Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Structural and evolutionary diversification of GATA zinc finger genes across jawed vertebrate species

Mangi Kim

aksrl0510@naver.com

Sangmyung University https://orcid.org/0009-0009-0227-4869

Research Article

Keywords: GATA zinc finger genes, Comparative genomics, Phylogenetic evolution, Modular diversification, Gene duplication and retention

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10940516/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Jawed vertebrates, also known as *Gnathostomata*, exhibit extensive genomic and phenotypic diversification driven by the evolution of complex transcriptional regulatory networks. GATA zinc finger (*zf-GATA*) genes represent an ancient regulatory family characterized by a conserved GATA DNA-binding domain. However, the evolutionary processes underlying their structural diversification across vertebrate lineages remain poorly understood. To address this knowledge gap, a genome-scale comparative and phylogenetic analysis of 2,809 *zf-GATA* genes from 181 jawed vertebrate species was performed to investigate how conserved molecular features and lineage-specific innovations have shaped the evolution of this TF family. Comparative analyses revealed that *zf-GATA* diversification has proceeded through a balance between preservation of the conserved GATA DNA-binding framework and modification of surrounding regulatory architectures. Across jawed vertebrates, the Cys-X_2-Cys-X_{17}-Cys-X_2-Cys GATA zinc-binding motif and key structural residues were strongly conserved, whereas protein architectures, conserved domain compositions, and gene repertoires exhibited lineage-specific diversification. Phylogenetic analyses further demonstrated that distinct *zf-GATA* groups have followed independent evolutionary trajectories while retaining a shared molecular framework, with gene duplication and lineage-specific retention contributing to the expansion and diversification of *zf-GATA* repertoires. Structural analyses supported the conservation of GATA domain organization despite extensive variation in overall protein architectures. Collectively, these findings provide an evolutionary framework demonstrating how TF families diversify through coordinated changes in conserved molecular modules, protein architectures, and genomic organization while maintaining essential DNA-binding functions.

1. Introduction

Jawed vertebrates represent a highly diverse group of vertebrates characterized by articulated jaws and paired appendages and exhibiting extensive ecological and evolutionary diversification (Brazeau and Friedman, 2015). As of 2021, approximately 70,000 extant vertebrate species had been catalogued, of which nearly 99% belong to jawed vertebrates, and large-scale genome sequencing efforts are rapidly expanding to establish comprehensive genomic resources across this diverse lineage (Rhie et al., 2021). These species exhibit remarkable variation in morphology, physiology, and ecological adaptation, shaped by heterogeneous environments and large-scale biogeographic processes (Jetz and Fine, 2012). Such phenotypic and ecological diversity is also reflected at the genomic level, manifesting as variation driven by environmental and developmental selective pressures, which in turn contributes to functional diversification of genomes (Mérot et al., 2020). These genomic changes have facilitated the diversification of gene repertoires and regulatory systems, ultimately leading to the evolution of diverse biomolecules, including enzymes, structural proteins, signaling molecules, and transcription factors (TFs) (Hellmuth and Stadler, 2024; Meadows and Lindblad-Toh, 2017).

Among these biomolecules, TFs are sequence-specific DNA-binding proteins that regulate gene expression and play central roles in controlling cell fate, tissue homeostasis, and other biological processes in jawed vertebrates (Vaquerizas et al., 2009; Wilkinson et al., 2017). Based on the structural characteristics of their DNA-binding domains, jawed vertebrate TFs can be divided into 73 distinct families (Shen et al., 2023), such as bZIP (Amoutzias et al., 2006), AP2 (Eckert et al., 2005), HSF (D. Westerheide et al., 2012), and GATA (Tang et al., 2014). This diversity reflects the need for sequence-specific recognition of a wide range of cis-regulatory elements and the coordination of complex gene regulatory networks underlying cellular identity and organismal development (Friedman et al., 2024; Hajheidari and Huang, 2022). Accordingly, TFs exert their regulatory functions by binding to cis-regulatory regions, including promoters and enhancers, to orchestrate context-dependent gene expression programs across diverse biological processes.

Within the diverse TF families, the jawed vertebrate GATA zinc finger (*zf-GATA*) genes contain one or two GATA domains that recognize the WGATAR (W = T/A, R = G/A) motif in target promoters (Shen et al., 2023). Among them, *GATA1-6* constitute an evolutionarily conserved group that plays crucial roles in the development and function of various tissues (Gillis et al., 2009; He et al., 2007). Within these, the *GATA1-3* subfamily is essential for hematopoiesis and neural development (Tsarovina et al., 2004), whereas *GATA4-6*, predominantly expressed in the heart, stomach, and gonads, directs cardiogenic endoderm and mesoderm differentiation (Tremblay et al., 2018). In teleost fishes, the teleost-specific third round of whole-genome duplication generated duplicated GATA paralogs, including *GATA1A/1B and GATA2A/2B* (Gillis et al., 2009). In addition, at least 12 jawed vertebrate *zf-GATA* genes have been identified, including *GATAD1/2A/2B/2AB* (Liang et al., 2021), *RERE/REREA-B* (Plaster et al., 2007), *MTA1-3* (Kumar and Wang, 2016), *ZGLP1* (Dong et al., 2019), *and TRPS1* (Saeki et al., 2026). These genes harbor GATA domains yet are involved in diverse biological processes, including transcriptional regulation, chromatin remodeling, and developmental processes, and may also modulate the activity of *GATA1-6* (Dong et al., 2019; Kumar and Wang, 2016; Liang et al., 2021; Plaster et al., 2007). For example, *RERE* functions as a nuclear co-regulator of *GATA4*, and the presence of similar cardiac phenotypes in deficient embryos suggests that *GATA4* acts as a downstream target of *RERE* during heart development (Kim et al., 2018). Together, these findings demonstrate that jawed vertebrate *zf-GATA* genes have diversified into functionally distinct regulators that contribute to transcriptional control and tissue-specific development.

Despite their essential roles in development and physiology, the evolutionary mechanisms underlying the structural diversification and relationships of *zf-GATA* genes across jawed vertebrate lineages remain poorly understood. To address this knowledge gap, a genome-scale comparative and phylogenetic analysis was performed using 2,809 *zf-GATA* genes from 181 jawed vertebrate species (Fig. 1). This study investigated how conserved molecular features and lineage-specific innovations have shaped *zf-GATA* evolution by integrating analyses of protein architectures, GATA domain characteristics, conserved domain organization, phylogenetic relationships, and structural features. The results provide insights into how *zf-GATA* genes have diversified through coordinated changes in gene repertoires, protein architectures, and conserved DNA-binding modules while maintaining the molecular framework required for GATA-mediated regulation. Collectively, this study establishes an evolutionary framework for understanding the balance between molecular conservation and structural diversification that has shaped the *zf-GATA* family across jawed vertebrates.

2. Materials and methods

2 – 1. Jawed vertebrate species selection and taxonomic information

AnimalTFDB4 (https://guolab.wchscu.cn/AnimalTFDB4) (Shen et al., 2023) provides TF annotations for 183 animal species based on genome assemblies and gene annotations from Ensembl Release 105 (Yates et al., 2026). Among these, 181 jawed vertebrate species were selected for this study (Supplementary Table S1). These species comprised a total of 3,885,519 predicted protein-coding genes (mean ± SD = 21,467 ± 5,867), 273,392 predicted TF-coding genes (mean ± SD = 1,510 ± 499), and 2,809 *zf-GATA* genes (mean ± SD = 15.5 ± 4.8).

In addition, NCBI Taxonomy IDs were collected for all selected jawed vertebrate species to obtain standardized taxonomic information (Schoch et al., 2020) (Supplementary Table S1). Furthermore, the overall phylogenetic relationships among jawed vertebrate lineages were interpreted based on a previously published time-calibrated phylogenetic tree (Irisarri et al., 2017). Accordingly, the 181 jawed vertebrate species were assigned to 13 groups: (i) *Primates* (n = 24), (ii) *Rodents* (n = 21), (iii) *Laurasiatheria* (n = 37), (iv) *Afrotheria* (n = 3), (v) *Xenarthra* (n = 2), (vi) *Marsupialia* (n = 5), (vii) *Yinotheria* (n = 1), (viii) *Amphibians* (n = 1), (ix) *Reptiles* (n = 13), (x) *Aves* (n = 12), (xi) *Coelacanths* (n = 1), (xii) *Actinopterygii* (n = 60), and (xiii) *Chondrichthyes* (n = 1). Groups (i-vii) represent major mammalian evolutionary lineages, whereas groups (viii-xiii) correspond to major non-mammalian jawed vertebrate lineages. Within *Mammalia*, these groups were defined according to major evolutionary lineages represented at different taxonomic ranks, including orders (*Primates and Rodents*), superorders (*Laurasiatheria, Afrotheria, and Xenarthra*), infraclasses (*Marsupialia*), and subclasses (*Yinotheria*). In contrast, *non-mammalian* vertebrates were represented at the class level. In this study, *Reptiles* refer only to non-avian reptiles; *Aves* are treated as a separate group. Based on the NCBI Taxonomy and the phylogenetic relationships among jawed vertebrate lineages, the selected species were further assigned to 53 taxonomic orders (Irisarri et al., 2017; Schoch et al., 2020).

2–2. Identification of jawed vertebrate genes

2–2. Identification of jawed vertebrate *zf-GATA* genes
A total of 2,809 *zf-GATA* genes from 181 jawed vertebrate species were retrieved from AnimalTFDB4 (Supplementary Table S2). AnimalTFDB4 identifies *zf-GATA* genes using profile hidden Markov model (HMM)-based searches (Eddy, 1998) against the Pfam database (version 35.0) (Finn et al., 2014). In AnimalTFDB4, *zf-GATA* genes were defined based on the presence of the GATA DNA-binding domain (Pfam accession: PF00320), which was identified using a family-specific HMM profile with an E-value threshold of 1 × 10^-4. For each gene, the longest protein isoform was analyzed using PfamScan, and domains with at least 70% coverage were retained for functional annotation and domain architecture characterization.

The protein-level *zf-GATA* dataset was subsequently consolidated into gene-level entries to remove transcript- and protein-isoform redundancy. Gene symbols and corresponding annotations were retained from the original AnimalTFDB4 records, which were derived from Ensembl and NCBI annotation resources. Among the 2,809 *zf-GATA* genes, 2,552 genes were assigned to 22 annotated *zf-GATA* gene groups, including *GATA1-6, GATA1A/1B/2A/2B, GATAD1/D2A/D2B/D2AB, RERE/REREA-B, MTA1-3, ZGLP1, and TRPS1*. The remaining 257 *zf-GATA* genes could not be assigned to any of the annotated gene groups and were retained as unclassified *zf-GATA* genes for subsequent analyses where appropriate. These AnimalTFDB4-derived *zf-GATA* annotations were subsequently subjected to independent domain-level re-annotation.

2–3. Re-annotation and refinement of jawed vertebrate genes

2–3. Re-annotation and refinement of jawed vertebrate *zf-GATA* genes
Independent domain-level re-annotation was performed to validate previous domain assignments and refine structural annotations of the 2,809 *zf-GATA* genes. Protein sequences of 2,809 *zf-GATA* genes were analyzed using InterProScan (version 5.74–105.0) (Jones et al., 2014), and Pfam domain matches were retrieved using the Pfam database (version 37.3) (Finn et al., 2014).

Among the 16 Pfam domain types identified, eight domain types (PF00320, PF01448, PF01426, PF17226, PF00249, PF05349, PF16563, and PF03154) were retained for subsequent comparative analyses of domain architecture. These domains corresponded to the GATA zinc finger, ELM2, BAH, MTA R1, Myb-like DNA-binding, GATA-type transcription activator N-terminal, P66A/P66B interaction, and Atrophin-1 family domains, respectively. The remaining eight domain types (PF00069, PF01391, PF02363, PF03451, PF05485, PF14970, PF23409, and PF23414) were excluded from downstream analyses because each was detected in three or fewer *zf-GATA* genes across the dataset. Due to their extremely low occurrence, these domains were considered low-frequency domain features and were not included in comparative domain architecture analyses. The retained domain types represented distinct domain regions within gene sequences, with no substantial overlap among their Pfam domain coordinates.

2–4. Quality assessment of jawed vertebrate genes

2–4. Quality assessment of jawed vertebrate *zf-GATA* genes
Among the 2,809 jawed vertebrate *zf-GATA* genes, dataset quality was assessed for 2,552 genes assigned to 22 annotated *zf-GATA* gene groups by evaluating the consistency of gene group classification and structural annotation (Supplementary Table S3). The remaining 257 unclassified *zf-GATA* genes were excluded because group-specific structural consistency could not be assessed.

First, protein length distributions were analyzed to assess gene group consistency and identify potential annotation inconsistencies, such as protein truncation or incomplete sequence models. Protein length deviations were identified using a gene group-specific interquartile range (IQR)-based approach, a commonly used non-parametric method for detecting distributional outliers in biological datasets (Kueckelhaus et al., 2024; Odamaki et al., 2016). Following the conventional 1.5 × IQR criterion, proteins with lengths outside the range of Q1 − 1.5 × IQR and Q3 + 1.5 × IQR were considered potential length deviations.

Second, domain architectures were evaluated to assess the structural organization of jawed vertebrate *zf-GATA* genes. For each *zf-GATA* gene group, the representative domain architecture was defined as the most frequently observed combination and organization of conserved domains within the group. Subsequently, individual *zf-GATA* genes were compared with their corresponding representative domain architecture. Genes consistent with the representative domain architecture were classified as structurally consistent, whereas those with divergent domain architectures were classified as domain architecture deviations. Genes lacking a detectable GATA domain (PF00320) in the re-annotation analysis were classified as unresolved because domain-level comparison could not be reliably performed.

Finally, GATA zinc finger motif (hereafter, GATA motifs) structures were evaluated based on the sequence integrity and organization of conserved Cys-based zinc-binding motifs according to the characteristic patterns of each *zf-GATA* gene group. For *GATA1-6 and GATA1A/1B/2A/2B* groups, genes were evaluated based on the presence of two Cys-X_2-Cys-X_{17}-Cys-X_2-Cys motifs. For *GATAD1/D2A/D2B/D2AB, MTA1-3, ZGLP1, and TRPS1* groups, genes were assessed based on the presence of a single Cys-X_2-Cys-X_{17}-Cys-X_2-Cys motif. For *RERE/REREA-B* groups, genes were evaluated based on the presence of a single Cys-X_2-Cys-X_{18}-Cys-X_2-Cys motif. Genes retaining the expected motif configuration were classified as structurally consistent, whereas those with motif-level alterations, including altered cysteine spacing, partial motifs, or loss of GATA motifs, were classified as motif-level deviations.

2–5. Classification of jawed vertebrate GATA motifs

GATA motifs within jawed vertebrate GATA domains were classified according to cysteine spacing patterns. Previous studies in plants and fungi have demonstrated that GATA motifs can be classified into multiple types based on variations in cysteine spacing patterns (Hu et al., 2025; Kim, 2024). In jawed vertebrates, GATA motifs were classified into eight types, including one predominant type and seven less frequently observed types. The predominant type was designated as Type IVa (Cys-X_2-Cys-X_{17}-Cys-X_2-Cys), representing the most frequently observed GATA motif configuration. The seven less frequently observed types included Type IVb (Cys-X_2-Cys-X_{18}-Cys-X_2-Cys), Type IV19 (Cys-X_2-Cys-X_{19}-Cys-X_2-Cys), Type IVe (Cys-X_2-Cys-$_{\leq 16,21\geq}$-Cys-X_2-Cys, where $n \leq 16$ or $n \geq 21$), Type IVp (partial motifs, e.g., Cys-X_2-Cys-X_{12}), atypical forms (e.g., Cys-X-Cys-X_{20}-Cys-X_2-Cys), Type IV4 (Cys-X_4-Cys-X_{18}-Cys-X_2-Cys), and Type IVc (Cys-X_2-Cys-X_{20}-Cys-X_2-Cys).

2–6. Co-occurrence analysis of GATA motifs and conserved domains across jawed vertebrate groups

To investigate the relationships between GATA motif types and conserved domains, co-occurrence patterns were analyzed across 13 taxonomic groups of jawed vertebrates. Among the 2,809 *zf-GATA* genes, 158 genes lacking a detectable GATA domain (PF00320) after re-annotation were excluded, and the remaining 2,651 genes were used for the co-occurrence analysis. For each taxonomic group, individual *zf-GATA* genes were decomposed into observed combinations of GATA motif types and domains co-occurring within each gene. Redundant combinations within individual *zf-GATA* genes were removed to retain unique domain-motif associations.

2–7. Phylogenetic analysis of representative gene groups in jawed vertebrates

2–7. Phylogenetic analysis of representative *zf-GATA* gene groups in jawed vertebrates
Phylogenetic relationships among 22 jawed vertebrate *zf-GATA* gene groups were analyzed using NGPhylogeny (https://ngphylogeny.fr) (Lemoine et al., 2019) (Supplementary Table S4). Representative protein sequences for each *zf-GATA* gene group were collected from five jawed vertebrate species: *Homo sapiens* (n = 13), *Danio rerio* (n = 6), *Denticeps clupeoides* (n = 1), *Laticauda laticaudata* (n = 1), *and Mus musculus* (n = 1). The Advanced (semi-automatic) workflow was employed. Protein sequences in FASTA format were aligned with MAFFT, and poorly aligned regions were trimmed using BMGE, both with default settings. Maximum-likelihood (ML) trees were then reconstructed using PhyML with 1,000 bootstrap replicates. Tree layout was examined in a phylogram-based linear format, with display support values and aligned text options enabled.

2–8. Three-dimensional structural modeling of jawed vertebrate genes

2–8. Three-dimensional structural modeling of jawed vertebrate *zf-GATA* genes
Three-dimensional structures of the same 22 representative jawed vertebrate *zf-GATA* genes used for the phylogenetic analysis (Section 2–7; Supplementary Table S4) were predicted using the AlphaFold framework (Bertoline et al., 2023). Protein sequences of the selected *zf-GATA* genes were used as input without truncation to preserve full-length structural context. AlphaFold produces a per-residue confidence score (pLDDT) ranging from 0 to 100. Regions with low pLDDT scores were considered potentially disordered or structurally unresolved.

2–9. Phylogenetic analysis of jawed vertebrate GATA motifs

Phylogenetic analysis of jawed vertebrate GATA motifs was performed to investigate the evolutionary relationships among annotated *zf-GATA* gene groups. Among the 2,809 *zf-GATA* genes identified in this study, 2,552 genes assigned to 22 *zf-GATA* gene groups were selected for motif-level phylogenetic analysis. A total of 3,608 GATA motifs derived from these group-assigned *zf-GATA* genes were collected and subjected to multiple sequence alignment using MAFFT v7.505 (Katoh and Standley, 2013). Subsequently, a maximum likelihood phylogenetic tree was constructed using IQ-TREE2 v2.0.7 (Minh et al., 2020) with the Dayhoff model, and branch support was evaluated using 1,000 ultrafast bootstrap replicates (-bb 1000). The resulting phylogenetic tree was visualized and annotated in the Interactive Tree of Life (iTOL) platform (https://itol.embl.de/) (Letunic and Bork, 2024).

3. Results

3 – 1. Validation and characterization of the jawed vertebrate zf-GATA gene repertoire

3-1-1. Independent validation of GATA domain annotations

In this study, 2,809 *zf-GATA* genes from 181 jawed vertebrate species were collected from AnimalTFDB4. To independently validate the domain annotations of the collected *zf-GATA* genes, all protein sequences were re-annotated using InterProScan based on the Pfam database. As a result, GATA domains were identified in 2,651 of the 2,809 *zf-GATA* genes (94.4%), indicating that the majority of the collected genes retained the characteristic GATA DNA-binding domain. The remaining 158 genes (5.6%) lacked detectable GATA domains, most of which belonged to the *GATAD1* gene group, suggesting that these cases may reflect differences in HMM-based detection sensitivity rather than complete loss of GATA zinc finger structures. Sequence inspection further showed that most of these genes retained a Cys-X_2-Cys-X_{17}-Cys-X_2-Cys motif, whereas a small number exhibited alternative configurations, including Cys-X_2-Cys-X_{18}-Cys-X_2-Cys and Cys-X_2-Cys-X_{19}-Cys-X_2-Cys motifs. Together, these findings support the reliability of the AnimalTFDB4-derived *zf-GATA* dataset and indicate that motif-level conservation can complement domain-based annotation for identifying divergent *zf-GATA* genes with reduced domain detectability.

3-1-2. Quality assessment of *zf-GATA* gene group assignments
Among the 2,809 *zf-GATA* genes, 2,552 genes were assigned to 22 *zf-GATA* gene groups and subjected to gene group-specific quality assessment. The reliability of the jawed vertebrate *zf-GATA* dataset was evaluated using three complementary criteria: protein length consistency, domain architecture conservation, and GATA motif preservation (Table 1). The remaining 257 unassigned genes were excluded because representative structural features could not be reliably defined.

Table 1
Summary of quality assessment of jawed vertebrate *zf-GATA* genes

QC category	Criterion	Pass	Not passed	Unresolved	Pass (%)
Protein length	Within gene-specific IQR range	2,211	341	-	86.6
Domain architecture	Consistent with representative gene-specific architecture	2,141	261	150	89.1*
GATA motif conservation	Conserved Cys-based zinc finger motif retained	2,451	101	-	96.0
Overall QC	**All three criteria satisfied**	**1,868**	**534**	**-**	**73.2**

* Pass percentages were calculated based on the total number of genes assigned to the 22 annotated *zf-GATA* gene groups (n = 2,552). Not passed indicates genes that did not meet the corresponding QC criterion. Unresolved genes lacked sufficient domain annotation evidence for direct architecture comparison.

First, protein length variation was assessed to identify potential inconsistencies within each *zf-GATA* gene group. Of the 2,552 *zf-GATA* genes, 2,211 genes (86.6%) exhibited protein lengths within the gene group-specific IQR range, whereas 341 genes (13.4%) showed length deviations beyond the defined thresholds. Across individual *zf-GATA* gene groups, the proportion of genes within the expected protein length range varied from 63.0% (*TRPS1*) to 100.0% (*GATA1B/2B*), with a mean ± SD of 86.6% ± 10.5%, indicating moderate variation among gene groups while maintaining overall protein length consistency. These findings indicate overall conservation of protein length with lineage-specific variation among *zf-GATA* genes.

Second, domain architecture was evaluated by comparing each gene's architecture with the most frequently observed architecture in its corresponding *zf-GATA* gene group. Within the evaluated dataset, 150 genes were classified as unresolved because GATA domains were not detected during re-annotation, preventing reliable domain architecture comparison. The remaining 2,402 genes included 2,141 genes (89.1%) with domain architectures consistent with their corresponding gene group representatives, whereas 261 genes exhibited altered domain organizations. Across individual *zf-GATA* gene groups, the proportion of genes with conserved domain architectures ranged from 35.9% (*REREA*) to 100.0% (*GATA1A/1B/2B, GATAD1, TRPS1, and ZGLP1*), with a mean ± SD of 89.1% ± 18.0%, indicating substantial variation in domain architecture conservation among *zf-GATA* gene groups. These findings demonstrate overall conservation of *zf-GATA* domain architectures with variable degrees of gene group-specific structural diversification.

Third, GATA motif conservation was assessed based on the presence and organization of conserved Cys-based motifs within *zf-GATA* genes. The analysis of the 2,552 *zf-GATA* genes showed that 2,451 genes (96.0%) retained conserved GATA motif structures, whereas 101 genes (4.0%) showed deviations from the representative motif organization. Across individual *zf-GATA* gene groups, the proportion of genes retaining conserved GATA motif structures ranged from 80.8% (*GATAD2A*) to 100.0% (*GATA1A/1B/2A, GATAD1/D2B, TRPS1, and ZGLP1*), with a mean ± SD of 96.0% ± 5.5%, indicating strong conservation of GATA motifs across diverse *zf-GATA* gene groups. These findings demonstrate strong conservation of the core Cys-based zinc-binding motif despite structural diversification among jawed vertebrate *zf-GATA* genes.

Integration of the three complementary quality assessment criteria showed that 1,868 of the 2,552 *zf-GATA* genes (73.2%) satisfied all three criteria, including conserved protein length characteristics, representative gene-specific domain architectures, and preserved GATA motifs. The remaining genes exhibited deviations in one or more structural features, primarily involving protein length variation, altered domain organization, or unresolved domain annotations. These deviations likely reflect a combination of lineage-specific structural diversification and limitations of sequence-based domain detection rather than widespread annotation errors. Collectively, these results demonstrate that the *zf-GATA* dataset preserves core structural features while capturing lineage-specific architectural diversity, providing a reliable foundation for subsequent comparative, structural, and phylogenetic analyses.

3 – 2. Structural landscape of jawed vertebrate *zf-GATA* genes

3-2-1. Taxonomic distribution of *zf-GATA* gene repertoires

The taxonomic distribution of 2,809 *zf-GATA* genes was examined across 181 jawed vertebrate species representing 13 major vertebrate groups and 53 taxonomic orders (Table 2 and Fig. 2a–2c). *Mammalian* groups generally exhibited comparable numbers of *zf-GATA* genes, ranging from an average of 10.0 ± 5.7 genes per species in *Xenarthra* to 14.6 ± 2.3 in *Marsupialia*, with *Primates, Rodents, Laurasiatheria, and Afrotheria* averaging between 12 and 14 genes per species. *Non-mammalian* groups also exhibited comparable numbers of *zf-GATA* genes, including *Amphibians* (14.0), *Reptiles* (12.8 ± 1.5), *Aves* (11.2 ± 2.1), *Coelacanths* (14.0), *and Chondrichthyes* (11.0). Among these *non-mammalian* groups, *Actinopterygii* contained the largest *zf-GATA* repertoire, averaging 20.0 ± 5.4 genes per species, substantially exceeding all other jawed vertebrate groups. Overall, *zf-GATA* gene abundance showed relatively conserved patterns across most jawed vertebrate groups, whereas *Actinopterygii* exhibited lineage-specific expansion of the *zf-GATA* repertoire.

Table 2
Distribution and abundance of jawed vertebrate *zf-GATA* genes

Group	Number of orders	Number of species	Total number of genes (Average ± SD/species)	Total number of TFs (Average ± SD/species)	Total number of *zf-GATAs* (Average ± SD/species)
Primates	1	24	492,217 (20,509 ± 1,279)	36,243 (1,510 ± 98)	342 (14.3 ± 1.6)
Rodents	3	21	410,724 (19,558 ± 2,177)	27,349 (1,302 ± 174)	282 (13.4 ± 2.5)
Laurasiatheria	5	37	716,807 (19,373 ± 2,240)	49,115 (1,327 ± 202)	505 (13.6 ± 2.6)
Afrotheria	3	3	52,665 (17,555 ± 2,162)	3,367 (1,122 ± 59)	37 (12.3 ± 2.5)
Xenarthra	2	2	35,104 (17,552 ± 7,296)	2,317 (1,159 ± 504)	20 (10.0 ± 5.7)
Marsupialia	3	5	96,685 (19,337 ± 2,469)	7,074 (1,415 ± 253)	73 (14.6 ± 2.3)
Yinotheria	1	1	17,417 (17,417 ± 0)	1,162 (1,162 ± 0)	11 (11.0 ± 0.0)
Amphibians	1	1	19,983 (19,983 ± 0)	1,368 (1,368 ± 0)	14 (14.0 ± 0.0)
Reptiles	4	13	248,720 (19,132 ± 1,951)	17,347 (1,334 ± 194)	166 (12.8 ± 1.5)
Aves	6	12	187,614 (15,635 ± 860)	11,843 (987 ± 105)	134 (11.2 ± 2.1)
Coelacanths	1	1	19,569 (19,569 ± 0)	1,165 (1,165 ± 0)	14 (14.0 ± 0.0)
Actinopterygii	22	60	1,568,599 (26,143 ± 7,774)	113,962 (1,899 ± 655)	1,200 (20.0 ± 5.4)
Chondrichthyes	1	1	19,415 (19,415 ± 0)	1,080 (1,080 ± 0)	11 (11.0 ± 0.0)
Total	**53**	**181**	**3,885,519 (21,467 ± 5,867)**	**273,392 (1,510 ± 499)**	**2,809 (15.5 ± 4.8)**

3-2-2. Diversity of domain architectures

To investigate structural diversification of jawed vertebrate *zf-GATA* genes, conserved protein domains and their combinations were analyzed across all genes. Among the 2,809 jawed vertebrate *zf-GATA* genes, 7,462 conserved domains were identified and classified into eight domain types (Fig. 2d). The GATA domain (PF00320), a defining feature of *zf-GATA* genes, was identified in 3,773 occurrences across the analyzed dataset. In addition, seven conserved domains were detected, including the ELM2 domain (PF01448; 699 occurrences), BAH domain (PF01426; 667 occurrences), MTA R1 domain (PF17226; 553 occurrences), Myb-like DNA-binding domain (PF00249; 531 occurrences), GATA-type transcription activator N-terminal domain (PF05349; 506 occurrences), P66A/P66B interaction domain (PF16563; 432 occurrences), and Atrophin-1 family domain (PF03154; 301 occurrences). These conserved domains represent additional structural modules associated with diverse regulatory functions, including chromatin remodeling, DNA binding, and protein–protein interactions. Overall, these findings demonstrate that jawed vertebrate *zf-GATA* genes share a conserved GATA domain framework while incorporating diverse additional structural modules associated with regulatory functions.

Based on domain combinations, the 2,809 jawed vertebrate *zf-GATA* genes were classified into six major and 45 minor domain architectures (Table 3 and Fig. 2e). The most prevalent architectures contained two tandem GATA domains, including the two GATA domain architecture (PF00320 + PF00320; n = 641) and the GATA-type transcription activator N-terminal domain combined with two GATA domains (PF05349 + PF00320 + PF00320; n = 470). The predominance of these architectures suggests that duplicated GATA domains represent a conserved structural framework maintained across jawed vertebrate *zf-GATA* genes. The single GATA domain architecture (PF00320 only; n = 374) also represented a major domain combination, indicating that *zf-GATA* genes can retain a simplified structural organization centered on a single GATA domain. Additional major architectures combined a single GATA domain with diverse regulatory domains, including PF01426 + PF01448 + PF00249 + PF00320 + PF17226 (n = 420), PF16563 + PF00320 (n = 417), and PF01426 + PF01448 + PF00320 + PF03154 (n = 78), highlighting structural diversification of *zf-GATA* proteins through the incorporation of regulatory modules beyond the core GATA domain. Overall, these findings demonstrate that jawed vertebrate *zf-GATA* genes conserve a GATA domain-centered framework while exhibiting architectural diversification through distinct domain organizations and combinations of regulatory modules.

Table 3
Domain architectures of jawed vertebrate *zf-GATA* genes

Group	Number of species	Number of *zf-GATAs*	Domain architecture						
			PF00320 + PF00320	PF05349 + PF00320+PF00320	PF01426 + PF01448+PF00249+ PF00320 + PF17226	PF16563 + PF00320	PF00320 only	PF01426 + PF01448+ PF00320 + PF03154	The other 45 minor architectures
Primates	24	342	75	61	69	50	46	20	21
Rodents	21	282	64	42	50	33	44	16	33
Laurasiatheria	37	505	107	84	84	68	81	12	69
Afrotheria	3	37	8	6	2	5	5	1	10
Xenarthra	2	20	5	3	2	3	2	1	4
Marsupialia	5	73	13	14	11	12	10	0	13
Yinotheria	1	11	3	3	2	2	1	0	0
Amphibians	1	14	3	3	2	2	2	0	2
Reptiles	13	166	36	35	22	25	22	3	23
Aves	12	134	29	28	25	15	14	0	23
Coelacanths	1	14	3	2	1	2	1	1	4
Actinopterygii	60	1,200	293	186	149	199	144	23	206
Chondrichthyes	1	11	2	3	1	1	2	1	1
Total	**181**	**2,809**	**641**	**470**	**420**	**417**	**374**	**78**	**409**

*** List of major 8 domains:** PF00320 (GATA zinc finger), PF01448 (ELM2 domain), PF01426 (BAH domain), PF00249 (Myb-like DNA-binding domain), PF17226 (MTA R1 domain), PF05349 (GATA-type transcription activator, N-terminal), PF16563 (Coiled-coil and interaction region of P66A and P66B with MBD2), and PF03154 (Atrophin-1 family).

To investigate lineage-specific patterns of domain architecture diversification, the distribution of major *zf-GATA* architectures was compared across 13 jawed vertebrate groups (Table 3). Across all groups, tandem GATA domain architectures (PF00320 + PF00320 and PF05349 + PF00320 + PF00320) were consistently observed, accounting for 17.8–27.3% and 14.3–27.3% of *zf-GATA* genes, respectively, highlighting the broad conservation of duplicated GATA domain-containing frameworks across lineages. The single GATA domain architecture was also widely distributed, accounting for 9.1–18.2% of *zf-GATA* genes, indicating the preservation of simplified GATA domain-centered organizations. In contrast, regulatory domain-associated architectures exhibited lineage-specific variation, with PF01426 + PF01448 + PF00249 + PF00320 + PF17226 (5.4–20.2%) and PF16563 + PF00320 (9.1–18.2%) showing broader distributions across groups, whereas PF01426 + PF01448 + PF00320 + PF03154 (0.0–9.1%) displayed more restricted occurrence. Collectively, jawed vertebrate *zf-GATA* genes retain conserved GATA-centered architectures with lineage-specific regulatory diversification. However, differences in taxonomic sampling and genomic coverage may influence these patterns, highlighting the need for broader lineage sampling.

3-2-3. Diversity and conservation of GATA motifs

GATA motifs are conserved four-cysteine zinc-binding structures within jawed vertebrate GATA domains that mediate DNA recognition and zinc coordination. To characterize the diversity of GATA motifs in jawed vertebrates, the distribution of motif configurations was analyzed among 3,931 GATA domains identified from 2,809 *zf-GATA* genes.

Among the 3,931 identified GATA domains, 3,608 (91.8%) contained the Type IVa motif (Cys-X_2-Cys-X_{17}-Cys-X_2-Cys), highlighting its status as the predominant conserved signature of jawed vertebrate GATA domains (Table 4 and Fig. 2f). In addition, diverse motif variants were identified, exhibiting expansions or reductions in the inter-cysteine spacing within the conserved GATA motif framework. For example, motif variants with altered inter-cysteine spacing were identified, including Type IVb (Cys-X_2-Cys-X_{18}-Cys-X_2-Cys; 222 motifs), Type IV19 (Cys-X_2-Cys-X_{19}-Cys-X_2-Cys; 46 motifs), Type IVc (Cys-X_2-Cys-X_{20}-Cys-X_2-Cys; 3 motifs), and Type IVe (Cys-X_2-Cys-$_{\leq 16,21\geq}$-Cys-X_2-Cys; 25 motifs), suggesting limited structural diversification within the conserved GATA motif framework. Furthermore, Type IVp motifs lacking complete Cys configurations (partial motifs; 20 motifs) and atypical forms retaining Cys residues but exhibiting distinct structural patterns (4 motifs) were also identified. Overall, these findings demonstrate that jawed vertebrate GATA motifs are highly conserved zinc-binding structures while permitting limited structural diversification through rare motif-level modifications.

Table 4
GATA motif types of jawed vertebrate *zf-GATA* genes

Group	Number of species	Number of *zf-GATAs*	Number of GATA domains	GATA Motif Types							
				IVa	IVb	IV19	IVe	IVp	Atypical form	IV4	IVc
Primates	24	342	479	456	23	0	0	0	0	0	0
Rodents	21	282	391	362	20	0	8	1	0	0	0
Laurasiatheria	37	505	699	657	31	0	7	3	1	0	0
Afrotheria	3	37	51	47	3	0	0	1	0	0	0
Xenarthra	2	20	28	27	1	0	0	0	0	0	0
Marsupialia	5	73	100	95	4	0	0	1	0	0	0
Yinotheria	1	11	17	17	0	0	0	0	0	0	0
Amphibians	1	14	20	18	1	0	0	0	0	0	1
Reptiles	13	166	237	225	11	0	0	0	1	0	0
Aves	12	134	192	179	12	0	0	1	0	0	0
Coelacanths	1	14	19	18	1	0	0	0	0	0	0
Actinopterygii	60	1,200	1,682	1,492	114	46	10	13	2	3	2
Chondrichthyes	1	11	16	15	1	0	0	0	0	0	0
Total	**181**	**2,809**	**3,931**	**3,608**	**222**	**46**	**25**	**20**	**4**	**3**	**3**

*** List of GATA motifs:** Type IVa ($CX_2CX_{17}CX_2C$), Type IVb ($CX_2CX_{18}CX_2C$), Type IV19 ($CX_2CX_{19}CX_2C$), Type IVe ($CX_2CX_{-16,\ 21}CX_2C$), Type IVp (Partial), atypical form (e.g., $CXCX_{20}CX_2C$), Type IV4 ($CX_4CX_{18}CX_2C$), and Type IVc ($CX_2CX_{20}CX_2C$).

To examine lineage-specific patterns of GATA motif diversification, the distribution of motif configurations was compared across 13 jawed vertebrate groups (Table 4 and Fig. 2f). The predominant Type IVa motif was highly conserved across all groups, accounting for 88.7–100.0% of GATA domains, with most lineages exhibiting frequencies above 90%. Type IVb was the most frequently observed variant after Type IVa, accounting for 3.6–6.8% of GATA domains in most groups, although it was absent in *Yinotheria*. Other motif variants were rarely detected across lineages, with Type IV19, Type IVe, Type IVp, atypical forms, Type IV4, and Type IVc collectively representing minor fractions of GATA domains. Notably, *Actinopterygii* showed the lowest Type IVa proportion (88.7%) and the greatest diversity of motif configurations, including the presence of Type IV19 (2.7%) and multiple rare motif variants. In contrast, *mammalian* groups generally maintained high Type IVa conservation (92.2–100.0%) with limited motif diversification. These findings demonstrate that GATA motifs are strongly conserved across jawed vertebrates, with structural diversification occurring mainly through rare modifications of the conserved Type IVa framework. However, expanded genomic sampling across underrepresented lineages will further improve comparative resolution of lineage-specific motif diversification patterns.

3-2-4. Conservation and divergence of GATA motifs

Despite their conserved four-cysteine zinc-binding framework, amino acid conservation and divergence within GATA motifs remain poorly characterized across jawed vertebrates. Characterizing residue-level conservation can distinguish positions subject to strong evolutionary constraints from those that tolerate evolutionary variation, thereby providing insight into the structural and functional evolution of GATA motifs. Therefore, amino acid variation within the predominant Type IVa motif was analyzed to identify conserved and divergent features of *zf-GATA* evolution.

In Type IVa, which accounts for 3,608 of 3,931 GATA motifs, the four core cysteine residues at C-1, C-4, C-22, and C-25 that constitute the Cys-X_2-Cys-X_{17}-Cys-X_2-Cys zinc-binding framework are positionally conserved across jawed vertebrates (Fig. 3). Beyond these invariant cysteine residues, several amino acid positions exhibit moderate to high conservation across jawed vertebrates. Highly conserved residues (> 80%) included polar uncharged (T-9), special property (G-18), and hydrophobic (W-12) residues, whereas moderately conserved residues (> 60%) comprised polar charged (R-13 and R-14), polar uncharged (N-3, T-6, and N-23), and hydrophobic (L-11 and A-24) residues. These findings indicate that Type IVa motif conservation is maintained through the coordinated preservation of diverse amino acid properties beyond the invariant cysteine framework. Because differences in taxonomic sampling and genomic coverage may limit the interpretation of lineage-specific patterns, group-level comparisons were not performed.

3-2-5. Integration of GATA motif and conserved domain relationships

To integrate motif-level and domain-level evolutionary features, co-occurrence patterns of GATA motifs and conserved domains were analyzed across jawed vertebrate *zf-GATA* genes. Of the 2,809 *zf-GATA* genes initially identified, 2,651 genes with detectable GATA domains after independent re-annotation were included in this analysis. These genes were decomposed into individual GATA motif-domain pairs, and duplicate pairs were removed, resulting in 6,136 unique motif-domain matches across 13 taxonomic groups.

Across the 13 jawed vertebrate groups, GATA motif-domain co-occurrence patterns were highly conserved, with only minor lineage-specific differences observed among domain groups (Fig. 4). Despite subtle differences among domain groups, a consistent overall pattern was observed in the distribution of GATA motif types. Five domains (PF00249, PF00320, PF05349, PF16563, and PF17226) exhibited very high Type IVa prevalence, ranging from 88.0% to 98.8%. Similarly, two domains (PF01426 and PF01448) showed moderately high Type IVa frequencies, ranging from 72.2% to 74.3%, with Type IVb motifs representing most of the remaining configurations (23.2% to 25.3%). In contrast, the PF03154 domain was strongly associated with Type IVb, which accounted for 96.1% of all observed motifs in this domain. These findings indicate that most conserved domains preferentially co-occur with Type IVa motifs, whereas PF03154 exhibits a distinct association with Type IVb.

3–3. Evolutionary diversification of jawed vertebrate *zf-GATA* genes

3-3-1. Evolutionary relationships and molecular diversification of *zf-GATA* gene groups

To investigate the evolutionary relationships and molecular diversification of *zf-GATA* gene groups across jawed vertebrates, representative protein sequences from 22 annotated *zf-GATA* gene groups were analyzed. Among the 2,809 jawed vertebrate *zf-GATA* genes, 2,552 genes were assigned to these gene groups based on conserved sequence features and domain organizations.

The phylogenetic analysis of representative *zf-GATA* protein sequences revealed distinct evolutionary relationships among the 22 *zf-GATA* gene groups (Fig. 5a). The *GATA1-6* gene groups formed a major conserved clade, within which *GATA4-6* constituted a distinct subclade, indicating their close evolutionary relationship. In addition, teleost-specific paralogous groups, including *GATA1A/1B and GATA2A/2B*, were positioned within the *GATA1-6* clade, consistent with lineage-specific duplication events that expanded *zf-GATA* gene repertoires in ray-finned fishes. Outside the *GATA1-6* clade, *TRPS1* formed a separate lineage adjacent to the *GATA1-6* group, whereas *GATAD1/D2A/D2B/D2AB and ZGLP1* clustered into another distinct clade. The *MTA1-3* gene group formed a separate lineage adjacent to the *GATAD/ZGLP1* clade, while *RERE/REREA-B* represented the most divergent lineage among the analyzed *zf-GATA* groups. This phylogenetic organization was consistent with differences in structural architecture among *zf-GATA* groups, with *GATA1-6* genes retaining conserved GATA domain organizations and other *zf-GATA* groups exhibiting distinct domain architectures.

The distribution of *zf-GATA* genes among the 22 gene groups revealed substantial variation in repertoire size across evolutionary lineages (Fig. 5b). Among the analyzed *zf-GATA* groups, *GATAD2B* represented the largest group with 190 genes, followed by *GATA3* (183), *GATA5* (178), *MTA1* (171), *MTA2* (168), *GATA4* (168), *MTA3* (160), *and GATA6* (158). In contrast, several groups contained relatively fewer members, including *GATA1B* (4) and *GATA2B* (11). *GATA1* contained 90 genes, while the teleost-specific paralogous groups *GATA1A and GATA2A* were represented by 67 and 62 genes, respectively, with *GATA1B and GATA2B* showing more limited retention. *GATAD1/D2A, RERE*, and other gene groups also exhibited substantial representation across jawed vertebrates, highlighting marked variation in gene group abundance. Overall, these differences indicate lineage-specific diversification of *zf-GATA* gene repertoires across jawed vertebrates.

The distribution of conserved domains among *zf-GATA* gene groups revealed distinct structural signatures corresponding to their phylogenetic relationships (Fig. 5c-5e). *GATA1-6* gene groups were characterized by conserved GATA domain frameworks, with most members containing tandem GATA domains, whereas *GATA4-6* additionally retained the GATA-type transcription activator N-terminal domain, indicating subgroup-specific structural features. Teleost-specific paralogous groups, including *GATA1A/1B and GATA2A/2B*, largely preserved the domain organizations of their corresponding *GATA1/2* groups. *GATAD2A/D2B/D2AB* were characterized by the presence of both the Myb-like DNA-binding and GATA domains, whereas *GATAD1 and ZGLP1* primarily retained a simplified GATA domain-centered architecture. *TRPS1* also exhibited a conserved single GATA domain organization without additional frequently detected domains. In contrast, *MTA1-3 and RERE* lineages displayed more complex architectures comprising multiple conserved domains. *MTA* proteins consistently contained combinations of the ELM2, BAH, MTA R1, and Atrophin-1 family domains, whereas *RERE* proteins exhibited variable combinations of ELM2, BAH, Myb-like DNA-binding, P66A/P66B interaction, and Atrophin-related domains. Together, these patterns demonstrate substantial structural diversification among *zf-GATA* gene groups while preserving lineage-specific domain organizations.

The distribution of GATA motif types revealed strong conservation of the Type IVa motif across most *zf-GATA* gene groups, with lineage-specific diversification observed in several groups (Fig. 5f). *GATA1-6, TRPS1, GATAD1, ZGLP1, and GATAD2B* predominantly retained the Type IVa motif, indicating broad conservation of the characteristic Cys-based zinc-binding configuration. In contrast, the *RERE* lineage exhibited a distinct motif composition, with *RERE and REREA/B* primarily containing the Type IVb motif rather than Type IVa. *GATAD2A and GATAD2AB* also displayed limited diversification from the Type IVa framework, including alternative motif configurations such as Type IVe, Type IVp, and Type IV4. Similarly, *MTA1-3* groups predominantly retained the Type IVa motif while containing occasional atypical and Type IVc variants. Overall, these patterns indicate that the Type IVa motif is broadly conserved across *zf-GATA* gene groups, whereas lineage-specific motif diversification has occurred in a subset of evolutionary lineages.

3-3-2. Three-dimensional structural diversification of *zf-GATA* proteins

To determine whether the molecular diversification identified among *zf-GATA* gene groups is reflected in higher-order protein organization, representative full-length protein structures were predicted (Fig. 6). Representative proteins from the 22 annotated *zf-GATA* gene groups were analyzed to compare the spatial organization of GATA domains and associated conserved domains across evolutionary lineages. Integrating these structural models with the phylogenetic and domain analyses enabled evaluation of how lineage-specific protein architectures are translated into distinct three-dimensional structural organizations.

GATA1-6 genes, together with the teleost-specific paralogous groups *GATA1A/1B and GATA2A/2B*, consistently contained two GATA domains positioned at separate locations within the primary sequence. Despite this separation, structural modeling showed that the N-terminal and C-terminal GATA domains were positioned in close spatial proximity in the folded proteins. This conserved structural arrangement indicates that the tandem GATA domains maintain a common three-dimensional organization across these *zf-GATA* groups despite sequence divergence during vertebrate evolution.

Among the single GATA domain-containing proteins, *TRPS1 and ZGLP1* exhibited markedly different structural organizations despite sharing the same GATA domain multiplicity. *ZGLP1* adopted a relatively simple architecture centered on a single GATA domain with limited additional structured regions. In contrast,

TRPS1 displayed a substantially more complex predicted structure containing multiple well-defined regions with consistently moderate to high prediction confidence (pLDDT), indicating that extensive structural organization is maintained beyond the GATA domain itself. These observations suggest that the structural complexity of *zf-GATA* proteins is influenced not only by GATA domain number but also by the organization of surrounding protein regions.

The GATAD family (*GATAD1/D2A/D2B/D2AB*) exhibited relatively simple overall structures despite containing the conserved P66A/P66B interaction domain (PF16563) in addition to the GATA domain. Compared with other multi-domain *zf-GATA* proteins, these proteins retained compact structural organizations with fewer extensively folded regions. This pattern indicates that the acquisition of a limited number of accessory domains does not necessarily produce highly complex three-dimensional architectures but instead contributes localized structural modules while preserving an overall compact protein organization.

In contrast, the *MTA1-3 and RERE/REREA-B* groups exhibited highly complex predicted structures containing multiple well-defined globular regions corresponding to their diverse conserved domains. Consistent with their domain architectures, these proteins displayed extensive high-confidence structural regions, reflecting the integration of multiple conserved domains into a single protein framework. Collectively, these findings demonstrate that lineage-specific diversification of *zf-GATA* genes identified at the sequence and domain levels is also reflected in their predicted three-dimensional protein organizations, providing structural evidence for the diversification of *zf-GATA* proteins across jawed vertebrates.

3-3-3. Evolutionary conservation and diversification of GATA motifs

To further resolve the evolutionary relationships underlying the structural diversification observed among full-length *zf-GATA* proteins, phylogenetic analysis was performed using GATA motifs rather than complete protein sequences. Because full-length proteins contain lineage-specific accessory domains that may influence phylogenetic reconstruction, analysis of the GATA motif alone enables evaluation of the evolutionary history of the conserved DNA-binding module independent of additional structural components. A total of 3,608 GATA motifs identified from 2,552 annotated *zf-GATA* genes were analyzed to assess the evolutionary conservation and diversification of the GATA motif across jawed vertebrates (Fig. 7).

The resulting phylogeny resolved several well-defined clades corresponding to major *zf-GATA* gene groups and revealed a hierarchical organization of GATA motif diversity (Fig. 7). Near the basal region of the tree, GATA motifs from the *MTA1-3* groups formed the earliest distinct lineage, followed by domains from the *RERE/REREA-B* groups. These were followed by the *GATAD1/D2A/D2B* groups, whereas the C-terminal GATA motifs of the *GATA1-3 and GATA4-6* gene groups formed a separate, well-supported lineage. In contrast, the N-terminal GATA motifs of the *GATA1-3 and GATA4-6* gene groups clustered together with the single GATA motif of *TRPS1 and ZGLP1*, forming another distinct lineage. This topology indicates that the two GATA motifs retained in the *GATA1-6* gene groups have followed distinct evolutionary trajectories rather than evolving as equivalent duplicated domains. The close association of the N-terminal GATA motifs with *TRPS1 and ZGLP1* further suggests that these motifs share greater sequence similarity with single-motif *zf-GATA* proteins than with their corresponding C-terminal GATA motifs, highlighting evolutionary diversification within the conserved GATA DNA-binding module. Together, these findings indicate that the conserved GATA motif has undergone lineage-specific diversification while maintaining distinct evolutionary relationships among major *zf-GATA* gene groups across jawed vertebrates.

4. Discussion

4 – 1. Evolutionary constraint and modular diversification shape the zf-GATA family in jawed vertebrates

The balance between preservation of essential molecular functions and diversification of regulatory properties shapes TF evolution (Voordeckers et al., 2015). Recent advances in comparative genomics have enabled large-scale evolutionary analyses of protein domains and TF families, revealing how conserved molecular modules diversify while maintaining essential biological functions (Kim, 2026; Xiao et al., 2025; Xue et al., 2026). The *zf-GATA* family is a notable example of this evolutionary principle, as its members retain a conserved GATA zinc finger domain required for DNA recognition yet exhibit substantial variation in protein organization and regulatory potential (Gillis et al., 2009; Kumar and Wang, 2016). However, the evolutionary framework explaining how conserved and lineage-specific features of *zf-GATA* proteins are maintained and diversified across jawed vertebrate lineages remains unclear.

In this study, a genome-scale comparative analysis of 2,809 *zf-GATA* genes from 181 jawed vertebrate species demonstrated that *zf-GATA* evolution has been shaped by hierarchical constraints acting at multiple structural levels. The core GATA DNA-binding framework, including conserved GATA domains and characteristic cysteine-based zinc-binding motifs, was strongly preserved across jawed vertebrate lineages, reflecting the essential requirements for DNA recognition and transcriptional regulation. In contrast, surrounding regulatory domains, domain architectures, and gene repertoires exhibited greater evolutionary diversification among vertebrate lineages. These findings suggest that *zf-GATA* proteins have evolved by preserving conserved DNA-binding modules while diversifying less constrained regulatory components in a lineage-specific manner. Such hierarchical diversification allows the expansion of regulatory potential without disrupting the essential molecular framework required for *zf-GATA* function.

Together, these findings highlight a pattern of constrained modular diversification in *zf-GATA* evolution. Conserved GATA domains provide a stable molecular foundation, whereas additional structural components contribute to lineage-specific expansion of regulatory potential (Bonchuk and Georgiev, 2024; Emerson and Thomas, 2009). By integrating comparative genomics, motif analysis, phylogenetics, and structural modeling, this study provides insights into how TF families diversify while preserving essential molecular features.

4 – 2. Gene duplication and lineage-specific retention contributed to *zf-GATA* repertoire diversification
Gene duplication followed by lineage-specific retention or loss represents a major evolutionary mechanism driving TF family expansion and diversification while preserving essential functional frameworks (Deem and Brisson, 2024). In this study, the overall number of *zf-GATA* genes was relatively conserved across most jawed vertebrate lineages, whereas *Actinopterygii* exhibited an expanded *zf-GATA* repertoire compared with other jawed vertebrate groups. This

lineage-specific increase is consistent with the evolutionary history of ray-finned fishes, which experienced additional whole-genome duplication events followed by lineage-dependent retention and loss of duplicated genes (Glasauer and Neuhauss, 2014). The presence of teleost-specific paralogous groups, including *GATA1A/1B and GATA2A/2B*, supports the contribution of duplication and selective retention to *zf-GATA* repertoire diversification. These paralogs are consistent with retention of duplicated copies generated by the teleost-specific third round of whole-genome duplication (Gillis et al., 2009), which expanded regulatory potential while preserving ancestral structural frameworks.

The retention patterns of duplicated *zf-GATA* genes provide insight into the evolutionary constraints acting on this gene family. Although *Actinopterygii* exhibited an expanded *zf-GATA* repertoire, these duplicated genes largely retained conserved GATA domains, characteristic motifs, and lineage-specific domain architectures. This observation suggests that gene duplication primarily increased the number of regulatory components available for evolutionary diversification while maintaining the molecular features required for *zf-GATA* function. However, variation in *zf-GATA* gene abundance among vertebrate lineages indicates that duplication alone does not determine gene repertoire size. Following duplication, duplicated genes are subject to lineage-specific retention pressures influenced by functional requirements, genomic context, and evolutionary history (Assis et al., 2024; Wilson and Liberles, 2023a). Thus, the diversification of the *zf-GATA* family appears to involve both the generation of additional gene copies and subsequent evolutionary filtering that determines which duplicated genes are retained, maintaining a balance between regulatory complexity and functional constraint.

The evolutionary patterns observed in *zf-GATA* genes reflect a broader principle underlying TF diversification. Expansion of TF families is often achieved through duplication and differential retention of existing genetic components rather than through the emergence of entirely novel molecular architectures. Retained duplicates can provide evolutionary flexibility by allowing regulatory specialization, subfunctionalization, or increased regulatory capacity while maintaining essential molecular frameworks (Kuzmin et al., 2022; Wilson and Liberles, 2023b). Thus, *zf-GATA* diversification exemplifies how TF families expand functional potential through the retention and modification of conserved components.

4 – 3. Constrained evolution of the GATA DNA-binding module and domain-mediated regulatory innovation

TF evolution is shaped by the balance between preserving core DNA-binding features and diversifying regulatory components (Spitz and Furlong, 2012). In *zf-GATA* proteins, the GATA DNA-binding domain serves as the essential structural module for sequence-specific DNA recognition, and its evolutionary conservation provides a framework for understanding how molecular constraints shape TF diversification (Lowry and Atchley, 2000). In this study, comparative analyses demonstrate that the GATA domain has remained highly conserved across jawed vertebrates, whereas surrounding protein regions have undergone substantial structural diversification.

In this study, analyses of GATA domains revealed that the Type IVa motif (Cys-X_2-Cys-X_{17}-Cys-X_2-Cys) represented the predominant configuration across jawed vertebrates. This strong conservation contrasts with the greater motif diversity observed in other major eukaryotic lineages, where Type IVb and Type IVc motifs (Cys-X_2-Cys-$X_{18–20}$-Cys-X_2-Cys) are prevalent in plants (Kim, 2024; Schwechheimer et al., 2022) and both Type IVa and Type IVb motifs (Cys-X_2-Cys-$X_{17–18}$-Cys-X_2-Cys) are commonly retained in fungi (Kim, 2026), indicating lineage-specific differences in the evolutionary constraints acting on the GATA zinc-binding framework. The strong conservation of the Type IVa motif and zinc-coordinating cysteine residues across jawed vertebrates indicates that maintenance of the canonical zinc-binding configuration represents a fundamental evolutionary constraint on GATA domain evolution. Although Type IVb and other rare variants occurred, they were largely restricted to minor modifications of inter-cysteine spacing or the conserved motif framework. Therefore, sequence variation within the GATA motif appears to be tolerated only within a restricted structural range that maintains the biochemical requirements for zinc-dependent DNA recognition.

Residue-level analysis further revealed that conservation within the GATA motif extends beyond the zinc-coordinating cysteine residues. Consistent with observations in other eukaryotic lineages (Kim, 2026, 2024; Virolainen et al., 2026), several non-cysteine positions showed substantial conservation across jawed vertebrates, indicating that evolutionary constraints also act on residues contributing to the structural stability and functional properties of the GATA domain. These conserved residues represented diverse amino acid properties, including polar uncharged, special property, hydrophobic, and polar charged residues (Yampolsky and Stoltzfus, 2005), suggesting that maintenance of the GATA motif relies on coordinated preservation of multiple biochemical features beyond the zinc-binding cysteine framework. Together, these patterns indicate that the Type IVa GATA motif has been maintained under strong purifying selection (Lowry and Atchley, 2000), with limited sequence variation permitted at specific positions while preserving the structural and biochemical requirements for DNA binding.

In contrast to the highly conserved GATA DNA-binding module, additional conserved domains displayed distinct lineage-specific distributions among *zf-GATA* gene groups. Domains associated with chromatin regulation, DNA interaction, and protein–protein interactions, including ELM2, BAH, MTA R1, Myb-like DNA-binding, P66A/P66B interaction, and Atrophin-related domains, were selectively incorporated into specific *zf-GATA* architectures. The restricted occurrence of these domains suggests that regulatory diversification has been achieved through modification of protein contexts surrounding the conserved GATA module (De Mendoza and Sebé-Pedrós, 2019). Thus, differences in *zf-GATA* protein organization are largely associated with variation in regulatory components rather than extensive modification of the DNA-binding core itself.

The co-occurrence analysis between GATA motifs and conserved domains further revealed that evolution of the GATA domain is influenced by its surrounding protein environment. Most conserved domains were predominantly associated with the Type IVa motif, indicating broad compatibility between the conserved GATA configuration and diverse regulatory architectures. However, the strong association between Type IVb motifs and the Atrophin-1 family domain suggests that specific structural contexts may permit or favor alternative GATA motif configurations. This non-random relationship indicates that motif variation is not independent from protein architecture but is constrained by interactions between the DNA-binding module and adjacent regulatory regions (Spitz and Furlong, 2012).

Three-dimensional structural analyses further supported the relationship between conserved GATA domains and diversified protein architectures. Despite sequence divergence, proteins containing multiple GATA domains retained conserved spatial arrangements around DNA-binding regions, suggesting functional constraints on higher-order organization. This interpretation is supported by functional studies of metazoan GATA TFs showing that multiple zinc fingers can cooperatively regulate DNA binding and transcriptional activity. For example, interactions between the N- and C-terminal zinc fingers of *GATA1* enhance DNA binding and transactivation (Merika and Orkin, 1993), whereas *GATA3* exhibits alternative cooperative binding modes depending on DNA motif spacing (Bates et al., 2008). These findings suggest that conserved organization of multiple GATA domains in specific *zf-GATA* proteins may reflect constraints associated with coordinated DNA recognition. In contrast, *zf-GATA* proteins with single GATA domains or additional regulatory domains displayed diverse architectures, which may reflect diversification through contextual modification while preserving the GATA DNA-binding framework.

Collectively, these findings indicate that the GATA domain represents a structurally constrained yet evolutionarily flexible molecular module. Across jawed vertebrates, preservation of the canonical Type IVa motif and conserved residues within the GATA domain reflects strong evolutionary constraints associated with zinc coordination and DNA recognition. In contrast, diversification of surrounding regulatory domains and protein architectures has provided opportunities for functional innovation without disrupting the conserved DNA-binding framework. Thus, *zf-GATA* evolution appears to have proceeded through a balance between molecular constraint and contextual diversification, allowing expansion of regulatory potential while maintaining the essential structural basis required for GATA-mediated DNA recognition.

4–4. Phylogenetic relationships reveal lineage-specific diversification patterns within the *zf-GATA* family

Phylogenetic analysis of representative full-length proteins revealed that the 22 annotated *zf-GATA* gene groups are organized into several well-supported evolutionary lineages corresponding to their structural organization. The *GATA1-6* proteins formed a major lineage characterized by conserved tandem GATA domain architectures, with *GATA4-6* constituting a closely related subclade. Teleost-specific paralogs (*GATA1A/1B and GATA2A/2B*) were nested within their respective *GATA1 and GATA2* lineages rather than forming independent clades, supporting their origin through lineage-specific duplication while retaining ancestral structural characteristics (Gillis et al., 2009). Outside the *GATA1-6* lineage, *TRPS1* occupied a separate position, whereas *GATAD1/D2A/D2B/D2AB* together with *ZGLP1* formed another distinct clade. *MTA1-3 and RERE/REREA-B* represented progressively more divergent lineages, consistent with their acquisition of additional conserved domains and increasingly complex protein architectures (Kumar and Wang, 2016; Plaster et al., 2007). Overall, the topology of the full-length protein phylogeny closely reflected differences in domain composition and protein organization, indicating that diversification of *zf-GATA* proteins has proceeded in parallel with structural innovation.

Phylogenetic reconstruction based on isolated GATA motifs provided a complementary perspective by focusing specifically on the evolution of the conserved DNA-binding module. Unlike the full-length protein phylogeny, which incorporates variation arising from both GATA domains and lineage-specific accessory domains, the GATA motif phylogeny resolved relationships independent of overall protein architecture. The GATA motifs of the *MTA, RERE, and GATAD* groups each formed distinct lineages, indicating that these proteins possess characteristic evolutionary signatures within their conserved DNA-binding regions in addition to their unique motif organizations. Notably, the two GATA motifs of *GATA1-6* proteins exhibited markedly different evolutionary histories. The C-terminal GATA domains formed a well-supported lineage distinct from the N-terminal domains, whereas the N-terminal GATA motifs clustered more closely with the single GATA motif of *TRPS1 and ZGLP1*. These findings indicate that the two GATA motifs of *GATA1-6* proteins have undergone asymmetric evolutionary diversification following their origin rather than evolving as equivalent duplicated modules.

Together, the full-length protein and GATA motif phylogenies provide complementary insights into the evolution of the *zf-GATA* family. The protein-level phylogeny primarily reflects diversification of overall domain architecture and structural organization, whereas the motif-level phylogeny resolves the evolutionary history of the conserved DNA-binding module itself. Their overall concordance indicates that lineage-specific diversification of *zf-GATA* proteins has occurred at multiple hierarchical levels through coordinated changes in protein architecture and the evolutionary trajectories of individual GATA motifs. These phylogenetic relationships establish an evolutionary framework for interpreting the structural diversity of the *zf-GATA* family while highlighting the long-term conservation of the GATA DNA-binding module across jawed vertebrates.

4–5. An integrated evolutionary framework for *zf-GATA* diversification in jawed vertebrates

The evolutionary patterns identified in this study indicate that *zf-GATA* diversification has emerged through coordinated changes across multiple biological levels rather than through independent modifications of individual components. Protein architecture, gene repertoire, and evolutionary relationships represent interconnected aspects of *zf-GATA* evolution, in which changes at one level were constrained and shaped by the evolutionary history of other levels. This integrated perspective explains how *zf-GATA* proteins have acquired structural and regulatory diversity while maintaining the conserved molecular framework required for DNA recognition.

At the molecular level, diversification of *zf-GATA* proteins has occurred through modification of a conserved structural framework (Lowry and Atchley, 2000). The persistence of the GATA DNA-binding module across jawed vertebrates provided a stable foundation upon which additional regulatory features could evolve. Changes in surrounding protein regions, including variation in domain composition and organization, generated diverse *zf-GATA* architectures while preserving the essential structural requirements for DNA binding. Thus, protein-level diversification appears to have proceeded through remodeling of regulatory contexts rather than extensive alteration of the conserved DNA-binding core.

These structural changes were further influenced by genome-level evolutionary processes that determined the distribution and abundance of *zf-GATA* genes across vertebrate lineages. Gene duplication provided additional copies that could contribute to regulatory diversification, whereas lineage-specific retention shaped which duplicated genes persisted during evolution (Glasauer and Neuhauss, 2014). Importantly, the retention of conserved structural characteristics among duplicated genes indicates that expansion of gene repertoires occurred within the constraints imposed by ancestral protein organization. Therefore, gene duplication and structural conservation operated together to increase evolutionary flexibility while maintaining functional continuity.

At the evolutionary relationship level, phylogenetic patterns demonstrate how these molecular and genomic processes have accumulated over evolutionary time. The distinct evolutionary trajectories of *zf-GATA* groups reflect the combined effects of protein architecture remodeling and gene repertoire changes, while the conserved relationships among GATA domains indicate persistent constraints on the DNA-binding module itself. These patterns suggest that *zf-GATA* diversification represents a coordinated evolutionary process in which genomic changes, protein innovation, and molecular conservation are interconnected rather than independent (Voordeckers et al., 2015).

Collectively, this integrated framework suggests that *zf-GATA* diversity in jawed vertebrates has been generated through progressive modification of conserved molecular systems. The preservation of the GATA DNA-binding framework enabled diversification of regulatory architectures and expansion of gene repertoires, while evolutionary constraints filtered structural changes compatible with *zf-GATA* function. Therefore, the *zf-GATA* family provides an example of how TF diversity can arise through the coordinated evolution of conserved molecular modules, protein architectures, and genomic organization while maintaining essential biological functions (De Mendoza and Sebé-Pedrós, 2019).

4–6. Annotation variability in jawed vertebrate genomes and its implications

The present study investigates the structural and evolutionary diversification of *zf-GATA* genes across 181 jawed vertebrate species. However, the potential impact of annotation variability among species remains an important consideration when interpreting large-scale comparative analyses. The analyzed gene datasets were obtained from Ensembl (Yates et al., 2026), which provides gene annotations based on publicly available genome assemblies deposited in the INSDC database (GenBank, ENA, and DDBJ) (Karsch-Mizrachi et al., 2024) rather than generating genome assemblies itself. Because these assemblies originate from diverse sources, they may differ in genome assembly quality and annotation completeness among species. Another important consideration is the approach used to define the *zf-GATA* gene set. AnimalTFDB classifies TFs primarily through Pfam profile HMM-based domain detection, providing a standardized framework for large-scale TF annotation (Shen et al., 2023). Nevertheless, divergent GATA domains, incomplete gene models, or fragmented assemblies may reduce domain detection sensitivity and cause misannotation or underrepresentation of some *zf-GATA* genes. Such technical variability may particularly affect the identification and structural characterization of *zf-GATA* genes, especially those with rare domain architectures or atypical GATA motif configurations.

Despite these potential limitations, the overall patterns identified in this study were largely consistent across jawed vertebrate lineages. To reduce potential annotation biases, independent domain-level re-annotation using InterProScan (Jones et al., 2014) and Pfam (Finn et al., 2014) was performed to validate the domain architectures of *zf-GATA* genes. The predominant structural features, including conserved major domain architectures, GATA motif configurations, and lineage-wide distribution patterns, were consistently observed across diverse taxonomic groups. These results suggest that the major evolutionary trends and structural characteristics identified in this study are unlikely to be solely attributable to annotation artifacts. Therefore, while species-specific variations require caution, the identified large-scale patterns provide a robust framework for understanding the structural and evolutionary diversification of *zf-GATA* genes in jawed vertebrates.

4–7. Limitations and future perspectives

Despite providing a comprehensive structural and evolutionary analysis of jawed vertebrate *zf-GATA* genes, this study has several limitations that should be considered when interpreting the observed evolutionary patterns. First, the taxonomic coverage of the present dataset remains constrained by the availability of genomic resources (Shen et al., 2023). Although 181 jawed vertebrate species representing major evolutionary lineages were analyzed, several deeply divergent or currently underrepresented lineages were represented by limited numbers of species. Expanding taxonomic sampling, particularly among early-diverging vertebrates and non-model lineages, will provide improved resolution of ancestral *zf-GATA* architectures, lineage-specific innovations, and diversification patterns (O'Leary et al., 2016; Rhie et al., 2021; Yuan et al., 2026).

Second, although independent domain-level re-annotation supported the reliability of the identified *zf-GATA* genes, variation in genome assembly quality and gene annotation completeness among species may influence the detection and classification of highly divergent or structurally unusual *zf-GATA* genes (Shen et al., 2023). In particular, the 257 unclassified *zf-GATA* genes identified in this study may represent highly divergent members or lineage-specific variants that could not be assigned to established gene groups using current annotation frameworks. Future studies integrating improved genome assemblies, refined profile models, and phylogenomic approaches will facilitate more comprehensive identification and classification of divergent *zf-GATA* genes (Ji et al., 2026).

Third, although gene repertoire patterns and the presence of teleost-specific paralogous groups support the contribution of duplication and lineage-specific retention to *zf-GATA* diversification, the precise evolutionary history of individual *zf-GATA* paralogs was not directly reconstructed in this study. Future analyses incorporating synteny conservation, gene tree–species tree reconciliation, and ancestral gene reconstruction will provide deeper insights into the timing, origin, and retention mechanisms of *zf-GATA* gene duplication events (Szöllősi et al., 2015).

Fourth, this study primarily focused on protein-level diversification, including domain organization, motif conservation, and predicted structural features. Whether lineage-specific structural diversification directly corresponds to functional diversification remains to be experimentally determined. Integrative analyses combining transcriptomic, epigenomic, and genome-wide DNA-binding datasets, together with experimental characterization of representative *zf-GATA* proteins, will be required to clarify how changes in protein architecture influence DNA-binding properties, regulatory networks, and biological functions (Romano and Miccio, 2020).

Finally, the three-dimensional structural analyses presented here were based on computational predictions, which provide valuable comparative insights but do not replace experimentally determined structures. Future structural and biochemical studies will be necessary to validate predicted domain arrangements

and determine how lineage-specific architectural differences influence molecular interactions and regulatory mechanisms (Kuhlman and Bradley, 2019; Varadi et al., 2022).

Together, these future directions will extend the evolutionary framework established in this study by integrating broader phylogenomic sampling, improved genome resources, functional genomics, and experimental structural analyses. Such multidisciplinary approaches will provide a more complete understanding of how conserved molecular modules and lineage-specific innovations have collectively shaped the diversification of the *zf-GATA* family across vertebrates.

5. Conclusion

This study establishes an evolutionary framework for understanding the diversification of *zf-GATA* genes across jawed vertebrates through large-scale comparative genomics, domain architecture analysis, GATA motif characterization, phylogenetic reconstruction, functional annotation, and structural modeling. The results demonstrate that *zf-GATA* genes have evolved through a balance between strong conservation of the GATA DNA-binding module and extensive diversification of surrounding regulatory architectures. Across jawed vertebrate lineages, conserved zinc-binding motifs and key structural features of the GATA domain have been maintained, whereas lineage-specific changes in domain organization, GATA domain multiplicity, and gene repertoire have contributed to the expansion of regulatory potential. Phylogenetic analyses further reveal that *zf-GATA* diversification has been shaped by coordinated evolutionary processes involving structural innovation, gene duplication, and lineage-specific retention. These findings provide insights into how conserved molecular modules are maintained while allowing functional diversification of TF families during vertebrate evolution. The evolutionary framework established here provides a foundation for future investigations into the molecular mechanisms underlying GATA-mediated gene regulation and the broader principles governing TF diversification.

Declarations

Author contributions

M.K. conducted all the work until the manuscript was finalized and published.

Data availability

All jawed vertebrate *zf-GATA* genes analyzed in this study are publicly available from AnimalTFDB4 (https://guolab.wchscu.cn/AnimalTFDB4) (Shen et al., 2023).

Acknowledgements

I sincerely thank the editor and the reviewers for their thorough review and valuable comments.

Competing interests

The authors declare no conflict of interest, financial or otherwise.

Ethical approval

Not applicable.

Informed consent

Not applicable.

Clinical trial number

Not applicable.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Amoutzias G, Veron A, Weiner J, Robinson-Rechavi M, Bornberg-Bauer E, Oliver S, Robertson D (2006) One Billion Years of bZIP Transcription Factor Evolution: Conservation and Change in Dimerization and DNA-Binding Site Specificity. Mol Biol Evol 24:827–835. https://doi.org/10.1093/molbev/msl211
2. Assis R, Conant G, Holland B, Liberles DA, O'Reilly MM, Wilson AE (2024) Models for the retention of duplicate genes and their biological underpinnings. F1000Research 12. 1400. https://doi.org/10.12688/f1000research.141786.2
3. Bates DL, Chen Y, Kim G, Guo L, Chen L (2008) Crystal Structures of Multiple GATA Zinc Fingers Bound to DNA Reveal New Insights into DNA Recognition and Self-Association by GATA. J Mol Biol 381:1292–1306. https://doi.org/10.1016/j.jmb.2008.06.072

4. Bertoline LMF, Lima AN, Krieger JE, Teixeira SK (2023) Before and after AlphaFold2: An overview of protein structure prediction. Front Bioinforma 3:1120370. https://doi.org/10.3389/fbinf.2023.1120370
5. Bonchuk AN, Georgiev PG (2024) C2H2 proteins: Evolutionary aspects of domain architecture and diversification. BioEssays 46, 2400052. https://doi.org/10.1002/bies.202400052
6. Brazeau MD, Friedman M (2015) The origin and early phylogenetic history of jawed vertebrates. Nature 520:490–497. https://doi.org/10.1038/nature14438
7. Westerheide D, Raynes S, Powell R, Xue C, Uversky BN, V (2012) HSF Transcription Factor Family, Heat Shock Response, and Protein Intrinsic Disorder. Curr Protein Pept Sci 13:86–103. https://doi.org/10.2174/138920312799277956
8. De Mendoza A, Sebé-Pedrós A (2019) Origin and evolution of eukaryotic transcription factors. Curr Opin Genet Dev 58–59:25–32. https://doi.org/10.1016/j.gde.2019.07.010
9. Deem KD, Brisson JA (2024) Problems with Paralogs: The Promise and Challenges of Gene Duplicates in Evo-Devo Research. Integr Comp Biol 64:556–564. https://doi.org/10.1093/icb/icae009
10. Dong Z, Zhang N, Liu Y, Xu W, Cui Z, Shao C, Chen S (2019) Expression analysis and characterization of zglp1 in the Chinese tongue sole (Cynoglossus semilaevis). Gene 683:72–79. https://doi.org/10.1016/j.gene.2018.10.003
11. Eckert D, Buhl S, Weber S, Jäger R, Schorle H (2005) The AP-2 family of transcription factors. Genome Biol 6:246. https://doi.org/10.1186/gb-2005-6-13-246
12. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763. https://doi.org/10.1093/bioinformatics/14.9.755
13. Emerson RO, Thomas JH (2009) Adaptive Evolution in Zinc Finger Transcription Factors. PLoS Genet 5:e1000325. https://doi.org/10.1371/journal.pgen.1000325
14. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. https://doi.org/10.1093/nar/gkt1223
15. Friedman MJ, Wagner T, Lee H, Rosenfeld MG, Oh S (2024) Enhancer–promoter specificity in gene transcription: molecular mechanisms and disease associations. Exp Mol Med 56:772–787. https://doi.org/10.1038/s12276-024-01233-y
16. Gillis WQ, St John J, Bowerman B, Schneider SQ (2009) Whole genome duplications and expansion of the vertebrate GATA transcription factor gene family. BMC Evol Biol 9:207. https://doi.org/10.1186/1471-2148-9-207
17. Glasauer SMK, Neuhauss SCF (2014) Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol Genet Genomics 289:1045–1060. https://doi.org/10.1007/s00438-014-0889-2
18. Hajheidari M, Huang SC (2022) Elucidating the biology of transcription factor–DNA interaction for accurate identification of cis-regulatory elements. Curr Opin Plant Biol 68:102232. https://doi.org/10.1016/j.pbi.2022.102232
19. He C, Cheng H, Zhou R (2007) GATA family of transcription factors of vertebrates: phylogenetics and chromosomal synteny. J Biosci 32:1273–1280. https://doi.org/10.1007/s12038-007-0136-7
20. Hellmuth M, Stadler PF (2024) The Theory of Gene Family Histories, in: Setubal, J.C., Stadler, P.F., Stoye, J. (Eds.), Comparative Genomics, Methods in Molecular Biology. Springer US, New York, NY, pp. 1–32. https://doi.org/10.1007/978-1-0716-3838-5_1
21. Hu D, Zhao R, Lin Y, Jiang C (2025) Evolution and Functional Diversity of GATA Transcription Factors in Filamentous Fungi: Structural Characteristics, Metabolic Regulation and Environmental Response. Microbiol Res 16:120. https://doi.org/10.3390/microbiolres16060120
22. Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire J-Y, Kupfer A, Petersen J, Jarek M, Meyer A, Vences M, Philippe H (2017) Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol 1:1370–1378. https://doi.org/10.1038/s41559-017-0240-5
23. Jetz W, Fine PVA (2012) Global Gradients in Vertebrate Diversity Predicted by Historical Area-Productivity Dynamics and Contemporary Environment. PLoS Biol 10:e1001292. https://doi.org/10.1371/journal.pbio.1001292
24. Ji HJ, Pertea M, Salzberg SL (2026) Annotating genomes at increased scale and resolution. Nat Rev Genet 27:429–441. https://doi.org/10.1038/s41576-026-00937-3
25. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031
26. Karsch-Mizrachi I, Arita M, Burdett T, Cochrane G, Nakamura Y, Pruitt KD, Schneider VA (2024) The international nucleotide sequence database collaboration (INSDC): enhancing global participation. Nucleic Acids Res. https://doi.org/10.1093/nar/gkae1058
27. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
28. Kim BJ, Zaveri HP, Jordan VK, Hernandez-Garcia A, Jacob DJ, Zamora DL, Yu W, Schwartz RJ, Scott DA (2018) RERE deficiency leads to decreased expression of GATA4 and the development of ventricular septal defects. Dis Model Mech 11:dmm031534. https://doi.org/10.1242/dmm.031534
29. Kim M (2026) Genome-wide comparative analysis of structural features in fungal GATA transcription factors: insights from 796 fungal species. Mycol Prog 25:35. https://doi.org/10.1007/s11557-026-02151-1
30. Kim M (2024) Comparative analysis of amino acid sequence level in plant GATA transcription factors. Sci Rep 14:29786. https://doi.org/10.1038/s41598-024-81159-7
31. Kueckelhaus J, Frerich S, Kada-Benotmane J, Koupourtidou C, Ninkovic J, Dichgans M, Beck J, Schnell O, Heiland DH (2024) Inferring histology-associated gene expression gradients in spatial transcriptomic studies. Nat Commun 15:7280. https://doi.org/10.1038/s41467-024-50904-x

32. Kuhlman B, Bradley P (2019) Advances in protein structure prediction and design. Nat Rev Mol Cell Biol 20:681–697. https://doi.org/10.1038/s41580-019-0163-x
33. Kumar R, Wang R-A (2016) Structure, expression and functions of MTA genes. Gene 582:112–121. https://doi.org/10.1016/j.gene.2016.02.012
34. Kuzmin E, Taylor JS, Boone C (2022) Retention of duplicated genes in evolution. Trends Genet 38:59–72. https://doi.org/10.1016/j.tig.2021.06.016
35. Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S, Gascuel O (2019) NGPhylogeny.fr: new generation phylogenetic services for non-specialists. Nucleic Acids Res 47:W260–W265. https://doi.org/10.1093/nar/gkz303
36. Letunic I, Bork P (2024) Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res 52:W78–W82. https://doi.org/10.1093/nar/gkae268
37. Liang Y, Zhang X, Liu Y, Wang L, Ye Y, Tan X, Pu J, Zhang Q, Bao X, Wei X, Li, Dongzhi, Kurita R, Nakamura Y, Li, Dali, Xu X (2021) GATA zinc finger domain-containing protein 2A (*GATAD2A*) deficiency reactivates fetal haemoglobin in patients with β-thalassaemia through impaired formation of methyl-binding domain protein 2 (MBD2)-containing nucleosome remodelling and deacetylation (NuRD) complex. Br J Haematol 193:1220–1227. https://doi.org/10.1111/bjh.17511
38. Lowry JA, Atchley WR (2000) Molecular Evolution of the GATA Family of Transcription Factors: Conservation Within the DNA-Binding Domain. J Mol Evol 50:103–115. https://doi.org/10.1007/s002399910012
39. Meadows JRS, Lindblad-Toh K (2017) Dissecting evolution and disease using comparative vertebrate genomics. Nat Rev Genet 18:624–636. https://doi.org/10.1038/nrg.2017.51
40. Merika M, Orkin SH (1993) Mol Cell Biol 13:3999–4010. https://doi.org/10.1128/mcb.13.7.3999-4010.1993. DNA-Binding Specificity of GATA Family Transcription Factors
41. Mérot C, Oomen RA, Tigano A, Wellenreuther M (2020) A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. Trends Ecol Evol 35:561–572. https://doi.org/10.1016/j.tree.2020.03.002
42. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015
43. Odamaki T, Kato K, Sugahara H, Hashikura N, Takahashi S, Xiao J, Abe F, Osawa R (2016) Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. BMC Microbiol 16:90. https://doi.org/10.1186/s12866-016-0708-5
44. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189
45. Plaster N, Sonntag C, Schilling TF, Hammerschmidt M (2007) REREa/Atrophin-2 interacts with histone deacetylase and Fgf8 signaling to regulate multiple processes of zebrafish development. Dev Dyn 236:1891–1904. https://doi.org/10.1002/dvdy.21196
46. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, Lee C, Ko BJ, Chaisson M, Gedman GL, Cantin LJ, Thibaud-Nissen F, Haggerty L, Bista I, Smith M, Haase B, Mountcastle J, Winkler S, Paez S, Howard J, Vernes SC, Lama TM, Grutzner F, Warren WC, Balakrishnan CN, Burt D, George JM, Biegler MT, Iorns D, Digby A, Eason D, Robertson B, Edwards T, Wilkinson M, Turner G, Meyer A, Kautt AF, Franchini P, Detrich HW, Svardal H, Wagner M, Naylor GJP, Pippel M, Malinsky M, Mooney M, Simbirsky M, Hannigan BT, Pesout T, Houck M, Misuraca A, Kingan SB, Hall R, Kronenberg Z, Sović I, Dunn C, Ning Z, Hastie A, Lee J, Selvaraj S, Green RE, Putnam NH, Gut I, Ghurye J, Garrison E, Sims Y, Collins J, Pelan S, Torrance J, Tracey A, Wood J, Dagnew RE, Guan D, London SE, Clayton DF, Mello CV, Friedrich SR, Lovell PV, Osipova E, Al-Ajli FO, Secomandi S, Kim H, Theofanopoulou C, Hiller M, Zhou Y, Harris RS, Makova KD, Medvedev P, Hoffman J, Masterson P, Clark K, Martin F, Howe, Kevin, Flicek P, Walenz BP, Kwak W, Clawson H, Diekhans M, Nassar L, Paten B, Kraus RHS, Crawford AJ, Gilbert MTP, Zhang G, Venkatesh B, Murphy RW, Koepfli K-P, Shapiro B, Johnson WE, Di Palma F, Marques-Bonet T, Teeling EC, Warnow T, Graves JM, Ryder OA, Haussler D, O'Brien SJ, Korlach J, Lewin HA (2021) Howe, Kerstin, Myers, E.W., Durbin, R., Phillippy, A.M., Jarvis, E.D., Towards complete and error-free genome assemblies of all vertebrate species. Nature 592, 737–746. https://doi.org/10.1038/s41586-021-03451-0
47. Romano O, Miccio A (2020) GATA factor transcriptional activity: Insights from genome-wide binding profiles. IUBMB Life 72:10–26. https://doi.org/10.1002/iub.2169
48. Saeki N, Kanai R, Tatsuta S, Kawaguchi S, Itoh M, Ohba S, Abe M (2026) Pathogenesis and potential therapeutic targets of trichorhinophalangeal syndrome; lessons obtained from animal studies. Dev Dyn 255:228–245. https://doi.org/10.1002/dvdy.70082
49. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020, baaa062. https://doi.org/10.1093/database/baaa062
50. Schwechheimer C, Schröder PM, Blaby-Haas CE (2022) Plant GATA Factors: Their Biology, Phylogeny, and Phylogenomics. Annu Rev Plant Biol 73:123–148. https://doi.org/10.1146/annurev-arplant-072221-092913
51. Shen W-K, Chen S-Y, Gan Z-Q, Zhang Y-Z, Yue T, Chen M-M, Xue Y, Hu H, Guo A-Y (2023) AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. Nucleic Acids Res 51:D39–D45. https://doi.org/10.1093/nar/gkac907
52. Spitz F, Furlong EEM (2012) Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13:613–626. https://doi.org/10.1038/nrg3207

53. Szöllősi GJ, Tannier E, Daubin V, Boussau B (2015) The Inference of Gene Trees with Species Trees. Syst Biol 64:e42–e62. https://doi.org/10.1093/sysbio/syu048
54. Tang Y, Wei Y, He W, Wang Y, Zhong J, Qin C (2014) GATA transcription factors in vertebrates: evolutionary, structural and functional interplay. Mol Genet Genomics 289:203–214. https://doi.org/10.1007/s00438-013-0802-4
55. Tremblay M, Sanchez-Ferras O, Bouchard M (2018) GATA transcription factors in development and disease. Development 145:dev164384. https://doi.org/10.1242/dev.164384
56. Tsarovina K, Pattyn A, Stubbusch J, Müller F, Van Der Wees J, Schneider C, Brunet J-F, Rohrer H (2004) Essential role of Gata transcription factors in sympathetic neuron development. Development 131:4775–4786. https://doi.org/10.1242/dev.01370
57. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10:252–263. https://doi.org/10.1038/nrg2538
58. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50, D439–D444. https://doi.org/10.1093/nar/gkab1061
59. Virolainen P, Pankova V, Nerezenko A, Chekunova E (2026) Structural Features of Algal and Fungal GATA Transcription Factors may Play a Role in Symbiosis. J Mol Evol 94:383–404. https://doi.org/10.1007/s00239-026-10310-x
60. Voordeckers K, Pougach K, Verstrepen KJ (2015) How do regulatory networks evolve and expand throughout evolution? Curr. Opin Biotechnol 34:180–188. https://doi.org/10.1016/j.copbio.2015.02.001
61. Wilkinson AC, Nakauchi H, Göttgens B (2017) Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity. Cell Syst 5:319–331. https://doi.org/10.1016/j.cels.2017.07.004
62. Wilson AE, Liberles DA (2023a) Expectations of duplicate gene retention under the gene duplicability hypothesis. BMC Ecol Evol 23:76. https://doi.org/10.1186/s12862-023-02174-2
63. Wilson AE, Liberles DA (2023b) Dosage balance acts as a time-dependent selective barrier to subfunctionalization. BMC Ecol Evol 23:14. https://doi.org/10.1186/s12862-023-02116-y
64. Xiao Y, Stolzer M, Wasserman L, Durand D (2025) Evolution of the Metazoan Protein Domain Repertoire Revealed by a Birth-Death-Gain Model. J Mol Evol 93:777–799. https://doi.org/10.1007/s00239-025-10286-0
65. Xue J-Y, Yang C-A, Hu S, Fan H-Y, Zhang Y-M, Shao Z-Q, Li L, Wang S, Wei T, Dong S, Liu Y, Li Z, Van De Peer Y (2026) Evolution of protein domains and protein domain combinations provides insights into the origin and diversification of land plants. Cell Genomics 6:101107. https://doi.org/10.1016/j.xgen.2025.101107
66. Yampolsky LY, Stoltzfus A (2005) The Exchangeability of Amino Acids in Proteins. Genetics 170:1459–1472. https://doi.org/10.1534/genetics.104.039107
67. Yates AD, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, Becker A, Bennett R, Berry A, Bhai J, Bhurji SK, Branco Lins PR, Brooks L, Budhanuru Ramaraju S, Campbell LI, Carbajo Martinez M, Carpenter J, Charkhchi M, Cortes LA, Davidson C, Dickson S, Dodiya K, Donaldson S, Houdaigui E, Naboulsi BE, Fairbrother-Browne T, Falola A, Fatima O, Martinez RG, Gurbich J, Hall T, Hardy H, Hollis M, Hunt Z, Kay T, Kaikala M, Lazar V, Lemos A, Lodha D, Mathlouthi D, Merino N, Merritt GA, Mirabueno R, Mushtaq LP, Nakib Hossain A, Pérez-Silva S, Piližota JG, Poppleton I, Prosovetskaia D, Raj I, Salam S, Saraf AIA, Sinha S, Sipos S, Sitnik B, Suner V, Surapaneni M-M, Tierney L, Urbina-Gómez JAS, Veidenberg D, Walsh A, Allen TA, Alvarez-Jarreta JM, Cheema J, Batista J, da Rocha J, De Silva NH, Tricomi FF, Giorgetti S, Ilsley GR, Keatley J, Loveland JE, Mudge JM, Naamati G, Tate J, Willhoft NL, Winterbottom A, Flint BR, Frankish A, Haggerty L, Hunt SE, Clark EL, Dyer SC, Freeberg MA, Martin RD (2026) Ensembl 2026. Nucleic Acids Res. 54, D1053–D1060. https://doi.org/10.1093/nar/gkaf1239
68. Yuan D, Ahamed A, Athar A, Devaraj R, Gupta D, Haseeb M, Ihsan M, Ivanov E, Kadhirvelu V, Khen A, Kumar M, Lathi A, Liyanage I, Meszaros L, O'Cathail C, Pauperio J, Paz R, Pesant S, Rahman N, Rajan J, Tutis I, Ventouratou M, Vijayaraja S, Waheed Z, Woollard P, Burdett T, Cochrane G, Sarkans U (2026) The European Nucleotide Archive in 2025. Nucleic Acids Res 54:D120–D127. https://doi.org/10.1093/nar/gkaf1295

Figures

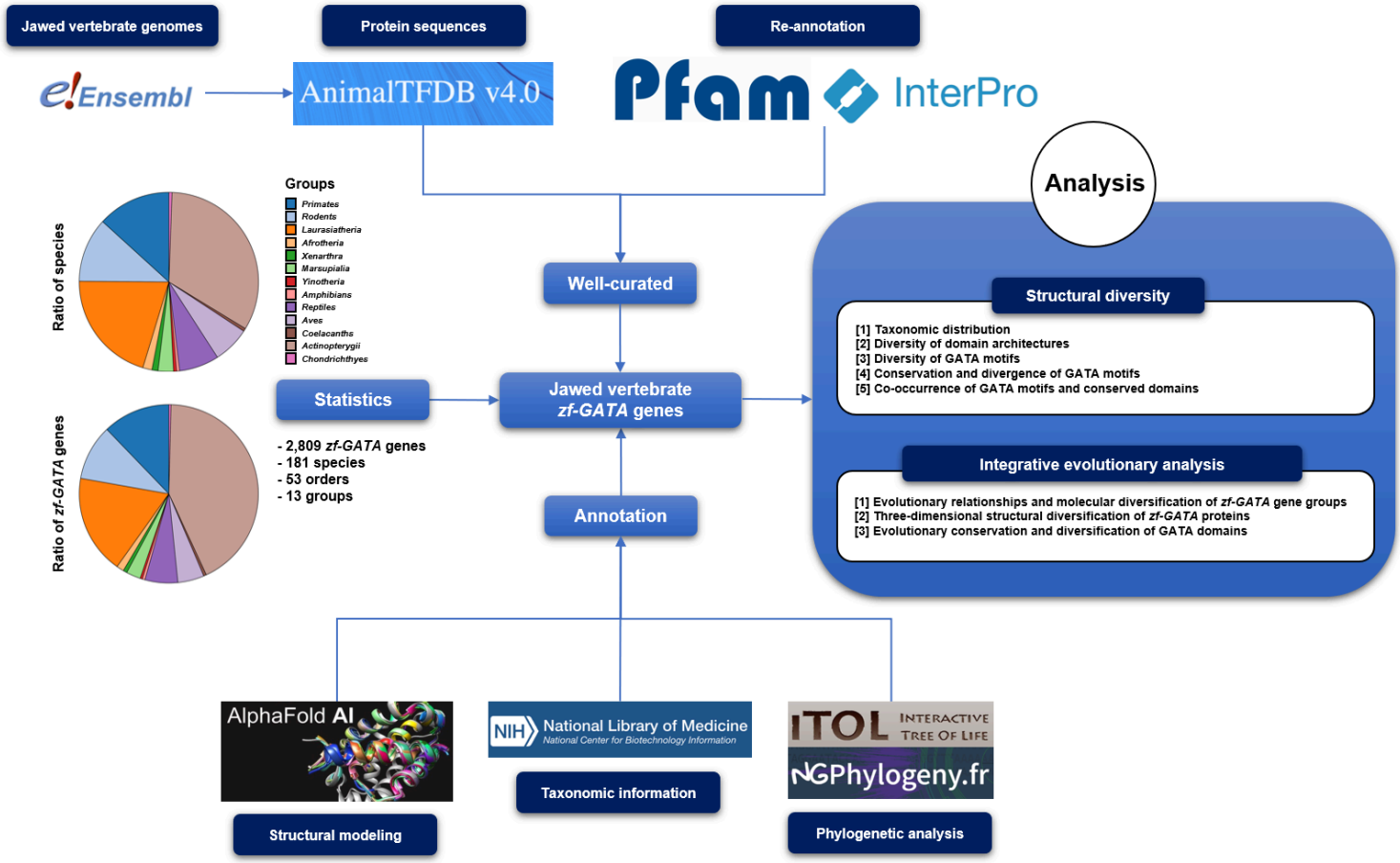


Figure 1

Fig 1. Schematic overview of the study workflow

The workflow summarizes the collection, refinement, and domain re-annotation of jawed vertebrate *zf-GATA* genes, together with dataset statistics. Subsequent analyses investigated the structural and evolutionary diversification of *zf-GATA* genes through comparative, phylogenetic, and structural approaches.

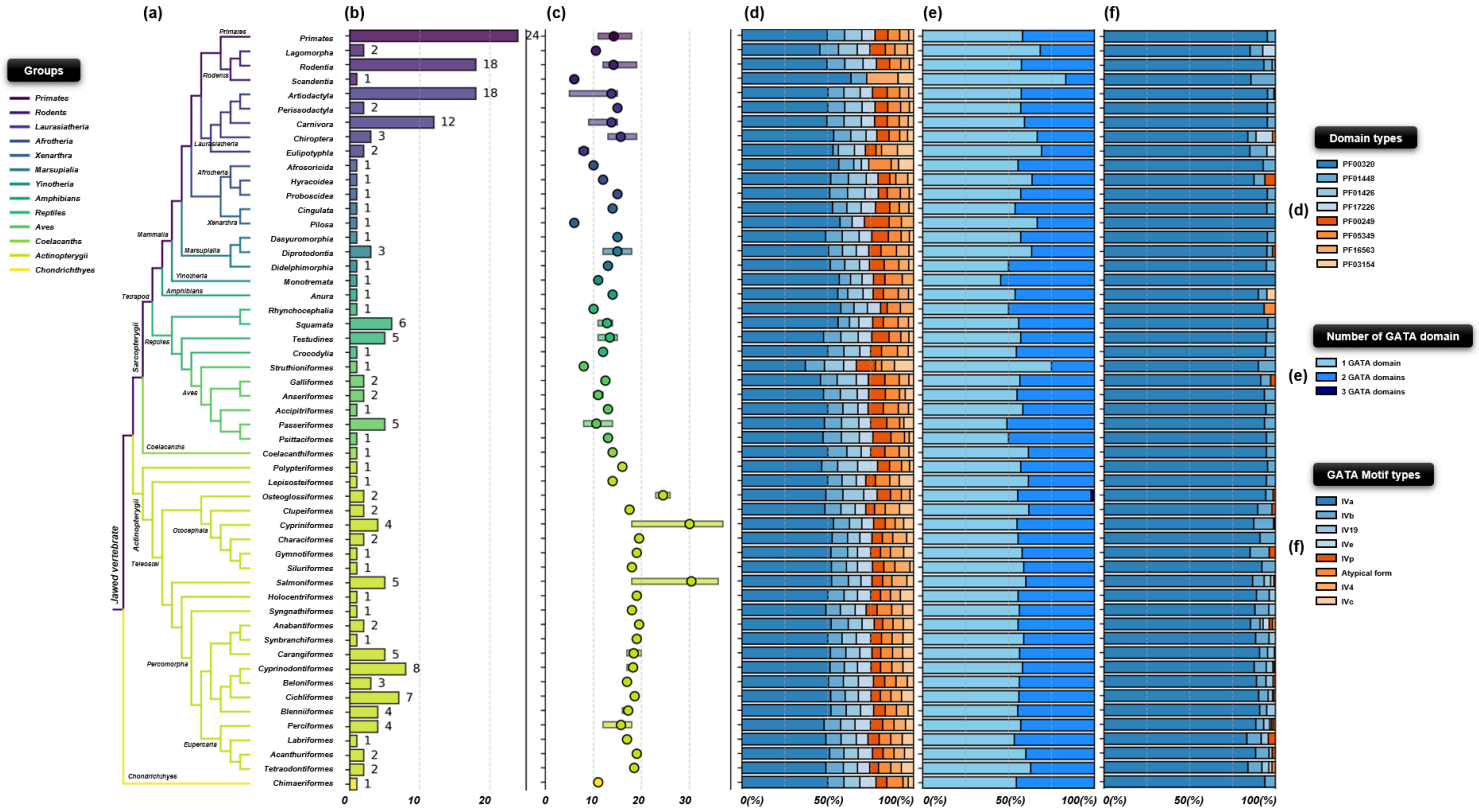


Figure 2

Fig 2. Taxonomic distribution and architectural features of jawed vertebrate *zf-GATA* genes

(a) Phylogenetic overview of the 13 jawed vertebrate groups and 53 taxonomic orders included in this study. **(c)** Variation in the number of *zf-GATA*genes per species across orders, summarized by minimum, mean, and maximum values. **(d)**Relative abundance of protein domains across orders. **(e)** Proportional distribution of *zf-GATA*genes with different GATA domain counts across orders. **(f)** Distribution of GATA motif types across orders.

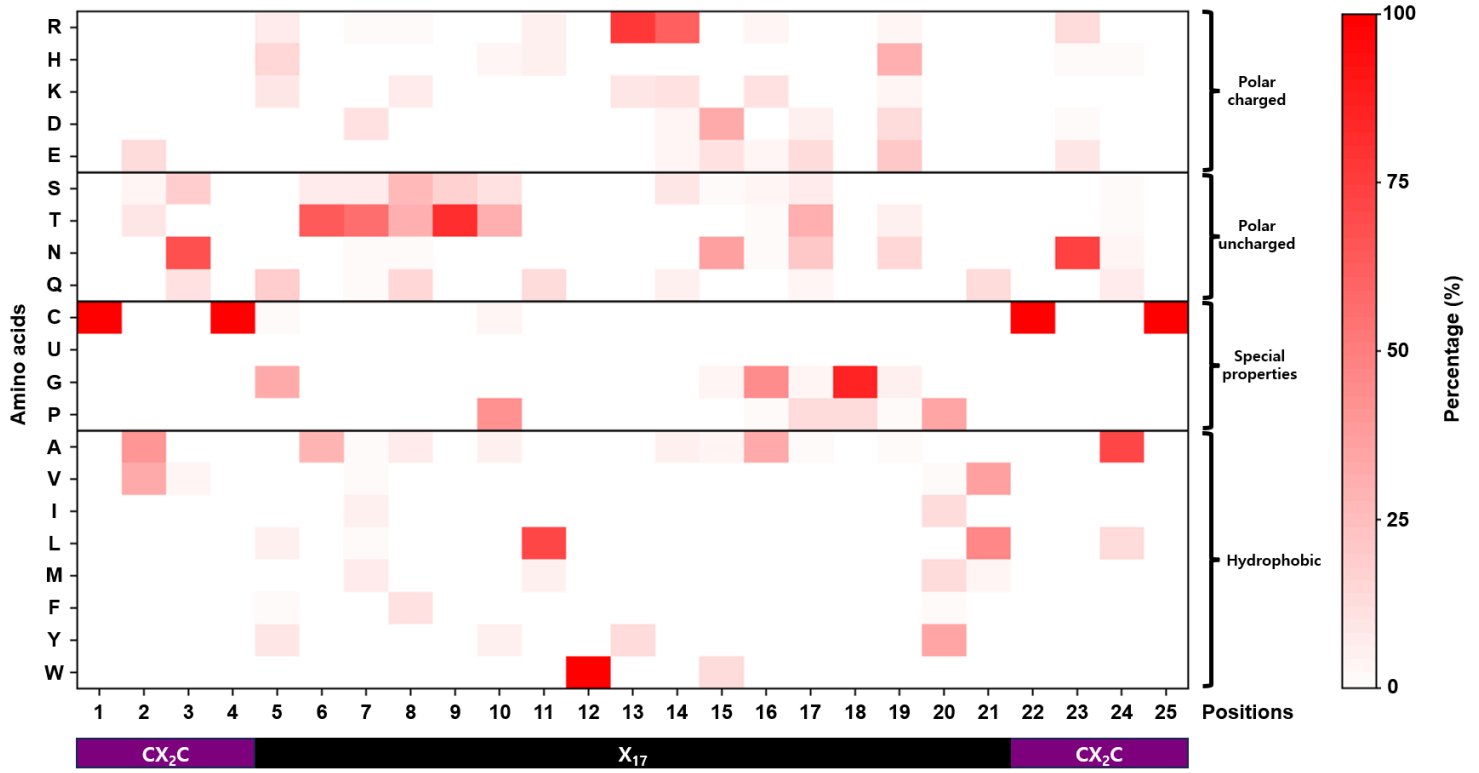


Figure 3

Fig 3. Amino acid sequence patterns of the GATA motif in jawed vertebrate *zf-GATA* genes

The x-axis denotes sequential positions within the Type IVa ($CX_2CX_{17}CX_2C$) motif. The y-axis categorizes amino acids into four physicochemical groups: charged polar, uncharged polar, residues with unique structural properties, and hydrophobic residues. Color intensity reflects the relative frequency of each amino acid at a given position, ranging from 0% (white) to 100% (red).

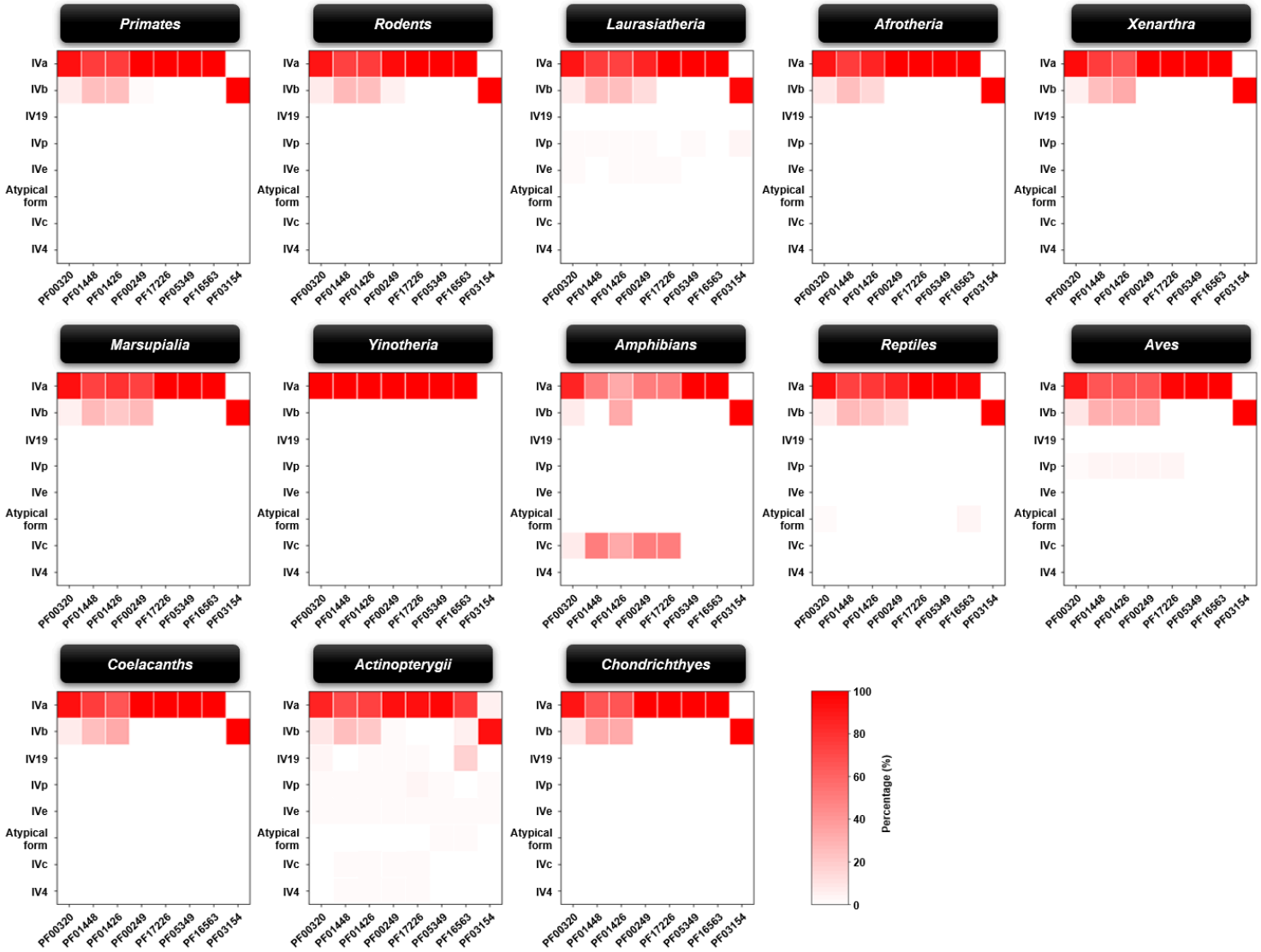


Figure 4

Fig 4. Co-occurrence patterns between jawed vertebrate GATA motif types and conserved domains

Heatmaps depicting the co-occurrence patterns between GATA motif types (y-axis) and both major and minor domains (x-axis) across 13 jawed vertebrate groups. Color intensity reflects the proportion of co-occurrence, ranging from 0% (white) to 100% (red).

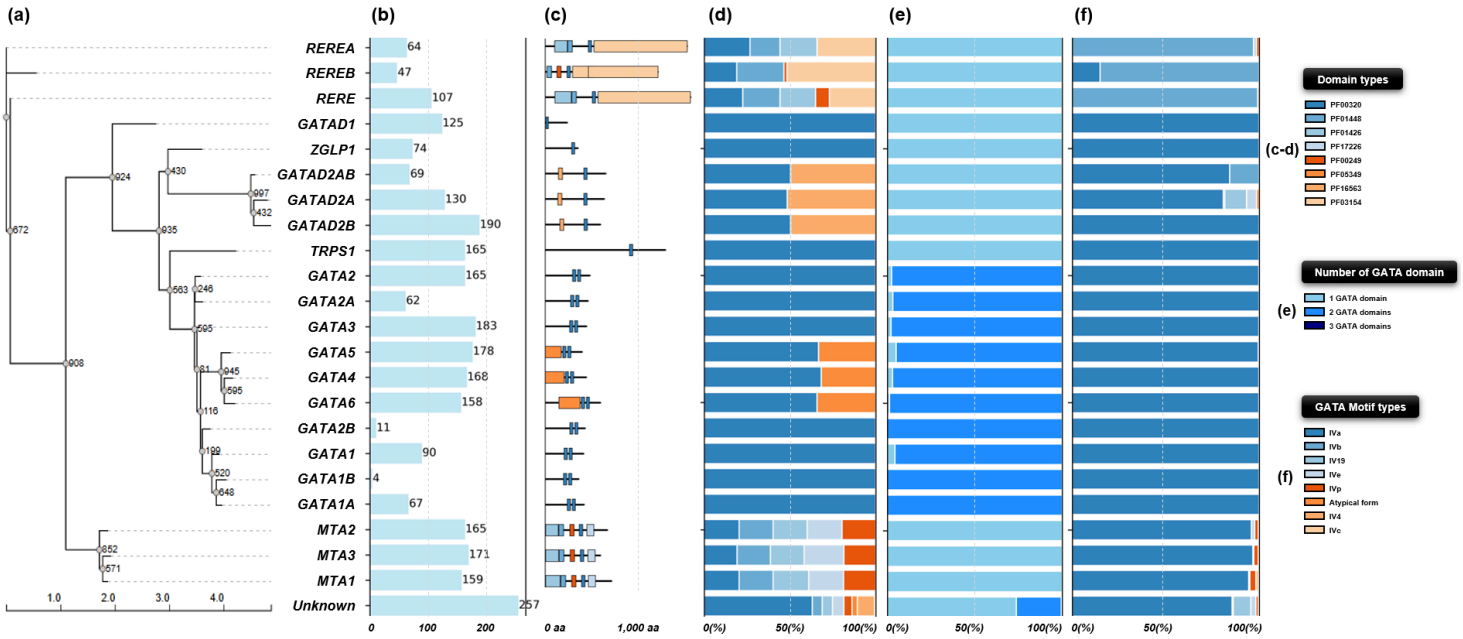


Figure 5

Fig 5. Phylogenetic relationships and structural characteristics of jawed vertebrate *zf-GATA*gene groups

(a) Phylogenetic relationships among representative genes from the 22 jawed vertebrate *zf-GATA* gene groups. **(b)** Number of *zf-GATA* genes in each gene group. **(c)**Domain architectures of representative genes from each *zf-GATA* gene group. **(d)**Relative abundance of protein domains in each *zf-GATA* gene group. **(e)** Proportional distribution of *zf-GATA*genes with different GATA domain counts in each gene group. **(f)** Distribution of GATA motif types in each *zf-GATA* gene group.

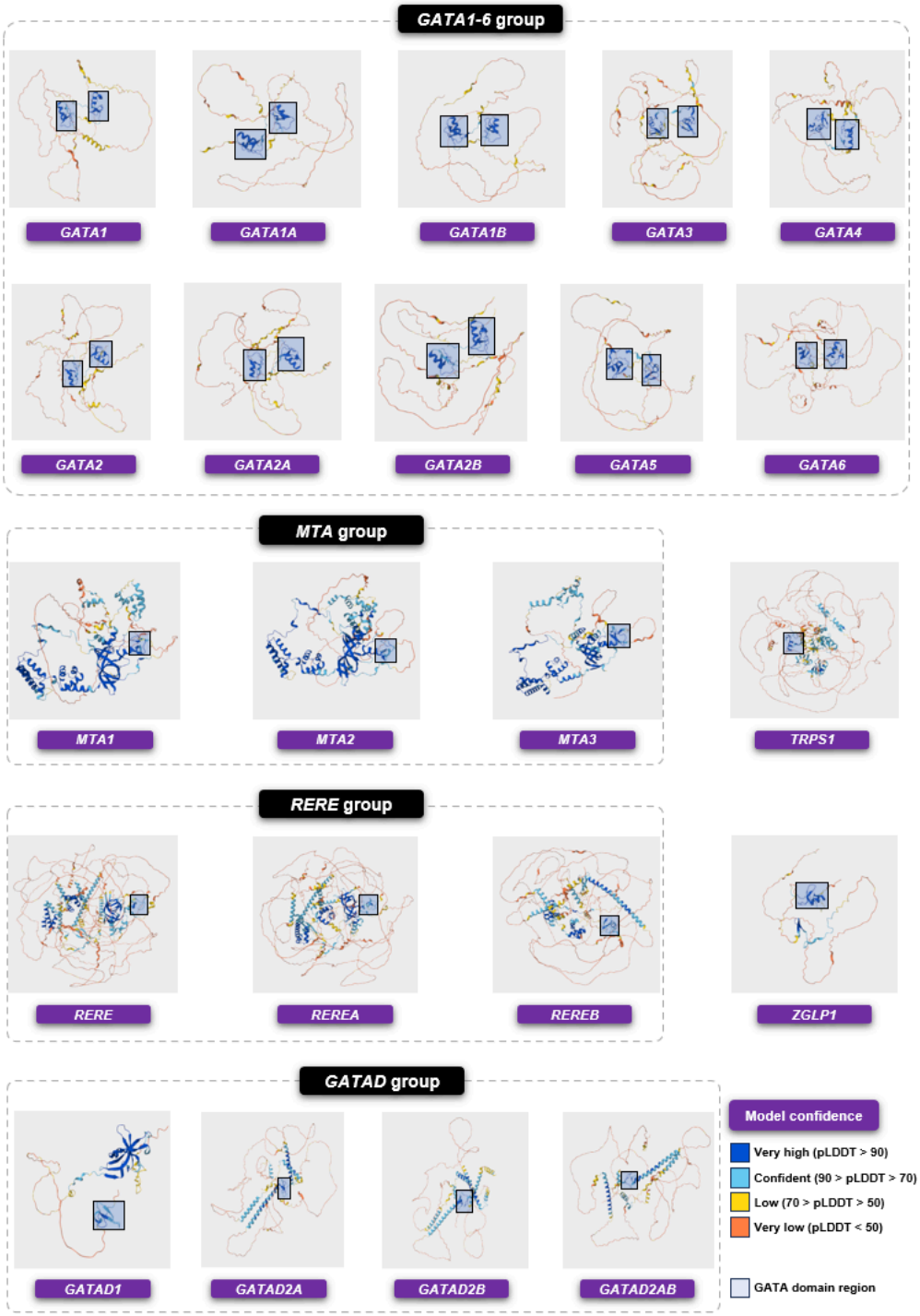


Figure 6

Fig 6. Three-dimensional structures of representative jawed vertebrate *zf-GATA* genes

Predicted structures of representative genes from the 22 jawed vertebrate *zf-GATA*gene groups are shown. The GATA domain region is highlighted with a semi-transparent blue box in each structure.

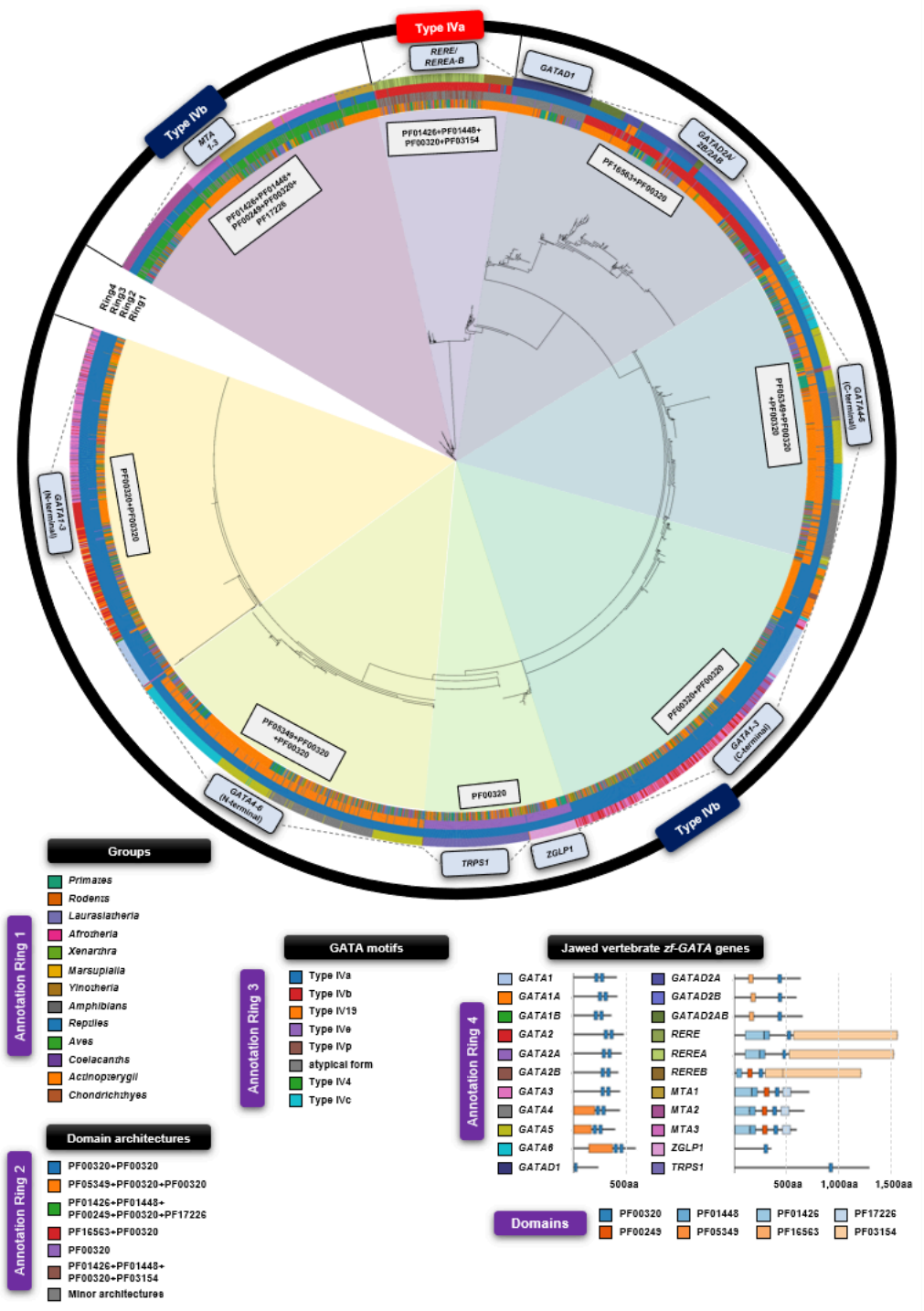


Figure 7

Fig 7. Phylogenetic relationships of jawed vertebrate GATA motifs

A maximum-likelihood phylogeny of 3,608 vertebrate GATA motifs was inferred using IQ-TREE2. Inner clades highlight representative *zf-GATA* genes with experimentally characterized biological functions. Concentric annotation rings provide layered information: Ring 1 denotes jawed vertebrate groups, Ring 2 represents domain architectures, Ring 3 indicates GATA motif types, and Ring 4 denotes individual *zf-GATA* gene groups.

Supplementary Files

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

- SupplementaryTableS1S4.xlsx