Published online 20 April 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 2 **1**
doi: 10.1093/nargab/lqab024

CladeOScope: functional interactions through the prism of clade-wise co-evolution

Tomer Tsaban [1,†], Doron Stupp [1,†], Dana Sherill-Rofe [1], Idit Bloch [1], Elad Sharon [1], Ora Schueler-Furman [2], Reuven Wiener [3] and Yuval Tabach [1,*]

[1]Department of Developmental Biology and Cancer Research, Institute for Medical Research Israel-Canada and Hadassah Medical School, The Hebrew University of Jerusalem, Jerusalem 9112001, Israel, [2]Department of Microbiology and Molecular Genetics, Institute for Medical Research Israel-Canada and Hadassah Medical School, The Hebrew University of Jerusalem, Jerusalem 9112001, Israel and [3]Department of Biochemistry and Molecular Biology, Institute for Medical Research Israel-Canada and Hadassah Medical School,The Hebrew University of Jerusalem, Jerusalem 9112001, Israel

Received September 10, 2020; Revised March 12, 2021; Editorial Decision March 15, 2021; Accepted March 18, 2021

ABSTRACT

Mapping co-evolved genes via phylogenetic profiling (PP) is a powerful approach to uncover functional interactions between genes and to associate them with pathways. Despite many successful endeavors, the understanding of co-evolutionary signals in eukaryotes remains partial. Our hypothesis is that 'Clades', branches of the tree of life (e.g. primates and mammals), encompass signals that cannot be detected by PP using all eukaryotes. As such, integrating information from different clades should reveal local co-evolution signals and improve function prediction. Accordingly, we analyzed 1028 genomes in 66 clades and demonstrated that the co-evolutionary signal was scattered across clades. We showed that functionally related genes are frequently co-evolved in only parts of the eukaryotic tree and that clades are complementary in detecting functional interactions within pathways. We examined the non-homologous end joining pathway and the UFM1 ubiquitin-like protein pathway and showed that both demonstrated distinguished co-evolution patterns in specific clades. Our research offers a different way to look at co-evolution across eukaryotes and points to the importance of modular co-evolution analysis. We developed the 'CladeOScope' PP method to integrate information from 16 clades across over 1000 eukaryotic genomes and is accessible via an easy to use web server at http://cladeoscope.cs.huji.ac.il.

INTRODUCTION

Phylogenetic profiling (PP) predicts functional interactions between genes (or their products) by measuring the similarity of their evolutionary profiles (i.e. presence and absence/loss) across different organisms. The phylogenetic profile of a gene is represented by a vector across all examined species of whether a gene ortholog is present or absent in each organism (1,2). Genes that are lost and retained together are considered to be co-evolved and are predicted to be functionally related such that when the function is needed, the genes are retained together and lost otherwise. It is well established that genes with similar phylogenetic profiles are substantially more likely to be functionally related (1–9). This can be used to annotate uncharacterized proteins to a putative function or pathway, based on the similarity of their PP with those of annotated proteins. When expanding PP comparisons to the entire genome, it is expected to reveal functional linkages on a genome-wide scale, elucidating both known and novel pathways and cellular systems. Previous studies identified unknown functional associations of genes and were employed to discover unknown disease-causing genes (3,10,11) and new members in pathways (12–14).

Originally, a binary representation of presence or absence was used to describe the PP of genes. This approach was successfully applied to predict functional interactions in prokaryotes (1,2,15–17). Nevertheless, the implementation of the binary representation to eukaryotes might not be trivial (18–20). Eukaryotic genes are not enclosed in operons, usually encode proteins far larger (21) with several domains, undergo massive splicing and have more paralogs as compared to prokaryotic protein-coding genes. Other potentially influential differences are due to the variation in mutation rates and generation time between eukaryotes and prokaryotes. Thus, different levels of sequence identity, with

*To whom correspondence should be addressed. Tel: +972 54 3973641; Fax: +972 2 6757333; Email: YuvalTab@ekmd.huji.ac.il
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

partial loss, suggest variable selective pressures. Variable sequence identity may hint to differences in functional similarity between orthologs, which also depends on the phylogenetic distance between species (3,4).

As variable conservation levels suggest differences in selection and therefore in function, several methods, such as normalized phylogenetic profiling (NPP) (3,4,14,22–24) and SVD-phy (7) have been suggested. These methods offer an alternative to the binary presence-or-absence scoring system using a continuous metric of conservation. Continuous methods aim to model small evolutionary changes in a non-discrete (binary) way and with higher resolution (3,7,8). These PP methods identify genes that co-evolved 'globally' across the entire tree of life, or in some cases across all eukaryotic species, hereafter referred to as 'all eukaryotes'.

However, the PP of a gene represents a complex signal that reflects the integration of genomic events and many evolutionary phenomena that happened across millions of years of evolution at different scales (both at the molecular, organismal and population levels). At the molecular level, new genes appeared, genes duplicated, diverged in sequence or were completely lost across clades. These events happened sporadically and through the process of natural selection contributed to the creation or destruction of functional associations between genes. For example, ancient gene duplications may lead to sub-functionalization, specialization or neo-functionalization of one or more of the resulting paralogs (25), which can then distinctively co-evolve with new genes and pathways. These phenomena can be manifested as a tight co-evolution of genes in a specific pathway in some clades as well as divergence in their evolution in other clades. These and similar processes may drive co-evolution but can also mask co-evolution signals. Thus, genes sharing similar phylogenetic profiles, despite millions of years of complex evolution even in part of the tree of life, may suggest functional interactions between them.

The variability between clades and the complexity of evolution make it reasonable to assume that co-evolution of genes might not be reflected by the PP signals across the full tree of life. These signals may instead be hidden 'locally' in specific clades (14,26,27). 'Clade-wise' PP analysis aims to detect *local* co-evolution signals. A clade in this context refers to a monophyletic group, essentially a 'branch' of the phylogenetic tree, consisting of species stemming from the same common ancestor (e.g. Primates, Mammals, Fungi, etc.).

It has been shown that PP analyses can benefit from clade-wise measurements (14,26). Shin *et al*. (26) found that by integrating local co-evolution across domains of life (eukaryotes, bacteria and archaea), one could identify functional interactions that were not identified when performing PP using the whole tree of life. Additionally, Sherill-Rofe and Rahat *et al*. (14) showed how local co-evolution can be used within the eukaryotic tree to identify candidate DNA repair genes.

The concept of clade-wise co-evolution holds the promise of a more accurate functional interaction prediction as these results emphasize the value of inspecting co-evolution in specific clades (and not only globally across all organisms). Clade-based approaches, previously hampered by the lack of available data, are poised to become more accurate and more specific given the exponential growth in the number of sequenced organisms. Furthermore, combining various clades should improve the predictive power and our understanding of the biology in cases where the relevant clade is unclear.

To date, these expectations have not been demonstrated thoroughly and at a large scale. Since the concept was first introduced, studies were only applied in very restricted cases, such as a comparison between domains of life (26) and analyses for a DNA repair pathway within a few eukaryotic clades, and limited to model a large group of genes (14). A comprehensive study of clade-based PP is still lacking to explore the benefits and predictivity of each clade, the difference between the clades and analyses of clade integration.

Here we explore a broader perspective for the integration of clade-wise analysis in PP. We analyzed the co-evolution patterns of 186 KEGG pathways in 66 clades and 1028 eukaryotic species genomes and demonstrated how clades vary in their ability to detect different events of co-evolution of functional interactions. We established that clades are complementary in the prediction of functional interactions and propose a method to integrate clade-wise phylogenetic profiles, which we term 'CladeOScope'. The CladeOScope method and its accompanying web server are presented for the prediction and analysis of functional interactions between human genes using clade-wise PP analyses. Our method can identify signals that are harder to detect by global PP methods. Examples are provided where clade analyses improved performance and identified additional pathway genes, such as the UFM1 pathway where the signal was found in an unbiased fashion in two clades—all eukaryotes and Alveolates.

MATERIALS AND METHODS

Normalized phylogenetic profiling

The NPP matrix was prepared as previously described (3,4,14,22–24). Specifically, proteomes for all species were downloaded from UniProt (June 2018 release, reviewed proteomes) (28). The reference (Human) proteome was also downloaded from UniProt (June 2018 release, reviewed proteomes) and proteins with lengths of <40 amino acids were excluded. In cases where multiple isoforms were annotated for the same protein, we retained only the longest isoform. Species were annotated according to NCBI Taxonomy (29). In total, the matrix contains 20 192 human genes and 1028 species.

The NPP for each human gene is based on the similarity between the query (human) protein and the BLAST (30,31) best hit in each of the different species. One directional BLAST was shown to be highly sensitive in identifying orthologs for PP (32).

The alignment is local, thus a low score may stem from partial alignment (e.g. of a specific domain of a protein). To reduce noise, bitscores less than a threshold t were clipped to t, where $t = 20.4$. This bitscore threshold is the minimal bitscore value across all species that corresponds to an E-value ≤ 0.05.

The bitscore of each best BLAST hit was first normalized by the bitscore of the query protein self-hit (33). Then the

log2 of the normalized bitscore was taken. The log scores were normalized by the level of conservation for all proteins in the given specific proteome by Z-scoring of all scores for a specific species (3,4). Thus we based our global analysis on a ∼20 000 genes × ∼1000 species NPP matrix, where each data point $x_{a,b}$ is the normalized phylogenetic profile (NPP) score for gene *a* in species *b,* as compared to human. Based on the resulting NPP matrix, we constructed a ∼20 000 × 20 000 correlation matrix containing the Pearson correlation between the phylogenetic profiles of every gene pair for each clade.

An identical process was used to construct the phylogenetic profile for *Caenorhabditis elegans* (*C. elegans*, taxid 6239) presented in Supplementary Figure S3. The canonical *C. elegans* proteome was retrieved from UniProt at the same time and was used to construct a PP matrix similarly to human.

Filtering non-conserved genes

Some of the protein-coding genes were conserved only in certain clades. The normalized phylogenetic profile of such genes is mostly clipped (as explained earlier) and has low information value. To eliminate this artifact, we excluded from the analysis genes with a bitscore < 40 in 90% or more of the species inspected in a clade. This was performed upon analyzing all eukaryotes as well as in the separate analyses of each clade.

Clade annotation and representative clades

Using the NCBI taxonomy (29), we annotated each species to all clades it belonged to. We then filtered the clades such that only clades with more than 20 species were retained, resulting in 66 clades (a full list of clades is available in Supplementary Table S2). For each of the 66 clades, as well as for the set of all eukaryotes, we constructed a gene-wise Pearson correlation matrix and filtered non-conserved genes as described above.

From these clades, we chose 16 representative clades spanning the eukaryotic tree. To define the clade combination for which CladeOScope calculates the score, clades were chosen based on three guiding principles: wide coverage, mutual exclusivity and uniformness in clade types. To achieve wide coverage, clades were chosen to span most of the eukaryotic tree. Mutual exclusivity was achieved by choosing non-nested clades such that each species belonged to as few clades as possible. Uniformness is attributed to choosing clades with similar depth in the tree, e.g. kingdoms or phyla. Additionally, uniformness refers to each species belonging to a similar number of chosen clades. All three principles are conceptually important to avoid over- or under-representation of species. Based on these principles, we defined a combination of 16 clades, in addition to the set including all eukaryotes: Chordata, Ecdysozoa, Platyhelminthes, Alveolates, Stramenopiles, Fungi, Viridiplantae, Mammalia, Archelosuria, Arthropoda, Nematoda, Basidiomycota, Ascomycota, Fungi incertae sedis, Liliopsida and Eudicotyledons (and see Supplementary Table S1 for further information).

Broad clade-wise co-evolution analysis of KEGG pathways

We sought to compare the ability of different clades to identify functional interactions between genes belonging to the same pathway by clade-wise co-evolution. We utilized KEGG pathways (34) downloaded from MSigDB (35) (http://software.broadinstitute.org/gsea/downloads.jsp; version 28.11.2018). For *C. elegans*, KEGG pathways were downloaded from the KEGG API (http://rest.kegg.jp/link/cel/pathway, retrieved at 10 November 2020) and matched to UniProt ids using the mapping available from KEGG (http://rest.kegg.jp/conv/cel/uniprot, retrieved at 10 November 2020). Overall, the data contained 186 KEGG pathways. For each KEGG pathway, we calculated the recall of pairwise interactions in the pathway among the top 5% of correlations in a given clade. The recall here is defined as the number of pairwise interactions passing a given threshold, divided by the total number of interactions between pairs of genes belonging to the same pathway. A recall value of of 1 indicates that we identified pairwise interactions between each pair of genes in the pathway at a given threshold.

We compared the scores in clades for each pathway and calculated how many clades outperformed the 1028 genomes that represent all eukaryotes, and how many times each clade was best performing for all pathways. The data were visualized as a heatmap (Figure 1). Each row depicts a KEGG pathway, each column depicts a clade, and each entry is the recall as described above. Heatmaps were produced with the R package ComplexHeatmap (36).

Multi-clade integration

To investigate the utility of using multiple clades for functional interaction prediction, we developed a heuristic measuring the unique contribution of each clade to pathways. For each clade, gene pairs were considered co-evolved if their correlation was among the top 5% of correlations for that clade out of all possible gene–gene pairs (20 192 × 20 192). For each pathway, we calculated the proportion of connections in the pathway found in each clade (considering all possible gene-pairs). To optimize a combination of several clades to yield a maximum of connections, we applied greedy optimization. The first step was to identify the clade detecting the largest number of 'unique connections' (not detected by other clades). Then we removed these connections from the overall pool and repeated the same step for the rest of the clades. Thus, for each pathway, we assembled a ranked list of clades according to the proportion of connections that it uniquely identified in the pathway.

'CladeOScope' method for clade integrated co-evolution prediction

The 'CladeOScope' approach is designed to integrate the phylogenetic signal from clades to predict functional interactions. For a given query gene, CladeOScope first calculates the Pearson correlation between the phylogenetic profile of the query and the profiles of all ∼20 000 genes separately in 16 clades as well as 'all eukaryotes' (all eukaryotic species in our NPP matrix).

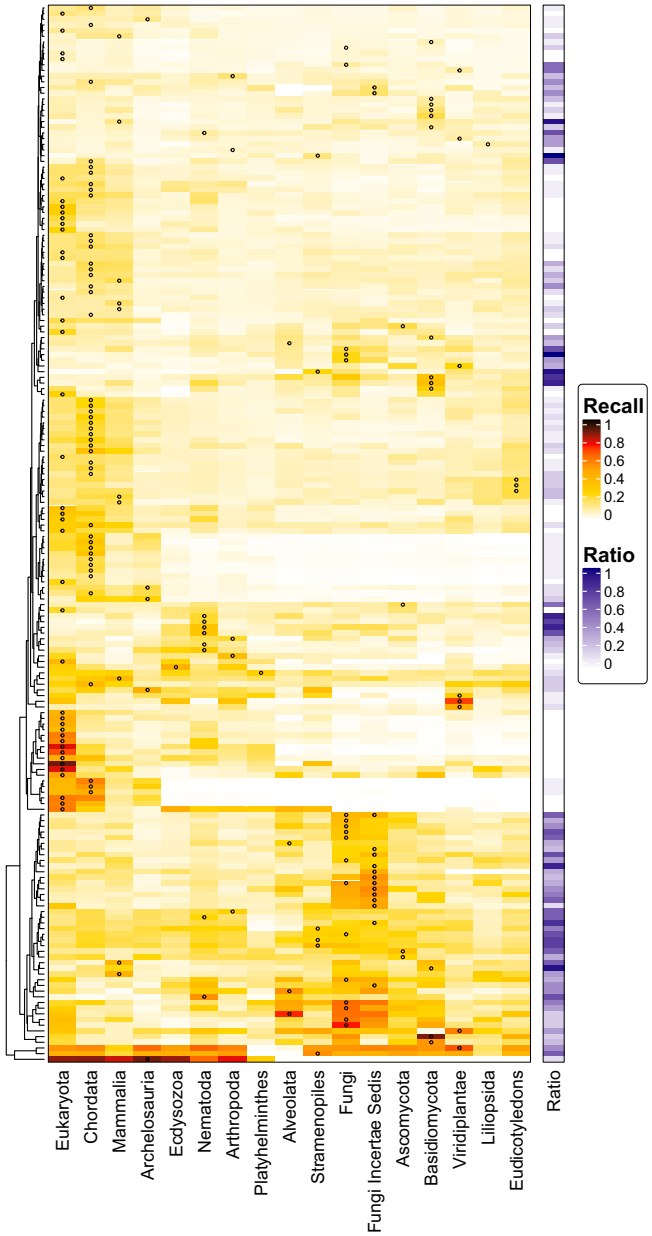

Figure 1. KEGG pathway prediction by clades. Heatmap demonstrating the performance of 17 clades (16 as defined in this study, as well as all eukaryotes) over 186 KEGG pathways. Each column depicts a clade while each row depicts a pathway. Each entry in the heatmap is colored by the percent of functional interactions in the pathway identified by the clade (see text). Dotted entries mark the best performing clade for each entry. The annotation bar 'Ratio' shows the fraction of clades that surpassed the score of all eukaryotes for each pathway (row).

CladeOScope then computes a *min rank* score, which is the best rank of correlation for each of the ∼20 000 genes across the clades. Thus, for a query gene q and a target gene h, the score is defined as follows: $S\ (q,h) = \min_c\ (\text{rank}(\rho_c(q,h)))$: where $\rho_c(q,h)$ is the Pearson correlation between the phylogenetic profiles of genes q and h in clade c, and $rank_c$ is the rank of this correlation in the vector of correlations between the query gene and all genes in clade c.

As ranks are not symmetrical (i.e. gene A may be ranked as the 10th most co-evolved for gene B, but gene B may be ranked only 30th for gene A), a geometric mean of the two ranks is used as a representative of the pair. As explained previously, non-conserved genes are excluded from the analysis (for each clade individually). Notably, since the score is based on the rank of the correlation, a lower score implies stronger evidence for co-evolution (the lower the rank, the higher the correlation).

In addition to the CladeOScope min rank score, for comparison we also computed the maximal correlation across all clades $\max(\rho_c(q,h))$, and a geometric mean across all clades.

Comparison to other methods

We benchmarked the ability of this integrated score, as well as the combination of clades, to predict functional interactions. Gene sets were gathered from CORUM (37) complexes, REACTOME (38) and KEGG (34) pathways. In addition to the KEGG pathways data described above, we downloaded CORUM complexes (https://mips.helmholtz-muenchen.de/corum/download/allComplexes.txt.zip; version 12 February 2019), and the Reactome database (reactome.org/download/current/ReactomePathways.gmt.zip; version 5 February 2019). CladeOScope was compared to the NPP method using all eukaryotes (3,4), as well as several other published PP methods including PrePhyloPro (8) and a binarized phylogenetic profile with hamming distance as the profile similarity (BPP hamming) (8,39). PrePhyloPro was calculated using the same BLAST data as the NPP as previously described, taking the BLAST E-Value as a measure of protein presence. BPP was similarly calculated from the BLAST *E*-value with a threshold of 10^{-3}.

CladeOScope web tool

The web server is implemented in HTML, CSS, JAVASCRIPT and R. It was built primarily using the R shiny package to embed R computations on both the server and client sides. The R packages used include iheatmapr (40) for interactive heatmaps, and ComplexHeatmap (36) for non-interactive heatmaps. The site is deployed using the Shiny Server software of R studio, hosted locally on a server of the Hebrew University of Jerusalem, and is available at http://cladeoscope.cs.huji.ac.il.

Paralog filtration

Paralogous genes present a known challenge in PP, both in the profile preparation step and in the analysis of co-evolved genes. Paralogous genes show similar phylogenetic profiles as they have a high sequence similarity. Although this is perhaps a true co-evolution signal, it is more easily captured by homology-based methods and thus of less interest in PP. Therefore, we highlight the existence of paralogs in the analysis and allow the user to filter out paralogous genes easily in the query or the results. This is done

by retaining the first gene of each pair of paralogous genes (discarding the second), based on the human paralogs lists as found in GeneCards (https://www.genecards.org/) (41), that are based on Homologene (https://www.ncbi.nlm.nih.gov/homologene/) (42) and Ensembl (https://www.ensembl.org/) (43).

RESULTS

Clades vary in their ability to detect co-evolution signals in different pathways

We determined whether PP performed using different clades improved the detection of known functional interactions as compared to using all eukaryotes. For the analysis, we examined the ability of 66 eukaryotic clades to detect functional interactions among members of 186 KEGG pathways. Genes in these pathways functionally interact by definition and are expected to show a co-evolution signal. For this analysis, we used 66 clades of 20 or more species at different levels of the eukaryotic tree (see 'Materials and Methods' section). For each clade we calculated the correlation matrix between all the genes and considered the top 5% gene pairs that had the highest correlations. The top 5% correlated gene pairs identified a significant proportion of functional interactions. Interestingly, different clades recover varying proportions of functional interactions in different pathways, with some clades performing better than others (Figure 1).

In some cases, different clades were grouped by biological context. For example, a cluster of metabolic pathways composed of 'KEGG Citrate Cycle TCA Cycle', 'KEGG Valine, Leucine and Isoleucine Degradation', 'KEGG Fatty Acid Metabolism' and 'KEGG Terpenoid Backbone Biosynthesis' was best predicted by Fungi, with 60–74% of the connections identified in Fungi for each pathway. However, this clade performed poorly for a cluster of different metabolic pathways. 'KEGG Glycosphingolipid Biosynthesis Lacto and Neolacto Series', 'KEGG Glycosphingolipid Biosynthesis Globo series' and 'KEGG Glycosphingolipid Biosynthesis Ganglio Series' detected around ∼0–7% of the pathway connections. These pathways in turn are well described by green plants (Viridiplantae), which detected up to 70% of the connections. Both of these clades, Fungi and Viridiplantae, performed poorly for a cluster of immunologic pathways of 'KEGG Graft Versus Host Disease', 'KEGG Asthma' and 'KEGG Type I Diabetes Mellitus'. For the latter cluster, while both Fungi and Viridiplantae detected ∼0–0.02% of the interactions in the pathways, they were well-detected by Chordata with 45–58% of the interactions.

Using only all eukaryotes at once is seldom optimal

Most previous PP approaches were based on using the full tree of life at once. We were interested to study the extent to which specific clades outperform functional interaction prediction in pathways as compared to using all eukaryotes. Analysis of 16 clades (see 'Materials and Methods' section) revealed that PP based on all eukaryotes had the best performance in only ∼20% of the pathways, (see row annotation bar in Figure 1). The Chordates clade (Chordata) had a better performance than all eukaryotes in ∼45% of the pathways, Fungi, Alveolates and Mammalia outperformed all eukaryotes in ∼35–25%, while the rest in ∼20% or less (Figure 2A). Platyhelminthes and Ecdysozoa surpassed eukaryotes the least, yet still for 11–12% of the pathways. This further suggests that the sole use of all eukaryotes may not be sufficient and that the addition of clade-specific information has a beneficial impact.

No single clade is optimal for all pathways

For each pathway, we computed which clade had the top score. In Figure 1, the dots (marking the best performing clade for each pathway) show a highly versatile pattern, suggesting that no one or two clades can be used to predict connections optimally in all KEGG pathways. Among the examined 66 clades (Figure 2B), those that scored the best for most pathways were vertebrates (∼18% of the pathways), followed by all eukaryotes (17.7%), fungi and metazoa (∼8% each), chordata (∼6%) and the rest for 5% or less. These results suggest that no specific clade captures all pathways. We then checked whether combining different clades had the potential to improve the detection of functional interactions.

Clades are complementary in predicting functional interactions

While different clades may detect a substantial proportion of a pathway's interactions, these interactions may overlap to various extents. Hence, we wanted to test whether using a variety of different clades could lead to complementary predictions such that, when combining co-evolved genes from several clades, one could better reconstruct the pathway. Combining information from different clades might have a significant effect on reconstructing a pathway as the best clade (which identifies the most unique pairwise connections) on average only predicts 31% of the connections in the pathway (Figure 3A). Using only all eukaryotes performed even worse, with around 15% of the connections identified (of a pathway) (Figure 3B). However, by combining the top five clades per pathway, it predicted up to 52% of connections at the fifth top clade as compared to 20% of connections in random gene sets (Figure 3A, orange). Overall, KEGG pathways identification benefits from further integrating the next clade in each of the top five clades (Figure 3A). By inspecting all 16 chosen clades, we found that this effect saturates at about the eighth best clade with ∼60% of connections identified per pathway on average (Figure 3B).

Some clades appear to be more informative than other clades and thus are listed more often as the top clades per pathway. Not surprisingly, the two prevailing clades are all eukaryotes and Chordata, as described above. However, some other clades such as Mammalia and Nematoda tend to be top-ranking clades as well (Figure 3C).

Recapitulation of pathway components using the combined local co-evolution approach

Recently we identified nine novel genes in the homologous recombination repair pathway using a simple clade-based

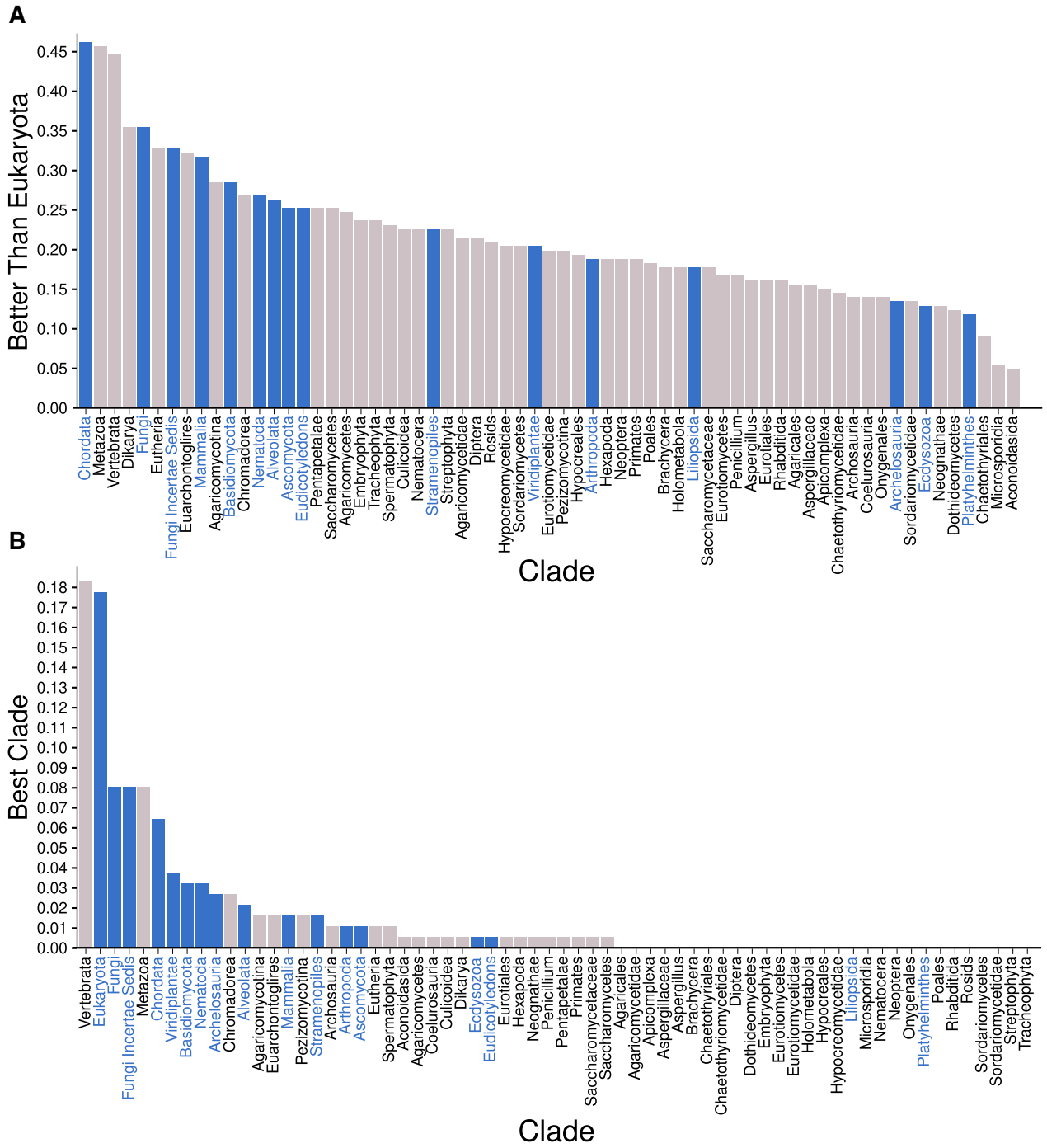

Figure 2. Comparison of clade prediction of KEGG pathways. (**A**) Different clades surpass the all eukaryotes score in predicting KEGG pathways. This plot demonstrates how in many pathways each clade scored higher than all eukaryotes. (**B**) The fraction of KEGG pathways for which each clade had the best score out of all 66 examined. For panels (A) and (B), the *x*-axis shows 66 examined clades ranked by performance, and clades selected for our method are marked in blue. The *y*-axis depicts the ratio of KEGG pathways in which each clade scored higher than all eukaryotes in panel (A) and the ratio of KEGG pathways for which each clade was the top scoring in panel (B).

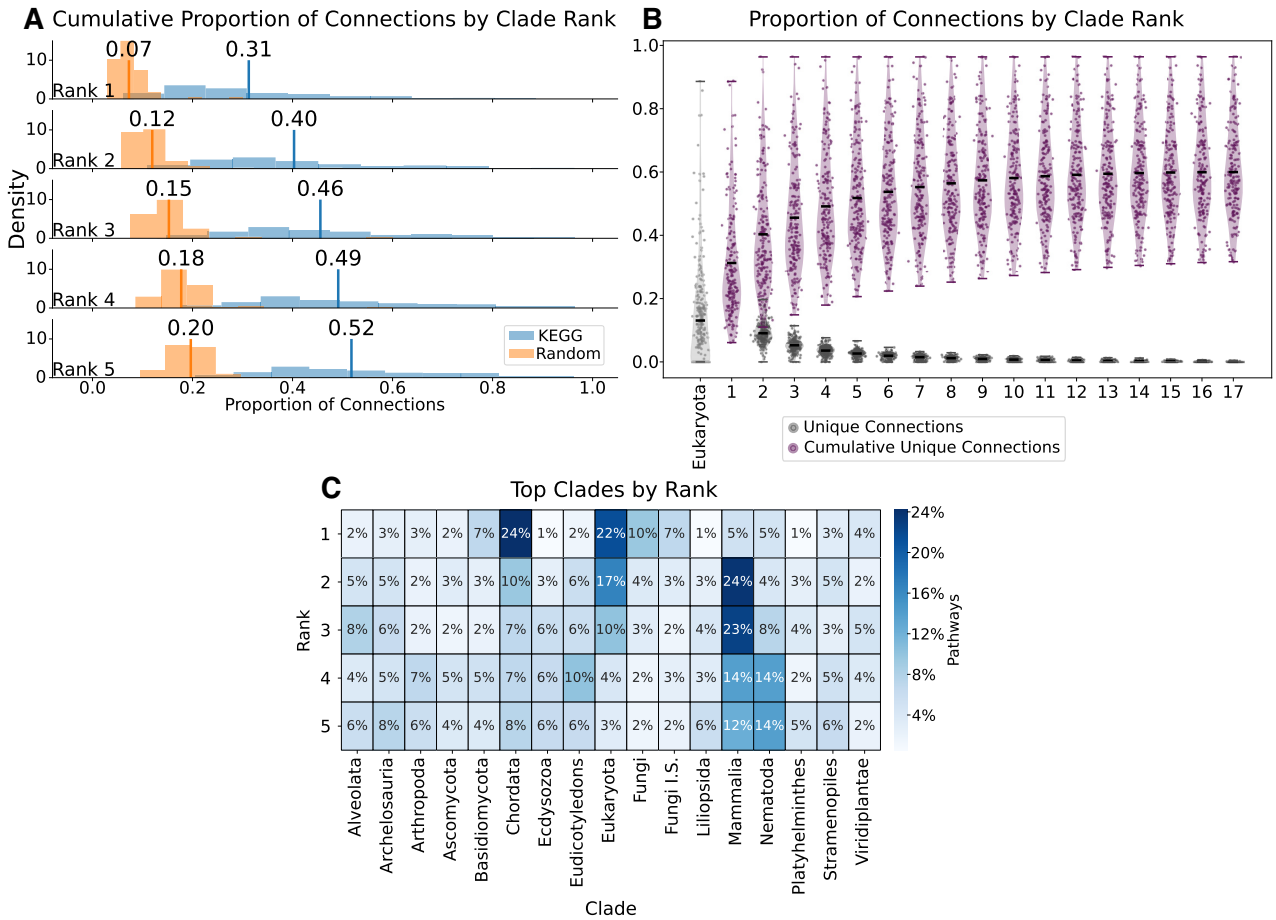

Figure 3. Clades are complementary in predicting functional interactions. Clades were used to predict functional interactions in KEGG pathways. (**A**) The recall for the top five clades per pathway (blue) compared to random gene sets (orange). The x-axis indicates the recall—i.e. the proportion of unique interactions identified (not identified by other clades). Histograms are ordered from top to bottom by the number of clades used for prediction. Vertical line is the mean of the distribution, with the value written above. (**B**) Performance of clades and groups of clades, ranked from best to worst per pathway, in predicting unique interactions (such that they are only predicted by a specific clade). All eukaryotes are shown for reference (light gray). For each rank, the proportion of unique connections (dark gray) and cumulative connections (purple) is shown. For each violin plot, the lines at the top and bottom are the min and max appropriately, while the black line in the middle is the mean. (**C**) Heatmap representing the percentage of pathways for which a specific clade (column) is ranked first to fifth (row).

PP approach (14). This pathway is important for repairing double-strands breaks and has a major role in cancer. In the present study, we demonstrated how the exploration of another DNA repair pathway, the non-homologous end joining pathway (NHEJ), might benefit from using clades. When inspected using all eukaryotes (top 5% interactions), NHEJ genes showed poor co-evolution with only the three DNA polymerase genes POLL, POLM and DNTT predicted to be co-evolved (Figure 4A). However, the top five clades (top 5% interactions in each clade, with redundancy) showed high connectivity between the 12 NHEJ genes and identified 68 interactions (51 unique interactions), 23-fold (17-fold unique) more than using all eukaryotes only. While most of the connections are predicted by the top clade, Nematoda in this case, some of these interactions seem to be specifically identified by other clades. One such example is the set of genes NHEJ1, FEN1, PRKDC and XRCC4, which are connected in Mammalia (Figure 4B).

However, this analysis integrates more interactions (top 5% of interactions in five clades) as compared to all eukaryotes (top 5% in a single clade). To compare the same number of interactions, we recapitulated the network using the CladeOScope method (see 'Materials and Methods' section). The interactions were scored by the minimal rank achieved for each pair across all clades. Thus the best (bottom) 5% of minimal ranks is taken. The CladeOScope-based analysis used the same number of interactions as in all eukaryotes but identified 26 interactions (eight times more than all eukaryotes) among the NHEJ genes (Figure 4C).

An additional pathway that is well reconstructed is glycosphingolipid biosynthesis, which underlies Tay Sachs disease. All eukaryotes identified a few disconnected subgroups of genes (Figure 4D). The top five clades connect all these genes, again revealing some subgroups of genes, such as HEXA, HEXB, FUT1 and FUT2, which are connected by Arthropoda (orange) and Nematoda (Figure 4E, turquoise). Figure 4F presents the network recapitulation using the CladeOScope method, which identified more interactions between the genes than all eukaryotes.

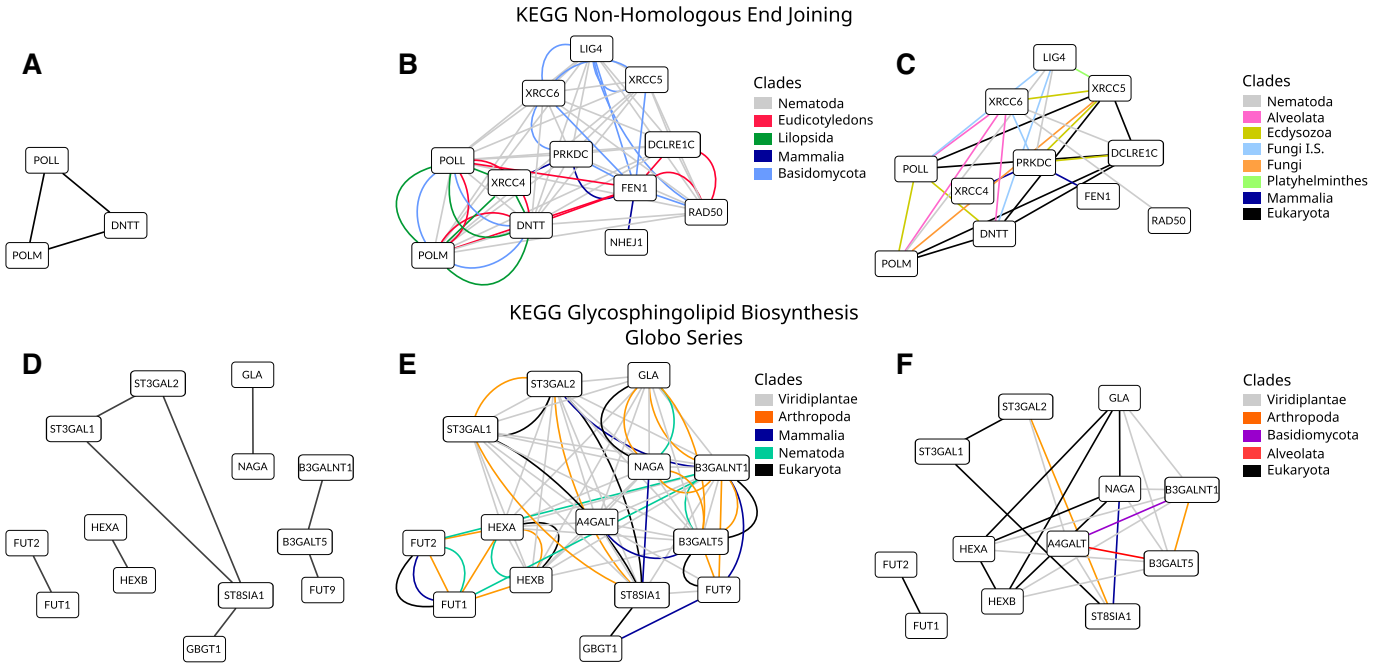

Figure 4. The utility of using clade-wise PP as demonstrated on specific pathways. The network of interaction between pathway genes is shown for two pathways, KEGG NHEJ (**A**–**C**) and KEGG glycosphingolipid biosynthesis globo series (**D**–**F**). For each pathway, the network spanned by interactions found in all eukaryotes is shown on the left (A and D, in black), the network spanned by the top five clades is shown in the middle (B and E, edges colored by clade) and on the right the network spanned by the CladeOScope method (C and F, based on minimal rank over all clades for each interaction; edges colored by clade). Light gray represents the top clade in each example by the top five combination method. On the right, a color legend is included to highlight the clades used for identification of connections.

The CladeOScope method for clade integrated phylogenetic profiling for functional interaction prediction

Different clades harbor co-evolution-based interaction signals for different genes. To harness the unique predictions of each clade for different pathways, we developed the CladeOScope method to provide a ranked list of the most correlated genes with the input gene, integrating data from different clades. We developed a score that ranks all the genes across all the clades. This score integrates 16 clades and all eukaryotes and represents how co-evolved a gene is with the query gene. To calculate it, we first calculate the Pearson correlation between each gene-pair in each clade. We then rank for each gene those genes that are the most correlated with it per clade. Finally, for each gene pair the score corresponds to the minimal (best) rank they achieve across all clades (see 'Materials and Methods' section). This scoring system was benchmarked and was shown to outperform other possible clade integration approaches such as the maximal correlation across all clades and a geometric mean across all clades, which were inferior to the min rank in performance (See Supplementary Figure S1).

We further experimented with different clade combinations (Supplementary Figure S2) and found that the combination shown in Supplementary Table S1 performs best. Supplementary Figure S2 presents its performance (Comb. 5) as compared to other existing approaches and several other clade combinations (Comb. 1–4, all satisfying the principles discussed above, see 'Materials and Methods' section).

To assess the ability of our and other PP approaches to predict functional interactions, we utilized known pathways from KEGG, REACTOME and protein complexes from CORUM (see 'Materials and Methods' section). We compared these methods using ROC curves (Figure 5A–C) and partial ROC curves (Figure 5D–F). Encouragingly, we found that CladeOScope could predict functional interactions with high performance, achieving an AUROC of 0.758 for KEGG on which it was optimized and a similar, albeit slightly reduced, performance for other databases; 0.725 for CORUM, and 0.692 in REACTOME. Examples of pathway recapitulation by CladeOScope were discussed above (see Figure 4C and F). These results demonstrate that CladeOScope outperformed other PP methods, both continuous and binary representation-based approaches, in utilizing gene co-evolution to predict functional interactions.

A similar analysis was performed for *C. elegans* KEGG pathways, showing ROC (Supplementary Figure S3A) and partial ROC curves (Supplementary Figure S3B). This analysis shows that the CladeOScope approach outperformed the other PP approaches as for human. While the difference is small, a different combination outperformed the rest for *C. elegans*. Specifically Comb. 0, where all 66 clades are used, had a slightly better performance than Comb. 5, which was described above as the best performing combination for human.

CladeOScope web tool

The primary aim of the CladeOScope web tool is to make the analysis of 1028 organisms and the 16 most informa-

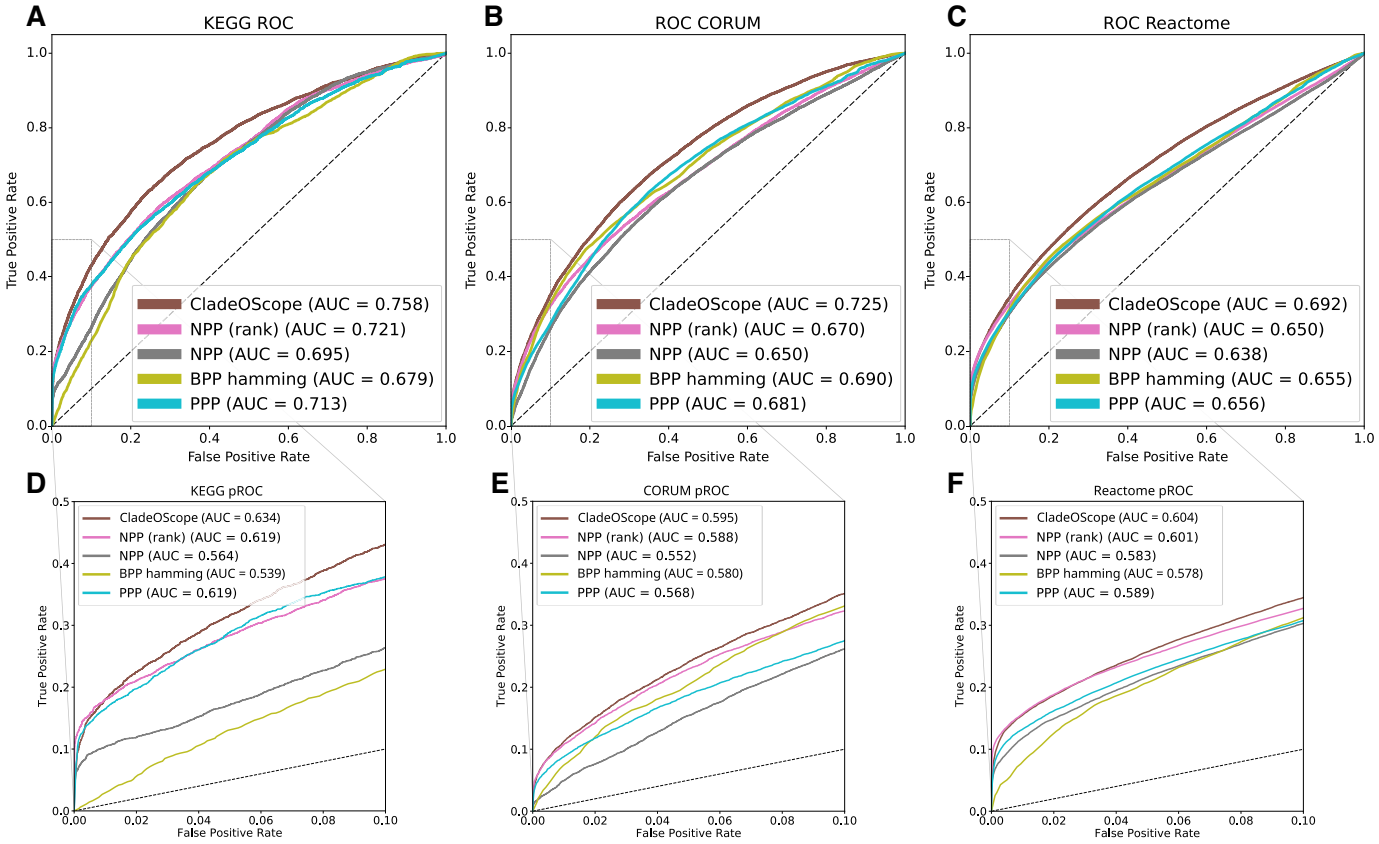

Figure 5. ROC curves for prediction of functional interactions. Prediction of functional interactions by the CladeOScope method was compared to the prediction using four other PP approaches—NPP with rank of correlation (NPP (rank)), NPP, Binarized PP with Hamming distance (BPP Hamming) and PrePhyloPro (PPP). The comparison was performed for predicting functional interactions (gene co-occurrence in KEGG pathways (**A** and **D**), CORUM complexes (**B** and **E**) and Reactome pathways (panels **C** and **F**)). Comparison is shown as ROC curves (A–C) with corresponding partial ROC curves where FPR < 0.1 (D–F, demarcated as dashed rectangle in A–C). TPR was adjusted for visibility. ROC: receiver operator characteristics; pROC: partial ROC; AUC: area under the curve; FPR: false positive rate; TPR: true positive rate.

tive clades simple and accessible. The goal is that every user, even without computational skills, will be able to map the evolution of their gene of interest and identify genes that most significantly co-evolved with it in each clade. It also provides a score that accumulates the information from all clades and points to the most promising candidates.

The interface is simple. The user submits a query gene and the web tool performs the aforementioned analyses and returns the list of co-evolved genes in 16 clades and the combined CladeOScope score. CladeOScope also includes a variety of useful visualizations of the co-evolutionary data, including a correlation heatmap and the clustered phylogenetic profiles of the query gene with its co-evolved genes. Additionally, the web tool enables pre-processing of gene-sets by paralog filtration for further study, e.g. for enrichment analysis. Paralogous genes are highly co-evolved by definition, however this signal is easily captured using homology-based methods and thus of less interest in PP. CladeOScope thus enables the optional filtering of paralogs present in a gene set to help identify functional interactions between non-paralogous genes. The webserver interface was used to identify the UFM1 components described below.

The UFM1 pathway can be accurately detected using CladeOScope

We present an example for using the CladeOScope web tool to predict genes associated with ubiquitin-like protein UFM1 (ubiquitin-fold modifier-1). Similar to ubiquitin, UFM1 modifies target proteins via a three-enzyme cascade involving E1, E2 and E3. However, in contrast to ubiquitin with tens of E2s and hundreds of E3s, UFM1 has a single E1 (UBA5), E2 (UFC1) and E3 (UFL1) (46). Currently little is known about the mechanism of UFL1 E3 ligase activity but recently it was shown that it functions in a complex with two additional proteins, DDRGK1 and CDK5RAP3 (47). Finally, like other post-translational modifications that are reversible, UFM1 is removed from the target protein by the UFM1 specific proteases, UFSP2 and UFSP1 (48). Taken together, although little is known about the role of protein modification by UFM1, this system includes seven proteins that, together, comprise the machinery needed for the deposition and removal of UFM1 from target proteins.

Querying UFL1 in CladeOScope single gene analysis identified 6/7 of its known partners, UFSP2, UFSP1, UFC1, UFM1, UBA5 and CDK5RAP3, in the top 15 genes in Alveolata and Alveolata only. The seventh part-

ner, DDRGK1, ranks 73rd, also in Alveolata (Figure 6B). In terms of the CladeOScope score, three were detected in the top 20, two in the top 50 and the rest in the top 100 results. Interestingly, when querying for UFC1, UFM1 or any other partner of the system, CladeOScope was able to detect all other partners specifically when combining the top scores of two clades, all eukaryotes and Alveolata (the top 100 in Alveolata intersected with the top 100 in all eukaryotes; see Figure 6A). The phylogenetic profile of this pathway (Figure 6C) identified Alveolates as having a strong signal as compared to other clades. These genes are lost throughout the Fungi clades (44) with only a partial signal in the Animalia clades. However, Alveolates show several concordant loss events across the pathway's genes. These loss events in the Plasmodium genus as well as several other species (demarcated in Figure 6D) contribute a strong co-evolutionary signal to the prediction of these genes as functionally interacting. Previous works have described the loss of UFM1 and other ubiquitin-like proteins in Alveolates and discussed their potential for therapeutics development (45).

In such cases, where other genes in the pathway are known, co-occurrence of known genes directed us to the relevant clade to search for functionally related genes. In different situations where the relevant clade cannot be identified, the CladeOScope algorithm suggests the relevant genes and clades. This two-level search further highlights the applicability of the CladeOScope web tool for research as it is very simple to use, accurate, and allows the user to explore new depths of information.

DISCUSSION

We built upon previous works that suggested the importance of clade-based PP in order to perform a thorough clade signal analysis and improve PP methods. We studied the complex co-evolution of eukaryotic genes under the hypothesis that different genes may show co-evolution signals in certain clades but not in others. We highlighted the evolutionary signal found in various clades of the eukaryotic tree of life. Moreover, we showed that clades differentially specialize in detecting functional interactions in different pathways as, for some functionally related genes, co-evolution is only detectable in some parts of the tree of life.

Overall, gene evolution is a complex process with an interplay between the evolution at genes, trait, organism and population levels, as well as co-evolution of species, and interaction with the changing environment. As such, while it is reasonable to assume that genes co-evolution is common across evolution, the assumption that it can be fully represented by a PP signal across all eukaryotes may be simplistic. Dealing with these complexities is the main challenge in understanding PP signals. Previous attempts to overcome these challenges in PP include the introduction of several metrics (6,8) or approaches (5). We showed that our clade-based analysis improved the predictive power of PP, with higher sensitivity and more accurate quantification of interactions. Furthermore, as the clade analysis is more biologically sound, it detected pathways that were previously not considered to be co-evolved. Using CladeOScope, we integrated co-evolution signals across clades and found that many pathways are better predicted with specific clades instead of using all eukaryotes. Moreover, no specific clade, including all eukaryotes, can cover the entire breadth of human pathways.

Our analysis showed that some clades are better suited to predict functional interactions between specific human genes. One hypothesis is that the nature of the pathway dictates the clade in which it will be detected. For example, we showed that some metabolic pathways were well predicted by distant species (Alveolates) while some immune pathways were found by closer species (Ecdysozoans). A different prism concerns the variability and thus informativeness of genes in specific clades. For example, some genes are found only in metazoans and thus are non-informative to study across all eukaryotes or distant species. In other cases, functionally related genes were tightly co-evolved in some clades, but their co-evolution dissolved in clades where the function was not relevant.

However, more complex evolutionary trajectories can also take part in local co-evolution. Gene duplication and sub-functionalization may lead one of the resulting paralogs to co-evolve with genes related to a different function, while in distant clades (prior to the duplication) they will share a single ortholog. As can be seen, these processes and others may both amplify the co-evolutionary signal in a specific clade, or mask it.

Our findings raise several questions for future research. First, we do not fully understand the meaning of co-evolution in a specific clade nor how to optimize its analysis. It remains unclear how to best combine these clade-specific signals. Furthermore, like other PP methods, we also observed high false-positive rates. This may be due to the inherent difficulties of PP methods in Eukaryotes in comparison to Prokaryotes where evolutionary distance is larger and genome architecture is different, i.e. operons, plasmids, and more common horizontal gene transfers etc. Further improvement may be achieved using machine learning methods. Additionally, other factors may affect performance of PP. For example, isoform selection for a gene may in turn affect the similarity calculated with respect to its orthologs and their isoforms. While the canonical isoform used in this study constitutes a sensible baseline, PP may benefit from a thorough isoform selection or harmonization scheme.

The primary goal of our research was to emphasize the importance of the co-evolution signal in the relevant clade for each pathway. The utilization of multiple clades as well as an increasing number of species allows for a more intricate exploration of pathway co-evolution. For cases where the appropriate clades for analysis are unknown, our method scores interactions by clade-relevance. This may bridge the gap between the use of PP to predict functional interactions and studies in comparative genomics in the co-evolution of specific pathways. For this purpose, we present our web tool, which enables the user to perform clade-wise co-evolutionary analyses throughout 16 eukaryotic clades.

Overall, the concept of searching for co-evolution in multiple clades is still in its infancy and further research is required to extract the maximum information from the teeming amount of genomic data. In addition, this research raises a set of new questions related to the crosstalk be-

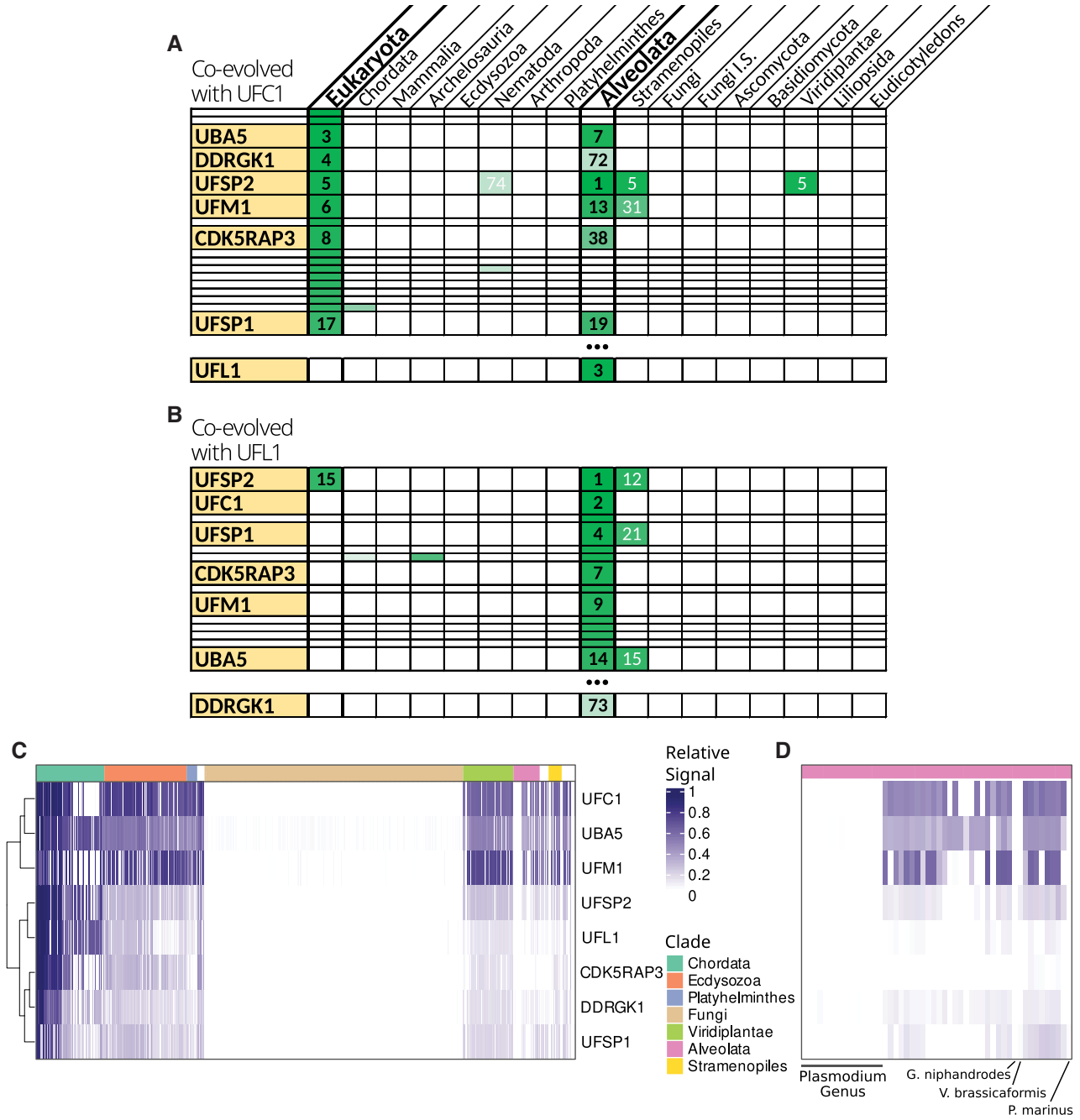

Figure 6. CladeOScope results for UFM1 pathway genes UFC1 and UFL1. (**A**) CladeOScope results for UFC1 gene as obtained by the web tool. It is clear that the genes of the pathway show a pattern of co-evolution in both alveolates and all eukaryotes. Most of the genes of the pathway were detected within the top 15 ranks while a few were detected lower in ranks 20–72. Each row depicts a gene with known genes of the pathway colored in yellow. Each column stands for a clade in which the gene was inspected. Values in cells indicate the rank of a gene in a clade (lower is better, 1 is best). Ranks greater than 100 were omitted and presented as a blank cell. Genes are sorted by ascending rank on all eukaryotes. (**B**) Similar results were obtained for the gene UFL1 of the UFM1 pathway. This time the only clade detecting the rest of the partners was alveolates with 6/7 in the top 14 ranks, and 7/7 in rank 73. Genes are sorted by ascending rank in Alveolata. Clade (column) order is shared across (A) and (B). (**C**) Phylogenetic profiles of genes in the UFM1 pathway. Color scale depicts the relative signal as a min–max gene-wise scaled profile. The profiles are self-hit normalized bitscores as described in the 'Materials and Methods' section. The top bar annotation describes the clades to which each species (column) belongs. (**D**) An enlarged view of the Alveolata clade; row (gene) order is preserved across (C) and (D).

tween gene PP, clades and species. We believe our work extends the understanding of co-evolution in the clade and global prism. It systematically and comprehensively explored clade-wise co-evolution of pathways and its broad application to functional interaction prediction. With ever-growing species sequencing data, these ideas will enhance our understanding of how human genes interact and evolve.

DATA AVAILABILITY

Code and a reproducible example are available under MIT license through GitHub: https://github.com/dst1/CladeOScope.

Data for Figures 1–4 and 6 are available through the source data file. Figures 5 and Supplementary Figures S1 and S2 can be generated by the reproducible code example. The PP matrices required to reproduce the method are available through: https://zenodo.org/record/4464120#.YEoFpHmxVPY.

Data are available under a CC0 license.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Author Contribution: T.T., D.S., D.S.R. and Y.T. conceived the idea for the work. Y.T., T.T. and D.S. developed the CladeOScope method. T.T. and D.S. performed all the analyses presented in this work. T.T. developed the CladeOScope web tool. T.T., D.S., I.B. and E.S. prepared the data for the web tool and analyses. T.T., D.S.R., D.S. and Y.T. designed the web tool. T.T., D.S., D.S.R., I.B., O.S.F., R.W. and Y.T. wrote the manuscript. Y.T. supervised the research. All authors read and commented on the manuscript.

FUNDING

This work was supported, in whole or in part, by the Israel Science Foundation, founded by the Israel Academy of Science and Humanities (grant number 1591/19 to Y.T and grant number 717/2017 to O.S.-F); Israel Innovative Authority (grant number 71420 to Y.T); SOYKA Pancreatic Cancer Project (grant number 5001346 to Y.T).
Conflict of interest statement. None declared.

REFERENCES

1. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl Acad. Sci. U.S.A.*, **96**, 4285–4288.
2. Date,S.V. and Marcotte,E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.*, **21**, 1055–1062.
3. Tabach,Y., Golan,T., Hernández-Hernández,A., Messer,A.R., Fukuda,T., Kouznetsova,A., Liu,J., Lilienthal,I., Levy,C. and Ruvkun,G. (2013) Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. *Mol. Syst. Biol.*, **9**, 692.
4. Tabach,Y., Billi,A.C., Hayes,G.D., Newman,M.a, Zuk,O., Gabel,H., Kamath,R., Yacoby,K., Chapman,B., Garcia,S.M. *et al.* (2013) Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature*, **493**, 694–698.
5. Li,Y., Calvo,S.E., Gutman,R., Liu,J.S. and Mootha,V.K. (2014) Expansion of biological pathways based on evolutionary inference. *Cell*, **158**, 213–225.
6. Dey,G., Jaimovich,A., Collins,S.R., Seki,A. and Meyer,T. (2015) Systematic discovery of human gene function and principles of modular organization through phylogenetic profiling. *Cell Rep.*, **10**, 993–1006.
7. Franceschini,A., Lin,J., von Mering,C., Jensen,L.J., Mering,C.V. and Jensen,L.J. (2016) SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics*, **32**, 1085–1087.
8. Niu,Y., Liu,C., Moghimyfiroozabad,S., Yang,Y. and Alavian,K.N. (2017) PrePhyloPro: phylogenetic profile-based prediction of whole proteome linkages. *PeerJ*, **5**, e3712.
9. Li,Y., Ning,S., Calvo,S.E., Mootha,V.K. and Liu,J.S. (2018) Bayesian Hidden Markov Tree Models for clustering genes with shared evolutionary history. *Ann. Stat.*, **46**, 1721–1741.
10. Arkadir,D., Lossos,A., Rahat,D., Snineh,M.A., Schueler-Furman,O., Nitschke,S., Minassian,B.A., Sadaka,Y., Lerer,I., Tabach,Y. *et al.* (2019) MYORG is associated with recessive primary familial brain calcification. *Ann. Clin. Transl. Neurol.*, **6**, 106–113.
11. Omar,I., Guterman-Ram,G., Rahat,D., Tabach,Y., Berger,M. and Levaot,N. (2018) Schlafen2 mutation in mice causes an osteopetrotic phenotype due to a decrease in the number of osteoclast progenitors. *Sci. Rep.*, **8**, 13005.
12. Avidor-Reiss,T., Maer,A.M., Koundakjian,E., Polyanovsky,A., Keil,T., Subramaniam,S. and Zuker,C.S. (2004) Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell*, **117**, 527–539.
13. Baughman,J.M., Perocchi,F., Girgis,H.S., Plovanich,M., Belcher-Timme,C.A., Sancak,Y., Bao,X.R., Strittmatter,L., Goldberger,O., Bogorad,R.L. *et al.* (2011) Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. *Nature*, **476**, 341–345.
14. Sherill-Rofe,D., Rahat,D., Findlay,S., Mellul,A., Guberman,I., Braun,M., Bloch,I., Lalezari,A., Samiei,A., Sadreyev,R. *et al.* (2019) Mapping global and local coevolution across 600 species to identify novel homologous recombination repair genes. *Genome Res.*, **29**, 439–448.
15. Škunca,N. and Dessimoz,C. (2015) Phylogenetic profiling: how much input data is enough? *PLoS One*, **10**, e0114701.
16. Sferra,G., Fratini,F., Ponzi,M. and Pizzi,E. (2017) Phylo_dCor: distance correlation as a novel metric for phylogenetic profiling. *BMC Bioinformatics*, **18**, 396.
17. Liu,C., Wright,B., Allen-Vercoe,E., Gu,H. and Beiko,R. (2018) Phylogenetic clustering of genes reveals shared evolutionary trajectories and putative gene functions. *Genome Biol. Evol.*, **10**, 2255–2265.
18. Singh,S. and Wall,D.P. (2008) Testing the accuracy of eukaryotic phylogenetic profiles for prediction of biological function. *Evol. Bioinforma.*, **4**, 217–223.
19. Snitkin,E.S., Gustafson,A.M., Mellor,J., Wu,J. and DeLisi,C. (2006) Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics*, **7**, 420.
20. Jothi,R., Przytycka,T.M. and Aravind,L. (2007) Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics*, **8**, 173.
21. Tiessen,A., Pérez-Rodríguez,P. and Delaye-Arredondo,L.J. (2012) Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Res. Notes*, **5**, 85.
22. Sadreyev,I.R., Ji,F., Cohen,E., Ruvkun,G. and Tabach,Y. (2015) PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. *Nucleic Acids Res.*, **43**, W154–W159.
23. Braun,M., Sharon,E., Unterman,I., Miller,M., Shtern,A.M., Benenson,S., Vainstein,A. and Tabach,Y. (2020) ACE2 co-evolutionary pattern suggests targets for pharmaceutical intervention in the COVID-19 pandemic. *iScience*, **23**, 101384.
24. Bloch,I., Sherill-Rofe,D., Stupp,D., Unterman,I., Beer,H., Sharon,E. and Tabach,Y. (2020) Optimization of co-evolution analysis through

phylogenetic profiling reveals pathway-specific signals. *Bioinformatics*, **36**, 4116–4125.
25. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
26. Shin,J. and Lee,I. (2015) Co-inheritance analysis within the domains of life substantially improves network inference by phylogenetic profiling. *PLoS One*, **10**, e0139006.
27. Dey,G. and Meyer,T. (2015) Phylogenetic profiling for probing the modular architecture of the human genome. *Cell Syst.*, **1**, 106–115.
28. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
29. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
30. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
31. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
32. Cheng,Y. and Perocchi,F. (2015) ProtPhylo: identification of protein–phenotype and protein–protein functional associations via phylogenetic profiling. *Nucleic Acids Res.*, **43**, W160–W168.
33. Enault,F., Suhre,K., Poirot,O., Abergel,C. and Claverie,J.-M. (2004) Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Res.*, **32**, W336–W339.
34. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
35. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdottir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
36. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
37. Giurgiu,M., Reinhard,J., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Ruepp,A. (2019) CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res.*, **47**, D559–D563.
38. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
39. Kensche,P.R., van Noort,V., Dutilh,B.E. and Huynen,M.A. (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *J. R. Soc. Interface*, **5**, 151–170.
40. Schep,A.N. and Kummerfeld,S.K. (2017) iheatmapr: Interactive complex heatmaps in R. *J. Open Source Softw.*, **2**, 359.
41. Safran,M., Dalah,I., Alexander,J., Rosen,N., Iny Stein,T., Shmoish,M., Nativ,N., Bahir,I., Doniger,T., Krug,H. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database*, **2010**, baq020.
42. Acland,A., Agarwala,R., Barrett,T., Beck,J., Benson,D.A., Bollin,C., Bolton,E., Bryant,S.H., Canese,K., Church,D.M. *et al.* (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **42**, D7–D17.
43. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
44. Daniel,J. and Liebau,E. (2014) The Ufm1 Cascade. *Cells*, **3**, 627–638.
45. Ponts,N., Yang,J., Chung,D.-W.D., Prudhomme,J., Girke,T., Horrocks,P. and Roch,K.G.L. (2008) Deciphering the ubiquitin-mediated pathway in apicomplexan parasites: a potential strategy to interfere with parasite virulence. *PLoS One*, **3**, e2386.
46. Komatsu,M., Chiba,T., Tatsumi,K., Iemura,S., Tanida,I., Okazaki,N., Ueno,T., Kominami,E., Natsume,T. and Tanaka,K. (2008) A novel protein-conjugating system for Ufm1, a ubiquitin-fold modifier. *EMBO J.*, **23**, 1977–1986.
47. Wu,J., Lei,G., Mei,M., Tang,Y. and Li,H. (2010) A novel C53/LZAP-interacting protein regulates stability of C53/LZAP and DDRGK domain-containing Protein 1 (DDRGK1) and modulates NF-kappaB signaling. *J. Biol. Chem.*, **285**, 15126–15136.
48. Kang,S.H., Kim,G.R., Seong,M., Baek,S.H., Seol,J.H., Bang,O.S., Ovaa,H., Tatsumi,K., Komatsu,M., Tanaka,K. *et al.* (2007) Two novel ubiquitin-fold modifier 1 (Ufm1)-specific proteases, UfSP1 and UfSP2. *J. Biol. Chem.*, **282**, 5256–5262.