Towards Big data Comparative Genomics

Rossier Victor

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Département de Biologie Computationnelle et d'Écologie et Évolution

Towards Big data Comparative Genomics

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Victor Rossier

Master en Science Moléculaires du Vivant de l'Université de Lausanne

Jury

Prof. Gilbert Greub, Président
Prof. Christophe Dessimoz, Directeur de thèse
Prof. Marc Robinson-Rechavi, Co-directeur de thèse
Dr Bastien Boussau, Expert
Prof. Robert Waterhouse, Expert

Lausanne
(2022)

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Gilbert	**Greub**
Directeur·trice de thèse	Monsieur	Prof.	Christophe	**Dessimoz**
Co-directeur·trice	Monsieur	Prof.	Marc	**Robinson-Rechavi**
Expert·e·s	Monsieur	Dr	Bastien	**Boussau**
	Monsieur	Prof.	Robert	**Waterhouse**

le Conseil de Faculté autorise l'impression de la thèse de

Victor Rossier

Maîtrise universitaire ès Sciences en sciences moléculaires du vivant, Université de Lausanne

intitulée

Towards Big data comparative genomics

Lausanne, le 4 novembre 2022

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Gilbert Greub

Abstract

Comparative genomics is a powerful approach to study evolution and discover the genetic basis of phenotypes. At the core of this approach lies the ability to differentiate comparable genes across species, the orthologs, from lineage-specific genes arising from gene duplications, the paralogs. However, the recent deluge of next generation sequencing data has turned genomics into a Big data discipline, thus fundamentally challenging comparative genomics methods, in particular the ones to infer orthologs and paralogs. On the other hand, the increasing number of genomes offers new opportunities for biological discovery, as each new sequenced species can be thought of as a privileged access to a unique evolutionary experience. Thus, in the first half of this thesis, I developed two comparative genomics methods to cope with some aspects of the velocity, volume and variety property of Big data. Then, in the second half, I capitalised on these new developments to study two biological systems that benefit particularly from increasing numbers of genomes. In chapter 2, I introduce OMAmer, a fast orthology assignment method based on alignment-free comparisons against gene families and subfamilies. OMAmer can process an entire human proteome (*i.e.* protein-coding gene repertoire) within a few minutes on a laptop and thus should provide the opportunity to close the gap between the increasing rate of genome sequencing and their integration in orthology databases. In chapter 3, I tackle the problem of visualising the evolutionary history of large gene families. To this end, I present Matreex, which combines phylogenetic profiles (for the compact view of gene distributions across species) with gene trees (for the evolutionary component). In chapter 4, I characterise the role of convergent gene duplications in animal venom evolution by contrasting the protein repertoires of 68 venomous and closely related non-venomous species. To this end, I use OMAmer and quality controls to integrate proteomes of heterogeneous quality into orthologous groups in a quick and robust way. In chapter 5, I generalize this comparative genomics approach for genotype-phenotype associations and apply it to seven convergent adaptations in birds. To this end, OMAmer was used to scale-up the inference of orthologs and paralogs for 363 recently released bird genomes. With this dense species sampling, I find convergent hemoglobin duplications in diving birds, which might be linked to the enhanced oxygen metabolism required for prolonged dives. Moreover, I observe hundreds of gene families with convergent gene losses associated with the loss of flight, some of which are associated with forelimb and feather development. I use Matreex to explore these families. Overall, I believe that this work represents a step closer towards Big data comparative genomics.

Résumé

La génomique comparative est une approche puissante pour étudier l'évolution et découvrir la base génétique des phénotypes. Au cœur de cette approche se trouve la capacité de différencier les gènes comparables d'une espèce à l'autre, les orthologues, des gènes spécifiques à une lignée issus de duplications de gènes, les paralogues. Cependant, le récent déluge de données de séquençage de nouvelle génération a transformé la génomique en une discipline de type Big Data, ce qui remet fondamentalement en question les méthodes de génomique comparative, en particulier celles permettant de prédire les orthologues et les paralogues. D'autre part, le nombre croissant de génomes offre de nouvelles possibilités de découverte biologique, car chaque nouvelle espèce séquencée peut être considérée comme un accès privilégié à une expérience évolutive unique. Ainsi, dans la première moitié de cette thèse, j'ai développé deux méthodes de génomique comparative pour faire face à certains aspects de la vélocité, du volume et de la variété des Big data. Ensuite, dans la seconde moitié, j'ai capitalisé sur ces nouveaux développements pour étudier deux systèmes biologiques qui bénéficient particulièrement de l'augmentation du nombre de génomes. Dans le chapitre 2, je présente OMAmer, une méthode rapide d'inférence d'orthologie basée sur des comparaisons sans alignement avec des familles et des sous-familles de gènes. OMAmer peut traiter un protéome (c'est-à-dire le répertoire des gènes codant pour les protéines) humain entier en quelques minutes sur un ordinateur portable et devrait donc permettre de combler le fossé entre le taux croissant de séquençage des génomes et leur intégration dans les bases de données d'orthologie. Dans le chapitre 3, j'aborde le problème de la visualisation de l'histoire évolutive des grandes familles de gènes. À cette fin, je présente Matreex, qui combine des profils phylogénétiques (pour une vue compacte de la distribution des gènes entre espèces) et des arbres génétiques (pour la composante évolutive). Dans le chapitre 4, je tente de caractériser le rôle des duplications de gènes convergents dans l'évolution des venins chez les animaux en contrastant les répertoires de protéines de 68 espèces venimeuses et d'espèces non venimeuses évolutivement proches. Dans ce but, J'utilise OMAmer et des contrôles de qualité pour intégrer des protéomes de qualité hétérogène dans des groupes orthologues de manière rapide et robuste. Dans le chapitre 5, OMAmer a été utilisé pour intensifier l'inférence d'orthologues et de paralogues pour 363 génomes d'oiseaux récemment publiés. Grâce à cet échantillonnage dense d'espèces, je caractérisé le rôle des duplications et des pertes de gènes pour sept adaptations convergentes chez les oiseaux. J'identifie notamment des duplications convergentes de l'hémoglobine chez les oiseaux plongeurs, qui pourraient être liées à

l'augmentation du métabolisme de l'oxygène nécessaire à des plongées prolongées. En outre, j'observe des centaines de familles de gènes présentant des contractions convergentes associées à la perte du vol, dont certaines sont associées au développement des membres antérieurs et des plumes. J'utilise Matreex pour explorer ces familles. Dans l'ensemble, je pense que cette thèse représente un pas de plus vers la génomique comparative Big data.

Acknowledgements

First of all, I would like to sincerely thank Christophe and Marc for giving me the opportunity to complete a PhD (a dream of mine), for their constant support and their expert supervision. I would also like to thank Nicolas Salamin for the free ticket he gave me to enter the academic world starting with an 8-month internship, then a Master's project and finally a PhD offer.

I would like to thank each of my colleagues in my two teams for their kindness and recurrent help during my PhD: Alex, Natasha, Tarcisio, Yannis, Clément Train, Sara, Dave, Adrian, Valentine, Julien, Sina, Natalia, Giulia, Sebastien, Anne, Darren, Tina, Amina and (post-pandemic:) Marina, Sina, Christa and Kenneth. A special thanks to Anna and David who also supervised me during my Master.

I would like to finish by thanking my family and my friends, who supported me during these five years and to my boys, Jules and Louis, who gave me the strength to complete this thesis.

Chapter 1

Introduction

Introduction

Big data refers to increasingly large and complex datasets, which can be characterised by three main components: *velocity*, *volume* and *variety* (Sagiroglu and Sinanc 2013). *Velocity* is a term applied to the rate of data generation or the frequency of data release (from batch to continuous stream). *Volume* refers to the size of the dataset and *variety* to its degree of heterogeneity. Each one of these Big data components challenges existing methods for storing, analysing and visualising data (Sagiroglu and Sinanc 2013). With the development of next generation sequencing (NGS) technologies, genomics has joined social media and astronomy as one of the main Big data disciplines (Stephens et al. 2015; Navarro et al. 2019). Specifically, genomics displays the largest increase in rate of data generation and the highest data heterogeneity, while having currently lower data volumes and less streaming data (Stephens et al. 2015; Navarro et al. 2019). Data analysis presents the main challenge for genomics as extracting relevant information from DNA sequences is a highly complex process involving multiple steps (Stephens et al. 2015). Although assembling DNA pieces into genomes and identifying genes is hard, comparing the genomes of every species pair is far more challenging (Stephens et al. 2015). For example, it has been estimated that we are still six orders of magnitude short in terms of computational power to calculate pairwise genome alignments between 2.5 Myo species (Stephens et al. 2015). The discipline in charge of that task is comparative genomics.

Homology provides the main criterion to compare biological characters. Originally defined as "the same organ in different animals under every variety of form and function" (Owen 1846), homology has been adapted to evolutionary biology to relate characters with shared ancestries. Although often displaying functional similarities (*e.g.* mammary glands in every mammal species), homologous characters can have different functions (*e.g.* human hands and bat wings). In practice however, homology is inferred through similarity such as the specific composition and ordering of hand bones evidences the homology between human hands and bat wings. Homology is applicable to all hierarchical levels of biology, including genes (Ochoterena et al. 2019). Since genes can be conserved across macro-evolutionary scales, comparative genomics mainly focus on genes as evolutionary and functional units (Altenhoff, Glover, and Dessimoz 2019). To read about complications related for instance to alternative splicing and partial homology, *see* (Koonin 2005; Gabaldón and Koonin 2013; Forslund et al. 2018; Linard et al. 2021).

Homology inference of genes is generally based on sequence similarity, with the assumption that homologous genes should have residual sequence similarity as a sign of their shared history (William R. Pearson 2013). Sequence similarity is typically computed with alignment methods, which identify conserved stretches of similar nucleotides or amino acids (Needleman and Wunsch 1970; T. F. Smith and Waterman 1981). To speed-up homology inference, BLAST aligns only the sequence pairs that share similar subsequences of size *k* (*k*-mers or seeds) (Stephen F. Altschul et al. 1990; S. F. Altschul et al. 1997). Indeed, given a precomputed table storing sequences at the indexes of their *k*-mers, all sequences with a given *k*-mer are accessible in constant time using the property of hashing (Leskovec, Rajaraman, and Ullman 2014). Recently, DIAMOND and MMSeqs2 have achieved considerable speed-ups by better handling computers cache memory (Buchfink, Xie, and Huson 2015; Steinegger and Söding 2017). However, the fastest approaches to measure sequence similarities (*i.e.* alignment-free approaches) only compare the *k*-mer contents of sequences (Zielezinski et al. 2017). In addition, despite being less sensitive than alignment-based methods, alignment-free approaches are robust to inversion mutations (Zielezinski et al. 2017).

Nonetheless, homology remains a hypothesis and sequence similarity only informs on the probability that it is correct. Thus, choosing an appropriate similarity threshold to reach a conclusion on homology remains a challenging problem. Indeed, two proteins can be similar due to random factors even in the absence of homology. Thus, a common approach to identify homologs starts by modelling a distribution of similarity scores between non-homologous proteins. Then, sequences with similarities falling in the tail of such distribution are likely homologs (Mitrophanov and Borodovsky 2006). For example, the BLAST *E*-value relies on the extreme value distribution (Gumbel) to model the distribution of local alignment scores assuming residues to be independently and identically distributed (Stephen F. Altschul et al. 1990; Karlin and Altschul 1990). However, these assumptions have been criticised for being unrealistic (W. R. Pearson 1998). For example, transmembrane proteins are more likely to share sequence stretches of hydrophobic amino acids than other proteins, which is not captured by the *E*-value. Thus, empirical distributions have been proposed for more realistic null models and typically result from non-homologous similarities computed from random pairs of real gene sequences, sometimes shuffled to remove remaining signals of homology (W. R. Pearson 1998).

Homology is further divided into orthology and paralogy, whether genes have started diverging at speciation or duplication events, respectively (Fitch 1970). This differentiation is useful for numerous applications (*reviewed in* Glover et al. 2019). For example, orthology is essential in reconstructing species trees because it underlies species evolutionary history. In addition, gene function predictors usually build upon the assumption that orthologs conserve their function longer than paralogs (Gabaldón and Koonin 2013). By contrast, paralogs are often associated with functional innovation and adaptation (Kuzmin, Taylor, and Boone 2021). Thus, correlating duplications (and losses) with phenotypes presents a promising avenue to unveil the molecular basis of phenotypes (S. D. Smith et al. 2020; Nagy et al. 2020).

Methods to infer orthology and paralogy are divided in two main categories (Altenhoff, Glover, and Dessimoz 2019). Tree-based methods extract orthology and paralogy from gene family trees (built using homologous genes). Graph-based approaches are faster and rely on the principle that, when comparing two species, orthologs are more similar than paralogs (paralogs originate from duplications that happened before the speciation). Following this principle, orthologs can be detected with best bidirectional hits (BBH), where pairs of reciprocally closest genes between two genomes are the orthologs (Overbeek et al. 1999). However, when duplications happen after the divergence of the two species, genes can have multiple co-orthologs in the other species (Sonnhammer and Koonin 2002), which are not identified by BBH. Thus, various methodological refinements have been developed to increase the sensitivity of BBH methods (*reviewed in* Altenhoff, Glover, and Dessimoz 2019).

By comparing genomes to identify similarities and differences, we can improve our understanding of gene functions and evolution as conserved genomic regions likely underlie essential functions, while highly divergent regions may result from lineage-specific adaptations (Stephan et al. 2022). In addition, the potential of comparative genomics has increased in light of the ever-increasing number of available genomes sequenced by ambitious initiatives (Rhie et al. 2021; Lewin et al. 2022, 2018). Indeed, we are accessing the largest available evolutionary experiment, where each new sequenced species carries valuable information on how evolution works (S. D. Smith et al. 2020; Blaxter et al. 2022). However, the challenges to comparative genomics posed by the entry of genomics into the Big Data era are equally high.

In this section, I explore these challenges and promising avenues within the 3Vs framework of Big data (*velocity*, *volume*, and *variety*). First, I discuss the concept of *velocity* in the context of the challenges posed to orthology inference by the exponential rate of available

genomes. I identify approaches that map sequences to specific patterns of gene families as the most promising to scale-up orthology inference. Second, I discuss the concept of *volume* when representing large gene families and highlight the advantages of hierarchical ortholog groups (HOGs). Third, I discuss the concept of *variety* around the issue of integrating heterogeneous genomes through the differential treatment of reference and "periphery" genomes.

Scaling-up orthology inference

Perhaps the greatest challenge for comparative genomics is the exponential rate at which genomic data is generated. Compared to other Big data fields, genomics displays the steepest curve, with a doubling rate of about every seven months (Stephens et al. 2015; Navarro et al. 2019). This trend is unlikely to decrease considering that the goal of sequencing all 1.5 Myo eukaryotic species before 2030 has been set (Lewin et al. 2018, 2022). This poses acute challenges to comparative genomics, in particular for the inference of orthologs.

Comparing repertoires of protein-coding genes or proteomes (using one protein isoform per gene) to identify orthologs and paralogs is the cornerstone of comparative genomics. However, exhaustive pairwise comparisons (all-against-all) scales inherently quadratically with the amount of input data. Nonetheless, the vast majority of orthology inference methods relies on all-against-all gene comparisons (Sonnhammer et al. 2014; Linard et al. 2021). Thus, combined with the ever-increasing rate of genome sequencing, current orthology databases manage to integrate only a small fraction of the available proteomes. For instance, orthology data is available for 2’496, 5’090 and 7’284 organisms in OMA, EggNOG and OrthoDB (Altenhoff et al. 2021; Huerta-Cepas et al. 2019; Zdobnov et al. 2021), while 27’412 and 177’157 organisms with complete and permanent draft genomes are currently (2022.08.17) referenced in the Genome Online Database (GOLD) (Mukherjee et al. 2021). Moreover, this discrepancy has increased over the past decade. For example, the number of genomes in OMA has increased only 2.5-fold since 2010 (Altenhoff et al. 2011), whereas the number of organisms referenced in GOLD with complete genomes and permanent drafts has increased 8-fold and 109-fold since the publication date (27.11.2010) of (Altenhoff et al. 2011). Indeed, in 2010, GOLD referenced 3’482 organisms with complete genomes and 1’624 with permanent drafts. Thus, without the development of faster and more scalable orthology inference methods, the gap will stretch out.

Three main strategies have been identified to scale-up orthology inference (Sonnhammer et al. 2014; Altenhoff, Glover, and Dessimoz 2019). The first relies on speeding up the sequence comparison process through optimal use of high-performance computers. The second relies on developing efficient algorithms for pairwise sequence comparisons, and the third relies on the key idea of reducing the number of pairwise comparisons.

All-against-all comparisons between stable gene models can be reused across time and shared across orthology resources (Sonnhammer et al. 2014). SIMAP is a database specifically designed to store such data and is used by EggNOG, among others (Arnold et al. 2014; Huerta-Cepas et al. 2019). OMA and OrthoDB use the same all-against-all data (Altenhoff et al. 2021; Zdobnov et al. 2021). In OMA standalone, users can download precomputed all-against-all data for species referenced in OMA (Altenhoff et al. 2019). Parallelising all-against-all comparisons has also become a standard (*e.g.* Ekseth, Kuiper, and Mironov 2014; Tabari and Su 2017; Kaduk and Sonnhammer 2017; Altenhoff et al. 2021). However, the increasing processing power and storage capabilities made possible by the reduced size of processors (as predicted by Moores' Law) is not sufficient to meet the demand from ever-increasing genomic data, combined with the quadratic nature of all-against-all comparisons.

Faster alternatives to Smith-Waterman and BLAST have been widely adopted to speed-up all-against-all comparisons (T. F. Smith and Waterman 1981; Stephen F. Altschul et al. 1990). SonicParanoid, OrthoDB and OrthoFinder use MMseq2 (Cosentino and Iwasaki 2019; Zdobnov et al. 2021; David M. Emms and Kelly 2019; Steinegger and Söding 2017). As a result, SonicParanoid achieved a 70x-1245x speed-up compared to the original InParanoid algorithm (Cosentino and Iwasaki 2019; Sonnhammer and Östlund 2015). Broccoli and OrthoFinder use DIAMOND (Derelle, Philippe, and Colbourne 2020; Buchfink, Xie, and Huson 2015). SwiftOrtho relies on an *ad-hoc* seed-and-extension algorithm using long *k*-mers in the seeding phase, which is nearly 30 times faster than BLAST (Hu and Friedberg 2019). Alignment-free approaches relying exclusively on *k*-mers have also been experimented despite their low sensitivity to detect distant homology (Mahmood et al. 2012; Miller, Pickett, and Ridge 2019). Nonetheless, despite all efforts to speed up the process of pairwise comparisons, all-against-all comparisons have a quadratic complexity and thus are not a long-term solution.

Thus, methods that attempt to reduce the algorithmic complexity by simply avoiding computing every single pairwise comparison are the most promising. To this end, one approach is to skip comparisons between distantly related species. Hieranoid took that idea to an extreme

by comparing only sister genomes and ancestral genomes along a guide species tree (Schreiber and Sonnhammer 2013; Kaduk and Sonnhammer 2017). Thus it scales linearly with the number of input genomes. Another approach consists of filtering pairwise alignments by precomputing candidate homologs with rough estimates of gene similarities. For example, JustOrthologs only compares the sequences of genes with similarly sized exons (Miller, Pickett, and Ridge 2019), while porthoDOM starts by clustering genes based on their domain architecture (Bitard-Feildel et al. 2015). Yet another way to avoid comparing every two pairs of sequences against each other is to exploit the transitive nature of homology (Wittwer et al. 2014; David Mark Emms and Kelly 2022). For example, knowing that the human insulin *INS* is homologous to both mouse insulins *ins1* and *ins2*, it is easy to deduce that *ins1* and *ins2* are also homologs.

The transitivity property of homology is exploited by approaches that map new sequences directly to gene families for accurate homology inference. Indeed, a match to a gene family avoids most false positive and negative homologs because the homologous relationships between gene family members have been accurately inferred in advance (David Mark Emms and Kelly 2022). By contrast, a BLAST search depends entirely on an arbitrary *E*-value threshold to delineate homologous matches. However, the main promise of mapping approaches lies in their ability to scale linearly with the number of query sequences. We have identified three types of promising mapping approaches to scale-up orthology inference.

“Closest sequence” approaches (*e.g.* EggNOG-mapper, TRAPID and OrthoDB [Cantalapiedra et al. 2021; Bucchini et al. 2021; Zdobnov et al. 2021]) assign the new sequence to the gene family of the most similar sequence identified in the reference database. Because this can represent millions of pairwise comparisons, these approaches typically allow users to narrow the search space by specifying a taxonomic scope (Cantalapiedra et al. 2021; Bucchini et al. 2021) or a set of reference species (Zdobnov et al. 2021). However, the linear complexity of "closest sequence" approaches depends on the assumption that the size of the databases remains constant over time. Although this is the case between releases, these databases grow at each release. In addition, "closest sequence" approaches generally lack specificity to distinguish orthologs from paralogs.

The second type of mapping approaches compares query sequences directly to models of gene families, whose number scales sublinearly with the number of genes as new genes can join existing families. Thus, a wide range of methods assign queries to families with pairwise alignments against Hidden Markov Models (HMMs) of reference families (Tang, Finn, and

Thomas 2019; Schreiber et al. 2014; David Mark Emms and Kelly 2022; El-Gebali et al. 2019). In practice however, they are slower than “closest sequence” approaches (Cantalapiedra et al. 2021). Recently, deep learning has been used to model gene families. By relying on convolutional neural networks to capture subsequence features of gene families, DeepFam and DeepNOG have achieved one order of magnitude speed-up compared to DIAMOND (Seo et al. 2018; Feldbauer et al. 2020; Buchfink, Xie, and Huson 2015). However, they are less sensitive and do not address the problem of distinguishing orthologs from paralogs, most likely due to the difficulty of modelling homologous groups with few members like gene subfamilies (Feldbauer et al. 2020).

The third type of mapping approaches relies on precise phylogenetic placements. Briefly, query sequences are first mapped to gene families using a “closest sequence” (David Mark Emms and Kelly 2022) or an HMM approach (Schreiber et al. 2014; Tang, Finn, and Thomas 2019). Then, each query is added to the multiple sequence alignments of the family using MAFFT (Katoh and Standley 2013) and placed on the gene family tree with EPA-ng for instance (Barbera et al. 2018). Although these approaches are highly accurate and correctly differentiate orthologs from paralogs, they are relatively slow. For example, a SHOOT search takes on average three times longer than a BLAST search (David Mark Emms and Kelly 2022).

Overall, mapping approaches are perhaps the most promising strategy to handle the high velocity of Big data genomics. However existing mapping approaches have various advantages and disadvantages. “Closest sequence” approaches are fast but not scalable in the long run. Deep learning approaches are fast and scalable but do not differentiate orthologs from paralogs, nor do "closest sequence" approaches. Although scalable and accurate, phylogenetic placement approaches are currently slower than BLAST. Thus, additional work is needed to achieve phylogenetically informed, rapid, and scalable orthology assignments. Even if these problems are solved, a key limitation that remains for all of these approaches is that they do not compare new sequences to each other and thus do not resolve orthology and paralogy between them.

Representing mega-large gene families

Genomics produces enormous amounts of data and their volume may soon be comparable to the one of social media and astronomy. (Stephens et al. 2015). In comparative genomics, the increasing number of available genomes challenges our representation of

orthology. In particular, it has raised the conceptual limitation of relying solely on pairwise evolutionary relationships when analysing hundreds or thousands of genomes (Dunn and Munro 2016; Fernández, Gabaldon, and Dessimoz 2020). Indeed, comparing more than two genomes requires generalising orthology and paralogy to sets of genes, or gene families. However, orthology is not a transitive relationship like homology (Altenhoff, Glover, and Dessimoz 2019) and thus two orthologs of the same gene are not necessarily orthologous to each other. For example, both rodent insulins *ins1* and *ins2* are orthologous to the primates insulin *INS* but *ins1* and *ins2* are paralogs since they result from a duplication event that occurred after the divergence of primates and rodents (Irwin 2021).

One approach for generalising orthology to gene families is to restrict downstream analyses to "strict" orthologous groups (OGs), where every two members are orthologs to each other (Fernández, Gabaldon, and Dessimoz, n.d.). Although these single-copy gene families may suffice as phylogenetic markers, they cannot be used to study paralogs, which are often associated with biological innovation (Kuzmin, Taylor, and Boone 2021). Moreover, single-copy gene families are biased towards returning to single-copy and thus enriched in essential functions (Dunn and Munro 2016; Waterhouse, Zdobnov, and Kriventseva 2011). Thus, it has been advocated for comparative genomics to focus more on representing evolution explicitly in a phylogenetic context rather than with pairwise or groupwise orthology and paralogy (Dunn and Munro 2016).

Reconciled gene trees represent the evolution of gene families explicitly by relating the genes at the tip of the tree with branches connected by speciation and duplication nodes. They are used by tree-based orthology inference approaches (Fuentes et al. 2021; Mi et al. 2021) as it is straightforward to extract pairwise orthology and paralogy from reconciled gene trees (Fernández, Gabaldon, and Dessimoz 2020). More importantly, precise gene evolutionary histories can be used to link duplications and losses with adaptations (Nagy et al. 2020). For example, mammals have lost two out of five visual opsin genes, likely due to an ancestral shift to a nocturnal lifestyle (Borges et al. 2018). However, gene trees are not suited for large numbers of genomes for two reasons. First, using gene trees to study gene repertoire evolution or coevolving families is difficult because most gene trees are not consistent with the species tree. Indeed, evolutionary processes like incomplete lineage sorting or horizontal gene transfers can change the order of speciation nodes in gene trees. Moreover, gene sequences may not carry enough information for accurate phylogenetic reconstruction (Graybeal 1998; Rokas et

al. 2003). As a result, several gene tree inference methods have integrated the species tree information (Thomas 2010; Boussau et al. 2013; Morel et al. 2020). Second, gene trees are computationally costly to infer and as a result Ensembl compara limits the size of gene families to maximum 1'500 genes before computing gene trees (Howe et al. 2021).

Hierarchical orthologous groups (HOGs) are promising alternatives to gene trees to exploit large numbers of genomes in comparative genomics (Altenhoff et al. 2013). One HOG is a group of genes descending from a common ancestral gene at a given taxonomic level (commonly referred as a gene family or subfamily), thus including both ortholog and paralog pairs. Collectively however, HOGs form nested hierarchical structures that depict the evolutionary history of gene families and subfamilies. Unlike gene trees, HOGs encode successive duplications as polytomies, lack branch lengths and most importantly rely on a common underlying species tree. Thus, HOGs are simplified models of gene families with potential to facilitate the genome-wide tracking of duplications (Glover et al. 2019). For example, HOGs enable to correlate the evolution of ancestral gene contents with adaptations (Train et al. 2019; Zajac et al. 2021). Tree-aware phylogenetic profilers can also benefit from consistent speciation orders across gene families to infer coevolving families (Moi et al. 2020). Reconstruction of ancestral synteny is another promising application of HOGs (Altenhoff et al. 2021). HOGs are also cheaper to compute than gene trees because they require only pairwise alignments and a species tree (Train et al. 2017). By contrast, most gene trees reconstruction methods rely on multiple sequence alignment and phylogenetic inference, which are two computationally expensive steps (Chor and Tuller 2005). Thus, HOGs with up to 120'366 members across 2'496 species (All.Dec2021 release) are available in OMA (Altenhoff et al. 2021). Overall, given their potential to integrate large numbers of genomes in numerous comparative genomics applications, HOGs have recently gained in popularity and are thus provided by many state-of-the-art orthology resources (Altenhoff et al. 2021; Kriventseva et al. 2019; Cantalapiedra et al. 2021; David M. Emms and Kelly 2019).

Visualisation is crucial when dealing with huge amounts of data as it enables building on our visual system to discover patterns in the data (Qu et al. 2019). This is particularly useful to draw hypotheses from complex data but requires adequate layouts to highlight relevant information and apply various degrees of summarisation otherwise (Nielsen et al. 2010). Due to their common underlying species tree, HOGs offer a promising framework for the visualisation of gene family evolutionary histories. For example, iHAM depicts gene family

and subfamily (*i.e.* HOG) memberships in a dynamic fashion when hovering over species tree nodes (Train et al. 2019). However, there are currently no tools leveraging HOGs to scale-up the visualisation of explicit gene evolutionary histories.

In Big data, simpler models are often preferred as they are more easily generalisable (Greene et al. 2014) and have greater potential to fully exploit the available data. For example, integrating mobile-phone data to model COVID-19 transmission instead of using more complex mathematical modelling improved predictions of epidemic trajectories (Chang et al. 2021; Ma and Lipsitch 2020). Likewise, HOGs should provide a great starting point to integrate larger numbers of genomes in comparative genomic pipelines. Nonetheless, refining the HOG model to include incomplete lineage sorting, horizontal gene transfers and other rare evolutionary events shall follow. For example, PANTHER infers a horizontal gene transfer when the number of implied losses exceeds a predefined threshold (Mi et al. 2021) and whole genome duplications would benefit from concerted HOG inferences (Altenhoff, Glover, and Dessimoz 2019).

Integrating genomes of heterogeneous quality

Compared to social media and astronomy, one specificity of genomics is its high data heterogeneity (Stephens et al. 2015). The increasing diversity of sequencing technologies (Navarro et al. 2019) and analysis pipelines for assembling and annotating genomes are the main causes of data heterogeneity. For example, the number of draft genomes overtook the one of complete genomes in GOLD soon after the popularisation of short-read technologies (Mukherjee et al. 2021). Although long-read technologies should enable a convergence toward high-quality assemblies (Rhie et al. 2021), their cost seems to remain prohibitive for large-scale genome sequencing. Indeed, most bird genomes were sequenced with short-read technologies (Bravo, Schmitt, and Edwards 2021) and the ones released in 2020 displayed a lower N50 (measure of genome contiguity) than the ones released in 2019 (Bravo, Schmitt, and Edwards 2021). Moreover, available protein-coding gene repertoires (proteomes) are also highly heterogeneous, which can have dramatic consequences on downstream analyses (Weisman, Murray, and Eddy 2022). One reason is the plethora of tools available to annotate genomes (Yandell and Ence 2012). Yet, most orthology inference methods deal poorly with the fragmented or fused gene models of draft quality genomes (Dalquen et al. 2013; Nevers, Defosset, and Lecompte 2020), in addition to being sensitive to cross-species contamination (Merchant, Wood, and Salzberg 2014). Moreover, incomplete proteomes are problematic as

best bidirectional hit (BBH) assumes proteome completeness (Nevers, Defosset, and Lecompte 2020). Thus, while waiting for cheaper long-read technologies and the homogenisation of genome annotation pipelines, orthology inference requires the development of robust approaches to integrate proteomes of heterogeneous quality, in particular from draft genomes.

The main strategy that has been used to increase the robustness of orthology inference is to treat differently proteomes with high quality metrics (*e.g.* contiguous underlying assembly, gene set completeness and high-quality gene models) from more recent and draft-quality proteomes (*e.g.* more fragmented assembly and gene models, incomplete protein set). High-quality reference proteomes are used for accurate orthology inference while draft ("periphery") proteomes are added afterward with mapping approaches. The main advantage is that the incomplete, fragmented or fused nature of draft proteomes does not disrupt the delineation of orthologous groups. HaMStR (Ebersberger, Strauss, and von Haeseler 2009) and OrthoGraph (Petersen et al. 2017) have exploited this idea to integrate gene sets obtained from expression data (inherently incomplete) to circumvent the completeness assumption of BBH. EggNOG relies on this idea to provide orthology knowledge for 5'090 organisms while preserving reliable orthologous groups (Powell et al. 2014).

Distinguishing reference from "periphery" proteomes requires extensive quality metrics to evaluate genome annotations. Although originally designed to measure assembly completeness, BUSCO has also established itself as the state-of-the-art to measure proteome completeness (Waterhouse et al. 2018). BUSCO relies on the OrthoDB orthology ressource to measure the proportion of universal and single-copy genes in a proteome (Zdobnov et al. 2021). As these genes are essential, they must be present once and any deviation from this expectation is suspicious with regard to proteome quality (Waterhouse, Zdobnov, and Kriventseva 2011; Waterhouse et al. 2018). However, measuring the overprediction or fragmentation levels of gene models currently relies on *ad-hoc* measures. For instance, OrthoInspector excludes proteomes with a high proportion of small proteins or of proteins that lack a start codon (Nevers et al. 2019). One promising approach would be to measure the deviation from the almost universal nature of protein length distributions (Nevers et al. 2021) that would suggest an enrichment of wrong or fragmented gene models.

The accuracy of mapping approaches highly benefits from the availability of closely related reference species in precomputed orthologous groups (Dylus et al. 2022). Thus, a balanced sampling of the tree of life should be prioritised and large-scale quality controls can

guide it (Feron and Waterhouse 2022). For example, the vertebrate genome project will start to generate one high quality genome for each of 260 vertebrate orders (Rhie et al. 2021). Identifying the optimal proteome quality thresholds to differentiate reference from "periphery" proteomes should also benefit from advances in the development of orthology benchmarks (Nevers et al. 2022).

Overall, integrating draft quality genomes into comparative analyses should not be overlooked as convergence towards long-read assemblies and homogeneous annotations remains far away. In the meantime, developing accurate and efficient proteome quality controls and improving the robustness of orthology inference should be prioritised.

Thesis objective and plan

With this thesis, my goal is to bring comparative genomics one step closer towards Big data, or at least towards exploiting its potential. To this end, I first present two tools to adapt orthology inference and gene family visualisation to the growing number of sequenced genomes. Then, I apply these methods in two biological systems with the goal of showcasing their potential.

In this chapter, I have introduced comparative genomics concepts that are required to understand my thesis. Then I have identified existing conceptual and methodological challenges in moving toward Big data in comparative genomics.

In chapter 2, I first demonstrate that orthology assignments obtained with closest sequence approaches like BLAST tend to result in over-specific assignments and thus lack precision. Then, to overcome this problem, I introduce OMAmer, an orthology assignment method based on alignment-free comparisons against gene families and subfamilies. I show that OMAmer is more precise, faster and scales better with the size of the reference database than closest sequence approaches.

In chapter 3, I introduce Matreex, a compact and reactive viewer for large gene families that provides new opportunities for discovery and communication in evolutionary biology. Briefly, Matreex combines gene trees for the evolutionary component and phylogenetic profiles that efficiently depict the distribution of genes across species. I illustrate Matreex with three biological applications.

In chapter 4, I attempt to characterise the role of convergent gene duplications in animal venom evolution by contrasting the protein repertoires of 68 venomous and closely related non-venomous species. I use OMAmer and quality controls to integrate proteomes of heterogeneous quality into orthologous groups in a quick and robust way.

In chapter 5, we use OMAmer to scale-up orthology inference for 363 bird genomes recently released by the Bird 10 '000 Genomes project. With this dense species sampling, I characterise the role of gene duplications and losses for seven convergent adaptations in birds. Notably, I identify convergent hemoglobin duplications in diving birds, which might be linked to the enhanced oxygen metabolism required for prolonged dives. Moreover, I observe hundreds of gene families with convergent contractions associated with the loss of flight, some of which are associated with forelimb and feather development. I use Matreex to explore these families.

In chapter 6, I discuss my methodological contributions towards Big data comparative genomics within the 3Vs framework of Big data: *velocity*, *volume* and *variety*. I also share my new thoughts (since the publication of chapter 2) on the existing limitations of OMAmer, its potential extensions and application to large-scale orthology inference. I conclude by reflecting on the limitations and perspectives of the two application chapters (4 and 5).

In parallel to my thesis work, I was involved in three collaborative projects that resulted or will result in a publication. First, I performed the orthology analyses for a study aiming to uncover the genetic consequences of asexuality in 10 stick insects (Jaron et al. 2022). Second, I participated in writing the manuscript of the last OMA paper (Altenhoff et al. 2021). Third, I collaborated on the development of OMArk, a tool to assess various measures of proteome quality built upon OMAmer (Nevers et al. *in prep*), which I applied in chapter 4.

References

Altenhoff, Adrian M., Manuel Gil, Gaston H. Gonnet, and Christophe Dessimoz. 2013. “Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs.” PloS One 8 (1): e53786.

Altenhoff, Adrian M., Natasha M. Glover, and Christophe Dessimoz. 2019. “Inferring Orthology and Paralogy.” Methods in Molecular Biology 1910: 149–75.

Altenhoff, Adrian M., Jeremy Levy, Magdalena Zarowiecki, Bartłomiej Tomiczek, Alex Warwick Vesztrocy, Daniel A. Dalquen, Steven Müller, et al. 2019. “OMA Standalone: Orthology Inference among Public and Custom Genomes and Transcriptomes.” Genome Research 29 (7): 1152–63.

Altenhoff, Adrian M., Adrian Schneider, Gaston H. Gonnet, and Christophe Dessimoz. 2011. “OMA 2011: Orthology Inference among 1000 Complete Genomes.” *Nucleic Acids Research* 39 (Database issue): D289–94.

Altenhoff, Adrian M., Clément-Marie Train, Kimberly J. Gilbert, Ishita Mediratta, Tarcisio Mendes de Farias, David Moi, Yannis Nevers, et al. 2021. “OMA Orthology in 2021: Website Overhaul, Conserved Isoforms, Ancestral Gene Order and More.” *Nucleic Acids Research* 49 (D1): D373–79.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” *Nucleic Acids Research* 25 (17): 3389–3402.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Arnold, Roland, Florian Goldenberg, Hans-Werner Mewes, and Thomas Rattei. 2014. “SIMAP--the Database of All-against-All Protein Sequence Similarities and Annotations with New Interfaces and Increased Coverage.” *Nucleic Acids Research* 42 (Database issue): D279–84.

Barbera, Pierre, Alexey M. Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, and Alexandros Stamatakis. 2018. “EPA-Ng: Massively Parallel Evolutionary Placement of Genetic Sequences.” *Systematic Biology*, August. https://doi.org/10.1093/sysbio/syy054.

Bitard-Feildel, Tristan, Carsten Kemena, Jenny M. Greenwood, and Erich Bornberg-Bauer. 2015. “Domain Similarity Based Orthology Detection.” *BMC Bioinformatics* 16 (May): 154.

Blaxter, Mark, John M. Archibald, Anna K. Childers, Jonathan A. Coddington, Keith A. Crandall, Federica Di Palma, Richard Durbin, et al. 2022. “Why Sequence All Eukaryotes?” *Proceedings of the National Academy of Sciences of the United States of America* 119 (4). https://doi.org/10.1073/pnas.2115636118.

Borges, Rui, Warren E. Johnson, Stephen J. O’Brien, Cidália Gomes, Christopher P. Heesy, and Agostinho Antunes. 2018. “Adaptive Genomic Evolution of Opsins Reveals That Early Mammals Flourished in Nocturnal Environments.” *BMC Genomics* 19 (1): 121.

Boussau, Bastien, Gergely J. Szöllosi, Laurent Duret, Manolo Gouy, Eric Tannier, and Vincent Daubin. 2013. “Genome-Scale Coestimation of Species and Gene Trees.” *Genome Research* 23 (2): 323–30.

Bravo, Gustavo A., C. Jonathan Schmitt, and Scott V. Edwards. 2021. “What Have We Learned from the First 500 Avian Genomes?” *Annual Review of Ecology, Evolution, and Systematics* 52 (1): 611–39.

Bucchini, François, Andrea Del Cortona, Łukasz Kreft, Alexander Botzki, Michiel Van Bel, and Klaas Vandepoele. 2021. “TRAPID 2.0: A Web Application for Taxonomic and Functional Analysis of de Novo Transcriptomes.” *Nucleic Acids Research* 49 (17): e101.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60.

Cantalapiedra, Carlos P., Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021. “eggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale.” *Molecular Biology and Evolution*, October. https://doi.org/10.1093/molbev/msab293.

Chang, Serina, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. 2021. “Mobility Network Models of COVID-19 Explain Inequities and Inform Reopening.” *Nature* 589 (7840): 82–87.

Chor, Benny, and Tamir Tuller. 2005. “Maximum Likelihood of Evolutionary Trees Is Hard.” In *Research in Computational Molecular Biology*, 296–310. Springer Berlin Heidelberg.

Cosentino, Salvatore, and Wataru Iwasaki. 2019. “SonicParanoid: Fast, Accurate and Easy Orthology Inference.” *Bioinformatics* 35 (1): 149–51.

Dalquen, Daniel A., Adrian M. Altenhoff, Gaston H. Gonnet, and Christophe Dessimoz. 2013. “The Impact of Gene Duplication, Insertion, Deletion, Lateral Gene Transfer and Sequencing Error on Orthology Inference: A Simulation Study.” *PloS One* 8 (2): e56925.

Derelle, Romain, Hervé Philippe, and John K. Colbourne. 2020. “Broccoli: Combining Phylogenetic and Network Analyses for Orthology Assignment.” *Molecular Biology and Evolution* 37 (11): 3389–96.

Dunn, Casey W., and Catriona Munro. 2016. “Comparative Genomics and the Diversity of Life.” *Zoologica Scripta* 45 (S1): 5–13.

Dylus, David, Adrian M. Altenhoff, Sina Majidian, Fritz J. Sedlazeck, and Christophe Dessimoz. 2022. “Read2Tree: Scalable and Accurate Phylogenetic Trees from Raw Reads.” *bioRxiv*. https://doi.org/10.1101/2022.04.18.488678.

Ebersberger, Ingo, Sascha Strauss, and Arndt von Haeseler. 2009. “HaMStR: Profile Hidden Markov Model Based Search for Orthologs in ESTs.” *BMC Evolutionary Biology* 9 (July): 157.

Ekseth, Ole Kristian, Martin Kuiper, and Vladimir Mironov. 2014. “orthAgogue: An Agile Tool for the Rapid Prediction of Orthology Relations.” *Bioinformatics* 30 (5): 734–36.

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, et al. 2019. “The Pfam Protein Families Database in 2019.” *Nucleic Acids Research* 47 (D1): D427–32.

Emms, David Mark, and Steven Kelly. 2022. “SHOOT: Phylogenetic Gene Search and Ortholog Inference.” *Genome Biology* 23 (1): 1–13.

Emms, David M., and Steven Kelly. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” *Genome Biology* 20 (1): 238.

Feldbauer, Roman, Lukas Gosch, Lukas Lüftinger, Patrick Hyden, Arthur Flexer, and Thomas Rattei. 2020. “DeepNOG: Fast and Accurate Protein Orthologous Group Assignment.” *Bioinformatics*, December. https://doi.org/10.1093/bioinformatics/btaa1051.

Fernández, Gabaldon, and Dessimoz. n.d. “Orthology: Definitions, Prediction, and Impact on Species Phylogeny Inference.” *Phylogenetics in the Genomic Era*. https://hal.archives-ouvertes.fr/hal-02535414/document.

Fernández, Rosa, Toni Gabaldon, and Christophe Dessimoz. 2020. “Orthology: Definitions, Prediction, and Impact on Species Phylogeny Inference.” *Phylogenetics in the Genomic Era*, 2.4:1–2.4:14.

Feron, Romain, and Robert M. Waterhouse. 2022. “Assessing Species Coverage and Assembly Quality of Rapidly Accumulating Sequenced Genomes.” *GigaScience* 11 (February). https://doi.org/10.1093/gigascience/giac006.

Fitch, Walter M. 1970. “Distinguishing Homologous from Analogous Proteins.” *Systematic Biology* 19 (2): 99–113.

Forslund, Kristoffer, Cecile Pereira, Salvador Capella-Gutierrez, Alan Sousa da Silva, Adrian Altenhoff, Jaime Huerta-Cepas, Matthieu Muffato, et al. 2018. “Gearing up to Handle the Mosaic Nature of Life in the Quest for Orthologs.” *Bioinformatics* 34 (2): 323–29.

Fuentes, Diego, Manuel Molina, Uciel Chorostecki, Salvador Capella-Gutiérrez, Marina Marcet-Houben, and Toni Gabaldón. 2021. “PhylomeDB V5: An Expanding Repository for Genome-Wide Catalogues of Annotated Gene Phylogenies.” *Nucleic Acids Research*, October. https://doi.org/10.1093/nar/gkab966.

Gabaldón, Toni, and Eugene V. Koonin. 2013. “Functional and Evolutionary Implications of Gene Orthology.” *Nature Reviews. Genetics* 14 (5): 360–66.

Glover, Natasha, Christophe Dessimoz, Ingo Ebersberger, Sofia K. Forslund, Toni Gabaldón, Jaime Huerta-Cepas, Maria-Jesus Martin, et al. 2019. “Advances and Applications in the Quest for Orthologs.” *Molecular Biology and Evolution* 36 (10): 2157–64.

Graybeal, A. 1998. “Is It Better to Add Taxa or Characters to a Difficult Phylogenetic Problem?” *Systematic Biology* 47 (1): 9–17.

Greene, Casey S., Jie Tan, Matthew Ung, Jason H. Moore, and Chao Cheng. 2014. “Big Data Bioinformatics.” *Journal of Cellular Physiology* 229 (12): 1896–1900.

Howe, Kevin L., Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, et al. 2021. “Ensembl 2021.” *Nucleic Acids Research* 49 (D1): D884–91.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K. Forslund, Helen Cook, Daniel R. Mende, et al. 2019. “eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses.” *Nucleic Acids Research* 47 (D1): D309–14.

Hu, Xiao, and Iddo Friedberg. 2019. “SwiftOrtho: A Fast, Memory-Efficient, Multiple Genome Orthology Classifier.” *GigaScience* 8 (10). https://doi.org/10.1093/gigascience/giz118.

Irwin, David M. 2021. “Evolution of the Insulin Gene: Changes in Gene Number, Sequence, and Processing.” *Frontiers in Endocrinology* 12 (April): 649255.

Jaron, Kamil S., Darren J. Parker, Yoann Anselmetti, Patrick Tran Van, Jens Bast, Zoé Dumas, Emeric Figuet, et al. 2022. “Convergent Consequences of Parthenogenesis on Stick Insect Genomes.” *Science Advances* 8 (8): eabg3842.

Kaduk, Mateusz, and Erik Sonnhammer. 2017. “Improved Orthology Inference with Hieranoid 2.” *Bioinformatics* 33 (8): 1154–59.

Karlin, S., and S. F. Altschul. 1990. “Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes.” *Proceedings of the National Academy of Sciences of the United States of America* 87 (6): 2264–68.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80.

Koonin, Eugene V. 2005. “Orthologs, Paralogs, and Evolutionary Genomics.” *Annual Review of Genetics* 39: 309–38.

Kriventseva, Evgenia V., Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. 2019. “OrthoDB v10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs.” Nucleic Acids Research 47 (D1): D807–11.

Kuzmin, Elena, John S. Taylor, and Charles Boone. 2021. “Retention of Duplicated Genes in Evolution.” Trends in Genetics: TIG, July. https://doi.org/10.1016/j.tig.2021.06.016.

Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. 2014. Mining of Massive Datasets. Cambridge University Press.

Lewin, Harris A., Stephen Richards, Erez Lieberman Aiden, Miguel L. Allende, John M. Archibald, Miklós Bálint, Katharine B. Barker, et al. 2022. “The Earth BioGenome Project 2020: Starting the Clock.” Proceedings of the National Academy of Sciences of the United States of America 119 (4). https://doi.org/10.1073/pnas.2115635118.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018. “Earth BioGenome Project: Sequencing Life for the Future of Life.” Proceedings of the National Academy of Sciences of the United States of America 115 (17): 4325–33.

Linard, Benjamin, Ingo Ebersberger, Shawn E. McGlynn, Natasha Glover, Tomohiro Mochizuki, Mateus Patricio, Odile Lecompte, et al. 2021. “Ten Years of Collaborative Progress in the Quest for Orthologs.” *Molecular Biology and Evolution* 38 (8): 3033–45.

Mahmood, Khalid, Geoffrey I. Webb, Jiangning Song, James C. Whisstock, and Arun S. Konagurthu. 2012. “Efficient Large-Scale Protein Sequence Comparison and Gene Matching to Identify Orthologs and Co-Orthologs.” *Nucleic Acids Research* 40 (6): e44.

Ma, Kevin C., and Marc Lipsitch. 2020. “Big Data and Simple Models Used to Track the Spread of COVID-19 in Cities.” Nature Publishing Group UK. November 10, 2020. https://doi.org/10.1038/d41586-020-02964-4.

Merchant, Samier, Derrick E. Wood, and Steven L. Salzberg. 2014. “Unexpected Cross-Species Contamination in Genome Sequencing Projects.” *PeerJ* 2 (November): e675.

Mi, Huaiyu, Dustin Ebert, Anushya Muruganujan, Caitlin Mills, Laurent-Philippe Albou, Tremayne Mushayamaha, and Paul D. Thomas. 2021. “PANTHER Version 16: A Revised Family Classification, Tree-Based Classification Tool, Enhancer Regions and Extensive API.” *Nucleic Acids Research* 49 (D1): D394–403.

Miller, Justin B., Brandon D. Pickett, and Perry G. Ridge. 2019. “JustOrthologs: A Fast, Accurate and User-Friendly Ortholog Identification Algorithm.” *Bioinformatics* 35 (4): 546–52.

Mitrophanov, Alexander Yu, and Mark Borodovsky. 2006. “Statistical Significance in Biological Sequence Analysis.” *Briefings in Bioinformatics* 7 (1): 2–24.

Moi, David, Laurent Kilchoer, Pablo S. Aguilar, and Christophe Dessimoz. 2020. “Scalable Phylogenetic Profiling Using MinHash Uncovers Likely Eukaryotic Sexual Reproduction Genes.” *PLoS Computational Biology* 16 (7): e1007553.

Morel, Benoit, Alexey M. Kozlov, Alexandros Stamatakis, and Gergely J. Szöllősi. 2020. “GeneRax: A Tool for Species Tree-Aware Maximum Likelihood Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss.” *Molecular Biology and Evolution*, June. https://doi.org/10.1093/molbev/msaa141.

Mukherjee, Supratim, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Jagadish Chandrabose Sundaramurthi, Janey Lee, Mahathi Kandimalla, I-Min A. Chen, Nikos C. Kyrpides, and T. B. K. Reddy. 2021. “Genomes OnLine Database (GOLD) v.8: Overview and Updates.” *Nucleic Acids Research* 49 (D1): D723–33.

Nagy, László G., Zsolt Merényi, Botond Hegedüs, and Balázs Bálint. 2020. “Novel Phylogenetic Methods Are Needed for Understanding Gene Function in the Era of Mega-Scale Genome Sequencing.” *Nucleic Acids Research* 48 (5): 2209–19.

Navarro, Fábio C. P., Hussein Mohsen, Chengfei Yan, Shantao Li, Mengting Gu, William Meyerson, and Mark Gerstein. 2019. “Genomics and Data Science: An Application within an Umbrella.” *Genome Biology* 20 (1): 109.

Needleman, S. B., and C. D. Wunsch. 1970. “A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins.” *Journal of Molecular Biology* 48 (3): 443–53.

Nevers, Yannis, Audrey Defosset, and Odile Lecompte. 2020. “Orthology: Promises and Challenges.” In *Evolutionary Biology—A Transdisciplinary Approach*, edited by Pierre Pontarotti, 203–28. Cham: Springer International Publishing.

Nevers, Yannis, Natasha Glover, Christophe Dessimoz, and Odile Lecompte. 2021. “Protein Length Distribution Is Remarkably Consistent across Life.” *bioRxiv*. https://doi.org/10.1101/2021.12.03.470944.

Nevers, Yannis, Tamsin E. M. Jones, Dushyanth Jyothi, Bethan Yates, Meritxell Ferret, Laura Portell-Silva, Laia Codo, et al. 2022. “The Quest for Orthologs Orthology Benchmark Service in 2022.” *Nucleic Acids Research*, May. https://doi.org/10.1093/nar/gkac330.

Nevers, Yannis, Arnaud Kress, Audrey Defosset, Raymond Ripp, Benjamin Linard, Julie D. Thompson, Olivier Poch, and Odile Lecompte. 2019. “OrthoInspector 3.0: Open Portal for Comparative Genomics.” *Nucleic Acids Research* 47 (D1): D411–18.

Nielsen, Cydney B., Michael Cantor, Inna Dubchak, David Gordon, and Ting Wang. 2010. “Visualizing Genomes: Techniques and Challenges.” *Nature Methods* 7 (3 Suppl): S5–15.

Ochoterena, Helga, Alexander Vrijdaghs, Erik Smets, and Regine Claßen-Bockhoff. 2019. “The Search for Common Origin: Homology Revisited.” *Systematic Biology* 68 (5): 767–80.

Overbeek, R., M. Fonstein, M. D’Souza, G. D. Pusch, and N. Maltsev. 1999. “The Use of Gene Clusters to Infer Functional Coupling.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (6): 2896–2901.

Owen, Richard. 1846. Lectures on the Comparative Anatomy and Physiology of the Vertebrate Animals: Delivered at the Royal College of Surgeons of England, in 1844 and 1846. Longman, Brown, Green, and Longmans.

Pearson, William R. 2013. “An Introduction to Sequence Similarity (‘homology’) Searching.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* Chapter 3 (June): Unit3.1.

Pearson, W. R. 1998. “Empirical Statistical Estimates for Sequence Similarity Searches.” *Journal of Molecular Biology* 276 (1): 71–84.

Petersen, Malte, Karen Meusemann, Alexander Donath, Daniel Dowling, Shanlin Liu, Ralph S. Peters, Lars Podsiadlowski, et al. 2017. “Orthograph: A Versatile Tool for Mapping Coding Nucleotide Sequences to Clusters of Orthologous Genes.” *BMC Bioinformatics* 18 (1): 111.

Powell, Sean, Kristoffer Forslund, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Jaime Huerta-Cepas, Toni Gabaldón, et al. 2014. “eggNOG v4.0: Nested Orthology Inference across 3686 Organisms.” *Nucleic Acids Research* 42 (Database issue): D231–39.

Qu, Zhonglin, Chng Wei Lau, Quang Vinh Nguyen, Yi Zhou, and Daniel R. Catchpoole. 2019. “Visual Analytics of Genomic and Cancer Data: A Systematic Review.” *Cancer Informatics* 18 (March): 1176935119835546.

Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. “Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species.” *Nature* 592 (7856): 737–46.

Rokas, Antonis, Barry L. Williams, Nicole King, and Sean B. Carroll. 2003. “Genome-Scale Approaches to Resolving Incongruence in Molecular Phylogenies.” *Nature* 425 (6960): 798–804.

Sagiroglu, Seref, and Duygu Sinanc. 2013. “Big Data: A Review.” In 2013 International Conference on Collaboration Technologies and Systems (CTS), 42–47.

Schreiber, Fabian, Mateus Patricio, Matthieu Muffato, Miguel Pignatelli, and Alex Bateman. 2014. “TreeFam v9: A New Website, More Species and Orthology-on-the-Fly.” *Nucleic Acids Research* 42 (D1): D922–25.

Schreiber, Fabian, and Erik L. L. Sonnhammer. 2013. “Hieranoid: Hierarchical Orthology Inference.” *Journal of Molecular Biology* 425 (11): 2072–81.

Seo, Seokjun, Minsik Oh, Youngjune Park, and Sun Kim. 2018. “DeepFam: Deep Learning Based Alignment-Free Method for Protein Family Modeling and Prediction.” *Bioinformatics* 34 (13): i254–62.

Smith, Stacey D., Matthew W. Pennell, Casey W. Dunn, and Scott V. Edwards. 2020. “Phylogenetics Is the New Genetics (for Most of Biodiversity).” *Trends in Ecology & Evolution* 35 (5): 415–25.

Smith, T. F., and M. S. Waterman. 1981. “Identification of Common Molecular Subsequences.” *Journal of Molecular Biology* 147 (1): 195–97.

Sonnhammer, Erik L. L., Toni Gabaldón, Alan W. Sousa da Silva, Maria Martin, Marc Robinson-Rechavi, Brigitte Boeckmann, Paul D. Thomas, Christophe Dessimoz, and Quest for Orthologs consortium. 2014. "Big Data and Other Challenges in the Quest for Orthologs." *Bioinformatics* 30 (21): 2993–98.

Sonnhammer, Erik L. L., and Eugene V. Koonin. 2002. "Orthology, Paralogy and Proposed Classification for Paralog Subtypes." *Trends in Genetics: TIG* 18 (12): 619–20.

Sonnhammer, Erik L. L., and Gabriel Östlund. 2015. "InParanoid 8: Orthology Analysis between 273 Proteomes, Mostly Eukaryotic." *Nucleic Acids Research* 43 (Database issue): D234–39.

Steinegger, Martin, and Johannes Söding. 2017. "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets." *Nature Biotechnology* 35 (11): 1026–28.

Stephan, Taylorlyn, Shawn M. Burgess, Hans Cheng, Charles G. Danko, Clare A. Gill, Erich D. Jarvis, Klaus-Peter Koepfli, et al. 2022. "Darwinian Genomics and Diversity in the Tree of Life." *Proceedings of the National Academy of Sciences of the United States of America* 119 (4). https://doi.org/10.1073/pnas.2115644119.

Stephens, Zachary D., Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. 2015. "Big Data: Astronomical or Genomical?" *PLoS Biology* 13 (7): e1002195.

Tabari, Ehsan, and Zhengchang Su. 2017. "PorthoMCL: Parallel Orthology Prediction Using MCL for the Realm of Massive Genome Availability." *Big Data Analytics* 2 (January). https://doi.org/10.1186/s41044-016-0019-8.

Tang, Haiming, Robert D. Finn, and Paul D. Thomas. 2019. "TreeGrafter: Phylogenetic Tree-Based Annotation of Proteins with Gene Ontology Terms and Other Annotations." *Bioinformatics* 35 (3): 518–20.

Thomas, Paul D. 2010. "GIGA: A Simple, Efficient Algorithm for Gene Tree Inference in the Genomic Age." *BMC Bioinformatics* 11 (June): 312.

Train, Clément-Marie, Natasha M. Glover, Gaston H. Gonnet, Adrian M. Altenhoff, and Christophe Dessimoz. 2017. "Orthologous Matrix (OMA) Algorithm 2.0: More Robust to Asymmetric Evolutionary Rates and More Scalable Hierarchical Orthologous Group Inference." *Bioinformatics* 33 (14): i75–82.

Train, Clément-Marie, Miguel Pignatelli, Adrian Altenhoff, and Christophe Dessimoz. 2019. "iHam and pyHam: Visualizing and Processing Hierarchical Orthologous Groups." *Bioinformatics* 35 (14): 2504–6.

Waterhouse, Robert M., Mathieu Seppey, Felipe A. Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2018. "BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics." *Molecular Biology and Evolution* 35 (3): 543–48.

Waterhouse, Robert M., Evgeny M. Zdobnov, and Evgenia V. Kriventseva. 2011. "Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in Vertebrates, Arthropods, and Fungi." *Genome Biology and Evolution* 3: 75–86.

Weisman, Caroline M., Andrew W. Murray, and Sean R. Eddy. 2022. "Mixing Genome Annotation Methods in a Comparative Analysis Inflates the Apparent Number of Lineage-Specific Genes." *Current Biology: CB* 32 (12): 2632–39.e2.

Wittwer, Lucas D., Ivana Piližota, Adrian M. Altenhoff, and Christophe Dessimoz. 2014. "Speeding up All-against-All Protein Comparisons While Maintaining Sensitivity by Considering Subsequence-Level Homology." *PeerJ* 2 (October): e607.

Yandell, Mark, and Daniel Ence. 2012. "A Beginner's Guide to Eukaryotic Genome Annotation." *Nature Reviews. Genetics* 13 (5): 329–42.

Zajac, Natalia, Stefan Zoller, Katri Seppälä, David Moi, Christophe Dessimoz, Jukka Jokela, Hanna Hartikainen, and Natasha Glover. 2021. “Gene Duplication and Gain in the Trematode Atriophallophorus Winterbourni Contributes to Adaptation to Parasitism.” *Genome Biology and Evolution* 13 (3). https://doi.org/10.1093/gbe/evab010.

Zdobnov, Evgeny M., Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Matthew Berkeley, and Evgenia V. Kriventseva. 2021. “OrthoDB in 2020: Evolutionary and Functional Annotations of Orthologs.” *Nucleic Acids Research* 49 (D1): D389–93.

Zielezinski, Andrzej, Susana Vinga, Jonas Almeida, and Wojciech M. Karlowski. 2017. “Alignment-Free Sequence Comparison: Benefits, Applications, and Tools.” *Genome Biology* 18 (1): 186.

Chapter 2

OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches

OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches

Victor Rossier[1,2,3], Alex Warwick Vesztrocy[1,2,3], Marc Robinson-Rechavi[4,5,*] and Christophe Dessimoz[1,2,3,5,6,*]

[1]Department of Computational Biology, University of Lausanne, Switzerland; [2]Center for Integrative Genomics, University of Lausanne, Switzerland; [3]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland; [4]Department of Ecology and Evolution, University of Lausanne, Switzerland; [5]Department of Genetics, Evolution, and Environment, University College London, UK; [6]Department of Computer Science, University College London, UK.

*Corresponding authors: Marc.Robinson-Rechavi@unil.ch & Christophe.Dessimoz@unil.ch

Abstract

Assigning new sequences to known protein families and subfamilies is a prerequisite for many functional, comparative and evolutionary genomics analyses. Such assignment is commonly achieved by looking for the closest sequence in a reference database, using a method such as BLAST. However, ignoring the gene phylogeny can be misleading because a query sequence does not necessarily belong to the same subfamily as its closest sequence. For example, a hemoglobin which branched out prior to the hemoglobin alpha/beta duplication could be closest to a hemoglobin alpha or beta sequence, whereas it is neither. To overcome this problem, phylogeny-driven tools have emerged but rely on gene trees, whose inference is computationally expensive.

Here, we first show that in multiple animal and plant datasets, 18 to 62% of assignments by closest sequence are misassigned, typically to an over-specific subfamily. Then, we introduce OMAmer, a novel alignment-free protein subfamily assignment method, which limits over-specific subfamily assignments and is suited to phylogenomic databases with thousands of genomes. OMAmer is based on an innovative method using evolutionarily-informed *k*-mers for alignment-free mapping to ancestral protein subfamilies. Whilst able to reject non-homologous family-level assignments, we show that OMAmer provides better and quicker subfamily-level assignments than approaches relying on the closest sequence, whether inferred exactly by Smith-Waterman or by the fast heuristic DIAMOND.

OMAmer is available from the Python Package Index (as omamer), with the source code and a precomputed database available at https://github.com/DessimozLab/omamer.

Introduction

Assigning new sequences to known protein families is a prerequisite for many comparative and evolutionary analyses (Glover *et al.*, 2019). Functional knowledge can also be transferred from reference to new sequences assigned in the same family (Gabaldón and Koonin, 2013).

However, when gene duplication events have resulted in multiple copies per species, multiple "subfamilies" are generated, which can make placing a protein sequence into the correct subfamily challenging. Gene subfamilies are nested gene families defined after duplication events and organized hierarchically into gene trees. For example, the epsilon and gamma hemoglobin subfamilies are defined at the placental level, and nested in the adult hemoglobin beta subfamily at the mammal level (Opazo *et al.*, 2008). Both belong to the globin family that originated in the LUCA (last universal common ancestor of cellular life).

Gene subfamily assignment is commonly achieved by looking for the most similar sequence (or "closest sequence", *see Discussion*) in a reference database, using a method such as BLAST or DIAMOND (Altschul *et al.*, 1990; Buchfink *et al.*, 2015), before assigning the query to the subfamily of the closest sequence identified. For example, EggNOG mapper uses reference subfamilies from EggNOG to functionally annotate millions of unknown proteins of genomes and metagenomes (Huerta-Cepas *et al.*, 2017, 2019). Briefly, each query is assigned to the most specific gene subfamily of its closest sequence, inferred using DIAMOND, with functional annotations then transferred accordingly.

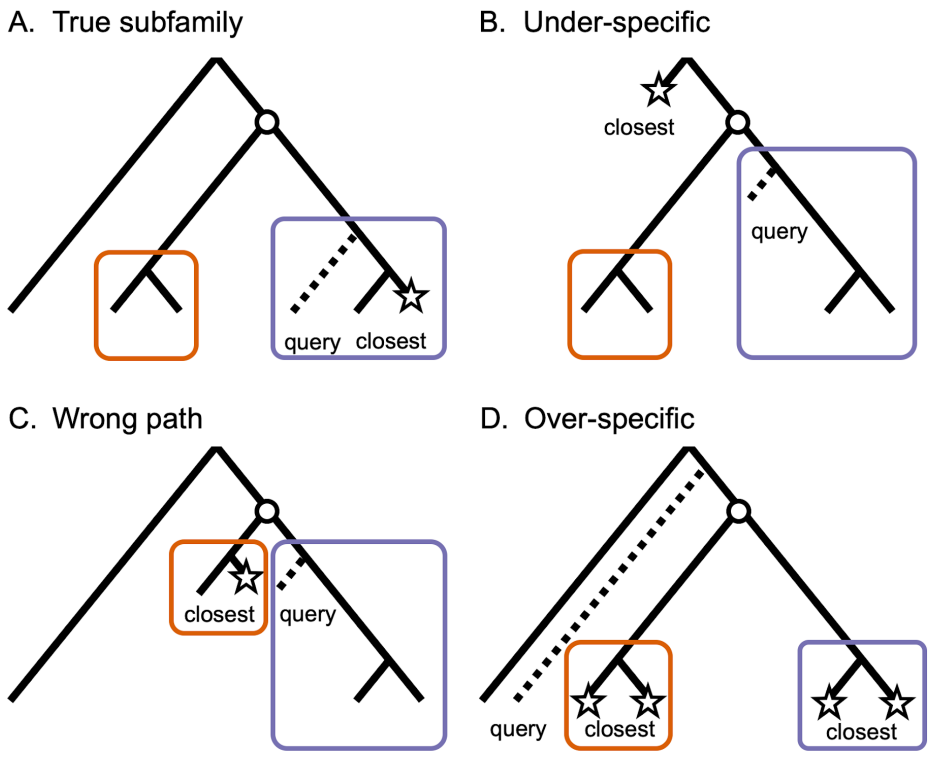


Fig. 1. The closest sequence to a query does not necessarily belong to the same subfamily. This figure conceptualizes the four possible closest sequence locations relative to the query. On each tree, the true position of the query is indicated by a dashed branch, while its closest sequence(s) in the family is indicated by a star. The circle represents a duplication event leading to two subfamilies depicted as color boxes. Scenario A is the only one in which the closest sequence is in the same subfamily as the query. Note that for scenarios B and C to happen, the rate of evolution needs to vary across the tree (departure from a "molecular clock"), whereas scenario D can even happen under a uniform rate of evolution.

However, ignoring the protein family tree can be misleading because a query sequence does not necessarily belong to the same subfamily as its closest sequence (Fig. 1). For instance, if the query branched out from a fast evolving subtree, its closest sequence might not belong to that subtree, but to a more general subfamily, or even not be classifiable in any known subfamily (Fig. 1. B). Or, in case of asymmetric evolutionary rates between sister subfamilies, the closest sequence might belong to a different subfamily altogether (Fig. 1. C). The prospect of observing these two scenarios is sustained by the long-standing observation that duplicated proteins experience accelerated and often asymmetric evolution (Conant and Wolfe, 2008; Sémon and Wolfe, 2007).

Moreover, the closest sequence to the query can belong to an over-specific subfamily even without any departure from the molecular clock in the family tree (Fig. 1. D). Such cases

may occur when the query branched out before the emergence of more specific (nested) subfamilies. Indeed, all known proteins from the same clade as the query can belong to nested subfamilies. Moreover, even when not all proteins belong to such nested subfamilies, the closest sequence may still belong to an over-specific subfamily by chance. Since duplications are common in evolution (Conant and Wolfe, 2008), finding such nested subfamilies as close relatives to the query divergence is expected to be common.

To avoid such errors, protein subfamily assignment tools relying on gene trees have been proposed (Schreiber *et al.*, 2014; Tang *et al.*, 2019). In short, these start by assigning queries to families with pairwise alignments against Hidden Markov profiles of reference families. Then, fine-grained assignments to subfamilies are performed with tree placement tools, which typically attempt to graft the query on every branch of the tree until maximizing a likelihood or parsimony score (Barbera *et al.*, 2018). However, gene tree inference is computationally expensive and therefore not scalable to the exponentially growing number of available sequences.

As a more scalable alternative to gene trees, the concept of hierarchical orthologous groups (HOGs) (Altenhoff *et al.*, 2013) provides a precise definition of the intuitive notion of protein families and subfamilies. Each HOG is a group of proteins descending from a single speciation event and organized hierarchically. Moreover, they collectively provide the evolutionary history of protein families and subfamilies, like gene trees. While the oldest HOG in the family hierarchy ("root-HOG") is the family itself, the other nested HOGs are its subfamilies. Thus, HOGs up to 100,000 members and covering thousands of species are available in large-scale phylogenomic databases (Altenhoff *et al.*, 2018; Huerta-Cepas *et al.*, 2019; Kriventseva *et al.*, 2019).

Here, we first demonstrate on six animal and plant proteomes (sets of proteins from a given species, see *Methods*) that 18 to 62% of assignments by closest sequence go to incorrect, mostly over-specific, subfamilies. To overcome this problem, we introduce OMAmer, a novel alignment-free protein subfamily assignment method, which limits over-specific subfamily assignments and is suited to phylogenomic databases with thousands of genomes. We show that OMAmer is able to assign proteins to subfamilies more accurately than approaches relying on the closest sequence, whether inferred exactly by Smith-Waterman or by the fast heuristic DIAMOND. Furthermore, we show that by adopting efficient alignment-free *k*-mer based analyses pioneered by metagenomic taxonomic classifiers such as Kraken or RAPPAS (Wood

and Salzberg, 2014; Linard *et al.*, 2019), and adapting them to protein subfamily-level classification, OMAmer is computationally faster and more scalable than DIAMOND.

Materials and methods

The OMAmer algorithm

In this section, we describe the two main algorithmic steps which make OMAmer more precise and faster than closest sequence approaches. First, to speed-up the protein assignment step, OMAmer preprocesses reference hierarchical orthologous groups (HOGs) into a *k*-mer table (Fig. 2). For each *k*-mer and family (root-HOG), this table stores the subfamily (sub-HOG) where the *k*-mer has most likely arisen (the most specific HOG containing all occurrences of the given *k*-mer within the root-HOG). Then, these evolutionarily-informed *k*-mers are used to yield more precise subfamily assignments by reducing over-specific assignments (Fig. 3).

k-mer table precomputation

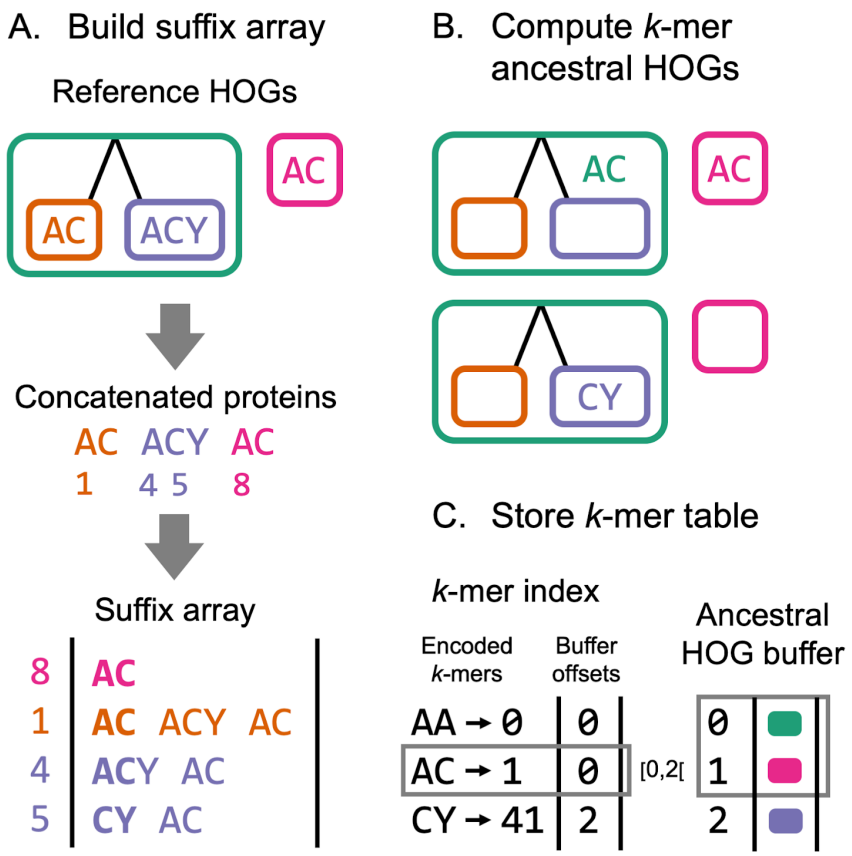


Fig. 2. OMAmer algorithm for compact *k*-mer table precomputation. A. To efficiently preprocess the *k*-mer table, a suffix array is first built from concatenated protein sequences of reference hierarchical orthologous groups (HOGs), encoding families (root-HOGs) and subfamilies (sub-HOGs). Numbers indicate suffix offsets in the concatenated protein array and bold characters highlight *k*-mers at the beginning of suffixes. B. The *k*-mer ancestral HOG (where the *k*-mer has arisen) is approximated within each root-HOG as the last common ancestor among HOGs with the given k-mer. For example, since both the orange and purple sub-HOGs contain the "AC" *k*-mer, the ancestral HOG for that *k*-mer is the green root-HOG. C. The compact *k*-mer table includes a *k*-mer index mapping to a buffer that stores each *k*-mer ancestral HOGs. Note that each offset of the index corresponds to a k-mer integer encoding. As illustrated with the grey boxes, the "AC" *k*-mer (encoded as "1") maps to the green and pink HOGs since these two lie within the [0,2[offset interval in the buffer.

To efficiently parse *k*-mer sets of reference HOGs, the suffix array (Manber and Myers, 1993) of all concatenated reference proteins is used as an intermediate data structure (Fig. 2. A.). There, all suffixes starting with a given *k*-mer are stored consecutively, which enables to quickly identify all HOGs containing the same k-mer using binary search.

Then, the *k*-mer ancestral HOG (where the *k*-mer has arisen) is approximated within each root-HOG as the last common ancestor (LCA) among HOGs containing the given *k*-mer (Fig. 2. B). Indeed, we assume that occurrences of the same *k*-mer in different members of a family mostly result from homology (*i.e.* same *k*-mer due to shared ancestry) rather than homoplasy (*i.e.* same *k*-mer arising independently). In the instances where the latter is true, the LCA approximation will favor overly general assignments. Thus, compared to the homoplasy assumption that would favor over-specific assignments, this approach is more conservative. Moreover, retaining a single ancestral HOG per *k*-mer and family reduces the memory footprint of the *k*-mer table.

Finally, to enable fast and memory efficient subfamily assignments, the resulting *k*-mer table is stored in the compressed sparse row format, consisting of two related arrays (Fig. 2. C). The *k*-mer index stores, at offsets corresponding to each *k*-mer integer encoding (*e.g.* 0 for AA, 1 for AC, etc.), offsets of the second array (the ancestral HOG buffer). There, the ancestral HOGs of each *k*-mer are stored consecutively. The formulae used to encode *k*-mers in integers is described in supplementary material.

Family and subfamily protein assignment

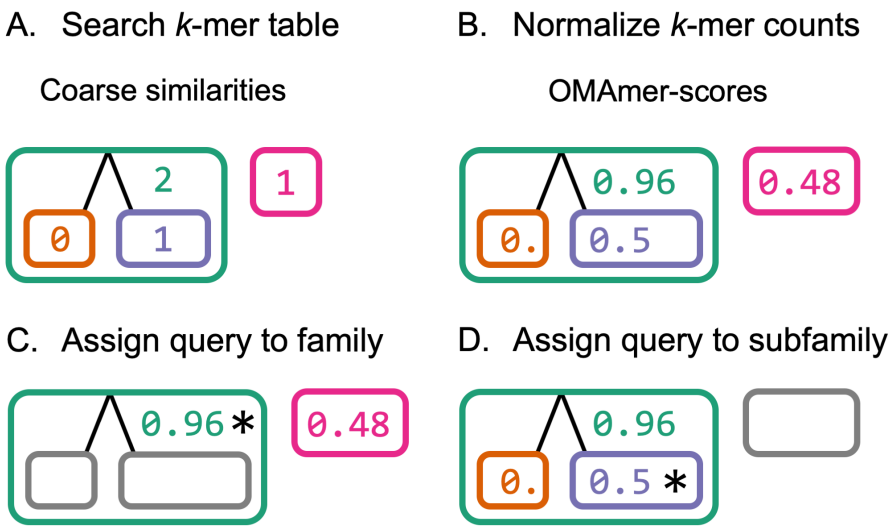


Fig. 3. OMAmer algorithm for protein assignment to family and subfamily. A. Coarse alignment-free similarities are obtained by searching the k-mer table. B. These similarities are normalized into OMAmer-scores to account for varying query lengths, composition biases, and sizes of reference HOGs. C. The family (root-HOG) with the highest HOG-score is retained (shown with an asterisk) D. The assignment is refined to the most specific subfamily on the highest scoring root to leaf path (shown with an asterisk).

The family (root-HOG) and subfamily (sub-HOG) protein assignment both rely on a common measure of similarity between the query protein and reference HOGs (the "OMAmer-score"). Essentially, this score captures the excess of similarity that is shared between the query and a given HOG, thus excluding the similarity with regions conserved in more ancestral HOGs. The OMAmer-score is computed in two main steps. First, a coarse alignment-free similarity is obtained by searching the *k*-mer table (Fig. 3. A). Second, this similarity is normalized to account for varying query lengths, composition biases, and sizes (number of different *k*-mers) of reference HOGs (Fig. 3. B). More details about the OMAmer-score computation are available in supplementary methods.

The protein is first assigned to the root-HOG with the highest OMAmer-score (Fig. 3. C). Indeed, at the family level, the OMAmer-score is analogous to other sequence similarity measures (*e.g.* alignment score) used to evaluate a probability of homology. Note that to speed-up the assignment, OMAmer-scores are computed only for the top 100 root-HOGs with the highest coarse alignment-free similarity.

Then, the assignment is refined to the most specific sub-HOG on the highest scoring root-to-leaf path within the predicted root-HOG (Fig. 3. D). Indeed, at the subfamily level, OMAmer-scores are only comparable when descending from the same parent since they capture an excess of similarity relative to that parent.

Finally, to reduce the risk of false positive assignments, thresholds on the OMAmer-score can be applied at both steps. At the family level, this avoids placing queries which have no homolog in the reference database. At the subfamily level, it penalizes more specific subfamilies to prevent over-specific assignments. Moreover, for applications where it is important to reject partial homologous matches (*e.g.* domain-level), OMAmer also outputs an "overlap-score" that measures the fraction of the query sequence overlapping with k-mers of reference root-HOGs (ignoring *k*-mers with multiple occurrences in the query sequence).

Benchmarking

In this section, we describe the experiments conducted to evaluate the accuracy of OMAmer compared to closest sequence methods: Smith-Waterman (Smith and Waterman, 1981) and DIAMOND (Buchfink *et al.*, 2015). Since placement in subfamilies initially requires accurate family-level assignments, we started by evaluating OMAmer at the family level (*i.e.* identifying the correct root-HOG). Second, to evaluate the impact of ignoring the phylogeny on subfamily assignments by closest sequences, we estimated the frequency of each closest sequence configuration ("true subfamily", "under-specific", "wrong-path" and "over-specific" [Fig. 1]). Third, we benchmarked subfamily-level assignments against closest sequence methods. Finally, we broke down the validation results of OMAmer by closest sequence configuration. The datasets and software parameters used in these experiments are described in supplementary methods.

Family-level validation

Positive query sets were constructed as the sets of proteins from a given species contained in reference hierarchical orthologous groups (HOGs). We call these sets of proteins "proteomes" in this work. The proteins of that species were removed from the reference database used before the *k*-mer index precomputation.

Since query proteins do not necessarily have homologous counterparts in the reference families (*e.g.* "orphan" genes, contamination, horizontal gene transfer), validating family

assignments also required negative sets of non-homologous queries. Therefore, negative query sets were built with two approaches, while always matching the size of their corresponding positive set. In the first approach, random proteins were simply simulated with UniProtKB amino acid frequencies (release 2020_01) (UniProt Consortium, 2019) and sequence lengths of positive queries. The second approach was designed to resemble events of contamination or of horizontal gene transfer. Each negative query was randomly selected from a unique clade-specific family lying outside the taxonomic scope of reference families. In practice, clade-specific families were randomly selected among HOGs without parent (root-HOGs) at a given taxonomic level.

The resulting family assignments were compared with the truth set, and classified into true positives (TPs), false negatives (FNs) and false positives (FPs) for various score thresholds. FPs included negative queries assigned to a family as well as positive queries assigned to the wrong family (their relative proportion is shown in Supp. Fig. 2). The remaining positive queries were divided into TPs and FNs depending on whether the score for their family of origin passed the threshold, or not. Finally, precision, recall and accuracy (F1) were computed from TPs, FNs and FPs (Supp. Table. 1), defined according to the score threshold.

In the following experiments, to assess subfamily-level assignment separately from family-level assignment, we focused on the query sequences assigned to the correct family (*i.e.* the set of TPs at the threshold where F1 is maximal [$F1_{max}$] for family assignment). Moreover, non-overlapping family-level TPs between methods being compared were further filtered out (sets of overlapping TPs are shown in Supp. Fig. 1.).

Quantification of subfamily assignment errors by closest sequences

We used Smith-Waterman local alignments as reference to find the closest sequence (Smith and Waterman, 1981; Wolf and Koonin, 2012). Indeed, being an exact algorithm, Smith-Waterman is guaranteed to find the highest scoring match, and it is the standard approach in the field (Wolf and Koonin, 2012). Then, we classified each query according to the location of its closest sequence (Fig. 1.) as follows: a “true subfamily” configuration arises when the most specific HOG of the closest sequence is the same as the query one. An “over-specific” configuration arises when the most specific HOG of the query is ancestral to the most specific HOG of the closest sequence. Conversely, an “under-specific” configuration arises when the most specific HOG of the closest sequence is ancestral to that of the query. The last

case is the “wrong-path” configuration, in which the most specific HOG of the query and of the closest sequence are in different parts of the family tree.

Subfamily-level validation

To assess TPs, FNs and FPs at this level we took the view that an assignment to a subfamily also implies assignment to its “parental” subfamilies (if there are any). For instance, let us consider a nested gene family of alcohol dehydrogenases. Under this view, an assignment to the specific “alcohol dehydrogenase 1C” is also implicitly an assignment to “alcohol dehydrogenase 1”, as well as to “alcohol dehydrogenase”. In this case, if a method incorrectly assigns the protein to the subfamily “alcohol dehydrogenase 1B”, in addition to counting a FP (the gene is not a true member of subfamily “B”) and a FN (the gene is missing from subfamily “C”), we also count one TP for correctly assigning to the parental sub-HOG “alcohol dehydrogenase 1”. In effect, the prediction is regarded as being only partially wrong. Note that there is no TP counted for correctly implying an assignment to the root-HOG (alcohol dehydrogenase), because the present analysis only seeks to assess within-family placement.

In addition, we repeated the analyses using a second approach taking the more stringent view that there are no implicit predictions of parental subfamilies, therefore no reward is given for partial correctness. Thus, in the previous example, there would be no TP counted—only one FP and one FN.

For both validation approaches, precision, recall and accuracy (F1) were computed from TPs, FPs, and FNs using the same formulae as at the family-level (Supp. Table 1).

Performance experiments

To benchmark the computational performance of OMAmer and DIAMOND, we measured real and CPU time, as well as the maximum resident set size (memory) using the GNU time command. All timing was performed on machines containing identical hardware (dual-socket Intel Xeon E5-2680, 64GB of RAM). Single threaded versions of both methods were used, with timing repeated 10 times in order to ensure stability.

Databases of increasing size (20 to 200 proteomes, in steps of 20) were generated from Metazoan proteomes, with each including all of the previous and an extra 20 randomly selected

species. The full proteomes of the initial 20 were used to query the databases of increasing size in order to gauge the scaling characteristics.

Software availability

OMAmer is available from the Python Package Index (as omamer), with the source code and a precomputed database available at https://github.com/DessimozLab/omamer.

Results

We first consider the problem of sequence placement at the overall family level (*i.e.* identifying the correct root hierarchical orthologous group, or "root-HOG", defined at either *Metazoa* or *Viridiplantae*). Then, we present our analyses of the subfamily placement problem in four parts: First, we quantify the different types of errors resulting from the closest sequence criterion. Second, we show that OMAmer overcomes many of these errors, resulting in higher accuracy than closest sequence approaches. Third, we show that this accuracy improvement is mainly achieved by avoiding over-specific sequence classification. And fourth, we compare the computational cost and scaling of OMAmer and DIAMOND.

At the overall family level, sequence placement is highly accurate

Query sequences must first be assigned to families before being placed within subfamilies. We evaluated this using DIAMOND and OMAmer, assessing the ability of the methods to either place a protein in its correct family, or to avoid placing a sequence with no homolog in the reference database (see *Methods*).

Both methods delivered similar and highly accurate results in placing platypus, spotted gar, and plant proteins ($F1_{max} > 0.9$; Supp. Fig. 2). The methods did not perform as well on the amphioxus proteome (OMAmer $F1_{max} = 0.81\text{-}0.84$; DIAMOND $F1_{max} = 0.86\text{-}0.88$; Supp. Fig. 2), but this is an outgroup to all other chordates in OMA, with a divergence of 600 MY (Peterson and Eernisse, 2016) to the closest species sampled (*i.e.* all vertebrates and urochordates) and with high levels of polymorphism which can result in alleles being misannotated as paralogs (Putnam *et al.*, 2008; Huang *et al.*, 2017; Kajitani *et al.*, 2019). Still, this first analysis indicates that, with reference proteomes within the same phylum, family-level protein assignments are highly accurate.

The closest sequence to a query is often not in the same subfamily

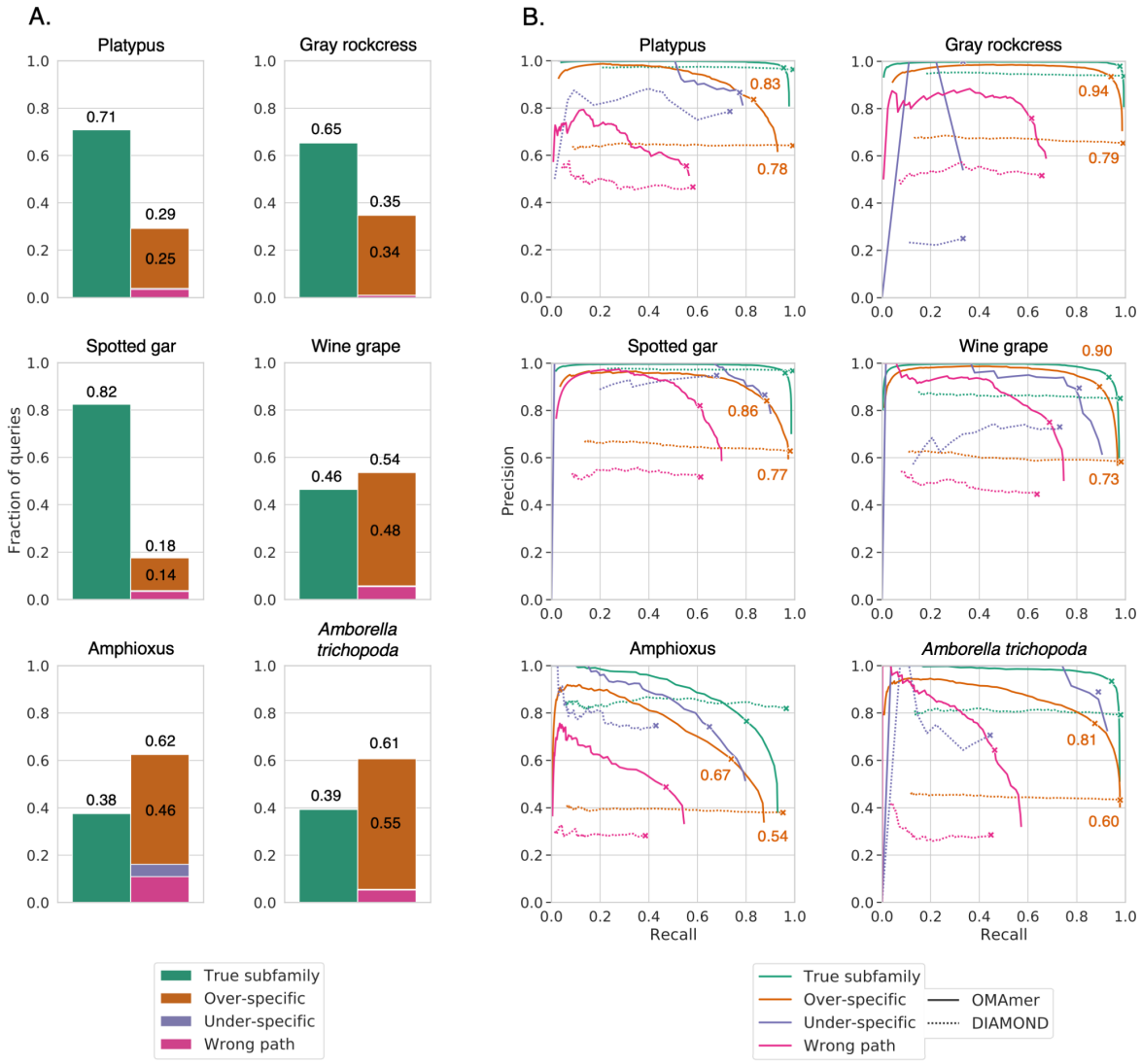


Fig. 4. Frequency of closest sequence configurations defined in Fig. 1 and OMAmer accuracy for each. A. The closest sequence to a query was often found in another subfamily. Smith-Waterman alignments were used as proxies for closest sequences. B. "Over-specific" configurations were especially well dealt with by OMAmer. Each curve displays the range of trade-offs between precision and recall when varying the threshold on the OMAmer-score and on the DIAMOND *E*-value. They were computed by breaking down queries by closest sequence configurations as in panel A, before the validation procedure itself. Crosses indicate the location of $F1_{max}$ values. "Over-specific" $F1_{max}$ values are specifically annotated.

For a large proportion of query sequences (18-62%), the closest counterpart (inferred as the highest scoring Smith-Waterman match, see *Methods*) belongs to a different subfamily (Fig. 4. A). In such cases, the closest sequence most often belongs to a more specific subfamily

(14-55% of all queries). These results highlight the need to account for the gene tree, especially in the presence of many nested subfamilies. Solving this problem is the primary aim of OMAmer.

OMAmer is more precise in subfamily placement

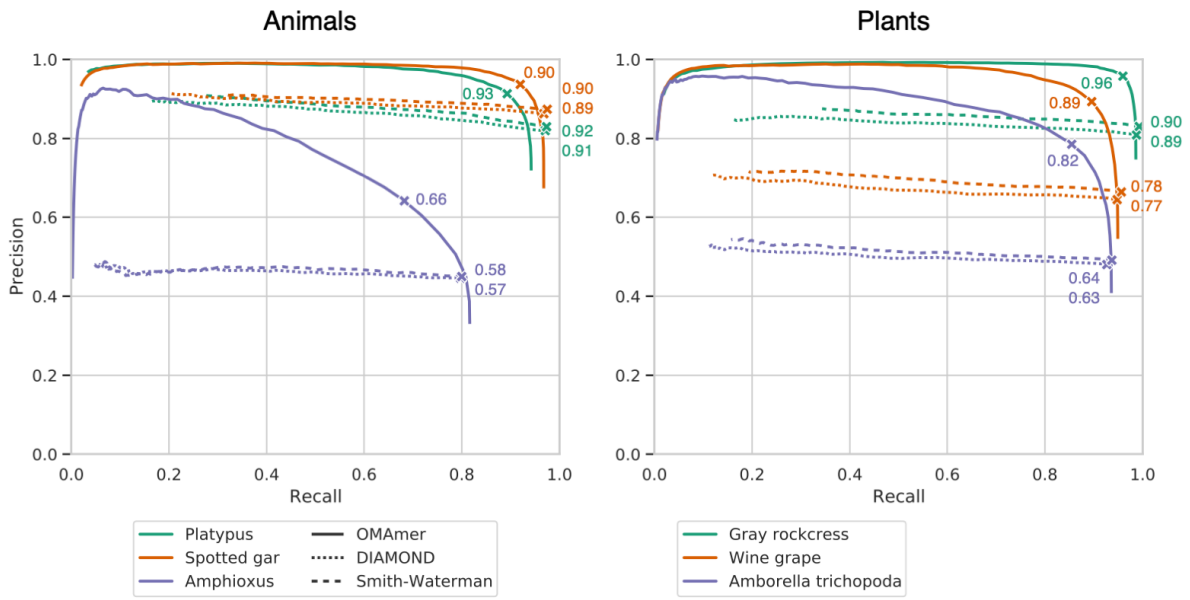


Fig. 5. Comparison of subfamily assignments with OMAmer and by closest sequence (DIAMOND and Smith-Waterman). Each curve displays the range of trade-offs between precision and recall when varying the threshold either on the OMAmer-score, on the DIAMOND *E*-value or on the Smith-Waterman alignment score. $F1_{max}$ values are indicated by crosses on each curve.

OMAmer systematically achieved, or equaled, the highest accuracy ($F1_{max}$) across species (Fig. 5.). Specifically, increases in $F1_{max}$ values between OMAmer and closest sequence methods ranged from 0.00 to 0.18. Moreover, OMAmer-score thresholds at $F1_{max}$ were generally congruent (ranging from 0.10 to 0.16), although it was lower for amphioxus (0.06).

Importantly, OMAmer provides a genuine precision-recall trade-off, providing users with the possibility of obtaining very high precision, at the cost of lower recall. There is no such possibility with closest sequence methods: varying the *E*-value and alignment-score thresholds has very limited impact on precision (Fig. 5). These results are consistent with a second and more stringent validation procedure that does not reward assignments to correct parental subfamilies (Supp. Fig. 3).

OMAmer deals especially well with over-specific closest sequences

As previously shown, over-specific placement is the most frequent mistake when only relying on assignments by closest sequences (Fig. 4. A). Since OMAmer was specifically designed to deal with such cases using evolutionarily-informed *k*-mers mapping toward ancestral subfamilies, we investigated whether this feature would explain OMAmer performance. Therefore, we reproduced the subfamily-level validation procedure with queries partitioned between the types of closest sequence configuration ("true subfamily", "under-specific", "wrong path" and "over-specific") depicted in Fig. 1. and quantified in Fig. 4 A.

As expected, OMAmer was systematically more accurate than DIAMOND for queries in the "over-specific" configuration (Fig. 4. B). Specifically, for these queries, increases in $F1_{max}$ values between OMAmer and DIAMOND ranged from 0.05 to 0.21. Moreover, OMAmer displayed a proportion of over-specific assignments (defined at $F1_{max}$) 0.07 to 0.37 lower than Smith-Waterman and DIAMOND (Supp. Fig. 5). In animals, this performance for queries in the "over-specific" configuration was achieved while sacrificing very little accuracy for queries in the "true subfamily" configuration (from 0.02 to 0.11, Fig. 4. B). In plants, OMAmer remained more accurate even for queries in the "true subfamily" configuration, with increases in $F1_{max}$ values between OMAmer and DIAMOND that ranged from 0.02 to 0.06. Queries in the "wrong-path" configuration were also placed more accurately by OMAmer, despite their small number. Finally, there were too few "under-specific" configurations to draw any conclusion.

Since DIAMOND is a closest sequence approach, like Smith-Waterman, it was expected to obtain precision values close to zero for queries in the "under-specific", "wrong path" and "over-specific" scenarios. However, this behaviour is not observed here because the validation procedure rewards assignments in correct parental subfamilies even when the predicted exact subfamily is incorrect. By contrast, the more stringent validation procedure that does not reward assignments to correct parental subfamilies does yield precision values close to zero (Supp. Fig. 4. B). Apart from this difference, the results of this section are consistent between the two validation procedures (Supp. Fig. 4. B and 5).

The occasional and counterintuitive positive correlation between precision and recall that can be observed with OMAmer at low recall values, seemed to appear only when a few FPs subfamilies remained predicted at high OMAmer-score thresholds, while the number of

TPs was steadily decreasing. Taking the example of the “wrong-path” Spotted gar proteins, 4 out of the 14 predicted subfamilies are FPs at recall of 0.02 obtained with the highest threshold value (0.99).

OMAmer run time scales better than DIAMOND with the number of reference proteomes

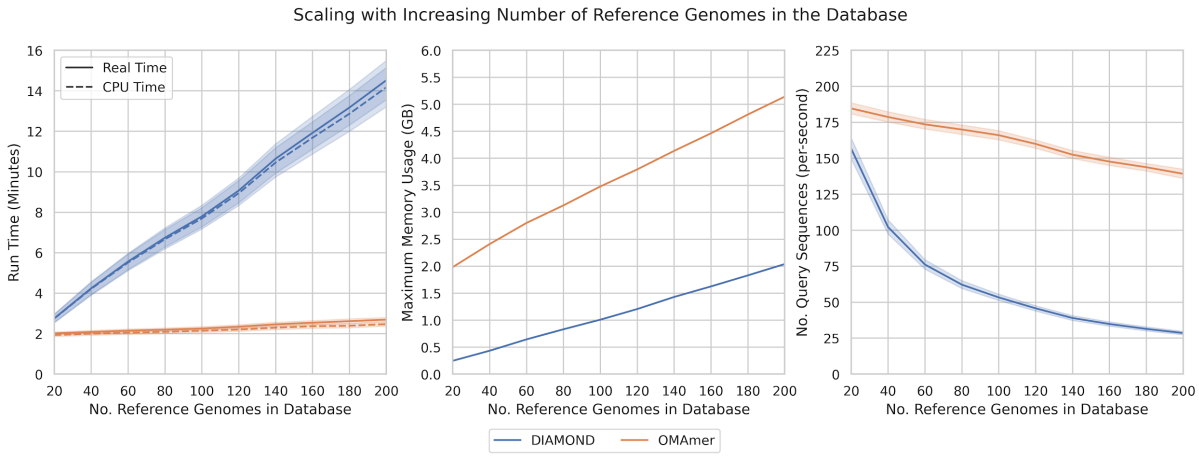


Fig. 6. Comparison of the computational performance of family and subfamily assignments between OMAmer and DIAMOND. OMAmer scales better than DIAMOND in terms of real and CPU time (total time on the left, and sequence/second on the right), but requires somewhat more memory (center) Error bars shown for 95% confidence interval, estimated using 10,000 bootstraps.

In an empirical scaling analysis, we varied the number of reference proteomes in the database whilst querying a number of full-proteomes (see *Methods* for details). OMAmer achieved better scaling than DIAMOND in terms of CPU and real time when increasing the number of reference proteomes in the database (Fig. 6, left). Both methods, however, exhibited a similar increase in maximum memory usage (Fig. 6, center), with OMAmer initially using over 2GB and DIAMOND using less than 256MB on a database of 20 reference proteomes. In order to achieve this performance, OMAmer only stores *k*-mers once per root-HOG. This does require extra computation, with the overhead being reflected in its memory usage and time to build the database (Supp. Fig. 6), taking between 15-20 minutes in comparison to 1-2 minutes for DIAMOND.

To put the timing into context, OMAmer is processing about 150 query sequences per second (Fig. 6, right). DIAMOND starts with a similar performance, before trailing off to less than 30 with the largest number of reference proteomes.

Discussion

In this study, we demonstrate that considering the phylogenetic relations between orthologous groups is essential for the problem of subfamily assignment. Indeed, although alignment-free, OMAmer generally outperforms closest sequence approaches, even when inferred by the exact Smith-Waterman algorithm. In particular, OMAmer systematically equaled or out-performed Smith-Waterman for the best precision-recall trade-off ($F1_{max}$).

However, the main advantage of OMAmer is its control over assignment precision through the setting of specific OMAmer-score thresholds that refrain over-specific placements. By contrast, relying on the closest sequence does not provide the ability for any precision-recall trade-off. Each assignment is bound to the most specific subfamily of the closest sequence and varying the *E*-value threshold has a large impact on recall but almost none on precision. Thus, while closest sequence approaches are useful for cases where high recall is the overriding priority, OMAmer is more flexible and applicable in a broad range of contexts.

In addition to providing robust subfamily assignments, OMAmer scales better than DIAMOND, in terms of run time, with the number of reference proteomes. This is achieved with alignment-free sequence comparisons against hierarchical orthologous groups (HOGs) instead of approximate alignments against protein sequences. Indeed, in addition to removing the computational burden of sequence alignment, merging sequence information in HOGs drastically reduces the number of comparisons. This is especially true since the number of reference HOGs increases more slowly than proteins with the number of reference proteomes.

Large-scale sequencing projects of genomes or metagenomes add difficulties such as chimeric assemblies or contaminations, thus mixing gene families from different species. OMAmer was designed as a starting point for the integration of such heterogeneous data. Thus, instead of constraining subfamily assignments along the known taxonomy of query proteomes, OMAmer performs taxonomically blind assignments. We hope that this feature will enable diverse applications of OMAmer. For example, the detection of contamination and horizontal gene transfers could be achieved by including all kingdoms in the OMAmer database and searching for incongruent placement regarding the query taxonomy. In particular, confidence measures similar to the "Alien index" (Gladyshev *et al.*, 2008) could be computed by subtracting the OMAmer-score of the highest-scoring taxonomically congruent HOG from the overall highest OMAmer-score potentially derived from a contaminant sequence. Other

promising applications are the binning of protein-level metagenomic assemblies (Steinegger *et al.*, 2019), and with some algorithmic adaptations, directly placing reads to skip genome assembly and annotation.

The OMAmer algorithm builds upon some key ideas of the metagenomic software Kraken, which classifies reads into the species taxonomy (Wood and Salzberg, 2014). Indeed, this task is analogous to protein subfamily assignments for two reasons. First, some prior knowledge, shaped as labelled reference sequences, is preprocessed before the assignment itself. Second, this prior knowledge is organized hierarchically in a tree graph. Thus, instead of relying on closest sequences, such methods of taxonomic classification exploit semi-phylogenetic information to improve their predictions. While MEGAN introduced the key idea of taking the LCA taxon among significant BLAST hits (Huson *et al.*, 2007), Kraken scaled up the approach by preprocessing LCA taxa in a database of taxonomically-informed *k*-mers (Wood and Salzberg, 2014).

While inspired by Kraken, the OMAmer algorithm features three key algorithmic innovations to fit the case of assigning proteins to subfamilies. The first difference lies in the types of events used to define clades or subtrees. Indeed, while taxa are defined by speciation nodes in Kraken, subfamilies are defined by duplication nodes in OMAmer. This is an important difference because duplication patterns are variable across protein families, whereas the reference taxonomy is the same for different genes and genomes in Kraken. Second, the dual problem of first placing sequences within families, followed by subfamily-level assignment is specific to OMAmer. Third, while Kraken relies on an arbitrary cut-off of one *k*-mer to avoid over-specific placements, OMAmer applies a user-defined threshold on the more refined OMAmer-score.

Beside closest sequence approaches, alignments to Hidden Markov Models (HMMs) have been extensively used for sequence to family or subfamily comparisons with tools such as HMMER3 (El-Gebali *et al.*, 2019; Mi *et al.*, 2019; Huerta-Cepas *et al.*, 2019; Ebersberger *et al.*, 2009). However, the use of HMMs is revealing a lack of scalability to phylogenomic database size. For instance, the developers of the EggNOG database reported that DIAMOND is considerably faster and achieves similar results to HMMER3 and have discontinued the use of HMMs in the latest EggNOG mapper release (Huerta-Cepas *et al.*, 2017, 2019). Moreover, maintaining subfamily HMM models can be problematic because it relies on ad-hoc criteria for subfamily delineation (*e.g.* curated, family-specific *E*-value thresholds in Pfam [El-Gebali

et al., 2019]). Finally, HMMs are tailored to detect remote homology rather than discriminating between specific subfamilies. Although this has benefited from hierarchically organized HMMs (Nguyen *et al.*, 2016), the family breakdown is used to improve family assignments rather than finding specific subfamilies.

Due to the rapid emergence of alignment-free methods, covering various biological problems ranging from phylogenetic inference to metagenomic taxonomic profiling (*reviewed in* [Zielezinski *et al.*, 2017]), the AFproject was launched to unite the benchmarking of these tools (Zielezinski *et al.*, 2019). However, the available datasets to benchmark protein sequence classification in that project are organized according to the SCOPE database (Fox *et al.*, 2014). There, each hierarchical level is either based on a degree of belief in homology among sets of proteins (families and superfamilies) or on structural similarities (folds and classes). By contrast, in this work, we seek to distinguish all subfamilies resulting from gene duplications, even recent ones yielding quite similar subfamilies. Of note, recent subfamilies can diverge in function (Naseeb *et al.*, 2017) and thus be important for annotation.

In this work, we used the most similar sequence (whether inferred exactly by Smith-Waterman or by the fast heuristic DIAMOND) as reference to find the closest sequence. Although the highest scoring local alignment is not always the closest sequence in a phylogenetic sense (Koski and Golding, 2001), this is a commonly used approximation for classifying large numbers of orthologs (Li *et al.*, 2003; Sonnhammer and Östlund, 2015; Huerta-Cepas *et al.*, 2019) and has shown to give similar results in simulation (Dalquen *et al.*, 2013) and empirical benchmarks (Altenhoff *et al.*, 2016)

Although placing proteins at the overall family level appears to be easier than at the subfamily level, we start to see some degradation with the amphioxus sequences (last common ancestor to vertebrates 600MYA [Peterson and Eernisse, 2016]). We expect further degradation for cases where query proteomes are even farther from the reference proteomes, because relying on *k*-mer exact matches is likely to be less sensitive than alignments such as provided by DIAMOND to detect distant homologs. Some avenues to increase OMAmer sensitivity in the absence of closely related reference species could be explored: the use of a reduced alphabet, which compresses the mutual information of sequences being compared (Edgar, 2004); or spaced seeds, *i.e.* non-contiguous *k*-mers, that have shown an increased sensitivity in metagenomics classification (Břinda *et al.*, 2015). On the other hand, adding such very distant proteomes is expected to be much rarer than adding proteomes to an already sampled clade.

This is especially true for the increase of sequences through projects such as i5k (insect genomes) (i5K Consortium, 2013) or the Vertebrate Genomes Project (Koepfli *et al.*, 2015), where duplications and thus subfamilies are common and a solid backbone of reference proteomes are available. OMAmer is especially well positioned to help classify the genes from such projects, which will present a challenge for slower or less precise methods.

Acknowledgements

We thank Adrian Altenhoff for fruitful discussions in the conception and validation of OMAmer and Yannis Nevers for proof-reading the manuscript. Computations were performed at the Vital-IT Center for high performance computing of the SIB Swiss Institute of Bioinformatics, as well as the Wally and Axiom clusters of the University of Lausanne.

Funding

This work was supported by the Swiss National Foundation grant No 167276, as part of the National Research Program 75 “Big Data”, as well as Swiss National Foundation grant No 183723.

Conflict of Interest: none declared.

References

Altenhoff,A.M. *et al.* (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS One*, **8**, e53786.

Altenhoff,A.M. *et al.* (2016) Standardized benchmarking in the quest for orthologs. *Nat. Methods*, **13**, 425–430.

Altenhoff,A.M. *et al.* (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.*, **46**, D477–D485.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Amborella Genome Project (2013) The Amborella genome and the evolution of flowering plants. *Science*, **342**, 1241089.

Barbera,P. *et al.* (2018) EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Syst. Biol.*

Betancur-R,R. *et al.* (2017) Phylogenetic classification of bony fishes. *BMC Evol. Biol.*, **17**, 162.

Břinda,K. *et al.* (2015) Spaced seeds improve k-mer-based metagenomic classification. *Bioinformatics*, **31**, 3584–3592.

Buchfink,B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

Conant,G.C. and Wolfe,K.H. (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.*, **9**, 938–950.

Dalquen,D.A. *et al.* (2013) The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. *PLoS One*, **8**, e56925.

Ebersberger,I. *et al.* (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol. Biol.*, **9**, 157.

Edgar,R.C. (2004) Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Res.*, **32**, 380–385.

El-Gebali,S. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.

Fox,N.K. *et al.* (2014) SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–9.

Gabaldón,T. and Koonin,E.V. (2013) Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.*, **14**, 360–366.

Gladyshev,E.A. *et al.* (2008) Massive horizontal gene transfer in bdelloid rotifers. *Science*, **320**, 1210–1213.

Glover,N. *et al.* (2019) Advances and Applications in the Quest for Orthologs. *Mol. Biol. Evol.*, **36**, 2157–2164.

Huang,S. *et al.* (2017) HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics*, **33**, 2577–2579.

Huerta-Cepas,J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.

Huerta-Cepas,J. *et al.* (2017) Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.*, **34**, 2115–2122.

Huson,D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.

i5K Consortium (2013) The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J. Hered.*, **104**, 595–600.

Jaillon,O. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Kajitani,R. *et al.* (2019) Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nat. Commun.*, **10**, 1702.

Koepfli,K.-P. *et al.* (2015) The Genome 10K Project: a way forward. *Annu Rev Anim Biosci*, **3**, 57–111.

Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.*, **52**, 540–542.

Kriventseva,E.V. *et al.* (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.*, **47**, D807–D811.

Li,L. *et al.* (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.

Linard,B. *et al.* (2019) Rapid alignment-free phylogenetic identification of metagenomic sequences. *Bioinformatics*.

Manber,U. and Myers,G. (1993) Suffix Arrays: A New Method for On-Line String Searches. *SIAM J. Comput.*, **22**, 935–948.

Mi,H. *et al.* (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.*, **47**, D419–D426.

Naseeb,S. *et al.* (2017) Rapid functional and evolutionary changes follow gene duplication in yeast. *Proc. Biol. Sci.*, **284**.

Nguyen,N.-P. *et al.* (2016) HIPPI: highly accurate protein family classification with ensembles of HMMs. *BMC Genomics*, **17**, 765.

Opazo,J.C. *et al.* (2008) Differential loss of embryonic globin genes during the radiation of placental mammals. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 12950–12955.

Peterson,K.J. and Eernisse,D.J. (2016) The phylogeny, evolutionary developmental biology, and paleobiology of the Deuterostomia: 25 years of new techniques, new discoveries, and new ideas. *Org. Divers. Evol.*, **16**, 401–418.

Putnam,N.H. *et al.* (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064–1071.

Schreiber,F. *et al.* (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res.*, **42**, D922–D925.

Sémon,M. and Wolfe,K.H. (2007) Consequences of genome duplication. *Curr. Opin. Genet. Dev.*, **17**, 505–512.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Sonnhammer,E.L.L. and Östlund,G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, **43**, D234–9.

Steinegger,M. *et al.* (2019) Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods*, **16**, 603–606.

Tang,H. *et al.* (2019) TreeGrafter: phylogenetic tree-based annotation of proteins with Gene Ontology terms and other annotations. *Bioinformatics*, **35**, 518–520.

UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.

Upham,N.S. *et al.* (2019) Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol.*, **17**, e3000494.

Willing,E.-M. *et al.* (2015) Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation. *Nat Plants*, **1**, 14023.

Wolf,Y.I. and Koonin,E.V. (2012) A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol. Evol.*, **4**, 1286–1294.

Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.

Zielezinski,A. *et al.* (2017) Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.*, **18**, 186.

Zielezinski,A. *et al.* (2019) Benchmarking of alignment-free sequence comparison methods. *Genome Biol.*, **20**, 144.

Supplementary material

Supplementary methods

k-mer integer encoding

The integer encoding of a *k*-mer x formed of k numerical characters x_i, ordered from $i = 1$ to $i = k$, from an alphabet A (A:0, C:1, …, Y:20) is defined as:

$$\sum_{i=1}^{k} x_i |A|^{k-i}$$

OMAmer-score computation

The coarse alignment-free similarity (1) is measured as the number of intersecting *k*-mers between the query *k*-mers Q and the HOG specific *k*-mers H. H includes *k*-mers specific to the HOG descendants but excludes the ones conserved in its ancestors. To compute (1), the number of intersecting *k*-mers between the query *k*-mers and each HOG ancestral *k*-mer set (the *k*-mers inferred to have arisen in the HOG) is retrieved from the precomputed *k*-mer table. Then, these counts are cumulated from leaves to root by adding the highest child HOG *k*-mer count to the current HOG count at each multifurcation.

$$|Q \cap H| \quad (1)$$

To account for the different sizes (number of different *k*-mers) of reference HOGs and the query composition bias, the expected number of shared *k*-mers between the query and the HOG observed in absence of homology, *i.e.* by chance, (2) is subtracted from (1) (3). OMAmer proposes a parametric (default OMAmer-score) and a non-parametric approach (sensitive OMAmer-score) to compute (2).

$$E(|Q \cap H|) \quad (2)$$

$$|Q \cap H| - E(|Q \cap H|) \quad (3)$$

In the parametric approach, (2) is calculated as the number of query *k*-mers $|Q|$ multiplied by the probability to observe one query *k*-mer x_q in H.

$$E(|Q \cap H|) = |Q| \, P(x_q \in H)$$

This probability is the inverse probability of not observing one x_q in H.

$$P(x_q \in H) = 1 - (1 - P(x_q))^{|H|}$$

The probability of observing $x_q$in a HOG of size one, *i.e.* with one *k*-mer, $P(x_q)$ is approximated as the mean frequency of query *k*-mers inside the *k*-mer table (the average fraction of HOGs containing each query *k*-mer x_i [remember that each *k*-mer can only be stored once per root-HOG]).

$$P(x_q) = \frac{1}{|Q|}\sum_{i=0}^{|Q|} freq(x_i)$$

In the non-parametric approach, (2) is simply (1) obtained from a random permutation of the query sequence. In an attempt to conserve some local composition bias, the permutation is performed by shuffling windows of size six in addition to shuffling individual amino acids within each such window. Note that this approach additionally corrects for HOG composition biases.

Finally, to make the OMAmer-score comparable across queries, (3) is divided by $|Q|$, from which was subtracted the number of query *k*-mers shared with more ancestral HOGs.

$$OMAmer{-}score{=}|Q \cap H| - E(|Q \cap H|)|Q| - |Q \cap ancestors(H)|$$

Datasets and software parameters

OMAmer was compared with two closest sequence methods lying at different extremes of the speed-accuracy tradeoff: DIAMOND (v0.9.14) and Smith-Waterman, respectively. Due to the computational cost of performing Smith-Waterman alignments, we used pre-computed alignments from OMA (January 2020) (Altenhoff *et al.*, 2018). DIAMOND databases were built with default parameters, and searches for the most similar sequence were performed with effectively no significance requirement (*E*-value set to 1e6). The OMAmer *k*-mer table was built with a *k*-mer size of 6.

OMAmer directly yields family and subfamily predictions. For Smith-Waterman and DIAMOND, each query was assigned to the family and most specific subfamily of its closest reference protein. To obtain multiple precision-recall values, predictions were computed for

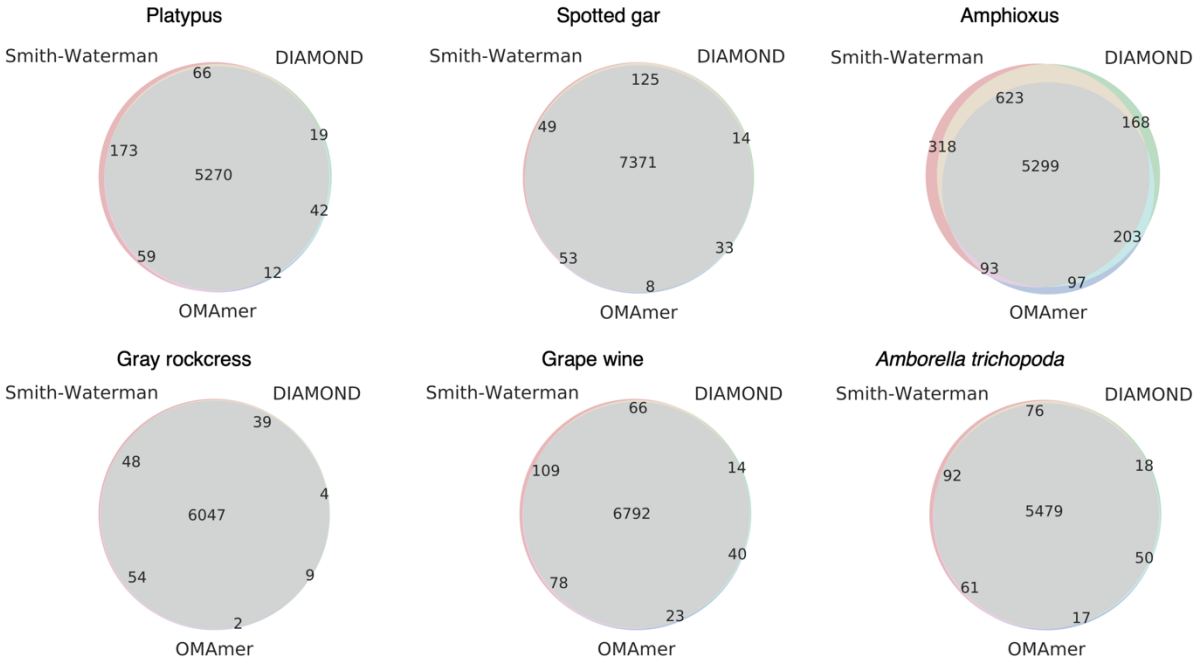
multiple score thresholds: *E*-values of 1e-322 to 1e6 for DIAMOND, alignment scores of 1 to 5,000 for Smith-Waterman and OMAmer-scores of 0 to 0.99.

To make family-level assignments comparable and well differentiated from subfamily-assignments, we selected HOGs from OMA (January 2020) defined at the *Metazoa* and *Viridiplantae* taxonomic levels as root-HOGs (families), and their sub-HOGs as subfamilies. To avoid low-confidence families, we further filtered out root-HOGs with less than six proteins. We picked *Metazoa* because it is one of the largest clades in OMA and *Viridiplantae* due to the high number of duplications and thus subfamilies in this clade. Note, due to the addition of *Branchiostoma lanceolatum* in the January 2020 OMA release, we removed it from the reference database used (before the *k*-mer index precomputation) to keep the same evolutionary distance existing between *Branchiostoma floridae* and reference proteomes of the previous OMA release (June 2019).

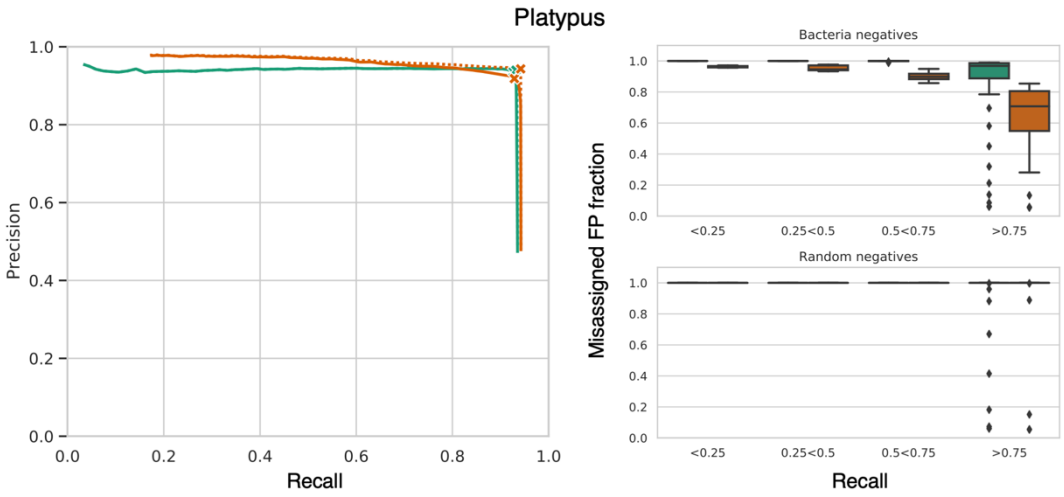
Then, we selected six species as experiment targets picked because they stand as outgroups of large clades in OMA and thus display some variability in divergence ages to reference species. Platypus, spotted gar and amphioxus were selected in *Metazoa*, while Gray rockcress, wine grape and *Amborella trichopoda* were chosen in *Viridiplantae* (Supp. Table 2). Clade-specific root-HOGs used to build the negative query set were picked at the *Bacteria* taxonomic level.

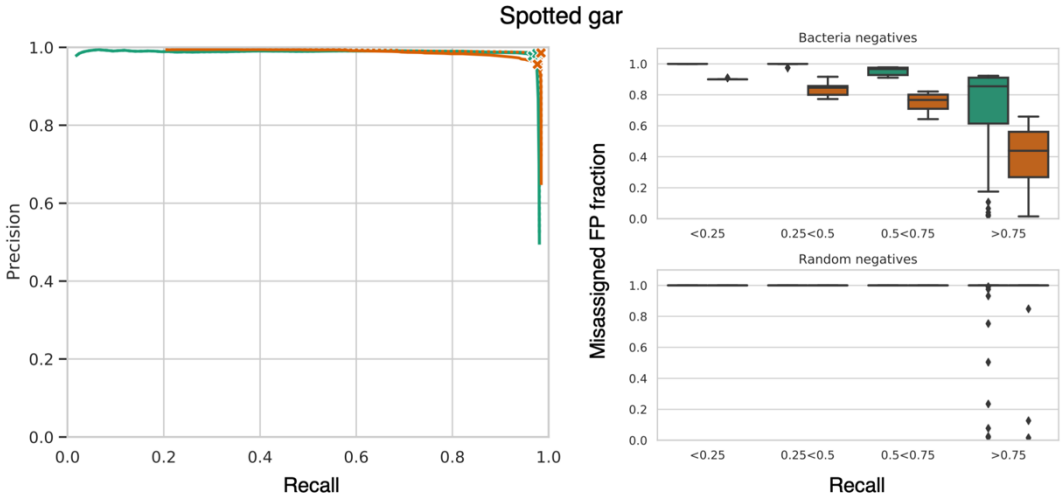
The *Metazoa* reference dataset included 1,309,488 proteins from 201 species organized in 235,983 HOGs and including 12,178 root-HOGs. The *Viridiplantae* reference dataset included 554,389 proteins from 63 species organized in 304,838 HOGs and including 8,652 root-HOGs. The query datasets (proteomes) included 5,811, 7,7227,387, 6,239, 7,219 and 5,931 proteins of platypus, spotted gar, amphioxus, gray rockcress, wine grape and *Amborella trichopoda* species, respectively. 4,952, 6,308, 5,803, 5,261, 5,712 and 4,057 queries belonged to a sub-HOG in addition to the root-HOG.

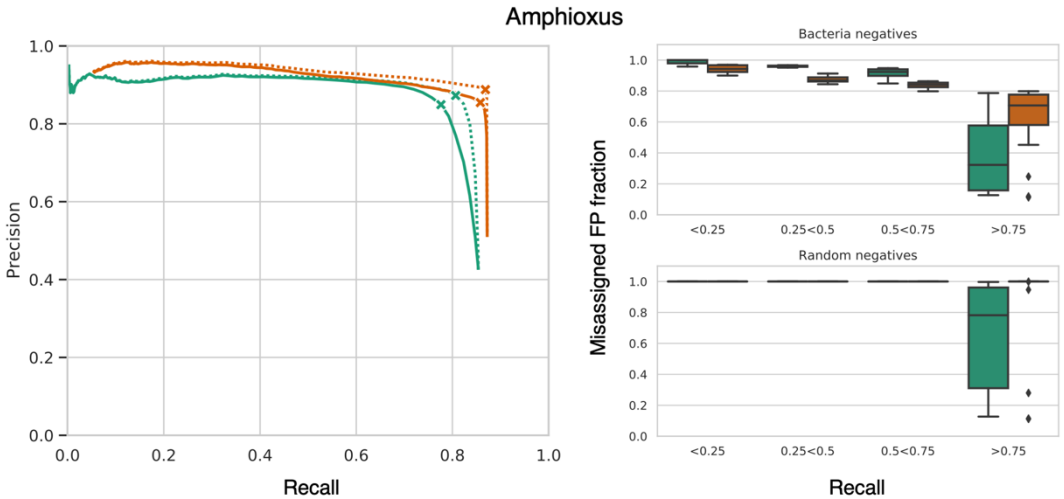
Supplementary figures and tables

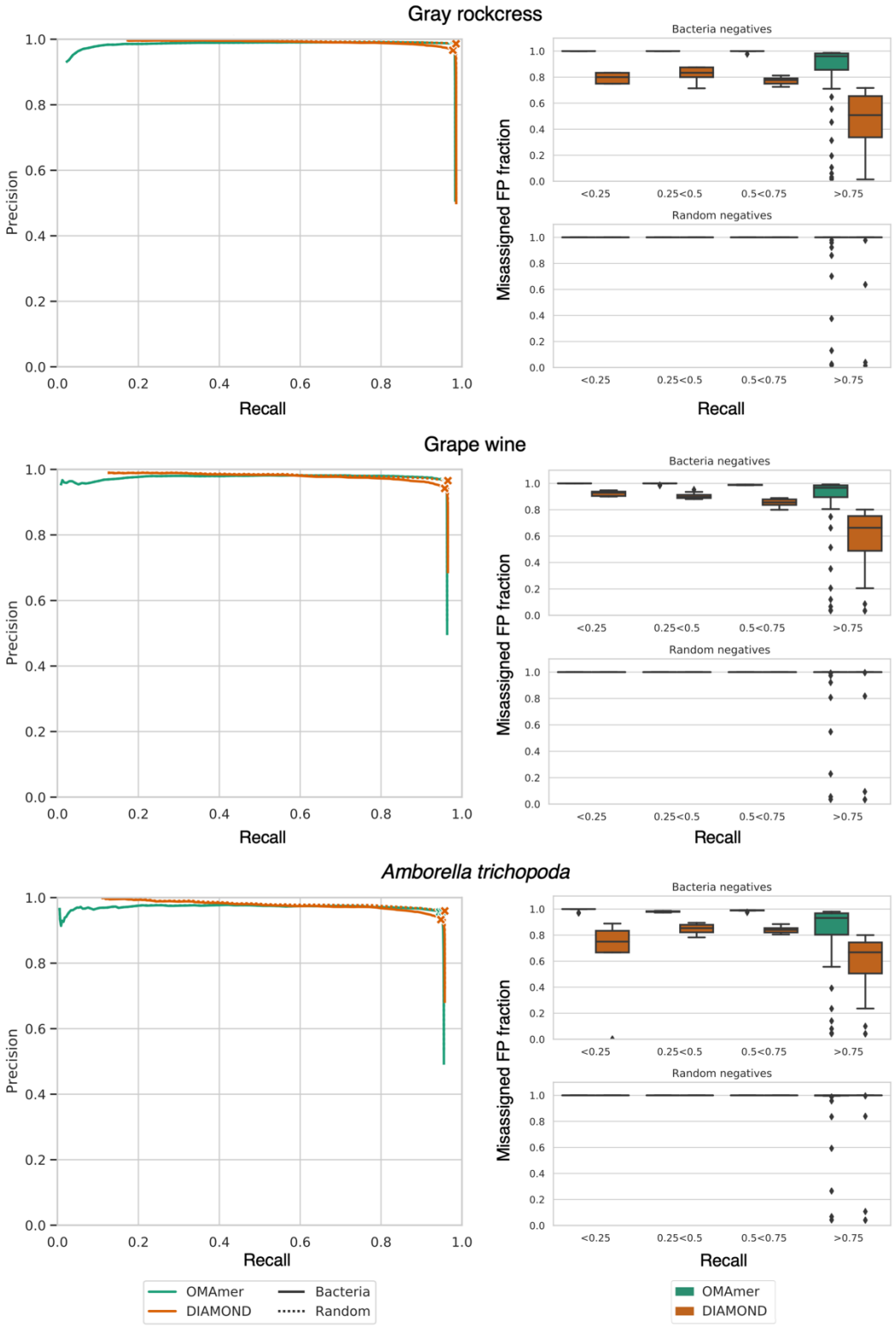


Supp. Fig. 1. Number of family-level TP queries overlapping between methods. TP sets were defined at F1max for DIAMOND and OMAmer and at the minimum score (1) for Smith-Waterman alignments. These queries were used to assess subfamily assignment.


Platypus
Precision
Recall
Misassigned FP fraction
Bacteria negatives
Random negatives
<0.25
0.25<0.5
0.5<0.75
>0.75

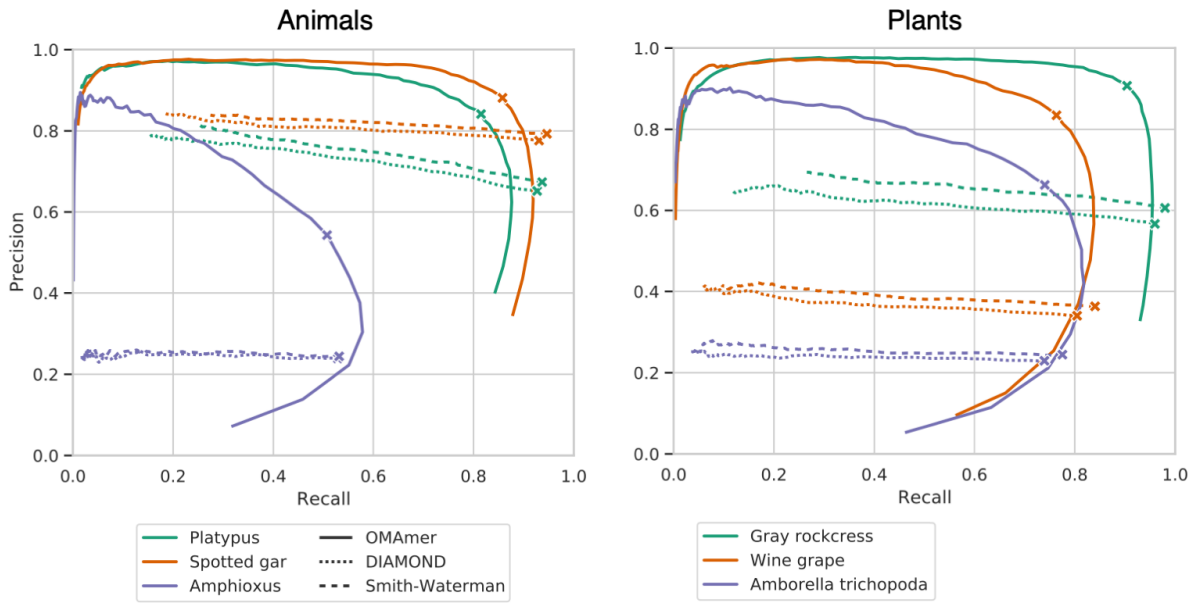

Spotted gar
Precision
Recall
Misassigned FP fraction
Bacteria negatives
Random negatives
<0.25
0.25<0.5
0.5<0.75
>0.75


Amphioxus
Precision
Recall
Misassigned FP fraction
Bacteria negatives
Random negatives
<0.25
0.25<0.5
0.5<0.75
>0.75



Supp. Fig. 2. Comparison of family assignments between OMAmer and DIAMOND across negative datasets. (Left) Each curve displays the range of trade-offs between precision and recall when varying the threshold on the OMAmer-score or on the DIAMOND *E*-value. The curves labeled *Bacteria* refer to analyses using bacteria-specific sequences as negatives whereas those labeled *Random* refer to

using random sequences as negatives. Crosses indicate the location of F1max values. (Right) Fraction of FPs coming from the misassignment of positive sequences.



Supp. Fig. 3. Comparison of subfamily assignments with OMAmer and by closest sequence (Smith-Waterman and DIAMOND). Each curve displays the range of trade-offs between precision and recall when varying the threshold either on the OMAmer-score, on the DIAMOND *E*-value or on the Smith-Waterman alignment score. These results were computed using the more stringent validation procedure. $F1_{max}$ values are annotated with crosses.

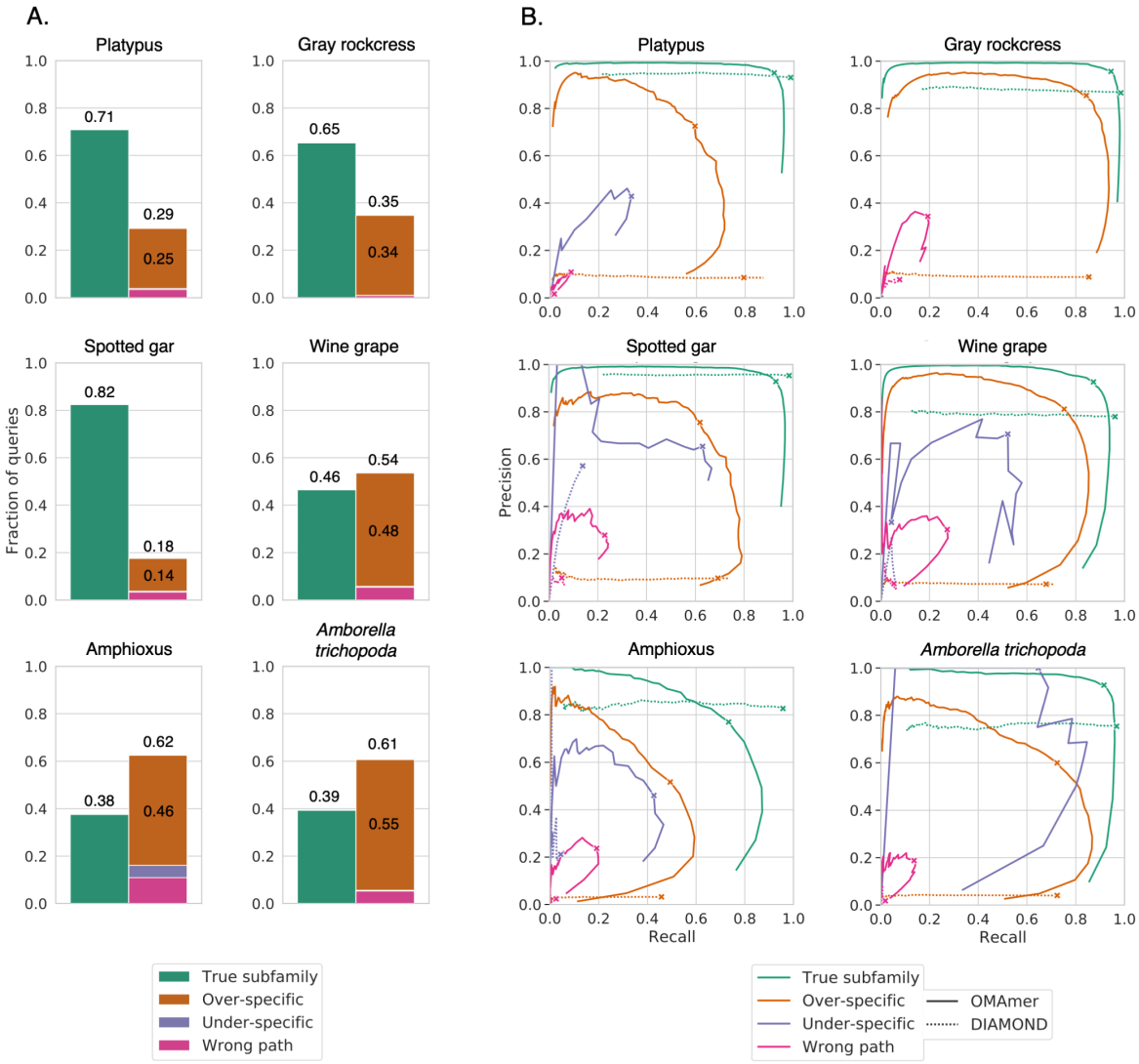
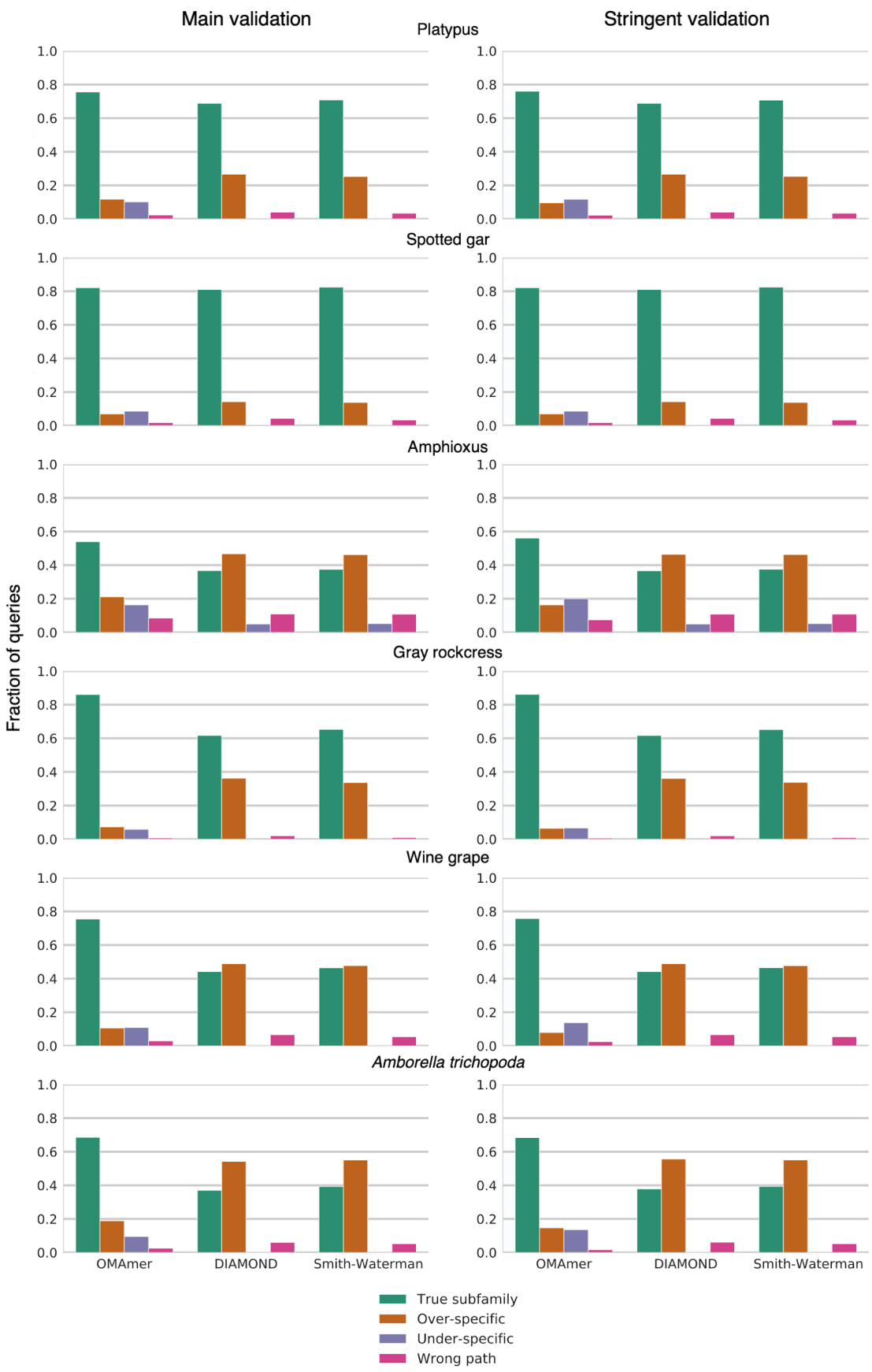
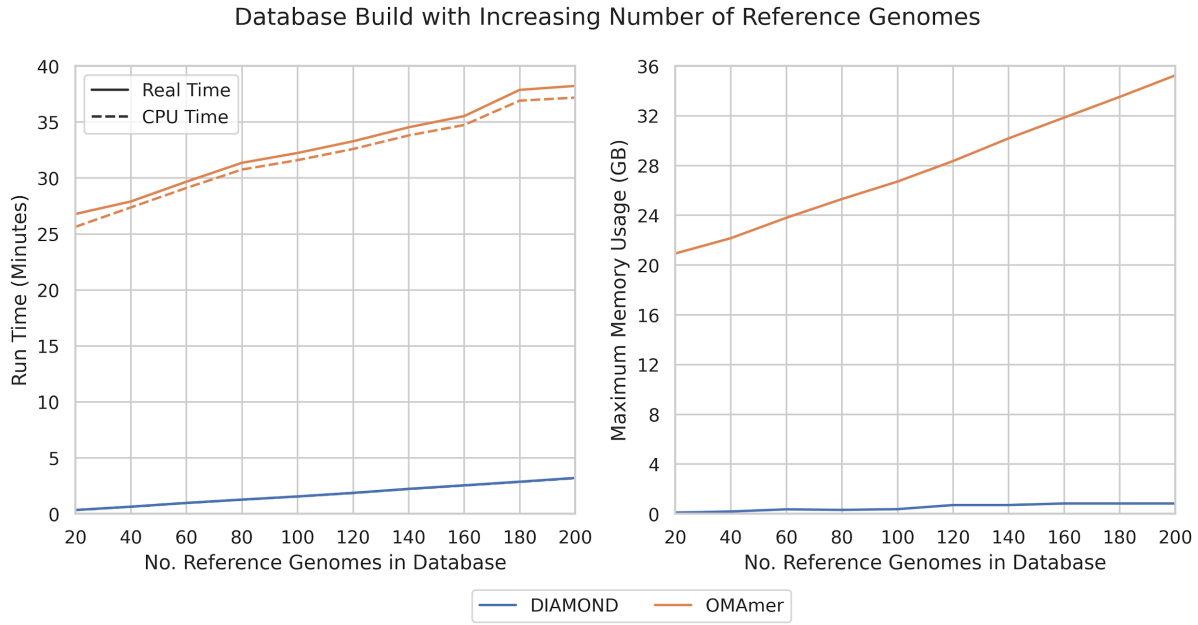


Fig. 4. Frequency of closest sequence configurations defined in Fig. 1 and OMAmer accuracy for each. A. The closest sequence to a query was often found in another subfamily. Smith-Waterman alignments were used as proxies for closest sequences. B. These results were computed using the more stringent validation procedure (*See methods*). Each curve displays the range of trade-offs between precision and recall when varying the threshold on the OMAmer-score and on the DIAMOND *E*-value. They were computed by breaking down queries by closest sequence configurations as in panel A, before the validation procedure itself. $F1_{max}$ values are annotated with crosses. Crosses indicate the location of $F1_{max}$ values. "Over-specific" $F1_{max}$ values are specifically annotated.



Supp. Figure 5. Partitioning of subfamily assignments at $F1_{max}$ into closest sequence configuration.



Supp. Figure 6. Run time (left) and maximum memory usage (right) during database build for DIAMOND and OMAmer. Whilst OMAmer is slower and requires more memory due to the increased pre-processing to enable fast lookup time, the increase in time and memory is linear with the number of reference genomes in the resulting database.

Supp. Table 1. Formulae of validation measures

Measure	Formula
Precision	$\frac{\#TPs}{(\#TP + \#FPs)}$
Recall	$\frac{\#TPs}{(\#TP + \#FNs)}$
Accuracy	$2x \frac{(precision * recall)}{(precision + recall)}$

#: number, TPs: true positives, FPs: false positives, FNs: false negatives.

Supp. Table 2. Species used as queries in benchmarks.

Species	Scientific name	LCA clade	Divergence age (mya)	Genome scaffold N50 (kb)
Spotted Gar	*Lepisosteus oculatus*	*Neopterygii*	320 (Betancur-R *et al.*, 2017)	6928 (Ensembl LepOcu1 assembly)
Platypus	*Ornithorhynchus anatinus*	*Mammalia*	250 (Upham *et al.*, 2019)	992 (Ensembl OANA5 assembly)
Amphioxus	*Branchiostoma floridae*	*Chordata*	600 (Peterson and Eernisse, 2016)	2600 (Putnam *et al.*, 2008)
Gray rockcress	*Arabis alpina*	*Brassicaceae*	27 (Willing *et al.*, 2015)	788 (Willing *et al.*, 2015)
Grape wine	*Vitis vinifera*	*Rosids*	>130 (Jaillon *et al.*, 2007)	2070 (Jaillon *et al.*, 2007)
Amborella trichopoda	*Amborella trichopoda*	*Magnioliopsida*	>160 (Amborella Genome Project, 2013)	4900 (Amborella Genome Project, 2013)

References

Amborella Genome Project (2013) The Amborella genome and the evolution of flowering plants. Science, 342, 1241089.

Betancur-R,R. et al. (2017) Phylogenetic classification of bony fishes. BMC Evol. Biol., 17, 162.

Jaillon,O. et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 449, 463–467.

Peterson,K.J. and Eernisse,D.J. (2016) The phylogeny, evolutionary developmental biology, and paleobiology of the Deuterostomia: 25 years of new techniques, new discoveries, and new ideas. Org. Divers. Evol., 16, 401–418.

Putnam,N.H. et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature, 453, 1064–1071.

Upham,N.S. et al. (2019) Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. PLoS Biol., 17, e3000494.

Willing,E.-M. et al. (2015) Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation. Nat Plants, 1, 14023.

Chapter 3

Matreex: compact and interactive visualisation of large gene families using hierarchical phylogenetic profiles

Matreex: compact and interactive visualisation of large gene families using hierarchical phylogenetic profiles

Victor Rossier, Clement Train, Yannis Nevers, Marc Robinson-Rechavi and Christophe Dessimoz

Introduction

Studying the evolutionary dynamics of gene families strongly benefits from appropriate visualisation tools. For example, we can draw evolutionary and functional hypotheses by visually correlating gene repertoires with adaptations or between families. Moreover, visualising the evolutionary history of a gene family provides the framework to generalise classical pairwise gene relationships (*e.g.* orthology and paralogy) to multiple species (Dunn and Munro 2016). However, the growing number of genomes sequenced and processed by comparative genomic pipelines results in increasingly larger gene families. For example, the OMA database provides families with up to 120'366 members across 2'496 species (All.Dec2021 release) (Altenhoff et al. 2021). Gene family visualisation tools able to integrate this large data volume are needed.

Gene trees labelled with duplications and speciations are typically used to depict the evolutionary history of gene families. However, existing gene tree viewers are not equipped to provide overviews of evolutionary trajectories required to study large gene families spanning thousands of taxa and dozens of subfamilies. To keep gene trees interpretable, most viewers merely rely on collapsing or trimming subtrees, by letting users dynamically expand the relevant ones, while collapsing others (Herrero et al. 2016; Mi et al. 2017; Nguyen et al. 2018; Fuentes et al. 2021). For example, the GeneView of Ensembl collapses by default all subtrees lying outside the lineage of the query gene and provides the option to collapse all nodes at a given taxonomic rank (Herrero et al. 2016). Similarly, the PhyloView of Genomicus displays the gene tree at a user-defined taxon and provides many customization features such as trimming outgroups (relative to the query gene) or duplication nodes (Nguyen et al. 2018). However, a collapsed or trimmed subtree is mostly uninformative, as its gene content and topology are not shown. Therefore, users can only choose between keeping a complete and often intractable gene tree or collapsing nodes and hiding the information of its children, with no middle ground. Moreover, these viewers are limited by their slow reactivity, which makes the exploration of large gene trees cumbersome. For example, a couple of seconds is needed to collapse a node in Ensembl GeneView or PhylomeDB, while any action brings the user back

to the top of the page in Genomicus PhyloView. Faster and more scalable web-based tools have been introduced to visualise large phylogenies of species or of viral genomes (Robinson et al. 2016; Turakhia et al. 2020), but they are not tailored to display gene families and also lack a way to summarise relevant information contained in the different relevant parts of the phylogenies.

Alternatively, gene families can be represented as vectors of gene copy numbers across species or phylogenetic profiles. Although these were initially developed to infer gene functions, as repeated co-occurrences provide evidence of interaction (Pellegrini et al. 1999), visualising these profiles has proven useful to illustrate the gene content of extant species (Musilova et al. 2021; Horn et al. 2022) or to compare likely coevolving families (van Dam et al. 2013; Nevers et al. 2017). Indeed, displaying the full gene repertoire of a species in the same column (or row) and all gene family members in the same row (or column) enables rapid visual identification of repeated and correlated gene presence and absences. The relevance of this kind of compact representation of gene families is evidenced by the large number of tools developed for that task (Sadreyev et al. 2015; Cromar et al. 2016; Tran et al. 2018; Tremblay et al. 2021; Ilnitskiy et al. 2022). However, unlike gene trees, phylogenetic profiles do not show evolutionary relationships among the genes; for instance, it is not possible to deduct from a profile alone whether two gene absences are the result of independent losses, or a single loss in a common ancestor.

Here, we introduce Matreex, an innovative viewer for large gene families that bridges the gap between these two typical representations of gene families: gene trees that provide their complete evolutionary picture but can be cumbersome to read and phylogenetic profiles that efficiently depict the distribution of genes across species but lack the evolutionary component. Matreex builds on the reactive framework from the Phylo.IO viewer (Robinson et al. 2016) and integrates phylogenetic profiles to summarise collapsed subtrees. Thus, it simplifies gene tree visualisation while reducing the information loss. The resulting highly compact and reactive visualisation of evolution enables Matreex to scale-up to the ongoing deluge of genomic data. We illustrate Matreex with three biological applications.

New Approach

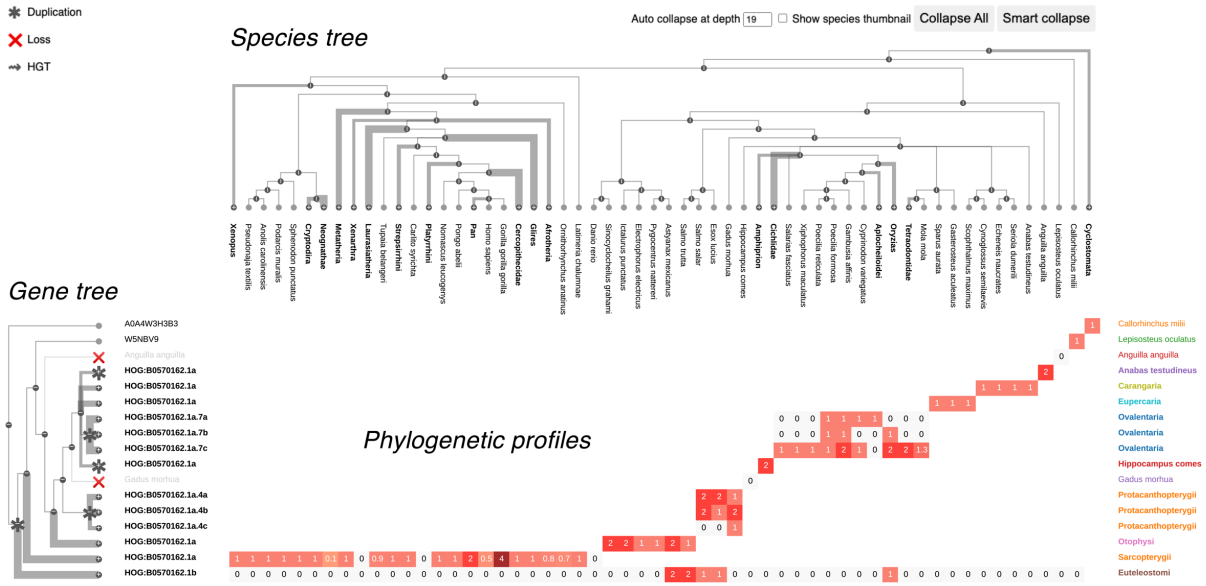


Fig 1. Matreex' layout consists of a gene tree, a species tree and a matrix of phylogenetic profiles. Branch thickness increases with the collapsed subtree size and cell colour darkness with the number of gene copies in the cell. The taxonomic levels of collapsed subtrees and phylogenetic profiles are annotated on the left. "Auto-collapse at depth" enables the automatic collapse of the species tree at a given depth from the root. "Show species thumbnail" enables displaying a taxon image (at present from Wikipedia) when hovering over a taxon. "Collapse All" and "Smart collapse" are two default views described in the main text. The examples shown are red-sensitive visual opsins (data from OMA All.Dec2021). Italic annotations do not belong to the Matreex layout but are added for figure clarity.

By enabling a compact and reactive visualisation of large gene families, Matreex provides the opportunity for new biological discoveries, for the production of paper figures, and for didactic support for teaching in evolutionary biology. Matreex complements the gene tree with a matrix of phylogenetic profiles and a species tree (Fig. 1). Thus, when collapsing a subtree to simplify the gene tree, the distribution of gene copy numbers across species remains available. This informs about ancestral events such as gene loss, duplication or transfer. Moreover, the species tree provides what is often a good proxy for the topology of these subtrees (Morel et al. 2020). Interestingly, the extreme case, where all subtrees without duplications are collapsed, provides the same information as a fully extended gene tree, but much more compact. This is achieved by removing most of the species tree redundancy from the gene tree. For example, displaying all information in the PANTHER (v.17) MutS family (762 genes) required 17 print screens with the PANTHER gene tree viewer versus four with Matreex (Supp. Figure 1). We provide rapid access to this view with Matreex' "Smart Collapse" option.

For genes families with a high number of duplication events, collapsing only subtrees without duplication is not enough and summarising them requires also collapsing subtrees with subfamilies (children of duplication nodes). In that case, the resulting phylogenetic profiles depict the combined gene content of each subfamily per species. For example, when collapsing a subtree in the insulin family, the profile will show only one copy for primates but two for rodents due to the existence of rodent-specific insulin subfamilies *Ins1* and *Ins2* (Irwin 2021). To deal with large gene families, Matreex includes the option of collapsing all subfamilies, including the root node (Matreex' "Collapse All"), as manually collapsing many nodes can be tedious. Starting from the family phylogenetic profile, the user can then unfold more and more specific subfamilies, thus revealing their species distributions and gene copy number variations. In particular, unfolding a node will reveal the gene tree topology until the next duplication nodes, which define the child subfamilies. Other subtrees will remain collapsed, as their topology is redundant with the species tree. This approach is user-friendly because it begins with a highly summarised view of the family before zooming into more specific subfamilies of interest.

Two main processes increase the size of gene families in practice: gene duplications and the increase in the number of species. The latter increases both with the number of available genomes and with the progress of orthology assessment methods and resources in handling a growing number of species (*e.g.* [Kriventseva et al. 2019; Altenhoff et al. 2021; Cantalapiedra et al. 2021; Rossier et al. 2021]). Thus, the ability to control which species (or taxa) to show and which to hide is key to achieve high levels of gene family customisation and compactness. For that task, Matreex relies on the interactive species tree that is displayed orthogonally to the gene tree. When collapsing a taxon in the species tree, all corresponding gene tree nodes are also collapsed, and the phylogenetic profiles are summarised. This is done by averaging the numbers of gene copies of the species collapsed. For example, collapsing the *Euarchontoglires* node (primates and rodents) would automatically merge *Ins1* and *Ins2* and average the insulin copy numbers of primates and rodents (*e.g.* 1.3 in PANTHER v.17). Moreover, to facilitate the exploration of large species trees, Matreex provides the option to collapse every taxon after a given node depth from the root.

Finally, Matreex implements several other design features to further facilitate the user experience. First, as scientific names can be quite obscure, images are displayed when hovering over taxon labels; at present Wikipedia images are used but other sources could be easily

implemented. Second, to highlight lineage-specific expansions, the matrix of phylogenetic profiles is displayed as a heatmap for which custom colours can be used to highlight specific clades.

Availability and implementation

Matreex is available from the Python Package Index (pip install matreex) with the source code and documentation available at https://github.com/DessimozLab/matreex. Matreex is implemented in JavaScript with the D3 library and wrapped in a Python module that supports the APIs of OMA and PANTHER (Kaleb et al. 2019; Mi et al. 2021), from which only the gene family or HOG identifier is required as input. Moreover, custom gene trees in JSON are supported when associated with a consistent species tree (gene tree-species tree conflicts are not supported by Matreex). The specifics of these formats are described in the github.

Applications

In this section, we illustrate how Matreex facilitates the analysis of gene families on three different use-cases with real biological applications.

The visual opsin gene repertoire correlates with adaptations in vertebrates

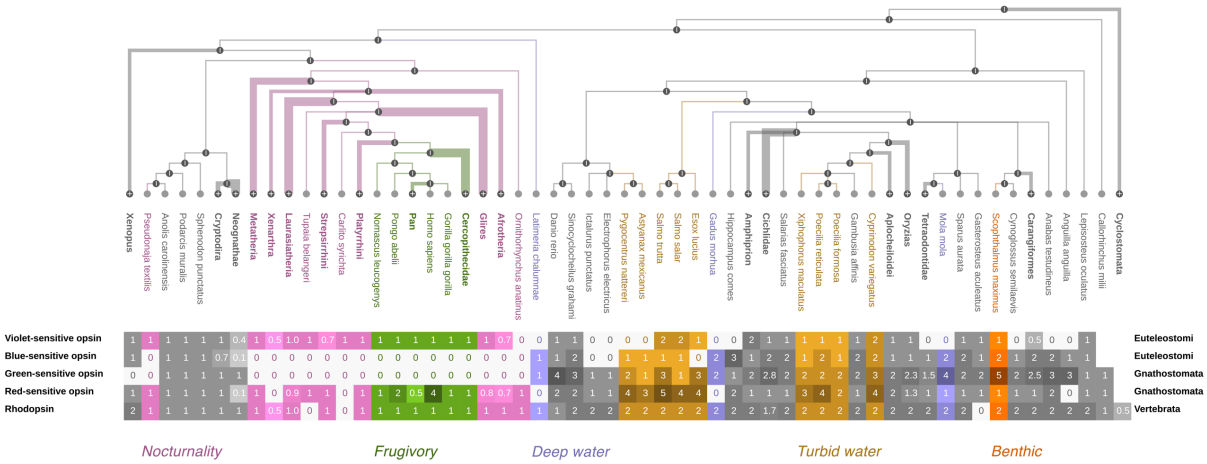


Figure 2. Visual opsin families in vertebrates (data from OMA All.Dec2021). Adaptations involved in textbook correlations with patterns of gene losses and duplications are annotated with separate colours. Nocturnality in snakes and mammals: loss of blue- and green-sensitive opsins. Frugivory in old-world primates: duplications of red-sensitive opsin. Deep-water: loss of violet- and red-sensitive opsins, duplications of blue- and green-sensitive opsins. Turbid water: duplications of red-sensitive

opsins. Benthic: duplications of green-sensitive opsins. Italic annotations do not belong to the Matreex layout but are added for figure clarity.

Matreex enables to quickly identify correlations between adaptations, or phenotypes, and variations in gene copy numbers. Indeed, the gene repertoire of a species or taxon is shown in a single column instead of being scattered across the gene tree. Thus, repeated losses or expansions associated with an adaptation are easy to identify. Moreover, Matreex further facilitates the task by depicting losses on a white background and expansions on a darker one. Clades can also be coloured by adaptation to highlight the correlation. Here, we illustrate this use-case with the textbook example of visual opsins (Graur and Li 1999; Paul G. Higgs and Teresa K. Attwood 2005). Due to its intuitive interpretation and aesthetically pleasing representation, we expect this usage to become particularly popular in outreach tasks including teaching and conference presentations.

The vertebrate ancestor had one rod opsin (Rhodopsin) for dim light vision and four cone opsins for a tetrachromatic vision, each sensitive to a specific range of light wavelength (Musilova et al. 2021). Specifically, the shortest wavelengths are absorbed by the violet-sensitive opsin (SWS1), followed by the blue- (SWS2) and green-sensitive (RH2) opsins for intermediate wavelengths. The red-sensitive (LWS) opsins absorb for the largest ones. By contrast, mammals and snakes lack the blue- and green-sensitive opsins, likely due to the nocturnal lifestyle of their ancestors (Borges et al. 2018; Katti et al. 2019). However, old-world primates (*Catarrhini*, including humans) regained a more complex colour vision by co-opting a red-sensitive opsin duplicate to absorb green wavelengths, which possibly gave primates a selective advantage for food and predator detection (Carvalho et al. 2017). Matreex shows clearly and at a glance both the losses in mammals and snakes (Fig. 2, pink), as series of white background zeroes, and the secondary amplification in *Homo sapiens* and *Pongo abelii* (Fig. 2, green), as darker cells with larger numbers of genes.

By contrast, the visual opsin repertoire of fishes is much more variable, likely due to the diversity of underwater light environments (Musilova et al. 2021) and this is immediately visible in the Matreex representation. In deep water, the light spectrum is shrinked to absorb only blue and green. Thus, deeper-living species are expected to lose red- and violet- sensitive opsins, while duplicating the green- and blue-sensitive ones to compensate for the lower photon abundance. Here, such an evolutionary pattern was detected in the cod (*Gadus morhua*, depth: 150-200m, max. 600m), the sunfish (*Mola mola*, depth 30-70m, max. 480m) and the coelacanth

(*Latimeria chalumnae*, depth: 180-250m, max. 700m) (Fig. 2, purple). Moreover, we found the most green-sensitive opsins (five) in the turbot flatfish (*Scophthalmus Maximus*), which could be an adaptation to deep benthic life (Wang et al. 2021) (Fig. 2, orange). Conversely, the light spectrum is shifted toward longer wavelengths in turbid water, thus favouring red-opsin duplications (Musilova et al. 2021). The present gene tree supports this assumption as we found the most red-opsin copies in fishes that live in the turbid freshwater and brackish habitats (Fig. 2, brown). In particular, five copies were detected in the brown trout (*Salmo trutta*) and four in the red piranha (*Pygocentrus nattereri*), the atlantic salmon (*Salmo salar*), the northern pike (*Esox lucius*), the guppy (*Poecilia reticulata*) and the pupfish (*Cyprinodon variegatus*).

Coevolution of the intraflagellar transport genes

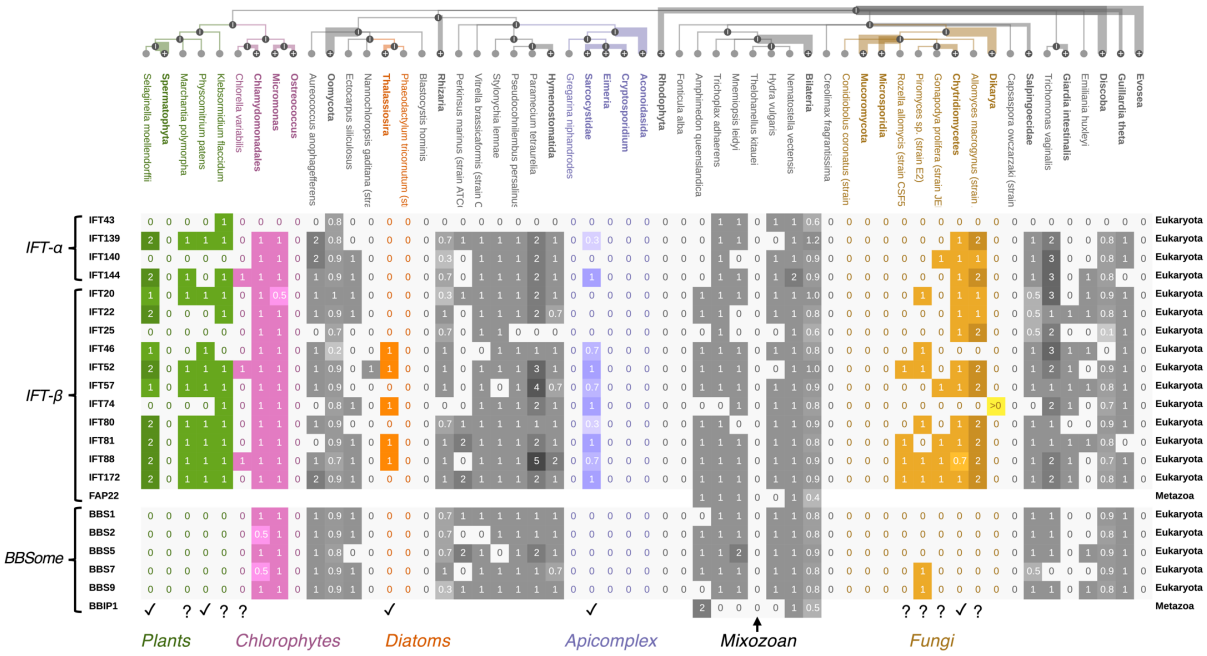


Figure 3. Intraflagellar transport gene families (data from OMA All.Dec2021). Colored clades display partial and complete IFT losses that fit the "last-in, first-out" hypothesis for gene module evolution. ✓ highlight partial IFT losses reported by van Dam et al. (2013) and **?**, the ones reported here. We also report the first evidence to our knowledge of a complete loss of IFT in the mixozoan *Thelohanellus kitauei*. Italic annotations, brackets, ✓ and **?** do not belong to the Matreex layout but are added for figure clarity.

Matreex enables to perform gene presence-absence analyses for dozens of non-homologous families spanning hundreds of species in a few minutes. Indeed, similarly to phylogenetic profile viewers, lists of gene families are valid inputs for Matreex. This is useful to visualise the result of a phylogenetic profile search (Altenhoff et al. 2021) or to study

coevolving gene families (*e.g.* involved in the same pathway). In this second application, we illustrate the latter by generalising a study on eukaryotic intraflagellar transport (IFT) genes from 622 species, compared to the 52 used originally (van Dam et al. 2013). Specifically, we used Matreex to simplify the task of contrasting our results with the literature and to propose new biological hypotheses.

Eukaryotic flagella (cilia) are involved in cell motility and sensory detection (Nevers et al. 2017). Their dysfunction is the cause of ciliopathies in humans (Badano et al. 2006). The IFT complex is essential to build and maintain the flagella. From an evolutionary perspective, IFT is a great example of the "last-in, first-out" hypothesis (van Dam et al. 2013), whereby modules added last are more dispensable and thus, lost first. Indeed, of the three IFT modules (IFT-α, IFT-β and BBSome), BBSome and IFT-α emerged from IFT-β duplications and their loss often precedes the complete loss of IFT and cilia. Thus, studying how ciliated eukaryotes cope with partial IFT loss is promising for the treatment of IFT-related human ciliopathies such as the Bardet–Biedl syndrome caused by BBSome alterations (Badano et al. 2006).

Repeated and correlated losses are visible at a glance in *Matreex* with columns of zeros on white backgrounds (Fig 3). As expected, complete loss of IFT complexes were detected in the main non-ciliated taxa (*e.g. Spermatophyta*, *Dikarya* or *Amoebozoa*). Moreover, due to the sheer number of used genomes, summarised in one easy to read figure, we were able to identify many other complete IFT losses. Although most were already established (*e.g. Fonticula alba*, *Creolimax fragrantissima*, *Capsaspora owczarzaki* (Torruella et al. 2015), *Entamoeba* (Wickstead and Gull 2007)), we report the first evidence to our knowledge of a complete loss of IFT in the mixozoan *Thelohanellus kitauei*, likely indicating the loss of the organelle in this species.

Then, we could first quickly confirm all established patterns of BBSome and IFT-α losses in species closely related to non-ciliated clades with complete IFT loss from (van Dam et al. 2013). Specifically, we detected the loss of BBSome in basal plants (*Selaginella moellendorffii* and the moss *Physcomitrella patens*) close to seed plants (*Spermatophyta*), in the apicomplexa *Sarcocystidae* (*Toxoplasma gondii* clade) close to *Aconoidasida* (*Plasmodium falciparum* clade) and in the basal fungi *Chytridiomycetes* (*Batrachochytrium dendrobatidis* clade) close to *Dikarya* and *Mucoromycota*. We also recovered the loss of BBSome and IFT-α in the diatoms *Thalassiosira* close to *Phaeodactylum tricornutum*. Secondly, we could identify other independent losses supporting the "last-in, first-out" hypothesis. In particular, we found

two losses of BBSome in the basal plants *Marchantia polymorpha* and *Klebsormidium flaccidum*. We also identified complete IFT losses in another three apicomplexa clades (*Eimeria*, *Cryptosporidium* and *Gregarina niphandrodes*) and two basal fungi clades (*Microsporidia* and *Conidiobolus coronatus*). Moreover, we found evidence for losses of BBSome and IFT-α in four basal fungi. While *Rozella allomycis* lacks all BBSome and IFT-α genes, *Piromyces sp.* and *Gonapodya prolifera* were found with merely one IFT-α and two BBSome genes, respectively. *Allomyces macrogynus* lacked BBSome. Finally, the presence of one IFT-α and two IFT-β genes in the chlorophytes *Chlorella variabilis* close to *Ostreococcus* provides a new candidate replicate for this "last-in, first-out" hypothesis. Its low number of IFT genes, which indicates dysfunctional cilia, could be due to the endosymbiont nature of *Chlorella variabilis* (Blanc et al. 2010).

When many gene families underwent duplications in the same species, the column attracts the eye as it becomes darker in Matreex. Thus, we identified four species with many duplicates of IFT-α and IFT-β genes. Although *Paramecium tetraurelia* and *Trichomonas vaginalis* have undergone whole genome duplications (Aury et al. 2006; Carlton et al. 2007), *Paramecium tetraurelia* IFT57 copies show evidence of subfunctionalization (Shi et al. 2018), while *Trichomonas vaginalis* displays specialised cilia that could have required the recruitment of additional IFT copies. Finally, to explain the retention of *S. moellendorffii* and *Allomyces macrogynus* duplicates, that have lost BBSome, we may speculate whether these extra copies could have been co-opted to replace the BBSome functions.

Origin and evolution of eukaryotic MutS genes

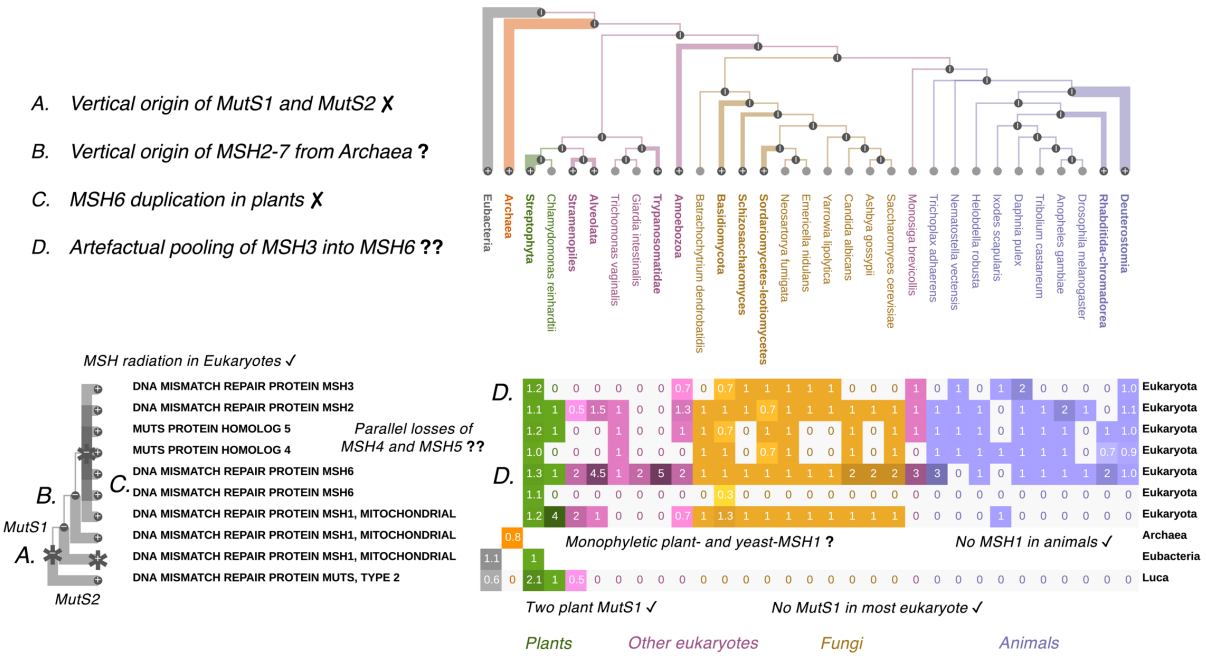


Figure 4. Detailed evolutionary analysis of the MutS family (data from PANTHER v.17). Established hypotheses on MutS evolution are annotated with a ✓ when supported by the data at hand and with a ✗ otherwise. Ongoing and new hypotheses are annotated with **?** and **??**, respectively. Italic annotations do not belong to the Matreex layout.

Matreex enables to precisely analyse the gene repertoire evolution of large multi-copy families. First, subfamily gene repertoires can be correlated among themselves or with adaptations similarly to the previous applications. Secondly, the gene tree enables to study the evolutionary relationships between phylogenetic profiles. This can be useful, for instance, to differentiate orthologous from paralogous profiles and to diagnose the underlying gene tree. In this last application, we performed a detailed analysis of the MutS family, whose evolutionary history remains largely under debate. Specifically, we used Matreex to simplify the task of systematically contrasting existing knowledge with the data at hand (Fig. 4). First, we used established hypotheses to diagnose the underlying gene tree. Secondly, we assessed which ongoing hypotheses were supported by it and, thirdly, we drew new hypotheses from visual patterns. Finally, we contextualised the results with functional and evolutionary knowledge from the literature to highlight the importance of such an approach.

MutS genes are involved in the DNA mismatch repair pathway (Liu et al. 2017; Mi et al. 2021). Although bacteria have multiple MutS genes, only MutS1 and MutS2 are found in both bacteria and eukaryotes. MutS2 is found only in photosynthetic eukaryotes and its transfer from cyanobacteria through the chloroplast endosymbiosis is well established (Lin et al. 2007).

Matreex clearly shows the absence of MutS2 in other eukaryotes and *Archaea*, as well as its two copies in plants (*Streptophyta*). However, PANTHER predicts a vertical origin of MutS2 from a pre-LUCA duplication followed by independent losses in *Archaea* and non-photosynthetic eukaryotes. Matreex shows this evolutionary trajectory with a fully connected phylogenetic profile for MutS2 and losses instead of empty cells for *Archaea* and most eukaryotes.

In contrast to MutS2, the origin of MutS1 remains under debate. Eukaryotic MutS1 genes (MSH2-7) were first thought to originate from the mitochondria endosymbiosis of an α-*Proteobacteria* (Lin et al. 2007) until the Asgard *Archaea* MutS1 was found to be more closely related to Eukaryotes than to α-*Proteobacteria* (Hofstatter and Lahr 2021). This implies the vertical origin of MSH2-7 from *Archaea*. Matreex clearly shows these orthologous relationships between archeal MutS1 and eukaryotic MSH2-7 because they form a monophyletic clade in the gene tree. Moreover, the archeal MutS1 profile does not overlap with those of MSH2-7, which indicates their orthology relationships.

MSH2-6 genes originated from duplications in the eukaryote ancestor, while MSH7 arose from a plant-specific duplication of MSH6 (Lin et al. 2007). Matreex clearly represents this radiation with a compact block of subfamily profiles, although PANTHER misclassified MSH7 as another eukaryote subfamily. This radiation presents a striking example of functional evolution through gene duplications. Indeed, while bacterial MutS genes form homodimers, the eukaryotic DNA mismatch repair pathway recruits three heterodimers that have specialised to bind specific DNA mismatches. Precisely, the mammal MSH2/6 and MSH2/3 bind up to 3 and 13 nucleotide indels, respectively (Muthye and Lavrov 2021), while the plant MSH2/7 prefers mismatches containing T and/or G/T, A/C, T/C, G/A, T/T, or A/A (Karthika et al. 2020). The functional specialisation of these MutS1 homologs, both at the subunit and heterodimer levels, probably improved the DNA mismatch repair system in eukaryotes. Moreover, MSH4 and MSH5 have lost the DNA repair function and are involved in meiotic recombination, thus presenting a striking example of neofunctionalization (Manhart and Alani 2016).

However, the origins of the plant- and yeast-MSH1 remain unclear. Although originally thought to descend from the same MutS ancestor as MSH2-7 (Lin et al. 2007), an acquisition of yeast-MSH1 in fungi along the mitochondrial endosymbiosis has been suggested (Hofstatter and Lahr 2021). Similarly, the plant-MSH1 could have been acquired from giant viruses or

along the chloroplast endosymbiosis (Wu et al. 2020; Hofstatter and Lahr 2021). The underlying gene tree supports the original hypothesis as we found the plant- and yeast-MSH1 genes in the same eukaryotic subfamily. Matreex simplified drawing this conclusion as both plant- and yeast-MSH1 belong to the same collapsed subtree and phylogenetic profile, indicating a monophyletic origin. Moreover, we recovered the absence of MSH1 in animals (with the exception of the tick), which has been recently linked with the exceptionally high evolutionary rates of their mitochondrial genes, as MSH1 is involved in repairing their sequences (Wu et al. 2020).

Matreex simplifies the identification of gene repertoire evolutionary patterns. Thus, we observed unexpected expansions of MSH6 in eukaryotes (*e.g. Alveolata*, *Trypanosomatidae*), yeast (*Saccharomyces cerevisiae*), nematodes and fruitfly (*Drosophila Melanogaster*). Then, by expanding the gene tree, we noticed many MSH3 genes misclassified as MSH6, which coincides with predicted MSH3 losses. For example, of the five *Trypanosomatidae* copies, two were surely misclassified MSH3 and MSH5 genes, and one was undefined. Thus, although the loss of MSH3 in nematodes and insects and the trypanosome-specific MSH8 subfamily are documented (Bell et al. 2004; Muthye and Lavrov 2021), we hypothesised that MSH6 is artefactually attracting other genes, in particular MSH3 ones, during phylogenetic reconstruction.

Finally, we observed repeated and correlated losses of MSH4 and MSH5 in fungi, fruit fly and other eukaryotes (annotated in pink). While losses in the latter are likely artefactual (Rzeszutek et al. 2022), *Schizosaccharomyces* and *D. Melanogaster* are known to have lost and replaced MSH4 and MSH5 for meiotic recombination (Kohl et al. 2012; Manhart and Alani 2016). Moreover, given that these two genes form an obligate complex, other correlated losses in fungi are plausible and could provide good candidates to study alternative meiotic recombination mechanisms.

Conclusion

In an era where the goal of sequencing all eukaryotic species before 2030 has been set (Lewin et al. 2022), it has become critical to develop new methods to represent this huge volume of upcoming data. Here, we introduce an innovative tool to scale the visualisation of gene families and illustrate its usefulness with three biological applications. First, using the textbook example of visual opsins, we revealed Matreex' potential to create easily interpretable

figures for outreach tasks. Secondly, by displaying 22 Intraflagellar gene families across 622 species cummulating 5'500 representatives, we showed how Matreex can be used for analyses of gene presence-absence and, notably, reported for the first time the complete loss of IFT in the mixozoan *Thelohanellus kitauei*. Finally, we demonstrated Matreex' usefulness in delving into precise evolutionary analyses of multi-copy gene families by combining the gene tree with phylogenetic profiles. Thus, we hope Matreex will become a valuable tool to gain insights into the evolution of increasingly large gene families.

References

Altenhoff AM, Train C-M, Gilbert KJ, Mediratta I, Mendes de Farias T, Moi D, Nevers Y, Radoykova H-S, Rossier V, Warwick Vesztrocy A, et al. 2021. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res.* 49:D373–D379.

Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* 444:171–178.

Badano JL, Mitsuma N, Beales PL, Katsanis N. 2006. The ciliopathies: an emerging class of human genetic disorders. *Annu. Rev. Genomics Hum. Genet.* 7:125–148.

Bell JS, Harvey TI, Sims A-M, McCulloch R. 2004. Characterization of components of the mismatch repair machinery in Trypanosoma brucei. *Mol. Microbiol.* 51:159–173.

Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S, Pangilinan J, Polle J, et al. 2010. The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell* 22:2943–2955.

Borges R, Johnson WE, O'Brien SJ, Gomes C, Heesy CP, Antunes A. 2018. Adaptive genomic evolution of opsins reveals that early mammals flourished in nocturnal environments. *BMC Genomics* 19:121.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* [Internet]. Available from: http://dx.doi.org/10.1093/molbev/msab293

Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, Wortman JR, Bidwell SL, Alsmark UCM, Besteiro S, et al. 2007. Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science* 315:207–212.

Carvalho LS, Pessoa DMA, Mountford JK, Davies WIL, Hunt DM. 2017. The Genetic and Evolutionary Drives behind Primate Color Vision. *Frontiers in Ecology and Evolution* 5:34.

Cromar GL, Zhao A, Xiong X, Swapna LS, Loughran N, Song H, Parkinson J. 2016. PhyloPro2.0: a database for the dynamic exploration of phylogenetically conserved proteins and their domain architectures across the Eukarya. *Database* [Internet] 2016. Available from: http://dx.doi.org/10.1093/database/baw013

van Dam TJP, Townsend MJ, Turk M, Schlessinger A, Sali A, Field MC, Huynen MA. 2013. Evolution of modular intraflagellar transport from a coatomer-like progenitor. *Proc. Natl. Acad. Sci. U. S. A.* 110:6943–6948.

Dunn CW, Munro C. 2016. Comparative genomics and the diversity of life. *Zool. Scr.* 45:5–13.

Fuentes D, Molina M, Chorostecki U, Capella-Gutiérrez S, Marcet-Houben M, Gabaldón T. 2021. PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. *Nucleic Acids Res.* [Internet]. Available from: http://dx.doi.org/10.1093/nar/gkab966

Graur, Li. 1999. Fundamentals of Molecular Evolution, 2nd edn Sinauer Associates. *Inc, Sunderland, Massachusetts, USA*.

Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, et al. 2016. Ensembl comparative genomics resources. *Database* [Internet] 2016. Available from: http://dx.doi.org/10.1093/database/bav096

Hofstatter PG, Lahr DJG. 2021. Complex Evolution of the Mismatch Repair System in Eukaryotes is Illuminated by Novel Archaeal Genomes. *J. Mol. Evol.* 89:12–18.

Horn T, Narov KD, Panfilio KA. 2022. Persistent parental RNAi in the beetle *Tribolium castaneum* involves maternal transmission of long double-stranded RNA. *Advanced Genetics*:2100064.

Ilnitskiy IS, Zharikova AA, Mironov AA. 2022. OrthoQuantum: visualizing evolutionary repertoire of eukaryotic proteins. *Nucleic Acids Res.* [Internet]. Available from: http://dx.doi.org/10.1093/nar/gkac385

Irwin DM. 2021. Evolution of the Insulin Gene: Changes in Gene Number, Sequence, and Processing. *Front. Endocrinol.* 12:649255.

Kaleb K, Vesztrocy AW, Altenhoff A, Dessimoz C. 2019. Expanding the Orthologous Matrix (OMA) programmatic interfaces: REST API and the OmaDB packages for R and Python. *F1000Res.* 8:42.

Karthika V, Babitha KC, Kiranmai K, Shankar AG, Vemanna RS, Udayakumar M. 2020. Involvement of DNA mismatch repair systems to create genetic diversity in plants for speed breeding programs. *Plant Physiology Reports* 25:185–199.

Katti C, Stacey-Solis M, Coronel-Rojas NA, Davies WIL. 2019. The Diversity and Adaptive Evolution of Visual Photopigments in Reptiles. *Frontiers in Ecology and Evolution* [Internet] 7. Available from: https://www.frontiersin.org/article/10.3389/fevo.2019.00352

Kohl KP, Jones CD, Sekelsky J. 2012. Evolution of an MCM complex in flies that promotes meiotic crossovers by blocking BLM helicase. *Science* 338:1363–1365.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 47:D807–D811.

Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, et al. 2022. The Earth BioGenome Project 2020: Starting the clock. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 119. Available from: http://dx.doi.org/10.1073/pnas.2115635118

Lin Z, Nei M, Ma H. 2007. The origins and early evolution of DNA mismatch repair genes--multiple horizontal gene transfers and co-evolution. *Nucleic Acids Res.* 35:7591–7603.

Liu D, Keijzers G, Rasmussen LJ. 2017. DNA mismatch repair and its many roles in eukaryotic cells. *Mutat. Res. - Rev. Mut. Res.* 773:174–187.

Manhart CM, Alani E. 2016. Roles for mismatch repair family proteins in promoting meiotic crossing over. *DNA Repair* 38:84–93.

Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, Thomas PD. 2021. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* 49:D394–D403.

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* 45:D183–D189.

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. 2020. GeneRax: A tool for species tree-aware maximum likelihood based gene family tree inference under gene duplication, transfer, and loss. *Mol. Biol. Evol.* [Internet]. Available from: http://dx.doi.org/10.1093/molbev/msaa141

Musilova Z, Salzburger W, Cortesi F. 2021. The Visual Opsin Gene Repertoires of Teleost Fishes: Evolution, Ecology, and Function. *Annu. Rev. Cell Dev. Biol.* 37:441–468.

Muthye V, Lavrov DV. 2021. Multiple Losses of MSH1, Gain of mtMutS, and Other Changes in the MutS Family of DNA Repair Proteins in Animals. *Genome Biol. Evol.* [Internet] 13. Available from: http://dx.doi.org/10.1093/gbe/evab191

Nevers Y, Prasad MK, Poidevin L, Chennen K, Allot A, Kress A, Ripp R, Thompson JD, Dollfus H, Poch O, et al. 2017. Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling. *Mol. Biol. Evol.* 34:2016–2034.

Nguyen NTT, Vincens P, Roest Crollius H, Louis A. 2018. Genomicus 2018: karyotype evolutionary trees and on-the-fly synteny computing. *Nucleic Acids Res.* 46:D816–D822.

Paul G. Higgs and Teresa K. Attwood. 2005. Bioinformatics and Molecular Evolution. BLACKWELL PUBLISHING

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96:4285–4288.

Robinson O, Dylus D, Dessimoz C. 2016. Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web. *Mol. Biol. Evol.* 33:2163–2166.

Rossier V, Vesztrocy AW, Robinson-Rechavi M, Dessimoz C. 2021. OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches. *Bioinformatics* [Internet]. Available from: http://dx.doi.org/10.1093/bioinformatics/btab219

Rzeszutek I, Swart EC, Pabian-Jewuła S, Russo A, Nowacki M. 2022. Early developmental, meiosis-specific proteins - Spo11, Msh4-1, and Msh5 - Affect subsequent genome reorganization in Paramecium tetraurelia. *Biochim. Biophys. Acta Mol. Cell Res.* 1869:119239.

Sadreyev IR, Ji F, Cohen E, Ruvkun G, Tabach Y. 2015. PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. *Nucleic Acids Res.* 43:W154–W159.

Shi L, Koll F, Arnaiz O, Cohen J. 2018. The Ciliary Protein IFT57 in the Macronucleus of Paramecium. *J. Eukaryot. Microbiol.* 65:12–27.

Torruella G, de Mendoza A, Grau-Bové X, Antó M, Chaplin MA, del Campo J, Eme L, Pérez-Cordón G, Whipps CM, Nichols KM, et al. 2015. Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Curr. Biol.* 25:2404–2410.

Tran N-V, Greshake Tzovaras B, Ebersberger I. 2018. PhyloProfile: dynamic visualization and exploration of multi-layered phylogenetic profiles. *Bioinformatics* 34:3041–3043.

Tremblay BJ-M, Lobb B, Doxey AC. 2021. PhyloCorrelate: inferring bacterial gene-gene functional associations through large-scale phylogenetic profiling. *Bioinformatics* [Internet]. Available from: http://dx.doi.org/10.1093/bioinformatics/btaa1105

Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, Hinrichs AS, Fernandes JD, Borges R, Slodkowicz G, et al. 2020. Stability of SARS-CoV-2 phylogenies. *PLoS Genet.* 16:e1009175.

Wang Y, Zhou L, Wu L, Song C, Ma X, Xu S, Du T, Li X, Li J. 2021. Evolutionary ecology of the visual opsin gene sequence and its expression in turbot (Scophthalmus maximus). *BMC Ecol Evol* 21:114.

Wickstead B, Gull K. 2007. Dyneins across eukaryotes: a comparative genomic analysis. *Traffic* 8:1708–1721.

Wu Z, Waneka G, Broz AK, King CR, Sloan DB. 2020. MSH1 is required for maintenance of the low mutation rates in plant mitochondrial and plastid genomes. *Proc. Natl. Acad. Sci. U. S. A.* 117:16448–16455.

Supplementary material

Matreex files of figures and code to reproduce them:

https://github.com/DessimozLab/matreex/tree/main/paper_figures

Supp. Figure 1:

https://github.com/DessimozLab/matreex/blob/main/paper_figures/SFig1.png

Chapter 4

Characterising the role of gene family expansions in animal venom evolution

Characterising the role of gene family expansions in animal venom evolution

Victor Rossier, Giulia Zancolli, Christophe Dessimoz, Marc Robinson Rechavi

Introduction

Convergent evolution, or the independent emergence of similar traits, is pervasive across the tree of life (Losos 2011). This implies a preeminent role of deterministic forces like natural selection or genetic constraints in evolution (Stern 2013; Sackton and Clark 2019). However, the frequent occurrence of evolutionary "one offs" also implies the opposite: a preeminent role of contingency and chance (Blount, Lenski, and Losos 2018). Moreover, the number of lineages that failed to evolve an adaptive trait and thus went extinct is unknown. This lack of denominator limits the ability to weigh the role of determinism and chance in evolution (Blount, Lenski, and Losos 2018). Understanding this problem would benefit from studying the molecular basis of convergent traits (Rosenblum, Parent, and Brandt 2014; Storz 2016). Indeed, due to the many-to-one mapping of genotypes to phenotypes, which gives multiple genetic options to evolve a phenotype, findings of molecular convergence support evolutionary predictability. Moreover, as the hierarchical levels of genetic similarity (residue, gene, pathway, etc.) are also affected by this many-to-one causality relationship, convergent changes at more specific levels also suggest fewer evolutionary paths (Losos 2011; Rosenblum, Parent, and Brandt 2014). For example, the sodium pump α-subunit associated with cardenolides resistance in insects repeatedly evolved the same substitutions (Sackton and Clark 2019; Dobler et al. 2012). This suggests that insects would repeat that phenotypic and genetic path if life's tape would replay (Gould 1990).

The increasing number of available genomes sequenced in the context of ambitious initiatives (i5K Consortium 2013; Koepfli et al. 2015) provides new opportunities to investigate the extent of molecular convergence. Indeed, the increasing number of phenotypically convergent replicates should provide larger statistical power to uncover and generalise the molecular underpinning of these traits (Sackton and Clark 2019). Moreover, densely sampled clades should enable better identification of ancestral genetic changes and mitigate methodological noise. In particular, large-scale analyses of convergent gene repertoire evolution become possible due to recent developments of reference-based orthology inference (Nagy et al. 2020; Rossier et al. 2021). By contrast, further algorithmic development to align

sequences, infer trees and test for selection are required to enable scalable scans of convergent substitutions. Similarly, large-scale comparative transcriptomics studies remain limited by the availability of expression data (Bastian et al. 2021).

However, the increasing diversity of sequenced organisms, sequencing technologies and analysis methods produces protein-coding gene repertoires (*i.e.* proteomes) of heterogeneous quality (Alkan, Sajjadian, and Eichler 2011; Feron and Waterhouse 2022). Although randomly distributed low-quality proteomes should not bias downstream analyses, examples of systematic biases of proteome quality have been reported. For example, 15% of genes were found missing in birds due to their high GC-content, which complicated genome assembly and annotation (Botero-Castro et al. 2017). Moreover, heterogeneity of genome annotation pipelines has been shown to have the potential to multiply by 15 the true number of lineage specific genes (Weisman, Murray, and Eddy 2022). Thus, extensive and fast proteome quality controls are required to integrate many proteomes of heterogeneous quality in comparative genomic pipelines.

Venom evolved independently in more than 100 animal clades comprising more than 200'000 species, making it one of the most convergent animal traits (Schendel et al. 2019; Zancolli and Casewell 2020). Thus, studying its molecular underpinning would especially benefit from large-scale genomic data. Venoms are cocktails of toxin proteins used mainly for predation and defence (Schendel et al. 2019). The classical model of toxin evolution starts with the recruitment of toxin genes from families with suitable biochemical features like the ability to be secreted (*i.e.* toxipotent families) (Fry et al. 2009; Barua, Koludarov, and Mikheyev 2021). Then, gene duplications are predicted to be selected first to increase venom production and second to complexify this cocktail through subfunctionalization (Kordis and Gubensek 2000; Chang and Duda 2012). Indeed, this fits the logic of an arms race with prey (venom for predation) or predators (venom for defence). Moreover, toxins have been convergently recruited from a limited number of gene families, thus suggesting molecular constraints (Fry et al. 2009; Zancolli and Casewell 2020). For example, Kallikreins have experienced convergent duplications in snakes, shrews and solenodons (Casewell et al. 2019; Barua, Koludarov, and Mikheyev 2021).

In addition to the co-option of physiological proteins into toxins, evolving the ability to deliver these toxins requires auto-resistance mechanisms, venom glands and delivery structures (Zancolli and Casewell 2020). In snakes, protein folding and modification functions were

found overrepresented in a gene regulatory network of over 3000 genes coexpressed with toxin genes (Barua and Mikheyev 2021). Recently, transcriptomes of non-homologous venom glands were found to display convergent expression profiles across eight independent animal venomous clades (Zancolli et al. 2022). Gene clusters coexpressed in venom glands were enriched in endoplasmic reticulum stress and unfolded protein response pathways. However, this convergent regulatory network rewiring of existing genes is probably not the only mechanism underlying this highly adaptive trait. For example, gene duplications can provide the potential to evolve new functions as one duplicate is expected to undergo less selective constraints (Sémon and Wolfe 2007; Conant and Wolfe 2008; Kuzmin, Taylor, and Boone 2021).

In this study, we attempt to characterise the role of gene repertoire evolution in convergent venom evolution using a dataset of 68 venomous animals and sister lineages spanning eight independent venom emergence events (Fig. 1, left). After fast reference-based orthology assignments and quality control of conserved protein sets, we specifically asked whether there were convergent gene family expansions in multiple venomous lineages. Although we found only limited evidence for molecular convergence at the level of gene family repertoires, convergently expanded families were notably enriched in functions related to secretion (exocytosis, Golgi to plasma membrane transport). On the other hand, families with toxin representatives or from pathways convergently expressed did not show convergent expansions. Finally, we observed an unexpectedly high number of convergent gene family contractions as a consequence of the mostly carnivorous diet of venomous species.

Results

Orthology assignments for proteomes of heterogeneous quality

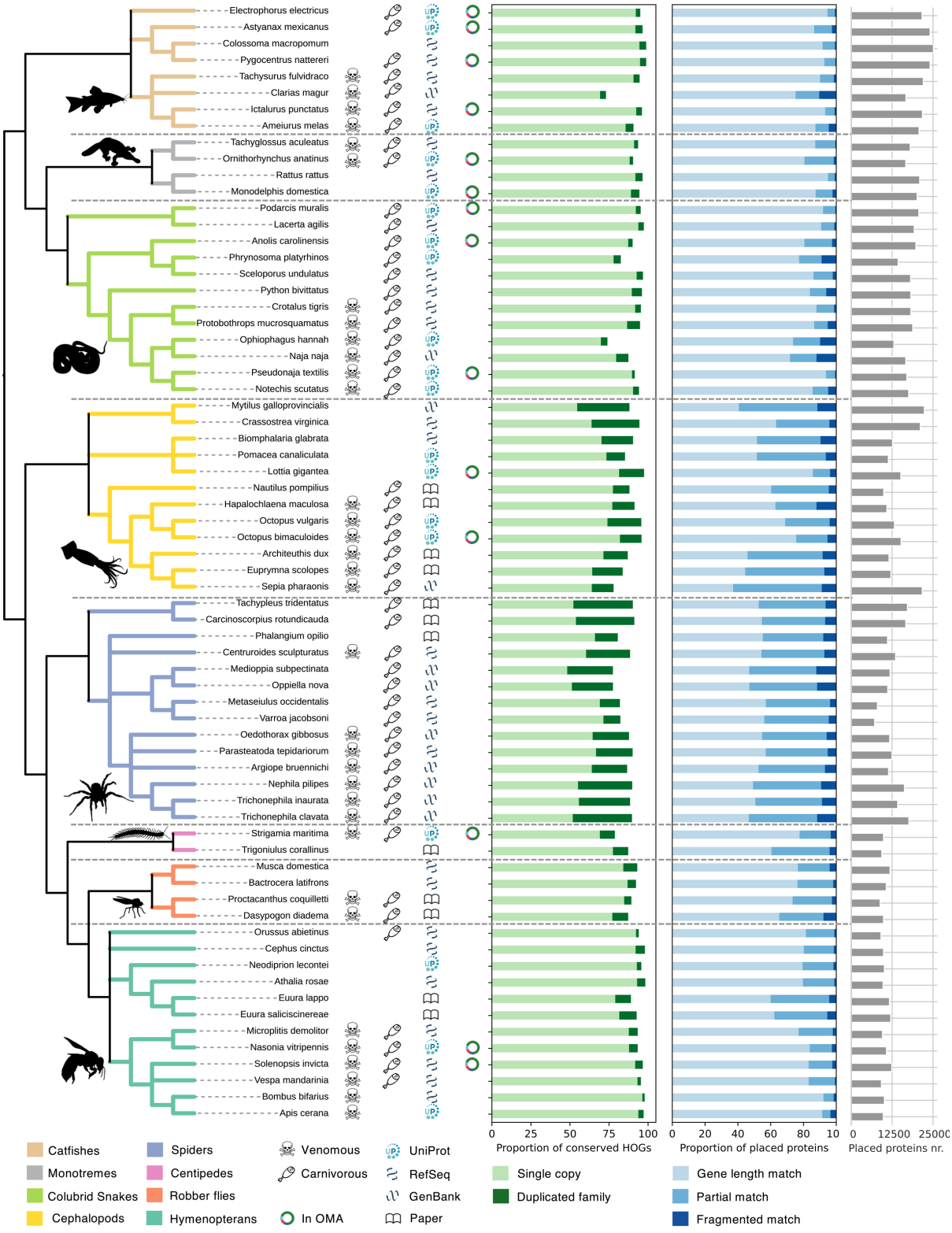


Figure 1. Taxonomic distribution, source and quality of data. Tree representation of the NCBI taxonomy of the selected species annotated by venomous clade (left). Venomous species are annotated with a skull and carnivorous species with a ham. Quality measures of conserved protein-coding gene repertoires computed with OMArk (right).

To integrate many protein-coding gene repertoires (proteomes) of heterogeneous quality in our comparative genomics analyses, we relied on the alignment-free orthology assignment method OMAmer (Rossier et al. 2021) and on extensive quality controls. Besides being fast, sequence assignment preserves the delineation of reference orthogroups, which enables the integration of low-quality proteomes without artefactually splitting or merging these groups. It is important to note that most venomous clades have at least one species which is already in the reference orthogroups (OMA symbol in Figure 1), limiting bias against the detection of venomous-specific in paralogs. Then, to control for potential biases coming from heterogeneous proteome quality without relying on drastic data filtering, we aimed to balance the proteome quality between venomous and outgroup species within each sampled clade. First, we minimised the difference between the sources of these paired sets of proteomes during the species selection procedure (Fig 1. left). Then, we measured proteome quality with OMArk to filter the proteomes causing imbalances (Nevers et al. *in prep*). Briefly, OMArk digests the orthology assignments of OMAmer to measure the proteome completeness and its proportion of wrong gene models.

The resulting dataset consisted of 68 venomous and closely related non-venomous species spanning eight bilaterian clades with independent venom emergence (Fig 1. left). Although the quality of these proteomes was highly variable between clades, the within-clade variation was much lower (Fig 1. right). For example, spiders displayed a high proportion of partial and fragmented matches, which could be explained by an overprediction of gene models and fragmented assemblies, respectively. But, because their closest non-venomous outgroups (mostly mites) displayed a similar pattern, the inference of spider gene family expansions and contractions should not be biased by proteome quality. Moreover, the limited quality imbalances still observed within clades between venomous and outgroup species were mostly balanced across clades (Supp. Fig. 1).

Do convergent gene family expansions, contractions or losses correlate with venom evolution?

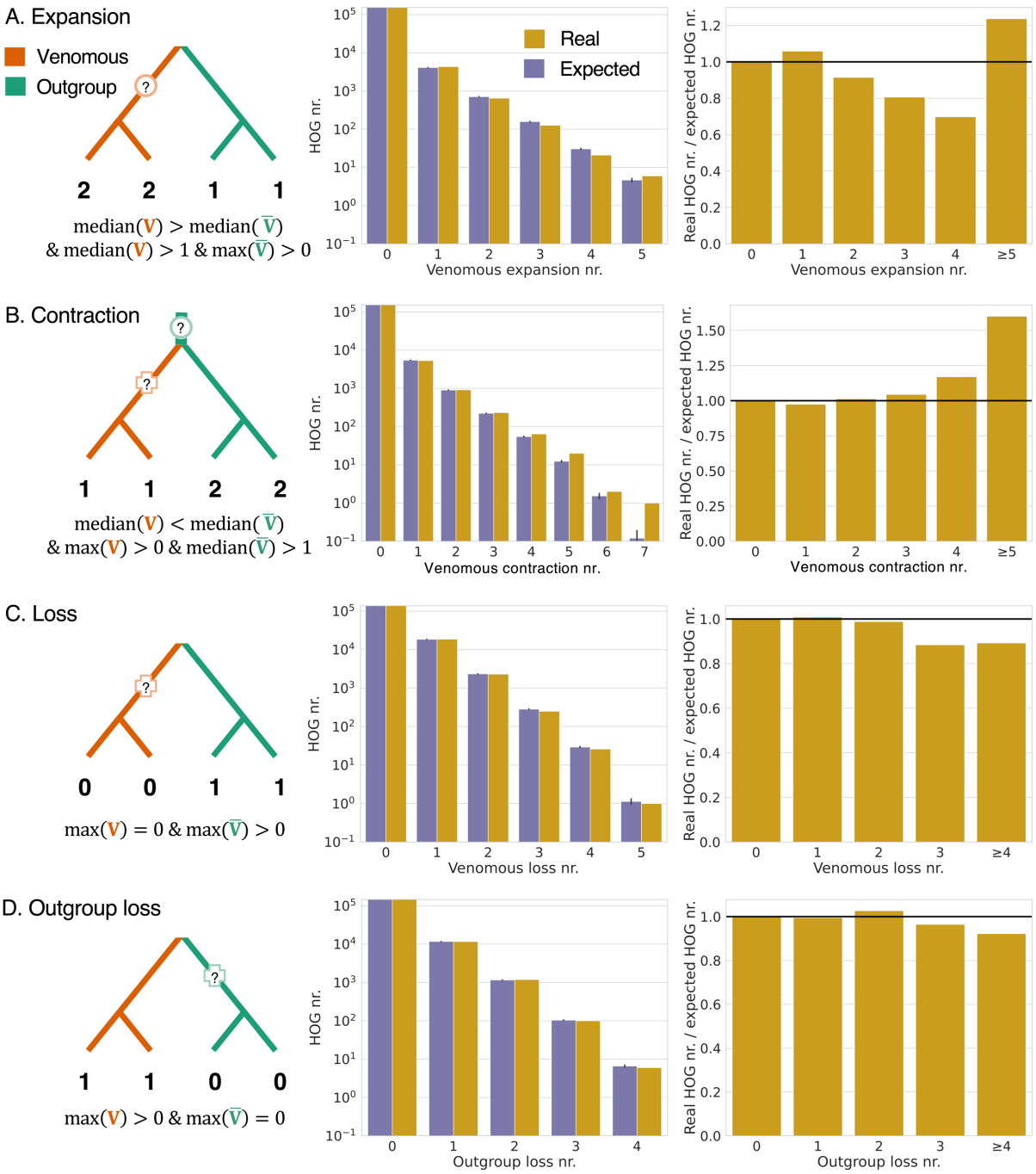


Figure 2. Classification of evolutionary events from gene family repertoires (left, middle) and quantification of these events (right). (Left) Bold numbers indicate extant copy numbers. The Hypothetical duplications and losses explaining the number of extant copies under the classified evolutionary event are annotated with circles and crosses, respectively. Formulae describes the classification conditions. (Middle) Yellow bars display the distribution of gene family expansions or contractions (convergent when >1). Violet bars display the expected distribution in absence of convergent phenotypes. (Right) Bars show the degree of enrichment (real over expected) for each number of events per family. The horizontal black line highlight enrichments (when bar is above) and depletions (when bar is below) of expansion and contraction numbers.

To characterise the role of gene family expansions, contractions, or losses in venom convergent evolution, we investigated whether each of these molecular mechanisms occurred in a non-random subset of gene families. To this end, we first classified each gene family in each venomous clade according to its evolution, based on gene copy numbers, into expansions, contractions, losses or outgroup losses (Fig. 2, left). Note that contractions cannot be distinguished from outgroup expansions with this approach. However, when gene repertoires are repeatedly smaller in venomous lineages than in outgroups, we expected contractions to be the more likely event as outgroups *a priori* do not share a common trait. Next, we modelled the expected number of families with a convergent event using a binomial model. Indeed, each expansion can be thought of as drawing a head among a number of coin tosses equal to the combined number of expansions in venomous and outgroups (*i.e.* contractions).

We found six out of 158'189 bilaterian families with convergent venomous expansions across five clades, 1.2 times more than expected (Fig 2, right). However, this result was not robust to jackknife resampling of clades (Supp. Fig. 2). Moreover, visualising the gene repertoire of these families revealed only small differences in copy numbers between venomous and outgroup species (Supp. Fig. 3). Finally, we found no function in the human representatives of these convergently expanded families that could be associated with venom (*e.g.* endoplasmic reticulum stress or unfolded protein response (Zancolli et al. 2022), Supp. Table 1). Then, we tested whether the 802 families with convergent expansions in at least two clades were involved in a non-random subset of pathways (Supp. Table 2). We found five out of ten significant GO terms with possible association with venom: exocytosis, tissue homeostasis, Golgi to plasma membrane transport, negative regulation of cytosolic calcium ion concentration and cellular response to mechanical stimulus.

Next, as we hypothesised that large gene family expansions should provide orthogonal evidence to the number of convergence occurrences for genotype-phenotype associations, we searched for families with unexpectedly large and convergent venomous expansions. The sum of venomous copy number medians of each clade (*i.e.* the number of ancestral venomous copies) was used to measure the size of expansions. We found 10 families cumulating 80 or more ancestral venomous copies, 1.6 times more than expected (Supp. Fig. 4). Of which, four also experienced convergent expansions and were involved in transcription regulation (Supp. Table 3). However, these four convergent and largely expanded families displayed signals of methodological artefacts as their species distributions were sparse and cephalopods displayed

over 200 copies per family (Supp. Fig. 5). This could reflect overly permissive orthology assignments or, more generally, a limitation of the heuristic nature of our orthology approach.

In contrast to convergent expansions, we observed a strong enrichment of convergent contractions (Fig. 2, right), which passed the jackknife resampling test (Supp. Fig. 2). The 87 families with at least four convergent contractions were enriched notably in metabolic functions: retinol metabolic process, lipid storage, digestion and L-ascorbic acid biosynthetic process (Supp. Table 4). Similarly the 1'232 families with convergent contractions (at least 2 clades) were enriched in lipid metabolic process, organic substance catabolic process, long-chain fatty acid metabolic process, glutathione metabolic process and response to nutrients (Supp. Table 5).

Are convergent contractions a genomic consequence of carnivory versus herbivory?

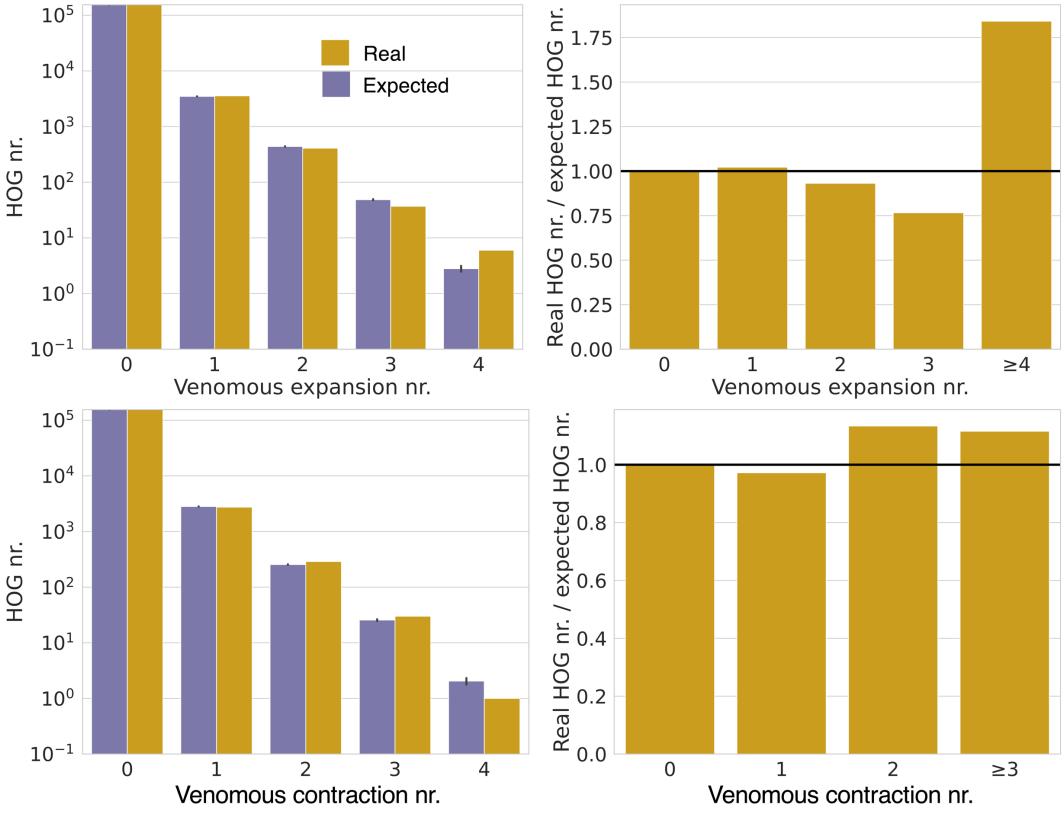


Figure 3. Quantification of expansions and contractions using the carnivorous species subset. (Left) Yellow bars display the distribution of gene family expansions or contractions (convergent when >1). Violet bars display the expected distribution in absence of convergent phenotypes. (Right) Bars show the degree of enrichment (real over expected) for each number of events per family. The horizontal black line highlight enrichments (when bar is above) and depletions (when bar is below) of events. The

horizontal black line highlight enrichments (when bar is above) and depletions (when bar is below) of expansion and contraction numbers.

Venomous species are more often carnivorous than outgroup species in our dataset (Fig. 1). Thus, to test whether the strong enrichment in convergent contractions is a consequence of carnivory in venomous lineages versus herbivory in outgroups, we repeated the analysis without herbivorous species. After balancing the number of venomous and outgroup species per clade, the resulting dataset consisted in five clades and 34 species (Supp. Dataset 1). We found almost no more enrichment of convergent contractions, and families with convergent contractions were not specifically enriched in metabolism processes (Supp. Table 6). Moreover, we found six families with four out of five convergent expansions, 1.8 times more than expected (Fig. 3). Of these, four did not overlap with previously identified families with five convergent expansions (Supp. Table 1).

Do toxins undergo convergent expansions after their recruitment?

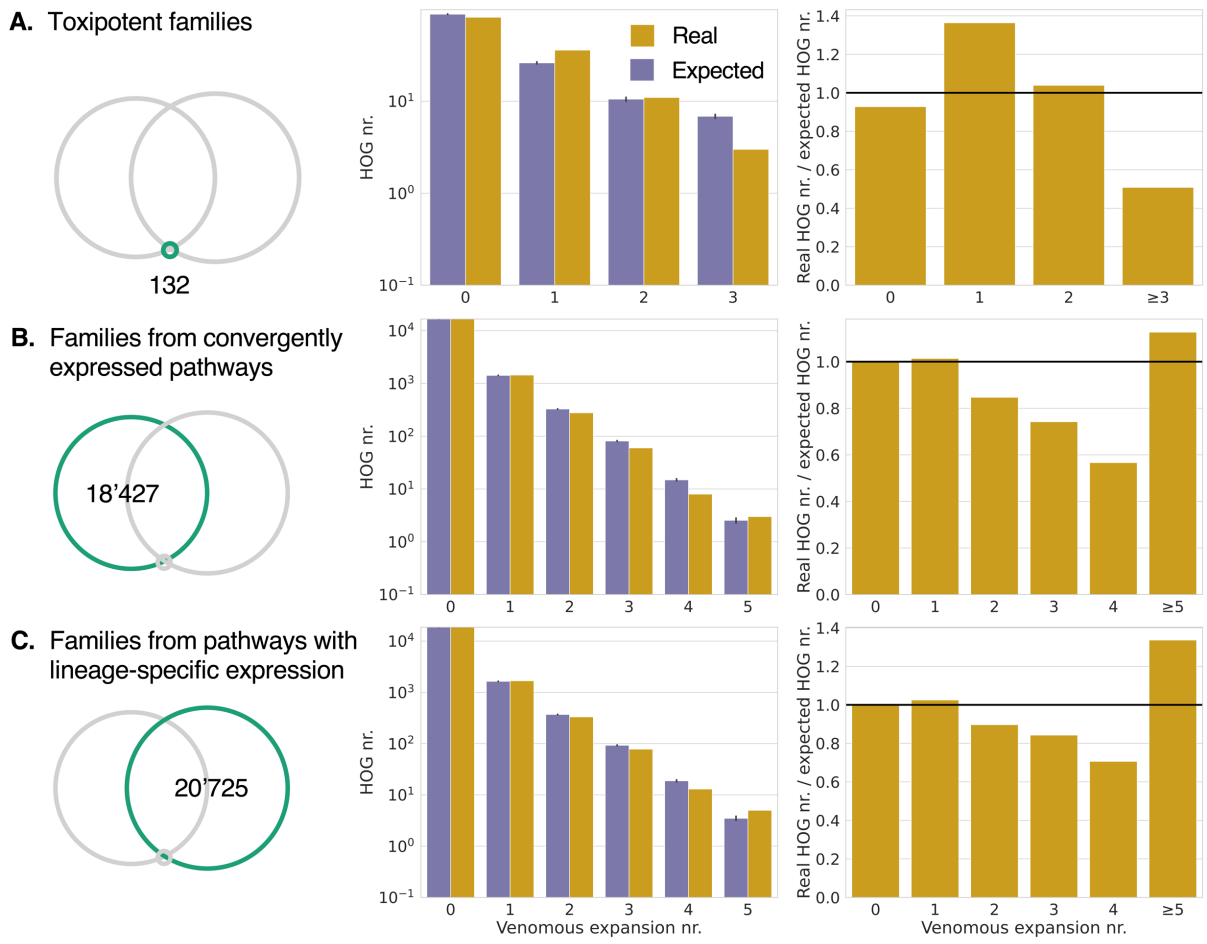


Fig. 4. Quantification of expansions on three subsets of gene families. (Left) Each subset is represented by the same circle in panels A, B and C. In each panel, the focal subset used for the analysis on the right distributions is highlighted in green. Circle and overlap sizes are proportional to the real sizes of these subsets and of their overlaps. For example, around half of the B subset is included in the C subset. All subsets are included in the 158'189 bilaterian families. (Middle) Yellow bars display the distribution of gene family expansions (convergent when >1). Violet bars display the expected distribution in absence of convergent phenotypes. (Right) Bars show the degree of enrichment (real over expected) for each number of events per family. The horizontal black line highlight enrichments (when bar is above) and depletions (when bar is below) of events. The horizontal black line highlight enrichments (when bar is above) and depletions (when bar is below) of expansion numbers.

As toxins are predicted to duplicate after their recruitment from a limited set of toxipotent families (Fry et al. 2009), we hypothesised that the latter should display convergent expansions in venomous lineages more often than by chance. Moreover, we expected toxipotent families with convergent expansions to be already described in the literature and, thus, use this analysis as a sanity check for our approach. Thus, we repeated the convergence analysis on 132 toxipotent families, which were defined with at least one match from a known toxin sequence. We found a depletion of convergent expansions in this subset, with only three families with three convergent expansions (Fig. 4, A). However, we identified spiders, robber flies and hymenopteran expansions in the phospholipase type III subfamily, which was previously described to have undergone convergent expansions after toxin recruitment (Fry et al. 2009). Moreover, the low variance of copy numbers within each venomous and outgroup lineage strengthens the evidence for ancestral duplications rather than methodological artefacts (Supp. Fig. 7).

Similarly to the full set analysis, we observed a strong enrichment of convergent contractions (Supp. Fig. 6). The most convergently evolving family was the type-B carboxylesterase / lipase family with six contractions.

Do families from convergently expressed pathways undergo convergent expansions?

Besides toxins, many other genes are expected to be involved in the ability to secrete venom (Zancolli and Casewell 2020). Thus, we repeated the analysis with a subset of gene families involved in convergently expressed pathways from (Zancolli et al. 2022). Indeed, as they are more likely to underlie venom evolution than random families, we hypothesised that

they should display molecular convergence also at other genetic levels. However, these 18'427 families displayed almost no enrichment of convergent gene family expansions (Fig. 4, B). By contrast, the 20'725 families involved in pathways with lineage-specific expression levels from (Zancolli et al. 2022) displayed a stronger enrichment (Fig. 4, C). In particular, the six families with five venomous expansions (Supp. Table 1.) belonged to that subset. This is surprising as these families were not identified to underlie convergent venom gland evolution.

Discussion

Modest role of convergent expansions in venom convergence

In this study, we aimed to characterise the role of convergent gene family expansions in venom evolution. To this end, we used a combination of data- and hypothesis-driven approaches to ask whether a non-random subset of families underwent convergent expansions in venomous lineages. We found limited evidence for molecular convergence at the level of gene family repertoire for three reasons. First, we identified an underwhelming number of bilaterian families with many convergent expansions: six families with five expansions out of eight venomous clades. Although this is slightly more than expected by chance, this is not supported statistically. Moreover, the signal was even weaker for families *a priori* associated with venom (convergently expressed in venom glands or toxipotent families). Second, the functions of these six families were not related to known adaptations required to produce venom and the copy number difference between venomous and outgroup species was generally small. Third, occurrences of convergent expansions in families with a large difference between venomous and outgroup copy numbers were found to be likely artefactual.

On the other hand, expansions spanning fewer clades (two to five) were enriched in several pathways likely related with venom evolution. Notably, two were related to secretion (exocytosis, golgi to plasma membrane) and one to the negative regulation of cytosolic calcium ion concentration, which might be linked to the signalling of venom gland activation (Luna et al. 2009). Moreover, the cellular response to mechanical stimulus could be related to the muscle contraction around the venom gland lumen, which is used to squeeze out the venom and stimulate venom replenishment. This suggests that convergent expansions in a few key pathways have played a role in venom evolution. Moreover, when correcting for the effect of carnivory versus herbivory, we observed a larger enrichment of convergent expansions (Fig. 3).

No evidence for convergence in venom-associated families

Although we recovered three clear convergent expansions in the phospholipase type III subfamily, we observed a drastic depletion of convergent expansions in toxipotent families. Thus, this supports a more modest role of convergent expansions than expected in these families similarly to the co-option of single-copy families observed in toxin evolution of parasitoid wasps (Martinson et al. 2017). Unexpectedly, we did not observe convergent expansions in Kallikreins as observed in previous studies (Casewell et al. 2019; Barua, Koludarov, and Mikheyev 2021). One possible explanation is that in our study, we controlled for local duplication rates by using closest outgroups and Kallikreins showed larger expansions in non-venomous mammals than in venomous ones (Barua, Koludarov, and Mikheyev 2021), which likely explain the lack of signal. However, this result should be taken with caution given that toxin sequences are small and fast evolving, complicating their orthology assignments (Zancolli and Casewell 2020). Thus, the depletion of expansions and enrichment of contraction could be due to the difficulty to assign toxins to toxipotent families. Moreover, families from pathways found convergently expressed in (Zancolli et al. 2022) displayed a lower enrichment of convergent expansions than the ones involved in pathways with lineage-specific expression levels. This suggests that convergent expansion is an orthogonal mechanism to convergent rewiring of gene regulatory networks.

Convergent contractions as a genomic consequence of diet

Although there are evidence of adaptive gene losses, most losses result from neutral evolution (Albalat and Cañestro 2016). Thus, the strong enrichment of convergent contractions involving hundreds of gene families that we observed is unlikely to be a consequence of venom evolution. These convergently contracted families were enriched in metabolism functions and, notably, the Type-B carboxylesterase / lipase family experienced the most contractions among toxipotent families. This family of detoxification enzymes was previously identified to have experienced large expansions in herbivorous beetles (Seppey et al. 2019). Thus, because venom is often used for predation, we hypothesised that we captured the genomic consequences of carnivory versus herbivory rather than the ability to secrete venoms. Indeed, when restricting the analysis to carnivorous species, we observed almost no more enrichment of convergent contractions (Fig. 3).

Specifically, because contractions were defined solely using the copy number difference between venomous and outgroup species (Fig. 2, left), they could be expansions in outgroups driven by the evolution of herbivory. Indeed, new detoxification and digestive enzymes are constantly required in the context of the arms race with plants and extensive gene family expansions have already been associated with herbivory (Seppey et al. 2019; Breeschoten et al. 2022). Nonetheless, whether these families experienced convergent venomous contractions or outgroup expansions, they are most likely associated with diet and thus not directly connected to venom evolution.

This unexpectedly high number of convergent outgroup expansions associated with herbivory had collateral consequences on the expected distribution of convergent venomous expansions. Because it inflated the number of trials in our binomial model (expansions plus contractions), the expected number of convergent venomous expansions was likely overestimated. Indeed, when restricting the analysis to carnivorous species, we observed a larger enrichment of convergent expansions (Fig. 3). This highlights the need for careful upstream species selection to avoid mixture effects from correlated phenotypes. On the other hand, we were able to identify this bias and show that convergent gene family evolution underlies both diet and venom evolution, which should of course be tested separately in follow-up studies.

Methods

Clade and species selection

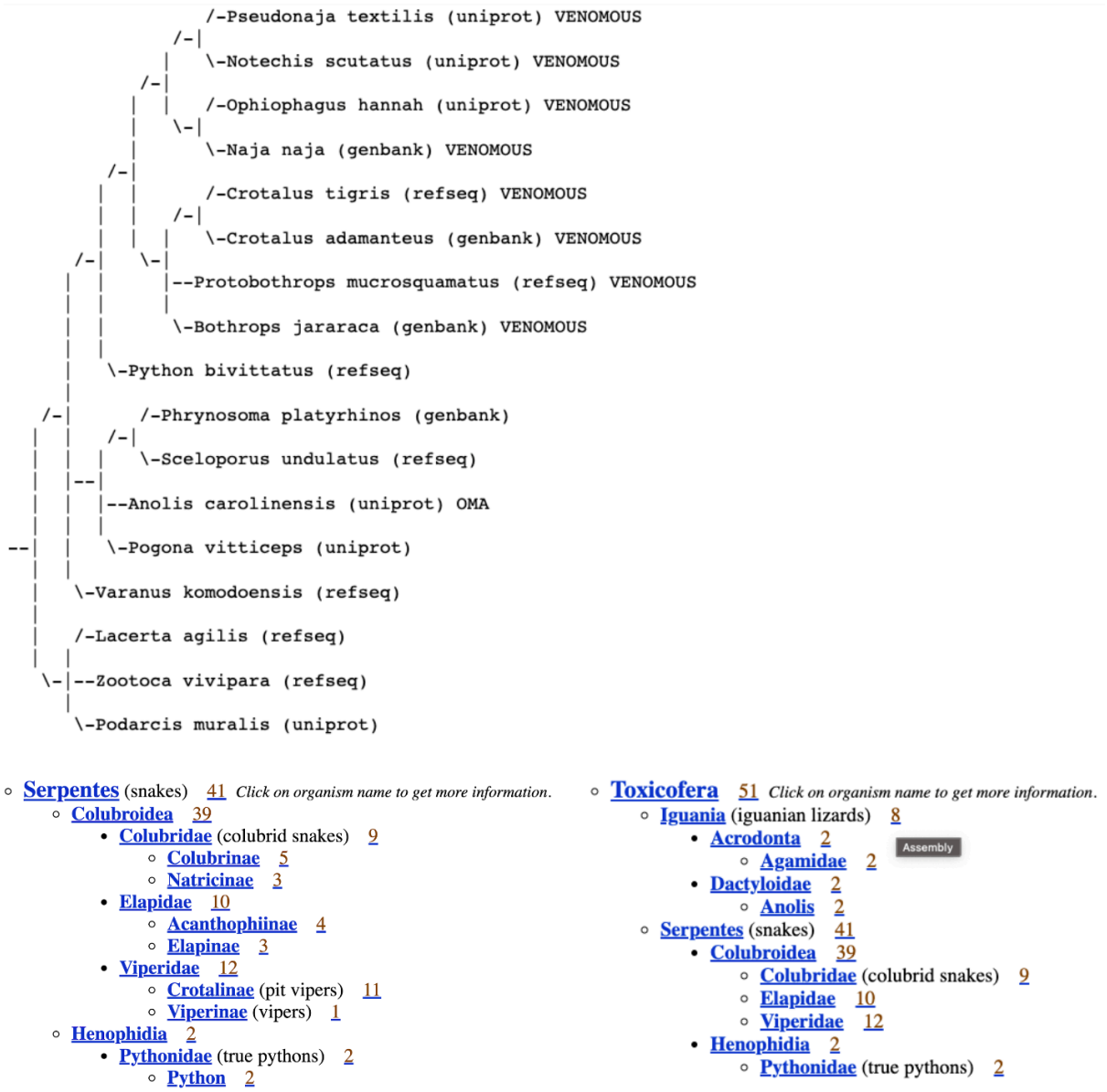


Figure 5. Semi-automated species selection procedures for venomous snakes (*Colubroidea*). Top: Annotated tree used to guide the selection procedure. In that example, the notoriously venomous Komodo varan and *Pogona vitticeps*, which display primitive venom glands, were not selected. Bottom: NCBI taxonomy of increasingly distant outgroups (*Henophidia* and *Iguania*) used to find proteomes absent from NCBI but available from the original genome papers (brown numbers indicate the number of assemblies).

We selected nine clades with specialised venomous glands, including the eight from Zancolli et al. (2022) (Supp. Dataset 1). Indeed, in contrast to less complex venom systems (secretory cells or tissues), venom glands are expected to have larger genomic consequences related to their high secretory loads and development (Schendel et al. 2019). Each sampled venomous clade was paired with close non-venomous outgroups to enable the estimations of clade-specific variations of gene repertoires. Then, to limit biases in these estimations, an equal

number of venomous and outgroup species were selected in each clade. For example, as the probability to detect a gene increases with the number of proteomes (*i.e.* species protein-coding gene repertoire), using systematically more outgroups would inflate the number of venomous gene losses. Moreover, to minimise methodological noise, we selected the maximum number of venomous species with an available proteome per clade. This should also improve the identification of shared variations of gene repertoire (derived from ancestral duplications and losses). Then, we prioritised the selection of outgroup species according to the following criteria. First, the more closely related species were selected as they are best to identify truly venomous-specific gene repertoire variations. Then, at equal evolutionary distances, we aimed to reduce potential proteome quality biases *a priori*. Thus, we attempted to balance the number of reference proteomes and sources (UniProt, RefSeq, GenBank, paper) between venomous and outgroup species for each clade. Indeed, we expected proteomes from reference species and from databases with higher quality standards to yield better orthology assignments. Finally, all else being equal, we maximised phylogenetic diversity to better disentangle ancestral events from lineage-specific ones.

To speed-up the species selection and limit subjective decisions, we used a semi-automated procedure. First, candidate venomous and outgroup species with available proteomes (2021.12) in UniProt, RefSeq and GenBank were identified based on the NCBI taxonomy using the ete3 toolkit (Huerta-Cepas, Serra, and Bork 2016) and the ncbi-genome-download python module. To guide the selection, these species were displayed in the context of a phylogenetic tree and annotated with their venomous status (venomous or outgroup), their reference status (in OMA or not) and their source database (Fig. 5, top). Then, we manually confirmed the venomous and non-venomous status of candidate venomous and outgroup species, respectively, using the species Wikipedia page, a google search (*e.g.* "*Naja naja* + venom") and the genome paper. When the number of venomous species was lower than five, we manually searched for proteomes of NCBI genomes from the original publication. Indeed, NCBI does not systematically provide the proteome. Finally, to maximise the number of closely related outgroup species, we exploited the same approach iteratively on increasingly more distant outgroup clades (Fig. 5, bottom).

Isoform filtering

Unnoticed isoforms are interpreted as paralogs by most orthology inference methods. Thus, we handled isoforms in two steps, starting by using existing isoform annotation when

available. UniProt proteomes were downloaded with one protein sequence per gene. The longest protein sequence of each gene was selected for NCBI (RefSeq and GenBank) proteomes using the "feature_table.txt" annotation file. For proteomes found in genome papers, selecting the longest isoform per gene was attempted by parsing FASTA headers. Secondly, 100% identical protein sequences were clustered with CD-HIT 4.8.1 and longest representatives were retained.

Reference-based orthology assignments

Orthogroups (*i.e.* gene families) were computed by assigning proteome sequences to precomputed metazoan families from the OMA database (2021.12 release) using OMAmer (version 0.2.1, sensitive option) (Altenhoff et al. 2021; Rossier et al. 2021). In particular, metazoan families with less than five species and with a species conservation (number of species with a gene divided by the number of species descending from the family root taxon) lower than 50% were removed as they are likely spurious and would slow down the sequence assignment step. The homologous family of a query protein can be missing due to the latter, to an incomplete set of reference families or the recent emergence of the gene. Thus, to minimise the number of false positive placements, which can happen in case of domain-level homology or by chance, we rejected partial and fragmented matches based on a score similar to other orthology inference approaches (Griesmann et al. 2018; Altenhoff et al. 2021). Specifically, we rejected any protein with an overlap length (length of sequence overlapping with *k*-mers of the reference family) smaller than a quarter of the sequence lengths median of its predicted subfamily. The resulting orthogroups consisted in the sets of proteins assigned to subfamilies that existed in the bilaterian ancestor (the last common ancestor of the selected species) or to more recent families. Thus, because we prioritised the number of proteins placed in these groups instead of the placement accuracy in specific bilaterian subfamilies, we disabled the OMAmer threshold that limits assignments to overly specific subfamilies.

Quality control of proteomes

Proteome quality measures were computed by running OMArk for each species on its set of proteins assigned to families (hereby its conserved proteome, *see previous section*) (Nevers et al. *in prep*, https://github.com/DessimozLab/OMArk). Indeed, the quality of unmapped proteins cannot impact downstream analysis. OMArk assesses proteome completeness by measuring the proportion of conserved gene families (with minimum 80% of

expected species) found in the protein set similarly to BUSCO (Simão et al. 2015). Moreover, a high proportion of these families found in multiple copies can indicate problematic isoform handling or many false positive gene models. The latter can also be concluded in case of high proportions of fragmented or partial matches identified by OMArk. Thus, to avoid biases in our analysis of convergence, we aimed to balance these measures between the venomous and outgroup species in each selected clade. For example, a high fraction of duplicates, fragmented matches and partial matches, as well as a higher completeness in the conserved proteomes of venomous species could inflate the proportion of predicted convergent venomous expansions. Thus, we filtered proteomes with very low values in one of these measures when none of the venomous or outgroup counterparts had a similar value (Supp. Dataset 1). We maintained the same number of species in venomous and outgroup clades following the species selection prioritisation described previously.

Inference of convergent expansions, contractions and losses

Each bilaterian family of each venomous clade was classified into one of four evolutionary events when displaying a copy number difference between venomous and outgroup species (Fig. 2, left). To avoid inferring expansions due to the lack of outgroup copies, discrete events (loss and outgroup loss) were classified separately from continuous ones (expansion and contraction). Because these events are exclusive as heads and tails of coins, a binomial function was used to model the probabilities of observing any event number per family.

$$P(e_f) = \left(\frac{e_f + \bar{e}_f}{e_f}\right) p^{e_f} (1-p)^{\bar{e}_f}$$

$$e_f = event\ nr.\ in\ family\ f\ (e.g. expansion)$$

$$\bar{e}_f = complementary\ event\ nr.\ in\ family\ f\ (e.g. contraction)$$

To account for clade-specific event frequencies (*e.g.* a clade having twice more expansions than contractions that would result in a higher probability to observe convergent expansions), the success probability p was calculated as follows.

$$p = \frac{1}{e_f + \bar{e}_f} \sum_{c=0}^{e_f + \bar{e}_f} \frac{e_c}{e_c + \bar{e}_c}$$

$$e_c = total\ nr.of\ event\ in\ clade\ c$$

$$\bar{e}_c = total\ nr.of\ complementary\ event\ in\ clade\ c$$

Expected distributions were simulated with 100 runs of this model per family.

Inference of large venomous expansions

A similar approach was used to model the probability to observe, in a gene family, a given sum of *ancestral venomous copy numbers* (approximated as the sum of clade median copy numbers). To account for clade- and family-specific proportions of ancestral venomous copy numbers, p was additionally weighted by the number of ancestral venomous copies of each clade in the family.

$$P(v_f) = \left(\frac{v_f + \bar{v}_f}{v_f}\right) p^{v_f}(1-p)^{\bar{v}_f}$$

$$v_f = ancestral\ venomous\ copy\ nr.in\ family\ f$$

$$\bar{v}_f = ancestral\ outgroup\ copy\ nr.in\ family\ f$$

$$p = \frac{1}{v_f + \bar{v}_f} \sum_{c=0}^{c} p_c v_{cf}$$

$$v_{cf} = ancestral\ venomous\ copy\ nr.of\ clade\ c\ in\ family\ f$$

$$p_c = \frac{v_c}{v_c + \bar{v}_c}$$

$$v_c = ancestral\ venomous\ copy\ nr.in\ clade\ c$$

$$\bar{v}_c = ancestral\ outgroup\ copy\ nr.in\ clade\ c$$

Expected distributions were simulated with 100 runs of this model per family.

Candidate venom families

To infer toxipotent families, we started by placing toxin sequences from UniProt-ToxProt (Jungo et al. 2012) in reference orthogroups as for orthology assignment. Then, each family with at least one toxin match was classified as a toxipotent family.

We refer to convergently expressed pathways for GO terms found significantly enriched in orthogroups with similar venom gland expression profiles from (Zancolli et al. 2022) (Supp. Dataset 1). Similarly, we refer to pathways with lineage specific expression profiles for GO terms found significantly enriched in orthogroups with lineage specific expression profiles (Supp. Dataset 1). Each family with at least one of these GO terms was classified as either a family from convergently expressed pathways or one from pathways with lineage-specific expression profiles.

GO enrichment tests

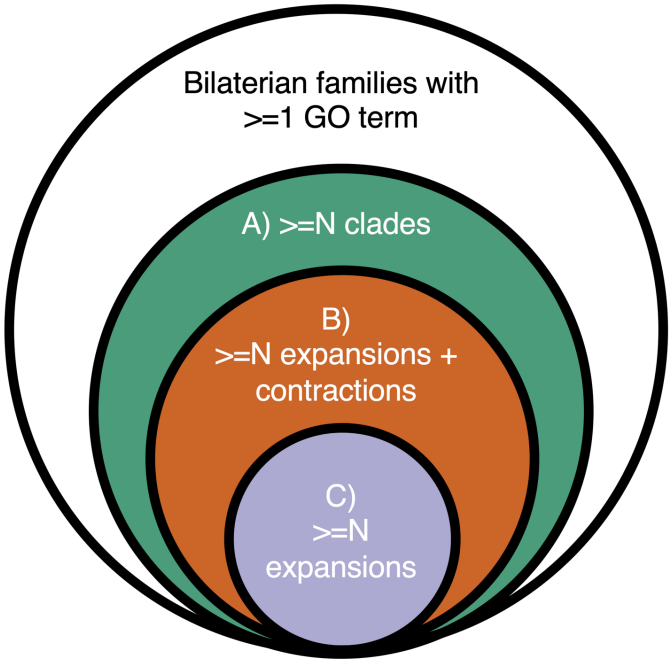


Fig. 6. Family subsets used for GO enrichment tests as foreground and background.

GO enrichment tests for biological processes were performed with TopGO (v. 2.44.0, weight.01 algorithm, fisher statistic test) on the go_basic.obo ontology (2021-Apr14 release) (Alexa and Rahnenfuhrer 2010). To minimise false negatives without decreasing precision, we ran enrichment tests with a permissive background and removed results found significant in a control test. The main test used families with at least *N* convergent events as foreground (Fig. 6, C) and families with at least *N* clade presences (*i.e.* at least one gene from the clade—what I call permissive) as background (Fig. 6, A). To control for a potential signal coming from

highly duplicated but non-convergent families, the control used a foreground that consisted of families with at least *N* venomous and outgroup events (*e.g.* expansion and contractions) but less than *N* venomous events (Fig. 6, B minus C). The control background was the same as the main test but without convergent families (Fig. 6, A minus C). All tests were performed on bilaterian families with at least one biological process GO term. Correction for multiple testing was performed using the qvalue R package (v. 2.24.0) (Dabney, Storey, and Warnes 2010). For each analysis, the top 20 most enriched GO terms with a non-significant control q-value (>0.05) were displayed.

References

Albalat, Ricard, and Cristian Cañestro. 2016. "Evolution by Gene Loss." Nature Reviews. Genetics 17 (7): 379–91.

Alexa, and Rahnenfuhrer. 2010. "topGO: Enrichment Analysis for Gene Ontology." R Package Version.

Alkan, Can, Saba Sajjadian, and Evan E. Eichler. 2011. "Limitations of next-Generation Genome Sequence Assembly." Nature Methods 8 (1): 61–65.

Altenhoff, Adrian M., Clément-Marie Train, Kimberly J. Gilbert, Ishita Mediratta, Tarcisio Mendes de Farias, David Moi, Yannis Nevers, et al. 2021. "OMA Orthology in 2021: Website Overhaul, Conserved Isoforms, Ancestral Gene Order and More." Nucleic Acids Research 49 (D1): D373–79.

Barua, Agneesh, Ivan Koludarov, and Alexander S. Mikheyev. 2021. "Co-Option of the Same Ancestral Gene Family Gave Rise to Mammalian and Reptilian Toxins." BMC Biology 19 (1): 268.

Barua, Agneesh, and Alexander S. Mikheyev. 2021. "An Ancient, Conserved Gene Regulatory Network Led to the Rise of Oral Venom Systems." Proceedings of the National Academy of Sciences of the United States of America 118 (14). https://doi.org/10.1073/pnas.2021311118.

Bastian, Frederic B., Julien Roux, Anne Niknejad, Aurélie Comte, Sara S. Fonseca Costa, Tarcisio Mendes de Farias, Sébastien Moretti, et al. 2021. "The Bgee Suite: Integrated Curated Expression Atlas and Comparative Transcriptomics in Animals." Nucleic Acids Research 49 (D1): D831–47.

Blount, Zachary D., Richard E. Lenski, and Jonathan B. Losos. 2018. "Contingency and Determinism in Evolution: Replaying Life's Tape." Science 362 (6415). https://doi.org/10.1126/science.aam5979.

Botero-Castro, Fidel, Emeric Figuet, Marie-Ka Tilak, Benoit Nabholz, and Nicolas Galtier. 2017. "Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds." Molecular Biology and Evolution 34 (12): 3123–31.

Breeschoten, Thijmen, Corné F. H. van der Linden, Vera I. D. Ros, M. Eric Schranz, and Sabrina Simon. 2022. "Expanding the Menu: Are Polyphagy and Gene Family Expansions Linked across Lepidoptera?" Genome Biology and Evolution 14 (1). https://doi.org/10.1093/gbe/evab283.

Casewell, Nicholas R., Daniel Petras, Daren C. Card, Vivek Suranse, Alexis M. Mychajliw, David Richards, Ivan Koludarov, et al. 2019. "Solenodon Genome Reveals Convergent Evolution of Venom in Eulipotyphlan Mammals." Proceedings of the National Academy of Sciences of the United States of America 116 (51): 25745–55.

Chang, Dan, and Thomas F. Duda Jr. 2012. "Extensive and Continuous Duplication Facilitates Rapid Evolution and Diversification of Gene Families." Molecular Biology and Evolution 29 (8): 2019–29.

Conant, Gavin C., and Kenneth H. Wolfe. 2008. “Turning a Hobby into a Job: How Duplicated Genes Find New Functions.” Nature Reviews. Genetics 9 (12): 938–50.

Dabney, Storey, and Warnes. 2010. “Qvalue: Q-Value Estimation for False Discovery Rate Control.” R Package Version.

Dobler, Susanne, Safaa Dalla, Vera Wagschal, and Anurag A. Agrawal. 2012. “Community-Wide Convergent Evolution in Insect Adaptation to Toxic Cardenolides by Substitutions in the Na,K-ATPase.” Proceedings of the National Academy of Sciences of the United States of America 109 (32): 13040–45.

Feron, Romain, and Robert M. Waterhouse. 2022. “Assessing Species Coverage and Assembly Quality of Rapidly Accumulating Sequenced Genomes.” GigaScience 11 (February). https://doi.org/10.1093/gigascience/giac006.

Fry, Bryan G., Kim Roelants, Donald E. Champagne, Holger Scheib, Joel D. A. Tyndall, Glenn F. King, Timo J. Nevalainen, et al. 2009. “The Toxicogenomic Multiverse: Convergent Recruitment of Proteins into Animal Venoms.” Annual Review of Genomics and Human Genetics 10: 483–511.

Gould, Stephen Jay. 1990. Wonderful Life: The Burgess Shale and the Nature of History. WW Norton & Company.

Griesmann, Maximilian, Yue Chang, Xin Liu, Yue Song, Georg Haberer, Matthew B. Crook, Benjamin Billault-Penneteau, et al. 2018. “Phylogenomics Reveals Multiple Losses of Nitrogen-Fixing Root Nodule Symbiosis.” Science 361 (6398). https://doi.org/10.1126/science.aat1743.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. “ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data.” Molecular Biology and Evolution 33 (6): 1635–38.

i5K Consortium. 2013. “The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment.” The Journal of Heredity 104 (5): 595–600.

Jungo, Florence, Lydie Bougueleret, Ioannis Xenarios, and Sylvain Poux. 2012. “The UniProtKB/Swiss-Prot Tox-Prot Program: A Central Hub of Integrated Venom Protein Data.” Toxicon: Official Journal of the International Society on Toxinology 60 (4): 551–57.

Koepfli, Klaus-Peter, Benedict Paten, Genome 10K Community of Scientists, and Stephen J. O’Brien. 2015. “The Genome 10K Project: A Way Forward.” Annual Review of Animal Biosciences 3: 57–111.

Kordis, D., and F. Gubensek. 2000. “Adaptive Evolution of Animal Toxin Multigene Families.” Gene 261 (1): 43–52.

Kuzmin, Elena, John S. Taylor, and Charles Boone. 2021. “Retention of Duplicated Genes in Evolution.” Trends in Genetics: TIG, July. https://doi.org/10.1016/j.tig.2021.06.016.

Losos, Jonathan B. 2011. “Convergence, Adaptation, and Constraint.” Evolution; International Journal of Organic Evolution 65 (7): 1827–40.

Luna, Milene S. A., Thiago M. A. Hortencio, Zulma S. Ferreira, and Norma Yamanouye. 2009. “Sympathetic Outflow Activates the Venom Gland of the Snake Bothrops Jararaca by Regulating the Activation of Transcription Factors and the Synthesis of Venom Gland Proteins.” The Journal of Experimental Biology 212 (Pt 10): 1535–43.

Martinson, Ellen O., Mrinalini, Yogeshwar D. Kelkar, Ching-Ho Chang, and John H. Werren. 2017. “The Evolution of Venom by Co-Option of Single-Copy Genes.” Current Biology: CB 27 (13): 2007–13.e8.

Nagy, László G., Zsolt Merényi, Botond Hegedüs, and Balázs Bálint. 2020. “Novel Phylogenetic Methods Are Needed for Understanding Gene Function in the Era of Mega-Scale Genome Sequencing.” Nucleic Acids Research 48 (5): 2209–19.

Rosenblum, Erica Bree, Christine E. Parent, and Erin E. Brandt. 2014. “The Molecular Basis of Phenotypic Convergence.” Annual Review of Ecology, Evolution, and Systematics 45 (1): 203–26.

Rossier, Victor, Alex Warwick Vesztrocy, Marc Robinson-Rechavi, and Christophe Dessimoz. 2021. “OMAmer: Tree-Driven and Alignment-Free Protein Assignment to Subfamilies Outperforms Closest Sequence Approaches.” Bioinformatics , March. https://doi.org/10.1093/bioinformatics/btab219.

Sackton, Timothy B., and Nathan Clark. 2019. “Convergent Evolution in the Genomics Era: New Insights and Directions.” Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 374 (1777): 20190102.

Schendel, Vanessa, Lachlan D. Rash, Ronald A. Jenner, and Eivind A. B. Undheim. 2019. “The Diversity of Venom: The Importance of Behavior and Venom System Morphology in Understanding Its Ecology and Evolution.” Toxins 11 (11). https://doi.org/10.3390/toxins11110666.

Sémon, Marie, and Kenneth H. Wolfe. 2007. “Consequences of Genome Duplication.” Current Opinion in Genetics & Development 17 (6): 505–12.

Seppey, Mathieu, Panagiotis Ioannidis, Brent C. Emerson, Camille Pitteloud, Marc Robinson-Rechavi, Julien Roux, Hermes E. Escalona, et al. 2019. “Genomic Signatures Accompanying the Dietary Shift to Phytophagy in Polyphagan Beetles.” Genome Biology 20 (1): 98.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” Bioinformatics 31 (19): 3210–12.

Stern, David L. 2013. “The Genetic Causes of Convergent Evolution.” Nature Reviews. Genetics 14 (11): 751–64.

Storz, Jay F. 2016. “Causes of Molecular Convergence and Parallelism in Protein Evolution.” Nature Reviews. Genetics 17 (4): 239–50.

Weisman, Caroline M., Andrew W. Murray, and Sean R. Eddy. 2022. “Mixing Genome Annotation Methods in a Comparative Analysis Inflates the Apparent Number of Lineage-Specific Genes.” Current Biology: CB 32 (12): 2632–39.e2.

Zancolli, Giulia, and Nicholas R. Casewell. 2020. “Venom Systems as Models for Studying the Origin and Regulation of Evolutionary Novelties.” Molecular Biology and Evolution 37 (10): 2777–90.

Zancolli, Giulia, Maarten Reijnders, Robert M. Waterhouse, and Marc Robinson-Rechavi. 2022. “Convergent Evolution of Venom Gland Transcriptomes across Metazoa.” Proceedings of the National Academy of Sciences of the United States of America 119 (1). https://doi.org/10.1073/pnas.2111392119.

Supplementary material

Supplementary tables

Supp. Table 1.

Family	Convergence level	Human gene representatives
HOG:B0613860.9djj.2100a.1520a	5 / 8	https://www.uniprot.org/uniprot/Q9VAT0 https://www.uniprot.org/uniprot/O00401 https://www.uniprot.org/uniprot/P42768
HOG:B0606155.2a	5 / 8	https://www.uniprot.org/uniprot/Q8NFP9 https://www.uniprot.org/uniprot/A0A494C1L5 https://www.uniprot.org/uniprot/B7Z0W8

HOG:B0622094.3a	5 / 8	https://www.uniprot.org/uniprot/P28715 https://www.uniprot.org/uniprot/Q4U2Q5 https://www.uniprot.org/uniprot/A0A494BZX4
HOG:B0596448	5 / 8	https://www.uniprot.org/uniprot/P20153 https://www.uniprot.org/uniprot/P19793 https://www.uniprot.org/uniprot/P48443
HOG:B0589384	5 / 8 4 / 5 (Carnivorous)	https://www.uniprot.org/uniprot/Q9W0T1 https://www.uniprot.org/uniprot/Q12830
HOG:B0634297.1b	5 / 8 4 / 5 (Carnivorous)	https://www.uniprot.org/uniprot/A0A0C4DG76 https://www.uniprot.org/uniprot/P00450 https://www.uniprot.org/uniprot/Q6MZM0
HOG:B0593984.3h	4 / 5 (Carnivorous)	https://www.uniprot.org/uniprotkb/P02462/entry https://www.uniprot.org/uniprotkb/P29400/entry
HOG:B0833930.8a	4 / 5 (Carnivorous)	https://www.uniprot.org/uniprotkb/P49643/entry
HOG:B0613891.8b.38e.20a.10a	4 / 5 (Carnivorous)	https://www.uniprot.org/uniprotkb/Q9NR22/entry https://www.uniprot.org/uniprotkb/Q99873/entry
HOG:B0602762	4 / 5 (Carnivorous)	https://www.uniprot.org/uniprotkb/Q9NRC6/entry

Supp. Table 2.

GO	Term	StudyCount	Enrichment	Qvalue	Control Qvalue	Pvalue	Control Pvalue
GO:0006887	exocytosis	43	2.25	0.001	0.637	1.90E-07	0.096
GO:0040017	positive regulation of locomotion	54	2.03	0.002	0.12	6.30E-07	0.003
GO:0001894	tissue homeostasis	34	2.42	0.002	0.634	7.10E-07	0.057
GO:0007419	ventral cord development	11	6.11	0.003	0.091	1.30E-06	0.002
GO:0098789	pre-mRNA cleavage required for polyadeny...	6	13.64	0.003	1	1.90E-06	0.434
GO:0006893	Golgi to plasma membrane transport	12	3.7	0.013	0.517	9.10E-06	0.041

GO:0040010	positive regulation of growth rate	6	10.91	0.013	0.517	9.20E-06	0.043
GO:0072347	response to anesthetic	5	12.5	0.032	0.263	2.60E-05	0.012
GO:0009408	response to heat	20	2.8	0.034	0.211	3.00E-05	0.007
GO:0051481	negative regulation of cytosolic calcium...	6	9.09	0.034	1	3.10E-05	1
GO:0071260	cellular response to mechanical stimulus	13	3.8	0.037	0.41	3.50E-05	0.027
GO:0048499	synaptic vesicle membrane organization	4	15.38	0.051	0.986	5.80E-05	0.248
GO:0010881	regulation of cardiac muscle contraction...	6	7.79	0.066	0.517	8.20E-05	0.041
GO:0048813	dendrite morphogenesis	33	3.17	0.068	1	9.40E-05	0.574
GO:0051984	positive regulation of chromosome segreg...	7	7.61	0.068	0.077	0.0001	0.001
GO:0055123	digestive system development	34	2.74	0.068	0.634	0.0001	0.072
GO:0033627	cell adhesion mediated by integrin	11	2.96	0.068	1	0.00011	1
GO:0007409	axonogenesis	76	2.37	0.08	1	0.00013	0.758
GO:0006334	nucleosome assembly	14	3.55	0.08	1	0.00014	0.272
GO:0051290	protein heterotetramerization	5	9.09	0.081	1	0.00015	0.613

Supp. Table 3.

Family	Venomous expansion nr.	Function

HOG:B0575242 https://omabrowser.org/oma/hog/HOG%3AB0575242/iham/	4	Regulation of transcription and odontogenesis https://www.uniprot.org/uniprotkb/P17026/entry https://www.uniprot.org/uniprotkb/Q9C0F3/entry
HOG:B0518831 https://omabrowser.org/oma/hog/HOG%3AB0518831/iham/	3	Regulation of transcription
HOG:B0490019 https://omabrowser.org/oma/hog/HOG%3AB0490019/iham/	4	Regulation of transcription
HOG:B0410500 https://omabrowser.org/oma/hog/HOG%3AB0410500/iham/	2	Regulation of transcription

Supp. Table 4

GO	Term	StudyCount	Enrichment	Qvalue	Control Qvalue	Pvalue	Control Pvalue
GO:0016339	calcium-dependent cell-cell adhesion via...	4	20	0.718	1	4.40E-05	1
GO:0097066	response to thyroid hormone	4	19.05	0.853	1	0.00013	1
GO:0042742	defense response to bacterium	14	6.73	0.853	1	0.00018	0.068
GO:0042572	retinol metabolic process	4	13.33	0.853	1	0.00021	0.241
GO:0072734	cellular response to staurosporine	2	66.67	0.853	1	0.00038	0.227
GO:0008340	determination of adult lifespan	8	4.49	0.853	1	0.00039	0.104
GO:0019915	lipid storage	6	6.25	0.853	1	0.00039	0.139
GO:0035725	sodium ion transmembrane transport	7	6.54	0.853	1	0.00042	0.065
GO:0007586	digestion	4	4.55	0.942	1	0.00052	1

GO:0019853	L-ascorbic acid biosynthetic process	2	50	0.942	1	0.00063	0.08
GO:0061959	response to (R)-carnitine	2	50	0.942	1	0.00063	1
GO:0045905	positive regulation of translational ter...	1	50	1	1	0.01607	1
GO:0045901	positive regulation of translational elo...	1	33.33	1	1	0.03189	0.545
GO:0032402	melanosome transport	1	4	1	1	0.22238	0.564
GO:0016339	calcium-dependent cell-cell adhesion via...	4	20	1	0.524	4.40E-05	0.001
GO:0097066	response to thyroid hormone	4	19.05	1	1	0.00013	1
GO:0042742	defense response to bacterium	14	6.73	1	1	0.00018	0.266
GO:0042572	retinol metabolic process	4	13.33	1	1	0.00021	0.614
GO:0072734	cellular response to staurosporine	2	66.67	1	1	0.00038	1
GO:0008340	determination of adult lifespan	8	4.49	1	1	0.00039	1

Supp. Table 5

GO	Term	StudyCount	Enrichment	Qvalue	Control Qvalue	Pvalue	Control Pvalue
GO:0001666	response to hypoxia	63	2.55	0	0.154	2.10E-09	0.002
GO:0006629	lipid metabolic process	218	2.23	0	0.135	1.10E-08	0.001
GO:0016241	regulation of macroautophagy	33	3.07	0	0.427	3.60E-08	0.012

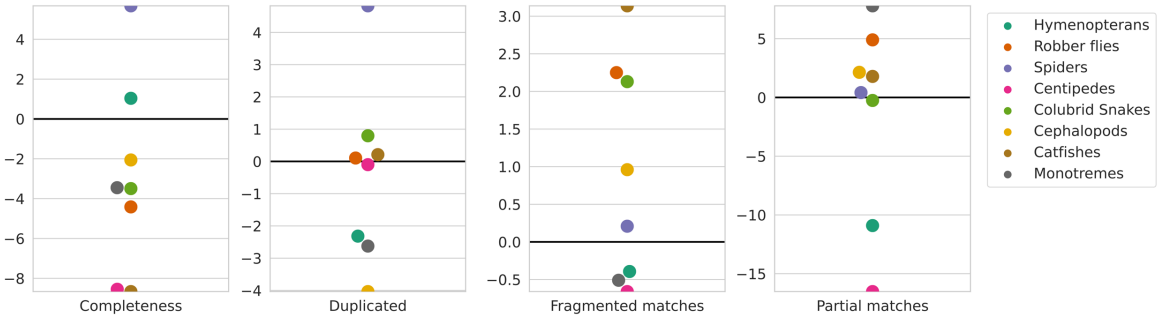
GO:1901575	organic substance catabolic process	301	2.11	0	1	4.20E-08	0.331
GO:0001676	long-chain fatty acid metabolic process	28	5.29	0	0.133	7.60E-08	0.001
GO:0051028	mRNA transport	31	3.3	0	1	2.10E-07	0.408
GO:0032355	response to estradiol	29	3.02	0	0.368	2.80E-07	0.009
GO:0006749	glutathione metabolic process	19	3.75	0.001	1	4.90E-07	0.462
GO:0035011	melanotic encapsulation of foreign targe...	10	7.63	0.001	0.181	5.70E-07	0.002
GO:0042593	glucose homeostasis	43	2.34	0.001	1	6.00E-07	0.236
GO:0002181	cytoplasmic translation	32	3.12	0.001	0.582	7.90E-07	0.025
GO:0050829	defense response to Gram-negative bacter...	27	2.89	0.001	0.149	1.50E-06	0.002
GO:0071356	cellular response to tumor necrosis fact...	33	2.53	0.001	1	1.50E-06	0.568
GO:0007595	lactation	14	4.39	0.002	1	2.10E-06	0.35
GO:0007584	response to nutrient	43	3.25	0.002	0.427	2.20E-06	0.013
GO:0003007	heart morphogenesis	50	2.03	0.002	0.427	2.60E-06	0.013
GO:0050878	regulation of body fluid levels	66	2.19	0.002	0.427	2.70E-06	0.013
GO:0046677	response to antibiotic	18	5.03	0.002	1	3.00E-06	0.935
GO:0051930	regulation of sensory perception of pain	14	4.83	0.002	0.733	3.00E-06	0.065
GO:0021762	substantia nigra development	12	4.9	0.002	0.262	3.20E-06	0.005

Supp. Table 6

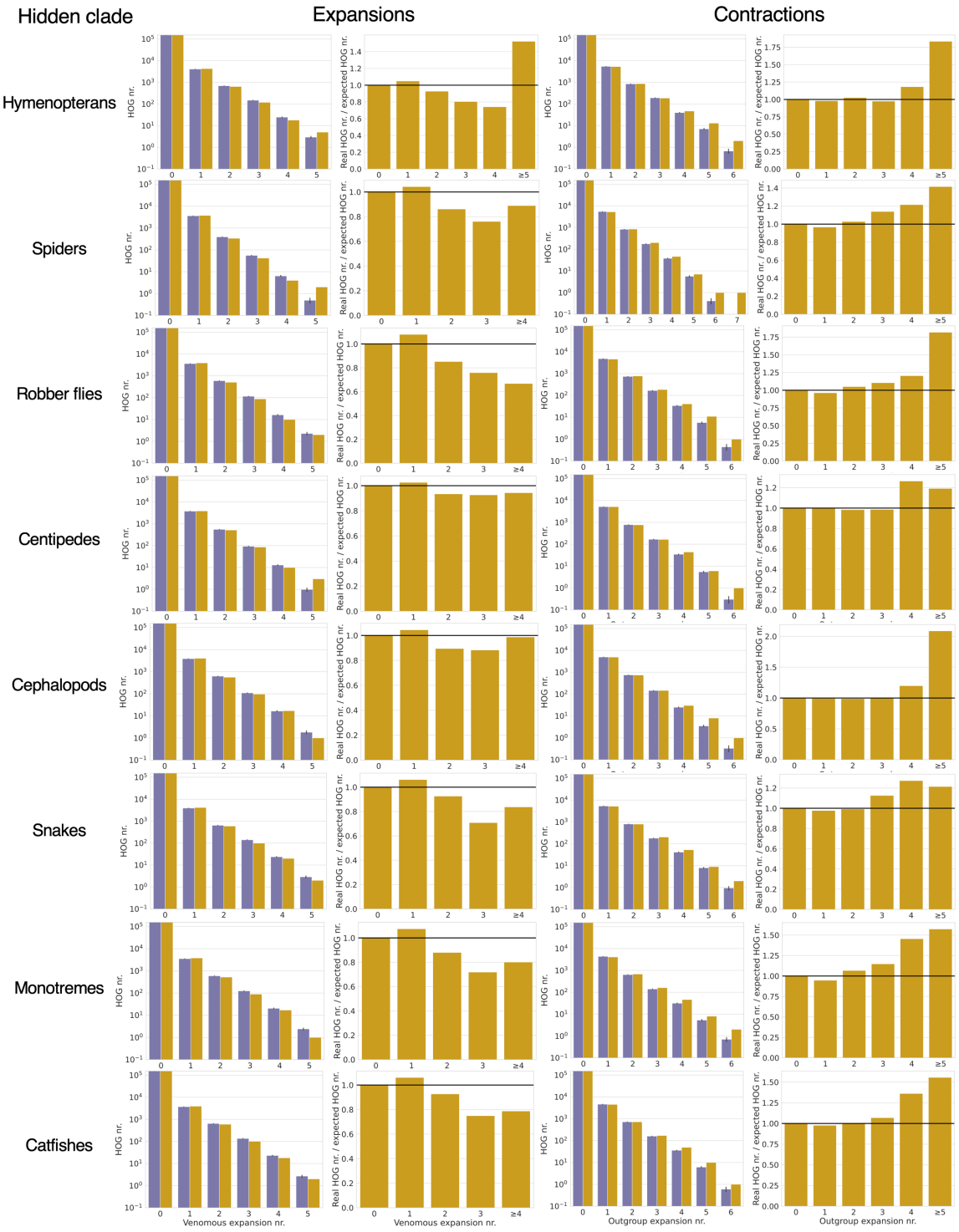
GO	Term	StudyCount	Enrichment	Qvalue	Control Qvalue	Pvalue	Control Pvalue
GO:0032355	response to estradiol	13	4.13	0.011	0.843	2.50E-06	0.029
GO:0030324	lung development	14	3.21	0.011	1	2.70E-06	0.101
GO:0006809	nitric oxide biosynthetic process	10	6.45	0.013	1	3.70E-06	1
GO:0071356	cellular response to tumor necrosis fact...	16	4.3	0.028	0.134	9.80E-06	0
GO:0071548	response to dexamethasone	9	7.09	0.029	1	1.20E-05	1
GO:0002931	response to ischemia	8	6.84	0.038	1	1.90E-05	0.252
GO:0045819	positive regulation of glycogen cataboli...	4	22.22	0.038	1	2.00E-05	0.529
GO:0050829	defense response to Gram-negative bacter...	12	4.72	0.059	1	3.40E-05	0.088
GO:0007568	aging	34	3.12	0.082	1	5.40E-05	0.424
GO:0032024	positive regulation of insulin secretion	8	4.4	0.082	1	5.70E-05	0.843
GO:0007204	positive regulation of cytosolic calcium...	17	2.83	0.082	0.675	6.20E-05	0.015
GO:0002023	reduction of food intake in response to ...	3	30	0.088	1	8.20E-05	1
GO:0010701	positive regulation of norepinephrine se...	3	30	0.088	1	8.20E-05	1
GO:0016042	lipid catabolic process	24	3.52	0.11	1	0.00011	0.101
GO:0035148	tube formation	13	2.69	0.13	1	0.00014	0.47

GO:0035924	cellular response to vascular endothelia...	6	5.13	0.13	1	0.00014	1
GO:0015015	heparan sulfate proteoglycan biosyntheti...	3	25	0.139	1	0.00016	0.16
GO:0032757	positive regulation of interleukin-8 pro...	6	6.67	0.2	0.392	0.00026	0.005
GO:0032722	positive regulation of chemokine product...	6	6.25	0.2	1	0.00027	0.15
GO:0002232	leukocyte chemotaxis involved in inflamm...	3	21.43	0.2	1	0.00028	0.208

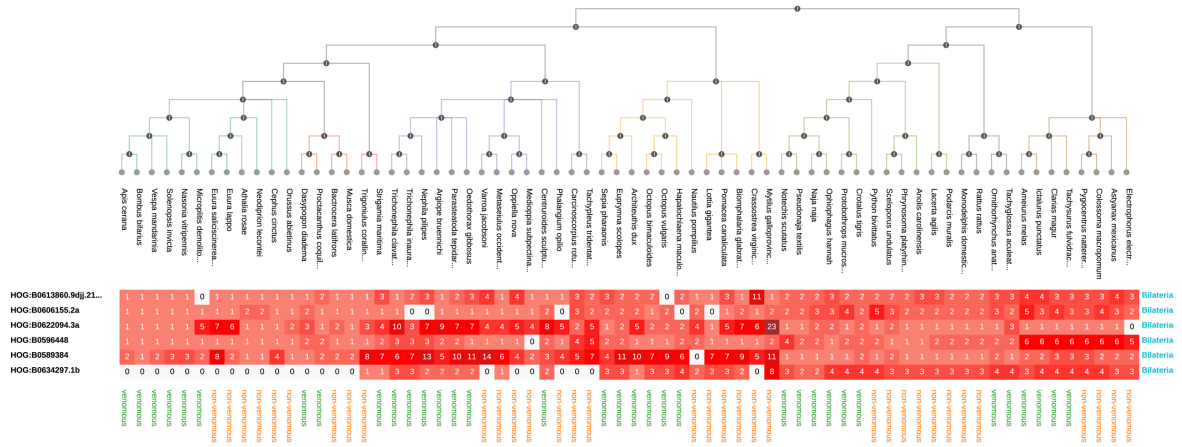
Supplementary figures



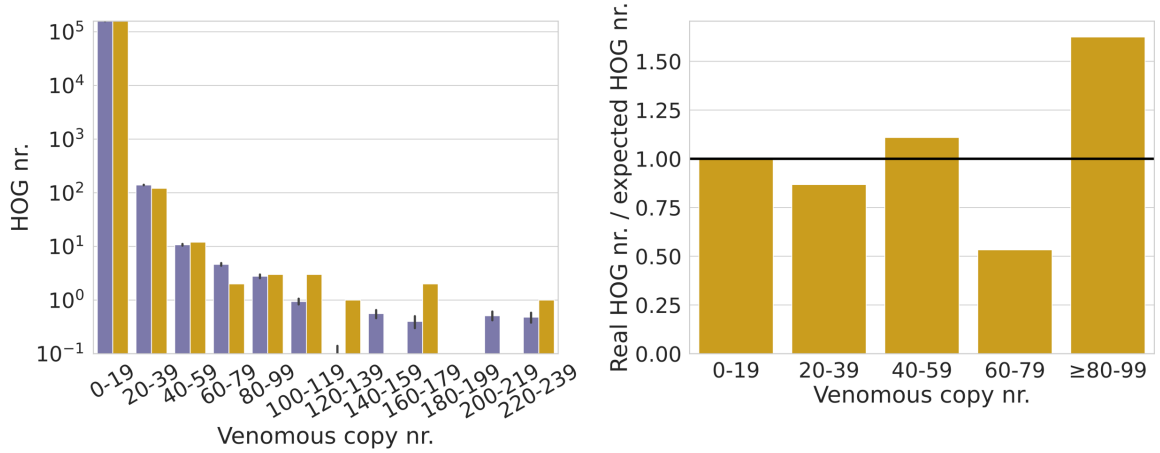
Supp. Figure 1. Quality balance of conserved proteomes.



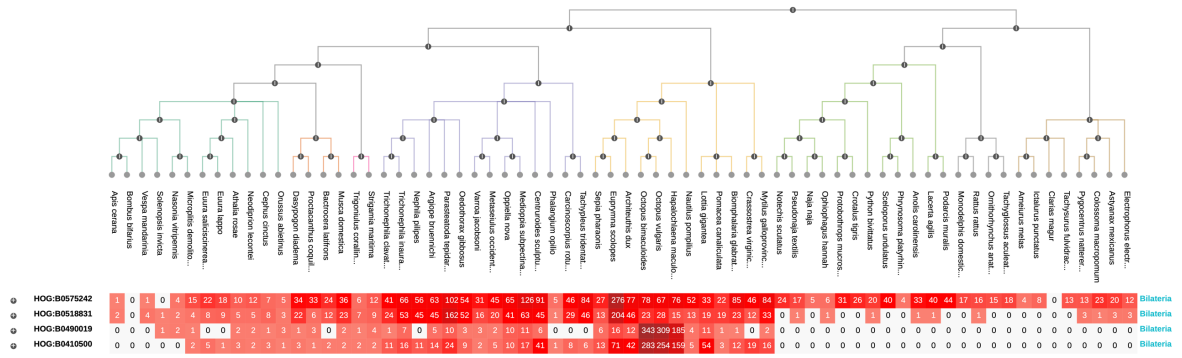
Supp. Figure 2. “Leave-one-out” robustness test.



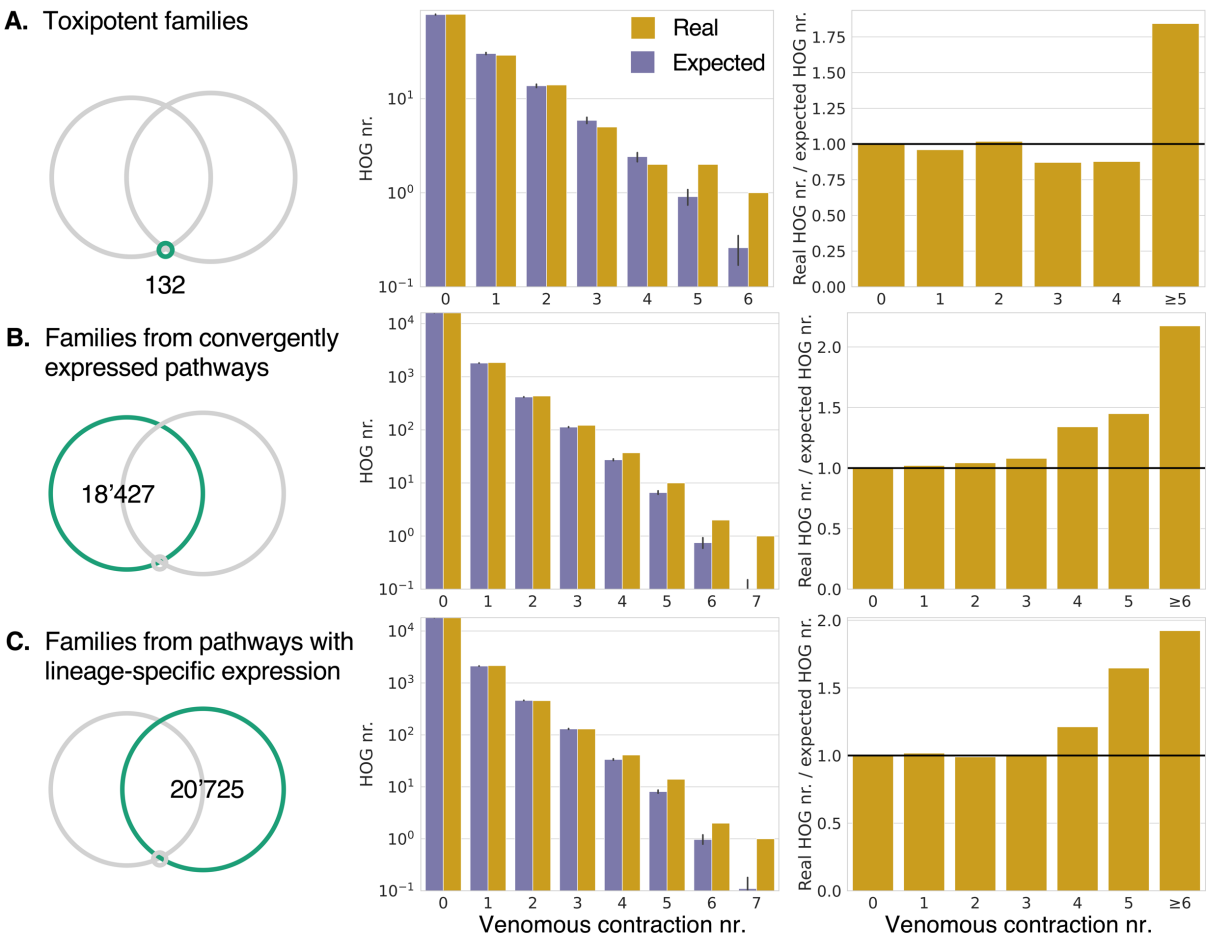
Supp. Figure. 3. Matreex visualisation of families with five convergent expansions.



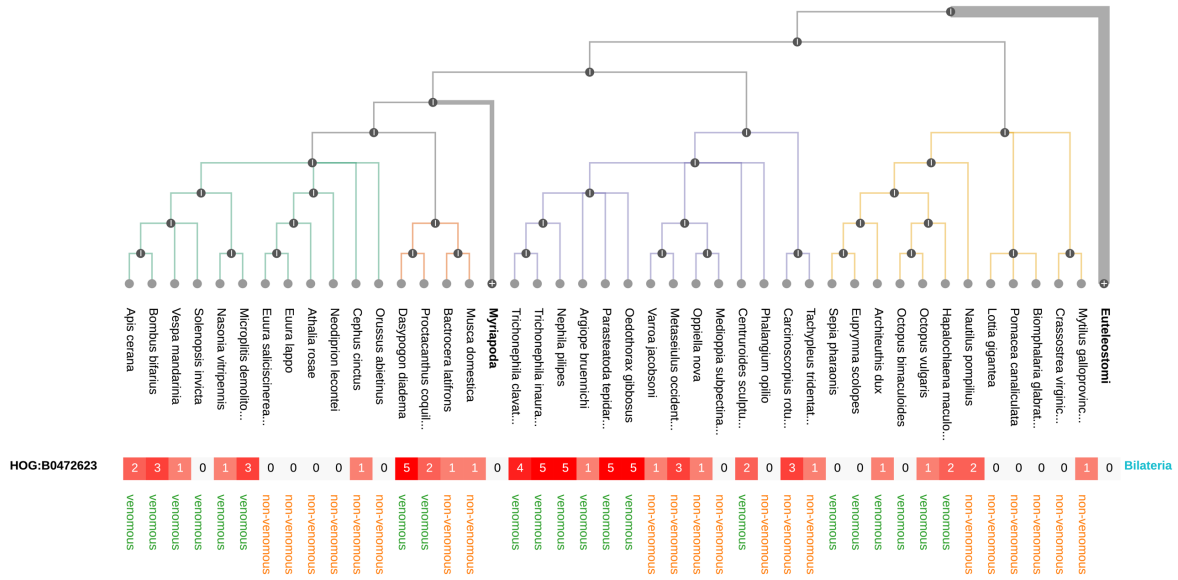
Supp. Figure 4. Families with large expansions.



Supp. Figure 5. Matreex view of families with large and convergent expansions.



Supp. Figure 6.



Supp. Figure 7. Matreex visualisation of Phospholipase type III subfamily.

Supp. Dataset 1.

https://docs.google.com/spreadsheets/d/1DBzMNe2pG8gjbr1ul9nmBzn2UY_SsAtuSlWq_UcSg3o/edit#gid=0

Chapter 5

Correlating gene content evolution with seven bird phenotypes

Correlating gene content evolution with seven bird phenotypes

Victor Rossier, Adrian Altenhoff, Alex Warwick Vesztrocy, Marc Robinson-Rechavi and Christophe Dessimoz

Introduction

The increasing number of sequenced genomes provides new opportunities to improve our functional understanding of genomes using macro-evolutionary approaches to identify the molecular basis of phenotypes (S. D. Smith et al. 2020; Nagy et al. 2020). Indeed, such approaches benefit specifically from an increased species sampling as their statistical power mostly depends on the availability of convergent phenotypic transitions (S. D. Smith et al. 2020). In contrast to population genetic approaches to map phenotypes to genotypes (*e.g.* GWAS), macro-evolutionary approaches enable the study of phenotypes that have been fixed across entire populations (*e.g.* the pouch of marsupials) (S. D. Smith et al. 2020). Briefly, they rely on explicit evolutionary relationships to correlate phenotypic transitions to molecular evolutionary changes using gene trees as backbones (Nagy et al. 2020; S. D. Smith et al. 2020). Duplications and losses that lead to gene family expansions and contractions are perhaps the lowest hanging fruit among evolutionary changes as they can be inferred solely based on gene trees. Such an approach has been used with success in fungi to reveal gene families associated with yeast-like forms (Nagy et al. 2014), white rot (Nagy et al. 2017) and multicellularity (Merényi et al. 2020). However, as gene trees are computationally costly, hierarchical orthologous groups (HOGs) provide a scalable alternative to study the genetic basis of convergent phenotypic transitions (Train et al. 2017; Zahn-Zabal, Dessimoz, and Glover 2020).

Focusing on recent and densely sampled clades is particularly promising to uncover phenotype-genotype associations at a macro-evolutionary scale for a number of reasons. First, closely related species have started diverging more recently and thus are more likely to reuse identical phenotypic or genetic solutions to solve similar problems (Rosenblum, Parent, and Brandt 2014). For example, because birds evolved beaks and lost fingers after their divergence from mammals ~300 MYA (Delsuc et al. 2018), woodpeckers and aye-aye lemurs rely on non-homologous organs (fingers and beaks) to fill the same ecological niche (extracting grubs from dead wood) (Blount, Lenski, and Losos 2018). Second, when clades have been sampled down to the genus or species level, a huge variety of phenotypic transitions becomes available. Thus,

the genetic basis of multiple phenotypes can be investigated at once and sometimes with multiple convergent replicates. Third, genomic data available within a clade is more likely to be homogeneous thanks to established community standards for sampling and analyses methods. Indeed, sequencing initiatives affiliated to the Earth BioGenome project (Lewin et al. 2022) are organised around distinctive clades (*e.g.* birds [Feng et al. 2020], arthropods [i5K Consortium 2013], vertebrates [Rhie et al. 2021]). For example, the Bird 10'000 Genomes (B10K) Project (re)annotated all genomes with the same protocol (Feng et al. 2020). Fourth, identifying homologous sequences is easier at smaller evolutionary scales due to the higher average similarity among sequences. Moreover, a denser sampling can provide the missing links to connect highly divergent but homologous sequences. Consequently, referenced-based assemblies, genome annotations, as well as orthology and phylogenetic inference should improve in quality within densely sampled clades. The same applies for phenotypic data, where homology between closely related structures is more easily identified.

Birds (*Aves*) are particularly promising to link phenotypes to genotypes at a macro-evolutionary scale. First, birds depict some of the best documented examples of convergent evolution starting with the convergent shape of bills in Darwin's finches or independent adaptations to arctic environments in penguins and auks (Stiller and Zhang 2019). More recently, a strong association between morphological forms and ecological niches has been established across all birds, which suggests a pervasive role of convergent evolution in birds (Pigot et al. 2020). Second, birds have a more compact genome than most vertebrates (Bravo, Schmitt, and Edwards 2021) and thus less genetic material for alternative molecular evolutionary roads (Rosenblum, Parent, and Brandt 2014). Third, birds are the second vertebrate clade with the most genomes in NCBI after bony fishes (512 vs. 638 genomes on January 1, 2020 [Bravo, Schmitt, and Edwards 2021]), while it is more densely sampled (5% vs. 3% of species). This mostly results from the effort of the B10K Project, which recently released a dataset of 363 bird genomes covering 92% of bird families (Feng et al. 2020). Moreover, phenotypic data covering six ecological variables and 11 continuous morphological traits has been recently released for ~10'000 bird species (*i.e.* AVONET [Tobias et al. 2022]).

Although convergence in conserved non-coding elements has been associated with beak morphology (Yusuf et al. 2020) and loss of flight (Sackton et al. 2019), advances in identifying the molecular basis of bird phenotypes remains limited (Bravo, Schmitt, and Edwards 2021). The fragmented nature of bird assemblies and the paucity of macro-

evolutionary phenotype-genotype association methods have been identified as potential explanatory factors (Stiller and Zhang 2019; S. D. Smith et al. 2020; Bravo, Schmitt, and Edwards 2021). Comparing a large number of genomes also poses considerable computational challenges. Indeed, typical comparative genomic approaches rely on pairwise comparison of genomes to infer homologous and orthologous genes, at a computational cost which grows quadratically with the number of genomes (Forslund et al. 2018; Linard et al. 2021).

In this study, we investigated the molecular basis of seven phenotypes that evolved convergently in birds, such as diving in penguins and auks, or loss of flights in ostriches and kiwis. To this end, we first developed fast-OMA, a more scalable version of OMA that starts with the assignment of protein sequences to reference HOGs to reduce the number of pairwise comparisons. Second, we applied this pipeline to infer hierarchical orthologous groups (HOGs) for 363 bird genomes recently released by the B10K initiative (Altenhoff et al. 2021; Feng et al. 2020) and third, searched for gene families with unexpected levels of convergent expansions or contractions in branches with phenotypic transitions. Notably, hemoglobin was found to be the family with the most expansions associated with diving, which supports an adaptation to breath-holding. Moreover, we observed hundreds of gene families with convergent contractions associated with the loss of flight. Interestingly, these families were enriched in regulators of bone morphogenetic proteins, which play a role in forelimb and feather development.

Results and discussion

We investigated seven phenotypes that evolved convergently multiple times in birds: diving, nocturnality, loss of flight, frugivory or nectarivory (sugar-based diet), herbivory, invertivore aerial lifestyle (*e.g.* swallow and swift) and arid adaptation. Phenotypic transitions were mapped on the bird phylogeny partly automatically with ancestral state reconstructions of AVONET ecological variables (*e.g.* trophic niche and habitat) (Tobias et al. 2022) and otherwise with literature searches.

Then, phenotypic transitions that correlated with an increased or reduced ancestral gene content (expansions or contractions) were recorded for each family (*i.e.* root-HOG). To control for family-specific duplication and loss rates, this difference was further required to be higher than in a close outgroup to count an expansion and or lower for contractions. For example, if hemoglobin alpha duplicated twice in the penguin ancestor (diving transition) and hemoglobin

beta was lost, while both hemoglobins were conserved in the ancestor of petrels and albatrosses (the selected outgroup branch for penguin), we would count an expansion. Then, to have a measure of unexpectedness, the resulting distributions of convergent expansion and contraction numbers across bird families were contrasted with null models. Precisely, we used the ratio between the observed and the expected number of families with a given number of expansions or contractions to infer unexpectedness. Then, families with unexpected levels of molecular convergence were investigated by testing for functional enrichments and visualised with Matreex (in the thesis: Chapter 3).

To minimise confirmation biases (Pavlidis et al. 2012), we focused on GO terms and gene families that fitted a precompiled set of hypotheses (copied in Supp. Table 2). This approach, which partly relies on prior knowledge, provides the added value of not exclusively relying on functions with significant q-values (p-values corrected for multiple testing). Indeed, as convergent genetic changes are expected to be the "needle in the haystack" of possible evolutionary roads towards phenotypic convergence, high significance levels were not expected (Stiller and Zhang 2019). Moreover, correcting for multiple testing for GO enrichment tests remains under debate in the community, in particular for decorrelation tests as used here (Alexa and Rahnenführer 2009; Hung et al. 2012). Briefly, the correlated natures of GO terms (all related in the GO graph) can lead to overly conservative corrections (Goeman and Mansmann 2008).

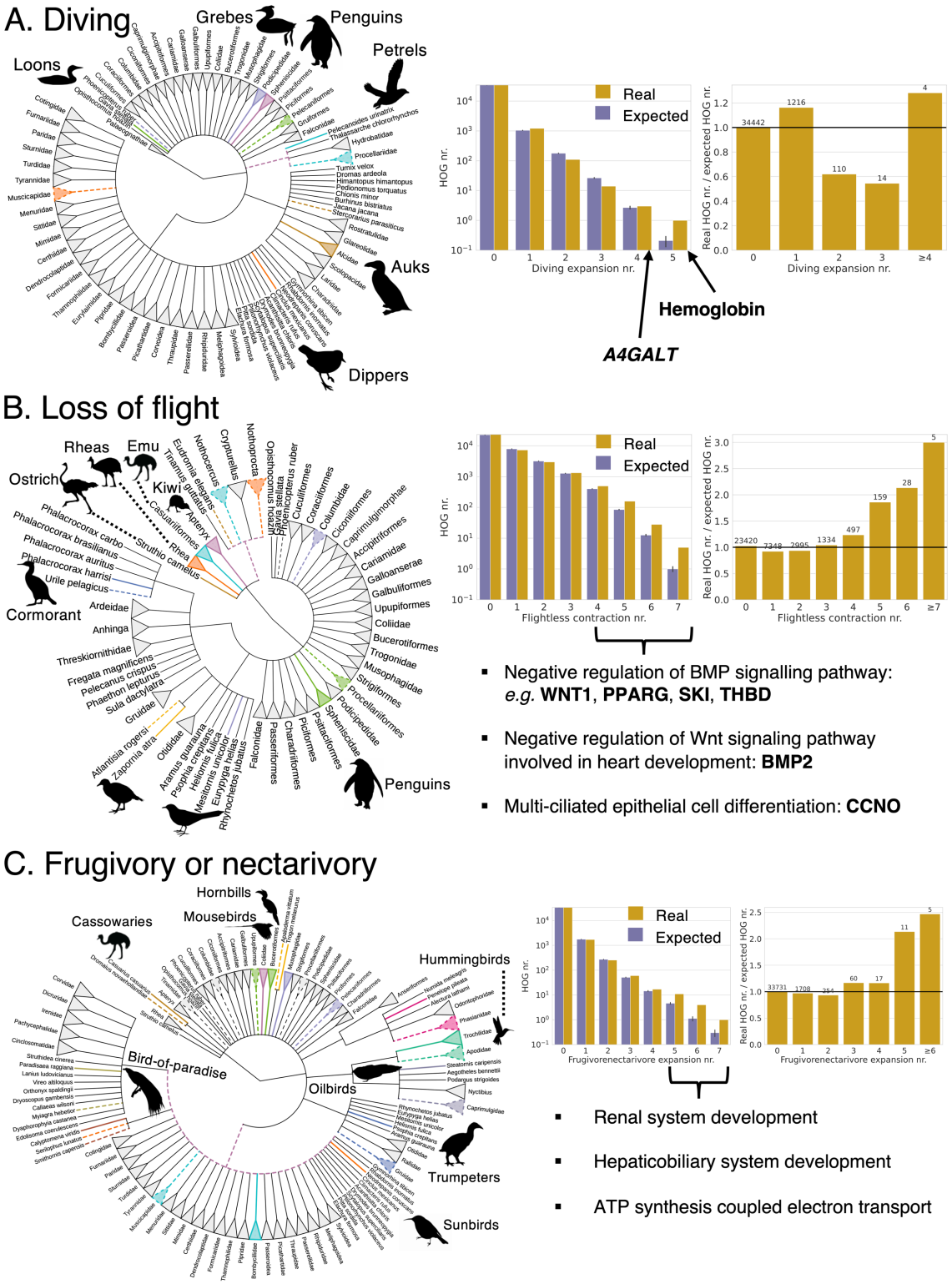


Figure 1. Gene family expansions and contractions correlate with three convergent bird phenotypes. (Left) NCBI taxonomy annotated with convergent phenotypic transitions (colored clades with solid lines), closest outgroups (coloured clades with dashed lines) and illustrated using PhyloPic silhouettes. (Right) The left histograms show observed (real) and expected distributions of expansion and contraction numbers across bird gene families. The right histograms show the ratio between these two distributions. Numbers on top of yellow bars are the numbers of families with the given level of

convergence (*e.g.* four families with at least four expansions that correlates with the transition to diving). A. The two convergently expanded families with expected functions are annotated. B-C. The expected functions enriched in convergently contracted or expanded families are annotated (with some example families in B.). Brackets highlight families tested for functional enrichment.

Diving correlates with convergent expansions in hemoglobin and the glycosphingolipid enzyme *A4GALT*

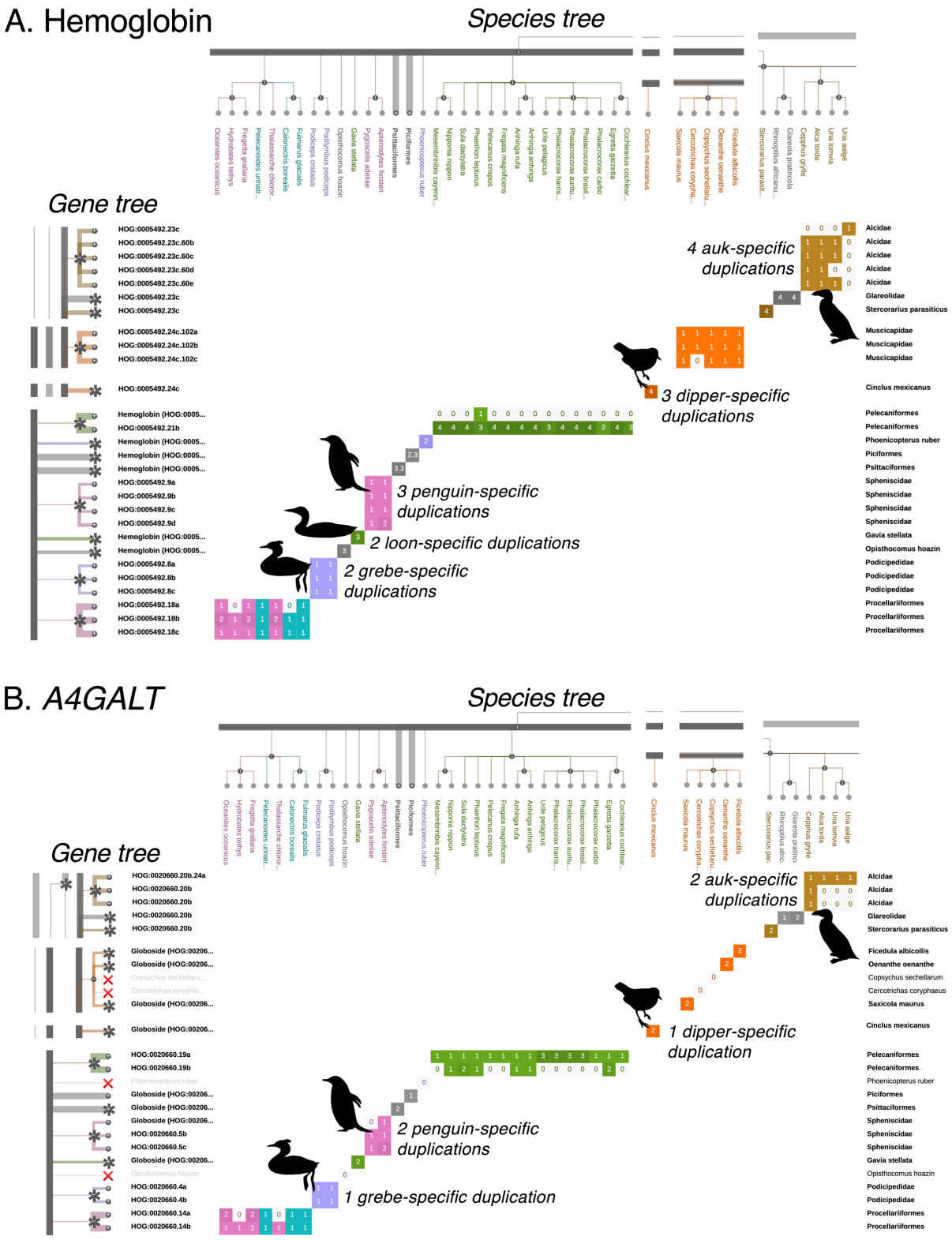


Figure 2. Convergent expansions associated with the transition to diving in Hemoglobin and 4-alpha-galactosyltransferase (*A4GALT*) families. Figure created using Matreex, whose layout consists of a gene tree (left), a species tree (top) and a matrix of phylogenetic profiles. Each diving clade and its corresponding outgroup are annotated with the same colour in tree branches and matrix' cells. Diving clades that underwent an expansion are annotated with a PhyloPic silhouette and their number of clade-

specific duplications. Discontinuities in both trees result from the removal of non-relevant clades in polytomic nodes.

Birds have evolved the ability to dive at depth repeatedly using two main propulsion mechanisms. Wing-propelled divers (*e.g.* penguins, awks, dippers) have adapted their wings for underwater propulsion. For example, penguins display modified stiff wing bones with scale-like feathers (Li et al. 2014). By contrast, foot-propelled divers (e.g. loons, grebes and *Pelecanoides* petrels) rely solely on their posteriorly positioned webbed-feet for propulsion (Gayk et al. 2018). Emperor penguins hold the record of the deepest (550m) and longest dive (22 min) (Kooyman and Ponganis 1998), while loons can reach depths of 60m and remain underwater for up to three minutes (Gayk et al. 2018). Thus, diving birds exhibit denser bones to overcome water buoyancy and an enhanced oxygen metabolism for holding their breath while diving (Li et al. 2014; Kooyman and Ponganis 1998). We also hypothesised that diving birds have adaptations related to vision and lipid metabolism (for thermoregulation).

We identified four families with at least four convergent expansions out of six species of diving birds, 1.2 times more than expected (Fig. 1). In particular, we identified five convergent expansions in the hemoglobin family (Fig. 2). Increasing hemoglobin production or its affinity to oxygen allows more oxygen to be stored in the body for breath-hold diving (Kooyman and Ponganis 1998). Even the dippers, which have recently diverged from non-diving songbirds (*Passeriformes*), display a higher hemoglobin concentration than non-diving birds (N. A. Smith et al. 2021). Moreover, positive selection has been observed in penguin and loon hemoglobins (Li et al. 2014; Gayk et al. 2018). Hemoglobin duplications could also increase oxygen concentration in the body through increased gene dosage or through subfunctionalization (Kuzmin, Taylor, and Boone 2021). The vertebrate hemoglobin α2β2 heterotetramer is a textbook example of this mechanism. Indeed, the specialisation of α- and β-hemoglobin enabled the hemoglobin α2β2 heterotetramer to bind and release oxygen molecules in a cooperative manner (the more oxygen bound to hemoglobin subunits, the higher their affinity for oxygen) (Pillai et al. 2020). Thus, binding is increasingly more efficient in pulmonary capillaries, which have a high oxygen concentration.

We also identified four expansions in the Lactosylceramide 4-alpha-galactosyltransferase (*A4GALT*) family. This enzyme catalyses the production of globotriaosylceramide, lactosylceramide and galactocerebroside, which are glycosphingolipids found in cell membranes and have functions in the immune system (Lingwood 2011; Wieland

Brown et al. 2013). Notably, genes involved in sphingolipid metabolism have been found to be positively selected in penguins (Pirri et al. 2021). In particular, the lactosylceramide and galactocerebroside catabolic enzyme, galactosylceramidase, was positively selected in the ancestor of the Adélie and emperor penguin (Li et al. 2014). Although glycosphingolipids comprise lipid molecules, they are not particularly involved in lipid storage or thermoregulation (Lingwood 2011). Nevertheless, their role in maintaining cell membrane fluidity in response to cold has been proposed in penguins (Pirri et al. 2021; Chattopadhyay and Jagannadham 2001).

A brief diagnostic of these two families using Matreex (Fig. 2) revealed that most duplications in diving clades seemed correctly placed as they did not imply too many losses in the resulting subfamilies. Nonetheless, we can mention three likely spurious hemoglobin and *A4GALT* subfamilies, each implying three losses out of four species. Moreover, the median sequence length of both families almost matched the one of their human ortholog, which most likely rules out the presence of fragments that could be misinterpreted as paralogs (Supp. Table 2). Moreover, these gene trees would be particularly interesting to test for positive selection in branches following these duplications to search for signals of neofunctionalization (Kuzmin, Taylor, and Boone 2021).

Loss of flight correlates with hundreds of convergent gene family contractions

Birds have lost the ability to fly on many occasions across at least 26 bird families (Roff 1994). In particular, the paraphyly of flightless ratites (ostriches, kiwis, rheas and cassowaries) with respect to the flying tinamous was recently established, thus implying convergent loss of flight (Sackton et al. 2019). Some phenotypes associated with the loss of flight include reduced wings, modified feathers and larger body sizes (Sackton et al. 2019; Burga et al. 2017).

We found 689 gene families with at least four (out of eight) convergent contractions in flightless clades, which is 1.2 to 3 times more than expected (Fig. 1). Notably, the “Negative regulation of BMP signalling pathway” was the fourth most significantly enriched GO term among these families (Supp. Table 3) and concerned 13 families. Bone morphogenetic proteins (BMP) are signalling proteins involved in various processes from embryonic development to adult homeostasis (*reviewed in* [Katagiri and Watabe 2016]). Specifically, they play a role in forelimb development and feather patterning (Newton and Smith 2021; Ho et al. 2019).

Moreover, convergent accelerated evolution of non-coding elements was identified near BMP genes in limbless lizards (Roscito et al. 2022) and near a BMP repressor in flightless ratites (Sackton and Clark 2019).

The degree of wing bone modifications varies across flightless species. Penguins have stiff wing joints for diving (Raikow, Bicanovsky, and Bledsoe 1988), while kiwis and cassowaries have a single wing finger remaining (Newton and Smith 2021). By contrast, rheas, ostriches, the galapagos cormorant and *Zapornia atra* have kept three wing fingers, although with reduced bones (Newton and Smith 2021; Burga et al. 2017; Gaspar, Gibb, and Trewick 2020). Among the 13 families involved in "Negative regulation of BMP signalling pathway" (BMP repressors), the flightless clades with the highest number of losses were penguins (13), kiwis (12), the ostrich (11) and cassowaries (10, Fig. 3). Except for the ostrich, these clades display the largest wing bone modifications. By contrast, rheas, *Mesitornis unicolor*, the galapagos cormorant and *Zapornia atra* displayed 9, 9, 7 and 5 losses, respectively (Fig. 3).

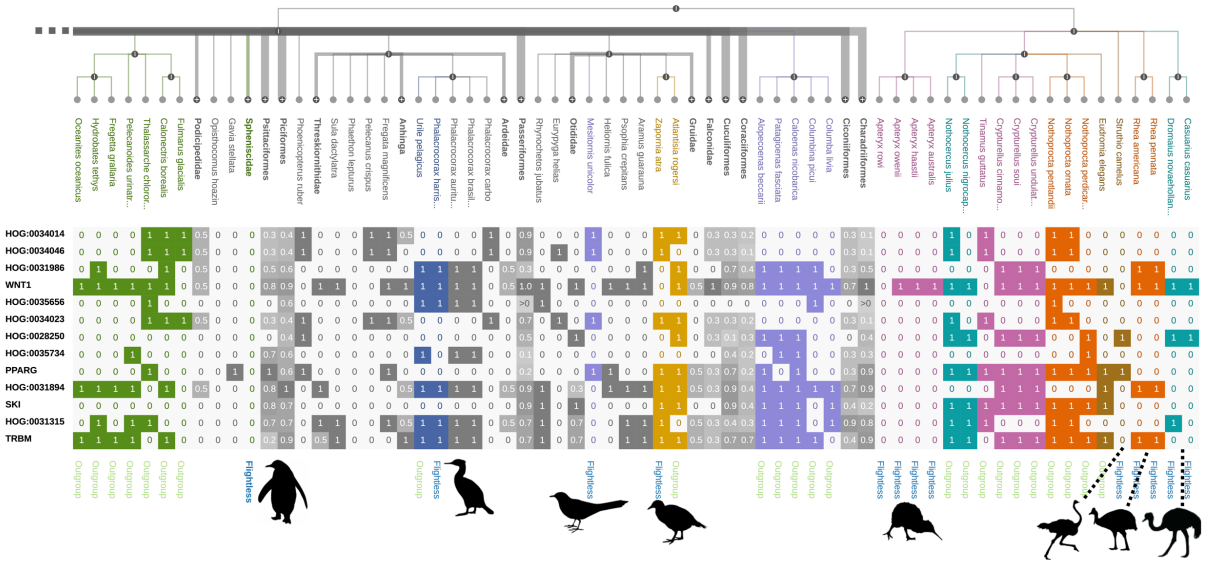


Figure 3. Convergently contracted families associated with loss of flight and involved in "Negative regulation of BMP signalling pathway". Figure created using Matreex, where each row is the presence-absence vector of one family sorted taxonomically (the species tree is displayed on top). Transitions to diving are annotated with "Flightless", PhyloPics and with a unique colour (shared with outgroups).

However, we also observed many losses of these BMP repressors in flying clades (Fig. 3). Thus, we hypothesised that our fast orthology inference approach missed short or fast-evolving genes. To briefly assess the extent of this potential methodological artefact, we took a closer look at the four most conserved families of BMP repressors in tinamous (the ratite

ingroup). While the protein sequences in these families are not particularly short (medians: 354, 104, 715 and 540), half of the missing proteins in the flightless clades were recovered with a BLASTP against the NCBI non-redundant database (Supp. Fig. 1). Although this suggests incomplete orthologous groups, many BMP repressors are still missing from flightless proteomes. Thus, the tendency of flightless species to undergo, on average, more BMP repressor losses remains. A systematic and sensitive search for these genes in the genomes of flightless species would be the next step to validate this result.

We also identified the bone morphogenetic protein 2 (BMP2) annotated with the enriched GO term "negative regulation of Wnt signalling pathway involved in heart development" (Supp. Table 3). BMP2 is a bone morphogenetic protein with an established role in osteoblast differentiation and bone formation (Katagiri and Watabe 2016). Moreover, BMP2 and BMP4 are the two BMPs identified near convergently evolving non-coding elements in limbless lizards (Roscito et al. 2022). However, only one loss was validated with BLAST (Supp. Fig. 1).

While diagnosing the four families of BMP repressors most conserved in timanous, we identified one of them to be the Wnt Family Member 1 (WNT1) with three BLAST-validated losses (Supp. Fig. 1). Like BMP proteins, Wnt proteins are developmental regulators ((Nusse 2012) for a review) involved in forelimb development and feather patterning (Newton and Smith 2021; Ho et al. 2019). Moreover, convergently evolving non-coding elements have been found near Wnt signalling genes in flightless ratites (Sackton et al. 2019). Specifically, Wnt1 plays a role in osteoblast function, bone development and bone homeostasis (Keupp et al. 2013; Laine et al. 2013). Finally, the loss of flight in galapagos cormorants has been linked to alterations in cilium genes responsible for limb shortening in human ciliopathies (Burga et al. 2017). Here, we identified the cyclin-O family required for "multi-ciliated epithelial cell differentiation" (Supp. Table 3) with three BLAST-validated contractions (Supp. Fig. 1).

Transition to frugivory or nectarivory correlates with convergent duplications in families involved in kidney development and cellular respiration

We observed repeated transitions to sugar-based diets (nectarivory or frugivory) (Supp. Fig. 2) and hypothesised correlated convergent evolution in gene families involved in sugar metabolism, taste, olfaction and kidney development (Chen and Zhao 2019; Wang et al. 2020).

We identified 16 families with five to seven convergent expansions out of 13, 2.2 to 2.5 times more than expected (Fig. 1). Notably, two families were involved in “renal system development” (Supp. Table 4), which may be related to a reduced ability to concentrate urines as water preservation is less important when eating fruits than insects. The reduced medullary thickness of kidneys in frugivorous bats testifies to such adaptation (Schondube, Herrera-M, and Martínez del Rio 2001). Moreover, one of these two families was also involved in “hepaticobiliary system development”, which plays a role in the storage of glycogen (the main molecule that stores glucose in the body). Finally, we identified a descendant term of cellular respiration (“ATP synthesis coupled electron transport”) enriched in families with at least three expansions (Supp. Table 5). Adenosine triphosphates (ATPs) are produced with glucose molecules during cellular respiration to provide energy for the body. The three families involved in this process are two NADH dehydrogenase subunits (6 and 4L) and the dnaj homolog subfamily c member 15. However, these results should be taken with caution as the five families discussed here contained particularly short proteins (Supp. Table 2).

Other traits displayed limited evidence for convergent gene repertoire evolution

We tested several other phenotypic and niche transitions: nocturnality, herbivory, invertivore aerial lifestyle (*e.g.* swallows and swifts) and desert habitat. Only the transition to nocturnality and herbivory were associated with unexpected numbers of convergent expansions or contractions (Supp. Table. 2). Furthermore, because most of the enriched functions in their convergently expanded families were neither expected nor significant, we did not dwell on these phenotypes. Briefly, one family with four nocturnal expansions was similar to beta-keratin proteins, which is the main component of feathers (although displaying a median sequence length of 89aa). This could be related to owls' softer feathers used for silent flight (Espíndola-Hernández et al. 2020). We also identified three convergent expansions in a large olfactory receptor family (2325 members) associated with herbivory. This might suggest an increased selection pressure on food selection as plants release toxins for defence.

Conclusion

In this study, we investigated the genomic basis of seven bird phenotypes by correlating gene family evolution with repeated phenotypic transitions. Testing that many phenotypes on a single dataset was possible due to the dense sampling of species as it provided many

independent transitions to various phenotypes. Even so, we only tested a small fraction of convergent bird phenotypes. For example, the transition to the raptor lifestyle (hawks, eagles and owls) has been reported to correlate with expansions in gene families involved in hearing, development and learning (Cho et al. 2019). Moreover, carnivorous mammals have been found repeatedly to lose genes related to glucose homeostasis and xenobiotic receptors (Hecker, Sharma, and Hiller 2019). Finally, evolving the ability to migrate for long distances is also pervasive among birds and is associated with numerous physiological adaptations, including those associated with circadian clocks to decide when to migrate and lipid-metabolism to sustain long flights (Fudickar, Jahn, and Ketterson 2021). In particular, the lack of resolution in the NCBI taxonomy prevented us from testing these phenotypes because too many transition branches were missing from this tree.

When testing multiple phenotypes, the chance to find a plausible biological story increases. Thus, we attempted to mitigate confirmation biases and biological storytelling (Pavlidis et al. 2012) by gathering literature knowledge *before* analysing the results. Although this approach could have been even more systematic and precise (e.g. at the level of GO terms and families), we expect it reduced the proportion of false positives.

To compute gene families and subfamilies (HOGs) for the 363 proteomes produced from the family-phase of the B10K project, we introduced a prototypic pipeline (fast-OMA) tuned for densely sampled clades. The key idea lies in restricting all-against-all pairwise alignments within precomputed homologous groups obtained with fast placements of protein sequences in reference HOGs.

Overall, our results highlight the potential of comparative evolutionary genomics for phenotype to genotype associations (S. D. Smith et al. 2020; Nagy et al. 2020), to unveil the function of genomic loci in an era of Big data genomics (Stephens et al. 2015).

Methods

Inferring bird HOGs

The 363 B10K proteomes (Feng et al. 2020) were assigned to reference HOGs from the OMA database (2021.04 release) using OMAmer (version 0.2.1, sensitive option) (Altenhoff et al. 2021; Rossier et al. 2021). In particular, we used every root-HOG with at least five members or with a species conservation larger than 50% and set the OMAmer threshold to 0

(*same rationale as in Chapter 4*). Homologous groups formed by protein sequences assigned to HOGs at the level of *Archelosauria* or to root-HOGs defined in *Archelosauria* descendants consisted of the input to reconstruct bird HOGs (*i.e.* including the 363 bird genomes) (Supp. Figure 3). Briefly, this step was performed by adapting the OMA pipeline to run all-against-all alignments and the GETHOGs algorithm for each homologous group separately (Altenhoff et al. 2013). Adrian Altenhoff and Alex Warwick Vesztrocy did these adaptations.

Mapping species names

NCBI species names including the 363 B10K species do not match the ones found in the AVONET dataset and the BirdTree phylogeny (Feng et al. 2020; Tobias et al. 2022; Jetz et al. 2012). Thus, to enable comparing changes in gene family sizes with phenotypic transitions annotated on this species tree, mapping species names between these two datasets was a prerequisite, at least for the 363 B10K species. First, to maximise the chance of finding the corresponding NCBI species, eBird and BirdLife synonyms were gathered for each BirdTree name using the AVONET dataset. Then, ete3 and RANT were used to map these species names to NCBI taxonomic identifiers (taxids) (Huerta-Cepas, Serra, and Bork 2016; Hosner et al. 2022). This approach recovered NCBI taxids for 9135/9993 BirdTree species and, most importantly, 362/363 B10K species. We identified the last species, *Phylloscopus sibilatrix*, to be *Rhadina sibilatrix* (taxid 2585818).

Ancestral state reconstructions and bird phylogeny

Ancestral state reconstructions (ASRs) of ecological niche variables (trophic niche, habitat) from the AVONET dataset (Tobias et al. 2022) were used to automate the identification of some phenotypic transitions (frugivory and nectarivory, herbivory, adaptation to aridity) (*see next section*). ASRs were computed with PastML (v. 1.9.34, MPPA + F81) on a time-calibrated phylogeny from BirdTree.org (first out of 10’000 trees from a Bayesian posterior distribution based on the Hackett backbone) (Ishikawa et al. 2019; Jetz et al. 2012). While ASRs used phenotypic information for the 9993 bird species of AVONET and BirdTree, a pruned version of this tree including only B10K species was used to infer phenotypic transitions (*see next section*). ASRs were visualised with ITOL (*e.g.* Supp. Fig. 2) (Letunic and Bork 2021).

Selecting phenotypic transitions

To identify the duplications and losses underlying the emergence of a given phenotype, we identified the branches on the species tree which underwent phenotypic transitions by reviewing the literature and using ASRs of AVONET ecological variables (Tobias et al. 2022). Diving, nocturnal and flightless clades were collected manually. In particular, species classified in the "aquatic dive" foraging niche from (Pigot et al. 2020) were used to identify most diving species. Other phenotypic transitions (frugivory and nectarivory, herbivory, invertivore aerial lifestyle and adaptation to aridity) were computed automatically based on ASRs. Specifically, branches with a phenotypic probability >0.5 and <0.5 in the parent branch were selected. Because this process was performed on the BirdTree and that HOGs were computed on the NCBI tree, only transitions nodes that mapped on the NCBI tree (with the same extant species) were kept. Because expansions and contractions were computed by contrasting duplications and losses between these branches undergoing phenotypic transition and outgroup branches, the NCBI branches with the best correspondence (most shared extant species) to the sister branches of the transition branch in the BirdTree were selected as outgroup branches. Visualisations of phenotypic transitions and outgroups on the NCBI taxonomy (Fig. 1) were computed with ITOL (Letunic and Bork 2021).

Inference of convergent expansions and contractions

To assess the extent of correlated evolution between phenotypes and gene family evolution, we counted the number of expansions and contractions in branches identified with a phenotypic transition in each bird family (root-HOG). An expansion was counted in a branch of the species tree when the difference between the number of duplications and losses in corresponding HOG branches was both larger than 0 (thus resulting in more ancestral copies) and larger than in the outgroup branch (thus excluding globally expanding families). Contractions were defined with reverse criteria. The same binomial models as in chapter 4 were used to generate expected distributions of expansion and contraction numbers across bird families.

GO enrichment tests

They were performed as in chapter 4.

Gene family functions

Gene family names were obtained by blasting the longest chicken representative (or longest family sequence in absence of a chicken one) against the NCBI non-redundant database. Similarly, the closest OMA HOG was obtained by searching the same sequence in the OMA browser (2021.12 release).

Validating gene losses

As a first step to validate gene losses associated with the loss of flight, we blasted (v. 2.13.0) the protein sequences of one tinamou species (*Nothoprocta pentlandii*) against the NCBI non-redundant database extending the number of reported hits to 5000 (and otherwise default parameters) (Altschul et al. 1990; NCBI Resource Coordinators 2018). Then, we filtered the BLASTP results with the taxa predicted to be lost.

Acknowledgements

We thank Natasha Glover for proof-reading the manuscript, Natalia Zajac for helping scan the literature of convergent bird phenotypes and Reto Burri for ideas of phenotypes.

References

Alexa, and Rahnenführer. 2009. "Gene Set Enrichment Analysis with topGO." *Bioconductor Improv*. https://mirrors.nju.edu.cn/bioconductor/3.2/bioc/vignettes/topGO/inst/doc/topGO.pdf.

Altenhoff, Adrian M., Clément-Marie Train, Kimberly J. Gilbert, Ishita Mediratta, Tarcisio Mendes de Farias, David Moi, Yannis Nevers, et al. 2021. "OMA Orthology in 2021: Website Overhaul, Conserved Isoforms, Ancestral Gene Order and More." *Nucleic Acids Research* 49 (D1): D373–79.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10.

Bingman, Verner P., and Emily M. Ewry. 2020. "On a Search for a Neurogenomics of Cognitive Processes Supporting Avian Migration and Navigation." *Integrative and Comparative Biology* 60 (4): 967–75.

Blount, Zachary D., Richard E. Lenski, and Jonathan B. Losos. 2018. "Contingency and Determinism in Evolution: Replaying Life's Tape." *Science* 362 (6415). https://doi.org/10.1126/science.aam5979.

Borges, Rui, João Fonseca, Cidália Gomes, Warren E. Johnson, Stephen J. O'Brien, Guojie Zhang, M. Thomas P. Gilbert, Erich D. Jarvis, and Agostinho Antunes. 2019. "Avian Binocularity and Adaptation to Nocturnal Environments: Genomic Insights from a Highly Derived Visual Phenotype." *Genome Biology and Evolution* 11 (8): 2244–55.

Bravo, Gustavo A., C. Jonathan Schmitt, and Scott V. Edwards. 2021. "What Have We Learned from the First 500 Avian Genomes?" *Annual Review of Ecology, Evolution, and Systematics* 52 (1): 611–39.

Burga, Alejandro, Weiguang Wang, Eyal Ben-David, Paul C. Wolf, Andrew M. Ramey, Claudio Verdugo, Karen Lyons, Patricia G. Parker, and Leonid Kruglyak. 2017. "A Genetic Signature of the

Evolution of Loss of Flight in the Galapagos Cormorant." *Science* 356 (6341). https://doi.org/10.1126/science.aal3345.

Chattopadhyay, M., and M. Jagannadham. 2001. "Maintenance of Membrane Fluidity in Antarctic Bacteria." *Polar Biology* 24 (5): 386–88.

Chen, Yan-Hong, and Huabin Zhao. 2019. "Evolution of Digestive Enzymes and Dietary Diversification in Birds." *PeerJ* 7 (April): e6840.

Cho, Yun Sung, Je Hoon Jun, Jung A. Kim, Hak-Min Kim, Oksung Chung, Seung-Gu Kang, Jin-Young Park, et al. 2019. "Raptor Genomes Reveal Evolutionary Signatures of Predatory and Nocturnal Lifestyles." *Genome Biology* 20 (1): 181.

Delsuc, Frédéric, Hervé Philippe, Georgia Tsagkogeorga, Paul Simion, Marie-Ka Tilak, Xavier Turon, Susanna López-Legentil, Jacques Piette, Patrick Lemaire, and Emmanuel J. P. Douzery. 2018. "A Phylogenomic Framework and Timescale for Comparative Studies of Tunicates." *BMC Biology* 16 (1): 39.

Espíndola-Hernández, Pamela, Jakob C. Mueller, Martina Carrete, Stefan Boerno, and Bart Kempenaers. 2020. "Genomic Evidence for Sensorial Adaptations to a Nocturnal Predatory Lifestyle in Owls." *Genome Biology and Evolution* 12 (10): 1895–1908.

Feng, Shaohong, Josefin Stiller, Yuan Deng, Joel Armstrong, Qi Fang, Andrew Hart Reeve, Duo Xie, et al. 2020. "Dense Sampling of Bird Diversity Increases Power of Comparative Genomics." *Nature* 587 (7833): 252–57.

Forslund, Kristoffer, Cecile Pereira, Salvador Capella-Gutierrez, Alan Sousa da Silva, Adrian Altenhoff, Jaime Huerta-Cepas, Matthieu Muffato, et al. 2018. "Gearing up to Handle the Mosaic Nature of Life in the Quest for Orthologs." *Bioinformatics* 34 (2): 323–29.

Fudickar, Adam M., Alex E. Jahn, and Ellen D. Ketterson. 2021. "Animal Migration: An Overview of One of Nature's Great Spectacles." *Annual Review of Ecology, Evolution, and Systematics* 52 (1): 479–97.

Gaspar, Julien, Gillian C. Gibb, and Steve A. Trewick. 2020. "Convergent Morphological Responses to Loss of Flight in Rails (Aves: Rallidae)." *Ecology and Evolution* 10 (13): 6186–6207.

Gayk, Zach G., Diana Le Duc, Jeffrey Horn, and Alec R. Lindsay. 2018. "Genomic Insights into Natural Selection in the Common Loon (Gavia Immer): Evidence for Aquatic Adaptation." *BMC Evolutionary Biology* 18 (1): 64.

Godoy-Vitorino, Filipa, Katherine C. Goldfarb, Ulas Karaoz, Sara Leal, Maria A. Garcia-Amado, Philip Hugenholtz, Susannah G. Tringe, Eoin L. Brodie, and Maria Gloria Dominguez-Bello. 2012. "Comparative Analyses of Foregut and Hindgut Bacterial Communities in Hoatzins and Cows." *The ISME Journal* 6 (3): 531–41.

Goeman, Jelle J., and Ulrich Mansmann. 2008. "Multiple Testing on the Directed Acyclic Graph of Gene Ontology." *Bioinformatics* 24 (4): 537–44.

Hecker, Nikolai, Virag Sharma, and Michael Hiller. 2019. "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* 116 (8): 3036–41.

Hosner, Peter A., Min Zhao, Rebecca T. Kimball, Edward L. Braun, and J. Gordon Burleigh. 2022. "Reconciling GenBank Names with Standardized Avian Taxonomies to Improve Linkage between Phylogeny and Phenotype." *bioRxiv*. https://doi.org/10.1101/2022.02.07.479408.

Ho, William K. W., Lucy Freem, Debiao Zhao, Kevin J. Painter, Thomas E. Woolley, Eamonn A. Gaffney, Michael J. McGrew, et al. 2019. "Feather Arrays Are Patterned by Interacting Signalling and Cell Density Waves." *PLoS Biology* 17 (2): e3000132.

Huelsmann, Matthias, Nikolai Hecker, Mark S. Springer, John Gatesy, Virag Sharma, and Michael Hiller. 2019. “Genes Lost during the Transition from Land to Water in Cetaceans Highlight Genomic Changes Associated with Aquatic Adaptations.” *Science Advances* 5 (9): eaaw6671.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. “ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data.” *Molecular Biology and Evolution* 33 (6): 1635–38.

Hung, Jui-Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. 2012. “Gene Set Enrichment Analysis: Performance Evaluation and Usage Guidelines.” *Briefings in Bioinformatics* 13 (3): 281–91.

i5K Consortium. 2013. “The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment.” *The Journal of Heredity* 104 (5): 595–600.

Ishikawa, Sohta A., Anna Zhukova, Wataru Iwasaki, and Olivier Gascuel. 2019. “A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios.” *Molecular Biology and Evolution* 36 (9): 2069–85.

Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann, and A. O. Mooers. 2012. “The Global Diversity of Birds in Space and Time.” *Nature* 491 (7424): 444–48.

Katagiri, Takenobu, and Tetsuro Watabe. 2016. “Bone Morphogenetic Proteins.” *Cold Spring Harbor Perspectives in Biology* 8 (6). https://doi.org/10.1101/cshperspect.a021899.

Keupp, Katharina, Filippo Beleggia, Hülya Kayserili, Aileen M. Barnes, Magdalena Steiner, Oliver Semler, Björn Fischer, et al. 2013. “Mutations in WNT1 Cause Different Forms of Bone Fragility.” *American Journal of Human Genetics* 92 (4): 565–74.

Kooyman, G. L., and P. J. Ponganis. 1998. “The Physiological Basis of Diving to Depth: Birds and Mammals.” *Annual Review of Physiology* 60: 19–32.

Kuzmin, Elena, John S. Taylor, and Charles Boone. 2021. “Retention of Duplicated Genes in Evolution.” *Trends in Genetics: TIG*, July. https://doi.org/10.1016/j.tig.2021.06.016.

Laine, Christine M., Kyu Sang Joeng, Philippe M. Campeau, Riku Kiviranta, Kati Tarkkonen, Monica Grover, James T. Lu, et al. 2013. “WNT1 Mutations in Early-Onset Osteoporosis and Osteogenesis Imperfecta.” *The New England Journal of Medicine* 368 (19): 1809–16.

Le Duc, Diana, Gabriel Renaud, Arunkumar Krishnan, Markus Sällman Almén, Leon Huynen, Sonja J. Prohaska, Matthias Ongyerth, et al. 2015. “Kiwi Genome Provides Insights into Evolution of a Nocturnal Lifestyle.” *Genome Biology* 16 (July): 147.

Le Duc, Diana, and Torsten Schöneberg. 2016. “Adaptation to Nocturnality - Learning from Avian Genomes.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 38 (7): 694–703.

Letunic, Ivica, and Peer Bork. 2021. “Interactive Tree Of Life (iTOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation.” *Nucleic Acids Research* 49 (W1): W293–96.

Lewin, Harris A., Stephen Richards, Erez Lieberman Aiden, Miguel L. Allende, John M. Archibald, Miklós Bálint, Katharine B. Barker, et al. 2022. “The Earth BioGenome Project 2020: Starting the Clock.” *Proceedings of the National Academy of Sciences of the United States of America* 119 (4). https://doi.org/10.1073/pnas.2115635118.

Li, Cai, Yong Zhang, Jianwen Li, Lesheng Kong, Haofu Hu, Hailin Pan, Luohao Xu, et al. 2014. “Two Antarctic Penguin Genomes Reveal Insights into Their Evolutionary History and Molecular Changes Related to the Antarctic Environment.” *GigaScience* 3 (1): 27.

Linard, Benjamin, Ingo Ebersberger, Shawn E. McGlynn, Natasha Glover, Tomohiro Mochizuki, Mateus Patricio, Odile Lecompte, et al. 2021. “Ten Years of Collaborative Progress in the Quest for Orthologs.” *Molecular Biology and Evolution*, April. https://doi.org/10.1093/molbev/msab098.

Lingwood, Clifford A. 2011. “Glycosphingolipid Functions.” *Cold Spring Harbor Perspectives in Biology* 3 (7). https://doi.org/10.1101/cshperspect.a004788.

McKechnie, Andrew E., Ben Smit, Maxine C. Whitfield, Matthew J. Noakes, William A. Talbot, Mateo Garcia, Alexander R. Gerson, and Blair O. Wolf. 2016. “Avian Thermoregulation in the Heat: Evaporative Cooling Capacity in an Archetypal Desert Specialist, Burchell’s Sandgrouse (Pterocles Burchelli).” *The Journal of Experimental Biology* 219 (Pt 14): 2137–44.

Merényi, Zsolt, Arun N. Prasanna, Zheng Wang, Károly Kovács, Botond Hegedüs, Balázs Bálint, Balázs Papp, Jeffrey P. Townsend, and László G. Nagy. 2020. “Unmatched Level of Molecular Convergence among Deeply Divergent Complex Multicellular Fungi.” *Molecular Biology and Evolution* 37 (8): 2228–40.

Merlin, Christine, Samantha E. Iiams, and Aldrin B. Lugena. 2020. “Monarch Butterfly Migration Moving into the Genetic Era.” *Trends in Genetics: TIG* 36 (9): 689–701.

Nagy, László G., Zsolt Merényi, Botond Hegedüs, and Balázs Bálint. 2020. “Novel Phylogenetic Methods Are Needed for Understanding Gene Function in the Era of Mega-Scale Genome Sequencing.” *Nucleic Acids Research* 48 (5): 2209–19.

Nagy, László G., Robin A. Ohm, Gábor M. Kovács, Dimitrios Floudas, Robert Riley, Attila Gácser, Mátyás Sipiczki, et al. 2014. “Latent Homology and Convergent Regulatory Evolution Underlies the Repeated Emergence of Yeasts.” *Nature Communications* 5 (July): 4471.

Nagy, László G., Robert Riley, Philip J. Bergmann, Krisztina Krizsán, Francis M. Martin, Igor V. Grigoriev, Dan Cullen, and David S. Hibbett. 2017. “Genetic Bases of Fungal White Rot Wood Decay Predicted by Phylogenomic Analysis of Correlated Gene-Phenotype Evolution.” *Molecular Biology and Evolution* 34 (1): 35–44.

NCBI Resource Coordinators. 2018. “Database Resources of the National Center for Biotechnology Information.” *Nucleic Acids Research* 46 (D1): D8–13.

Newton, Axel H., and Craig A. Smith. 2021. “Regulation of Vertebrate Forelimb Development and Wing Reduction in the Flightless Emu.” *Developmental Dynamics: An Official Publication of the American Association of Anatomists* 250 (9): 1248–63.

Nusse, Roel. 2012. “Wnt Signaling.” *Cold Spring Harbor Perspectives in Biology* 4 (5). https://doi.org/10.1101/cshperspect.a011163.

Pavlidis, Pavlos, Jeffrey D. Jensen, Wolfgang Stephan, and Alexandros Stamatakis. 2012. “A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans.” *Molecular Biology and Evolution* 29 (10): 3237–48.

Pigot, Alex L., Catherine Sheard, Eliot T. Miller, Tom P. Bregman, Benjamin G. Freeman, Uri Roll, Nathalie Seddon, Christopher H. Trisos, Brian C. Weeks, and Joseph A. Tobias. 2020. “Macroevolutionary Convergence Connects Morphological Form to Ecological Function in Birds.” *Nature Ecology & Evolution* 4 (2): 230–39.

Pillai, Arvind S., Shane A. Chandler, Yang Liu, Anthony V. Signore, Carlos R. Cortez-Romero, Justin L. P. Benesch, Arthur Laganowsky, Jay F. Storz, Georg K. A. Hochberg, and Joseph W. Thornton. 2020. “Origin of Complexity in Haemoglobin Evolution.” *Nature*, May, 1–6.

Pirri, Federica, Lino Ometto, Silvia Fuselli, Flávia A. N. Fernandes, Lorena Ancona, Céline Le Bohec, Lorenzo Zane, and Emiliano Trucchi. 2021. “Selection-Driven Adaptation to the Extreme Antarctic Environment in the Emperor Penguin.” *bioRxiv*. https://doi.org/10.1101/2021.12.14.471946.

Potter, Joshua H. T., Rosie Drinkwater, Kalina T. J. Davies, Nicolas Nesi, Marisa C. W. Lim, Laurel R. Yohe, Hai Chi, et al. 2021. “Nectar-Feeding Bats and Birds Show Parallel Molecular Adaptations in Sugar Metabolism Enzymes.” *Current Biology: CB* 31 (20): 4667–74.e6.

Raikow, Robert J., Lesley Bicanovsky, and Anthony H. Bledsoe. 1988. “Forelimb Joint Mobility and the Evolution of Wing-Propelled Diving in Birds.” *The Auk* 105 (3): 446–51.

Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. “Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species.” *Nature* 592 (7856): 737–46.

Rocha, Joana L., Raquel Godinho, José C. Brito, and Rasmus Nielsen. 2021. “Life in Deserts: The Genetic Basis of Mammalian Desert Adaptation.” *Trends in Ecology & Evolution* 36 (7): 637–50.

Roff, Derek A. 1994. “The Evolution of Flightlessness: Is History Important?” *Evolutionary Ecology* 8 (6): 639–57.

Roscito, Juliana Gusson, Katrin Sameith, Bogdan Mikhailovich Kirilenko, Nikolai Hecker, Sylke Winkler, Andreas Dahl, Miguel Trefaut Rodrigues, and Michael Hiller. 2022. “Convergent and Lineage-Specific Genomic Differences in Limb Regulatory Elements in Limbless Reptile Lineages.” *Cell Reports* 38 (3): 110280.

Rosenblum, Erica Bree, Christine E. Parent, and Erin E. Brandt. 2014. “The Molecular Basis of Phenotypic Convergence.” *Annual Review of Ecology, Evolution, and Systematics* 45 (1): 203–26.

Rossier, Victor, Alex Warwick Vesztrocy, Marc Robinson-Rechavi, and Christophe Dessimoz. 2021. “OMAmer: Tree-Driven and Alignment-Free Protein Assignment to Subfamilies Outperforms Closest Sequence Approaches.” *Bioinformatics* , March. https://doi.org/10.1093/bioinformatics/btab219.

Sackton, Timothy B., and Nathan Clark. 2019. “Convergent Evolution in the Genomics Era: New Insights and Directions.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 374 (1777): 20190102.

Sackton, Timothy B., Phil Grayson, Alison Cloutier, Zhirui Hu, Jun S. Liu, Nicole E. Wheeler, Paul P. Gardner, et al. 2019. “Convergent Regulatory Evolution and Loss of Flight in Paleognathous Birds.” *Science* 364 (6435): 74–78.

Schoenjahn, Jonny, Chris R. Pavey, and Gimme H. Walter. 2022. “Low Activity Levels Are an Adaptation to Desert-Living in the Grey Falcon, an Endotherm That Specializes in Pursuing Highly Mobile Prey.” *Journal of Thermal Biology* 103 (January): 103108.

Schondube, J. E., L. G. Herrera-M, and C. Martínez del Rio. 2001. “Diet and the Evolution of Digestion and Renal Function in Phyllostomid Bats.” *Zoology* 104 (1): 59–73.

Seppey, Mathieu, Panagiotis Ioannidis, Brent C. Emerson, Camille Pitteloud, Marc Robinson-Rechavi, Julien Roux, Hermes E. Escalona, et al. 2019. “Genomic Signatures Accompanying the Dietary Shift to Phytophagy in Polyphagan Beetles.” *Genome Biology* 20 (1): 98.

Sheard, Catherine, Montague H. C. Neate-Clegg, Nico Alioravainen, Samuel E. I. Jones, Claire Vincent, Hannah E. A. MacGregor, Tom P. Bregman, Santiago Claramunt, and Joseph A. Tobias. 2020. “Ecological Drivers of Global Gradients in Avian Dispersal Inferred from Wing Morphology.” *Nature Communications* 11 (1): 2463.

Smith, N. Adam, Krista L. Koeller, Julia A. Clarke, Daniel T. Ksepka, Jonathan S. Mitchell, Ali Nabavizadeh, Ryan C. Ridgley, and Lawrence M. Witmer. 2021. “Convergent Evolution in Dippers (Aves, Cinclidae): The Only Wing-Propelled Diving Songbirds.” *Anatomical Record* , November. https://doi.org/10.1002/ar.24820.

Smith, Stacey D., Matthew W. Pennell, Casey W. Dunn, and Scott V. Edwards. 2020. “Phylogenetics Is the New Genetics (for Most of Biodiversity).” *Trends in Ecology & Evolution* 35 (5): 415–25.

Stephens, Zachary D., Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. 2015. “Big Data: Astronomical or Genomical?” *PLoS Biology* 13 (7): e1002195.

Stiller, Josefin, and Guojie Zhang. 2019. “Comparative Phylogenomics, a Stepping Stone for Bird Biodiversity Studies.” *Diversity* 11 (7): 115.

Tobias, Joseph A., Catherine Sheard, Alex L. Pigot, Adam J. M. Devenish, Jingyi Yang, Ferran Sayol, Montague H. C. Neate-Clegg, et al. 2022. “AVONET: Morphological, Ecological and Geographical Data for All Birds.” *Ecology Letters* 25 (3): 581–97.

Toda, Yasuka, Meng-Ching Ko, Qiaoyi Liang, Eliot T. Miller, Alejandro Rico-Guevara, Tomoya Nakagita, Ayano Sakakibara, et al. 2021. “Early Origin of Sweet Perception in the Songbird Radiation.” *Science* 373 (6551): 226–31.

Train, Clément-Marie, Natasha M. Glover, Gaston H. Gonnet, Adrian M. Altenhoff, and Christophe Dessimoz. 2017. “Orthologous Matrix (OMA) Algorithm 2.0: More Robust to Asymmetric Evolutionary Rates and More Scalable Hierarchical Orthologous Group Inference.” *Bioinformatics* 33 (14): i75–82.

Wang, Kai, Shilin Tian, Jorge Galindo-González, Liliana M. Dávalos, Yuzhi Zhang, and Huabin Zhao. 2020. “Molecular Adaptation and Convergent Evolution of Frugivory in Old World and Neotropical Fruit Bats.” *Molecular Ecology* 29 (22): 4366–81.

Wieland Brown, Laura C., Cristina Penaranda, Purna C. Kashyap, Brianna B. Williams, Jon Clardy, Mitchell Kronenberg, Justin L. Sonnenburg, Laurie E. Comstock, Jeffrey A. Bluestone, and Michael A. Fischbach. 2013. “Production of α-Galactosylceramide by a Prominent Member of the Human Gut Microbiota.” *PLoS Biology* 11 (7): e1001610.

Yusuf, Leeban, Matthew C. Heatley, Joseph P. G. Palmer, Henry J. Barton, Christopher R. Cooney, and Toni I. Gossmann. 2020. “Noncoding Regions Underpin Avian Bill Shape Diversification at Macroevolutionary Scales.” *Genome Research* 30 (4): 553–65.

Zahn-Zabal, Monique, Christophe Dessimoz, and Natasha M. Glover. 2020. “Identifying Orthologs with OMA: A Primer.” *F1000Research* 9 (27): 27.

Supplementary material

Supplementary tables

Supp. Table 1. Hypothesis table.

Phenotypic or ecological niche transition	Phenotypic/physiological (top) and molecular (bottom) a priori knowledge	biological functions/processes and gene families hypothesised to undergo molecular changes +: expansions predicted -: loss/contraction predicted
Diving	Penguins have developed scalelike feathers for thermoregulation and waterproof propriety, specific visual sensitivity and eye lens for underwater predation (Li et al. 2014). To overcome buoyancy, diving birds increase their body density with wettable feathers and solid bones (Gayk et al. 2018). Goon adaptations to diving include compensations in oxygen and energy metabolism (dive-induced local hypoxia) and elevated solute exchange (Gayk et al. 2018).	Feather Vision Lipid secretion Bone development Oxygen metabolism
	Penguins have many keratinocyte beta-keratin genes (but higher than other aquatic birds) and two other feather-related genes under positive selection (EVPL, DSG1).	Feather keratins (+) Energy metabolism

	Pseudogenization of Rh2 in penguins and positive selection in genes involved in phototransduction and visual perception (CNGB1, MYO3A, UACA, CRB1, CRY2, MYO3B) (Li et al. 2014). Positively selected genes in loons associated with solute exchange and ATP metabolism. GNB1 potentially involved in low-light signal transduction and HMOX1 involved in oxygen respiration (Gayk et al. 2018). More: Transition to water of cetaceans (Huelsmann et al. 2019).	Correlation with nocturnality
Wing-propelled diving	Moreover, penguins have stiff wing joints and reduced distal wing musculature (Li et al. 2014).	Forelimb development
Leg-propelled diving	Locomotion in foot-propelled aquatic birds associated with hindlimb and pelvic girdle morphology. Webbing or keratin lobes around toes. Goons exhibit posteriorly positioned feet for foot-propelled diving (Gayk et al. 2018).	Hindlimb development
Nectarivory + frugivory	Increased activity of maltase and sucrase in bats associated with transitions to nectarivory and frugivory from insectivory. Olfaction and taste are important to select food (Wang et al. 2020). Reduced relative medullary thickness of kidneys (Wang et al. 2020). Sugar instead of fat and trehalase (sugar in insect haemolymph). Reduced relative medullary thickness of kidney (Chen and Zhao 2019). Hummingbirds detect sugar through modifications to the ancestral savory receptor heterodimer (T1R1-T1R3) (Toda et al. 2021).	Sugar (+), Lipid (-) and trehalose (-) metabolism Olfaction Taste Kidney development Feather colours
	Chitinase genes (CHIAs), the trehalase gene (Treh) and the amylase gene (AMY) are largely correlated with dietary changes in mammals. Alanine–glyoxylate aminotransferase (AGT) may have help for adaptations to diet in birds and mammals (Wang et al. 2020). OR1/3/7 and OR2/13 are lined to frugivory. Three bitter taste receptor genes (Tas2r11, Tas2r18 and Tas2r67) were all lost in obligate frugivorous bats. Carbohydrate metabolism adaptation expected. Frugivorous bat Artibeus lituratus exhibits high insulin sensitivity and elevated glucose tolerance (Wang et al. 2020).	Digestive enzymes Olfactory receptors Bitter taste receptors (-) Insulin/Glucagon

Nectarivory	High sugar intake without diabetes? Behavioural and phenotypic traits for accessing the nectar of tubular flowers includes hovering flight, elongated rostra, and hyper-extensible tongues (Potter et al. 2021).	Glucose metabolism Beak and tongue development
	In nectar–feeding bats and hummingbirds, parallel enrichment of positively selected genes associated with carbohydrate metabolism and chemical stimulus involved in sensory perception of sweet taste (Potter et al. 2021).	Sweet taste perception
Flightlessness	Ratites show forelimb reduction, reduced pectoral muscle mass associated with the absence of the sternal keel, and feather modifications, as well as generally larger body size (Sackton et al. 2019).	Limb development Correlation with large body size
	Convergence in regulatory region evolution near genes associated with sequence-specific DNA binding, limb morphogenesis, Wnt signalling, and regulation of epithelial cell proliferation (Sackton et al. 2019). In flightless galapagos cormorant, deletion of a regulatory domain of a transcription factor involved in limb growth in chicken (Burga et al. 2017). Flightless galapagos cormorant has an unfunctional IFT122 that is associated with small limbs in humans ciliopathies (Burga et al. 2017).	DNA binding limb morphogenesis Wnt signalling regulation of epithelial cell proliferation Cilia (-)
Long-distance migrants	Phenotypes related to distance include, increased appetite and energy storage (fat and glucose), wing morphology (pointedness predictor of migration (Sheard et al. 2020)), alteration of muscle fibers, metabolism and oxygen delivery, as well as reproductive quiescence and reduced fear of the unknown (Fudickar, Jahn, and Ketterson 2021). Phenotypes related to timing and navigation includes circadian clocks, skylight cues, and magnetic sensing (Merlin, Iiams, and Lugena 2020).	Energy metabolism and storage Wing development Oxygen delivery Dopamin? Circadian clock Correlation with nocturnality
	Gene associated with migration includes Clock and Adcyap1, which are involved in circadian rhythms, VPS13A that could be related in motor control in migratory flight (lysosomal degradation and lipid transfer → remove reactive oxygen species resulting from a prolonged migration?) and LRP8 that could play a role in learning and memory (navigation, homing) (Bingman and Ewry 2020). In butterflies, FBXO45 (E3 ubiquitin ligase complex) is selectively expressed in the nervous system and regulates neurotransmission and collagen type IV α1 could be linked to flight efficiency regulation during long-distance migration (Merlin, Iiams, and Lugena 2020).	Learning and memory Removal of reactive oxygen

Raptor lifestyle	Adaptations for hunting, killing, and/or eating meat: highly developed sensory systems, efficient circulatory and respiratory systems, and exceptional flight capabilities necessary to capture prey (Cho et al. 2019). Like other raptors, owls have cryptic plumage coloration, reversed sexual size dimorphism as well as acute vision and hearing + Forward-looking eyes, claws, and curved beaks (Espíndola-Hernández et al. 2020).	Forelimbs and feathers Circulatory and respiratory system Vision and Hearing (+) Feather colour Facial development Beak and claws Protein and lipid metabolism
	Ancestral branches of the two-three raptor lineages showed an expansion of gene families (RHCE, CENPQ, SFTPA1, TFF2, PARL) associated with sensory perception of sound, regulation of anatomical structure morphogenesis, postsynaptic density and specialisation, and learning functions (Cho et al. 2019).	Morphology (beak, claws, wings, face) Learning
Nocturnality	A nocturnal lifestyle generally involves adaptations related to the sensory system, circadian rhythms, and plumage colour patterns (Espíndola-Hernández et al. 2020). Visual sensitivity can be enhanced or reduced (Le Duc and Schöneberg 2016). In mammals, nocturnality is associated with energetic metabolism optimised for sun radiation-independent body temperature regulation and low energy metabolism (Le Duc et al. 2015).	Vision Hearing (+) Olfaction (+) Tactility (+) Circadian rhythm Feather colour Energy metabolism
	Pseudogenization of Rh2 in owls, adaptive signature in opsin family and visual, non-visual response in vertebrate, 25 eye-development genes coevolving and inactivation of opsin genes and accelerated evolution of genes related to mitochondrial function, and energy expenditure in kiwi (Borges et al. 2019), (Stiller and Zhang 2019). Owls might have evolved a special type of DNA packaging in the retina, similar to what has been found in the rods of nocturnal mice and primates (Espíndola-Hernández et al. 2020).	Opsins Eye development Energy metabolism DNA packaging
Herbivory	Detoxification gene families to neutralise plant secondary compounds includes cytochrome P450 monooxygenases (P450s), carboxylesterases (CEs), UDP-glycosyltransferases (UGTs), and glutathione S-transferases (GSTs). ATP-binding cassette (ABCs) transporters to evacuate these and endopeptidases, such as cysteine (CYSs), and serine (SERs) proteases, as well as more specific enzymes such as glycoside hydrolases (GHs) to break down polysaccharide molecules. Other adaptation to phytophagy includes chemoreceptors and specialised mouthparts (Seppey et al. 2019). Hoatzin exhibits special anatomical adaptation for herbivory: modification of sternum and pectoral	Detoxification (+) Digestive enzymes (+) Taste/Chemoreception (+) Beak morphology Gastrointestinal morphology Correlation with flightlessness

	girdle to accommodate the filled crop, as well as modified gastrointestinal system (Godoy-Vitorino et al. 2012).	
Polar circle habitat	Penguins have developed enhanced thermoregulation with scalelike feathers and improved fat storage. Moreover, seasonal changes of daylight in Antarctica could affect visual and non-visual phototransduction abilities of penguins (Li et al. 2014). Important reorganization of the cardiovascular system in polar bears to adapt to the Arctic environment [51] Tibetan antelope exhibits signals of positive selection and gene-family expansion in genes associated with energy Metabolism → more search about arctic extreme environment!	Feather Lipid metabolism Phototransduction Circadian rhythm Correlation with nocturnality
	In penguins, pseudogenization of OPSP that is involved in circadian rhythm, positive selection in genes related to lipid metabolism (e.g. FASN) and 17 forelimb-related genes harbouring non-neutral penguin-specific amino acid changes (e.g. EVC2 and EVC) (Li et al. 2014).	
Desert habitat (Maintain temperature while preserving water, food and water scarcity)	In mammals, convergent adaptations were found in fat metabolism (to cope with sparse food and water supplies), in insulin signaling/response (e.g. glucose and serum urate transporters) for its role in reducing energy demands during starvation and in retention of water at the kidney with endocrine systems (vasopressin and/or RAAS-mediated osmoregulation) and the arachidonic acid metabolism (convergence with birds!). Also, association with thyroid-induced metabolism, salt metabolism and prevention of high blood pressure (Rocha et al. 2021). Sandgrouse birds show a unique feather morphology and behaviours that allow adult birds to transport water in their belly feathers to their chicks (McKechnie et al. 2016). Sparse feathers (Schoenjahn, Pavey, and Walter 2022).	Fat metabolism Insulin Arachidonic acid metabolism Thyroid-induced metabolism, Salt metabolism Blood pressure Feather
	Camels show differences in copy number of CYP2 that is associated with water reabsorption in the kidney. BMP2, a gene involved in fat-cell differentiation in several tissues (Rocha et al. 2021).	
Invertivore Aerial (e.g. swallows and swifts)	Swallows (family: Hirundinidae) and swifts (family: Apodidae) are both specialised to feed on flying insects, exhibiting similar morphology including long wings and very short legs (Heers and Dial 2015).	Limb development

Supp. Table 2. Result table. In the "GO enrichment" column, numbers in parentheses HOG ids and in the "HOGs" column, the first number is the same HOG id. The bold name is the gene name from the best blast hit (see methods). Closest OMA HOGs and human members (UniProt) were sometimes linked. NrMemberGenes: number of genes in the bird-OMA family. MedianSeqLen: median protein length of the bird-OMA family members. Each trait clade with a convergent event (expansion or contraction) reported.

https://docs.google.com/spreadsheets/d/1bSyWV0Z4qKUce3KpygYCMxxUNtV8HLUuhRmQVZodY-E/edit?usp=sharing

Trait	Event number enrichment	GO enrichment	HOGs
Diving	4 HOGs with >=4 convergent expansions, 1.2 times more than expected	>hydrogen peroxide catabolic process (176010) >glycosphingolipid metabolic process (107793)	176010 lactosylceramide 4-alpha-galactosyltransferase https://omabrowser.org/oma/hog/HOG:B0559855/Sarcopterygii/iham/ https://www.uniprot.org/uniprotkb/Q9NPC4/entry MedianSeqLen: 354 NrMemberGenes: 520 Alcidae Cinclus_mexicanus Spheniscidae Podicipedidae 107793 hemoglobin subunit epsilon https://omabrowser.org/oma/hog/HOG%3AB0569893/iham/ https://www.uniprot.org/uniprotkb/P68871/entry#sequences MedianSeqLen: 147 NrMemberGenes: 1131 Alcidae Cinclus_mexicanus Spheniscidae Gavia_stellata Podicipedidae
Wing-propelled diving	5 HOGs with >=3 convergent expansions, 1.75 times more than expected	>regulation of vesicle-mediated transport (24711) >regulation of lipoprotein particle clearance (24711)	24711 HNRPK https://www.uniprot.org/uniprotkb/P61978/entry https://omabrowser.org/oma/hog/HOG%3AB0597354/iham/ NrMemberGenes: 485 MedianSeqLen: 417 Alcidae Cinclus_mexicanus Pelecanoides_urinatrix
Nocturnal	3 HOGs with >=4 convergent expansions, 2.5 times more than expected	1 / 3 HOG with >0 GO term	155749 beta-keratin-related protein-like https://omabrowser.org/oma/hog/HOG%3AB0533270/iham/ NrMemberGenes: 410 MedianSeqLen: 89 Apteryx Burhinus_bistriatus Strigiformes Cochlearius_cochlearius 206754 Fibrinogen-like protein 1-like protein https://omabrowser.org/oma/hog/HOG%3AB0569197/iham/ NrMemberGenes: 461 MedianSeqLen: 270 Caprimulgimorphae Burhinus_bistriatus Strigiformes Cochlearius_cochlearius
Loss of flight	689 HOGs with 4 to 7 convergent expansions, 1.2 to 3 times more than expected	>negative regulation of cell cycle >negative regulation of BMP signaling pathway ("bone morphogenetic protein (BMP) signaling … patterning of the skeletal system") https://en.wikipedia.org/wiki/Bone_morphogenetic_protein https://anatomypubs.onlinelibrary.wiley.com/doi/full/10.1002/dvdy.288 "Involved in forelimb development"	171253 regulatory factor X-associated protein https://omabrowser.org/oma/hog/HOG%3AB0555188/iham/ https://www.uniprot.org/uniprotkb/O00287/entry NrMemberGenes: 176 MedianSeqLen: 239 Struthio_camelus Rhea Casuariiformes Apteryx Zapornia_atra Spheniscidae Mesitornis_unicolor 147862

		https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000132 “Patterning of feather” >(embryonic eye morphogenesis) >(neuron projection morphogenesis) >negative regulation of Wnt signaling pathway involved in heart development (147862) >multi-ciliated epithelial cell differentiation (107379) >cardiac muscle hypertrophy	BMP2 https://omabrowser.org/oma/hog/HOG%3AB0594354/iham/ Human BMP2, 4, 5, 6, 7 NrMemberGenes: 230 MedianSeqLen: 363 Struthio_camelus Zapornia_atra Spheniscidae Mesitornis_unicolor 107379 cyclin-O https://omabrowser.org/oma/hog/HOG%3AB0573289/iham/ https://www.uniprot.org/uniprotkb/P22674/entry NrMemberGenes: 255 MedianSeqLen: 261 Rhea Zapornia_atra Spheniscidae Mesitornis_unicolor Four BMP families: 145560 WNT1 NrMemberGenes: 309 MedianSeqLen: 354 Struthio_camelus Zapornia_atra Spheniscidae Mesitornis_unicolor 288951 PPARG NrMemberGenes: 132 MedianSeqLen: 104 Rhea Casuariiformes Apteryx Spheniscidae 44761 SKI https://omabrowser.org/oma/hog/HOG%3AB0580337/iham/ https://www.uniprot.org/uniprotkb/P12755/entry NrMemberGenes: 235 MedianSeqLen: 715 Struthio_camelus Rhea Casuariiformes Apteryx Mesitornis_unicolor 74461 Thrombomodulin https://omabrowser.org/oma/hog/HOG%3AB0560018/iham/ NrMemberGenes: 245 MedianSeqLen: 540 Struthio_camelus Casuariiformes Apteryx Spheniscidae Mesitornis_unicolor
Loss of flight	48 HOGs with 3 to 4 convergent expansions, 2.2 to 3.5 times more than expected	*Nothing relevant to my knowledge*	
Frugivory and Nectarivory	16 HOGs with 5 to 7 convergent expansions, 2.2 to 2.5 more than expected	>cytoskeleton organisation (q-value: 0) (104696 139981 155749 226181 280625 37248 88285) >hepaticobiliary system development (280625) > ATP synthesis coupled electron transport (105561) (>non-canonical Wnt signaling pathway (218004)https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4037346/) > renal system development (280625 218004) >nephron (epithelium) development (280625)	280625 PO21 https://omabrowser.org/oma/hog/HOG%3AB0561215/iham/ NrMemberGenes: 445 MedianSeqLen: 130 Trochilidae Steatornis_caripensis Coliidae Calyptomena_viridis Bombycillidae 218004 hypothetical protein NrMemberGenes: 1600 MedianSeqLen: 100 Penelope_pileata Steatornis_caripensis Paradisaea_raggiana Bucerotiformes Trogon_melanurus Psophia_crepitans
Frugivory and Nectarivory	93 HOGs with 3 to 7 convergent expansions, 1.2 to 2.5 more than expected	>ATP synthesis coupled electron transport (105561 173655 19276)	105561 NADH dehydrogenase subunit 6 NrMemberGenes: 640 MedianSeqLen: 78 Trochilidae Calyptomena_viridis Paradisaea_raggiana Bombycillidae Trogon_melanurus Musophagidae 173655 dnaj homolog subfamily c member 15

			NrMemberGenes: 209 MedianSeqLen: 126 Steatornis_caripensis Bucerotiformes Trogon_melanurus Musophagidae 19276 NADH dehydrogenase subunit 4L NrMemberGenes: 773 MedianSeqLen: 98 Trochilidae Trogon_melanurus Psophia_crepitans Musophagidae
Herbivory	11 HOGs with 3 convergent expansions, 5 times more than expected	*Nothing relevant to my knowledge*	53786 olfactory receptor 14J1-like https://omabrowser.org/oma/hog/HOG%3AB0534754/iham/ NrMemberGenes: 2325 MedianSeqLen: 155 Galloanserae Thinocorus_orbignyianus Opisthocomus_hoazin 206754 EW135 NrMemberGenes: 461 MedianSeqLen: 270 Galloanserae Thinocorus_orbignyianus Opisthocomus_hoazin

Supp. Table 3. Flightless >=4 contractions.

GO	Term	StudyCount	Enrichment	Qvalue	Control Qvalue	Pvalue	Control Pvalue
GO:0006357	regulation of transcription by RNA polym...	119	1.59	0.121	0.436	0.000016	0
GO:0045786	negative regulation of cell cycle	17	1.14	0.121	1	0.000018	0.034
GO:0045944	positive regulation of transcription by ...	70	1.63	0.598	1	0.00015	0.707
GO:0030514	negative regulation of BMP signaling pat...	13	4.53	0.598	1	0.00018	0.062
GO:0048048	embryonic eye morphogenesis	8	5	0.966	1	0.00035	0.015
GO:1902895	positive regulation of pri-miRNA transcr...	6	4.72	1	1	0.00148	0.736
GO:0048812	neuron projection morphogenesis	30	1.3	1	1	0.00815	1
GO:0003308	negative regulation of Wnt signaling pat...	1	4.55	1	1	0.19932	0.117
GO:1903251	multi-ciliated epithelial cell different...	1	4	1	1	0.22844	1
GO:0032402	melanosome transport	1	2.5	1	1	0.33476	0.55
GO:0048814	regulation of dendrite morphogenesis	3	1.75	1	1	0.57288	1
GO:0048813	dendrite morphogenesis	6	1.6	1	1	0.65708	1
GO:0010862	positive regulation of pathway-restricte...	1	0.76	1	1	0.7369	1

GO:0009755	hormone-mediated signaling pathway	7	1.29	1	1	1	1
GO:0043367	CD4-positive, alpha-beta T cell differen...	1	0.62	1	1	1	1
GO:0003300	cardiac muscle hypertrophy	4	2.34	1	1	1	1
GO:0032409	regulation of transporter activity	3	0.51	1	1	1	1
GO:0003306	Wnt signaling pathway involved in heart ...	1	2.5	1	1	1	1
GO:0003307	regulation of Wnt signaling pathway invo...	1	4	1	1	1	1
GO:0001773	myeloid dendritic cell activation	1	1.25	1	1	1	1

Supp. Table 4. Frugivory >=5 expansions.

GO	Term	StudyCount	Enrichment	Qvalue	Control Qvalue	Pvalue	Control Pvalue
GO:0007010	cytoskeleton organization	7	4.93	0	1	0.000000031	0.035
GO:0002930	trabecular meshwork development	1	Inf	1	1	0.0029	1
GO:0015074	DNA integration	1	50	1	1	0.0151	1
GO:0007259	receptor signaling pathway via JAK-STAT	2	22.22	1	1	0.0332	0.009
GO:0098609	cell-cell adhesion	3	4.41	1	1	0.282	1
GO:0006820	anion transport	1	0.55	1	1	1	1
GO:0061008	hepaticobiliary system development	1	12.5	1	1	1	1
GO:0042773	ATP synthesis coupled electron transport	1	25	1	1	1	1
GO:0045597	positive regulation of cell differentiat...	1	1.56	1	1	1	0.842
GO:0035567	non-canonical Wnt signaling pathway	1	20	1	1	1	1
GO:0042775	mitochondrial ATP synthesis coupled elec...	1	25	1	1	1	1
GO:0045935	positive regulation of nucleobase-contai...	2	1.18	1	1	1	1
GO:0072001	renal system development	2	6.9	1	1	1	1
GO:0045937	positive regulation of phosphate metabol...	2	2.67	1	1	1	1
GO:0048522	positive regulation of cellular process	5	1.14	1	1	1	1

GO:0048523	negative regulation of cellular process	3	0.78	1	1	1	1
GO:0003205	cardiac chamber development	1	6.67	1	0.308	1	0
GO:0072006	nephron development	1	7.14	1	1	1	1
GO:0072009	nephron epithelium development	1	8.33	1	1	1	1
GO:0006810	transport	3	0.88	1	1	1	1

Supp. Table 5. Frugivory >=3 expansions.

GO	Term	StudyCount	Enrichment	Qvalue	Control Qvalue	Pvalue	Control Pvalue
GO:0015074	DNA integration	4	50	0.015	0.978	1.10E-06	0
GO:0090502	RNA phosphodiester bond hydrolysis, endo...	5	15.62	0.106	1	1.50E-05	0.046
GO:0007010	cytoskeleton organization	12	1.57	0.124	1	2.70E-05	0
GO:0006278	RNA-dependent DNA biosynthetic process	4	13.33	0.745	1	0.00022	0.005
GO:0042773	ATP synthesis coupled electron transport	3	14.29	1	1	0.0093	1
GO:0002930	trabecular meshwork development	1	100	1	1	0.01427	1
GO:0042776	mitochondrial ATP synthesis coupled prot...	1	20	1	1	0.04678	1
GO:1900016	negative regulation of cytokine producti...	1	16.67	1	1	0.05587	1
GO:0030150	protein import into mitochondrial matrix	1	14.29	1	1	0.06488	1
GO:0022904	respiratory electron transport chain	4	14.29	1	1	0.07114	1
GO:0060350	endochondral bone morphogenesis	1	2.7	1	1	0.30905	1
GO:0030155	regulation of cell adhesion	4	1.37	1	1	0.60603	1
GO:0051493	regulation of cytoskeleton organization	3	0.96	1	1	0.6544	0.489
GO:0030154	cell differentiation	13	0.75	1	1	0.70003	1
GO:0006605	protein targeting	2	2.13	1	1	1	1
GO:0042775	mitochondrial ATP synthesis coupled elec...	2	10	1	1	1	1
GO:0030029	actin filament-based process	3	0.7	1	1	1	1

GO:0051495	positive regulation of cytoskeleton orga...	1	1.15	1	1	1	0.585
GO:0090501	RNA phosphodiester bond hydrolysis	5	8.47	1	1	1	1
GO:0034754	cellular hormone metabolic process	2	5.88	1	1	1	0.579

Supplementary figures

BMP2: Struthio_camelus **Zapornia_atra** Spheniscidae Mesitornis_unicolor

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
hypothetical protein FQV18_0015065 [Eudyptula novaehollandiae]	Eudyptula novaehollandiae	676	676	100%	0.0	95.33%	363	KAF1483946.1
hypothetical protein FQV23_0002127 [Spheniscus humboldti]	Spheniscus humboldti	673	673	100%	0.0	95.05%	363	KAF1406203.1
hypothetical protein FQV10_0008254 [Eudyptes schlegeli]	Eudyptes schlegeli	673	673	100%	0.0	95.05%	363	KAF1551302.1
hypothetical protein FQV07_0001227 [Pygoscelis papua]	Pygoscelis papua	657	657	100%	0.0	92.93%	367	KAF1444164.1
hypothetical protein N308_11910 [Struthio camelus australis]	Struthio camelus australis	550	550	77%	0.0	99.29%	281	KFV80794.1
PREDICTED: bone morphogenetic protein 2 [Struthio camelus australis]	Struthio camelus australis	550	550	77%	0.0	99.29%	313	XP_009672296.1
PREDICTED: bone morphogenetic protein 2 [Aptenodytes forsteri]	Aptenodytes forsteri	548	548	85%	0.0	91.35%	330	XP_019328456.1
hypothetical protein AS27_11318 [Aptenodytes forsteri]	Aptenodytes forsteri	545	545	77%	0.0	98.93%	281	KFM09694.1
PREDICTED: bone morphogenetic protein 2 [Pygoscelis adeliae]	Pygoscelis adeliae	543	543	77%	0.0	98.93%	386	XP_009317903.1
hypothetical protein N332_06382 [Mesitornis unicolor]	Mesitornis unicolor	542	542	77%	0.0	98.58%	281	KFQ39983.1
PREDICTED: bone morphogenetic protein 2 [Mesitornis unicolor]	Mesitornis unicolor	540	540	76%	0.0	98.57%	280	XP_010189327.1
Bone morphogenetic protein 4 [Pygoscelis adeliae]	Pygoscelis adeliae	451	451	93%	2e-155	65.50%	350	KFW66317.1

Cyclin-O: **Rhea Zapornia_atra** Spheniscidae **Mesitornis_unicolor**

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Cyclin-O [Eudyptes chrysocome]	Eudyptes chrysocome	373	373	99%	5e-126	70.34%	295	KAF1641707.1
Cyclin-O [Eudyptes pachyrhynchus]	Eudyptes pachyrhynchus	363	363	97%	2e-122	70.53%	287	KAF1595568.1
Cyclin-O [Eudyptes moseleyi]	Eudyptes moseleyi	360	360	93%	3e-121	72.22%	274	KAF1591109.1
Cyclin-O [Eudyptes sclateri]	Eudyptes sclateri	360	360	96%	4e-121	70.25%	284	KAF1521486.1
Cyclin-O [Eudyptes filholi]	Eudyptes filholi	341	341	82%	2e-114	76.01%	248	KAF1638536.1
Cyclin-O [Spheniscus humboldti]	Spheniscus humboldti	320	320	79%	3e-106	75.19%	240	KAF1397111.1
Cyclin-O [Eudyptula novaehollandiae]	Eudyptula novaehollandiae	306	306	67%	8e-101	82.06%	225	KAF1475159.1
Cyclin-O [Eudyptes robustus]	Eudyptes robustus	305	305	67%	3e-100	81.61%	225	KAF1622946.1
Cyclin-O [Spheniscus magellanicus]	Spheniscus magellanicus	304	304	67%	5e-100	81.98%	225	KAF1418567.1
Cyclin-O [Pygoscelis papua]	Pygoscelis papua	289	289	67%	4e-94	78.17%	231	KAF1449014.1
Cyclin-O [Pygoscelis antarcticus]	Pygoscelis antarcticus	288	288	64%	5e-94	81.69%	214	KAF1461190.1

WNT1: **Struthio_camelus Zapornia_atra** Spheniscidae **Mesitornis_unicolor**

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Protein Wnt-1 [Eudyptula minor]	Eudyptula minor	724	724	100%	0.0	98.59%	354	KAF1524834.1
Protein Wnt-1 [Eudyptes schlegeli]	Eudyptes schlegeli	723	723	100%	0.0	98.31%	354	KAF1569387.1
Protein Wnt-1 [Eudyptes robustus]	Eudyptes robustus	710	710	100%	0.0	96.90%	355	KAF1571216.1
Protein Wnt-1 [Pygoscelis papua]	Pygoscelis papua	689	689	94%	0.0	98.81%	335	KAF1674515.1
Protein Wnt-1 [Spheniscus humboldti]	Spheniscus humboldti	687	687	94%	0.0	98.51%	335	KAF1403159.1
Protein Wnt-1 [Megadyptes antipodes antipodes]	Megadyptes antipodes antipodes	687	687	94%	0.0	98.21%	335	KAF1483573.1
PREDICTED: proto-oncogene Wnt-3 [Struthio camelus australis]	Struthio camelus australis	316	316	99%	1e-105	45.63%	355	XP_009683604.1

PPARG: **Rhea** Casuariiformes Apteryx Spheniscidae

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PPARG protein [Casuarius casuarius]	Casuarius casuarius	231	231	99%	1e-71	98.06%	483	NXE52231.1
PREDICTED: peroxisome proliferator-activated receptor gamma [Aptenodytes forsteri]	Aptenodytes forsteri	224	224	99%	3e-69	99.03%	475	XP_009285335.1
PREDICTED: peroxisome proliferator-activated receptor gamma [Apteryx mantelli mantelli]	Apteryx mantelli mantelli	224	224	99%	3e-69	99.03%	475	XP_013803408.1
Peroxisome proliferator-activated receptor gamma [Spheniscus magellanicus]	Spheniscus magellanicus	224	224	99%	3e-69	99.03%	483	KAF1404057.1
Peroxisome proliferator-activated receptor gamma [Pygoscelis antarcticus]	Pygoscelis antarcticus	224	224	99%	4e-69	99.03%	483	KAF1455903.1
Peroxisome proliferator-activated receptor gamma [Eudyptula albosignata]	Eudyptula albosignata	224	224	99%	4e-69	99.03%	483	KAF1542866.1
peroxisome proliferator-activated receptor gamma isoform X2 [Dromaius novaehollandiae]	Dromaius novaehollandiae	223	223	99%	9e-69	98.06%	475	XP_025955336.1
peroxisome proliferator-activated receptor gamma isoform X1 [Dromaius novaehollandiae]	Dromaius novaehollandiae	223	223	99%	1e-68	98.06%	488	XP_025955335.1
PREDICTED: peroxisome proliferator-activated receptor alpha isoform X2 [Apteryx mantelli mantelli]	Apteryx mantelli mantelli	147	147	99%	4e-40	63.81%	391	XP_013803070.1

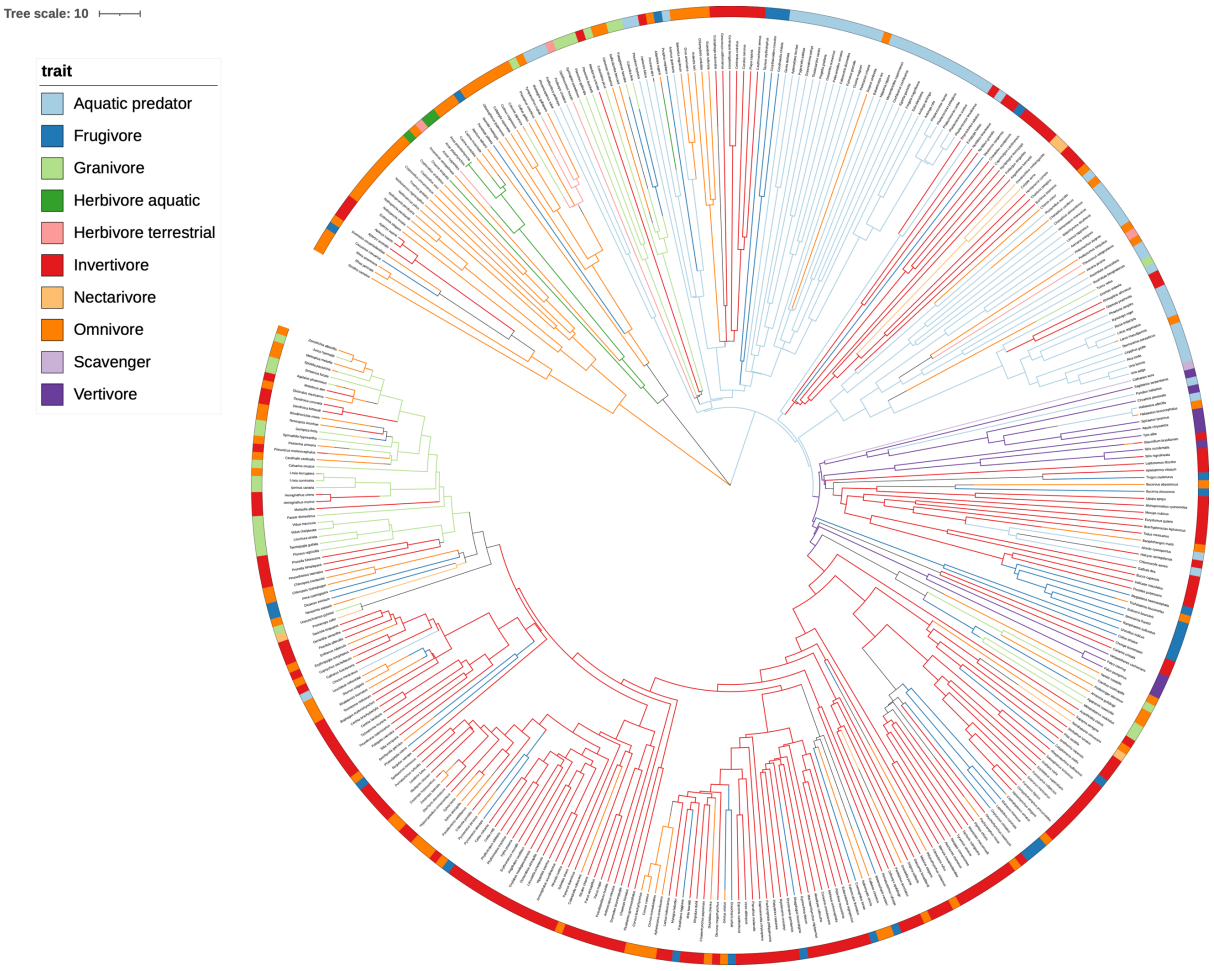
SKI: Struthio_camelus **Rhea** Casuariiformes Apteryx **Mesitornis_unicolor**

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
ski oncogene isoform X3 [Apteryx rowi]	Apteryx rowi	1138	1138	100%	0.0	96.77%	712	XP_025941240.1
ski oncogene isoform X1 [Dromaius novaehollandiae]	Dromaius novaehollandiae	1137	1137	100%	0.0	96.35%	712	XP_025952316.1
ski oncogene isoform X2 [Apteryx rowi]	Apteryx rowi	1127	1127	100%	0.0	94.51%	729	XP_025941239.1
ski oncogene isoform X1 [Apteryx rowi]	Apteryx rowi	1121	1121	100%	0.0	92.11%	748	XP_025941236.1
ski oncogene isoform X2 [Dromaius novaehollandiae]	Dromaius novaehollandiae	1017	1017	91%	0.0	96.46%	665	XP_025952317.1
Ski oncogene [Struthio camelus australis]	Struthio camelus australis	774	774	72%	0.0	94.57%	516	KFV79548.1

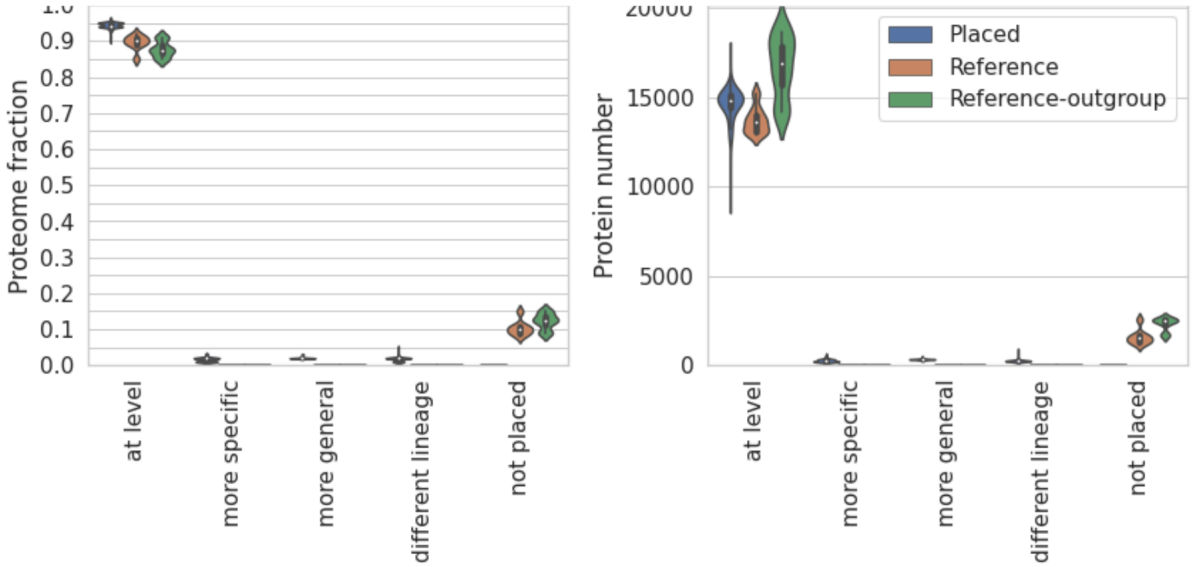
TRBM: **Struthio_camelus** Casuariiformes **Apteryx** Spheniscidae **Mesitornis_unicolor**

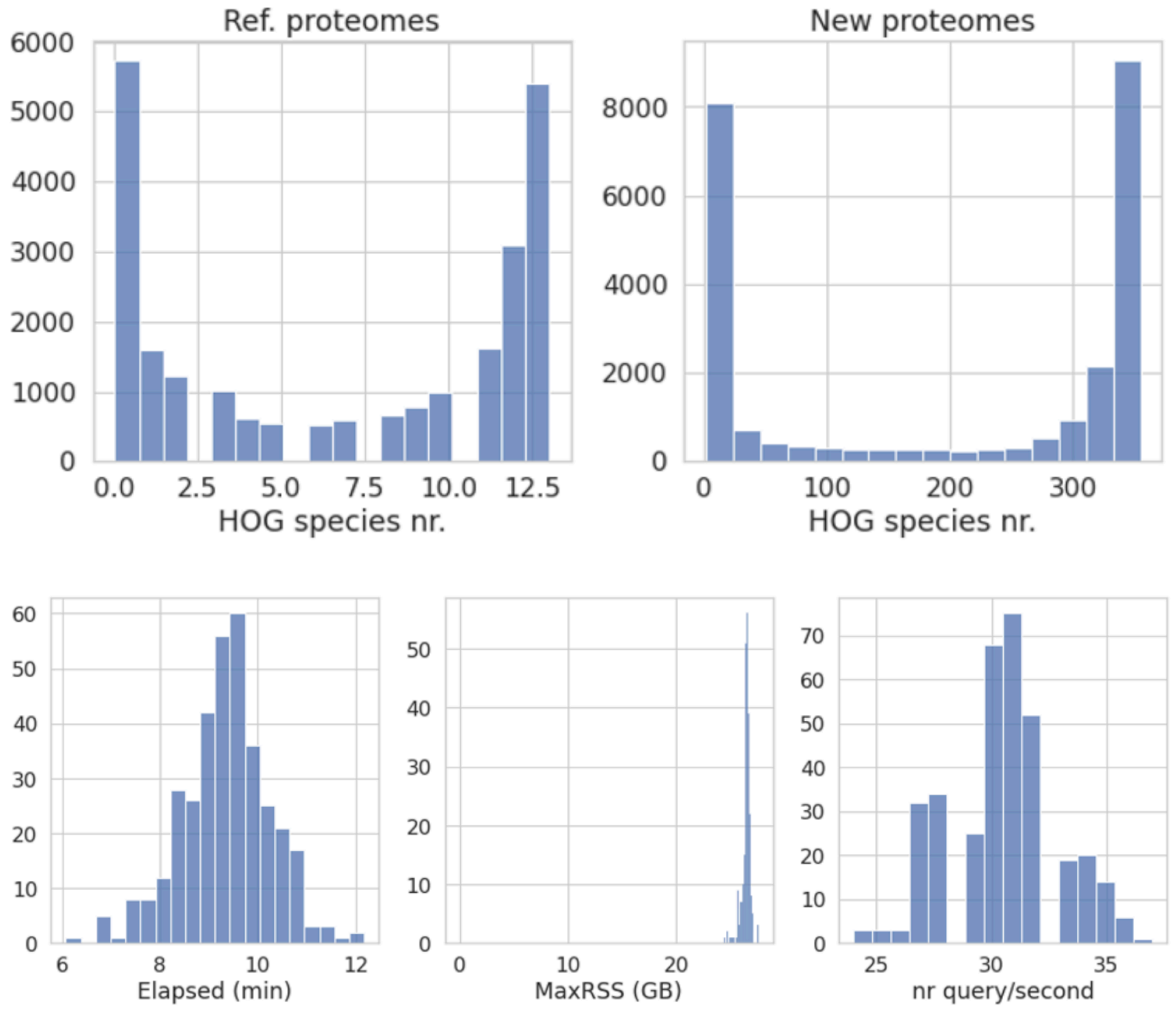
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
uncharacterized protein LOC112986660 [Dromaius novaehollandiae]	Dromaius novaehollandiae	789	1010	94%	0.0	76.05%	1152	XP_025961821.1
Thrombomodulin [Eudyptes sclateri]	Eudyptes sclateri	663	663	93%	0.0	66.22%	528	KAF1511882.1
Thrombomodulin [Eudyptes chrysocome]	Eudyptes chrysocome	663	663	93%	0.0	66.22%	528	KAF1652009.1
Thrombomodulin [Eudyptes schlegeli]	Eudyptes schlegeli	662	662	93%	0.0	66.22%	528	KAF1539978.1
Thrombomodulin [Eudyptes filholi]	Eudyptes filholi	662	662	93%	0.0	66.35%	547	KAF1628102.1
Thrombomodulin [Eudyptula minor]	Eudyptula minor	661	661	93%	0.0	66.22%	526	KAF1559469.1
Thrombomodulin [Eudyptes robustus]	Eudyptes robustus	660	660	93%	0.0	65.83%	528	KAF1641999.1
Thrombomodulin [Spheniscus humboldti]	Spheniscus humboldti	659	659	93%	0.0	66.22%	528	KAF1406188.1
Thrombomodulin [Eudyptula novaehollandiae]	Eudyptula novaehollandiae	659	659	93%	0.0	66.03%	527	KAF1483976.1
Thrombomodulin [Spheniscus magellanicus]	Spheniscus magellanicus	658	658	93%	0.0	66.03%	528	KAF1427262.1
thrombomodulin [Apteryx rowi]	Apteryx rowi	454	511	59%	2e-154	75.88%	321	XP_025915988.1
PREDICTED: thrombomodulin [Struthio camelus australis]	Struthio camelus australis	358	408	59%	7e-117	70.45%	290	XP_009671459.1

Supp. Figure 1. BLAST results. Flightless clades missing the genes are in bold.



Supp. Figure 2.





Supp. Figure 3. Results from OMAmer placement used to infer HOGs with OMA. (Top) 90-95% of bird proteins (~15K) were placed in reference bird-HOGs. By contrast, reference bird-species have 90% of their proteins in bird-HOGs. (Middle) The number of species in resulting homologous groups peaked at 363. (Bottom) Placing one bird proteome in the whole OMA database took a maximum of 12 min and 30GB of RAM.

Chapter 6

Discussion

Discussion

Comparative genomics is a powerful approach to study evolution and discover the genetic basis of phenotypes. However, the recent deluge of next generation sequencing (NGS) data has turned genomics into a Big data discipline (Stephens et al. 2015), thus fundamentally challenging comparative genomics methods, in particular the ones to infer orthologs and paralogs. On the other hand, the increasing number of sequenced genomes offers new opportunities for discovery. Thus, in the first half of this thesis, I developed two comparative genomics methods to cope with some aspects of the *velocity*, *volume* and *variety* property of Big data. Then, in the second half, I capitalised on these new developments to study two biological systems that benefit particularly from increasing numbers of genomes. In this section, I first frame my methodological contributions within the 3Vs of Big data before reflecting on the limitations and perspectives of the two application chapters.

Velocity

In Big data genomics, velocity mainly refers to the rate of data generation (Navarro et al. 2019). In chapter 1, approaches that map new sequences to reference gene families were identified as the most promising strategy to scale-up orthology inference to the increasing rate of genome sequencing. However, state-of-the-art mapping approaches were found to be either phylogenetically-aware (thus precise) and scalable but slow (*e.g.* SHOOT [Emms and Kelly 2022]) or fast and scalable but not precise (*e.g.* DeepNOG [Feldbauer et al. 2020]). Thus, in chapter 2, I developed OMAmer, a mapping approach aiming to combine these three properties. By modelling gene families and subfamilies with hierarchical orthologous groups (HOGs) and limiting over-specific placements through the setting of similarity thresholds, OMAmer is more precise than mapping approaches relying on closest sequences. Moreover, OMAmer performs alignment-free comparisons directly against HOGs (and not against sequences). Thus, in addition to being extremely fast, the number of comparisons scales sublinearly with the number of reference genomes. Moreover, the algorithmic complexity should continue to decrease with the growing knowledge of HOG *k*-mer spaces that is directly linked to the increasing integration of new genomes in reference databases. Indeed, the number of additional computing operations required by the addition of a new genome should decrease as each new genome should had less novel *k*-mer to the reference database. In this subsection,

I share my new thoughts (since the publication of the chapter) on the existing limitations of OMAmer, its potential extensions and application to large-scale orthology inference.

In its current state, the main drawback of OMAmer is its lack of sensitivity when query and reference species are distantly related, most likely a consequence of the choice of relying solely on alignment-free comparisons (Zielezinski et al. 2017). Indeed, DIAMOND was systematically either on par or more sensitive than OMAmer for family-level assignments, although combining *k*-mers of homologous sequences probably mitigated this effect. Nonetheless, sensitivity is probably the less relevant property for a mapping approach when considering that reference orthology databases will release increasingly denser coverages of the tree of life. In addition, although OMAmer was shown to be more precise than closest sequence approaches, OMAmer surely remains less precise than phylogenetic placements approaches relying on molecular evolutionary models (Schreiber et al. 2014; Tang, Finn, and Thomas 2019; Emms and Kelly 2022). For example, these methods should better deal with the varying evolutionary rates of subfamilies that is expected every time one duplicate undergoes a relaxed selection pressure (Kuzmin, Taylor, and Boone 2021). By contrast, by relying on a single similarity threshold, OMAmer shall miss fast-evolving subfamilies and predict too many slow-evolving ones. Thus, since the publication of OMAmer, I have identified four avenues to increase OMAmer's accuracy, each based on a different source of information.

First, including models of sequence evolution to extend the *k*-mer space of query sequences or reference HOGs has the potential to increase OMAmer's sensitivity. One possibility would be to search for similar *k*-mers instead of exact matches, like BLAST and MMSeqs2 do (Stephen F. Altschul et al. 1990; Steinegger and Söding 2017). Thus, given the desired level of sensitivity, increasingly more distant *k*-mers from query *k*-mers could be generated using a substitution matrix (*e.g.* BLOSUM62) and searched against reference gene families. Alternatively, more sophisticated evolutionary models accounting for the phylogeny could be used to extend the *k*-mer spaces of reference families with the most likely *k*-mers to have evolved from each HOG, similarly to the metagenomics taxonomic classifier RAPPAS (Linard, Swenson, and Pardi 2019). However, this would require associating multiple sequence alignments (MSAs) and phylogenetic trees to HOGs, which OMA does not provide (Altenhoff et al. 2021).

Second, conserved subsequence orders between genes provide strong evidence for homology. Thus, exploiting this information should yield better homology prediction than

relying merely on comparing unordered *k*-mer sets. For example, Gapped BLAST and MMSeqs2 require two consecutive *k*-mer matches on the same diagonal (separated by the same number of residues in the two sequences) before the alignment process (S. F. Altschul et al. 1997; Steinegger and Söding 2017). Moreover, these tools use this idea to reduce the similarity threshold on *k*-mer matches for a higher sensitivity, without sacrificing specificity. Finally, conserved *k*-mer lists encode the differentiation between homologous and repetitive *k*-mers, which would enable OMAmer to exploit more refined alignment-free statistics integrating *k*-mer frequencies (Zielezinski et al. 2017; Lippert, Huang, and Waterman 2002).

Third, the species tree should also provide another valuable source of information to improve the accuracy of orthology assignments since several methods of orthology and gene tree inference already exploit this idea (Thomas 2010; Altenhoff et al. 2013; Boussau et al. 2013; Morel et al. 2020). Indeed, knowing in advance where the query sequence diverged from the species tree strongly reduces the set of possible HOGs during orthology assignment. For example, a primates sequence cannot belong to the *Murinae*-specific subfamilies *ins1* and *ins2*, while a *Murinae* sequence can but cannot belong merely in the *INS* family (Irwin 2021). Thus, stopping the placement procedure when it reaches a HOG involving the reference taxon closest to the query taxon should provide a better criterion for avoiding overly specific placements, while limiting overly general placements. Although its effect on accuracy remains to be assessed, this option has been added to OMAmer after its publication. However, generalising this idea to taxonomic mixtures for applications such as detecting contamination or metagenomics would require a more sophisticated integration of the taxonomic information. One promising avenue comes again from metagenomics. To estimate species abundance from reads placed in a reference species tree, Bracken redistributes reads by calculating their Bayesian probability of having been placed in a particular taxonomic level *given* that they belong to a particular species (Lu et al. 2017). Specifically, the prior probability that a read comes from a particular species and the probability that it was placed at a particular taxonomic level given that it comes from a particular species are calculated using empirical distributions obtained by placing reference species. Thus, since Bracken relies on Kraken taxonomic assignments, I believe OMAmer could benefit from Bracken's ideas in the same way that OMAmer was inspired by Kraken (Wood and Salzberg 2014). For example, one could imagine refining the OMAmer score, which guides placement in the HOG hierarchy, by weighing it with similar probabilities. Thus, scores for species enriched with specific placements would be downweighed, while scores for species with overly general placements would be upweighted.

Fourth, conserved gene orders (or synteny) between the query and ancestral reference genomes could also complement sequence similarity. Synteny is already used in orthology inference and ancestral synteny inference is under development in OMA (Linard et al. 2021; Altenhoff et al. 2021). This information could, for instance, help discriminate subfamilies that have recently diverged but not enough to be differentiated with *k*-mers. In an extreme case where HOG A (locus 1) and HOG B (locus 2) are both equally similar to the query genes A (locus 1) and B (locus 2), one could accurately assign gene A to HOG A and gene B to HOG B by maintaining the synteny between the HOGs and the queries.

Integrating the ever-increasing number of available genomes into orthology databases like OMA remains the main goal of OMAmer. However, orthology assignment approaches provide incomplete evolutionary relationships because they do not resolve orthology and paralogy between query sequences. Thus, although the resulting extended orthologous groups are useful for many applications (*e.g.* chapter 4), explicit evolutionary relationships encoded in gene trees or HOGs have stronger potential for biological discoveries (*see* chapter 1). To bridge the gap between orthology assignments and HOGs, I expect large-scale orthology inference to become an iterative three step process. First, existing pipelines of orthology inference shall be run on the highest quality and taxonomically most diverse subset of reference proteomes to build accurate reference HOGs. Second, "periphery" proteomes, which represent the bulk of the data, shall be mapped on these reference HOGs using efficient mapping approaches such as OMAmer. Third, mapped sequences should be integrated into reference HOGs and the whole process could repeat in an iterative manner. By breaking-down all-against-all alignments within each HOG, this process should scale at least sub-quadratically with the number of proteomes. In chapter 5, we presented a prototype of such an approach in which we computed an OMA instance for 363 bird genomes by decomposing the computation of all-versus-all alignments in each bird HOG independently. Nonetheless, more sophisticated developments would be required to extend the approach to the whole tree of life, for instance by computing pairwise alignments while traversing HOGs from leaves to roots.

Volume

Efficient visualisations are required to interpret large volumes of complex data (Qu et al. 2019) and as highlighted in chapter 1, there is a lack of such tools to visualise large gene families. Thus, in chapter 3, I built up on the HOG model and the Phylo.IO codebase (Robinson, Dylus, and Dessimoz 2016) to provide a compact and reactive visualisation for

large gene families. Briefly, Matreex achieves compactness by associating phylogenetic profiles (for compact view of gene distribution across species) to gene trees (for the evolutionary component). I apply Matreex in chapter 5 to diagnose gene families that depict duplications and losses correlated with bird phenotypes.

Variety

My direct contribution to the variety characteristic of Big data also lied in the development of OMAmer. Indeed, mapping approaches provide a robust solution for the integration of low-quality proteomes (*see* chapter 1). Indirectly, however, OMAmer is the basis for the new tool, OMArk (Nevers et al. *in prep*), which aims to assess multiple aspects of proteome quality, including completeness like BUSCO (Waterhouse et al. 2018) but also proportions of false and fragmented gene models. In chapter 4, I leveraged the robustness of OMAmer to integrate low-quality proteomes into orthologous groups and OMArk's quality controls to balance the quality between venomous and outgroup proteomes, thus limiting genome annotation biases.

Applications

In the second half of my thesis, I attempted to capitalise on these methodological developments with two pilot studies toward Big data genomics. In these two studies, I focused on investigating the genetic basis of convergent phenotypes, which should provide larger statistical power (Sackton and Clark 2019). Moreover, they illustrate nicely the potential of Big data in comparative genomics since the number of available convergent replicates should correlate with the number of sequenced genomes. In both studies, I focused on gene family expansions and contractions as these types of genetic changes can be inferred directly from orthologous groups or HOGs (by contrast to changes in sequence evolution or gene expression, for instance).

In chapter 4, I tried to characterise the role of convergence at the level of gene family expansions in animal venom evolution by contrasting the gene repertoires of 68 venomous and closely related non-venomous species. Although I found little evidence for extensive levels of convergence, I found gene families with at least two gene family expansions in venomous clades to be enriched in venom associated functions. This suggested a more modest role of convergent duplications underlying venom evolution. However, several factors suggest a possible lack of power in this analysis. First, the initial dataset was biased by unbalanced

distributions of herbivores and carnivores between venomous and outgroups, which lead to an underestimation of the true level of convergence. Second, the dataset was relatively small, which limited the ability to deal with noise statistically. Indeed, comparing small groups of low-quality proteomes can yield false positives and negatives simply by chance, while larger groups provide better estimates of the mean. One reason for this small dataset was the lack of an integrated proteome database, which complicated the task of gathering proteomes. For example, NCBI assemblies often lack protein files. Moreover, collecting phenotypic data required exhaustive manual curation (*e.g.* systematic verification of venomous status) due to the lack of an integrated phenotypic database for venomous species. This also limited the ability to scale-up the dataset collection. Third, we relied on an *ad hoc* approach to infer gene family expansions and contractions based only on extant copy numbers as the fast-OMA pipeline used in chapter 5 was not available yet. This approach cannot easily distinguish expansion in one clade (*e.g.* venomous) from contraction in the other (*e.g.* outgroup). Moreover it cannot differentiate ancestral expansions or contractions, which are more likely to be causal, from events that have arose after the phenotypic transitions.

In chapter 5, to alleviate these limitations, I tried another strategy to capitalise on the fast orthology assignments from OMAmer. This strategy that relied on changing the focus of the study from a phenotype of interest (venom) to a clade of interest (birds) had several advantages. First, gathering the dataset was straightforward since 363 bird proteomes were just released from the second phase of the Bird 10'000 genome project (Feng et al. 2020). Second, these proteomes had been annotated, thus limiting potential annotation biases from proteome heterogeneity (Weisman, Murray, and Eddy 2022). Third, they had a single protein isoform per gene. Fourth, phenotypic data was available in the recently released AVONET dataset (Tobias et al. 2022). Fifth, the dataset was much larger, which enabled me to investigate multiple convergent phenotypes. Finally, I was able to improve the pipeline to detect convergent expansions and contractions due to the recent development of the fast-OMA pipeline. Indeed, I was able to count expansions and contractions based on the ancestral duplications and losses instead of extant copy numbers. This approach has a greater potential to identify causal mutations, which are more closely related to the phenotype (Nagy et al. 2020).

However, despite this progress, I identified two main areas for improvement in Chapter 5. First, using more resolved species trees instead of the NCBI taxonomy to reconstruct HOGs would increase the number of sampled phenotypic transitions. For example, most branches

depicting a transition to a migratory behaviour were missing. Moreover, polytomies likely disturb the inference of duplications and losses. As a result, we are planning a new instance of bird-OMA using a more resolved species tree. Second, the macro-evolutionary approach I developed to identify phenotype to genotype associations has great potential for improvement. For example, correlating gain and loss of phenotypes to evolutionary changes in the same analysis can provide statistical power even in absence of phenotypic convergence (Nagy et al. 2017). Moreover, other genetic changes could be correlated with phenotypic transitions using HOGs as backbones such as changes in domain content, gene order (synteny), sequence evolutionary rates or dN/dS (Nagy et al. 2020).

In conclusion, this thesis brings us one step closer toward Big data comparative genomics.

References

Altenhoff, Adrian M., Manuel Gil, Gaston H. Gonnet, and Christophe Dessimoz. 2013. "Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs." *PloS One* 8 (1): e53786.

Altenhoff, Adrian M., Clément-Marie Train, Kimberly J. Gilbert, Ishita Mediratta, Tarcisio Mendes de Farias, David Moi, Yannis Nevers, et al. 2021. "OMA Orthology in 2021: Website Overhaul, Conserved Isoforms, Ancestral Gene Order and More." *Nucleic Acids Research* 49 (D1): D373–79.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10.

Boussau, Bastien, Gergely J. Szöllosi, Laurent Duret, Manolo Gouy, Eric Tannier, and Vincent Daubin. 2013. "Genome-Scale Coestimation of Species and Gene Trees." *Genome Research* 23 (2): 323–30.

Emms, David Mark, and Steven Kelly. 2022. "SHOOT: Phylogenetic Gene Search and Ortholog Inference." *Genome Biology* 23 (1): 1–13.

Feldbauer, Roman, Lukas Gosch, Lukas Lüftinger, Patrick Hyden, Arthur Flexer, and Thomas Rattei. 2020. "DeepNOG: Fast and Accurate Protein Orthologous Group Assignment." *Bioinformatics* , December. https://doi.org/10.1093/bioinformatics/btaa1051.

Feng, Shaohong, Josefin Stiller, Yuan Deng, Joel Armstrong, Qi Fang, Andrew Hart Reeve, Duo Xie, et al. 2020. "Dense Sampling of Bird Diversity Increases Power of Comparative Genomics." *Nature* 587 (7833): 252–57.

Irwin, David M. 2021. "Evolution of the Insulin Gene: Changes in Gene Number, Sequence, and Processing." *Frontiers in Endocrinology* 12 (April): 649255.

Kuzmin, Elena, John S. Taylor, and Charles Boone. 2021. "Retention of Duplicated Genes in Evolution." *Trends in Genetics: TIG*, July. https://doi.org/10.1016/j.tig.2021.06.016.

Linard, Benjamin, Ingo Ebersberger, Shawn E. McGlynn, Natasha Glover, Tomohiro Mochizuki, Mateus Patricio, Odile Lecompte, et al. 2021. “Ten Years of Collaborative Progress in the Quest for Orthologs.” *Molecular Biology and Evolution* 38 (8): 3033–45.

Linard, Benjamin, Krister Swenson, and Fabio Pardi. 2019. “Rapid Alignment-Free Phylogenetic Identification of Metagenomic Sequences.” *Bioinformatics* , January. https://doi.org/10.1093/bioinformatics/btz068.

Lippert, Ross A., Haiyan Huang, and Michael S. Waterman. 2002. “Distributional Regimes for the Number of K-Word Matches between Two Random Sequences.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (22): 13980–89.

Lu, Jennifer, Florian P. Breitwieser, Peter Thielen, and Steven L. Salzberg. 2017. “Bracken: Estimating Species Abundance in Metagenomics Data.” *PeerJ Computer Science* 3 (January): e104.

Morel, Benoit, Alexey M. Kozlov, Alexandros Stamatakis, and Gergely J. Szöllősi. 2020. “GeneRax: A Tool for Species Tree-Aware Maximum Likelihood Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss.” *Molecular Biology and Evolution*, June. https://doi.org/10.1093/molbev/msaa141.

Nagy, László G., Zsolt Merényi, Botond Hegedüs, and Balázs Bálint. 2020. “Novel Phylogenetic Methods Are Needed for Understanding Gene Function in the Era of Mega-Scale Genome Sequencing.” *Nucleic Acids Research* 48 (5): 2209–19.

Nagy, László G., Robert Riley, Philip J. Bergmann, Krisztina Krizsán, Francis M. Martin, Igor V. Grigoriev, Dan Cullen, and David S. Hibbett. 2017. “Genetic Bases of Fungal White Rot Wood Decay Predicted by Phylogenomic Analysis of Correlated Gene-Phenotype Evolution.” *Molecular Biology and Evolution* 34 (1): 35–44.

Navarro, Fábio C. P., Hussein Mohsen, Chengfei Yan, Shantao Li, Mengting Gu, William Meyerson, and Mark Gerstein. 2019. “Genomics and Data Science: An Application within an Umbrella.” *Genome Biology* 20 (1): 109.

Qu, Zhonglin, Chng Wei Lau, Quang Vinh Nguyen, Yi Zhou, and Daniel R. Catchpoole. 2019. “Visual Analytics of Genomic and Cancer Data: A Systematic Review.” *Cancer Informatics* 18 (March): 1176935119835546.

Robinson, Oscar, David Dylus, and Christophe Dessimoz. 2016. “Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web.” *Molecular Biology and Evolution* 33 (8): 2163–66.

Sackton, Timothy B., and Nathan Clark. 2019. “Convergent Evolution in the Genomics Era: New Insights and Directions.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 374 (1777): 20190102.

Schreiber, Fabian, Mateus Patricio, Matthieu Muffato, Miguel Pignatelli, and Alex Bateman. 2014. “TreeFam v9: A New Website, More Species and Orthology-on-the-Fly.” *Nucleic Acids Research* 42 (D1): D922–25.

Steinegger, Martin, and Johannes Söding. 2017. “MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets.” *Nature Biotechnology* 35 (11): 1026–28.

Stephens, Zachary D., Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. 2015. “Big Data: Astronomical or Genomical?” *PLoS Biology* 13 (7): e1002195.

Tang, Haiming, Robert D. Finn, and Paul D. Thomas. 2019. “TreeGrafter: Phylogenetic Tree-Based Annotation of Proteins with Gene Ontology Terms and Other Annotations.” *Bioinformatics* 35 (3): 518–20.

Thomas, Paul D. 2010. “GIGA: A Simple, Efficient Algorithm for Gene Tree Inference in the Genomic Age.” *BMC Bioinformatics* 11 (June): 312.

Tobias, Joseph A., Catherine Sheard, Alex L. Pigot, Adam J. M. Devenish, Jingyi Yang, Ferran Sayol, Montague H. C. Neate-Clegg, et al. 2022. “AVONET: Morphological, Ecological and Geographical Data for All Birds.” *Ecology Letters* 25 (3): 581–97.

Waterhouse, Robert M., Mathieu Seppey, Felipe A. Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2018. “BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics.” *Molecular Biology and Evolution* 35 (3): 543–48.

Weisman, Caroline M., Andrew W. Murray, and Sean R. Eddy. 2022. “Mixing Genome Annotation Methods in a Comparative Analysis Inflates the Apparent Number of Lineage-Specific Genes.” *Current Biology: CB* 32 (12): 2632–39.e2.

Wood, Derrick E., and Steven L. Salzberg. 2014. “Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments.” *Genome Biology* 15 (3): R46.

Zielezinski, Andrzej, Susana Vinga, Jonas Almeida, and Wojciech M. Karlowski. 2017. “Alignment-Free Sequence Comparison: Benefits, Applications, and Tools.” *Genome Biology* 18 (1): 186.