Matreex: compact and interactive visualisation 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, Switzerland; [2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland; [3]Department of Ecology and Evolution, University of Lausanne, Switzerland;

*Corresponding author: Christophe.Dessimoz@unil.ch

Abstract

Studying gene family evolution strongly benefits from insightful visualisations. However, the ever-growing number of sequenced genomes is leading to increasingly larger gene families, which challenges existing gene tree visualisations. 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 visualisation of gene families. Matreex's key idea is to use "phylogenetic" profiles, which are dense representations of gene repertoires, to minimise the information loss when collapsing subtrees. We illustrate Matreex 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 multi-copy gene families. Secondly, 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.

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 more than 100'000 members across more than 2’500 species (Altenhoff et al. 2021). Thus, gene family visualisation 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 labelled 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 is 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. 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

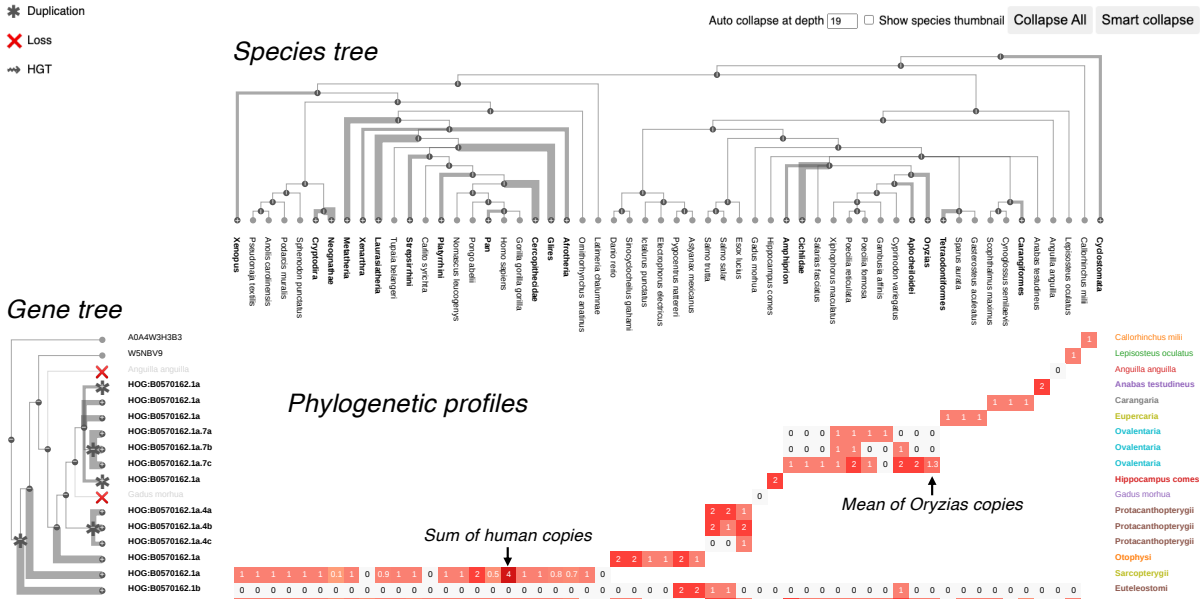


Figure 1. Matreex's layout consists of a gene tree, a species tree and a matrix of phylogenetic profiles. Gene tree labels represent gene subfamily memberships (OMA HOGs in this case) for collapsed nodes, gene ids for leaves, and taxon or species names for lost genes (implied from the species tree). Branch thickness increases with the collapsed subtree size and cell colour 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.

To enable compact and reactive visualisation 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 compact. 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 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, the profile of the collapsed *Sarcopterygii* subtree of the red-sensitive visual opsin (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 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 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 summarised. 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 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 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.

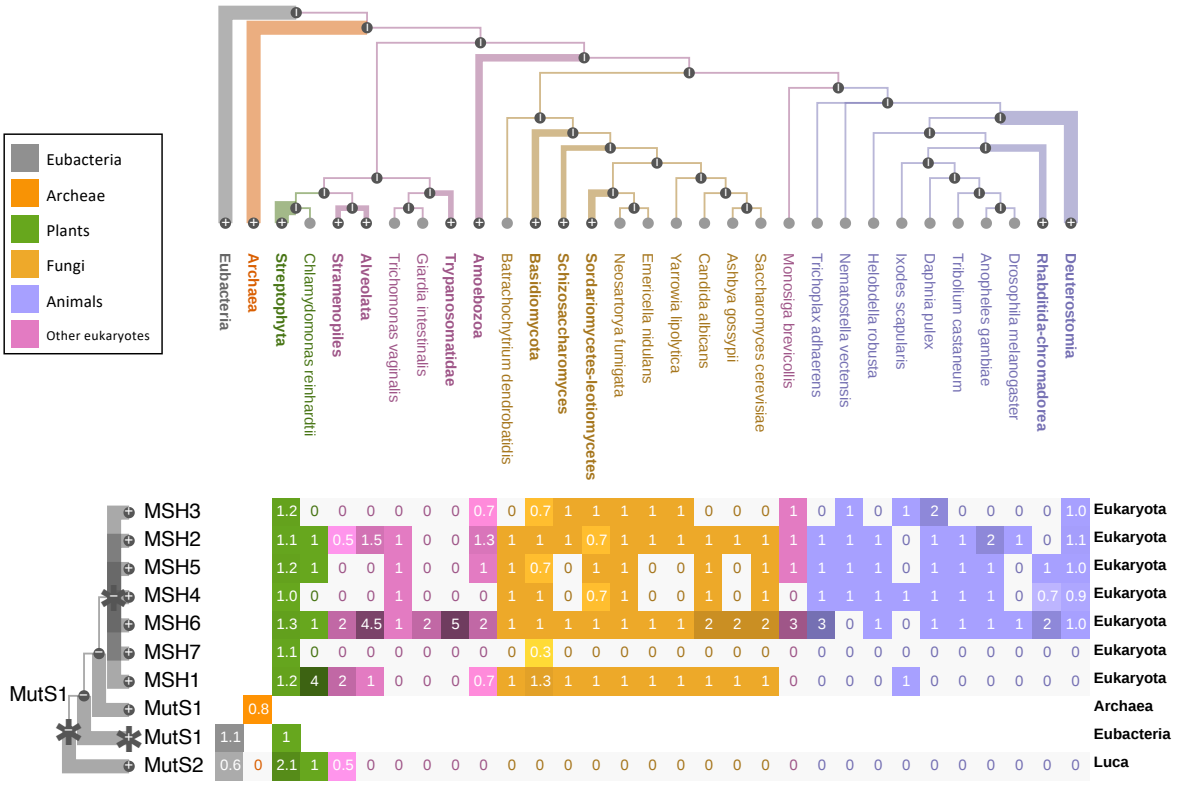
Matreex is both a command-line tool and a python library that produces html output files, which can be viewed in standard web browsers and therefore easily shared. It is implemented in JavaScript with the D3 library and wrapped in a Python module that supports the OMA and PANTHER APIs (Kaleb et al. 2019; Mi et al. 2021). For these two databases, only the gene family identifier, or list of gene family identifiers, is required as input. Moreover, 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). To obtain a tree fulfilling this condition, we suggest using a gene tree inference method that already attempts to minimise conflicts between gene trees and species trees (Boussau et al. 2013; Noutahi et al. 2016; Morel et al. 2020), and to replace remaining conflicting regions with polytomies.

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

A. Matreex view of the eukaryotic MutS family



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	Monophyletic plant- and yeast-MSH1 (Lin et al. 2007)	Debated. The alternative hypothesis proposes MSH1 acquisition from giant viruses (Wu et al. 2020) or 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

Figure 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.

Matreex enables users to analyse precisely the gene repertoire evolution of large multi-copy families. First, subfamily gene repertoires can be correlated among themselves or with adaptations. 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. 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 visualisation. 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 (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 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* (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 radiation 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 lost and the current state of the literature, this pattern most likely reflects a methodological artefact rather than a true evolutionary scenario.

193 However, the origins of the plant- and fungi-MSH1 remain unclear. Although originally thought to
194 descend from the same MutS ancestor as MSH2-7 (Fig. 2, H5) (Lin et al. 2007), an acquisition of MSH1
195 in fungi along with the mitochondrial endosymbiosis has been suggested (Hofstatter and Lahr 2021).
196 Similarly, the plant-MSH1 could have been acquired from giant viruses or along the chloroplast
197 endosymbiosis (Wu et al. 2020; Hofstatter and Lahr 2021). The underlying gene tree supports the
198 original hypothesis as we found the plant- and fungi-MSH1 genes in the same eukaryotic subfamily.
199 Matreex helped to draw this conclusion as both plant- and fungi-MSH1 belong to the same collapsed
200 subtree and phylogenetic profile, indicating a monophyletic origin. Moreover, we recovered the
201 absence of MSH1 in animals (with the exception of the tick *Ixodes scapularis* whose copy likely
202 originated from transfer or contamination from the genome of its *Rickettsia* endosymbiont), which
203 has been recently linked with the exceptionally high evolutionary rates of their mitochondrial genes,
204 as MSH1 is involved in repairing their sequences (Wu et al. 2020).

205 Matreex simplifies the identification of gene repertoire evolutionary patterns. Thus, we observed
206 unexpected expansions of MSH6 in eukaryotes (*e.g. Alveolata*, *Trypanosomatidae*), yeast
207 (*Saccharomyces cerevisiae*), nematodes and fruitfly (*Drosophila Melanogaster*). Then, by expanding
208 the gene tree, we noticed many MSH3 genes misclassified as MSH6, which coincides with predicted
209 MSH3 losses. For example, of the five *Trypanosomatidae* copies, two were surely misclassified MSH3
210 and MSH5 genes, and one was undefined. Thus, although the loss of MSH3 in nematodes and insects
211 and the trypanosome-specific MSH8 subfamily are documented (Bell et al. 2004; Muthye and Lavrov
212 2021), we hypothesised that MSH6 is artefactually attracting other genes, in particular MSH3 ones,
213 during phylogenetic reconstruction (Fig. 2, H6).

214 Finally, we observed repeated and correlated losses of MSH4 and MSH5 in fungi, fruit fly and other
215 eukaryotes (Fig. 2, H7). While losses in the latter are likely artefactual (Rzeszutek et al. 2022),
216 *Schizosaccharomyces* and *D. Melanogaster* are known to have lost and replaced MSH4 and MSH5 for
217 meiotic recombination (Kohl et al. 2012; Manhart and Alani 2016). Moreover, given that these two
218 genes form an obligate complex, other correlated losses in fungi are plausible and could provide good
219 candidates to study alternative meiotic recombination mechanisms.

Coevolution of the intraflagellar transport genes

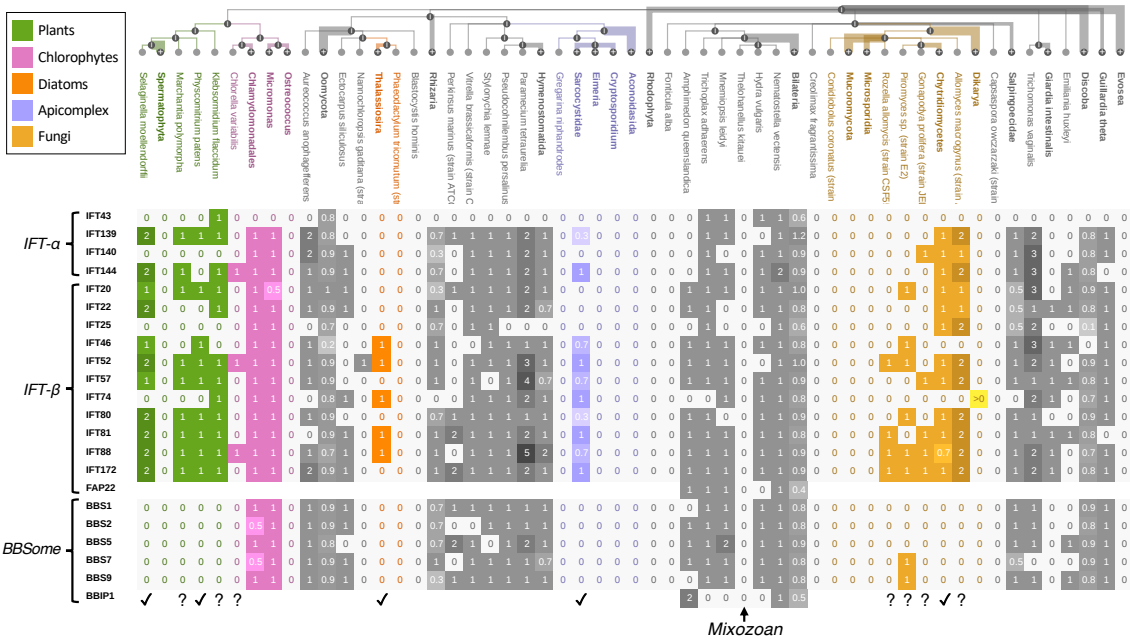


Figure 3. Intraflagellar transport gene families (data from OMA All.Dec2021). Colored clades display partial and complete IFT losses that fit the "last-in, first-out" hypothesis for gene module evolution. ✓ highlight partial IFT losses reported by van Dam et al. (2013) and ?, the ones reported here.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

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

250 Then, we could first quickly confirm all established patterns of BBSome and IFT-α losses in species
251 closely related to non-ciliated clades with complete IFT loss from (van Dam et al. 2013). Specifically,
252 we detected the loss of BBSome in basal plants (*Selaginella moellendorffii* and the moss *Physcomitrella*
253 *patens*) close to seed plants (*Spermatophyta*), in the apicomplexa *Sarcocystidae* (*Toxoplasma gondii*
254 clade) close to *Aconoidasida* (*Plasmodium falciparum* clade) and in the basal fungi *Chytridiomycetes*
255 (*Batrachochytrium dendrobatidis* clade) close to *Dikarya* and *Mucoromycota*. We also recovered the
256 loss of BBSome and IFT-α in the diatoms *Thalassiosira* close to *Phaeodactylum tricornutum*. Secondly,
257 we could identify other independent losses supporting the "last-in, first-out" hypothesis. In particular,
258 we found two losses of BBSome in the basal plants *Marchantia polymorpha* and *Klebsormidium*
259 *flaccidum*. We also identified complete IFT losses in another three apicomplexa clades (*Eimeria*,
260 *Cryptosporidium* and *Gregarina niphandrodes*) and two basal fungi clades (*Microsporidia* and
261 *Conidiobolus coronatus*). Moreover, we found evidence for losses of BBSome and IFT-α in four basal
262 fungi. While *Rozella allomycis* lacks all BBSome and IFT-α genes, *Piromyces sp.* and *Gonapodya*
263 *prolifera* were found with merely one IFT-α and two BBSome genes, respectively. *Allomyces*
264 *macrogynus* lacked BBSome. Finally, the presence of one IFT-α and two IFT-β genes in the
265 chlorophytes *Chlorella variabilis* close to *Ostreococcus* provides a new candidate replicate for this
266 "last-in, first-out" hypothesis. Its low number of IFT genes, which indicates dysfunctional cilia, could
267 be due to the endosymbiont nature of *Chlorella variabilis (Blanc et al. 2010)*.

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

The visual opsin gene repertoire correlates with adaptations in vertebrates

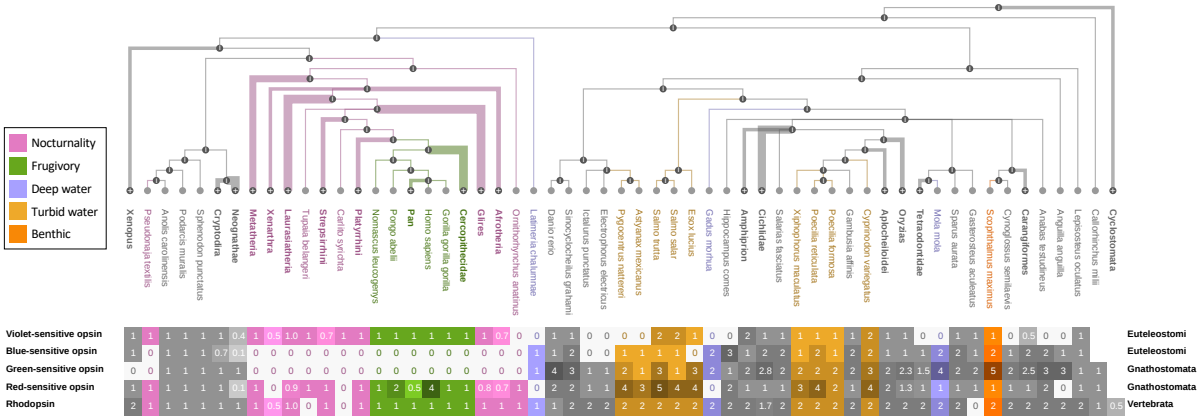


Figure 4. Visual opsin families in vertebrates (data from OMA All.Dec2021). Adaptations involved in textbook correlations with patterns of gene losses and duplications are annotated with separate colours. Nocturnality in snakes and mammals: loss of blue- and green-sensitive opsins. Frugivory in old-world 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. Ecological niche legends on the left do not belong to the Matreex layout but were added for figure clarity.

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 grey background and expansions on a darker one. Clades can also be coloured 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; Paul G. Higgs and Teresa K. 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 colour 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 grey 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 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. 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 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. 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 visualisation of gene families and illustrate its usefulness with three biological applications. First, we demonstrated Matreex's usefulness in delving into precise evolutionary analyses of multi-copy gene families by combining the gene tree with phylogenetic profiles. Secondly, 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 *Thelohanellus 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.

Funding

333 This work was supported by the Swiss National Foundation [167276] as part of the National Research
334 Program 75 'Big Data', as well as Swiss National Foundation [205085].

335 References

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

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

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

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

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

349 Borges R, Johnson WE, O'Brien SJ, Gomes C, Heesy CP, Antunes A. 2018. Adaptive genomic evolution
350 of opsins reveals that early mammals flourished in nocturnal environments. *BMC Genomics*
351 19:121.

352 Boussau B, Szöllosi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of
353 species and gene trees. *Genome Res.* 23:323–330.

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

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

360 Carvalho LS, Pessoa DMA, Mountford JK, Davies WIL, Hunt DM. 2017. The Genetic and Evolutionary

361 Drives behind Primate Color Vision. *Frontiers in Ecology and Evolution* 5:34.

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

366 van Dam TJP, Townsend MJ, Turk M, Schlessinger A, Sali A, Field MC, Huynen MA. 2013. Evolution of
367 modular intraflagellar transport from a coatomer-like progenitor. *Proc. Natl. Acad. Sci. U. S. A.*
368 110:6943–6948.

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

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

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

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

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

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

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

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

388 Katti C, Stacey-Solis M, Coronel-Rojas NA, Davies WIL. 2019. The Diversity and Adaptive Evolution of

389 Visual Photopigments in Reptiles. *Frontiers in Ecology and Evolution* [Internet] 7. Available from:
390 https://www.frontiersin.org/article/10.3389/fevo.2019.00352

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

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

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

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

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

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

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

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

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

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

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

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

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

425 Noutahi E, Semeria M, Lafond M, Seguin J, Boussau B, Guéguen L, El-Mabrouk N, Tannier E. 2016.
426 Efficient Gene Tree Correction Guided by Genome Evolution. *PLoS One* 11:e0159559.

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

429 Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions
430 by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.*
431 96:4285–4288.

432 Penel S, de Vienne DM. 2022. tidy tree: A New Layout for Phylogenetic Trees. *Mol. Biol. Evol.*
433 [Internet] 39. Available from: http://dx.doi.org/10.1093/molbev/msac204

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

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

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

442 Sadreyev IR, Ji F, Cohen E, Ruvkun G, Tabach Y. 2015. PhyloGene server for identification and
443 visualization of co-evolving proteins using normalized phylogenetic profiles. *Nucleic Acids Res.*
444 43:W154–W159.

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

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

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

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

455 Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, Hinrichs AS, Fernandes JD,
456 Borges R, Slodkowicz G, et al. 2020. Stability of SARS-CoV-2 phylogenies. *PLoS Genet.*
457 16:e1009175.

458 Wang Y, Zhou L, Wu L, Song C, Ma X, Xu S, Du T, Li X, Li J. 2021. Evolutionary ecology of the visual
459 opsin gene sequence and its expression in turbot (Scophthalmus maximus). *BMC Ecol Evol*
460 21:114.

461 Wickstead B, Gull K. 2007. Dyneins across eukaryotes: a comparative genomic analysis. *Traffic*
462 8:1708–1721.

463 Wu Z, Waneka G, Broz AK, King CR, Sloan DB. 2020. MSH1 is required for maintenance of the low
464 mutation rates in plant mitochondrial and plastid genomes. *Proc. Natl. Acad. Sci. U. S. A.*
465 117:16448–16455.

466 Xu S, Dai Z, Guo P, Fu X, Liu S, Zhou L, Tang W, Feng T, Chen M, Zhan L, et al. 2021. ggtreeExtra:
467 Compact Visualization of Richly Annotated Phylogenetic Data. *Mol. Biol. Evol.* 38:4039–4042.