이 학 박 사 학 위 논 문

Bioinformatic approaches to understand macroevolution among different vertebrate lineages

척추동물아문 내 다른 계통 간 대진화를
이해하기 위한 생물정보학적 접근

2022년 8월

서울대학교 대학원
생물정보협동과정 생물정보학전공
이 철

이 학 박 사 학 위 논 문

Bioinformatic approaches to understand macroevolution among different vertebrate lineages

척추동물아문 내 다른 계통 간 대진화를
이해하기 위한 생물정보학적 접근

2022년 8월

서울대학교 대학원
생물정보협동과정 생물정보학전공
이 철

Bioinformatic approaches to understand macroevolution among different vertebrate lineages

By

Chul Lee

Supervisor: Professor Heebal Kim

August, 2022

Interdisciplinary Program in Bioinformatics

Seoul National University

척추동물아문 내 다른 계통 간
대진화를 이해하기 위한 생물정보학적 접근

지도교수 김 희 발

이 논문을 이학박사 학위논문으로 제출함
2022년 8월

서울대학교 대학원
생물정보협동과정 생물정보학전공
이 철

이철의 이학박사 학위논문을 인준함

2022년 8월

위 원 장	김 선	(Seal)
부위원장	김 희 발	(Seal)
위 원	Erich D. Jarvis	(Seal)
위 원	윤 철 희	(Seal)
위 원	조 서 애	(Seal)

Abstract

Bioinformatic approaches to understand macroevolution among different vertebrate lineages

Chul Lee

Interdisciplinary Program in Bioinformatics

The Graduate School

Seoul National University

Bioinformatics aims to improve the quality of life of mankind by decoding molecular mechanisms of biological phenomena based on digitalized sequence information of various species. It generally begins with a construction of reference genomes representing each species and moves on downstream analyses for microevolution within species and macroevolutions between species. Although short-read sequencing technologies initiated genomics era, the short read assemblies had critical problems for lower continuity and erroneous gene annotations causing mis-interpretations. Long read sequencing technologies improved assembly continuities fundamental to chromosome-level scaffolds and corrected false annotations. Following up the paradigm shift from short-reads to long-reads, here, I performed a series of bioinformatic analyses to understand macroevolutions of various vertebrate species from reference genome construction to comparative genome approaches.

Chapter 1 summarized the general background of this dissertation. First, it described the paradigm shift of the reference genome constructions achieving chromosome-scale scaffolds. Next, comparative genomic approaches for specific traits were summarized.

Chapter 2, as a case of constructing a reference genome, illuminated a chromosome-level reference genome of giant-fin mudskipper, an endemic species in republic of Korea. Based on the four latest genome sequencing technologies (Pacbio CLR, 10X Genomics linked reads, Bionano optical mapping, and Arima Genomics Hi-C) in the international cooperation with the Vertebrate genomes project, it improved the 100-fold longer continuity (Scaffold N50) with a total of 25 chromosomal-level scaffolds compared to that of the previous genome. In addition, a total of 24,744 genes were annotated with Pacbio Isoseq transcriptome data.

In Chapter 3, as a case of combining the reference genome quality evaluation method and comparative genomic analyses, a method was developed to explore the chromosomal evolution between vertebrate species in distant lineages focusing on the BUSCO genes. In addition, it suggested methods for detecting false loss and duplication errors that cause problems in downstream analyses in reference genomes of various vertebrate lineages, such as, mammals, birds, and fishes, and revealed how those kinds of errors occurred.

In Chapter 4, as a case using the existing comparative genomic approaches, the molecular mechanisms of terrestrial adaptation and limb emergence were identified by applying the series of analyses for apormorphic evolution of the monophyletic lineage of lobed-fin fishes including coelacanths and human.

In Chapter 5, as a case developing a new comparative genomic approach, the rule of amino acid convergence was proposed and candidate genes related to vocal learning were discovered through the multi-omic analyses for convergent evolution between polyphyletic lineages of vocal learning bird and control groups.

Among the major findings of this study based on the bioinformatics approaches from the reference genome construction to comparative genomic researches, telomere sequence distributions on chromosomes and the principles of amino acid convergence would be a standard for comparisons in various lineages. In

addition, the systemized comparative genomic approaches that identified candidate genes involved in limb development and vocal learning may be utilized to discover new candidate genes associated with various useful traits of living things in the world.

Keywords: Reference genome assembly, Vertebrate Genomes Project, False gene losses, False gene gains, Apomorphic evolution, Convergent evolution

Student Number: 2015-30991

Contents

List of Tables

List of Figures

Chapter 1. Literature Review

1.1. Paradigm shift in reference genome constructions

Reference genome sequences are fundamental to bioinformatic applications in various scholar and industry fields, such as, biology, heath medicine, agriculture, and ecology. The first generation of reference genome assemblies of human [1] and representative model species in various lineages such as *Caenorhabditis elegans* [2], *Arabidopsis thaliana* [3], and *Mus musculus* [4], were initiated with Sanger sequencing technologies (read length 700-1000 base pair, bp) and their chromosome genetic maps. Although Sanger-based whole genome shotgun sequencing needed huge cost estimated as 1$/base in decade-long projects [5], the pioneering reference genomes opened genomics era with genome projects to understand micro-evolution within each species, such as, the human 1000 genome project [6].

The next generation sequencing (NGS) technologies, such as, Illumina platform (100-150 bp), rapidly decreased sequencing cost to 1$/read based on sequencing by synthesis. The high-throughput technologies explosively launched international genome consortiums and genome projects to construct reference genomes for various species or clades, such as, pig genome project [7] and Bird 10K genome project [8]. The accumulations of reference genomes provided unprecedent opportunities to understand macroevolution across species or clades by providing control data sets, but these shorter read assemblies without supports of chromosome genetic maps caused lower-quality problems including erroneous fragmentations of most chromosomes into thousands of pieces. Moreover, many genes in the short read-based reference genomes various species which are involved in traits were missing or duplicated totally or partially resulting in misinterpretation indistinct for real biological variations or errors in assemblies [9].

As a game changer, long-read sequencing technologies for the contig assembly process, such as, Pacbio continuous long reads (CLR, 1000-60000bp), hugely improved the continuity of genome assemblies with similar cost [9]. Additionally, there were new technologies for scaffolding processes, such as, 10X Genomics linked reads for phasing, Bionano optical mapping recognizing genomic landscapes of specific sequences as a probe in long molecules from 150000 bp to

multi-megabase pairs, and Arima Genomics Hi-C replaceable for chromosome genetic maps by reflecting 3D structures of chromosomes. By combining the above multiple technologies including long read sequencing, Vertebrate Genomes Project (VGP) suggested standard assembly pipelines as promising solutions towards complete and error-free genome assemblies and successfully constructed high quality reference genome assemblies for 16 vertebrate species achieving chromosome-level scaffolds [10]. These chromosome-scale reference genomes provided unprecedent opportunities to understand chromosome evolution across vertebrate lineages, so I developed a new tool to analyze and visualize chromosomal rearrangements between species with synteny of singleton orthologous gene sets.

It was enough to demonstrate longer reads can generate longer assemblies and to find several examples of better gene contents corrected in the new long read assemblies which were erroneously missing or duplicated in previous short read genome assemblies. As the first VGP collaboration in South Korea, here, I applied the VGP standard assembly pipeline version 1.6 to generate high quality reference genome of a Korean endemic species, Korean giant-fin mudskipper (*periophthalmus magnuspinnatus*).

However, there was absent for any systemized method to evaluate gene content quality by comparing the different versions of reference genomes of same species. My team developed two methods by combining existing comparative genomics approaches to detect erroneous regions not only in the prior genome assemblies but also in the new one [11,12]. In this thesis, I described my contributions in both studies to generalize the tendency of false missing and duplications in other short read assemblies of vertebrates.

1.2. Comparative genomics for specific traits

"What does make us human?" It is the main question that I had started the master and Ph.D. courses. I believe that language and tool developments are important key traits to build prosperous civilizations of mankind, vocal learning and limb developments were regarded as fundamental traits of language and tool

developments, respectively.

Vocal learning is a specific ability to imitate sound of same or other species. It is rarely observed in a few animals, such as human, some bats, dolphins, whales, elephants, seals, songbirds, parrots, and hummingbirds [13,14]. To understand how to get vocal learning ability, there were various comparative approaches between vocal learners and non-learners. Comparative genomic approaches could detect two major types of variants: regulatory variants for gene expression alterations (heterometry, heterotopy, and heterochrony) and coding variants for gene product alterations (heterotypy) [15]. As representative examples of gene product alterations related to language and vocal learning, Lai *et al.* found an amino acid substitution (R553H) on *FOXP2* gene which could explain a hereditary language disorder without any obvious neurological, anatomical, or physiological cause in KE family in United Kingdom[16]. As a follow-up study, Enard et al. found two human-specific amino acid substitutions on *FOXP2* gene mutually exclusive to several vocal non-learning animals, such as, chimpanzee, gorilla, mouse, and chicken [17]. Over the *FOXP2* gene, Zhang et al. performed genome-wide approaches to detect amino acid substitutions specific to vocal learning birds mutually exclusive to vocal non-learning birds, but they did not explain direct relationships between the substitutions and vocal learning ability [8]. By considering convergence at molecular level, Parker et al. identified convergent amino acid substitutions specific to echolocating animals which are also regarded as vocal learners and found the convergences associated with numerous genes for hearing or deafness [18]. However, it faced critical debates for the genome-wide convergent amino acid substitutions were frequently observed on similar sensory genes in the closest control set [19].

To detect more reliable candidate genes and variants, here, I applied systemized approaches for gene product alterations with multiple lines of evidence especially for site-wise positive selection on amino acid substitutions [20]. Additionally, I attempted to discover basic rules of molecular convergences by investigating phylogenetic features and its underlying variants at codon and nucleotide level.

This chapter is under review in *Scientific data*
as a partial fulfillment of Chul Lee's Ph.D. program

Chapter 2. Chromosome-level genome assembly of Periophthalmus magnuspinnatus: an indigenous mudskipper in the Yellow Sea

2.1. Abstract

Giant-fin mudskipper, *Periophthalmus magnuspinnatus* (PM), is an important euryhaline fish for evolutionarily and ecologically. It lives endemically on coastal mudflats of the Yellow Sea, adapted to life both in and out of water, and has a potential as a bio-indicator to monitor environmental changes. The previous Sanger-based reference genome of PM provided a resource to understand molecular mechanisms of its land adaptation. However, it was too fragmented to analyse chromosome structures. As part of the Vertebrate Genomes Project, here I generated a *de novo* chromosome-scale genome assembly of PM (fPerMag1) by using multiple sequencing technologies: PacBio CLR, 10X linked reads, Bionano optical maps, and Arima Hi-C paired reads. I assembled a 753 Mb genome with 25 chromosomes, which is 100-fold more contiguous than the previous assembly. Of these chromosomes, 60% included telomeric repeats at the 5' and 3' ends. I detected a total 27,880 genes based on the NCBI annotation and the additional annotation that included long-read transcriptome data. The new fPerMag1 assembly provides unprecedented opportunities to investigate chromosomal evolution across *Gobiidae* fishes.

2.2. Introduction

Mudskippers, *Oxudercidae* subfamily in *Gobiidae* family, occupy an important ecological niche, and are therefore useful models to understand both aquatic and terrestrial adaptations. This fish lineage has amphibious abilities, such as breathing air and walk-like behaviour on land. On land they use their mouth and throat to breathe, and under water they use their gills[21,22]. Although their side pectoral fins are anatomically different from limbs of tetrapod animals, it is functionally similar to legs of human and other animals that walk upright and leap on land, including coastal mudflats[23]. Mudskippers are also regarded as important biological indicators of pollutions in coastal ecosystems. They have high tolerance for various types of pollutants, so they can be used to investigate the environmental pollutions of their habitats[24].

The giant-fin mudskipper, *Periophthalmus magnuspinnatus* (PM), is one of *Periophtalmus* species which is known to adapt primarily to terrestrial environments compared to the other main genera in Gobiidae family[25] (**Figure 2.1a, b**). This fish lives endemically in the Yellow Sea[26,27] (**Figure 2.1c, d**). This sea has one of the largest intertidal mudflats in the world, which is an important stopover habitat of migratory shorebirds of the East Asian-Australasian Flyway[28]. However, this species suffers from environmental changes associated with the rapid loss of tidal wetlands[29] and increased pollutants[30,31]. This species could be useful to study molecular mechanisms of adaptive traits for both land and water habitats and to use it as the bio-indicator of changing ecosystems of the Yellow Sea. Investigations into these areas would benefit from a high-quality reference genome sequences for the PM.

Figure 2.1. Morphological features of the Korean giant-fin mudskipper and sampling location. (a) Morphology of the sequenced *Periophthalmus magnuspinnatus* individual (fPerMag1) with a big 1st dorsal fin and a distinct horizontal line in the middle of 2nd dorsal fin. (b) Habitat of the mudskipper nearby a *Suaeda* plant (seepweed) on the mudflats of the Yellow sea. (c, d) Geographic locations of the individual of the fPerMag1 assembly.

2.3. Materials and methods

Sample collection, species identification, and tissue isolations

Six adult PMs were caught in the Seon-Du ri 4 port, Gil-Sang myun, Gang-Hwa gun, Incheon, the Republic of Korea (37.604181°N, 126.480635°E) based on morphological species identification considering PM-specific features: a big 1st dorsal fin and a distinct horizontal line in the middle of 2nd dorsal fin. I placed them in plastic box with sea water and kept them alive for transport to the lab.

For molecular species identification of the six individuals, small 5mm chunks of the tail fins of each individual were cut, and were placed in test tubes and then in liquid nitrogen. The fin tissue samples were rinsed with distilled water and brought to the room temperature (25°C) for DNA extraction. Genomic DNAs were extracted from the tail fin tissue samples using the MFX-6100 automated DNA extraction system (Toyobo, Osaka, Japan) with MagExtractor genome DNA purification kit (Toyobo, Osaka, Japan). The extracted DNAs were examined by electrophoresis with 1% agarose gel, and the concentration were quantified with a NanoVue spectrophometer (GE Healthcare, Little Chalfont, UK). Partial sequences of mitochondrial cytochrome oxidase subunit I (COI) gene were amplified by polymerase chain reaction (PCR) using the universal primers VF2_t1 (5'-CGCCTGTTTATCAAAAACAT-3') and FishR2_t1 (5'-ACTTCAGGGTGACCGAAGAATCAGAA-3') (Ward et al. 2005). PCRs were carried out in 20 μl containing 1μl extracted DNA, 2.5U of ExTaq (Takara Bio, Tokyo, Japan), 2 μl of 10X ExTaq buffer, 1.6 μlof dNTP mixture (10 mM), and 10 pmol of each primer. Amplifications were performed using a ABI Veriti thermal cycler (Applied Biosystems, CA, USA) in the following conditions: initial denaturation at 94℃ for 7 min, 35 cycles of denaturation at 94℃ for 1 min, annealing at 52℃ for 1min, and extension at 72℃ for 1min, and final extension at 72℃ for 7 min. PCR products were visualized by electrophoresis with 1.5% agarose gel and purified using a Expin™ PCR SV purification kit (GeneAll, Seoul, South Korea) according to the manufacturer's instructions. Purified PCR products were resolved

on an ABI 3730 automated DNA capillary sequencer (Applied Biosystems, CA, USA) with a BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems, CA, USA).

Sequencing was performed by the sanger sequencing method, whereupon the 650 bp fragment of COI gene was obtained from each sample. These sequences were identified by comparing to reference sequences in the GenBank database using BLAST algorithm (http://balst.ncbi.nlm.nih.gov.Blast.cgi). All samples had a COI gene that exhibited high similarity with *Periophthalmus magnuspinnatus* (KT951744), ranged from 99.9% to 100%.

In order to prepare tissue samples for the new genome assembly of PM, the largest individual (length=9 cm) was selected, and the remaining fish brought back to the habitat and released. This largest fish was anaesthetized by immersion in 0.05% 2-phenoxyethanol (≥99%, Sigma-Aldrich) for 5 min. After anesthetization, the brain, liver, gill, ovary (female), and muscle tissues were dissected as small 5mm chunks from the individual on a water-ice block, and then were frozen in liquid nitrogen immediately. All of tissues were stored in the cryogenic refrigerator (-80°C) before DNA and RNA extraction.

Genomic DNA extraction and sequencing libraries

I used the gill tissue to generate high molecule weight DNA, Bionano Prep™ Animal Tissue DNA Isolation Fibrous Tissue Protocol was applied. All of genomic raw data were sequenced by Single Molecule Real-Time (SMRT) of Pacific bioscience (PacBio) continuous long reads (CLR)[32], optical mapping of Bionano genomics[33], and HiSeq of Illumina with library constructions using linked reads of 10x genomics[34] and Hi-C of Arima genomics[35] by following each protocol.

Chromosome-level assembly based on 4 types of genomic data

fPerMag1 genome was assembled with the VGP assembly standard pipeline version 1.6[10] (**Figure 2a**). Based on PacBio CLR data, I generated the primary contigs (c1) and alternative haplotigs (c2) by using FALCON[36] and FALCON-Unzip[37]. To discard false duplications in the primary contig set (c1→p1), I ran Purge_Dups[38]. Using 10X Genomics linked reads, I generated the first primary scaffolds (p1→s1) by using scaff10x[39]. For the Bionano optical maps and the s1 assembly, I applied Bionano

Solve[40] with the DLE-1 one-enzyme non-nicking approach and generated the s2 assembly (s1→ s2). I then aligned Arima Genomics Hi-C reads, which reflects the 3D structures of each chromosome into the genome assembly (s2→s3), to the s2 scaffolds[41] and used including Salsa2[42] to scaffold them further. To polish any base errors, in the s3 assembly, I applied Arrow (smrtanalysis 5.1.0.26412) with PacBio CLR reads (s3→t1) and FreeBayes[43] with linked reads (t1→t2-3), respectively. The resulting primary assembly and alternative haplotigs were named 'fPerMag1.pri.asm' and 'fPerMag1.alt.asm'. Lastly, I manually curated the automated assembly using gEVAL[44,45] (https://vgp-geval.sanger.ac.uk/index.html) to remove remaining contamination and false haplotype duplications. After 390 manual interventions (break and joins) to correct structural errors, the scaffold number was reduced by 56%, increasing the scaffold N50 by 4%. The curation process identified and named 25 chromosomes-level scaffolds accounting 99.5% of the assembly sequence.

Summary plots of genome assemblies of PM

To compare fPerMag1 and previous PM assemblies, a dot plot was generated with D-genies[46] using the fPerMag1 primary assembly as a reference, pre.PM assembly as query, and default options. To summarize genomic features of fPerMag1 assembly, a circos plot was generated with OmicsCircos package[47] in R version 3.5.3[48].

Telomeres at 5' and 3' ends of chromosomes

To investigate telomeric repeats conserved in vertebrates[49], I developed a custom script (Python version 3.7.3) to identify the '(TTAGGG)n' sequence and its complimentary '(CCCTAA)n' sequence (n≥2) in the fPerMag1 primary genome assembly, with an output in 'bed' format. I used bedtools (v2.26.0) with the 'intersect –wa' option to check whether the telomeric sequences overlapped with repeat sequences detected by RepeatMasker version 4.0.8 (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/009/829/125/GCF_009829125.1_fPerMag1.pri/GCF_009829125.1_fPerMag1.pri_rm.out.gz). For chromosomal and unplaced scaffolds with telomeric repeats, I manually validated the telomeric repeats by visualize of 30kbp regions on the 5' and 3' ends of each scaffold, using IGV[50].

Transcriptomic RNA extraction and gene annotations

I applied the NCBI Eukaryotic Genome Annotation Pipeline v8.4[51] without

transcriptome data of the fPerMag1 individual. For additional annotation, I generated PacBio Iso-Seq transcriptomes of 5 type tissues (brain, ovary, muscle, liver, and gill) of the same individual used for the fPerMag1 assembly. I extracted RNA by using the Iso-Seq SMRTbell library by following its protocol. The additional annotation was generated by AUGUSTUS[52], following the protocol for PacBio IsoSeq[53]. I used default options for each step (**Alternate Protocol 1**: *GENERATING TRAINING GENE STRUCTURES FROM PROTEINS* based on the NCBI annotation release 100, **Alternate Protocol 4**: *TRAINING AUGUSTUS FOR A NEW SPECIES*, **Alternate Protocol 6**: *GENERATING HINTS FROM IsoSeq DATA*, and **Basic Protocol 4**: *RUNNING AUGUSTUS WITH HINTS*).

Data Records

All of raw data, intermediates, and the final chromosome-level assembly of fPerMag1 assembly were deposited in the genome ark of the Vertebrates Genome Projects (https://vgp.github.io/genomeark/Periophthalmus_magnuspinnatus/) and the NCBI database (GCF_009829125.1).

Technical Validation

To validate improvement of assembly quality of fPerMag1, I performed Benchmarking Universal Single-Copy Orthologs (BUSCO v5.2.2) analysis[54] with following options: "-l vertebrata_odb10 -m genome -c 5 --augustus_species human".

Code Availability

The VGP assembly standard pipeline version 1.6 is available (https://github.com/VGP/vgp-assembly). The scripts and raw data for statistics and the additional annotation are available at (https://github.com/chulbioinfo/fPerMag1).

2.4. Results and Discussion

The previous PM genome assembly (GenBank: GCA_000787105.1) provided insights into terrestrial adaptive traits of mudskippers, such as the immune system, ammonia excretion, aerial vision, and response to hypoxia[25,55]. However, the assembly has the limitation of short read and non-phased assemblies, including false duplications and false breakages of scaffolds causing mis-annotations[11,12,56]. As a result, this prior Sanger-based assembly was highly fragmented with 26,060 scaffolds and a low scaffold N50 of 0.296 Mb (**Table 2.1**), making not suitable for genome-wide analyses, including on structural variants at the chromosome level.

Here, I generated a *De Novo* chromosome-scale assembly of PM (fPerMag1, GenBank: GCA_009829125.1) by following the Vertebrate Genomes Project (VGP) standard pipeline v1.6[10], which uses a combination of 4 sequencing technologies: PacBio Continuous Long Reads (CLR), 10X Genomics linked reads, Bionano optical maps, and Arima Genomics Hi-C paired reads (**Figure 2.2a, Table 2.1, 2.2**). GenomeScope[57] based kmers analyses on the 10X Genomics linked reads estimated its genome size as ~634 Mbp, but the fPerMag1 assembly is approximately ~753 Mbp. After scaffolding, I identified 25 chromosome-level scaffolds supported by the Hi-C data, with 99.5% of the assembled bases assigned to chromosomes. The repeat content was increased at the 5' and 3' ends of each chromosome, which were conserved vertebrate telomere sequences, (TTAGGG)n[49] in 60% (15 out of 25) of the chromosomes (**Figure 2.2b, Table 2.3**). As an example, chromosome 10 showed long conserved telomeric simple repeats of 12 and 15 kbp at both 5' and 3' ends, respectively (**Figure 2.2d**). On the other hand, 6 unplaced scaffolds also had telomeric repeats at 5' or 3' ends (**Figure 2.2b, Table 2.3**) indicating that they were not yet placed into specific chromosomes.

Table 2.1. Summary of assembly statistics of giant-fin mudskipper genomes.

	Previous PM	***fPerMag1***
GenBank Accession ID	GCA_000787105.1	GCA_009829125.1
Technology	Sanger	Pacbio/ Bionano/ 10x*/ Arima Hi-C*
Genome coverage	77x	90x/ 962x/ 95x/ 125x
Assembly software	Soapdenovo v. 2.04	VGP standard 1.6 pipeline
Number of contigs	76,770	822
N50 of contigs (Mb)	0.028	2.3
Number of scaffolds	26,060	124
N50 of scaffolds (Mb)	0.296	32.9
Total length (Mb)	701.7	752.6
Number of chromosomes	N/A	25
Chromosome length (Mb)	N/A	749
BUSCO (n=3,354)	C:93.8%[S:93.4%,D:0.4%], F:3.9%,M:2.3%	C:96.6%[S:95.9%,D:0.7%], F:1.2%,M:2.2%

Table 2.2. Genomic raw data of the fPerMag1 assembly.

	Pacbio SMRT SubReads	***Bionano optical map***	***10x Genomics linked reads****	***Arima Hi-C****
Approximate Coverage	90x	962x	95x	125x
Download Size (Gbp)	121.200	1.419	34.457	57.519
Download Link	aws s3 --no-sign-request sync s3://genomeark/species/Periophthalmus_magnuspinnatus/fPerMag1/genomic_data/pacbio/ . --exclude "*scraps.bam* --exclude "*ccs.bam*"	aws s3 --no-sign-request sync s3://genomeark/species/Periophthalmus_magnuspinnatus/fPerMag1/genomic_data/bionano/ .	aws s3 --no-sign-request sync s3://genomeark/species/Periophthalmus_magnuspinnatus/fPerMag1/genomic_data/10x/ .	aws s3 --no-sign-request sync s3://genomeark/species/Periophthalmus_magnuspinnatus/fPerMag1/genomic_data/arima/ .

Table 2.3. Summary of repeats with vertebrate telomeric sequences, (TTAGGG)n, in fPerMag1 primary assembly.

Chr. name	*Scaffold name*	*Len. of scaffold bp*	*Len. of (TTAGGG)n bp*	*Telomeric repeats in 30kbp at 5' end bp (# repeats)*	*Telomeric repeats in 30kbp at 3' end bp (# repeats)*
1	NC_047126.1	36,052,970	4,536	0 (0)	0 (0)
2	NC_047127.1	19,540,339	354	0 (0)	0 (0)
3	NC_047128.1	36,957,123	6,984	4,579 (2)	1,374 (2)
4	NC_047129.1	35,021,165	11,994	0 (0)	11,608 (1)
5	NC_047130.1	34,011,676	684	0 (0)	0 (0)
6	NC_047131.1	33,835,733	1,914	5,757 (21)	0 (0)
7	NC_047132.1	34,908,682	2,940	0 (0)	45 (1)
8	NC_047133.1	29,687,151	462	0 (0)	0 (0)
9	NC_047134.1	36,171,850	15,366	4,471 (2)	9,688 (2)
10	NC_047135.1	36,988,977	33,756	12,250 (2)	15,743 (1)
11	NC_047136.1	30,435,919	13,062	13,762 (1)	0 (0)
12	NC_047137.1	19,532,153	14,568	0 (0)	15,507 (1)
13	NC_047138.1	32,620,125	14,274	1,343 (3)	0 (0)
14	NC_047139.1	32,379,725	432	0 (0)	0 (0)
15	NC_047140.1	32,865,169	444	0 (0)	0 (0)
16	NC_047141.1	33,977,497	20,886	3,580 (17)	18,997 (1)
17	NC_047142.1	31,521,007	5,988	5,245 (4)	0 (0)
18	NC_047143.1	28,378,236	10,572	0 (0)	7,042 (1)
19	NC_047144.1	29,583,991	1,374	0 (0)	140 (3)
20	NC_047145.1	28,068,090	5,292	0 (0)	2,745 (12)
21	NC_047146.1	33,436,419	486	0 (0)	0 (0)
22	NC_047147.1	29,018,424	402	0 (0)	0 (0)
23	NC_047148.1	24,170,445	5,418	0 (0)	0 (0)
24	NC_047149.1	27,914,776	16,038	0 (0)	16,596 (1)
25	NC_047150.1	2,028,439	384	0 (0)	0 (0)
-	NW_022986699.1	63,897	15,474	10,669 (1)	4,846 (3)
-	NW_022986717.1	37,404	3,582	9,668 (35)	2,936 (14)
-	NW_022986752.1	15,992	1,338	2,751 (9)	3,832 (2)
-	NW_022986775.1	5,726	432	1,946 (7)	0 (0)
-	NW_022986778.1	3,561	828	1,200 (2)	0 (0)
-	NW_022986786.1	156,916	10,512	0 (0)	10,641 (1)

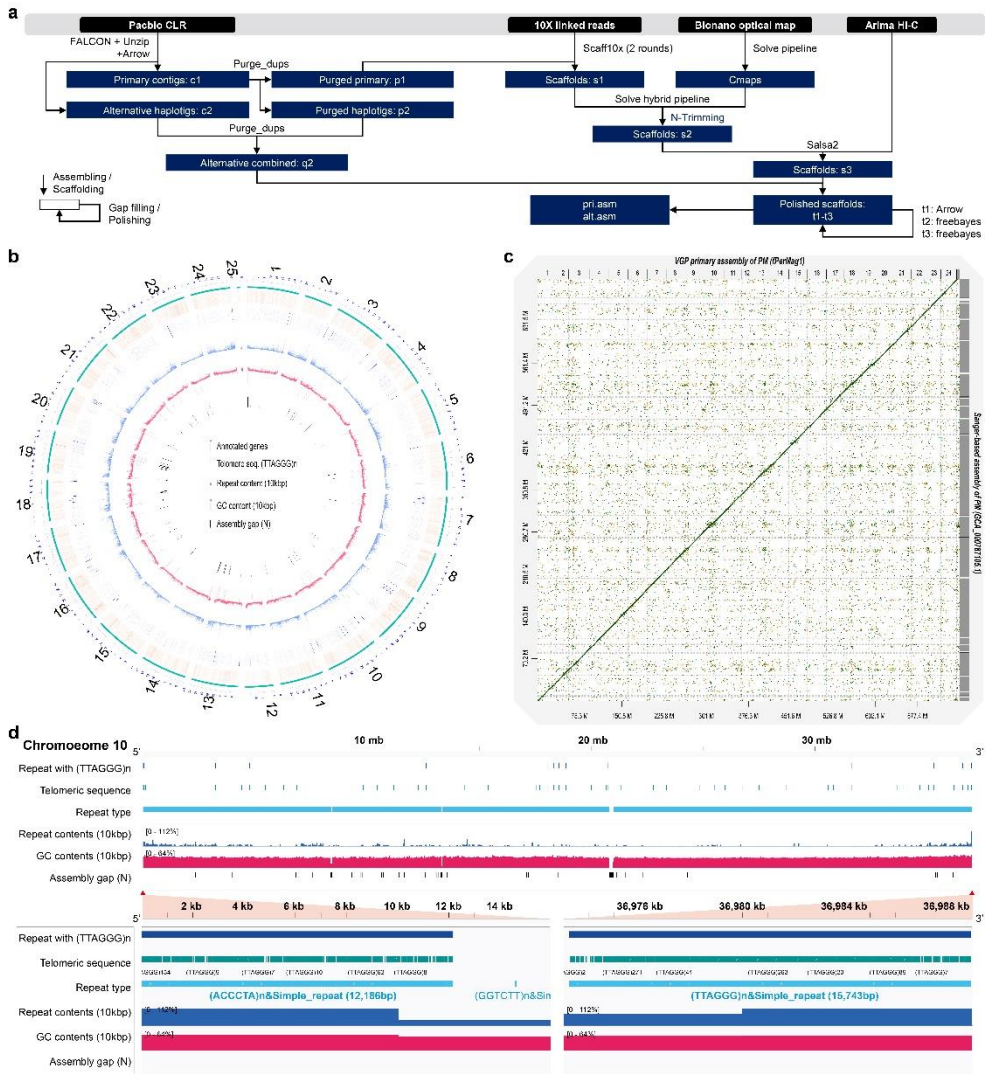

Figure 2.2. Workflow and summary of chromosome-level assembly of the Korean giant-find mudskipper (fPerMag1). (a) Schematic overview of VGP standard assembly 1.6 pipeline, modified from Rhie, et al. [10]. Black boxes in the grey shadow box indicate raw data of 4 sequencing technologies. Texts nearby strait arrows and the 'U' turned arrow indicate assembling or scaffolding programs, and polishing programs, respectively. Blue boxes with arrows indicate outputs of each assembly step. (b) Circos plot for assembled chromosomes of the fPerMag1 primary assembly. From the out-layer: black numbers, chromosome names; orange bars, with 5% transparency, location of genes in the complete annotation (n= 27,880) merged from NCBI and additional annotations; blue scale bars, sequence lengths (Mbp) of each chromosomal scaffold; blue bars with 25% transparency, conserved vertebrate

telomere sequence (TTAGGG)n (n>=2); blue and red lines, repeat rates and GC rates in 10Kbp window, respectively; black bar graph, rates of assembly gaps in 10Kbp window. This circus plot was generated with the OmicCircos package [47] in R version 3.5.3 [48]. (c) Dotplot for the previous sanger-based assembly and the new fPerMag1 primary assembly. This plot was generated with D-genies [46]. (d) Chromosome 10 highlighting the conserved vertebrate telomeric repeats, (TTAGGG)n. Red triangles indicate telomeric regions at the 5' and 3' ends (16kb window) and zoom in below. Blue, turquoise, and sky blue bars indicate telomeric repeat region, telomeric sequences, and all repeats, respectively, detected by RepeatMasker. Blue and red bar graphs indicate repeat and GC contents with 10kb windows. Black bars indicate assembly gaps (N).

Although the previous PM assembly aligned well to the new fPerMag1 assembly, the previous PM assembly was much more fragmented (**Figure 2.2c**). The fPerMag1 assembly was 100 times more contiguous with a contig N50 of 2.3 Mbp and scaffold N50 of 32.9 Mbp relative to a contig N50 of 0.028 Mbp and scaffold N50 of 0.296 Mbp of the previous assembly. The improved continuity of fPerMag1 assembly was mainly based on the CLR reads for contigs and with secondaly on the scaffolding steps with Bionano optical maps increasing scaffold N50 from 4.96 Mbp to 25.9 Mbp (**Table 2.4**).

To validate quality improvements of fPerMag1 assembly compared to the previous PM assembly, I conducted BUSCO analyses[54] for both assemblies. The complete BUSCO genes increased from 93.8% (S: 93.4%, D 0.4%) to 96.6% (S:95.9%, D: 0.7%) and fragmented BUSCO genes were decreased from 3.9% to 1.2% in the new assembly (**Table 2.1**).

By applying the NCBI Eukaryotic Genome Annotation, I identified 24,742 genes including 21,306 protein coding genes, but also found the previous PM assembly had more protein coding genes, 22,256 (**Table 2.5**). In order to detect more genes, I generated long read transcriptomes from multiple-type tissues of the same individual as the fPerMag1 assembly using Pacbio Iso-Seq (**Table 2.6**) and applied an additional annotation using AUGUSTUS based on the transcriptome data. I identified 3,438 additional protein coding genes supported by the RNA read mapping, which were mutually exclusive to the NCBI annotation (**Figure 2.3, Table 2.5**).

I believe my high-quality assembly fPerMag1 provides opportunities to identify sequence and structural variants related to land adaptation of PM, to compare populations of distant coastal regions to trace changing ecosystems of the Yellow Sea, and to start cytogenomics to decode chromosomal evolution across *Gobiidae* fishes.

Table 2.4. Summary of intermediates of the fPerMag1 assembly.

Assembly ID	*Assembly_level*	*Input*	*Total lengths*	*# Contigs or # Scaffolds*	*Max*	*N50*
c1	Contigs	Pacbio	979,057,940	2,488	8,646,240	1,042,800
p1	Purged_contigs	c1 + Pacbio	749,730,448	1,160	8,646,240	1,323,184
s1	Scaffolds 1	p1 + 10x	749,796,148	503	27,164,672	4,962,451
s2	Scaffolds 2	s1 + Bionano	772,959,812	340	34,410,915	25,943,279
s3	Scaffolds 3	s2 + Hi-C	772,995,812	279	37,210,474	31,615,909
t1	Polished_scaffolds 1	s3 + Pacbio	773,067,407	279	37,214,669	31,618,202
t2	Polished_scaffolds 2	t1 + 10x	773,026,920	279	37,212,511	31,616,735
t3	Polished_scaffolds 3	t2 + 10x	773,021,550	279	37,212,213	31,616,471

Table 2.5. Summary of gene annotations of giant-fin mudskipper.

	Previous PM	***fPerMag1***	
Annotation version	Ensembl (ver. 105.1)	NCBI annotation (rel.100)	Additional annotation
Source	Ensembl resources	NCBI resources	Pacbio Isoseq
# transcript reads	0	0	163,001
Annotation software	Ensembl Gene Annotation (e!94) Fish Clade (Full genebuild)	NCBI Eukaryotic Genome Annotation Pipeline (Gnome)	AUGUSTUS
# genes	24,197	24,442	3,438
# protein coding genes	22,256	21,306	3,438

Table 2.6. Transcriptome raw data of fPerMag1 assembly.

Pacbio SMRT Iso-Seq reads

Tissue	Ovary	Muscle	Liver	Gill	Brain
Download Size (Mbp)	22.7	50.4	18.2	29.4	22.3
NCBI SRA Accession	SRX8147373	SRX8147372	SRX8147371	SRX8147370	SRX8147369

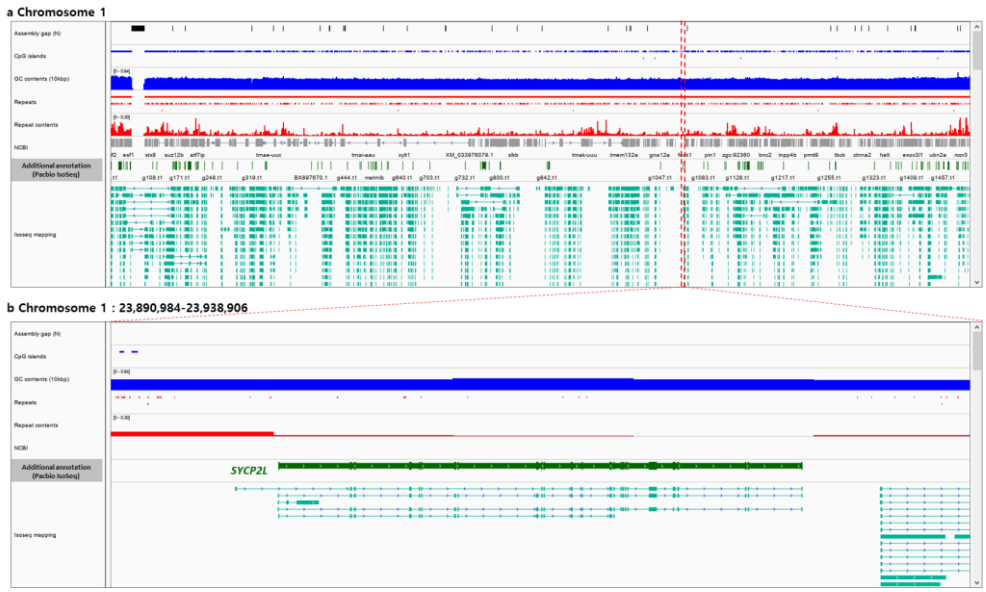

Figure 2.3. Updated gene annotation based on PacBio Iso-Seq transcriptomes. (a) Example genes in chromosome 1 that were newly detected by the additional annotation based on the PacBio Iso-Seq transcriptome data from 5 types of tissue of the same individual used for the fPerMag1 assembly. (b) *SYCP2L* gene (g1057), a representative example on chromosome 1. Grey and green structures show annotated genes in the NCBI annotation (release 100) and the additional annotation generated by AUGUSTUS with PacBio Iso-Seq, respectively. Turquoises indicates mapped reads of PacBio Iso-Seq transcriptomes of the fPerMag1 individual.

This chapter was published in *BioRxiv* and *Nature*
as a partial fulfillment of Chul Lee's Ph.D. program

Chapter 3. Comparative genomic approaches to detect erroneous genes in reference genomes and to visualize chromosomal evolutions of vertebrates

3.1. Abstract

High-quality and complete reference genome assemblies are fundamental for the application of genomics to biology, disease, and biodiversity conservation. However, such assemblies are available for only a few non-microbial species. To address this issue, the international Genome 10K (G10K) consortium has worked over a five-year period to evaluate and develop cost-effective methods for assembling highly accurate and nearly complete reference genomes. Here, I present lessons learned from generating new chromosome-level reference genomes of 16 species that represent six major vertebrate lineages. my new assemblies correct substantial errors by adding falsely missing sequences and by removing false duplications in some of the best historical reference genomes and can prevent misinterpretations for their biological effects on various traits. I discover that the missing errors were related to high GC and repeat contents leading failure of sequencing that do not originate from individual differences. Conclusively, these false missing and duplications repeatedly occurred in other short-read based genome assemblies of vertebrates. I reveal chromosome rearrangements that are specific to lineages by developing a method, ChrOrthLink. My findings provide unprecedented insights to chromosomal evolution across vertebrate lineages and discover reasons of wide-spread false gene losses and gains which are now rectified in the Vertebrate Genomes Project reference genomes.

3.2. Introduction

Reference genome sequences are fundamental to bioinformatic applications in various fields, such as, biology, heath medicine, agriculture, and ecology. The first generation of genome projects to build reference genomes were initiated with Sanger sequencing technologies (read length 700-1000 base pair, bp) and chromosome genetic maps for human [1] and representative model species in various lineages such as *Caenorhabditis elegans* [2], *Arabidopsis thaliana* [3], and *Mus musculus* [4]. About 100 reference genomes of vertebrates were published by 2010 mostly using Sanger reads. The next generation sequencing (NGS) technologies, such as, Illumina platform (100-150 bp), rapidly increased the throughput based on sequencing by synthesis and gradually increased the number of published reference genomes to about 700 by 2018 [58].

These accumulations of short read based reference genomes provided unprecedent opportunities to understand macroevolution across species or clades by providing control data sets, but these shorter read assemblies caused lower-quality problems with fragmentation errors [9]. Moreover, many genes in the short read-based reference genomes various species which are involved in traits were missing or duplicated totally or partially resulting in misinterpretation indistinct for real biological variations or errors in assemblies [9,59-61].

As a promising solution, longer read sequencing technologies, such as, Pacbio continuous long reads (CLR, 1000-60000bp) [62], hugely improved the continuity in contig assembly process [9]. In addition, new technologies were developed for scaffolding processes, such as, 10X Genomics linked-reads, Bionano Genomics optical mapping, and Arima Genomics Hi-C reads [63-65]. Vertebrate Genomes Project (VGP) attempted to suggest optimized pipelines by combining above technologies and constructed chromosome-level reference genome assemblies [10]. It successfully demonstrated that longer reads can generate better continuity of assemblies.

However, it was still ambiguous whether longer reads could generate better

gene contents of assemblies than short reads or not. Here, I developed two systemized methods using genome-wide alignment program, CACTUS, to detect and quantity erroneous regions in the prior genome assemblies. I also confirmed the general tendency of false gene losses and gains in previous short read assemblies of various species in the vertebrate lineage. Additionally, in order to find novel discoveries for chromosome evolution based on chromosome-level reference genomes generated by VGP, I developed a new tool to analyze and to visualize chromosomal rearrangements across 6 major vertebrate lineages.

3.3. Materials and methods

False gene annotation in previous assemblies of same species

I detected evidence of erroneous coding sequences in previous assemblies of the zebra finch, platypus, and climbing perch for the genes which are related to specific complex traits[66,67] or, included in the BUSCO gene set[26,27]. To identify the erroneous annotations, such as false duplications or truncated sequences due to missamblies, I collected exon sequences from the VGP annotation of the genes and performed blastn v2.6.0+ searches[37] against both the previous and VGP assembly, with options -task blastn, -perc_identity 90, and -evalue 0.00001. Among the hits found from the blast search, I defined false duplications of an exon when duplicated hits within the same scaffold were found on the previous assembly only. Also, I detected truncated exons, where the length of the blast hit was shorter than the length of query exon. For visualization, I used Gene Structure Display Server 2.0+[38] and manually modified the display in order to handle small discrepancies between elements. For the intuitive visualization of platypus' vitellogenin-2 gene, I visualized only the scaffolds with more than three blast hits of the previous assembly.

Falsely duplicated MTOR genes in other reference genome assemblies

To test for possible false duplications of the MTOR gene in other published genome assemblies of vertebrates, I extracted 449 RefSeq annotated genomes of 330 vertebrate species from NCBI, and found 38 assemblies have the original *MTOR* gene and at least 1 *MTOR*-like genes, respectively. I parsed the genic sequences of each gene from each assembly, aligned them for each species by using LAST [69], checked the *MTOR*-like harboring scaffolds were fully aligned to parts of the genic region of the *MTOR* genes, calculated proportions (>50%) of lengths of *MTOR*-like genes per scaffold with the duplicated genes, and considered the qualities and quantities of sequencing reads used to the generate assemblies. Following the above steps, I identified 4 assemblies of 4 species (*Bubalus bubalis*, *Tinamus guttatus*, *Scleropages formosus* and *Bufo gargarizans*) that have scaffolds with duplicated

MTOR-like genes. For the 4 candidate assemblies, I mapped raw sequencing reads used to generate each assembly and applied purge_dups and assessed whether they are false duplications. I discovered 2 assemblies out of the 4 assemblies (i.e. species; white-throated tinamou [*Tinamus guttatus*] and domestic water buffalo [*Bubalus bubalis*]), contained erroneous scaffolds with falsely duplicated *MTOR*-like genes.

Distinguishing falsely missing regions from individual differences versus technical errors

To distinguish between assembly differences versus biological individual differences for the platypus and climbing perch, I performed mapping of prior Sanger and Illumina reads onto each VGP genome assembly by using minimap2[68] (v2.22-r1105-dirty) with the options: -ax map-pb and -ax sr for the Sanger reads of the prior platypus and Illumina paired-reads of the prior climbing perch, respectively. I calculated read depths of the prior reads mapped onto the VGP assemblies, and output it in 'psl' format using igvtools[69] (v2.11.1) with the option: -count. In parallel, to analyze prior assembly gaps, I converted cactus genome alignments formats between the prior and VGP assemblies of each species from '.hal' to '.maf' of '.psl' by using HAL[70]. Using a custom python script (https://github.com/chulbioinfo/FalseGeneLoss), I investigated proportions of nucleotide sites of VGP assemblies homologous to missing regions in the previous assemblies that were supported by the prior reads with 1x depth cutoff or were aligned to prior assembly gaps ('N').

As a secondary measure, I searched for Benchmarking Universal Single-Copy Orthologs (BUSCO) in the prior assemblies. I assumed that deleted regions in highly conserved genes would more likely reflect incomplete assemblies rather than individual differences in a species. I performed BUSCO analyses[54] (version 5.2.2) on the prior and VGP genome assemblies of platypus and climbing perch with options: -l vertebrata_odb10 -m genome --augustus_species human. I checked the intersections between prior missing BUSCOs and VGP complete BUSCOs, identified overlaps between the lists of missing BUSCOs only in previous assemblies and the lists of missing genes and missing exons, and selected representative

examples. Finally, I manually checked signatures of sequencing errors (depth drops with a few mapped reads and fragmented scaffolds, respectively) as evidence to exclude the possibility of individual differences.

For the prior missing BUSCO gene of the platypus, I analyzed basepair-wise conservation scores calculated by Phylop based on 100-way multiz genome-wide alignments of 100 vertebrates and confirmed the absence of matching regions in the prior platypus assembly. Additionally, I checked GC content of *COQ6* of the platypus and other vertebrates (hg38, mm39, GCF_004126475.2, GCF_000002295.2, GCF_004115215.2, GCF_003957565.2, GCF_007399415.2, GCF_901001135.1, latCha1, GCF_010909765.2, tetNig2, fr3, oryLat2, and gasAcu1 of human, mouse, pale spear-nosed bat, opossum, platypus, zebra finch, Goode's desert tortoise, two-lined caecilian, coelacanth, thorny skate, tetraodon, fugu, medaka, and stickleback, respectively) in UCSC genome browser[71].

Chromosome evolution analyses

As the species divergence were too high to generate a complete genome-to-genome alignment, I estimated chromosome orthology between species by using BUSCO genes. I used the BUSCO gene annotations generated using the vertebrata_odb9 database for the 16 VGP species (mLynCan4, mRhiFer1, mPhyDis1, mOrnAna1, bCalAnn1, bTaeGut1, bStrHab1, rGopEvg1, aRhiBiv1, fGouWil2, fAstCal1, fArcCen1, fCotGob3, fMasArm1, fAnaTes1, and sAmbRad1), and additionally performed the same BUSCO analysis on the primary assembly of the human genome reference (GRCh38.p12). I used ChrOrthLink (https://github.com/chulbioinfo/chrorthlink) to identify and visualize shared 'complete singleton BUSCO genes', which defines 1:1 orthologous chromosomal regions in all species. Among the total gene set, I identified 1,147 vertebrate BUSCO genes that were present and highly conserved as single copy in all 16 VGP species and human assemblies. The transcription start position of each gene was used to link orthologous chromosomes between different species and visualized using genoPlotR v3.5.3[72]. I also calculated the average number of chromosomes that have orthologous

segments between human or skate to all other lineages. All input data and scripts are available on github: https://github.com/chulbioinfo/chrorthlink.

3.4. Results and discussion

Erroneous gene annotations in previous assemblies compared to the new VGP reference genomes of same species

An example of a whole gene heterotype false duplication in the RefSeq annotation of the prior zebra finch Sanger-based reference[51] is the BUSCO gene *SPC25* of the NDC80 kinetochore complex[70], which correctly had only one haplotype copy in the VGP primary assembly (**Figure 3.1a** and **Table 3.1**) and the other in the VGP alternate assembly. The GABA receptor *GABRG2* with specialized gene expression in vocal learning circuits[71] had a partial tandem duplication of four of its 10 exons, resulting in an annotated partial false gene duplication as two adjacent genes (*GABRG2* and *GABRG2*-like) in the prior Sanger-based zebra finch assembly (**Figure 3.1b**). The vitellogenin-2 (*VTG2*) gene, an important component of egg-yolk in all egg-laying species[72], was distributed across 14 contigs in three different scaffolds, two that received two corresponding *VTG2*-like gene locus (LOC) annotations and the third that was mistakenly included as part of the intron of another gene (*Calpain-13*) and that had an inverted non-tandem false exon duplication (red), all together causing false amino acid sequences in five exons (blue), in the prior Sanger-based platypus assembly[43] (**Figure 3.1c**). The BUSCO *YIPF6* gene, associated with inflammatory bowel disease[73], was split between two different scaffolds and, thus, not annotated and presumed to be a gene loss in the prior Illumina-based climbing perch assembly[74] (**Figure 3.1b**). Each of these genes is now present on one long contig, with no gaps and no false gene-region gains or losses in the VGP assemblies, validated in reliable blocks with support from two or more sequencing platforms (**Table 3.1**).

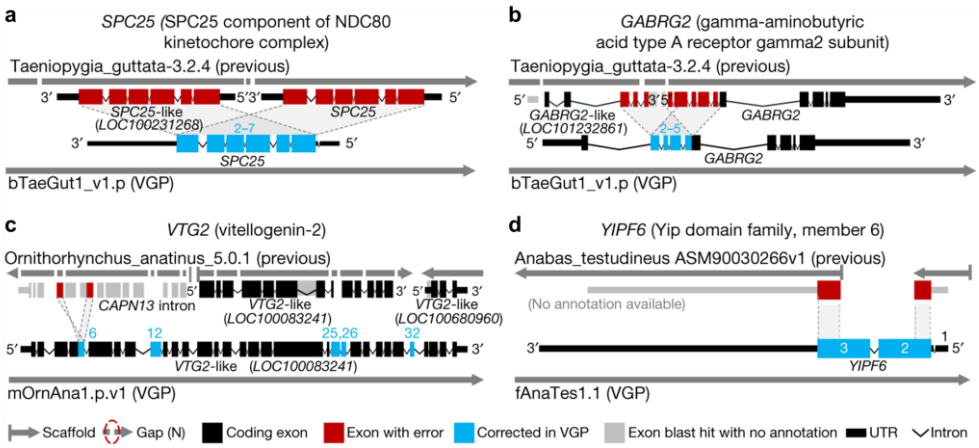

Figure 3.1. Example assembly errors and associated annotation errors in previous (old) reference assemblies corrected in the new VGP assemblies. Both haplotypes of *SPC25* (**a**) were erroneously duplicated on two different contigs, annotating one as *SPC25-like*. The 5' end part of *GABRG2* (**b**) was erroneously annotated as a separate *GABRG2-like* protein coding gene, due to false duplication of exons 2–5. The *VTG2* gene (**c**) was annotated on 3 scaffolds as part of 3 separate genes, two *VTG2-like* and an intron of *CANP13*. *YIPF6* (**d**) was partially missing in the previous assembly due to truncated exon sequences at the scaffold ends. No gene annotation was available for the previous climbing perch assembly. **i**, Gene synteny around the *VTR2C* receptor in the platypus shows completely missing genes (*NUDT16*), truncated and duplicated *ARHGAP4*, and many gaps in the prior Sanger-based assembly compared with the filled in and expanded gene lengths in the new VGP assembly. All examples shown here showed support from at least two technologies across these regions, while the prior assemblies showed hallmarks of misassembly.

Table 3.1. Assemblies and transcripts used to find false exon duplication in previous references. All locus found here were inside a reliable blocks.

Genome	Category	Previous	VGP
Zebra finch	Accession	Taeniopygia_guttata-3.2.4 (GCF_000151805.1)	bTaeGut1_v1.p (GCF_008822105.2)
Fig. 3.2a	Gene annotation	*SPC25 & SPC25-like (LOC100231268)*	*SPC25*
	Transcript ID	XM_012574872.1 (rna9554) XM_002198158.3 (rna9550)	rna-XM_012574872.2
	Gene locus		NC_044219.1:13,659,189-13,663,009
	Reliable block locus	na	NC_044219.1:10,602,325-38,045,260 (Super_Scaffold_7)
Fig. 3.2b	Gene annotation	*GABRG2 & GABR2G2-like (LOC101232861)*	*GABRG2*
	Transcript ID	XM_012575408.1 (rna12930) XM_012575403.1 (rna12929)	rna-XM_030284101.1
	Coordinates		*NC_044225.1:2,970,857-3,030,035*
	Reliable block locus	na	*NC_044225.1:2,268,121-6,009,197 (Super_Scaffold_13)*
Platypus	Accession	Ornithorhynchus_anatinus_5.0.1 (GCF_000002275.2)	mOrnAna1.p.v1 (GCF_004115215.1)
Fig. 3.2c	Gene annotation	*VTG2-like (LOC100083241), VTG2-like (LOC100680960), & CAPN13 intron*	*VTG2*
	Transcript ID	rna-XM_016225321.1 rna-XM_003429627.3	rna-XM_029063584.1
	Coordinates		*NC_041731.1:103,823,950-103,887,329*
	Reliable block locus	na	*NC_041731.1:25,491,142-104,433,552 (Super_Scaffold_4)*
Climbing perch	Accession	ASM90030266v1 (GCA_900302665.1)	fAnaTes1.2 (GCF_900324465.1)
Fig. 3.2d	Gene annotation	*None*	*YIPF6*
	Transcript ID	na	rna-XM_026349816.1
	Coordinates		NC_046630.1:1,721,730-1,724,982
	Reliable block locus	na	NC_046630.1:1,132,484-20,956,182 (Super_Scaffold_8_ctg1)

Specific categories of genes have higher levels of false duplications

Out of falsely duplicated genes in the previous assemblies of zebra finch, Anna's hummingbird, and platypus, *MTOR* gene in all short read assemblies were partially duplicated (**Figure 3.2a, b**, **Figure 3.3a, b**), which regulates growth, metabolism, signaling, and disease with the kinase domain using ATP. Further, by applying purge_dups, I found false gene gains of *MTOR* in other vertebrate species genome assemblies, including the white-throated tinamou and domestic water buffalo (**Figure 3.2c,d**). These assemblies were generated with Illumina short reads only. Their *MTOR*-like harboring scaffolds and the homologous regions in original *MTOR* genes showed read coverages drops to the haploid-level.

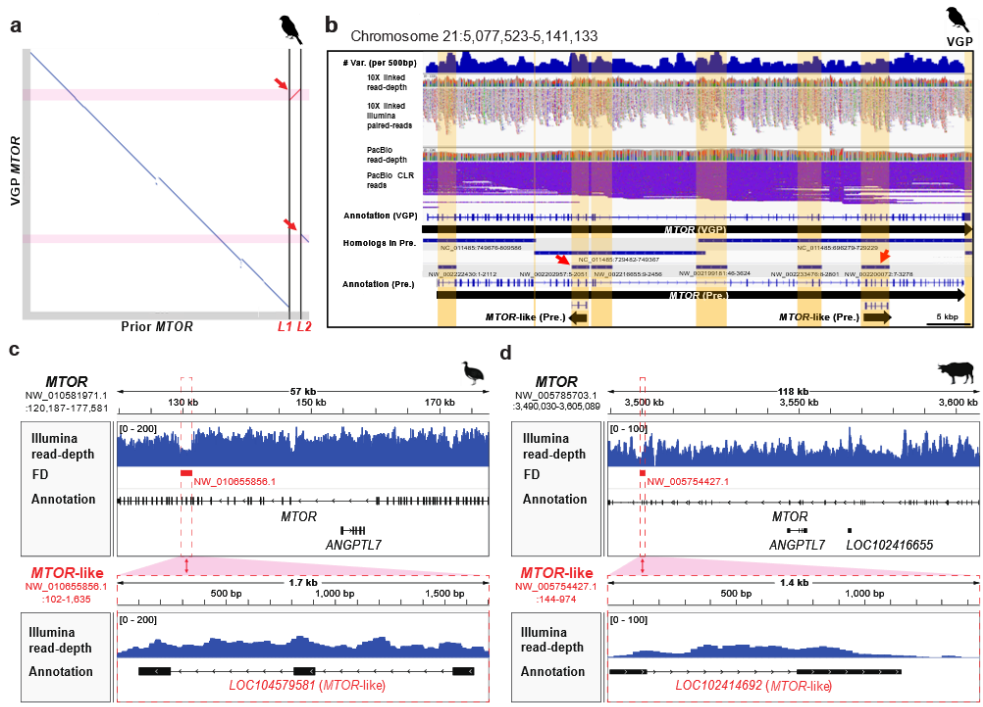

Figure 3.2. False duplication of the *MTOR* gene in vertebrate assemblies. a, Alignment dot plot of the *MTOR* genes in the previous and VGP assemblies of the zebra finch. The alignment of two *MTOR*-like genes in the previous assembly is next to lines L1 and L2 and highlighted in pink**. b,** Genome landscape of the *MTOR* gene in the VGP assembly. Heterozygosity density within 500bp windows is shown at the top. The homologous regions of the previous assembly are represented with blue bars above each genomic position label. The falsely duplicated scaffolds including the *MTOR*-like gene in the previous assembly are shown with red arrows. **c,** False gene gains of the *MTOR* gene in white-throated tinamou (GCF_000705375.1) and **d,** water buffalo (GCF_000471725.1) assemblies. Scaffolds with false duplications (FD) of *MTOR*-like genes were aligned to parts of the original *MTOR* gene and indicated as red dot boxes in each panel.

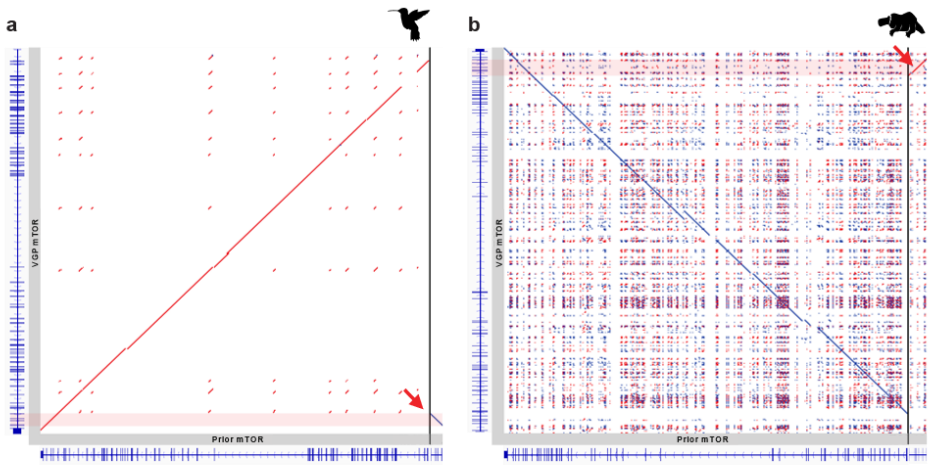

Figure 3.3. False duplications of the *MTOR* gene in the prior hummingbird (a) and the platypus (b) assemblies. Alignment dot-plot shows *MTOR* gene alignment between the previous and the VGP assemblies. The alignment of the *MTOR*-like gene in each previous assembly is marked by a red arrow. The blue bars represent the exons of *MTOR* and *MTOR*-like genes. The platypus *MTOR* region is more repetitive than in the other species.

Missing genomic regions have higher GC- and repeat-content

When I separated the genomic sequences into partitions, there was a clear dramatic higher proportion of missing sequences in CpG rich islands and repeat regions (**Figure 3.4**).

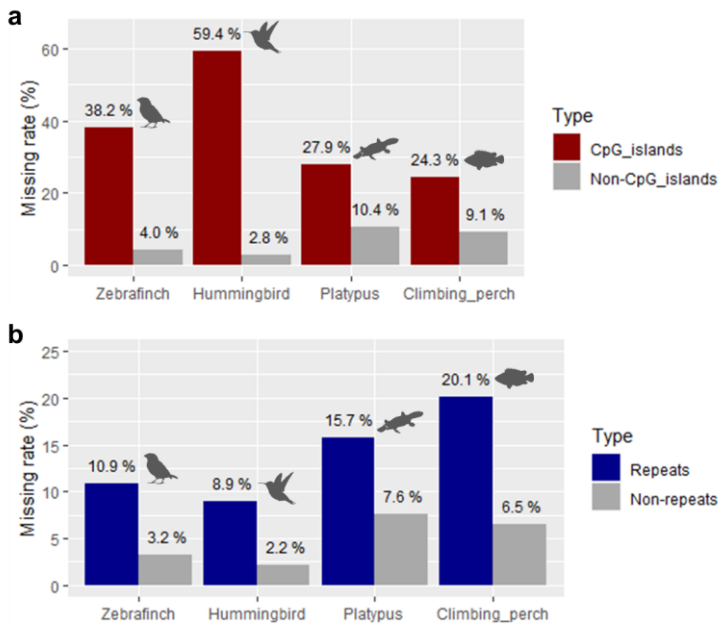

Figure. 3.4. Proportion, GC-content, and repeat-content of missing regions in prior assemblies found in VGP assemblies. (**a**) Missing rates in prior assemblies for CpG islands and non-CpG regions. (**b**) Missing rates in prior assemblies for repeats and non-repeated regions.

False gene losses in previous annotations of avian species

I next examined individual genes for various types of false gene losses, considering biological functions and contexts. The dopamine receptor D1B gene (*DR1DB* also called *DRD5*) is upregulated at higher levels in several vocal learning brain regions of songbirds, hummingbirds, and humans [73]. Previously I reported that the zebra finch *DR1DB* was mis-annotated due to missing GC-rich promoter sequences, resulting in false inference of exon and intron structure on the single exon gene [10] (**Figure 3.5a, top**). I identified a similar pattern of error in the prior Anna's hummingbird assembly (**Figure 3.5a, bottom**). Raw read mapping of the previous data showed that the promoter region in which a GC-rich CpG island exists was not sufficiently sequenced in the previous assembly, and this region contained regulatory sequences revealed by chromatin accessibility maps based on ATAC seq signals (**Figure 3.5a, top**). This missing sequence affected the annotation of the *DRD1B* gene in both bird species, leading to annotation of a false intron and exon in the upstream sequence. Here I clearly identified that the zebra finch and hummingbird *DRD1B* gene has a single exon, as reported in some other birds previously [73].

The second missing example is Calcium-dependent secretion activator 2 gene (*CADPS2*) which regulates the exocytosis of vesicles filled with neurotransmitters and neuropeptides in neurons [74] and shows specialized upregulated expression in several forebrain vocal learning nuclei of songbirds [75]. Thus, there has been interest in identifying the regulatory region responsible for this upregulation. I discovered a GC-rich 5' exon and upstream regulatory region, the latter with differential ATAC-Seq signals in the robust nucleus of the arcopallium (RA) song nucleus versus surrounding neurons, that were missing in the prior assembly of zebra finch (**Figure 3.5b**). This resulted in a false annotation of gene structure in the prior assembly, where the first non-GC-rich intron was misannotated as the regulatory region and two initial exons. In the Anna's hummingbird, I identified a similar error in the 5' upstream part of *CADPS2* gene. The first GC-rich exon was a CpG island that failed to be sequenced in the previous assembly (**Figure 3.5b**). Unlike Sanger and Illumina platforms in the previous assemblies, all missing GC-rich regions of the genes were newly detected in the VGP assemblies (**Figure 3.5a, b**).

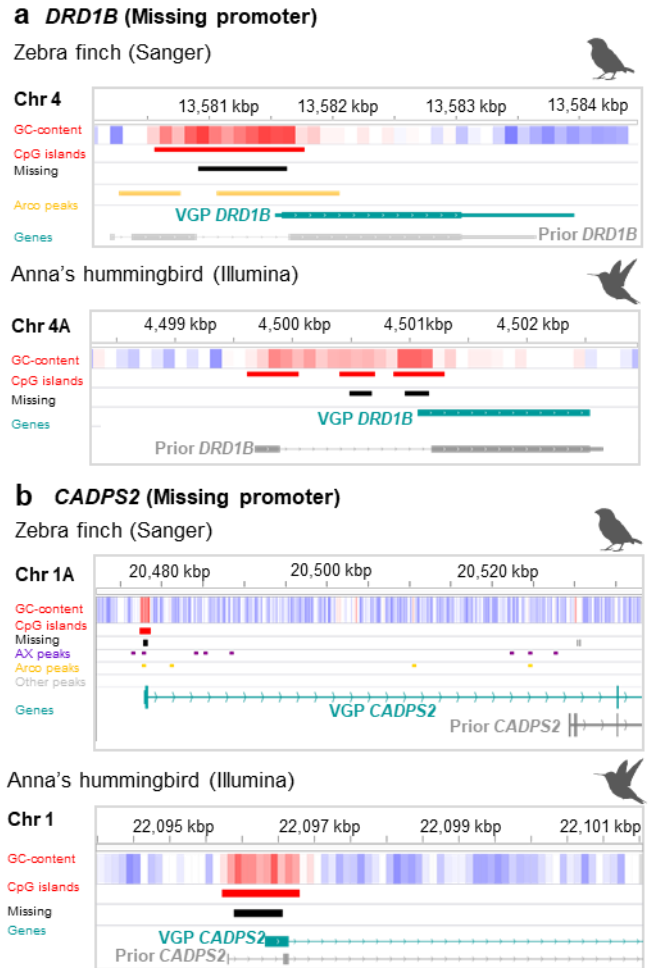

Figure 3.5. Effects of false gene losses in the previous assemblies on annotations in zebrafinch and Anna's hummingbird. *DRD1B* (**a**) and *CADPS2* (**b**) were missing 5' UTRs, CpG islands of promoter regions, and some coding sequence in the prior assemblies, resulting in the false understanding of the genes' structures and false annotations. In the zebra finch, the missing regions of both genes are inferred regulatory regions based on open chromatin ATAC peaks unique to Area X (AX) and arcopallium (Arco) compared to striatum brain regions, respectively.

False gene losses in previous annotations of a mammalian species

The *ADAM* metallopeptidase domain 7 (*ADAM7*) gene is highly conserved across mammals [76], is involved in spermatozoa secretions in the epididymis, including in platypus [77], has a metalloprotease domain regulated by several critical cysteine residues [78]. *ADAM7* in the prior platypus sanger assembly was fragmented into two scaffolds (NC_009098 and NW_001790718) and its prior annotation falsely missed six 5' exons, which included the critical catalytic cysteine residue (**Figure 3.6, Figure 3.7**). *ADAM7* in the VGP platypus assembly includes the critical cysteine residue (Cys50; **Figure 3.7b**), which is homologous with the human Cys170 and of other mammals (**Figure 3.7c, d**). This finding indicates that erroneous fragmentation in the prior assembly caused an annotation error for falsely missing exons with biologically important residues.

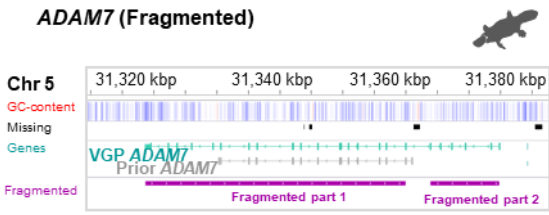

Figure 3.6. Effects of false gene loss in the previous assembly on annotation in platypus. *ADAM7* was fragmented on different two scaffolds and its N-terminal 6 exons were missed in the prior annotation.

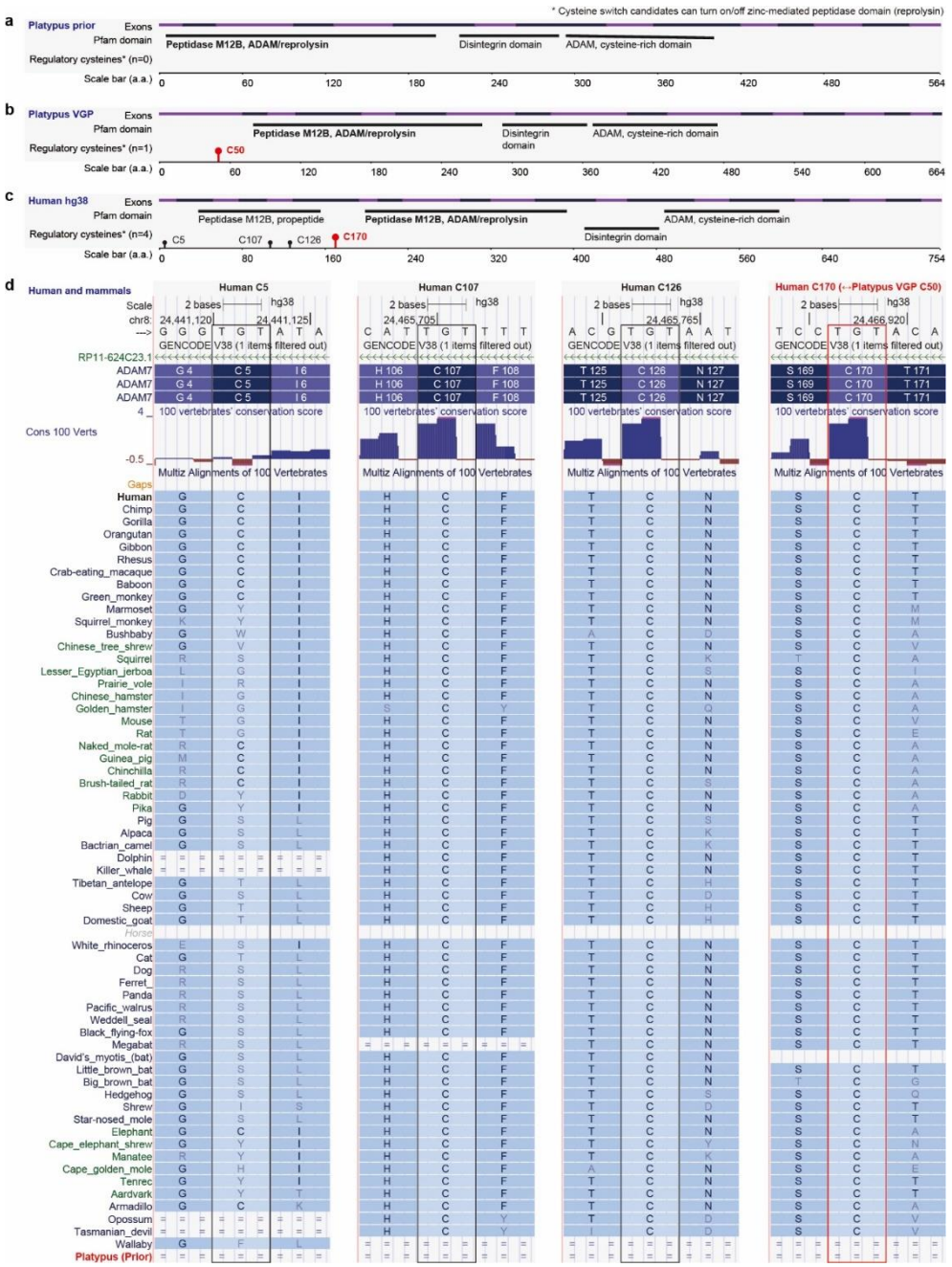

Figure 3.7. Functional domains and conserved cysteine switch of *ADAM7* missing in the prior platypus assembly. a, Protein coding region summary of *ADAM7* in the previous platypus assembly and annotation showing missing sequences in the 5' six exons. **b,** Protein summary of *ADAM7* in the VGP platypus assembly and annotation of correcting the missing errors. **c,** Protein summary of *ADAM7* in GRChg38 human assembly and annotation. The critical cysteine switch

in the VGP platypus (C50) is homologous to human C170 in the gene-wide peptide alignment by Clustral W (red bold). Data collection and visualizations is from ENSEMBL [79]. **d,** Conservation of critical cysteine regulators located in front of the zinc-medicated catalytic domain (reprolysin) in *ADAM7*. Data visualization from UCSC genome browser [71].

Falsely missing regions distinguished from individual variations

Because the zebra finch and hummingbird prior and VGP assemblies are from the same individuals, the missing regions in the prior assemblies compared to VGP assemblies can't be due to biological variation between individuals. However, for the platypus and climbing perch, since they are from different individuals, the missing regions in the prior assemblies could include biological variation between individuals. I think this unlikely explains most of the missing genomic regions, especially for the platypus, considering it would require one the prior individual having lost over 2 chromosomes' worth of genetic material (> 200 Mb), and selectively in GC-rich and repetitive regions, biased towards protein coding gene promoters. Further many of the missing regions in the prior assemblies are in assembly gaps, supporting missing sequence as opposed to biological variation. It is also unlikely that the platypus and climbing perch are different from the zebra finch and hummingbird in this regard. Nevertheless, for the platypus and climbing perch I sought additional measures to validate that most of the differences are not due to biological heterozygosity differences of massive gene losses.

First, I found the prior raw sequence data that went into the previous platypus and climbing perch assemblies from the NCBI trace archives, aligned them to the VGP assemblies, and checked the prior read depths in the VGP regions homologous to the missing regions in the prior assemblies. If the prior individual genome had true deletions, I would expect no reads from those regions mapping to the VGP assemblies. Additionally, if a missing region is within assembly gaps in the previous assemblies, such gaps indicate the potential existence of the sequence in the previous individual's genomes. Based on above analyses for prior reads and assembly gaps, I found 37.3% of the missing regions in the prior platypus individual and 65.9% in the prior climbing perch individual had prior reads that mapped to the VGP selected individuals (**Figure 3.8a**). However, the read depth was low on these prior missing regions of the assembly, which could explain why they were not assembled.

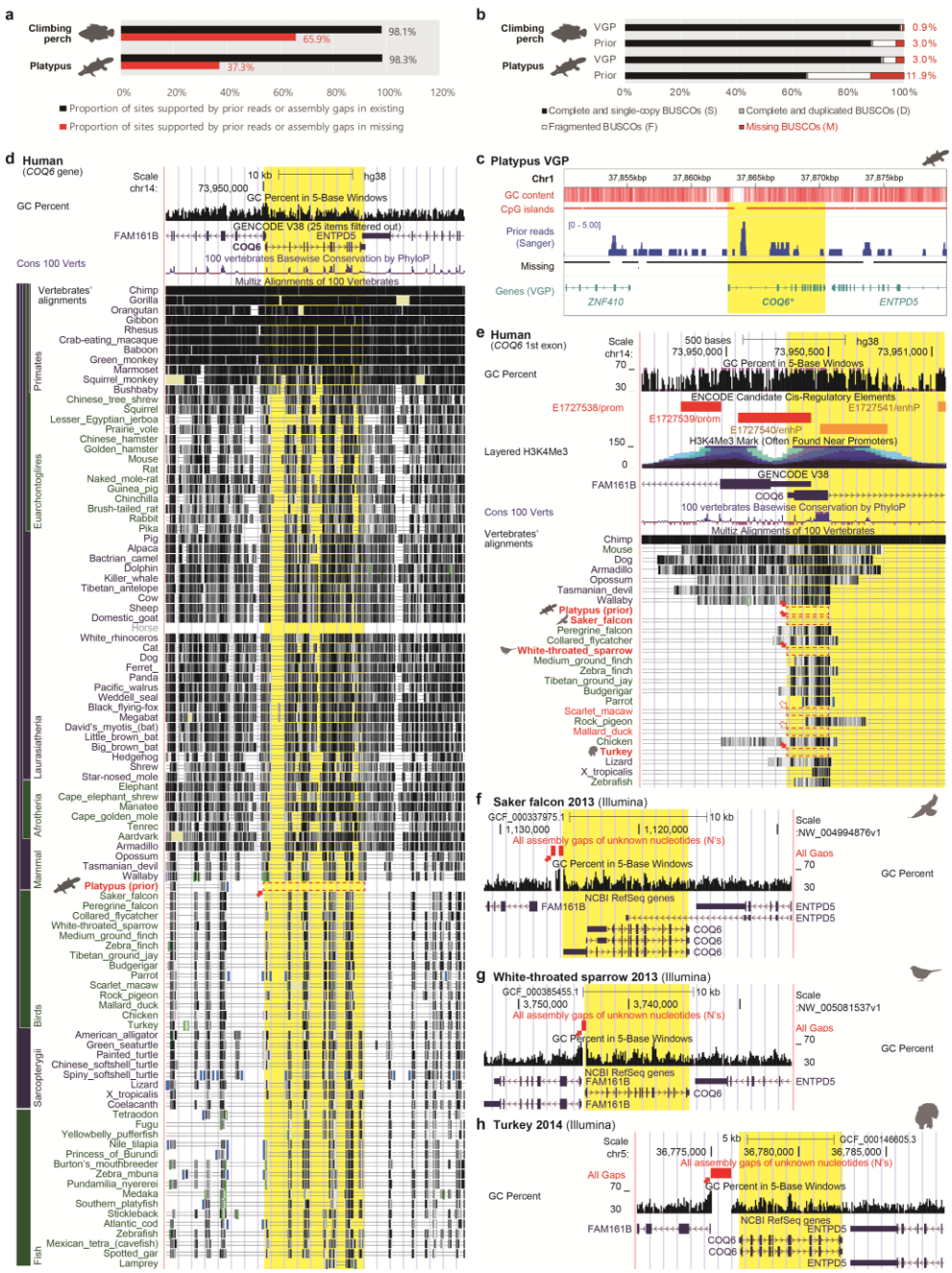

Figure 3.8. *COQ6* is an example gene that is falsely missing due to sequence and assembly errors in a highly divergent GC-rich ortholog. a, Proportions of sites supported by prior reads or assembly gaps in missing or existing regions in prior assemblies. Red and black colors indicate missing and existing regions, respectively. **b,** BUSCO comparisons between prior and VGP genome assemblies of platypus and climbing perch originating from different assemblies but also different platypus

individuals. Red color indicates the percentages of missing BUSCO genes in each genome. **c,** Genomic features and prior read depths on the *COQ6* gene and its neighbor genes. Prior reads were generated with the Sanger platform. Prior missing BUSCO gene, *COQ6*, marked as bold and asterisk with yellow highlight. **d,** *COQ6* was highly conserved in vertebrates except in the previous assembly of platypus. **e,** Missing first exon and promoter of *COQ6* in the prior assembly of platypus and several genome assemblies of birds. The GC-rich regions nearby the first exon were regarded as promoters, based on histone modification (H3K27Ac). Filled red arrows and red boxes indicate species with missing errors on the regions validated with data in the UCSC genome browser. Unfilled red arrows and red dashed boxes indicate species with candidates of missing and scaffolding errors. **f-h,** Missing errors supported by assembly gaps on the 5' GC-rich region of *COQ6* in Illumina-based genome assemblies of saker falcon, white-throated sparrow, and turkey, respectively. Filled red arrows and red boxes indicate gaps near 5' GC-rich regions.

Falsely missing genes conserved in vertebrates

Next, I focused on specific genes, particularly the universally conserved single-copy ortholog genes (BUSCO) found across all vertebrate species [54]. Being "universally conserved", missing BUSCO genes could be regarded as more likely to be the result of errors in assemblies rather than real biological variation. I discovered higher proportions of missing BUSCO genes in the prior platypus and climbing perch assemblies, supporting their lower qualities (**Figure 3.8b**). I examined more closely the case of a BUSCO gene that was completely missing in the prior platypus assembly, Coenzyme Q6, Monooxygenase (*COQ6*), and found that the entire gene was present in the VGP assembly but was GC-rich in the platypus with spotty Sanger raw read coverage in the prior assembly, indicating sequencing errors (**Figure 3.8b**). The spotty read coverage also indicates that the regions of 0 coverage are unlikely biological variations within the gene. In the 100 vertebrate UCSC genome alignment [71], the gene was more complete in 98 other species, with the exception of the horse, due to an apparent alignment error in UCSC Genome Browser (**Figure 3.8c**, **Figure 3.9**). Remarkably, I found the platypus has evolved a much higher species-specific GC-content in *COQ6* (**Figure 3.10, Figure 3.11**). I also discovered that most tetrapods, including human, have sequence conservation with high GC content in the 1st exon of *COQ6* and its promoter, supported by histone modification data (**Figure 3.8d**). However, Illumina-based genome assemblies of five birds (saker falcon, white-throated sparrow, scarlet macaw, mallard duck, and turkey) missed this first exon and the promoter. Three of these birds (saker falcon, white-throated sparrow, and turkey) showed assembly gaps indicating absence of sequencing reads overlapping the missing 5' region of *COQ6* (**Figure 3.8e-g**). Human also showed a conserved high GC content in the promoter and 1st exon (**Figure 3.10**). These findings suggest that falsely missing regions are associated with GC-rich regions with low read coverage and/or sequence errors, of various tetrapod vertebrate genome assemblies generated with Sanger or Illumina platforms, and that the platypus had evolved a much higher GC-content for this gene, reducing sequencing and assembly for the entire gene specifically in the platypus.

I previously reported on another vertebrate BUSCO gene, Yip1 Domain

Family Member 6 (*YIPF6*), as missing two exons and the 3' UTR in the prior climbing perch assembly [10]. Here, I precisely delineated the 5' missing region (2 exons), as it was due to the gene being split on two different scaffolds (OMLL01016988 and OMLL01012084) in the prior assembly (**Figure 3.12**). When mapping prior reads from the prior individual to the VGP assembly, there were two GC-rich regions of low coverage, one of which was not assembled, and another region of 0 coverage without any gap in the prior assembly, which could represent a real biological indel difference for this part of the gene between individuals.

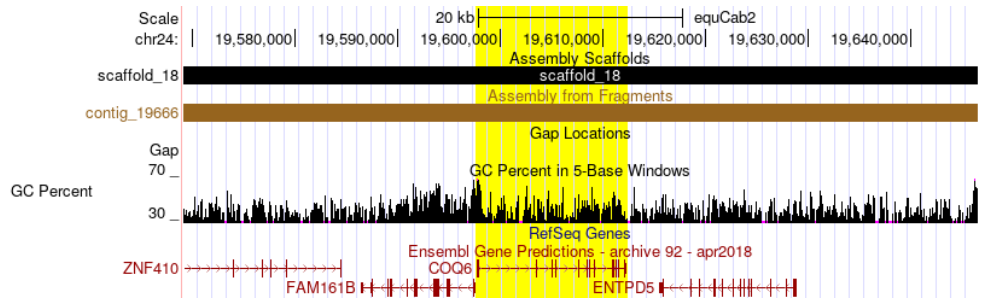

Figure 3.9. *COQ6* and its neighbor genes in the prior horse genome assembly (equCab2, 2007). Yellow highlight indicates the genic region of *COQ6*. Data visualization from UCSC genome browser [71].

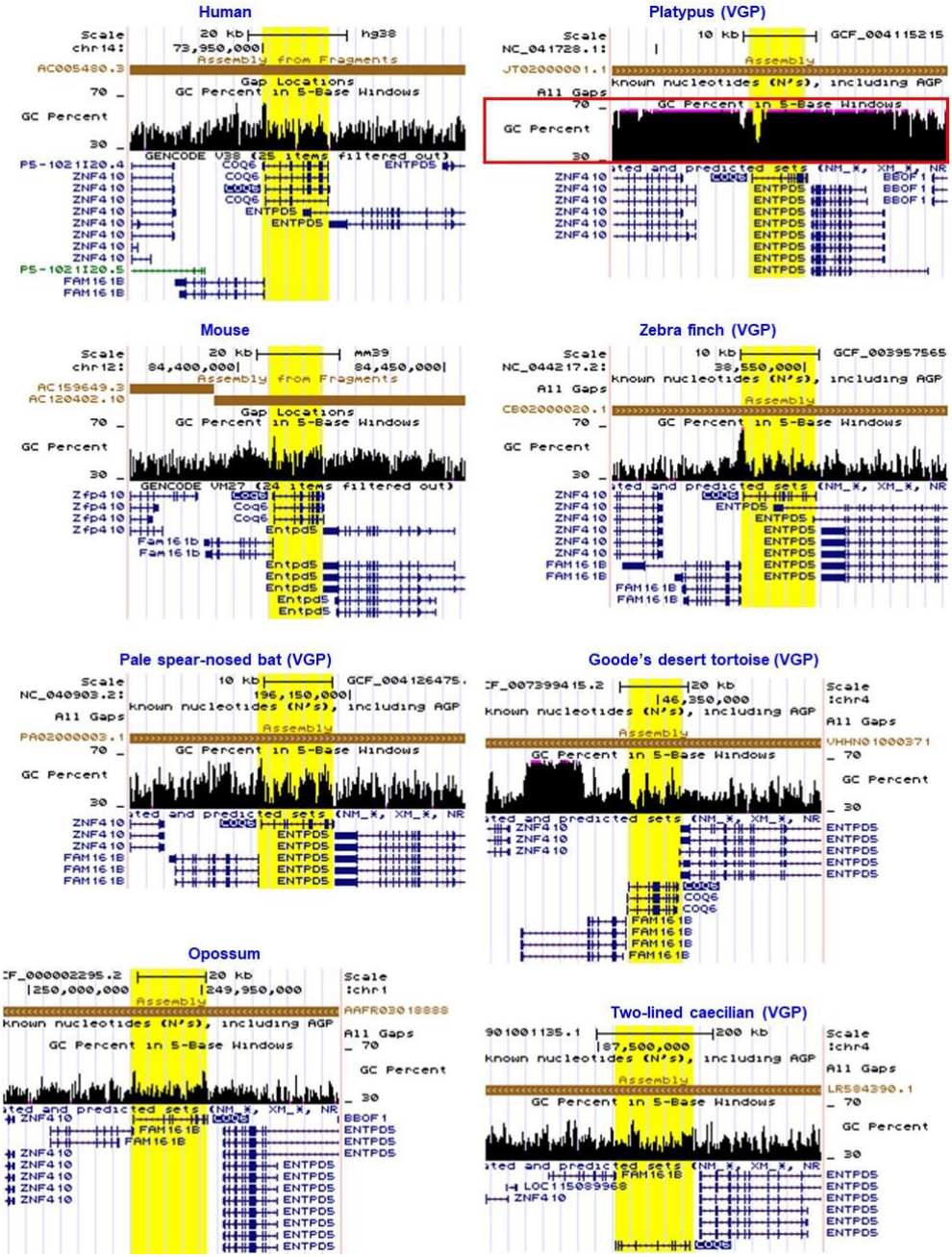

Figure 3.10. Species-specific high GC content in *COQ6* of platypus compared to 7 species representative of other tetrapod lineages. Yellow highlighted columns indicate genic regions of *COQ6* of each species. Red box highlights the region of high GC content broadly over 70% in the platypus. Displays generated in the UCSC browser [71].

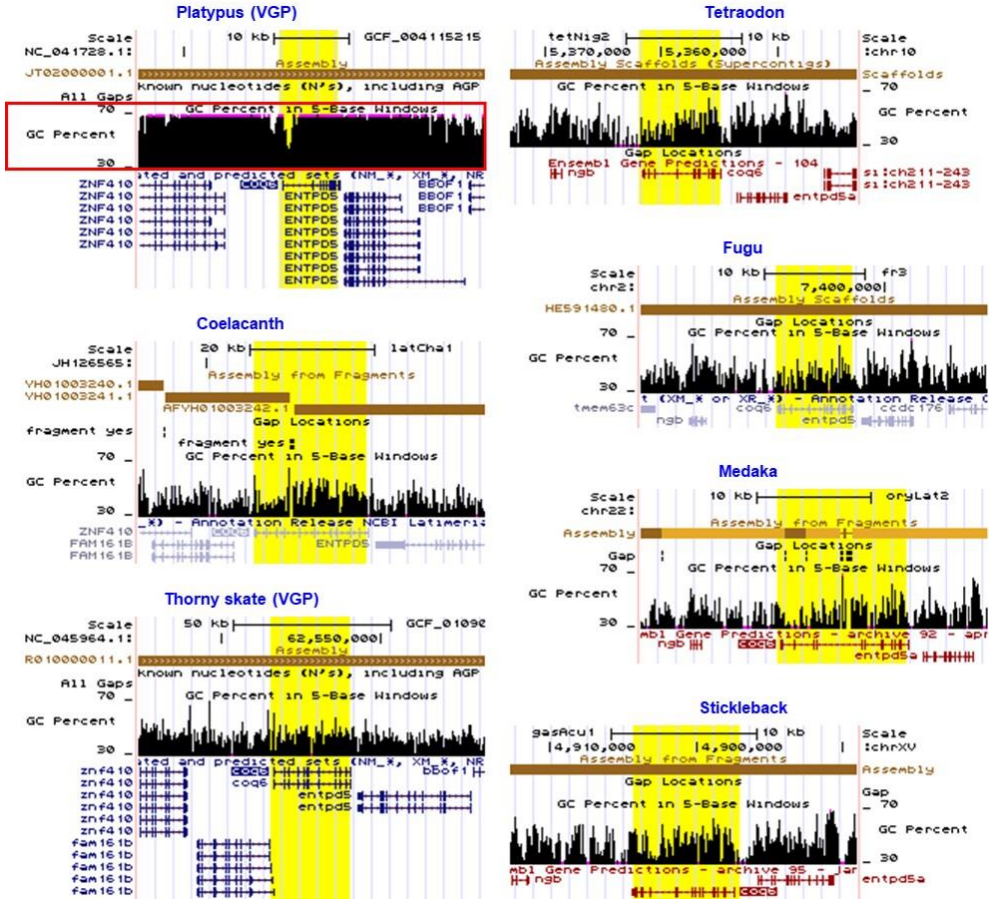

Figure 3.11. Species-specific high GC content in *COQ6* of the platypus compared to representatives of fish lineages. Yellow highlighted columns indicate genic regions of *COQ6* of each species. Red box highlights the region of high GC content broadly over 70% in the platypus. Displays generated in the UCSC browser [71].

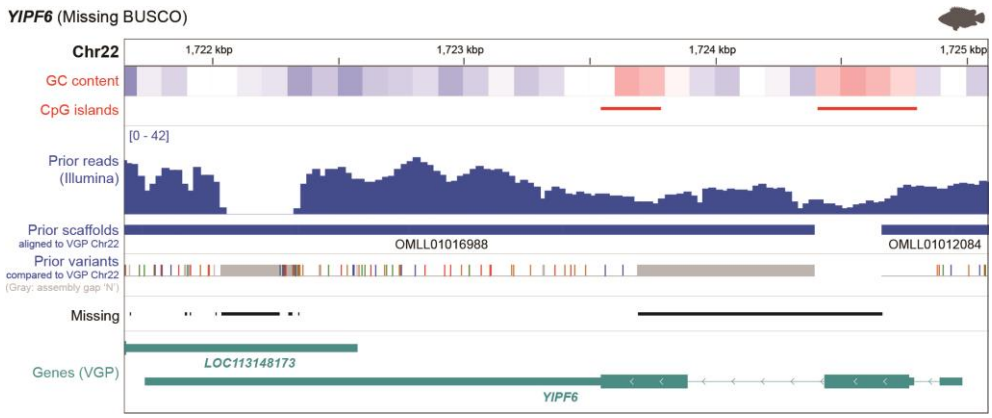

Figure 3.12. Example gene *YIPF6* with false missing sequences in the previous climbing perch assembly. The climbing perch prior genome assembly had erroneously missing regions caused by sequencing and assembly errors in a BUSCO gene, *YIPF6*. The row of prior variants shows the nucleotide substitutions from the aligned region in the VGP assembly: green, red, orange, blue, and gray colors indicating A, T, G, C, and N (assembly gap) in the prior assembly.

Chromosomal evolution of vertebrates

I used the more complete chromosome assemblies to determine if I could reveal new insights into chromosome evolution among vertebrates. Given that more than 430 million years (My) of divergence among the species sequenced makes it difficult to generate high coverage whole-genome alignments, I focused my initial analyses on 1,147 highly conserved BUSCO vertebrate genes shared among the assemblies of all 16 VGP species and the human reference (GRCh38). I found chromosome orthology between all species, but with different proportional relationships. Human chromosomes (1-22, and X) mapped to a lower average number of 3.7 (±1.3) chromosomes in other mammals, compared to 5.6 (±2.2) chromosomes in the amphibian, and to 9.6 (±3.3) chromosomes in teleost fishes (**Fig. 3.13, Table 3.2**). Despite belonging to the fish lineage and having a very high repeat content, the skate chromosome arrangement was more conserved with tetrapod vertebrates, mapping to 2.9 (±1.4) chromosomes on average compared to 4.8 (±2.5) in teleost fishes (**Table 3.3**). These findings indicate that, along with the GC-content reduction, the teleost lineage experienced more massive chromosome rearrangements since divergence from their most recent common ancestor with tetrapods, consistent with a proposed higher rearrangement rate in Teleostei[83].

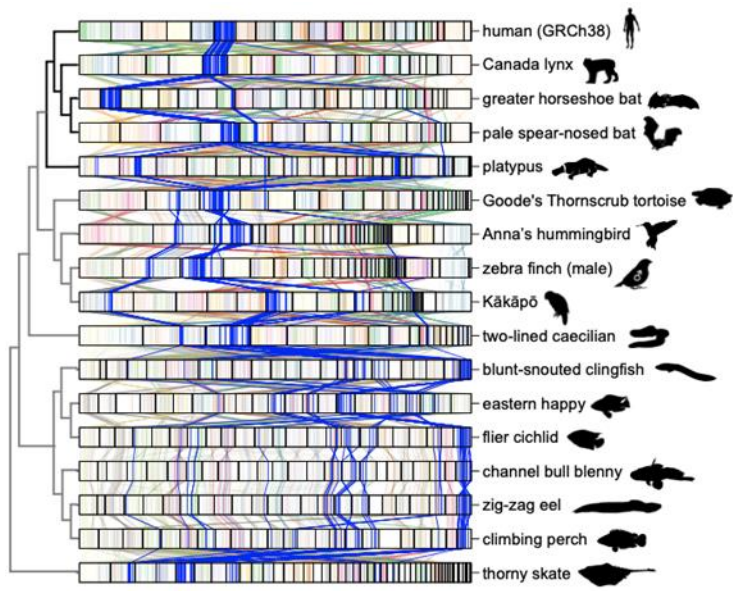

Figure 3.13. Chromosome synteny maps across the species sequenced based on BUSCO gene alignments. Chromosome sizes (bar lengths) are normalized to genome size, to make visualization easier. Genes (lines) are colored according to the locations in chromosomes of the human genome; the homologs of genes in human chromosome 6 are in dark blue, as an example, and the other chromosomes are lighter shades of different colors. The cladogram from the TimeTree database[92].

Table 3.2. Average number of chromosome segments in each lineage and clades mapped to human and thorny skate chromosomes. For each chromosome in the reference, number of chromosomes where identical BUSCO genes were found in the query genome assembly is shown.

		Num. of chromosomes mapped per clade							Mammals				Reptile	Birds			Amphibian	Teleost fishes						Skate
Reference	Chr.	Mammals	Reptile	Birds	Amphibian	Teleost fishes	Non-Teleost	Skate	mLynCan4	mRhiFer1	mPhyDis1	mOrnAna1	rGopEvg1	bCalAnn1	bTaeGut1	bStrHab1	aRhiBiv1	fGouWil2	fAstCal1	fArcCen1	fCotGob3	fMasArm1	fAnaTes1	sAmbRad1
Human(GRCh38)	1	6.0	9	9.7	10	16.0	8.3	11	5	4	6	9	9	11	9	9	10	16	15	17	17	16	15	11
	2	6.5	6	7.0	7	14.3	7.3	13	3	8	4	11	6	7	7	7	7	16	12	15	15	14	14	13
	3	5.5	8	7.0	8	14.5	6.8	9	3	4	3	12	8	7	7	7	8	15	14	15	15	15	13	9
	4	4.5	3	2.7	5	10.3	4.0	6	3	5	4	6	3	3	2	3	5	10	9	12	13	9	9	6
	5	4.5	4	4.0	5	11.7	4.5	6	2	4	3	9	4	4	4	4	5	10	11	13	13	12	11	6
	6	5.0	3	3.7	6	13.2	5.0	10	2	4	5	9	3	3	4	4	6	13	12	13	15	13	13	10
	7	3.5	4	4.0	6	12.0	4.2	6	2	4	3	5	4	4	4	4	6	11	12	12	13	12	12	6
	8	4.8	4	5.7	5	13.0	5.3	8	3	5	4	7	4	6	5	6	5	13	12	16	13	12	12	8
	9	3.3	4	5.0	6	11.3	4.5	7	2	1	2	8	4	5	5	5	6	12	9	15	11	10	11	7
	10	3.3	3	3.0	6	8.5	3.6	5	3	4	2	4	3	3	3	3	6	7	9	8	11	8	8	5
	11	4.5	3	5.3	6	9.7	4.9	6	3	6	4	5	3	6	5	5	6	10	9	11	9	10	9	6
	12	3.8	3	3.3	12	10.2	4.7	7	2	3	4	6	3	3	4	3	12	10	9	10	12	10	10	7
	13	1.8	1	1.0	2	8.3	1.5	2	1	1	2	3	1	1	1	1	2	9	8	8	9	8	8	2
	14	4.0	3	3.3	5	6.5	3.9	5	4	4	4	4	3	4	3	3	5	7	7	6	6	6	7	5
	15	2.8	3	4.0	5	7.3	3.8	7	2	2	2	5	3	5	4	3	5	7	6	9	7	8	7	7
	16	2.8	4	3.0	4	9.5	3.4	6	2	2	2	5	4	2	4	3	4	10	8	11	11	8	9	6
	17	2.5	4	5.0	5	6.5	4.1	7	1	2	1	6	4	6	5	4	5	6	6	6	8	6	7	7
	18	3.3	3	3.0	5	7.0	3.5	5	3	2	2	6	3	4	3	2	5	8	6	7	7	7	7	5
	19	2.5	2	2.0	2	2.2	2.2	2	2	2	3	3	2	2	2	2	2	2	3	2	2	2	2	2
	20	3.8	7	5.3	7	9.2	5.1	6	2	2	2	9	7	5	5	6	7	10	7	9	12	9	8	6
	21	2.3	2	2.0	4	6.2	2.3	2	2	2	2	3	2	2	2	2	4	6	6	7	6	6	6	2
	22	2.0	2	2.0	4	5.8	2.4	4	2	2	2	2	2	2	2	2	4	5	5	6	7	6	6	4
	X	2.3	3	3.0	4	7.8	2.8	3	2	2	2	3	3	3	3	3	4	7	7	7	9	9	8	3
	Avg.	**3.7**	**3.8**	**4.1**	**5.6**	**9.6**	**4.3**	6.2	2.4	3.3	3.0	6.1	3.8	4.3	4.0	4.0	5.6	9.6	8.8	10.2	10.5	9.4	9.2	6.2
	S.D.	**1.3**	**1.9**	**2.0**	**2.2**	**3.3**	**1.6**	2.8	0.9	1.7	1.2	2.7	1.9	2.2	1.9	2.0	2.2	3.6	3.0	3.9	3.7	3.3	3.1	2.8

		Num. of chromosomes mapped per clade						Mammals					Reptile	Birds			Amphibian	Teleost fishes						Skate
Reference	Chr.	Mammals	Reptile	Birds	Amphibian	Teleost fishes	Non-Teleost	Human	mLynCan4	mRhiFer1	mPhyDis1	mOrnAna1	rGopEvg1	bCalAnn1	bTaeGut1	bStrHab1	aRhiBiv1	fGouWil2	fAstCal1	fArcCen1	fCotGob3	fMasArm1	fAnaTes1	sAmbRad1
	1	8.6	4	4.0	3	11.3	6	8	8	10	7	10	4	4	5	3	3	10	10	13	15	10	10	na
	2	8.8	3	3.0	4	5.7	6	9	7	13	9	6	3	3	3	3	4	7	5	5	6	5	6	na
	3	4.8	3	2.3	3	8.7	3.7	4	3	5	4	8	3	2	3	2	3	8	7	13	8	8	8	na
	4	4.6	1	1.0	3	6.7	3	4	5	5	3	6	1	1	1	1	3	8	6	7	7	6	6	na
	5	6.8	4	3.3	3	7.3	5.1	6	8	7	6	7	4	3	3	4	3	8	6	8	8	7	7	na
	6	5.2	1	1.3	2	8.2	3.3	5	4	5	6	6	1	1	2	1	2	8	8	8	9	9	7	na
	7	2.8	2	2.0	1	5.8	2.3	2	2	4	2	4	2	2	2	2	1	5	5	7	8	5	5	na
	8	5.6	3	2.0	3	8.5	4	6	5	6	6	5	3	2	2	2	3	10	7	7	11	8	8	na
	9	2.4	2	2.3	3	4.5	2.4	3	2	2	1	4	2	2	3	2	3	5	5	4	3	5	5	na
	10	4.6	3	3.7	4	6.0	4.1	3	5	5	5	5	3	4	4	3	4	6	6	8	6	6	4	na
	11	1.6	1	1.0	2	3.2	1.4	1	1	3	1	2	1	1	1	1	2	3	3	2	5	3	3	na
	12	2.6	2	2.0	3	4.0	2.4	3	3	3	2	2	2	2	2	2	3	4	4	4	4	4	4	na
	13	4.6	2	3.3	5	4.8	4	4	4	5	5	5	2	4	3	3	5	5	6	5	5	4	4	na
	14	6.8	3	3.0	5	10.8	5.1	6	6	6	7	9	3	3	3	3	5	10	10	12	13	11	9	na
	15	1.4	1	1.0	4	4.3	1.5	2	1	1	1	2	1	1	1	1	4	4	5	4	5	4	4	na
	16	2.2	2	2.7	1	4.2	2.2	2	2	2	2	3	2	2	4	2	1	4	4	4	5	4	4	na
	17	1.8	2	1.7	4	3.3	2	3	1	1	1	3	2	1	1	3	4	3	2	5	5	2	3	na
	18	4.6	3	3.0	4	3.2	3.9	4	4	6	5	4	3	3	3	3	4	3	3	3	3	4	3	na
	19	5.6	3	3.0	7	5.5	4.7	6	4	6	5	7	3	3	3	3	7	4	5	8	6	5	5	na
	20	2.8	3	3.0	2	4.3	2.8	4	2	4	2	2	3	2	4	3	2	4	4	6	4	4	4	na
	21	4.8	2	2.0	5	3.5	3.7	5	3	5	4	7	2	2	2	2	5	2	4	4	5	3	3	na
Skate(sAmbRad1)	22	4.6	3	3.7	3	6.5	4	5	3	5	5	5	3	5	3	3	3	7	7	7	6	6	6	na
	23	5.6	5	5.3	6	7.2	5.5	5	5	5	5	8	5	6	5	5	6	7	6	8	8	7	7	na
	24	3.0	2	2.0	2	3.5	2.5	3	4	3	3	2	2	2	2	2	2	4	3	3	5	3	3	na
	25	2.6	2	2.3	3	5.3	2.5	3	2	2	3	3	2	3	2	2	3	7	5	5	5	5	5	na
	26	1.4	1	1.0	1	3.0	1.2	1	1	1	1	3	1	1	1	1	1	3	3	3	3	3	3	na
	27	3.0	3	3.0	2	4.5	2.9	3	3	3	3	3	3	3	3	3	2	4	4	4	4	5	6	na
	28	4.4	4	4.3	5	8.3	4.4	4	5	5	3	5	4	4	4	5	5	8	8	8	9	8	9	na
	29	1.4	1	1.0	1	1.2	1.2	1	1	1	2	2	1	1	1	1	1	1	2	1	1	1	1	na
	30	2.0	2	2.0	2	2.0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	na
	31	2.4	2	2.3	3	2.5	2.4	3	2	2	3	2	2	2	3	2	3	2	3	2	4	2	2	na
	32	1.6	1	1.0	1	5.0	1.3	1	2	1	1	3	1	1	1	1	1	5	4	8	4	4	5	na
	33	1.6	1	1.0	1	5.2	1.3	1	1	3	2	1	1	1	1	1	1	6	5	5	5	5	5	na
	34	1.6	1	1.0	2	2.7	1.4	1	1	2	2	2	1	1	1	1	2	3	2	5	2	2	2	na
	37	1.6	1	1.0	2	3.2	1.4	1	2	2	1	2	1	1	1	1	2	3	3	3	4	3	3	na
	38	1.4	1	1.0	1	2.2	1.2	2	1	1	1	2	1	1	1	1	1	2	2	2	3	2	2	na
	39	2.6	2	2.3	2	4.3	2.4	2	2	3	3	3	2	2	3	2	2	4	5	5	4	4	4	na
	41	2.0	2	2.0	2	1.8	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	na
	43	2.0	2	2.0	2	2.0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	na
	44	1.6	1	1.0	1	2.3	1.3	2	1	2	2	1	1	1	1	1	1	2	2	2	3	3	2	na
	46	2.0	2	2.0	2	2.0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	na
	Avg.	**3.4**	**2.2**	**2.2**	**2.8**	**4.8**	**2.9**	3.3	3.0	3.7	3.2	4.0	2.2	2.2	2.3	2.2	2.8	4.8	4.6	5.3	5.3	4.6	4.5	
	S.D.	**2.0**	**1.0**	**1.1**	**1.5**	**2.5**	**1.4**	2.0	1.9	2.5	2.0	2.3	1.0	1.2	1.2	1.0	1.5	2.6	2.2	3.0	3.0	2.4	2.3	

This chapter was published in *Marine genomics*
as a partial fulfillment of Chul Lee's Ph.D. program

Chapter 4. Coelacanth-specific adaptive genes give insights into primitive evolution for water-to-land transition of tetrapods

4.1. Abstract

Coelacanth is a group of extant lobe-finned fishes in *Sarcopterygii* that provides evolutionary information for the missing link between ray-finned fish and tetrapod vertebrates. Its phenotypes, different from actinopterygian fishes, have been considered as primitive terrestrial traits such as cartilages in their fatty fins which are homologous with the humerus and femur. To investigate molecular evolution of coelacanth which led to its divergence into *Sarcopterygii*, I compared its protein coding sequences with 11 actinopterygian fishes. I identified 47 genes under positive selection specific to coelacanth, when compared to *Holostei* and *Teleostei*. Out of these, *NCDN* and 14 genes were associated with spatial learning and nitrogen metabolism, respectively. In homeobox gene superfamily, I identified coelacanth-specific amino acid substitutions, and also observed that one of replacements in *SHOX* was shared with extant tetrapods. Such molecular changes may cause primordial morphological change in the common ancestor of sarcopterygians. These results suggest that certain genes such as *NCDN, MMS19, TRMT1, ALX1, DLX5* and *SHOX* might have played a role in the evolutionary transition between aquatic and terrestrial vertebrates

4.2. Introduction

Coelacanth, the name derived from its characteristic hollow caudal fin, was first described in 1839 from the fossil records (Agassiz, 1844). Abundance of the fossils from the Early Devonian to the Late Cretaceous sediments implied that the fish flourished during the period. However, drastic disappearance of the post-Devonian coelacanth fossils implies that its population rapidly declined with Cretaceous–Paleogene (K–Pg) mass extinction. Therefore, scientific community was shocked at unexpected report of living coelacanth *Latimeria chalumnae* in east coast of South Africa in 1938, and *Latimeria menadoensis* in Indonesia (Erdmann et al., 1998, Smith, n.d). Coelacanth initially gained the title 'living fossil' after this first observation due to its morphological similarity to its ancient form in fossil record, and the fact that it is sole survivor in Actinistia, a group mostly consisted of fossil lobe-finned fishes in Sarcopterygii. The term was considered appropriate for decades, but controversy over appropriateness of the term recently have been aroused. The morphological similarity between extant coelacanth and the fossil record had been one of the reason why coelacanth was called 'living fossil,' but as the diverse shape of coelacanth was reported (Friedman and Coates, 2006, Wendruff and Wilson, 2012), coelacanth's morphological conservation has become questionable. In addition, with coelacanth being observed in the diverse shape among the actinistians, it was suggested that coelacanth-specific evolution has been accumulated after the divergence from the most recent common ancestor (MRCA) of *Sarcopterygii* (Bockmann et al., 2013).

For all the dispute, coelacanth gives essential information to trace back the origin of tetrapod limbs, which is one of the key events influenced landing of vertebrates. Coelacanth forms a clade with lungfish and tetrapods which are classified into the sarcopterygians, sharing conserved skeletons in fleshy fins or derivative, vertebrate limbs. Coelacanth possesses a muscular lobed-fins composed with cartilages, including one homologous to humerus and femur which articulates fins to pectoral or pelvic girdle, which is an intermediate form of actinopterygians

and tetrapods (Francillon et al., 1973). The phenotype related to water-to-land transition originates from the genetic factors shared among the sarcopterygian clade, which makes it important to analyze its genomic sequence. As lungfish turned out to be closer relative of tetrapods than coelacanth, it became more meaningful to analyze coelacanth genome to investigate the first emergence of landing-related traits different from Actinopterygii.

Comparative genomics serves as a valuable tool to find out genomic features related to common or specific traits between different species. In coelacanth, comparing common sequence shared with other vertebrates revealed genetic factors that may have adaptively evolved while the landing-related traits emerged in their MRCA. For example, island I region of the *HoxD* gene cluster is conserved within Sarcopterygii but not in Actinopterygii, which has indispensable role in developing autopod of mouse (Fromental-Ramain et al., 1996). Not only the island I region, but also several conserved noncoding elements (CNEs) which are located in regulatory regions of key genes for limb development such as *bmp7*, *grem1*, *shh*, and *gli3* were reported (Nikaido et al., 2013, Zuniga et al., 2012). Especially, based on the first construction of coelacanth reference genome, the adaptation of vertebrates to land environment were determined by comparing it with other bony vertebrate genomes, such as, conserved limb enhancers in *HoxD* locus, amino acid differences in homeobox genes related to organism's basic body plan between coelacanth, ray-finned fishes, and tetrapods (Amemiya et al., 2013). In addition, one of the genes related to nitrogen waste metabolism which may be necessary in non-aquatic habitats, Carbamoyl phosphate synthase I (*CPS1*), was subjected to positive selection on branches leading to tetrapods and to amniotes, respectively (Amemiya et al., 2013).

By sorting out the type of point mutation whether synonymous or nonsynonymous, ratio between the frequency of each mutation can be calculated (dN/dS) to deduce type of selection that a gene went through (Yang and Bielawski, 2000). Synonymous substitution does not affect the phenotype, so it is free from the selective pressure and occurs at constant rate. On contrary, frequency of nonsynonymous mutation (dN) rises when the diversifying evolution takes place for

example, by exposure to the new environment. In dolphin, positively selected genes (PSGs) enriched based on branch-site model, provided better understanding for its aquatic adaptation, like echolocation and fat storage (McGowen et al., 2012). However, dN/dS analysis in coelacanth has been applied only to small sets of genes, such as a gene cluster or coelacanth specific retrocopies (Du and He, 2015, Zapilko and Korsching, 2016). In this study, I describe the result of genome-wide search of PSGs in coelacanth associated with this species specific adaptation to the aquatic habitat nearby the ocean floor or primordial changes of the most common ancestor of *Sarcopterygii* to affect landing of tetrapods. Hierarchical clustering of the discovered genes according to their biological function elucidated the group function of PSGs specific to coelacanth. In particular, I observed the genes significantly clustered into nitrogen-metabolism process which involves conversion of ammonia into urea. Moreover, through analyzing specific amino acid substitution within genes crucial to the limb development that is shared by coelacanth and tetrapods but absent in ray-finned fish lineage, this study implies the importance of these genetic features for vertebrate landing.

4.3. Materials and methods

Reference genome sequences and tree topology

In order to investigate coelacanth-specific PSGs that may be advantageous for water-to-land transitions of vertebrates, I collected reference genome sequences of *Osteichthyes*, including coelacanth (*Latimeria chalumnae*), 1 holostean fish (*Lepisosteus oculatus*), 10 teleostean fishes (*Astyanax mexicanus*, *Danio rerio*, *Gadus morhua*, *Gasterosteus aculeatus*, *Oreochromis niloticus*, *Oryzias latipes*, *Poecilia formosa*, *Takifugu rubripes*, *Tetraodon nigroviridis*, and *Xiphophorus maculatus*), and 4 tetrapod vertebrates (*Anolis carolinensis*, *Homo sapiens*, *Pelodiscus sinensis*, and *Xenopus tropicalis*), from BioMart in ENSEMBL database release 86 (Yates et al., 2016) (Table 1). For building a reliable cladogram to scan for genes under positive selection on a specific branch, I searched a golden standard ENSEMBL tree built by using Dendroscope 3 program in ENSEMBL Compara (Vilella et al., 2009) (Fig. 1).

Table 4.1. Versions of reference sequences of species.

Common name	***Scholar name***	***Class***	***Infraclass***	***Order***	***Reference version***
Amazon molly	*Poecilia formosa*	*Actinopteri*	*Teleosteiei*	*Cyprinodontiformes*	Poecilia_formosa-5.1.2
Cave fish	*Astyanax mexicanus*	*Actinopteri*	*Teleosteiei*	*Characiformes*	AstMex102
Cod	*Gadus morhua*	*Actinopteri*	*Teleosteiei*	*Gadiformes*	gadMor1
Fugu	*Takifugu rubripes*	*Actinopteri*	*Teleosteiei*	*Tetraodontiformes*	FUGU 4.0
Medaka	*Oryzias latipes*	*Actinopteri*	*Teleosteiei*	*Beloniformes*	HdrR
Platyfish	*Xiphophorus maculatus*	*Actinopteri*	*Teleosteiei*	*Cyprinodontiformes*	Xipmac4.4.2
Stickleback	*Gasterosteus aculeatus*	*Actinopteri*	*Teleosteiei*	*Gasterosteiformes*	BROAD S1
Tetraodon	*Tetraodon nigroviridis*	*Actinopteri*	*Teleosteiei*	*Tetraodontiformes*	TETRAODON 8.0
Tilapia	*Oreochromis niloticus*	*Actinopteri*	*Teleosteiei*	*Perciformes*	Orenil1.0
Zebrafish	*Danio rerio*	*Actinopteri*	*Teleosteiei*	*Cypriniformes*	GRCz10
Spotted gar	*Lepisosteus oculatus*	*Actinopteri*	*Holostei*	*Lepisosteiformes*	LepOcu1
Coelacanth	*Latimeria chalumnae*	*Sarcopterygii*		*Coelacanthiformes*	LatCha1
Anole Lizard	*Anolis carolinensis*	*Reptilia*		*Squamata*	AnoCar2.0
Chinese softshell turtle	*Pelodiscus sinensis*	*Reptilia*		*Testudines*	PelSin_1.0
Human	*Homo sapiens*	*Mammalia*		*Primates*	GRCh38.p7
Xenopus	*Xenopus tropicalis*	*Amphibia*		*Anura*	JGI 4.2

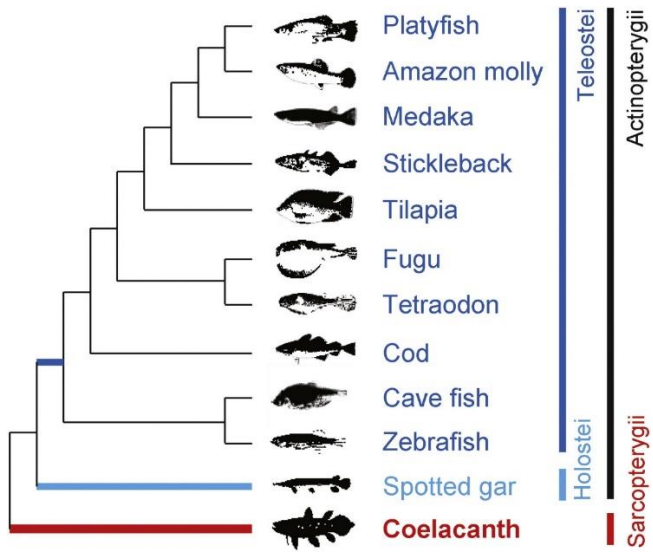

Figure 4.1. Cladogram of *Osteichthyes* family. Bold lines in the tree indicate the most recent ancestral branches of each lineage. Blue, skyblue, and red indicates *Teleostei*, *Holostei*, and coelacanth lineages, respectively.

Orthologous gene set alignments

Multiple sequence alignments of suitable coding gene sets were prepared for detection of positive selection with the following steps. Firstly, to exclude possibility of functional changes caused by gene expansion (gain and loss of genes), I focused on genes that show one to one orthologues in 12 fishes. Using coelacanth genome as a representative dataset, I found 4160 coding gene sets in ENSEMBL Biomart (Kinsella et al., 2011). Secondly, I filtered out 28 genes with sequence lengths which are not multiple of 3. After filtering these genes, I aligned 4132 gene sets by using PRANK (Löytynoja and Goldman, 2008) with two options; '-codon' for codon-wise alignments and '-F' for the most accurate alignments to identify homologous sites in each species. Finally, to exclude regions with poorly scored alignment caused by indels and mismatch, I trimmed 4132 alignments by using GBlocks (Talavera and Castresana, 2007) with one option '-t = c' for codon-wise adjustments. Finally, I prepared conserved coding sequence alignments of 3538 genes.

PSGs specific to coelacanth

To identify genes responsible for the evolution of coelacanth, I screened for the molecular signatures under episodic adaptive evolution. This was done by calculating dN (number of non-synonymous substitutions per number of non-synonymous sites of each gene), dS (number of synonymous substitutions per number of synonymous sites of each gene), and dN/dS (ratio of number of non-synonymous substitutions per number of non-synonymous sites to number of synonymous substitutions per number of synonymous sites of each gene) values of 3538 orthologous genes from 12 fishes excluding 4 tetrapods as an outgroup. In order to detect accurate selection signatures and to estimate site-wise selection on the latest ancestral branch of each lineage of coelacanth, spotted gar, and Teleostei fishes in the species tree (Fig. 1), 'branch-site model' based on 'CodeML' in PAML program (version 4.8) (Yang, 2007) was performed with 3 options; 'model = 2' for 2 or more dN/dS ratios for branches, 'NSsites = 2' to detect sites under positive selection on a foreground branch, and 'CodonFreq = 2' to calculate codon frequencies based on 'F3X4'. Based on estimated parameters from the test, I compared maximum likelihoods of null and alternative models by using likelihood ratio test (LRT,

$D = 2 * \Delta l$). The statistical significances were calculated by using chi-square test and false discovery rate (FDR) was used for multiple test correction using R program (version 3.2.3.) (Team, 2013). Consequently, I identified sites under positive selection on each lineage with posterior probability. PSGs were detected with strict filtering criteria (dN/dS value of class 2 of foreground branch > 1, $D > 0$, and adjusted $p < 0.05$). After identification of significant PSGs, I checked posterior probability of each gene (> 0.95) to find specific sites under positive selection (site class 2) based on the Bayes empirical Bayes (BEB) inference. Finally, PSGs specific to coelacanth were identified through comparing PSGs of coelacanth, *Holostei*, and *Teleostei*.

Conserved domain search

To determine whether sites under positive selection are located in functional domains of each gene, I performed domain analysis by using Batch web C-Search tool in NCBI (Marchler-Bauer et al., 2011). Peptide sequences of PSGs unique to coelacanth were used as a query set, and following options were applied: Data source: CDSEARCH/cdd v3.15; Expected value: 0.01; Composition-corrected scoring: Applied; Low-complexity regions: Not filtered.

Gene ontology analysis

To check the group functions of PSGs specific to coelacanth, I applied gene ontology analysis with gene set enrichment tests by using DAVID functional annotation (Huang et al., 2009). To compare with other fishes, zebrafish was used as a representative background model. The cutoff of statistical significance of enrichment test was applied as the default p-value < 0.1, due to the small number of coelacanth-specific PSGs. I summarized gene ontology of biological process based on hierarchical clustering with 'hclust' function in R (version 3.2.3.) (Team, 2013).

Protein-protein interaction network analysis

To investigate interactions among genes, Search Tool for the Retrieval of Interacting Genes (STRING) online database (http://string-db.org/) was used (Szklarczyk et al., 2014). STRING provides direct (physical) and indirect (functional) associations among genes based on multiple resources (Szklarczyk et al., 2014). I searched interactions between 5 genes of urea cycle and 14 coelacanth-specific PSGs of

nitrogen compound metabolic process to generate a network with the following options: Organism: *Danio rerio*; Active interaction sources: Text-mining, Experiments, Databases, Co-expression, Neighborhood, Gene fusion, and Co-occurrence; minimum required interaction score: medium confidence (0.4).The network was visualized using Cytoscape 3.3.0 (Shannon et al., 2003).

Amino acid changes specific to coelacanth

Target-specific amino acid substitutions (TAAS) analysis (Zhang et al., 2014) was conducted to find mutually exclusive amino acid substitutions between coelacanth and other fishes. The TAAS module and a codon translator were written and executed by Python (version 2.7.9., htttp://www.python.org). For one of homeobox genes, *SHOX*, I conducted additional TAAS analysis with 100 way multiz-alignment of 100 vertebrates (Blanchette et al., 2004) in UCSC genome browser (Meyer et al., 2013).

4.4. Results

Coelacanth is an important species to use for investigating the adaptation of tetrapod to land environment. To identify genetic features which led to the evolution of coelacanth, I investigated PSGs by scanning the whole coding regions of coelacanth genome. Based on the Ensembl database (release 86) (Yates et al., 2016), I collected coding sequences of coelacanth and control sets, from 1 holostean fish, 10 teleostean fishes, and 4 tetrapod vertebrates (Table 1). The four-limbed animals were used as an outgroup in comparative genomic approaches to understand primitive evolution shared among finned and limbed sarcopterygians. The cladogram construction was based on the species tree in Ensembl Compara. The cladogram construction was based on the species tree in Ensembl Compara (Vilella et al., 2009) (Fig. 1). Focusing on the genomic data set of coelacanth, I searched one to one orthologues conserved in all of 12 fishes to exclude duplicated or lost gene in Ensembl Biomart (Kinsella et al., 2011). To match homologous codons of each gene, I aligned coding sequences by using PRANK program (Löytynoja and Goldman, 2008). I filtered out indel and divergent regions with poor alignment scores by using Gblocks program (Talavera and Castresana, 2007) to prevent artifacts of dN/dS analysis due to missing data or alignment error. After alignment and trimming of the multiple sequences, I obtained conserved alignments of 992,062 codons in 3538 genes.

Positive selection on functional domains of coelacanth

In order to detect positive selection specifically experienced by coelacanth excluded from ray-finned fishes, I performed dN/dS analysis using the branch-site model A in codeML of PAML package (Yang, 2007) that can estimate the varying dN/dS (ω) values among different sites and lineages. I scanned whole one to one orthologous gene sets by focusing on the most recent ancestral branches of coelacanth and other two fish lineages, independently (Fig. 1, see Materials and Methods). Out of 3538 genes, 2.3% (81 genes with 800 sites), 4.2% (150 genes with 829 sites) and 10.2% (362 genes with 1039 sites) were under positive selection on coelacanth, Holostei, and Teleostei, respectively (adjusted p-value < 0.05, $\omega_2 > 1$, BEB > 0.95). Out of

these genes, I identified 47 PSGs unique to coelacanth compared to holostean and teleostean lineages (Fig. 2A). To determine if these sites were located in functional regions in each protein, I conducted NCBI conserved domain search (Marchler-Bauer et al., 2011). All of 47 PSGs specific to coelacanth consisted of 122 functional domains. However, only 34 PSGs contained 52 domains with 159 sites under positive selection. Out of these 34 PSGs, neurochondrin (*NCDN*)showed the highest number of positively selected sites of 23 harboring in functional domains (Fig. 2B).

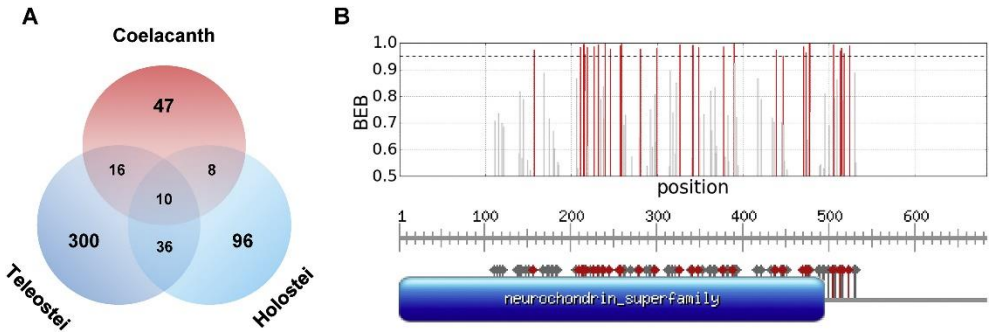

Figure 4.2. Positively selected genes on *Teleostei*, *Holostei*, and coelacanth. (A) Venn diagram of the number of genes under positive selection on each lineage (dN/dS > 1, FDR < 0.05, Posterior probability > 0.95). Red, blue and skyblue indicate the number of PSGs on coelacanth, Teleostei, and Holostei lineage. (B) Distribution of posterior probabilities of dN/dS analysis on NCDN gene. X-axis: positions in the peptide sequence of coelacanth, Y-axis: score calculated by bayes empirical bayes (BEB); Black dash line: threshold of statistical significance (BEB = 0.95); Red bar: BEB > 0.95; Grey bar: 0.5 < BEB ≤ 0.95; Bottom of the graph indicate the conserved domain (blue box) and sites under positive selection (red pin: BEB > 0.95, grey pin: 0.5 < BEB ≤ 0.95).

Functional annotation and protein network of PSGs

To estimate the function of 47 PSGs combination uniquely identified in coelacanth, I performed functional annotation analysis by using DAVID (Huang et al., 2009). These genes were enriched in 4 major clusters of biological processes; nitrogen compound metabolic process (NCMP), metabolic process, spindle organization, and cellular transition metal ion homeostasis (Fig. 3). Out of these, NCMP included interconversion of nitrogenous organic matter and ammonium, which is a key process in adapting to the changing environment during water-to-land transition. In the previous study, Amemiya et al. found that *CPS1* gene, which is involved in ammonium conversion, was accelerated in of MRCA of tetrapods and MRCA of amniotes by adaptation to land (Amemiya et al., 2013). Out of 14 PSGs of NCMP, 4 genes -*DDX11*, *DDX49*, *MMS19*, and *TRMT1*- showed protein interactions with 2 genes -*ARG2* and *CPS1*- of urea cycle (Fig. 4). Out of these 4 genes, both of *MMS19* and *TRMT1* showed the highest numbers of residues under positive selection on functional domains among NCMP genes. These genes were also directly associated with *ARG2* and *CPS1*.

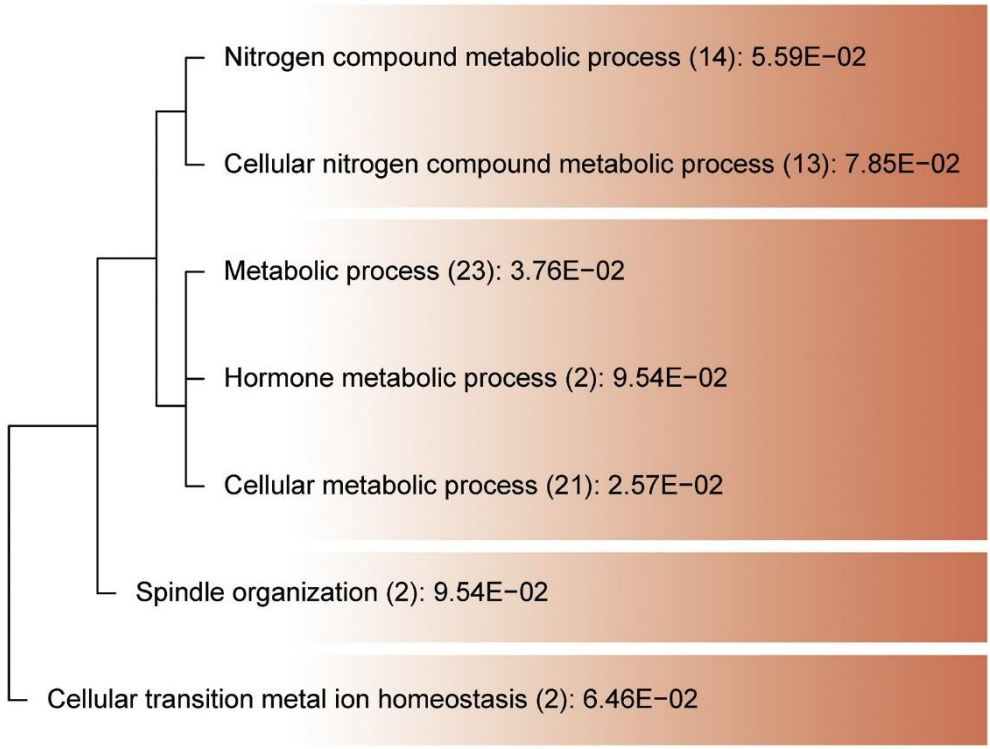

Figure 4.3. Enriched GO term of coelacanth-specific PSGs. Four clusters of biological processes divided in red shades.

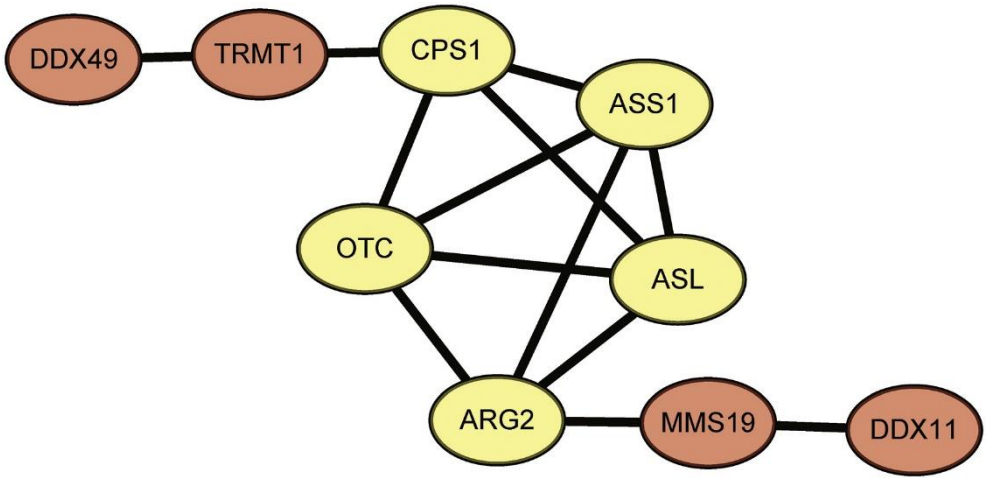

Figure 4.4. Protein-protein interaction networks among genes of urea cycle and coelacanth-specific PSGs of nitrogen compound metabolic process. Red and yellow circles indicate coelacanth-specific PSGs and genes of urea cycle, respectively.

Non-synonymous substitutions on homeobox gene superfamily

In previous study (Amemiya et al., 2013), the genetic alteration on the regulatory regions of *HOX* genes, associated with morphological developments, were investigated in order to discover molecular evolution of limb emergence in tetrapod. However, it was not discovered that genetic alterations on coding regions causing gene product alterations were responsible for anatomical changes of the MRCA of lobe-finned fishes and four-legged vertebrates different from actinopterygian fishes. dN/dS analysis explains molecular evolutionary history of coelacanth based on non-synonymous and synonymous mutations, but it does not identify amino acid substitutions specific to coelacanth which may lead to changed functions of the resulting protein products. So, I conducted target-specific amino acid substitutions (TAAS) analysis (Zhang et al., 2014) to identify coelacanth-specific variation in homeobox gene superfamily.

Within 3538 conserved one to one orthologues, 43 genes were *HOX* gene superfamily. Out of these, 40 genes showed 603 amino acid substitutions specific to coelacanth compared to ray-finned fishes. Including 4 outgroup species in tetrapod vertebrate lineage, I found only 35 genes which contained 300 coelacanth-specific substitutions showing the same information as that of tetrapod. All of 35 genes did not show strong statistical significance; however, 6 of them showed higher likelihood values in alternative model than the null model ($D > 0$), which may be the evidence of positive selection on parts of the genes. Out of these 6 genes, 3 genes showed 4 coelacanth-specific amino acid with significant posterior probability (BEB > 0.95). Especially, *SHOX* gene included the top number of amino acid substitutions. One of the amino acid in *SHOX* gene, serine was shared between coelacanth and some of tetrapod animals as opposed to that of ray-finned fishes, leucine.

Focusing on *SHOX* gene, I collected and aligned amino acid sequences of 100 vertebrates (83 tetrapod species and 17 fishes including coelacanth) in UCSC genome browser (Meyer et al., 2013). *SHOX* gene was present in 81 vertebrates, but was absent in 19 species (Fig. 5). In the candidate site, all of *Sarcopterygii*, including

tetrapods and coelacanth, showed different non-synonymous substitutions (asparagine, serine, threonine, and glycine) from *Actinopterygii* (leucine) (Fig. 5).

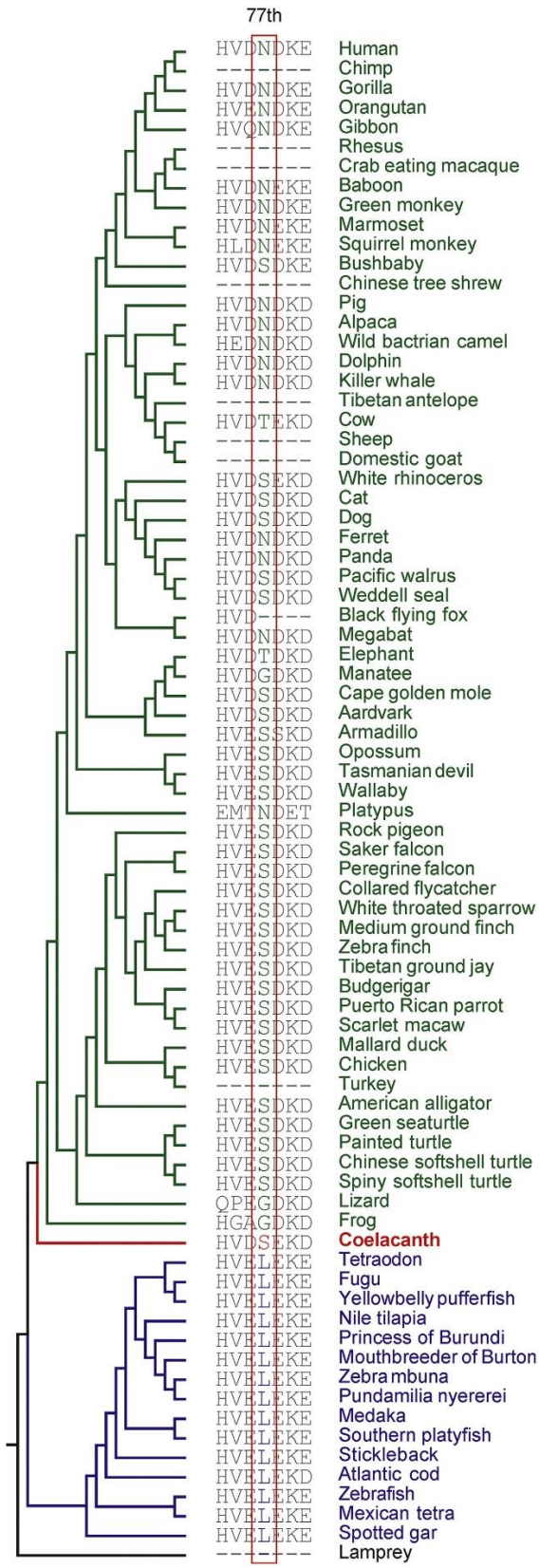
77th
HVDNDKE Human
---+--- Chimp
HVDNDKE Gorilla
HVENDKE Orangutan
HVQNDKE Gibbon
---+--- Rhesus
---+--- Crab eating macaque
HVDNEKE Baboon
HVDNDKE Green monkey
HVDNEKE Marmoset
HLDNEKE Squirrel monkey
HVDSDKE Bushbaby
---+--- Chinese tree shrew
HVDNDKD Pig
HVDNDKD Alpaca
HEDNDKD Wild bactrian camel
HVDNDKD Dolphin
HVDNDKD Killer whale
---+--- Tibetan antelope
HVDTEKD Cow
---+--- Sheep
---+--- Domestic goat
HVDSEKD White rhinoceros
HVDSDKD Cat
HVDSDKD Dog
HVDNDKD Ferret
HVDNDKD Panda
HVDSDKD Pacific walrus
HVDSDKD Weddell seal
HVD---- Black flying fox
HVDNDKD Megabat
HVDTDKD Elephant
HVDGDKD Manatee
HVDSDKD Cape golden mole
HVDSDKD Aardvark
HVESSKD Armadillo
HVESDKD Opossum
HVESDKD Tasmanian devil
HVESDKD Wallaby
EMTNDET Platypus
HVESDKD Rock pigeon
HVESDKD Saker falcon
HVESDKD Peregrine falcon
HVESDKD Collared flycatcher
HVESDKD White throated sparrow
HVESDKD Medium ground finch
HVESDKD Zebra finch
HVESDKD Tibetan ground jay
HVESDKD Budgerigar
HVESDKD Puerto Rican parrot
HVESDKD Scarlet macaw
HVESDKD Mallard duck
HVESDKD Chicken
---+--- Turkey
HVESDKD American alligator
HVESDKD Green seaturtle
HVESDKD Painted turtle
HVESDKD Chinese softshell turtle
HVESDKD Spiny softshell turtle
QPEGDKD Lizard
HGAGDKD Frog
HVDSEKD Coelacanth
HVELEKE Tetraodon
HVELEKE Fugu
HVELEKE Yellowbelly pufferfish
HVELEKE Nile tilapia
HVELEKE Princess of Burundi
HVELEKE Mouthbreeder of Burton
HVELEKE Zebra mbuna
HVELEKE Pundamilia nyererei
HVELEKE Medaka
HVELEKE Southern platyfish
HVELEKE Stickleback
HVELEKD Atlantic cod
HVELEKE Zebrafish
HVELEKE Mexican tetra
HVELEKE Spotted gar
---+--- Lamprey

Figure 4.5. Amino acid substitutions specific to coelacanth and tetrapod mutually exclusive to fishes on *SHOX* gene. Red box in peptide alignments indicates the site with coelacanth and tetrapod specific amino acid replacement compared to other fishes. Numbers on top of alignment indicate positions of peptide sequence of human. In amino acid alignments and common names, green, red, and blue indicate tetrapod, coelacanth, and other fishes, respectively. Tree and alignment are from UCSC genome browser database (Meyer et al., 2013).

4.5. Discussion

Coelacanth genome has been known to give insight into the process of landing in vertebrates' evolutionary history. In the track of vertebrate's lineage showing their transition from water to land, several important characteristics for terrestrial adaptation appears. Reduced number of skeletal structures in limbs as they become larger, reinforced with muscular support, is a tendency which is regarded more beneficial by using limbs to move through tangled plants in shallow water or lifting the body against gravity in terrestrial environment. Different from the aquatic environment where nitrogenous wastes are excreted in form of ammonia, showing cellular toxicity and requiring large amount of water to remove, terrestrial life-forms should convert the ammonia into urea or uric acid, to limit the water expenditure. As one species of sarcopterygians with such landing-related trait's appearance, coelacanth has been researched to understand primitive evolution which makes sarcopterygians different from actinopterygians.

In this study, I found advantageous genetic alterations of coelacanth by using two comparative genomics approaches within bony vertebrates, *Osteichthyes* (Fig. 1). Firstly, I performed dN/dS analysis to identify 47 genes with significant sites under positive selection in coelacanth compared to ray-finned fishes, and I found *NCDN* gene which contained the most positively selected sites (Fig. 2). The functional annotation for these PSGs showed significant 4 biological process clusters including nitrogen compound metabolic process with 14 PSGs specific to coelacanth (Fig. 3). Out of these 14 PSGs, *MMS19* and *TRMT1* directly interact with *ARG2* and *CPS1* related to urea cycle (Fig. 4). Furthermore, I found coelacanth specific amino acid substitutions based on TAAS analysis for 43 homeobox superfamily genes which are known to be associated with limb emergence in tetrapods. As a result, *ALX1, DLX5* and *SHOX* were identified based on the LRT scores (*D*). *SHOX* gene consisted of the highest number of sites with coelacanth-specific amino acid residues which are estimated to have received positive selection. Moreover, one of these

substitutions in *SHOX* showed coelacanth and tetrapod specific information which is different from that of Actinopterygian lineages (Fig. 5). For the 3 phenotypes mentioned earlier which are related with coelacanth specific genotypes, I conducted a detailed search on biological functions.

For spatial learning, I identified neurochondrin (*NCDN*) showing the top number of sites located in its functional domains (Fig. 2B). *NCDN* was reported as a candidate gene involved in improved spatial learning process (Schweitzer et al., 2006). It was suggested that modulation of *NCDN* residue like palmitoylation have an essential role in its specific endosomal targeting (Schwaibold and Brandt, 2008, Shinozaki et al., 1997). I estimate that evolution of the gene including the alteration of residues in functional domain was advantageous for coelacanth to understand complex pattern of the marinal landform. Since coelacanth lives near the ocean floor in the deep sea, there is a possibility that coelacanth independently acquired the characteristic after the divergence from the tetrapods. However, if the alteration was inherited from the MRCA of tetrapods and coelacanth relative to *Actinopterygii*, it may indicate that the trait is beneficial for landing.

Under water, most fishes can easily release nitrogen wastes as ammonium through gill. However, land animals require a safe mechanism to discharge nitrogen wastes, for example, by excreting them as non-toxic nitrogenous organic matters. I discovered 2 candidate genes under positive selection on coelacanth different from ray-finned fishes, which are directly associated with nitrogen metabolism converting ammonia into other types of nitrogen compounds. Interestingly, *CPS1*, a key gene in urea cycle (Amemiya et al., 2013), was directly associated with *TRMT1* as a co-expressed gene (Fig. 4). The tRNA Methyltransferase 1 (*TRMT1*) is known to regulate tRNA processing and gene expression. *TRMT1* contains two major domains: Zinc finger domain to bind tRNA and methyltransferase domain to regulate translation of gene (Liu and Stråby, 2000). Positively selected sites specific to coelacanth on *TRMT1* were located in the methyltransferase domain. I suggest that heterotypic *TRMT1* could be related to alterations of gene expression or translation process of *CPS1*. On the other hand, *MMS19* showed direct interaction with *ARG2* which is related to urea cycle. *MMS19* nucleotide excision

repair homolog was involved in one of chaperones participating in the cytoplasmic Iron-Sulfur cluster protein assembly, which is vital for all living cell (van Wietmarschen et al., 2012). Based on the interaction between *ARG2* and *MMS19*, I predicted that *MMS19* could affect the function of *ARG2* by helping the formation of the protein structure. However, the possibility that mutations on *TRMT1* and *MMS19* affecting genes other than *CPS1* and *ARG2* was not considered in the current comparative genomics approach. Since coelacanth still inhabits the ocean, it is possible for these genes of coelacanth to evolve in order to cancel out adaptive mutation in *CPS1* for energy saving by maintaining ammonia-excretion system. Therefore, it is important to validate relationships between nitrogen waste metabolism and these candidate genes.

Lobed-fins in coelacanth are one of the major characteristics relevant with water-to-land transition process, their proximal domains having common ancestry with tetrapod limbs' stylopod and zeugopod (Yano and Tamura, 2013). Coelacanth possesses several cartilages in their fins, including a component homologous to land vertebrates' humerus and femur, which is not found in actinopterygians. Considering evolution of terrestrial vertebrates' limb was the result of lengthened and enlarged lobed-fins with enlarged endoskeletons and muscular support, the difference in appendage anatomy in coelacanth compared to actinopterygians give primary information to date back early process of limb emergence. In the molecular level, evolution in homeobox genes and their regulatory elements have been known to largely participate in limb emergence, with several models explaining how the mutation in the gene cluster made evolution in vertebrate limbs (Coates and Cohn, 1999, Tabin and Laufer, 1993). In sarcopterygians including coelacanth, *HoxA* and *HoxD* cluster specifying segments in limb which were emerged from 4-fold *Hox* gene cluster duplication followed by functional diversification, is one of the evolution in homeobox genes related to limb emergence (Coates, 1994). Conserved non-coding elements (CNEs) of *HoxD* cluster in coelacanth were analyzed to catch variation responsible for the change in gene expression in previous study, focusing on regulatory elements of genes (Amemiya et al., 2013). Also, T-box genes which are group of transcription factors to control homeobox genes' expression are closely

related to the limb evolution, as the *Tbx2/3/4/5* genes forms two tightly linked genes in the evolutionary lineage (Ruvinsky and Gibson-Brown, 2000).

I focused on alteration in gene products caused by mutation in protein coding region related to limb emergence rather than the gene expansion or change in gene expression as mentioned above. I suggested 3 candidate genes in the homeobox gene superfamily. The first candidate is Aristaless-related homeobox gene (**Alx1**), which is related to development of craniofacial and scapular bones as the body part 'arista' indicates a bristle arising from its head in drosophila. Model organism studied well to describe the gene's effect on phenotype is mouse, having abnormality in craniofacial and scapular bones when mutation occurs, having relation with its incomplete functioning during development as described in Mouse Genome Database (MGD) (Eppig et al., 2014, Kuijper et al., 2005, Qu et al., 1999, Zhao et al., 1996). *Alx1* contains conserved upstream sequence that serves as a binding site of *Pbx1* and *Emx2* to control scapular blade development (Capellini et al., 2010). Also, *Alx1* (*Cart1*) mutant mouse showed slight loss of anterior blade bone, and truncated clavicle as *Alx4* gene additionally being lost (Kuijper et al., 2005). Based on the reported phenotypic change in scapular bones articulate with proximal limbs, it would be worth to test whether the relationship between mutation in *Alx1* and phenotypic impact as further study.

Another candidate, *Dlx5* belongs to Distal-less homeobox (*Dlx*) gene family which is related to limb development in broad range of animals including vertebrates (Panganiban and Rubenstein, 2002). High conservation of *Dlx* gene family across species suggests its crucial role in development, especially related with appendage growth (Stock et al., 1996). Though recent studies have demonstrated their additional developmental roles including craniofacial morphogenesis, neurogenesis and hematopoiesis, Dlx5 gene as role for normal development of limbs, digits and other craniofacial bones like a mandible (Depew et al., 2002, Kraus and Lufkin, 2006, Merlo et al., 2002). Coupled with *Dlx6*, *Dlx5* plays vital role in mammalian limb development, having epistasis over *Msx2* homeodomain transcription factor which also participates in the appendicular skeletal development (Robledo et al., 2002, Vieux-Rochas et al., 2013). Also,

Dlx5 was suggested to be a candidate gene for split hand/foot malformations (*SHFM*) in human, deduced from the patients with nonsense mutation in exonic region.

Short stature homeobox gene *SHOX* is the last candidate, an X-linked gene firstly described from Turner syndrome patients' distinctive abnormality, short stature. Idiopathic short stature and shorten limbs in the patients stem from haploinsufficiency of normal *SHOX* as an X chromosome becomes absent. Critical role of *SHOX* in limb growth and development appears in several human disorders with shorten, malformed limbs like Léri-Weill dyschondrosteosis, Turner syndrome and Langer mesomelic dysplasia as the amino acid substitution or deletion in *SHOX* occurs (Barca-Tierno et al., 2011, Fukami et al., 2005, Rao et al., 1997, Superti-Furga et al., 1998). As the general feature of homeobox genes, *SHOX* also shows high conservation from mammals to fish and flies. Therefore, lack of functional *SHOX* protein disrupts normal bone growth in many vertebrates, not only for human. In zebrafish, pectoral fin-bud is one of major part that *SHOX* is predominantly expressed, and the blockage of *SHOX* expression results in disruption in normal bone development (Sawada et al., 2015). In chicken, artificial overexpression of *SHOX* in their limbs significantly increased the length of skeletal elements (Tiecke et al., 2006) Also, considering *SHOX* influence the bone development from the early stage of embryogenesis by controlling downstream genes like *CTGF* and *FGFR3* which are involved in limb development, mutation in single site of *SHOX* gene can alter the pathway, possibly affect interaction with those proteins (Beiser et al., 2014, Decker et al., 2011).

In conclusion, I analyzed coelacanth genome to gain insights into the evolutionary process that affected landing of sarcopterygians compared to actinopterygians. I tried to show that molecular evolution specific to the lobe-finned fish different from ray-finned fishes can provide meaningful information about the primitive evolution of sarcopterygians related to water-to-land transition. Based on comparative genomics approaches, I identified key candidate genes (*NCDN, MMS19, TRMT1, Alx1, Dlx5, and SHOX*) leading the episodic adaptive evolution for primitive changes in the sarcopterygian clade to influence on water-to-land transition of tetrapods. However, biological validations with genome editing technologies are

required to verify the causality between candidate genes and change in phenotype for adaptation in land. I expect that these novel candidates will give insights into evolutionary history of coelacanth and tetrapod adjusting to life ashore.

This chapter was submitted to *Molecular biology and evolution*
as a partial fulfillment of Chul Lee's Ph.D. program

Chapter 5. Amino Acid Convergences between Independent Lineages in Birds Give Evolutionary Insights into Avian Vocal Learning

5.1. Abstract

Vocal learning, a convergent trait to imitate sounds heard and an important component of spoken-language, is rarely observed in Mammalia and Aves. Molecular convergences of several vocal learning mammals were already discovered, but that of vocal learning birds still remains as an evolutionary enigma. By analyzing avian genomes, here I investigated whether three avian vocal learning clades have amino acid convergences that could explain their specialized trait. I identified single amino acid variants (SAVs) of avian vocal learners and of control sets designed for most species combinations from three independent lineages similar to vocal learning birds, and classified SAVs into convergent and divergent SAVs (ConSAVs and DivSAVs) by considering their ancestral substitutions. I illuminated frequencies of ConSAVs are proportional to products of the most recent common ancestral branches, and confirmed the number of ConSAVs of vocal learning clades in birds did not exceed that of several control sets. I also found amino acid convergences in birds were originated from independent nucleotide substitutions at different sites in each codon. However, gene with ConSAVs of vocal learning birds were uniquely enriched in 'learning' functions, and a subset of ConSAV genes under positive selection were supported by specialized gene expressions in brain subdivisions. Top candidate learning genes, including *DRD1B* and *PRKAR2B*, converged on the cAMP signaling pathway. These results provide insights into molecular mechanisms of the convergent evolution of the vocal learning trait in birds.

5.2. 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 [80,81]. Mutant mice humanized for the two SAVs of *FOXP2* showed more advanced learning abilities [82] and alterations of cortico-basal ganglia circuits [83-85], which play critical roles in spoken-language [86]; and mice containing a heterozygous missense mutation that causes speech syllable apraxia in humans also showed syllable sequencing deficits [87,88].

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 [14,86,89-92]. 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 [14,90,91,93]. 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 [92,94]. Even though songbirds and parrots are relatively closely related, the closest lineage to songbirds [94], 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 [95]. Some of the positively selected genes were in ion channels, which are known to control neurological function, behavior and disease [95]. However, the comparison was made narrowly between only one vocal learner (zebra finch) with one vocal non-learner (chicken), which is a very distant [94] 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 [8], 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 [90,91,93]. I also found mutually exclusive amino acid substitutions unique to vocal learners, using a novel method (Target-specific Amino Acid Substitutions [TAAS] analysis) [8]. However, the study overlooked several viewpoints reported around that time for molecular convergences [96,97]; it did not separately test for convergent versus divergent amino acid substitutions; it did not test for preponderance of convergences and divergences over proper control sets of species; and it did not test for possible influences of close phylogenetic relationships.

Here, I investigated basic rules of molecular convergences and their biological functions in various combinations of avian species, including vocal learners. I improved and developed computational methods to identify convergent and divergent substitutions among species from polyphyletic lineages, and tested whether vocal learning birds have more molecular convergences or divergences than control sets. I discovered phylogenetic features associated with the number of convergent and divergent substitutions among species beyond those of previous studies [19,97], and the underlying nucleotide variant changes associated with these amino acid substitutions. I found a preponderance of higher changes in avian vocal learning clades when considering their most recent common ancestor branch lengths, and I 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.

5.3. Materials and methods

Multiple sequence alignments of singleton orthologous genes in birds

In my preliminary studies, the Avian Phylogenomics Project (now the Bird10K project) defined 8,295 singleton orthologous gene sets across 48 avian species, and constructed the phylogenetic avian family tree consisting of 34 orders [8,94,98]. This 1:1 orthologous gene set was identified by reciprocal best blast hits and synteny, using two species as a reference: chicken and zebra finch. They were then aligned across all species using SATé+MAFFT and SATé+Prank, for both nucleotide and amino acid sequences. Alignment frameshift errors were corrected when translating into amino acid sequence alignments. As results, 4,519,041 amino acids and 13,557,123 nucleotides were detected as homologous sites. In my previous analyses for amino acid substitutions, I used Gblocks [99] to remove poorly scored alignments with sequence divergences and columns with gaps in at least one species included. However, here I found that this was too aggressive, removing 65% of the whole regions of aligned sequences. For example, vocal learner-specific amino acid substitutions of *DRD1B* was excluded because of gaps in one of outgroup species (Lizard). Therefore, I used whole regions of alignments without the trimming step in the current study.

Detection of convergent variants

I initially developed an algorithm to find amino acid substitutions specific to a group of species, called Target-specific Amino Acid Substitution (TAAS) analysis [8]. It could not detect insertion/deletions specific to a group of species. In this study, I improved the algorithm of the previous analysis and applied ancestral sequence reconstructions to find convergent variants at amino acid, codon, and nucleotide levels, and named it as convergent variant finder (ConVarFinder). The ConVarFinder analysis focuses on identifying molecular convergences specific to multi-species from polyphyletic lineages, while TAAS analysis ignored phylogenetic relationships between species of the group with an interest. First, it

identified mutually exclusive variants at amino acid, codon, and nucleotide levels between a target group of species relative to all other species tested (single amino acid variant [SAV], single codon variant [SCV], and single nucleotide variant [SNV]) by analyzing each homologous site in codon sequence alignments of 48 birds. To focus on point mutations, I excluded continuous variants potentially regarded as structural variants. Examples of SAVs and SCVs were summarized and visualized by using WebLogo (v2.8.2) [100,101].

Next, the mutually exclusive variants were classified as 4 types based on equality or inequality of sequence information in each group and the type 1 and 2 variants with same sequence information and the type 3 and 4 with different sequence information in target species were mainly classified as identical and different variants at each level (iSAV, iSCV, iSNV, dSAV, dSCV, and dSNV). In parallel, it analyzed evolutionary histories of the mutually exclusive variants from ancestors to terminal taxa with their phylogenetic tree. The ancestral sequences were estimated by RAxML (version 8.2.12) [102] for codon substitutions with '-f A -m GTRCAT -p 12345' options and for indels converted as binary sequences with '-f A -m BINCAT -p 12345' options. The RAxML usually removed the codon sites consisting of all gaps ('---' or 'NNN') in all species, so I trimmed the reduced sequences when I merged the codon and indel sequences by using a custom python script. Based on the 'RAxML_marginalAncestralStates' and 'RAxML_nodeLabelledRootedTree' outputs, I checked the substitutions on the most recent common ancestral (MRCA=origin) branches of each clade of target species and classified their evolutionary directions as convergences or divergences. The source codes of ConVarFinder and estimated ancestral sequences are accessible at the following link (https://github.com/chulbioinfo/ConVarFinder).

Control sets of species combinations from three independent lineages

Considering that I have 6 vocal learning species I calculated all 6 species combinations of 47 birds in the avian family tree excluding Rifleman, which was 10,737,573 combinations. Of these, 8,239 combinations of 6 species originated from 3 independent lineages including 3 vocal learning clades (songbirds, parrots,

hummingbirds). From these combinations without the set of vocal learners, I designed 2 main types of control sets: all control sets from the 8,238 set of 6 species with 3 independent origins; and core control sets consisting of 59 possible convergent combinations of species that have a similar phylogenetic history to vocal learners, but contained 6 species originated from 2 clades out of 3 vocal learning clades and 1 vocal non-learning clade.

Correlation tests

To check statistical significances of correlations between various features I discovered in this study, such as, convergences and divergences at amino acid level (ConSAVs VS DivSAVs), I calculated Spearman rank correlation coefficient as:

$$rho = \frac{\sum(x' - m_{x'})(y' - m_{y'})}{\sqrt{\sum(x' - m_{x'})^2 \sum(y' - m_{y'})^2}}$$

where x' and y' are each rank of x and y, respectively; and $m_{x'}$ and $m_{y'}$ correspond to the means of rank(x) and rank(y), respectively. By using 'cor.test' function with the option method = "spearman" in R package (ver. 3.5.1), I tested correlations between ConSAVs and DivSAVs in the multiple combinations of species (e.g. a set of avian vocal learners, 8,238 all control sets, and 59 core control sets). After then, I performed linear regression analysis for modeling the relationship between ConSAVs and DivSAVs based on 'lm' function, and visualized it with 'plot', 'points', and 'abline' function in R package (ver. 3.5.1) [48]. I also performed Bonferroni Outlier Test to check whether the number of convergent variants of vocal learners or other species combinations is an outlier, as determined by residuals from regression model with the 'outlierTest' function in R package (ver. 3.5.1) [48,103]; option for limitation of the max number of outliers as 3: 'n.max=3'. I applied the correlation and outlier tests among various features including the frequencies of molecular variants and phylogenetic features of species combinations.

Phylogenetic features related to the number of molecular convergences

I performed multiple clade-wise comparisons of at least 3 polyphyletic clades to find relationships between convergent variants and various phylogenetic features. Using the branch lengths of the avian total evidence phylogenetic tree from Jarvis et al [94], I calculated four types of phylogenetic branch measures for convergent groups of species: product of origin branch lengths (POB), product of terminal branch lengths (PTB), distance between terminal branches (DTB), and distance between terminal nodes (DTN). POB was calculated by multiplying lengths of most recent common ancestral (MRCA=origin) branches of each target clade and PTB as branch lengths of terminal taxa. DTB was calculated as a summation adding lengths of all branches between the MRCA node of the 47 birds and each terminal taxon, whereas the DTN was calculated as the summation between the MRCA node and the most recent ancestral nodes of each terminal taxon. The source code to calculate each phylogenetic feature is accessible at the following link (Https://github.com/chulbioinfo/ConVarFinder).

PCA and ML tree analyses for Rifleman

With the SAV and ConSAV sites found in vocal learners, Rifleman was added and principle component analysis (PCA) was performed using the method as implemented in JalView[104]. Focusing on the 148 AVL-SAV and 24 AVL-ConSAV sites, pairwise scores between bird species was computed by summing the substitution scores from BLOSUM62. Then, I performed spectral decomposition of the score matrix to obtain principal component (PC) vector and eigenvalue of the respective vectors. Sorting the PCs in the descending order of eigen values, I defined the first two vectors as PC1 and PC2. The PCA biplot was computed using these two vectors. For the maximum likelihood (ML) tree, I constructed it using MEGA [105], and selected the JTT model, on the part of the amino acid sequence alignment of all AVL-SAV sites or AVL-ConSAV sites.

Gene ontology functional annotations and gene network analyses

To investigate if there were enriched functions of genes with molecular variants in the vocal learning set (n=1) and control sets (n=8,238), I summarized 53,058 lists of

genes with 1 or more variants considering combinations of 3 types (all, convergent, and divergent variants) at 3 levels (amino acid, codon, and nucleotide levels) specific to each set. I conducted Gene Ontology (GO) analysis by using g:Profiler (v 0.3.5.) [106] with the default option. and ClueGO (ver. 2.3.3.) [107] in Cytoscape [108] with the following options: GO BiologicalProcess-GOA (released in 08.04.2016); all of GO tree interval; all of GO Term/Pathway selection; multiple testing correction by Bonferroni (adjusted p-value < 0.05); and default options of others. After then, I tested whether the number of genes is correlated with the number of significant GO terms and the significances of GO terms, by applying regression analyses using 'lm' function. I visualized the results with 'plot', 'points', and 'abline' functions in the R package (ver. 3.5.1) [48].

After then, focusing on 2 lists enriched for learning process: AVL-ConSAV gene list and DivSCV and DivSNV gene lists of a control set (different codon convergences specific to Dalmatian pelican, little egret, houbara bustard, red-crested turaco, white-throated tinamou, and ostrich), I searched networks between the enriched genes for learning by analyzing protein-protein interactions among convergent genes by using CluePedia ver. 1.3.3. [109] in Cytoscape [108], selecting the following databases: STRING-ACTIONS_v10.0 (released in 07.05.2015); activation v10.0; binding v10.0; catalysis v10.0; expression v10.0; inhibition v10.0; ptmod v10.0, and reaction v10.0. Sequences of the convergent variants of gene lists of vocal learners and a control set associated with learning were summarized and visualized by WebLogo (v2.8.2) [100,101].

Fixed differences of vocal learner-specific amino acid variants within populations of zebra finch and chicken

ConVarFinder analysis was performed with the assumption that a haploid sequences identified are representative of the species. However, variation is also prevalent within a species. More than 20 million (20,739,045) and 1.6 million (1,661,545) variants have been reported in chicken (n=9,586) and zebra finch (n=1,257), respectively, according to Ensembl database release 84 [110,111]. Hence, I performed additional analysis to check if the AVL-SAV sequences I identified not due to within

species variation. Local alignment was conducted for the CDS sequences containing AVL-SAVs using BLAST (ver. 2.8.1) [112] to find the position of SAV on the chromosome sequence of chicken (Galgal4) and zebra finch (taeGut3.2.4) according to Ensembl database release 84[110]. Fixation of sequence in a species was assessed by comparing the chromosomic position of all AVL-SAVs with the polymorphism data of chicken and zebra finch obtained from Ensembl dbSNP build 145 and 139 of chicken and zebra finch, respectively [113]. AVL-SAVs overlapping with polymorphism was considered polymorphic.

I also performed additional fixation analyses on several genes amplified by PCR from red blood cells in blood of zebra finch (n = 3 males and 3 females) and chicken (n = 3 males and 3 females). The *DRD1B (=DRD5)* gene was cloned from genomic DNA by using zebra finch specific primers (forward 5'-GCC CTG CGT CAG TGA GAC CA-3' and reverse 5'-CCG CCA GCC CCC TGT ATG AC-3') and white-leghorn chicken specific primers (forward 5'-CAG ATC TCC CCC GAC CCC GA-3' and reverse 5'-GGC AAC AAT GCC GCC TGG AG-3'). The PCR reaction was conducted a total volume of 20 ul containing 100 ng genomic DNA, 10x PCR buffer, 0.4 µl dNTP (10 mM each), 10 pmol of each primer, and 0.5 U Taq polymerase (BioFACT) in the following thermocycling conditions: 2 min at 95°C, followed by 35 cycles of 20 s at 95°C, 40 s at 60°C, 2 min at 72°C, and, finally, 5 min at 72°C. The PCR products were cloned into the pGEM-T easy vector (Promega) and sequenced using an ABI Prism 3730 XL DNA Analyzer (Thermo Fisher–Applied Bio- systems).

Positive selection on polyphyletic lineages

The *dN* (the rate of non-synonymous substitution), *dS* (the rate of synonymous substitution) and ω = *dN/dS* were estimated along each branch of the phylogenetic tree and across sites by using the branch-site model A, implemented in codeml within PAML ver. 4.6 [20] with F3X4 codon frequencies. I assumed the vocal learning trait in birds was originated from the most recent common ancestral branches of each vocal learning clade. Log likelihood ratio test (LRT, *D* value) was performed to compare the null hypothesis with a fixed ω (model 2) and an alternative hypothesis

with an estimated ω (model 2). Orthologs with ω2 Foreground > 1 and number of accelerated sites (BEB > 0.5) > 0 were retained (branches tested for positive selection are referred to as "foreground" branches and all other are referred to as "background" branches).

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

Specialized gene expression in song learning nuclei and singing-regulated genes

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

1) DEG_2014: A data set of 1,849 differentially expressed genes between song nuclei (RA, HVC, LMAN, and Area X) from Whiteney et al.[114] and Pfenning et al.[114], were I selected those that had expression in one nucleus different from all others (NUC VS other NUCs) .

2) DEG_2019: A data set of 1,148 differentially expressed genes between a song nucleus relative to its surrounding brain region (NUC VS SUR) that were obtained using the micro-dissected method (Gedman et al in preparation).

3) DEG_2020: A data set of differentially expressed genes obtained by the laser capture microscope (LCM) (Gedman et al in preparation) in 5 different comparisons: (a) 2,065 differentially expressed genes among four song nuclei (RA, HVC, LMAN, and Area X) relative to the surrounding brain regions (NUC VS SUR), (b) 4,148 differentially expressed genes between a song nuclei relative to another song nuclei (NUC VS NUC), (c) 3,308 differentially expressed genes between a surrounding region of a song nucleus relative to another surrounding region (SUR VS SUR), (d) 1,942 differentially expressed genes among a song nucleus relative to the other song nuclei (NUC VS other NUCs), and (e) 1,388 differentially expressed genes among a

surrounding region of a song nucleus relative to the other surrounding regions (SUR VS other SURs).

4) SRG_2014: A data set of 1,108 singing-regulated genes in zebra finch by using microarray approaches from Whiteney et al.[114].

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

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

Institutional review for animal cares and experiments

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

5.4. Results

Amino acid convergences specific to avian vocal learning clades

Based on the 48 genomes of avian species [8] spanning most orders in their phylogenetic tree [94], I 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 (**Figure. 5.1A**). Rifleman, a close relative of songbirds and sub-oscines, was initially excluded because of the uncertainty of its vocal learning ability, although assumed to be a vocal non-learner [94].

To understand molecular convergences related to the vocal learning trait in birds, I developed a new method to detect as avian vocal learner-specific convergent variants by improving the algorithm of TAAS analysis and applying the ancestral sequence reconstructions (**Figure. 5.1B, 5.2, 5.3**). I named the new approach as 'Convergent Variant Finder (ConVarFinder)' analysis and performed it for 4,519,041 homologous amino acid sites in multiple sequence alignments of 8,295 orthologous genes used as a standard of core orthologous gene sets of 48 avian species [8]. Out of these homologous sites, 148 sites (0.0033%) detected in 135 genes (1.6%) contained single amino acid variants (SAVs) of vocal learning birds mutually exclusive to vocal non-learning birds (**Table 5.1**). The vocal learner-specific SAVs (VL-SAVs) were logically classified into four types based on equality or inequality of sequence information (SI) within each group of vocal learning and non-learning birds, respectively (**Figure 5.1B**). Out of 148 VL-SAV sites, 24 sites (16%) showed identical SAVs (iSAVs; type 1 and 2 SAVs) and 124 sites (84%) showed different SAVs (dSAVs; type 3 and 4 SAVs) within avian vocal learners (**Table 5.1**). For example, the 253rd site of *B3GNT2* was a Type 1 (iSAV) 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 4 (iSAV) site with glutamine, valine, and leucine (Q, V, and L) observed in avian vocal learners and isoleucine and alanine (I and A) in all vocal non-learners (**Figure 5.1C**).

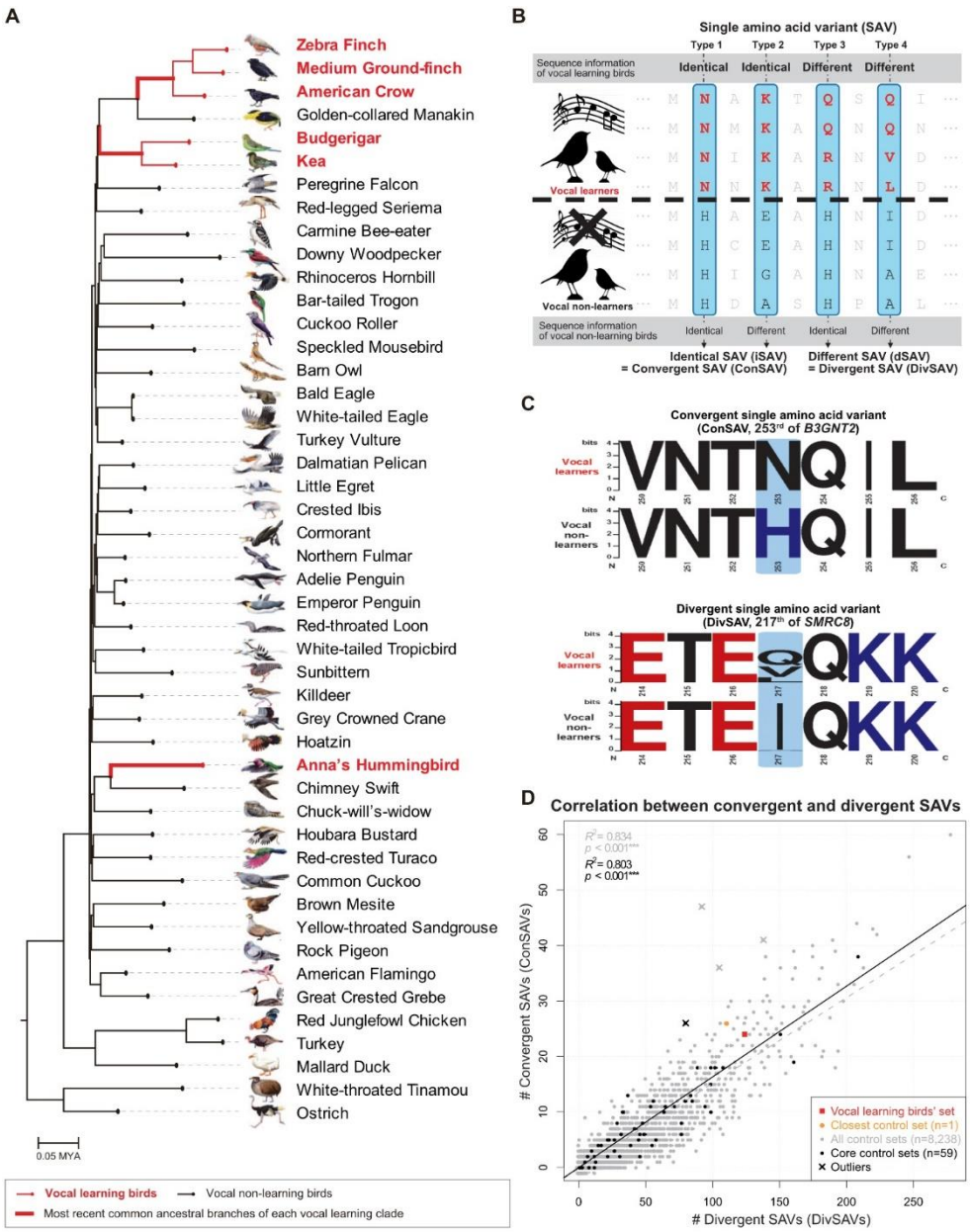

Figure 5.1. Amino acid convergences of avian vocal learning clades do not show the top-predominance compared to control sets. (**A**) Avian family tree and genomes analyzed. The branch lengths of the 48 birds is estimated from the RAxML tree of Jarvis, et al.[94]. Red, avian vocal learning lineages. MRCA (origin branch) of each vocal learning clade is indicated as a bold red line. (**B**) Illustration of the four types of single amino acid variants (SAV, sky blue-colored boxes) and their sequence information in vocal learning birds versus vocal non-learning birds classifying them

into identical and different SAVs (iSAVs and dSAVs, respectively). The iSAVs and dSAVs were perfectly matched with convergent and divergent SAVs (ConSAVs and DivSAVs) defined by substitutions at most recent common ancestral branches of each clade of target species (**Table 5.2**). (**C**) Example cases of an convergent SAV (ConSAV) site in *B3GNT2* and a divergent SAV (DivSAV) site in *SMRC8*. (**D**) Correlation plots between amino acid convergences (ConSAVs; y-axis) and divergences (DivSAV; x-axis) of control species sets of 6 species originated from 3 independent lineages.

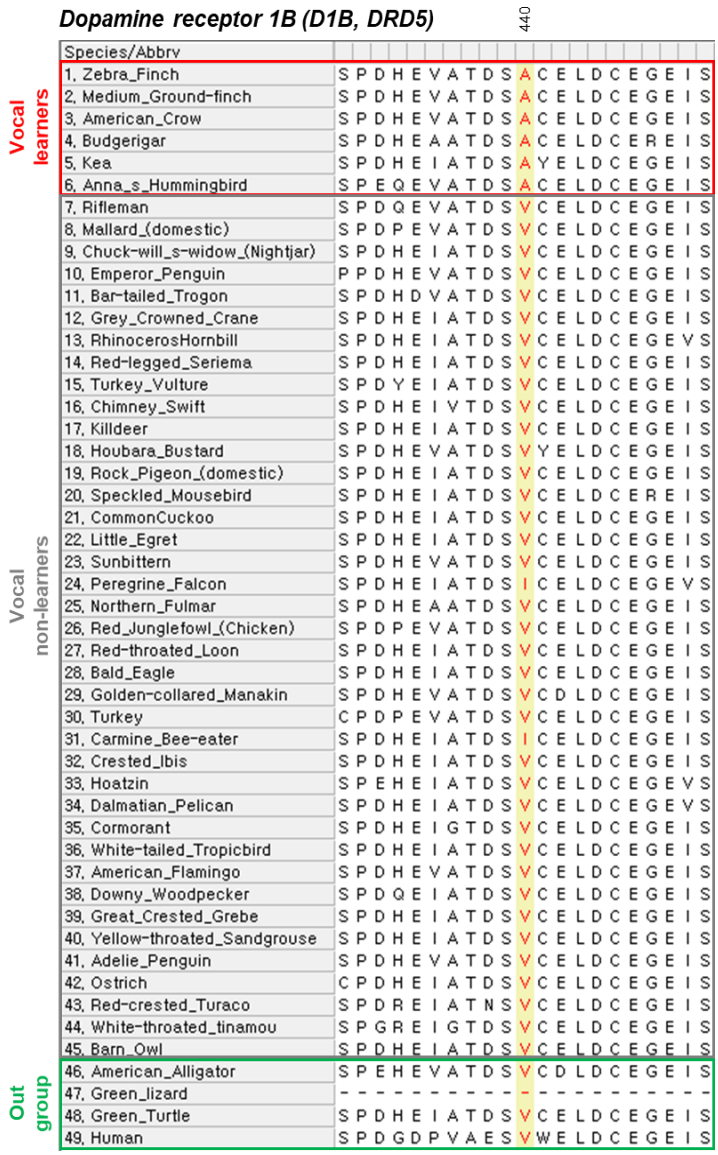

Figure 5.2. Example of a trimmed region with a low alignment score caused by a regional deletion in an outgroup species. Note the outgroup lizard has missing sequence, which would have caused the entire sequence be removed for all species using G-blocks, and thus the convergent site in vocal learning birds (yellow) would have not been discovered.

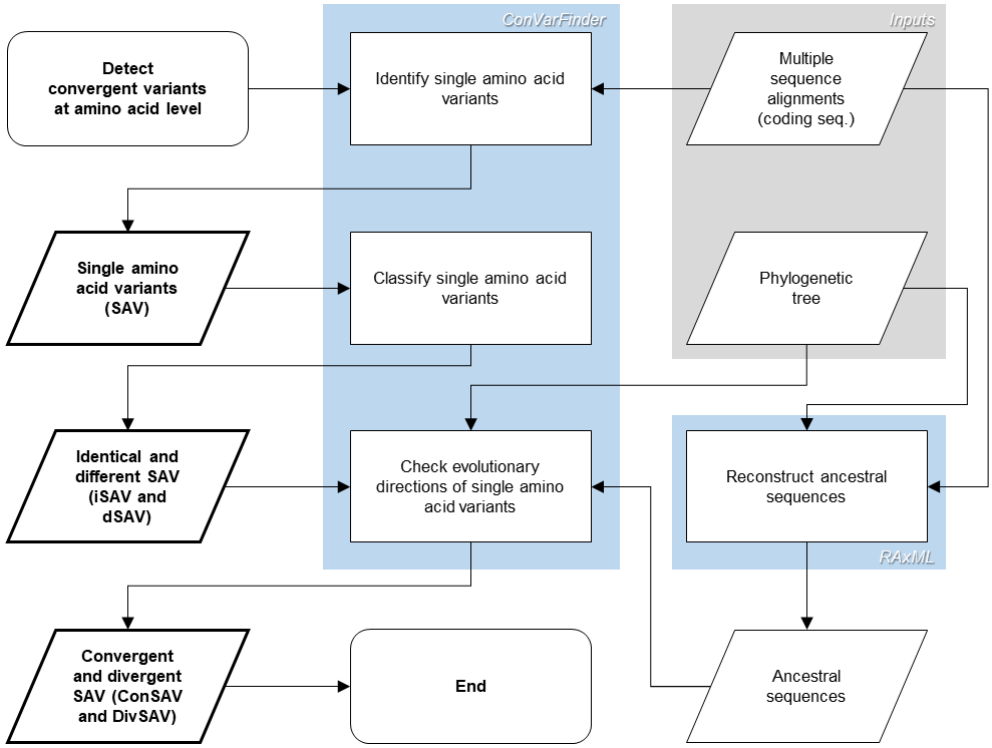

Figure 5.3. Flow chart of convergent variant finder (ConVarFinder).

Table 5.1. Avian vocal learner-specific single amino acid variants and its supporting evidence. Partial data sheet with amino acid information.

Symbol	Pos_AA	AA_targets	AA_others	Type_AASub_target	Type_AASub_others	Type_AASub	AA_MRCAbranches	EvoDir_MRCA2Ter_AA
ABCG2	330	L,T,A	E,R,I,K,V	d	d	4	L>L/V>T/V>A	cDIV
ACP2	92	D,Q	E,G,A	d	d	4	D>D/D>D/A>Q	cDIV
AKAP1	55	R,S,Q	D,H,K,N,-,G	d	d	4	S>S/K>Q/K>R	cDIV
ALKBH4	195	W,S	C,L,Y,A,P,F	d	d	4	S>S/S>S/W>W	cDIV
ALPK1	675	C,V,N	D,A,-,G,S	d	d	4	C>C/S>V/G>N	cDIV
ANKFN1	234	V,A	G,L,M,-	d	d	4	G>V/V>V/M>V	cDIV
ANKRD16	234	G,S,A	D,-,N	d	d	4	G>G/S>S/D>A	cDIV
ARMC6	53	P,-	Q	d	i	3	P>P/P>P/P>P	cDIV
ASH1L	61	D,S	R,-,N	d	d	4	S>S/->D/N>D	cDIV
B3GNT2	253	N	H	i	i	1	H>N/N>N/H>N	cCON
BIRC7	11	P,M	I,L,A,S,-,T,V	d	d	4	P>P/T>M/T>M	cDIV
BMP2K	334	-,A	P,T,N	d	d	4	A>A/T>A/A>A	cDIV
BRCA2	422	R,K,-,G,S	D,A	d	d	4	G>R/->-/D>K	cDIV
BRIP1	913	G,T	I,A,-,F,V	d	d	4	G>G/T>T/I>T	cDIV
C12orf35	170	P,H,Y,G	E,K,Q	d	d	4	P>P/Q>H/G>G	cDIV
C12orf35	604	R,-	E,G	d	d	4	E>-/E>-/G>R	sDIV
C12orf55	315	D,W,-	C,H,Y,F,S,Q	d	d	4	->-/Y>W/->D	cDIV
C3orf67	426	D,H,R,N,T	L,I,K,-,P,V	d	d	4	->T/->R/->N	cDIV
C8B	489	L	F,-,V	i	d	2	F>L/V>L/L>L	cCON
C9ORF152	53	F,M,T,V	L,K	d	d	4	I>V/F>F/L>F	cDIV
CCDC13	216	P,K,A	R,T,-	d	d	4	A>A/K>K/P>P	cDIV
CCDC69	43	E,D,S	H,K,N,-,T	d	d	4	E>E/D>D/N>S	cDIV
CD86	91	P,N,T,V	E,D,H,K	d	d	4	K>T/V>V/H>N	cDIV
CDCA7	256	L,S	A,-,P,T,V	d	d	4	L>L/L>L/S>S	cDIV
CFAP70	469	G,R,H	I,A,-,F,T,M,V	d	d	4	I>H/G>G/R>R	cDIV
CFLAR	391	R,L,S	I,T,-,V	d	d	4	S>S/I>L/I>S	cDIV
CHGB	252	G,A	E,D,N	d	d	4	D>G/G>G/D>G	cDIV
CLBA1	241	G,-,A	D,N	d	d	4	A>A/D>G/->-	cDIV
CLUL1	231	I,-,V	D,H,A	d	d	4	V>V/V>V/D>-	cDIV
COL6A3	171	L,T	I,M,-	d	d	4	M>L/L>L/T>T	cDIV
COL6A3	185	R,N,Q	K,-	d	d	4	K>Q/Q>Q/N>N	cDIV
COL6A3	1526	L,M	V	d	i	3	V>L/L>L/M>M	cDIV
CXorf21	173	P,T,A	C,F,-,S	d	d	4	S>T/A>A/S>P	cDIV

Symbol	Pos_AA	AA_targets	AA_others	Type_AASub_target	Type_AASub_others	Type_AASub	AA_MRCAbranches	EvoDir_MRCAZTer_AA
DERA	252	Y	H,I,N,-,F	i	d	2	Y>Y/Y>Y/I>Y	cCON
DLGAP5	533	K,S,A	E,D,-,G	d	d	4	K>K/S>S/E>A	cDIV
DNAH10	330	I,T,V	L,-,S	d	d	4	T>T/T>T/S>T	cDIV
DNAH10	1297	N,A	D,H,Q	d	d	4	N>N/N>N/P>A	cDIV
DNAH10	2268	N,Q	C,R,H,K,Y,T	d	d	4	Q>Q/H>N/Q>Q	cDIV
DNAH10	3382	H,T	S,-,N	d	d	4	T>T/H>H/T>T	cDIV
DNAH3	2947	Q	E,R,K,N,-	i	d	2	R>Q/Q>Q/R>Q	cCON
DRC7	627	D,I,V	R,K,A,T,M,S	d	d	4	I>I/S>D/I>V	cDIV
DRC7	657	G,D,H	E,R,K,N,S	d	d	4	E>D/D>H/R>D	cDIV
DRD5	416	A	I,V	i	d	2	A>A/A>A/V>A	cCON
E2F8	463	G,E,A	H,N,-,V,T,S	d	d	4	A>A/G>G/E>E	cDIV
EFHB	87	P,T,S	C,R,H,Y,-,G,Q	d	d	4	P>T/P>P/S>S	cDIV
EFHB	245	P,E,K,N	G,-,A	d	d	4	P>P/A>D/->K	cDIV
EFHB	431	F,Y,S	E,L,I,A,-,G,V	d	d	4	L>Y/F>F/A>S	cDIV
EFHC1	77	E,V	P,T,-,A	d	d	4	V>V/V>V/P>E	cDIV
ENPP1	723	P,D,S	T,V,A	d	d	4	S>S/D>D/P>P	cDIV
ENSGALT00000010226	413	K,S	R,H,Y,N,-	d	d	4	S>S/S>S/N>K	cDIV
ENSGALT00000012528	233	D,S,Q	H,K,N	d	d	4	Q>Q/R>S/D>D	cDIV
ENSGALT00000015652	665	V,A	L,M,-,I	d	d	4	V>A/V>V/S>A	cDIV
ENSGALT00000017732	652	Z,H,Y	L,N,W,-,T,S,Q	d	d	4	H>H/S>Z/Q>Z	cDIV
ENSGALT00000025242	99	E,-,S	G,D	d	d	4	D>-/G>S/D>E	cDIV
ENSGALT00000027531	902	P,E,I	R,K,A,T,M,S,Q	d	d	4	P>P/P>I/A>E	cDIV
ENSGALT00000030336	164	T,V	P,G,A	d	d	4	P>V/T>T/T>T	cDIV
ENSGALT00000032705	82	K,N,Q	R,I,D,-,G,S	d	d	4	K>K/Q>Q/G>N	cDIV
ENSGALT00000032989	246	R,H,L	K,Q	d	d	4	H>H/R>R/Q>L	cDIV
ENSGALT00000036845	48	S	G,R,K,E	i	d	2	R>S/S>S/R>S	cCON

Symbol	Pos_AA	AA_targets	AA_others	Type_AASub_target	Type_AASub_others	Type_AASub	AA_MRCAbranches	EvoDir_MRCAZTer_AA
ENSGALT00000037492	133	I,F	L,S,-,P,V	d	d	4	F>F/L>I/I>I	cDIV
ENSGALT00000039593	87	W,N	E,R,K,-,P,Z,Q	d	d	4	W>W/W>W/N>N	cDIV
FANCI	415	N,A	I,T,-	d	d	4	N>N/T>A/A>A	cDIV
FBXO48	59	S	-,A	i	d	2	->S/S>S/->S	cCON
FGFBP1	166	M	I,K,N,A,-,V,T,S	i	d	2	M>M/M>M/I>M	cCON
FNDC1	1034	S	G	i	i	1	S>S/S>S/G>S	cCON
GDPD4	106	L,V	I,-	d	d	4	I>V/I>L/V>V	cDIV
GDPD4	408	G,S	D,N	d	d	4	N>S/N>S/G>G	cDIV
GDPD4	436	E,H,S	G,D,N	d	d	4	N>H/S>S/E>E	cDIV
GPATCH1	429	E,D,-	G,V,S,A	d	d	4	E>E/E>D/A>D	cDIV
GPLD1	280	D,-,S	G	d	i	3	G>D/G>-/G>D	cDIV
GPR35	157	D,-,Q	E,K,V	d	d	4	D>D/E>Q/D>-	cDIV
HAUS8	63	G,N	E,D,S	d	d	4	S>N/N>N/G>G	cDIV
HEATR6	400	C,P,M,-	N,T,S,A	d	d	4	T>M/P>P/T>P	cDIV
HEPH	1098	L,A	H,Y,-,P,S	d	d	4	F>L/L>L/P>A	cDIV
HMGXB3	269	D	E,-	i	d	2	D>D/D>D/E>D	cCON
IBA57	267	P	L,A,-,G,S	i	d	2	S>P/P>P/S>P	cCON
IBSP	259	N	D,A,S,-,G,V	i	d	2	D>N/N>N/N>N	cCON
IFT88	299	E,L,Y	H,-,Q	d	d	4	E>E/L>L/H>Y	cDIV
INPP5E	160	G,S	T,-,A	d	d	4	S>S/S>S/A>G	cDIV
ITFG3	311	D,-,Q	E,R,G	d	d	4	Q>Q/Q>Q/D>D	cDIV
KCNS3	490	M,V,A	I,T,S	d	d	4	V>V/A>A/T>M	cDIV
KIAA0391	306	L,S,A	P,T,-	d	d	4	S>S/P>L/P>L	cDIV
KIAA1841	479	E,Q	G,R,-	d	d	4	Q>Q/E>E/G>E	cDIV
KIF27	476	D,M,V,A	E,K,-,G	d	d	4	E>V/->A/E>M	cDIV
LARP1B	393	C,H,S	G,R,-	d	d	4	C>C/G>R/R>S	cDIV
LPO	287	D,H,Q	-,N	d	d	4	D>D/N>Q/H>H	cDIV
LRRC8A	92	I,M,S	T,A	d	d	4	S>S/I>I/A>M	cDIV
LRRN4	475	H	R,F,Y,S	i	d	2	H>H/H>H/Y>H	cCON
LYVE1	96	I,T,Q	K,Y,V	d	d	4	Q>Q/T>T/I>I	cDIV
LZTFL1	155	H,Y,S	C,R,-	d	d	4	R>H/H>Y/S>S	cDIV
MEI4	255	G,R,T	N,S	d	d	4	R>R/G>G/S>T	cDIV
MFSD4B	243	W,S	C,H,Y	d	d	4	W>W/S>S/Y>S	cDIV
MTFR1	103	T	P,-,G,A	i	d	2	A>T/A>T/A>T	sPAR
MUM1	123	P,F	L	d	i	3	F>F/F>F/F>P	cDIV

Symbol	Pos_AA	AA_targets	AA_others	Type_AASub_target	Type_AASub_others	Type_AASub	AA_MRCAbranches	EvoDir_MRCAZTer_AA
NBN	407	L,I,-	V,A	d	d	4	V>I/I>I/V>L	cDIV
NDC1	454	I,K	S,N	d	d	4	K>K/K>K/S>I	cDIV
NOLC1	341	P,L	-,Q	d	d	4	P>P/P>P/Q>L	cDIV
OTOA	859	G,A	F,T,-,S	d	d	4	A>G/A>A/A>A	cDIV
OTUD3	112	S,A	T,-	d	d	4	A>A/S>S/A>A	cDIV
PAG1	49	Y,N	H,-,Q	d	d	4	H>N/Y>Y/Y>Y	cDIV
PDZD8	482	T,A	P,S,Q	d	d	4	A>A/P>T/T>T	cDIV
PHACTR2	457	G,-	E,A,V,T,S	d	d	4	V>G/G>G/V>-	cDIV
PIK3R4	671	C	R	i	i	1	C>C/C>C/C>C	cCON
PLEKHO1	229	L,T	I,V,A	d	d	4	T>T/T>T/V>L	cDIV
PRKAR2B	32	V	I,-	i	d	2	V>V/V>V/V>V	cCON
PTPRB	914	K,M	T,V,A	d	d	4	K>K/M>M/V>M	cDIV
REST	617	F,Y	C,H,G,S	d	d	4	Y>Y/Y>Y/F>F	cDIV
REXO1	535	P,S	A,-,T,G,V	d	d	4	S>S/A>P/S>S	cDIV
RIOK1	507	V,F,N,A	E,D,K,-,Q	d	d	4	H>N/E>V/V>F	cDIV
RPAP1	746	Y,-,S	C,R,L,H,D,A,Q	d	d	4	->-/S>S/Q>Y	cDIV
SACS	2254	I,N	T,A	d	d	4	T>N/N>N/T>I	cDIV
SCAMP2	106	D	-,N	i	d	2	N>D/D>D/D>D	cCON
SERPINB6	111	I,S	L,F,-,A	d	d	4	->S/->S/A>I	sDIV
SESN1	126	T	M,-,V,A	i	d	2	A>T/A>T/T>T	cCON
SETD4	19	R,K	-,Q	d	d	4	Q>K/Q>K/Q>R	sDIV
SH3BP2	217	P,S	G,T,N,A	d	d	4	A>P/P>P/A>S	cDIV
SMCR8	215	L,V,Q	I,A	d	d	4	Q>Q/V>V/I>V	cDIV
SMPD3	307	C	Y,-	i	d	2	C>C/C>C/Y>C	cCON
SPAG16	124	L,K,T	R,M,V	d	d	4	L>L/K>K/M>T	cDIV
SPART	409	L,I,K,F,M	C,H,Y,-,Z,V	d	d	4	V>L/I>I/F>L	cDIV
SPG11	1876	P,L,A	C,-,S,Q	d	d	4	P>A/P>P/S>P	cDIV
SYNJ2	438	R,Q	H	d	i	3	H>Q/H>Q/H>R	cDIV
TANC1	1619	V	L,I,K,A,-,P,T,M	i	d	2	V>V/V>V/V>V	cCON
TASOR	694	G,P,S	T,-,A	d	d	4	G>G/P>P/T>S	cDIV
TCOF1	279	S,-,V,A	P,L,Q	d	d	4	V>V/V>V/L>S	cDIV
TCTE3	80	E,K,N	G,D,H	d	d	4	D>N/N>N/G>E	cDIV
TDP2	268	E	R,K,T,Q	i	d	2	E>E/E>E/R>E	cCON
TDRD9	789	M,V	L,-	d	d	4	V>V/M>M/M>M	cDIV
TDRD9	984	H,K	S,D,-,N	d	d	4	K>K/H>H/K>K	cDIV
TICRR	328	N,A	L,M,T,V	d	d	4	A>A/A>A/N>N	cDIV

Symbol	Pos_AA	AA_targets	AA_others	Type_AASub_target	Type_AASub_others	Type_AASub	AA_MRCAbranches	EvoDir_MRCAZTer_AA
TMEM209	180	-,S	N,A,P,T,V	d	d	4	S>S/S>-/T>-	cDIV
TNFRSF1A	251	R,H,K,Y	I,N	d	d	4	N>R/N>K/Y>Y	cDIV
TNS3	951	P,E,S	D,L,R,I,K,A,-,F,V	d	d	4	S>S/S>E/S>S	cDIV
TP53I3	354	D,Q,A	E,R,L,K	d	d	4	T>A/D>D/L>Q	cDIV
TPCN2	65	E,-,N	R,K,T	d	d	4	E>E/N>N/K>-	cDIV
TPCN2	114	K,-,A	E	d	i	3	K>K/A>A/E>-	cDIV
TRAFD1	445	G,P,T	E,D,A,-,V	d	d	4	G>G/->T/A>P	cDIV
TREM2	206	G,D,S	H,-,N	d	d	4	G>G/S>S/N>D	cDIV
TSEN2	248	M,S	R,I,A,T,G,V	d	d	4	G>M/M>M/I>S	cDIV
TTC37	752	E,T,A	D,N,-,G,S	d	d	4	T>T/A>A/D>E	cDIV
URB2	106	K	E,A,G,Q	i	d	2	K>K/K>K/E>K	cCON
USP4	263	R,S,A	I,-,T,M,V	d	d	4	A>A/A>S/R>R	cDIV
WDR77	284	G,P,N	R,S	d	d	4	G>G/N>N/P>P	cDIV
WDR78	224	E,F,A	P,-,S	d	d	4	S>E/S>F/S>A	sDIV
XPC	434	P	C,R,H	i	d	2	P>P/P>P/P>P	cCON
ZBTB49	192	G	L,I,A,M,V	i	d	2	A>G/G>G/V>G	cCON
ZC3H6	1124	P,S,A	I,T,-,N	d	d	4	P>P/P>P/T>S	cDIV
ZDHHC1	455	F,M,S	D,I,L,-,T,V	d	d	4	M>M/I>S/L>F	cDIV

Table 5.1. Avian vocal learner-specific single amino acid variants and its supporting evidence. Partial data sheet with codon information.

Symbol	Pos_codon	Codon_targets	Codon_others	Type_CodonSub_target	Type_CodonSub_other	Type_CodonSub	EvoDir_Codon
ABCG2	988	CTA,GCA,CTG,ACA	GAA,GTA,AGA,AAA,ATA	d	d	4	cDIV
ACP2	274	GAT,GAC,CAG	GGG,GCG,GCC,GCT,GAG	d	d	4	cDIV
AKAP1	163	AGT,CAG,AGG	AAG,CAT,AAA,---,GAT,GGT,AAT,AAC	d	d	4	cDIV
ALKBH4	583	TGG,TCC	TTC,GCC,CTG,CCC,TGC,TAC,TTT	d	d	4	cDIV
ALPK1	2023	TGC,AAT,GTT,TGT	AGT,GCT,GAC,---,GAT,GGT	d	d	4	cDIV
ANKFN1	700	GTG,GCG	ATG,CTG,GGG,---	d	d	4	cDIV
ANKRD16	700	GCC,GGC,AGC	GAC,---,AAC	d	d	4	cDIV
ARMC6	157	CCA,---,CCG,CCT	CAG	d	i	3	cDIV
ASH1L	181	GAT,AGC	CGC,AAT,---,AAC	d	d	4	cDIV
B3GNT2	757	AAT	CAT,CAC	i	d	2	cCon
BIRC7	31	ATG,CCA,CCT,CCG	TCT,CTT,GTT,ATT,GCC,AGT,CTG,ACG,GCT,CTC,---,ACC,GTC,ACT,ATC	d	d	4	cDIV
BMP2K	1000	---,GCT	ACC,CCT,ACT,AAC	d	d	4	cDIV
BRCA2	1264	AAG,AGT,CGT,---,GGT	GAT,GAC,GCT,GCC	d	d	4	cDIV
BRIP1	2737	GGG,ACT	GTT,GCC,---,TTT,ATT,ATC	d	d	4	cDIV
C12orf35	508	CCA,TAT,CAT,GGA,CCG	GAA,AAA,CAA,CAG	d	d	4	cDIV
C12orf35	1810	AGA,---	GAA,GGT,GGA,GAG	d	d	4	sDIV
C12orf55	943	TGG,GAC,---	TCT,TTC,TAT,TCC,TCA,CAG,TGC,TGT,TAC,CAC	d	d	4	cDIV

Symbol	Pos_codon	Codon_targets	Codon_others	Type_CodonSub_target	Type_CodonSub_other	Type_CodonSub	EvoDir_Codon
C3orf67	1276	CAT,AGA,ACA,GAC,AAT	GTT,AAG,CTG,AAA,CTC,---,CCT,GTC,ATT	d	d	4	cDIV
C8B	1465	TTG	GTC,TTC,TTT,---	i	d	2	cCon
C9ORF152	157	TTC,GTA,ATG,ACA,TTT	AAA,TTA,CTA,TTG	d	d	4	cDIV
CCDC13	646	AAA,CCA,GCT	AGA,ACG,ACA,---,ACC	d	d	4	cDIV
CCDC69	127	GAC,AGC,GAG	AAG,CAC,AAA,---,ACC,AAT,AAC	d	d	4	cDIV
CD86	271	ACT,CCT,GTT,AAC	AAG,CAT,GAC,GAG,GAT	d	d	4	cDIV
CDCA7	766	TCT,CTT	GTT,GCT,---,CCT,ACT	d	d	4	cDIV
CFAP70	1405	GGC,AGG,CAC	TTC,GTT,GCC,ATG,ACG,GCT,ACT,---,ACC,TTT,ATT	d	d	4	cDIV
CFLAR	1171	AGT,AGA,TTA	GTT,ATT,---,ACC,ACT,ATC	d	d	4	cDIV
CHGB	754	GGC,GGT,GCT	GAA,GAC,GAG,GAT,AAT	d	d	4	cDIV
CLBA1	721	GGT,---,GCT	GAT,AAT,AAC	d	d	4	cDIV
CLUL1	691	GTC,---,ATC	GAT,GAC,GCC,CAC	d	d	4	cDIV
COL6A3	511	CTG,TTG,ACG	ATG,---,ATC	d	d	4	cDIV
COL6A3	553	AAT,CAG,CGG	---,AAG	d	d	4	cDIV
COL6A3	4576	CTA,ATG,CTG	GTG,GTC	d	d	4	cDIV
CXorf21	517	GCA,CCC,GCT,ACA,ACC	TCT,TTC,TCC,TGC,---	d	d	4	cDIV
DERA	754	TAC	CAC,---,TTT,ATC,AAC	i	d	2	cCon
DLGAP5	1597	AAA,TCC,TCA,GCT	GAA,GGA,GAC,---	d	d	4	cDIV

Symbol	Pos_codon	Codon_targets	Codon_others	Type_CodonSub_target	Type_CodonSub_other	Type_CodonSub	EvoDir_Codon
DNAH10	988	ACA,GTG,ATA	TCG,TCA,TTA,---	d	d	4	cDIV
DNAH10	3889	GCC,AAT,AAC	CAA,CAT,CAG,GAC,CAC	d	d	4	cDIV
DNAH10	6802	AAT,CAG	AAG,TAT,CAT,CGT,TGT,ACC,CAC	d	d	4	cDIV
DNAH10	10144	CAT,ACC	AGT,AGC,---,AAT,AAC	d	d	4	cDIV
DNAH3	8839	CAA,CAG	GAA,AGG,AAG,AGA,AAA,---,AAC	d	d	4	cDIV
DRC7	1879	GTG,GAC,GTC,ATC	GCG,AGG,AAG,TCG,AGA,ATG,ACG,AGC,ACA,ACC	d	d	4	cDIV
DRC7	1969	GAT,CAT,GGT,GAC	AAG,AGT,AGA,CGT,AAA,GAG,AAT,AAC	d	d	4	cDIV
DRD5	1246	GCC	GTA,GTT,GTG,GTC,ATC	i	d	2	cCon
E2F8	1387	GCC,GGT,GAA	AGT,CAT,ACT,---,AAT,GTC,AAC	d	d	4	cDIV
EFHB	259	TCT,ACC,CCT	TAT,CGG,CAT,CGT,CAG,TGC,---,TGT,GGT	d	d	4	cDIV
EFHB	733	GAA,AAA,AAT,CCA	GCG,GCA,GCT,---,GGA	d	d	4	cDIV
EFHB	1291	TCT,TTT,TAT	GAA,CTT,GTT,GCT,---,GGT,ATT	d	d	4	cDIV
EFHC1	229	GTC,GAG	GCC,CCC,---,ACC	d	d	4	cDIV
ENPP1	2167	GAT,TCT,CCT	ACT,ACC,GTT,GCT	d	d	4	cDIV
ENSGALT00000010226	1237	AAG,AGC	CGC,---,AAC,TAC,AAT,CAC	d	d	4	cDIV
ENSGALT00000012528	697	GAC,CAA,CAG,AGC	AAA,AAC,CAC	d	d	4	cDIV
ENSGALT00000015652	1993	GCA,GTG,GTA	CTA,ATG,ATA,---,TTA	d	d	4	cDIV
ENSGALT00000017732	1954	CAT,TAG,TAT,CAC	TGG,AGT,CTG,CAG,---,AAT,ACT	d	d	4	cDIV

Symbol	Pos_codon	Codon_targets	Codon_others	Type_CodonSub_target	Type_CodonSub_other	Type_CodonSub	EvoDir_Codon
ENSGALT00000025242	295	AGT,GAA,---,GAG	GAT,GGT,GAC	d	d	4	cDIV
ENSGALT00000027531	2704	CCA,ATT,GAG	CAA,GCG,AGG,AAG,TCG,ATG,GCA,ACG,AAA,ACA,ACT	d	d	4	cDIV
ENSGALT00000030336	490	GTG,ACT	GCG,GCC,GCA,GGC,GCT,CCT,GGT	d	d	4	cDIV
ENSGALT00000032705	244	AAA,CAA,CAG,AAC	GGG,AGT,AGA,GGC,CGT,---,GAC,GGA,GGT,ATT	d	d	4	cDIV
ENSGALT00000032989	736	CAT,CTG,AGG	AAA,CAA,CAG	d	d	4	cDIV
ENSGALT00000036845	142	AGT,AGC	GAA,GGG,AGG,AGA,AAA	d	d	4	cDIV
ENSGALT00000037492	397	TTC,ATC	TCT,CTT,TCC,CTG,CTC,---,CCT,GTC	d	d	4	cDIV
ENSGALT00000039593	259	TGG,AAC	AGG,AAG,TAG,CAG,---,GAG,CCG	d	d	4	cDIV
FANCI	1243	GCC,AAC	ACC,---,ATC	d	d	4	cDIV
FBXO48	175	TCC	GCC,GCA,GCG,---	i	d	2	cCon
FGFBP1	496	ATG	GTT,AAG,AGT,GCT,ACT,---,ACC,AAT,ATT,ATC	i	d	2	cCon
FNDC1	3100	AGC	GGC,GGG,GGA	i	d	2	cCon
GDPD4	316	CTC,GTC	ATA,---,ATC	d	d	4	cDIV
GDPD4	1222	AGT,GGG,AGC	GAT,GAC,AAC	d	d	4	cDIV
GDPD4	1306	AGT,GAA,AGC,CAC	GAT,GGC,GAC,AAC	d	d	4	cDIV
GPATCH1	1285	GAT,GAC,---,GAG	GGG,GTT,AGT,GCC,GGC,GCT,GGT	d	d	4	cDIV
GPLD1	838	GAC,---,AGC	GGC,GGT,GGA	d	d	4	cDIV
GPR35	469	GAT,CAA,GAC,---	GAA,AAA,GTC	d	d	4	cDIV

Symbol	Pos_codon	Codon_targets	Codon_others	Type_CodonSub_target	Type_CodonSub_other	Type_CodonSub	EvoDir_Codon
HAUS8	187	GGT,AAT,AAC	AGT,GAC,GAT,GAA	d	d	4	cDIV
HEATR6	1198	CCA,ATG,TGC,---,CCT	GCC,ACG,AGC,GCT,ACC,AAT,ACT	d	d	4	cDIV
HEPH	3292	CTA,CTT,TTG,GCT	TCT,TAT,CAT,CCC,---,CCT,CCG	d	d	4	cDIV
HMGXB3	805	GAC	---,GAG	i	d	2	cCon
IBA57	799	CCA,CCG	TCT,GGG,TCG,TCA,GCA,---,TTG	d	d	4	cDIV
IBSP	775	AAT,AAC	GCG,GTT,AGT,GGC,AGC,---,GAT,GGT	d	d	4	cDIV
IFT88	895	GAA,CTT,TAT	CAT,CAG,---,CAC	d	d	4	cDIV
INPP5E	478	TCG,TCC,GGG	GCG,GCC,GCA,ACG,ACA,---,ACC	d	d	4	cDIV
ITFG3	931	GAC,---,CAG	GGG,GAG,CGG	d	d	4	cDIV
KCNS3	1468	GCA,ATG,GTA	TCA,ACG,ACA,ATA,ACT	d	d	4	cDIV
KIAA0391	916	GCC,TCC,CTC	CCC,---,CCG,ACC	d	d	4	cDIV
KIAA1841	1435	GAA,CAA	CGA,AGA,GGC,---,GGA	d	d	4	cDIV
KIF27	1426	GCA,ATG,GAC,GTA	GAA,GGG,AAG,---,GAG	d	d	4	cDIV
LARP1B	1177	CAT,AGT,TGT	AGG,GGC,---,GGA,GGT	d	d	4	cDIV
LPO	859	GAT,GAC,CAG,CAC	AAT,---,AAC	d	d	4	cDIV
LRRC8A	274	TCG,TCC,ATA,ATG	GCG,GCC,GCA,ACG,ACA,ACC	d	d	4	cDIV
LRRN4	1423	CAT,CAC	TAT,TCG,CGC,TAC,TTT	d	d	4	cDIV
LYVE1	286	ACA,ATA,CAA	AAA,AAG,GTA,TAT	d	d	4	cDIV

Symbol	Pos_codon	Codon_targets	Codon_others	Type_CodonSub_target	Type_CodonSub_other	Type_CodonSub	EvoDir_Codon
LZTFL1	463	CAT,TCT,TAT	AGA,CGT,TGC,---,TGT	d	d	4	cDIV
MEI4	763	GGC,ACC,AGG	AGT,AGC,AAC	d	d	4	cDIV
MFSD4B	727	TGG,TCT	TAC,CAT,TAT,TGT	d	d	4	cDIV
MTFR1	307	ACA,ACC	GCC,GGC,CCC,GCT,---	d	d	4	cDIV
MUM1	367	TTC,TTT,CCT	CTA,CTG,TTA,TTG	d	d	4	cDIV
NBN	1219	ATA,---,CTG	GCA,GTG,GTA	d	d	4	cDIV
NDC1	1360	ATC,AAG	AGT,AAT,AGC,AAC	d	d	4	cDIV
NOLC1	1021	CTG,CCA,CCG	CAA,CAG,---	d	d	4	cDIV
OTOA	2575	GGT,GCT	TCT,ACG,---,TTT,ACT	d	d	4	cDIV
OTUD3	334	GCC,TCC	ACC,---,ACG	d	d	4	cDIV
PAG1	145	AAT,TAT	CAT,---,CAG,CAC	d	d	4	cDIV
PDZD8	1444	GCA,ACA	CAA,CCA,TCA,CCT,CCG	d	d	4	cDIV
PHACTR2	1369	GGG,---	GCG,TCG,GCA,GTG,GCT,ACA,GAG	d	d	4	cDIV
PIK3R4	2011	TGT	CGC,CGT,CGG	i	d	2	cCON
PLEKHO1	685	CTC,ACC	GCC,GTC,ATC	d	d	4	cDIV
PRKAR2B	94	GTA	ATA,---	i	d	2	cCON
PTPRB	2740	ATG,AAG	GCG,GCA,GTG,ACG,GCT	d	d	4	cDIV

Symbol	Pos_codon	Codon_targets	Codon_others	Type_CodonSub_target	Type_CodonSub_other	Type_CodonSub	EvoDir_Codon
REST	1849	TTT,TAT	TCT,CAT,TGC,TGT,GGT	d	d	4	cDIV
REXO1	1603	TCC,CCC,TCA	GCG,GCC,GCA,GTG,GCT,ACA,---,GGA	d	d	4	cDIV
RIOK1	1519	GTG,GCG,AAT,TTT	GAA,AAG,CAG,---,GAC,GAG	d	d	4	cDIV
RPAP1	2236	AGT,---,TAT	CTT,CAA,CAT,CGT,CAG,GCT,TGT,GAT,CAC	d	d	4	cDIV
SACS	6760	AAT,ATT	ACC,ACT,GCT	d	d	4	cDIV
SCAMP2	316	GAT,GAC	AAT,---,AAC	d	d	4	cDIV
SERPINB6	331	TCC,ATC	TTC,GCC,CTG,CTC,---,TTA,TTG	d	d	4	sDIV
SESN1	376	ACG	GCG,GCA,ATG,GTG,---	i	d	2	cCon
SETD4	55	AAG,CGG	CAA,CAG,---	d	d	4	sDIV
SH3BP2	649	TCC,CCC	GCG,GCC,GCA,GGC,ACC,ACT,AAC	d	d	4	cDIV
SMCR8	643	CAA,GTA,CTG,GTG,CAG	ATA,GCG,ATT,ATC	d	d	4	cDIV
SMPD3	919	TGC	TAC,---,TAT	i	d	2	cCon
SPAG16	370	CTG,ACG,AAG	ATG,GTG,AGG	d	d	4	cDIV
SPART	1225	CTA,ATG,AAA,TTA,TTT,ATT	GTA,GTT,TAT,TAA,CAT,---,TGT,TAC	d	d	4	cDIV
SPG11	5626	GCA,CTG,CCC,CCG	TCT,CAA,TCC,TGC,---	d	d	4	cDIV
SYNJ2	1312	CAA,CAG,CGT	CAT,CAC	d	d	4	cDIV
TANC1	4855	GTG	GCG,CCA,AAG,ATG,CTG,ATC,ACG,ACA,---,CCG	i	d	2	cCON
TASOR	2080	AGT,GGT,CCT	GCT,ACA,---,ACC,ACT	d	d	4	cDIV

Symbol	Pos_codon	Codon_targets	Codon_others	Type_CodonSub_target	Type_CodonSub_other	Type_CodonSub	EvoDir_Codon
TCOF1	835	GTT,GCA,GTG,TCA,---	CCA,CTA,CTG,CAG,CCG	d	d	4	cDIV
TCTE3	238	AAA,AAT,GAG	GAT,CAT,GGC,GGT	d	d	4	cDIV
TDP2	802	GAA	AAA,AGA,CAA,ACA	i	d	2	cCon
TDRD9	2365	ATG,GTG,GTT	CTG,TTA,---,TTG	d	d	4	cDIV
TDRD9	2950	CAT,AAG,CAC	AGT,---,GAT,AAT,AAC	d	d	4	cDIV
TICRR	982	GCA,AAT	GTA,ATG,GTG,CTG,ACG	d	d	4	cDIV
TMEM209	538	TCG,TCT,---	GTT,GCC,GCA,CCC,ACG,GCT,CCT,ACT,AAC	d	d	4	cDIV
TNFRSF1A	751	CAT,AAA,CGT,TAT	AAT,ATC	d	d	4	cDIV
TNS3	2851	GAA,TCT,TCC,CCC,CCG	CTT,GCG,GTT,GCC,AGA,GCT,AAA,GAC,---,TTT,ATC	d	d	4	cDIV
TP53I3	1060	GCC,GAC,CAG	CTG,CGC,AAG,GAG	d	d	4	cDIV
TPCN2	193	GAA,AAC,---,GAG	AGG,AAG,AAA,ACA,ACC	d	d	4	cDIV
TPCN2	340	AAA,GCA,---	GAA	d	i	3	cDIV
TRAFD1	1333	CCC,GGT,ACC	GAA,GCC,GCA,GTG,GCT,GAC,---	d	d	4	cDIV
TREM2	616	AGT,GGT,GGA,GAT	CAT,AAT,---,AAC	d	d	4	cDIV
TSEN2	742	ATG,AGC	GCC,AGA,GGC,GCT,ATA,ACC,GTC,ATC	d	d	4	cDIV
TTC37	2254	GCC,ACC,GAA	GGG,GGC,AGC,GAC,GGA,---,GAT,AAC	d	d	4	cDIV
URB2	316	AAA,AAG	GAA,CAA,GCA,GGA,GAG	d	d	4	cDIV
USP4	787	AGA,GCA,TCA,GCT	GTA,ATG,ACA,ATA,---,GTC,ATT	d	d	4	cDIV

Symbol	Pos_codon	Codon_targets	Codon_others	Type_CodonSub_target	Type_CodonSub_other	Type_CodonSub	EvoDir_Codon
WDR77	850	GGC,CCC,AAT	AGG,AGT,AGA,CGT,AGC,CGC	d	d	4	cDIV
WDR78	670	GCA,TTC,GAG	TCG,TCA,CCA,---	d	d	4	sDIV
XPC	1300	CCT	CGG,CAT,CGT,CGC,TGC,TGT	i	d	2	cCON
ZBTB49	574	GGG,GGA	GTA,GTT,GCC,ATG,GCA,GTG,ATA,TTA	d	d	4	cDIV
ZC3H6	3370	GCC,TCC,CCC	ATT,---,ACC,AAT,ACT,AAC	d	d	4	cDIV
ZDHHC1	1363	AGT,ATG,TTT,TCC	---,GAC,TTG,ACC,GTC,ATT,ATC	d	d	4	cDIV

Table 5.1. Avian vocal learner-specific single amino acid variants and its supporting evidence. Partial data sheet with positive selection.

Symbol	Pos_AA	AA_targets	AA_others	[PAML] Adjusted *p*-value	[PAML] Likelihood Ratio Test	[PAML] dN/DS of MRCA branches	[PAML] Posterior Probability (BEB)
ABCG2	330	L,T,A	E,R,I,K,V				
ACP2	92	D,Q	E,G,A				
AKAP1	55	R,S,Q	D,H,K,N,-,G				
ALKBH4	195	W,S	C,L,Y,A,P,F				
ALPK1	675	C,V,N	D,A,-,G,S				
ANKFN1	234	V,A	G,L,M,-				
ANKRD16	234	G,S,A	D,-,N				
ARMC6	53	P,-	Q	1.19.E-03	13.2	588.8	0.999**
ASH1L	61	D,S	R,-,N	3.24.E-01	0.8	1.7	0.991**
B3GNT2	253	N	H	2.75.E-01	1.1	3.3	0.999**
BIRC7	11	P,M	I,L,A,S,-,T,V				
BMP2K	334	-,A	P,T,N	4.23.E-01	0.3	1.9	0.971*
BRCA2	422	R,K,-,G,S	D,A				
BRIP1	913	G,T	I,A,-,F,V				
C12orf35	170	P,H,Y,G	E,K,Q				
C12orf35	604	R,-	E,G				
C12orf55	315	D,W,-	C,H,Y,F,S,Q	4.19.E-02	5.3	3.1	0.512
C3orf67	426	D,H,R,N,T	L,I,K,-,P,V				
C8B	489	L	F,-,V	2.76.E-01	1.1	2.1	0.954*
C9ORF152	53	F,M,T,V	L,K				
CCDC13	216	P,K,A	R,T,-	1.05.E-01	3.2	30.2	0.502
CCDC69	43	E,D,S	H,K,N,-,T	5.00.E-01	0.0	1.3	0.510
CD86	91	P,N,T,V	E,D,H,K				
CDCA7	256	L,S	A,-,P,T,V				
CFAP70	469	G,R,H	I,A,-,F,T,M,V				
CFLAR	391	R,L,S	I,T,-,V	6.70.E-02	4.1	4.3	0.742
CHGB	252	G,A	E,D,N				
CLBA1	241	G,-,A	D,N				
CLUL1	231	I,-,V	D,H,A				
COL6A3	171	L,T	I,M,-	7.58.E-13	56.7	569.4	0.923
COL6A3	185	R,N,Q	K,-	7.58.E-13	56.7	569.4	0.790
COL6A3	1526	L,M	V	7.58.E-13	56.7	569.4	0.830
CXorf21	173	P,T,A	C,F,-,S				

Symbol	Pos_AA	AA_targets	AA_others	[PAML] Adjusted *p*-value	[PAML] Likelihood Ratio Test	[PAML] dN/DS of MRCA branches	[PAML] Posterior Probability (BEB)
DERA	252	Y	H,I,N,-,F				
DLGAP5	533	K,S,A	E,D,-,G	3.34.E-10	44.3	5.9	1.000**
DNAH10	330	I,T,V	L,-,S				
DNAH10	1297	N,A	D,H,Q				
DNAH10	2268	N,Q	C,R,H,K,Y,T				
DNAH10	3382	H,T	S,-,N				
DNAH3	2947	Q	E,R,K,N,-				
DRC7	627	D,I,V	R,K,A,T,M,S	2.06.E-04	17.0	3.8	0.714
DRC7	657	G,D,H	E,R,K,N,S				
DRD5	416	A	I,V	2.74.E-01	1.2	3.7	0.500
E2F8	463	G,E,A	H,N,-,V,T,S				
EFHB	87	P,T,S	C,R,H,Y,-,G,Q	1.67.E-05	22.3	4.9	0.963*
EFHB	245	P,E,K,N	G,-,A	1.67.E-05	22.3	4.9	0.983*
EFHB	431	F,Y,S	E,L,I,A,-,G,V	1.67.E-05	22.3	4.9	0.795
EFHC1	77	E,V	P,T,-,A	6.70.E-02	4.0	2.6	0.980*
ENPP1	723	P,D,S	T,V,A				
ENSGALT00000010226	413	K,S	R,H,Y,N,-				
ENSGALT00000012528	233	D,S,Q	H,K,N	6.32.E-02	4.4	6.4	0.986*
ENSGALT00000015652	665	V,A	L,M,-,I	7.45.E-03	9.2	8.8	0.672
ENSGALT00000017732	652	Z,H,Y	L,N,W,-,T,S,Q				
ENSGALT00000025242	99	E,-,S	G,D	8.56.E-02	3.6	7.1	0.968*
ENSGALT00000027531	902	P,E,I	R,K,A,T,M,S,Q				
ENSGALT00000030336	164	T,V	P,G,A	1.63.E-02	7.6	5.3	0.982*
ENSGALT00000032705	82	K,N,Q	R,I,D,-,G,S				
ENSGALT00000032989	246	R,H,L	K,Q	2.50.E-02	6.6	11.1	0.932
ENSGALT00000036845	48	S	G,R,K,E				

Symbol	Pos_AA	AA_targets	AA_others	[PAML] Adjusted *p*-value	[PAML] Likelihood Ratio Test	[PAML] dN/DS of MRCA branches	[PAML] Posterior Probability (BEB)
ENSGALT00000037492	133	I,F	L,S,-,P,V	8.17.E-04	14.1	43.2	0.825
ENSGALT00000039593	87	W,N	E,R,K,-,P,Z,Q				
FANCI	415	N,A	I,T,-	4.93.E-02	4.9	4.9	0.943
FBXO48	59	S	-,A				
FGFBP1	166	M	I,K,N,A,-,V,T,S				
FNDC1	1034	S	G	6.37.E-02	4.3	6.7	0.981*
GDPD4	106	L,V	I,-	3.30.E-01	0.8	1.3	0.999**
GDPD4	408	G,S	D,N	3.30.E-01	0.8	1.3	0.961*
GDPD4	436	E,H,S	G,D,N				
GPATCH1	429	E,D,-	G,V,S,A				
GPLD1	280	D,-,S	G				
GPR35	157	D,-,Q	E,K,V				
HAUS8	63	G,N	E,D,S				
HEATR6	400	C,P,M,-	N,T,S,A	2.73.E-47	217.1	53.3	0.630
HEPH	1098	L,A	H,Y,-,P,S				
HMGXB3	269	D	E,-	2.97.E-01	1.0	2.3	0.995**
IBA57	267	P	L,A,-,G,S				
IBSP	259	N	D,A,S,-,G,V				
IFT88	299	E,L,Y	H,-,Q	3.51.E-03	10.9	3.1	0.993**
INPP5E	160	G,S	T,-,A				
ITFG3	311	D,-,Q	E,R,G				
KCNS3	490	M,V,A	I,T,S	2.75.E-01	1.2	11.5	0.885
KIAA0391	306	L,S,A	P,T,-				
KIAA1841	479	E,Q	G,R,-				
KIF27	476	D,M,V,A	E,K,-,G				
LARP1B	393	C,H,S	G,R,-				
LPO	287	D,H,Q	-,N	3.55.E-01	0.6	1.2	0.995**
LRRC8A	92	I,M,S	T,A				
LRRN4	475	H	R,F,Y,S				
LYVE1	96	I,T,Q	K,Y,V	2.21.E-01	1.6	6.0	0.648
LZTFL1	155	H,Y,S	C,R,-				
MEI4	255	G,R,T	N,S	4.08.E-02	5.4	5.8	0.952*
MFSD4B	243	W,S	C,H,Y	2.04.E-01	1.8	6.6	0.929

Symbol	Pos_AA	AA_targets	AA_others	[PAML] Adjusted *p*-value	[PAML] Likelihood Ratio Test	[PAML] dN/DS of MRCA branches	[PAML] Posterior Probability (BEB)
MTFR1	103	T	P,-,G,A	1.01.E-01	3.2	10.1	0.521
MUM1	123	P,F	L	1.97.E-02	7.1	4.0	0.998**
NBN	407	L,I,-	V,A	1.46.E-02	7.9	10.0	0.947
NDC1	454	I,K	S,N				
NOLC1	341	P,L	-,Q	2.86.E-03	11.4	7.3	0.983*
OTOA	859	G,A	F,T,-,S				
OTUD3	112	S,A	T,-	1.13.E-01	3.0	3.0	0.985*
PAG1	49	Y,N	H,-,Q	2.08.E-02	7.0	7.5	0.932
PDZD8	482	T,A	P,S,Q				
PHACTR2	457	G,-	E,A,V,T,S				
PIK3R4	671	C	R	4.93.E-02	4.9	10.4	0.997**
PLEKHO1	229	L,T	I,V,A				
PRKAR2B	32	V	I,-	3.57.E-06	25.4	295.8	0.999**
PTPRB	914	K,M	T,V,A				
REST	617	F,Y	C,H,G,S				
REXO1	535	P,S	A,-,T,G,V				
RIOK1	507	V,F,N,A	E,D,K,-,Q	3.61.E-03	10.8	22.0	0.713
RPAP1	746	Y,-,S	C,R,L,H,D,A,Q				
SACS	2254	I,N	T,A	5.00.E-01	0.1	1.3	0.998**
SCAMP2	106	D	-,N	6.70.E-02	4.1	3.7	0.998**
SERPINB6	111	I,S	L,F,-,A				
SESN1	126	T	M,-,V,A				
SETD4	19	R,K	-,Q	1.78.E-01	2.1	3.9	0.931
SH3BP2	217	P,S	G,T,N,A				
SMCR8	215	L,V,Q	I,A				
SMPD3	307	C	Y,-	3.37.E-02	6.0	14.3	0.994**
SPAG16	124	L,K,T	R,M,V	3.01.E-01	1.0	4.1	0.889
SPART	409	L,I,K,F,M	C,H,Y,-,Z,V				
SPG11	1876	P,L,A	C,-,S,Q				
SYNJ2	438	R,Q	H				
TANC1	1619	V	L,I,K,A,-,P,T,M				
TASOR	694	G,P,S	T,-,A	1.56.E-01	2.4	2.3	0.930
TCOF1	279	S,-,V,A	P,L,Q	1.97.E-02	7.2	8.6	0.863
TCTE3	80	E,K,N	G,D,H	2.21.E-01	1.7	2.1	0.527

Symbol	Pos_AA	AA_targets	AA_others	[PAML] Adjusted *p*-value	[PAML] Likelihood Ratio Test	[PAML] dN/DS of MRCA branches	[PAML] Posterior Probability (BEB)
TDP2	268	E	R,K,T,Q				
TDRD9	789	M,V	L,-	2.43.E-01	1.5	2.2	0.924
TDRD9	984	H,K	S,D,-,N				
TICRR	328	N,A	L,M,T,V				
TMEM209	180	-,S	N,A,P,T,V				
TNFRSF1A	251	R,H,K, Y	I,N	3.93.E-02	5.6	35.9	0.930
TNS3	951	P,E,S	D,L,R,I,K,A,- ,F,V				
TP53I3	354	D,Q,A	E,R,L,K				
TPCN2	65	E,-,N	R,K,T				
TPCN2	114	K,-,A	E	6.52.E-02	4.2	6.7	0.982*
TRAFD1	445	G,P,T	E,D,A,-,V	3.20.E-01	0.9	1.6	0.902
TREM2	206	G,D,S	H,-,N	5.00.E-01	0.1	1.3	0.791
TSEN2	248	M,S	R,I,A,T,G,V				
TTC37	752	E,T,A	D,N,-,G,S				
URB2	106	K	E,A,G,Q				
USP4	263	R,S,A	I,-,T,M,V				
WDR77	284	G,P,N	R,S				
WDR78	224	E,F,A	P,-,S	7.45.E-03	9.2	19.2	0.959*
XPC	434	P	C,R,H				
ZBTB49	192	G	L,I,A,M,V				
ZC3H6	1124	P,S,A	I,T,-,N				
ZDHHC1	455	F,M,S	D,I,L,-,T,V				

Table 5.1. Avian vocal learner-specific single amino acid variants and its supporting evidence. Partial data sheet with specialized gene expression.

Symbol	DEG2014 (NUC vs other NUCs)	DEG2019 (NUC vs SUR)	DEG2020 (NUC vs SUR)	DEG2020 (NUC vs other NUCs	DEG2020 (SUR vs other SURs)	Singing 2014
ABCG2						
ACP2						
AKAP1						
ALKBH4						
ALPK1						
ANKFN1			HVC Down, LMAN Down	AreaX Up, LMAN Down	RA Down	
ANKRD16						
ARMC6				AreaX Down		1
ASH1L						1
B3GNT2	Ax Up	RA Down		AreaX Up	AreaX Up	1
BIRC7			LMAN Down			
BMP2K						
BRCA2						
BRIP1						
C12orf35	Ax Up					
C12orf55						
C3orf67						
C8B						
C9ORF152						
CCDC13						
CCDC69						
CD86						
CDCA7						
CFAP70				AreaX Up	AreaX Up, RA Down	
CFLAR						
CHGB	Ax Down	AreaX Down, HVC Up, LMAN Down	HVC Up, LMAN Down	AreaX Down, HVC Up	AreaX Down	1
CLBA1						
CLUL1			LMAN Up	LMAN Up		

Symbol	DEG2014 (NUC vs other NUCs)	DEG2019 (NUC vs SUR)	DEG2020 (NUC vs SUR)	DEG2020 (NUC vs other NUCs	DEG2020 (SUR vs other SURs)	Singing 2014
COL6A3	HVC_RA Up				RA Up	1
CXorf21						
DERA			HVC Up, LMAN Up			
DLGAP5						
DNAH10						
DNAH3						
DRC7						1
DRD5	Ax Up	AreaX Up	LMAN Down	AreaX Up, LMAN Down	AreaX Up	
E2F8						
EFHB						
EFHC1						
ENPP1		RA Up	LMAN Up, RA Up	AreaX Down	AreaX Down	
ENSGALT00000010226		AreaX Up, LMAN Up				
ENSGALT00000012528		AreaX Up				
ENSGALT00000015652						
ENSGALT00000017732						
ENSGALT00000025242		RA Up				
ENSGALT00000027531						
ENSGALT00000030336						
ENSGALT00000032705						
ENSGALT00000032989						
ENSGALT00000036845						

Symbol	DEG2014 (NUC vs other NUCs)	DEG2019 (NUC vs SUR)	DEG2020 (NUC vs SUR)	DEG2020 (NUC vs other NUCs	DEG2020 (SUR vs other SURs)	Singing 2014
ENSGALT00000037492						
ENSGALT00000039593						
FANCI		RA Up				
FBXO48						
FGFBP1						
FNDC1		RA Up		RA Up		
GDPD4		LMAN Up	LMAN Up	LMAN Up	RA Down	
GPATCH1				AreaX Up		
GPLD1			LMAN Up			
GPR35						
HAUS8			LMAN Down			
HEATR6						
HEPH						
HMGXB3	Ax Up			AreaX Up		
IBA57				AreaX Down		
IBSP				AreaX Down		
IFT88						
INPP5E	Ax Up		LMAN Up			
ITFG3	Ra Up					
KCNS3					AreaX Down	
KIAA0391						
KIAA1841		LMAN Up	LMAN Up			
KIF27			LMAN Down	AreaX Up		
LARP1B						
LPO						
LRRC8A			LMAN Up			
LRRN4						
LYVE1				RA Up		
LZTFL1						
MEI4			AreaX Up	HVC Down		

Symbol	DEG2014 (NUC vs other NUCs)	DEG2019 (NUC vs SUR)	DEG2020 (NUC vs SUR)	DEG2020 (NUC vs other NUCs	DEG2020 (SUR vs other SURs)	Singing 2014
MFSD4B	Ax Up					
MTFR1			LMAN Up	AreaX Down		
MUM1		AreaX Down				
NBN				AreaX Up		
NDC1						
NOLC1						
OTOA						
OTUD3			LMAN Up			
PAG1			LMAN Up			
PDZD8				AreaX Up		
PHACTR2	HVC LMAN Up	HVC Down	LMAN Down			
PIK3R4	Ax Up			AreaX Up		
PLEKHO1		LMAN Down	LMAN Down			
PRKAR2B	Ra Down			AreaX Up, RA Down	AreaX Up	
PTPRB						
REST				AreaX Up		
REXO1						
RIOK1						
RPAP1			LMAN Down	AreaX Up		
SACS	Ax Down		HVC Up, LMAN Up		RA Up	
SCAMP2						
SERPINB6						
SESN1		HVC Down				
SETD4						
SH3BP2						
SMCR8						
SMPD3	Ax Up			AreaX Up, RA Down	AreaX Up, RA Down	
SPAG16						
SPART						

Symbol	DEG2014 (NUC vs other NUCs)	DEG2019 (NUC vs SUR)	DEG2020 (NUC vs SUR)	DEG2020 (NUC vs other NUCs	DEG2020 (SUR vs other SURs)	Singing 2014
SPG11	Ax Up					
SYNJ2						
TANC1			HVC Up			
TASOR				AreaX Up		
TCOF1	Ax Up		LMAN Down			
TCTE3						
TDP2						
TDRD9						
TICRR				AreaX Up	RA Down	
TMEM209						
TNFRSF1A	HVC LMAN Up	HVC Up	HVC Up			
TNS3	Ax Up		LMAN Down	AreaX Up	AreaX Up	1
TP53I3						
TPCN2						
TRAFD1						
TREM2				AreaX Down		
TSEN2	Ax Up					1
TTC37			HVC Up, LMAN Up	RA Down		
URB2						
USP4		AreaX Up	LMAN Up			1
WDR77						
WDR78		HVC Up				
XPC				AreaX Up		
ZBTB49				AreaX Up		
ZC3H6	Ax Up			AreaX Up		
ZDHHC1						

In parallel, to define the VL-SAVs as molecular convergences and divergences by estimating those evolutionary directions from ancestral states to existing species, I performed ancestral sequence reconstructions with the RAxML [115]. Based on amino acid changes from the most recent common ancestors (MRCA) of each vocal learning clade to 6 terminal nodes in birds, the 24 iSAVs and 124 dSAVs were classified as 24 convergent SAVs (ConSAVs; 1 simple parallel and 23 complex convergent SAVs) and 124 divergent SAVs (DivSAVs; 4 simple divergent and 120 complex divergent SAVs), respectively (**Table 5.1**). As an example of vocal learner-specific amino acid convergences (**Figure 5.4A**), the 103rd residue of *MTFR1* was a type 2 SAV site with an identical amino acid variants as threonine (T) of vocal learners substituted from another amino acid (alanine, A) which was estimated as MRCA sequences of three independent lineages of vocal learners. Distinguished from a narrow meaning of 'convergent substitutions' that the same descendent amino acid independently originated from different ancestral amino acids, the iSAV with threonine (T) could be defined as 'parallel substitutions' from same ancestral amino acid (A). To simplify this issue like the previous studies [19,116], I called the both types of variants as molecular convergences (**Table 5.1**). As another example of amino acid divergences (**Figure 5.4B**), the 224th site of *WDR78* was a type 4 SAV site with different SI as glutamic acid, phenylalanine, and alanine (E, F, and A) observed in vocal learning birds and with different SI as proline, serine and deletion (P, S, and '-') in vocal non-learning birds. These different amino acids (E, F, and A) were divergently substituted from the amino acid (S) estimated as MRCA sequences of vocal learning clades.

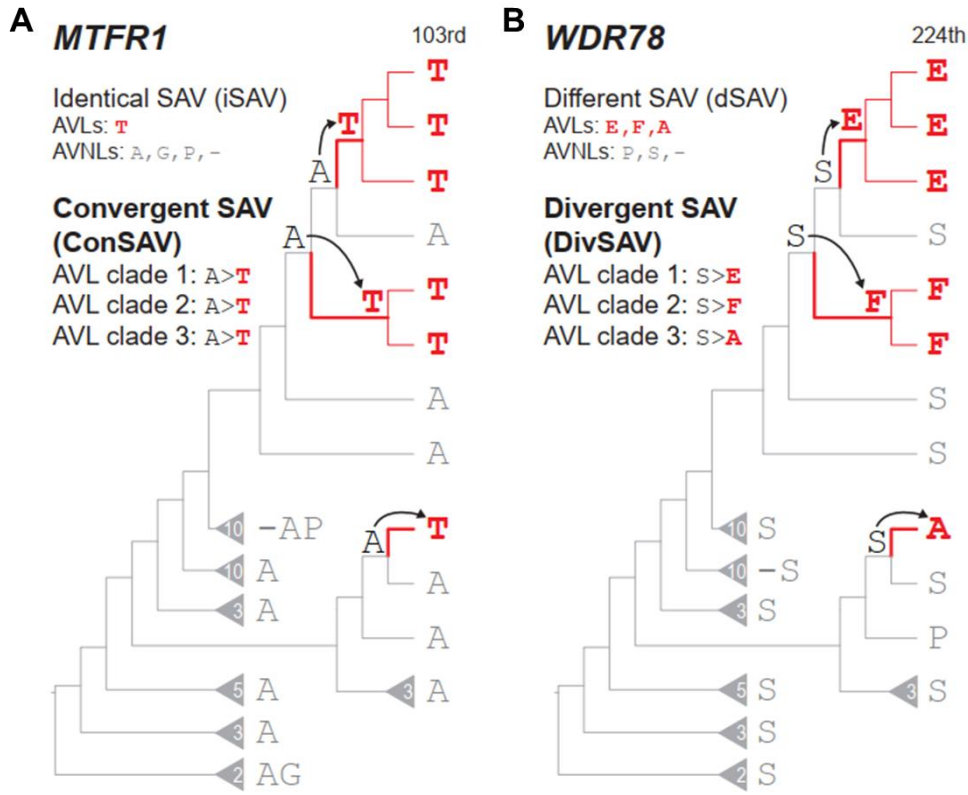

Figure 5.4. Examples of convergent and divergent single amino acid variants (ConSAVs and DivSAVs). (**A**) Example gene with amino acid convergences, *MTFR1*. (**B**) Example gene with amino acid divergences, *WDR78*. The most recent common ancestral (MRCA=origin) branches marked as bold red lines. Black arrows indicate the ancestral substitutions at each origin branch of vocal learning clades. Black, red, and grey characters indicate amino acid sequences of ancestors of each MRCA node of vocal learning clades, amino acid sequences of vocal learning clades, and amino acid sequences of vocal non-learning birds, respectively. Grey triangles and the numebers on them indicate clades of vocal non-learning birds and the number of species in each clade.

Avian vocal learners did not show a preponderance of amino acid convergences

I 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, I designed 2 types of clade-specific control sets (**Figure 5.5**): 1) all controls consisting of 8,238 different species combinations with 6 target species from 3 independent lineages without considering any traits; 2) Out of these 8,238 control sets, 59 core controls consisting of all possible sets given the phylogeny with 6 target species having at least 2 vocal learning clades and 1 non-learning clade. Out of 59 core control sets, the closest control set for the set of vocal learning birds included songbirds and parrots as 2 vocal learning clades and swift as a vocal non-learning clade which is a close relative species of hummingbirds. I conducted the ConVarFinder analysis for this total of 8,238 control species combinations, and identified various SAVs of each control set.

As an extension to the previous studies on convergent evolution in reptile and mammalian lineages [19,96,97] that tested pair-wise combinations of two species, I found strong correlations between amino acid convergences and divergences tested in higher dimensional combinations of species (**Figure 5.1D**). Although higher than the expectation according to the regression with all control sets and core control sets, the number of convergent substitutions in vocal learning birds was not an outlier (adjusted $p > 0.05$, Bonferroni Outlier Test [103]) from the trend observed in the control sets. Several outliers did exist among the control sets, with the highest residual being 32.46 in one of the all control sets (4 Passeriformes, budgerigar, and falcon), and 17.61 in a core control set (3 songbirds, Anna's hummingbird, and 2 land fowls; **Figure 5.1D**). These species combinations of 2 control sets with the highest residuals, however, do not share any known convergent traits as far as I are aware. 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 convergent traits.

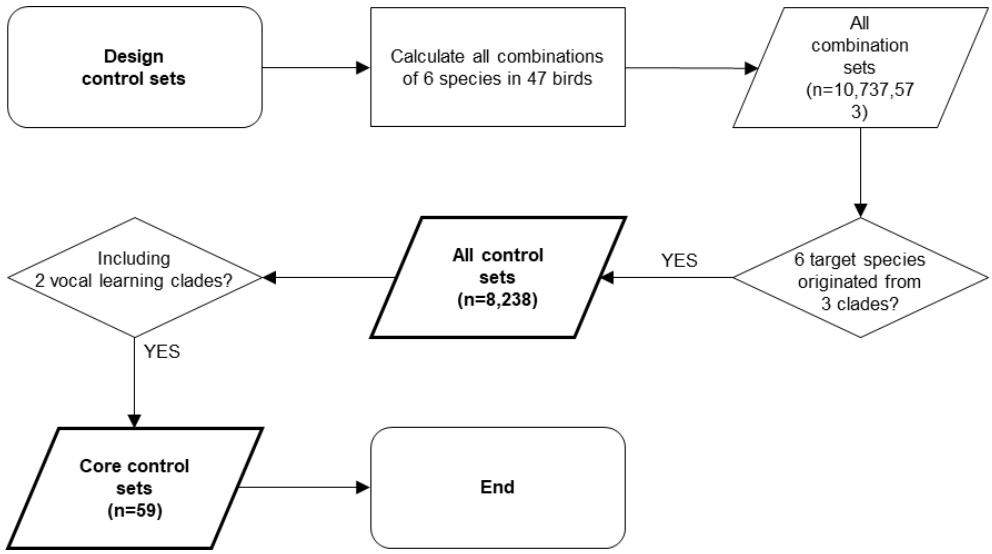

Figure 5.5. Flow chart to design control sets.

Amino acid convergences are associated with product of origin branch lengths in birds

I sought a measure of molecular convergence that controls for phylogenetic relationships. According to previous studies on mammalian or drosophila genomes [116], and vertebrate mitochondrial genomes [97], 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, I took into consideration for 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 (**Figure 5.6A**). I observed strong and significant correlations between ConvSAVs and the product of MRCA branch length lengths (POB) for both all control sets and core control sets (**Figure 5.6B**). The correlation was also observed for both SAVs and DivSAVs (**Figure 5.6C, D**). Much weaker correlations of three types of SAVs were observed with the product of terminal branches (PTB), distances among terminal branches (DTB) and terminal nodes (DTN) than that of POB (**Figure 5.6**). Like the ConSAV versus DivSAV correlation analyses (**Figure 5.1D**), the avian vocal learners were not a significant outlier relative to all and core control sets in correlations between three types of SAVs and POB (**Figure 5.6**). 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 the frequencies of amino acid convergences of vocal learning birds are under this trend relative to other species combinations.

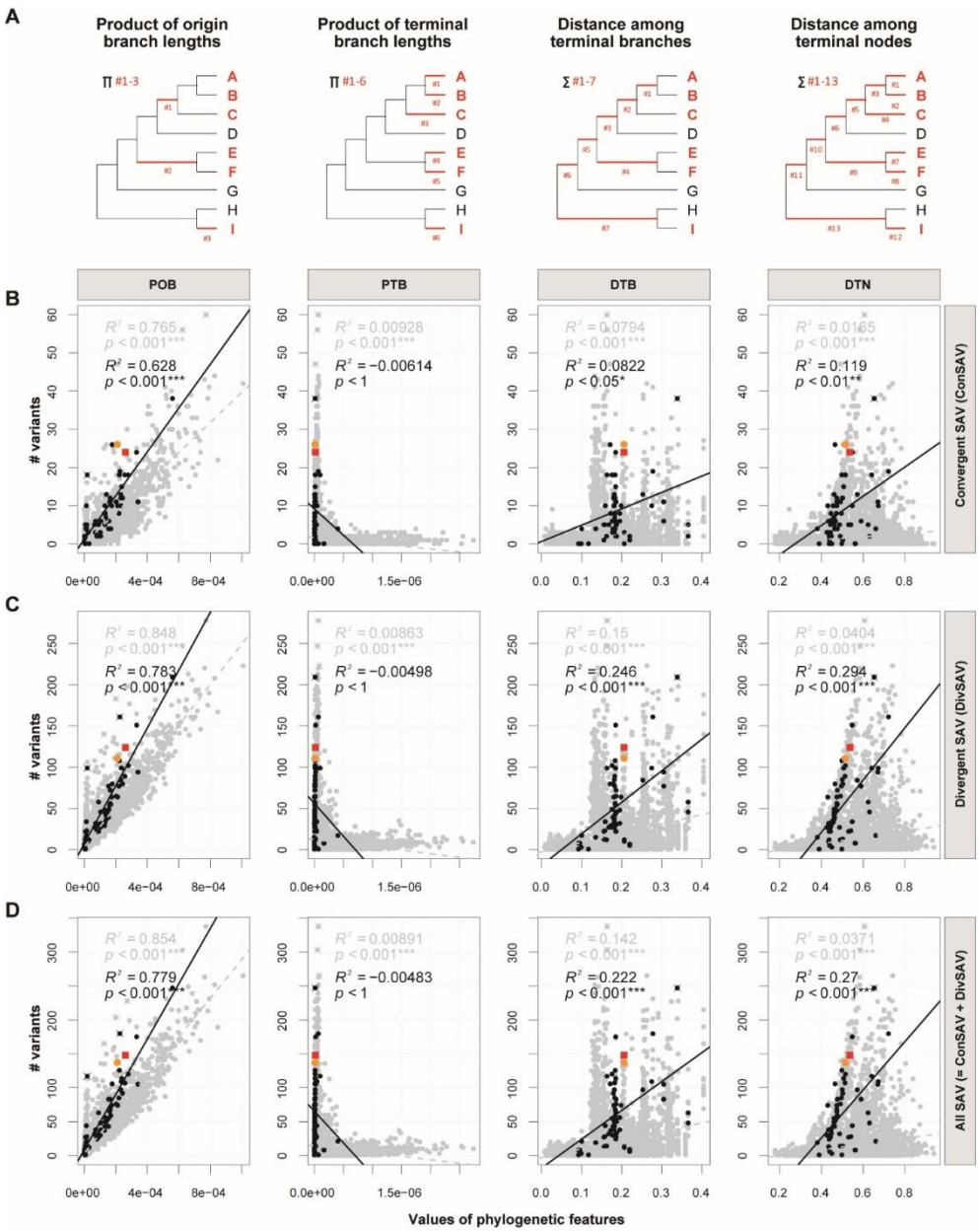

Figure 5.6. Amino acid convergence amount is positively correlated to the product of origin branch lengths. (**A**) Four types of phylogenetic tree features: product of origin branch lengths (POB); product of terminal branch lengths (PTB); distance among terminal branches (DTB); and distance among terminal nodes (DTN). In the example type of tree branches, red lines show the branches used for the calculations and red texts show the species clades that have a convergent trait. (**B-D**) Shown are regression analyses of three types of single amino acid variants

(SAVs) for convergent SAVs (ConSAVs), divergent SAVs (DivSAVs), and total SAVs (ConSAVs + DivSAVs) in the vocal learning set and control sets of avian species with each type of phylogenetic features.

Amino acid convergences can arise from complex molecular sources at codon and nucleotide levels

To investigate what types of codon and nucleotide variants can cause SAVs, I modified the algorithm for amino acid variants to detect single codon variants (SCVs) made up of 3-nucleotides and single nucleotide variants (SNVs) in those codons (**Figure 5.7A, B**). Similar to the SAV analysis, the codon and nucleotide variants were classified into convergences and divergences (**Figure 5.7A, B**). Theoretically, an amino acid variant can originate from single nucleotide variants at a homologous codon position (SNV) or complex multiple nucleotide variants which are not mutually exclusive between target and the other species groups (No-SNV) (**Figure 5.7C**). However, some SNVs can also give rise to no change in the amino acid, namely synonymous substitutions (**Figure 5.7D**). I checked for overlaps among variants specific to vocal learning birds and control sets 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, I found 600 SCV sites specific to avian vocal learners (AVL-SCVs) and the SCVs were fully overlapped with 148 SAVs and 165 SNVs (**Figure 5.7E**). Out of these 600 AVL-SCVs, 56 (15.7%) showed nonsynonymous SNVs and 98 (9.0%) showed complex nonsynonymous No-SNVs, resulting in 148 AVL-SAVs among vocal learners (**Figure 5.7E**). The remaining SCVs consisted of 111 (18.5%) synonymous SNVs and 341 (56.8%) complex synonymous No-SNVs (**Figure 5.7E**). An example of a AVL-SAV caused by a nonsynonymous SNV is in the 253rd codon site of *B3GNT2*, where all vocal learners had the same convergent nucleotide (A), codon (AAT), and amino acid (Asparagine, N) sequence mutually exclusive to all vocal non-learners (e.g. C; CAT or CAC; and Histidine, H; **Figure 5.7F**). An example of a AVL-SAV caused by complex synonymous No-SAVs is the 475th site of *LRRN4*, where all of vocal learners showed amino acid convergence (AVL-ConSAV) to Histidine (H), while their divergent codons consist of CAC or CAT for vocal learners with non-exclusive nucleotide variants for vocal non-learners (**Figure 5.7F**). In the all and core control species sets with at least one SCVs (n=8,109 and 59, respectively),

although I found different total numbers of SCVs and their corresponding SAVs and SNVs, their relative proportions (%) were similar to each other and to that of vocal learners (**Figure 5.7E**); about 1/3 of SAVs of control sets originated from SNVs at each homologous nucleotide site, while 2/3 originated from complex non-exclusive nucleotide changes at different nucleotide sites in each codon (**Figure 5.7E**). These findings suggest that amino acid convergences originate not only from simple single nucleotide substitutions at each homologous site but also from complex nucleotide variants without any mutual exclusivity between target and the other species group.

Next, to trace evolution of molecular sources causing amino acid convergences, I checked the proportions of amino acid convergences (ConSAVs) originated from convergent or divergent variants at codon and nucleotide levels. For vocal learning birds, out of 24 amino acid convergences (AVL-ConSAV) sites, 15 (62.5%) were caused by codon convergences (ConSCVs) and 9 (37.5%) codon divergences (DivSCVs) (**Figure 5.7G**), while 17 (70.8%) were caused by nucleotide convergences (ConSNVs), 1 (4.2%) nucleotide divergence (DivSNVs), and 6 (25%) complex non-exclusive nucleotide variants (No-SNVs) (**Figure 5.7H**). For medians of all and core control sets with at least one ConSAVs (n=2,826 and 53, respectively), out of amino acid convergences of controls (Ctrl-ConSAVs), almost half and half were caused by that of codon convergences and divergences, respectively (**Figure 5.7G**), while most ConSAVs (80% and 71.4%) were caused by nucleotide convergences (ConSNV) (**Figure 5.7H**). These findings suggest that a convergent feature can emerge from its underlying variants under convergent or divergent evolution.

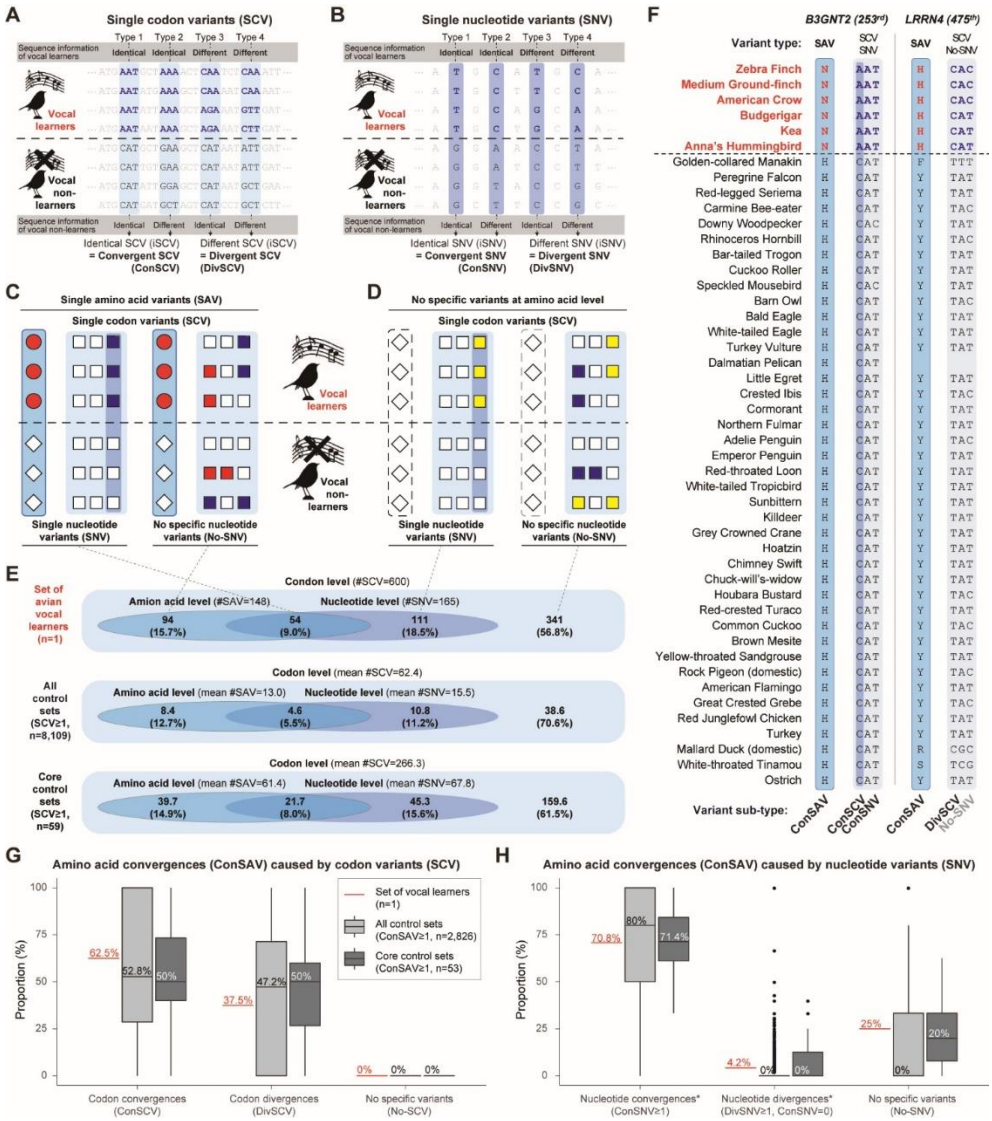

Figure 5.7. Amino acid convergences emerged from complex molecular sources at codon and nucleotide levels. (**A**) Concept of four types of single codon variants (SCV). (**B**) Concept of four types of single nucleotide variants (SNV). (**C**) Concept of convergent single amino acid variants (ConSAVs) explained by SCV. Left case, SAVs caused by SCVs with SNV at a homologous nucleotide site. Right case, ConSAVs caused by SCVs with complex non-exclusive variants at different sites at the codon (No-SNVs). (**D**) Concept of SCV explained by synonymous substitutions between species, which are those do not cause amino acid changes. Left case, synonymous SCVs with SNV at a homologous nucleotide site. Right case, SCVs with multiple nucleotide variants at different sites (No-SNVs). (**E**) Venn diagrams of

the different subsets of SCVs caused by the four types of nucleotide substitutions outlined in (**C**) and (**D**), in avian vocal learners (n=1), all control sets of species (n=8,238), and the core control sets (n=59). (**F**) Examples of identical amino acid convergences among vocal learners (ConSAVs) originating from ConSNVs at the same site (in *B3GNT2*) or No-SNVs (in *LRRN4*). Red text, avian vocal learners. Sky blue boxes, sites with SAVs; Dark sky blue box, SNV; Light sky blue boxes, SCVs.

Various types of sequence variants are best explained by the product of MRCA branch lengths

Next, I performed correlation tests between nine types of sequence variants (three types of convergences [all, convergent, and divergent] at three levels [amino acid, codon, and nucleotide]) and four types of phylogenetic features (POB, PTB, DTB, and DTN). As expected, all nine types of convergent variants (SAVs, ConSAVs, DivSAVs, SCVs, ConSCVs, DivSCVs, SNVs, ConSNVs, and DivSNVs) were highly correlated with each other, in both all control sets (**Figure 5.8**) or core control species sets (**Figure 5.9**). 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 (**Figure 5.8, 5.9**). Correlations were overall weaker in the core control sets of species, presumably due to a smaller number of species combinations than the all control sets. Like three types of SAVs compared to POB (**Figure 5.6**), the residuals of the numbers of vocal learner-specific variants at other two levels calculated from the regression line with POB values still did not exceed in both of all and core control sets (**Figure 5.8, 5.9**).

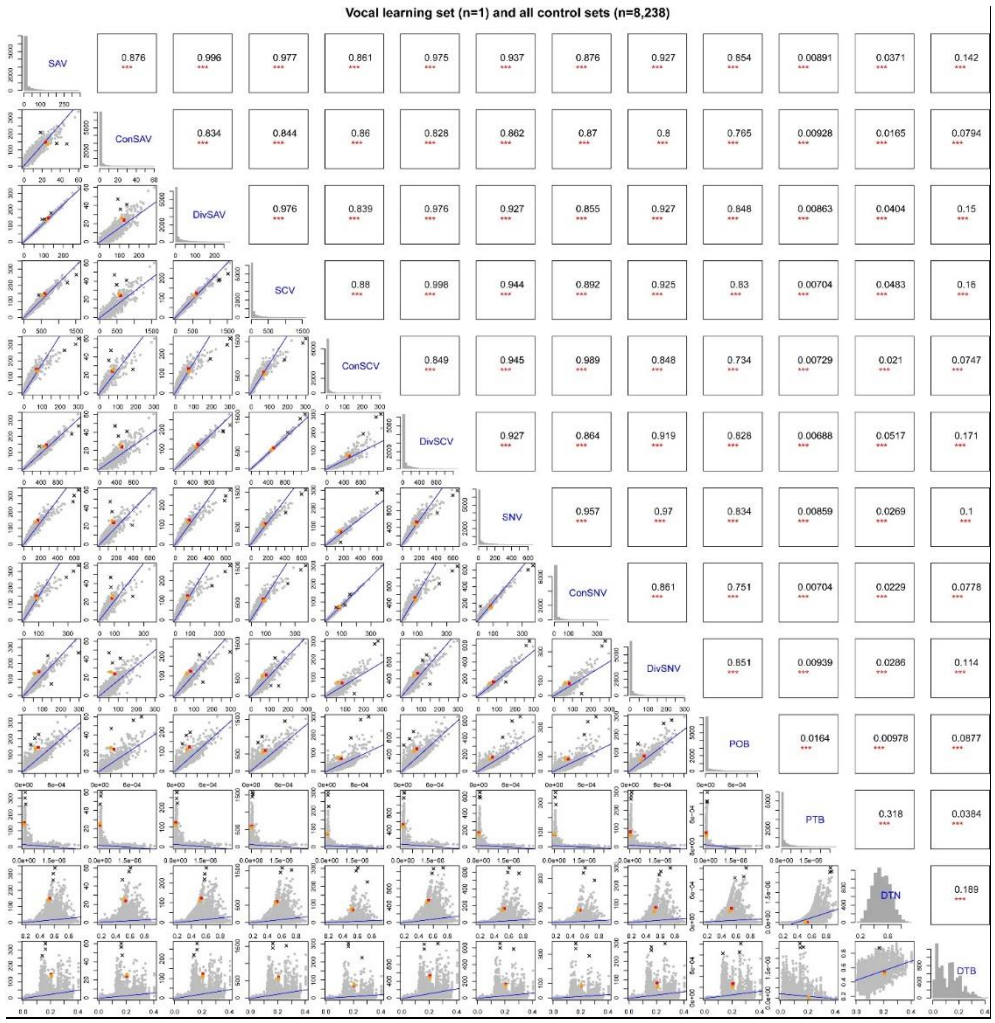

Figure 5.8. Codon and nucleotide variants are also proportional to the product of origin branch lengths in all control sets. *p* values and Adjusted R^2 of correlations are visualized at upper diagonal matrix (p<0.05*, p<0.01**, and p<0.001***). Histograms of frequencies of each convergent variant and 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, orange, and red spots indicate all control sets (n=8,237), the set of the closest control set (n=1), and the set of avian vocal learners (n=1), respectively. Black 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, SAV = single amino acid variants, ConSAV = convergent

SAV, DivSAV = divergent SAV, SCV = single codon variants, ConSCV = convergent SCV, DivSCV = divergent SCV, SNV = single nucleotide variants, ConSNV = convergent SNV, DivSNV = divergent SNV.

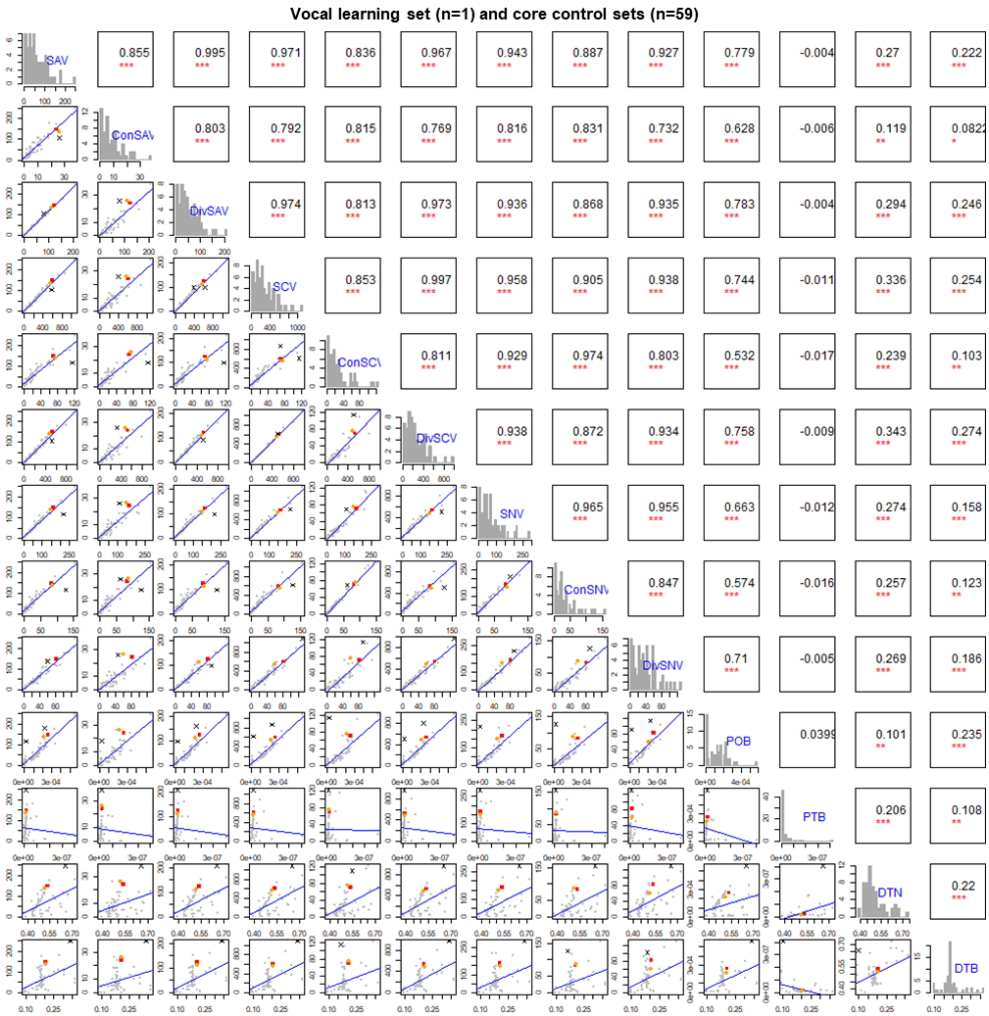

Figure 5.9. Codon and nucleotide variants are also proportional to the product of origin branch lengths in core control sets. *p* values and Adjusted R^2 of correlations are visualized at upper diagonal matrix (p<0.05*, p<0.01**, and p<0.001***). Histograms of frequencies of each convergent variant and 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, orange, and red spots indicate core control sets (n=58), the set of the closest control set (n=1), and the set of avian vocal learners (n=1), respectively. Black 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, SAV = single amino acid variants, ConSAV = convergent

SAV, DivSAV = divergent SAV, SCV = single codon variants, ConSCV = convergent SCV, DivSCV = divergent SCV, SNV = single nucleotide variants, ConSNV = convergent SNV, DivSNV = divergent SNV.

Post hoc analyses of vocal learner-specific 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[94]. Although rifleman was excluded from the initial ConVarFinder search, I can ask whether its sequences match those of vocal non-learners as assumed. I applied principal component analysis (PCA) and phylogenetic analysis for the 148 AVL-SAV sites and the subset of 24 AVL-ConSAV sites (**Figure 5.10A, B**). PC1 and PC2 accounted for 53% and 66% of the total variances of the AVL-SAV sites and AVL-ConSAV sites, respectively. The vocal learning birds clustered away from the vocal non-learning group as expected. For the AVL-SAV sites, rifleman clustered with the vocal non-learners (**Figure 5.10A**). For the AVL-ConSAV subset, rifleman was separate from the two groups, but was still closer to vocal non-learners (**Figure 5.10B**). Phylogenetic analyses of these AVL-SAVs and AVL-ConSAVs were 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 (**Figure 5.10A, B**). These results support the assumption that rifleman is a vocal non-learner.

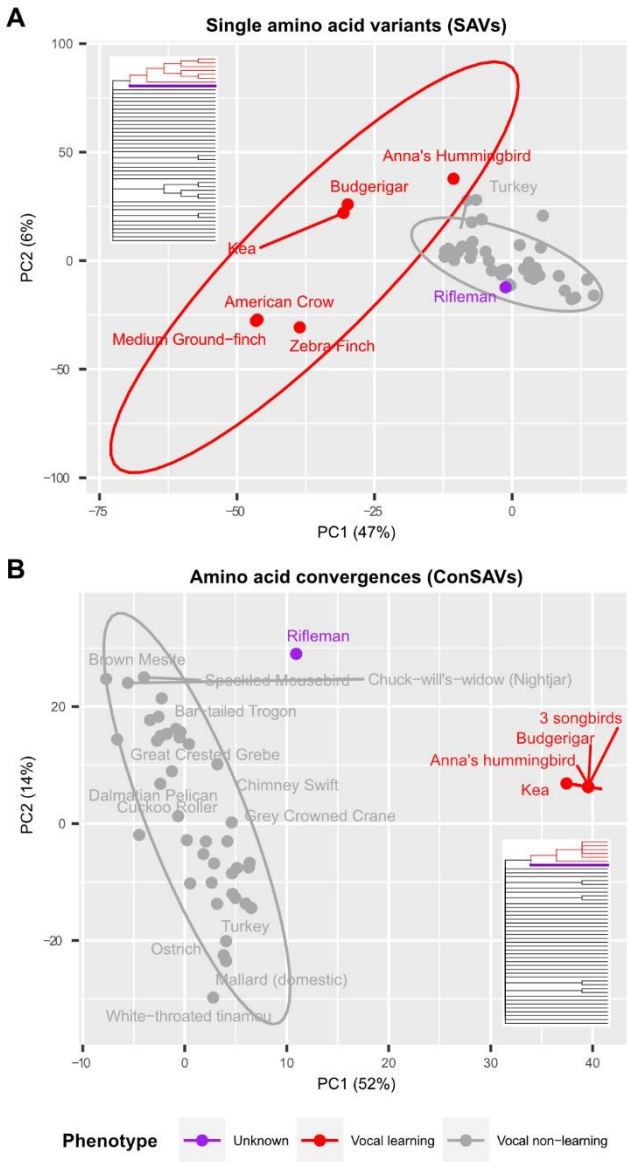

Figure 5.10. Rifleman amino acid profile similar to vocal non-learning birds. (**A**) Principle component analysis (PCA) and consensus tree of 148 AVL-SAV sites. (**B**) PCA and consensus tree of 24 AVL-ConSAV sites. Red, avian vocal learners; Grey, avian vocal non-learners; Purple, rifleman.

Biological functions of genes with amino acid convergences

To investigate the biological functions of genes with various types of variants in vocal learners and in control sets, I performed gene ontology (GO) analyses for 53,058 gene lists with 1 or more variants. Among them, at least one significant (adjusted $p < 0.05$) GO term was found for 7,901 gene lists (14.9%). I further found a positive correlation between the number of significant GO terms and the number of genes with convergent variants in each list (**Figure 5.11A**); however, I found weaker negative correlation between the average of adjusted *p* value of significant GO terms and the number of genes with convergent variants (**Figure 5.11B**).

In vocal learning birds, I did not find any GO enrichment for the total AVL-SAV gene list. However, the AVL-ConSAVs gene list 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*; **Figure 5.11C**). The amino acid convergences (AVL-ConSAVs) of *DRD1B*, *PRKAR2B*, and *TANC1* also showed codon convergences (AVL-ConSCVs) in all vocal learners, while that of *LRRN4* showed different synonymous codon changes (AVL-DivSCVs; **Figure 5.11D,F**). Out of the 8,238 control species combinations, only one control had 2 gene lists with Ctrl-DivSCVs and Ctrl-DivSNVs showed significant enrichment for 'learning' (GO:0007612, both adjusted *p* values = 0.02); the associated set of species (**Figure 5.11E**) did not include any vocal learners, but a convergent variant in *LRRN4* contributed to this functional enrichment (**Figure 5.11F**). The findings indicate that convergent genes in vocal learners do function in the brain and for learning.

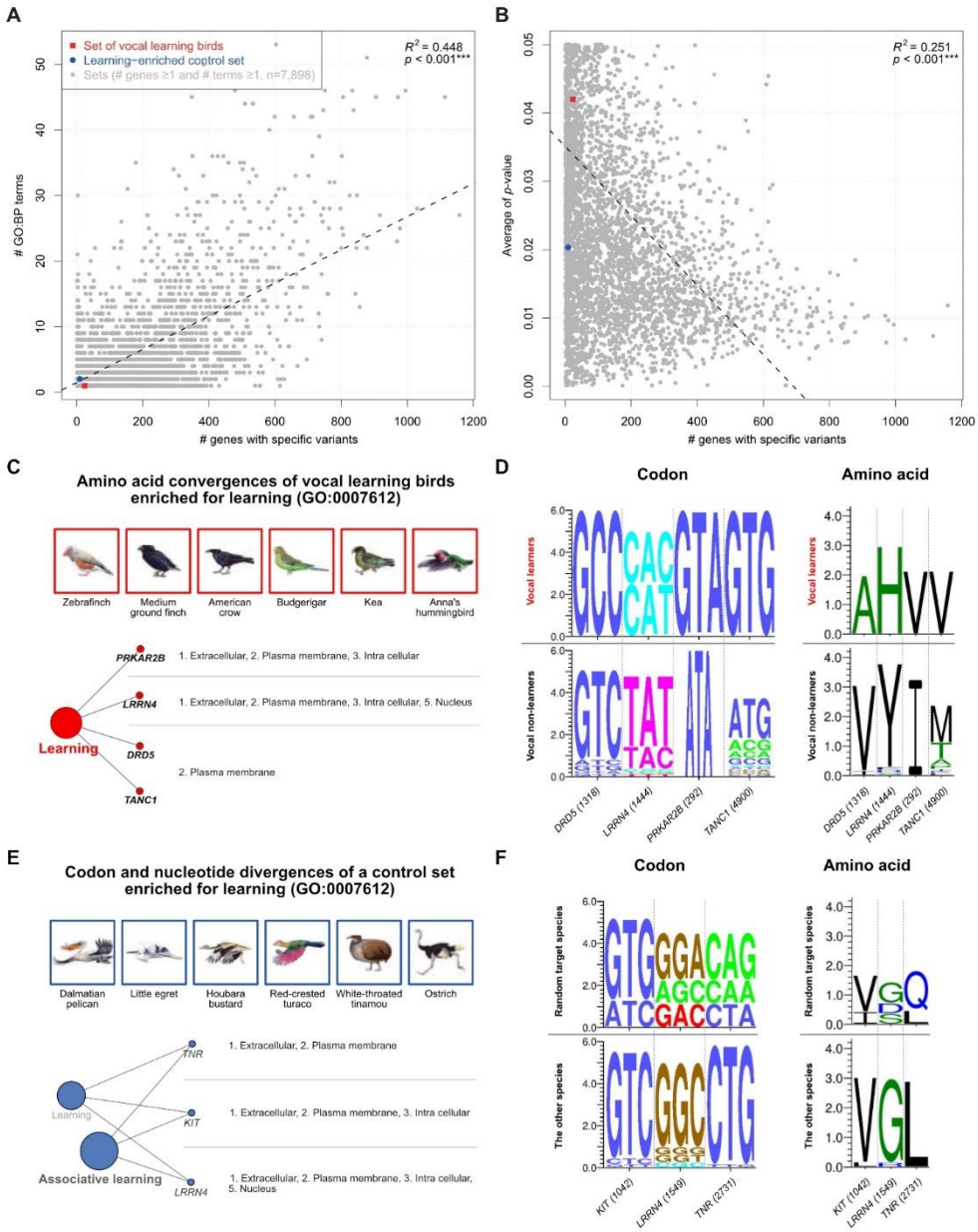

Figure 5.11. Genes with amino acid convergences of vocal learning birds distinctly enriched for a biological function, learning. (**A**) Correlation plot between the number of significantly enriched GO terms and the number of genes with 1 or more variants in each species data set with 9 types (SAVs, ConSAVs, DivSAVs, SCVs, ConSCVs, DivSCVs, SNVs, ConSNVs, and DivSNVs) (n=53,058). (**B**) Correlation plot between averages of p-values of the same data sets. (**C**) Gene ontology analysis for learning associated genes with amino acid

convergences (ConSAVs) of vocal learners (*adj. p* < 0.05). (**D**) Codon and amino acid logos of AVL-ConSAV genes associated with learning. (**e**) Gene ontology analysis for learning associated genes with codon and nucleotide divergences (DivSCVs and DivSNVs) of a control set (*adj. p* < 0.05). (**F**) Codon and amino acid logos of Ctrl-DivSCV and DivSNV genes associated with learning.

Fixation and positive selection of amino acid convergences in vocal learning birds

I checked for additional evidence of whether the SAVs in vocal learners are reliable, by checking for sequence assembly artifacts and SNPs within species. I 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). At the 148 AVL-SAV 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 (**Figure 5.12**). I 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 (**Figure 5.13**). These findings indicate that the vast majority (99.3%) of the single amino acid variants I identified in vocal learners are the result of true species-specific variants.

In addition, to consider positive selection on the AVL-SAVs, I performed dN/dS analysis with the branch-site model for SAV genes in the avian vocal learning species and in their closest control set (songbirds, parrots, and swifts) (**Figure 5.14; Table 5.1**). I found that under half of amino acid convergences showed signs of positive selection (Likelihood ratio value (D) > 0, dN/dS (ω_2.) values of foreground branches >1) in the vocal learning birds (10 of 24 genes, 41.7%) and the closest control set (12 of 26 genes, 46%), with 12.5% (3) and 23% (6) being statistically significant, respectively (adjusted p values (FDR) < 0.05, posterior probability > 0.5). These findings suggest that a subset of genes with amino acid convergences 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.

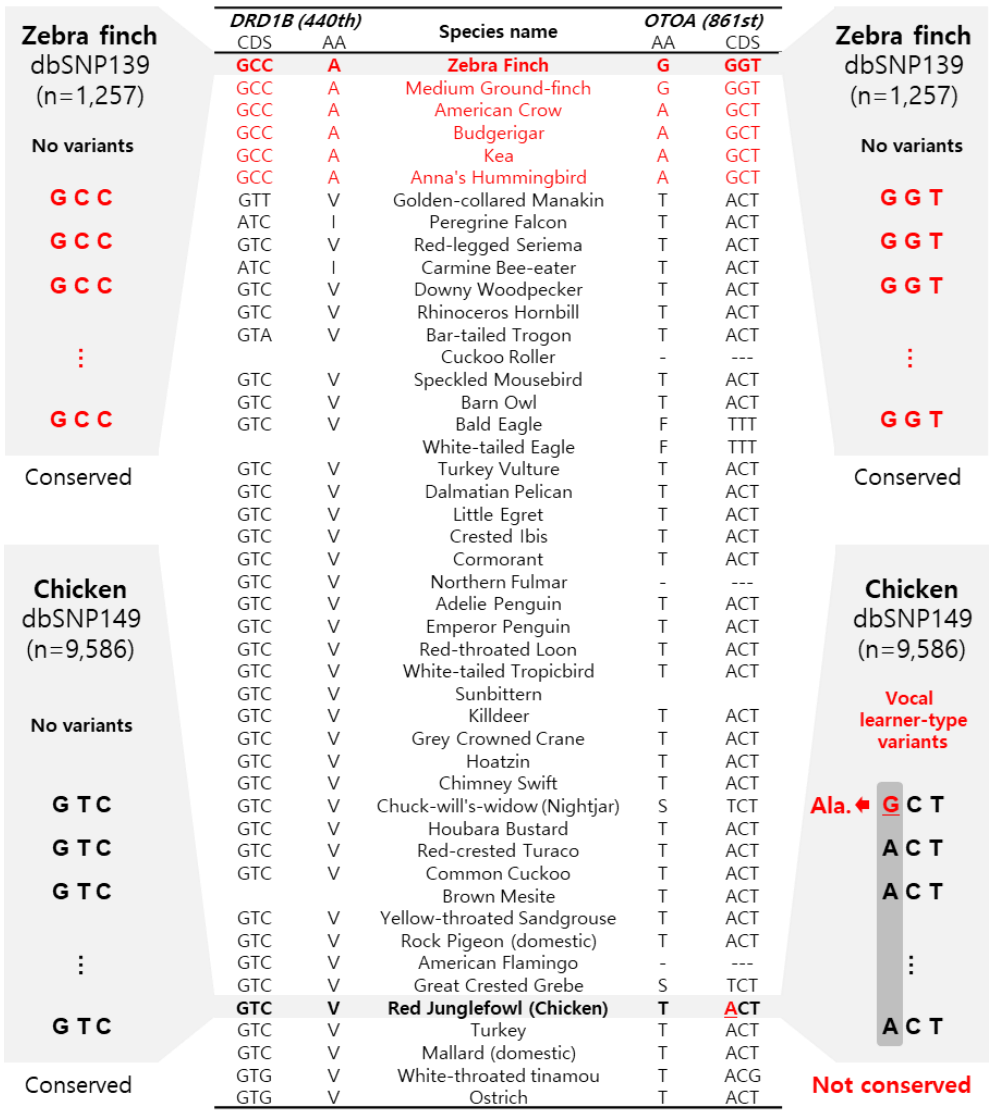

Figure 5.12. Examples of fixed and unfixed differences within each population. The central table indicate convergent single amino acid variants of vocal learning birds in *DRD1B* and *OTOA*. Numbers in parentheses indicate positions in peptide alignments of each gene. Bold characters in the species name column indicate representative species of vocal learners and non-learners which are marked as red and black, respectively. Amino acid and codon column show amino acids and codons of each species at the AVL-SAV sites of each gene. Blank and '- (gap)' indicate absence of orthologous gene in the species' genome and deletions in the species. Under bar at the first site of the AVL-SAV site in *OTOA* gene of chicken indicates a nonsynonymous SNP in chicken population (dbSNP149, number of samples =

9,586). Except for the case of *OTOA* gene of chicken, all of AVL-SAV sites are conserved within zebra finch population (dbSNP139, number of samples = 1,257) and chicken population without any nonsynonymous substitutions.

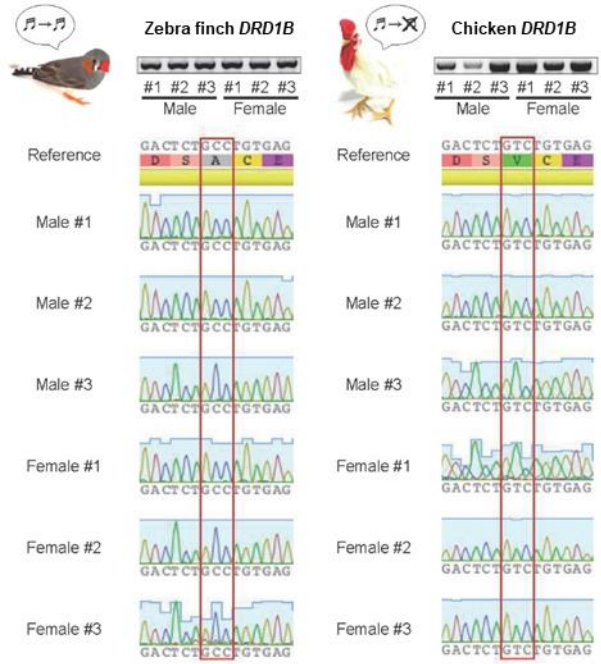

Figure 5.13. Fixed differences of the avian vocal learner-specific amino acid convergences (AVL-ConSAVs) in *DRD1B*. Shown are sequences determined from PCR reactions from individual animals. All of 3 male and 3 female samples of zebra finch and chicken showed fixation of the vocal learner-type codon (GCC) and vocal non-learner-type codon (GTC) at the AVL-ConSAV site in *DRD1B*, respectively.

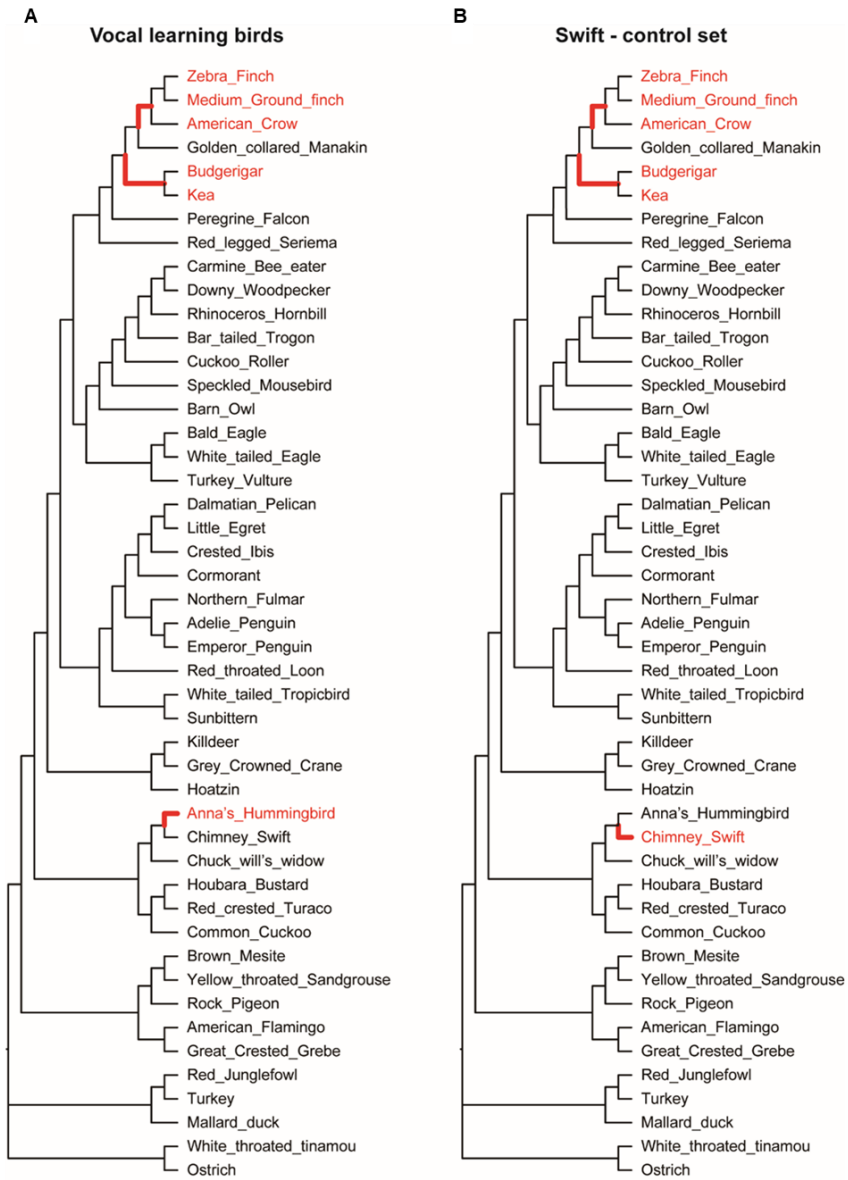

Figure 5.14. Evolutionary models of positive selection on avian vocal learner set and their closest relative set (Swift). (**A**) parsimonious hypothesis to get independent gains of vocal learning ability for each vocal learning clade. (**B**) parsimonious hypothesis for positive selection on species of the closest control set like vocal learners' set. Red characters indicate target species of each set. Bold branches indicate the most recent common ancestral (=origin) branches of each clade of target species which are assumed as foreground branches under positive selection.

Genes with amino acid convergences are specialized in vocal learning and the associated brain subdivisions

I next tested if the AVL-SAV genes are expressed in vocal learning brain circuits. I analyzed 8 brain transcriptome data sets, which include genes that show singing-regulated expression (increased or decreased) in song learning nuclei of songbirds [117] (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 [93,117], DEG_2019: micro-dissected RNA sequencing in 2019 [118], and DEG_2020: laser capture microscope with RNA sequencing in 2020 [118]; **Figure 5.15, Table 5.1**). Based on above data sets, I summarized 6 types of DEG lists: DEG 2014 and DEG 2020 of NUC vs other NUCs, DEG 2019 and DEG 2020 of NUC vs SURs, DEG 2020 of SUR vs other SURs, and Singing-related genes 2014. Relative to the average of all genes (8,295 avian orthologous genes) measured, I found no enrichment of AVL-SAV genes (up to 27.3%) among the singing regulated genes or song nuclei specialized genes relative to the surrounding brain regions, whether positively selected or not (**Figure 5.16A**). However, in two independent transcriptome experiments, I found 60-100% of AVL-ConSAV genes under positive selection 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 (**Figure 5.16A**). These enrichments were not found for the AVL-DivSAV genes, not for any positively selected gene set in the closest related control set (**Figure 5.16A**). 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 (DEG_2020 of NUC vs other NUCs; **Figure 5.16B, Table 5.1**).

Out of 24 AVL-ConSAV genes, a total of 8 genes (33.3%) 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 5.2**). These include two genes, *DRD1B* and *PRKAR2B*, revealed in the GO analyses for learning functions (**Figure 5.11C, D**). Further *DRD1B* has specialized up-regulation specific to adult Area X compared to its surrounding striatum (**Figure 5.16C** and **Table 5.2**) [119-121].

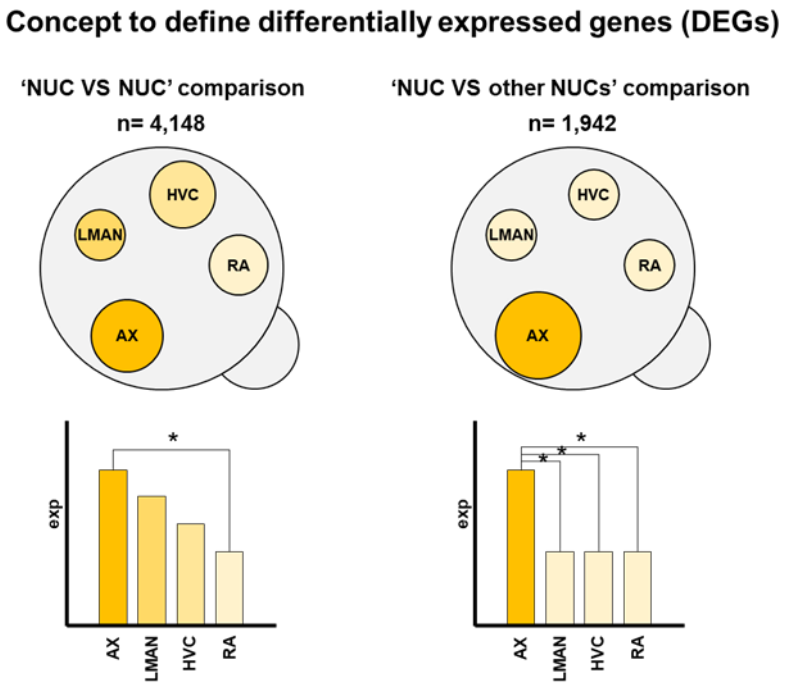

Figure 5.15. Different concepts to define differentially expressed genes in song nuclei. 'NUC VS NUC' comparison: gene supported by a significantly differential expression between a song nucleus relative to at least one of other song nuclei. 'NUC VS other NUCs' comparison: gene supported by 3 significantly differential expression among a song nucleus relative to the other song nuclei.

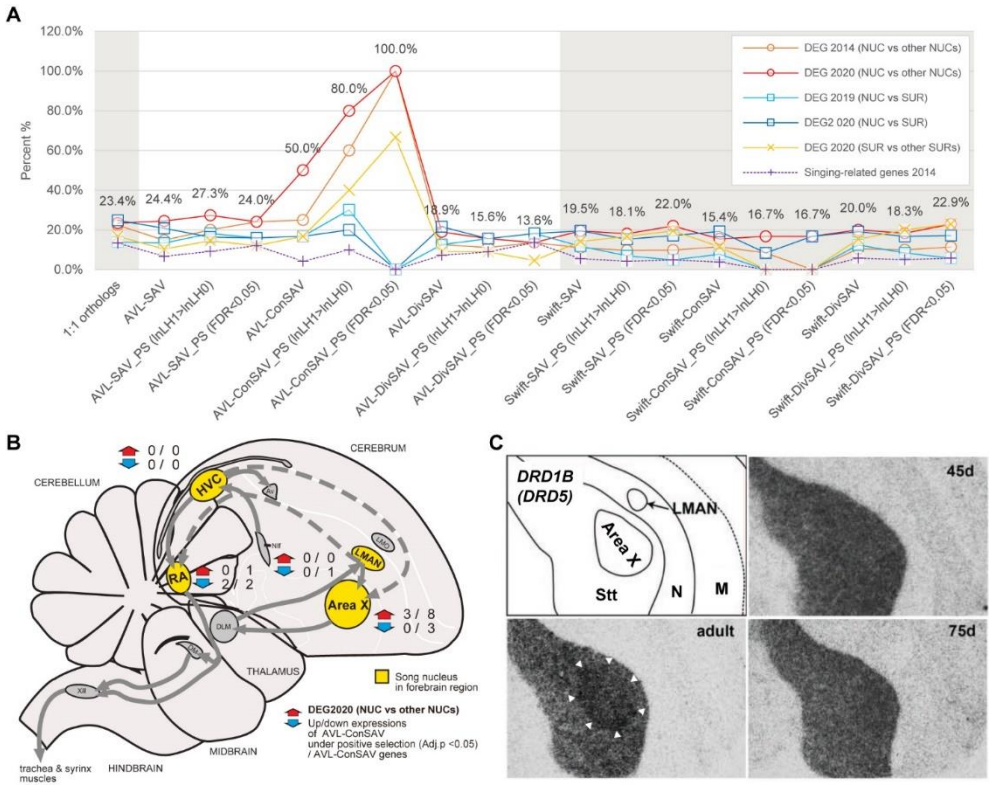

Figure 5.16. Genes with amino acid convergences under positive selection expressed differentially in song nuclei. (**A**) Proportions of singing-related genes (SRG) or differentially expressed genes (DEG) in song learning nuclei or adjacent brain subdivisions of the zebra finch brain (y-axis) that have 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 microarray, micro-dissected RNA sequencing[118], and laser capture microscope RNA sequencing[118] data sets from 2014, 2019, and 2020 analyses, respectively. 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 the numbers of the subset of AVL-ConSAV genes under positive selection / the numbers of all set of AVL-ConSAV genes supported by up

and down regulated expressions in ‘NUC vs other NUCs’ in the DEG_2020 data source. (**C**) *DRD1B* (=*DRD5*) 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. [119].

Table 5.2. Candidate genes related to the avian vocal learning trait with amino acid convergences 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.

Symbol	*B3GNT2*	*DRD1B*	*FNDC1*	*HMGXB3*	*MTFR1*	*PIK3R4*	*PRKAR2B*	*SMPD3*
Pos A.A.	253	416	1034	269	103	671	32	307
A.A. AVLs	N	A	S	D	T	C	V	C
A.A. AVNLs	H	IV	G	E-	AGP-	R	I-	Y-
dN/dS (ω) F.G.	3.3	3.7	6.7	2.3	10.1	10.4	295.8	14.3
Likeli-hood ratio (D)	1.1	1.2	4.3	1	3.2	4.9	25.4	6
Adj. p (FDR)	2.8.E-01	2.7.E-01	6.4.E-02	3.0.E-01	1.0.E-01	4.9.E-02	3.6.E-06	3.4.E-02
Posterior probability (BEB)	0.999	0.5	0.981	0.995	0.521	0.997	0.999	0.994
DEG 2014 (NUC vs other NUCs)	Ax Up	Ax Up		Ax Up		Ax Up	Ra Down	Ax Up
DEG 2019 (NUC vs SUR)	RA Down	AreaX Up	RA Up					
DEG 2020 (NUC vs SUR)		LMAN Down			LMAN Up			
DEG 2020 (NUC vs other NUCs)	AreaX Up	AreaX Up,LMAN Down	RA Up	AreaX Up	AreaX Down	AreaX Up	AreaX Up,RA Down	AreaX Up,RA Down
DEG 2020 (SUR vs other SURs)	AreaX Up	AreaX Up					AreaX Up	AreaX Up,RA Down
Singing-related gene 2014	+							

5.5. Discussions

As the primary structure of proteins, amino acid substitutions can contribute to various traits including human language [122-124]. my findings give us new insights into convergent evolution of amino acid substitutions of polyphyletic lineages in birds, and possible influence on the convergent trait, vocal learning. I discovered correlations between the frequency of amino acid convergence (ConSAVs) with the product of MRCA (=origin) branch lengths (POB). These ConSAVs originate from underlying complex variants as not only convergences but also divergences at codon and nucleotide levels. Remarkably, although vocal learners did not have a higher preponderance of various types of their specific variants including amino acid convergences above background levels, I find that a subset of the ConSAV sites and the associated genes have been positively selected upon and have specialized expression between different brain subdivisions. To explain my findings, I propose a hypothesis of selection on a background of convergent substitutions for convergent traits.

Improving the algorithm of my previous method, I developed a new method 'ConVarFinder' to find molecular convergences potentially associated with convergent traits. It can detect mutually exclusive variants between target species and the other species like previous one and includes a new function to classify them as convergent and divergent variants based on sequence information of terminal and ancestral nodes. Although I traced substitutions from ancestors of each clade to terminal taxa to reflect their evolutionary histories precisely, molecular convergences could be defined with an identical variant of terminal taxa in polyphyletic lineages different from the sequences of the other species by focusing on existing species. This assumption was supported by all types of identical variants (iSAVs, iSCVs, and iSNVs) in species combinations from 3 independent clades that I classified as convergent variants (ConSAVs, ConSCVs, and ConSNVs) by analyzing ancestral substitutions at MRCA branches of each clade. It suggests that

ancestral sequence reconstructions would be skipped to simplify methods to identify convergent substitutions between independent lineages.

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 I find correlations between the phylogenetic feature, POB, and convergent variants in species from multiple independent lineages, where other analyses and studies have attempted and failed to find [116,125-127]. my positive selection, functional association, and gene expression analyses suggest that selection occurs on some of these amino acid convergences to contribute to evolving novel, convergent traits, in my case vocal learning. 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 genes with amino acid convergences in vocal learning clades and specialized expression specific to a vocal learning nucleus and the associated surrounding brain subdivision was both intriguing and perplexing to us. If anything, I 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 I believe is real, as I 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 [128], but my findings suggest that they could synergistically influence evolution of each other. Studies in my 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, I suggest 8 key candidate genes for associations with the vocal learning ability in birds (**Table 5.2**).

When searching for convergent substitutions among species, I believe my 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 [96] or mammals [19]. I further find that such a relationship exists in higher dimensional combinations of species from multiple independent lineages, but this type of analyses does not control for species relationships. Several other studies found that the rate or number of convergent substitutions decreases with increasing genetic distance between two lineages [97,116]. my findings with the product of the MRCA branch lengths in 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-ConSAV sites includes the *DRD1B* dopamine receptor associated with learning [129], and *LRRN4* that affects long lasting memory [120,130], fundamental traits of vocal learning [131]. *TANC1* regulates dendritic spines and spatial memory [132]. At the mechanistic level, *DRD1B*, through its G-protein, regulates activity of adenylyl cyclase's synthesis of cAMP in the cell membrane [121,133]; *PRKAR2B*, or Protein Kinase cAMP-dependent Type II Regulatory Subunit Beta, is an enzyme that activates cAMP-dependent protein kinase (PKA) inside the cell [134]. Additionally, one of the most well-known genes that PKA inhibits is involved in learning, including vocal learning [135], 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 [136]. 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 (**Figure 5.17**).

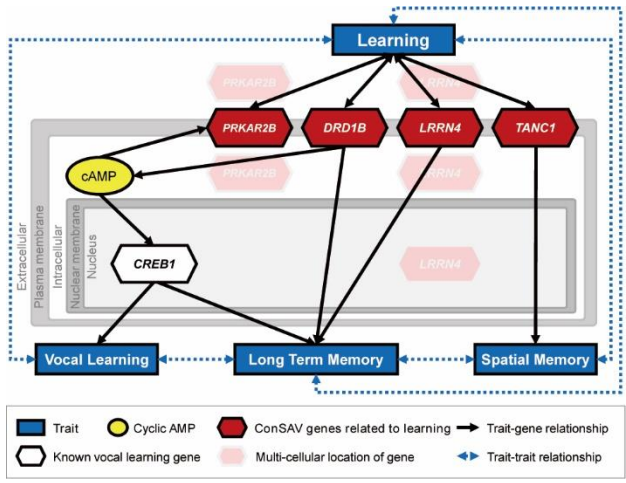

Figure 5.17. Candidate genes converge on Cyclic AMP-based vocal learning pathway. Red hexagons indicate learning genes with amino acid convergences (ConSAVs) specific avian vocal learners. Transparent red hexagons indicate muti-cellular location of the candidate genes. White hexagon indicates vocal learning gene, *CREB1*. Yellow circle indicates cyclic AMP (cAMP). Grey-scale rectangular indicates cellular positions of the genes and cAMP. Blue rectangular indicates traits related to learning genes. Black arrows and dashed blue arrows indicate trait-gene relationships and trait-trait relationships, respectively.

I analyzed the standard singleton orthologous gene sets and their alignments of protein coding sequences in 48 avian genome assemblies generated from the initial phase of avian phylogenetics project [8,94]. These short-read-based genome assemblies do have most protein coding genes assembled, but do not have all repetitive and GC-rich regions assembled [137]. Using multiple long-read sequencing technologies that can correct the above errors including falsely missing genes or exons [9,12], recent genome projects are generating comparably high-quality reference genomes of various lineages. The improving and expanding genome assemblies can give us opportunities to validate my findings for rules of molecular convergences based on more precise orthologous gene sets in avian and other lineages.

Although my study illuminated novel findings, it spurs on ideas for future studies. Vocal learning species could share other convergent traits besides vocal learning [14,86,138,139], and the identified genes could be associated with these other traits. The basic rules of convergent evolution I discovered in protein coding regions leave open the possibility that similar or different rules apply to non-coding regions. ConSAVs 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. I identified new candidate genes and specific nucleotide variants that can be genetically manipulated when the technology is more advanced [82,140] to test possible causal roles in the evolution and function of vocal learning. It will be useful to determine if the convergent rules I identified here are specific to birds, or are more widespread across life forms.

General discussion

This dissertation consists of a series of bioinformatic approaches from reference genome constructions to comparative genomic analyses to understand molecular mechanisms of macro-evolution across species in various lineages.

Reference genome assemblies are being exponentially increased through ongoing improvements of sequencing technologies and assembly algorithms. Although the 1st and 2nd generations of sequencing technologies had opened the genomics era and had initiated its expansion, the previous reference genomes generated with the short-read based reads of the Sanger and Illumina technologies had critical problems such as, fragmentations, missing sequences, and false duplications. The international consortium, Vertebrate Genomes Project (VGP), is constructing the better-quality genome assemblies with combinations of recent sequencing and assembly technologies to solve the prior errors. As the first VGP collaboration in Republic of Korea, I generated the chromosome-scale genome assembly satisfying the VGP-platinum-quality of Korean giant-fin mudskipper, *Periophthalmus magnuspinnatus*, which is an indigenous fish in the Yellow sea. Compared to the previous assembly (GCA_000787105.1), it achieved the 100-fold longer continuity, corrected erroneous fragmentations, missing, and duplications of highly conserved genes in vertebrates, and detected more protein coding genes. Moreover, its genome sequences were approximately 753 Mbp and 99.5% of the assembled bases were assigned to 25 chromosomes. Out of these chromosomes, 60% included telomeric repeats at the 5' or 3' ends. The new assembly validated the usage of the VGP standard assembly pipeline 1.6 to generate chromosomal assemblies and introduced the improved sequencing and assembly technologies to the academic society in the South Korea.

This kind of chromosome-level reference genomes can provide unprecedent opportunities to investigate molecular evolution of chromosomes within or across species. Base-wise genome-wide alignment programs detected evolutionary breakpoints which were conserved chromosomal recombinations specific to a same family, order or class different from their closely relative lineages.

However, there was a challenge to align reference genomes of distant species in vertebrates because of their divergent sequences. To investigate chromosomal rearrangements between species in vertebrates, I developed a new method, Chromosomal Orthologous link (ChrOrthLink). It uses highly conserved singleton orthologous genes detected by BUSCO analysis and traces their synteny on chromosomes across species. I preformed the analysis for 16 vertebrates species, visualized the chromosomal rearrangements marking singleton complete BUSCO genes on their chromosomes, and found that tetrapod animals had chromosomal conservations with the cartridge fish lineage compared to ray-fined fish lineage. Although this approach has a limited resolution to analyze 'gene-wise' synteny blocks ignoring their neighbor genes which are not conserved, I believe this new strategy can make blueprints of chromosomal evolution across diverse species in various lineages over vertebrates.

Many researchers tried to find solutions to improve qualities of genome assemblies, but they were overlooked what genes were mis-assembled in previous generations. I suggested a new usage of genome-wide alignments to detect various types of errors in a genome assembly by comparing different versions of genome assemblies of same species. I designed back-to-back studies to quantify 'false gene losses' and 'false gene gains' of 4 vertebrate species which had previous assemblies of same species of 1st release VGP reference genomes, and mainly contributed following findings. Missing errors were prevalent on CpG islands and repeat regions relative to non-CpG island and non-repeat regions. Raw-read mapping used in prior assemblies to new assemblies can validate errors did not originated from biases of individual differences in these genome assembly comparisons of same species. It empathizes the necessity of preservations of raw sequencing data or bio-samples to validate improvements of next genome assemblies in future. As examples of false missing and false duplications, *COQ6* and *mTOR* genes highly conserved in vertebrates were fully or partially missing or duplicated in short-read-based reference genomes of various species in mammals, birds, and fishes. These findings demonstrated genomic factors causing technical limitations of previous sequencing and assembly technologies and support the new utilizations of the comparative

genomic approaches to evaluate the assembly quality.

At the beginning stage of genomics, a few reference genomes of human and model species were available to find species-specific variants in comparative analyses to investigate molecular evolution specific to in a new genome assembly find variants. The ongoing accumulations of reference genomes of various species give us opportunities to understand molecular evolutions for species and lineages. Comparative genomic approaches for accumulative genome assemblies of various species could identify novel candidate genes associated with interesting traits, such as the limb-emergence in the *Sarcopterygii* clade and vocal learning abilities in the *Aves* clades.

As a representative candidate variant for limb emergence, I found the amino acid substitution on *SHOX* gene. At the site, coelacanth had serine under positive selection substituted from leucine of *Actinopterygii* species and it was shared with several tetrapod species in *Sarcopterygii*. For vocal learning, I discovered the convergent amino acid substitutions on *DRD1B* and *PRKAR2B* genes unique to vocal learning birds mutually exclusive to non-learning birds. These convergent amino acid patterns of avian vocal learners were uniquely observed in the candidate gene. These amino acid convergences were under positive selection and conserved within zebrafinch population. The convergent genes of vocal learners were enriched for learning and were supported by differential gene expression related to vocal learning in sub-brain regions. I believe these candidates provide new insights for molecular mechanisms to understand fundamental traits of human evolution.

References

1 Olsen, U. D. J. G. I. H. T. B. E. P. P. R. P. W. S. S. T. D. N. C. J.-F. *et al.* Initial sequencing and analysis of the human genome. *nature* **409**, 860-921 (2001).

2 Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546-567 (1996).

3 de, A. G. I. g. t. o. g. g. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *nature* **408**, 796-815 (2000).

4 Simon 3, E. B. I. B. E. G. N. K. A. M. E. R. A. G. S. G. S. A. U.-V. A. W. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).

5 Weissenbach, J. The rise of genomics. *Comptes rendus biologies* **339**, 231-239 (2016).

6 Consortium, G. P. A map of human genome variation from population scale sequencing. *Nature* **467**, 1061 (2010).

7 Groenen, M. A. *et al.* Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**, 393-398 (2012).

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

9 Korlach, J. *et al.* De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *Gigascience* **6**, gix085 (2017).

10 Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737-746 (2021).

11 Ko, B. J. *et al.* Widespread false gene gains caused by duplication errors in genome assemblies. *bioRxiv* (2021).

12 Kim, J. *et al.* False gene and chromosome losses affected by assembly and sequence errors. *bioRxiv* (2021).

13 Nottebohm, F. The origins of vocal learning. *The American Naturalist* **106**, 116-140 (1972).

14 Nowicki, S. & Searcy, W. A. The evolution of vocal learning. *Current opinion in neurobiology* **28**, 48-53 (2014).

15 Usui, N. & Konopka, G. Decoding the molecular evolution of human cognition using comparative genomics. *Brain, behavior and evolution* **84**, 103-116 (2014).

16 Lai, C. S., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F. & Monaco, A. P. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**, 519-523 (2001).

17 Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**, 869-872 (2002).

18 Parker, J. *et al.* Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228-231 (2013).

19 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).

20 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood.

Molecular biology and evolution **24**, 1586-1591 (2007).
21 Park, J. Structure of the skin of an air-breathing mudskipper, Periophthalmus magnuspinnatus. *Journal of fish biology* **60**, 1543-1550 (2002).
22 Park, J., Kim, I. & Lee, Y. A study on the vascularization and structure of the epidermis of the air-breathing mudskipper, Periophthalmus magnuspinnatus (Gobiidae, Teleostei), along different parts of the body. *Journal of Applied Ichthyology* **22**, 62-67 (2006).
23 Pace, C. & Gibb, A. C. Mudskipper pectoral fin kinematics in aquatic and terrestrial environments. *Journal of Experimental Biology* **212**, 2279-2286 (2009).
24 Ansari, A. A., Trivedi, S., Saggu, S. & Rehman, H. Mudskipper: A biological indicator for environmental monitoring and assessment of coastal waters. *Journal of Entomology and Zoology Studies* **2**, 22-33 (2014).
25 You, X. *et al.* Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *Nature communications* **5**, 5594 (2014).
26 Lee, Y.-J., Choi, Y. & Ryu, B.-S. A taxonomic revision of the genus Periophthalmus (Pisces: Gobiidae) from Korea with description of a new species. *Korean Journal of Ichthyology* **7**, 120-127 (1995).
27 Wang, Z. A long-term misidentified new record species of Gobiidae from China-Periophthalmus magnuspinnatus. *Acta Zootaxon* **31**, 906-910 (2006).
28 Studds, C. E. *et al.* Rapid population decline in migratory shorebirds relying on Yellow Sea tidal mudflats as stopover sites. *Nature Communications* **8**, 14895, doi:10.1038/ncomms14895 (2017).
29 Murray, N. J., Clemens, R. S., Phinn, S. R., Possingham, H. P. & Fuller, R. A. Tracking the rapid loss of tidal wetlands in the Yellow Sea. *Frontiers in Ecology and the Environment* **12**, 267-272 (2014).
30 Jiang, X., Teng, A., Xu, W. & Liu, X. Distribution and pollution assessment of heavy metals in surface sediments in the Yellow Sea. *Marine pollution bulletin* **83**, 366-375 (2014).
31 Li, G. *et al.* Heavy metals distribution and contamination in surface sediments of the coastal Shandong Peninsula (Yellow Sea). *Marine pollution bulletin* **76**, 420-426 (2013).
32 Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138 (2009).
33 Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotechnol* **30**, 771-776, doi:10.1038/nbt.2303 (2012).
34 Zheng, G. X. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology* **34**, 303 (2016).
35 Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268-276, doi:10.1016/j.ymeth.2012.05.001 (2012).
36 Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies

from long-read SMRT sequencing data. *Nature methods* **10**, 563-569 (2013).
37 Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods* **13**, 1050-1054 (2016).
38 Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896-2898 (2020).
39 Ning, Z. H., E. . Scaff10X https://github.com/wtsi-hpag/Scaff10X.
40 Bionano Genomics, I. Bionano Software Downloads. https://bionanogenomics.com/support/software-downloads/.
41 Arima Genomics, I. Arima Genomics Mapping Pipeline. https://github.com/ArimaGenomics/mapping_pipeline.
42 Ghurye, J. *et al.* Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS computational biology* **15**, e1007273 (2019).
43 Bishara, A. *et al.* Read clouds uncover variation in complex regions of the human genome. *Genome research* **25**, 1570-1580 (2015).
44 Chow, W. *et al.* gEVAL—a web-based browser for evaluating genome assemblies. *Bioinformatics* **32**, 2508-2510 (2016).
45 Howe, K. *et al.* Significantly improving the quality of genome assemblies through curation. *GigaScience* **10**, giaa153 (2021).
46 Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).
47 Hu, Y. *et al.* OmicCircos: a simple-to-use R package for the circular visualization of multidimensional omics data. *Cancer informatics* **13**, CIN. S13495 (2014).
48 Team, R. C. R: A language and environment for statistical computing. (2013).
49 Meyne, J., Ratliff, R. L. & Moyzis, R. K. Conservation of the human telomere sequence (TTAGGG)n among vertebrates. *Proceedings of the National Academy of Sciences* **86**, 7049-7053, doi:doi:10.1073/pnas.86.18.7049 (1989).
50 Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178-192 (2013).
51 Françoise Thibaud-Nissen, P., Alexander Souvorov, PhD, Terence Murphy, PhD, Michael DiCuccio, MD, and Paul Kitts, PhD. Eukaryotic Genome Annotation Pipeline. *The NCBI Handbook [Internet]. 2nd edition.* (2013).
52 Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic acids research* **32**, W309-W312 (2004).
53 Hoff, K. J. & Stanke, M. Predicting genes in single genomes with AUGUSTUS. *Current protocols in bioinformatics* **65**, e57 (2019).
54 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-

3212, doi:10.1093/bioinformatics/btv351 (2015).
55 Kim, J., Lee, C., Yoo, D. & Kim, H. Genetic Adaptations in Mudskipper and Tetrapod Give Insights into Their Convergent Water-to-Land Transition. *Animals* **11**, 584 (2021).
56 Korlach, J. *et al.* De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience* **6**, doi:10.1093/gigascience/gix085 (2017).
57 Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202-2204, doi:10.1093/bioinformatics/btx153 (2017).
58 Rice, E. S. & Green, R. E. New approaches for genome assembly and scaffolding. *Annu Rev Anim Biosci* **7**, 17-40 (2019).
59 Kelley, D. R. & Salzberg, S. L. Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome biology* **11**, 1-11 (2010).
60 Phillippy, A. M., Schatz, M. C. & Pop, M. Genome assembly forensics: finding the elusive mis-assembly. *Genome biology* **9**, 1-13 (2008).
61 Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome biology* **4**, 1-10 (2003).
62 Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics* **13**, 278-289 (2015).
63 Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome research* **27**, 757-767 (2017).
64 Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature biotechnology* **30**, 771-776 (2012).
65 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* **326**, 289-293 (2009).
66 Lovell, P. V. *et al.* ZEBrA: Zebra finch Expression Brain Atlas—A resource for comparative molecular neuroanatomy and brain evolution studies. *Journal of Comparative Neurology* **528**, 2099-2131 (2020).
67 Robinson, R. For mammals, loss of yolk and gain of milk went hand in hand. *PLoS biology* **6**, e77 (2008).
68 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
69 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26 (2011).
70 Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341-1342, doi:10.1093/bioinformatics/btt128 (2013).
71 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006, doi:10.1101/gr.229102 (2002).
72 Guy, L., Roat Kultima, J. & Andersson, S. G. genoPlotR: comparative

gene and genome visualization in R. *Bioinformatics* **26**, 2334-2335 (2010).

73 Haug-Baltzell, A., Jarvis, E. D., McCarthy, F. M. & Lyons, E. Identification of dopamine receptors across the extant avian family tree and analysis with other clades uncovers a polyploid expansion among vertebrates. *Frontiers in neuroscience* **9**, 361 (2015).

74 Speidel, D. *et al.* CAPS1 regulates catecholamine loading of large dense-core vesicles. *Neuron* **46**, 75-88 (2005).

75 Lovell, P. V., Clayton, D. F., Replogle, K. L. & Mello, C. V. Birdsong "transcriptomics": neurochemical specializations of the oscine song system. *PloS one* **3**, e3440 (2008).

76 Bahudhanapati, H., Bhattacharya, S. & Wei, S. Evolution of Vertebrate Adam Genes; Duplication of Testicular Adams from Ancient Adam9/9-like Loci. *PLOS ONE* **10**, e0136281, doi:10.1371/journal.pone.0136281 (2015).

77 Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175 (2008).

78 Wart, H. E. V. & Birkedal-Hansen, H. The cysteine switch: a principle of regulation of metalloproteinase activity with potential applicability to the entire matrix metalloproteinase gene family. *Proceedings of the National Academy of Sciences* **87**, 5578-5582, doi:doi:10.1073/pnas.87.14.5578 (1990).

79 Howe, K. L. *et al.* Ensembl 2021. *Nucleic acids research* **49**, D884-D891 (2021).

80 Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. **418**, 869 (2002).

81 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).

82 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).

83 Enard, W. *et al.* A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell* **137**, 961-971 (2009).

84 Reimers-Kipping, S., Hevers, W., Pääbo, S. & Enard, W. Humanized Foxp2 specifically affects cortico-basal ganglia circuits. *Neuroscience* **175**, 75-84 (2011).

85 Enard, W. FOXP2 and the role of cortico-basal ganglia circuits in speech and language evolution. *Current opinion in neurobiology* **21**, 415-424 (2011).

86 Jarvis, E. D. Evolution of vocal learning and spoken language. *Science* **366**, 50-54, doi:10.1126/science.aax0287 (2019).

87 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).

88 Castellucci, G. A., McGinley, M. J. & McCormick, D. A. Knockout of

Foxp2 disrupts vocal development in mice. *Scientific reports* **6**, 23305 (2016).
89 Nottebohm, F. The origins of vocal learning. *American Naturalist*, 116-140 (1972).
90 Jarvis, E. D. Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Sciences* **1016**, 749-777 (2004).
91 Petkov, C. I. & Jarvis, E. Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Frontiers in evolutionary neuroscience* **4**, 12 (2012).
92 Cahill, J. A. *et al.* Positive selection in noncoding genomic regions of vocal learning birds is associated with genes implicated in vocal learning and speech functions in humans. *Genome research* **31**, 2035-2049 (2021).
93 Pfenning, A. R. *et al.* Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**, 1256846 (2014).
94 Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320-1331 (2014).
95 Warren, W. C. *et al.* The genome of a songbird. *Nature* **464**, 757-762 (2010).
96 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).
97 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).
98 Jarvis, E. D. *et al.* Phylogenomic analyses data of the avian phylogenomics project. *GigaScience* **4**, s13742-13014-10038-13741 (2015).
99 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* **17**, 540-552 (2000).
100 Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18**, 6097-6100 (1990).
101 Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome research* **14**, 1188-1190 (2004).
102 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).
103 Fox, J. *et al.* Package 'car'. *Vienna: R Foundation for Statistical Computing* (2012).
104 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191 (2009).
105 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms.

Molecular biology and evolution **35**, 1547-1549 (2018).
106 Raudvere, U. *et al.* g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research* **47**, W191-W198 (2019).
107 Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091-1093 (2009).
108 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).
109 Bindea, G., Galon, J. & Mlecnik, B. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics*, btt019 (2013).
110 Yates, A. *et al.* Ensembl 2016. *Nucleic acids research* **44**, D710-D716 (2016).
111 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic acids research* **46**, D754-D761 (2018).
112 Madden, T. in *The NCBI Handbook [Internet]. 2nd edition* (National Center for Biotechnology Information (US), 2013).
113 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308-311 (2001).
114 Pfenning, A. R. *et al.* Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**, 1256846, doi:10.1126/science.1256846 (2014).
115 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
116 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).
117 Whitney, O. *et al.* Core and region-enriched networks of behaviorally regulated genes and the singing genome. *Science* **346**, 1256780 (2014).
118 al., G. e. in preparation. (2020).
119 Kubikova, L., Wada, K. & Jarvis, E. D. Dopamine receptors in a songbird brain. *Journal of Comparative Neurology* **518**, 741-769 (2010).
120 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).
121 Rangel-Barajas, C., Coronel, I. & Florán, B. Dopamine receptors and neurodegeneration. *Aging and disease* **6**, 349 (2015).
122 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).
123 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).
124 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).
125 Speed, M. P. & Arbuckle, K. Quantification provides a conceptual basis for convergent evolution. *Biological Reviews* **92**, 815-829 (2017).
126 Storz, J. F. Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews Genetics* **17**, 239, doi:10.1038/nrg.2016.11 (2016).
127 Rittschof, C. C. & Robinson, G. E. in *Current Topics in Developmental Biology* Vol. 119 (ed Virginie Orgogozo) 157-204 (Academic Press, 2016).
128 Carroll, S. B. Evolution at two levels: on genes and form. *PLoS Biol* **3**, e245, doi:10.1371/journal.pbio.0030245 (2005).
129 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).
130 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).
131 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).
132 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).
133 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).
134 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).
135 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).
136 Kandel, E. R. The molecular biology of memory: cAMP, PKA, CRE, CREB-1, CREB-2, and CPEB. *Molecular brain* **5**, 14 (2012).
137 Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv* (2020).
138 Naguib, M. & Riebel, K. in *Biocommunication of animals* 233-247 (Springer, 2014).
139 Mason, N. A. *et al.* Song evolution, speciation, and vocal learning in passerine birds. *Evolution* **71**, 786-796, doi:10.1111/evo.13159 (2017).
140 Liu, W.-c. *et al.* Human mutant huntingtin disrupts vocal learning in transgenic songbirds. *Nature neuroscience* **18**, 1617 (2015).

요약 (국문초록)

척추동물아문 내 다른 계통 간 대진화를 이해하기 위한 생물정보학적 접근

이 철
협동과정 생물정보학전공
서울대학교 대학원 자연과학대학

생물정보학은 디지털화된 유전서열정보를 토대로 다양한 생명현상의 원리를 규명하고 이를 활용해 인류의 삶의 질을 향상하는 것을 목적으로 할 것이다. 생물정보학적 연구는 각 종을 대표하는 표준유전체 구축으로 일반적으로 시작되고 미소 혹은 대진화에 대한 후속 연구를 진행한다. 비록 짧은 단편 해독 기술이 유전체 시대를 열었지만, 짧은 단편의 조립은 낮은 연결성이나 오류가 포함된 유전자 주석 등의 심각한 문제들을 가진다. 긴 단편 해독 기술은 염색체 수준의 주석 (scaffolds)에 필수적인 보다 긴 컨티그 (contig) 조립을 생산할 수 있다. 짧은 단편에서 긴 단편으로 변화하는 페러다임에 발 맞추어, 본 논문은 표준유전체 구축에서 비교유전체 분석까지 이어지는 일련의 생물정보학적 분석에

대한 집약적 연구를 수행했으며, 이는 다양한 척추동물 종들의 대진화를 이해하는 것이 목적이다.

제 1 장에서는 연구의 일반적인 배경지식을 정리하였다. 첫째로, 염색체 수준의 주석을 달성한 표준유전체 구축의 페러다임 변화를 설명했다. 다음으로, 특이적 형질에 관련된 분자 진화를 규명하는 비교유전체 분석 방법 및 사례를 정리했다.

제 2 장에서는 표준유전체를 구축한 사례로서, 대한민국의 고유종인 큰볏말뚝망둥어의 염색체 수준 표준유전체를 구축했다. 척추동물 유전체 프로젝트와 국제 협력을 통해 4 가지 최신 유전체 해독기술들 (Pacbio CLR, 10X Genomics linked reads, Bionano optical mapping, 그리고 Arima Genomics Hi-C)을 활용하여, 기존 표준유전체와 비교해 연결성 (continuity, Scaffold N50 기준)이 약 100 배 향상되고 총 25 개의 염색체를 가진 고품질 표준유전체를 완성했다. 또한, Pacbio Isoseq 전사체 데이터를 유전자 주석에 활용하여 총 24,744 개의 유전자를 발굴했다.

제 3 장에서는 표준유전체 품질 평가 방법과 비교유전체학적 분석을 접목한 사례로서, 분화 시기가 오래된 종 간에도 BUSCO 유전자를 활용해 염색체 수준의 진화 양상을 탐색하는 방법과 척추동물 내에서 사례를 제시했다. 또한, 포유류, 조류, 어류 등 다양한 척추동물의 표준유전체에서 후속 분석 상의 문제를 야기하는 허위 소실 및 중복 오류를 탐색하는 방법과 사례를 제시하고 발생원인을 밝혔다.

제 4 장에서는 기존의 비교유전체학적 분석을 적용한 사례로서, 실러캔스를 포함하는 육기아강 단계통 파생적 진화에 대한 분석을 통해 육상 적응 및 사지 출현의 분자 기작을 규명했다.

제 5 장에서는 새로운 비교유전체학적 분석을 적용한 사례로서, 발성학습 조류 및 대조군 각각의 다계통 수렴 진화에 대한 분석을 통해

아미노산 수렴의 진화적 법칙을 제안하고 발성 학습에 연관된 후보 유전자를 발굴했다.

이러한 표준유전체 구축에서부터 비교유전체 분석으로 이어지는 생물정보학적 접근을 통해 규명된 주요 연구결과 중에, 염색체 상 텔로미어 서열 분포 및 아미노산 수렴 진화의 원리는 척추동물 외에 다른 분류 군에서도 비교될 기준이 될 수 있을 것으로 기대된다. 또한, 사지 발달 및 발성 학습에 연관된 후보 유전자를 발굴한 비교유전체학적 접근법은 전 세계 다양한 생물들의 다양한 유용 형질에 연관된 유용 유전자를 발굴하는데 활용될 수 있을 것이다.

주요어: 표준유전체 조립, 척추동물 유전체 프로젝트, 허위 유전자 소실, 허위 유전자 중복, 파생적 진화, 수렴 진화

학번: 2015-30991