Matreex: Compact and Interactive Visualization for Scalable Studies of Large Gene Families

Victor Rossier [1,2,3], Clement Train[1], Yannis Nevers [1,2], Marc Robinson-Rechavi [2,3], and Christophe Dessimoz [1,2,*]

[1]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

[2]SIB Swiss Institute of Bioinformatics, Comparative Genomics, Lausanne, Switzerland

[3]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

*Corresponding author: E-mail: Christophe.Dessimoz@unil.ch.

Accepted: May 03, 2024

Abstract

Studying gene family evolution strongly benefits from insightful visualizations. However, the ever-growing number of sequenced genomes is leading to increasingly larger gene families, which challenges existing gene tree visualizations. Indeed, most of them present users with a dilemma: display complete but intractable gene trees, or collapse subtrees, thereby hiding their children's information. Here, we introduce Matreex, a new dynamic tool to scale up the visualization of gene families. Matreex's key idea is to use "phylogenetic" profiles, which are dense representations of gene repertoires, to minimize the information loss when collapsing subtrees. We illustrate Matreex's usefulness with three biological applications. First, we demonstrate on the MutS family the power of combining gene trees and phylogenetic profiles to delve into precise evolutionary analyses of large multicopy gene families. Second, by displaying 22 intraflagellar transport gene families across 622 species cumulating 5,500 representatives, we show how Matreex can be used to automate large-scale analyses of gene presence–absence. Notably, we report for the first time the complete loss of intraflagellar transport in the myxozoan *Thelohanellus kitauei*. Finally, using the textbook example of visual opsins, we show Matreex's potential to create easily interpretable figures for teaching and outreach. Matreex is available from the Python Package Index (pip install Matreex) with the source code and documentation available at https://github.com/DessimozLab/matreex.

Key words: gene evolution, visualization, software tool, tree reconciliation, phylogenetic profile.

Significance

In an era where the goal of sequencing all eukaryotic species has been set, it has become critical to find ways to represent this huge volume of upcoming data. In particular, studying gene family evolution strongly benefits from insightful visualizations of their complex histories of duplications and losses. However, existing tools merely rely on gene trees and present users with an insoluble dilemma: display complete but intractable gene trees, or collapse subtrees, thereby hiding swaths of information. In this article, we introduce Matreex, a new dynamic tool to scale up the visualization of gene families. Matreex's key idea is to use "phylogenetic" profiles to minimize the information loss when collapsing subtrees.

Introduction

Studying the evolutionary dynamics of gene families strongly benefits from appropriate visualization tools. For example, we can draw evolutionary and functional hypotheses by visually correlating gene repertoires with adaptations or between families. Moreover, visualizing the evolutionary history of a gene family provides the framework to generalize 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 more than 100,000 members across more than 2,500 species (Altenhoff et al. 2021). Thus, gene family visualization tools able to integrate this large volume of data and exploit its full potential are needed. Although many tools can represent large gene trees (Xu et al. 2021; Penel and de Vienne 2022), few are interactive, which is essential for users to explore large gene families.

Gene trees labeled with duplications and speciations are typically used to depict the evolutionary history of gene families. However, existing interactive 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 visualize 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 summarize 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), visualizing 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 summarize collapsed subtrees. Thus, it simplifies gene tree visualization while reducing the information loss. The resulting highly compact and reactive visualization of evolution enables Matreex to scale up to the ongoing deluge of genomic data. Moreover, it provides the opportunity for new biological discoveries, for the production of paper figures, and for didactic support for teaching in evolutionary biology. We illustrate Matreex with three biological applications.

New Approach

To enable compact and reactive visualization of large gene families, 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 in the corresponding row. This keeps information about ancestral events such as gene loss, duplication, or transfer. Moreover, the species tree displayed orthogonally from the gene tree provides what is often a good proxy for the topology of these subtrees (Morel et al. 2020). Remarkably, the extreme case, where all subtrees without duplications and congruent with the species tree are collapsed, provides the same information as a fully extended gene tree but much more compactly. Indeed, these subtrees do not need to be displayed because their topology is explicit in the species tree. We provide rapid access to this view with Matreex's "Smart Collapse" option.

For gene families with a high number of duplication events, collapsing only subtrees without duplication is not enough and summarizing them requires also collapsing subtrees with subfamilies (children of duplication nodes). In that case, the resulting phylogenetic profiles depict the

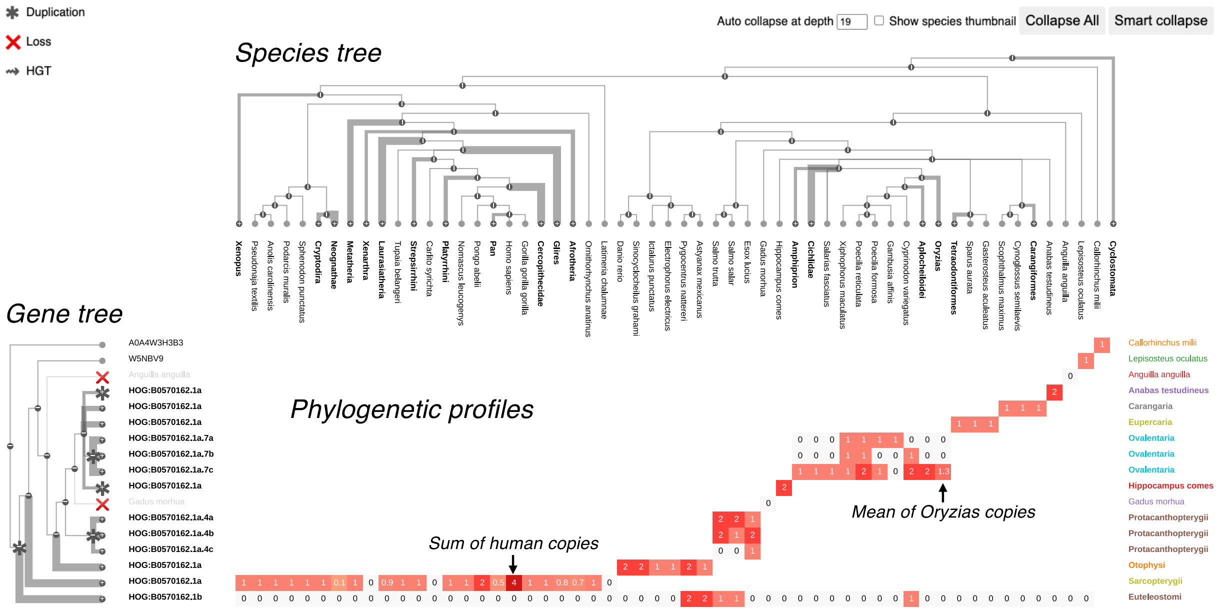


Fig. 1.—Matreex's layout consists of a gene tree, a species tree and a matrix of phylogenetic profiles. Gene tree labels represent gene subfamily memberships (OMA HOGs in this case) for collapsed nodes, gene ids for leaves, and taxon or species names for lost genes (implied from the species tree). Figures in the phylogenetic profiles represent the average number of in-paralogs of the clade species. For a given profile, clades that have lost their genes are displayed with zeros on a gray background, while clades that are outgroups of the corresponding subtree remain empty. Branch thickness increases with the collapsed subtree size and cell color darkness with the number of in-paralogs in the cell. The taxonomic levels of collapsed subtrees and phylogenetic profiles are annotated on the right. "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 release). Italic annotations do not belong to the Matreex layout but were added for figure clarity.

combined gene content of each subfamily per species. For example, the profile of the collapsed *Sarcopterygii* subtree of the red-sensitive visual opsin (long-wavelength-sensitive [LWS]) family shows four copies for humans due to the existence of primates-specific subfamilies (Fig. 1). To deal with large gene families, Matreex includes the option of collapsing all subfamilies, including the root node (Matreex's "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 summarized 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 allow users to zoom on taxa of interest, while achieving high levels of gene family compactness. For that task, Matreex provides control over which taxa are displayed through its interactive species tree. When collapsing a taxon in the species tree, all corresponding gene tree nodes are also collapsed, and the phylogenetic profiles are summarized. This is done by averaging the numbers of in-paralogs of the species descending from the collapsed node. For example, collapsing the *Oryzias* node in the LWS family automatically merges *Oryzias*-specific subfamilies and averages the copy numbers of *Oryzias* species (Fig. 1). 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 improve 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 colors 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, the documentation, and code to reproduce the below figures available at https://github.com/DessimozLab/matreex.

Matreex is both a command-line tool and a Python library that produces html output files, which can be viewed in standard web browsers and, therefore, easily shared. It is implemented in JavaScript with the D3 library and wrapped in a Python module that supports the OMA and PANTHER APIs (Kaleb et al. 2019; Mi et al. 2021). For these two databases, only the gene family identifier, or list of gene family identifiers, is required as input.

To visualize gene trees from other sources, users can also upload their own gene tree in JSON (format described in the GitHub). However, similarly to OMA HOGs or PANTHER gene trees, Matreex requires input gene trees to be consistent with their associated species trees (where speciations follow the same order in both trees). While this may be seen as a limitation of Matreex—reducing its application scope, the assumption that duplications and losses occur on branches of the species tree ensures Matreex's scalability to very large gene trees. Indeed, only one species tree needs to be displayed at the top to navigate the gene tree.

For users who wish to visualize their own genomes with Matreex, one option is to use OMA standalone (Altenhoff et al. 2019). In short, OMA standalone allows you to combine all-against-all alignments exported from the OMA database with custom genomes. The resulting OrthoXML files can then be converted to JSON using ETE3 for visualizing in Matreex (Huerta-Cepas et al. 2016). Similarly, FastOMA can be used to quickly recompute orthologous groups with additional genomes (Majidian et al. 2024).

Applications

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

Origin and Evolution of Eukaryotic MutS Genes

Matreex enables users to analyze precisely the gene repertoire evolution of large multicopy families. First, subfamily gene repertoires can be correlated among themselves or with adaptations. Second, the gene tree enables the study of evolutionary relationships between phylogenetic profiles. This can be useful, for instance, to differentiate orthologous from paralogous profiles and to evaluate the quality of 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. 2). First, we evaluated whether some established hypotheses were further evidenced or challenged by the examined MutS gene tree. Second, we assessed which still-debated hypotheses were supported by this tree, and, third, we formulated new hypotheses by searching for patterns in this new visualization. Finally, we contextualized the results with functional and evolutionary knowledge from the literature to highlight the importance of such an approach.

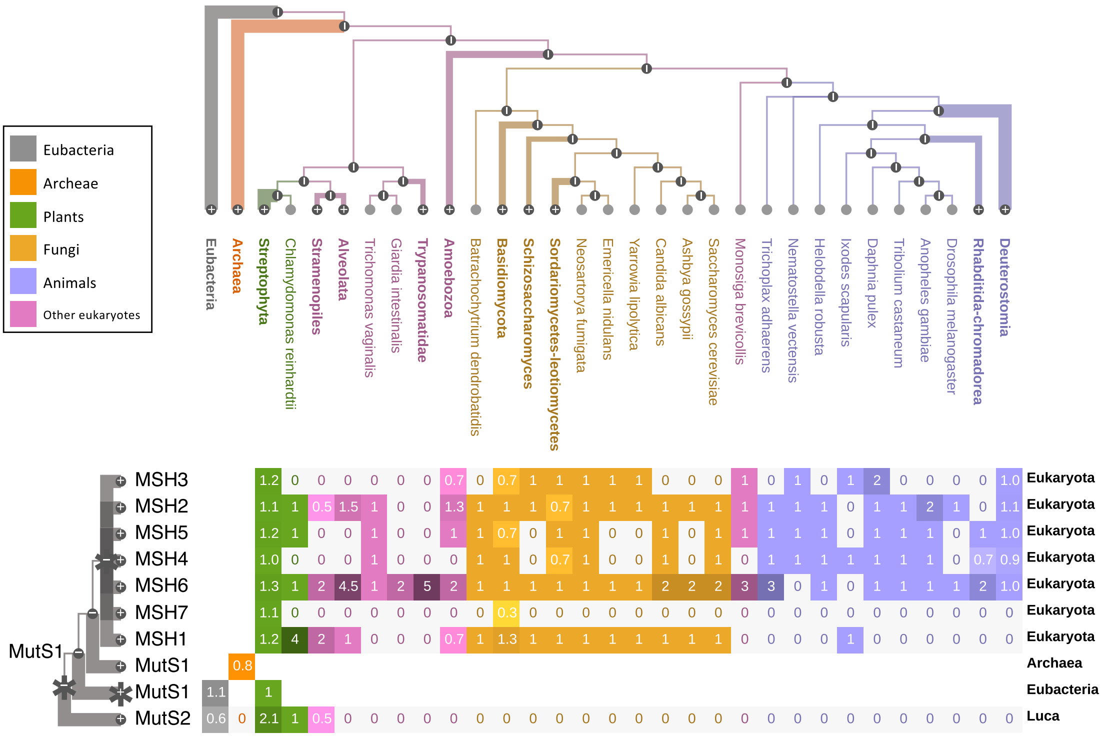
MutS genes are involved in the DNA mismatch repair pathway (Liu et al. 2017; Mi et al. 2021). Although bacteria have multiple MutS genes, only MutS1 and MutS2 are found in both bacteria and eukaryotes. MutS2 is found only in photosynthetic eukaryotes and its transfer from cyanobacteria through the chloroplast endosymbiosis is well established (Fig. 2, H1) (Lin et al. 2007). Matreex shows clearly 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 nonphotosynthetic 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. This literal interpretation of the gene tree is unlikely given the large number of gene losses required to support this scenario. However, this example illustrates the usefulness of Matreex in quality control applications, for instance by comparing the structure of the gene tree with gene distributions.

In contrast to MutS2, the origin of MutS1 remains under debate. Eukaryotic MutS1 genes (MSH2-7) were first thought to originate from the mitochondrial 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* (Fig. 2, H2). Matreex shows clearly these orthologous relationships between archeal MutS1 and eukaryotic MSH2-7 because they form a monophyletic clade in the gene tree and their profiles do not overlap.

MSH2-6 genes originated from duplications in the eukaryote ancestor (Fig. 2, H3), while MSH7 arose from a plant-specific duplication of MSH6 (Fig. 2, H4) (Lin et al. 2007). Matreex clearly represents this rapid gene family expansion with a compact block of subfamily profiles, although PANTHER supports MSH7 to be another eukaryote subfamily. This is visible with Matreex by its phylogenetic profile mainly displaying zeros instead of empty cells. Given the improbable number of implied gene losses and the current state of the literature, this pattern most likely reflects a methodological artifact rather than a true evolutionary scenario.

However, the origin of MSH1 in eukaryotes remains unclear. Although originally thought to descend from the same bacterial MutS ancestor as MSH2-7 (Fig. 2, H5) (Lin et al. 2007), a second acquisition of MSH1 genes in

a Matreex view of the eukaryotic MutS family



	Eubacteria	Archaea	Streptophyta	Chlamydomonas reinhardtii	Stramenopiles	Alveolata	Trichomonas vaginalis	Giardia intestinalis	Trypanosomatidae	Amoebozoa	Batrachochytrium dendrobatidis	Basidiomycota	Schizosaccharomyces	Sordariomycetes-leotiomycetes	Neosartorya fumigata	Emericella nidulans	Yarrowia lipolytica	Candida albicans	Ashbya gossypii	Saccharomyces cerevisiae	Monosiga brevicollis	Trichoplax adhaerens	Nematostella vectensis	Helobdella robusta	Ixodes scapularis	Daphnia pulex	Tribolium castaneum	Anopheles gambiae	Drosophila melanogaster	Rhabditida-chromadorea	Deuterostomia	
MSH3			1.2	0	0	0	0	0	0	0.7	0	0.7	1	1	1	1	1	0	0	0	1	0	1	0	1	2	0	0	0	0	1.0	Eukaryota
MSH2			1.1	1	0.5	1.5	1	0	0	1.3	1	1	1	0.7	1	1	1	1	1	1	1	1	1	1	0	1	1	2	1	0	1.1	Eukaryota
MSH5			1.2	1	0	0	1	0	0	1	1	0.7	0	1	1	0	0	1	0	1	1	1	1	1	0	1	1	1	0	1	1.0	Eukaryota
MSH4			1.0	0	0	0	1	0	0	0	1	1	0	0.7	1	0	0	1	0	1	0	1	1	1	1	1	1	1	0	0.7	0.9	Eukaryota
MSH6			1.3	1	2	4.5	1	2	5	2	1	1	1	1	1	1	1	2	2	2	3	3	0	1	0	1	1	1	1	2	1.0	Eukaryota
MSH7			1.1	0	0	0	0	0	0	0	0	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Eukaryota
MSH1			1.2	4	2	1	0	0	0	0.7	1	1.3	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	Eukaryota
MutS1		0.8																														Archaea
MutS1	1.1		1																													Eubacteria
MutS2	0.6	0	2.1	1	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Luca

b Hypotheses on MutS evolution

ID	Name	Status	Gene tree support
H1	MutS2 originated from the transfer of a cyanobacteria MutS1 to photosynthetic eukaryotes (Lin et al. 2007).	Established	No, the gene tree supports a vertical origin of MutS1 and MutS2.
H2	Vertical origin of eukaryotic MutS1 genes (MSH2-7) from Archaea (Hofstatter and Lahr 2021)	Debated. The alternative hypothesis proposes MutS1 eukaryotic genes to have been transferred from an α-Proteobacteria (Lin et al. 2007).	Yes
H3	MutS1 radiation in Eukaryotes (Lin et al. 2007)	Established	Yes
H4	MSH7 evolved in plants from the duplication of MSH6 (Lin et al. 2007).	Established	No, the gene tree implies MSH7 to be another eukaryote subfamily.
H5	Common origin of MSH1 and MSH2-7 genes (Lin et al. 2007)	Debated. The alternative hypothesis proposes MSH1 acquisition along organelle endosymbioses (Hofstatter and Lahr 2021).	Yes
H6	Artefactual pooling of MSH3 into MSH6	New	Yes
H7	Parallel losses of MSH4 and MSH5	Established for Schizosaccharomyces and D. Melanogaster (Kohl et al. 2012; Manhart and Alani 2016). New for other fungi and eukaryotes.	Yes

Fig. 2.—Detailed evolutionary analysis of the eukaryotic MutS family. a) Matreex view of the MutS gene family (gene tree from PANTHER v.17). Clade legends and gene family names do not belong to the Matreex layout but were added for figure clarity. b) Hypotheses on MutS evolution discussed in the text with their level of support in the literature and in the examined gene tree. An orange, green, or blue background indicates, respectively, a conflict with the literature, no conflict with the literature, or a new hypothesis from this work.

eukaryotes from cyanobacterial endosymbiosis has been suggested (Hofstatter and Lahr 2021). The underlying gene tree supports the original hypothesis as we found all eukaryotic MSH1 genes in the same subfamily (not including the "plant MSH1" gene, which is known to be of entirely independent origins (Ogata 2011)). Matreex helped to draw this conclusion as these eukaryotic MSH1 genes belong to the same collapsed subtree and phylogenetic profile, indicating a monophyletic origin. Moreover, we recovered the absence of MSH1 from most bilaterian animal lineages (Bell et al. 2004; Muthye and Lavrov 2021), with the exception of the tick *Ixodes scapularis* whose copy likely originated from transfer or contamination from the genome of its *Rickettsia* endosymbiont), 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* and *Trypanosomatidae*), yeast (*Saccharomyces cerevisiae*), nematodes, and fruit fly (*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 hypothesized that MSH6 is artifactually attracting other genes, in particular MSH3 ones, during phylogenetic reconstruction (Fig. 2, H6).

Finally, we observed repeated and correlated losses of MSH4 and MSH5 in fungi, fruit fly, and other eukaryotes (Fig. 2, H7). While losses in the latter are likely artifactual (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.

Coevolution of the IFT Genes

Matreex enables users to perform gene presence–absence analyses for dozens of nonhomologous families spanning hundreds of species in a few minutes. This is useful to visualize 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 generalizing 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 light gray backgrounds (Fig. 3). As expected, complete loss of IFT complexes was detected in the main nonciliated taxa (e.g. *Spermatophyta*, *Dikarya*, or *Amoebozoa*). Moreover, due to the sheer number of used genomes, summarized 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], and *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 nonciliated 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*. Second, 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 *C. 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;

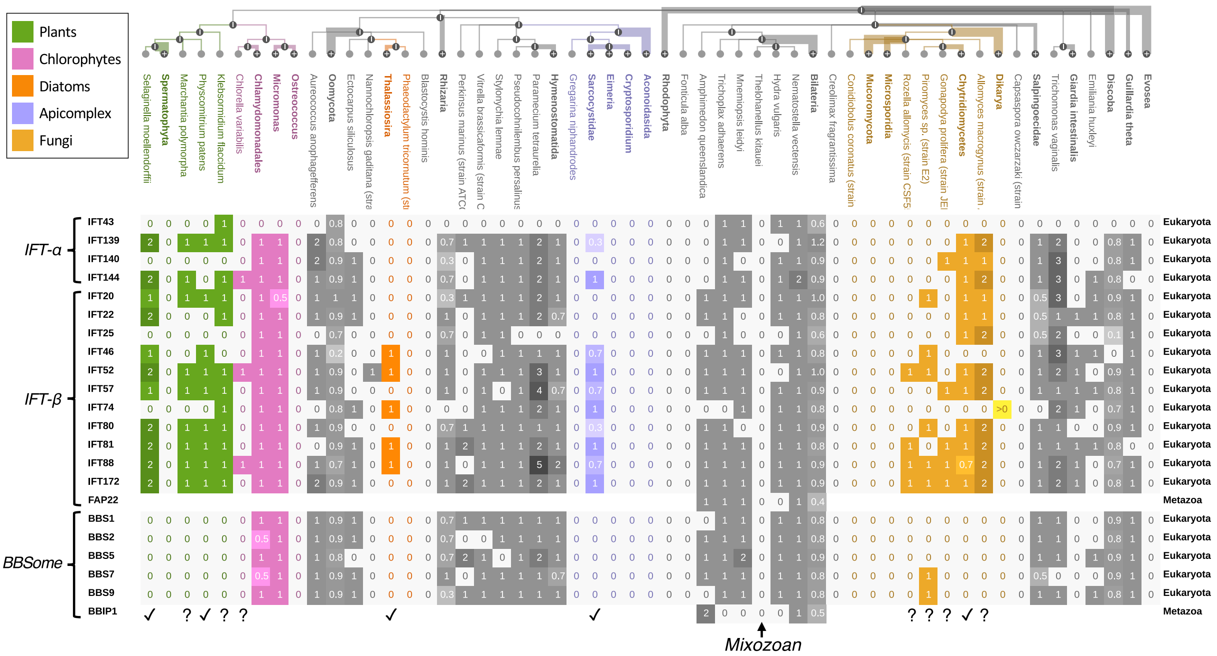


FIG. 3.—IFT gene families (data from OMA all.Dec2021). Colored clades display partial and complete IFT losses that fit the "last-in, first-out" hypothesis for gene module evolution. ✓ highlights partial IFT losses reported by van Dam et al. (2013) and ? the ones reported here. To our knowledge, we are the first to report a complete loss of IFT in the mixozoan *T. kitauei*. The species tree is unresolved because it comes from the OMA database, which is derived from the NCBI taxonomy (Schoch et al. 2020). Clade legends, Italic annotations, brackets, ✓, and ? symbols do not belong to the Matreex layout but were added for figure clarity.

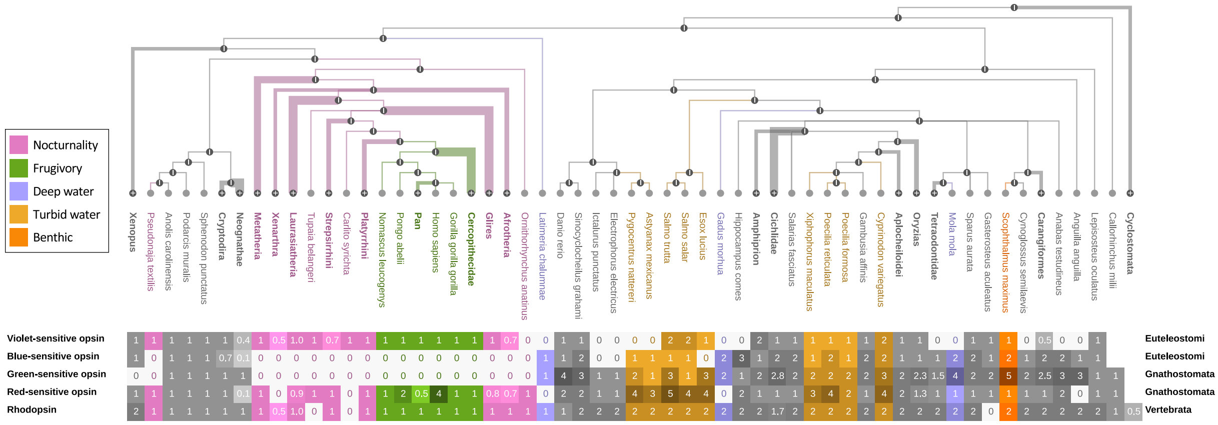
Carlton et al. 2007), *P. tetraurelia* IFT57 copies show evidence of subfunctionalization (Shi et al. 2018), while *T. vaginalis* displays specialized cilia that could have required the recruitment of additional IFT copies. Finally, to explain the retention of *S. moellendorffii* and *A. macrogynus* duplicates that have lost BBSome, we may speculate whether these extra copies could have been co-opted to replace the BBSome functions.

The Visual Opsin Gene Repertoire Correlates with Adaptations in Vertebrates

Matreex enables to quickly identify correlations between adaptations, or phenotypes, and variations in gene copy numbers. Matreex further facilitates the task by depicting losses on a light gray background and expansions on a darker one. Clades can also be colored to highlight the correlation (using the Matreex library examples available on GitHub). Here, we illustrate this feature with the textbook example of visual opsins (Graur and Li 1999; Higgs and Attwood 2005). Because Matreex's representation facilitates an intuitive interpretation of the data, 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 color vision by co-opting a red-sensitive opsin duplicate to absorb green wavelengths, which possibly helped primates to identify edible fruits (Carvalho et al. 2017). Matreex shows clearly and at a glance both the losses in mammals and snakes (Fig. 4, pink), as series of light gray background zeroes, and the secondary amplification in *Homo sapiens* and *Pongo abelii* (Fig. 4, 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 shrunk 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 to 200 m,



	Xenopus	Pseudonaja textilis	Anolis carolinensis	Podarcis muralis	Sphenodon punctatus	Cryptodira	Neognathae	Metatheria	Xenarthra	Laurasiatheria	Tupaia belangeri	Strepsirrhini	Carlito syrichta	Platyrrhini	Nomascus leucogenys	Pongo abelii	Pan	Homo sapiens	Gorilla gorilla gorilla
Violet-sensitive opsin	1	1	1	1	1	1	0.4	1	0.5	1.0	1	0.7	1	1	1	1	1	1	1
Blue-sensitive opsin	1	0	1	1	1	0.7	0.1	0	0	0	0	0	0	0	0	0	0	0	0
Green-sensitive opsin	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Red-sensitive opsin	1	1	1	1	1	1	0.1	1	0	0.9	1	1	0	1	1	2	0.5	4	1
Rhodopsin	2	1	1	1	1	1	1	1	0.5	1.0	0	1	0	1	1	1	1	1	1

	Cercopithecidae	Glires	Afrotheria	Ornithorhynchus anatinus	Latimeria chalumnae	Danio rerio	Sinocyclocheilus grahami	Ictalurus punctatus	Electrophorus electricus	Pygocentrus nattereri	Astyanax mexicanus	Salmo trutta	Salmo salar	Esox lucius	Gadus morhua	Hippocampus comes	Amphiprion	Cichlidae	Salarias fasciatus
Violet-sensitive opsin	1	1	0.7	0	0	1	1	0	0	0	0	2	2	1	0	0	2	1	1
Blue-sensitive opsin	0	0	0	0	1	1	2	0	0	1	1	1	1	0	2	3	1	2	2
Green-sensitive opsin	0	0	0	0	1	4	3	1	1	2	1	3	1	3	2	1	2	2.8	2
Red-sensitive opsin	1	0.8	0.7	1	0	2	2	1	1	4	3	5	4	4	0	2	1	1	1
Rhodopsin	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	1.7	2

	Xiphophorus maculatus	Poecilia reticulata	Poecilia formosa	Gambusia affinis	Cyprinodon variegatus	Aplocheiloidei	Oryzias	Tetraodontidae	Mola mola	Sparus aurata	Gasterosteus aculeatus	Scophthalmus maximus	Cynoglossus semilaevis	Carangiformes	Anabas testudineus	Anguilla anguilla	Lepisosteus oculatus	Callorhinchus milii	Cyclostomata	
Violet-sensitive opsin	1	1	1	1	2	1	1	0	0	1	1	1	0	0.5	0	0	1			Euteleostomi
Blue-sensitive opsin	1	2	1	2	2	1	1	1	2	2	1	2	1	2	2	1	1			Euteleostomi
Green-sensitive opsin	2	2	2	2	3	2	2.3	1.5	4	2	2	5	2	2.5	3	3	1	1		Gnathostomata
Red-sensitive opsin	3	4	2	1	4	2	1.3	1	1	1	1	1	1	1	2	0	1	1		Gnathostomata
Rhodopsin	2	2	2	2	2	2	2	2	2	2	0	2	2	2	2	2	2	1	0.5	Vertebrata

Fig. 4.—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 colors. 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. Ecological niche legends on the left do not belong to the Matreex layout but were added for figure clarity.

max. 600 m), the sunfish (*Mola mola*, depth 30 to 70 m, max. 480 m), and the coelacanth (*Latimeria chalumnae*, depth: 180 to 250 m, max. 700 m) (Fig. 4, 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, as previously suggested (Wang et al. 2021) (Fig. 4, orange). Conversely, the light spectrum is shifted toward longer wavelengths in turbid water, thus favoring 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. 4, 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*).

Conclusion

At a time when 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 introduced an innovative tool to scale the visualization of gene families and illustrate its usefulness with three biological applications. First, we demonstrated Matreex's usefulness in delving into precise evolutionary analyses of multicopy gene families by combining the gene tree with phylogenetic profiles. Second, by displaying 22 intraflagellar gene families across 622 species cumulating 5,500 representatives, we showed how Matreex can be used for analyses of gene presence–absence and reported for the first time the complete loss of IFT in the mixozoan *T. kitauei*. Finally, using the textbook example of visual opsins, we demonstrated Matreex's potential to create easily interpretable figures for outreach tasks. Thus, we hope Matreex will become a valuable tool to gain insights into the evolution of increasingly large gene families.

Acknowledgments

The authors thank Dan Sloan for helpful feedback on the MutS gene family section.

Funding

This work was supported by the Swiss National Science Foundation (SNSF) (167276) as part of the National Research Program 75 "Big Data" and the Swiss National Science Foundation regular project funding scheme (205085).

Data Availability

Matreex is available from the Python Package Index (pip install Matreex) with the source code and documentation available at https://github.com/DessimozLab/matreex.

Literature Cited

Altenhoff AM, Levy J, Zarowiecki M, Tomiczek B, Warwick Vesztrocy A, Dalquen DA, Müller S, Telford MJ., Glover NM, Dylus D, et al. OMA standalone: orthology inference among public and custom genomes and transcriptomes. Genome Res. 2019:29(7):1152–1163. https://doi.org/10.1101/gr.243212.118.

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. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. Nucleic Acids Res. 2021:49(D1):D373–D379. https://doi.org/10.1093/nar/gkaa1007.

Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 2006:444(7116):171–178. https://doi.org/10.1038/nature05230.

Badano JL, Mitsuma N, Beales PL, Katsanis N. The ciliopathies: an emerging class of human genetic disorders. Annu Rev Genomics Hum Genet. 2006:7(1):125–148. https://doi.org/10.1146/annurev.genom.7.080505.115610.

Bell JS, Harvey TI, Sims A-M, McCulloch R. Characterization of components of the mismatch repair machinery in *Trypanosoma brucei*. Mol Microbiol. 2004:51(1):159–173. https://doi.org/10.1046/j.1365-2958.2003.03804.x.

Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S, Pangilinan J, Polle J, et al. The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell. 2010:22(9): 2943–2955. https://doi.org/10.1105/tpc.110.076406.

Borges R, Johnson WE, O'Brien SJ, Gomes C, Heesy CP, Antunes A. Adaptive genomic evolution of opsins reveals that early mammals flourished in nocturnal environments. BMC Genomics. 2018: 19(1):121. https://doi.org/10.1186/s12864-017-4417-8.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021:38(12):5825–5829. https://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. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. Science. 2007:315(5809):207–212. https://doi.org/10.1126/science.1132894.

Carvalho LS, Pessoa DMA, Mountford JK, Davies WIL, Hunt DM. The genetic and evolutionary drives behind primate color vision. Front Ecol Evol. 2017:5:34. https://doi.org/10.3389/fevo.2017.00034.

Cromar GL, Zhao A, Xiong X, Swapna LS, Loughran N, Song H, Parkinson J. PhyloPro2.0: a database for the dynamic exploration of phylogenetically conserved proteins and their domain architectures across the eukarya. Database. 2016:baw013. https://doi.org/10.1093/database/baw013.

Dunn CW, Munro C. Comparative genomics and the diversity of life. Zool Scr. 2016:45(S1):5–13. https://doi.org/10.1111/zsc.12211.

Fuentes D, Molina M, Chorostecki U, Capella-Gutiérrez S, Marcet-Houben M, Gabaldon T. PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. Nucleic Acids Res. 2021:50:D1062–D1068. https://doi.org/10.1093/nar/gkab966.

Graur D, Li WH. Fundamentals of molecular evolution. 2nd edn. Sunderland, Massachusetts, USA: Sinauer Associates Inc; 1999.

Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SM, Amode R, Brent S, et al. Ensembl comparative genomics resources. Database. 2016:bav096. https://doi.org/10.1093/database/bav096.

Higgs PG, Attwood TK. Bioinformatics and molecular evolution. Oxford, UK: Blackwell Publishing Ltd; 2005.

Hofstatter PG, Lahr DJG. Complex evolution of the mismatch repair system in eukaryotes is illuminated by novel archaeal genomes. J Mol Evol. 2021:89(1-2):12–18. https://doi.org/10.1007/s00239-020-09979-5.

Horn T, Narov KD, Panfilio KA. Persistent parental RNAi in the beetle Tribolium castaneum involves maternal transmission of long double-stranded RNA. Advanced Genetics. 2022:3:2100064. https://doi.org/10.1002/ggn2.202100064.

Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016:33(6): 1635–1638. https://doi.org/10.1093/molbev/msw046.

Ilnitskiy IS, Zharikova AA, Mironov AA. OrthoQuantum: visualizing evolutionary repertoire of eukaryotic proteins. Nucleic Acids Res. 2022:50(W1):W534–W540. https://doi.org/10.1093/nar/gkac385.

Kaleb K, Vesztrocy AW, Altenhoff A, Dessimoz C. Expanding the orthologous matrix (OMA) programmatic interfaces: REST API and the OmaDB packages for R and Python. F1000Res. 2019:8: 42. https://doi.org/10.12688/f1000research.17548.2.

Katti C, Stacey-Solis M, Coronel-Rojas NA, Davies WIL. The diversity and adaptive evolution of visual photopigments in reptiles. Front Ecol Evol. 2019:7, Available from: https://www.frontiersin.org/article/10.3389/fevo.2019.00352. https://doi.org/10.3389/fevo.2019.00352.

Kohl KP, Jones CD, Sekelsky J. Evolution of an MCM complex in flies that promotes meiotic crossovers by blocking BLM helicase. Science. 2012:338(6112):1363–1365. https://doi.org/10.1126/science.1228190.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019:47(D1):D807–D811. https://doi.org/10.1093/nar/gky1053.

Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, et al. The Earth BioGenome Project 2020: starting the clock. Proc Natl Acad Sci U S A. 2022:119(4):e2115635118. https://doi.org/10.1073/pnas.2115635118.

Lin Z, Nei M, Ma H. The origins and early evolution of DNA mismatch repair genes–multiple horizontal gene transfers and co-evolution. Nucleic Acids Res. 2007:35(22):7591–7603. https://doi.org/10.1093/nar/gkm921.

Liu D, Keijzers G, Rasmussen LJ. DNA mismatch repair and its many roles in eukaryotic cells. Mutat Res - Rev Mut Res. 2017:773: 174–187. https://doi.org/10.1016/j.mrrev.2017.07.001.

Majidian S, Nevers Y, Kharrazi AY, Vesztrocy AW, Pascarelli S, Moi D, Glover N, Altenhoff AM, Dessimoz C. 2024. Orthology inference at scale with FastOMA. bioRxiv 577392, https://doi.org/10.1101/2024.01.29.577392, 31 January 2024, preprint: not peer reviewed.

Manhart CM, Alani E. Roles for mismatch repair family proteins in promoting meiotic crossing over. DNA Repair (Amst). 2016:38:84–93. https://doi.org/10.1016/j.dnarep.2015.11.024.

Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, Thomas PD. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. 2021:49(D1):D394–D403. https://doi.org/10.1093/nar/gkaa1106.

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017:45(D1):D183–D189. https://doi.org/10.1093/nar/gkw1138.

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. GeneRax: a tool for species tree-aware maximum likelihood based gene family tree inference under gene duplication, transfer, and loss. Mol Biol Evol. 2020:37(9):2763–2774. https://doi.org/10.1093/molbev/msaa141.

Musilova Z, Salzburger W, Cortesi F. The visual opsin gene repertoires of teleost fishes: evolution, ecology, and function. Annu Rev Cell Dev Biol. 2021:37(1):441–468. https://doi.org/10.1146/annurev-cellbio-120219-024915.

Muthye V, Lavrov DV. Multiple losses of MSH1, gain of mtMutS, and other changes in the MutS family of DNA repair proteins in animals. Genome Biol Evol. 2021:13(9):evab191. https://doi.org/10.1093/gbe/evab191.

Nevers Y, Poidevin L, Chennen K, Allot A, Kress A, Ripp R, Thompson JD, Dollfus H, Poch O, Lecompte O. Insights into ciliary genes and evolution from multi-level phylogenetic profiling. Mol Biol Evol. 2017:34(8):2016–2034. https://doi.org/10.1093/molbev/msx146.

Nguyen NTT, Vincens P, Roest Crollius H, Louis A. Genomicus 2018: karyotype evolutionary trees and on-the-fly synteny computing. Nucleic Acids Res. 2018:46(D1):D816–D822. https://doi.org/10.1093/nar/gkx1003.

Ogata H, Ray J, Toyoda K, Sandaa R-A, Nagasaki K, Bratbak G, Claverie J-M. Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment. The ISME Journal. 2011:5(7):1143–1151. https://doi.org/10.1038/ismej.2010.210.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999:96(8): 4285–4288. https://doi.org/10.1073/pnas.96.8.4285.

Penel S, de Vienne DM. Tidy tree: a new layout for phylogenetic trees. Mol Biol Evol. 2022:39(10):msac204. https://doi.org/10.1093/molbev/msac204.

Robinson O, Dylus D, Dessimoz C. Phylo.io: interactive viewing and comparison of large phylogenetic trees on the web. Mol Biol Evol. 2016:33(8):2163–2166. https://doi.org/10.1093/molbev/msw080.

Rossier V, Vesztrocy AW, Robinson-Rechavi M, Dessimoz C. OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches. Bioinformatics. 2021:37(18):2866–2873. https://doi.org/10.1093/bioinformatics/btab219.

Rzeszutek I, Swart EC, Pabian-Jewuła S, Russo A, Nowacki M. Early developmental, meiosis-specific proteins - Spo11, msh4-1, and msh5 - affect subsequent genome reorganization in *Paramecium tetraurelia*. Biochim Biophys Acta Mol Cell Res. 2022:1869(6): 119239. https://doi.org/10.1016/j.bbamcr.2022.119239.

Sadreyev IR, Ji F, Cohen E, Ruvkun G, Tabach Y. PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. Nucleic Acids Res. 2015:43(W1): W154–W159. https://doi.org/10.1093/nar/gkv452.

Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, et al. NCBI taxonomy: a comprehensive update on curation, resources and tools. Database. 2020:2020:baaa062. https://doi.org/10.1093/database/baaa062.

Shi L, Koll F, Arnaiz O, Cohen J. The ciliary protein IFT57 in the macronucleus of paramecium. J Eukaryot Microbiol. 2018:65(1):12–27. https://doi.org/10.1111/jeu.12423.

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. Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi. Curr Biol. 2015:25(18): 2404–2410. https://doi.org/10.1016/j.cub.2015.07.053.

Tran N-V, Greshake Tzovaras B, Ebersberger I. PhyloProfile: dynamic visualization and exploration of multi-layered phylogenetic profiles. Bioinformatics. 2018:34(17):3041–3043. https://doi.org/10.1093/bioinformatics/bty225.

Tremblay BJ-M, Lobb B, Doxey AC. PhyloCorrelate: inferring bacterial gene-gene functional associations through large-scale phylogenetic profiling. Bioinformatics. 2021:37(1):17–22. https://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. Stability of SARS-CoV-2 phylogenies. PLoS Genet. 2020:16(11):e1009175. https://doi.org/10.1371/journal.pgen.1009175.

van Dam TJP, Townsend MJ, Turk M, Schlessinger A, Sali A, Field MC, Huynen MA. Evolution of modular intraflagellar transport from a coatomer-like progenitor. Proc Natl Acad Sci U S A. 2013: 110(17):6943–6948. https://doi.org/10.1073/pnas.1221011110.

Wang Y, Zhou L, Wu L, Song C, Ma X, Xu S, Du T, Li X, Li J. Evolutionary ecology of the visual opsin gene sequence and its expression in turbot (*Scophthalmus maximus*). BMC Ecol Evol. 2021:21(1):114. https://doi.org/10.1186/s12862-021-01837-2.

Wickstead B, Gull K. Dyneins across eukaryotes: a comparative genomic analysis. Traffic. 2007:8(12):1708–1721. https://doi.org/10.1111/j.1600-0854.2007.00646.x.

Wu Z, Waneka G, Broz AK, King CR, Sloan DB. MSH1 is required for maintenance of the low mutation rates in plant mitochondrial and plastid genomes. Proc Natl Acad Sci U S A. 2020:117(28): 16448–16455. https://doi.org/10.1073/pnas.2001998117.

Xu S, Dai Z, Guo P, Fu X, Liu S, Zhou L, Tang W, Feng T, Chen M, Zhan L, et al. ggtreeExtra: compact visualization of richly annotated phylogenetic data. Mol Biol Evol. 2021:38(9):4039–4042. https://doi.org/10.1093/molbev/msab166.

Associate editor: Barbara Holland