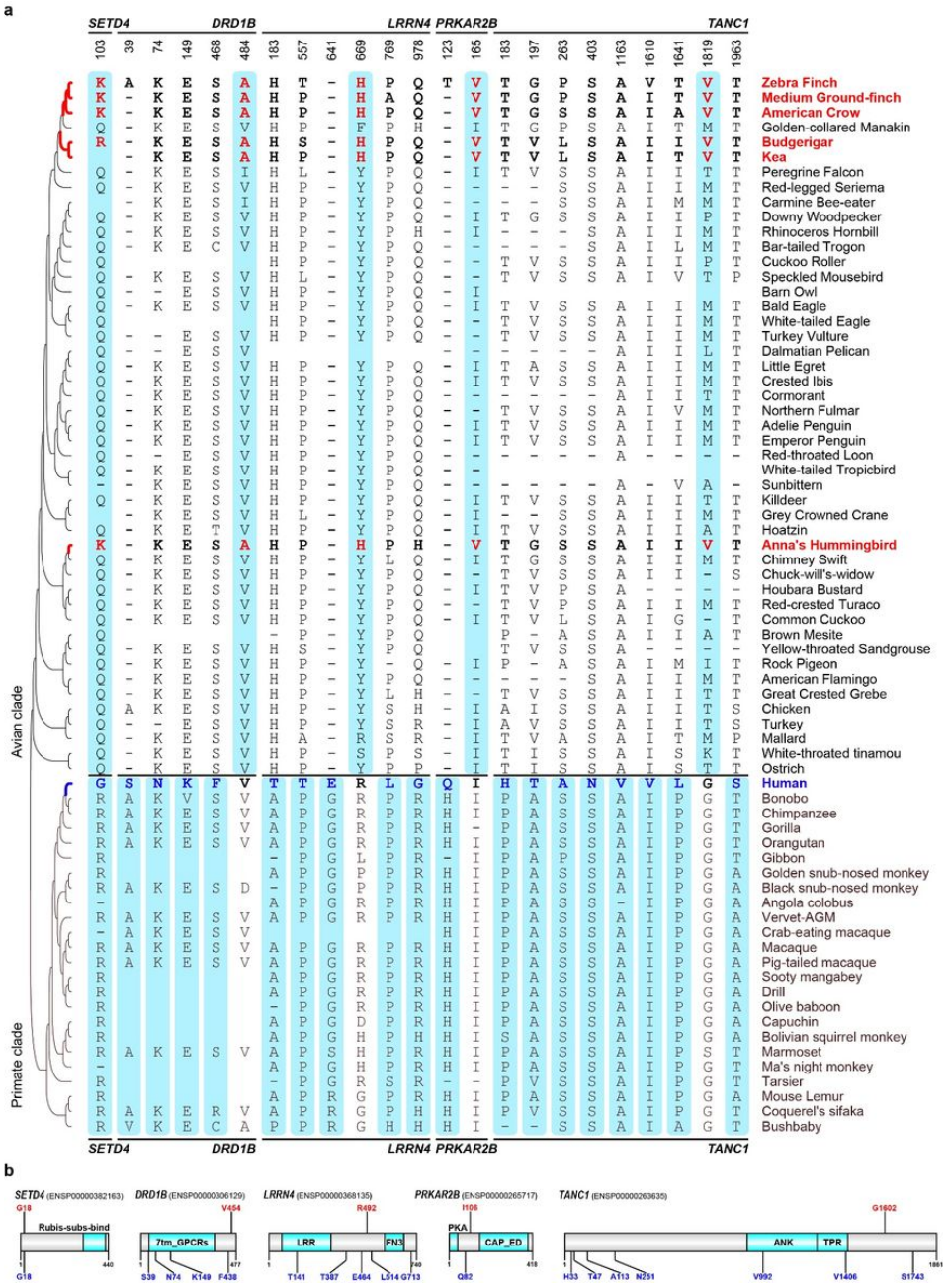

Figure 9

Candidate genes with avian vocal learner and human specific substitutions. (a) Red texts, vocal learning bird-specific amino acid substitutions compared to vocal non-learning birds. Blue texts, human-specific amino acid substitutions compared to non-human primates. Amino acid of avian vocal learners and human marked as red and blue bold letters. Cladogram of birds and primates from the Jarvis et al tree15 and Ensembl tree53, respectively. (b) Amino acid positions of avian vocal learner-specific amino acid substitutions in zebra finch and human-specific amino acid substitutions in human in the same genes. Rubis-subs-bind: Rubisco LSMT substrate-binding. 7tm_GPCRs: seven-transmembrane G protein-coupled receptor superfamily. LRR: Leucine-rich repeat (LRR) protein. FN3: Fibronectin type 3 domain. PKA: Dimerization/Docking domain of the Type II beta Regulatory subunit of cAMP-dependent protein kinase. CAP_ED: effector domain of the CAP family of transcription factors. ANK: ankyrin repeats, and TPR: Tetratricopeptide (TPR) repeat.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- VocalLearningpaperv151SupplementaryInformation20201029.docx
- VocalLearningpaperv151SupplementaryInformation20201029.docx
- VocalLearningpaperv151SupplementaryData120201029.xlsx
- VocalLearningpaperv151SupplementaryData120201029.xlsx
- VocalLearningpaperv151SupplementaryData220201029.xlsx
- VocalLearningpaperv151SupplementaryData220201029.xlsx
- VocalLearningpaperv151SupplementaryData320201029.xlsx
- VocalLearningpaperv151SupplementaryData320201029.xlsx