Title: Chromosomal rearrangements and segmental deletions drive gene loss in squamates

Buddhabhushan Girish Salve, Sonal H, Nagarjun Vijay

buddhabhushan21@iiserb.ac.in, sonal24@iiserb.ac.in, nagarjun@iiserb.ac.in

Computational Evolutionary Genomics Lab, Department of Biological Sciences, IISER Bhopal, Bhauri, Madhya Pradesh, India

Author for Correspondence: Nagarjun Vijay, nagarjun@iiserb.ac.in

Abstract

Background

Genomic rearrangements, including segmental deletions, duplications, translocations, and inversions of DNA segments, can contribute to gene losses, thereby reshaping genome architecture and potentially resulting in functional consequences. In squamates, karyotypic evolution mainly involves chromosome number reduction through fusions and microchromosome to macrochromosome translocations, although fissions have also contributed to diversification in several lineages. Despite these dynamics, the evolutionary processes and underlying genetic mechanisms driving chromosomal rearrangements and associated gene losses in squamates remain poorly understood.

Results

In this study, we analysed chromosome/scaffold-level assemblies of 265 squamates, corroborated by short-read, long-read, and transcriptomic data. We found multiple lines of evidence for the absence of 53 genes in the squamate lineage. Notably, increased genomic rearrangement activity appears to be associated with the squamate-specific loss of these genes, including *IL34*, *STAP1*, *LAPTM5*, and *TNIP2*, which are key regulators of macrophage activation and polarisation. Most of the missing genes were intact in tuatara and other vertebrate species. Furthermore, we find that many of the genes missing from squamates are organised in syntenic clusters and are involved in essential immune functions—raising important questions about whether their absence in squamates reflects true gene loss or simply results from incomplete assemblies.

Conclusions

Together, our comparative genomic analyses highlight that the loss of crucial genes in squamate lineages has occurred primarily through inter- and intrachromosomal rearrangements, including segmental deletions. These findings offer insights into the evolutionary loss of genes involved in macrophage differentiation and inform the development of novel pharmaceutical approaches for modulating immune responses.

Keywords

segmental deletions, macrophage polarisation, gene loss, karyotype evolution, squamates

Background

Genomic rearrangements, including deletions, inversions, translocations, and duplications, play a major role in shaping genome architecture and can potentially contribute to gene loss [1–3]. Such structural changes often disrupt gene integrity by breaking coding regions, altering regulatory landscapes, or relocating genes into heterochromatic or transcriptionally silent regions [4, 5]. For example, large-scale deletions can physically remove entire genes, whereas rearrangements such as translocations or inversions may separate essential gene components or place them under incompatible regulatory control, rendering them non-functional [3, 6]. Moreover, rearrangements mediated by repetitive elements can create unstable genomic regions prone to further structural variation, thereby accelerating gene erosion [7–11]. These processes are particularly evident in rapidly evolving lineages, such as birds and mammals, where gene loss contributes to lineage-specific adaptation and diversification [12–15].

The recent dramatic increase in the number of high-quality genome assemblies of diverse organisms has facilitated studies on lineage-specific gene losses [16–19]. The NCBI RefSeq annotation of orthologues for multiple vertebrate species and the availability of short- and long-read sequencing data allow the confirmation of gene losses [20, 21]. However, most studies linking gene loss and genomic rearrangements in vertebrate evolution have focused on birds [3, 22], mammals [10] and fishes [23], and the association of such chromosomal rearrangement and gene loss have not been systematically performed in squamates.

Squamate reptiles are a diverse monophyletic group of vertebrates classified into lizard and snake families, which diverged from tuatara approximately 250 million years ago (MYA) [24, 25]. With more than 11,000 described species, they are more species-rich than birds [26]. Squamates have emerged as excellent systems for studying genomic rearrangement [27], karyotype [28], genome size evolution [29], and unique biology such as tissue regeneration and limb degeneration [30, 31]. Yet, the dearth of genome assemblies has been a bottleneck for large-scale comparative genomic studies [32]. The recent availability of high-quality genomes across the squamate tree of life has opened new avenues to look at lineage-specific genome dynamics, such as sex chromosome evolution [33], changes in the repeat landscape [34], and GC content heterogeneity [35]. Notably, the genome size of squamates is diverse, with a trend toward genome size reduction [36]. In birds and bats, genome size reduction has reduced the gene repertoire with syntenic clusters of lost genes [36, 37]. However, the impact of these factors on the gene content of squamates is unexplored.

To address this gap, we focussed on gene loss in squamates and the role of genomic rearrangements. Our preliminary genome-wide screening of vertebrates revealed that 53 key genes with conserved functions are missing in squamates. Among these genes, previous studies have reported the absence of *PLAAT1*, *SLC24A1* and *IL34* in squamates and associated them with adaptation to low-light environments [38], changes in the visual system [39] and myeloid cell-type evolution [40], respectively. These studies suggest the evolutionary loss of these genes based on a search of the genome assemblies and failed to recover any remnants. Here, we leverage the dramatic increase in the availability of high-quality genomes for squamates coupled with a comprehensive search of genomic short- and long-read and transcriptomic data to rule out inadequacies in the genome assembly posing as gene loss [41–43]. We began our search by looking for genes annotated in the five outgroup species (i.e., Western clawed frog (*Xenopus tropicalis*), American alligator (*Alligator mississippiensis*), chicken (*Gallus gallus*), green sea turtle (*Chelonia mydas*), and human (*Homo sapiens*)) and missing in all annotated squamate genomes. Furthermore, we looked for direct (search of genomes, raw read datasets, statistical and phylogenetic signal of gene loss) and indirect evidence (co-evolution of the receptor) supporting the loss of these genes. We found that gene loss in squamates occurred in conserved syntenic clusters, similar to the pattern observed in birds [37]. The candidate genes we identified need to be evaluated with better-quality genome assemblies, and their functional impact should be assessed through comparative functional assays.

Results

In summary, our analysis identified 53 genes that are missing in the squamate lineage but retained in the five outgroup species (Additional file 1: Table S1). These missing genes are primarily associated with large-scale segmental deletions and chromosomal rearrangements, indicating a genomic basis for their loss. These genes tend to be of intermediate GC (39.56%

to 64.72%) and belong to multi-gene families with multiple paralogues. Despite extensive efforts, we could not recover them from genome assemblies, raw genomic reads, or transcriptomic datasets. Notably, four of the 53 missing genes are implicated in macrophage activation and polarisation.

Comparative analysis reveals large-scale gene loss in squamates

Comparative analysis revealed that 53 genes are missing in all 16 squamate species analysed, while 1-to-1 orthologues are annotated in outgroup lineages (see method section; Additional file 1: Table S1). Pairwise genome alignments (at these 53 gene loci) between chicken (*Gallus gallus*) and 27 other vertebrate species revealed that the squamate (~280 MYA) genomes consistently contain a higher number of unaligned 1 Kb windows compared to tuatara (*Sphenodon punctatus*; ~280 MYA), Testudines (~261 MYA), and Crocodilians (~245 MYA), despite these groups having comparable evolutionary divergence from chicken (**Fig. 1**; Additional file 1: Table S2; pairwise divergence time obtained from Timetree website [44]). We found that the number of unaligned 1 Kb windows is significantly higher in squamates compared to Testudines (Pairwise Wilcoxson adjusted $p < 0.0001$, n = 265), Crocodylia (adjusted $p = 0.00012$, n = 53), and Sphenodontia (adjusted $p = 0.035$, n = 53). However, no significant difference was observed between squamates and coelacanth (adjusted $p = 0.481$, n = 53) (Additional file 1: Table. S2 and Additional file 2: Fig. S1). The inconsistent occurrence of a higher number of unaligned windows in non-squamate species for some genes could be due to assembly gap and/or high sequence divergence. The aligned region could also be from the paralogues regions. We found a similar pattern in the case of whole genome alignments with humans as the focal species (Additional file 2: Fig. S2). A BLASTn search of 265 squamate genomes revealed that most missing genes are absent from squamates. Of the 53 genes, 21 have query coverage (search on squamate genomes using human orthologues as a query) of ≤ 25 %, 23 genes between 25 - 74 %, and 9 genes have coverage of ≥ 75 % (*SUV39H2*, *FREM2*, *DNM3*, *RBBP7*, *MOB1A*, *SSTR4*, *WNT2*, *LCP1* and *CAV3*; Additional file 1: Table S3). Our bioinformatic analysis (using BioMart paralogue annotation) revealed two or more paralogues for each gene having query coverage > 75 % (Additional file 1: Table S4). Furthermore, we assembled gene sequences using raw whole-genome sequencing datasets. Phylogenetic analysis of the retrieved sequences revealed that they cluster with closely related paralogues rather than the expected orthologues (Additional file 1: Table S5).

Gene absence occurs in conserved syntenic blocks and chromosomal rearrangements

We examined the genomic locations of the missing genes in the chicken genome and found that they are distributed across both macro- and microchromosomes (**Fig. 2**). This contrasts with previous studies on gene loss in birds, where missing genes were predominantly located on microchromosomes [37]. Most of the genes are missing singly, while we found clusters of missing genes on chromosomes 1 (*LCP1*, *RUBCNL*, *SIAH3*, *LRCH1*, *ARL11*, *DLEU7*, *STOML3*, *INTS6* and *FREM2*) and 4 (*TBC1D14*, *MAN2B2*, *GPR78*, *GRK4*, *TNIP2*, *KDM3A*) of chicken. We also noted 7 clusters of two or more genes within ~ 2 Mb of each other in the chicken genome. The cross-species comparison of syntenic loci revealed that these regions are prone to both intra- and inter-chromosomal rearrangements in squamate species such as the central bearded dragon (*Pogona vitticeps*), brown anole (*Anolis sagrei*), and terrestrial garter snake (*Thamnophis elegans*) (**Fig. 2**). In several cases, these loci are also mapped to telomeric regions. In contrast, comparable rearrangements were not observed in other reptiles, including the American alligator (*Alligator mississippiensis*) and the green sea turtle (*Chelonia mydas*), suggesting squamate-specific genomic rearrangements.

Orthologues of missing genes are GC-neutral, conserved and paralogue-enriched

It has been previously reported that missing genes in genome assemblies are often associated with high GC content [41, 42, 45, 46]. However, in our analysis, the mean GC content of the missing genes (calculated across vertebrate orthologues) ranged from 39.56% to 64.72% (**Fig. 3**). Notably, nineteen genes—*CAV3* (55.21%), *NOTCH2* (55.30%), *CYYR1* (55.43%), *DIP2A* (55.47%), *UNKL* (55.80%), *RIPPLY3* (56.06%), *ADRA1B* (56.48%), *IL34* (56.76%), *WNT2* (56.90%), *TNIP2* (57.83%), *SLC9A5* (58.12%), *MATK* (58.23%), *SIAH3* (58.52%), *DLEU7* (58.67%), *ARL11* (58.92%), *HSD17B1* (61.35%), *YBX3* (61.65%), *SH2D2A* (64.47%), and *SSTR4* (64.72%)—had GC content exceeding 55% (Additional file 1: Table S5).

Immune genes often evolve under strong selective pressure due to the ongoing evolutionary arms race between the host immune system and invading pathogens. This rapid evolution results in low amino acid conservation among orthologous proteins across species, limiting the effectiveness of sequence similarity-based tools such as BLASTn for detecting these genes [46–49]. Therefore, some of these genes may not be truly absent but may be present as highly diverged orthologues. We observed that 25 missing genes including *CYYR1*, *TMEM273*, *OLAH*, *PLAAT1*, *RIPPLY3*, *IL34*, *HHLA2*, *IFTAP*, *UTS2B*, *IL13RA2*, *SLC24A1*, *HSD17B1*, *TNIP2*, *RUBCNL*, *GPR78*, *SH2D2A*, *IL5RA*, *UNKL*, *ARL11*, *ZNF438*, *MAN2B2*, *DLEU7*, *MATK*, *KREMEN1*, and *GRK4* share less than 70% sequence identity with their human orthologues (Additional file 1: Table S5), suggesting that moderate to high sequence divergence may contribute to their loss or misannotation in squamate genomes.

Multi-gene families are often prone to inactivation or deletion, consistent with the birth-and-death model of gene family evolution [12, 23, 50, 51]. We found that 39 of the missing genes belong to multi-gene families and have at least one paralogue, except for the genes *CYYR1*, *DLEU7*, *IFTAP*, *IL26*, *IL34*, *LAPTM5*, *OLAH*, *PLAAT1*, *RIPPLY3*, *RUBCNL*, *TMEM273*, *TNIP2*, *UTS2B*, and *ZNF438*, which appear to be single-copy genes without detectable paralogues. (**Fig. 4**; Additional file 1: Table S4 and S5). Comparative genome alignment between human and tuatara reveals that several genes were intact in the ancestral lineage with conserved gene synteny, highlighting their loss as specific to squamates (Additional File 1: Table S6 and S7). Most of the missing genes lacked evidence of transcription in the RNA-seq datasets, with some genes having spurious mapping. We found evidence for segmental deletion in the case of 12 genes, namely, *GRK4*, *HHLA2*, *HSD17B1*, *IL34*, *IL5RA*, *LAPTM5*, *MOB1A*, *RBBP7*, *RIPPLY3*, *SLC24A1*, *SLC9A5* and *TBC1D14* (Additional file 1: Table S5).

Notably, genes such as *KDM3A* (GC content = 45.3335 ± 1.88272; Seq. identity = 81.22 ± 12.46; Max. Coverage in genome = 42; Hybpiper-based paralogue: *KDM3B*), *SUV39H2* (GC content = 42.62 ± 2.81; Seq. identity = 81.66 ± 11.62; Max. Coverage in genome = 77; Hybpiper-based paralogue: *SUV39H1*), and *RBBP7* (GC content = 45.91 ± 2.06; Seq. identity = 74.31 ± 9.65; Max. Coverage in genome = 82; Hybpiper-based paralogue: *RBBP4*), involved in chromatin remodelling, are missing in squamates (Additional file 1: Table S1-S5 and S8).

In summary, the comparison of gene orthologues revealed that many missing genes have moderate to high GC and high sequence divergence in ~ 47 % of identified missing genes, hindering detection. Most belong to multi-gene families, with ancestral presence confirmed in tuatara, indicating squamate-specific loss. The preliminary screen suggested an equipollency between genuine gene loss and the possibility that these genes belong to the genomic dark matter. However, based on our bioinformatic analysis, we infer that genes with GC content below 55 % and high sequence conservation across vertebrates are more likely to represent true losses, as our sequence similarity-based searches consistently failed to recover them.

Comparative genomic evidence for *IL34* loss via segmental deletion in squamates

As reported previously [40], our comparative genomic analysis confirmed the absence of the *IL34* gene in squamates despite its widespread conservation across other vertebrates. Synteny analysis revealed that *COG4* and *SF3B3* typically flank *IL34* on the left and *MTSS2* and *VAC14* on the right (**Fig. 5A**; Additional file 1: Table S9). Notable exceptions are observed in species such as the zebrafish (*Danio rerio*), green sea turtle (*Chelonia mydas*), American alligator (*Alligator mississippiensis*), and tiger rattlesnake (*Crotalus tigris*), which have *IONP2*, *PARD6A*, *BBS2*, and *CNOT1* respectively on the right flank of the *IL34* locus. The missing annotation of the *IL34* gene in squamates could potentially result from assembly artefacts. To rule out this possibility, we examined the syntenic locus of *IL34* in squamate species such as the common wall lizard (*Podarcis muralis*) and the Indian cobra (*Naja naja*) using PacBio and Oxford Nanopore long-read data. In both cases, overlapping reads span the *IL34* gene locus, indicating that the assemblies are correct at this locus (**Fig. 6**). BLASTn analysis of squamate genome assemblies and raw read datasets failed to recover any sequence corresponding to the *IL34* gene, further supporting its absence in this lineage and ruling out the possibility of its translocation to another genomic region (Additional file 1: Table S3). Next, we compared intergenic distances by examining squamate species for evidence of segmental deletions compared to non-squamate species (see Methods). Wilcoxon rank-sum tests (alternative = "less") consistently yielded adjusted *p*-values < 0.001 for all combinations of gene distance ratios involving *SF3B3–MTSS2*, indicating significant contraction in this region in squamates (**Fig. 5B**; Additional file 1: Table S5; Additional file 2: Fig. S3). In contrast, other gene combinations did not significantly differ across all comparisons. Additionally, evidence for segmental deletion was supported by phylogenetic logistic regression analyses, with significant results from both logistic_MPLE (slope = 0.2; *p*-value = 0.0022) and logistic_IG10 (slope = 0.2; *p*-value = 0.0018) (**Fig. 5C**; Additional file 1: Table

S5). Additionally, we found intrachromosomal rearrangement of the *IL34* locus in some of the Viperidae snake species (Additional file 2: Fig. S4).

Our bioinformatic analysis, integrating synteny, long-read mapping-based genome assembly verification, BLASTn searches, transcriptome searches, intergenic distance comparisons, and phylogenetic regression, robustly confirm the loss of the *IL34* gene in squamates.

Co-evolution of *CSF1R* with *IL34* gene loss in the squamate lineage

Our previous results from multifaceted approaches concurrently infer loss of the *IL34* gene from squamates. Next, we investigated the evolutionary consequences of *IL34* gene loss on its receptor —*CSF1R* in squamates. Previous studies have suggested that *IL34* remains relatively unchanged, with detectable co-variation with *CSF1R*, whereas no such correlation is observed for *CSF1* [Garceau et al. [52], reviewed in [53]]. However, this hypothesis remains untested in lineages where *IL34* is lost. The availability of high-quality genome assemblies of squamates and *IL34* gene loss opens the opportunity to address this interesting evolutionary question─how two ligands evolve with a single receptor? Using selection analysis tools, we detected signatures of relaxed selection acting on *CSF1R* in several squamate species. We found squamate species such as Western terrestrial garter snake (*Thamnophis elegans*; *p*-values = 0.0003, k-value = 0.63), komodo dragon (*Varanus komodoensis*; *p*-value = 0.0053, k-value = 0.62), and central bearded dragon (*Pogona vitticeps*; *p*-value = 0.0047, k-value = 0.59) show signatures of relaxed selection by RELAX (with *p*-values <0.05 and k-value <0) and/or codeml (**Fig. 7**; Additional file 1: Table S10). Notably, signatures of relaxed selection were not pervasive across the screened vertebrate lineages, suggesting lineage-specific evolutionary changes. Several sites within the extracellular domain of *CSF1R* were identified as positively selected by both MEME and FEL analyses, with squamate species exhibiting a relatively higher number of positively selected sites (Additional file 1: Table S10). Together, the results of comparative sequence analyses revealed lineage-specific modifications in *CSF1R* in squamates, potentially reflecting compensatory evolution following *IL34* gene loss. These findings provide indirect but compelling evidence for *IL34* gene loss in squamates and support the hypothesis of receptor adaptation following the loss of one of its ligands.

Four genes regulating macrophage activation and polarisation are absent in squamates

The activation of the *CSF1R* is mediated by its two ligands, *IL34* and *CSF1*—apex macrophage survival cytokines, which together constitute the *IL34*/*CSF1*/*CSF1R* axis—a key regulatory module governing the proliferation, differentiation, and survival of tissue macrophages, osteoclasts, Langerhans cells of the skin, Paneth cells of the small intestine, brain microglia, and B cell-stimulating mononuclear phagocytes. This axis is also central to macrophage polarisation [53–61]. Our previous results confirmed the loss of *IL34* in squamate reptiles. We identified three more genes, *LAPTM5*, *STAP1*, and *TNIP2*, that are missing from squamates (**Fig. 8A**).

Synteny analysis revealed a conserved gene order for these genes across multiple vertebrate species. For *LAPTM5*, the flanking genes *SDC3* (left) and *MATN1* (right) were conserved in human (*Homo sapiens*), green sea turtle (*Chelonia mydas*), Western clawed frog (*Xenopus tropicalis*), American alligator (*Alligator mississippiensis*), tuatara (*Sphenodon punctatus*), green anole (*Anolis carolinensis*), Eastern brown snake (*Pseudonaja textilis*), and chicken (*Gallus gallus*) (**Fig. 8B**; Additional file 1: Table S9). In the case of *STAP1*, the genes *CENPC* (left) and *UBA6* (right) were found to be consistently syntenic. We found evidence of segmental deletion for both genes, *STAP1* and *LAPTM5* (Additional file 2: Fig. S5). For *TNIP2*, conserved synteny was observed with *ADD1* and *SH3BP2* on the left flank and *FAM193A* and *RNF4* on the right in species such as human, green sea turtle, Western clawed frog, American alligator, and tuatara. However, in squamate reptiles, synteny is disrupted. For example, in green anole, the flanking genes of *TNIP2*—*ADD1* and *SH3BP2*—are located on chromosome 5, whereas *FAM193A* and *RNF4* are found on chromosome 4. Similarly, in the Indian cobra, *ADD1* is on chromosome 4 and *RNF4* on chromosome 3 (**Fig. 8B**; Additional file 2: Fig. S6). This inter-chromosomal rearrangement of *TNIP2*'s syntenic block in squamates may underlie its loss, consistent with structural genomic changes driving gene erosion. The missing genes *IL34* (56.76 ± 6.81 %), *LAPTM5* (56.76 ± 6.81), *STAP1* (43.01 ± 2.77 %), and *TNIP2* (57.83 ± 7.30 %; exhibit high GC in mammals) do not have extreme GC-content (**Fig. 8C**; Additional file 1: Table S5). The GO enrichment analysis reveals the involvement of these genes in multiple pathways, such as macrophage activation and migration, as well as the regulation of the immune system and cytokine-mediated signalling pathways (**Fig. 8D**). Our results suggest that the gene repertoire involved in the macrophage regulatory mechanism of squamates differs from that of other vertebrates with fundamental implications for macrophage activation and polarisation.

Discussion

Our study identifies large-scale gene losses across squamates, with 53 genes consistently missing in all 265 squamate genomes screened. The whole-genome alignments revealed that genes missing from squamates exhibit a significantly higher number of unaligned genomic regions than tuatara, testudines, and crocodilians at these loci in chicken. Among the 53 missing genes, 21 had < 25% query coverage, 23 had intermediate coverage of 25–74%, and 9 had ≥75% coverage in the genomes of squamates. These genes are dispersed across macro- and microchromosomes in the chicken genome and are typically lost individually, while some are missing in clusters. The mean GC content of the missing genes ranges from 39.56 % to 64.72 %, is within a GC-neutral range and suggests that GC bias is unlikely to cause them to be missing. Additionally, many of these genes exhibit less than 70% sequence identity to their human orthologues, indicating substantial evolutionary sequence divergence. The bioinformatic analysis revealed evidence for segmental deletion and genomic rearrangements. Despite extensive efforts, we could not recover these genes from available genomic or transcriptomic resources, including searches in high-quality genome assemblies, analyses of raw genomic sequencing reads, and examination of RNA-seq data across multiple squamate species. The consistent absence across these datasets suggests that these genes are potentially lost rather than missing due to assembly gaps and/or annotation errors and/or sequencing limitations. Our comparative genomic analyses confirmed the complete loss of *IL34* in squamates via segmental deletion and intrachromosomal rearrangements at the *IL34* locus in some *Viperidae* snakes. We also identified squamate-specific modifications in *CSF1R*, suggesting possible compensatory changes following *IL34* loss. Notably, several missing genes—including *IL34*, *LAPTM5*, *STAP1*, and *TNIP2*—are key regulators of immune processes such as macrophage activation, migration, and immune response modulation.

Evolutionary loss of conserved genes in squamates: patterns and limitations

In squamates, we discovered that segmental deletions and intra- and inter-chromosomal genomic rearrangements are plausible causes of these missing genes. In contrast with previously documented gene losses in birds, which are primarily found on microchromosomes, these genes are dispersed throughout different macro- and micro-chromosomes rather than clustered on microchromosomes in the chicken genome [37]. We observed that some gene losses occured in clusters, with two or more adjacent genes missing from the same genomic region. For *IL34*, both segmental deletions and intra-chromosomal rearrangements were observed, suggesting that these loci lie in structurally unstable regions. The pattern of *TNIP2* presence in the vicinity of the telomeric region of chicken chromosome 4 compared to interchromosomal rearrangement in squamates suggests that gene loss may be facilitated by prevalence of the gene in highly variable and rapidly evolving regions such as telomeric and sub-telomeric regions [62, 63].

Previous studies [37, 46, 47, 64] have shown that some genes initially reported as missing in birds were later recovered using transcriptomic data [43], with their absence attributed to high GC content and limitations in earlier genome assemblies [42]. We incorporated GC content analysis to avoid similar errors, searched short- and long-read datasets, and examined RNA-seq data to reduce false gene absence in squamates. The GC content of missing genes in squamates is less than that reported for the missing GC-rich genes in birds, which were later found to be "dark matter" missing in prior genome assemblies [41, 42, 65–67]. While lineage-specific GC content shifts in squamates cannot be ruled out, the presence of most missing genes in tuatara weakens the likelihood of a widespread increase in GC content in squamates. Furthermore, we assembled gene sequences by performing BLASTn searches using raw whole-genome sequencing datasets. Phylogenetic analysis of the retrieved sequences revealed that they clustered with closely related paralogues rather than the expected orthologues. We verified genome assemblies for conserved synteny and ruled out assembly artefacts as the cause of missing genes, suggesting that structural rearrangements, rather than assembly artefacts, may play a role in gene loss. Many of the missing genes, including several interleukins, show less than 70% identity to their human orthologues and are primarily linked to immune functions (for gene-specific detailed references, see Additional file 1: Table S8). A plausible explanation is the involvement of these genes in host-pathogen arms races, where rapid evolution driven by changing selective pressures can lead to high sequence divergence and positive selection [68–71]. Consequently, these genes may still exist in squamates but in highly diverged forms, making them difficult to detect using standard sequence homology-based approaches. Furthermore, we could not accurately estimate deletion sizes, as the genes are absent in all the examined squamates, and the nearest outgroup with intact genes, tuatara, diverged ~250 MYA [25]. This deep divergence limits the precise delineation of deletion boundaries, unlike in previous studies, where segmental deletions could be more clearly defined [6, 48].

The discovery of squamate-specific gene losses offers insights into their evolutionary importance, supporting the idea that gene loss can be adaptive and extending the "less-is-more" hypothesis to squamate [62, 72–76]. Many missing genes have multiple paralogues; in some cases, their apparent absence may result from assembly limitations. However, if the gene loss is true, their functions may be compensated by the paralogues. Thus, high-quality, haplotype-resolved telomere-to-telomere chromosome-level assemblies are crucial to distinguish gene loss from missing genes in incomplete assemblies. Additionally, most missing genes form conserved syntenic clusters in non-squamate vertebrates, suggesting their loss in squamates likely involved block deletions, potentially contributing to genome size reduction [36]. Future genome assemblies should aim to assess the presence or absence of these genes to understand the mechanisms and evolutionary consequences of their loss.

Functional implications of missing genes

Earlier studies reported the absence of *PLAAT1*, *SLC24A1*, and *IL34* in squamates and associated their loss with adaptations to low-light environments, modifications in the visual system, and the evolution of myeloid cell types, respectively [38–40]. Our methodology has extended the search genome-wide and identified these genes as missing, reinforcing the reliability of our gene loss detection approach and highlighting the need for detailed functional investigations of additional missing genes. A literature survey revealed that the 53 genes missing from squamates are linked to diverse biological functions, including immunity and inflammation, cell growth, apoptosis, cancer, development, cardiovascular and metabolic processes, musculoskeletal formation, sensory function, and core cellular mechanisms like autophagy, protein degradation, and transcriptional regulation (for gene-specific detailed references, see Additional file 1: Table S8). The involvement of these genes in fundamental physiological and immune functions raises critical questions about how these pathways operate in their absence, offering a valuable opportunity to explore compensatory mechanisms and lineage-specific adaptations. The loss of these genes may play a pivotal role in evolutionary genetics by influencing genome size and complexity, affecting the rate of evolution and adaptation, offering insights into gene and pathway essentiality, and contributing to phenotypic diversity in squamates [12]. Furthermore, we found that the genes *KDM3A*, *SUV39H2,* and *RBBP7*, which are missing in squamates, are associated with the Gene Ontology (GO) term "chromatin remodeling", and are involved in DNA replication, recombination, and repair. The absence of these genes suggests a potential impact on chromatin architecture, with possible consequences for genome accessibility and stability in squamates [3]. The genes *FREM2*, *WNT2* and *NOTCH2* involved in eye development are missing and may be associated with changes in the visual system in squamates [30, 38, 39].

The present study reports the loss of four genes─ *IL34*, *STAP1*, *LAPTM5* and *TNIP2*- involved in macrophage activation and polarisation. *IL34* expression is observed in tissues such as the skin, brain, kidneys, and testes and is modulated in diseases like Alzheimer's, cancer and HBV viral infection [77–79]. The macrophages derived from *IL34* exhibit more anti-viral activity than those from *CSF1* [57, 80, 81]. Despite having a normal phenotype, *IL34*$^{-/-}$ mice exhibit impaired responses to skin antigens and viral infections, highlighting its physiological importance [59, 82]. These findings suggest that squamates may be functionally compromised owing to the loss of *IL34,* potentially rendering them more susceptible to pathogenic infections. Furthermore, during metamorphosis, *CSF1R* gene expression increases in the muscle tissue of regressing tails in metamorphosing animals, consistent with previous studies demonstrating the crucial roles of macrophages in tail and body muscle remodelling during metamorphosis (Additional file 2: Fig. S7) [83, 84]. Our comparison of RNA-seq expression data from metamorphosing animals revealed that *IL34* is not abundant, and the ligand-receptor pair *CSF1-CSF1R* is expressed during the inflammatory phase of tail regeneration [85]. In mice, *STAP1* deficiency impairs TCR-mediated T cell activation [86], *TNIP2* acts as a regulator of NF-κB signalling [87, 88], and *LAPTM5* inhibits HIV-1 progeny infectivity by transporting viral Env to the lysosome for degradation in macrophages [89]. However, the role of these genes in reptiles remains largely unexplored and needs functional validation. The limited availability of tools for assessing immune function in free-living reptiles presents a challenge in functionally testing this scenario. These findings offer new insights into the evolutionary reduction in the gene repertoire regulating macrophage differentiation, which may help develop novel pharmaceutical strategies for the immune modulation of monocytes and macrophages.

Conclusions

This study systematically identified and confirmed that 53 genes are missing in the squamate lineage but retained in outgroup species, suggesting lineage-specific gene loss events primarily associated with large-scale segmental deletions and chromosomal rearrangements. Notably, four missing genes are involved in macrophage activation and polarisation─*IL34*, *STAP1*, *LAPTM5* and *TNIP2*, while others play key roles in immunity, development, metabolism, cellular processes, and neuronal and musculoskeletal functions. The missing genes are GC-neutral and are members of multi-gene families with several paralogues. Our study provides direct and indirect evidence of potential gene losses in the squamate lineage. Despite comprehensive searches across genome assemblies, raw reads, and transcriptomes, none could be recovered, supporting their possible loss in squamates. Further studies with the aim of generating high-quality genomes of squamates are needed to identify the causative forces that triggered these gene losses and their final confirmation.

Methods

Identification of genes missing from squamates

We retrieved the list of all 20,595 protein-coding genes annotated in the human genome from NCBI. The orthologues for each human gene across 679 annotated vertebrate genomes were obtained using the datasets command-line utility of NCBI (parameters: --ortholog all) [90]. To identify candidate missing genes specific to squamates, we focused on genes consistently identified as missing (no orthologue is annotated in NCBI) across all 16 annotated squamate genomes. As a filtering criterion to ensure that these genes were not universally absent or pseudogenised or annotation artefact, we retained only those genes that were classified as intact (annotated without the "Low quality protein" tag of NCBI) in multiple distantly related species, namely Western clawed frog (*Xenopus tropicalis*), American alligator (*Alligator mississippiensis*), chicken (*Gallus gallus*), green sea turtle (*Chelonia mydas*), and human (*Homo sapiens*) allowing to confidently infer squamate-specific missing genes by comparing across a broad phylogenetic tree of vertebrates (Additional file 1: Table S1).

Quantifying cross-species alignment conservation

To validate the absence of missing genes in squamates, we quantified the alignment coverage of diverse vertebrates and compared squamates vs non-squamate orders. For this, we analysed pairwise whole-genome alignments between chicken (GRCg7b) and 27 other vertebrate species, including squamates, using precomputed LASTZ alignments from the Ensembl Comparative database (https://ftp.ensembl.org/pub/current_maf/ensembl-compara/pairwise_alignments/) [91]. The multiple alignment format (MAF) files for each species were downloaded and processed using MafFilter v1.3.1 [92] with the OutputCoordinates filter enabled (output_src_size=yes) to extract conserved syntenic blocks between chicken and the focal species. The aligned regions in close proximity were merged using BEDTools v2.27.1 merge (parameters: -d 5) [93]. To quantify the presence or absence of genomic regions, we created non-overlapping 1 Kb windows across the chicken genome using BEDTools makewindows with (parameters: -w 1000) based on the GRCg7b reference genome. Next, we used BEDTools intersect with the -wao option to compute overlap between each 1 Kb window and the aligned regions for each focal species. Finally, to assess gene-wise conservation, we intersected the annotated coordinates of squamate-specific missing protein-coding genes with the 1 Kb windows. For each gene and species pair, we calculated the number of 1 Kb windows within the gene with no alignment and normalised it by the total gene length (in Kb) to compute an alignment score. Finally, the resulting matrix (gene × species) of normalised aligned window counts was visualised as a heatmap using the tidyheatmap v1.11.6 R package [94]. A similar plot was generated for humans as a reference species. Pairwise Wilcoxon rank-sum tests (alternative = "greater") were performed to compare the number of unaligned 1 Kb windows between squamates and other orders. Multiple testing correction was applied using the "BH" method, and adjusted p-values were used to assess significance (Additional file 1: Table S2).

Analysis of evolutionary rearrangement and chromosomal mapping of missing genes

To investigate the chromosomal distribution and rearranged genomic regions of missing genes across squamates, we mapped them to chicken karyotypes. We used GENESPACE v1.3.1 [95] to understand patterns of synteny and orthology across chicken (*Gallus gallus*), American alligator (*Alligator mississippiensis*), green sea turtle (*Chelonia mydas*), central bearded dragon *(Pogona vitticeps*), brown anole (*Anolis sagrei*), terrestrial garter snake (*Thamnophis elegans*) genomes. A comparative framework enabled the visualisation of conserved syntenic blocks and rearranged genomic regions linked to missing genes in squamates. We used BEDTools nuc to calculate GC content across 100 bp sliding windows. To identify sequences associated with non-B DNA-forming motifs, we used the GFA v2 program [96]. The resulting outputs were used to plot the GC content and the presence of non-B DNA motifs across chicken chromosomes. Additionally, the missing gene loci were compared between genome assemblies of chicken (*Gallus gallus*) and green anole (*Anolis carolinensis*) using the NCBI Comparative Genome Viewer [97].

BLASTn search of missing genes in squamate genomes and raw read datasets

To increase the power of screening for missing genes in squamate genomes, we performed BLASTn [98] searches (parameters: -evalue 0.05 -outfmt '17 SQ', 1, 6, and 7) across 265 squamate genome assemblies, including both chromosome-level and scaffold-level assemblies, using human (*Homo sapiens*) gene sequences as queries (Additional file 1: Table S11). The BLASTn output in format '17 SQ' was used to generate IGV reports [99], which were then screened to assess the presence of gene sequences in the squamate assemblies. Furthermore, we screened publicly available short- and long-read DNA and RNA sequencing data from PacBio (HiFi and Revio), Oxford Nanopore, and Illumina sequencing technologies—from squamate species (Additional file 1: Table S12). We employed BLASTn (parameters: -evalue 0.05 -max_target_seqs 10000) with query sequences of missing genes of the five outgroup species. Additionally, we used HybPiper v2.3.2 [100] to recover genes from raw read datasets. To identify paralogues, the sequence obtained from HybPiper was analysed using IQ-TREE v2.3.6 [101], along with coding DNA sequences from 39 species that produced BLASTn hits to the extracted sequence (parameters: -evalue 0.001)(Additional file: Table S12). We used the ‘ape’ package to identify the closest phylogenetic branches to the HybPiper recovered gene sequence [102].

Furthermore, we used Ensembl BioMart to obtain paralogue, GC content, sequence identity, and transcript count data for human and chicken orthologues of missing squamate genes.

To explore potential transcriptional evidence for missing genes, we mapped RNA-seq data from the central bearded dragon (*Pogona vitticeps*) of the testis, muscle, lung, liver, kidney, eye, heart, and brain tissues to the chicken genome using minimap2 v2.29-r1283 [103] (parameters: -ax splice -uf -k14) and FLAIR v2.2.0 [104] (Additional file 1: Table S12). We then screened for the presence of transcripts corresponding to candidate missing genes and visualised the read coverage and splice junctions using IGV-report, allowing us to assess transcriptional signals for these genes in central bearded dragon (*Pogona vitticeps*) despite their apparent genomic absence.

Identification of 1-to-1 orthologues, assembly verification and GC content of missing genes

We confirmed gene synteny of lost genes and identified 1-to-1 orthologues across eight vertebrate species, namely Western clawed frog (*Xenopus tropicalis*), American alligator (*Alligator mississippiensis*), chicken (*Gallus gallus*), green sea turtle (*Chelonia mydas*), green anole (*Anolis carolinensis*), Eastern brown snake (*Pseudonaja textilis*), tuatara (*Sphenodon punctatus*) and human (*Homo sapiens*) using NCBI genome viewer [21] and ENSEMBL [91]. To assess the integrity of the missing gene locus in the chromosome/scaffold-level assembly of Indian cobra (*Naja naja*) and common wall lizard (*Podarcis muralis*). We used the *scan_alignment* and *alignment_plot* options in the Klumpy v1.0.10 tool [105]. A required BAM file was generated by mapping long-read sequencing data from squamates species (Additional file 1: Table S12) with Minimap2 v2.17-r941 [103] (parameters: -ax map-pb or -ax map-hifi or -ax map-ont). The resulting alignment plots were examined for the presence of continuous and overlapping reads spanning the missing gene loci.

The missing gene annotations in the genome and/or missing gene sequence from the genome assembly of species could be due to its high GC content and/or high sequence divergence [41, 42, 106]. Therefore, we calculated GC content across vertebrate orthologues using seqkit (parameters: fx2tab --name –gc) and sequence identity with respect to humans using EMBOSS-needle [107].

Squamate lineage-specific evidence of segmental deletion

a. Dataset preparation

We developed an integrated analysis pipeline encompassing orthologue retrieval, synteny mapping, intergenic distance measurement, and gap detection to investigate segmental deletions underlying missing genes. Firstly, syntenic genes flanking each focal gene were identified from the human genome assembly (GRCh38.p14) using the NCBI gene feature table. Four neighbouring syntenic genes were extracted (two on both sides of the focal gene), and orthologues were retrieved across vertebrate genomes using NCBI datasets (parameters: --include product-report --ortholog all). The longest isoform was selected for each orthologue based on protein length and species, where orthologues of all four syntenic genes were annotated. We calculated pairwise intergenic distances using BEDTools closest. Gene pairs on different chromosomes were flagged as potential Evolutionary Breakpoint Regions (EBRs). A comprehensive gene distance matrix with complete orthologous information was assembled for all species. Next, to detect gaps in the locus, genomic intervals spanning the outermost flanking genes were merged with BEDTools merge -d 100000000, and the corresponding sequences were retrieved using NCBI efetch. These regions were scanned for assembly gaps using the Klumpy v1.0.10 find_gaps subprogram, and the total number of gaps was recorded per species. Each species' taxonomy classification (class and order) was obtained using NCBI Taxonomy resources (esearch and efetch). Finally, all data—including intergenic distances, gap counts, and taxonomic info—were consolidated into a unified summary table. A subset of species with complete data was used for downstream visualisation (heatmap) and phylogenetic logistic regression analysis implemented in R. Based on the presence of intact synteny and absence of gaps, species were classified as having either an intact locus (non-squamates) or a potentially deleted locus (squamates).

b. Statistical analysis of gene distance and phylogenetic signal

i. Ratio-Based Comparative Analysis

To evaluate whether changes in intergenic distances between the focal gene (genes which are missing in squamates) and its flanking genes are associated with segmental deletion, we first excluded species exhibiting large-scale genomic disruptions such as gaps and EBRs. We calculated six pairwise ratios representing the remaining species' upstream, downstream, and across-gene intergenic distances. We conducted Wilcoxon rank-sum tests for each ratio with the alternative hypothesis set to “less” to compare distributions between species with intact versus deleted gene loci. The resulting *p*-values were adjusted for multiple testing using the FDR method, and significant comparisons were visualised as a heatmap using the pheatmap package in R [108].

ii. Phylogenetic Logistic Regression

To quantify the phylogenetic association between relative gene distance and gene loss, we performed phylogenetic logistic regression using the “phylolm” R package [109]. Squamates (0) and non-squamates (1) were used as the binary variable. The independent variable was the relative intergenic distance for the focal gene, computed as:

$$\underset{\longleftarrow}{\boldsymbol{G_{L2}}} \qquad \underset{\longrightarrow}{\boldsymbol{G_{L1}}} \qquad \underset{\Longrightarrow}{\boldsymbol{focal\ gene}} \qquad \underset{\longleftarrow}{\boldsymbol{G_{R1}}} \qquad \underset{\longrightarrow}{\boldsymbol{G_{R2}}}$$

Where

G_{L1}: Gene on the left flank of a lost gene on 1^{st} position

G_{L2}: Gene on the left flank of a lost gene on 2^{nd} position

G_{R1}: Gene on the right flank of a lost gene on 1^{st}position

G_{R2}: Gene on the right flank of a lost gene on 2^{nd}position

$focal\ gene$: missing gene in squamates

The direction of the arrow indicates gene orientation

$$Relative\ distance\ in\ percentage\ = \left(\frac{d_{focal}}{d_{upstream} + d_{focal} + d_{downstream}}\right) \times 100$$

Where

d_{focal} = distance between G_{L1} and G_{R1}

$d_{downstream}$ = distance between G_{L1} and G_{L2}

$d_{upstream}$ = distance between G_{R1} and G_{R2}

Two methods were used for model fitting: logistic_MPLE and logistic_IG10 (Additional file 1: Table S5).

Signatures of receptor-ligand co-evolution as indirect evidence of *IL34* loss in squamates

To explore the evolutionary impact of *IL34* loss in squamates, we investigated the *IL34–CSF1–CSF1R* signalling axis, focusing on the functional interplay between these ligands (*IL34*, *CSF1*) and their shared receptor *CSF1R*. *CSF1R* is a type III receptor tyrosine kinase critical for developing, surviving, and functioning myeloid lineage cells [61]. Previous studies have shown evolutionary co-variation between *IL34* and *CSF1R*, but not between *CSF1* and *CSF1R* [Garceau et al. [52] and reviewed in [53]]. Based on this, we hypothesised that *IL34* loss may influence the evolution and function of *CSF1R* in squamates. We leveraged ENSEMBL annotation for *CSF1R* across vertebrates and retrieved the coding DNA sequence (CDS) of the *CSF1R* with a conserved exon phase (Additional file: Table S10). Multiple sequence alignments (MSAs) of intact *CSF1R* from Squamates and vertebrate clades were generated using PRANK v.170427 in the Guidance2 suite [110, 111]. The MSAs were trimmed using trimAI v1.2 [112] and used to assess the strength of selection. Further, the alignments and phylogenetic tree obtained from the Timetree website (https://timetree.org/) were used to investigate the strength of selection [44]. We relied on the dN/dS-based method of Hyphy v2.5.48 and PAML v4.9f [113, 114]. Our selection suite consists of RELAX, BUSTED, aBSERL, MEME, and FEL frameworks of branch-site and site models of Hyphy, along with the null (M0) and branch (branch-free and branch-neutral) models of codeml [115–121]. To investigate whether the gene evolves under relaxed selection, we specified one branch as the foreground and all other branches as the background at a time in the RELAX model of the HyPhy program [68] and codeML of the PAML program [69]. A significant result of $k > 1$ indicates intensified selection along test branches, while $k < 1$ suggests relaxed selection, with significance determined at $p < 0.05$. We also implemented aBSREL and BUSTED, which are branch and gene-wide tests, respectively, to test for signatures of selection on branches of the genes. Site-specific models, such as FEL and MEME, were used to identify sites evolving under positive and negative selection, and ggplot2 was used for visualisation in the R program [122].

Gene loss linked to macrophage activation and polarisation

CSF1 and *IL34* are apex macrophage survival cytokines [53, 57]. Therefore, we investigated additional genes involved in macrophage polarisation to assess broader immune regulatory shifts in squamates. Based on our search "macrophage" word from QuickGO (https://www.ebi.ac.uk/QuickGO/api/index.html#!/gene_ontology), we identified four genes─*IL-34*, *LAPTM5*, *STAP1*, and *TNIP2* missing in squamates. We used Morpheus, https://software.broadinstitute.org/morpheus, to create a heatmap. Previous studies have shown that *CSF1R* expression increases in the muscle tissue of regressing tails during metamorphosis [83, 84, 123], suggesting a role in tissue remodelling and immune modulation. To test whether similar expression patterns occur during tail regeneration, we analysed publicly available RNA-seq datasets from the regenerating tails of Western clawed frog (*Xenopus tropicalis*; with intact *IL34*), green anole (*Anolis carolinensis*; with *IL34* loss) and tokay gecko (*Gekko gecko*; with *IL34* loss; Additional file 2: Fig. S7). Transcript abundance was quantified using kallisto v0.51.1 [124], and expression levels were obtained in transcripts per million (TPM). This approach allowed us to assess macrophage-related gene expression's conservation and functional relevance during regeneration across species with and without *IL34*.

Data availability

The data and relevant code for this study are available on GitHub:
https://github.com/CEGLAB-Buddhabhushan/Missing_genes_in_squamata.git

Abbreviations

EBR: Evolutionary Breakpoint Region

HBV: Hepatitis B Virus

MYA: Million Years Ago

TCR: T Cell Receptor

PacBio: Pacific Biosciences

ONT: Oxford Nanopore Technologies

FDR: False Discovery Rate

BH: Benjamini–Hochberg correction

NCBI: National Center for Biotechnology Information

BLASTn: Basic Local Alignment Search Tool (nucleotide)

MAF: Multiple Alignment Format

IGV: Integrative Genomics Viewer

RNA-seq: RNA sequencing

Mb: Megabase

Kb: Kilobase

References

1. Burssed B, Zamariolli M, Bellucco FT, Melaragno MI. Mechanisms of structural chromosomal rearrangement formation. Mol Cytogenet. 2022;15.

2. Claude SJ, Park S, Park SJ. Gene loss, genome rearrangement, and accelerated substitution rates in plastid genome of Hypericum ascyron (Hypericaceae). BMC Plant Biol. 2022;22.

3. Huang Z, De O. Furo I, Liu J, Peona V, Gomes AJB, Cen W, et al. Recurrent chromosome reshuffling and the evolution of neo-sex chromosomes in parrots. Nat Commun. 2022;13.

4. Harewood L, Fraser P. The impact of chromosomal rearrangements on regulation of gene expression. Hum Mol Genet. 2014;23:R76–82.

5. Stewart NB, Rogers RL. Chromosomal rearrangements as a source of new gene formation in Drosophila yakuba. PLOS Genet. 2019;15:e1008314.

6. Salve BG, Kurian AM, Vijay N. Concurrent loss of ciliary genes WDR93 and CFAP46 in phylogenetically distant birds. 2023. https://doi.org/10.1098/rsos.230801.

7. Stankiewicz P, Shaw CJ, Withers M, Inoue K, Lupski JR. Serial segmental duplications during primate evolution result in complex human genome architecture. Genome Res. 2004;14:2209–20.

8. Hinsch H, Hannenhalli S. Recurring genomic breaks in independent lineages support genomic fragility. BMC Evol Biol. 2006;6:1–12.

9. Emanuel BS, Shaikh TH. Segmental duplications: an "expanding" role in genomic instability and disease. Nat Rev Genet. 2001;2:791–800.

10. Shinde SS, Sharma S, Teekas L, Sharma A, Vijay N. Recurrent erosion of COA1/MITRAC15 exemplifies conditional gene dispensability in oxidative phosphorylation. Sci Reports 2021 111. 2021;11:1–16.

11. Meredith RW, Gatesy J, Cheng J, Springer MS. Pseudogenization of the tooth gene enamelysin (MMP20) in the common ancestor of extant baleen whales. Proc R Soc B Biol Sci. 2011;278:993–1002.

12. Albalat R, Cañestro C. Evolution by gene loss. Nat Rev Genet 2016 177. 2016;17:379–91.

13. Blumer M, Brown T, Freitas MB, Destro AL, Oliveira JA, Morales AE, et al. Gene losses in the common vampire bat illuminate molecular adaptations to blood feeding. Sci Adv. 2022;8:6494.

14. Sharma V, Hecker N, Roscito JG, Foerster L, Langer BE, Hiller M. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. Nat Commun 2018 91. 2018;9:1–9.

15. Guijarro-Clarke C, Holland PWH, Paps J. Widespread patterns of gene loss in the evolution of the animal kingdom. Nat Ecol Evol 2020 44. 2020;4:519–23.

16. Koepfli KP, Paten B, O'brien SJ, Antunes A, Belov K, Bustamante C, et al. The Genome 10K Project: a way forward. Annu Rev Anim Biosci. 2015;3:57–111.

17. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. Phylogenomic analyses data of the avian phylogenomics project. Gigascience. 2015;4.

18. Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, et al. Six reference-quality genomes reveal evolution of bat adaptations. Nature. 2020;583:578–84.

19. Stiller J, Feng S, Chowdhury AA, Rivas-González I, Duchêne DA, Fang Q, et al. Complexity of avian evolution revealed by family-level genomes. Nat 2024 6298013. 2024;629:851–60.

20. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2007;35 SUPPL. 1.

21. Rangwala SH, Kuznetsov A, Ananiev V, Asztalos A, Borodin E, Evgeniev V, et al. Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). Genome Res. 2021;31:159–69.

22. Friocourt F, Lafont AG, Kress C, Pain B, Manceau M, Dufour S, et al. Recurrent DCC gene losses during bird evolution. Sci Reports 2017 71. 2017;7:1–11.

23. Kato A, Pipil S, Ota C, Kusakabe M, Watanabe T, Nagashima A, et al. Convergent gene losses and pseudogenizations in multiple lineages of stomachless fishes. Commun Biol. 2024;7:408.

24. Simões TR, Caldwell MW, Tałanda M, Bernardi M, Palci A, Vernygora O, et al. The origin of squamates revealed by a Middle Triassic lizard from the Italian Alps. Nature. 2018;557:706–9.

25. Gemmell NJ, Rutherford K, Prost S, Tollis M, Winter D, Macey JR, et al. The tuatara genome reveals ancient features of amniote evolution. Nat 2020 5847821. 2020;584:403–9.

26. Simões TR, Pyron RA. THE SQUAMATE TREE OF LIFE. Bull Museum Comp Zool. 2021;163:47–95.

27. Mezzasalma M, Macirella R, Odierna G, Brunelli E. Karyotype Diversification and Chromosome Rearrangements in Squamate Reptiles. Genes 2024, Vol 15, Page 371. 2024;15:371.

28. Waters PD, Patel HR, Ruiz-Herrera A, Alvarez-Gonzalez L, Lister NC, Simakov O, et al. Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. Proc Natl Acad Sci U S A. 2021;118:e2112494118.

29. Pinto BJ, Gamble T, Smith CH, Wilson MA. A lizard is never late: Squamate genomics as a recent catalyst for understanding sex chromosome and microchromosome evolution. J Hered. 2023;114:445–58.

30. Emerling CA. Genomic regression of claw keratin, taste receptor and light-associated genes provides insights into biology and evolutionary origins of snakes. Mol Phylogenet Evol. 2017;115:40–9.

31. Roscito JG, Sameith K, Kirilenko BM, Hecker N, Winkler S, Dahl A, et al. Convergent and lineage-specific genomic differences in limb regulatory elements in limbless reptile lineages. Cell Rep. 2022;38:110280.

32. Gable SM, Mendez JM, Bushroe NA, Wilson A, Byars MI, Tollis M. The State of Squamate Genomics: Past, Present, and Future of Genome Research in the Most Speciose Terrestrial Vertebrate Order. Genes (Basel). 2023;14:1387.

33. Modi WS, Crews D. Sex chromosomes and sex determination in reptiles: Commentary. Curr Opin Genet Dev. 2005;15:660–5.

34. Gable SM, Bushroe NA, Mendez JM, Wilson A, Pinto BJ, Gamble T, et al. Differential Conservation and Loss of Chicken Repeat 1 (CR1) Retrotransposons in Squamates Reveal Lineage-Specific Genome Dynamics Across Reptiles. Genome Biol Evol. 2024;16.

35. Alföldi J, Di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, et al. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature. 2011;477:587–91.

36. Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci U S A. 2017;114:E1460–9.

37. Lovell P V, Wirthlin M, Wilhelm L, Minx P, Lazar NH, Carbone L, et al. Conserved syntenic clusters of protein coding genes are missing in birds. Genome Biol. 2014;15:565.

38. Drabeck DH, Wiese J, Gilbertson E, Arroyave J, Stiassny MLJ, Alter SE, et al. Gene loss and relaxed selection of plaat1 in vertebrates adapted to low-light environments. Proc R Soc B. 2024;291.

39. Gower DJ, Fleming JF, Pisani D, Vonk FJ, Kerkkamp HMI, Peichl L, et al. Eye-Transcriptome and Genome-Wide Sequencing for Scolecophidia: Implications for Inferring the Visual System of the Ancestral Snake. Genome Biol Evol. 2021;13.

40. Pinheiro D, Mawhin MA, Prendecki M, Woollard KJ. In-silico analysis of myeloid cells across the animal kingdom reveals neutrophil evolution by colony-stimulating factors. Elife. 2020;9:1–20.

41. Kim J, Lee C, Ko BJ, Yoo DA, Won S, Phillippy AM, et al. False gene and chromosome losses in genome assemblies caused by GC content variation and repeats. Genome Biol. 2022;23:1–29.

42. Hron T, Pajer P, Pačes J, Bartůněk P, Elleder D. Hidden genes in birds. Genome Biol. 2015;16:164.

43. Bornelöv S, Seroussi E, Yosefi S, Pendavis K, Burgess SC, Grabherr M, et al. Correspondence on Lovell et al.: identification of chicken genes previously assumed to be evolutionarily lost. Genome Biol. 2017;18:112.

44. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34:1812–9.

45. Hara Y, Kuraku S. The impact of local genomic properties on the evolutionary fate of genes. Elife. 2023;12.

46. Rohde F, Schusser B, Hron T, Farkašová H, Plachý J, Härtle S, et al. Characterization of chicken tumor necrosis factor-α, a long missed cytokine in birds. Front Immunol. 2018;9 APR:605.

47. Janusova S, Krchlikova V, Hron T, Elleder D, Stepanek O. Identification of GC-rich LAT genes in birds. PLoS One. 2023;18 4 April:11–4.

48. Salve BG, Sharma S, Vijay N. Evolutionary diversity of CXCL16-CXCR6: Convergent substitutions and recurrent gene loss in sauropsids. Immunogenet 2024 765. 2024;76:397–415.

49. Brocker C, Thompson D, Matsumoto A, Nebert DW, Vasiliou V. Evolutionary divergence and functions of the human interleukin (IL) gene family. Hum Genomics. 2010;5:30.

50. Nei M, Gu X, Sitnikova T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A. 1997;94:7799–806.

51. Nei M, Rooney AP. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 2005;39 Volume 39, 2005:121–52.

52. Garceau V, Smith J, Paton IR, Davey M, Fares MA, Sester DP, et al. Pivotal Advance: Avian colony-stimulating factor 1 (CSF-1), interleukin-34 (IL-34), and CSF-1 receptor genes and gene products. J Leukoc Biol. 2010;87:753–64.

53. Baghdadi M, Umeyama Y, Hama N, Kobayashi T, Han N, Wada H, et al. Interleukin-34, a comprehensive review. J Leukoc Biol. 2018;104:931–51.

54. Droin N, Solary E. Editorial: CSF1R, CSF-1, and IL-34, a “menage a trois” conserved across vertebrates. J Leukoc Biol. 2010;87:745–7.

55. Ovchinnikov DA, DeBats CEE, Sester DP, Sweet MJ, Hume DA. A conserved distal segment of the mouse CSF-1 receptor promoter is required for maximal expression of a reporter gene in macrophages and osteoclasts of transgenic mice. J Leukoc Biol. 2010;87:815–22.

56. Hume DA, Gutowska-Ding MW, Garcia-Morales C, Kebede A, Bamidele O, Trujillo AV, et al. Functional evolution of the colony-stimulating factor 1 receptor (CSF1R) and its ligands in birds. J Leukoc Biol. 2020;107:237–50.

57. Lin H, Lee E, Hestir K, Leo C, Huang M, Bosch E, et al. Discovery of a Cytokine and Its Receptor by Functional Screening of the Extracellular Proteome. Science (80-). 2008;320:807–11.

58. Wang Y, Colonna M. Interkeukin-34, a cytokine crucial for the differentiation and maintenance of tissue resident macrophages and Langerhans cells. Eur J Immunol. 2014;44:1575–81.

59. Greter M, Lelios I, Pelczar P, Hoeffel G, Price J, Leboeuf M, et al. Stroma-Derived Interleukin-34 Controls the Development and Maintenance of Langerhans Cells and the

Maintenance of Microglia. Immunity. 2012;37:1050–60.

60. Yamane F, Nishikawa Y, Matsui K, Asakura M, Iwasaki E, Watanabe K, et al. CSF-1 receptor-mediated differentiation of a new type of monocytic cell with B cell-stimulating activity: its selective dependence on IL-34. J Leukoc Biol. 2014;95:19–31.

61. Stanley ER, Chitu V. CSF-1 receptor signaling in myeloid cells. Cold Spring Harb Perspect Biol. 2014;6.

62. Neves F, Marques JP, Areal H, Pinto-Pinho P, Colaço B, Melo-Ferreira J, et al. TLR7 and TLR8 evolution in lagomorphs: different patterns in the different lineages. Immunogenetics. 2022;74:475–85.

63. Baird DM. Telomeres and genomic evolution. Philos Trans R Soc B Biol Sci. 2018;373.

64. Ko BJ, Lee C, Kim J, Rhie A, Yoo DA, Howe K, et al. Widespread false gene gains caused by duplication errors in genome assemblies. Genome Biol. 2022;23:1–26.

65. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet 2018 196. 2018;19:329–46.

66. Li M, Sun C, Xu N, Bian P, Tian X, Wang X, et al. De Novo Assembly of 20 Chicken Genomes Reveals the Undetectable Phenomenon for Thousands of Core Genes on Microchromosomes and Subtelomeric Regions. Mol Biol Evol. 2022;39.

67. Seroussi E, Cinnamon Y, Yosefi S, Genin O, Smith JG, Rafati N, et al. Identification of the Long-Sought Leptin in Chicken and Duck: Expression Pattern of the Highly GC-Rich Avian leptin Fits an Autocrine/Paracrine Rather Than Endocrine Function. Endocrinology. 2016;157:737–51.

68. Neves F, Abrantes J, Steinke JW, Esteves PJ. Maximum-likelihood approaches reveal signatures of positive selection in IL genes in mammals. Innate Immun. 2014;20:184–91.

69. Neves F, Abrantes J, Almeida T, De Matos AL, Costa PP, Esteves PJ. Genetic characterization of interleukins (IL-1α, IL-1β, IL-2, IL-4, IL-8, IL-10, IL-12A, IL-12B, IL-15 and IL-18) with relevant biological roles in lagomorphs. Innate Immun. 2015;21:787–801.

70. Sironi M, Cagliani R, Forni D, Clerici M. Evolutionary insights into host-pathogen interactions from mammalian sequence data. Nat Rev Genet. 2015;16:224–36.

71. Tenthorey JL, Emerman M, Malik HS. Evolutionary Landscapes of Host-Virus Arms Races. Annu Rev Immunol. 2022;40 Volume 40, 2022:271–94.

72. Wang X, Grus WE, Zhang J. Gene Losses during Human Origins. PLOS Biol. 2006;4:e52.

73. Helsen J, Voordeckers K, Vanderwaeren L, Santermans T, Tsontaki M, Verstrepen KJ, et al. Gene Loss Predictably Drives Evolutionary Adaptation. Mol Biol Evol. 2020;37:2989–3002.

74. Brandis G, Hughes D. The SNAP hypothesis: Chromosomal rearrangements could emerge from positive Selection during Niche Adaptation. PLoS Genet. 2020;16.

75. Olson M V. When Less Is More: Gene Loss as an Engine of Evolutionary Change. Am J Hum Genet. 1999;64:18–23.

76. van der Loo W, Magalhaes MJ, de Matos AL, Abrantes J, Yamada F, Esteves PJ.

Adaptive Gene Loss? Tracing Back the Pseudogenization of the Rabbit CCL8 Chemokine. J Mol Evol. 2016;83:12–25.

77. Baghdadi M, Endo H, Tanaka Y, Wada H, Seino K ichiro. Interleukin 34, from pathogenesis to clinical applications. Cytokine. 2017;99:139–47.

78. Walker DG, Tang TM, Lue LF. Studies on colony stimulating factor receptor-1 and ligands colony stimulating factor-1 and interleukin-34 in Alzheimer's disease brains and human microglia. Front Aging Neurosci. 2017;9 AUG:271250.

79. Cheng ST, Tang H, Ren JH, Chen X, Huang AL, Chen J. Interleukin-34 inhibits hepatitis B virus replication in vitro and in vivo. PLoS One. 2017;12:e0179605.

80. Hossainey MRH, Hauser KA, Garvey CN, Kalia N, Garvey JM, Grayfer L. A perspective into the relationships between amphibian (Xenopus laevis) myeloid cell subsets. Philos Trans R Soc B Biol Sci. 2023;378.

81. Lopez Ruiz V, Robert J. The amphibian immune system. Philos Trans R Soc B. 2023;378.

82. Wang Y, Szretter KJ, Vermi W, Gilfillan S, Rossini C, Cella M, et al. IL-34 is a tissue-restricted ligand of CSF1R required for the development of Langerhans cells and microglia. Nat Immunol. 2012;13:753–60.

83. Grayfer L, Edholm ES, Robert J. Mechanisms of amphibian macrophage development: characterization of the Xenopus laevis colony-stimulating factor-1 receptor. Int J Dev Biol. 2014;58:757–66.

84. Nishikawa A, Murata E, Akita M, Kaneko K, Moriya O, Tomita M, et al. Roles of macrophages in programmed cell death and remodeling of tail and body muscle of Xenopus laevis during metamorphosis. Histochem Cell Biol. 1997;109:11–7.

85. Xu C, Hutchins ED, Tokuyama MA, Wilson-Rawls J, Kusumi K. Transcriptional analysis of scar-free wound healing during early stages of tail regeneration in the green anole lizard, Anolis carolinensis. J Immunol Regen Med. 2020;7:100025.

86. Kagohashi K, Sasaki Y, Ozawa K, Tsuchiya T, Kawahara S, Saitoh K, et al. Role of Signal-Transducing Adaptor Protein-1 for T Cell Activation and Pathogenesis of Autoimmune Demyelination and Airway Inflammation. J Immunol. 2024;212:951–61.

87. Banks CAS, Boanca G, Lee ZT, Eubanks CG, Hattem GL, Peak A, et al. TNIP2 is a Hub Protein in the NF-κB Network with Both Protein and RNA Mediated Interactions. Mol Cell Proteomics. 2016;15:3435–49.

88. Van Huffel S, Delaei F, Heyninck K, De Valck D, Beyaert R. Identification of a Novel A20-binding Inhibitor of Nuclear Factor-κB Activation Termed ABIN-2. J Biol Chem. 2001;276:30216–23.

89. Zhao L, Wang S, Xu M, He Y, Zhang X, Xiong Y, et al. Vpr counteracts the restriction of LAPTM5 to promote HIV-1 infection in macrophages. Nat Commun 2021 121. 2021;12:1–14.

90. O'Leary NA, Cox E, Holmes JB, Anderson WR, Falk R, Hem V, et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. Sci Data 2024 111. 2024;11:1–10.

91. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic Acids Res. 2023;51:D933–41.

92. Dutheil JY, Gaillard S, Stukenbrock EH. MafFilter: A highly flexible and extensible multiple genome alignment files processor. BMC Genomics. 2014;15:1–10.

93. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

94. Mangiola S, Papenfuss AT. tidyHeatmap: an R package for modular heatmap production based on tidy principles. J Open Source Softw. 2020;5:2472.

95. Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, et al. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. Elife. 2022;11.

96. Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, et al. Non-B DB v2.0: A database of predicted non-B DNA-forming motifs and its associated tools. Nucleic Acids Res. 2013;41.

97. Rangwala SH, Rudnev D V., Ananiev V V., Oh DH, Asztalos A, Benica B, et al. The NCBI Comparative Genome Viewer (CGV) is an interactive visualization tool for the analysis of whole-genome eukaryotic alignments. PLOS Biol. 2024;22:e3002405.

98. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

99. Robinson JT, Thorvaldsdottir H, Turner D, Mesirov JP. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). Bioinformatics. 2023;39.

100. Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, et al. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl Plant Sci. 2016;4.

101. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37:1530–4.

102. Paradis E, Blomberg S, Bolker B, Brown J, Claramunt S, Claude J, et al. Analyses of Phylogenetics and Evolution [R package ape version 5.8-1]. CRAN Contrib Packag. 2024. https://doi.org/10.32614/CRAN.PACKAGE.APE.

103. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

104. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun. 2020;11.

105. Madrigal G, Minhas BF, Catchen J. Klumpy: A tool to evaluate the integrity of long-read genome assemblies and illusive sequence motifs. Mol Ecol Resour. 2024;:e13982.

106. Beauclair L, Ramé C, Arensburger P, Piégu B, Guillou F, Dupont J, et al. Sequence properties of certain GC rich avian genes, their origins and absence from genome assemblies: Case studies. BMC Genomics. 2019;20:734.

107. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open

Software Suite. Trends Genet. 2000;16:276–7.

108. Kolde R. pheatmap: Pretty Heatmaps. CRAN Contrib Packag. 2010. https://doi.org/10.32614/CRAN.PACKAGE.PHEATMAP.

109. Ives AR, Garland T. Phylogenetic Logistic Regression for Binary Dependent Variables. Syst Biol. 2010;59:9–26.

110. Sela I, Ashkenazy H, Katoh K, acids research TP-N, undefined 2015. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. academic.oup.com. 2015;43:7–14.

111. Löytynoja A. Phylogeny-aware alignment with PRANK. Methods Mol Biol. 2014;1079:155–70.

112. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.

113. Kosakovsky Pond SL, Frost SDW, Muse S V. HyPhy: Hypothesis testing using phylogenies. Bioinformatics. 2005.

114. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24:1586–91.

115. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol. 2015;32:1342–53.

116. Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, et al. Gene-wide identification of episodic selection. Mol Biol Evol. 2015;32:1365–71.

117. Kosakovsky Pond SL, Frost SDW. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005;22:1208–22.

118. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLOS Genet. 2012;8:e1002764.

119. Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. Mol Biol Evol. 2015;32:820–32.

120. Poon AFY, Lewis FI, Kosakovsky Pond SL, Frost SDW. An Evolutionary-Network Model Reveals Stratified Interactions in the V3 Loop of the HIV-1 Envelope. PLOS Comput Biol. 2007;3:e231.

121. Yang Z, Nielsent R. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. Mol Biol Evol. 2002;19:908–17.

122. Wickham H. ggplot2. 2016. https://doi.org/10.1007/978-3-319-24277-4.

123. Su Y, Cao Y, Liu C, Xu Q, Li N, Lan M, et al. Inactivating IL34 promotes regenerating muscle stem cell expansion and attenuates Duchenne muscular dystrophy in mouse models. Theranostics. 2023;13:2588.

124. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 2016 345. 2016;34:525–7.

125. Ge SX, Jung D, Jung D, Yao R. ShinyGO: A graphical gene-set enrichment tool for

animals and plants. Bioinformatics. 2020;36:2628–9.

Acknowledgements

We thank the Council of Scientific & Industrial Research (CSIR) and the University Grant Commission (UGC) for a fellowship to BGS and SH, respectively. We used BioRender (https://biorender.com) to arrange figures and images of species.

Funding

This article was funded by the Department of Biotechnology, Ministry of Science and Technology, India (grant no. BT/11/IYBA/2018/03) and Science and Engineering Research Board (grant no. ECR/2017/001430) provided funds for procuring computational resources (i.e. Har Gobind Khorana Computational Biology cluster) used.

Author information

Authors and Affiliations

Buddhabhushan Girish Salve (buddhabhushan21@iiserb.ac.in), Sonal H (sonal24@iiserb.ac.in), Nagarjun Vijay (nagarjun@iiserb.ac.in)

Computational Evolutionary Genomics Lab, Department of Biological Sciences, IISER Bhopal, Bhauri, Madhya Pradesh, India

Contributions

BGS and NV. conceived and designed the study; BGS, SH, and NV performed research and analysed data; and BGS and NV wrote the manuscript. All authors have read and approved the final version of the manuscript and agree with the order of presentation of the authors.

Corresponding author

Correspondence to Nagarjun Vijay (nagarjun@iiserb.ac.in)

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Figure legends

Fig. 1

Evidence for the absence of missing genes in squamates using whole-genome alignments and BLASTn analysis. The heatmap analysis of whole-genome alignments between chicken (*Gallus gallus*) and 27 other vertebrate species to assess the presence/absence of 53 missing genes in squamates and other species. The values used to plot the heatmap are normalised (by gene length) 1 Kb genomic windows from chicken that failed to align with each species genome alignment, highlighting lineage-specific alignment gaps (Additional file 1: Table S2). Species are grouped by taxonomic class and order (left side of the heatmap), while gene names are shown at the bottom. A yellow-to-purple gradient represents the degree of non-alignment, with purple indicating higher proportions of unaligned regions. The violin plots (at the top) show BLASTn query coverage for each of the 53 missing genes searched into squamate genomes, further supporting their absence. The red dashed line marks query coverage of 70 %. Lower coverage values indicate reduced or no detectable sequence in squamates.

Fig. 2

Gene losses localise to discrete chromosomal sites, are in syntenic blocks, and are prone to chromosomal rearrangement. The left panel shows the genomic locations of 53 missing genes mapped across chicken chromosomes (black horizontal bars scaled by genomic position on the x-axis). Missing genes are indicated by light green dots labelled with gene names. GC content, calculated in 100 bp windows, is represented by a gradient from light grey (low GC) to red (high GC). Below the GC track, the presence of predicted non-B DNA motifs is indicated by colour-coded bars. The right panel depicts chromosomal rearrangements across representative sauropsid species (*Gallus gallus*, *Alligator mississippiensis*, *Chelonia mydas*, *Pogona vitticeps*, *Anolis sagrei*, and *Thamnophis elegans).* The chromosome numbers are mentioned above each grey-coloured rectangle, and coloured ribbons (as defined in the left panel) represent conserved synteny blocks connecting orthologous regions, with seven regions of interest (ROI_1 to ROI_7) highlighted. These regions are implicated in the loss of two or more genes**.** The chicken orthologues of the human genes *IFTAP*, *PLAAT1*, and *TMEM273* correspond to *C11orf74*, *HRASLS*, and *C10orf128*, respectively.

Fig. 3

Orthologues of missing genes have moderate GC content and are evolutionarily conserved. The violin plots depict the distribution of four metrics across all available orthologues for each of the 53 genes missing in the squamate lineage but retained in outgroups. The x-axis represents individual gene names, and the y-axis shows values for each metric. Red dashed lines in each panel indicate threshold values used to assess potential gene loss indicators. Median values are shown in red text, and sample sizes are indicated in black. The top panel shows the maximum query coverage obtained for the focal gene in BLASTn search in 265 squamates genomes. The red dashed line indicated a threshold of 70 % query coverage. The second panel shows the sequence identity (in %) between the human orthologue gene sequence and the sequence of other vertebrate species. Genes below the 70% identity line are more likely to be lost or highly diverged. The red dashed line indicated a threshold of 70 % sequence identity. The third panel represents the distribution of the GC content of each gene across its orthologues. A horizontal red dashed line is drawn at 55%, marking the GC-content level. The fourth panel indicate local GC-richness across genes. The threshold line marks GC stretches of 3.936.

Fig. 4

Missing genes often belong to multi-gene families of paralogues. Human genes corresponding to squamate-missing genes are plotted based on their GC-content (x-axis) and paralogues (y-axis) as annotated in Ensembl. Each point represents a paralogue, with the size of the light-coloured filled circle indicating the percentage sequence identity between the missing gene and its paralogue. Dashed lines indicate an approximate GC content of 55% for both axes (vertical red for the missing gene and horizontal blue for the paralogues). This representation highlights that genes lost in squamates often belong to multi-gene families with moderate GC content.

Fig. 5

Evidence for *IL34* gene loss via segmental deletion. **A.** Micro-gene synteny of *IL34* in 24 vertebrate representative species. The phylogenetic branches of species with *IL34* lost are depicted with dashed red, while blue phylogenetic branches represent the 3rd round of whole genome duplication. The green-coloured thunderbolt marks the gene loss event in squamates. Arrows represent genes, with their direction indicating gene orientation and gene names labelled inside each arrow. **B.** The heatmap shows the reduction in gene distance at the *IL34* locus (between *SF3B3-MTSS2*) in squamates. The values inside the box represent adjusted p-values (significant differences between squamates and non-squamates, calculated using pairwise Wilcox one-tailed test (alternative = “less”). **C.** Phylogenetic logistic regression analyses. The relative genomic distance between flanking genes (x-axis, in %) and the status of the *IL34* gene (y-axis) are correlated using two methods (logistic_IG10 and logistic_MPLE) to evaluate the segmental deletion of the *IL34* gene across vertebrates (n□=□179). The species tree was obtained from the Timetree website (https://timetree.org/) and annotated in iTOL (https://itol.embl.de/). Species images are from https://BioRender.com.

Fig. 6

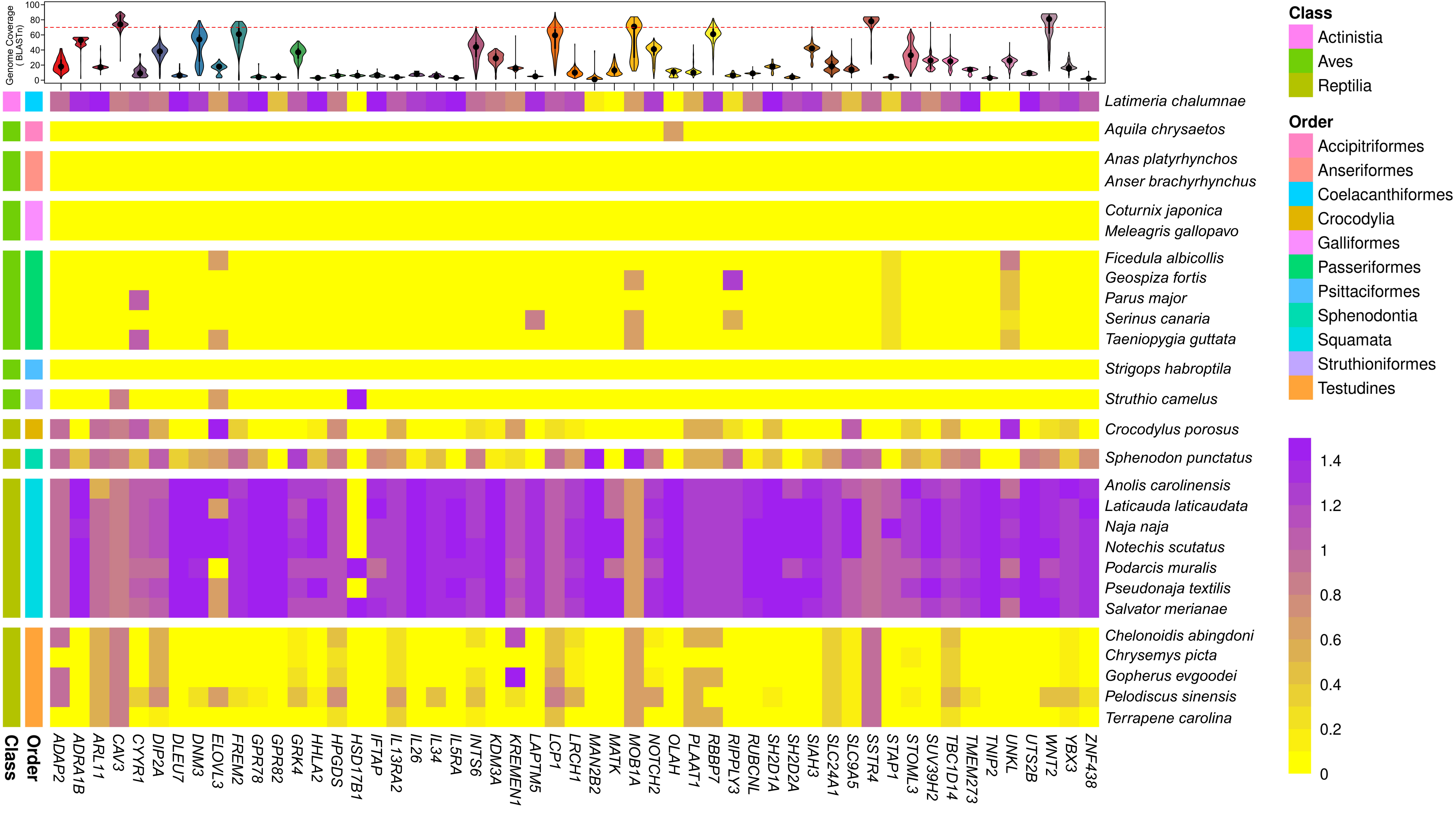
Validation of genome assembly at *IL34* gene syntenic locus in common wall lizard and Indian cobra genomes using long-read sequencing. UCSC Genome Browser snapshots of the *IL34* syntenic region are shown for the common wall lizard (*Podarcis muralis*; panel A, chr8:39,120,000–39,125,000) and the Indian cobra (*Naja naja*; panel B, chr4:770,000–780,000). The top portion of each panel displays GC content (black line), annotated syntenic genes (*MTSS2*/*SF3B3* in *P. muralis*, E2320_000953/958 in *N. naja*), and RepeatMasker annotations (including SINEs, LINEs, and low-complexity elements). The bottom portion of both panels shows long-reads from PacBio and Nanopore spanning the entire region. Grey and yellow segments indicate forward- and reverse-strand alignments, respectively. Transparent red and blue boxes mark exons of flanking syntenic genes. SRA (Short Read Archive) accession numbers for the datasets used to generate the BAM files are indicated next to the plots.

Fig. 7

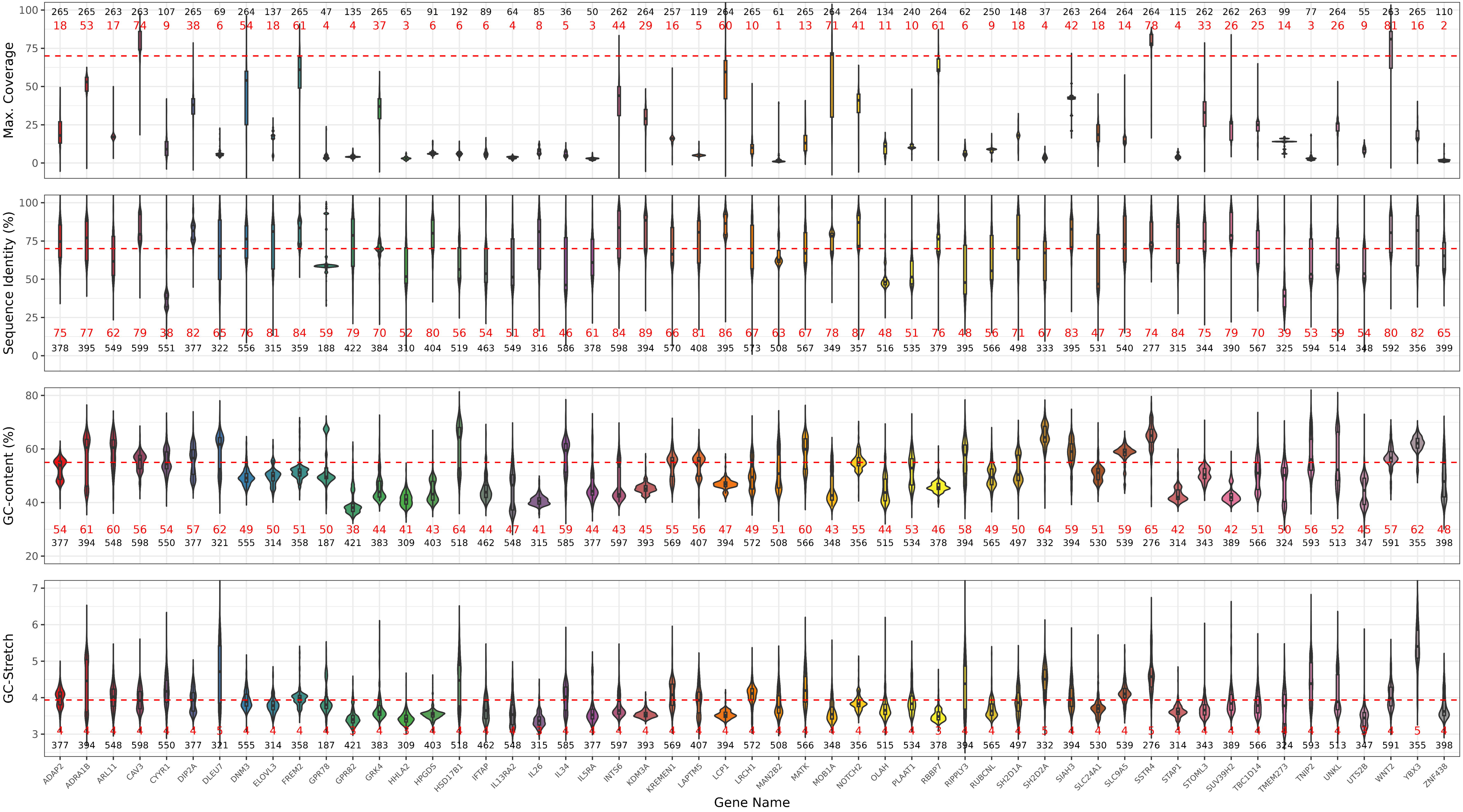
Indirect evidence for *IL34* loss: Co-evolution of receptor *CSF1R*. Selection analysis of the *CSF1R* coding sequence was conducted across 32 vertebrate species using both branch-based (RELAX, aBSREL, and CodeML) and site (FEL and MEME) models. The phylogenetic tree (left) illustrates the evolutionary relationships among species, with squamate lineages highlighted in red text and red box. Selection inferences such as relaxed/intensified and positive selection, based on RELAX, CodeML, and aBSREL analyses, are shown for each species. The central panel displays the distribution of positively and negatively selected codon sites along the gene, with codon positions on the x-axis and selection signals from focal species (foreground) against the rest of the species in the background. The right panel summarises the number of positively selected sites detected by FEL and MEME for each species, visualised using ggplot2 [122].

Fig. 8.

Squamate-specific loss of four genes involved in macrophage activation and polarisation. **A.** Heatmap showing the gene status across 16 squamates and five outgroup vertebrate species. Each row represents a gene along with gene ID, and each column represents a species. The green colour indicates an intact gene, the pink indicates a missing gene (lacking annotations), and the red represents low-quality protein annotation by NCBI. The genes involved in macrophage activation and polarisation ─*IL34*, *LAPTM5*, *STAP1*, and *TNIP2*—are consistently missing in squamate while intact in outgroup species. **B.** Gene synteny of the *LAPTM5*, *STAP1*, and *TNIP2* genes across representative vertebrate lineages reveals conserved gene order. Gene names are shown at the top, with arrows indicating gene orientation. A pink cross marks genes absent in the respective species. Vertical black lines between genes denote placement on different scaffolds. In post-chromosomal rearrangement (post-CR) species such as green anole and Indian cobra, *TNIP2* is lost and neighbouring syntenic genes are relocated. In contrast, pre-CR species like chicken and painted turtle retain the ancestral syntenic arrangement. These patterns suggest that *TNIP2* loss may be associated with lineage-specific chromosomal rearrangements. **C.** GC content across vertebrate species. The red dots denote the mean GC content and values mentioned in blue text, and the white dots indicate the median. The y-axis represents GC content in percentage. **D.** Gene Ontology (GO) enrichment analysis for biological processes/pathways for *IL34*, *LAPTM5*, *STAP1*, and *TNIP2* genes. We used ShinyGO v0.82 for gene-set enrichment analysis [125].


Genome Coverage (BLASTn)
100
80
60
40
20
0
Latimeria chalumnae
Aquila chrysaetos
Anas platyrhynchos
Anser brachyrhynchus
Coturnix japonica
Meleagris gallopavo
Ficedula albicollis
Geospiza fortis
Parus major
Serinus canaria
Taeniopygia guttata
Strigops habroptila
Struthio camelus
Crocodylus porosus
Sphenodon punctatus
Anolis carolinensis
Laticauda laticaudata
Naja naja
Notechis scutatus
Podarcis muralis
Pseudonaja textilis
Salvator merianae
Chelonoidis abingdoni
Chrysemys picta
Gopherus evgoodei
Pelodiscus sinensis
Terrapene carolina
Class
Order
ADAP2
ADRA1B
ARL11
CAV3
CYYR1
DIP2A
DLEU7
DNM3
ELOVL3
FREM2
GPR78
GPR82
GRK4
HHLA2
HPGDS
HSD17B1
IFTAP
IL13RA2
IL26
IL34
IL5RA
INTS6
KDM3A
KREMEN1
LAPTM5
LCP1
LRCH1
MAN2B2
MATK
MOB1A
NOTCH2
OLAH
PLAAT1
RBBP7
RIPPLY3
RUBCNL
SH2D1A
SH2D2A
SIAH3
SLC24A1
SLC9A5
SSTR4
STAP1
STOML3
SUV39H2
TBC1D14
TMEM273
TNIP2
UNKL
UTS2B
WNT2
YBX3
ZNF438
Class
Actinistia
Aves
Reptilia
Order
Accipitriformes
Anseriformes
Coelacanthiformes
Crocodylia
Galliformes
Passeriformes
Psittaciformes
Sphenodontia
Squamata
Struthioniformes
Testudines
1.4
1.2
1
0.8
0.6
0.4
0.2
0

chr1
SUV39H2
WNT2
IL26
YBX3
HHLA2
CYYR1
RIPPLY3
GPR82
RBBP7
RUBCNL
LRCH1
DLEU7
ARL11
INTS6
SIAH3
FREM2
LCP1
STOML3
chr2
ZNF438
OLAH
chr3
SSTR4
chr4
IL13RA2
SH2D1A
HPGDS
STAP1
GPR78
TNIP2
TBC1D14
GRK4
KDM3A
MAN2B2
chr5
C11orf74
chr6
ELOVL3
C10orf128
chr7
DIP2A
chr8
NOTCH2
DNM3
chr9
HRASLS
UTS2B
chr10
SLC24A1
chr11
SLC9A5
IL34
chr12
IL5RA
CAV3
chr13
ADRA1B
chr14
UNKL
chr15
KREMEN1
chr18
ADAP2
chr22
MOB1A
chr23
LAPTM5
chr25
SH2D2A
chr27
HSD17B1
chr28
MATK
Chromosome
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
Genomic Position (bp)
90 Mbp
Gallus gallus
1
4
6
8
9
11
12
Alligator mississippiensis
1
2
6
5
7
10
12
Chelonia mydas
1
4
7
8
9
12
Pogona vitticeps
5
1
3
4
10
2
Anolis sagrei
5
1
3
4
8
2
Thamnophis elegans
3
6
11
9
15
10
14
2
Chromosomes scaled by physical position
GC-content
0.75
0.50
0.25
0.00
Type of Non-B site
A_Phased_Repeat
Direct_Repeat
G_Quadruplex_Motif
Inverted_Repeat
Mirror_Repeat
Short_Tandem_Repeat
Z_DNA_Motif
Regions of interest
REGION_1
REGION_2
REGION_3
REGION_4
REGION_5
REGION_6
REGION_7


Max. Coverage
100
75
50
25
0
265 265 263 263 107 265 69 264 137 265 47 135 265 65 91 192 89 64 85 36 50 262 264 257 119 264 265 61 265 264 264 134 240 264 62 250 148 37 263 264 264 264 115 262 262 263 99 77 264 55 263 265 110
18 53 17 74 9 38 6 54 18 61 4 4 37 3 6 6 6 4 8 5 3 44 29 16 5 60 10 1 13 71 41 11 10 61 6 9 18 4 42 18 14 78 4 33 26 25 14 3 26 9 81 16 2
Sequence Identity (%)
100
75
50
25
0
75 77 62 79 38 82 65 76 81 84 59 79 70 52 80 56 54 51 81 46 61 84 89 66 81 86 67 63 67 78 87 48 51 76 48 56 71 67 83 47 73 74 84 75 79 70 39 53 59 54 80 82 65
378 395 549 599 551 377 322 556 315 359 188 422 384 310 404 519 463 549 316 586 378 598 394 570 408 395 573 508 567 349 357 516 535 379 395 566 498 333 395 531 540 277 315 344 390 567 325 594 514 348 592 356 399
GC-content (%)
80
60
40
20
54 61 60 56 54 57 62 49 50 51 50 38 44 41 43 64 44 47 41 59 44 43 45 55 56 47 49 51 60 43 55 44 53 46 58 49 50 64 59 51 59 65 42 50 42 51 50 56 52 45 57 62 48
377 394 548 598 550 377 321 555 314 358 187 421 383 309 403 518 462 548 315 585 377 597 393 569 407 394 572 508 566 348 356 515 534 378 394 565 497 332 394 530 539 276 314 343 389 566 324 593 513 347 591 355 398
GC-Stretch
7
6
5
4
3
4 4 4 4 4 4 5 4 4 4 4 3 4 3 4 4 4 4 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 4 4 4 5 4 4 4 5 4 4 4 4 4 4 4 3 4 5 4
377 394 548 598 550 377 321 555 314 358 187 421 383 309 403 518 462 548 315 585 377 597 393 569 407 394 572 508 566 348 356 515 534 378 394 565 497 332 394 530 539 276 314 343 389 566 324 593 513 347 591 355 398
ADAP2
ADRA1B
ARL11
CAV3
CYYR1
DIP2A
DLEU7
DNM3
ELOVL3
FREM2
GPR78
GPR82
GRK4
HHLA2
HPGDS
HSD17B1
IFTAP
IL13RA2
IL26
IL34
IL5RA
INTS6
KDM3A
KREMEN1
LAPTM5
LCP1
LRCH1
MAN2B2
MATK
MOB1A
NOTCH2
OLAH
PLAAT1
RBBP7
RIPPLY3
RUBCNL
SH2D1A
SH2D2A
SIAH3
SLC24A1
SLC9A5
SSTR4
STAP1
STOML3
SUV39H2
TBC1D14
TMEM273
TNIP2
UNKL
UTS2B
WNT2
YBX3
ZNF438
Gene Name

GC-content = 55.30 %
GC-content = 55.56 %
Human paralogue transcript ID (ranked by GC content)
Gene stable ID (ranked by GC content)
0
50
100
150
200
0
10
20
30
40
NOTCH2
KREMEN1
ADAP2
SLC9A5
MATK
SSTR4
MAN2B2
DNM3
DNM3
STAP1
SUV39H2
HHLA2
ADRA1B
GRK4
GPR78
HSD17B1
UNKL
SH2D2A
DIP2A
SIAH3
WNT2
RBBP7
KDM3A
HPGDS
IL13RA2
SH2D1A
CAV3
YBX3
IL5RA
MOB1A
SLC24A1
ELOVL3
STOML3
FREM2
LCP1
INTS6
GPR82
ARL11
TBC1D14
LRCH1
IL5RA
GPR82
Gene.name
ADAP2
ADRA1B
ARL11
CAV3
DIP2A
DNM3
ELOVL3
FREM2
GPR78
GPR82
GRK4
HHLA2
HPGDS
HSD17B1
IL13RA2
IL5RA
INTS6
KDM3A
KREMEN1
LCP1
LRCH1
MAN2B2
MATK
MOB1A
NOTCH2
RBBP7
SH2D1A
SH2D2A
SIAH3
SLC24A1
SLC9A5
SSTR4
STAP1
STOML3
SUV39H2
TBC1D14
UNKL
WNT2
YBX3
% Sequence identity (Human gene vs Paralog)
0–25%
25–50%
50–75%
75–100%

	COG4-SF3B3	COG4-MTSS2	COG4-VAC14	**SF3B3-MTSS2**	SF3B3-VAC14	MTSS2-VAC14
COG4-SF3B3	NA	0.02	1.00	0.00	1.00	1.00
COG4-MTSS2	1.00	NA	1.00	0.00	1.00	1.00
COG4-VAC14	0.15	0.00	NA	0.00	0.92	0.33
SF3B3-MTSS2	1.00	1.00	1.00	NA	1.00	1.00
SF3B3-VAC14	0.04	0.00	1.00	0.00	NA	0.23
MTSS2-VAC14	0.34	0.01	1.00	0.00	1.00	NA

A
Scale
5 kb
GCF_004329235.1
chr8:
39,120,000
39,125,000
All gaps of unknown nucleotides (N's), including AGP annotated gaps
Assembly
SIRZ01000008.1
GC Percent in 5-Base Windows
GC Percent
70
30
CBI RefSeq genes, curated and predicted sets (NM_*, XM_*, NR_*, XR_*, NP_* or YP
MTSS2
MTSS2
MTSS2
MTSS2
MTSS2
MTSS2
SF3B3
NCBI RefSeq other annotations (not NM_*, NR_*, XM_*, XR_*, NP_* or YP_*)
CpG Islands on All Sequence (Islands < 300 Bases are Light Green)
CpG: 22
RepeatMasker Repetitive Elements
TGGAG)n#Simple_repeat
Squam2
(CTGT)n
CR1-3_Croc
Simple Tandem Repeats by TRF
TTTTTACATGAGTAGA
Genomic Intervals Masked by WindowMasker + SDust
WM + SDust
Common wall lizard (Podarcis muralis)
PacBio
SRR8468524
NC_041319.1: 39,114,790-39,128,268
17
18
19
20
21
22
23
24
25
26
27
39.12M
39.12M
Nanopore
SRR8468506
NC_041319.1: 39,114,790-39,128,268
6
7
8
9
10
11
12
13
-39,128,268
39.12M
39.12M
B
Scale
5 kb
GCA_009733165.1
chrMIC_4:
770,000
775,000
780,000
All gaps of unknown nucleotides (N's), including AGP annotated gaps
Assembly
SOZL01000530.1
GC Percent in 5-Base Windows
GC Percent
70
30
Your Sequence from Blat Search
Blat Sequence
Gene models submitted to NCBI
E2320_000953
E2320_000958
E2320_000953
CpG Islands on All Sequence (Islands < 300 Bases are Light Green)
CpG: 51
Repeating Elements by RepeatMasker
SINE
DNA
Simple
Simple Tandem Repeats by TRF
GAAGG
Genomic Intervals Masked by WindowMasker + SDust
WM + SDust
Indian Cobra (Naja naja)
PacBio
SRR10418082
SRR10418083
SRR10418084
SRR10418085
SRR10418086
SRR10418087
SRR10418088
SRR10418089
CM019161.1: 768,622-781,041
13
14
15
16
17
18
19
773.62K
E2320_000958
778.62K
Nanopore
SRR10418058
SRR10418059
SRR10418060
SRR10418061
SRR10418062
SRR10418065
SRR10418066
SRR10418067
SRR10418068
SRR10418069
SRR10418070
SRR10418071
SRR10418072
SRR10418073
SRR10418074
CM019161.1: 768,622-781,041
6
7
8
9
10
11
12
773.62K
E2320_000958
778.62K

Danio rerio
Oryzias melastigma
Poecilia formosa
Poecilia reticulata
Latimeria chalumnae
Xenopus tropicalis
Rana temporaria
Hyla sarda
Bufo bufo
Ornithorhynchus anatinus
Phascolarctos cinereus
Loxodonta africana
Homo sapiens
Mus musculus
Rhinolophus ferrumequinum
Orcinus orca
Equus caballus
Canis lupus
Sphenodon punctatus
Rhineura floridana
Thamnophis elegans
Varanus komodoensis
Pogona vitticeps
Anolis carolinensis
Dermochelys coriacea
Caretta caretta
Alligator mississippiensis
Gallus gallus
Cygnus olor
Calypte anna
Taeniopygia guttata
Rissa tridactyla
24 6
8 5
14 17
18 16
43 20
53 21
41 22
44 36
42 30
41 16
26 17
40 29
22 14
33 19
40 21
22 16
19 14
22 9
61 37
52 36
62 39
61 39
40 26
36 21
30 30
36 31
31 21
37 25
29 27
24 14
43 24
12 7
0
250
500
750
1000
Codon
-60
-40
-20
0
20
40
No. of Sites
Relaxed selection (RELAX)
Relaxed selection (codeml)
Positive selection (aBSREL)
Intensified selection (RELAX)
Intensified selection (codeml)
non- significant
Positive sites (FEL)
Negative sites (FEL)
Positive sites (MEME)
No. of Positive sites (FEL)
No. of Positive sites (MEME)

A
Pantherophis_guttatus
Pituophis_catenifer_annectens
Thamnophis_elegans
Thamnophis_sirtalis
Ahaetulla_prasina
Crotalus_tigris
Protobothrops_mucrosquamatus
Notechis_scutatus
Pseudonaja_textilis
Erythrolamprus_reginae
Python_bivittatus
Candoia_aspera
Anolis_sagrei
Anolis_carolinensis
Sceloporus_undulatus
Pogona_vitticeps
Varanus_komodoensis
Elgaria_multicarinata_webbii
Podarcis_muralis
Podarcis_raffonei
Zootoca_vivipara
Lacerta_agilis
Rhineura_floridana
Tiliqua_scincoides
Hemicordylus_capensis
Sphaerodactylus_townsendi
Euleptes_europaea
Heteronotia_binoei
Gekko_japonicus
Eublepharis_macularius
Alligator_mississippiensis
Gallus_gallus
Chelonia_mydas
Homo_sapiens
Xenopus_tropicalis
id
IRAK3-11213
TLR8-51311
PGLYRP4-57115
MMP8-4317
IL18-3606
BPI-671
ACKR2-1238
IL34-146433
STAP1-26228
TNIP2-79155
LAPTM5-7805
SLAMF1-6504
SSC5D-284297
CD84-8832
SFTPD-6441
DMBT1-1755
TLR7-51284
IFNG-3458
SOAT1-6646
CKLF-51192
SLAMF8-56833
Missing
Intact
LQP
B
PUM1
SDC3
LAPTM5
MATN1
PTPRU
Human
Green sea turtle
Tropical clawed frog
American alligator
Tuatara
Green anole
Eastern brown snake
Chicken
ATP5IF1
ZNF593
EPHA5
CENPC
STAP1
UBA6
GNRHR
TMPRSS11C
COQ2
TMPRSS11D
ADD1
SH3BP2
TNIP2
FAM193A
RNF4
PPP2R2C
WFS1
Post-CR
Pre-CR
Indian cobra
Green anole
Chr 4
Chr 3
Chr 5
Chr 4
Chicken
Chr 4
Painted turtle
Chr 2
C
IL34
LAPTM5
STAP1
TNIP2
GC content (in %)
40
50
60
70
80
48.39
63.07
52.9
60.35
57.66
42.98
51.85
44.58
56.13
48.45
41.96
46.32
41.14
41.94
46.43
50.53
50.28
49.83
65.05
49.83
Amphibia
Aves
Crocodylia
Mammalia
Testudines
D
Positive reg. of macrophage activation
Reg. of macrophage chemotaxis
Macrophage chemotaxis
Reg. of macrophage migration
Positive reg. of protein tyrosine kinase activity
Macrophage migration
Reg. of antigen receptor-mediated signaling pathway
Cytokine-mediated signaling pathway
Immune response-regulating signaling pathway
Cellular response to cytokine stimulus
Response to cytokine
Positive reg. of immune system proc.
Positive reg. of signal transduction
Response to biotic stimulus
Positive reg. of cell communication
Positive reg. of signaling
Defense response
Immune response
0
100
200
300
400
Fold Enrichment
N. of Genes
2.0
2.5
3.0
3.5
4.0
-log10(FDR)
2.75
3.00
3.25
3.50