PrePhyloPro: phylogenetic profile-based prediction of whole proteome linkages

Yulong Niu[1,2,3], Chengcheng Liu[4], Shayan Moghimyfiroozabad[5], Yi Yang[2] and Kambiz N. Alavian[1,3,5]

[1] Department of Medicine, Division of Brain Sciences, Imperial College London, London, United Kingdom
[2] Key Lab of Bio-resources and Eco-environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, Sichuan, China
[3] School of Medicine, Department of Internal Medicine, Endocrinology, Yale University, New Haven, CT, United States of America
[4] Department of Periodontics, West China Hospital of Stomatology, Sichuan University, Chengdu, China
[5] Department of Biology, The Bahá'í Institute for Higher Education (BIHE), Tehran, Iran

ABSTRACT

Direct and indirect functional links between proteins as well as their interactions as part of larger protein complexes or common signaling pathways may be predicted by analyzing the correlation of their evolutionary patterns. Based on phylogenetic profiling, here we present a highly scalable and time-efficient computational framework for predicting linkages within the whole human proteome. We have validated this method through analysis of 3,697 human pathways and molecular complexes and a comparison of our results with the prediction outcomes of previously published co-occurrency model-based and normalization methods. Here we also introduce PrePhyloPro, a web-based software that uses our method for accurately predicting proteome-wide linkages. We present data on interactions of human mitochondrial proteins, verifying the performance of this software. PrePhyloPro is freely available at http://prephylopro.org/phyloprofile/.

Subjects Bioinformatics, Genomics
Keywords Linkage prediction, Whole proteome, Phylogenetic profile

Submitted 13 April 2017
Accepted 28 July 2017
Published 28 August 2017

Corresponding author
Kambiz N. Alavian,
k.alavian@imperial.ac.uk

Academic editor
Shawn Gomez

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.3712**

INTRODUCTION

The development of sequencing technologies has facilitated the access to whole genomic information from numerous organisms. Despite successful small-scale attempts to identify protein–protein interactions in a limited number of model organisms (*Ewing et al., 2007*; *Li et al., 2004*; *Tarassov et al., 2008*), determining genome-wide linkages remains a challenge. Phylogenetic profiling, by comparing genome sequences across different species, makes it possible to explore whole-proteome protein linkages (*Pellegrini et al., 1999*). This method is based on the assumption that functionally related proteins are likely to have evolved in a correlated manner. Several studies have successfully employed phylogenetic profiling to identify novel members of protein complexes (*Avidor-Reiss et al., 2004*; *Dey et al., 2015*; *Gabaldon, Rainey & Huynen, 2005*), expand known pathways (*Li et al., 2014*), and analyze non-coding elements (*Tabach et al., 2013a*).

Based on occurrence information across different species, two main groups of prediction algorithms have been proposed (*Kensche et al., 2008*). In the first group, the phylogenetic

profiles of paired proteins are directly compared by a "co-occurrency" method such as Hamming distance (*Cheng & Perocchi, 2015*; *Pellegrini et al., 1999*), Pearson correlation coefficient (*Glazko & Mushegian, 2004*), Jaccard similarity (*Brilli et al., 2008*; *Jaccard, 1912*; *Yamada, Kanehisa & Goto, 2006*), Fisher's exact test (*Barker & Pagel, 2005*), and mutual information (*Huynen et al., 2000*; *Wu, Kasif & DeLisi, 2003*). Other normalization methods, including singular value decomposition (SVD) (*Franceschini et al., 2015*; *Psomopoulos, Mitkas & Ouzounis, 2013*) and normalized phylogenetic profile (NPP) (*Sadreyev et al., 2015*; *Tabach et al., 2013a*; *Tabach et al., 2013b*) of phylogenetic profiles before calculating the co-occurrence, have been proposed to reduce the rate of false positive predictions. Although the co-occurrence-based methods do not correct for the effect of phylogenetic bias or the non-independence of the profile values, they are widely used for predicting functional linkages mainly due to being very time-efficient. The second group is comprised of "model-based" approaches such as collapsing of subtree (*Von Mering et al., 2005*), tree-kernel (*Vert, 2002*), maximum likelihood (*Barker & Pagel, 2005*), and parsimony methods (*Barker, Meade & Pagel, 2007*). To account for the statistical non-independence of the profile values, the model-based methods use the phylogenetic tree to correlate the evolutionary processes (*Kensche et al., 2008*). Recently modifications of these methods have used sophisticated statistical models to infer gene gain and loss across a wide range of eukaryotic organisms (*Dey et al., 2015*; *Li et al., 2014*). These methods are dependent on the reliability of phylogeny and require lengthy computational times. Many of these phylogenetic profiling algorithms are not user-friendly and have low computational efficiency resulting in high false positive rates.

As increasing numbers of sequenced genomes have become available, a number of phylogenetic profile databases and tools for visualization of *Eukarya* phylogenetic profiles have been developed (*Cheng & Perocchi, 2015*; *Cromar et al., 2016*; *Ott et al., 2012*; *Sadreyev et al., 2015*; *Szklarczyk et al., 2015*). After considering the computational efficiency and prediction power of the current methods and tools, here we propose a method and online tool, called PrePhyloPro (PPP), which combines multiple co-occurrency measures and utilizes top rank thresholds to determine potential linkages. To identify human whole-proteome functional linkages, we constructed a comprehensive phylogenetic profile using 972 different species. We evaluated PPP with positive and negative reference datasets based on known human pathways and protein complexes. In comparison to conventional phylogenetic profiling methods, this method presented overall improvement, i.e., higher sensitivity and enhanced specificity in the receiver operating characteristic (ROC) curves. Moreover, an analysis of biological features of the predicted protein links from 3,697 human pathways and complexes, resulted in 21.7% overall true positive rate when the top rank was set as 400. We also developed a web-based server based on PPP to acquire and visualize human whole proteome predicted linkages.

RESULTS

Prediction of whole proteome functional linkages

To construct comprehensive phylogenetic profiles, we included a wide range of eukaryotic and prokaryotic organisms with at least one organism in every known *Class* or *Phylum*,

resulting in 972 different species (Table S1). We then implemented a new phylogenetic profiling method, named PPP, by combining multiple co-occurrency measures. Apart from physically interacting protein pairs, the predicted linkages potentially represent related components (sensors, regulators, and regulons) of signalling pathways and subunits of protein complexes.

To assess the performance of PPP, we compared this technique with eight conventional phylogenetic profiling methods, which were divided into three categories. The first category, comprised of co-occurrency methods, including Jaccard similarity (''Jaccard''), Pearson correlation coefficient (''Cor''), mutual information (''MI'') and Hamming distance (''Hamming''), relied on the evolutionary similarity or distance (*Glazko & Mushegian, 2004*; *Kensche et al., 2008*). The second category represented the gain and loss relationships of two proteins with additional phylogeny; we used maximum likelihood (''Tree'') (*Barker & Pagel, 2005*), Dollo parsimony distance (''Dollo'') (*Kensche et al., 2008*) as representatives of this category. The third category combined co-occurrency methods with normalized phylogenetic profiles such as NPP (*Sadreyev et al., 2015*; *Tabach et al., 2013a*; *Tabach et al., 2013b*) and SVD (*Franceschini et al., 2015*; *Psomopoulos, Mitkas & Ouzounis, 2013*). Because of the evolutionary conservation of protein complexes, the correlations of subunits in the same complex have been widely used as validation datasets (*Barker & Pagel, 2005*; *Kensche et al., 2008*; *Ta, Koskinen & Holm, 2011*; *Zhou et al., 2006*). We retrieved subunit composition information for 1,604 human protein complexes from the ''comprehensive resource of mammalian protein complexes (CORUM)'' database (*Ruepp et al., 2010*) and generated multiple control datasets (Tables S2–S4).

Performance of PPP in predicting known linkages

ROC curves were plotted for the analysis methods after applying a series of relaxed thresholds. False positive rate (FPR) and true positive rate (TPR, also known as sensitivity) were calculated and represented in the x- and y-axis, respectively. A larger area under the curve (AUC) of ROC would indicate better reliability of the method. We observed that the AUC of PPP was the largest (0.73) in comparison to other conventional approaches. ''Jaccard'' had the third largest AUC (0.71), as this coefficient was one of the important tools used in PPP. In comparison to ''Jaccard'', PPP had enhanced sensitivity upon increasing the FPR. For example, by changing the FPR to 0.20, the sensitivity of ''Jaccard'' was 0.46, whereas the sensitivity of PPP increased to 0.53, showing a noticeable improvement (Fig. 1A). ''Cor'' and ''MI'', two similar correlation methods, had close AUCs (0.66 and 0.68, respectively). Interestingly, in comparison to PPP, ''MI'' displayed slightly higher sensitivity for the FPR values between 0.53 and 0.81. ''Hamming'' achieved a relatively low AUC of 0.62 (Fig. 1A). To determine the accuracy of positive predictions, i.e., potential functional linkages, we calculated the precision and recall (PR) for each method. In agreement with ROC curves, PPP showed a lower rate of decrease in the precision as the recall increased, indicating higher prediction of true positives comparing to the conventional methods. Similarly, more true positive predictions were detected by ''MI'' for the recall between 0.20 and 0.49 (Fig. 1B). Our results showed PPP identified overall more true linkages than each individual measure.

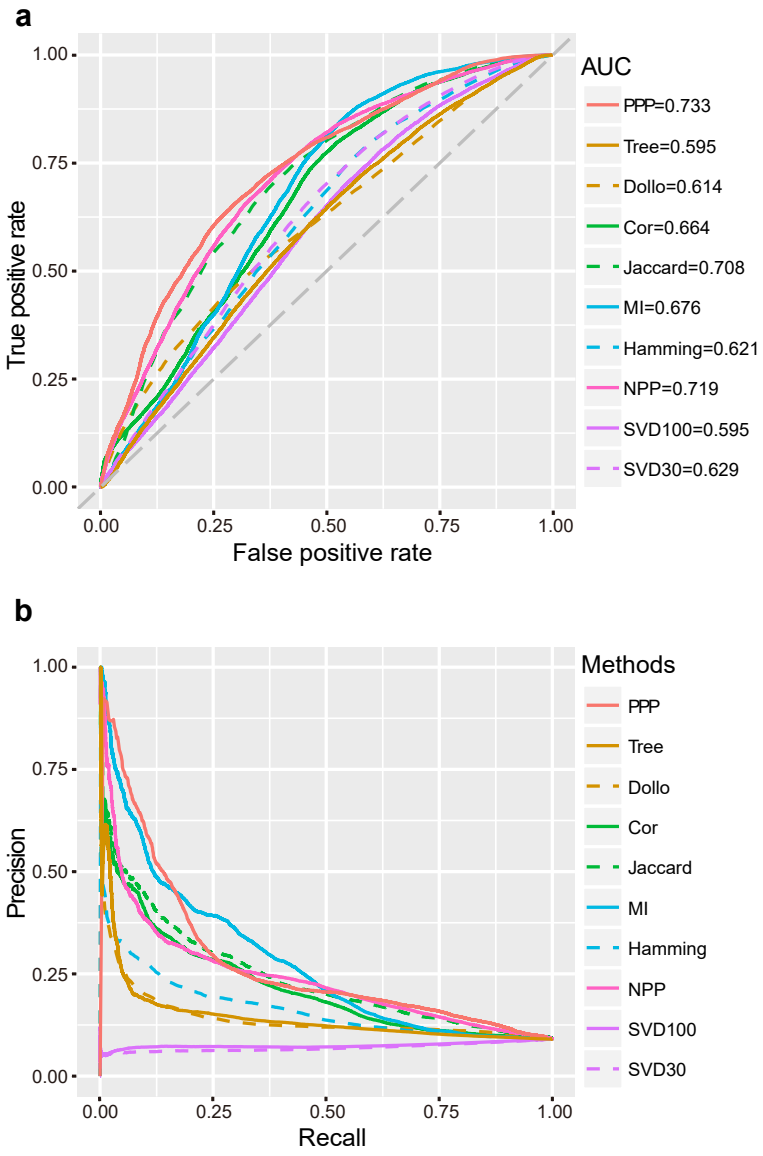

Figure 1 **Performance of PPP.** ROC curves (A) and PR curves (B) of PPP compared with Jaccard similarity ("Jaccard"), Pearson correlation coefficient ("Cor"), mutual information ("MI"), Hamming distance ("Hamming"), maximum likelihood ("Tree"), Dollo parsimony distance ("Dollo"), NPP normalization ("NPP"), and SVD normalization using all ("SVD100") or top 30% ("SVD30") of the unitary matrix, on a dataset comprising 57,114 positive linkages and 571,140 random protein pairs. The gray diagonal dash line is the random guess line.

Recent studies have suggested that the model-based methods have higher discriminative power and better performance (*Barker, Meade & Pagel, 2007*; *Barker & Pagel, 2005*; *Dey et al., 2015*; *Zhou et al., 2006*). Other studies have questioned the superior performance of these methods mainly due to their reliance on the correctness of the annotation of genomes, which may not always be the case (*Kensche et al., 2008*). We, nevertheless, included the "Tree" (*Barker & Pagel, 2005*) and "Dollo" (*Kensche et al., 2008*) as representatives by

using the likelihood ratio (LR) and the parsimony distance as measures, respectively. Our study found that the AUC and the decrease rate of precision were both lower in "Tree" and "Dollo" compared to PPP (Fig. 1). Our phylogenetic profiles included a wide range of both eukaryotic and prokaryotic species. The decreased precision rate of the model-based methods may be due to being arbitrarily applied to the phylogenetic profiles across a broad evolutionary scenario.

Sophisticated pre-processing and normalization methods are recently proposed to use sequence alignment bit scores instead of binary phylogenetic profiles, to accurately reflect the evolutionary relationships. For example, NPP (*Sadreyev et al., 2015*; *Tabach et al., 2013a*; *Tabach et al., 2013b*) and SVD (*Franceschini et al., 2015*; *Psomopoulos, Mitkas & Ouzounis, 2013*) combined the z-score and truncated unitary matrix with "Cor" and Euclidean distance, respectively. Our results showed that NPP achieved a significantly better performance than individual "Cor" correlation measures that had the second largest AUC (0.72). Similarly, setting the top percentage of unitary matrix as 30% ("SVD30"), would result in a higher AUC than that of the L_p-norm based methods, including "Hamming" (Fig. 1).

To further examine the performance of PPP, we constructed another negative reference dataset with a different random seed in our program, confirming the reliability of the predicted linkages (Figs. S1A, S1B). To further validate the predicted linkages by PPP, we applied rebuilt positive linkages excluding large complexes (Figs. S1C, S1D) and an independent validation dataset described by *Ta, Koskinen & Holm (2011)* (Figs. S1E, S1F). The AUC of PPP was the largest in both cases, suggesting the robustness of this method. Furthermore, we validated predicted protein pairs that were present in a wide range of species by using the results from the MatrixMatchMaker (MMM) method (*Bezginov et al., 2013*; *De Juan, Pazos & Valencia, 2013*; *Rodionov et al., 2011*; *Tillier & Charlebois, 2009*) (Table S6). Among the MMM protein pairs, the ones with more homology showed a higher hit rate at stringent top ranks, indicating increased true linkage detection by PPP when both proteins of each pair are present in a wide range of species (Fig. S2).

Evaluating predicted linkages in human pathways and complexes

To evaluate the efficiency of PPP in predicting known linkages, we first generated a list of predicated linkages involved in human pathways or complexes based on five databases. We used Kyoto Encyclopedia of Genes and Genomes (KEGG), BioCarta, Reactome and NCI/Nature Pathway Interaction Database (NCI), for pathways analysis and CORUM to establish connections of proteins within complexes. The list included known linkages in 241, 247, 1,393 and 212 pathways in the KEGG, BioCarta, Reactome and NCI databases, respectively. We also included 1,604 different complexes through CORUM (Table S5). With the threshold of top interactions set as 400 (i.e., the top 400 protein pair phylogenetic profile correlations/similarities), PPP achieved an overall prediction rate of 21.7%. The method predicted more than 50% of the interactions in several pathways with at least 30 known linkages (Fig. 2A). The high ratios of predicted (PPP) and known/original (databases) links indicated the reliability of our approach. For example, the mitogen-activated protein kinases (MAPK) signalling pathway is comprised of a total of 17 known links, 15 of which

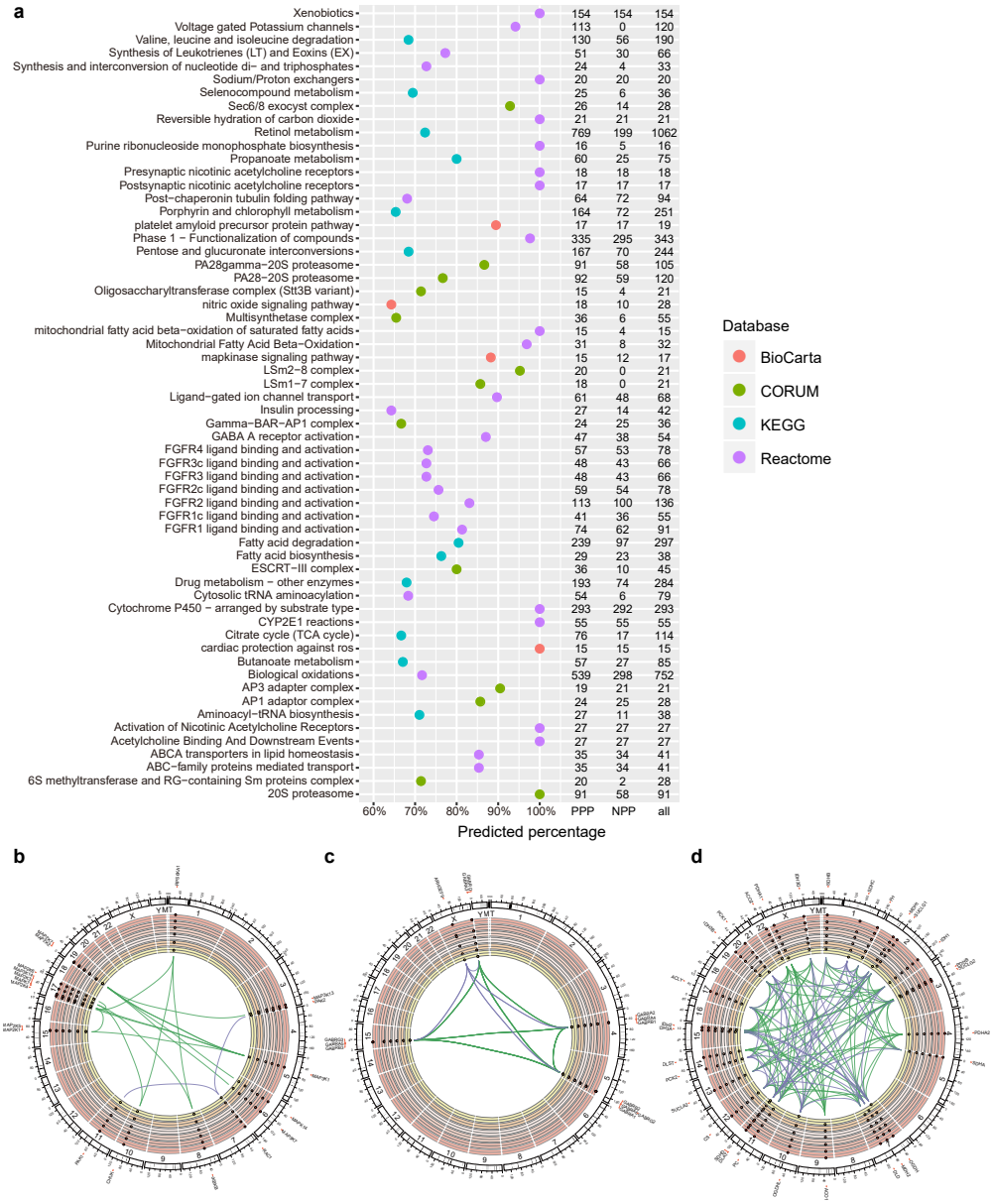

Figure 2 **PPP predicted linkages in human pathways and complexes.** (A) Top pathways and complexes with high predicted percentage (>50% and the number of predicted links is at least 15) in BioCarta, KEGG, Reactome, and CORUM databases. The number of PPP predicted (threshold 400, sensitivity 0.97), NPP predicted (threshold 0.73, sensitivity 0.97), and original linkages is presented on the right. (B–D) Selected Circos visualization of predicted linkages in the MAPK signaling pathway (B), GABA A receptor activation (C), and TCA citrate cycle (D). The outer ring shows the ideogram of human karyotype plus the mitochondria genome. The next six rings, coloured with yellow to dark red, show the present percentage for each protein. From outer to inner rings, points in each ring indicate the percentage in three Kingdom-size groups in *Eukaryota* (*Animals*, *Plants*, *Fungi*, and *Protists*), *Bacteria*, and *Archaea*. In percentage rings, the vertical axis ranges from 0 to 1, and the axis direction is from inner to outer space. The dash central line indicates the 50th percentile. In the centre, arcs indicate correlated relationships between paired proteins. The green arcs represent successfully predicted links according to the corresponding database. The purple arcs indicate false negative links. Outside the ideogram, proteins are annotated with symbols, and marked at corresponding genomic positions.

were identified by PPP (Figs. 2A, 2B). We performed Circos visualization (*Krzywinski et al., 2009*), representing predicted links in three different pathways or complexes: MAPK signalling pathway (Fig. 2B), γ-aminobutyric acid (GABA) A receptor activation (Fig. 2C) and tricarboxylic acid (TCA) cycle (Fig. 2D).

PPP successfully identified the linkages within protein families, especially when members of the family share a common evolutionary pattern. It has been shown that the members of the MAPK signaling pathway arose at the dawn of eukaryotic evolution (*Glatz et al., 2013*) and may have orthologies in some bacterial species (*Miller et al., 2010*; *Pereira, Goss & Dworkin, 2011*). Our phylogenetic profiling confirmed the distribution of two members of this family, MAP2K and MAP3K, in almost all eukaryotes and some prokaryotes. This common evolutionary pattern was the basis for the detected linkages within the MAPK by PPP (Fig. 2B). Similarly, linkages between α subunits (GABRA1 to GABRA6), β subunits (GABRB1 to GABRB3), and γ subunits (GABRG2 and GABRG3) of the GABA A receptor were detected because of the similarity in their phylogenetic profile and their presence only in animals (Fig. 2C). By design, PPP uses co-occurrence as the main criterion for detecting linkages. The dissimilar phylogenetic distribution of two proteins, therefore, would translated into absence of interaction. This might result in the presence of false negatives, due to the involvement of evolutionary modules in pathways or complexes (*Li et al., 2014*; *Pellegrini et al., 1999*). For example, we failed to observe the linkages between ARHGEF9 and the other GABA A subunits, because homologs of α/β/γ subunits were exclusively present in *Metazoa* while ARHGEF9 homologs were also detected in *Fungi* (Fig. 2C). Modularity could be explained based on the selection for adaptation rate, where common evolutionary rates could force certain genes to evolve together and to maintain an interaction, preventing other genotypes (with similar phylogenetic profiles) from being included based on the difference in their rates of adaptation (*Wagner, 1996*).

Similarly, unlike the rest of the MAPK signalling pathway members, homologs of RAC1 were present mainly in eukaryotes, resulting in false negatives with respect to interaction with the other members of the MAPK pathway (Fig. 2B). Likewise, since PPP was based on calculating co-occurrences, it limited the correctly predicted linkages within the TCA cycle to the known evolutionary modules (*Li et al., 2014*) (for examples in "ACO1/ACO2/CS/DLST/SUCLA1/SUCLA2" and in "IDH3B/IDH3G/IDH3A"). Most of the true linkages to PCK1 and PCK2 were missed due to the same reason (Fig. 2D). Overall, our data suggests that PPP is suitable for predicting interactions between proteins that share common homologous distributions, but might be limited in detecting linkages between proteins that belong to different evolutionary modules in human pathways or complexes.

Input of PrePhyloPro

We implemented PPP and comprehensive phylogenetic profiles into an intuitive and easy to use web-based software "PrePhyloPro" for whole proteome linkages prediction. PrePhyloPro could be used for detecting novel (physical) protein-protein interactions, for predicting new components of biological complexes, and suggesting potential new linkages in signaling pathways or metabolic processes.

PrePhyloPro is designed as a user-friendly tool that requires three steps. The first step is choosing algorithm parameters including the top rank threshold, BLAST *E*-value threshold, and the reference organisms. The top rank is the number of linkages (with the highest correlation coefficient values) corresponding to each queried protein. In the front page three options (0.001, 0.0005, and 0.0001) are provided for the BLAST *E*-value threshold. This threshold is applied to choose homologies among the 972 species in constructing phylogenetic profiles. Using a smaller top rank and BLAST *E*-value would result in less but more reliable linkages. Currently the PrePhyloPro software provides linkages prediction in two model organisms, *Homo sapiens* and *Arabidopsis thaliana*. Prediction in more species will be available, as future updates will be applied to the software. The second step is to set the size of protein names for the output plots as the default setting may not be suitable for studies with high or low number of queried proteins. In the last step, a table of query proteins in the "txt" or "csv" format will be uploaded onto the website for linkage prediction. The markup colours and aliases for query proteins could be added to this table to optimize visualization of output figures. Examples of input files are provided online.

Output of PrePhyloPro

To minimize the processing time for each query, we have already calculated the co-occurrency for all protein pairs under the three BLAST *E*-value thresholds, and have saved them in backend databases. In total, PrePhyloPro takes less than 1 min to determine whole proteome search for linkages of 20 candidate (query) proteins. PrePhyloPro returns an integrated webpage including output figures and tables. As an example of an input protein set, we have used subunits of the human F1Fo ATP synthase (Fig. 3). The outputs for this set include the phylogenetic profile plot and correlation matrix of input proteins. In the phylogenetic profile plot, the top panel represents the 972 surveyed species, which are divided into six major taxa (*Animals*, *Plants*, *Protists*, *Fungi*, *Archaea*, and *Bacteria*). The left panel represents the input proteins (marked with user-defined colours) combined with a cluster dendrogram measured by Euclidean distances. The blue and gray bars in the phylogenetic profile plot correspond to presence or absence of homologies, respectively (Fig. 3A). The correlation matrix, as a complement to the phylogenetic profile plot, shows the Pearson correlation coefficient of paired input proteins that are colour-coded from blue (no correlation between any given profile pairs) to red (highly correlated profiles) (Fig. 3B). These two figures not only demonstrate homologous distributions of input proteins among an array of eukaryotic and prokaryotic organisms, but also indicate the evolutionary relationships within a query set. For example, the F1Fo ATP synthase is proposed to have evolved from at least two major parts, i.e., the catalytic core (F1) and the membrane-bound subunits (Fo) (*Falk & Walker, 1988*; *Mulkidjanian et al., 2007*; *Rak, Gokova & Tzagoloff, 2011*). As anticipated, PrePhyloPro clustered a majority of the subunits into two groups, corresponding to the F1 and Fo components (Figs. 3A, 3B).

The output of PrePhyloPro also includes the visualization of predicted linkages. We used the circosJS package (*Girault, 2017*) to create an interactive Circos plot, which integrates the chromosome location, homologous distribution, values of co-occurrence, and linkages. With the threshold of top rank and the BLASTP *E*-value set to 20 and 0.001, respectively, the

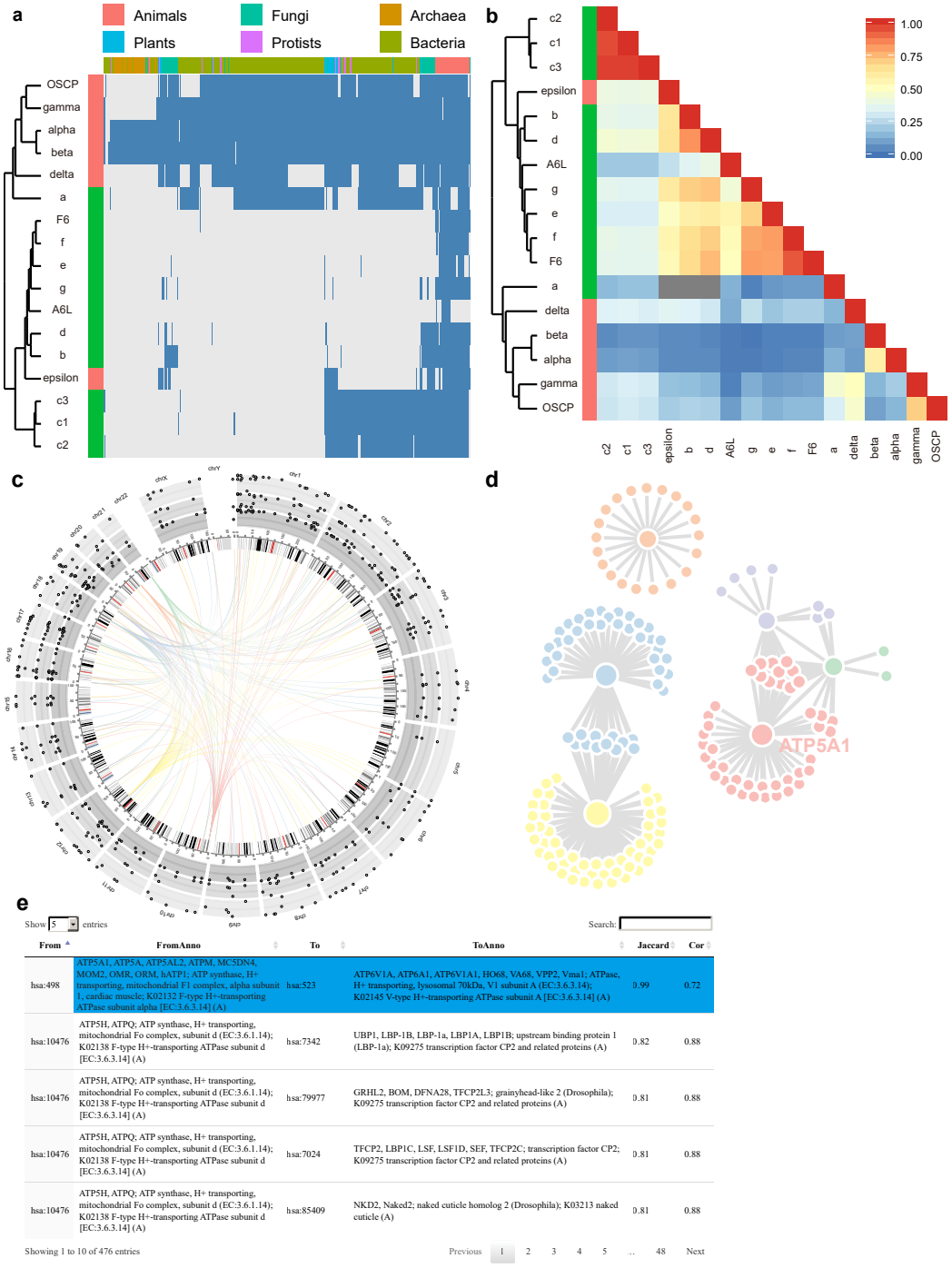

Figure 3 Outputs of PrePhyloPro using human F1Fo ATP synthase subunits as input proteins. (A–B) The phylogenetic profile plot (A) and the correlation matrix (B) of the F1Fo ATP synthase. The left colour bar indicates subunits of F1 (red) and Fo (green) regions. (C–D) The D3 interactive Circos plot (C) and the network (D) of predicted linkages of subunits in the F1 region. (E) The numeric table of predicted linages. The linkage between the α subunit and ATP6V1A is highlighted.

predicted linkages of the F1 subunits of the ATP synthase (α, β, γ, δ, ε, and OSCP subunit) were displayed in Fig. 3C. An ideogram of normal karyotype plus the mitochondrial genome is plotted in the central ring. The outer three rings with grey background show the percentage of present homologies for each protein in *Bacteria*, *Archaea*, and *Eukaryota* (from the outer to the inner ring). Hovering over the points will show the corresponding present percentages. In the centre, connecting arcs represent the predicted links for the 6 subunits in F1 ATP synthase with user-defined colours (In Fig. 3C: α (blue), β (yellow), γ (red), δ (orange), ε (purple), and OSCP (green)). Hovering over the arcs will indicate the linkage partners, their chromosomal locations and Jaccard/Cor values.

To directly visualize network topology, PrePhyloPro generates an interactive linkage network using JavaScript D3 (*Gandrud, Allaire & Kent, 2015*). Nodes with more linked partners have a bigger size. Hovering over one node will enlarge its size and brings up its corresponding gene symbol. As an example, the α subunit (ATP5A1) of ATP synthase (Fig. 3D). In contrast to static visualization, the D3 network provides more interactive features, like dragging or pulling one node from a crowded group, which is particularly helpful for a large network with overlapping nodes.

The outputs of PrePhyloPro contain a table summarizing the predicted linkages. The "from" and "to" columns are composed of standard protein IDs of input proteins and the predicted interacting proteins, respectively. The proteins symbols and descriptions are also included in this table. The last two columns display the Jaccard similarity and Pearson correlation coefficient values sorted in a decreasing order. Hovering over one cell of the table highlights its row in a blue background. Other features include a search box at the right corner, the option for adjusting the number of entries (at the left corner), and the sort option (at the top of each column) (Fig. 3E). Output results and figures can be downloaded to local devices. The downloaded folder includes high quality figures. The correlation matrix and prediction linkages are stored in numeric tables that can be used for further analysis and validation.

PrePhyloPro identified the known linkages between the α/β subunits and between the γ/δ subunits of the F1 component of the synthase (Figs. 3C, 3E). The linkage between the α subunit and ATP6V1A (highlighted in Fig. 3E), a subunit of human *V*-type ATP synthase, was also identified by PrePhyloPro. These two proteins are believed to have evolved from the same ancestor by gene duplication (*Iwabe et al., 1989*; *Shih & Matzke, 2013*). Moreover, PrePhyloPro showed strong linkages (Jaccard similarity > 0.99) between the α/β subunits and ATP-binding cassette transporters (ABC) family members (Table S7). Recent studies have shown functional linkages between the two sets of proteins. ABCB7 and nuclear genes of ATP synthase are both significantly down-regulated in SOD2 deficient erythroblasts under oxidative stress (*Martin et al., 2011*), while mutations of ABCD1 lead to the oxidation of α/β subunits and defects in oxidative phosphorylation (*Lopez-Erauskin et al., 2013*). These studies suggest a possible regulatory relationship within ATP synthase and ABC family. Another interesting predicted partner of α/β subunits was AFG3L2 (Jaccard similarity > 0.99). Mutations in the *Saccharomyces cerevisiae* homolog of AFG3L2, AFG3, inhibit the assembly of ATP synthase, suggesting a similar role of AFG3L2 in human (*Guzelin, Rep & Grivell, 1996*; *Paul & Tzagoloff, 1995*). Moreover, PrePhyloPro predicted

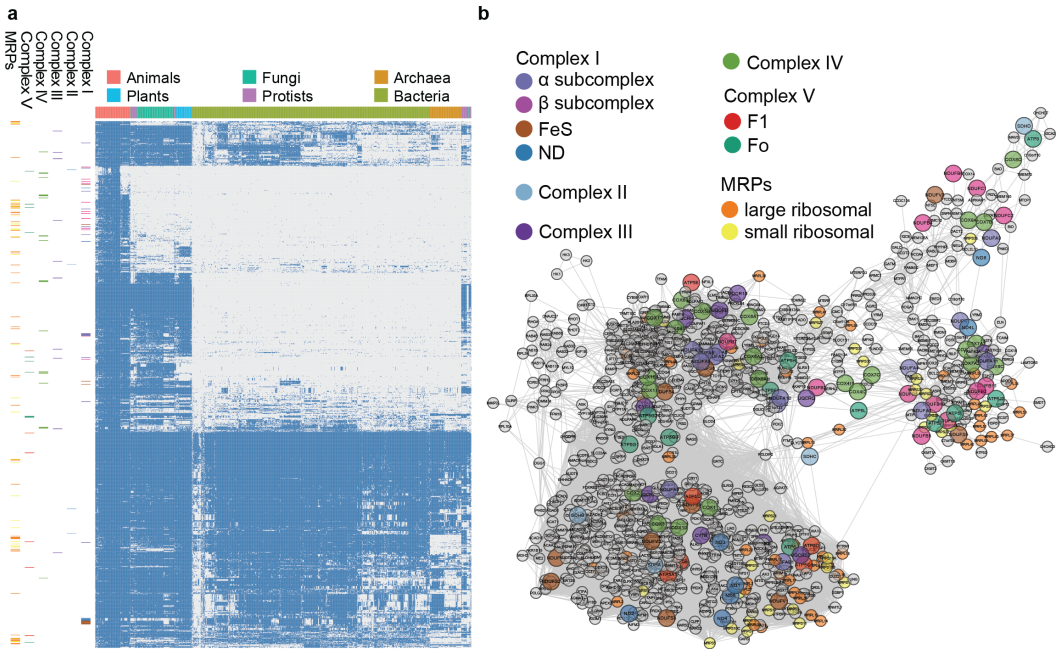

Figure 4 Phylogenetic profile and network visualization of human mitochondrial proteome. (A) Phylogenetic profile plot for 1,006 human mitochondrial proteins across 972 fully sequenced organisms. Blue and grey squares indicate the gene gain and loss, respectively. Hierarchical cluster is applied to both the organisms (columns) and proteins (rows). The organisms are organised into 6 taxa: *Animals*, *Plants*, *Fungi*, *Protista*, *Bacteria*, and *Archaea*. On the left, each small band indicates the corresponding subunits in complex I to V, and MRPs. (B) Predicted linkage network for human mitochondria proteins.

the high correlations (Jaccard similarity >0.95) between α/β subunits and adenylate kinase isoforms (AKs) (Table S7). Consistently, AKs maintain the cellular energy balance and collaborate in ATP synthesis, especially through coupling of the mitochondrial resident AK2 with the OXPHOS activity (*Klepinin et al., 2016*). Additionally, the transcription of AK2 and α/β subunit are both enhanced by triiodothyronine (*Severino et al., 2011*). It has also been suggested that the transcription of α subunit and AK1 is regulated by PGC-1α, a master regulator of metabolism (*Lucas et al., 2014*). These studies confirm the potential of PrePhyloPro in predicting linkages based on co-evolution of proteins.

Inferring evolutionary relationships from phylogenetic profiles of mitochondrial proteins

To evaluate the large-scale prediction power of PrePhyloPro, we used the entire human mitochondrial proteome containing 1,006 mitochondrial proteins (*Pagliarini et al., 2008*) as the input set. PrePhyloPro returned the list of predicted linkages and the phylogenetic profile plot of mitochondrial proteins (Fig. 4A). Only predicted linkages between a pair of mitochondrial proteins were selected for the purposes of visualization (Fig. 4B). Interestingly, after mapping the oxidative phosphorylation complexes (complex I to complex V) and mitochondrial ribosomal proteins (MRPs), we noticed that the subunits within a complex are dispersedly distributed in the phylogenetic profile figure (Figs. 4A, 4B). In agreement with previous studies (*Li et al., 2014*; *Pagliarini et al., 2008*),

this indicated that mitochondrial proteins have originated from multiple modules during evolution. PrePhyoPro detected closer linkages among members of known evolutionary modules (*Li et al., 2014*) of mitochondrial complexes. For example, COX1, COX2, ATP6, ND1, ND3, and SDHA that are widely present among eukaryotic and prokaryotic species are gathered tightly in a subnetwork (Fig. S3). Confirming this result, a recent study showed that dietary lipid affects the expression of COX1, COX2, ATP6, and ND1 by common transcription factors such as the peroxisome proliferator-activated receptor (*Eya et al., 2015*). On the other hand, the group consisting of COX7A1, NDUFB1, ND6, and NDUFA1 was exclusively present in the *Metazoa* and was concentrated in a different subnetwork (Fig. 4B and Fig. S3). These results suggest that, in addition to detecting protein linkages, PrePhyloPro provides insight into the evolutionary relationships between paired proteins. Deciphering the evolutionary relationships within a query set would be useful for further exploring biological functions in pathways and complexes.

DISCUSSION

In this study, we implemented a phylogenetic profiling method named PPP to predict whole proteome linkages. PPP combined multiple co-occurrences and used top ranks to select most likely linkages. This method excluded solo proteins that have no connections with other proteins in prediction results, even when a stringent threshold was set to achieve more reliable linkages. Moreover, PPP displayed robustness in comparison to other conventional approaches. We are aware that factors, such as control datasets or the species chosen to construct phylogenetic profiles and trees, may contribute to the poor performance of some methods. For example, in comparison to PPP, "MI" displayed higher sensitivity in the ranges corresponding to large FPR, which may be due to the discrimination power of mutual information. But because "MI" is limited in making a distinction between the anti-correlating and correlating protein pairs (*Steuer et al., 2002*), positive predictions established with a high "MI" threshold might include negatively correlated pairs.

In our test datasets, model-based methods exhibited lower predictive power than co-occurrence methods. A major limitation of model-based methods is the underestimation of paired proteins that are both present in a wide range of species, resulting in lack of true positive predictions. For example, the α and β subunits of the F1Fo ATP synthase are two co-occurring proteins, as they are present in almost all living species (*Gogarten et al., 1989*). They physically interact with each other to form an $\alpha_3\beta_3$ hexamer (*Rubinstein, Walker & Henderson, 2003*). "Tree" and "Dollo", however, yielded an extremely low value indicating no linkages. The "Tree" method showed a rapid increase in precision to the maximum and remained constant at more stringent LR thresholds, suggesting higher prediction power of "Tree" with carefully chosen LRs (*Barker & Pagel, 2005*). Dollo parsimony distance is limited in eukaryotes, where horizontal gene transfers are rare events (*Barker, Meade & Pagel, 2007*; *Kensche et al., 2008*). Thus, model-based methods would have a better performance in predicting functional linkages specific to a *Class* or a *Phylum* rather than linkages that are conserved across a wide range of species.

Although PPP showed improvements in ROC curves, two kinds of inaccuracies still existed, which could be corrected by taking additional steps. The false negatives occurred

when actually linked partners had different evolutionary rates. For example evolutionary constrains are not common in signalling and transcriptional pathways (*Dey et al., 2015*). Other than protein co-evolution based approaches, methods like weighted gene co-expression network analysis (WGCNA) is appropriate to detect proteins that share similar co-expression patterns (*Langfelder & Horvath, 2008*; *Liu et al., 2015*). On the other hand, in addition to setting stringent thresholds for the top rank, using network algorithms to filter PPP outputs might help in reducing the rate of false positives. This approach could include selecting hubs connecting to more input proteins by using centrality measurements (degree and betweenness).

Several online tools exist for phylogenetic profiling, for example STRING (version 10) using SVD (*Szklarczyk et al., 2015*), PhyloGene using NPP (*Sadreyev et al., 2015*), and ProtPhylo using "Hamming" (*Cheng & Perocchi, 2015*). PrePhyloPro based on PPP is a complementary online tool for whole proteome linkage predictions. PrePhyloPro includes several visualization methods, including the interactive Circos plot integrated metadata of the homology distribution, genome locations, occurrency values and prediction linkages.

METHODS

Phylogenetic profiling

Protein sequences and annotation information from 972 different species were retrieved from the KEGG database (*Kanehisa et al., 2006*), including 276 eukaryotic, 614 bacterial, and 82 archaea organisms, as well as mitochondrial and chloroplast proteins. BLASTP (*Camacho et al., 2009*) was used to comparing 20,127 human proteins sequences with selected species. To construct the homology matrix, we used BLASTP *E*-value as the criteria, in which 1 denoted that homologies of human proteins found in the corresponding species, otherwise 0.

Four independent co-occurrence methods were used to evaluate the correlated relationship between a pair of proteins (*Glazko & Mushegian, 2004*; *Kensche et al., 2008*). For each pair of proteins across n species, for example $X, Y \in \{0,1\}^n$, the Jaccard similarity is defined from co-occurrence of presences:

$$J(X,Y) = \frac{\left|\left\{i | x_i = 1 \cap y_i = 1\right\}\right|}{\left|\left\{i | x_i = 1 \cup y_i = 1\right\}\right|}. \tag{1}$$

The Pearson correlation coefficient is:

$$cor(X,Y) = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{(n-1)S_X S_Y} \tag{2}$$

where $\overline{X}$ is the sample mean of X, and S_x is sample standard deviations of X.

The mutual information is:

$$I(X,Y) = \sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{3}$$

where $p(x)$ is the probability of a symbol (0 or 1) appears in X.

The L_p-norm is defined as:

$$d_L = \left[\sum_{i=1}^{N} |x_i - y_i|^p\right]^{1/p} \quad (4)$$

where $p = 1$ is the Hamming distance.

We applied two phylogenetic tree based method, maximum likelihood (*Barker & Pagel, 2005*) and Dollo parsimony distance (*Kensche et al., 2008*), as representations of model-based methods. In order to reduce the computational time, as well as achieving a balanced phylogenetic tree, we decreased two third of the bacterial species. The phylogenetic tree was constructed based on the small ribosomal RNA (16S/18S) sequences downloaded from the SILVA database (release 119) (*Quast et al., 2013*). After reducing the redundancy of ribosomal RNA sequences, a total of 522 species were selected including 243 in *Eukaryota*, 201 in *Bacteria*, and 78 in *Archaea*. The truncated ribosomal RNA sequences were aligned using the MAFFT program (*Katoh & Standley, 2013*), and the phylogenetic tree was generated by the RAxML program with default parameters (*Stamatakis, 2014*). LRs of the maximum likelihood method were calculated by the BayesTrait (*Barker & Pagel, 2005*). In the Dollo parsimony method, the gain/loss state is firstly reconstructed for each node of the phylogenetic tree, and then the Dollo parsimony distance is calculated as:

$$d_{\text{Dollo}}(X, Y) = \sum_{i \in \text{branches}} \left|(anc(x_i) - desc(x_i)) - \left(anc\left(y_i\right) - desc\left(y_i\right)\right)\right| \quad (5)$$

where $anc(x_i)$ and $desc(x_i)$ are the ancestral and descendant's state of a branch (*Kensche et al., 2008*).

The NPP was used to normalize phylogenetic profiles, in which processed BLASTP bit scores were included (*Sadreyev et al., 2015*; *Tabach et al., 2013a*; *Tabach et al., 2013b*). In a bit score profile P with n proteins across m species, for each bit score, it was set as 1 if lower than the 70. Then for each protein n_i, if its number of homologous across m organisms was lower than a threshold, e.g., 12, the protein is removed because of its poor conservation. Next the bit score p_{i_j} was normalized as $\log 2\left(p_{ij}/p_{\max i}\right)$, where $p_{\max i}$ was the maximum bit score in the i-th row. The last step was to normalize bit scores across species. Specifically, the bit score p_{i_j} was normalized as $\left(p_{ij} - \mu_j\right)/\sigma_j$, which was also known as the z-score, where μ_j and σ_j were the mean and the standard deviation of the j-th column, respectively. Compared to the original profile P, the NPP normalized profile $P'\left(n' \times m\right)$ had the same organism number m, but may contain less proteins.

Another normalization method was called SVD (*Franceschini et al., 2015*; *Psomopoulos, Mitkas & Ouzounis, 2013*). In a bit score profile P with n proteins across m species, for each bit score, it was set as 0 if lower than the 60. Then the bit score p_{i_j} was normalized as $p_{ij}/p_{\max i}$, where $p_{\max i}$ was the maximum bit score in the i-th row. The next step was to SVD of the profile following $P = U\sum V'$, where U was the unitary matrix. The profile P' was defined as the top trimming columns of U. Because the SVD predictions are sensitive to the "trimming" parameter (top percentages of the unitary matrix), we set this parameter as 100% and 30%. Similar to the second step of NPP, poor conserved proteins were marked in the original profile and removed in $P'\left(n' \times m'\right)$. The last step was Euclidean normalization

of species in P' as $p_{ij}/\sqrt{\sum_{i=1}^{m'} p_{ij}^2}$. The SVD normalized profile P' may have less organisms and proteins than those in the original profile P.

The Pearson correlation coefficient and Euclidean distance ($p = 2$ in Eq. 4) were applied to measure co-occurrence of paired proteins as described in NPP and SVD, respectively.

PPP method

We combined co-occurrences to implement a new method named PPP to improve the prediction efficiency. PPP was inspired by "solo proteins" that did not link to any other proteins, when we conducted whole proteome linkages prediction. However, solo proteins may be not in existence, considering the huge size of protein interactions representing complex biological activities (*Stumpf et al., 2008*). One solo protein occurred when correlated relationships of this protein with others were all lower than a pre-defined threshold in co-occurrence methods. Especially, a stringent threshold set to get higher reliable linkages always yielded more solo proteins. Thus, instead of setting an arbitrary threshold of co-occurrency, PPP chose top-ranking T linkages for each protein.

We illustrated PPP by predicting whole proteome linkages among n proteins from a phylogenetic profile across m species. The first step was to roughly exclude linkages with negative correlations. For a given protein n_i, the Pearson correlation with the other proteins was calculated and denoted as $cor\,(l) \in V_{cor}$. Similarly, the Jaccard similarity vector V_J was generated. We excluded proteins with negative Pearson correlations, because a negative number would imply the presence of one protein and the absence of another in a pair, which could not be considered as a functional link under an evolutionary scenario. In the next step, we recorded the rank in decreasing order of each element in the vector V_J and denoted as $rank\,(l) \in V_{J-\mathrm{rank}}$. Finally, top T proteins were considered to be functionally linked to the proteins

$$L_i = \{l | cor\,(l) > 0 \wedge rank\,(l) < T\} \tag{6}$$

where T ranges from 1 to $n-1$. Thus, for a given protein, it has at least one partner owning the largest co-occurrency. A smaller T yielded less but more reliable linkages.

The whole proteome functional linkages were calculated for proteins as the same procedural $L = \{L_1, \ldots, L_n\}$. The predicted linkages were considered as a symmetric relationship, which means that we neglected the linkage direction between paired proteins.

Control datasets

We retrieved 1,604 human complexes from the CORUM database (*Ruepp et al., 2010*). To generate the positive references, we chose paired proteins in same complexes. The control dataset consisted a total of 57,114 positive linkages (Table S2) and 571,140 negative linkages (Table S3), which were constructed by randomly selecting two proteins located in different complexes. Moreover, the negative linkages were not allowed from the complexes that resided in the same subcellular position. To avoid an arbitrary judgment, we could randomly choose multiple negative linkage lists (Table S4). Because protein pairs in large complexes contributed a large proportion in positive linkages and bias the results, we excluded complexes containing more than 40 subunits and generated another control

dataset with 29,189 positive linkages and 291,890 negative linkages. Moreover, we used the control datasets from *Ta, Koskinen & Holm (2011)*, which included 26,525 positive linkages and a same number of negative ones.

Similarities or distances in co-occurrence and normalization methods, LRs in the maximum likelihood method, distances in the Dollo parsimony method, and top rank in PPP were used to generate series of thresholds. Under each threshold, the number of TP true positives (TP) and true negative (TN) represented the positive and negative predicted linkages, respectively. In contrast, false positive (FP) and false negative (FN) were erroneously detected as positive and negative linkages under certain thresholds. In ROC curves, we represented the FPR and TPR in the x-axis and y-axis, respectively, as:

$$FPR = 1 - specificity = 1 - \frac{TN}{N} \tag{7}$$

$$TPR = sensitivity = \frac{TP}{P}. \tag{8}$$

P and N represent the total number of positive and negative reference links, respectively. In PR curves, we defined the precision and recall as:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN}. \tag{10}$$

To validate the linked proteins that were present across a wide range of species, we compared the PPP predicted linkages with the results generated by the MMM method, which evolves the largest common submatrix between paired proteins (*Bezginov et al., 2013*; *Rodionov et al., 2011*; *Tillier & Charlebois, 2009*). With the MMM threshold set as 12, a total of 6,422 co-evolved protein pairs were retrieved (Table S6). We chose proteins that had 95%, 85%, 75%, and 65% homologies present percentage across 972 species, respectively, and then we re-generated the MMM protein pairs accordingly. The hit rates were calculated as N_{top}/N_{MMM}, where N_{top} denoted the number of PPP predicted linkages under certain top rank threshold (ranging from 1 to 3,000) and N_{MMM} was the number of MMM re-generated protein pairs.

Selection and vitalization of functional gene-sets and mitochondria correlation network

To evaluate the biological features of our whole proteome predicted functional linkages, we chose four different biological gene-sets databases: KEGG (*Kanehisa et al., 2006*), Biocarta, NCI/Nature Pathway Interaction Database (NCI) (*Schaefer et al., 2009*), and Reactome (*Fabregat et al., 2016*). The R/Bioconductor package "graphite" (*Sales et al., 2012*) re-constructed the pathway topology into 2,093 different protein-protein interaction networks. We calculated the predicted percentages for each gene-sets using our whole proteome functional linkages list. Three representative gene-sets as MAPK signalling

pathway, GABA A receptor activation, and TCA citrate cycle, were visualized by an integrated Circos plot (*Krzywinski et al., 2009*).

To generate the human mitochondria correlation network with our prediction results, we retrieved 1,006 human mitochondria related proteins (*Pagliarini et al., 2008*), which was visualized by Cytoscape (*Shannon et al., 2003*). Five protein complexes in oxidative phosphorylation system (OXPHOS), including complex I (NADH-ubiquinone oxidoreductase), complex II (succinate-ubiquinone oxidoreductase), complex III (ubiquinol-cytochrome c reductase), complex IV (cytochrome c oxidase), and complex V (F1Fo ATP synthase) were highlighted, as well as MRPs.

Statistical analysis

The pROC package was used to perform the typical ROC analysis (*Robin et al., 2011*). The SVD normalization was carried out by the SVD-Phy package (*Franceschini et al., 2015*). The rest programming tasks were conducted using the open-source R Project (http://www.r-project.org/).

ACKNOWLEDGEMENTS

We are grateful to Nicolas Girault for creating the excellent circosJS package. We thank Philip Kensche for the valuable discussion of Dollo parsimony, Andrew Meade for providing source codes of the BayesTrait program, and Youquan Bu for the suggestions on narrowing down the predicted linkages. We thank Elisabeth Tillier for sending us the MMM predicted linkages. We also thank Kwangbom Choi for the constructive comments.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

YN was supported by the China Scholarship Council. This work was supported by the Imperial College London, Department of Medicine, Division of Brain Sciences funds to KNA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
China Scholarship Council.
Imperial College London, Department of Medicine, Division of Brain Sciences.
Division of Brain Sciences.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Yulong Niu conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

- Chengcheng Liu performed the experiments, reviewed drafts of the paper.
- Shayan Moghimyfiroozabad analyzed the data, reviewed drafts of the paper.
- Yi Yang analyzed the data, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Kambiz N. Alavian conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:
http://prephylopro.org/phyloprofile/.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3712#supplemental-information.

REFERENCES

Avidor-Reiss T, Maer AM, Koundakjian E, Polyanovsky A, Keil T, Subramaniam S, Zuker CS. 2004. Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell* **117**:527–539 DOI 10.1016/S0092-8674(04)00412-X.

Barker D, Meade A, Pagel M. 2007. Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics* **23**:14–20 DOI 10.1093/bioinformatics/btl558.

Barker D, Pagel M. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLOS Computational Biology* **1**:e3 DOI 10.1371/journal.pcbi.0010003.

Bezginov A, Clark GW, Charlebois RL, Dar VU, Tillier ER. 2013. Coevolution reveals a network of human proteins originating with multicellularity. *Molecular Biology and Evolution* **30**:332–346 DOI 10.1093/molbev/mss218.

Brilli M, Mengoni A, Fondi M, Bazzicalupo M, Lio P, Fani R. 2008. Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. *BMC Bioinformatics* **9**:551 DOI 10.1186/1471-2105-9-551.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421 DOI 10.1186/1471-2105-10-421.

Cheng Y, Perocchi F. 2015. ProtPhylo: identification of protein-phenotype and protein-protein functional associations via phylogenetic profiling. *Nucleic Acids Research* **43**:W160–W168 DOI 10.1093/nar/gkv455.

Cromar GL, Zhao A, Xiong X, Swapna LS, Loughran N, Song H, Parkinson J. 2016. PhyloPro2.0: a database for the dynamic exploration of phylogenetically conserved proteins and their domain architectures across the Eukarya. *Database* **2016**:1–10 DOI 10.1093/database/baw013.

De Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein co-evolution. *Nature Reviews Genetics* **14**:249–261 DOI 10.1038/nrg3414.

Dey G, Jaimovich A, Collins SR, Seki A, Meyer T. 2015. Systematic discovery of human gene function and principles of modular organization through phylogenetic profiling. *Cell Reports* **10**:993–1006 DOI 10.1016/j.celrep.2015.01.025.

Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M, Taylor R, Dharsee M, Ho Y, Heilbut A, Moore L, Zhang S, Ornatsky O, Bukhman YV, Ethier M, Sheng Y, Vasilescu J, Abu-Farha M, Lambert JP, Duewel HS, Stewart II, Kuehl B, Hogue K, Colwill K, Gladwish K, Muskat B, Kinach R, Adams SL, Moran MF, Morin GB, Topaloglou T, Figeys D. 2007. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular Systems Biology* **3**:1–17 DOI 10.1038/msb4100134.

Eya JC, Ukwuaba VO, Yossa R, Gannam AL. 2015. Interactive effects of dietary lipid and phenotypic feed efficiency on the expression of nuclear and mitochondrial genes involved in the mitochondrial electron transport chain in rainbow trout. *International Journal of Molecular Sciences* **16**:7682–7706 DOI 10.3390/ijms16047682.

Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P. 2016. The Reactome pathway knowledgebase. *Nucleic Acids Research* **44**:D481–D487 DOI 10.1093/nar/gkv1351.

Falk G, Walker JE. 1988. DNA sequence of a gene cluster coding for subunits of the F0 membrane sector of ATP synthase in Rhodospirillum rubrum. Support for modular evolution of the F1 and F0 sectors. *Biochemical Journal* **254**:109–122 DOI 10.1042/bj2540109.

Franceschini A, Lin J, Von Mering C, Jensen LJ. 2015. SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics* **32**:1085–1087 DOI 10.1093/bioinformatics/btv696.

Gabaldon T, Rainey D, Huynen MA. 2005. Tracing the evolution of a large protein complex in the eukaryotes, NADH:ubiquinone oxidoreductase (Complex I). *Journal of Molecular Biology* **348**:857–870 DOI 10.1016/j.jmb.2005.02.067.

Gandrud C, Allaire J, Kent R. 2015. networkD3: D3 JavaScript network graphs from R. R package version 0.2.8.

Girault N. 2017. circosJS: a d3 library to build circular graphs. *Available at https://github.com/nicgirault/circosJS* (accessed on 18 June 2017).

Glatz G, Gogl G, Alexa A, Remenyi A. 2013. Structural mechanism for the specific assembly and activation of the extracellular signal regulated kinase 5 (ERK5) module. *Journal of Biological Chemistry* **288**:8596–8609 DOI 10.1074/jbc.M113.452235.

Glazko GV, Mushegian AR. 2004. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biology* **5**:1–13 DOI 10.1186/gb-2004-5-5-r32.

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, Konishi J, Denda K, Yoshida M. 1989. Evolution of

the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **86**:6661–6665 DOI 10.1073/pnas.86.17.6661.

Guzelin E, Rep M, Grivell LA. 1996. Afg3p, a mitochondrial ATP-dependent metalloprotease, is involved in degradation of mitochondrially-encoded Cox1, Cox3, Cob, Su6, Su8 and Su9 subunits of the inner membrane complexes III, IV and V. *FEBS Letters* **381**:42–46 DOI 10.1016/0014-5793(96)00074-9.

Huynen M, Snel B, Lathe 3rd W, Bork P. 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Research* **10**:1204–1210 DOI 10.1101/gr.10.8.1204.

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proceedings of the National Academy of Sciences of the United States of America* **86**:9355–9359 DOI 10.1073/pnas.86.23.9355.

Jaccard P. 1912. The distribution of the flora of the alpine zone. *New Phytologist* **11**:37–50 DOI 10.1111/j.1469-8137.1912.tb05611.x.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research* **34**:D354–D357 DOI 10.1093/nar/gkj102.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–780 DOI 10.1093/molbev/mst010.

Kensche PR, Van Noort V, Dutilh BE, Huynen MA. 2008. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *Journal of the Royal Society Interface* **5**:151–170 DOI 10.1098/rsif.2007.1047.

Klepinin A, Ounpuu L, Guzun R, Chekulayev V, Timohhina N, Tepp K, Shevchuk I, Schlattner U, Kaambre T. 2016. Simple oxygraphic analysis for the presence of adenylate kinase 1 and 2 in normal and tumor cells. *Journal of Bioenergetics & Biomembranes* **48**:531–548 DOI 10.1007/s10863-016-9687-3.

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

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**:559 DOI 10.1186/1471-2105-9-559.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M. 2004. A map of the interactome network of the metazoan C. elegans. *Science* **303**:540–543 DOI 10.1126/science.1091403.

Li Y, Calvo SE, Gutman R, Liu JS, Mootha VK. 2014. Expansion of biological pathways based on evolutionary inference. *Cell* **158**:213–225 DOI 10.1016/j.cell.2014.05.034.

Liu C, Niu Y, Zhou X, Xu X, Yang Y, Zhang Y, Zheng L. 2015. Cell cycle control, DNA damage repair, and apoptosis-related pathways control pre-ameloblasts differentiation during tooth development. *BMC Genomics* **16**:592 DOI 10.1186/s12864-015-1783-y.

Lopez-Erauskin J, Galino J, Ruiz M, Cuezva JM, Fabregat I, Cacabelos D, Boada J, Martinez J, Ferrer I, Pamplona R, Villarroya F, Portero-Otin M, Fourcade S, Pujol A. 2013. Impaired mitochondrial oxidative phosphorylation in the peroxisomal disease X-linked adrenoleukodystrophy. *Human Molecular Genetics* **22**:3296–3305 DOI 10.1093/hmg/ddt186.

Lucas EK, Dougherty SE, McMeekin LJ, Reid CS, Dobrunz LE, West AB, Hablitz JJ, Cowell RM. 2014. PGC-1alpha provides a transcriptional framework for synchronous neurotransmitter release from parvalbumin-positive interneurons. *Journal of Neuroscience* **34**:14375–14387 DOI 10.1523/JNEUROSCI.1222-14.2014.

Martin FM, Xu X, Von Lohneysen K, Gilmartin TJ, Friedman JS. 2011. SOD2 deficient erythroid cells up-regulate transferrin receptor and down-regulate mitochondrial biogenesis and metabolism. *PLOS ONE* **6**:e16894 DOI 10.1371/journal.pone.0016894.

Miller M, Donat S, Rakette S, Stehle T, Kouwen TR, Diks SH, Dreisbach A, Reilman E, Gronau K, Becher D, Peppelenbosch MP, Van Dijl JM, Ohlsen K. 2010. Staphylococcal PknB as the first prokaryotic representative of the proline-directed kinases. *PLOS ONE* **5**:e9057 DOI 10.1371/journal.pone.0009057.

Mulkidjanian AY, Makarova KS, Galperin MY, Koonin EV. 2007. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nature Reviews Microbiology* **5**:892–899 DOI 10.1038/nrmicro1767.

Ott A, Idali A, Marchais A, Gautheret D. 2012. NAPP: the nucleic acid phylogenetic profile database. *Nucleic Acids Research* **40**:D205–D209 DOI 10.1093/nar/gkr807.

Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, Walford GA, Sugiana C, Boneh A, Chen WK, Hill DE, Vidal M, Evans JG, Thorburn DR, Carr SA, Mootha VK. 2008. A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134**:112–123 DOI 10.1016/j.cell.2008.06.016.

Paul MF, Tzagoloff A. 1995. Mutations in RCA1 and AFG3 inhibit F1-ATPase assembly in *Saccharomyces cerevisiae*. *FEBS Letters* **373**:66–70 DOI 10.1016/0014-5793(95)00979-J.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America* **96**:4285–4288 DOI 10.1073/pnas.96.8.4285.

Pereira SF, Goss L, Dworkin J. 2011. Eukaryote-like serine/threonine kinases and phosphatases in bacteria. *Microbiology and Molecular Biology Reviews* **75**:192–212 DOI 10.1128/MMBR.00042-10.

Psomopoulos FE, Mitkas PA, Ouzounis CA. 2013. Detection of genomic idiosyncrasies using fuzzy phylogenetic profiles. *PLOS ONE* **8**:e52854 DOI 10.1371/journal.pone.0052854.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013. The SILVA ribosomal RNA gene database project: improved

data processing and web-based tools. *Nucleic Acids Research* **41**:D590–D596 DOI 10.1093/nar/gks1219.

Rak M, Gokova S, Tzagoloff A. 2011. Modular assembly of yeast mitochondrial ATP synthase. *EMBO Journal* **30**:920–930 DOI 10.1038/emboj.2010.364.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**:77 DOI 10.1186/1471-2105-12-77.

Rodionov A, Bezginov A, Rose J, Tillier ER. 2011. A new, fast algorithm for detecting protein coevolution using maximum compatible cliques. *Algorithms for Molecular Biology* **6**:1–9 DOI 10.1186/1748-7188-6-17.

Rubinstein JL, Walker JE, Henderson R. 2003. Structure of the mitochondrial ATP synthase by electron cryomicroscopy. *EMBO Journal* **22**:6182–6192 DOI 10.1093/emboj/cdg608.

Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. 2010. CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Research* **38**:D497–D501 DOI 10.1093/nar/gkp914.

Sadreyev IR, Ji F, Cohen E, Ruvkun G, Tabach Y. 2015. PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. *Nucleic Acids Research* **43**:W154–W159 DOI 10.1093/nar/gkv452.

Sales G, Calura E, Cavalieri D, Romualdi C. 2012. graphite—a bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics* **13**:20 DOI 10.1186/1471-2105-13-20.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. 2009. PID: the pathway interaction database. *Nucleic Acids Research* **37**:D674–D679 DOI 10.1093/nar/gkn653.

Severino V, Locker J, Ledda-Columbano GM, Columbano A, Parente A, Chambery A. 2011. Proteomic characterization of early changes induced by triiodothyronine in rat liver. *Journal of Proteome Research* **10**:3212–3224 DOI 10.1021/pr200244f.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**:2498–2504 DOI 10.1101/gr.1239303.

Shih PM, Matzke NJ. 2013. Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proceedings of the National Academy of Sciences of the United States of America* **110**:12355–12360 DOI 10.1073/pnas.1305813110.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313 DOI 10.1093/bioinformatics/btu033.

Steuer R, Kurths J, Daub CO, Weise J, Selbig J. 2002. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18(Suppl 2)**:S231–S240 DOI 10.1093/bioinformatics/18.suppl_2.S231.

Stumpf MP, Thorne T, De Silva E, Stewart R, An HJ, Lappe M, Wiuf C. 2008. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America* **105**:6959–6964 DOI 10.1073/pnas.0708078105.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, Von Mering C. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* **43**:D447–D452 DOI 10.1093/nar/gku1003.

Ta HX, Koskinen P, Holm L. 2011. A novel method for assigning functional linkages to proteins using enhanced phylogenetic trees. *Bioinformatics* **27**:700–706 DOI 10.1093/bioinformatics/btq705.

Tabach Y, Billi AC, Hayes GD, Newman MA, Zuk O, Gabel H, Kamath R, Yacoby K, Chapman B, Garcia SM, Borowsky M, Kim JK, Ruvkun G. 2013a. Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature* **493**:694–698 DOI 10.1038/nature11779.

Tabach Y, Golan T, Hernandez-Hernandez A, Messer AR, Fukuda T, Kouznetsova A, Liu JG, Lilienthal I, Levy C, Ruvkun G. 2013b. Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. *Molecular Systems Biology* **9**:1–17 DOI 10.1038/msb.2013.50.

Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW. 2008. An *in vivo* map of the yeast protein interactome. *Science* **320**:1465–1470 DOI 10.1126/science.1153878.

Tillier ER, Charlebois RL. 2009. The human protein coevolution network. *Genome Research* **19**:1861–1871 DOI 10.1101/gr.092452.109.

Vert JP. 2002. A tree kernel to analyse phylogenetic profiles. *Bioinformatics* **18(Suppl 1)**:S276–S284 DOI 10.1093/bioinformatics/18.suppl_1.S276.

Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research* **33**:D433–D437 DOI 10.1093/nar/gki005.

Wagner GP. 1996. Homologues, natural kinds and the evolution of modularity. *American Zoologist* **36**:36–43 DOI 10.1093/icb/36.1.36.

Wu J, Kasif S, DeLisi C. 2003. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* **19**:1524–1530 DOI 10.1093/bioinformatics/btg187.

Yamada T, Kanehisa M, Goto S. 2006. Extraction of phylogenetic network modules from the metabolic network. *BMC Bioinformatics* **7**:130 DOI 10.1186/1471-2105-7-130.

Zhou Y, Wang R, Li L, Xia X, Sun Z. 2006. Inferring functional linkages between proteins from evolutionary scenarios. *Journal of Molecular Biology* **359**:1150–1159 DOI 10.1016/j.jmb.2006.04.011.