OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae614
https://doi.org/10.1093/bib/bbae614
Case Study

A versatile pipeline to identify convergently lost ancestral conserved fragments associated with convergent evolution of vocal learning

Xiaoyi Li[1,2,3], Kangli Zhu[3], Ying Zhen[2,3,4,*]

[1]School of Life Sciences, Fudan University, 220 Handan Road, Yangpu District, Shanghai 200433, China
[2]Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences and Research Center for Industries of the Future, Westlake University, 600 Dunyu Road, Xihu District, Hangzhou, Zhejiang 310030, China
[3]Westlake Laboratory of Life Sciences and Biomedicine, 600 Dunyu Road, Xihu District, Hangzhou, Zhejiang 310030, China
[4]Institute of Biology, Westlake Institute for Advanced Study, 18 Shilongshan Road, Xihu District, Hangzhou, Zhejiang 310024, China

[*]Corresponding author. Tel: 0086-17764573607. E-mail: zhenying@westlake.edu.cn

Abstract

Molecular convergence in convergently evolved lineages provides valuable insights into the shared genetic basis of converged phenotypes. However, most methods are limited to coding regions, overlooking the potential contribution of regulatory regions. We focused on the independently evolved vocal learning ability in multiple avian lineages, and developed a whole-genome-alignment-free approach to identify genome-wide Convergently Lost Ancestral Conserved fragments (CLACs) in these lineages, encompassing noncoding regions. We discovered 2711 CLACs that are overrepresented in noncoding regions. Proximal genes of these CLACs exhibit significant enrichment in neurological pathways, including glutamate receptor signaling pathway and axon guidance pathway. Moreover, their expression is highly enriched in brain tissues associated with speech formation. Notably, several have known functions in speech and language learning, including ROBO family, SLIT2, GRIN1, and GRIN2B. Additionally, we found significantly enriched motifs in noncoding CLACs, which match binding motifs of transcriptional factors involved in neurogenesis and gene expression regulation in brain. Furthermore, we discovered 19 candidate genes that harbor CLACs in both human and multiple avian vocal learning lineages, suggesting their potential contribution to the independent evolution of vocal learning in both birds and humans.

Keywords: vocal learning; conserved fragments; glutamate receptor pathway; axon guidance pathway; molecular convergence

Introduction

Convergent evolution is the independent evolution of similar traits across distantly related lineages. Convergence at molecular level of species with convergently evolved traits provides a great opportunity to dissect the genetic basis of such traits and to comprehend the constraints of evolution. However, identifying molecular convergence has been challenging. Previous efforts primarily focused on protein coding regions, identifying convergence of specific amino acid substitutions, amino acid profile changes, or relative evolution rate changes in coding regions [1–3]. Regulatory evolution could have significant contribution to phenotypic evolution, however, due to relatively poor annotation of regulatory elements in noncoding regions and the less conservative nature of noncoding regions, few methods could examine molecular convergence including noncoding regions.

Vocal learning is the unique ability to modify and acquire new sounds through imitation, and plays a crucial role in communication and social interaction [4–7]. Vocal learning ability convergently evolved in limited lineages of birds and mammals, among which songbirds and humans are the most advanced vocal learners [8]. The learning and production of vocalization is a complex process, involving many specialized brain regions, which are referred to as song nuclei and as song system collectively. Non-vocal learners such as chicken and chimpanzee lack or have less-developed song systems [9–12]. In human, several brain regions are essential for vocalization, including Broca's area, laryngeal motor cortex, frontal gyrus, and superior gyrus [13–15]. Song system of zebra finch has also been extensively studied. It includes Direct Vocal-Motor Pathway (DMP) and Anterior Forebrain Pathway (AFP). DMP consists of four major song nuclei, *i.e.* HVC, RA, nXIIts, and LMC. AFP includes area X, DLM, and LMAN. These brain regions are responsible for different processes in vocal learning [12, 13, 16, 17] (Fig. S1).

Understanding the genetic basis of vocal learning can provide important insights into the evolution and development of human speech and language, and may contribute to a better understanding of speech and language disorders. Previous studies took a variety of approaches to explore the genetic basis of vocal learning. Mutations in genes and pathways have been linked to human speech disorders, *i.e.* 55 genes in MalaCards database [18], including ROBO1 [19, 20]. Many studies examined gene expression patterns in song and speech related brain regions

Received: April 30, 2024. **Revised:** October 10, 2024. **Accepted:** November 12, 2024

in vocal learners, and found that axon guidance genes SLIT and ROBO have convergent specialized expression [21–24]. Genome-wide gene expression of four zebra finch song nuclei during singing behavior was examined, and convergent transcriptional specialization in song-related brain areas of humans and vocal learning birds were profiled [25, 26]. Constitutive markers of each song nucleus for zebra finch were also identified [27].

Evolutionary signatures in vocal learning lineages have also been used to identify candidate genes related to vocal learning ability. FOXP2 coding region was found to have undergone positive selection in the human lineage, although this finding remains a subject of debates [28–30] and no positive selection signal was detected in FOXP2 coding region in vocal learning birds [31]. A couple of studies utilized the convergent evolution of vocal learning in multiple lineages. Coding sequences that evolved at accelerated rate or were under positive selection were identified in three mammalian vocal learners, but no shared gene was found [32]. This study also found 73 genes with amino acid convergence, including several axon guidance related genes such as ROBO1. Another study searched for positively selected regions across aligned whole genomes of vocal learning birds and identified enriched signals in noncoding regions [33]. The basis of this study is the recent advancements in whole genome alignment (WGA) methods that greatly increases the alignable region across genomes [34–37]. However, WGA could still be relatively time- and resource- consuming and only alignable regions across all taxa have been examined.

In this study, we developed a new pipeline to identify genome-wide molecular convergence in the form of loss of ancestral conserved fragments in both coding and noncoding regions. We have applied this method to the convergent evolution of vocal learning ability in avian lineages of hummingbirds, parrots and songbirds. Our specific objectives were to (i) identify genome-wide molecular convergence in both coding and noncoding regions in three avian vocal learning lineages; (ii) examine the patterns of molecular convergence, including relative contribution of coding and noncoding regions; (iii) explore the possible functional consequences of identified molecular convergence; and (iv) identify potential molecular convergence between human and avian vocal learners.

Materials and methods

Identification of ancestral conserved fragments and CLACs

We chose high-quality genomes from phase I and phase II of the Bird 10,000 Genomes Project [38, 39] to avoid false discovery of loss of ancestral conserved fragments due to low genome quality and coverage. For each non-vocal learning order, the species with the highest N50 was selected as representative species to ensure relatively unbiased selection at the order level. For vocal learning orders, all species with genome scaffold N50 > 2 Mb were used, and the final species set has 2–3 representative species per family in songbirds (Table S1). This results in 34 vocal learning species from three orders, *i.e.* Anna's Hummingbird in Caprimulgiformes, Budgerigar in Psittaciformes and 32 songbird species in oscine Passeriformes, as well as 24 species from 19 orders that are incapable of vocal learning, including six non-oscine Passeriformes and one non-hummingbird Caprimulgiformes. Chicken genome (Galgal6.0), as a basal non-vocal learner, was used as the reference genome.

We adopted *phyluce* pipeline with modification, which is for identifying ultra-conserved elements across species for phylogenomic studies [39], to identify ancestral conserved fragments (ACFs) across species (https://phyluce.readthedocs.io) (Fig. S2). For each genome other than the reference chicken, we first generated 100 bp short reads by tiling genome with 5 bp step size using a custom script. These reads cover the whole genome evenly, at 20× depth. These 100 bp reads were aligned to the chicken genome using *stampy* with parameters '*–maxbasequal 93 –substitutionrate 0.05*' [40]. Reads mapped to multiple regions were filtered. Coordinates of genomic regions covered by unique mapped reads were extracted, sorted and merged (within 20 bp) using *BEDTools*, and such regions were considered as conserved between reference genome and the corresponding species. ACFs were defined as conserved regions shared by a specific group of basal species, *i.e.* all non-vocal learning species in this study. Regions less than 10 bp or with >25% masked sites in chicken genome annotation were further filtered. ACFs across non-vocal learning species served as a proxy of functionally important fragments.

We then identify ACFs that are lost in vocal learners. If an ACF is lost independently in multiple vocal learning lineages, we consider it to be associated with vocal learning ability. Here we selected ACFs lost in more than two of the three vocal learning lineages, and for the songbird lineage, ACFs lost in more than 16 of the 32 songbirds were considered lost. We name such fragments as Convergently Lost Ancestral Conserved fragments (CLACs) (Fig. S2). Scripts are available at https://github.com/lixiaoyi-12/FCM.

Noncoding CLACs annotations

To explore the potential functions of noncoding CLACs, we examined if they contain H3K27ac, H3K4me1, or H3K27me3 ChIP-seq peaks identified from chicken in *Seki et al* [41], as well as peaks of ChIP-seq and ATAC-seq in *Sackton et al* [42]. Peaks of the same epigenetic marker from different developmental stages were merged, and noncoding CLACs that contain these ChIP-seq peaks were identified using *BEDTools intersect*. Genomic Association Tester was used to assess significance of the overlap between noncoding CLACs with ChIP-seq peaks [43], using chicken genome as background.

SEA suite was used to predict enriched known sequence motifs in noncoding CLACs [44], with E-value smaller than 0.01. Background nucleotide frequencies were calculated in the corresponding genomic regions of the Galgal6.0 genome via *get-markov* based on 0-order model. MEME suite was used to *de novo* predict motifs in noncoding CLACs with fixed length 8 bp using the same background frequencies [45]. TOMTOM was used to compare the predicted motifs with known motif databases with a threshold q-value <0.05 [45] (*e.g.* JASPAR 2020 [46]).

Results

We developed a new method to identify ACFs in both coding and noncoding regions across the genome that are convergently lost in multiple target lineages with derived convergent traits. Whole genome alignment requires extensive computational resources and has long processing time, and thus making it difficult to incorporate newly sequenced or updated genomes. Therefore, we first used a versatile WGA-free approach to identify ACFs across genomes of species with ancestral traits. Specifically, we tiled short reads from each target genome and mapped them back to a basal reference genome using a short-read aligner [39, 47]. Regions of the reference genome covered by uniquely mapped reads were considered to be conserved between the reference

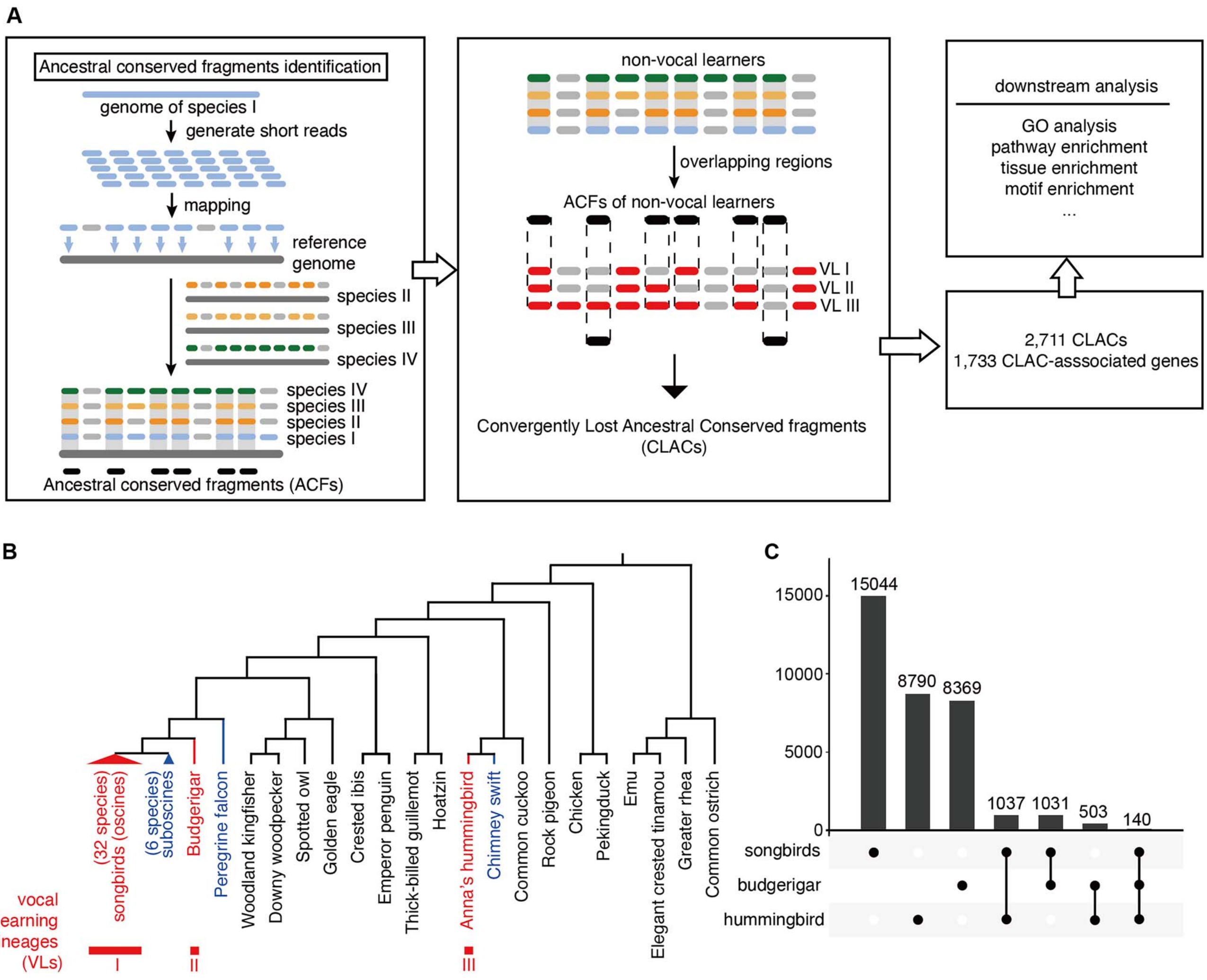

Figure 1. **Genomes and pipeline used in the paper.** (A) Summary of our pipeline to identify convergently lost ancestral conserved fragments (CLACs) underlying convergently evolved vocal learning ability. See Fig. S2 for a more detailed flowchart. (B) Phylogeny of avian species included in our analysis. The phylogenetic relationship is modified from *Feng et al.* 2020. Vocal learning lineages and sister lineages used as the control set are highlighted. (C) Number of ancestral conserved fragments lost in vocal learning lineages. The histogram shows corresponding number of ancestral conserved fragments lost in one, two and all vocal learning lineages. The black dots at the bottom indicate the fragments are lost in the lineage.

genome and the examined genome. Shared conserved regions across all species that have the ancestral trait, *i.e.* non-vocal learning, were identified as ACFs (Fig. 1A; Fig. 2S). We use this conservation to define important functional regions across genomes with ancestral trait in both coding and noncoding regions. Then we take advantage of the convergent evolution system to identify ACFs that have been lost independently in multiple vocal learning lineages, which may be associated with evolution of novel vocal learning ability rather than species-specific evolution. Our approach uses presence or absence of alignment as a binary proxy of levels of conservation. Deletions or lineage-specific non-alignable regions are included in the search for convergent signals, which are usually filtered out from classic WGA-based methods. In addition, compared with previous methods to identify molecular convergence of specific amino acids or changes of evolutionary rates in protein coding regions, our method could easily be applied to the whole genome including both coding and noncoding regions. WGA-free approach allows us to identify and define conserved fragments versatilely in customized group of genomes.

Identification of genome-wide ancestral conserved fragments in non-vocal learning birds

We first identified genome-wide ACFs across distantly related species of birds that do not have vocal learning ability. We selected 58 high-quality bird genomes, including 34 species with advanced vocal learning ability from three lineages, and 24 species incapable of vocal learning (Fig. 1B; Table S1). The genomes are selected to represent diverse major avian clades, and have relative high quality to minimize the false identification of lost conserved fragments due to low genome quality (see Methods). The 24 non-vocal learners cover most of avian lineages with available genomes, including the basal lineages. We obtained 433,014 genomic regions that are conserved across all 24 non-vocal learners, ~113.86 Mb (10.69%) of the chicken genome. Majority of the identified ACFs are less than 1 kb with a median length of 209 bp (Fig. 2A). The ACFs distribute unevenly across chromosomes, with a smaller proportion in small chromosomes such as chromosome 16, 25, 31, and 32 (Fig. 2B). One exception is sex chromosome Z, which has unproportionally less ACFs, mostly driven by a non-oscine passeriform, white-breasted antbird.

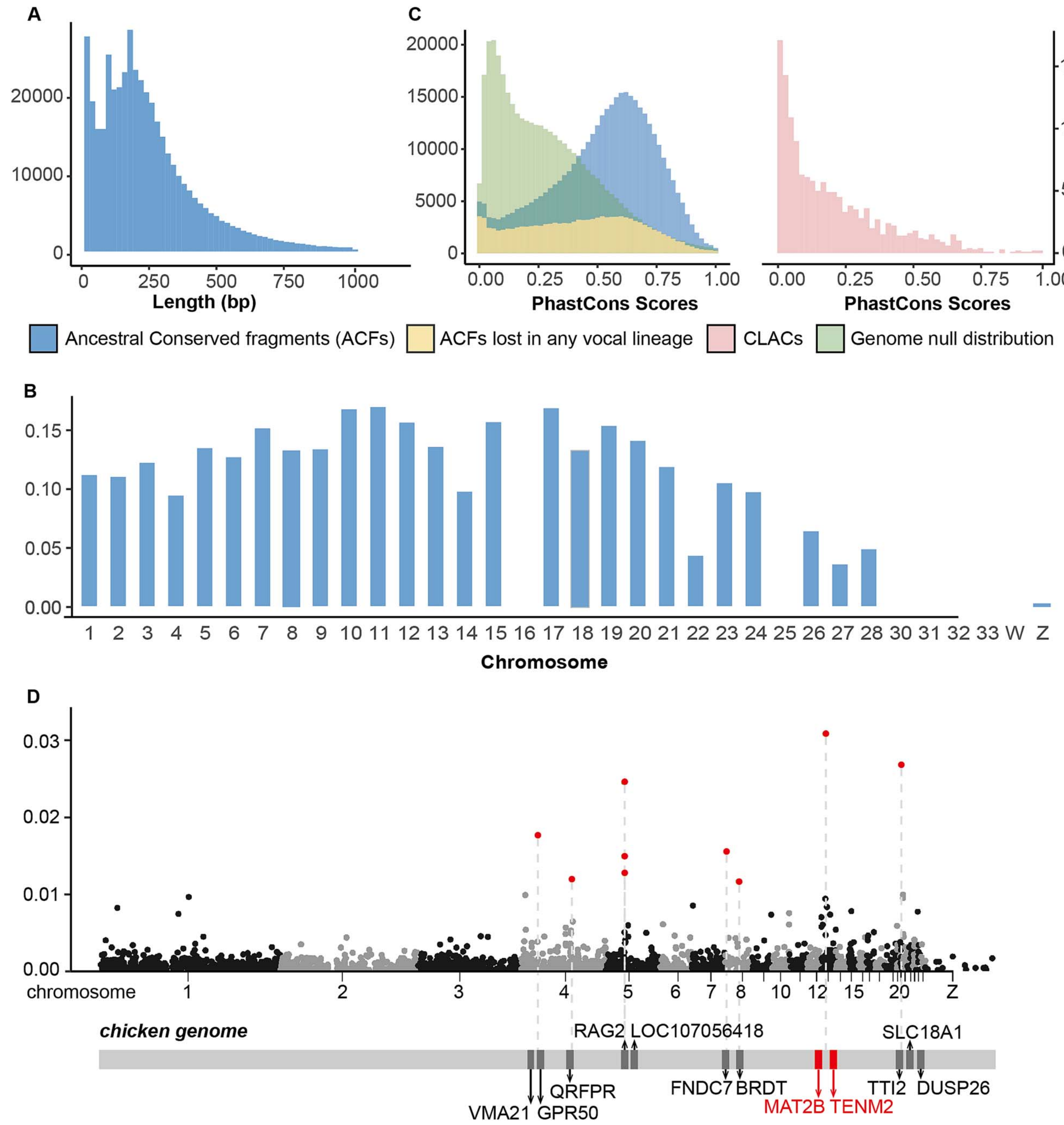

Figure 2. **Summary of identified ancestral conserved fragments.** (A) Length distribution of identified conserved fragments (<=1 kb). (B) Chromosome distribution of ancestral conserved fragments in chicken genome. The proportions were calculated by dividing the total length of ancestral conserved fragments on a specific chromosome by corresponding chromosome length. (C) Distribution of PhastCons scores of different groups of conserved fragments. (D) Length proportion of CLACs in genomic windows of 100 kb. Below shows the position of genes around the hotspot regions.

Ancestral conserved fragments lost in multiple vocal learning bird lineages

Highly conserved fragments across distantly related taxa suggest that they are under strong evolutionary constraint, indicative of functional importance. Fast evolution or loss of such region in specific lineage may associate with relaxation of purifying selection or positive selection and consequently changes of ancestral functionality. Among the three avian lineages that independently evolved vocal learning ability [7], we find 2711 ACFs that are shared across all non-vocal learners but lost in at least two vocal learning lineages, which we name CLACs (Fig. 1). As a control, we apply this pipeline to a set of sister taxa to the vocal learners, *i.e.* peregrine falcon, chimney swift, and non-oscines species in Passeriformes (Fig. 1B). We find that only 58 ACFs are lost in at least two of these sister groups, representing the number of CLACs expected by random chance given similar phylogenetic distances. We note this could be an underestimate due to unequal terminal branch lengths leading to vocal learning and control lineages. Nevertheless, none of CLACs in the control group overlaps with that in vocal learning lineages, and the number is much smaller than the number in vocal learners, suggesting convergent loss of ACFs in vocal learners exceed background noise and may be enriched with real biological signals related to the evolution of vocal learning ability.

We calculated the alignment-based conservation of each ACF using PhastCons score reported from 48 avian species [48, 49]. As expected, we find that CLACs are enriched in lower PhastCons scores compared to all ACFs (Fig. 2C). To determine whether there are hotspots of CLACs across the genome, we examined the spatial distribution of CLACs and ACFs in 100 kb genomic windows. We found that CLACs have several chromosomal hotspots, compared with relatively even spatial distribution of ACFs (Fig. 2D; Fig. S3). The genomic window that contain the largest length proportion of CLACs locates between TENM2 and MAT2B genes in chromosome 13. This region has been previously reported as hotspot region with accelerated evolution [33]. TENM2 is involved in neuron development, and has been identified as a candidate gene to modulate hippocampal structure and affect learning ability [50].

Enrichment of CLACs in noncoding regions, brain related pathways and tissues

We explore the potential functional consequences of CLACs based on chicken genome annotations. Approximately 25.0%, 39.5% and 31.7% of all ACFs are located in coding, intronic and intergenic regions, respectively (Fig. 3A; Table S2). Compared with chicken genome annotation that has ~2.8% coding regions, ACFs are highly enriched in coding regions, as expected. Interestingly, CLACs are highly deficit in coding regions (3.44%), compared with all ACFs (Fig. 3A), suggesting that noncoding regions including introns and intergenic regions may contribute more to the evolution of vocal learning ability.

We identified closest gene to each CLAC and found 1733 such CLAC-associated genes. Using genes associated with all ACFs as background in GO analysis, CLAC-associated genes are most significantly enriched in many neurological processes, such as glutamate receptor signaling pathway (enrichment fold = 4.23; FDR = 6.73×10^{-3}), axon guidance (enrichment fold = 2.05; FDR = 4.86×10^{-2}) and neuron projection guidance (enrichment fold = 2.05; FDR = 4.76×10^{-2}) (Fig. 3B; Table S3). Several biological processes related to cell–cell adhesion are significantly enriched, with the most significant term cell–cell adhesion mediated by cadherin (enrichment fold = 4.8; FDR = 1.33×10^{-3}), which is consistent with previous findings that cadherin play a crucial role in brain development by regulating cell–cell adhesion [51, 52] and dynamic expression of cadherins regulates vocal development in songbirds [53]. CLAC-associated genes are enriched in molecular functions including glutamate receptor activity (enrichment fold = 4.48; FDR = 3.05×10^{-2}), and many neuron-related cellular components (Fig. 3B; Table S3), including presynaptic membrane (enrichment fold = 3.06; FDR = 1.08×10^{-2}), postsynapse density (enrichment fold = 2.80; FDR = 4.29×10^{-5}), asymmetric synapse (enrichment fold = 2.77; FDR = 7.00×10^{-5}) and neuron to neuron synapse (enrichment fold = 2.62; FDR = 1.36×10^{-4}). These enrichment patterns are not found in the outgroup control, and are consistent with the known neurological basis of vocal learning ability [27, 54, 55].

Using IPA pathway enrichment analysis [56], the CLAC-associated genes are significantly enriched in several neurological pathways (Table S4), including synaptogenesis signaling pathway (p-value = 1.26×10^{-19}), glutamate receptor signaling pathway (Fig. 4A; p-value = 2.52×10^{-18}), neuropathic pain signaling in dorsal horn neurons (p-value = 5.00×10^{-14}), CREB signaling in neurons (p-value = 7.94×10^{-13}) and axon guidance signaling pathway (Fig. 4B; p-value = 3.16×10^{-11}). Glutamate receptor signaling pathway is implicated in the process of learning, memory and synaptic plasticity [57]. Several previous studies have highlighted axon guidance pathway playing an important role in human vocal learning [20, 58], and convergent expression patterns of this pathway have been reported in vocal learning birds and mammals [24].

Majority of the CLAC-associated genes overlap with previously reported vocal learning related genes using distinct strategies (Fig. 3C). 999 CLAC-associated genes were previously found to express in the four major forebrain regions of DMP in the song system of songbirds by *Whitney et al* [25]. There are 915 CLAC-associated genes found to convergently express in brains of vocal learning birds and humans by *Pfenning et al* [26]. 401 CLAC-associated genes have been reported as constitutive markers of song nuclei in contrast with adjacent brain regions by *Lovell et al* [27]. 258 CLAC-associated genes are detected by all three studies (Fig. 3C), including GRIN1, member of glutamate receptor signaling pathway, as well as PLXNA1, ROBO1, and SLIT2, members of axon guidance pathway. In addition to comparative transcriptome studies, using in situ hybridization data of 380 genes in zebra finch brain from www.zebrafinchatlas.org [27], 60 CLAC-associated genes could be confirmed to show significant differential expression in at least one song nucleus compared with their adjacent brain regions, including ROBO2 and GRIN2B (Fig. 3D).

We further examined the tissue enrichment pattern of these CLAC-associated genes [59], and found them significantly more expressed in several brain regions, using genes associated with all ACFs as background (Table S5). Majority of tissues with the most significant FDRs are brain regions, including superior frontal gyrus (FDR = 1.42×10^{-8}), corpus callosum (FDR = 8.57×10^{-8}) and prefrontal cortex (FDR = 4.30×10^{-7}). 1135 CLAC-associated genes are expressed in superior frontal gyrus. Superior frontal gyrus is the brain region responsible for the formation of working memory and cognitive functions [60], and it is interconnected with Broca's area, the speech production center located in inferior frontal gyrus through frontal aslant tract [61]. 1168 CLAC-associated genes have expression in prefrontal cortex, which is located in the front part of the frontal lobe and is associated with human speech production and language comprehension [62]. Corpus callosum connects the two brain hemispheres and has been reported to associate with brain lateralization of vocal learning [63]. Loss or injury of corpus callosum leads to impairments in verbal performance and dyslexia [64–67]. All results are relatively robust to aligner and mapping stringency (Supplementary text; Table S6; Fig. S4). The above functional enrichment patterns support the functional relevance of CLACs to vocal learning.

Noncoding CLACs enrich in motifs associated with neurogenesis

Majority of CLACs locate in noncoding regions in chicken genome (Fig. 3A), suggesting an important role of regulatory changes in the evolution of vocal learning ability. Due to limited knowledge on most noncoding regions, understanding the direct functional consequences of noncoding changes has been challenging. We first *de novo* predicted enriched DNA motifs in noncoding CLACs sequences with customized background nucleotide content of chicken with MEME suite, which find statistically over-represented motifs in target sequences [45] (Fig. S5). We identified a motif *RGCAGCTG* which assembles binding motifs of several known transcriptional factors, including MYOG (Q-value = 1.05×10^{-2}) related to nervous system development, and ASCL2 (Q-value = 1.69×10^{-3}) that plays a role in neuronal precursors fate determination and nervous system development [68] (Fig. 4D; Table S7). We also examined if there are any enrichment of known DNA motifs in noncoding CLACs using

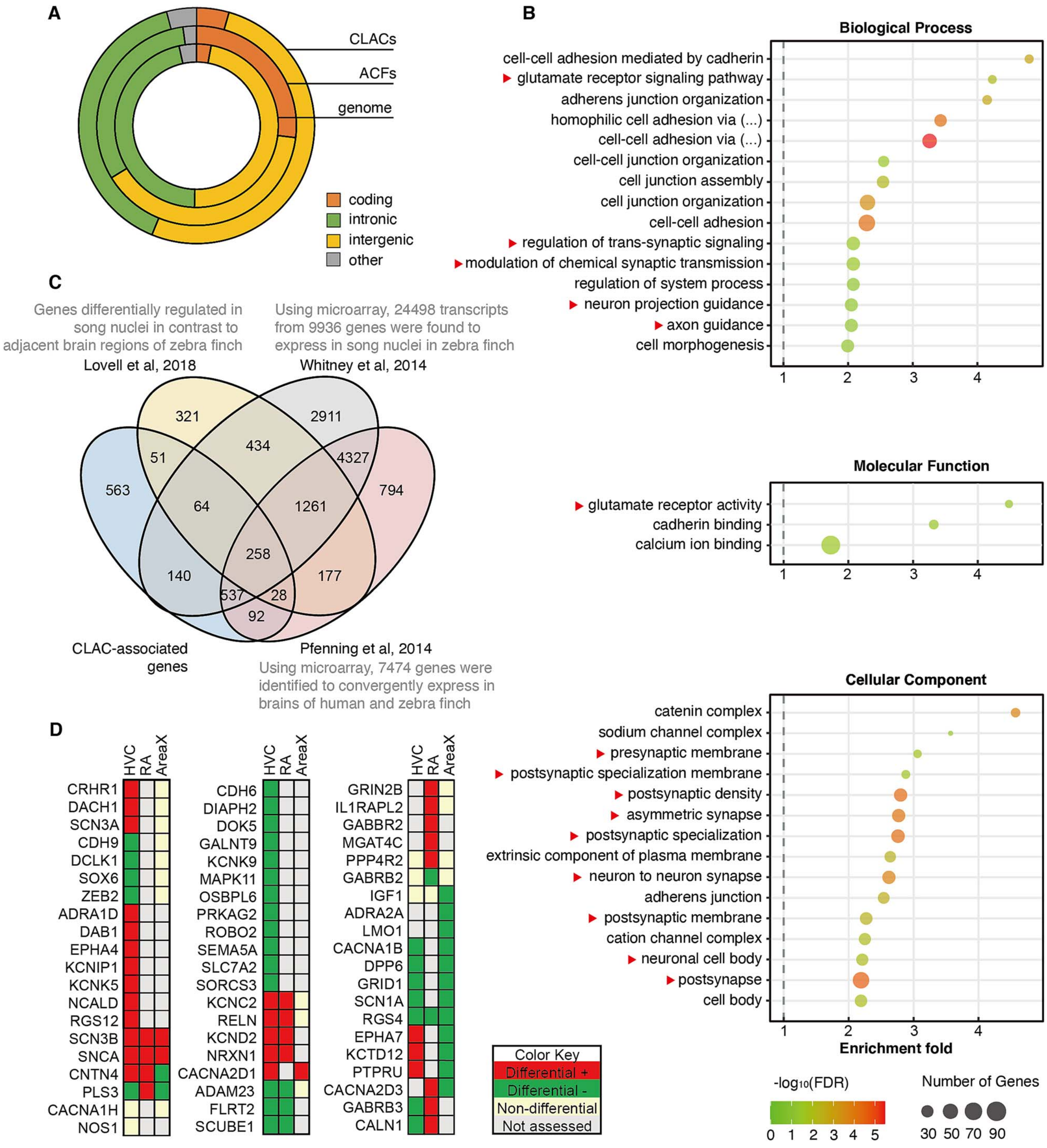

Figure 3. **Functional annotations and enrichment patterns of CLACs.** (A) Percentages of ancestral conserved fragments and CLACs in different functional categories. (B) Significantly enriched GO terms for CLAC-associated genes. Terms related to brain and neuron activities are highlighted with arrowheads. Top 15 pathways with highest fold enrichment and FDR < 0.05 were shown. (C) Comparison of CLAC-associated genes with previously reported genes related to avian vocal learning ability. (D) CLAC-associated genes with significant differential expression in at least one song nucleus compared with their adjacent brain regions using in situ hybridization data in zebra finch brain from www.zebrafinchatlas.org [27].

SEA suite, which identifies known motifs that are statistically more over-represented in target sequences compared with control sequences [44]. In addition to ASCL2 (E-value = 1.03×10^{-31}) and MYOG (E-value = 2.21×10^{-19}), binding motifs of several genes implicated in nervous system development were found, including MYF5 (E-value = 1.49×10^{-22}), PTF1A (E-value = 6.07×10^{-22}), NEUROD1 (E-value = 5.01×10^{-10}), zinc finger protein ZIC2 (E-value = 1.71×10^{-11}), and ZIC3 (E-value = 2.15×10^{-11}) (Table S8). NEUROD1 has been reported to involved in the development of nervous system and neurogenesis [69]. ZIC transcription factor family is known to play important roles in various neuronal developmental processes, such as neurogenesis, organogenesis of central nervous system, and the development of cerebellum [70–72].

Additionally, we compared CLACs with previously reported epigenetic markers in chicken, which are candidate regulatory

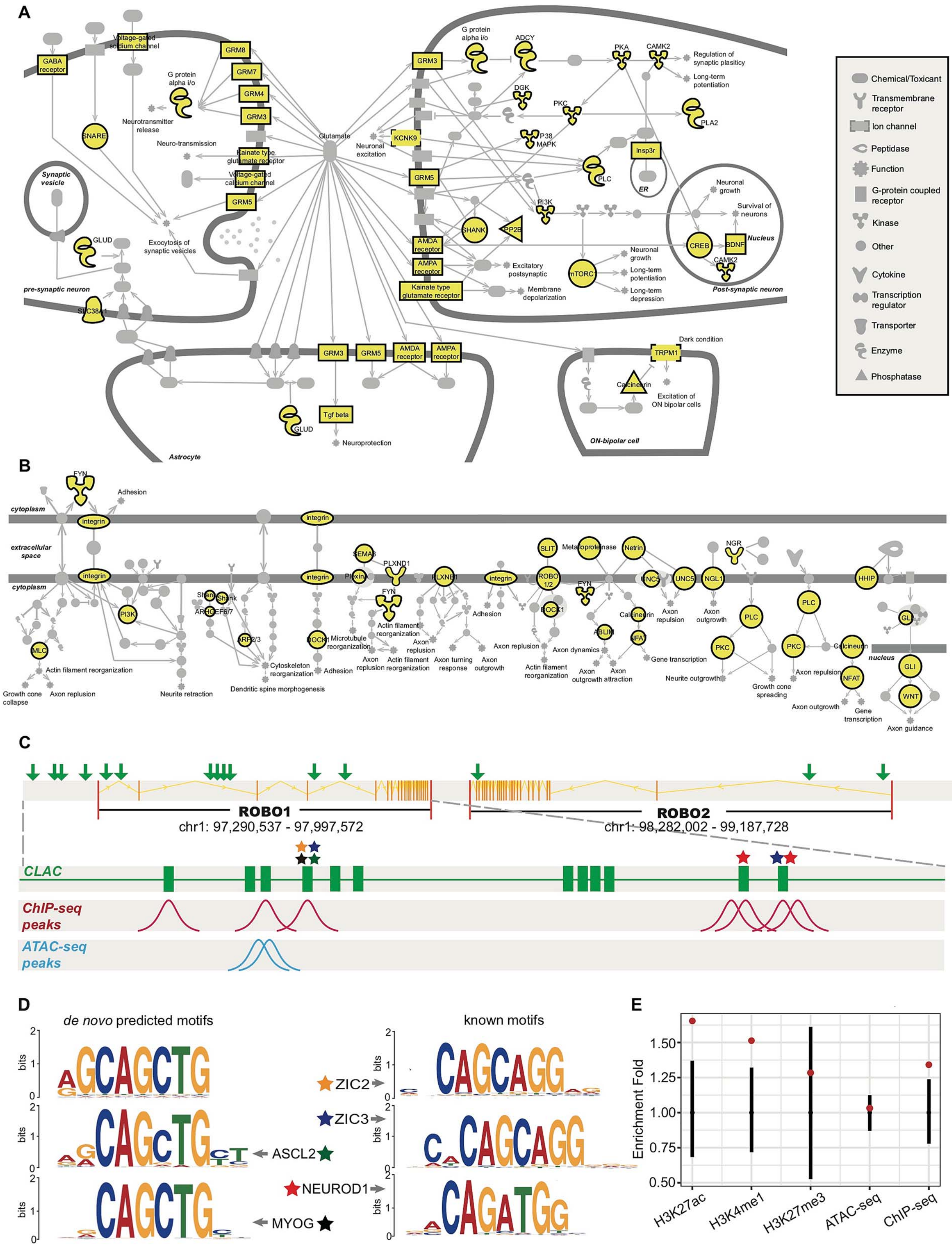

Figure 4. **CLACs are most significantly enriched in neurological pathways**, including (A) glutamate receptor signaling pathway and (B) axon guidance signaling pathway. Simplified to highlight CLAC-associated genes. (C) Genomic region containing ROBO1 and ROBO2. CLACs are highlighted with green arrows and bars. Peaks of ChIP-seq and ATAC-seq overlapping CLACs and transcription factor binding motifs (colored stars matching panel D) were displayed. (D) Enriched motifs in noncoding CLACs, which assemble binding motifs of transcription factors. (E) Enrichment of noncoding CLACs with reported epigenetic markers.

elements. Compared with histone markers in key embryonic developmental stages of chicken [41], there are 202 noncoding CLACs contain ChIP-seq peaks of at least one of H3K27ac, H3K4me1, or H3K27me3 epigenetic markers. Using all ACFs as background, noncoding CLACs are enriched in the ChIP-seq peaks of H3K27ac (Q-value = 3.30×10^{-3}), H3K4me1 (Q-value = 6.30×10^{-3}), and in H3K27me3 peaks, although not statistically significant (Fig. 4E). The intronic CLAC associated with GRIN2B overlaps with peaks of the histone marker H3K4me1. Four CLACs associated with ROBO1 also have multiple histone markers H3K4me1 and H3K27ac (Fig. 4C). Additionally, compared with ChIP-seq and ATAC-seq in eight tissues during the course of chicken development [42], we found 125 noncoding CLACs overlap with ATAC-peaks (Q-value = 1.02×10^{-2}), and 379 overlap with ChIP-peaks (Q-value = 3.97×10^{-2}) (Fig. S6). Two intergenic CLACs associated with ROBO1 have both ChIP-peaks and ATAC-peaks (Fig. 4C). Changes in these putative regulatory regions may be associated with the development of vocal learning ability through modification in the regulation of GRIN2B and ROBO1.

Human included as a vocal learner

Songbirds are often used as model organisms to explore the neurological and molecular basis of vocal learning behavior. Human is the most advanced mammalian vocal learner that independently evolved this ability. Previous studies suggest the neurological basis of human and zebra finch may share similarity, but certain parts of the related nucleus are not homologous [73]. To explore whether there is convergent genetic basis between human and vocal learning birds, we take one step further and include human in our search for convergent loss of ACFs associated with vocal learning ability, using chimpanzee as the non-vocal learning sister group. Specifically, we tiled genomes of human and chimpanzee, and mapped to chicken genome. We identified 111,488 ACFs shared by all avian non-vocal learners and chimpanzee. We found 19 of these ACFs were convergently lost in multiple avian vocal learners and human. One CLAC associates with NCAM2, which play a role in axonal projection and also regulates plasticity of synapses to influence learning process [74]. Another CLAC locates 30 kb upstream SLC6A1 gene, which encodes an important GABA reporter associated with neurodevelopmental disorders, and expresses in brain cortex and in the developing brain [75, 76]. One CLAC overlaps with ADCY8, which involved in synaptogenesis signaling pathway [77, 78]. Another convergently lost region is located in the intron of LRRC4C, a member of netrin G family of axon guidance molecules and is related to neurodevelopmental disorders [79]. Our results suggest these genes as candidates that might contribute to the evolution of vocal learning in both birds and human.

Discussion

We develop a versatile strategy to identify genome-wide coding and noncoding fragments conserved in avian non-vocal learners but are lost convergently in multiple derived vocal learning lineages. Loss of ancestral conserved fragments suggests change of ancestral functionality and multiple independent loss exclusively in vocal learning lineages suggest these may associate with the evolution of novel vocal learning ability. We find 2711 such CLACs, which are highly enriched in noncoding regions. Their associated genes are significantly enriched in neurological biological pathways including glutamate receptor signaling pathway and axon guidance signaling pathway, and the expression of these genes are enriched in brain tissues that have been previously implicated in vocal learning ability, supporting the functional relevance of identified CLACs to vocal learning. We further include human, an advanced vocal learner in our analysis, and find several candidate genes that may contribute to the evolution of vocal learning ability in both birds and human. Our method could be applied to other convergent evolution systems to identify both coding and noncoding molecular convergence underlying the evolution of focal traits.

Noncoding regulatory regions may contribute significantly to the evolution of vocal learning ability in birds

Previous methods for identification of molecular convergence mainly focused on protein coding regions [1–3]. Our method enables us to identify molecular convergence in both coding and noncoding regions. Notably, majority of CLACs identified are from noncoding regions of genome. Our results, consistent with another recent study that used WGA to detect accelerated genomic regions of vocal learning birds [33], highlight the potentially important contribution of noncoding regulatory changes to the evolution of the vocal learning ability.

The functional relevance of CLACs are supported by pathway and tissue enrichment patterns, as well as differential expression of CLAC-associated genes in song nuclei of vocal learning birds using previously published comparative transcriptomes and in situ hybridization data [26, 27] (Fig. 3D). Future direct functional validation could use reporter assays to compare the regulatory activity of orthologous sequences of candidate CLACs in neural cell cultures [78], or in transgenic animals where temporal and spatial regulatory activity of CLAC could be examined [80].

As comparisons to our results, we used existing methods on coding genes to detect convergent shifts in evolutionary rates, as well as gene family expansion in multiple vocal learning lineages. We found 224 genes and two gene families, respectively (Supplementary text). However, these genes do not have enrichment in any GO term or pathway, and do not overlap with CLAC-associated genes we identified, suggesting different genes have undergone convergent evolution at protein level or regulatory level.

Avian vocal learning may involve multiple neurological pathways

There have always been tremendous interests to understand the neurological and genetic basis of vocal learning ability, thus we have a wealth of information from previous studies using distinct data and approaches. The consistency with previous findings support the functional relevance of CLACs to vocal learning. Avian CLAC-associated genes are significantly enriched in several neurological biological processes and pathways. Of these, glutamate receptor signaling pathway has been implicated in Alzheimer's disease, schizophrenia and many other brain disorders [81]. Glutamate receptor gene families are important neurotransmitter receptors families, and play roles in synaptic plasticity and transmission. Previous study has examined expression of these glutamate receptors in the brain of zebra finch, and found that GRIN1 has significantly higher expression in Area X and GRIN2B shows lower expression in HVC, LMAN, and DLM, relative to the surrounding brain subdivisions [82] (Fig. S1).

CLAC-associated genes are significantly enriched in axon guidance pathway, including ROBO1, ROBO2, and SLIT2. Multiple CLACs locates in proximity of ROBO and SLIT genes (Fig. 4C). SLIT2 encodes a ligand of ROBO proteins. ROBO-SLIT genes have been reported to play important roles in the evolution of vocal learning ability [24, 58]. They were also reported to have convergent specialized expression in brains of vocal learners [83], convergent amino acid substitutions across three vocal learning mammals [32], and evolve at accelerated rates in vocal learning birds [33]. Mutations in ROBO1 have been linked to human speech disorders [20]. Recent study on genes related to language impairment and developmental dyslexia have found ROBO1 and ROBO2 were under positive selection in human lineage [84]. Consistent with these evidences from previous studies, we highlight regulatory changes in ROBO1, ROBO2, SLIT2, and 62 other genes in axon guidance pathway may play an important role in the evolution of vocal learning in birds and humans.

Additionally, synaptogenesis signaling pathway and CREB signaling in neurons also have been previously linked to vocal learning ability. Synaptogenesis delay has been linked to deficit of speech processing in human infants [85]. CREB plays a role in dopamine release in brain [86], which may function in vocal learning process for zebra finch pupil [87].

Candidate genes involved in avian vocal learning

Notably, 12 CLACs are associated with ROBO1, with four and eight CLACs locate upstream and in the intron regions, respectively (Fig. 4C). One CLAC locates only 2344 bp away from GRIN1,which plays a role in glutamate receptor signaling pathway, synaptogenesis signaling pathway and CREB signaling in neurons. Mutations in GRIN1 have been found to associate with neurodevelopmental disorders such as jumbled speech and schizophrenia [88, 89]. Another candidate gene GRIN2B has one CLAC in its intron. Polymorphism in GRIN2B is associated with thinking and speech disorders, such as verbal fluency and abstract thinking [90]. GRIN2B is also a molecular marker of dopaminergic and glutamatergic synaptic plasticity during singing process [91], which has high expression in song nuclei LMAN during vocal learning and plays a role in song plasticity [92]. ATP13A4 has been associated with childhood apraxia of speech and is highly expressed in brain region responsible for language [93, 94]. NRXN1 has been reported in many neurodevelopmental disorders including autism, schizophrenia, also associated with features of speech ability [95]. PDE7B, with one CLAC only 1378 bp upstream, has been identified as candidate dyslexia genes [96]. CXCL12 has been related to neuroinflammation and memory deficit meditation [97] and has a CLAC only 250 bp 5′ upstream.

Convergence between human and vocal learning birds

Humans and songbirds both have remarkable ability of advanced vocal learning. Prior research indicates that despite some similarity in the neurological underpinning of vocal learning between human and zebra finch, some related brain nuclei are analogous [98]. Whether there could be any molecular convergence between vocal learning of these lineages is not clear. Our study found that 19 of identified ACFs were convergently lost in multiple avian vocal learners and humans. Several are associated with genes with known neurological functions in learning process and brain development, including NCAM2 [74], SLC6A1 [75, 76], ADCY8 [77], and LRRC4C [79]. Additionally, five of CLAC-associated genes identified in avian vocal learners are among the 55 genes reported to relate to human speech disorder in the MalaCards database, *i.e.* ATP13A4, GRIN1, GRIN2B, NRXN1, and ROBO1. Furthermore, there are 915 CLAC-associated genes found to convergently express in brains of vocal learning birds and humans by Pfenning et al [26]. These results suggest that there may be molecular convergence to some extent at regulatory element level, gene level or pathway level that contributes to the evolution of novel vocal learning ability in both humans and birds.

Advantages and limitations of our method

Our WGA-free method have several advantages. First, this method does not rely on WGA that could be influenced by genome rearrangement (*i.e.* structure variants, inversion and transpositions) across distantly-related genomes. Second, this method is computationally efficient and could easily incorporate new or updated genomes without the need to re-process the other genomes. Third, deletions or lineage-specific non-alignable regions are naturally included in the search for convergent signals, which are usually filtered out in classic WGA-based methods. Lastly, this method should be easily applied to decipher molecular convergence of other convergently evolved traits.

There are also a few limitations. First, we need relatively high-quality genomes to avoid false discovery of CLACs due to low genome quality. Bird genome project strategically sequenced genomes with good representation across avian groups, however, the genome availability might introduce selection bias. With more high-quality genomes available at unprecedent speed, updated genomes and new genomes could be readily incorporated using our versatile pipeline to further narrow down candidate genomic regions. Second, mapping short reads tiled from a target species to the reference genome could have spurious signals, and should be taken with caution if used alone. However, the requirements that ACFs are shared by all non-vocal learning taxa and that CLACs are convergently lost in multiple vocal learning taxa should remove majority of these spurious signals. Third, we used presence or absence of alignment as a binary estimate of conservation, thus we could not estimate evolutionary rate parameters due to the lack of alignment, neither could distinguish positive selection and relaxation of purifying selection. Lastly, future functional validation is needed to examine how the change of identified candidate sequences in regulatory activity contribute to evolution and development of vocal learning.

Key Points

- We present a versatile pipeline for identifying genome-wide molecular convergence in the form of loss of ancestral conserved fragments in both coding and noncoding regions.
- Applying this pipeline to the convergent evolution of vocal learning ability in birds, we found such molecular convergence was significantly enriched in noncoding regions, near genes in glutamate receptor signaling pathway, axon guidance pathway and genes expressed in brain.
- We further found molecular convergence that may contribute to the evolution of vocal learning in both birds and humans.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Acknowledgements

We thank Jianzhi Zhang and Jian Yang for helpful comments on our manuscript. We thank Cai Li and Guojie Zhang for sharing conservation score data from the avian phylogenomic project. We thank Qi Zhou for sharing pipeline to process whole genome alignments. We thank Westlake Supercomputer Center for HPC and computation assistance.

Author contributions

Y.Z. conceived of and supervised the study. X.L. and K.Z. analyzed the data. X.L. and Y.Z. wrote the manuscript.

Conflict of interest

None declared.

Funding

This work was supported by National Natural Science Foundation of China grants [32370662 (YZ), 32100484 (KZ)]; Zhejiang Provincial Natural Science Foundation of China grants [LQ20C040001 (KZ)]; 'Pioneer' and 'Leading Goose' R&D Program of Zhejiang [2024SSYS0032]; and Research Center for Industries of the Future (RCIF) at Westlake University.

Data Accessibility Statement

This study did not generate new sequencing data.

Scripts, and coordinates of all ACFs and CLACs are available on GitHub (https://github.com/lixiaoyi-12/FCM).

Benefit-Sharing Statement

Benefits from this research accrue from the sharing of our data and results on public databases as described above.

Competing Interest Statement

We declare that none of the authors have competing interests.

References

1. Kowalczyk A, Meyer WK, Partha R. *et al.* RERconverge: an R package for associating evolutionary rates with convergent traits. *Appl Bioinformatics* 2019;**35**:4815–7. https://doi.org/10.1093/bioinformatics/btz468.
2. Zhang J, Kumar S. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol Biol Evol* 1997;**14**:527–36. https://doi.org/10.1093/oxfordjournals.molbev.a025789.
3. Rey C, Guéguen L, Sémon M. *et al.* Accurate detection of convergent amino-acid evolution with PCOC. *Mol Biol Evol* 2018;**35**:2296–306. https://doi.org/10.1093/molbev/msy114.
4. Janik VM, Slater PJB. Vocal learning in mammals, Advances in the Study of Behavior 1997;59–99. https://doi.org/10.1016/S0065-3454(08)60377-0.
5. Jarvis ED. Learned birdsong and the neurobiology of human language. *Ann N Y Acad Sci* 2004;**1016**:749–77. https://doi.org/10.1196/annals.1298.038.
6. Fitch WT, Huber L, Bugnyar T. Social cognition and the evolution of language: constructing cognitive phylogenies. *Neuron* 2010;**65**:795–814. https://doi.org/10.1016/j.neuron.2010.03.011.
7. Petkov CI, Jarvis ED. Birds, primates, and spoken language origins: Behavioral phenotypes and neurobiological substrates. *Front Evol Neurosci* 2012;**4**:12.
8. Martins PT, Boeckx C. Vocal learning: beyond the continuum. *PLoS Biol* 2020;**18**:e3000672. https://doi.org/10.1371/journal.pbio.3000672.
9. Jarvis ED, Güntürkün O, Bruce L. *et al.* Avian brains and a new understanding of vertebrate brain evolution. *Nat Rev Neurosci* 2005;**6**:151–9. https://doi.org/10.1038/nrn1606.
10. Brainard MS, Doupe AJ. Auditory feedback in learning and maintenance of vocal behaviour. *Nat Rev Neurosci* 2000;**1**:31–40. https://doi.org/10.1038/35036205.
11. Bolhuis JJ, Gahr M. Neural mechanisms of birdsong memory. *Nat Rev Neurosci* 2006;**7**:347–57. https://doi.org/10.1038/nrn1904.
12. Bolhuis JJ, Eda-Fujiwara H. Birdsong and the brain: the syntax of memory. *Neuroreport* 2010;**21**:395–8. https://doi.org/10.1097/WNR.0b013e3283385949.
13. Simonyan K, Horwitz B. Laryngeal motor cortex and control of speech in humans. *Neuroscientist* 2011;**17**:197–208. https://doi.org/10.1177/1073858410386727.
14. Flinker A, Korzeniewska A, Shestyuk AY. *et al.* Redefining the role of Broca's area in speech. *Proc Natl Acad Sci U S A* 2015;**112**:2871–5. https://doi.org/10.1073/pnas.1414491112.
15. Binder JR. Current controversies on Wernicke's area and its role in language. *Curr Neurol Neurosci Rep* 2017;**17**:58. https://doi.org/10.1007/s11910-017-0764-8.
16. Riters LV. The role of motivation and reward neural systems in vocal communication in songbirds. *Front Neuroendocrinol* 2012;**33**:194–209. https://doi.org/10.1016/j.yfrne.2012.04.002.
17. Saravanan V, Hoffmann LA, Jacob AL. *et al.* Dopamine depletion affects vocal acoustics and disrupts sensorimotor adaptation in songbirds. *eNeuro* 2019;**6**:ENEURO.0190–19.2019. https://doi.org/10.1523/ENEURO.0190-19.2019.
18. Rappaport N, Twik M, Plaschkes I. *et al.* MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res* 2017;**45**:D877–87. https://doi.org/10.1093/nar/gkw1012.
19. Bates TC, Luciano M, Medland SE. *et al.* Genetic variance in a component of the language acquisition device: ROBO1 polymorphisms associated with phonological buffer deficits. *Behav Genet* 2011;**41**:50–7. https://doi.org/10.1007/s10519-010-9402-9.
20. Hannula-Jouppi K, Kaminen-Ahola N, Taipale M. *et al.* The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia. *PLoS Genet* 2005;**1**:e50. https://doi.org/10.1371/journal.pgen.0010050.
21. Haesler S, Wada K, Nshdejan A. *et al.* FoxP2 expression in avian vocal learners and non-learners. *J Neurosci* 2004;**24**:3164–75. https://doi.org/10.1523/JNEUROSCI.4369-03.2004.
22. Hara E, Rivas MV, Ward JM. *et al.* Convergent differential regulation of parvalbumin in the brains of vocal learners. *PloS One* 2012;**7**:e29457. https://doi.org/10.1371/journal.pone.0029457.
23. Horita H, Kobayashi M, Liu WC. *et al.* Specialized motor-driven dusp1 expression in the song systems of multiple lineages of vocal learning birds. *PloS One* 2012;**7**:e42173. https://doi.org/10.1371/journal.pone.0042173.
24. Wang R, Chen CC, Hara E. *et al.* Convergent differential regulation of SLIT-ROBO axon guidance genes in the brains of vocal learners. *J Comp Neurol* 2015;**523**:892–906. https://doi.org/10.1002/cne.23719.
25. Whitney O, Pfenning AR, Howard JT. *et al.* Core and region-enriched networks of behaviorally regulated genes and the singing genome. *Science* 2014;**346**:1256780. https://doi.org/10.1126/science.1256780.

26. Pfenning AR, Hara E, Whitney O. *et al.* Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* 2014;**346**:1256846. https://doi.org/10.1126/science.1256846.
27. Lovell PV, Huizinga NA, Friedrich SR. *et al.* The constitutive differential transcriptome of a brain circuit for vocal learning. *BMC Genomics* 2018;**19**:231. https://doi.org/10.1186/s12864-018-4578-0.
28. Enard W, Przeworski M, Fisher SE. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 2002;**418**:869–72. https://doi.org/10.1038/nature01025.
29. Atkinson EG, Audesse AJ, Palacios JA. *et al.* No evidence for recent selection at FOXP2 among diverse human populations. *Cell* 2018;**174**:1424–1435.e15. https://doi.org/10.1016/j.cell.2018.06.048.
30. Zhang J, Webb DM, Podlaha O. Accelerated protein evolution and origins of human-specific features: FOXP2 as an example. *Genetics* 2002;**162**:1825–35. https://doi.org/10.1093/genetics/162.4.1825.
31. Webb DM, Zhang J. FoxP2 in song-learning birds and vocal-learning mammals. *J Hered* 2005;**96**:212–6. https://doi.org/10.1093/jhered/esi025.
32. Wang R. *Dissecting the Genetic Basis of Convergent Complex Traits Based on Molecular Homoplasy*. Durham, NC: Duke University, 2011.
33. Cahill AJ, Armstrong J, Deran A. *et al.* Positive selection in non-coding genomic regions of vocal learning birds is associated with genes implicated in vocal learning and speech functions in humans. *Genome Res* 2021;**31**:2035–49. https://doi.org/10.1101/gr.275989.121.
34. Armstrong J, Hickey G, Diekhans M. *et al.* Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature* 2020;**587**:246–51. https://doi.org/10.1038/s41586-020-2871-y.
35. Leonard AS, Crysnanto D, Mapel XM. *et al.* Graph construction method impacts variation representation and analyses in a bovine super-pangenome. *Genome Biol* 2023;**24**:124. https://doi.org/10.1186/s13059-023-02969-y.
36. Kille B, Balaji A, Sedlazeck FJ. *et al.* Multiple genome alignment in the telomere-to-telomere assembly era. *Genome Biol* 2022;**23**:182. https://doi.org/10.1186/s13059-022-02735-6.
37. *Sequence Alignment (I)*, in *Algorithms in Bioinformatics*. 2021. p. 51–97.
38. O'Leary NA, Wright MW, Brister JR. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**:D733–45. https://doi.org/10.1093/nar/gkv1189.
39. Faircloth BC. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 2016;**32**:786–8. https://doi.org/10.1093/bioinformatics/btv646.
40. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 2011;**21**:936–9. https://doi.org/10.1101/gr.111120.110.
41. Seki R, Li C, Fang Q. *et al.* Functional roles of Aves class-specific cis-regulatory elements on macroevolution of bird-specific features. *Nat Commun* 2017;**8**:14229. https://doi.org/10.1038/ncomms14229.
42. Sackton TB, Grayson P, Cloutier A. *et al.* Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* 2019;**364**:74–8. https://doi.org/10.1126/science.aat7244.
43. Heger A, Webber C, Goodson M. *et al.* GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* 2013;**29**:2046–8. https://doi.org/10.1093/bioinformatics/btt343.
44. Bailey TL, Grant CE. SEA: simple enrichment analysis of motifs. *BioRxiv* 2021. August 24.
45. Bailey TL, Boden M, Buske FA. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;**37**:W202–8. https://doi.org/10.1093/nar/gkp335.
46. Fornes O, Castro-Mondragon JA, Khan A. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2020;**48**:D87–92. https://doi.org/10.1093/nar/gkz1001.
47. Faircloth BC. Identifying conserved genomic elements and designing universal bait sets to enrich them. *Methods in Ecology and Evolution* 2017;**8**:1103–12. https://doi.org/10.1111/2041-210X.12754.
48. Siepel A, Bejerano G, Pedersen JS. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;**15**:1034–50. https://doi.org/10.1101/gr.3715005.
49. Zhang G, Li B, Li C. *et al.* Comparative genomic data of the avian Phylogenomics project. *GigaScience* 2014;**3**:26. https://doi.org/10.1186/2047-217X-3-26.
50. Delprato A, Bonheur B, Algéo MP. *et al.* Systems genetic analysis of hippocampal neuroanatomy and spatial learning in mice. *Genes Brain Behav* 2015;**14**:591–606. https://doi.org/10.1111/gbb.12259.
51. Hirano S, Takeichi M. Cadherins in brain morphogenesis and wiring. *Physiol Rev* 2012;**92**:597–634. https://doi.org/10.1152/physrev.00014.2011.
52. Punovuori K, Malaguti M, Lowell S. Cadherins in early neural development. *Cell Mol Life Sci* 2021;**78**:4435–50. https://doi.org/10.1007/s00018-021-03815-9.
53. Eiji Matsunaga KO. Expression analysis of cadherins in the songbird brain: relationship to vocal system development. *J Comp Neurol* 2008;**508**:329–42. https://doi.org/10.1002/cne.21676.
54. Lovell PV, Clayton DF, Replogle KL. *et al.* Birdsong "transcriptomics": neurochemical specializations of the oscine song system. *PloS One* 2008;**3**:e3440. https://doi.org/10.1371/journal.pone.0003440.
55. Wada K, Howard JT, McConnell P. *et al.* A molecular neuroethological approach for identifying and characterizing a cascade of behaviorally regulated genes. *Proc Natl Acad Sci U S A* 2006;**103**:15212–7. https://doi.org/10.1073/pnas.0607098103.
56. Kramer A, Green J, Pollard J Jr. *et al.* Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 2014;**30**:523–30. https://doi.org/10.1093/bioinformatics/btt703.
57. Fairless R, Bading H, Diem R. Pathophysiological ionotropic glutamate signalling in Neuroinflammatory disease as a therapeutic target. *Front Neurosci* 2021;**15**:741280. https://doi.org/10.3389/fnins.2021.741280.
58. Lei H, Yan Z, Sun X. *et al.* Axon guidance pathways served as common targets for human speech/language evolution and related disorders. *Brain Lang* 2017;**174**:1–8. https://doi.org/10.1016/j.bandl.2017.06.007.
59. Komljenovic A, Roux J, Robinson-Rechavi M. *et al.* BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. *F1000Res* 2016;**5**:2748. https://doi.org/10.12688/f1000research.9973.1.
60. Boisgueheneuc F, Levy R, Volle E. *et al.* Functions of the left superior frontal gyrus in humans: a lesion study. *Brain* 2006;**129**:3315–28. https://doi.org/10.1093/brain/awl244.

61. Dick AS, Garic D, Graziano P. *et al.* The frontal aslant tract (FAT) and its role in speech, language and executive function. *Cortex* 2019;**111**:148–63. https://doi.org/10.1016/j.cortex.2018.10.015.
62. Gabrieli JD, Poldrack RA, Desmond JE. The role of left prefrontal cortex in language and memory. *Proc Natl Acad Sci U S A* 1998;**95**: 906–13. https://doi.org/10.1073/pnas.95.3.906.
63. Hinkley LB, Marco EJ, Brown EG. *et al.* The contribution of the corpus callosum to language lateralization. *J Neurosci* 2016;**36**: 4522–33. https://doi.org/10.1523/JNEUROSCI.3850-14.2016.
64. Nakamura H, Watanabe Y. Isthmus organizer and regionalization of the mesencephalon and metencephalon. *Int J Dev Biol* 2005;**49**:231–5. https://doi.org/10.1387/ijdb.041964hn.
65. Unterrainer MJ, Meier AM, Wranek U. *et al.* Verbal performances and their relation to the corpus callosum depending on gender. *Klin Neuroradiol* 1998;**8**:22–9. https://doi.org/10.1007/BF03044065.
66. Erickson RL, Paul LK, Brown WS. Verbal learning and memory in agenesis of the corpus callosum. *Neuropsychologia* 2014;**60**: 121–30. https://doi.org/10.1016/j.neuropsychologia.2014.06.003.
67. Kershner JR. Neurobiological systems in dyslexia. *Trends Neurosci Educ* 2019;**14**:11–24. https://doi.org/10.1016/j.tine.2018.12.001.
68. Liu MH, Cui YH, Guo QN. *et al.* Elevated ASCL2 expression is associated with metastasis of osteosarcoma and predicts poor prognosis of the patients. *Am J Cancer Res* 2016;**6**:1431–40.
69. Cho JH, Tsai MJ. The role of BETA2/NeuroD1 in the development of the nervous system. *Mol Neurobiol* 2004;**30**:035–48. https://doi.org/10.1385/MN:30:1:035.
70. Aruga J. The role of Zic genes in neural development. *Mol Cell Neurosci* 2004;**26**:205–21. https://doi.org/10.1016/j.mcn.2004.01.004.
71. Ali RG, Bellchambers HM, Arkell RM. Zinc fingers of the cerebellum (Zic): transcription factors and co-factors. *Int J Biochem Cell Biol* 2012;**44**:2065–8. https://doi.org/10.1016/j.biocel.2012.08.012.
72. McMahon AR, Merzdorf CS. Expression of the zic1, zic2, zic3, and zic4 genes in early chick embryos. *BMC Res Notes* 2010;**3**:167. https://doi.org/10.1186/1756-0500-3-167.
73. Colquitt BM, Merullo DP, Konopka G. *et al.* Cellular transcriptomics reveals evolutionary identities of songbird vocal circuits. *Science* 2021;**371**:eabd9704. https://doi.org/10.1126/science.abd9704.
74. Parcerisas A, Ortega-Gascó A, Pujadas L. *et al.* The hidden side of NCAM family: NCAM2, a key cytoskeleton organization molecule regulating multiple neural functions. *Int J Mol Sci* 2021;**22**:10021. https://doi.org/10.3390/ijms221810021.
75. Minelli A, Brecha NC, Karschin C. *et al.* GAT-1, a high-affinity GABA plasma membrane transporter, is localized to neurons and astroglia in the cerebral cortex. *J Neurosci* 1995;**15**:7734–46. https://doi.org/10.1523/JNEUROSCI.15-11-07734.1995.
76. Mermer F, Poliquin S, Rigsby K. *et al.* Common molecular mechanisms of SLC6A1 variant-mediated neurodevelopmental disorders in astrocytes and neurons. *Brain* 2021;**144**:2499–512. https://doi.org/10.1093/brain/awab207.
77. Sanabra C, Mengod G. Neuroanatomical distribution and neurochemical characterization of cells expressing adenylyl cyclase isoforms in mouse and rat brain. *J Chem Neuroanat* 2011;**41**:43–54. https://doi.org/10.1016/j.jchemneu.2010.11.001.
78. Gu Z, Pan S, Lin Z. *et al.* Climate-driven flyway changes and memory-based long-distance migration. *Nature* 2021;**591**: 259–64. https://doi.org/10.1038/s41586-021-03265-0.
79. Maussion G, Cruceanu C, Rosenfeld JA. *et al.* Implication of LRRC4C and DPP6 in neurodevelopmental disorders. *Am J Med Genet A* 2017;**173**:395–406. https://doi.org/10.1002/ajmg.a.38021.
80. Dutrow EV, Emera D, Yim K. *et al.* Modeling uniquely human gene regulatory function via targeted humanization of the mouse genome. *Nat Commun* 2022;**13**:304. https://doi.org/10.1038/s41467-021-27899-w.
81. Willard SS, Koochekpour S. Glutamate, glutamate receptors, and downstream signaling pathways. *Int J Biol Sci* 2013;**9**:948–59. https://doi.org/10.7150/ijbs.6426.
82. Wada K, Sakaguchi H, Jarvis ED. *et al.* Differential expression of glutamate receptors in avian neural pathways for learned vocalization. *J Comp Neurol* 2004;**476**:44–64. https://doi.org/10.1002/cne.20201.
83. Lovell PV, Wirthlin M, Kaser T. *et al.* ZEBrA: Zebra finch expression brain atlas—a resource for comparative molecular neuroanatomy and brain evolution studies. *J Comp Neurol* 2020;**528**: 2099–131. https://doi.org/10.1002/cne.24879.
84. Mozzi A, Forni D, Clerici M. *et al.* The evolutionary history of genes involved in spoken and written language: beyond FOXP2. *Sci Rep* 2016;**6**:22157. https://doi.org/10.1038/srep22157.
85. Novitskiy N, Chan PHY, Chan M. *et al.* Deficits in neural encoding of speech in preterm infants. *Dev Cogn Neurosci* 2023;**61**:101259. https://doi.org/10.1016/j.dcn.2023.101259.
86. Wang H, Xu J, Lazarovici P. *et al.* cAMP response element-binding protein (CREB): a possible Signaling molecule link in the pathophysiology of schizophrenia. *Front Mol Neurosci* 2018;**11**:255. https://doi.org/10.3389/fnmol.2018.00255.
87. Tanaka M, Sun F, Li Y. *et al.* A mesocortical dopamine circuit enables the cultural transmission of vocal behaviour. *Nature* 2018;**563**:117–20. https://doi.org/10.1038/s41586-018-0636-7.
88. Lemke JR, Geider K, Helbig KL. *et al.* Delineating the GRIN1 phenotypic spectrum: a distinct genetic NMDA receptor encephalopathy. *Neurology* 2016;**86**:2171–8. https://doi.org/10.1212/WNL.0000000000002740.
89. Bhardwaj T, Ahmad I, Somvanshi P. Systematic analysis to identify novel disease indications and plausible potential chemical leads of glutamate ionotropic receptor NMDA type subunit 1, GRIN1. *J Mol Recognit* 2023;**36**:e2997. https://doi.org/10.1002/jmr.2997.
90. Mangano GD, Riva A, Fontana A. *et al.* De novo GRIN2A variants associated with epilepsy and autism and literature review. *Epilepsy Behav* 2022;**129**:108604. https://doi.org/10.1016/j.yebeh.2022.108604.
91. So LY, Munger SJ, Miller JE. Social context-dependent singing alters molecular markers of dopaminergic and glutamatergic signaling in finch basal ganglia area X. *Behav Brain Res* 2019;**360**: 103–12. https://doi.org/10.1016/j.bbr.2018.12.004.
92. Chakraborty M, Chen LF, Fridel EE. *et al.* Overexpression of human NR2B receptor subunit in LMAN causes stuttering and song sequence changes in adult zebra finches. *Sci Rep* 2017;**7**:942. https://doi.org/10.1038/s41598-017-00519-8.
93. Kwasnicka-Crawford DA, Carson AR, Roberts W. *et al.* Characterization of a novel cation transporter ATPase gene (ATP13A4) interrupted by 3q25-q29 inversion in an individual with language delay. *Genomics* 2005;**86**:182–94. https://doi.org/10.1016/j.ygeno.2005.04.002.
94. Worthey EA, Raca G, Laffin JJ. *et al.* Whole-exome sequencing supports genetic heterogeneity in childhood apraxia of speech. *J Neurodev Disord* 2013;**5**:29. https://doi.org/10.1186/1866-1955-5-29.
95. Brignell A, St John M, Boys A. *et al.* Characterization of speech and language phenotype in children with NRXN1 deletions. *Am J Med Genet B Neuropsychiatr Genet* 2018;**177**:700–8. https://doi.org/10.1002/ajmg.b.32664.

96. Buonincontri R, Bache I, Silahtaroglu A. *et al.* A cohort of balanced reciprocal translocations associated with dyslexia: identification of two putative candidate genes at DYX1. *Behav Genet* 2011;**41**:125–33. https://doi.org/10.1007/s10519-010-9389-2.

97. Mai CL, Tan Z, Xu YN. *et al.* CXCL12-mediated monocyte transmigration into brain perivascular space leads to neuroinflammation and memory deficit in neuropathic pain. *Theranostics* 2021;**11**:1059–78. https://doi.org/10.7150/thno.44364.

98. Zhang Y, Zhou L, Zuo J. *et al.* Analogies of human speech and bird song: from vocal learning behavior to its neural basis. *Front Psychol* 2023;**14**:1100969. https://doi.org/10.3389/fpsyg.2023.1100969.

Briefings in Bioinformatics, 2025, **26(1)**, bbae614
https://doi.org/10.1093/bib/bbae614
Case Study