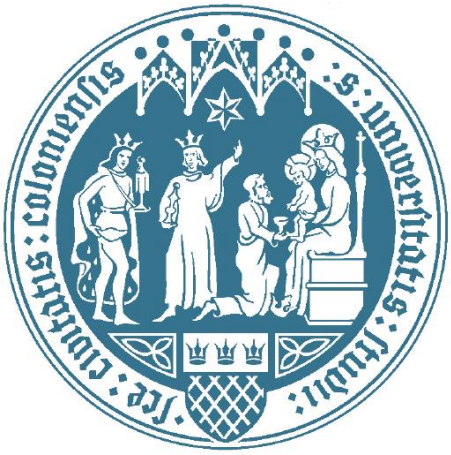
Origin and fate of duplicated genes

Three case studies from fish, birds and beetles

INAUGURAL-DISSERTATION

Zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln



vorgelegt von
Katja Maria Palitzsch
aus Quedlinburg
angenommen im Jahr 2025

Abstract

In the framework of evolution, genomes are highly dynamic systems, shaped by a large variety of mechanisms of gene gain, diversification and loss driving adaptation to environmental conditions of living organisms. This thesis provides a comparative analysis of the origin, fate, and evolutionary significance of duplicated genes, illustrated through three case studies in fish, birds, and beetles. The introductory chapters summarize the theory on molecular evolution, focusing on mechanisms such as *de-novo* gene emergence, gene duplication, and multigene family dynamics. Population-scale genomic analyses reveal extraordinary copy number variation and population-specific expansion within large gene families. In zebrafish, NLR immune genes show both massive copy number diversification and striking sequence homogenization, reflecting ongoing forces of duplication and concerted evolution. Studies of beetle odorant receptor genes uncover a highly dynamic and population-structured gene repertoire, displaying patterns of birth-and-death evolution and tandem gene duplication events. A detailed examination of the BORIS (CTCFL) gene across bird species demonstrates recurrent and lineage-specific gene loss within neognathous birds, associated with relaxed selective constraints and the accumulation of repetitive elements. Collectively, these studies demonstrate how gene duplication, diversification, and loss interact as intertwined processes that generate genomic novelty and functional diversity, forming the foundation of organismal adaptation. The thesis highlights the importance of population-level and comparative approaches for understanding the evolutionary mechanisms underlying genome complexity and adaptation across taxa.

Zusammenfassung

Genome sind hoch dynamische Systeme, die im Rahmen evolutionärer Prozesse von einer Reihe an Entstehungs-, Diversifizierungs- und Verlustmechanismen geformt werden. Diese Dynamik ist ein Antrieb für Anpassungsprozesse an Umweltbedingungen. Die vorliegende Arbeit liefert eine vergleichende Analyse zu Ursprung, Schicksal und evolutionärer Bedeutung duplizierter Gene, illustriert durch drei Beispiele in Fischen, Vögeln und Käfern. Die einleitenden Kapitel geben einen Überblick über Theorien molekularer Evolution, mit Schwerpunkt auf Modellen zur Genenstehung und zum Genverlust. Es werden *de-novo* Gen-Entstehung und Gen-Duplikation sowie spontaner Genverlust und Pseudogensierung erläutert. Dabei liegt der Fokus auf evolutionären Dynamiken von Multigen-Familien. Die genomischen Analysen auf Populationseben zeigen ausserordentliche Variabilität innerhalb von Genfamilien,- sowohl auf Populationsebene als auch auf Ebene von Individuen. Diese Variabilität kann neben der Anzahl von Genkopien auch die Zusammensetzung des Gen-Repertoires betreffen. So zeigt die Familie der NLR-Immungene in Zebrafischen eine hohe Diversität in der Anzahl der Kopien bei gleichzeitig grosser Sequenzhomogenität, die auf eine hohe Zahl rezenter Duplikations-Erreigneisse in Kombination mit konzertierter Evolution hindeutet. Am Beispiel von Odorant-Rezeptorgenen in Käfern wird demonstriert, dass hier auf Populationsebene ein hoch strukturiertes Gen-Repertoire vorliegt, welches Anzeichen von Evolution durch fortlaufende Gen-Entstehung und Gen-Verluste aufweist. Anhand des Verfalles des BORIS-Gens in Vögeln wird schlussendlich exemplarisch deutlich, wie, angetrieben durch reduzierten Selektionsdruck und die Ansammlung repetetiver Elemente, wiederkehrende, Linien-spezifische Mutationsereignisse in Neukiefervögeln zum fortschreitenden Verlust des Genes führen. Zusammengefasst geben die drei Untersuchungen einen Einblick in den Ablauf möglicher adaptiver Prozesse durch das dynamische Ineinandergreifen von Entstehungs-, Diversifizierungs- und Verlustereignissen.

Contents

1 Introduction

1.1 Evolution from a Molecular Genetic Perspective - from Morphology to Molecular Genetics

Before the introduction of the first gene sequence–based quantitative assessment of evolutionary relationships among major groups of organisms (**??**) (reviewed in **?**), early evolutionary biologists, including Charles Darwin (fig. 1), inferred phylogenetic relationships solely from the morphological traits of living organisms.

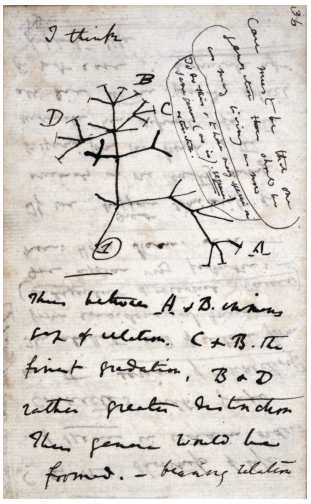


Figure 1: Page 36 from the *Notebooks on Transmutation of Species* by Charles Darwin, showing a sketch of a phylogenetic tree with the notes: *I think*
Case must be that one generation then should be as many living as now. (To do this & to have many species in same genus (as is) requires extinction.)
Thus between A & B. immense gap of relation, C & B. the finest gradation, B & D rather greater distinction Thus genera would be formed.— bearing relation (Charles Darwin 1837 , Public domain, via Wikimedia Commons, transcript by `https://darwin-online.org.uk`)

.

To date sequence data do not entirely replace morphological information as a resource for investigating phylogenetic relationships (**??**). However, the rapid increase in DNA, RNA, and protein sequence data since the 1970s has enabled the reconstruction of not only relationships among groups and species but also the evolutionary history of individual genes and gene families among and within species with high statistical confidence. In this dissertation, the evolutionary history and population structure of gene families are

analyzed, with three representative case studies serving as examples: transcription factors, an immune gene family, and the family of odorant receptor genes.

1.2 Structure of the Thesis

This thesis is organized to provide a comprehensive understanding of the origin and evolutionary fate of duplicated genes, guided by both theoretical concepts and empirical data from selected case studies in fish, birds, and beetles. The introductory chapters establish the foundational concepts in genome evolution.

First I review *de-novo* gene emergence and gene/genome duplication as the fundamental mechanisms contributing to genomic innovation. These chapters outline how new genetic material arises, the various mutational and selective forces involved, and the broader implications of these mechanisms for genome complexity and adaptation. Subsequently, the work explores the evolutionary dynamics of multigene families, comparing models such as concerted evolution and the birth-and-death process, and explaining their respective roles in generating gene family diversity.

The heart of the thesis consists of three detailed case studies, each illustrating a distinct trajectory of gene family evolution. The impacts of gene duplication, diversification, and loss in natural populations are demonstrated based on zebrafish NLR immune genes, beetle odorant receptors, and the decay of the BORIS gene in birds. These studies provide insights into the interplay between gene family turnover and organismal adaptation, with particular attention to population-level diversity and lineage-specific evolutionary events.

Finally, the thesis concludes with a synthesis and outlook section, summarizing key findings, discussing their implications for evolutionary biology, and suggesting directions for future research on genome and gene family evolution.

2 Fundamental Mechanisms of Genome Evolution

The central difference between living biological systems and abiotic chemical reaction units is their ability to reproduce autonomously. Therefore proteins and nucleic acids are mutually dependent. Proteins are not able to reproduce independently, nor are nucleic acids able to replicate without the catalytic interaction with proteins. Ribonucleic acids, which have both, catalytic (e.g. rRNA) and coding properties (e.g. mRNA), can be regarded

as intermediate stages and mediators in the process of early genome evolution. Therefore it is assumed that the first self-replicating proto-genomes consisted of ribonucleotides. Compared to ribonucleic acids, desoxibonucleic acids are characterized by higher stability. Additionally, due to their ability to form complementary double strands, they are accessible for damage repair mechanisms that theoretically can be carried out without loss of information (summarized in **?**).

Nevertheless, repair mechanisms are erroneous to a certain extent, which can lead to mutations. The impact of mutations on genome evolution has been discussed since the 1960s. A first attempt to resolve the role of mutation was made by **?**, who presented a method for estimating the divergence times between species with their hypothesis of the molecular clock. This is based on the assumption that the mutation rate of DNA and protein sequences is constant over time in all species and that the genetic differences are therefore proportional to the divergence time of any two species.

This is complemented by the neutral theory of adaptive evolution, which was presented by **?**, who argued that the rate of molecular evolution, measured as nucleotide substitutions, is much higher than could be sustained by positive selection alone. He proposes that the majority of these substitutions are selectively neutral, having no significant effect on fitness. Consequently, random genetic drift, rather than natural selection, becomes the dominant force driving the fixation of most mutations at the molecular level. This paradigm shift, explains the observed high levels of genetic variation and the rapid accumulation of molecular changes in populations.

The neutral theory was later extended by **?**, who proposed that not only strictly neutral mutations but also slightly deleterious ones can be fixed through random genetic drift. Ohta suggested that such mutations are more likely to become fixed in small populations or during bottlenecks, thereby accelerating the rate of molecular evolution under these conditions.

Those classical theoretical models of speciation and population genetics were built on concepts dating back to Darwins recognition of heritable diversity and classical Mendelian genetics. Limited methods for measuring genetic variability up to the 1970s restricted empirical validation (summarized in **?**).

By now, continuously advancing sequencing technologies generate ongoing expanding datasets from an increasing number of species and populations, greatly enhance our understanding of genome-wide evolutionary patterns within and between species.

Population genomic studies in model organisms like *Drosophila* have demonstrated that genetic diversity is shaped by the combined effects of mutation, recombination, natural selection, demography, and linkage. Empirical findings highlight a strong correlation

between recombination rate and genetic variability (summarized in **?**), widespread signatures of both purifying and adaptive selection, pervasive impacts of linked selection, and distinct evolutionary patterns on sex chromosomes, such as those predicted by the faster-X hypothesis (**????**). Further theoretical and empirical work emphasizes Hill–Robertson interference (**?**), which describes how selection at linked loci reduces the efficacy of adaptation and overall genetic variation. Characteristic genomic signatures of selection, including selective sweeps and background selection, have motivated the development of statistical tests such as Tajima's D (**?**) and the McDonald–Kreitman test (**?**) (summarized in **?** and **?**).

While much of classical and modern evolutionary genetics has focused on how mutation, recombination, selection, and drift shape existing genetic variation, long-term evolutionary innovation relies equally on the origin of entirely new genetic material. Beyond the modulation of pre-existing diversity, genomes are continuously remodeled through processes that introduce novel sequences and functions. Two of the most important mechanisms driving this innovation are the emergence of *de-novo* genes from non-coding regions and the duplication of existing genes, followed by their functional divergence. Together, these processes provide the raw material upon which selection can act, thereby expanding the functional repertoire of organisms and fueling adaptive evolution.

2.1 *De-Novo* Gene Emergence

De-novo genes, often referred to as orphan genes in earlier literature (as discussed in **???**), were first identified on a large scale in yeast (**?**). Since these initial observations, accumulating evidence from an increasing number of taxa has highlighted *de-novo* gene emergence as a widespread and significant mechanism contributing to the origin of novel genetic material across species (**?**). *De-novo* genes are defined as genes that arise from ancestrally non-genic DNA. They are often species-specific, polymorphic, and short-lived within populations (**?**).

The origin of *de-novo* genes has been explained by two models. The first, known as the "expression first" model, suggests that non-coding regions of the genome are transcribed as long non-coding RNAs. Through the acquisition of mutations, these transcripts can begin to code for short peptides, forming proto-genes that may develop into fully functional genes. The "ORF first" model proposes that open reading frames (ORFs) are already present within non-coding regions and only require the acquisition of regulatory elements to enable their transcription and translation. Once these regulatory signals are in place, these latent ORFs can be expressed and give rise to functional *de-novo* genes (**??**).

The detection of *de-novo* genes is impeded by limited sequence homology, an issue further compounded by assembly errors and sequencing gaps. It can be improved through synteny-based approaches, long read sequencing approaches and optimized assembly strategies (**??**). Young *de-novo* genes are typically short, consisting of a low number of exons and short ORFs, show enrichment of repetitive elements, preferentially use optimal codons, and display low overall but high tissue-specific expression levels (**???????**). Cases of testis biased expression and of enrichment on the X chromosome were observed in *Drosophila* (**??**), with both, adaptive pressures and neutral processes shaping their trajectories. *De-novo* gene expression can arise through adoption of nearby regulatory elements or 3D genome interactions. They can contribute to diverse functions including reproduction, stress response, development and metabolism (**???**). The protein products of *de-novo* genes are frequently intrinsically disordered or very small (microproteins), although some acquire stable secondary or tertiary structures (summarized in **?**). Key open questions remain about what determines persistence, how they integrate into networks, and how evolutionary forces differ across tissues, species, and immune contexts (reviewed in **???**)

2.2 Gene Duplication

Alongside *de-novo* gene emergence, gene duplication constitutes another principal source of novel genes. In contrast to *de-novo* birth, duplicates originate from preexisting coding sequences and therefore more closely resemble established genes in structure and sequence. Gene duplicates often follow evolutionary trajectories similar to those of their ancestral genes (**?**).

Gene duplication is common across diverse organisms and serves as a key driver of genetic complexity and variation (**??**). Duplication rates are similar to mutation rates that for example lead to single-nucleotide polymorphisms (**?**).

Gene duplications originate through diverse mechanisms and are broadly classified into two types: small-scale duplications (SSD) and whole-genome duplications (WGD) (**?**). SSD entail the duplication of individual genes or chromosomal segments, usually via gene conversion, tandem duplication and/or transposition (**??**) (fig. 2A,B). In contrast, WGD involves duplication of the entire genome resulting in transient or persistent auto- or allopolyploidy (fig. 2C). It is now widely accepted that the metazoan evolution included at least three major WGD events as described by the 2R respective 3R hypothesis which

posits that the first round of WGD occured in an ancestor of jawless vertebrates after the divergence from invertebrates (**?**), the second before the origin of mammals (**?**) and a third, fish-specific WGD (3R) occurred in the ancestor of all teleost fish (**?**).

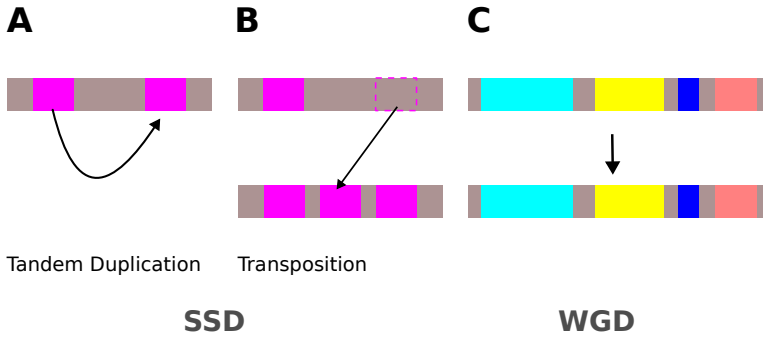


Figure 2: **Mechanisms of gene duplication:** Small-scale duplications (SSD), through gene conversion unequal recombination or mediated by transposable elements may occur as (A) tandem duplication, or (B) transposition. (C) Whole-genome duplication (WGD) might arise from meiosis errors and can result in transient or permanent polyploidy.

Because deleterious mutations are much more common than beneficial ones, the classical model predicts that most duplicate genes will most likely lose function and become pseudogenes or disappear immediately after duplication (**??**). However, many duplicate pairs avoid non-functionalization, following several evolutionary trajectories once they arise (**?**). Sometimes, both gene copies undergo hypo-functionalization, reducing their expression, which makes both copies essential to maintain dosage balance. In some cases, duplicate genes partition the ancestral functions through a process known as sub-functionalization, whereby each copy aquires responsibility for distinct aspects of the original gene's role, making both necessary for full functionality (**?**).

Beyond sub-functionalization, duplicated genes can also diverge through alternative evolutionary outcomes, including neo-functionalization and the accumulation of neutral variation. Neo-functionalization occurs when one duplicate develops a novel beneficial function, making the loss of either gene disadvantageous and favoring their long-term retention. An illustrative example of neofunctionalization is described for *Drosophila*, where the majority of young duplicate genes are retained through the acquisition of novel functions (**?**). These novel functions predominantly arise in the child copy, frequently exhibiting initial expression in the testes. This pattern highlights the combined effects of positive selection and testes-driven innovation as major forces driving the evolutionary preservation and functional diversification of duplicate genes.

Duplicated genes may also accumulate neutral variation, differences in expression or function that are not under strong selective pressure, similar to the variation observed among single-copy genes (summarized in **??**).

Experimental and genomic evidence shows that duplications can rapidly drive adaptation since gene duplication underlies a wide range of adaptive responses across taxa. For example, duplications of stress-related genes enhance survival and fitness under heat stress in *Escherichia coli* (**?**). In yeast and plants polyploidy, rather than single-gene duplication, facilitates adaptation to high salinity, conferring increased resilience to osmotic stress (**??**). In microbes, fungi and plants copy number variations of metal-resistance and transporter genes enable survival in environments with toxic concentrations of heavy metals and arsenites (**????**). Similarly, amplification of drug-resistance genes, such as pfmdr1, in the malaria parasite *Plasmodium falciparum*, promotes rapid adaptation to Chloroquin treatments (**?**). In Antarctic Cod (*Dissostichus mawsoni*), adaptive duplication of antifreeze glycoprotein and related genes is essential for cold tolerance, with such duplications occurring at much higher frequency than in closely related species (**?**).

Furthermore polyploidy, is increasingly recognized as being favored under stressful or changing environmental and biotic conditions, with stress responses playing a key role in the long-term establishment and success of polyploid organisms compared to nonpolyploids (reviewed in **?**).
These findings indicate that genes involved in environmental interactions, requiring high expression levels, or functioning within complex pathways are particularly susceptible to adaptive duplication. While some duplications may be deleterious, accumulating evidence for adaptive events challenges classical neutral models and underscores the need for ecological genomics research leveraging copy number variation to elucidate the role of duplications in adaptation (**?**).

Through successive cycles of duplication and divergence, gene copies can evolve distinct functions, develop new expression profiles or regulatory properties. An additional perspective on the evolutionary role of gene duplication is that an increased copy number might also be deleterious due to its impact on genome stability and cellular efficiency (**?**).

In summary gene duplication can be regarded as a central source of genetic novelty. The

described mechanisms of successive duplication and divergence cycles may ultimately give rise to multigene families that exemplify the evolutionary impact of gene duplication.
The following chapter will explore the structure, diversity, and evolutionary dynamics of gene families.

3 Multigene Families

Groups of genes that originated from a common ancestral gene through duplication events, display sequence homology and, in the broadest sense, share related functions are referred to as multigene families (**?**). They are found across both eukaryotic and prokaryotic genomes, and their prevalence tends to increase with genome size (**?**). These families can be broadly categorized into two types: those with uniform members, such as histone and ribosomal RNA genes, in which family members are highly similar to each other, and those with higher functional diversity, e.g. immunoglobulins, T-cell receptors, chemosensory receptors or major histocompatibility complex (MHC) genes, which display considerable genetic variation among their members and encompass a wide range of functional diversity (**?????**).
Multigene families provide a powerful framework for investigating evolutionary processes by enabling detailed analysis of their origin, expansion, and divergence in sequence, expression, and function, thus offering unique models to connect genetic changes with adaptive traits (**???**).

3.1 Mechanisms of Multigene Familiy Evolution

The evolutionary mechanisms of multigene families are discussed since the early 1960, when the "divergent evolution" model was proposed (**?**) (fig. 3 A). It proposes that gene families gradually diverge to acquire new functions. This model was challenged by studies of tandemly arranged gene families, which exhibited greater sequence homogeneity within species than between species, giving rise to the “concerted evolution" model, where gene family members evolve together via mechanisms like unequal crossing over and gene conversion (summarized in **?**) (fig. 3B). The concerted evolution model dominated for decades, but the emergence of extensive genomic data in the 1990s revealed unexpectedly high intraspecific diversity, contradicting strict homogenization. These observations, in

conjunction with the presence of pseudogenes and atypical phylogenetic patterns, gave rise to the formulation of the "birth-and-death" model of evolution, which posits that members of gene families are continually generated and eliminated, thereby contributing to genetic and functional diversification. (reviewed in **???**).

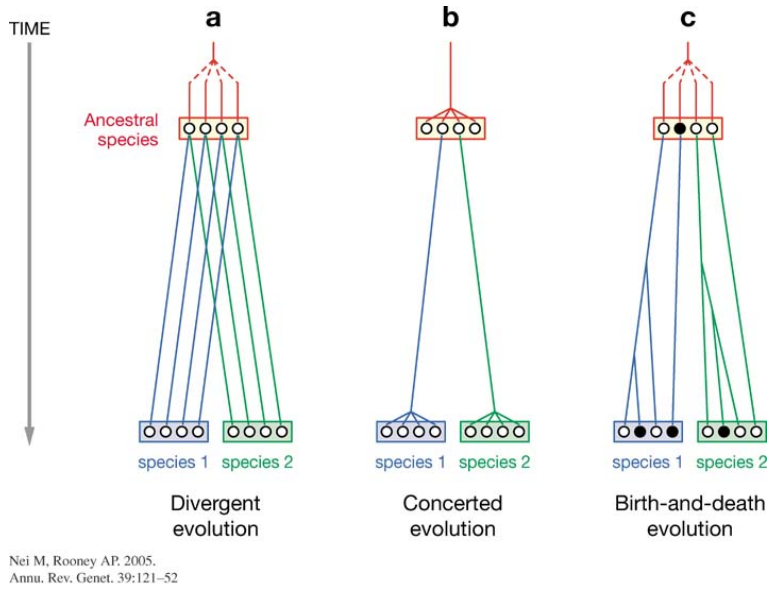


Figure 3: Schematic illustration of three principal models of multigene family evolution: (a) Divergent evolution, where gene copies in descendant species gradually accumulate differences; (b) Concerted evolution, characterized by sequence homogenization within species due to mechanisms like gene conversion and unequal crossing over, resulting in greater similarity among gene copies within a species than between species; and (c) Birth-and-death evolution, in which new gene copies arise while others are lost or become pseudogenes, leading to both sequence diversification and variable gene family size across species. Empty circles: functional genes, filled circles: pseudogenes, Adapted from **?**.

3.2 Concerted Evolution

In the concerted evolution model, all members of a gene family evolve together instead of independently. Mutations spread throughout the entire gene array via repeated unequal crossover or gene conversion events. This mechanism results in high sequence uniformity among gene family members within a species (**???**). Concerted evolution is governed by parameters like mutation rate, gene conversion rate, and population size, which together influence sequence similarity among gene copies. Gene conversion and unequal crossing over drive homogenization, while mutation introduces diversity. The equilibrium between these forces depends on rates of recombination and genetic drift. Purifying selection further reduces the mutational load in families with uniform members. The balance between mutation, homogenization, selection, and drift shapes diversity within these families (summarized in **?** and **?**). Empirical studies have shown that concerted evolution can be highly variable, sometimes affecting only coding regions or acting differently on certain domains (**???**) and often interacting with functional diversification (summarized in **?**). Concerted evolution is prevalent in eukaryotic organisms, where tandemly repeated genes like ribosomal RNA, U2 splicosomal small nuclear RNA (U2 snRNA), histones and ubiquitins undergo homogenization, resulting in high sequence similarity among repeats within a species (summarized in **?**). The U2 snRNA locus in primates exemplifies this, having maintained tandem organization and concerted evolution for over 35 million years, with large scale sequence deletions efficiently spreading to all copies (**?**).
Concerted evolution thus maintains high sequence uniformity within gene families, shaping their function and evolutionary potential across generations.

3.3 The Birth-and-Death-Model

Although concerted evolution was long considered the default model, accumulating genomic evidence showed that birth-and-death evolution, shaped by gene duplication, loss, mutation, and selection, is another mechanism driving the long-term dynamics of multigene families (summarized in **???**). This finding was initiated by the observation of unusual evolutionary patterns in some large families, such as the MHC genes (**?**). MHC (major histocompatibility complex) genes present peptides to T lymphocytes to induce immune responses, with class I genes divided into highly polymorphic (Ia) and less polymorphic (Ib). The exceptional diversity of Ia genes provides hosts with broad protection against rapidly evolving pathogens (**?**).

The coexistence of highly polymorphic Ia and less polymorphic Ib genes and the presence of distinct monophyletic clades in phylogenetic trees contradicted the hypothesis of pure concerted evolution that is characterized by continuous sequence homogenization. Although gene conversion, recombination, and exon exchange had been documented at MHC loci, their impact on polymorphism were considered to be limited. Therefore previous studies proposed that nucleotide substitution and positive selection are the primary drivers of MHC polymorphism, with gene conversion playing only a minor role (**?**). In contrast, **?** used mathematical simulations to show that the combined effects of selection and recombination can maintain both allelic and copy number diversity in tandem gene arrays, such as the MHC gene family.

The birth-and-death model represents a mechanism of multigene family evolution that is characterized by the two hallmarks interspecific gene clustering and the presence of pseudogenes. In this model, gene duplicates may persist and diverge, degenerate into pseudogenes, or be deleted right after duplication. (**?**)(reviewed in **?**).

The birth-and-death process is not isolated but operates in interaction with other evolutionary forces, which makes the process of expansion and contraction of gene families particularly complex (**????**). Thus, evolutionary patterns may be obscured not only by recent duplication, but also by gene loss resulting from strong purifying selection or rapid gene turnover, potentially mimicking concerted evolution or lateral gene transfer (reviewed in **?**).
Recently a duplication–selection model, following the principle of Haigh's mutation–selection framework of Muller's ratchet (**??**) was suggested, where mutation is replaced by gene duplication to study gene copy number variation in a population. According to this model gene copy number increases via duplication while individual fitness declines with higher copy numbers. It depicts two scenarios in finite populations where random genetic drift gradually eliminates individuals with fewer gene copies, resulting in rapid, self-accelerating copy accumulation under the Compound Copy Model (CCM), or slower, steady accumulation under the Single Template Model (STM) unless strong epistasis or recombination halts the process (**?**). The compound copy model (CCM) fits the observed data well for large gene families but essential mechanisms like unequal recombination and whole genome duplication are not incorporated. Therefore additional models, that besides unequal recombination, also integrate various selection schemes were introduced by **?**.

Overall, the described mechanisms of concerted and birth-and-death evolution can be regarded as complex, dynamically intertwined processes that, in combination with additional parameters, contribute to the shaping of gene families. Distinguishing which specific processes have led to the particular configuration of a gene family is very challenging based solely on sequence analyses, as discussed in **?**.

3.4 Pseudogenisation and Gene Loss

According to the birth-and-death model, gene loss, alongside gene duplication, is a crucial evolutionary force in multigene families. Gene loss can arise through spontaneous deletion events, such as unequal crossing-over or the activity of transposable or viral elements that physically remove the gene entirely. It can also result from gradual pseudogenisation, where an initial loss-of-function mutation (e.g., nonsense, frameshift, missense, splice-site, or regulatory changes) leads to the accumulation of additional disabling mutations until the gene becomes nonfunctional (reviewed in **?**)
Although patterns of gene loss vary widely among lineages, gene loss is a pervasive phenomenon, shaping genomes and contributing significantly to interspecies genetic and phenotypic differences of nearly all life forms (reviewed in **?**). Comparative genomic analyses have shown that, in some cases, gene loss happens at even higher rates than gene gain (**??**).
This widespread and sometimes accelerated loss of genes raises the question of why organisms can tolerate the absence of so many genetic elements without severe fitness consequences. Although early authors such as **??** considered pseudogenisation and gene loss primarily as consequences of dispensability, large-scale knockout studies have shown that many genes appear highly dispensable, a pattern that more likely reflects the mutational robustness of biological systems than a genuinely high proportion of dispensable genes. This robustness can arise either from genetic redundancy (paralogues or functionally convergent analogues) or from alternative pathways embedded within scale-free gene and protein interaction networks (reviewed in **?**), and thus does not necessarily indicate that the affected genes are redundant or even disadvantageous.

The likelihood of degradation and loss might also depend on the origin of a gene. As described in chapter 2.1, young *de-novo* genes are considerably more prone to loss-of-function (LoF) mutations than older genes, leading to a higher death rate compared with older genes or those originating from duplication events. Comparative analyses indicate

that specific sequence and expression features contribute to their evolutionary persistence. Higher GC content, longer gene length, and elevated expression levels are positively correlated with conservation, whereas increased microsatellite density predisposes genes to degeneration (**?**). Notably, *de-novo* genes with male-biased expression, particularly in the testes, exhibit enhanced resistance to premature inactivation. This differential stability accounts for the frequent enrichment of *de-novo* genes in testes and helps explain why male-biased transcription is a recurring hallmark of those *de-novo* genes that survive beyond their early stages (reviewed in **?**).

Rather than being eliminated entirely, dying genes often degrade gradually via loss-of-function mutations and persist as pseudogenes in the genome (**?**). They can therefore be regarded as genomic archives that allow the reconstruction of the evolutionary history of specific genes and gene families. Furthermore there is increasing evidence, that pseudogenes are not always strictly non-functional but, in some cases, fulfill regulatory roles or even acquire novel functions and can, in fact, contribute to adaptation (**?????**).
Loss-of-function (LoF) mutations arise from four main genetic changes. Nonsense single-nucleotide polymorphisms can introduce premature stop codons. Splice site mutations disrupt normal mRNA processing. Frameshift mutations alter the coding sequence, and loss of initiation codons prevents gene transcription or translation (summarized in **?**). LoF mutations can be advantageous in certain contexts. For instance, LoF in the SLC30A8 gene reduces type 2 diabetes risk in humans (**?**), while LoF in the mismatch repair gene MSH2 in the fungal pathogen *Cryptococcus deuterogattii* increases drug resistance (**?**).
Pseudogenes can encode functional proteins, as demonstrated by the human PGK2 retrocopy (**?**). Additionally, they can regulate their parental genes through various RNA mediated mechanisms, including antisense interactions, the production of small interfering RNAs (siRNAs), and functioning as microRNA sponges (**???**) (reviewed in **?**).

4 Evolutionary Innovation Through Gene Family Dynamics: Case Studies in Vertebrate and Insect Adaptation

The evolutionary fate of duplicated genes, and the diversity they generate, lies at the heart of molecular evolutionary theory and is a recurrent theme throughout this thesis. The introductory chapters laid the theoretical groundwork by explaining how evolutionary

innovation arises not only from rare *de-novo* gene emergence, but more fundamentally through frequent gene duplication and divergence, giving rise to multigene families whose structure and variability underpin organismal adaptation. Multigene family evolution is driven by an interplay between mechanisms such as concerted evolution and the birth-and-death process, shaping both genetic homogenization and diversity. Ribosomal RNA and U2 snRNA gene families, exemplify concerted evolution by their extraordinarily high sequence similarity within species. Immune gene families and chemoreceptor arrays are prominent examples of birth-and-death evolution generating both conserved core genes and rapidly diversifying members. The case studies presented here draw upon three gene families from fish, beetles, and birds to illustrate these principles in the three different, vertebrate and non-vertebrate biological systems.

4.1 Copy, Paste, and Stay the Same: Zebrafish NLRs as a Concert of Duplication and Homogenization

The attached paper "Copy number variation and population-specific immune genes in the model vertebrate zebrafish" reveals high copy number diversity in combination with high sequence homogeneity of nucleotide-binding domain leucine-rich repeat (NLR) immune genes in zebrafish.

To study zebrafish NLR genes we generated targeted single-molecule real-time (SMRT) sequencing to analyze almost 100 zebrafish from various wild and laboratory populations. This approach revealed a high diversity of unique NLR immune genes, greatly exceeding previous counts from reference genomes. Individual fish harbored from 100 to over 550 NLR copies, with wild populations containing up to four times as many as laboratory strains. Strikingly little nucleotide diversity was detected within these genes. Most copies were either monomorphic or showed very few polymorphisms. Most NLRs were rare and found in only single or small subsets of individuals. Only a tiny fraction was identified in all members of a population.
Gene duplication is highlighted in the chapters 2 and 3 as a primary driver of genetic novelty and expansion of multigene families. Our findings indicate that the zebrafish NLR gene family is most likely subject to ongoing large-scale gene duplications, resulting in hundreds of highly similar gene copies per individual, likely via tandem duplications and unequal crossing-over particularly on chromosome 4. This high copy number variation

can be observed especially in wild populations.

As discussed in chapter 3, concerted evolution is a process whereby repeated unequal crossing-over and gene conversion lead to accumulation of gene copies with exceptionally high sequence similarity. Concerted evolution as predominant mechanism is commonly observed in multigene families with uniform function. The zebrafish NLR family displays pronounced sequence similarity, suggesting, that concerted evolution might be a substantial evolutionary mechanism in this immune gene family.

This evolutionary trajectory in combination with the population specific differences in the wild populations reflects the interplay between duplication, homogenization, and selective pressures and exemplify principal models of multigene family evolution described in chapters 2 and 3.

4.2 Scents and Sensibility: How Beetle Odorant Receptors Evolve by Birth-and-Death and Genomic Jazz

Besides NLR immune gene clusters, chemosensory receptor families such as odorant receptors (ORs) provide excellent models for investigating the evolutionary fates of duplicated genes. Chemosensory receptors, in particular, exemplify dynamic expansion and contraction showing signatures of evolution predominantly driven by birth-and-death processes, as reviewed and summarized by **??** and **?**. In different *Drosophila* species **?** modeled the evolutionary dynamics of odorant receptor genes and identified substantial gene gains and losses, suggesting that birth-and-death evolution predominantly shapes the odorant receptor gene repertoires within *Drosophila* species.

In our study, "Striking Variability in the Odorant Receptor Repertoire of the Darkling Beetle Carchares macer within and between Populations", we conducted a population-scale analysis of odorant receptor (OR) genes in nearly 100 short-read genomes from wild samples of the tenebrionid beetle species *Carchares macer* from five distinct sampling sites. Insect ORs were previously assigned to seven subgroups (**?**). *C. macer* exhibits a lineage-specific expansion in one of the subgroups, accompanied by a complete absence of

representatives in three other groups. Additionally we observed pronounced presence/absence polymorphisms at both individual and population levels. The identification of around 100 annotated OR genes, with extrapolation suggests a species-wide repertoire approaching around 120 OR genes in this species, whereby the OR repertoire per individual consists of only 40 to 60 OR genes. Only a small subset of OR genes is universally present in all analyzed genomes, while the majority displays variable frequencies across individuals.
The observed indications of clustering of the *C. macer* OR genes and their association with tandem duplication further support the roles of structural rearrangements in shaping multigene family architecture as described for OR repertoires in other insect species (**?**).

Our results show that OR gene family variability in is not limited to differences between species, but can also be observed within species. Each population showcases a unique OR gene repertoire fingerprint, often featuring genes that are exclusive or markedly enriched in specific localities. Significant statistical variation in OR gene numbers between populations underscores the diversity of the *C. macer* OR gene repertoire. This findings suggest, that birth-and-death evolutionary processes, as described in chapter 3, might not only shape the OR repertoires across but also within species.

4.3 Gone, Split, or Falling Apart: Watching the Death of BORIS in Bird Genomes

Finally the attached manuscript "Decay of the CTCF paralog BORIS in neognathous birds" provides a comprehensive, data-rich example of evolution by gene death, a phenomenon introduced in the introductory section 3.3 of this thesis, as a key aspect of genome evolution alongside gene duplication and *de-novo* gene emergence.

The study investigates the evolutionary fate of the BORIS (Brother of Regulator of Impritned Sites) gene across 59 bird species. BORIS, also known as CTCFL, arose as duplicate of the essential chromatin organizer CCCTC-binding factor (CTCF) during early vertebrate evolution and is conserved across mammals and reptiles (**??**). Contrary to earlier claims that birds lack BORIS, the analysis reveals that BORIS is indeed present in the avian lineage but has frequently undergone severe, recurrent degradation, including

point mutations, loss of exons, and fragmentation, in neognathous birds, while remaining intact in paleognathous birds. These independent degradation events are specific to the BORIS locus, occur multiple times across the avian tree, and are correlated with an accumulation of lineage-specific repetitive elements and a relaxed of purifying selection in the affected lineages. The process is thus not a simple binary loss, but a spectrum of ongoing pseudogenisation and molecular decay. An additional observation is a correlation between BORIS gene loss and shifts in sperm and genital morphology. Although the link between reproductive traits and adaptive, or at least tolerated, gene loss in this context remains highly speculative, it may be worth further investigation given BORIS's reported role in sperm development in mammals.

As described in chapter 3.3 of the thesis, gene death, represented as comlete and sudden gene loss or gradual pseudogenisation, is a fundamental evolutionary process. The mechanisms underlying gene degradation such as elevated rates of mutation, relaxed selection, due to changes in selective constraints or redundancy after duplication, and the accumulation of repetitive elements that further degrade or destabilize loci are mirrored in the fate of the BORIS gene within neognathous bird species. In contrast to the NLR and OR genes, the evolutionary fate of BORIS was examined at the interspecies level rather than within populations, indicating that the evolutionary processes involved occur on substantially longer timescales than those observed for ORs and NLRs. Nevertheless the study illustrates an ongoing case of gene loss, exemplifying a gene that arose through duplication, diversified, and acquired a new function, yet degrades on one branch while persisting on other branches. The study exemplifies a case of gene loss that proceeds through a gradual process rather than an abrupt deletion, reflecting a complex interplay between mutation, selection, and genomic context.
The example of BORIS exemplifies that gene death can be a dynamic, recurrent driver of genome evolution, shaping functional repertoires not just by what is gained but also by what is discarded or rendered nonfunctional. Our findings substantiate the role of gene loss as a creative force in evolutionary biology.

5 Publications

5.1 Copy number variation and population-specific immune genes in the model vertebrate zebrafish

Authors: Yannick Schäfer, Katja Palitzsch, Maria Leptin, Andrew R Whiteley, Thomas Wiehe and Jaanus Suurväli

Status: Published in eLife, DOI: https://doi.org/10.7554/eLife.98058

Authors contribution: The significant contributions of the author of this thesis to the paper *Copy number variation and population-specific immune genes in the model vertebrate zebrafish* include the development and adaptation of the extraction procedure for high molecular weight (HMW) genomic DNA from zebrafish samples collected from wild populations. She was directly responsible for extracting DNA from all wild samples, ensuring the integrity and suitability of the genetic material for downstream analyses. Further, she performed target-capture and -enrichment and prepared DNA libraries for PacBio SMRT sequencing, facilitating the generation of high-quality sequencing data used for the study. These experimental efforts were complemented by contributions to the investigation and methodology. The author laid the foundation for further data analysis by developing and validating initial sequence processing and analysis pipelines, and was closely involved in development and validation of further technical approaches, further analysis and investigation within the project. Finally, the author contributed to the writing, the review and editing of the manuscript.

RESEARCH ARTICLE

Copy number variation and population-specific immune genes in the model vertebrate zebrafish

Yannick Schäfer[1], Katja Palitzsch[1], Maria Leptin[1], Andrew R Whiteley[2], Thomas Wiehe[1]*, Jaanus Suurväli[1,3]*

[1]Institute for Genetics, University of Cologne, Cologne, Germany; [2]WA Franke College of Forestry and Conservation, University of Montana, Missoula, United States; [3]Department of Biological Sciences, University of Manitoba, Winnipeg, Canada

Abstract Copy number variation in large gene families is well characterized for plant resistance genes, but similar studies are rare in animals. The zebrafish (*Danio rerio*) has hundreds of NLR immune genes, making this species ideal for studying this phenomenon. By sequencing 93 zebrafish from multiple wild and laboratory populations, we identified a total of 1513 NLRs, many more than the previously known 400. Approximately half of those are present in all wild populations, but only 4% were found in 80% or more of the individual fish. Wild fish have up to two times as many NLRs per individual and up to four times as many NLRs per population than laboratory strains. In contrast to the massive variability of gene copies, nucleotide diversity in zebrafish NLR genes is very low: around half of the copies are monomorphic and the remaining ones have very few polymorphisms, likely a signature of purifying selection.

***For correspondence:** twiehe@uni-koeln.de (TW); jaanus.suurvali@gmail.com (JS)

Competing interest: The authors declare that no competing interests exist.

Funding: See page 13

Preprinted: 24 August 2023
Received: 28 March 2024
Accepted: 03 June 2024
Published: 04 June 2024

Reviewing Editor: Detlef Weigel, Max Planck Institute for Biology Tübingen, Germany

Editor's evaluation

This useful study employs a sequence capture approach to characterize the diversity of NLR sequences in wild zebrafish populations. The authors provide solid evidence that wild zebrafish populations harbor several thousand NLR genes in total, with individual fish having a few hundred NLR gene copies.

Introduction

The innate immune system of an organism provides the first defense line against pathogens. Immune genes tend to evolve quickly and are often associated with a high degree of genetic variability. Many genes and proteins of the immune system are lineage-specific (limited to specific groups of animals, plants, or other taxa), while others have defense roles in a wide range of species. In particular, proteins containing a large nucleotide-binding domain followed by smaller repeats have an immune function in animals, plants, fungi, and bacteria alike (***Ting et al., 2008***; ***Jones and Dangl, 2006***; ***Uehling et al., 2017***; ***Gao et al., 2022***). In animals, these repeats are usually leucine-rich repeats (LRRs) and the proteins themselves are classified as NLRs (nucleotide binding domain leucine-rich repeat containing, also known as NOD-like receptors). They have a multitude of functions: some act as pathogen sensors or transcription factors (***Almeida-da-Silva et al., 2023***), others are components or modulators of inflammasomes, large protein complexes that are assembled within cells as part of the response to biological or chemical danger (***Almeida-da-Silva et al., 2023***).

eLife digest Humans and other animals have immune systems that protect them from bacteria, viruses and other potentially harmful microbes. Members of a family of genes known as the NLR family play various roles in helping to recognize and destroy these microbes. Different species have varying numbers of NLR genes, for example, humans have 22 NLRs, but fish can have hundreds. 400 have been found in the small tropical zebrafish, also known as zebra danios.

Zebrafish are commonly used as model animals in research studies because they reproduce quickly and are easy to keep in fish tanks. Much of what we know about fish biology comes from studying strains of those laboratory zebrafish, including the 400 NLRs found in a specific laboratory strain. Many NLRs in zebrafish are extremely similar, suggesting that they have only evolved fairly recently through gene duplication. It remains unclear why laboratory zebrafish have so many almost identical NLRs, or if wild zebrafish also have lots of these genes.

To find out more, Schäfer et al. sequenced the DNA of NLRs from almost 100 zebrafish from multiple wild and laboratory populations. The approach identified over 1,500 different NLR genes, most of which, were previously unknown. Computational modelling suggested that each wild population of zebrafish may harbour up to around 2,000 NLR genes, but laboratory strains had much fewer NLRs. The numbers of NLR genes in individual zebrafish varied greatly – only 4% of the genes were present in 80% or more of the fish. Many genes were only found in specific populations or single individuals.

Together, these findings suggest that the NLR family has expanded in zebrafish as part of an ongoing evolutionary process that benefits the immune system of the fish. Similar trends have also been observed in the NLR genes of plants, indicating there may be an evolutionary strategy across all living things to continuously diversify large families of genes. Additionally, this work highlights the lack of diversity in the genes of laboratory animals compared with those of their wild relatives, which may impact how results from laboratory studies are used to inform conservation efforts or are interpreted in the context of human health.

Plants have their own NLRs that are structurally similar to the ones from animals and also carry out central functions in the immune response (***Urbach and Ausubel, 2017***; ***Yue et al., 2012***). Their diversity has been extensively characterized in several species, including the thale cress (*Arabidopsis thaliana*), and vastly different repertoires have been found from different strains or individuals (***Van de Weyer et al., 2019b***). NLR repertoires can also be referred to as NLRomes, and a species-wide repertoire is called the 'pan-NLRome'.

Most knowledge about NLRs in animals comes from studies of humans and rodents, but their NLR repertoires (20–30 genes) are smaller than those of many other species such as the purple sea urchin, the sponge *Amphimedon queenslandica,* and many fish (***Hibino et al., 2006***; ***Yuen et al., 2014***; ***Suurväli et al., 2022***). However, even in mice one NLR (*Nlrp1*) has different copy numbers in different laboratory strains, ranging from 2 to 5 (***Lilue et al., 2018***). In many fishes, studies have reported NLR repertoires in the range of 10–50 genes (e.g., ***Rajendran et al., 2012***; ***Li et al., 2016***). In others, hundreds of NLRs are present, including in the model species zebrafish (*Danio rerio*) (***Stein et al., 2007***; ***Laing et al., 2008***; ***Tørresen et al., 2018***; ***Adrian-Kalchhauser et al., 2020***; ***Suurväli et al., 2022***). The zebrafish reference genome contains nearly 400 NLR genes, two-thirds of which are located on the putative sex chromosome (chromosome 4), in a genomic region associated with extensive haplotypic variation (***Howe et al., 2013***; ***Howe et al., 2016***; ***McConnell et al., 2023***; ***Anderson et al., 2012***).

The majority of fish NLRs represent a fish-specific subtype that was originally labeled NLR-C (***Laing et al., 2008***), although they can be further divided into at least six groups based on structural similarities and sequence of conserved exons (***Howe et al., 2016***; ***Adrian-Kalchhauser et al., 2020***). A schematic structure of proteins encoded by zebrafish NLR-C genes is presented in ***Figure 1A***. All of them possess a FISNA domain (fish-specific NACHT-associated domain), which precedes the nucleotide-binding domain NACHT and is encoded by the same large exon near the N-terminus of the protein (***Howe et al., 2016***). FISNA-NACHT is in some cases preceded by the effector domain PYD, but this is encoded by a separate exon (***Howe et al., 2016***). Additionally, many NLR-C proteins have a B30.2

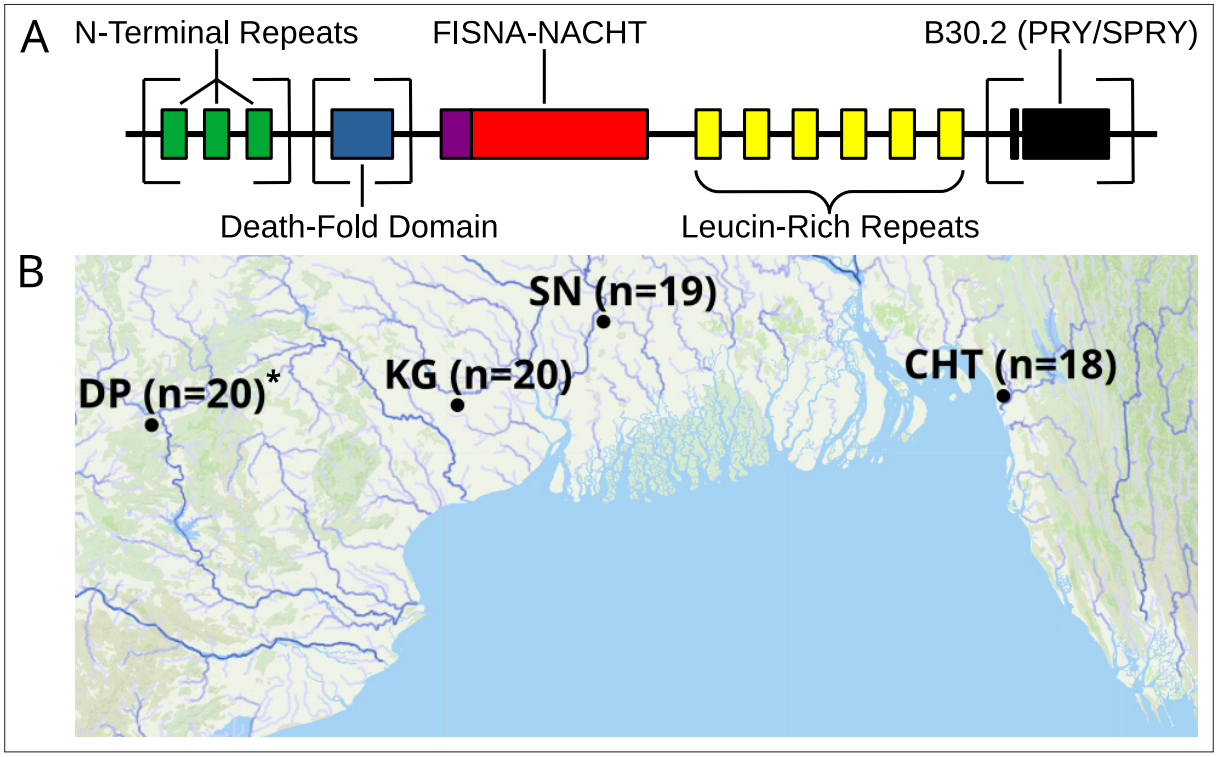


Figure 1. Structure of zebrafish NLRs and a map showing the origin of wild zebrafish samples. (**A**) Generalized, schematic representation of the domain architecture of an NLR-C protein. Each box represents a translated exon. The N-terminal repeats, the death-fold domain, as well as the B30.2 domain only occur in subsets of NLR-C genes. The number of N-terminal repeats and leucine-rich repeats can vary. Domains that can be either present or absent in different NLRs are surrounded by square brackets. (**B**) Sampling sites for wild zebrafish. All sites are located near the Bay of Bengal. Final sequenced sample sizes are indicated in parentheses. The map is based on geographic data collected and published by AQUASTAT from the Food and Agriculture Organization of the United Nations (***FAO, 2021***). The population DP is marked with an asterisk because its analysis and results are presented only in figure supplements.

domain (also known as PRY/SPRY) at the C-terminal end, separated from FISNA-NACHT by multiple introns and exons containing the LRRs (***Figure 1A***; ***Howe et al., 2016***). The B30.2 domain functions through protein–protein interaction (***Woo et al., 2006***) and is also found in a variety of other genes such as the large family of TRIM ubiquitin ligases (***van der Aa et al., 2009***; ***Howe et al., 2016***; ***Suurväli et al., 2022***) that are often also involved in immunity.

It is not known why fishes possess so many NLRs, how they evolve, and how much within-species genetic variability they have. The previously observed repeated expansions and contractions of this family suggest it to have a high rate of gene birth and death (***Suurväli et al., 2022***). Studies have shown that viral and bacterial infections can induce the expression of specific fish NLRs (reviewed in ***Chuphal et al., 2022***). Some of these have PYD or CARD domains and can even form inflammasomes similar to mammalian NLRs (***Kuri et al., 2017***; ***Li et al., 2018b***). A species-wide inventory of major NLR exons in a model species such as zebrafish would provide valuable insights into the evolution and diversity of this large immune gene family.

Results

By adapting and modifying a protocol that combines bait-based exon capture with PacBio SMRT technology (***Witek et al., 2016***), we successfully generated circular consensus sequence (CCS) data for targeted parts of the immune repertoire from 93 zebrafish (of initial 96), representing four wild populations (***Figure 1B***) and two laboratory strains. With this approach, we aimed to sequence all exons in zebrafish that encode the nucleotide-binding FISNA-NACHT domains and all exons that encode B30.2 domains. Samples of one wild population (DP) suffered from poor sequence coverage and had

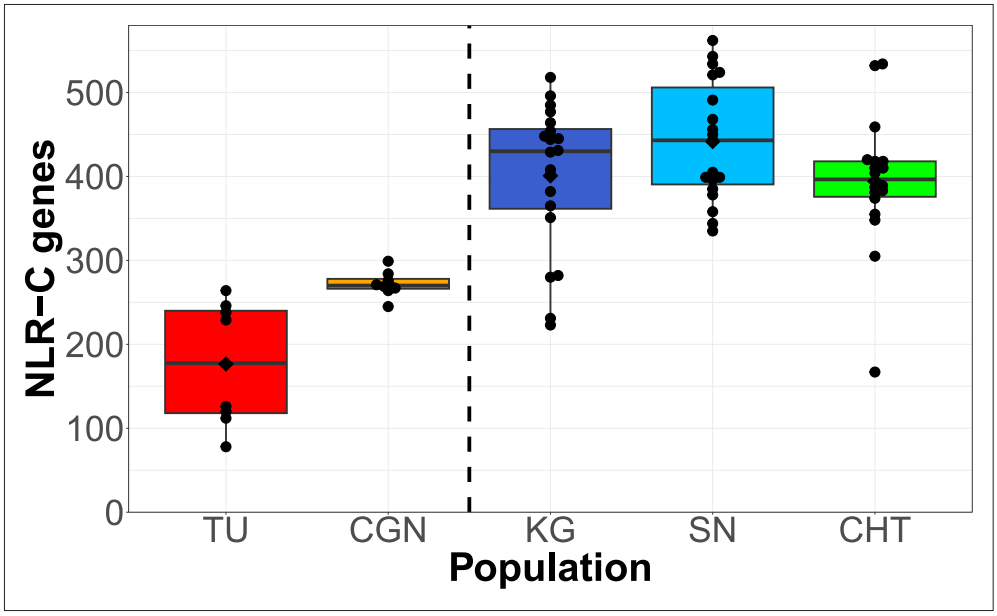


Figure 2. Total counts of NLRs found per individual, shown for each population. Black diamonds on the box plots denote means, horizontal lines denote medians. Left side: two laboratory strains; right side: three wild populations.

The online version of this article includes the following source data and figure supplement(s) for figure 2:

Source data 1. Source tables for ***Figure 2*** and its supplements.

Source data 2. Sequences and target locations of RNA baits.

Figure supplement 1. Sequencing and assembly statistics of circular consensus sequence (CCS) reads from NLR exons.

Figure supplement 2. Assembled NLRs in the reference genome GRCz11.

Figure supplement 3. Identification of B30.2 domains associated with zebrafish NLRs.

to be excluded from downstream analyses in order to avoid bias in interpretation. Results involving this population are only shown in figure supplements and not in the main figures.

Our protocol used PCR with primers targeting ligated adapters to amplify the below-nanogram amounts of genomic DNA obtained from exon capture. This limited our fragment sizes to the lengths of what the polymerase was able to amplify. Zebrafish NLRs can have their exons spread out across tens of kilobases, so that we cannot know which exons belong to the same gene. However, we were able to use captured sequence surrounding the targeted exons to distinguish among near-identical coding sequences and separate NLR-associated B30.2 domains from B30.2 elsewhere in the genome.

The zebrafish pan-NLRome

We used an orthology clustering approach on NLR sequences assembled from all populations to create a reference set of NLRs (a pan-NLRome). This resulted in the identification of 1513 unique FISNA-NACHT containing sequences and 567 for NLR-associated B30.2 (NLR-B30.2). Nearly 10% of the sequences (145 FISNA-NACHT and 64 NLR-B30.2) contained pre-mature stop codons that were at least 10 amino acids from the end and led to early truncation of the protein. In total, 101 of the 1513 FISNA-NACHT were preceded by an exon containing the N-terminal effector domain PYD. Nearly all of those (97 out of 101) were found in group 1 NLR-C genes identified by the presence of the characteristic sequence motif GIAGVGKT (***Howe et al., 2016***). Since the combination of FISNA and NACHT is only present in NLR-C, its count of 1513 can be considered equal to the total number of NLR-C genes in the data. We found each individual zebrafish to have 100–550 NLR genes from the pan-NLRome in at least one copy (***Figures 2 and 3***), and only 50–75% of these have a high-quality match in the GRCz11 reference genome (***Figure 2—figure supplement 2***). In general, laboratory zebrafish had less NLRs than wild samples (***Figure 2***). The number and length of CCS reads and assembled contigs (both prior to orthology clustering) are presented in ***Figure 2—figure supplement 1***.

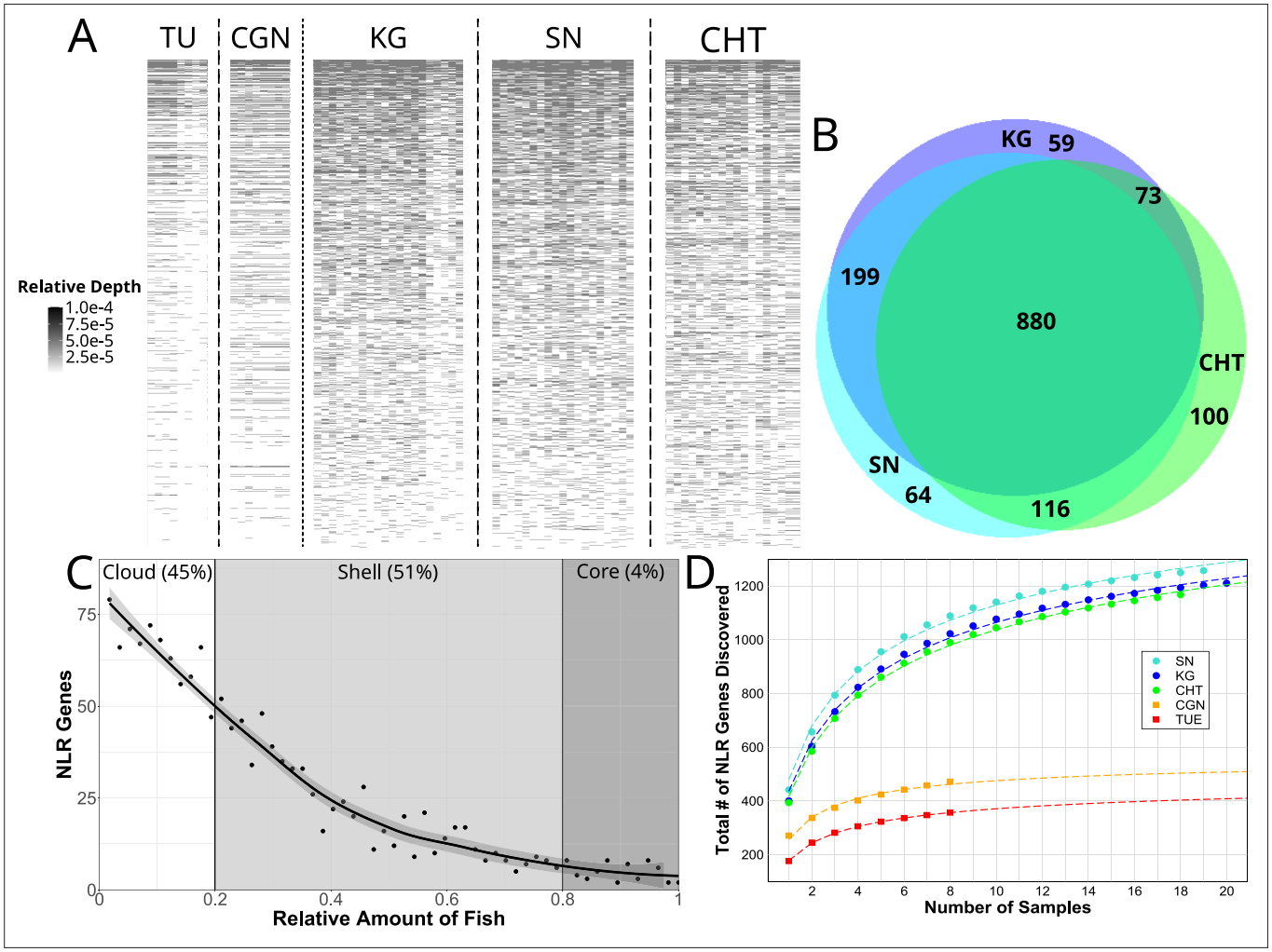


Figure 3. Copy number variation of NLR genes. (**A**) Sequence data from each individual zebrafish (vertical axis) was aligned to FISNA-NACHT exon sequences of the pan-NLRome (horizontal axis). Grayscale intensity shows, for each NLR, the proportion of NLR-aligning data in each given fish that matches this specific gene. Darker gray indicates a higher likelihood of this NLR being represented in multiple copies in the particular individual. Light gray indicates a single copy, white indicates absence. For clarity, only the 1235 FISNA-NACHT exons for which at least one fish had a minimum of 10 reads mapped to it are shown. (**B**) Numbers of pan-NLRome sequences (based on FISNACHT diagnosis) found in all three, two, or only one wild population. (**C**) Relative numbers of fish in which pan-NLRome sequences were found in wild populations. 'Core' pan-NLRome: genes which are found in at least 80% of the sample (from a total of 57 wild fish); 'shell': genes in at least 20%; 'cloud': rare genes found in less than 20% of the sample. (**D**) Observed and estimated sizes of population-specific pan-NLRomes. Data points (filled circles and squares) show the average number of totally discovered NLR genes (as identified via their FISNA-NACHT domain) when investigating x fish. The dashed line is obtained by non-linear fit of the data to the function given in ***Equation 2***. For all populations, the hypothetical pan-NLRome size – when extrapolating $x \to \infty$ – is finite (see ***Table 1***).

The online version of this article includes the following source data and figure supplement(s) for figure 3:

Source data 1. Source tables for ***Figure 3*** and its supplements.

Figure supplement 1. Comparison of copy number variation in FISNA-NACHT and NLR-B30.2 exons.

Figure supplement 2. Copy number variation of NLR genes, including the DP population.

Whereas FISNA-NACHT is only found in NLRs, B30.2 domains are also found in other gene families. In addition to the 567 NLR-B30.2 domains, we also found 732 B30.2 domains not associated with NLRs. We were able to distinguish between them by utilizing the sequence of a short highly conserved 47 bp exon that appears to precede B30.2 in NLRs, but not in other genes (***Figure 2—figure supplement 3***). Each individual zebrafish possesses 20–180 NLR-B30.2s n at least one copy (***Figure 3—figure supplement 1***).

Table 1. Values of fitted parameters and saturation limits for FISNA-NACHT and NLR-B30.2 exons, by population.

Population	FISNA-NACHT				NLR-B30.2			
-	α	β	Limit	Quantile*	α	β	Limit	Quantile*
TU	178.274	1.43356	519.548	118	53.8579	1.40774	164.73	164
CGN	257.207	1.62786	569.367	23	78.7156	1.61283	177.246	25
DP	309.14	1.01231	25284	2930†	69.3609	0.87454	∞	na
KG	436.761	1.2152	2288.41	2060	145.715	1.1418	1113.23	6.41e6
SN	479.892	1.26093	2152.12	3907	145.548	1.10183	1514.35	3.75e9
CHT	416.712	1.18893	2451.81	1.12e5	135.677	1.11911	1218.54	1.41e8

*Sample size required to capture 90% of the population's pan-NLRome.
†DP required sample size refers to only 10% (instead of 90%) of its hypothetical pan-NLRome size.

Copy number variation in the pan-NLRome

Aligning CCS reads to the pan-NLRome revealed a considerable amount of variability in the proportion of reads mapping to them, both between and within populations (***Figure 3A***). This can be interpreted as the gene being present in different copy numbers. Furthermore, each NLR had its own distinct pattern of copy number variation, although generally the highest copy numbers were observed for the wild populations KG, SN, and CHT (***Figure 3A***). We also observed some sequencing batch-related differences, but the copy numbers differed even between individuals sequenced in the same batch.

Of the 1513 unique FISNA-NACHT and 567 NLR-B30.2 sequences, 880 FISNA-NACHT and 346 NLR-B30.2 (59 and 57%, respectively) were detected in at least one individual from all wild populations (***Figure 3B***, ***Figure 3—figure supplement 1***).

There were also NLR sequences shared between just two wild populations, and some were restricted to a single population (***Figure 3B***). Moreover, we observed a lot of variability in the distributions of gene copies among fish within populations (***Figure 3C***). Only around 4% of the genes in the pan-NLRome were found in 80%, or more, of the wild fish. They constitute the core NLRome (***Van de Weyer et al., 2019a***). Most genes (51%) were found in the so-called shell of the pan-NLRome (20–80% of fish). Almost as many (45%) are found in a few fish (less than 20% of the sample) only. Although 60% of NLR genes occur in all wild populations, only 4% are omnipresent, that is, are in the core pan-NLRome. Thus, there is considerable variation in the NLR repertoires of individuals from the same population.

The total number of NLRs identified in a number x of individual fish can be fitted to a harmonic function (***Medini et al., 2020***). Using this function (see 'Materials and methods'), we estimated the sizes of the NLRomes of the populations (***Figure 3D***) and found a total of 520 and 570 NLRs in the laboratory strains TU and CGN, respectively (***Table 1***). For the wild populations, we estimated four times as many: 2283 in KG, ,896 in SN, and 2452 in CHT.

Differences from the reference genome

NLRs sequenced in this study were often different from those present in the reference genome GRCz11. Even NLRs sequenced from the strain that the reference genome itself is based on (TU) did not always align well to it. When the exon itself did align, the intronic sequences surrounding it could often be very different from the reference. In numbers, only around 75% of NLRs occurring in TU fish aligned to the reference genome GRCz11 with high mapping qualities (***Figure 2—figure supplement 2A***). This number dropped even lower elsewhere – from 60–65% of NLRs in CGN which aligned well to the reference, down to only around 50% for the wild populations. The majority of NLRs that did not map well had a very poor mapping quality of 1 (***Figure 2—figure supplement 2B***). Moreover, there were 9 FISNA-NACHT and 10 NLR-associated B30.2 in the pan-NLRome which did not map anywhere in the reference genome.

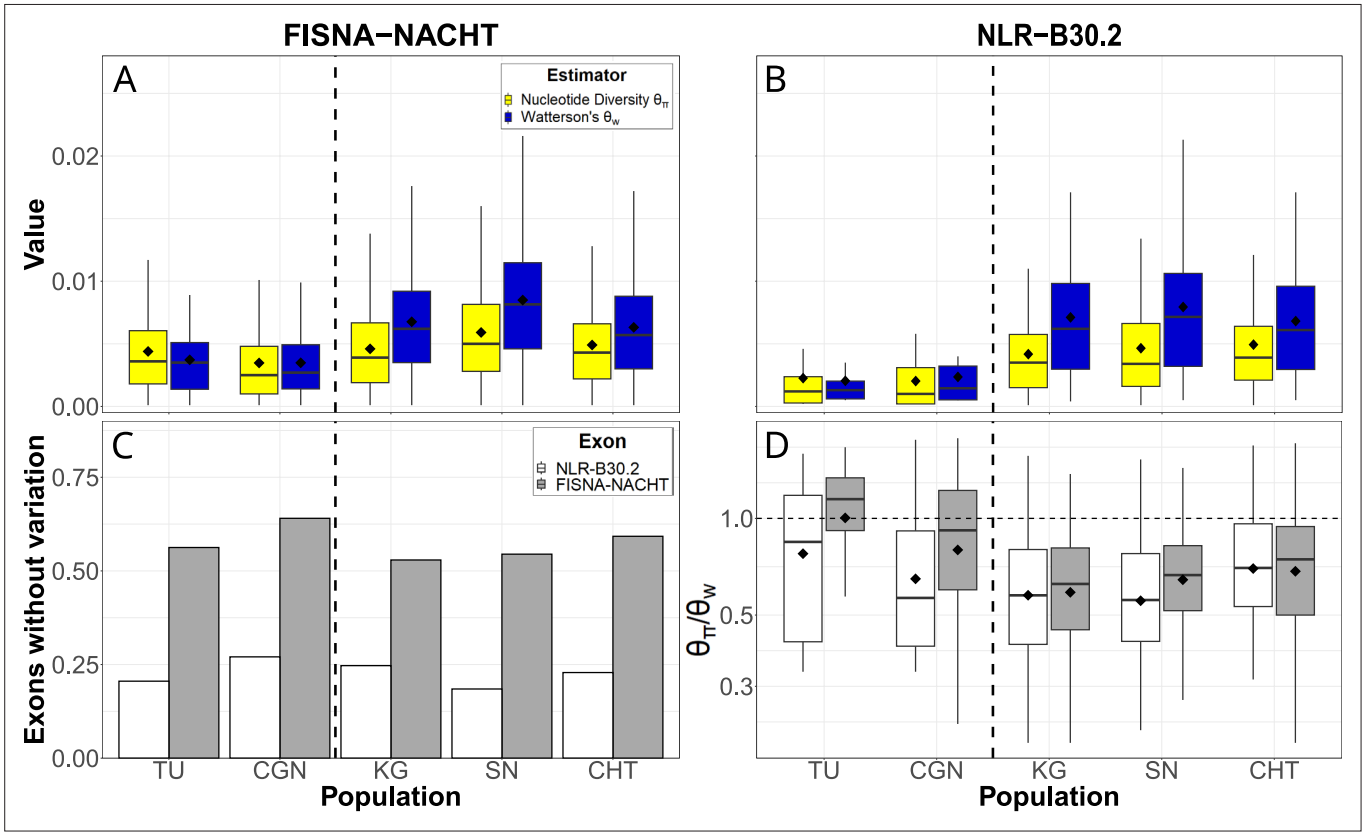


Figure 4. Single-nucleotide variation in NLR exons. Pairwise nucleotide diversity (θ_π) and Watterson's estimator of the scaled mutation rate (θ_w) for FISNA-NACHT (**A**) and NLR-associated B30.2 (**B**) exons. (**C**) Proportion of exons without any single nucleotide polymorphisms. (**D**) Ratio of θ_π/θ_w. Only exons with at least one single-nucleotide polymorphism are shown. The dotted, horizontal line marks a ratio of 1, the expected value under neutrality and constant population size. The black diamonds on box plots denote means, horizontal lines denote medians.

The online version of this article includes the following source data and figure supplement(s) for figure 4:

Source data 1. Source tables for ***Figure 4*** and its supplement.

Figure supplement 1. Single-nucleotide polymorphisms of different NLR exons shown by population, including DP.

Purifying selection on single-nucleotide variants

We used the pan-NLRome as a reference for identifying single-nucleotide polymorphisms in the data. NLR sequence diversity was rare, with a large fraction of exons not having any variants in any of the populations. If variants were present, nucleotide diversity (θ_π) was up to 0.016 and Watterson's estimator (θ_w) up to 0.021 (***Figure 4A and B***). In laboratory strains, genetic variability of FISNA-NACHT exceeded that of B30.2, but no such pattern was observed for wild populations. B30.2 exons of laboratory strains were also less variable than B30.2 from wild zebrafish (***Figure 4B***). The proportion of exons without any polymorphisms was much higher among FISNA-NACHT than among B30.2 (***Figure 4C***). The majority of variable NLR exons had θ_π/θ_w ratios of less than 1 (***Figure 4D***), indicating an excess of rare alleles.

Discussion

We sequenced and assembled the FISNA-NACHT and B30.2 exons of hundreds of NLRs from 93 zebrafish. We were able to capture the diversity of this gene family in three wild populations and two laboratory strains, and produced lower coverage NLR data for an additional wild population (DP). Analyzing the 73 zebrafish from populations other than DP, we found evidence that each genome from a wild individual contains only a fraction of more than 1500 identified NLR copies. The number of NLRs found per individual, each with one or more copies, ranged from around 100–550. Some of

the lower counts were likely underestimated due to low sequencing depths in specific samples. Since all samples from population DP suffered from low read depth, their analysis is only shown in figure supplements. As targeted sequencing based on bait-capture requires sufficient homology between bait and target, diverged NLR exons may have been missed in our approach. This affects B30.2 exons even more than FISNA-NACHT exons because they are much shorter. However, the observed slow increase in newfound NLR gene copies per sequenced individual after the first few individuals indicates that not many NLRs were missed. The sizes of NLR repertoires differ between zebrafish individuals in the three wild populations.

Nonlinear fitting of NLR counts to ***Equation 2*** suggested that the investigated populations all possess closed pan-NLRomes with roughly 500–600 NLRs in the laboratory strains and around 2000 NLRs in the wild populations. The total numbers of NLRs with a B30.2 exon are about 170 in the laboratory strains and between 1100 and 1500 for the wild populations (***Table 1***). To explore the entire NLRome of wild populations, large samples are needed: based on the curve-fitting results, we estimate that capturing 90% of the NLRome may require up to several hundred thousand fish (***Table 1***). Orthogroup clustering with the data from DP resulted in 47 FISNA-NACHT exons which did not occur in any other population. Our results suggest that the pan-NLRome of the entire species must be vastly larger than what we have been able to detect with our limited sample sizes from a limited number of populations. Geographically distant populations – for example, in Nepal or the Western Ghats (***Whiteley et al., 2011***) – likely harbor many more NLRs which are not present in the populations we sequenced.

Although a few zebrafish assemblies are available in addition to the reference genome, for instance, the fDanRer4.1 assembly from the Tree of Life Initiative (GCA_944039275.1), none of those provide a suitable framework for mapping and analyzing NLRs on their own. One of the hindrances is the fact that the majority of NLR genes are located on the notoriously difficult to assemble long arm of chromosome 4, which harbors plenty of structural variation (***McConnell et al., 2023***). Furthermore, large multi-copy gene families are difficult to analyze. Read mapping and counting of copies in a particular genome is not trivial. Any downstream analysis which relies on clearly distinguishing paralogous and orthologous comparisons becomes fuzzy, if not impossible. Still, improving sequencing technology and the rising interest in pan-genomic studies ***Bayer et al., 2020***; ***Sherman and Salzberg, 2020***; ***Liao et al., 2023*** have already started to transform the data structures in which genomes are stored, away from a single-reference genome-based view, toward graph-based genome networks. Whether the promise of a thereby improved inventory of structural variation of a species holds up remains still to be seen. Anyway, as shown for the zebrafish NLRs, the availability of a single high-quality reference genome is certainly not sufficient neither to identify nor to understand the diversity of large gene families.

Properties of the zebrafish NLRome

We have previously demonstrated a substantial reduction in single-nucleotide variation in zebrafish laboratory strains compared to wild populations (***Suurväli et al., 2020***). Here, we showed that the copy numbers of the NLRome and their variation are also heavily reduced. The most obvious explanation for this observation is the recent population bottleneck which marks the establishment of laboratory strains. The reduction in copy number variation in the major histocompatibility complex (MHC) locus in a population of greater prairie-chicken was attributed to a recent bottleneck as well (***Eimes et al., 2011***). Additionally, the reduced amount of pathogenic challenges in a laboratory environment could lead to a steady loss of expendable genes. For these reasons, one has to exercise caution when extending conclusions from immune-related studies on laboratory zebrafish to wild zebrafish. The same caution should also be exercised when extending results from laboratory organisms to other species, including human.

Studies have shown that even mammals have hundreds of genes with diverse molecular functions that are affected by copy number variation, even though it rarely involves full genes (***Kooverjee et al., 2023***; ***Zarrei et al., 2015***). One example of the latter is the MHC locus, which harbors varying numbers of gene copies between closely related species of ruminants (***He et al., 2024***) and has haplotype-specific copies in mice (***Lilue et al., 2018***). However, the vast number of NLRs in zebrafish combined with presence/absence variation (***McConnell et al., 2023***) and high rates of duplication exceeds what has been found in other animals so far. A comparable situation can be found in the

NLR genes of the thale cress (*A. thaliana*). Our predicted number of NLRs in a zebrafish population is on the same scale as the 2127 NLRs found in the thale cress NLRome (**Van de Weyer et al., 2019b**). Moreover, copy numbers also vary greatly between *A. thaliana* accessions (**Lee and Chae, 2020**). A total of 464 conserved, high-confidence orthogroups were identified in *A. thaliana*, 106–143 of which were defined as the core NLRome because they were found in a subset comprising at least 80% of the accessions (**Van de Weyer et al., 2019b**). In wild zebrafish, we found a set of 880 NLR genes which were detected in at least one individual from three wild populations, but only 58 NLRs were found in the vast majority (more than 80%) of wild individuals. Although structural similarities of NLRs in plants and animals are thought to be the result of convergent evolution (**Yue et al., 2012**), it appears that the similarities extend to their evolutionary trajectories. However, the overall number of gene copies as well as the variation in copy numbers within populations and in individual gene repertoires are more extreme in zebrafish than in *A. thaliana*.

We postulate that as immune genes, many NLR genes are likely shared between populations because they provide a fitness advantage in the defense against common pathogens. The additional NLRs shared among only some of the wild populations and the population-specific NLRs may represent local adaptations to ecological niches. Additionally, there could be functional redundancy within the NLRome, so that different individuals have different NLRs with the same functional role. In general, the fact that hundreds of NLR gene copies are maintained in zebrafish, together with a signature of purifying selection, suggests that the evolution of these genes is far from neutral. Although the expression of fish NLRs is often induced by pathogen exposure (reviewed in **Chuphal et al., 2022**), the exact function of most zebrafish NLR-C genes remains unclear. It is possible that some of them participate in the formation and activity of inflammasomes (**Li, 2018a**; **Valera-Pérez et al., 2019**; **Lozano-Gil et al., 2022**, **Kuri et al., 2017**), but we only found the N-terminal effector domains (CARD or PYD) that are typically involved in this function (**Petrilli et al., 2005**) in a small subset of NLR-C genes.

Although we mainly used the counts of FISNA-NACHT orthogroups to estimate total numbers of NLRs, we also analyzed the B30.2 exons of NLR-C genes. In general, NLR-associated B30.2 exons exhibit patterns of copy number variation that are similar to those seen for FISNA-NACHT. For example, about half of the B30.2 sequences are found in all wild populations, similar to the set of 880 FISNA-NACHT exon sequences conserved among populations.

What drives the copy number differences?

There are at least two mechanisms which could contribute to the extensive copy number variation seen among zebrafish populations: first, it could be attributed to a high degree of haplotypic variation. Large DNA fragments contain different sets of genes and gene copies, similar to the zebrafish MHC loci (**McConnell et al., 2014**). Extensive haplotypic variation occurs on the long arm of chromosome 4, the location containing over two-thirds of all NLRs in zebrafish (**McConnell et al., 2023**). Such segregating haplotype blocks would explain the existence of the core NLRome, but not the frequent presence of genes that occur only in a single individual.

Alternatively or additionally, the evolution of NLR-C genes could be driven by duplication events (**Cannon et al., 2004**) and gene conversion (**Laing et al., 2008**). Gene duplications can be caused by unequal recombination, transposon activity, or whole genome/chromosome duplications (**Magadum et al., 2013**; **Kapitonov and Jurka, 2007**). The arrangement of NLR-B30.2 genes in clusters on the long arm of chromosome 4 suggests that tandem duplication via unequal crossing-over (**Otto et al., 2022**) played the most important role in the expansion. Since there are many transposable elements on the long arm of chromosome 4 (**Howe et al., 2013**), it would be reasonable to assume that at least some of them have assisted in the local expansion and transfer of NLR exons and genes to chromosomes other than chromosome 4. Since our targeted sequencing approach does not elucidate the genomic arrangement of the NLR gene copies and many of them do not have recognizable orthologs in the reference genome, we cannot draw further conclusions about the role of tandem arrays in their evolutionary trajectory.

It is tempting to speculate that chromosome 4 could be a source of NLRs which continuously generates new copies. However, gene gains must be balanced by gene loss to maintain a stable genome size. NLR-C genes may be lost via accumulation of random mutations due to a lack of selective pressure and loss-of-function mutations, but they may also be lost through unequal recombination. This

mechanism would allow only NLR genes contributing to the functionality of the immune system to be kept, while others would disappear.

In the similarly evolving plant NLRs, tandem duplication is thought to be the primary driver of NLR gene expansion (***Cannon et al., 2004***), but they are also often associated with transposable elements. If the diversity of unrelated NLR genes in such distantly related species is driven by common molecular mechanisms, then the same mechanisms might also act on NLRs of other phylogenetic clades and even on unrelated large gene families, such as odorant receptors (***Mombaerts, 1999***).

Conclusion

This study showcases an example of the evolutionary dynamics affecting very large gene families. The sheer amount of copy number variation that appears to be present in a single gene family of zebrafish is staggering, with different individuals each having numerous genes that are not present in all others. This can only be caused by diversity-generating mechanisms that are active even now. In this study, we have laid the groundwork for future studies investigating the molecular basis and evolutionary mechanisms contributing to the diversity of large, vertebrate gene families.

Materials and methods

Key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Strain (*Danio rerio*)	Cologne zebrafish; CGN; KOLN	Other		8 Cologne fish, AG Hammerschmidt, University of Cologne
Strain (*D. rerio*)	Tübingen zebrafish; TU	Other		8 Tübingen fish, AG Hammerschmidt, University of Cologne
Biological sample (*D. rerio*)	DP	Other		20 wild fish, Dandiapalli, India (22.22155, 84.79430)
Biological sample (*D. rerio*)	CHT	Other		20 wild fish, Chittagong, Bangladesh (22.47400, 91.78300)
Biological sample (*D. rerio*)	KG	Other		20 wild fish, Leturakhal, India (22.26189 87.27881)
Biological sample (*D. rerio*)	SN	Other		20 wild fish, Santoshpur, India (22.93765 88.55311)
Sequence-based reagent	Baits; RNA baits; hybridization baits	Daicel Arbor Biosciences	Cat# Mybaits-1-24	Sequences available in ***Figure 2—source data 2***
Commercial assay or kit	MagAttract HMW DNA Kit	QIAGEN	Cat# 67563	
Commercial assay or kit	NucleoSpin Tissue Kit	MACHEREY-NAGEL	Cat# 740952.50	
Commercial assay or kit	NEBNext Ultra II DNA Library Prep Kit	New England Biolabs	Cat# E7645L	
Sequence-based reagent	NEBNext Multiplex Oligos for Illumina	New England Biolabs	Cat# E7335L	Index Primers Set 1
Commercial assay or kit	Kapa HiFi Hotstart Readymix	Kapa Biosystems	Cat# 07958935001	
Commercial assay or kit	PreCR Repair Mix	New England Biolabs	Cat# M0309L	
Commercial assay or kit	SMRTbell Template Prep Kit 1.0-SPv3	Pacific Biosciences	Cat# 100-991-900	
Other	GRCz11	NCBI RefSeq	RefSeq:GCF_000002035.6	Zebrafish reference genome

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Other	M220 miniTUBE, Red	Covaris	Cat# 4482266	Used to shear DNA on Covaris ultrasonicator
Other	DB MyOne Streptavidin C1	Thermo Fisher Scientific	Cat# 65001	Used to retrieve bait-bound DNA fragments
Other	AMPure XP	Beckman Coulter	Cat# A63881	Size selection beads
Other	Ampure PB	Pacific Biosciences	Cat# 100-265-900	PacBio-compatible size selection beads
Software, algorithm	lima	Pacific Biosciences	lima:v1.0.0; lima:v1.8.0; lima:v1.9.0; lima:v1.11.0	
Software, algorithm	ccs	Pacific Biosciences	ccs:v4.2.0	
Software, algorithm	pbmarkdup	Pacific Biosciences	pbmarkdup:v1.0.0	
Software, algorithm	pbmm2	Pacific Biosciences	pbmm2:v1.3.0	
Software, algorithm	samtools	https://doi.org/10.1093/bioinformatics/btp352	samtools:v1.7	
Software, algorithm	EMBOSS	https://doi.org/10.1016/s0168-9525(00)02024-2	EMBOSS:v6.6.0.0	
Software, algorithm	HMMER	https://doi.org/10.1093/bioinformatics/btt403	HMMER:v3.2.1	
Software, algorithm	blastn	https://doi.org/10.1186/1471-2105-10-421	blastn:v2.11.0+	
Software, algorithm	hifiasm	https://doi.org/10.1038/s41592-020-01056-5	hifiasm:v0.15.4-r347	
Software, algorithm	get_homologues	https://doi.org/10.1128/AEM.02411-13	get_homologues:x86_64–20220516	
Software, algorithm	deepvariant	https://doi.org/10.1038/nbt.4235	deepvariant:r1.0	
Software, algorithm	GLnexus	https://doi.org/10.1101/343970	Glnexus:v1.2.7–0-g0e74fc4	
Software, algorithm	vcftools	https://doi.org/10.1093/bioinformatics/btr330	vcftools:v0.1.16	

Samples

Wild zebrafish from four sites in India and Bangladesh (***Figure 1B***) had been collected in the frame of other projects (e.g., ***Whiteley et al., 2011***; ***Shelton et al., 2020***). Laboratory zebrafish from the Tübingen (TU) and Cologne (CGN) strains were provided by Dr. Cornelia Stein from the Hammerschmidt laboratory (Institute for Zoology, University of Cologne). All samples were stored in 95% ethanol until use. Tail fins from 20 fish per wild population and 8 fish per laboratory strain were used as starting material for the subsequent steps.

DNA extraction, exon capture, and sequencing

Genomic DNA was extracted with kits from QIAGEN (MagAttract HMW kit) and MACHEREY-NAGEL (Nucleospin Tissue Kit), followed by shearing with red miniTUBEs on the Covaris M220 ultrasonicator. Nicks in the DNA were repaired with PreCR Repair Mix (New England Biolabs). Samples were barcoded with the NEBNext Ultra II DNA Library Prep Kit, then pooled together in batches of four or eight (details provided in Appendix 1). RNA baits for the exon capture (Daicel Arbor Biosciences) were custom-designed to target immune genes of interest (mainly NLRs, but also some others) based on version GRCz10 of the reference genome. Bait sequences and target locations are available in ***Figure 2—source data 2***. Exon capture and PacBio library preparation were both done according to a protocol adapted from ***Witek et al., 2016***. Libraries were sequenced at the Max Planck-Genome-Centre Cologne, with PacBio Sequel and Sequel II. Additional details are provided in Appendix 1.

Read processing, mapping, and clustering

Raw sequences were de-multiplexed with lima. Consensus sequences of DNA fragments with at least three passes (CCS reads) were inferred with ccs, followed by PCR duplicate removal with pbmarkdup. All read mapping was done with pbmm2 (v.1.3.0), a PacBio wrapper for minimap2 (***Li, 2018a***). lima, ccs, pbmarkdup, and pbmm2 were all provided by Pacific Biosciences. Mapped files were processed and filtered with samtools (v1.7) (***Li et al., 2009***). De novo assemblies were generated with hifiasm (v0.15.4-r347) (***Cheng et al., 2021***). Tools from the HMMER suite (v3.2.1) (***Wheeler and Eddy, 2013***) were used to detect the presence of NLR-associated sequences. Contigs containing FISNA-NACHT or B30.2 were sorted into orthoclusters using get_homologues (build x86 64–20220516) (***Contreras-Moreira and Vinuesa, 2013***) and blastn (v2.11.0+) (***Altschul et al., 1990***). Orthoclusters for which pbmm2 did not align any CCS reads to the representative sequence with at least 95% identity were excluded from further analyses. Further details are provided in Appendix 1.

Modeling

To estimate the full size of each population's NLR repertoire, we calculated the increment in the total number of identified NLR exon sequences when adding sequence data from one additional individual of a population to a set of already surveyed individuals. As noted earlier (***Medini et al., 2020***), these increments are well approximated by a power-law decay.

Briefly, given a sample of n individuals, there are

$$w_n(x) = \binom{n}{x-1}(n-(x-1)) = \binom{n}{x}x \tag{1}$$

ways to choose $x-1$ individuals from the entire sample and add another – not yet chosen – one. For each x, we calculated the increment in the number of identified exon sequences and averaged over all possible choices of individuals. Summation of the average increments yields the total number of exons identified with x individuals, as plotted in ***Figure 3D***. Then, we fitted the nonlinear function

$$y = \alpha H(x,\beta) \tag{2}$$

where $H(x,\beta)$ is the generalized harmonic number with parameter β, that is,

$$H(x,\beta) = \sum_{k=1}^{x} \frac{1}{k^{\beta}} \tag{3}$$

It represents the sum of increments, decaying according to a power-law, with parameters α (intercept) and β (decay rate). Importantly, if $\beta > 1$, the series in ***Equation 3*** converges and its limit may be interpreted as the size of a *closed* NLRome. The NLRome is *open*, if $\beta \leq 1$. Values of the fitted parameters and saturation limits are presented in ***Table 1***.

Genetic diversity

Single-nucleotide genotypes in each fish were identified from the.bam output of pbmm2 by using deepvariant (r1.0) (***Poplin et al., 2018***) with the PacBio model. Joint genotyping of the individual samples was done with glnexus (v1.2.7–0-g0e74fc4) (***Yun et al., 2021***) with its deepvariant-specific setting. Per-site θ_π of the NLR exons was calculated with vcftools (v0.1.16) (***Danecek et al., 2011***). Watterson's estimator of the scaled mutation rate is

$$\theta_w = \frac{S}{H(n-1,1)\,l} \tag{4}$$

where S is the number of segregating sides seen in a sample of n aligned sequences, each of size l (here, 1761 bp for the FISNA-NACHT exons and 540 bp for the B30.2 exons).

Under neutrality (all alleles confer the same fitness to an individual) and constant population size over time, one expects equality $\theta_\pi = \theta_w$.

Data visualization

Plots and heat maps were created in RStudio (v2022.07.2) with R (v4.2.1) using ggplot2 (v3.3.6) or xmgrace (v5.1.25; https://plasma-gate.weizmann.ac.il/Grace/). Venn diagrams were created via

BioVenn (***Hulsen et al., 2008***; ***Figure 3B***) and ggvenn (v0.1.9) (***Figure 1A***). Final processing of the images was done in Inkscape (v1.1.2; https://inkscape.org/).

Acknowledgements

The authors are grateful to Emilia Martins and Anuradha Bhat for their contributions in wild sample collection, and to Cornelia Stein and the laboratory of Matthias Hammerschmidt for laboratory samples. We thank Bruno Hüttel and Max Planck-Genome-Centre Cologne for all the advice with library construction and for sequencing. We are also thankful for the contributions of Philipp Schiffer (help with writing the initial project proposal), Lisa Vogelsang (assistance with laboratory work), and Robert Fürst and Anna Rottmann (management of computational infrastructure). We further thank anonymous peer-reviewers for constructive feedback which helped us improve the manuscript. This work was funded by a grant to TW and ML in the frame of the priority program SPP-1819 of the German Research Foundation (DFG). JS was additionally supported by a National Sciences and Engineering Council of Canada (NSERC) Discovery Grant to Colin Garroway.

Additional information

Funding

Funder	Grant reference number	Author
Deutsche Forschungsgemeinschaft	SPP1819	Maria Leptin Thomas Wiehe

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Yannick Schäfer, Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing – original draft, Writing – review and editing; Katja Palitzsch, Investigation, Methodology, Writing – review and editing; Maria Leptin, Conceptualization, Resources, Supervision, Funding acquisition, Validation, Project administration, Writing – review and editing; Andrew R Whiteley, Resources, Investigation, Writing – review and editing; Thomas Wiehe, Conceptualization, Resources, Formal analysis, Supervision, Funding acquisition, Visualization, Methodology, Project administration, Writing – review and editing; Jaanus Suurväli, Data curation, Software, Formal analysis, Supervision, Investigation, Visualization, Methodology, Writing – original draft, Writing – review and editing

Author ORCIDs

Yannick Schäfer http://orcid.org/0000-0002-5264-8816
Katja Palitzsch http://orcid.org/0000-0002-6292-4925
Maria Leptin http://orcid.org/0000-0001-7097-348X
Andrew R Whiteley http://orcid.org/0000-0002-8159-6381
Thomas Wiehe http://orcid.org/0000-0002-8932-2772
Jaanus Suurväli http://orcid.org/0000-0003-0133-7011

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.98058.sa1
Author response https://doi.org/10.7554/eLife.98058.sa2

Additional files

Supplementary files

- MDAR checklist

Data availability

NLR reads are available in the NCBI Sequence Read Archive (BioProject PRJNA966920). Scripts are available on GitHub (https://github.com/YSchaefer/pacbio_zebrafish, copy archived at ***Schaefer, 2024***). Sequences of the hybridization baits are provided as a source dataset.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
University of Cologne, Yannick Schaefer	2023	Targeted PacBio Sequencing of Zebrafish NLR Exons	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA966920	NCBI BioProject, PRJNA966920

The following previously published dataset was used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Genome Reference Consortium	2017	Genome assembly GRCz11	https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002035.6/	NCBI Assembly, GCF_000002035.6

References

Adrian-Kalchhauser I, Blomberg A, Larsson T, Musilova Z, Peart CR, Pippel M, Solbakken MH, Suurväli J, Walser JC, Wilson JY, Alm Rosenblad M, Burguera D, Gutnik S, Michiels N, Töpel M, Pankov K, Schloissnig S, Winkler S. 2020. The round goby genome provides insights into mechanisms that may facilitate biological invasions. *BMC Biology* **18**:11. DOI: https://doi.org/10.1186/s12915-019-0731-8, PMID: 31992286

Almeida-da-Silva CLC, Savio LEB, Coutinho-Silva R, Ojcius DM. 2023. The role of NOD-like receptors in innate immunity. *Frontiers in Immunology* **14**:1122586. DOI: https://doi.org/10.3389/fimmu.2023.1122586, PMID: 37006312

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**:403–410. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2

Anderson JL, Rodríguez Marí A, Braasch I, Amores A, Hohenlohe P, Batzel P, Postlethwait JH. 2012. Multiple sex-associated regions and a putative sex chromosome in zebrafish revealed by RAD mapping and population genomics. *PLOS ONE* **7**:e40701. DOI: https://doi.org/10.1371/journal.pone.0040701

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. 2020. Plant pan-genomes are the new reference. *Nature Plants* **6**:914–920. DOI: https://doi.org/10.1038/s41477-020-0733-0

Cannon SB, Mitra A, Baumgarten A, Young ND, May G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biology* **4**:1–21. DOI: https://doi.org/10.1186/1471-2229-4-10, PMID: 15171794

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**:170–175. DOI: https://doi.org/10.1038/s41592-020-01056-5, PMID: 33526886

Chuphal B, Rai U, Roy B. 2022. Teleost NOD-like receptors and their downstream signaling pathways: A brief review. *Fish and Shellfish Immunology Reports* **3**:100056. DOI: https://doi.org/10.1016/j.fsirep.2022.100056, PMID: 36419601

Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology* **79**:7696–7701. DOI: https://doi.org/10.1128/AEM.02411-13, PMID: 24096415

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Research* **14**:1188–1190. DOI: https://doi.org/10.1101/gr.849004, PMID: 15173120

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* **27**:2156–2158. DOI: https://doi.org/10.1093/bioinformatics/btr330, PMID: 21653522

Eimes JA, Bollmer JL, Whittingham LA, Johnson JA, VAN Oosterhout C, Dunn PO. 2011. Rapid loss of MHC class II variation in a bottlenecked population is explained by drift and loss of copy number variation. *Journal of Evolutionary Biology* **24**:1847–1856. DOI: https://doi.org/10.1111/j.1420-9101.2011.02311.x, PMID: 21605219

FAO. 2021. AQUASTATS: Rivers of South and East Asia. https://data.apps.fao.org/catalog/dataset/dc2a5121-0b32-482b-bd9b-64f7a414fa0d [Accessed April 5, 2023].

Gao LA, Wilkinson ME, Strecker J, Makarova KS, Macrae RK, Koonin EV, Zhang F. 2022. Prokaryotic innate immunity through pattern recognition of conserved viral proteins. *Science* **377**:eabm4096. DOI: https://doi.org/10.1126/science.abm4096, PMID: 35951700

He K, Liang C, Ma S, Liu H, Zhu Y. 2024. Copy number and selection of MHC genes in ruminants are related to habitat, average life span and diet. *Gene* **904**:148179. DOI: https://doi.org/10.1016/j.gene.2024.148179, PMID: 38242373

Hibino T, Loza-Coll M, Messier C, Majeske AJ, Cohen AH, Terwilliger DP, Buckley KM, Brockton V, Nair SV, Berney K, Fugmann SD, Anderson MK, Pancer Z, Cameron RA, Smith LC, Rast JP. 2006. The immune gene repertoire encoded in the purple sea urchin genome. *Developmental Biology* **300**:349–365. DOI: https://doi.org/10.1016/j.ydbio.2006.08.065

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, McLaren S, Sealy I, Caccamo M, Churcher C, Scott C, Barrett JC, Koch R, Rauch G-J, White S, Chow W, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**:498–503. DOI: https://doi.org/10.1038/nature12111, PMID: 23594743

Howe K, Schiffer PH, Zielinski J, Wiehe T, Laird GK, Marioni JC, Soylemez O, Kondrashov F, Leptin M. 2016. Structure and evolutionary history of a large family of NLR proteins in the zebrafish. *Open Biology* **6**:160009. DOI: https://doi.org/10.1098/rsob.160009, PMID: 27248802

Hulsen T, de Vlieg J, Alkema W. 2008. BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **9**:1–6. DOI: https://doi.org/10.1186/1471-2164-9-488, PMID: 18925949

Jones JDG, Dangl JL. 2006. The plant immune system. *Nature* **444**:323–329. DOI: https://doi.org/10.1038/nature05286

Kapitonov VV, Jurka J. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends in Genetics* **23**:521–529. DOI: https://doi.org/10.1016/j.tig.2007.08.004

Kooverjee BB, Soma P, van der Nest MA, Scholtz MM, Neser FWC. 2023. Copy number variation discovery in south african nguni-sired and bonsmara-sired crossbred cattle. *Animals* **13**:2513. DOI: https://doi.org/10.3390/ani13152513, PMID: 37570321

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Research* **19**:1639–1645. DOI: https://doi.org/10.1101/gr.092759.109, PMID: 19541911

Kuri P, Schieber NL, Thumberger T, Wittbrodt J, Schwab Y, Leptin M. 2017. Dynamics of in vivo ASC speck formation. *The Journal of Cell Biology* **216**:2891–2909. DOI: https://doi.org/10.1083/jcb.201703103, PMID: 28701426

Laing KJ, Purcell MK, Winton JR, Hansen JD. 2008. A genomic view of the NOD-like receptor family in teleost fish: identification of A novel NLR subfamily in zebrafish. *BMC Evolutionary Biology* **8**:1–15. DOI: https://doi.org/10.1186/1471-2148-8-42, PMID: 18254971

Lee RRQ, Chae E. 2020. Variation Patterns of NLR Clusters in *Arabidopsis thaliana* Genomes. *Plant Communications* **1**:100089. DOI: https://doi.org/10.1016/j.xplc.2020.100089, PMID: 33367252

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**:2078–2079. DOI: https://doi.org/10.1093/bioinformatics/btp352, PMID: 19505943

Li J, Chu Q, Xu T. 2016. A genome-wide survey of expansive NLR-C subfamily in miiuy croaker and characterization of the NLR-B30.2 genes. *Developmental & Comparative Immunology* **61**:116–125. DOI: https://doi.org/10.1016/j.dci.2016.03.011

Li H. 2018a. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**:3094–3100. DOI: https://doi.org/10.1093/bioinformatics/bty191, PMID: 29750242

Li J, Gao K, Shao T, Fan D, Hu C, Sun C, Dong W, Lin A, Xiang L, Shao J. 2018b. Characterization of an NLRP1 Inflammasome from zebrafish reveals a unique sequential activation mechanism underlying inflammatory caspases in ancient vertebrates. *The Journal of Immunology* **201**:1946–1966. DOI: https://doi.org/10.4049/jimmunol.1800498

Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, Buonaiuto S, Chang XH, Cheng H, Chu J, Colonna V, Eizenga JM, Feng X, Fischer C, Fulton RS, Garg S, et al. 2023. A draft human pangenome reference. *Nature* **617**:312–324. DOI: https://doi.org/10.1038/s41586-023-05896-x, PMID: 37165242

Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, Bennett R, Chow W, Collins J, Collins S, Czechanski A, Danecek P, Diekhans M, Dolle D-D, Dunn M, Durbin R, Earl D, Ferguson-Smith A, Flicek P, Flint J, Frankish A, et al. 2018. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nature Genetics* **50**:1574–1583. DOI: https://doi.org/10.1038/s41588-018-0223-8, PMID: 30275530

Lozano-Gil JM, Rodríguez-Ruiz L, Tyrkalska SD, García-Moreno D, Pérez-Oliva AB, Mulero V. 2022. Gasdermin E mediates pyroptotic cell death of neutrophils and macrophages in a zebrafish model of chronic skin inflammation. *Developmental & Comparative Immunology* **132**:104404. DOI: https://doi.org/10.1016/j.dci.2022.104404

Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. 2013. Gene duplication as a major force in evolution. *Journal of Genetics* **92**:155–161. DOI: https://doi.org/10.1007/s12041-013-0212-8

McConnell SC, Restaino AC, de Jong JLO. 2014. Multiple divergent haplotypes express completely distinct sets of class I MHC genes in zebrafish. *Immunogenetics* **66**:199–213. DOI: https://doi.org/10.1007/s00251-013-0749-y, PMID: 24291825

McConnell SC, Hernandez KM, Andrade J, de Jong JLO. 2023. Immune gene variation associated with chromosome-scale differences among individual zebrafish genomes. *Scientific Reports* **13**:7777. DOI: https://doi.org/10.1038/s41598-023-34467-3, PMID: 37179373

Medini D, Donati C, Rappuoli R, Tettelin H. 2020. The Pangenome: A data-driven discovery in biology. Tettelin H, Medini D (Eds). *The Pangenome*. Springer. DOI: https://doi.org/10.1007/978-3-030-38281-0

Mombaerts P. 1999. Molecular biology of odorant receptors in vertebrates. *Annual Review of Neuroscience* **22**:487–509. DOI: https://doi.org/10.1146/annurev.neuro.22.1.487, PMID: 10202546

Otto M, Zheng Y, Wiehe T. 2022. Recombination, selection, and the evolution of tandem gene arrays. *Genetics* **221**:iyac052. DOI: https://doi.org/10.1093/genetics/iyac052, PMID: 35460227

Petrilli V, Papin S, Tschopp J. 2005. The inflammasome. *Current Biology* **15**:R581. DOI: https://doi.org/10.1016/j.cub.2005.07.049

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**:983–987. DOI: https://doi.org/10.1038/nbt.4235

Rajendran KV, Zhang J, Liu S, Kucuktas H, Wang X, Liu H, Sha Z, Terhune J, Peatman E, Liu Z. 2012. Pathogen recognition receptors in channel catfish: I. Identification, phylogeny and expression of NOD-like receptors. *Developmental & Comparative Immunology* **37**:77–86. DOI: https://doi.org/10.1016/j.dci.2011.12.005

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The european molecular biology open software suite. *Trends in Genetics* **16**:276–277. DOI: https://doi.org/10.1016/S0168-9525(00)02024-2

Schaefer Y. 2024. Pacbio_Zebrafish. swh:1:rev:f0ca8f4882960af1bbf86c4965aa83a7c9d46eed. Software Heritage. https://archive.softwareheritage.org/swh:1:dir:3f46915fd8d0ba244d03271a7f6d43015131377e;origin=https://github.com/YSchaefer/pacbio_zebrafish;visit=swh:1:snp:1d6f5528bc0d6956e17b73828fc208b3cc543938;anchor=swh:1:rev:f0ca8f4882960af1bbf86c4965aa83a7c9d46eed

Shelton DS, Shelton SG, Daniel DK, Raja M, Bhat A, Tanguay RL, Higgs DM, Martins EP. 2020. Collective behavior in wild zebrafish. *Zebrafish* **17**:243–252. DOI: https://doi.org/10.1089/zeb.2019.1851, PMID: 32513074

Sherman RM, Salzberg SL. 2020. Pan-genomics in the human genome era. *Nature Reviews. Genetics* **21**:243–254. DOI: https://doi.org/10.1038/s41576-020-0210-7, PMID: 32034321

Stein C, Caccamo M, Laird G, Leptin M. 2007. Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. *Genome Biology* **8**:1–23. DOI: https://doi.org/10.1186/gb-2007-8-11-r251, PMID: 18039395

Suurväli J, Whiteley AR, Zheng Y, Gharbi K, Leptin M, Wiehe T. 2020. The laboratory domestication of zebrafish: From diverse populations to inbred substrains. *Molecular Biology and Evolution* **37**:1056–1069. DOI: https://doi.org/10.1093/molbev/msz289, PMID: 31808937

Suurväli J, Garroway CJ, Boudinot P. 2022. Recurrent expansions of B30.2-associated immune receptor families in fish. *Immunogenetics* **74**:129–147. DOI: https://doi.org/10.1007/s00251-021-01235-4

Thioulouse J, Dray S, Dufour AB, Siberchicot A, Jombart T, Pavoine S. 2018. *Multivariate Analysis of Ecological Data with Ade4*. Springer. DOI: https://doi.org/10.1007/978-1-4939-8850-1

Ting JP-Y, Lovering RC, Alnemri ES, Bertin J, Boss JM, Davis BK, Flavell RA, Girardin SE, Godzik A, Harton JA, Hoffman HM, Hugot J-P, Inohara N, Mackenzie A, Maltais LJ, Nunez G, Ogura Y, Otten LA, Philpott D, Reed JC, et al. 2008. The NLR gene family: a standard nomenclature. *Immunity* **28**:285–287. DOI: https://doi.org/10.1016/j.immuni.2008.02.005, PMID: 18341998

Tørresen OK, Brieuc MSO, Solbakken MH, Sørhus E, Nederbragt AJ, Jakobsen KS, Meier S, Edvardsen RB, Jentoft S. 2018. Genomic architecture of haddock (Melanogrammus aeglefinus) shows expansions of innate immune genes and short tandem repeats. *BMC Genomics* **19**:1–17. DOI: https://doi.org/10.1186/s12864-018-4616-y

Uehling J, Deveau A, Paoletti M. 2017. Do fungi have an innate immune response? An NLR-based comparison to plant and animal immune systems. *PLOS Pathogens* **13**:e1006578. DOI: https://doi.org/10.1371/journal.ppat.1006578

Urbach JM, Ausubel FM. 2017. The NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events. *PNAS* **114**:1063–1068. DOI: https://doi.org/10.1073/pnas.1619730114, PMID: 28096345

Valera-Pérez A, Tyrkalska SD, Viana C, Rojas-Fernández A, Pelegrín P, García-Moreno D, Pérez-Oliva AB, Mulero V. 2019. WDR90 is a new component of the NLRC4 inflammasome involved in *Salmonella* Typhimurium resistance. *Developmental & Comparative Immunology* **100**:103428. DOI: https://doi.org/10.1016/j.dci.2019.103428

van der Aa LM, Levraud J-P, Yahmi M, Lauret E, Briolat V, Herbomel P, Benmansour A, Boudinot P. 2009. A large new subset of TRIM genes highly diversified by duplication and positive selection in teleost fish. *BMC Biology* **7**:1–23. DOI: https://doi.org/10.1186/1741-7007-7-7, PMID: 19196451

Van de Weyer AL, Monteiro F, Furzer OJ, Nishimura MT, Cevik V, Witek K, Jones JDG, Dangl JL, Weigel D, Bemm F. 2019a. The *Arabidopsis thaliana* Pan-NLRome. *bioRxiv*. DOI: https://doi.org/10.1101/537001

Van de Weyer AL, Monteiro F, Furzer OJ, Nishimura MT, Cevik V, Witek K, Jones JDG, Dangl JL, Weigel D, Bemm F. 2019b. A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. *Cell* **178**:1260–1272. DOI: https://doi.org/10.1016/j.cell.2019.07.038, PMID: 31442410

Wheeler TJ, Eddy SR. 2013. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**:2487–2489. DOI: https://doi.org/10.1093/bioinformatics/btt403

Whiteley AR, Bhat A, Martins EP, Mayden RL, Arunachalam M, Uusi-Heikkilä S, Ahmed ATA, Shrestha J, Clark M, Stemple D, Bernatchez L. 2011. Population genomics of wild and laboratory zebrafish (*Danio rerio*). *Molecular Ecology* **20**:4259–4276. DOI: https://doi.org/10.1111/j.1365-294X.2011.05272.x, PMID: 21923777

Witek K, Jupe F, Witek AI, Baker D, Clark MD, Jones JDG. 2016. Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing. *Nature Biotechnology* **34**:656–660. DOI: https://doi.org/10.1038/nbt.3540

Woo J-S, Imm J-H, Min C-K, Kim K-J, Cha S-S, Oh B-H. 2006. Structural and functional insights into the B30.2/SPRY domain. *The EMBO Journal* **25**:1353–1363. DOI: https://doi.org/10.1038/sj.emboj.7600994, PMID: 16498413

Yue J, Meyers BC, Chen J, Tian D, Yang S. 2012. Tracing the origin and evolutionary history of plant nucleotide-binding site–leucine-rich repeat (NBS-LRR) genes. *New Phytologist* **193**:1049–1063. DOI: https://doi.org/10.1111/j.1469-8137.2011.04006.x

Yuen B, Bayes JM, Degnan SM. 2014. The characterization of sponge NLRs provides insight into the origin and evolution of this innate immune gene family in animals. *Molecular Biology and Evolution* **31**:106–120. DOI: https://doi.org/10.1093/molbev/mst174, PMID: 24092772

Yun T, Li H, Chang P-C, Lin MF, Carroll A, McLean CY. 2021. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**:5582–5589. DOI: https://doi.org/10.1093/bioinformatics/btaa1081, PMID: 33399819

Zarrei M, MacDonald JR, Merico D, Scherer SW. 2015. A copy number variation map of the human genome. *Nature Reviews Genetics* **16**:172–183. DOI: https://doi.org/10.1038/nrg3871

Appendix 1

Supplementary methods

DNA extraction

High molecular weight (HMW) DNA from laboratory zebrafish was extracted from caudal fin clips using the QIAGEN MagAttract HMW DNA extraction kit. HMW DNA from wild zebrafish was extracted from caudal fin clips using the NucleoSpin Tissue Kit from MACHEREY-NAGEL with the following adjustments. Tissues other than muscle were removed before DNA extraction with forceps. The incubation time of the Proteinase K treatment was changed from 1 to 3 hr to 10–15 min. An RNAse A treatment step was included by incubating with 400 µg RNAse A (Sigma-Aldrich) for 2 min at room temperature. All DNA samples were quantified and quality checked with Qubit 3.0 (Thermo Fisher Scientific), 0.8% agarose gels, and the 4200 TapeStation Electrophoresis System (Agilent Technologies). DNA extraction failed for one of the 20 CHT samples, but was successful for the other 95 fin clips.

Shearing and barcoding

HMW DNA was sheared into 1.5–6 kb fragments with the red miniTUBEs of the Covaris M220 ultrasonicator. Quality control after shearing was performed using the 4200 TapeStation Electrophoresis System (Agilent Technologies). The obtained DNA fragments were size selected with 0.4× AMPure XP beads (Beckmann Coulter Inc) to exclude fragments smaller than 1.5 kb. For wild zebrafish samples, a DNA damage repair step was included in order to repair any possible DNA damage resulting from long periods of storage (particularly important for the older CHT samples). The repair step was carried out with PreCR Repair Mix (New England Biolabs).

DNA fragments were barcoded with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) and NEBNext Multiplex Oligos for Illumina, Index Primers Set 1 (New England Biolabs). The manufacturer's standard protocol was followed until the amplification step for the enrichment of barcode-ligated fragments. At this stage, the recommended amplification protocol (PCR program) was modified to suit large DNA fragments (***Appendix 1—table 2***) and the high-fidelity Kapa polymerase (Kapa HiFi Hotstart Readymix, Kapa Biosystems) was used. The resulting barcoded DNA was purified and size-selected two more times, first with 0.5× AMPure XP beads and then with 0.4× AMPure XP beads. The amount of DNA was quantified with Qubit and quality checked with gel electrophoresis on a 0.8% agarose gel. The samples were then pooled with each pooled sample containing barcoded DNA of either four fish (CGN, TU, first library of each wild population) or eight fish (the remaining libraries of the wild populations).

NLR capture with hybridization baits

Target enrichment was carried out according to MYbaits manual version 3.02 by using the MYbaits customized target enrichment kit for Next Generation Sequencing (MYcroarray, now part of Daicel Arbor Biosciences). The bait set contained nearly 20,000 unique 120 bp biotinylated RNA molecules in equimolar amounts. Most of the baits were designed to specifically bind to the FISNA-NACHT and B30.2 exons in the genome, but we also targeted other genes of interest. Bait hybridization and target enrichment for each pooled sample were performed according to the MYbaits manual version 3.02, with half the amount of baits and reagents used for the four-fish pools than for the eight-fish pools. Following an overnight incubation of the pooled DNA samples with RNA baits, bait-bound DNA fragments were extracted from the solution with DB MyOne Streptavidin C1 beads (Thermo Fisher Scientific). The enriched libraries were subsequently amplified with P5 and P7 primers (Illumina) by running 26 cycles of the program described in ***Appendix 1—table 2***. If the DNA yield was less than 1000 ng afterward (measured by Qubit), five more PCR cycles were added. Enrichment success was evaluated by qPCR, using 5× HOT FIREPol EvaGreen qPCR Mix Plus (ROX) (Solis BioDyne) and primers specific for the FISNA-NACHT exons from each of the four groups of NLRs ***Appendix 1—table 3 and 4***. The gene il1 was used as a single copy control. All primers were custom-ordered from biomers.net GmbH. The qPCR experiment was deemed successful if a strong enrichment could be seen for all NLR groups, weaker enrichment for il1, and no enrichment for the random intron. After this, the sample was selected for subsequent PacBio library construction and purified with 0.7× Ampure PB Beads (Pacific Biosciences).

Library construction and sequencing

The final libraries were prepared with the SMRTbell Template Prep Kit 1.0-SPv3 (Pacific Biosciences). At the ligation step, the recommended amount of PacBio adapters was increased from 1 to 5 µl per 40 µl total reaction volume and the reaction was incubated overnight at room temperature. For the SN and CHT libraries in pools of eight (see ***Appendix 1—table 1***), barcoded PacBio adapters were used instead of regular ones. The product codes for barcodes were BC1001 and BC1002 for CHT, BC 1003 and BC1004 for SN.

The first libraries (TU, CGN, 4 DP and 4 KG samples) were size selected to 2–8 kb with the BluePippin pulsed field electrophoresis system (Sage Science). The following libraries were size selected to 1.5–8 kb.

All sequencing was done at the Max Planck-Genome-Centre Cologne. All TU, CGN, DP, and KG zebrafish, as well as four CHT and four SN samples were sequenced with 1M v2 SMRT Cells of the Sequel instrument (Pacific Biosciences). The rest of the samples (all with barcoded adapters) were multiplexed together and sequenced with an 8M SMRT Cell of the much higher throughput Sequel 2 instrument (Pacific Biosciences). One of the already sequenced SN samples (SN24) was also resequenced in this run as it yielded no data in the first one. Furthermore, Pacific Biosciences upgraded their kits with a superior polymerase after we had sequenced TU, CGN and the first four samples of each wild population; all samples other than those were sequenced with their LR (long run) polymerase.

An overview of the sequencing is presented in ***Appendix 1—table 1***.

Read processing and assembly

Raw data were de-multiplexed and stripped of primer/adapter sequences with lima from Pacific Biosciences. For the samples sequenced with the PacBio Sequel I, the parameters –enforce-first-barcode –split-bam-named –W 100 were used with lima v1.0.0 for the runs without the LR polymerase. For Sequel runs with the LR polymerase, lima v1.8.0 and v1.9.0 were used with the same parameters. To remove PacBio barcodes from the data produced on Sequel II, lima v1.11.0 was used with parameters –split-bam-named –peek-guess and for the subsequent removal of NEBNext barcodes, the parameters were changed to –enforce-first-barcode –split-bam-named –peek-guess. Consensus sequences of all DNA fragments with a minimum of three passes (henceforth referred to as CCS reads) were calculated using ccs (v4.2.0, Pacific Biosciences) with default parameters. PCR duplicates were identified and flagged with pbmarkdup (v1.0.0, Pacific Biosciences) with default parameters, then excluded from downstream analyses. Any chimeric reads containing a primer sequence in the middle were identified with blastn (v2.11.0+) (***Altschul et al., 1990***) and removed. The filtered CCS reads were assembled into contigs for each fish separately using hifiasm (v0.15.4-r347) (***Cheng et al., 2021***) with default parameters.

NLR identification

To obtain a list of NLR gene positions in the reference genome, we first extracted known NLR locations from Ensembl. In addition, the reference genome was translated in all frames using transeq (from EMBOSS:6.6.0.0) (***Rice et al., 2000***) and searched for further NLRs using hmmsearch from hmmer (v3.2.1) (***Wheeler and Eddy, 2013***), without bias correction and with the hidden Markov model (HMM) profiles for zf_FISNA-NACHT and zf_B30.2 from ***Adrian-Kalchhauser et al., 2020***. Each position in which the zf_FISNA-NACHT model found a hit with a maximum i-Evalue of $1e-200$ and a minimum alignment length of 500 aa was considered a FISNA-NACHT exon. The filtering thresholds for B30.2 exons were an i-Evalue of $1e-5$ and a minimum alignment length of 150 aa. This approach was used both during bait design and as a preparatory step for the first round of read filtering.

To distinguish CCS reads of NLR genes from other CCS reads, the CCS reads of each fish were mapped against the reference genome GRCz11 using pbmm2 (v1.3.0) with preset ccs. CCS reads which mapped within a known NLR gene or one found with our HMM-based approach with any mapping quality were considered potential NLR reads and used as input for subsequent steps.

De novo assembled contigs containing NLR exon sequences were identified by translating all contigs of each fish in all frames with transeq (from EMBOSS:6.6.0.0) and subsequently searching for FISNA-NACHT and B30.2 domains using hmmsearch from hmmer (v3.2.1) without bias correction and the HMMs zf_FISNA-NACHT and zf_B30.2 again. The HMM-based approach was chosen for the contigs in particular because we assumed that there would be NLR sequences in the data which

are absent in the reference genome and therefore might not be mapped. The approach enabled us to include all FISNA-NACHT and B30.2 exon data found by ***Adrian-Kalchhauser et al., 2020*** in our searches, optimizing the search sensitivity.

By examining all NLRs annotated in the reference genome, we found a highly conserved 47 bp exon preceding B30.2 to be present in most NLR-B30.2 genes (NLRs containing a B30.2 domain), but not in other NLRs nor in most other B30.2-containing genes. B30.2 exons from NLRs were distinguished from B30.2 elsewhere in the genome by generating a HMM for the 47 bp exon based on the blast hits and searching the contigs for matches to this model with hmmsearch from hmmer (v3.2.1). The model was created with hmmbuild from hmmer (v3.2.1).

The FISNA-NACHT and B30.2 orthoclusters were postprocessed after get_homologues as follows: whenever an orthocluster contained more than one contig, a consensus sequence for the cluster was created from all those contigs with cons from EMBOSS:6.6.0.0. These consensus sequences and the contig sequences of the singleton clusters made up the representative sequences of the orthoclusters. Some representative sequences were reversed with revseq from EMBOSS:6.6.0.0 so that all exons were in the same orientation. The representative sequences were then blasted against each other using blastn (v2.12.0+) with default parameters and output format 6. In cases in which 98% and at least 3 kb of a representative sequence matched another with at least 98% identity, the two clusters they represented were fused into a new cluster by combining their contigs and generating a new consensus sequence from them. This process was conducted twice and reduced the number of FISNA-NACHT clusters from the initial 4743 to 2008 and B30.2 clusters from 14,879 to 2,635.

The bam files produced by mapping the NLR reads of each fish separately to the representative sequences of the orthoclusters were filtered using samtools (v1.7) (***Li et al., 2009***). If the representative sequence had at least one primary alignment (SAM flag 0 or 16) with length >1 kb, mapping quality 60, and no more than nine soft-clipping bases at both ends of the mapped read, the orthocluster was assumed to occur in the respective fish.

Circular genome plots were created with circos (v 0.69–8) (***Krzywinski et al., 2009***) running on Perl 5.036000. Principal component analysis of scaled NLR counts per individual was conducted and plotted with the R packages ade4 (v1.7-22) and adegraphics (v1.0–21) (***Thioulouse et al., 2018***).

Appendix 1—table 1. Sequencing scheme for the zebrafish samples.
Libraries sequenced after the introduction of an improved (long run) sequencing chemistry are marked with LR. Samples that yielded no data after sequencing are marked with asterisks.

Individuals	Library	Sequencer
TU01, TU02, TU03, TU06	TU L1	Sequel
TU08, TU10, TU12, TU14	TU L2	Sequel
CGN1, CGN2, CGN3, CGN4	CGN L1	Sequel
CGN5, CGN6, CGN7, CGN8	CGN L2	Sequel
DP07, DP09, DP10, DP12	DP L1	Sequel
DP15, DP20, DP23, DP24, DP25, DP28, DP31, DP34	DP L2	Sequel (LR)
DP03, DP05, DP13, DP16, DP21, DP29, DP31, DP33	DP L3	Sequel (LR)
KG35, KG41, KG42, KG43	KG L1	Sequel
KG03, KG05, KG07, KG12, KG14, KG15, KG18, KG19	KG L2	Sequel (LR)
KG20, KG22, KG24, KG26, KG29, KG32, KG33, KG44	KG L3	Sequel (LR)
SN21, SN23, (SN24*), SN26	SN L1	Sequel
SN03, SN04, SN08, SN09, SN10, SN11, SN12, SN24	SN L2	Sequel II (LR)
SN13, SN14, SN15, SN16, SN17, SN18, SN19, SN20	SN L3	Sequel II (LR)
CHT19, CHT23, CHT26, CHT28	CHT L1	Sequel
CHT01 - CHT07, (CHT13*)	CHT L2	Sequel II (LR)

Appendix 1—table 1 Continued on next page

Appendix 1—table 1 Continued

Individuals	Library	Sequencer
CHT08, CHT10 - CHT12, CHT14 - CHT16, (SN25*)	CHT L3	Sequel II (LR)

Appendix 1—table 2. PCR program used for barcoding.
For library amplification, the same program was used with 26 or 31 cycles.

Step	Temperature (°C)	Duration	
Initialization	98	4 min	
Denaturation	98	30 s	
Annealing	65	30 s	{x 12}
Elongation	72	12 min	
Final elongation	72	20 min	
Storage	4	∞	

Appendix 1—table 3. qPCR program for the evaluation of enrichment efficiency.

Step	Temperature (°C)	Duration	
Initialization	95	12 min	
Denaturation	95	15 s	
Annealing	65	20 s	{x 40}
Elongation	72	20 s	

Appendix 1—table 4. Sequences of qPCR primers used for evaluation of target enrichment.

Gene	Direction	Sequence
il1	+	5′-tgg-tga-acg-tca-tca-tcg-cc-3′
il1	-	5′-tcc-agc-acc-tct-ttt-tct-cca-a-3′
foxo6 intron	+	5′-agt-tct-gtg-tgg-gaa-cag-gg-3′
foxo6 intron	-	5′-gtg-cat-ctt-tag-cgt-tgg-ct-3′
NLR group 1	+	5′-cct-gac-aca-ggt-caa-caa-aac-a-3′
NLR group 1	-	5′-gat-tgt-ctt-ttc-ctt-cag-ccc-ag-3′
NLR group 2	+	5′-tgg-att-ggg-ctg-aag-gga-aa-3′
NLR group 2	-	5′-agg-ttc-agt-cct-tta-gtc-tct-gg-3′
NLR group 3	+	5′-ctg-ctg-gag-gtg-aaa-gat-cag-ac-3′
NLR group 3	-	5′-gat-tgt-tga-gca-gtg-agc-agg-a-3′
NLR group 4	+	5′-tac-ctg-gac-aag-aca-aag-cca-3′
NLR group 4	-	5′-ctc-ctt-ctc-ttc-agc-cca-gtc-3′

5.2 Striking Variability in the Odorant Receptor Repertoire of the Darkling Beetle *Carchares macer* within and between Populations

Authors: Katja Palitzsch, Sigrun Korsching, Rosa-Stella Mbulu, Reinhard Predel and Thomas Wiehe
Status: in preparation for submission

Authors contribution: The author of this thesis conributed to the manuscript *Striking Variability in the Odorant Receptor Repertoire of the Darkling Beetle Carchares macer within and between Populations* by development and adaptation of the extraction procedure for high molecular weight (HMW) genomic DNA from ethanol-stored Carchares macer beetles, optimizing protocols to ensure the integrity of genetic material for Illumina whole genome sequencing and performed the extraction of HMW gDNA from these beetle samples, which formed the basis for subsequent genomic studies. Furthermore she managed the downstream processing of Illumina whole-genome sequencing data, including the critical steps of read trimming, genome assembly, and comprehensive quality assessment of the genomic assemblies, curated, organized and validated the datasets for analysis. The author conducted comprehensive data analyses, applying bioinformatics approaches to annotate odorant receptor genes and assess their variability, presence–absence, and population-specific characteristics. Beyond experimental and analytical work, the author played a leading role in conceptualizing and writing the manuscript.

Striking Variability in the Odorant Receptor Repertoire of the Darkling Beetle *Carchares macer* within and between Populations

Katja Palitzsch, Sigrun Korsching, Rosa Stella Mbulu,
Reinhard Predel, Thomas Wiehe

September 10, 2025

Abstract

A key function that drives ecological adaptation and diversification is the ability to detect and discriminate chemical signals. The odorant receptor genes are a central component of the chemosensory system in insects. Beetles (Coleoptera) display exceptional OR diversity, with both broad and lineage-specific expansions reflecting differences in ecological niches. So far insect OR gene repertoires have been mostly characterized at the species level and potential intra-specific variation across populations remains largely unexplored. Here, we have addressed this question using the darkling beetle *Carchares macer*, which is endemic to the Namib Desert, where it is adapted to aridity. We generated and analyzed *de-novo* genome assemblies from multiple geographically separated *C. macer* populations. Genome assemblies are highly complete (96–99% BUSCOs), with nearly identical genome sizes of about 168 Mbp. Using a stringent bioinformatics pipeline, we annotated a total of 91 OR genes, including a single conserved ORCO gene and a large lineage-specific OR expansion of 68 ORs unique to *C. macer*. Other ORs clustered four of seven reference coleopteran OR subfamilies. Substantial variability in both the composition and number of OR genes was observed at the individual and population levels. Only two OR genes were present across all individuals, and half of all ORs were only present in less than 60% of samples. Several OR genes displayed population-specific distributions or were exclusive to single populations. The number of OR genes per individual ranged from 40 to 60, with significant differences among populations. An accumulation model suggested that the *C. macer* OR repertoire in total comprises a number of $\sim$ 120 OR genes. Many OR candidates were found in genomic clusters, consistent with tandem gene duplication as a key mechanism shaping OR diversity. These results reveal pronounced individual and population-level variation in the OR gene repertoire of *Carchares macer*, underlining highly dynamic evolution within this gene family. Our findings demonstrate that extraordinary chemosensory gene diversity and turnover can occur within a single beetle species, highlighting population differentiation as an important aspect of chemosensory evolution in insects.

1 Introduction

The ability to detect and discriminate chemical signals (chemo-sensation) is a fundamental function of living organisms, mediating behaviors such as host seeking, communication, mating, and avoidance (Gadenne et al., 2016; Fleischer et al., 2018). The first step in the signal transduction cascade leading to these behaviors universally involves binding of the chemical signal to a receptor. In arthropods, three major families of chemosensory receptors have been identified: ionotropic receptors (IRs), gustatory receptors (GRs), and odorant receptors (ORs), reviewed in Robertson (2019). IRs are related to mammalian ionotropic glutamate receptors (Croset et al., 2010), GRs represent an ancient and widely distributed receptor family across arthropods and other protostomes, while ORs are unique to insects (Croset et al., 2010; Brand et al., 2018; Yan et al., 2020; Wicher and Miazzi, 2021). ORs are related to GRs, and are considered part of an expanded GR superfamily (Wicher and Miazzi, 2021; Eyun et al., 2017).

Insect ORs likely originated approximately 440 million years ago in early insects (Zygentoma) (Brand et al., 2018; Yan et al., 2020). Despite the same name and a shared structure of seven-transmembrane domains (7TM) they are phylogenetically unrelated to vertebrate ORs, which are G protein-coupled receptors (Buck and Axel, 1991; Mombaerts, 1999). Insect ORs are ligand-gated ion channels and exhibit an inverted topology compared to vertebrate ORs, with an intracellular N-terminus and extracellular C-terminus (Benton et al., 2006; Butterwick et al., 2018). A highly conserved OR co-receptor (ORCO) (Brand et al., 2018) and a variable OR subunit form heteromeric OR/ORCO complexes, which mediate direct, odorant-gated ion exchange, allowing for rapid odor detection independently of second-messenger signaling pathways (Benton et al., 2006; Sato et al., 2008; Butterwick et al., 2018). The heteromeric channel architecture enables high sequence diversity and specialized odor tuning (Butterwick et al., 2018).

Insect ORs are highly divergent and rapidly evolving (Balart-García et al., 2024). The evolution of the insect OR gene family is driven by mechanisms such as gene duplication, sequence divergence together with functional modulation, pseudogenization, and gene loss. Comparative genomics therefore frequently reveals patterns of birth-and-death evolution, as well as significant expansions or contractions across different lineages (reviewed by Robertson (2019); Andersson et al. (2019); Balart-García et al. (2024)). Aside from the conserved ORCO gene, the number of OR genes varies greatly among species. For example, approximately 60 OR genes have been identified in Drosophila (Robertson et al., 2003), 80–130 in mosquitoes (Fox et al., 2001; Hill et al., 2002), 163 in honey bees (Robertson and Wanner, 2006) and around 300 - 500 in ants (Zhou et al., 2015; Cohanim et al., 2018; McKenzie and Kronauer, 2018). In contrast, only four ORs have been found in the dragonfly *Ladona fulva* (Brand et al., 2018; Yan et al., 2020). The genesis of the OR gene family parallels the evolution of insects into terrestrial environments which requires detection of volatile odorants (Brand et al., 2018; Yan et al., 2020; Wicher and Miazzi, 2021). The dynamic evolution within the OR family might reflect ongoing adaptations to different ecological niches (Benton, 2015). A correlation between the breadth of the ecological niche and/or host range and the size of the OR gene repertoires has been suggested, with specialists having fewer ORs than polyphagous species (Andersson et al., 2019; Balart-García et al., 2021, 2024; Cohanim et al., 2018).

Indeed, the polyphagous grain beetle *Tribolium castaneum* possesses a rather large repertoire of approximately 256 OR genes (Engsontia et al., 2008). *Tribolium* belongs to the large and diverse family of darkling beetles (Tenebrionidae). It is the only species from this family for which the OR repertoire has been investigated so far. All other known beetle OR repertoires are somewhat smaller, ranging from 32 to 182, and come from nine non-tenebrionid families (Mitchell et al., 2020). Taking all ten species together, nine monophyletic OR subfamilies (1, 2a, 2b, 3, 4, 5a, 5b, 6, 7) can be distinguished. Some are lost from some species, while others are massively expanded in other species (Mitchell et al., 2020).

This rapid evolution within a single insect order raises the question, to what extent differences of the OR repertoire might exist even within populations in a single species. To the best of our knowledge this question so far has not been addressed for insect ORs. Here we use *Carchares macer*, another tenebrionid, to approach this question. Like *Tribolium castaneum*, and many other species in this family, *Carchares macer* shares the ability to cope with extreme heat and drought (Cloudsley-Thompson and Chadwick, 1964). Although they are remarkably efficient at minimizing water loss, the specific mechanisms underlying this trait are not yet fully understood (Cloudsley-Thompson and Chadwick, 1964; Draney, 1993).

Carchares macer is endemic to the Namib Desert, where it lives in association with vegetation, feeds on detritus, and utilizes nocturnal humidity as its primary water source. In the present study, we annotated and analyzed the OR repertoiree in close to one hundred newly generated and de-novo assembled genomes of *Carchares macer*. We report that nearly half of the OR genes are found only in few individuals, showing an extreme variability in the OR repertoire size and composition exceeding all that has been reported previously for vertebrate ORs

2 Methods

2.1 Sampling

Carchares macer beetles were collected from five populations at the western Namib desert. The populations are distributed across a 160 km transect with an individual distance of 40 km. The collection was performed as part of a larger collaborative research effort (`https://sfb1211.uni-koeln.de/`) (for GPS coordinates of sampling locations see supplementary Table 1). All samples were stored at 95% Ethanol on -80°. 20 beetles per population were used for the subsequent steps.

2.2 DNA-Extraction and Sequencing

High molecular weight genomic DNA was extracted using Ethanol precipitation method (supplementary tables 2 and 3).
Population data were generated using an *Illumina* paired end sequencing approach with a read length of 150 bp and 20-fold coverage. Library preparation and sequencing were performed at the Cologne Center for Genomics (CCG).

2.3 Genome Assembly and Quality Assessment

To perform a comprehensive genome assembly and quality assessment, the following procedure was executed: First, adapters were removed from the sequencing data using `Trimmomatic` (Bolger et al., 2014). The trimmed paired-end data were used as input for the `ABySS` Genome assembler (Jackman et al., 2017). The quality and completeness of the assembled genomes were evaluated using `quast` (https://github.com/ablab/quast) and Benchmarking Universal Single-Copy Orthologs `BUSCO` with application of the insecta-odb10 library consisting of 1367 HMM profiles from insect single copy orthologs. (Manni et al., 2021) The genome of 1 individual (NT06-24) was excluded from the presence-absence analysis (chapter 3.4), as no ORCO gene could be detected. Given the essential role of ORCO, its absence is likely the result of a sequencing or assembly artifact, rendering this genome less reliable for comparative analysis.

2.4 Estimation of evolutionary population distances

To estimate the pairwise evolutionary distance of individuals and populations, all assembled genomes were processed with `andi` (Haubold et al., 2014) using verbose options (-v) and joining (-j). Based on the resulting distance matrix, a phylogenetic tree (neighbor joining tree) was generated with `iTol` (Letunic and Bork, 2021). The tree was visualized in `figtree` and graphically edited in `figtree` and `inkscape`.

2.5 Data Mining

A combined approach of `BLASTp` and HMM profile search was used to identify OR genes in the genome of *C. macer*.

Wokflow

- `BLASTp`, query: Collection of beetle ORs published in Engsontia et al. (2008) and Mitchell et al. (2020), (supplementary Table 4) filtered for length ($\geq$ 330AA) and presence of 6-7 TMHs

- filtering ($\geq$ 200AA, 4-8 TMHs) for candidates, add newly identified candidates to the query set ($\rightarrow$ advanced query set)

- recursive `BLASTp` searches were run with the advanced queryset and subsequent filtering($\geq$ 300AA, 5-8 TMHs). Newly identified ORs were added to the query set and the circle was repeated until no new results could be generated

- generation of HMM profile using the validated *C. macer* OR candidates and resolve ultra-long hits ($\geq$ 500AA, $\geq$ 8$TMHs$) via HMM search of the corresponding genomic region.

- clustering the OR candidates with the references from the initial OR query set and a collection of beetle gustatory receptor genes (supplementary Table 5) as outgroup.

For all `BLASTp` searches a pre-defined e-value cutoff of 1e-50 was applied. Higher e-values did not result in additional OR candidates.

TMH search Search for transmembrane helices (TMHs) in the identified OR candidates was done with `TMHMM 2.0` (`https://services.healthtech.dtu.dk/services/TMHMM-2.0/`) and `DeepTMHMM - 1.0`(`https://services.healthtech.dtu.dk/services/DeepTMHMM-1.0/`)

Multiple sequence alignments Multiple sequence alignments were calculated with the `MAFFT` v7.304b `E-INSI` algorithm (Katoh and Standley, 2013).

Phylogenetic trees Maximum likelihood pylogenetic trees were generated using the function `pml_bb` of the R package `phangorn 2.12.1` (Schliep, 2011). Trees where exported to Newick format and visualized and formated in `figtree` (`https://tree.bio.ed.ac.uk/software/figtree/`) and `inkscape` (`https://inkscape.org/`).

2.6 Estimation of the cumulative OR gene repertoire and comparison across populations

To estimate the total number of olfactory receptor (OR) genes within and across *C. macer* populations, an extrapolated accumulation analysis was performed. We looked at two scenarios: (1) within and (2) between subpopulations. In (1), we randomly chose an individual from a subpopulation and counted its OR genes. Then, we sequentially and randomly added further genomes from this subpopulation and added the number of newly found OR genes to the cumulative count. To estimate the variance of the total count, we repeated this random selection process 100 times for each subpopulation. In (2), we pooled all subpopulations and chose a random sample of 20 genomes from the pooled population. Counting was done as in (1). A nonlinear fit was applied using an accumulation function of the form:

$$y = \gamma + \alpha \cdot \text{Harmonic}(x, \beta),$$

where $\alpha, \beta, \gamma \geq 0$ and

$$\text{Harmonic}(x, \beta) = \sum_{k=1}^{x} \frac{1}{k^{\beta}}.$$

When $x \to \infty$, the function has a finite limit, if $\beta > 1$. We take this asymptotic value as estimate of the total number of OR genes in the respective population. In this case we call its "ORome" *closed*. When there is no finite limit ($\beta \leq 1$), the ORome is *open* – with terminology borrowed from pan-genome analysis (Tettelin and Medini, 2020).

2.7 Population-wise comparison of OR gene counts

The number of identified OR genes per individual was compared between populations. Gene counts were subject to one-way analysis of variance (ANOVA), followed by Tukey's Honest Significant Difference (HSD) test for pairwise comparisons. Statistical tests were carried out with R. The results were visualized with the R packages `dplyr`, `ggpubr` and `ggplot2`.

2.8 Identification of local clustering patterns

IDs of scaffolds, containing OR candidates were extracted from `AUGUSTUS` gff-output using `awk`. Data sorting and rearrangement were performed using the Python package `pandas`. Plots were generated with the R-package `pheatmap` and the Python package `matplotlib`. Figures were arranged and formatted with `inkscape`.

3 Results

3.1 Genome assembly metrics

To generate whole genome population data, we applied Illumina paired-end sequencing. This resulted in a total of 92 genome assemblies. The size of the largest scaffold varied between populations, ranging from approximately 2 to more than 7 Mbp (Fig 1, supplementary Figure 1). The corresponding N50 values ranged between 25 and 75 kb (Fig 1), indicating moderate variation in the contiguity of the assemblies. Total assembly lengths, serving as genome size estimates, consistently measured around 168 Mbp (supplementary Figure 1).

Assembly completeness was evaluated using *BUSCO* with the insect-specific reference database insecta-odb10, which comprises 1367 conserved single-copy orthologs from 75 insect genomes. The proportion of complete and single-copy BUSCOs detected per sample ranged from 96% to 99% (Fig 1, supplementary Figure 1), indicating high completeness across all assemblies.

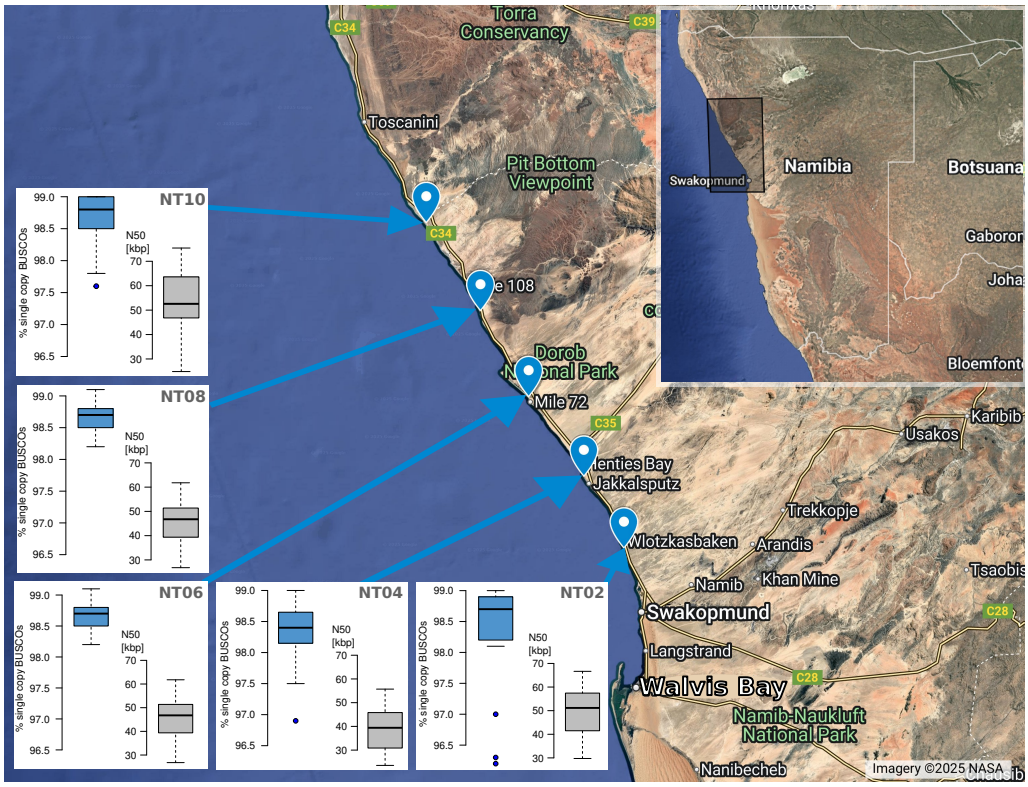


Figure 1: **Sampling locations and Illumina genome assembly statistics for populations NT02 to NT10.** Sampling locations (supplementary Table 1) for populations NT02–NT10 in Namibia are shown with associated genome assembly metrics. Blue boxes represent the percentage of single-copy BUSCOs, and grey boxes show N50 scaffold lengths (kbp) for all samples of the population, indicating high assembly completeness and variable contiguity. Sampling sites are approximately 40km apart. NT = Namibia Transect, The map was generated with Google My Maps

3.2 Clustering of *C. macer* genomes reflects geographic origin

We used the alignment-free clustering program `andi` (Haubold et al., 2014) to estimate evolutionary distance within and between subpopulations. Results were visualized as a Neighbor-Joining tree and show a clear signal of clustering which reflects the geographic origin of the sampled individuals (Fig 2). The deepest split separates NT02 – the southern most sample – from all other subpopulations. The youngest splits are between NT10 – the most northern population – and the NT08 population, followed by the split of NT04 from NT06. The observed tree structure is consistent with the idea of a south to north colonization, with (rare) migration between some of the neighboring subpopulations (between NT08 and NT10 or between NT04 and NT06).

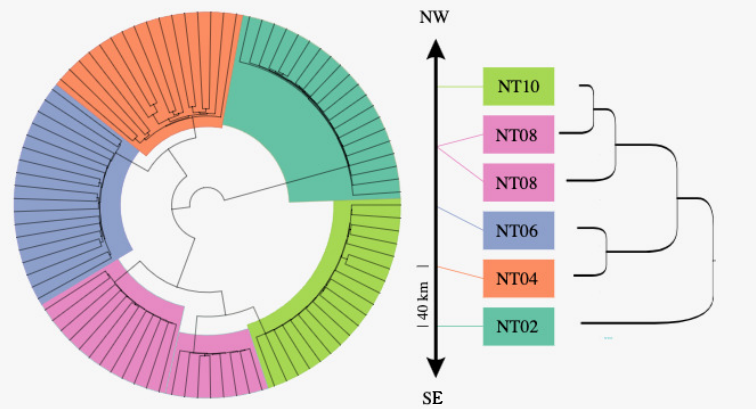


Figure 2: **Phylogeny of the five *Carchares macer* populations:** Based on evolutionary distances among all genomes from the five populations (NT02, NT04, NT06, NT08, NT10), applying *andi* (Haubold et al., 2014). The resulting pairwise distances were used to construct a neighbor-joining tree (clustering precision: CP = 0.9826). The geographic distance between adjacent populations is ∼40 km.

3.3 The *C. macer* OR repertoire consists of about hundred OR Genes

For *de novo* identification of olfactory receptor (OR) genes we employed iterative cycles of BLASTp searches combined with stringent filtering criteria. Only sequences exceeding 300 amino acids in length and exhibiting five to eight predicted transmembrane helices were retained. To validate the candidates, we performed clustering with known OR sequences (reference ORs) and applied a maximum likelihood approach that also ensured the separation from gustatory receptor (GR) genes of Coleoptera.

The BLASTp search was repeated recursively until no additional OR candidates could be detected. This pipeline yielded a total repertoire of 91 OR candidates. All identified OR candidates (in the following referred to as , CmacORs = *Carchares macer* - ORs) cluster with previously described coleopteran OR genes and are clearly separated from GR genes (bootstrap support = 0.97; Fig 3A). Among these, one candidate could be confidently identified as the olfactory co-receptor ORCO, based on its close phylogenetic relationship (bootstrap support = 1) to TcOR1, the ORCO gene of the tenebrionid beetle *Tribolium castaneum*. Our phylogenetic analysis indicates that the ORCO clade was the first to split from the GR outgroup and to give rise to all other OR lineages (Fig 3 A). In addition to ORCO, 23 CmacOR candidates clustered with perviously described coleopteran OR genes, while additional 68 sequences formed a distinct, lineage-specific expansion unique to *C. macer* (bootstrap support = 0.99) (Fig 3A). According to Mitchell et al. (2020), coleopteran OR genes can be classified into seven distinct subgroups. This subgroup structure is also reflected in our OR phylogeny, with the individual subgroups represented to varying extent within the *C. macer* OR repertoire. CmacOR candidates were not detected in the subgroups 3, 6, and 7. Subgroups 1, 4, and 5 include five, one, and four candidates, respectively. Subgroup 2 contains the highest number of assigned OR genes, including 81 candidates in total (Fig 3B,C). Of these, 68 are lineage-specific CmacORs that form a distinct clade, which is phylogenetically assigned to subgroup 2 with strong bootstrap support (0.97; Fig 3B,C).

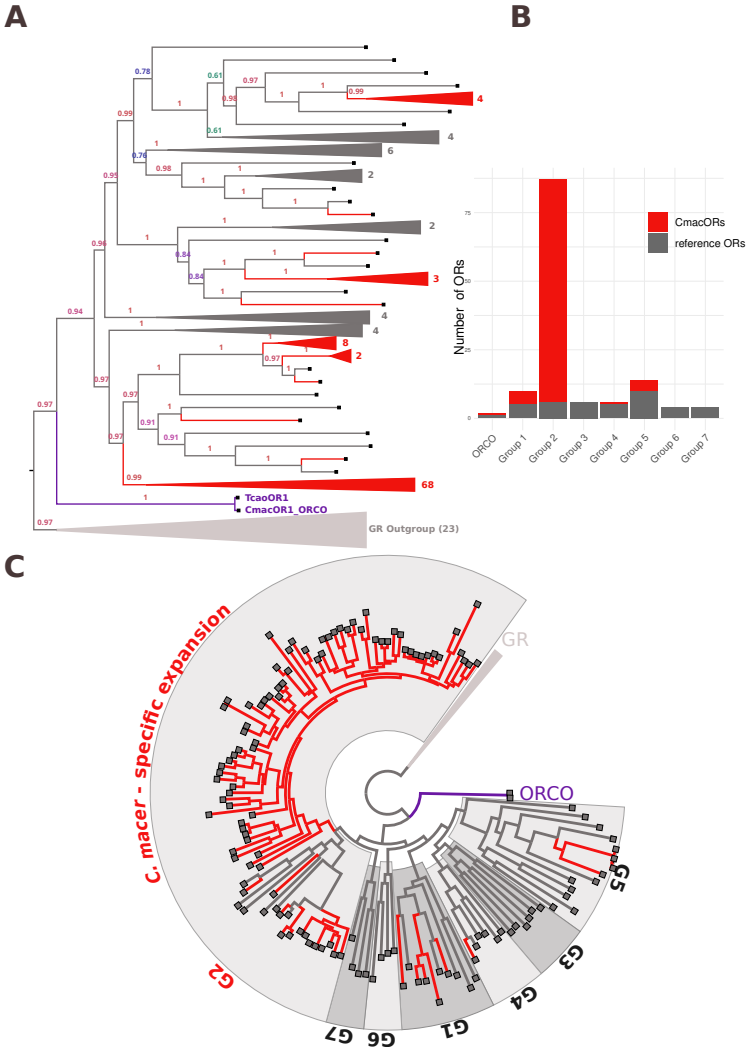


Figure 3: **Trans-species phylogeny of the odorant receptor repertoire (OR repertoire)** A) Maximum-likelihood tree including one representative sequence from each *C. macer* OR gene, derived from 92 individuals from five populations (shown in red). The *C. macer* ORCO candidate (CmacOR1-ORCO) clusters with the reference ORCO (TcOR1) from *Tribolium castaneum* (ORCOs shown in purple). Coleopteran odorant receptor reference genes (supplementary table 4), used as BLAST queries were included for comparison (dark gray). Coleopteran gustatory receptors (GRs; light gray; supplementary Table 5) were used as the outgroup to root the tree. Branch labels indicate bootstrap support values (based on 100 replicates). Tip labels show the number of ORs contained within collapsed branches. B) Stacked bar plots showing the number of CmacORs per subgroup (red upper bars) in comparison to the corresponding number of reference ORs (gray lower bars). C) Group assignment of identified *C. macer* OR genes (highlighted as red branches) to the seven coleopteran OR gene subgroups (G1 - G7) described by Mitchell et al. (2020) (see also supplementary Figure 2).

3.4 The OR gene repertoire of *C. macer* exhibits substantial individual variation

To investigate individual variation in OR gene repertoires, we generated a presence–absence matrix (PAM), indicating for each individual whether a given OR ortholog is present (1) or absent (0) (Fig 4A). The OR orthologs (Fig 4A, x-axis) are sorted by groups as depicted in Fig 3B. The composition of the OR gene repertoire varies considerably among

individuals. To additionally quantify the distribution and relative abundance of each OR genes in the dataset, we calculated an empirical cumulative distribution function (ECDF) illustrating the number of individuals possessing the gene (Fig 4B).

Observed values range from a minimum of 1 to a maximum of 91 individuals per gene, with a mean of 47.24. Half of the genes display counts of 50 or fewer (median = 50, Fig 4B). A small subset of 10 genes exhibits very low counts between one and three, while at the upper extreme, 27 genes reach high counts between 80 and 91 (Fig 4A,B). only two genes, including ORCO, are present in all 91 analyzed genomes. The ECDF shows a relatively even spread between the lower and upper quartiles, indicating substantial variation in gene representation. The interquartile range (IQR) can be visually estimated, with 25% of the data falling below approximately 25 counts per ortholog and 75% below about 75 counts. Overall, these results highlight the high diversity among individual OR repertoires.

Relating gene counts to group assignment, we found that, with the exception of CmacOR69 (group 5), which was present in only four individuals—representatives of groups 1, 4, and 5, with counts ranging from 26 to 90, occur at medium to high frequencies (Fig 4 A). This supports their conserved character, as previously demonstrated in Fig 3C. Group 2, which includes species-specific expansions, comprises both the rarest and most common representatives (Fig 4A). This pattern suggests that the expansion of group 2 began shortly after speciation and continues to the present.

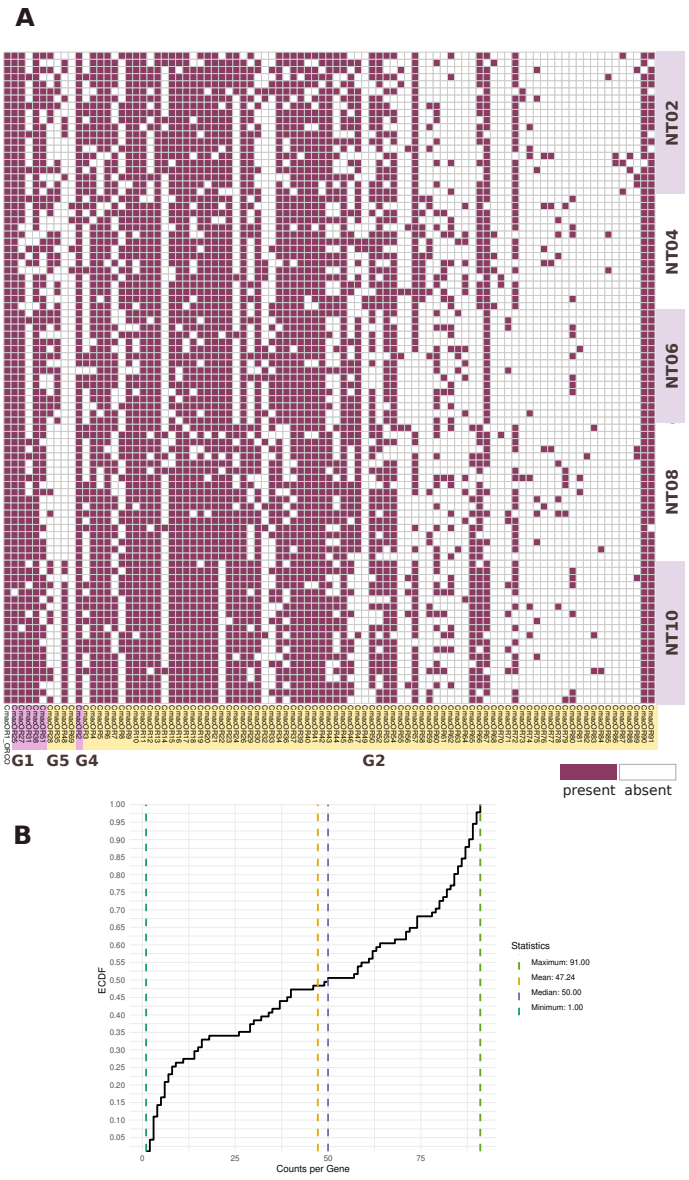


Figure 4: **A) Presence–absence matrix (PAM) of OR genes across 91 individuals from five populations (NT02–NT10).** The heatmap displays the distribution of identified OR genes across individuals, clustered by OR subgroups (G1 - G5) (Mitchell et al. (2020)). Filled cells indicate gene presence, while empty cells represent gene absence, illustrating the variability in OR gene repertoires among samples. B)Empirical cumulative distribution function (ECDF) of gene counts of gene counts. The black step curve illustrates the cumulative proportion of genes as a function of observed counts per gene. Vertical dashed lines represent descriptive statistics: minimum (green, 1), mean (orange, 47.24), median (blue, 50.00), and maximum (green, 91). The y-axis denotes the fraction of genes with counts below or equal to each threshold.

3.5 The distribution of OR genes on genomic scaffolds reveals patterns of local clustering

Chromosomal clustering of OR genes has been well documented in vertebrates (Niimura and Nei, 2005) and also reported in insect OR repertoires (Cohanim et al., 2018). However, the use of Illumina short-read sequencing limits the ability to generate chromosome-scale assemblies, thereby constraining the detection of local gene clusters. To identify

potential OR gene clustering, we examined the genomic scaffolds for recurring patterns in OR gene arrangement that might suggest tandem organization. For 32 of the 91 CmacOR genes, clustering with other OR genes on the same scaffold is observed in over 50% of their occurrences (Fig 5A, > 50% clustered appearance). Twenty out of 91 CmacOR genes appeared exclusively as single OR on a scaffold (Fig 5A, 0% clustered appearance). To evaluate the extent to which differences in clustered occurrence are associated with scaffold lengths, we statistically tested for a correlation between scaffold lengths and cluster sizes (Fig 5B). The Kruskal-Wallis test revealed a highly significant difference in cluster sizes among the twelve groups analyzed (test statistic = 120.65, df = 11, p = 1.34e-20).

This indicates that the distribution of cluster sizes varies systematically rather than randomly between groups. To examine whether these differences are related to the lengths of the underlying scaffolds, a Spearman rank correlation was additionally performed. The resulting correlation coefficient ($\rho = 0.171$, p = 1.97e-15) indicates a significant but weak positive correlation between scaffold length and cluster size (Fig 5B). Given the weak association between scaffold length and cluster size, it is necessary to examine scaffold lengths across the individual cluster groups. For instance, the scaffolds on which single OR genes are located range in length from 1.7 kbp to 434 kbp (Fig 5B). In comparison, the scaffold harboring the largest cluster (12 OR genes) is 134.5 kbp long, while the scaffold containing the second-largest cluster (11 OR genes) measures 71.5 kbp.

Overall, these observations support the conclusion that cluster size can be partially, but not entirely, explained by scaffold length. This indicates that at least some of the *C. macer* OR genes might be localized in clusters within the genome.

This is supported by the observation that the cluster composition reveals recurring patterns (Fig 5C). For instance, ORs 11, 12, 13, 15, 16, 21, and 91 frequently co-occur on the same scaffold. A similar pattern is observed for ORs 9, 10, 17, 18, 19, 20, 45, 46, and 47. Another prominent cluster comprises ORs 06, 29, 36, 40, 41, 42, 43, and 44 (Fig 5C). Together, these findings indicate that certain OR genes tend to be physically clustered within the genomes analyzed.

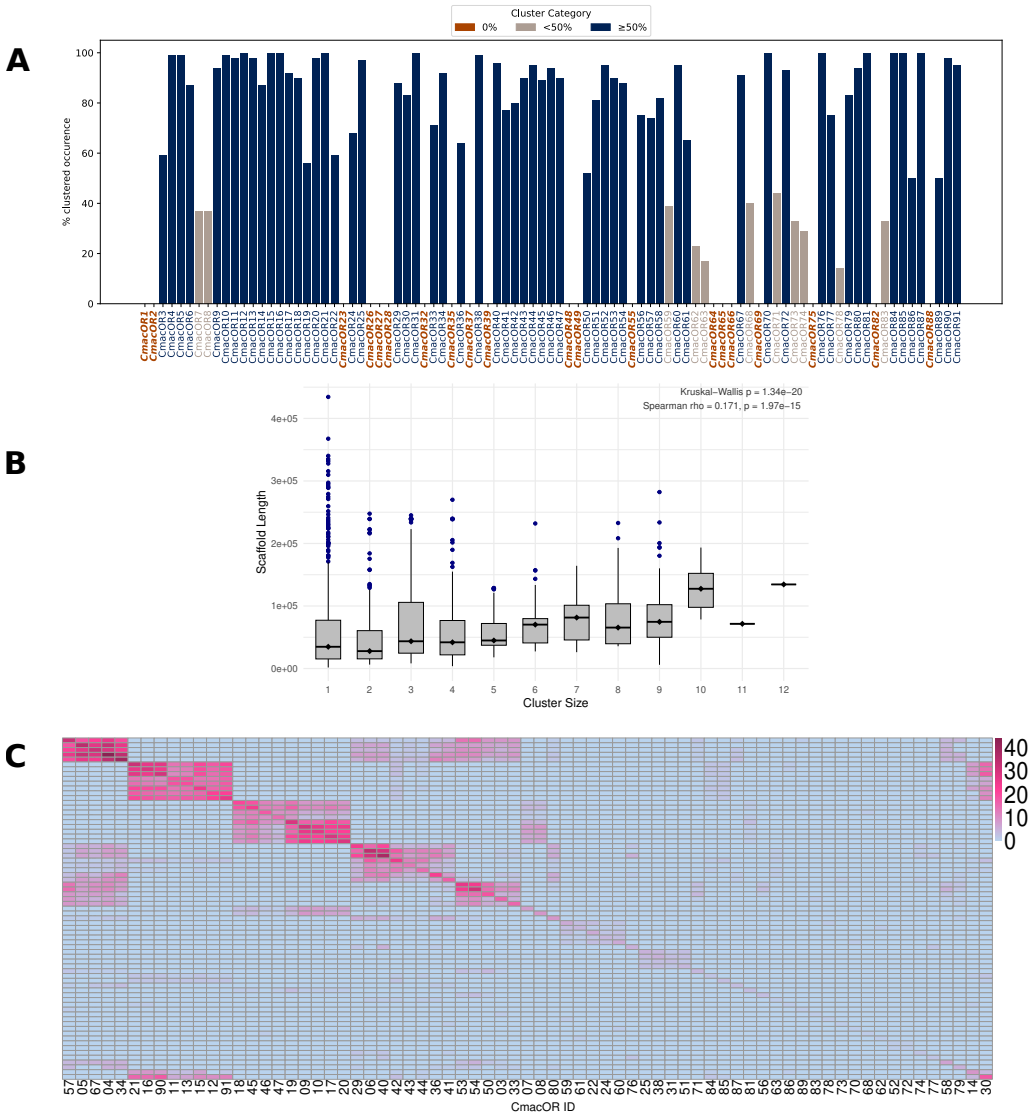


Figure 5: **Patterns of local clustering of OR genes across 91 *C. macer* samples.** A) OR-containing scaffolds were identified from BLASTp output files, and all OR genes were mapped to their respective scaffolds for each sample. For each OR gene, we determined whether it appeared exclusively in clusters (≥ 2 ORs per scaffold), exclusively as a singleton, or in both configurations across samples. The proportion of clustered occurrences was calculated for each gene across all samples. Blue bars and x-axis labels represent genes that occur in clusters in $> 50\%$ of samples. Grey bars correspond to genes with $< 50\%$ clustered occurrence. Orange labels indicate OR genes that were observed exclusively as singletons (0% clustering). B) Distribution of scaffold lengths across different cluster sizes. Black diamonds indicate the median scaffold length per cluster size. Statistical results from the Kruskal-Wallis test and Spearman correlation are displayed in the top-right corner, highlighting significant associations between cluster size and scaffold length. C) Recurrent clustering patterns were identified by aligning all OR clusters across samples and co-occurring OR genes are grouped accordingly. The x- and y-axes display the same ordered set of CmacOR gene identifiers. Legend = Number of joint occurrences.

3.6 OR gene repertoires differ in both number and composition across populations

In addition to analyzing the individual structure of the OR repertoires, we also investigated whether population-specific characteristics exist in the distribution and composi-

tion of OR repertoires. Therefore we first examined the frequency of individual OR genes across five populations and assessed whether certain genes represent population-specific candidates. We specifically calculated the relative abundance of each OR gene within the total OR repertoire of each population .

Several such population-specific candidates were identified. For example, OR genes 52, 66, and 72 are present in 40–60% of individuals from populations NT02, NT04, NT08, and NT10, but are entirely absent from population NT06 (Fig 6A). Another case of complete absence is CmacOR22, which is not found in any individuals of the NT10 population, whereas it occurs in 60–80% of individuals in the remaining populations (Fig 6A). An example of a gene exclusive to a single population is CmacOR58, which is found only in NT04 (around 80%) and is absent from all other populations (Fig 6A). We also identified genes with low frequencies in a single population and little to no presence in others, for instance, CmacORs 49, 68, 69, 70, 77, and 85 constitute approximately 20% of the OR repertoire in NT04 but are either sporadically present or entirely absent in the other populations (Fig 6A). Comparisons at both the individual and population levels reveal substantial variability in the composition of the OR repertoire. Moreover, our data demonstrate that each individual or population possesses only a subset of the species-wide OR repertoire. To assess the distribution of OR gene counts per individual across populations, we quantified the number of OR genes per beetle within each population and compared these values using analysis of variance (ANOVA), followed by Tukey's post hoc test.

The total number of OR genes per individual ranges from 40 to 60, with pairwise comparisons revealing significant differences between certain populations. The most pronounced difference in OR gene counts was observed between populations NT06 (mean = 45.12, median = 44.5) and NT10 (mean = 50.65, median = 50), with a highly significant p-value ($p = 1.3e\text{-}6$) (Figure 7C). A similarly significant difference was detected between NT02 (mean = 46.25, median = 47) and NT10 ($p = 5e\text{-}5$) (Fig 6C). In contrast, populations NT04 (mean = 49.06, median = 50) and NT08 (mean = 47.63, median = 48) showed no significant difference in OR gene counts ($p = 0.58$), nor did NT04 and NT10 ($p = 0.47$) (Fig 6C). Taken together, these results indicate that the number of OR genes per individual fluctuates around 50 but can vary significantly between different populations. Overall, these findings demonstrate that the high variability in the OR gene repertoire is not limited to the individual level but extends to the population level, allowing us to describe a distinct population-specific OR repertoire fingerprint. Given the described individual and population-specific variability, it is essential to consider to what extent the current dataset captures the full OR repertoire of *Carchares macer*, and how the number of detectable OR genes may be affected by sample size. To evaluate the relationship between sample size and the number of detected OR genes, we applied a nonlinear accumulation model to our data, where genomes are sequentially added. This approach allows us to estimate whether OR gene discovery approaches saturation or continues to increase with additional individuals. A random selection of 20 samples from all populations was included for comparison with the population-specific curves. The limiting value (lim), representing the predicted total number of distinct OR genes detectable with infinite sampling, varies across populations from 73.34 to 118.00 (Fig 6B). This range suggests a potential increase in the number of identifiable OR genes with expanded sampling, from 1.2-fold up to approximately 1.6-fold, as reflected by the corresponding β values (Fig 6B). The fitted curve based on randomly pooled individuals yields a limiting value

of 110.14 and a β of 1.383 which is higher than in most individual populations, with the exception of NT08 (Fig 6). These results indicate that sequencing additional genomes continues to reveal new OR genes, although the discovery rate decreases and gradually approaches a plateau. The extrapolation suggests that the total number of unique OR genes in *Carchares macer* may approach approximately 120 with exhaustive sampling. (Fig 6B).

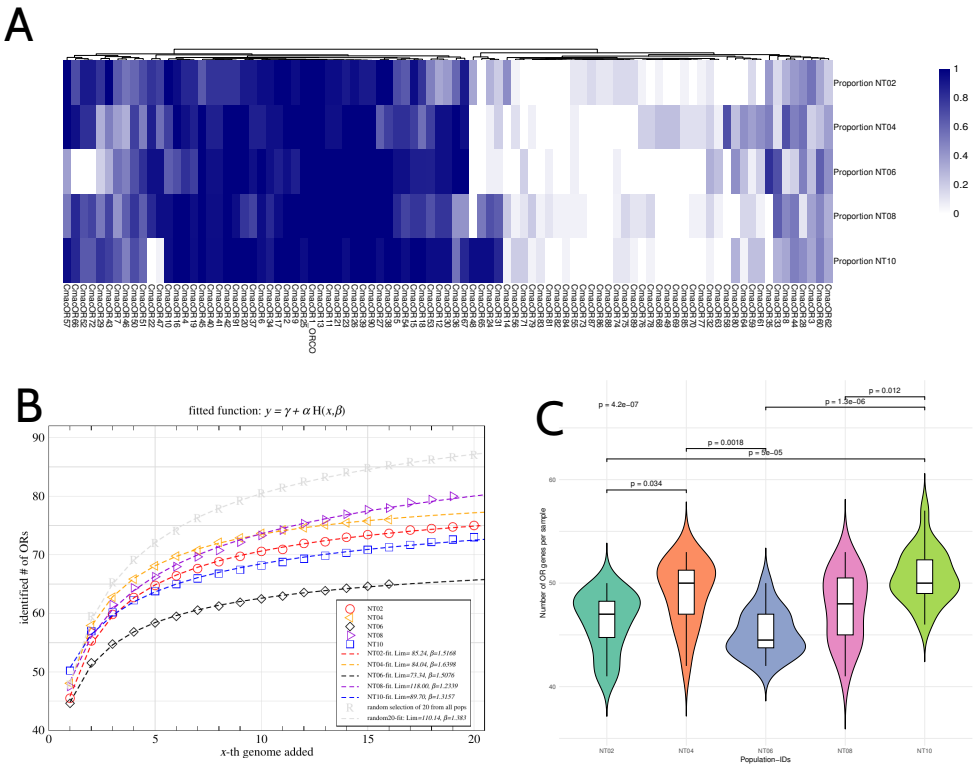


Figure 6: **Population-specific OR gene profiles across five *C. macer* populations. (A)** Relative abundance of single genes based on the OR gene matrix (OR repertoire) for each population. Gene counts were aggregated per population and normalized by the number of samples in the respective group. **(B)** Extrapolated cumulative number of identified ORs: with a sequential addition of genomes from the distinct populations NT02 to NT10 and a pan-population consisting of 20 randomly chosen samples. Nonlinear fits to an accumulation function ($y = \gamma + \alpha$ Harmonic(x, β)), where Harmonic$(x, \beta) = \sum_{k=1,\ldots,x} 1/k^{\beta}$. This function has a finite limit, if $\beta > 1$. It represents the estimated total number of OR genes in the respective populations. **(C)** Comparison of the numbers of identified OR genes per population. The counts of OR genes per sample were analyzed with ANOVA followed by a Tukey-test for pairwise significance. Significance lines and associated p-values are shown only for the pairs with significant differences. For additional p-values see supplementary Table 6.

4 Discussion

The present study provides a comprehensive analysis of the odorant receptor (OR) gene repertoire (OR repertoire) in the Namib desert beetle *Carchares macer*, revealing substantial individual and population-level variation, as well as evidence for extensive lineage-specific expansions.

A critical first step in this study was the generation of high quality short read genome assemblies for *C. macer* beetles, collected from five geographically separated populations. A high degree of completeness and coverage is essential given the highly dynamic nature of chemoreceptor gene families, which are known for rapid birth-and-death processes (Benton, 2015; Robertson, 2019). Reliable detection of orthologs and lineage-specific expansions depends on minimizing assembly gaps that could obscure gene content or misinterpret presence/absence variation. Therefore, the generation of high-quality assemblies was of particular importance. Our assembly statistics revealed consistently high completeness for 91 assemblies, with genome sizes of approximately 168 Mbp. BUSCO analyses using the insecta-odb10 ortholog database yielded scores of 96% to 99%, indicating that nearly all expected single-copy insect genes were identified in each assembly. N50 values between 25 and 75 kb, although indicative of the moderate scaffold contiguity characteristic of Illumina short-read sequencing, are consistent with those reported for published beetle genomes assembled exclusively from Illumina raw reads without long-read polishing ($\sim$ 50 - 10 kbp) (as summarized in Li et al. (2019)), and therefore can be considered appropriate to enable robust gene annotation of high variable gene families such as OR genes.

To analyze the phylogenetic divergence between the individuals, we applied an neighbor-joining approach inferred from genome-wide evolutionary distances (`andi`) and observed a pronounced clustering among the beetles *C. macer* from the five sampling sites, collected along a $\sim$ 160 km transect. This finding highlights substantial genetic structuring across the species' geographic range. It resolves both deep divergences, such as the earliest lineage split represented by the southernmost NT02 population, and more recent separations, as exemplified by the close phylogenetic relatedness and apparent derivation of NT10 from NT08. These patterns are consistent with a scenario of progressive northward expansion, possibly encouraged by historical environmental changes, habitat connectivity, or climatic events that facilitated range shifts. The degree of divergence between neighboring populations (e.g., NT04 and NT06) and the identification of NT10 as a likely subpopulation of NT08 raise intriguing questions about ongoing gene flow, adaptation to local environments, and the potential influence of physical or ecological barriers. Nevertheless, the five groups are clearly distinct from one another and can therefore be regarded as separate populations, which serve as the foundation for our main population-scale analysis of the odorant receptor repertoire in *Carchares macer*.

Within this framework of differentiated populations, maximum likelihood analyses support a conserved presence of the canonical insect odorant-receptor-co-receptor gene (ORCO), which shows strong homology and clear phylogenetic affinity to reference coleopteran ORCOs such as TcOR1 from *Tribolium castaneum*. This is consistent with the widely conserved, essential role of ORCO in insect olfaction and mirrors findings across beetles

and other insect orders (Benton et al., 2006; Sato et al., 2008; Wicher, 2015).

Beyond ORCO, the population-scale analysis of the *C. macer* odorant receptor (OR) gene repertoire reveals a dynamic pattern typical for rapid gene family evolution under the birth-and-death model, with high rates of gene gain, turnover, and loss (Nei and Rooney, 2005). Twentythree OR genes cluster with previously described coleopteran OR subgroups (Mitchell et al., 2020), suggesting a shared core of evolutionarily stable OR lineages. However, the large majority of OR genes reside in a lineage-specific expansion of 68 genes related to subgroup 2. This subgroup itself is strongly represented as well, whereas three other subgroups (3, 6, and 7) are completely absent, resulting in a markedly asymmetric partitioning of C. macer ORs into coleopteran subgroups. Such lineage-specific expansions, coupled with the loss of other OR branches, exemplify the birth-and-death process, where gene families are continuously shaped by episodic duplications and deletions, resulting in both stable and highly dynamic gene subgroups (Nei and Rooney, 2005; Eirín-López et al., 2012). These patterns align with evolutionary trajectories reported for the OR gene family in other insect species (Benton, 2015; Balart-García et al., 2024; Andersson et al., 2019; Robertson, 2019; Sánchez-Gracia et al., 2009; Zhou et al., 2015; Mitchell et al., 2020)
Lineage-specific expansions originating from a single OR sub-group are common features of insect OR gene repertoires. For instance, group 5 OR genes are highly expanded in *Tribolium castaneum* group 3 is expanded in both *Calosoma scrutator* and *Nicrophorus vespilloides*, group 1 in *Onthophagus taurus* and group 7 in *Dendroctonus ponderosae*. Expansions of group 2, as seen in *Carchares macer* also occur in *Priacma serrata* and *Agrilus planipennis* (Mitchell et al., 2020).
Unexpectedly, a very pronounced additional evolutionary fluidity was revealed by the sample-specific analysis of the OR gene repertoire. A striking individual variation was observed across the analyzed samples. The vast majority of OR genes are present at intermediate frequencies, and only two genes, including the conserved ORCO, are universally represented. This pattern again points to high diversity and suggests a flexible genetic architecture. Analysing the phylogenetic subgroups separately, we find that groups 1, 4, and 5 contain many genes at moderate to high frequencies, supporting their conservation and functional importance.
In contrast, group 2 displays a full spectrum of representation, from rare to nearly ubiquitous, consistent with an ongoing process of gene family expansion and turnover likely initiated after speciation.
Incidentally, both the complete representation of ORCO and the differences in representation between subgroups show that the short read sequencing approach used here does not lead to noticable undercounts of OR genes in the individual samples. ORCO, as necessary co-receptor for all OR genes, is expected to be present in all samples and was indeed found in all samples consistent with a loss rate below 1 percent. Similarly, undercounts should not distinguish between subgroups.

Population-level analysis further reveals distinct "OR repertoire fingerprints," with significant differences among populations in both overall OR gene numbers and the presence or absence of specific genes. Some genes are highly prevalent or even exclusive to particular populations while others appear scattered within or across populations. Statistically significant differences in the mean OR gene number per individual between

several population pairs suggest that processes such as genetic drift, local adaptation, or unique demographic histories might shape the composition and size of the OR gene repertoire. We used an accumulation analysis to infer the complete OR repertoire of *C. macer* from the observed repertoire of 91 genes. This extrapolation suggests the full *C. macer* OR repertoire may comprise up to 120 distinct genes. This amounts to about half of the repertoire size of *Tribolium castaneum* ($\sim$ 260 OR genes) (Engsontia et al., 2008; Mitchell et al., 2020). Previous studies have indicated a correlation between chemosensory gene content and host range in beetles (Andersson et al., 2019). Detailed functional studies, alongside investigations into the lifestyle and dietary preferences of *C. macer*, could provide valuable insights into the factors driving the observed differences between *C. macer* and *T. castaneum*.
Our analyses clearly show that each individual possesses only a subset of the species' entire OR gene repertoire. This finding emphasizes that odor detection and environmental interaction are highly individualized, and underscores the necessity of population-scale studies for a comprehensive understanding of chemoreception in insects. It will be interesting to see the extent of individual and population diversity in the large repertoire of *T. castaneum*. The variation between individuals and populations we observe in *C. macer*, suggests that current estimates for the OR repertoires of insect species might likely underestimate their true genetic diversity and evolutionary dynamics.

In addition to mapping the qualitative and quantitative distribution of OR genes, we tested for local genomic clustering. Indeed the genomic distribution of some OR genes in *Carchares macer* reveals distinct patterns of local clustering, suggesting that physical proximity of OR loci may play a role in gene family organization and evolution. Despite the limitations inherent in short-read sequencing and scaffold-based mapping, more than a third of the analyzed OR genes show a pronounced tendency to co-occur in clusters within genomic scaffolds, whereas a subset of 20 genes is consistently found as single ORs on one scaffold. Statistical analyses show that cluster size depends only weakly on scaffold length, implying that the observed clustering or lack thereof near accurately reflects the genomic organization. This is further supported by recurring patterns in cluster composition, such as repeated co-localization of specific OR genes in different samples. These observations parallel findings of OR gene clusters in vertebrates and other insects (Niimura and Nei, 2005; Cohanim et al., 2018). To validate and fully resolve the extent of local clustering, it is necessary to analyze a high-quality genome from *C. macer* generated using a long-read sequencing approach.

The variation in OR gene content among individuals and populations, described here for the first time, indicates that the chemosensory system of the desert beetle *Carchares macer* is highly dynamic. The prevalence of rare and population-specific ORs and significant differences in gene number likely enable rapid responses to changing environments, while a conserved core of ORs, including ORCO, highlights the constant necessity of certain chemosensory functions. The finding that only a few ORs are consistently present in all individuals, while most exhibit restricted distributions, suggests that olfactory individuality is common within the species.

Further studies are required to evaluate whether such individuality is a shared characteristic among tenebrionid beetles or other insect species. In combination with functional

studies, such as ligand assays or expression profiling, it will be possible to clarify how olfactory individuality relates to ecological and phylogenetic constraints, and how specific OR variants influence behavior and ecological adaptation.

In summary, the current study provides a robust and detailed characterization of the odorant receptor gene repertoire in the Namib desert beetle *Carchares macer*, combining high-quality genomic assemblies with population-scale comparative analyses. The findings reveal pronounced genetic structuring, extensive lineage-specific expansions, and considerable inter-individual diversity, illustrating the dynamic evolutionary landscape of chemoreceptor gene families in this desert-adapted species. These results emphasize the importance of comprehensive and population-level approaches for uncovering the complexity of gene family evolution and the mechanisms underlying sensory adaptation in extreme environments.

Acknowledgements: This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Projektnummer 268236062-SFB 1211 and by the Cologne Center for Genomics (CCG) where Illumina whole genome sequencing was carried out. We thank Álvaro Zúñiga Reinoso for his valuable expertise and support in conducting the beetle sampling.

References

Andersson, M. N., Keeling, C. I., and Mitchell, R. F. (2019). Genomic content of chemosensory genes correlates with host range in wood-boring beetles (Dendroctonus ponderosae, Agrilus planipennis, and Anoplophora glabripennis). *BMC Genomics*, 20(1):690–.

Balart-García, P., Bradford, T. M., Beasley-Hall, P. G., Polak, S., Cooper, S. J., and Fernández, R. (2024). Highly dynamic evolution of the chemosensory system driven by gene gain and loss across subterranean beetles. *Molecular Phylogenetics and Evolution*, 194:108027.

Balart-García, P., Cieslak, A., Escuer, P., Rozas, J., Ribera, I., and Fernández, R. (2021). Smelling in the dark: Phylogenomic insights into the chemosensory system of a subterranean beetle. *Molecular Ecology*, 30(11):2573–2590.

Benton, R. (2015). Multigene family evolution: Perspectives from insect chemoreceptors. *Trends in Ecology & Evolution*, 30(10):590–600.

Benton, R., Sachse, S., Michnick, S., and Vosshall, L. (2006). Atypical membrane topology and heteromeric function of Drosophila odorant receptors in vivo. *PLoS Biology*, 4:e20.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics (Oxford, England)*, 30:2114–20.

Brand, P., Robertson, H., Lin, W., Pothula, R., Klingeman, W., Jurat-Fuentes, J., and Johnson, B. (2018). The origin of the odorant receptor gene family in insects. *eLife Sciences*, 7.

Buck, L. and Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, 65:175–87.

Butterwick, J. A., del Mármol, J., Kim, K. H., Kahlson, M. A., Rogow, J. A., Walz, T., and Ruta, V. (2018). Cryo-EM structure of the insect olfactory receptor Orco. *Nature*, 560(7719):447–452.

Cloudsley-Thompson, J. L. and Chadwick, M. J. (1964). *Arid Lands, Their Flora and Fauna: Life in Deserts.* Dufour, Philadelphia.

Cohanim, A. B., Amsalem, E., Saad, R., Shoemaker, D., and Privman, E. (2018). Evolution of olfactory functions on the fire ant social chromosome. *Genome Biology and Evolution*, 10:2947–2960.

Croset, V., Cummins, S., and Benton, R. (2010). Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *Journal of Neurogenetics*, 24:30–31.

Draney, M. L. (1993). The subelytral cavity of desert tenebrionids. *The Florida Entomologist, vol. 76, no. 4, 1993, pp. 539–49. JSTOR, https://doi.org/10.2307/3495783*, 76(4):539–549.

Eirín-López, J. M., Rebordinos, L., Rooney, A. P., and Rozas, J. (2012). The birth-and-death evolution of multigene families revisited. *Genome dynamics*, 7:170–96.

Engsontia, P., Sanderson, A., Cobb, M., Walden, K., Robertson, H., and Brown, S. (2008). The red flour beetle's large nose: An expanded odorant receptor gene family in Tribolium castaneum. *Insect Biochemistry and Molecular Biology*, 38:387–97.

Eyun, S.-I., Soh, H. Y., Posavi, M., Munro, J. B., Hughes, D. S. T., Murali, S. C., Qu, J., Dugan, S., Lee, S. L., Chao, H., Dinh, H., Han, Y., Doddapaneni, H., Worley, K. C., Muzny, D. M., Park, E.-O., Silva, J. C., Gibbs, R. A., Richards, S., and Lee, C. E. (2017). Evolutionary history of chemosensory-related gene families across the arthropoda. *Molecular Biology and Evolution*, 34:1838–1862.

Fleischer, J., Pregitzer, P., Breer, H., and Krieger, J. (2018). Access to the odor world: olfactory receptors and their role for signal transduction in insects. *Cellular and Molecular Life Sciences*, 75(3):485–508.

Fox, A. N., Pitts, R. J., Robertson, H. M., Carlson, J. R., and Zwiebel, L. J. (2001). Candidate odorant receptors from the malaria vector mosquito Anopheles gambiae and evidence of down-regulation in response to blood feeding. *Proceedings of the National Academy of Sciences*, 98(25):14693–14697.

Gadenne, C., Barrozo, R. B., and Anton, S. (2016). Plasticity in insect olfaction: To smell or not to smell? *Annual Review of Entomology*, 61(Volume 61, 2016):317–333.

Haubold, B., Klötzl, F., and Pfaffelhuber, P. (2014). Andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, 31(8):1169–1175.

Hill, C. A., Fox, A. N., Pitts, R. J., Kent, L. B., Tan, P. L., Chrystal, M. A., Cravchik, A., Collins, F. H., Robertson, H. M., and Zwiebel, L. J. (2002). G protein-coupled receptors in Anopheles gambiae. *Science (New York, N.Y.)*, 298:176–8.

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., Jahesh, G., Khan, H., Coombe, L., Warren, R. L., and Birol, I. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a bloom filter. *Genome Research*, 27:768–777.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30:772–80.

Letunic, I. and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49(W1):W293–W296.

Li, F., Zhao, X., Li, M., He, K., Huang, C., Zhou, Y., Li, Z., and Walters, J. R. (2019). Insect genomes: progress and challenges. *Insect Molecular Biology*, 28(6):739–758.

Manni, M., Berkeley, M. R., Seppey, M., and Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols*, 1(12):e323.

McKenzie, S. K. and Kronauer, D. J. C. (2018). The genomic architecture and molecular evolution of ant odorant receptors. *Genome Research*, 28:1757–1765.

Mitchell, R. F., Schneider, T. M., Schwartz, A. M., Andersson, M. N., and McKenna, D. D. (2020). The diversity and evolution of odorant receptors in beetles (coleoptera). *Insect Molecular Biology*, 29:77–91.

Mombaerts, P. (1999). Seven-transmembrane proteins as odorant and chemosensory receptors. *Science*, 286(5440):707–711.

Nei, M. and Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*, 39(1):121–152. PMID: 16285855.

Niimura, Y. and Nei, M. (2005). Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. *Gene*, 346:13–21.

Robertson, H. M. (2019). Molecular evolution of the major arthropod chemoreceptor gene families. *Annual Review of Entomology*, 64(1):227–242. PMID: 30312552.

Robertson, H. M. and Wanner, K. W. (2006). The chemoreceptor superfamily in the honey bee, Apis mellifera: expansion of the odorant, but not gustatory, receptor family. *Genome Research*, 16:1395–403.

Robertson, H. M., Warr, C. G., and Carlson, J. R. (2003). Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 100(suppl_2):14537–14542.

Sánchez-Gracia, A., Vieira, F. G., and Rozas, J. (2009). Molecular evolution of the major chemosensory gene families in insects. *Heredity*, 103(3):208–216.

Sato, K., Pellegrino, M., Nakagawa, T., Nakagawa, T., Vosshall, L. B., and Touhara, K. (2008). Insect olfactory receptors are heteromeric ligand-gated ion channels. *Nature*, 452(7190):1002–1006.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics (Oxford, England)*, 27:592–3.

Tettelin, H. and Medini, D. (2020). *The pangenome: Diversity, dynamics and evolution of genomes*. Springer Nature.

Wicher, D. (2015). Olfactory signaling in insects. *Progress in Molecular Biology and Translational Science*, 130:37–54.

Wicher, D. and Miazzi, F. (2021). Functional properties of insect olfactory receptors: ionotropic receptors and odorant receptors. *Cell and Tissue Research*, 383:7–19.

Yan, H., Jafari, S., Pask, G., Zhou, X., Reinberg, D., and Desplan, C. (2020). Evolution, developmental expression and function of odorant receptors in insects. *The Journal of Experimental Biology*, 223:jeb208215.

Zhou, X., Rokas, A., Berger, S. L., Liebig, J., Ray, A., and Zwiebel, L. J. (2015). Chemoreceptor evolution in hymenoptera and its implications for the evolution of eusociality. *Genome Biology and Evolution*, 7:2407–16.

Supplementary Material

GPS-Coordinates of the Sampling Locations in Namibia

Table 1: Table S5: GPS coordinates for sampling locations NT02–NT10.

Location	GPSy (Latitude)	GPSx (Longitude)
NT02	-22.4398219585419	14.4546589907259
NT04	-22.1658029593527	14.2904379777610
NT06	-21.8661510106176	14.0655640047044
NT08	-21.5415410231799	13.8683160301298
NT10	-21.2080950010568	13.6482220049948

DNA Extraction Protocol

Table 2: Buffers, solutions, and reagents used in the DNA extraction protocol.

Type	Composition
Lysis buffer (pH 8.0)	10 mM Tris-HCl; 400 mM NaCl; 100 mM EDTA
Buffer for Proteinase K solution (pH 8.0)	1% SDS; 4 mM EDTA
Additional reagents	10% SDS; 5 M NaCl; 99.8% Ethanol; RNAse A; 10 mM Tris-HCl (pH 8.4)

Table 3: Step-by-step DNA extraction procedure, including quality control.

Step	**Description**
Sample preparation and lysis	Add 80 μl SDS (10%) to 1200 μl lysis buffer; add 20 μl Proteinase K to 180 μl Proteinase K solution immediately before use; remove elytra and open beetle; rinse in 99.8% ethanol and air-dry 2–3 min; submerge in lysis-buffer-SDS mixture and shock-freeze in liquid N_2; crush with pestle, add remaining lysis buffer; add Proteinase K mixture, incubate overnight at 56°C.
RNAse A treatment	Centrifuge overnight lysed sample at 4500 rpm for 10 min; transfer supernatant to fresh tube; add 5 μl RNAse A, incubate 30 min at RT; add 10 μl Proteinase K, incubate 30 min at 56°C; split sample equally into two 2 ml tubes.
DNA precipitation and elution	Add 240 μl 5 M NaCl to each tube, mix; add 1.2 ml ice-cold 99.8% EtOH, mix; incubate overnight at -20°C; centrifuge 13,000 rpm 15 min; wash pellet 3× with 70% EtOH, centrifuge each 15 min; air-dry pellet, dissolve in 50 μl 10 mM Tris-HCl pH 8.4; dissolve overnight at 4°C.
Reference and QC	Protocol: *10x Genomics® Sample Preparation Demonstrated Protocol • Rev A* `https://assets.ctfassets.net/an68im79xiti/3oGwQ5kl6UyCocGgmoWQie/768ae48be4f99b1f984e21e409e801fd/CG000145_SamplePrepDemonstratedProtocol_-DNAExtractionSingleInsects.pdf`; DNA quality assessed with Qubit 3.0, 0.8% agarose gels, and Nanodrop spectrophotometer.

Initial beetle OR query set and gustatory receptors outgroup

Table 4: A collection of beetle OR genes, from the OR repertoires published in Engsontia et al. (2008) and Mitchell et al. (2020) was used as initial query set for the initial BLASTp search and as references for the clustering with the newly identified OR genes of *C. macer* in the ML tree. The reference ORs represent a comprehensive selection from all OR subgroups (Mitchell et al., 2020) that survived a filtering for a sequence length of ($\geq$ 330 AA) and the presence of 6–7 TMHs.

Species	**Family**	**Reference OR ID**
Anoplophora glabripennis	Cerambycidae	AglaOR100, AglaOR116, AglaOR127
Agrilus planipennis	Buprestidae	AplaOR22CTE, AplaOR44NTE
Calosoma scrutator	Carabidae	CscrOR31
Dendroctonus ponderosae	Curculionidae	DponOR22, DponOR45FIX, DponOR54FIX
Leptinotarsa decemlineata	Chrysomelidae	LdecOR19, LdecOR55, LdecOR73
Nicrophorus vespilloides	Silphidae	NvesOR13, NvesOR46NTE, NvesOR47, NvesOR51
Onthophagus taurus	Scarabaeidae	OtauOR172, OtauOR185, OtauOR196, OtauOR203, OtauOR36, OtauOR51, OtauOR56
Priacma serrata	Cupedidae	PserOR121NTE, PserOR14
Tribolium castaneum	Tenebrionidae	TcOR108, TcOR128, TcOR153, TcOR167, TcOR195, TcOR263, TcOR275FIX, TcOR295, TcOR316, TcOR325, TcOR339, TcOR36, TcOR58, TcOR72, TcOR90, TcOr1

Table 5: Accession numbers and gene IDs and species names for gustatory receptors used as outgroup

Accession Number	Gene ID	Species
AKC58579.1	gustatory receptor 2	Anomala corpulenta
APC94341.1	gustatory receptor 14	Pyrrhalta aenescens
EEZ97769.2	gustatory receptor 144	Tribolium castaneum
EEZ99384.1	gustatory receptor 87	Tribolium castaneum
EEZ99385.1	gustatory receptor 88	Tribolium castaneum
EFA04712.1	gustatory receptor 6	Tribolium castaneum
EFA07615.1	gustatory receptor 118	Tribolium castaneum
EFA07633.1	gustatory receptor 155	Tribolium castaneum
KAI4457571.1	invertebrate gustatory receptor	Holotrichia oblita
KAI4466609.1	invertebrate gustatory receptor	Holotrichia oblita
KAK9710413.1	7tm Chemosensory receptor	Popillia japonica
NP_001137601.1	gustatory receptor	Tribolium castaneum
QBB73005.1	gustatory receptor	Protaetia brevitarsis
QBB73007.1	gustatory receptor	Protaetia brevitarsis
QBB73009.1	gustatory receptor	Protaetia brevitarsis
RZC39789.1	gustatory receptor for sugar taste 64a	Asbolus verrucosus
WKF45111.1	gustatory receptor 3	Podabrus annulatus
WKF45114.1	gustatory receptor 6	Podabrus annulatus
WKF45117.1	gustatory receptor 9	Podabrus annulatus
XP_008194199.3	gustatory receptor for sugar taste 64a	Tribolium castaneum
XP_022905524.1	gustatory receptor 68a-like	Onthophagus taurus
XP_022910663.1	gustatory and pheromone receptor 39a-like	Onthophagus taurus
XP_022912957.1	gustatory receptor for bitter taste 66a-like	Onthophagus taurus
XP_025832380.1	gustatory receptor for sugar taste 64e-like	Agrilus planipennis
XP_044254899.1	gustatory receptor for sugar taste 64e-like	Tribolium madens
XP_044256029.1	gustatory and pheromone receptor 39a-like	Tribolium madens
XP_044766785.1	gustatory receptor 68a-like	Coccinella septempunctata
XP_050518348.1	gustatory receptor 5a for trehalose	Diabrotica virgifera virgifera
XP_063930601.1	gustatory and pheromone receptor 32a-like	Zophobas morio
XP_065163157.1	gustatory receptor for sugar taste 64e-like	Dalotia coriaria
XP_065173417.1	gustatory receptor 68a-like	Dalotia coriaria

Assembly Quality - Scaffold Lengths

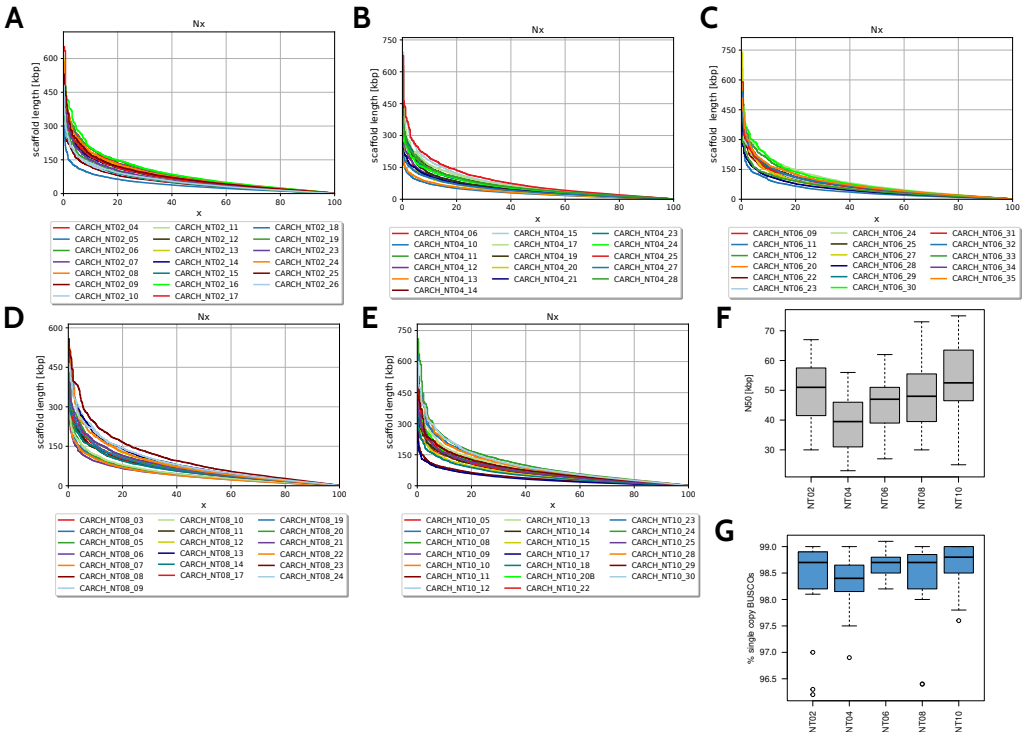


Figure 1: Assembly statistics for five populations (NT02, NT04, NT06, NT08, and NT10): (A-E) Scaffold length distributions sorted by population A) NT02, B) NT04, C) NT06, D) NT08 and E) NT10. (F+G) assembly quality metrics including N50 values (F) and the percentage of single-copy BUSCOs (G).

C. macer ORs Group Assignment

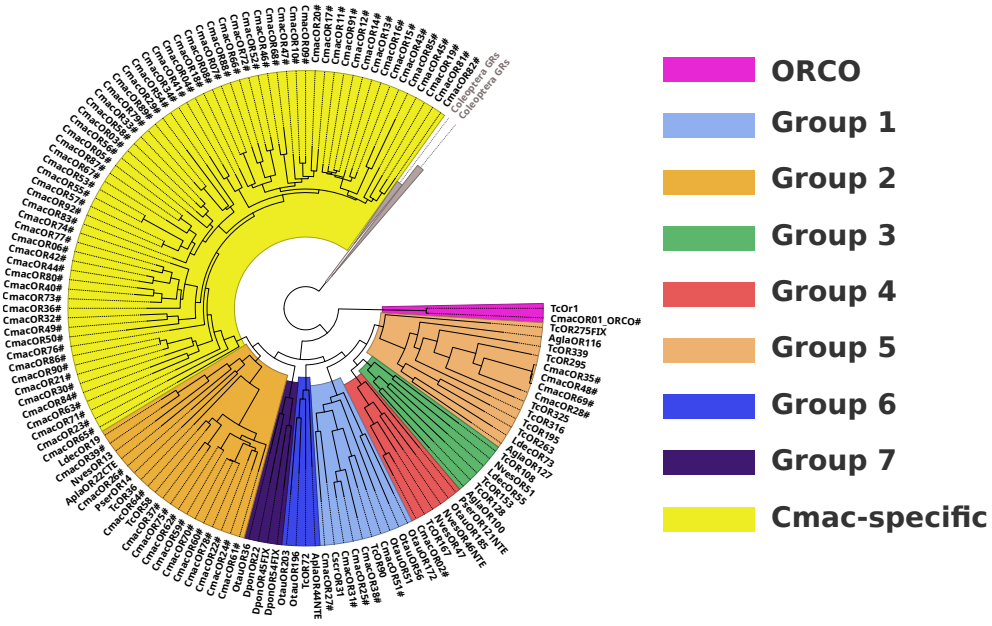


Figure 2: **Maximum likelyhood tree of the *C. macer* OR repertoire**: *C. macer* OR = CmacOR, labeled with '' were clustered with coleopteran reference ORs (supplementary Table 4). Colepteran gustatory receptors (GRs, supplementary Table 5) were used as outgroup. Group assignments according to Mitchell et al. (2020) are highlighted. ORCO = Odorant Receptor Co-Receptor.

Pairwise Population Comparison of OR gene counts per Individual

Table 6: **Pairwise population comparison of the numbers of OR genes per sample** The numbers of OR genes per sample in the populations NT02–NT10 were analyzed with ANOVA followed by a Tukey-test. The table shows the results of the Tukey-test for all pairwise comparisons.

Comparison	diff	lwr	upr	p adj
NT04–NT02	2.812500	0.1395572	5.485443	0.0341234
NT06–NT02	1.125000	-3.7979428	1.547943	0.7667774
NT08–NT02	1.381579	-1.1714393	3.934597	0.5602305
NT10–NT02	4.400000	1.8799253	6.920075	0.0000496
NT06–NT04	-3.937500	-6.7550291	-1.119971	0.0017773
NT08–NT04	-1.430921	-4.1349457	1.273104	0.5816462
NT10–NT04	-1.587500	-4.260443	1.0854428	0.4670871
NT08–NT06	2.506579	-0.1974457	5.210604	0.0825520
NT10–NT06	5.525000	2.8520572	8.197943	0.0000013
NT10–NT08	3.018421	0.4654028	5.571439	0.0121725

4.1

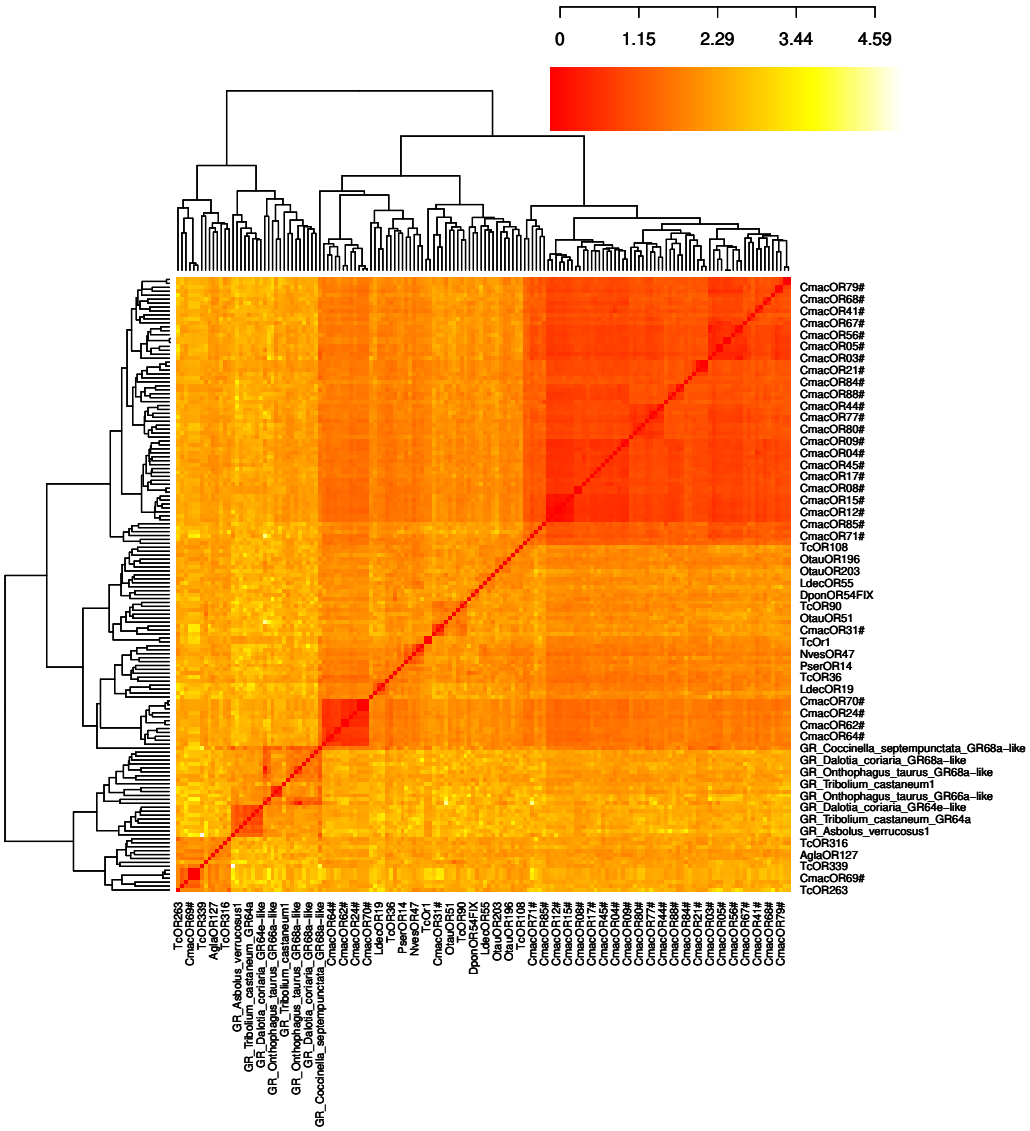


Figure 3: Pairwise distance matrix generated with a maximum-likelihood approach including one representative sequence from each *C. macer* OR gene, derived from 92 individuals from five populations. Coleopteran odorant receptor reference genes (supplementary table 4), used as BLAST queries were included for comparison. Coleopteran gustatory receptors (GRs; supplementary Table 5) were used as the outgroup. The matrix depicts evolutionary distances between gene sequences, with values representing the estimated number of substitutions per site according to the best-fitting amino acid model (LG+G(4)). Receptor gene IDs are indicated on both axes. The color scale ranges from zero (identical sequences) to 4.59 substitutions per site, highlighting the spectrum of sequence divergence within and between receptor families

5.3 Decay of the CTCF paralog BORIS in neognathous birds

Authors: Katja Palitzsch, Thomas Wiehe and Peter Heger

Status: Preprint (doi: https://doi.org/10.1101/2025.02.12.637905), in preprations for submission

Authors contribution: The substantial contributions of the author of this thesis to the manuscript "Decay of the CTCF paralog BORIS in neognathous birds" include the key laboratory work for validation of the BORIS status in Gallus gallus via PCR and Sequencing. Furthermore, the author performed the comparative data analysis to investigate the status of BORIS among 59 bird species applying bioinformatical approaches such as HMM-profile search and sequence curation. She furthermore conducted dN/dS analyses to assess evolutionary constraints, contributed to the annotation and analysis of repetitive elements and performed the analysis for correlation of sperm morphological data with the extent of BORIS gene degradation. Finally, the author played a leading role in conceptualizing, writing and editing the manuscript.

Decay of the CTCF paralog BORIS in neognathous birds

Katja Palitzsch[1], Thomas Wiehe[1*], Peter Heger[1,2*]

[1]Institute for Genetics, University of Cologne, Zülpicher Straße 47a, Cologne, 50674, NRW, Germany.
[2]Regional Computing Centre (RRZK/ITCC), University of Cologne, Weyertal 121, Cologne, 50931, NRW, Germany.

*Corresponding author(s). E-mail(s): twiehe@uni-koeln.de;
peter.heger@uni-koeln.de;
Contributing authors: k.palitzsch@uni-koeln.de;

Abstract

BORIS (brother of the regulator of imprinted sites), the paralog of the genome organizer CTCF, originated at least 318 million years ago (Mya), in the ancestor of amniotes (mammals, reptiles, and birds). Based on results from chicken (*Gallus gallus*), the gene was thought to be absent from birds. Using comparative genomics of 59 bird species, we show that birds possess BORIS, but frequently experience severe degradation of the gene, as observed in *Gallus gallus*. The degradation events are restricted to neognathous birds, specific for the BORIS coding sequence, and occur multiple times independently on different branches. They comprise a wide range of molecular decay, from individual point mutations to the inactivation and/or loss of particular zinc fingers, to the almost complete disintegration of the gene. The decay is accompanied by relaxed evolutionary constraints on BORIS codons across neognathous birds and coincides with the accumulation of species-specific repetitive elements in degenerate loci. BORIS represents a case of a presently ongoing, convergent, and specific gene loss within a lineage. As possible explanation, we propose a link between the loss of BORIS and a shift in sperm and/or genital morphology during the evolution of Neognathae.

Keywords: CTCFL/BORIS, pseudogenization, Neognathae/Paleognathae, evolution, reproduction

Significance Statement

The gene BORIS was believed to be absent from birds. However, genome analysis of 59 bird species reveals its presence, though it often underwent severe degradation in neognathous birds. These independent degradation events affect the BORIS coding sequence and range from point mutations to near complete gene disintegration. This decay correlates with relaxed evolutionary constraints and species-specific accumulation of repetitive elements. A potential link between BORIS loss and changes in sperm or genital morphology during neognathous bird evolution is suggested.

Introduction

Isolated as a protein that binds to three regularly spaced CCCTC repeats upstream of the transcription start site, the CCCTC-binding factor CTCF was first described as transcriptional repressor of the c-*myc* oncogene in chicken [1]. In the meantime, it has become clear that CTCF is a highly conserved gene with very important roles in animal biology: As ubiquitously expressed DNA-binding protein with eleven zinc fingers (ZFs), CTCF is a key factor in 3D chromatin organization, genome partitioning, and gene regulation [2–6]. It operates through its unique ability to establish independent domains of gene expression known as topologically associating domains (TADs) [7, 8], a hallmark of cell-type specific gene expression and organismal development [9–14]. CTCF emerged in the ancestor of bilaterian animals, 540 million years ago. It is present in most bilaterians, such as insects, molluscs, and vertebrates, but absent from non-bilaterian animals and other eukaryotes [15, 16].

While CTCF is a single-copy gene in most bilaterians, including protostomes and non-vertebrate deuterostomes [16], some species and/or lineages experienced CTCF duplication. Prominent duplication events have been postulated in the ancestor of amniotes (mammals, reptiles, birds) and during early vertebrate evolution when CTCF gave rise to its paralog CTCFL (CTCF-like protein) or BORIS (brother of the regulator of imprinted sites) [17, 18]. BORIS is located within a synteny block highly conserved in amniotes. In humans, this synteny block is situated on chromosome 20 and spans seven genes, SPO11 (Meiotic recombination protein SPO11), RAE1

(RNA export 1 homolog), RBM38 (RNA binding motif protein 38), CTCFL (BORIS), PCK1 (Phosphoenolpyruvate carboxykinase 1), ZBP1 (Z-DNA binding protein 1), and PMEPA1 (Prostate transmembrane protein, androgen induced 1) [19]. The direct neighbours of BORIS, RBM38 and PCK1, are ancient genes and regulate gene expression at the post-transcriptional level (RBM38: [20, 21]) or perform basic metabolic tasks (PCK1: [22, 23]). In contrast, CTCF's neighbour genes, RIPOR1 (RHO family interacting cell polarization regulator 1) and CARMIL2 (capping protein regulator and myosin 1 linker 2), exhibit highly dissimilar sequence signatures and molecular functions [24, 25], suggesting that CTCF duplicated and transposed to its new genomic location as an isolated gene.

As a result of their evolutionary history, CTCF and BORIS share a high degree of homology in their central zinc finger region. This part of the protein is characterized by 74 % amino acid identity, similar DNA binding properties, and a conserved genomic structure [17, 26–28]. In both paralogs, the ZF region is composed of seven exons, and each of these carries information for one and a half, or two (exons 3, 7, 8), zinc fingers [17, 26]. Despite these similarities, BORIS and CTCF differ substantially in their *N*- and *C*-terminal domains [26, 29]. As a consequence, the two proteins interact with different partner proteins and have distinct developmental functions and expression patterns [17, 26, 30, 31]. While CTCF is ubiquitously expressed from early development on [32–37], BORIS expression in mammals seems to be restricted to germline cells [17, 26, 38] where it regulates genes with a role in spermatogenesis and spermatid differentiation [39–41]. Due to its function as a transcriptional regulator of critical genes, amplification of the BORIS locus and/or misexpression of BORIS target genes are involved in the etiology of a number of cancers. According to its testis-specific expression and the irregular expression in cancer tissues, BORIS is classified as a member of the cancer-testis (CT) genes [for review, see: 42, 43].

With the growing amount of genomic data from a wide variety of organisms throughout all kingdoms of life [for review, see 44], evidence has accumulated that gene loss is a powerful evolutionary force, similar to evolution by gene duplication [45]. There are numerous examples of gene loss throughout the animal kingdom, including

the regression of vision and pigmentation in Mexican cavefish [46–48], the loss of vitamin C synthesis in primates and other vertebrates [49–51], or the loss of developmental regulators such as Hox genes and CTCF in nematodes [52, 53]. In particular, there is increasing evidence that species that share a similar ecological niche experience the loss of the same gene(s) in a convergent manner. Systematic computational screens revealed that, for example, carnivorous and herbivorous mammals, animals with low visual acuity, or aquatic mammals encounter convergent loss of the same genes independently [54–56]. Expanding on these studies, we describe here numerous mutational states that characterize the ongoing loss of a putative gene regulatory protein, BORIS, from a distinct lineage of birds. We support our main findings by analyses of repeat element content and selective forces in the BORIS locus and conclude that this locus is destined for extinction in neognathous birds.

Results

Absence of a functional BORIS gene in *Gallus gallus*

The CTCF paralog BORIS, also known as CTCF-like (CTCFL) [26], originated in the ancestor of amniotes 318 Mya [17] and therefore is expected to exist in the genomes of reptiles, mammals, and birds. However, apart from a small fragment with similarity to zinc finger one (ZF I) [17], previous studies failed to detect BORIS in the chicken genome despite its presence in other bird and amniote species [19]. To resolve the conflicting findings, we investigated by a combined *in vitro* and *in silico* approach if there is a functional BORIS gene in the *Gallus gallus* genome. Utilizing computational searches with a hidden Markov model for the BORIS/CTCFL coding sequence, we identified the 36 AA CTCFL fragment described by Hore et al. [17] and three additional genomic fragments with similarity to BORIS/CTCFL in the *Gallus gallus* genome (Figure 1; Table 1).

As expected from synteny information in other amniotes [19], all four ORFs were located between the genes PCK1 and RBM38 where the BORIS locus is situated. The pieces corresponded to ZF I (HMM hit 2, as identified by Hore et al. [17]), the majority of ZF II (HMM hit 3), and a conserved protein sequence directly after

ZF XI, separated into two exons (HMM hits 1 and 4). Together, they comprised 128 AA or 19.3 % of BORIS' expected size (663 AA in humans) and spanned $\sim$ 3.4 kb in the genome, with no identifiable traces of the missing parts of the gene in the corresponding genomic region. To confirm that the degenerate configuration derived *in silico* reflects the situation *in vivo*, we PCR-amplified and sequenced from chicken genomic DNA a 3.7 kb region extending from the detected ZF I fragment (HMM hit 2) to the conserved region after ZF XI (HMM hit 4; Figure 1; Table 1). We would expect to find remnants of the highly characteristic CTCF/CTCFL zinc finger domains [15] in the amplified DNA if the gene was present in *Gallus gallus*. However, the re-sequenced DNA closely matched the genome assembly and thus did not contain additional parts of BORIS missing from the assembly (Figure 1D). In *G. gallus*, the region from the start of HMM hit 2 to the end of HMM hit 1 corresponds to exons 3 and 9 (AA 206–554; see Figure 1C and Table 1) of the 591 AA BORIS HMM and covers 1277 bp in the genome assembly. In contrast, the corresponding region (BORIS AA 206–554) of the closely related galliform *Numida meleagris* extends over seven exons and 6136 bp, suggesting that several central exons of BORIS are missing in *G. gallus*. Together, these data confirm the correctness of the *Gallus gallus* genome assembly at the BORIS locus and demonstrate the absence of a functional BORIS gene in this species despite the retention of four exon fragments in the original syntenic context.

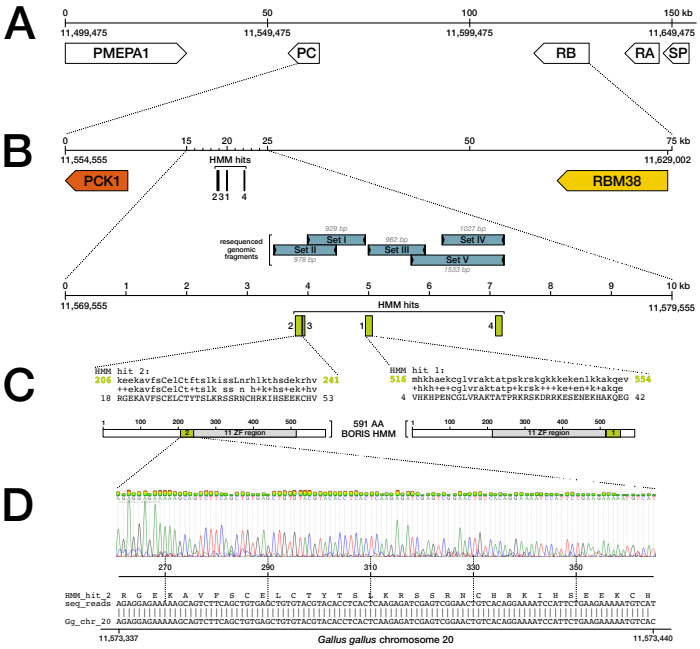


Figure 1: Presence of non-functional BORIS fragments within the *Gallus gallus* BORIS syntenic region. A: Syntenic block of the five genes PMEPA1, PCK1 (PC), RBM38 (RB), RAE1 (RA), and SPO11 (SP), surrounding the BORIS/CTCFL locus on *Gallus gallus* chromosome 20 (position 11 499 475 to 11 653 697 in genome assembly version GRCg6a). Order and arrangement of the genes are conserved in amniotes [19]. Gene sizes and distances are drawn to scale, gene orientation is indicated by arrowheads. **B:** The 75 kb region between the BORIS neighbour genes PCK1 and RBM38 with four BORIS fragments identified by HMM searches (green; HMM hits 1–4) and five re-sequenced genomic regions, covering a 3.7 kb area (blue Set I to Set V; for PCR/sequencing primers, see Table 2). HMM hits are numbered according to Table 1. Drawn to scale. **C:** Alignment details of HMM hits 1 and 2, as reported by `HMMSEARCH` [57]. The identified genomic fragments (lower sequence; position on ORF indicated by numbers) are aligned to a 591 AA hidden Markov model of BORIS (upper sequence, match on HMM indicated by green numbers). Conserved residues between identified ORF and BORIS HMM are reported in the central line. Corresponding hit regions within the BORIS HMM are marked in green in the cartoon below. ZF region: zinc finger region of BORIS. **D:** Exemplary sequencing detail: Chromatogram of a Set II sequencing read across HMM hit 2 (top) and alignment between the obtained sequence/ORF and the corresponding genomic region on *G. gallus* chromosome 20 (bottom).

The degeneration of BORIS is recent, recurrent, and restricted to neognathous birds

To determine if BORIS is also prone to degenerative events in birds other than *Gallus gallus*, we compiled a comprehensive dataset of 59 bird genomes, including Paleognathae and other Galliformes (Supplementary Tables S1, S2), and examined in detail

Table 1: BORIS fragments within the *Gallus gallus* PCK1–RBM38 syntenic region. Hit number and ORF-ID of four *Gallus gallus* ORFs with similarity to BORIS, as revealed by HMM searches (columns one and two). ORFs are derived from chromosome 20 of the *Gallus gallus* genome assembly, version GRCg6a (NC_006107.5, contig 11628597). Column three displays the corresponding ORF sequence. The actual HMM search hit region, corresponding to a BORIS fragment, is positioned within square brackets and highlighted in italics. Rightmost columns indicate the region of the 591 AA BORIS HMM to which the detected fragments display similarity, and the corresponding similarity measure (E value). For graphical representation, see Figure 1.

Hit	ORF-ID	Sequence	E value	HMM
1	11554241_660	QRG[*VHKHPENCGLVRAKTATPRKRSKDRRKESENEKHAKQEG*]NQ	$2.6e^{-13}$	516..554
2	11554241_630	RVKSIFCISWSQCVFVL[*RGEKAVFSCELCTYTSLKRSSRNCHRKI HSEEKCHV*]SQGFSNSCSPAEPRECPYRYLLRVDIPCYLLPAVLNQ QASDRLRGVFCDHSALRSSTSFVLSSTLRGCGLFAILCPLLPF	$3.8e^{-13}$	206..241
3	11554241_626	SQSFVSHGLNVSLFSEEKKQSSAVSCVRTPHSRDRVGTVTGKSILK KNV[*TCLKAFQTAALLQNHVNVHTG*]ICCV	$3.3e^{-10}$	244..264
4	11554241_730	SPQYLIFCKT[*DLELFPDVSTVKSEHCAREIAPHLEGTEGTALKAE DR*]STL	$5.1e^{-06}$	553..588

the state of the BORIS gene in these species. As for *Gallus*, we extracted from the genomes the syntenic region surrounding the BORIS locus, from the genes PCK1 to RBM38. We then translated the corresponding sequences into six reading frames and scanned the resulting ORFs with a 591 AA BORIS hidden Markov model (HMM) to detect with high specificity ORFs similar to the BORIS protein, independently of potentially missing or erroneous genome annotations. With the help of this method, we identified and annotated across all bird species ORFs of BORIS' highly conserved zinc finger region. Our results show that the BORIS zinc finger region is complete and intact in the eleven paleognathous species of our collection, spanning all known families of this monophyletic clade [58] (Figure 2). In contrast, we detected a large number of degenerative events in the BORIS genes of neognathous birds (Figure 2). Mapped onto a bird phylogeny, our results revealed that (i) BORIS mutations are detectable only in neognathous birds but are absent from Paleognathae; (ii) changes in BORIS comprise a wide spectrum of molecular decay: from point mutations affecting zinc-complexing residues—and therefore DNA binding capabilities [59]—(e. g. in *Columba livia*, *Amazona aestiva*, *Nipponia nippon*) to frame shifts and stop codons (e. g. in

Nestor notabilis, *Tauraco erythrolophus*, *Calidris pugnax*), to the loss of individual zinc fingers or exons (e. g. in Passeriformes, *Cariama cristata*, or *Cuculus canorus*), to the severe disintegration of the entire genomic locus, as in *Acanthisitta chloris*, *Dryobates pubescens*, or *Gallus gallus*; (iii) different bird lineages acquired distinct modifications of the BORIS locus, on a global scale across the dataset as well as within monophyletic clades (e. g. Columbaves) or between sister species (e. g. *Dryobates pubescens* vs. *Merops nubicus*); (iv) mutations preferentially affect the *C*-terminal part of the BORIS zinc finger region (30 species with mutations in ZF VII–XI) while changes in the N-terminal part are less prevalent (14 species with mutations in ZF I–V; Figure 2). Although, among the birds in our set, the *G. gallus* BORIS locus is most severely affected by degeneration, substantial damage to BORIS is also observed in several other species (e. g. *Acanthisitta chloris* or *Dryobates pubescens*). Together, we identified 26 independent BORIS mutational profiles in our tree of 48 neognathous species, and therefore mutations in every other species (54.2 %; Figure 2).

To investigate whether degradation events of the BORIS gene are detectable in other amniotes as well, we used the same work flow (scanning ORFs of the BORIS syntenic region with our BORIS HMM) and analyzed the mutational state of BORIS in representative sets of the two other amniote clades, in mammals and reptiles. The corresponding results clearly show that BORIS is intact in all investigated mammalian ($n = 38$) and reptilian ($n = 16$) species (Supplementary Figures S1,S2) and confirm that specifically neognathous birds experienced a loss of this gene, as its integrity in paleognathous birds had suggested (Figure 2).

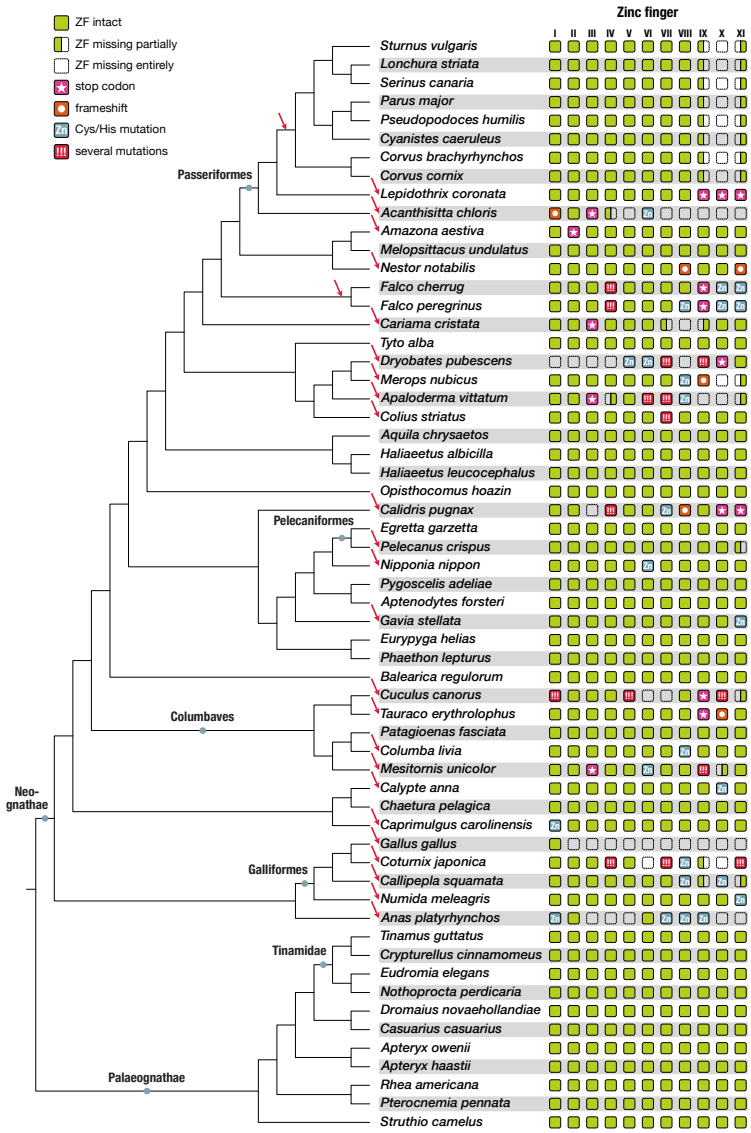


Figure 2: Convergent degeneration of CTCFL/BORIS in neognathous birds. Left: Cladogram depicting the relationships of 59 bird species. Tree topology after https://phylotastic.org/ and Prum et al. [58], Wang et al. [60], Weir et al. [61]. Major bird lineages are indicated by grey dots. **Right:** Presence and mutational state of the eleven CTCFL/BORIS C_2H_2 zinc fingers (ZF I–XI), mapped onto the bird phylogeny (left). Intact and complete zinc fingers are depicted as green squares, damaged zinc fingers are classified by icons (see legend at top left). Red arrows indicate independent degeneration events within a given lineage.

Degenerative events are specific for the BORIS locus

If the observed losses were a result of chromosomal rearrangements around the BORIS locus in Neognathae, we could expect that BORIS' neighbour genes suffer from such modifications as well. To test this possibility, we investigated the completeness and integrity of the two BORIS flanking genes, PCK1 and RBM38, across the bird phylogeny. As proxy for intactness we selected the two genes' coding sequence and subsequently collected these from the proteomes of different bird species by `BLAST` searches at NCBI. We then created multiple sequence alignments of the two proteins (PCK1 and RBM38), each consisting of sequences from at least 40 species. Finally, we inferred from these alignments the mutational state of the two neighbour genes across birds.

Neighbour gene RBM38 (RNA-binding protein 38) is a 215 AA RNA-binding protein with a role in transcript stabilization. It is highly conserved in birds and other species. Regardless of mutations at the BORIS locus, the RBM38 gene is complete and undamaged in all examined birds (Supplementary File S1). Similarly, alignments of PCK1 (Cytosolic phosphoenolpyruvate carboxykinase 1), the rate-limiting enzyme in gluconeogenesis (622 AA in birds), show a strong conservation of its protein sequence across birds, without signs of damage or deletion (Supplementary File S2). In particular, species with defects in BORIS do maintain intact neighbour genes RBM38 and PCK1, e. g. *Gallus*, *Acanthisitta*, or *Dryobates*. These observations indicate that the degenerative events observed in neognathous birds are restricted to and specifically affect the BORIS locus.

Accumulation of species-specific repetitive elements in degenerate BORIS loci

Genomic regions free of coding sequence or regulatory function may evolve with little selective constraint and accumulate nucleotide substitutions, insertions, or deletions at a faster rate than sections under purifying selection. In particular, unconstrained regions may tolerate the insertion of repetitive elements.

To test whether degenerate BORIS loci are susceptible to repeat insertion, we carried out a thorough analysis of repeat element abundance in 18 bird species with intact and degenerate BORIS loci (Supplementary Table S3) by combining a de novo and a library-based approach for repeat detection. First, we generated a comprehensive bird de novo repeat library from the joined output of independent RepeatModeler (detection of transposable elements, tandem repeats, and LTRs) and Mitetracker (detection of MITEs—miniature inverted-repeats) runs on all 18 genomes. We then carried out RepeatMasker analyses on all genomes, using the combined de novo library. In addition, a second RepeatMasker run on all 18 genomes utilized a standard RepeatMasker library (RepBase Release 20181026) for repeat detection (results are summarized in Supplementary File S3).

In agreement with previous studies [62, 63], we find that several repeat classes are highly abundant in birds on a genome-wide scale (SINEs, LINEs, CR1 and LTR elements, DNA transposons, Small RNA) while others could not be detected at all (CRE/SLACS, R1/LOA/Jockey, BEL/Pao, Ty1/Copia, En-Spm, or PiggyBac elements; Supplementary Figure S3). For most species, the RepeatMasker library-based pipeline reveals more repetitive elements in total than does the de novo pipeline (Supplementary Figure S3). A few repeat classes, however, could only be detected by the de novo library (e. g. the R2/R4 class, «Other», and Simple Repeats; Supplementary Figure S4), demonstrating the value of a two-tiered strategy for repeat identification. Second, we find that highly abundant retroelements, in particular the classes «LINE», «CR1», «LTR element», and «retroviral element», are more prevalent in neognathous birds than in paleognathous species (Supplementary Figure S4), supporting the view that the evolution of Neognathae is accompanied by an expansion of retroelements that utilize RNA intermediates. In contrast, DNA-based repeat elements are more prevalent in paleognathous birds despite their generally lower abundance (Penelope, DNA transposons, Tourist/Harbinger, Simple Repeats; Supplementary Figure S4), suggesting that Neognathae and Paleognathae differ in

their mechanisms of repetitive element control. *Dryobates'* unusually high repeat content is in line with previous findings [63], demonstrating the robustness of our repeat detection pipeline (Supplementary Figure S3).

Next, we analysed in detail the repeat landscape within the BORIS syntenic region of the 18 bird species. From species with intact BORIS coding region we deduced that complete BORIS loci extend over ∼13 kb in birds. We then identified BORIS marker exons in all bird genomes and aligned the 13 kb region accordingly, thereby obtaining genomic coordinates of the supposed consensus BORIS locus. When we looked for repeat elements present within these coordinates, we found that both, repeat counts per kb of genomic sequence and total repeat counts, were elevated in degenerate BORIS loci of neognathous birds, as compared to intact loci (Figure 3A,B), although statistical tests with a significance threshold of $p = 0.05$ narrowly failed to recover significant differences (Mann-Whitney U test: statistic: 12.00, P value = 0.09; Kruskal-Wallis test: H statistic: 3.125, P value = 0.077). In contrast, repeat contents of intact BORIS loci were almost identical between the two bird groups (Mann-Whitney U test: statistic: 15.00, P value = 0.86; Kruskal-Wallis test: H statistic: 0.077, P value = 0.782), implying a true, albeit not significant difference in repeat counts between intact and degenerate BORIS loci. Of 90 distinct repeat elements that populate the BORIS loci of the 18 analysed species, 68 (75.6 %) are present only once in a single locus of a single species, 18 elements (20.0 %) occur in two or three copies (often within the same locus), and four (4.4 %) elements occur 5 to 8 times (Supplementary File S4). The most abundant repetitive elements (5 to 8 counts) are MER131 («medium reiterated frequency repeat» 131), a conserved interspersed repeat common in Euteleostomi [64]); MIR1_Amn («mammalian-wide interspersed repeats»), an ancient family of tRNA-derived SINEs (short interspersed nuclear elements) [65]; and two repeats detected by our de novo library (Drypub_rnd-1_family-156 and Strtur_ltr-1_family-80; Supplementary File S4). According to their conservation across species, all these elements share a long history and originated in the ancestors of birds or earlier. On the other hand, most elements that occur exactly once are direct repeats (DRs; 45 of 68 unique

elements or 66.2 %), in particular CR1 repeats common in neognathous birds (Supplementary Figure S4) and other amniotes [66]. Detailed analyses reveal that all direct repeats within BORIS loci are restricted to a single species (Supplementary File S4), suggesting that they emerged in their host and are evolutionarily young compared to the more ancient elements mentioned above. When we looked at the abundance of ancient and young repeats, we found that the latter are significantly overrepresented in degenerate BORIS loci (Kruskal-Wallis test: H statistic: 4.109, P value = 0.043; Figure 3C,D). In contrast, the number of ancient repeats is not significantly different in intact and degenerate BORIS loci (Kruskal-Wallis test: H statistic: 0.433, P value = 0.510). Thus, degenerate BORIS loci emerged independently in closely related neognathous birds (Figure 2) and contain a large number of evolutionarily young repetitive sequences compared to intact loci (Figure 3), arguing that BORIS degeneration and repeated invasion of its locus are species-specific processes that take place simultaneously and possibly influence each other.

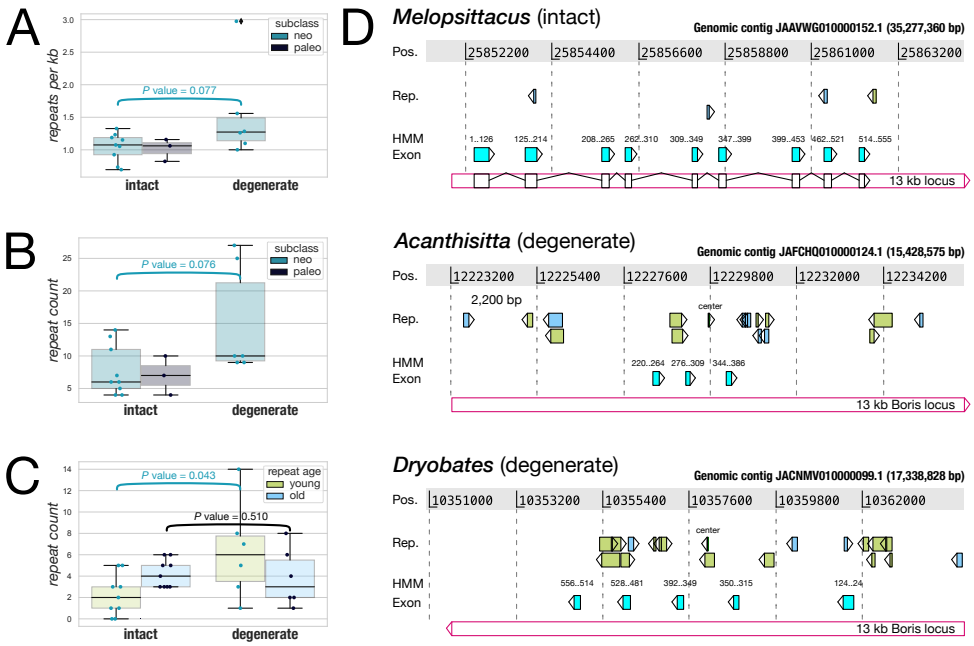


Figure 3: Accumulation of repetitive elements in degenerate BORIS loci. A: Boxplots showing the total number of repeat elements per kb detected on the genomic contig carrying BORIS, in species with intact (left; n=12) and degenerate BORIS loci (right; n=6: *Acanthisitta chloris*, *Anas platyrhynchos*, *Calidris pugnax*, *Cuculus canorus*, *Dryobates pubescens*, *Gallus gallus)*. **B:** Boxplots showing the number of repeat elements detected within the 13 kb BORIS locus in species with intact (left) and degenerate (right) BORIS. **C:** Boxplots showing the number of ancient and young (species-specific) repeat elements in intact vs. degenerate BORIS loci. **D:** Representative intact and degenerate BORIS loci of three bird species, annotated with repeat elements (ancient: light-blue, young/species-specific: yellow) and BORIS exons (cyan). Genomic contigs carrying the respective BORIS locus are referenced at the top right. Four labels at the left indicate (i) the genomic position of an element on the respective contig («Pos.», highlighted as grey box), (ii) the location and orientation of repetitive elements («Rep.»), (iii) the region on the 591 AA BORIS HMM where exons matched («HMM»), and (iv) the location and orientation of BORIS exons as detected by HMM search («Exon»). Red boxes at the bottom display the 13 kb BORIS consensus region. Drawn to scale.

Relaxed selection on BORIS codons in neognathous birds

The recurrent degeneration of BORIS in neognathous birds indicates that the gene might become dispensable in this bird lineage. We should therefore be able to observe a difference in the selective pressure on BORIS between Neognathae and Paleognathae,

across which BORIS is consistently well conserved (Figure 2). As a measure of selective constraint we examined the nonsynonymous vs. synonymous substitution rate (dN/dS) in the two bird lineages. To this end we created a codon-based multiple sequence alignment of the BORIS zinc finger region across the avian phylogeny. We included in the alignment ten paleognathous and 19 neognathous birds with intact, full-length coding sequence, but excluded species with degenerate BORIS. Then, we determined the dN/dS ratios in pairwise comparisons of all possible sequence combinations using the maximum likelihood approach implemented in `CODEML` [67]. We found that comparisons within paleognathous species (PP) had a low dN/dS ratio (0.001 to 0.25 in 45 unique comparisons; mean: 0.069), as expected for a conserved gene under functional constraint. On the other hand, dN/dS ratios between neognathous birds (NN: 0.099 to 0.699 in 171 comparisons; mean: 0.204) and between neo- and paleognathous species (NP: 0.069 to 0.35 in 190 comparisons; mean: 0.165) were markedly higher, suggesting a relaxation of selective constraint by higher nonsynonymous substitution rates in the neognathous lineage (Figure 4A). When we calculated dN/dS ratios in an analogous way for two control genes, PCK1 (18 neognathous, five paleognathous species) and CTCF (18 neognathous, eight paleognathous species), we observed that their dN/dS ratios were uniformly low across the bird tree, i. e. across NN, NP, and PP comparisons (Figure 4B). Statistical tests underscore these results: while there is no significant difference between NN and PP comparisons for the genes PCK1 (H statistic: 0.507; P value = 0.476) and CTCF (H statistic: 3.428; P value = 0.064), dN/dS values for the BORIS coding sequence are significantly higher in NN comparisons than in PP comparisons (H statistic: 210.478; P value = 1.08e-47), as determined by nonparametric Kruskal-Wallis tests (Figure 4B). Thus, BORIS is maintained as a functional gene under purifying selection in Paleognathae. In contrast, functional constraint is relaxed (higher dN/dS values) specifically in neognathous birds, even in those that still possess an intact BORIS zinc finger domain.

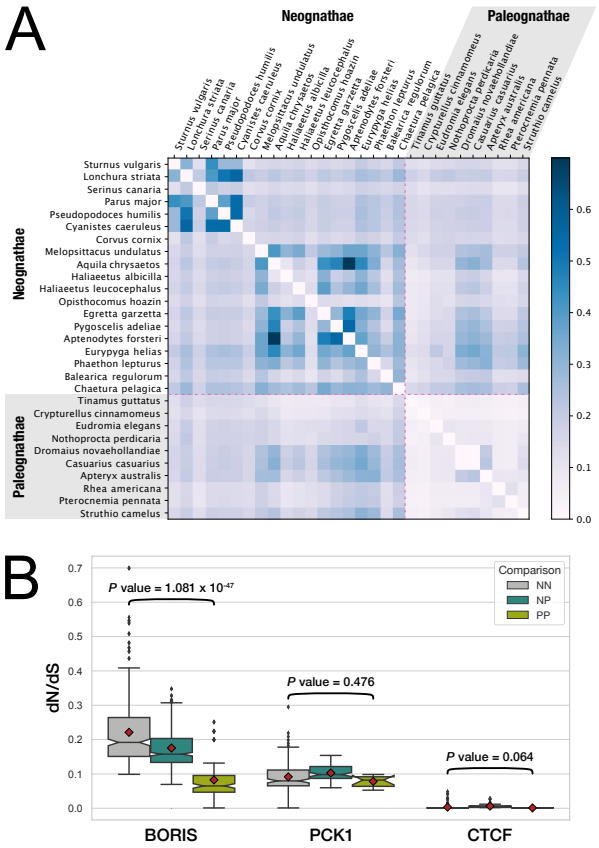


Figure 4: Maximum likelihood analysis of BORIS codon evolution in neognathous and paleognathous birds. A: Heatmap of dN/dS ratios derived from pairwise sequence comparisons of ten paleognathous and 19 neognathous bird species. Paleognathous birds are highlighted by a grey background and a pink dashed line. dN/dS ratios were determined by `CODEML` [67] using the ZF coding region of intact BORIS genes. Note that dN/dS values are plotted in logarithmic scaling. **B:** Boxplot summaries of dN/dS ratios for BORIS (same data as in heatmap above) and two control genes, PCK1 and CTCF, on the basis of pairwise comparisons between neognathous and paleognathous species (neognath-neognath = NN, grey; neognath-paleognath = NP, bluegreen; paleognath-paleognath = PP, yellowgreen). Red diamonds denote the mean of a dataset. Note that dN/dS results for the ZF region of the BORIS paralog CTCF are extremely low and uniform in all analysed comparisons, preventing the formation of a box. *P*-values < 0.05 indicate a significant difference between dN/dS values derived from NN vs. those derived from PP pairwise comparisons (Kruskal-Wallis nonparametric tests).

While these findings establish the ongoing relaxation of selective pressure on BORIS across Neognathae, they do not allow to dissect the relative contributions of individual phylogenetic branches to the overall signal. Also, pairwise comparisons are not independent from each other and may be distorted by an extreme inflation of the dN/dS ratio in only one, or a few, branches [68]. To obtain a more detailed picture, we mapped the substitution events inferred by `CODEML` onto a consensus bird phylogeny, counting nonsynonymous and synonymous substitutions per branch. In paleognathous birds, we find on most internal branches (20 out of 21) much lower counts of nonsynonymous than synonymous substitutions. Only on the branch leading to the genus Apteryx the counts are similar to each other. In contrast, within Neognathae, nonsynonymous counts are similar to or even larger than synonymous ones on 16 out of 37 branches (Figure 5). Importantly, we find such an excess of nonsynonymous substitutions across the entire neognathous avian tree, suggesting that relaxed constraint on BORIS appears to be a general phenomenon in these birds and not the result of few, particularly fast evolving species that might bias the overall view. Still, one extreme signal with many more nonsynonymous than synonymous substitutions can be detected on the branch leading to Psittacopasserae (parrots and passerines [69]), indicating that substantial changes in the BORIS coding region occurred during the evolution of this monophyletic clade (Figure 5). This may point to a phase of accelerated evolution in birds around 60 Mya and is in line with difficulties to resolve the phylogenetic placement of passerines [69].

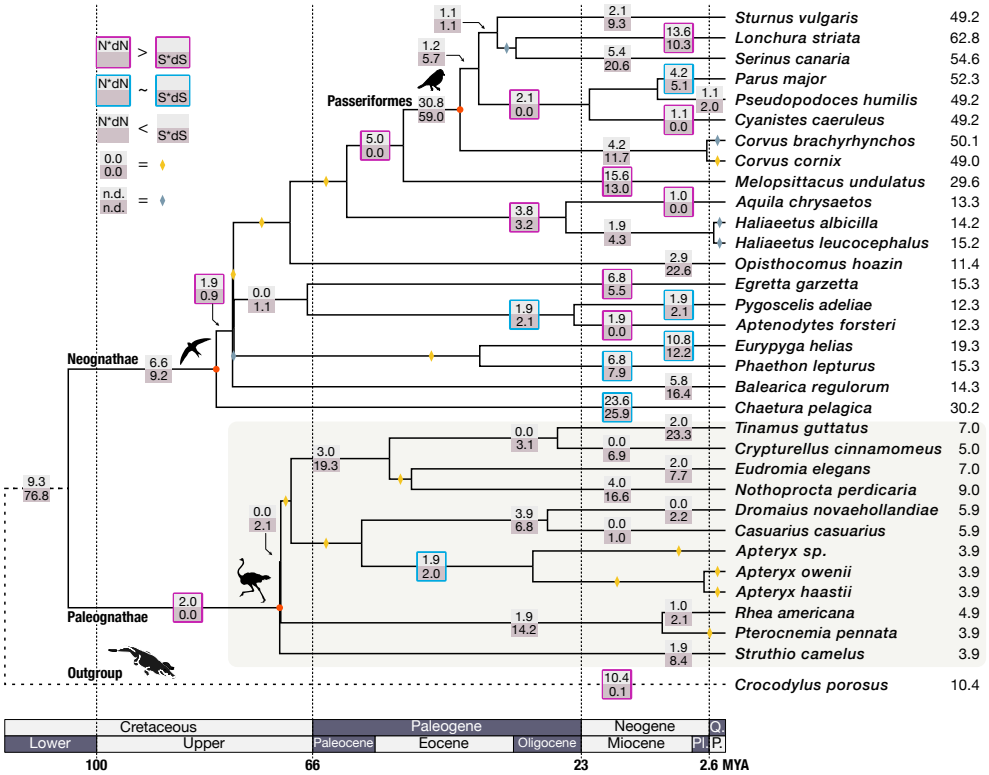


Figure 5: Branch-specific mapping of BORIS codon evolution in birds. Branch-specific mapping of substitution events onto a consensus phylogeny of 12 paleognathous birds, 20 neognathous birds, and the outgroup species *Crocodylus porosus*. Tree topology and time scale are derived from http://datelife.opentreeoflife.org/. Species silhouettes for major clades (red dots) were downloaded from https://www.phylopic.org/. Branch numbers indicate $N \times dN$ (upper) and $S \times dS$ (lower) values as obtained from `CODEML` [67], using the codon-aligned ZF region of intact BORIS genes as input. Pink frames highlight branches where $N \times dN > S \times dS$. Blue frames indicate branches with $N \times dN \simeq S \times dS$. Yellow diamonds identify branches with $N \times dN = S \times dS = 0.0$. Grey diamonds label branches with undetermined $N \times dN$, $S \times dS$ values. Paleognathous birds are highlighted by a grey background. Pl: Pliocene, P: Pleistocene, Q: Quaternary.

Altered sperm morphology in neognathous birds

In mammals, BORIS is expressed in testes and is associated with sperm development and differentiation [39–41]. In contrast, reptiles express BORIS in gonads *and* in somatic tissues according to a comparative study [17]. So far, nothing is known about expression patterns and BORIS functionality in birds. Assuming that BORIS is also important for sperm development in birds and reptiles, as is in mammals, sperm from Neognathae may differ from sperm from Paleognathae/Reptilia, reflecting the molecular difference of BORIS conservation between Neognathae and Paleognathae on a phenotypic level. To test this idea, we extracted measurements from the Sperm Morphology Database [70] on the length of sperm heads, mid pieces, and flagella for Aves and Reptilia. Since the database provides only two datasets for paleognathous birds, we combined this group and Reptilia to a larger dataset and compared it to the Neognathae (Figure 6).

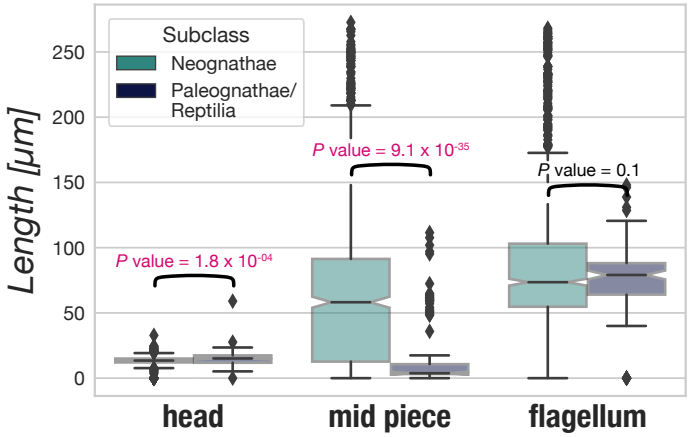


Figure 6: Comparison of morphological characteristics of sperm from Neognathae and Paleognathae/Reptilia. The lengths of sperm heads (left), mid pieces (middle), and flagella (right) from neognathous birds (554 species) are compared to the corresponding data of paleognathous birds and reptiles (122 species). All data are taken from the SpermTree Database (https://spermtree.org/ [70]). *P*-values < 0.05 are highlighted in red and indicate that a morphological trait differs significantly between Neognathae and Paleognathae/Reptilia, as determined by Kruskal-Wallis tests.

Our comparisons reveal that sperm head lengths differ slightly between Neognathae and Paleognathae/Reptilia (Neognathae: 14.13±6.49 µm; Paleognathae/Reptilia: 15.41 ± 5.53 µm; Figure 6), as do flagellum lengths (Neognathae: 101.86 ± 58.77 µm; Paleognathae/Reptilia: 79.23 ± 19.49 µm; Figure 6). In contrast, there are significant differences in the length of mid pieces between the two groups. The mid pieces of sperm from Neognathae are clearly larger (81.92 ± 66.99 µm) than those of Paleognathae/Reptilia (18.53 ± 27.75 µm; Figure 6). This is confirmed by non-parametric statistical tests (Kruskal-Wallis) that recover significant differences in the length of sperm heads and mid pieces between Neognathae and Paleognathae/Reptilia (head: H statistic: 13.997, P-value = 1.83e-04; mid piece: H statistic: 151.285, P-value = 9.08e-35; flagellum: H statistic: 2.462, P-value = 0.116).

Due to the low abundance of paleognathous species, the difference in sperm mid piece length reported here essentially reflects a difference between Neognathae (554 data points) and Reptilia (120 data points; compare Figure 6 to Supplementary Figure S5A). However, additional comparisons between Paleognathae and Reptilia and between Paleognathae and Neognathae corroborate that mid piece and flagellum lengths of Paleognathae are well below the interquartile range of neognaths and reptiles (Supplementary Figure S5B, C), suggesting that the two branches of birds indeed show a significant difference in the length of their sperm mid pieces.

These findings indicate that a relaxed selective pressure on BORIS and its tendency to degrade are correlated with an increase in the length of sperm heads and mid pieces in Neognathae. Whether these alterations of sperm morphology in neognathous birds are caused by a change in BORIS function or expression needs to be determined in future experiments.

Discussion

Here we report that the BORIS gene, a paralog of the transcription factor CTCF, is conserved in paleognathous birds, but is degrading in many species of its sister clade, the neognathous birds. Loss of BORIS cannot be explained by a single event in the

neognathous ancestor. Instead, our comparison of genomic sequences suggests a series of independent mutational damages since the Paleognathae and Neognathae split about 108 Mya. We were able to identify traces of the BORIS gene in the genomes of all analyzed Neognathae and found no case of complete absence. However, we discovered a wide range of partial losses, with damage from mild to severe. We conclude from these findings that the degradation of BORIS is a recurrent and still ongoing process. Although the BORIS gene has apparently remained intact in some Neognathae, our analysis of nonsynonymous vs. synonymous substitutions suggests that purifying selection is reduced in Neognathae compared to Paleognathae – a likely consequence of the compromised function of BORIS as a common characteristic across the entire clade. Still, we cannot exclude some functionality in at least some neognathous species. In the human germline and in human cancer cells, 23 alternatively spliced transcripts of BORIS with varying sets of zinc fingers were identified [83]. Therefore, one could expect that mutations leading to a reduced number of zinc fingers, as observed in the Passeriformes, do not necessarily result in a loss of function. However, in species such as *Acanthisitta chloris*, *Amazona aestiva*, or *Apaloderma vittatum*, the presence of premature stop codons or frameshifts in the zinc-finger domain strongly suggests a loss of function. Transcriptome studies could help to decide this question. Searching currently available transcriptomes from six passerine birds and five different tissues each [84], did not reveal evidence that BORIS mRNA is present in these species (results not shown).

The process of gene damage must have started soon after the split of Neognathae and Paleognathae about 100 million years ago and continues to the present. This is supported by our finding that young, species-specific repetitive elements are overrepresented within the degrading BORIS loci of neognathous species. Accordingly, we consider the BORIS gene as a gene currently undergoing pseudogenization in Neognathae. Besides mutation and gene duplication, loss of genes through pseudogenization is an important driver of evolution [44, 85–87]. There are numerous examples for lineage-specific events of convergent progressive pseudogenization due to, or followed by, altered selective pressure. While in some cases, the origin for the altered

selection is quite obvious [54, 88–91], it is much more obscure or unknown in others [92–94], for instance in BORIS. Taking a closer view on the similarities and differences between the two infra-classes of birds might provide insights.

The taxonomic classification of birds into Neognathae and Paleognathae is based on morphological differences of the Pterygoid-Palatinum Complex (PPC) of the adult skull [97]. In contrast to Paleognathae ($\sim$ 50 known species), Neognathae exhibit a high taxonomic diversity, encompassing approximately 10 000 species, and are characterized by greater cranial kinesis, the movement of skull bones relative to each other, which may have facilitated their extensive diversification as well as their enhanced vocal capabilities [98–100]. In songbirds, it has been shown that the ability of vocal learning is associated with a specific gene expression pattern in the brain [104]. Due to its descent from a transcriptional regulator and the conserved structure of the zinc finger domain, BORIS likely acts as a transcriptional regulator in birds. However, there is currently no indication that BORIS is involved in transcriptional regulation in neuronal tissue.

Beyond differences in cranial morphology, Paleognathae and Neognathae also differ in the morphology of their sex chromosomes. Sex determination in birds relies on ZW chromosomes with heterogamy in females (ZW) and homogamy in males (ZZ). While most Paleognathae have retained ZW homomorphism with extensive and well-conserved pseudoautosomal regions and recombination rates comparable to autosomes, Neognathae display a pronounced ZW heteromorphism with small, gene-poor W chromosomes [101–103]. BORIS is known for its testis-specific expression in mammals [26]. For instance, it has been shown that BORIS regulates other testis-specific genes during spermatogenesis in mice [41, 83, 105]. Therefore, it seems plausible to consider a relationship between the varying characteristics of W chromosomes and the conservation status of BORIS between Paleognathae and Neognathae. However, the origin of BORIS from a duplicate of CTCF does not support a link to sex chromosomes, as both CTCF and BORIS are genuinely autosomal genes.

To further explore a possible involvement of BORIS in the reproductive strategies of birds, we asked whether Neognathae and Paleognathae differ in sperm morphology.

We compared the morphological characteristics of sperm in species with and without conserved BORIS. We could show that larger sperm midpiece in neognathous species correlates with altered selective pressure on the BORIS gene compared to Paleognathae and reptiles with a conserved BORIS gene. Genes encoding proteins involved in reproduction are known for their rapid evolution [106–108]. Previous authors suggested post-copulatory sexual selection as the main driver behind this rapid evolution, as reproductive proteins involved in sperm competition tend to evolve specifically fast [109]. This applies particularly to species with promiscuous mating systems [110–112], where sexual selection drives diversity in sperm morphology and function, including variations in the size and structure of the midpiece, which are crucial for motility and successful fertilization [113–117].

Sperm competition and post-copulatory sexual selection are often accompanied by promiscuous mating styles [109–111]. Despite the social monogamy observed during parental care in many neognathous species, genetic analyses of clutches in a variety of bird species have shown that extra-pair paternity is common among Paleognathae and Neognathae [118, 119]. Thus, sperm competition can be assumed to occur throughout the entire avian phylogeny, regardless of whether BORIS is conserved or not, which makes an association between the conservation status of BORIS and sperm competition less likely. Immler et al. [120] could show that the duration of sperm storage in the female reproductive tract also plays a role in the evolution of the sperm. In pheasants, sperm size traits are negatively associated with the duration of sperm storage, indicating that prolonged sperm-female interaction might influence sperm evolution [120].

One notable reproductive difference between Paleognathae and most Neognathae is the presence or absence of external genitalia. While Paleognathae retain external genitalia, most neognathous species have lost this trait [121]. Only three percent of avian species, belonging to two main clades, have retained the ancestral copulatory organ: the Paleognathae and Anseriformes. All other birds have lost the penis-like structure often referred to as intromittend organ (IO) [122–125]. There is a conserved developmental stage of external genital development among all amniotes that suggests

a single evolutionary origin of amniote external genitalia [126]. This implies that, in analogy to the presence or absence of the BORIS gene, the presence of an IO is the ancestral state and the loss of an IO is a derived state. The alteration in copulatory anatomy might require different sperm characteristics. Assuming that BORIS exhibits a similar expression pattern and functionality in birds as in mammals, it would be plausible to assume that the reduced selective pressure on BORIS in Neognathae could be functionally related to the altered genital morphology. However, whether there is actually a causal relationship between BORIS degradation, the absence of an IO, and altered sperm morphology remains highly speculative at this point. Clarification can be expected from dedicated expression studies with a focus on reproductive tissues.

Materials and Methods

Extraction of genomic DNA

We purchased fresh liver from chicken (*Gallus gallus*) at the local supermarket and cut it into cubes of 2 cm edge length. To avoid contamination with DNA from other poultry, the portions were transferred to 2.8 % sodium hypochlorite solution (NaClO, as in «DanKlorix») for 10 min and washed 3× in PBS (phosphate buffered saline). The pieces were stored in plastic tubes at −20 °C. For DNA extraction, the outer layer of the frozen tissue was removed, and samples with a volume of ∼1 mm^3 were collected from the inside material. After homogenisation in 180 µL lysis buffer, extraction of the genomic DNA was carried out using the Macherey & Nagel NucleoSpin™ Tissue kit. DNA extraction was performed according to the manufacturer's specifications, with the modification that, after addition of 25 µL Proteinase K, we incubated the sample for 20 min, instead of 1 h to 2 h, at 56 °C and 300 rpm. Prior to elution, the silica column was centrifuged at 13 000 *g* for 5 min and dried for another 5 min. The DNA was eluted in 100 µL Tris (10 mM, pH 8.4) and quantified with a NanoDrop™ 2000 spectrophotometer (ThermoFisher Scientific). DNA quality was verified on a 1.0 % agarose gel.

PCR amplification and sequencing

We designed five sets of PCR primers to obtain from chicken genomic DNA five overlapping fragments of a 3.7 kb region targeting the *G. gallus* BORIS locus (see Table 2; Figure 1). To amplify the selected fragments, we used standard PCR conditions: 1. Initial denaturation (94 °C, 3 min), 2. Denaturation (94 °C, 30 s), 3. Annealing (50 °C, 3 s), 4. Elongation (72 °C, 3 min), 5. Final extension (72 °C, 10 min), with 36 cycles of steps 2–4. The amplified DNA was quantified with a NanoDrop™ 2000 spectrophotometer (ThermoFisher Scientific) and separated on a 1.0 % agarose gel. Amplicons of the expected size were excised from the gel, extracted using the Macherey & Nagel NucleoSpin™ gel and PCR clean-up kit, and sequenced in forward and reverse orientation using the primers displayed in Table 2. Sanger sequencing was performed at Eurofins Genomics, 85560 Ebersberg, Germany. Invdividual sequencing reads were assembled using the phred/phrap/consed package [71, 72]. The resulting contigs were aligned to the *Gallus gallus* reference genome with `MUSCLE` [73]. Alignments were visualized with `SEAVIEW` [74].

Table 2: Primer list for the PCR-based amplification of genomic DNA fragments. Columns two and three show start and end coordinates of the respective primer on chromosome 20 of the *G. gallus* genome assembly, version GRCg6a. Column five indicates the expected amplicon size for the respective primers.

Name	Start	End	Sequence	[bp]
1FW (Set I)	11 573 553	11 573 572	CAGCAGGCAAGTGACAGACT	929
1RW	11 574 480	11 574 499	TTATGCTAGCAGCGTGGTGT	
2FW (Set II)	11 573 050	11 573 069	CCTCTCTTGGAGGTGGGAGT	978
2RW	11 574 008	11 574 027	ATTTACGGGGAGGTTCCTGC	
3FW (Set III)	11 574 533	11 574 552	GGCCAAAACTGCTACACCCA	962
3RW	11 575 474	11 575 494	ACCTTTCTGTCTGCTGGCATT	
4FW (Set IV)	11 575 765	11 575 784	AGGCCTAGGTGTCCTCAAAC	1027
4RW	11 576 771	11 576 791	TGGCTGTGATGGACTGACATC	
4.1FW (Set V)	11 575 259	11 575 279	GTAGCTCTTGAGCACTGAGCA	1533
4RW	11 576 771	11 576 791	TGGCTGTGATGGACTGACATC	

Data collection and generation of ORFs

After identifying bird genomic contigs/scaffolds containing BORIS or—if this was unsuccessful—its neighbour genes PCK1 and RBM38 by NCBI `BLAST` (https://blast.ncbi.nlm.nih.gov/Blast.cgi), using human and bird proteins as queries (BORIS: sp|Q8NI51|CTCFL_HUMAN; PCK1: sp|P05153|PCKGC_CHICK; RBM38: sp|Q5ZJX4|RBM38_CHICK), we downloaded from the NCBI database the respective sequences of 59 bird species under the accessions listed in Supplementary Tables S1 and S2.

The 59 sequences were translated into six open reading frames (ORFs) using the Emboss tool `GETORF` with parameters «`-minsize 90 -find 0`» to include all regions between Stop codons of at least 30 AA length [75]. This produced a data set of 5000 to 10 000 ORFs per species, our raw material for the identification and annotation of BORIS fragments and of BORIS neighbor genes in subsequent steps.

Multiple sequence alignment

Multiple sequence alignments were carried out using the `MAFFT` v7.304b «`einsi`» algorithm [76] or `MUSCLE` [73] with default parameters.

HMM model and HMM searches

To generate a hidden Markov model (HMM) representative for BORIS from diapsid amniotes (Sauria), to which birds belong, we selected 16 BORIS protein sequences from seven reptilian and nine bird species (Supplementary Table S4). We verified the orthology of these sequences to the protein BORIS in phylogenetic analyses [see 19], generated a multiple sequence alignment from the orthologs using `MUSCLE` [73], and built a 591 AA BORIS HMM from the alignment using `HMMer` version 3.1b2 [57]. With the resulting BORIS hidden Markov model, we scanned the ORFs derived from the BORIS syntenic region (spanning genes PCK1 to RBM38; see above) of 59 bird species and annotated ORFs with similarity to BORIS in the corresponding genomic sequence. On the basis of these results, we obtained the status of each individual BORIS zinc finger in each bird species and mapped it onto a consensus bird phylogeny.

Identification of PCK1 and RBM38 orthologs

Using *Columba livia* PCK1 (tr|A0A2I0M656|A0A2I0M656_COLLI) and RBM38 (tr|A0A2I0M622|A0A2I0M622_COLLI) as queries, we performed `BLASTP` searches restricted to the taxon «Aves» (Taxonomy ID: 8782) at the NCBI databases (https://blast.ncbi.nlm.nih.gov/Blast.cgi). We downloaded PCK1 blast hits from 76 bird species and RBM38 blast hits from 44 species in fasta format, aligned the corresponding sequences with `MUSCLE` [73] and visually inspected the alignments using `SEAVIEW` [74].

Identification and analysis of repetitive sequences

For comprehensive repeat detection in bird genomes, we combined de novo and library-based identification of repeat elements. For de novo identification, we ran `RepeatModeler` version 2.0.1 [77] with parameters «`-engine rmblast`» and «`-LTRStruct`» on 18 selected bird whole genome sequences (Supplementary Table S3). Typically, several hundred repeat elements per species were detected de novo. We concatenated all repeat elements of the 18 species and eliminated redundancy by clustering with `CD-HIT` [78] and `VSEARCH` [79]. Clustering reduced the number of repeat elements from 7372 (overall) to 6666 distinct repeat clusters at 90 % identity threshold. In addition, we identified MITEs (Miniature inverted-repeat transposable elements) in the 18 bird genomes using `MITE Tracker` [80] with parameter «`--mite_max_len 1000`». We concatenated the 5988 individual MITEs of 18 species and obtained 4557 MITE clusters with 90 % identity threshold. We then combined the clustered `RepeatModeler` and `MITE Tracker` results to construct a comprehensive bird de novo repeat library. We used the combined de novo library and a standard repeat library (Dfam release 3.2; https://dfam.org/home) to mask repeats in the 18 bird genomes in independent `RepeatMasker` runs. A typical `RepeatMasker` run contained the following parameters: «`-engine rmblast -pa 3 -s -lib $LIB -dir $DATADIR -xsmall -gff -xm -nolow input.fasta`». After repeat-masking, we collected for each bird species the repeat content within 100 kb around the BORIS zinc finger region (ZF region start ±50 kb), roughly corresponding to the PCK1 to RBM38 syntenic

region. Finally, we analysed in detail repeat contents of the ∼13 kb BORIS consensus locus in selected species using custom scripts and a local `JUPYTERLAB` instance (python 3.10.9, jupyter server 2.4.0, jupyterlab 3.6.1, pandas 1.5.3, seaborn 0.12.2, and matplotlib 3.7.1).

Calculation of substitution rates

To investigate the ratio of nonsynonymous vs. synonymous mutations in the BORIS coding region, we first generated a codon-based multiple sequence alignment of BORIS coding sequence from 19 neognathous and 10 paleognathous birds with an intact gene using `MACSE` (Multiple Alignment of Coding SEquences) release v2.03 [81]. Then we removed poorly aligned positions and divergent regions of the alignments using `GBLOCKS` Version 0.91b with parameters «`-t=c -b2="(#seq/2)+1+0.5" -b4=2 -b5=a -d=y`», with «#seq» specifying the number of sequences in the alignment [82]. Next, we quantified the number of synonymous and nonsynonymous substitutions on each branch using `CODEML` of the `PAML` package (parameters: `runmode = -2; seqtype = 1; CodonFreq = 2; NSsites = 0`) [67] . The tree topology passed to `CODEML` corresponded to the literature-based consensus phylogeny depicted in Figure 2. Finally, we stored pairwise dN/dS ratios from the `CODEML` output in a table and used a `JUPYTER` notebook instance (python 3.10.9, jupyter server 2.4.0, jupyterlab 3.6.1, numpy 1.24.2, and matplotlib 3.7.1) for transforming the data into numpy arrays and plotting. The counts of nonsynonymous and synonymous substitutions ($N \times dN$, $S \times dS$, respectively) are obtained by multiplying the rates (dN and dS) with the number of nonsynonymous and synonymous sites (N and S), as given in the `CODEML` output table.

Comparison of sperm morphological data

To compare morphological characteristics of sperm cells between neognathous and paleognathous birds, we extracted measurements on the length of sperm head, mid piece, and flagellum for Aves and Reptilia from the Sperm Tree Database (file «spermtree_01_21_22.xlsx» on https://spermtree.org/database/ [70]). Since the database contains data for only two paleognathous species, we combined Paleognathae

(two species) and Reptilia (120 species) into a single dataset and compared it to sperm length measurements from Neognathae (554 species). Using pandas 1.5.3 dataframes, seaborn 0.12.2, and matplotlib 3.7.1 under python 3.10.9, we created boxplots to visualize the data. In addition, we carried out non-parametric Kruskal-Wallis tests to determine if differences in the measurements between Neognathae, Paleognathae, and Reptilia are statistically significant, using the above mentioned python framework and the scipy package (v1.10.1).

Acknowledgements. This research was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)-Projektnummer 268236062-SFB 1211 and by the Regional Computing Centre of the University of Cologne (RRZK/ITCC) through access to the HPC system CHEOPS (Cologne High Efficient Operating Platform for Science) on which repeat identification was carried out. We thank Yichen Zheng for helpful discussions of this manuscript.

Dedication. This work is dedicated to the memory of Kamel Jabbari (1966-2020).

References

[1] Lobanenkov VV, Nicolas RH, Adler VV, Paterson H, Klenova EM, Polotskaja AV, et al. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene. 1990;5(12):1743–53.

[2] Filippova GN. Genetics and epigenetics of the multifunctional protein CTCF. Current Topics in Developmental Biology. 2008;80:337–60.

[3] Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009;137(7):1194–211.

[4] Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of Chromosomal Domains by Loop Extrusion. Cell Reports. 2016 May;15:2038–2049. https://doi.org/10.1016/j.celrep.2016.04.085.

[5] Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nature Reviews Genetics. 2018 Dec;19:789–800. https://doi.org/10.1038/s41576-018-0060-8.

[6] Braccioli L, de Wit E. CTCF: a Swiss-army knife for genome organization and transcription regulation. Essays in Biochemistry. 2019 Apr;63:157–165. https://doi.org/10.1042/EBC20180069.

[7] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012 Apr;485:376–380. https://doi.org/10.1038/nature11082.

[8] Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148(3):458–72.

[9] Narendra V, Rocha PP, An D, Raviram R, Skok JA, Mazzoni EO, et al. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science. 2015 Feb;347(6225):1017–1021. https://doi.org/10.1126/science.1262088.

[10] Acemel RD, Maeso I, Gómez-Skarmeta JL. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. Wiley Interdiscip Rev Dev Biol. 2017 May;6. https://doi.org/10.1002/wdev.265.

[11] Amândio AR, Beccari L, Lopez-Delisle L, Mascrez B, Zakany J, Gitto S, et al. Sequential in *cis* mutagenesis in vivo reveals various functions for CTCF sites at the mouse HoxD cluster. Genes & Development. 2021 Nov;35:1490–1509. https://doi.org/10.1101/gad.348934.121.

[12] Labade AS, Salvi A, Kar S, Karmodiya K, Sengupta K. Nup93 and CTCF modulate spatiotemporal dynamics and function of the HOXA gene locus during differentiation. Journal of Cell Science. 2021 Dec;134. https://doi.org/10.1242/jcs.259307.

[13] Ushiki A, Zhang Y, Xiong C, Zhao J, Georgakopoulos-Soares I, Kane L, et al. Deletion of CTCF sites in the SHH locus alters enhancer-promoter interactions and leads to acheiropodia. Nature Communications. 2021 Apr;12:2282. https://doi.org/10.1038/s41467-021-22470-z.

[14] Chen LF, Long HK. Topology regulatory elements: From shaping genome architecture to gene regulation. Current Opinion in Structural Biology. 2023 Nov;83:102723. https://doi.org/10.1016/j.sbi.2023.102723.

[15] Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T. The chromatin insulator CTCF and the emergence of metazoan diversity. Proceedings of the National Academy of Sciences of the United

States of America. 2012;109(43):17507–12.

[16] Heger P, Zheng W, Rottmann A, Panfilio KA, Wiehe T. The genetic factors of bilaterian evolution. eLife. 2020 Jul;9. https://doi.org/10.7554/eLife.45530.

[17] Hore TA, Deakin JE, Marshall Graves JA. The evolution of epigenetic regulators CTCF and BORIS/CTCFL in amniotes. PLoS Genetics. 2008;4(8):e1000169.

[18] Kadota M, Yamaguchi K, Hara Y, Kuraku S. Early vertebrate origin of CTCFL, a CTCF paralog, revealed by proximity-guided shark genome scaffolding. Scientific Reports. 2020 Sep;10:14629. https://doi.org/10.1038/s41598-020-71602-w.

[19] Jabbari K, Heger P, Sharma R, Wiehe T. The Diverging Routes of BORIS and CTCF: An Interactomic and Phylogenomic Analysis. Life. 2018 Jan;8. https://doi.org/10.3390/life8010004.

[20] Shu L, Yan W, Chen X. RNPC1, an RNA-binding protein and a target of the p53 family, is required for maintaining the stability of the basal and stress-induced p21 transcript. Genes & Development. 2006 Nov;20:2961–2972. https://doi.org/10.1101/gad.1463306.

[21] Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Letters. 2008 Jun;582:1977–1986. https://doi.org/10.1016/j.febslet.2008.03.004.

[22] Yoo-Warren H, Monahan JE, Short J, Short H, Bruzel A, Wynshaw-Boris A, et al. Isolation and characterization of the gene coding for cytosolic phosphoenolpyruvate carboxykinase (GTP) from the rat. Proceedings of the National Academy of Sciences of the United States of America. 1983 Jun;80:3656–3660. https://doi.org/10.1073/pnas.80.12.3656.

[23] Seenappa V, Joshi MB, Satyamoorthy K. Intricate Regulation of Phosphoenolpyruvate Carboxykinase (PEPCK) Isoforms in Normal Physiology and Disease. Current Molecular Medicine. 2019;19:247–272. https://doi.org/10.2174/1566524019666190404155801.

[24] Lv Z, Ding Y, Cao W, Wang S, Gao K. Role of RHO family interacting cell polarization regulators (RIPORs) in health and disease: Recent advances and prospects. International Journal of Biological Sciences. 2022;18:800–808. https://doi.org/10.7150/ijbs.65457.

[25] Lanier MH, Kim T, Cooper JA. CARMIL2 is a novel molecular connection between vimentin and actin essential for cell migration and invadopodia formation. Molecular Biology of the Cell. 2015 Dec;26:4577–4588. https://doi.org/10.1091/mbc.E15-08-0552.

[26] Loukinov DI, Pugacheva E, Vatolin S, Pack SD, Moon H, Chernukhin I, et al. BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(10):6806–11.

[27] Sleutels F, Soochit W, Bartkuhn M, Heath H, Dienstbach S, Bergmaier P, et al. The male germ cell gene regulator CTCFL is functionally different from CTCF and binds CTCF-like consensus sites in a nucleosome composition-dependent manner. Epigenetics & Chromatin. 2012 Jun;5:8. https://doi.org/10.1186/1756-8935-5-8.

[28] Pugacheva EM, Rivero-Hinojosa S, Espinoza CA, Méndez-Catalá CF, Kang S, Suzuki T, et al. Comparative analyses of CTCF and BORIS occupancies uncover two distinct classes of CTCF binding genomic regions. Genome Biology. 2015 Aug;16:161. https://doi.org/10.1186/s13059-015-0736-8.

[29] Klenova EM, Morse HC 3rd, Ohlsson R, Lobanenkov VV. The novel BORIS + CTCF gene family is uniquely involved in the epigenetics of normal biology and cancer. Seminars in Cancer Biology. 2002;12(5):399–414.

[30] Nishana M, Ha C, Rodriguez-Hernaez J, Ranjbaran A, Chio E, Nora EP, et al. Defining the relative and combined contribution of CTCF and CTCFL to genomic regulation. Genome Biology. 2020 May;21:108. https://doi.org/10.1186/s13059-020-02024-0.

[31] Del Moral-Morales A, Salgado-Albarrán M, Sánchez-Pérez Y, Wenke NK, Baumbach J, Soto-Reyes E. CTCF and Its Multi-Partner Network for Chromatin Regulation. Cells. 2023 May;12. https://doi.org/10.3390/cells12101357.

[32] Filippova GN, Fagerlie S, Klenova EM, Myers C, Dehner Y, Goodwin G, et al. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. Molecular and Cellular Biology. 1996;16(6):2802–13.

[33] Filippova GN, Lindblom A, Meincke LJ, Klenova EM, Neiman PE, Collins SJ, et al. A widely expressed transcription factor with multiple DNA sequence specificity, CTCF, is localized at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers. Genes Chromosomes Cancer. 1998;22(1):26–36.

[34] Moon H, Filippova G, Loukinov D, Pugacheva E, Chen Q, Smith ST, et al. CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. EMBO Rep. 2005;6(2):165–70.

[35] Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, et al. Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. Cell. 2012;148(1-2):335–48.

[36] Hug CB, Vaquerizas JM. The Birth of the 3D Genome during Early Embryonic Development. Trends in Genetics. 2018 Dec;34:903–914. https://doi.org/10.1016/j.tig.2018.09.002.

[37] Watanabe K, Fujita M, Okamoto K, Yoshioka H, Moriwaki M, Tagashira H, et al. The crucial role of CTCF in mitotic progression during early development of sea urchin. Development, Growth & Differentiation. 2023 Sep;65:395–407. https://doi.org/10.1111/dgd.12875.

[38] Monk M, Hitchins M, Hawes S. Differential expression of the embryo/cancer gene ECSA(DPPA2), the cancer/testis gene BORIS and the pluripotency structural gene OCT4, in human preimplantation development. Molecular Human Reproduction. 2008 Jun;14:347–355. https://doi.org/10.1093/molehr/gan025.

[39] Suzuki T, Kosaka-Suzuki N, Pack S, Shin DM, Yoon J, Abdullaev Z, et al. Expression of a testis-specific form of Gal3st1 (CST), a gene essential for spermatogenesis, is regulated by the CTCF paralogous gene BORIS. Molecular and Cellular Biology. 2010 May;30:2473–2484. https://doi.org/10.1128/MCB.01093-09.

[40] Kosaka-Suzuki N, Suzuki T, Pugacheva EM, Vostrov AA, Morse HC, Loukinov D, et al. Transcription factor BORIS (Brother of the Regulator of Imprinted Sites) directly induces expression of a cancer-testis antigen, TSP50, through regulated binding of BORIS to the promoter. The Journal of Biological Chemistry. 2011 Aug;286:27378–27388. https://doi.org/10.1074/jbc.M111.243576.

[41] Rivero-Hinojosa S, Pugacheva EM, Kang S, Méndez-Catalá CF, Kovalchuk AL, Strunnikov AV, et al. The combined action of CTCF and its testis-specific paralog BORIS is essential for spermatogenesis. Nature Communications. 2021 Jun;12:3846. https://doi.org/10.1038/s41467-021-24140-6.

[42] Soltanian S, Dehghani H. BORIS: a key regulator of cancer stemness. Cancer Cell International. 2018;18:154. https://doi.org/10.1186/s12935-018-0650-8.

[43] Debaugny RE, Skok JA. CTCF and CTCFL in cancer. Current Opinion in Genetics & Development. 2020 Apr;61:44–52. https://doi.org/10.1016/j.gde.2020.02.021.

[44] Albalat R, Cañestro C. Evolution by gene loss. Nature Reviews Genetics. 2016 Jul;17:379–391. https://doi.org/10.1038/nrg.2016.39.

[45] Ohno S. Evolution by gene duplication. Springer-Verlag Berlin Heidelberg; 1970.

[46] Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, et al. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nature Genetics. 2006 Jan;38:107–111. https://doi.org/10.1038/ng1700.

[47] Protas M, Conrad M, Gross JB, Tabin C, Borowsky R. Regressive evolution in the Mexican cave tetra, Astyanax mexicanus. Current Biology. 2007 Mar;17:452–454. https://doi.org/10.1016/j.cub.2007.01.051.

[48] Gross JB, Borowsky R, Tabin CJ. A novel role for Mc1r in the parallel evolution of depigmentation in independent populations of the cavefish Astyanax mexicanus. PLoS Genetics. 2009 Jan;5:e1000326. https://doi.org/10.1371/journal.pgen.1000326.

[49] Burns JJ. Missing step in man, monkey and guinea pig required for the biosynthesis of L-ascorbic acid. Nature. 1957 Sep;180:553. https://doi.org/10.1038/180553a0.

[50] Nishikimi M, Koshizaka T, Ozawa T, Yagi K. Occurrence in humans and guinea pigs of the gene related to their missing enzyme L-gulono-gamma-lactone oxidase. Archives of Biochemistry and Biophysics. 1988 Dec;267:842–846. https://doi.org/10.1016/0003-9861(88)90093-8.

[51] Drouin G, Godin JR, Pagé B. The genetics of vitamin C loss in vertebrates. Current Genomics. 2011 Aug;12:371–378. https://doi.org/10.2174/138920211796429736.

[52] Aboobaker AA, Blaxter ML. Hox gene loss during dynamic evolution of the nematode cluster. Current Biology. 2003;13(1):37–40.

[53] Heger P, Marin B, Schierenberg E. Loss of the insulator protein CTCF during nematode evolution. BMC Molecular Biology. 2009;10(1):84.

[54] Hecker N, Sharma V, Hiller M. Convergent gene losses illuminate metabolic and physiological changes in herbivores and carnivores. Proceedings of the National Academy of Sciences of the United States of America. 2019 Feb;116:3036–3041. https://doi.org/10.1073/pnas.1818504116.

[55] Huelsmann M, Hecker N, Springer MS, Gatesy J, Sharma V, Hiller M. Genes lost during the transition from land to water in cetaceans highlight genomic changes associated with aquatic adaptations. Science Advances. 2019 Sep;5:eaaw6671. https://doi.org/10.1126/sciadv.aaw6671.

[56] Indrischek H, Hammer J, Machate A, Hecker N, Kirilenko B, Roscito J, et al. Vision-related convergent gene losses reveal SERPINE3's unknown role in the eye. eLife. 2022 Jun;11. https://doi.org/10.7554/eLife.77999.

[57] Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755–63.

[58] Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, et al. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 2015

Oct;526:569–573. https://doi.org/10.1038/nature15697.

[59] Nakahashi H, Kieffer Kwon KR, Resch W, Vian L, Dose M, Stavreva D, et al. A genome-wide map of CTCF multivalency redefines the CTCF code. Cell Reports. 2013 May;3:1678–1689. https://doi.org/10.1016/j.celrep.2013.04.024.

[60] Wang N, Kimball RT, Braun EL, Liang B, Zhang Z. Assessing phylogenetic relationships among Galliformes: a multigene phylogeny with expanded taxon sampling in Phasianidae. PLoS One. 2013;8:e64312. https://doi.org/10.1371/journal.pone.0064312.

[61] Weir JT, Haddrath O, Robertson HA, Colbourne RM, Baker AJ. Explosive ice age diversification of kiwi. Proceedings of the National Academy of Sciences of the United States of America. 2016 Sep;113:E5580–E5587. https://doi.org/10.1073/pnas.1603795113.

[62] Kapusta A, Suh A. Evolution of bird genomes—a transposon's-eye view. Annals of the New York Academy of Sciences. 2017 Feb;1389:164–185. https://doi.org/10.1111/nyas.13295.

[63] Carotti E, Tittarelli E, Canapa A, Biscotti MA, Carducci F, Barucca M. LTR Retroelements and Bird Adaptation to Arid Environments. International Journal of Molecular Sciences. 2023 Mar;24. https://doi.org/10.3390/ijms24076332.

[64] Jurka J. Repbase update: a database and an electronic journal of repetitive elements. Trends in Genetics. 2000 Sep;16:418–420. https://doi.org/10.1016/s0168-9525(00)02093-x.

[65] Jurka J, Zietkiewicz E, Labuda D. Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. Nucleic Acids Research. 1995 Jan;23:170–175. https://doi.org/10.1093/nar/23.1.170.

[66] Suh A, Churakov G, Ramakodi MP, Platt RN, Jurka J, Kojima KK, et al. Multiple lineages of ancient CR1 retroposons shaped the early genome evolution of amniotes. Genome Biology and Evolution. 2014 Dec;7:205–217. https://doi.org/10.1093/gbe/evu256.

[67] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 2007 Aug;24:1586–1591. https://doi.org/10.1093/molbev/msm088.

[68] Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. Pairwise comparisons across species are problematic when analyzing functional genomic data. Proceedings of the National Academy of Sciences of the United States of America. 2018 Jan;115:E409–E417. https://doi.org/10.1073/pnas.1707515115.

[69] Suh A, Paus M, Kiefmann M, Churakov G, Franke FA, Brosius J, et al. Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. Nature Communications. 2011 Aug;2:443. https://doi.org/10.1038/ncomms1448.

[70] Fitzpatrick JL, Kahrl AF, Snook RR. SpermTree, a species-level database of sperm morphology spanning the animal tree of life. Scientific Data. 2022 Jan;9:30. https://doi.org/10.1038/s41597-022-01131-w.

[71] Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research. 1998;8(3):186–94.

[72] Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Research. 1998;8(3):195–202.

[73] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.

[74] Galtier N, Gouy M, Gautier C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Computer Applications in the Biosciences. 1996;12(6):543–8.

[75] Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics. 2000;16(6):276–7.

[76] Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research. 2005;33(2):511–8.

[77] Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences of the United States of America. 2020 Apr;117:9451–9457. https://doi.org/10.1073/pnas.1921046117.

[78] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006 Jul;22(13):1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

[79] Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584. https://doi.org/10.7717/peerj.2584.

[80] Crescente JM, Zavallo D, Helguera M, Vanzetti LS. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics. 2018 Oct;19:348. https://doi.org/10.1186/s12859-018-2376-y.

[81] Ranwez V, Harispe S, Delsuc F, Douzery EJP. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PloS One. 2011;6:e22594. https://doi.org/10.1371/journal.pone.0022594.

[82] Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution. 2000 Apr;17:540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

[83] Pugacheva EM, Suzuki T, Pack SD, Kosaka-Suzuki N, Yoon J, Vostrov AA, et al. The structural complexity of the human BORIS gene in gametogenesis and cancer. PloS One. 2010 Nov;5:e13872. https://doi.org/10.1371/journal.pone.0013872.

[84] Hao Y, Xiong Y, Cheng Y, Song G, Jia C, Qu Y, et al. Comparative transcriptomics of 3 high-altitude passerine birds and their low-altitude relatives. Proceedings of the National Academy of Sciences. 2019;116(24):11851–11856.

[85] Helsen J, Voordeckers K, Vanderwaeren L, Santermans T, Tsontaki M, Verstrepen KJ, et al. Gene Loss Predictably Drives Evolutionary Adaptation. Molecular Biology and Evolution. 2020 07;37(10):2989–3002. https://doi.org/10.1093/molbev/msaa172. https://academic.oup.com/mbe/article-pdf/37/10/2989/33830490/msaa172.pdf.

[86] 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 Mar;9:1215. https://doi.org/10.1038/s41467-018-03667-1.

[87] Wen ZY, Kang YJ, Ke L, Yang DC, Gao G. Genome-Wide Identification of Gene Loss Events Suggests Loss Relics as a Potential Source of Functional lncRNAs in Humans. Molecular Biology

and Evolution. 2023 May;40. https://doi.org/10.1093/molbev/msad103.

[88] Liu A, He F, Shen L, Liu R, Wang Z, Zhou J. Convergent degeneration of olfactory receptor gene repertoires in marine mammals. BMC Genomics. 2019 Dec;20:977. https://doi.org/10.1186/s12864-019-6290-0.

[89] Liu J, Shu M, Liu S, Xue J, Chen H, Li W, et al. Differential MC5R loss in whales and manatees reveals convergent evolution to the marine environment. Development Genes and Evolution. 2022 Aug;232:81–87. https://doi.org/10.1007/s00427-022-00688-1.

[90] Meredith RW, Gatesy J, Murphy WJ, Ryder OA, Springer MS. Molecular decay of the tooth gene Enamelin (ENAM) mirrors the loss of enamel in the fossil record of placental mammals. PLoS Genetics. 2009 Sep;5:e1000634. https://doi.org/10.1371/journal.pgen.1000634.

[91] Meredith RW, Gatesy J, Cheng J, Springer MS. Pseudogenization of the tooth gene enamelysin (MMP20) in the common ancestor of extant baleen whales. Proceedings Biological Sciences. 2011 Apr;278:993–1002. https://doi.org/10.1098/rspb.2010.1280.

[92] Sharma V, Walther F, Hecker N, Stuckas H, Hiller M. Convergent Losses of TLR5 Suggest Altered Extracellular Flagellin Detection in Four Mammalian Lineages. Molecular biology and evolution. 2020 03;37. https://doi.org/10.1093/molbev/msaa058.

[93] Águeda Pinto A, Alves LQ, Neves F, McFadden G, Jacobs BL, Castro LFC, et al. Convergent Loss of the Necroptosis Pathway in Disparate Mammalian Lineages Shapes Viruses Countermeasures. Frontiers in Immunology. 2021;12:747737. https://doi.org/10.3389/fimmu.2021.747737.

[94] Velova H, Gutowska-Ding M, Burt D, Vinkler M. Toll-Like Receptor Evolution in Birds: Gene Duplication, Pseudogenization, and Diversifying Selection. Molecular biology and evolution. 2018 06;35. https://doi.org/10.1093/molbev/msy119.

[95] Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nature Reviews Genetics. 2020 Mar;21(3):191–201.

[96] Kovalenko T, Patrushev L. Pseudogenes as Functionally Significant Elements of the Genome. Biochemistry (Moscow). 2018 11;83:1332–1349. https://doi.org/10.1134/S0006297918110044.

[97] Pycraft W. On the Morphology and Phylogeny of the Palæognathæ (Ratitæand Crypturi) and Neognathæ (Carinatæ). The Transactions of the Zoological Society of London. 2010 07;15:149 – 290. https://doi.org/10.1111/j.1096-3642.1900.tb00023.x.

[98] Gussekloo S, Berthaume M, Pulaski D, Westbroek I, Waarsing J, Heinen R, et al. Functional and evolutionary consequences of cranial fenestration in birds. Evolution. 2017 02;71. https://doi.org/10.1111/evo.13210.

[99] Gussekloo SWS, Zweers GA. The Paleognathous Pterygoid-Palatinum Complex. a True Character? Netherlands Journal of Zoology. 1999;49(1):29 – 43. https://doi.org/10.1163/156854299X00038.

[100] Hu H, Sansalone G, Wroe S, Mcdonald P, O'Connor J, li Z, et al. Evolution of the vomer and its implications for cranial kinesis in Paraves. Proceedings of the National Academy of Sciences. 2019 09;116:201907754. https://doi.org/10.1073/pnas.1907754116.

[101] Grejo Setti P, Deon G, Santos R, Goes C, Garnero A, Gunski R, et al. Evolution of bird sex chromosomes: a cytogenomic approach in Palaeognathae species. BMC Ecology and Evolution. 2024 04;24:51. https://doi.org/10.1186/s12862-024-02230-5.

[102] Schartl M, Schmid M, Nanda I. Dynamics of vertebrate sex chromosome evolution: from equal size to giants and dwarfs. Chromosoma. 2016 06;125. https://doi.org/10.1007/s00412-015-0569-y.

[103] Griffin D, Kretschmer R, Srikulnath K, Singchat W, O'Connor R, Romanov M. Insights into avian molecular cytogenetics—with reptilian comparisons. Molecular Cytogenetics. 2024 10;17:24. https://doi.org/10.1186/s13039-024-00696-y.

[104] Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, et al. Convergent transcriptional specializations in the brains of humans and song-learning birds. Science. 2014;346(6215):1256846. https://doi.org/10.1126/science.1256846. https://www.science.org/doi/pdf/10.1126/science.1256846.

[105] Suzuki M, Gerstein M, Yagi N. Stereochemical basis of DNA recognition by Zn fingers. Nucleic Acids Research. 1994;22(16):3397–405.

[106] Swanson WJ, Vacquier VD. The rapid evolution of reproductive proteins. Nature Reviews Genetics. 2002 Feb;3:137–144. https://doi.org/10.1038/nrg733.

[107] Wilburn DB, Swanson WJ. From molecules to mating: Rapid evolution and biochemical studies of reproductive proteins. Journal of proteomics. 2016 Mar;135:12–25.

[108] Dapper AL, Wade MJ. Relaxed Selection and the Rapid Evolution of Reproductive Genes. Trends in genetics : TIG. 2020 Sep;36:640–649.

[109] Ramm SA, Oliver PL, Ponting CP, Stockley P, Emes RD. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. Molecular Biology and Evolution. 2008 Jan;25:207–219. https://doi.org/10.1093/molbev/msm242.

[110] Schumacher J, Rosenkranz D, Herlyn H. Mating systems and protein-protein interactions determine evolutionary rates of primate sperm proteins. Proceedings Biological Sciences. 2014 Jan;281:20132607. https://doi.org/10.1098/rspb.2013.2607.

[111] Gozashti L, Corbett-Detig R, Roy SW. Evolutionary rates of testes-expressed genes differ in monogamous and promiscuous Peromyscus species. bioRxiv. 2021;https://doi.org/10.1101/2021.04.21.440792. https://www.biorxiv.org/content/early/2021/04/29/2021.04.21.440792.full.pdf.

[112] Kopania EEK, Thomas GWC, Hutter CR, Mortimer SME, Callahan CM, Roycroft E, et al. Molecular evolution of male reproduction across species with highly divergent sperm morphology in diverse murine rodents. bioRxiv. 2023;https://doi.org/10.1101/2023.08.30.555585. https://www.biorxiv.org/content/early/2023/09/03/2023.08.30.555585.full.pdf.

[113] Martinez G, Garcia C. Sexual selection and sperm diversity in primates. Molecular and Cellular Endocrinology. 2020 Dec;518:110974. https://doi.org/10.1016/j.mce.2020.110974.

[114] Anderson MJ, Dixson AF. Sperm competition: motility and the midpiece in primates. Nature. 2002 Apr;416:496. https://doi.org/10.1038/416496a.

[115] Rowe M, Albrecht T, Cramer ERA, Johnsen A, Laskemoen T, Weir JT, et al. Postcopulatory sexual selection is associated with accelerated evolution of sperm morphology. Evolution. 2015 Apr;69:1044–1052. https://doi.org/10.1111/evo.12620.

[116] Omotoriogun TC, Laskemoen T, Rowe M, Albrecht T, Bowie RCK, Sedláček O, et al. Variation in sperm morphology among Afrotropical sunbirds. Ibis. 2016;158(1):155–166. https://doi.org/https://doi.org/10.1111/ibi.12334. https://onlinelibrary.wiley.com/doi/pdf/10.1111/ibi.12334.

[117] Carballo L, Battistotti A, Teltscher K, Lierz M, Bublat A, Valcu M, et al. Sperm morphology and evidence for sperm competition among parrots. Journal of Evolutionary Biology. 2019 Aug;32:856–867. https://doi.org/10.1111/jeb.13487.

[118] Brouwer L, Griffith SC. Extra-pair paternity in birds. Molecular Ecology. 2019 Nov;28:4864–4882. https://doi.org/10.1111/mec.15259.

[119] Valdez DJ. An updated look at the mating system, parental care and androgen seasonal variations in ratites. General and Comparative Endocrinology. 2022 Jul;323-324:114034. https://doi.org/10.1016/j.ygcen.2022.114034.

[120] Immler S, Saint-Jalme M, Lesobre L, Sorci G, Roman Y, Birkhead TR. The evolution of sperm morphometry in pheasants. Journal of Evolutionary Biology. 2007 May;20:1008–1014. https://doi.org/10.1111/j.1420-9101.2007.01302.x.

[121] Briskie JV, Montgomerie R. Sexual Selection and the Intromittent Organ of Birds. Journal of Avian Biology. 1997;28(1):73–86.

[122] Herrera AM, Shuster SG, Perriton CL, Cohn MJ. Developmental basis of phallus reduction during bird evolution. Current Biology. 2013 Jun;23:1065–1074. https://doi.org/10.1016/j.cub.2013.04.062.

[123] Brennan PLR. Genital evolution: cock-a-doodle-don't. Current Biology. 2013 Jun;23:R523–R525. https://doi.org/10.1016/j.cub.2013.04.035.

[124] Brennan PLR. Bird genitalia. Current Biology. 2022 Oct;32:R1061–R1062. https://doi.org/10.1016/j.cub.2022.09.015.

[125] Hooper R, Maher K, Moore K, McIvor G, Hosken D, Thornton A. Ultimate drivers of forced extra-pair copulations in birds lacking a penis: jackdaws as a case-study. Royal Society Open Science. 2024 Mar;11:231226. https://doi.org/10.1098/rsos.231226.

[126] Sanger TJ, Gredler ML, Cohn MJ. Resurrecting embryos of the tuatara, Sphenodon punctatus, to resolve vertebrate phallus evolution. Biology Letters. 2015 Oct;11. https://doi.org/10.1098/rsbl.2015.0694.

[127] Sackton TB, Grayson P, Cloutier A, Hu Z, Liu JS, Wheeler NE, et al. Convergent regulatory evolution and loss of flight in paleognathous birds. Science. 2019 Apr;364:74–78. https://doi.org/10.1126/science.aat7244.

Supplementary information

Supplementary figures

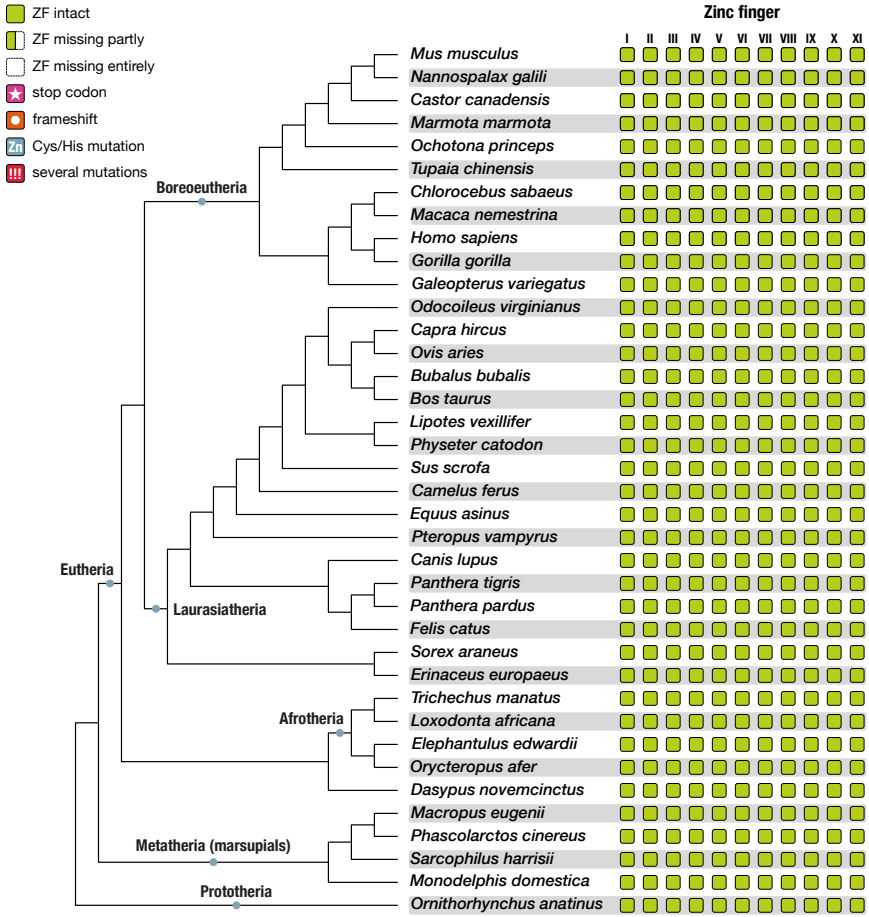


Figure S1: Mammalian BORIS genes possess intact zinc fingers. Left: Cladogram depicting the relationships of 38 mammalian species. Tree topology after https://phylotastic.org/ and Prum et al. [58], Wang et al. [60]. Major mammalian lineages are indicated by grey dots. **Right:** Presence and state of the eleven C_2H_2 zinc fingers of CTCFL/BORIS (ZF I–XI), mapped onto the mammalian phylogeny (left). Intact and complete zinc fingers are depicted as green squares, damaged zinc fingers are classified by icons (see legend).

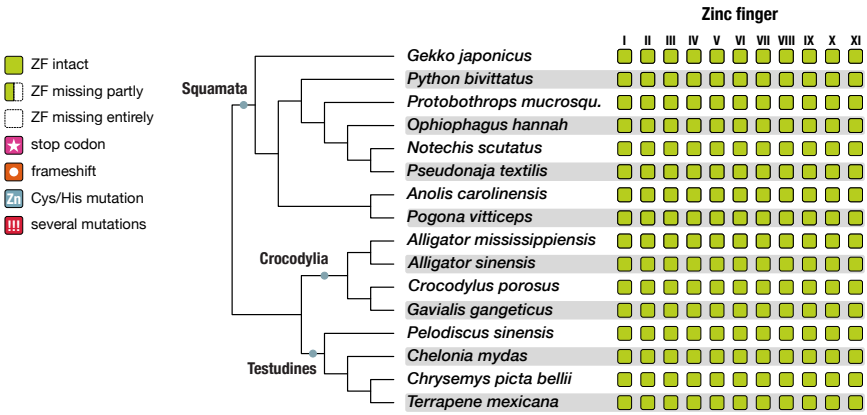


Figure S2: Reptilian BORIS genes possess intact zinc fingers. Left: Cladogram depicting the relationships of 16 reptilian species. *Protobothrops mucrosqu.*: *Protobothrops mucrosquamatus*. Tree topology after https://phylotastic.org/ and Prum et al. [58], Wang et al. [60]. Major reptilian lineages are indicated by grey dots. **Right:** Presence and state of the eleven C_2H_2 zinc fingers of CTCFL (ZF I–XI), mapped onto the reptilian phylogeny (left). Intact and complete zinc fingers are depicted as green squares, damaged zinc fingers are classified by icons (see legend).

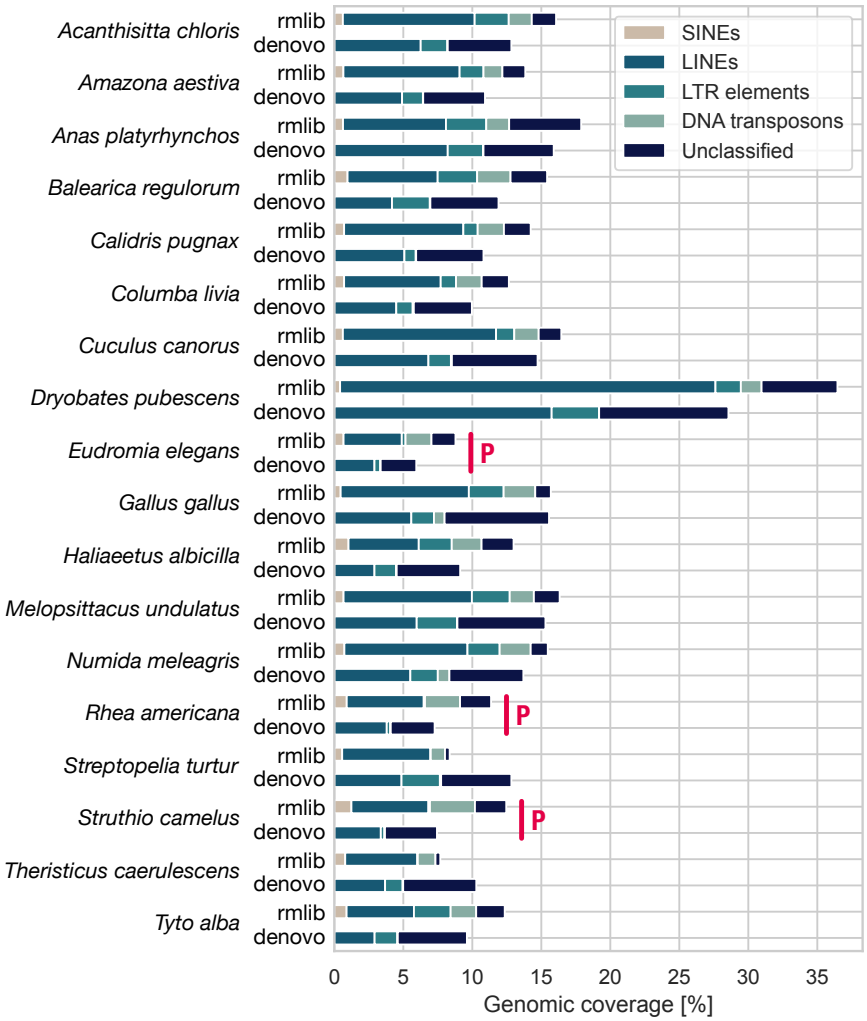


Figure S3: Percentage of frequent repeat element classes in selected bird genomes. Stacked bar graph depicting the genomic coverage (in percent) of the five most abundant repeat element classes in 18 bird species (in alphabetical order). For each species, data from two independent repeat detection approaches are plotted, using the RepeatMasker-based repeat library («rmlib») and the de novo repeat library («denovo»). Three paleognathous bird species are highlighted by a red «P», all other species are neognathous. As data basis, we used parsed RepeatMasker summary output files for all 18 species (Supplementary Table S3), separately generated for de novo- and RepeatMasker library-based detection.

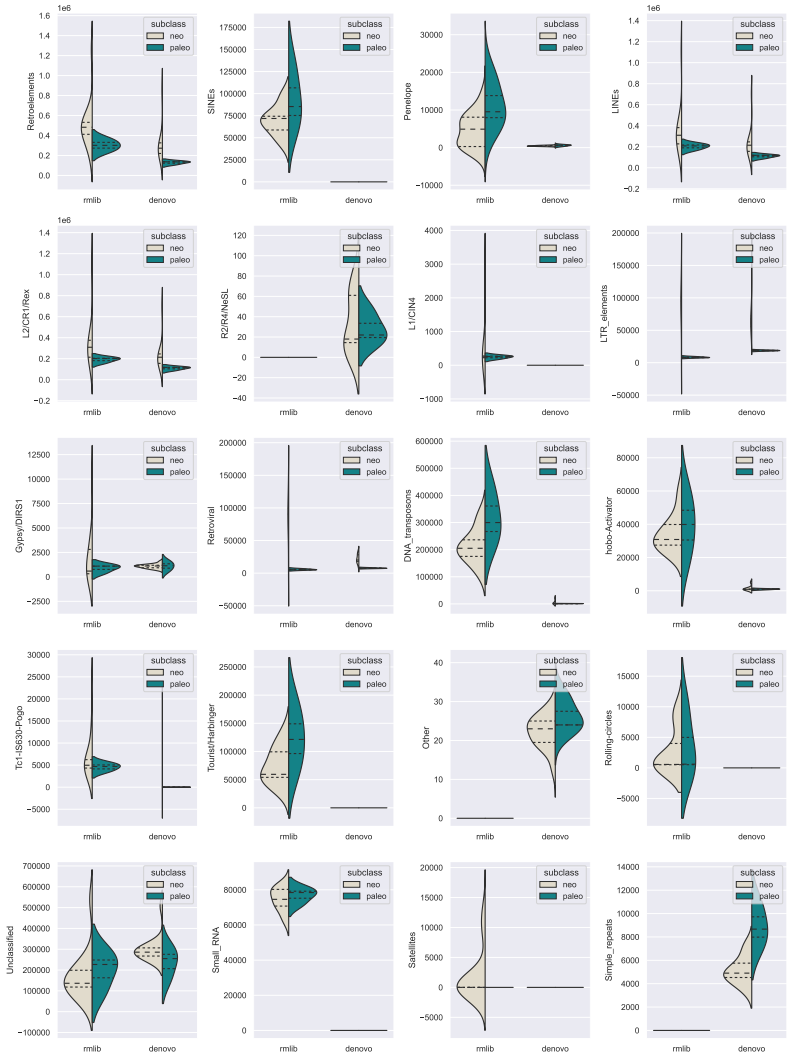


Figure S4: Repeat element landscape of selected birds. Violin plots depicting the distribution of repeat element abundance for 20 different repeat classes, as indicated in y-axis labels. Each subplot represents data from two independent repeat detection approaches, using the RepeatMasker-based repeat library (at the left) and the de novo repeat library (at the right). The RepeatMasker-based and de novo violin plots themselves are split to display differences between neognathous (left: creme) and paleognathous birds (right: teal). As data basis, we used parsed RepeatMasker summary output files for 15 neognathous and three paleognathous bird species (Supplementary Table S3), separately generated for de novo- and RepeatMasker library-based detection. Several repeat classes could not be identified by either RepeatMasker library and are omitted from the plot (e. g. CRE/SLACS, R1/LOA/Jockey, BEL/Pao, Ty1/Copia, En-Spm, or PiggyBac elements). Dashed lines represent interquartile range and median.

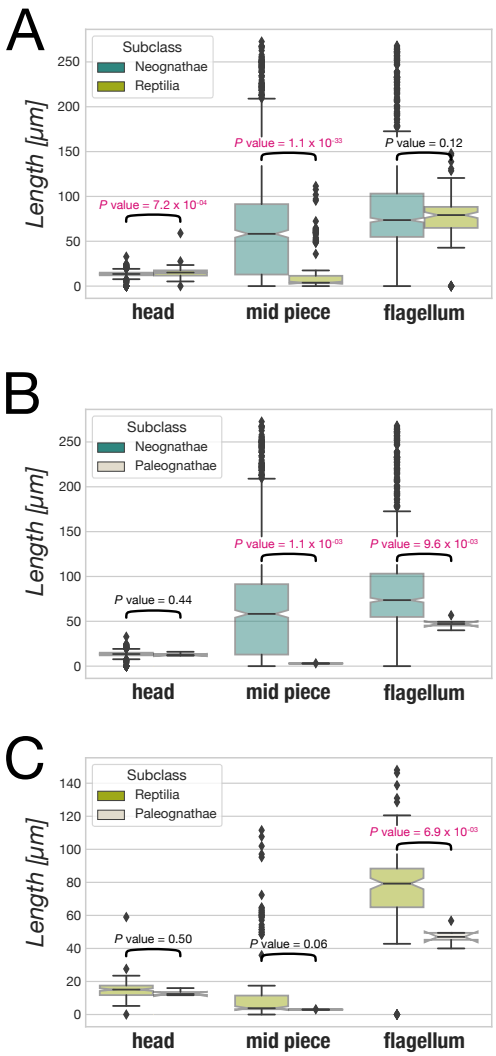


Figure S5: Comparison of morphological characteristics of sperm from Neognathae, Paleognathae, and Reptilia. The lengths of sperm heads (left), mid pieces (middle), and flagella (right) from neognathous birds (554 species) are compared to the corresponding data of Reptilia (120 species, **A**) and Paleognathae (2 species, **B**). In addition, measurements from Reptilia are compared with those of paleognathous birds (**C**). All data are taken from the SpermTree Database (https://spermtree.org/ [70]). *P* values < 0.05 are highlighted in red and indicate that a morphological trait differs significantly between the compared groups, as determined by Kruskal-Wallis tests.

Supplementary tables

Table S1: List of 59 investigated bird species. Species information and accession numbers of genomic scaffolds containing the BORIS syntenic region from 59 bird species, part 1. Symbols in column 4 denote if a scaffold includes the syntenic block from genes PCK1 to RBM38 (<>) or is partially complete with only the PCK1 (<) or the RBM38 end (>) present. $: taken from Sackton et al. [127].

Species	Lineage	Accession	Scaffold
Acanthisitta chloris	Neognathae; Passeriformes	NW_019776521.1	<>
Amazona aestiva	Neognathae; Psittaciformes	LMAW01000002.1	<>
Anas platyrhynchos	Neognathae; Anseriformes	NC_051792.1	<>
Apaloderma vittatum	Neognathae; Trogoniformes	NW_009709789.1	<>
Aptenodytes forsteri	Neognathae; Sphenisciformes	NW_008794747.1	<>
Apteryx haastii$	Paleognathae; Apterygiformes	PTFD01000001.1	<>
Apteryx owenii$	Paleognathae; Apterygiformes	PTFC01000003.1	<>
Aquila chrysaetos	Neognathae; Accipitriformes	NW_010972709.1	<>
Balearica regulorum	Neognathae; Gruiformes	NW_010747151.1	<>
Calidris pugnax	Neognathae; Charadriiformes	NW_015090780.1	<>
Callipepla squamata	Neognathae; Galliformes	MCFN01000399.1	<>
Calypte anna	Neognathae; Apodiformes	NW_007620763.1	<>
Caprimulgus carolinensis	Neognathae; Caprimulgiformes	JMFU01067627.1	<>
Cariama cristata	Neognathae; Cariamiformes	NW_009636176.1	<>
Casuarius casuarius$	Paleognathae; Casuariiformes	PTFA01000086.1	>
Chaetura pelagica	Neognathae; Apodiformes	NW_009953486.1	<>
Colius striatus	Neognathae; Coliiformes	NW_010701813.1	<>
Columba livia	Neognathae; Columbiformes	NW_004973200.1	<>
Corvus brachyrhynchos	Neognathae; Passeriformes	NW_008236180.1	<>
Corvus cornix	Neognathae; Passeriformes	NW_018113990.1	<>
Coturnix japonica	Neognathae; Galliformes	NC_029535.1	<>
Crypturellus cinnamomeus$	Paleognathae; Tinamiformes	PTEZ01000102.1	<>
Cuculus canorus	Neognathae; Cuculiformes	NW_009243347.1	<>
Cyanistes caeruleus	Neognathae; Passeriformes	NW_019776521.1	<>
Dromaius novaehollandiae$	Paleognathae; Casuariiformes	PTEY01000246.1	<
Egretta garzetta	Neognathae; Pelecaniformes	NW_009260622.1	<>
Eudromia elegans$	Paleognathae; Tinamiformes	PTEX01000028.1	<>
Eurypyga helias	Neognathae; Gruiformes	JJRO01094590.1	<>
Falco cherrug	Neognathae; Falconiformes	NW_004994897.1	<>
Falco peregrinus	Neognathae; Falconiformes	NW_004930514.1	<>
Gallus gallus	Neognathae; Galliformes	NC_006107.5	<>
Gavia stellata	Neognathae; Gaviiformes	NW_009295348.1	<>
Haliaeetus albicilla	Neognathae; Accipitriformes	NW_009767762.1	<>
Haliaeetus leucocephalus	Neognathae; Accipitriformes	NW_010972709.1	<>
Lepidothrix coronata	Neognathae; Passeriformes	NW_016690229.1	<>

Table S2: List of 59 investigated bird species, continued. Species information and accession numbers of genomic scaffolds containing the BORIS syntenic region from 59 bird species, part 2. Symbols in column 4 denote if a scaffold includes the syntenic block from genes PCK1 to RBM38 (<>) or is partially complete with only the PCK1 (<) or the RBM38 end (>) present. $: taken from Sackton et al. [127].

Species	Lineage	Accession	Scaffold
Lonchura striata	Neognathae; Passeriformes	NW_018657563.1	<>
Melopsittacus undulatus	Neognathae; Psittaciformes	NC_034427.1	<>
Merops nubicus	Neognathae; Coraciiformes	JJRJ01051273.1	<>
Mesitornis unicolor	Neognathae; Gruiformes	NW_010159074.1	<>
Nestor notabilis	Neognathae; Psittaciformes	NW_009919459.1	<>
Nipponia nippon	Neognathae; Pelecaniformes	NW_009000645.1	<>
Nothoprocta perdicaria$	Paleognathae; Tinamiformes	PTEW01000004.1	>
Numida meleagris	Neognathae; Galliformes	NC_034427.1	<>
Opisthocomus hoazin	Neognathae; Opisthocomiformes	NW_009901057.1	<>
Parus major	Neognathae; Passeriformes	NC_031788.1	<>
Patagioenas fasciata	Neognathae; Columbiformes	LSYS01003456.1	<>
Pelecanus crispus	Neognathae; Pelecaniformes	JJRG01105365.1	<>
Phaethon lepturus	Neognathae; Pelecaniformes	NW_010546123.1	<>
Dryobates pubescens	Neognathae; Piciformes	NW_009664594.1	<>
Pseudopodoces humilis	Neognathae; Passeriformes	NW_005087575.1	<>
Pterocnemia pennata$	Paleognathae; Rheiformes	PTJI01000154.1	<>
Pygoscelis adeliae	Neognathae; Sphenisciformes	NW_008825076.1	<>
Rhea americana$	Paleognathae; Rheiformes	PTEV01000005.1	<>
Serinus canaria	Neognathae; Passeriformes	NW_007931143.1	<>
Struthio camelus	Paleognathae; Struthioniformes	NW_009271896.1	<>
Sturnus vulgaris	Neognathae; Passeriformes	NW_014650517.1	<>
Tauraco erythrolophus	Neognathae; Musophagiformes	NW_010041494.1	<>
Tinamus guttatus	Paleognathae; Tinamiformes	NW_010578138.1	<>
Tyto alba	Neognathae; Strigiformes	JJRD01144017.1	<>

Table S3: List of 18 bird species used for repeat analysis. Species information and NCBI nucleotide accession numbers of 18 genome assemblies used for repeat analysis are shown.

Species	Lineage	Accession
Acanthisitta chloris	Neognathae; Passeriformes	JAFCHQ000000000
Amazona aestiva	Neognathae; Psittaciformes	JAESHV000000000
Anas platyrhynchos	Neognathae; Anseriformes	JACEUM000000000
Balearica regulorum	Neognathae; Gruiformes	JAAIYC000000000
Calidris pugnax	Neognathae; Charadriiformes	LDEH00000000
Columba livia	Neognathae; Columbiformes	AKCR00000000
Cuculus canorus	Neognathae; Cuculiformes	JAGIYT000000000
Dryobates pubescens	Neognathae; Piciformes	JACNMV000000000
Eudromia elegans	Paleognathae; Tinamiformes	PTEX00000000
Gallus gallus	Neognathae; Galliformes	AADN00000000
Haliaeetus albicilla	Neognathae; Accipitriformes	VZSQ00000000
Melopsittacus undulatus	Neognathae; Psittaciformes	JAAVWG000000000
Numida meleagris	Neognathae; Galliformes	JABXER000000000
Rhea americana	Paleognathae; Rheiformes	PTEV00000000
Streptopelia turtur	Neognathae; Columbiformes	CABFKC000000000 ?
Struthio camelus	Paleognathae; Struthioniformes	JJRT00000000
Theristicus caerulescens	Neognathae; Pelecaniformes	JAJGSR000000000
Tyto alba	Neognathae; Strigiformes	JAEUGV000000000

Table S4: Sequences from diapsid amniotes used to generate the BORIS hidden Markov model. Origin, NCBI accession number, and length of the nine bird and seven reptilian protein sequences used for HMM generation are shown.

Species	**Lineage**	**Seq-ID**	**[AA]**
	Birds		
Amazona aestiva	Aves; Psittaciformes	LMAW01000002.1	465
Coturnix japonica	Aves; Galliformes	XP_015737575.1	401
Empidonax traillii	Aves; Passeriformes	XP_027737053.1	508
Gavia stellata	Aves; Gaviiformes	KFV41672.1	234
Lepidothrix coronata	Aves; Passeriformes	XP_017670709.1	451
Meleagris gallopavo	Aves; Galliformes	XP_019477849.1	612
Numida meleagris	Aves; Galliformes	XP_021272595.1	545
Pseudopodoces humilis	Aves; Passeriformes	XP_014109495.1	459
Serinus canaria	Aves; Passeriformes	XP_030088730.1	446
	Reptiles		
Alligator mississippiensis	Archelosauria; Crocodylia	KYO42050.1	551
Anolis carolinensis	Lepidosauria; Squamata	XP_016850798.1	581
Chelonia mydas	Testudines; Cryptodira	XP_027678081.1	665
Crocodylus porosus	Archelosauria; Crocodylia	XP_019403720.1	510
Pelodiscus sinensis	Testudines; Cryptodira	XP_014430204.2	664
Pogona vitticeps	Lepidosauria; Squamata	XP_020653141.1	554
Python bivittatus	Lepidosauria; Squamata	XP_025022470.1	672

Supplementary files

1. File S1 — Multiple sequence alignment of RBM38 proteins from 54 birds:
   Supplementary_File1.aln (Multiple sequence alignment in Clustalw's .aln format)
2. File S2 — Multiple sequence alignment of PCK1 proteins from 76 birds:
   Supplementary_File2.aln (Multiple sequence alignment in Clustalw's .aln format)
3. File S3 — 18 birds repeat masker summary:
   Supplementary_File3.xlsx (Spreadsheet in .xlsx format)
4. File S4 — 18 birds repeats BORIS loci:
   all_boris_repeats_table.tsv (Table in .tsv format)

6 Discussion

The introductory section of this thesis offers an overview of the mechanisms underlying gene duplication and the evolutionary forces that drive the diversification of gene families. These concepts are illustrated through three case studies, each highlighting distinct perspectives on the potential fates of duplicated genes, considering both short-term dynamics at the population level and long-term outcomes at the species level.

In zebrafish, the remarkable expansion and homogenization of the NLR gene family, driven by ongoing tandem duplications, promote gene conversion events that generate an extensive repertoire of highly similar immune genes. The zebrafish NLR-ome exhibits extensive copy number variation, particularly in wild populations, underscoring how rapid gene family evolution can directly influence population structure and adaptive capacity over short evolutionary timescales. **?** compared the observed status of the zebrafish NLR gene family with models of different evolutionary dynamics and demonstrated that the substantial copy number variation in wild populations is best accounted for by the compound copy model (CCM), in which multiple gene copies act as templates for duplication in each generation, driving rapid increases in gene copy number and promoting sequence homogenization.
Alltogether these findings might be an indication for the pronounced impact of concerted evolution, as discussed in chapter 3.2, on the development of this gene family.
The available data do not permit an evaluation of NLR gene loss in the analyzed wild populations. Accordingly, the role of gene loss in shaping the zebrafish NLR-ome remains undetermined. This question could be addressed through analyses of high-quality genomes generated by long-read sequencing and could be initiated with the investigation of publicly available reference assemblies, adressing quality and quantity of NLR pseudogene repertoires in zebrafish genomes. Even when accounting for the observed differences between laboratory and wild populations, such studies may provide tentative evidence regarding the potential role of functional losses in NLR gene evolution.

The study of odorant receptor (OR) repertoires in the tenebrionid beetle *Carchares macer* offers insights into both short-term population-level dynamics within *C. macer* and long-term species-level evolution through comparison with OR repertoires of other beetle species published by previous authors. The *C. macer* OR repertoire displays extensive individual and population specific diversity.Notably, the finding that only a minority of OR genes are universally present across populations reflects the pan-genome concept (**?**)

of the OR gene family. All identified OR genes cluster with previously characterized OR genes from other beetle species and constitute a distinct group that is separate from the ancestral gustatory receptors. The *C. macer* ORs can be assigned to subgroups defined in previous studies (**??**), with one subgroup showing a lineage-specific expansion, while three of the seven previously reported subgroups lack *C. macer* OR genes. In addition, the conserved ORCO was identified in all analyzed genomes. The presence of ORCO across all samples, alongside with the consistent group assignments, indicates that our results provide a broad representation of the OR repertoire in *C. macer*.

Previous authors like **???** or **?** proposed that birth-and-death evolution predominantly shapes chemosensory receptor gene repertoires, due to observations of substantial gene gains and losses within this families.

Our findings, that a single subgroup shows expansion while others are entirely absent, are consistent with a common pattern in beetle OR repertoires (**?**) and may indicate gene loss events during the evolution of OR genes in this species. Furthermore, the striking individual diversity of OR gene repertoires is consistent with the hypothesis that the birth-and-death evolutionary pattern might also be extendable to the individual level.

Previous authors have identified significant proportions of OR pseudogenes in insect genomes (**???**) supporting the hypothesis that insect OR evolution is predominantly driven by birth-and death dynamics. Analogous to the case of zebrafish NLR genes, the investigation of pseudogenes might offer further insights into the roles of gene loss and pseudogenisation in shaping the evolution of the OR repertoire.

Overall our findings for zebrafish NLRs and beetle ORs challenge assumptions based solely on single reference individuals, emphasizing the importance of broad population-level comparative genomic approaches to fully capture the dynamic nature of evolution of this gene families.

Moreover, the thesis highlights the importance of methodological tailoring to the biological characteristics of the target gene family. For example, while Illumina short-read assemblies turned out to be sufficient for the odorant receptor genes, the extreme sequence similarity of zebrafish NLR genes required targeted long-read (SMRT) sequencing to ensure accurate copy resolution. Although it could be argued that copy number variations of OR genes in *C. macer*, similar to zebrafish NLRs, could still be detected in genomes generated with a long read sequencing approach, the results regarding the OR gene analysis in *C. macer* are consistent with patterns of OR gene repertoires observed in other coleopteran species (**?**).

Analyses of pseudogenes for both Zebrafish NLRs and *C. macer* ORs may likewise yield

further insights into on what extent pseudogenisation and gene loss might contribute to the evolutionary dynamics in both gene families. Therefore it may be informative to consider documented cases of pseudogenisation.

A well-studied example is the long-term pseudogenisation of the BORIS gene in birds, which provides the third case, investigated in the framework of this project. The BORIS (CTCFL) gene in birds exemplifies a process of pseudogenisation over an evolutionary time scale of around 100 Mio years. It reveals recurrent, lineage-specific degradation events in neognathous bird species, while displaying a high level of conservation in palaeognathous birds. The decay correlates with relaxed selective pressures and the accumulation of repetitive elements. Regarding the gene families discussed above, it would be of interest to investigate whether, and to what extent, similar patterns of progressive degradation occur, and to compare these with the pattern observed in BORIS. Such an analysis could further elucidate similarities and differences between degradation processes operating at the population level, representing shorter evolutionary timescales, and those at the species level, reflecting longer timescales.

In summary the three case studies demonstrate a complex dynamic interplay between gene duplication, diversification, and loss, showing that these mechanisms act not in isolation but as intertwined drivers of genetic novelty, where genomes are fluid systems with gene birth, expansion, and death interacting constantly underscoring that the evolutionary fate of gene duplicates is far from linear or uniform. Rather it is context-dependent, with different fates in different lineages.

A key question that persists is why OR gene repertoires display such remarkable individual and population scale diversity, whereas NLR genes exhibit high copy number variation, combined with high sequence similarity, despite both playing important roles in interactions with complex environmental factors, which are pathogens in the case of NLRs and volatile molecules in the case of ORs. The fact that pathogens are themselves subject to evolutionary pressure to adapt in competition with their hosts, while volatile molecules are influenced by complex biotic and abiotic dynamics, could provide important indications for future research directions aimed at achieving a deeper understanding of the dynamics of gene family evolution.
Future research could deepen the integration of functional genomics to clarify the roles of newly duplicated and degenerating genes in shaping phenotypes and adaptation. It

should also expand ecological and time-series population genomics to connect gene family dynamics with ongoing environmental changes. Furthermore, investigations should examine the interplay between gene family evolution and structural as well as regulatory genome variation, and how these processes contribute to macroevolutionary diversification. Furthermore the closer examination of gene–pseudogene proportions within the discussed gene families could provide valuable additional data for improving theoretical evolutionary models.