BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

KinOrtho: a method for mapping human kinase orthologs across the tree of life and illuminating understudied kinases

Liang-Chin Huang[1], Rahil Taujale[1], Nathan Gravel[2], Aarya Venkat[3], Wayland Yeung[1], Dominic P. Byrne[4], Patrick A. Eyers[4] and Natarajan Kannan[1,3]*

*Correspondence:
nkannan@uga.edu
[3] Department
of Biochemistry
and Molecular Biology,
University of Georgia, 120
Green St., Athens, GA 30602,
USA
Full list of author information
is available at the end of the
article

Abstract

Background: Protein kinases are among the largest druggable family of signaling proteins, involved in various human diseases, including cancers and neurodegenerative disorders. Despite their clinical relevance, nearly 30% of the 545 human protein kinases remain highly understudied. Comparative genomics is a powerful approach for predicting and investigating the functions of understudied kinases. However, an incomplete knowledge of kinase orthologs across fully sequenced kinomes severely limits the application of comparative genomics approaches for illuminating understudied kinases. Here, we introduce KinOrtho, a query- and graph-based orthology inference method that combines full-length and domain-based approaches to map one-to-one kinase orthologs across 17 thousand species.

Results: Using multiple metrics, we show that KinOrtho performed better than existing methods in identifying kinase orthologs across evolutionarily divergent species and eliminated potential false positives by flagging sequences without a proper kinase domain for further evaluation. We demonstrate the advantage of using domain-based approaches for identifying domain fusion events, highlighting a case between an understudied serine/threonine kinase TAOK1 and a metabolic kinase PIK3C2A with high co-expression in human cells. We also identify evolutionary fission events involving the understudied OBSCN kinase domains, further highlighting the value of domain-based orthology inference approaches. Using KinOrtho-defined orthologs, Gene Ontology annotations, and machine learning, we propose putative biological functions of several understudied kinases, including the role of TP53RK in cell cycle checkpoint(s), the involvement of TSSK3 and TSSK6 in acrosomal vesicle localization, and potential functions for the ULK4 pseudokinase in neuronal development.

Conclusions: In sum, KinOrtho presents a novel query-based tool to identify one-to-one orthologous relationships across thousands of proteomes that can be applied to any protein family of interest. We exploit KinOrtho here to identify kinase orthologs and show that its well-curated kinome ortholog set can serve as a valuable resource for illuminating understudied kinases, and the KinOrtho framework can be extended to any protein-family of interest.

Huang *et al. BMC Bioinformatics*   (2021) 22:446

Page 2 of 25

Background

Since the completion of the human genome project, thousands of species have been fully sequenced [1], providing a broader coverage of species diversity across the tree of life. "Moonshot" approaches, such as the Earth BioGenome Project (EBP), aim to catalog, and then characterize, genomes across eukaryotic biodiversity during the next decade [2]. The acquisition of genomic (and their associated proteomic) datasets enables the accurate prediction of protein functions through ever-deeper evolutionary analysis of related sequences [3]. Protein kinases transfer the gamma phosphate group from ATP to an expanding subset of amino acids in their regulatory targets [4]. They can be distinguished from other mechanistically related enzymes, such as metabolic and glycan-modifying kinases [5]. Protein kinases represent one of the largest druggable families of signaling proteins abnormally regulated in various human diseases, including most human cancers [6–8]. The human genome encodes nearly 550 protein kinase-related genes (collectively referred to as the human kinome) that have been broadly classified into major groups and families [9, 10]. A majority of the human kinome members have been functionally characterized in multiple model organisms; however, nearly 30% of human kinases remain understudied, despite multi-organism knowledge of their primary sequence [11–13]. These are collectively referred to as "dark" kinases based on a subset of metrics such as the number of published papers (Jensen PubMed score [14] < 50 and PubTator score [15] < 150) and grant funding (no R01). Many of the understudied kinases, such as RIO and NEK families, contain clear orthologs in a majority of eukaryotic genomes, suggesting essential (rate-limiting) biological functions across life [16–18]. A major focus of the Illuminating the Druggable Genome (IDG; https://commonfund.nih.gov/idg/index) consortium is to characterize the functions of these understudied proteins as a conceptual starting point for developing new drugs for a wide range of diseases such as cancer, neurodegenerative and autoimmune disorders that are associated with abnormal kinome signaling [19].

Comparative genomics is a powerful approach for inferring gene functions and is based on the assumption that genes descended from the same ancestor are likely to retain commonly shared functions [20, 21]. These gene descendants are called orthologs and paralogs, two major types of homologs related to speciation and duplication events, respectively [22]. Paralogs can be further categorized as in-paralogs and out-paralogs: the former arises from duplication after speciation, while the latter arises from duplication before speciation [23]. The concept of "one-to-one" orthologous relationships (one protein in one species versus one protein in the other species) has been extended to "one-to-many" or "many-to-many" relationships and are collectively termed orthologous groups [24]. Co-orthologs are defined as a pair of genes from the same orthologous group but different species [25]. Given the importance of these relationships for functional analysis, several orthology inference methods have been developed. We have previously used these approaches to analyze canonical protein kinases and pseudokinases, including a broad survey of pseudoenzymes [26],

pseudokinases [27], and a variety of understudied kinases whose biological function remains unknown, despite conservation in various eukaryotic lineages [28–30].

Current orthology inference methods can be broadly classified into two major categories: tree-based methods [31–34] and graph-based methods [35–43]. Tree-based methods, such as EnsemblCompara [31], construct reconciled trees based on gene trees and corresponding species trees and then infer the type of evolutionary event (speciation/duplication) that represents each internal node of the tree. By contrast, graph-based methods, such as OrthoMCL [35], avoid building trees and identify hypothetical orthologous relationships (orthologs/paralogs) typically by two main steps: graph construction and clustering. Graph-based methods represent proteins as nodes and the relationships connecting the nodes (sequence similarity, for example) as edges. The nodes in the graph are then clustered into orthologous groups by different strategies [44–46]. Tree-based methods are generally more accurate than graph-based methods, depending on the accuracy of species trees [46, 47]. However, tree-based methods are computationally expensive, limiting the exploration of thousands of species across the tree of life [47]. In contrast, graph-based methods are faster, but the increased speed is generally achieved at the cost of reduced sensitivity. When applied to large datasets, the performance of graph-based methods is comparable to tree-based methods and, in some cases, even better than tree-based methods [48].

Most orthology inference methods rely on time-consuming all-vs-all sequence similarity searches across full-length gene or protein sequences across entire genomes. As such, these methods are not designed for focused analysis on individual gene families. Within large protein families, such as the protein kinase superfamily, traditional orthology inference methods display high false-positive rates since they do not consider the conservation of known functional domains. Some existing methods identify sequences as putative orthologs that almost certainly lack the classical bilobal kinase domain. In contrast, domain-based methods, such as Hierarchical grouping of Orthologous and Paralogous Sequences (HOPS) [49], FlowerPower [50], Domain based Detection of Orthologs (DODO) [51], Microbial Genome Database (MBGD) [52], and Domainoid [53], are tailored to identify evolutionary relationships based on functionally relevant regions, notably domains, of a protein. However, the performances of these methods are reliant on the annotation of domains based on prior knowledge, thereby making it challenging to identify novel domains and relationships. A good example of this is the discovery of atypical kinases with very low sequence identity compared to search sequences, such as the atypical SelO kinase [54].

To address the above challenges in orthology prediction, we developed KinOrtho as a complementary approach for efficient and accurate identification of human kinase orthologs across ~ 17,000 species, extending well beyond the 15 model organisms defined in the seminal study of Manning and colleagues [9] and a recently updated kinase-centric database with kinases from 2000 species [55]. KinOrtho is query-based and achieves increased sensitivity by combining similarities in the commonly conserved protein kinase domain and flanking regulatory domains. This enables us to develop one-to-one orthology relationships that provide a finer resolution of orthologs across species than previous efforts. We apply KinOrtho towards identifying putative proteins involving fusion or fission events, or so-called "Rosetta Stone protein" [56]. By integrating

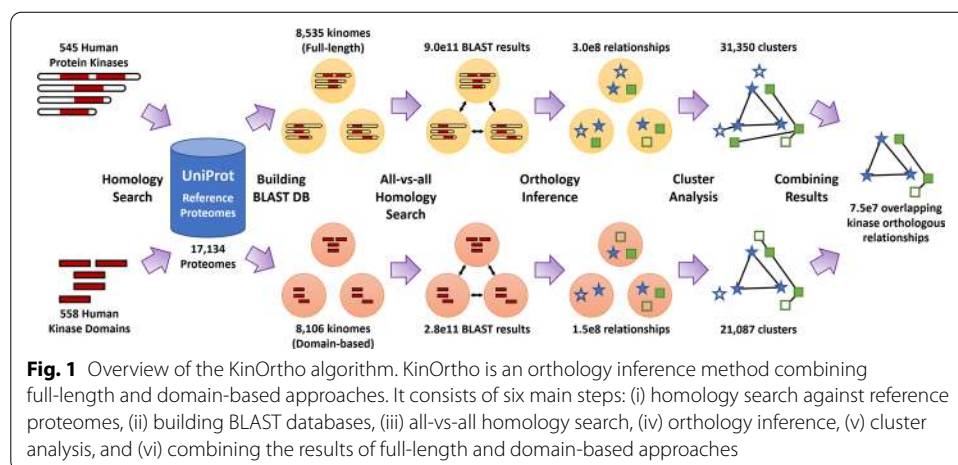Huang *et al. BMC Bioinformatics*   (2021) 22:446

Page 4 of 25

evolutionary information with gene expression patterns, we identify a potential functional association between an understudied kinase "Serine/threonine-protein kinase TAO1" (TAOK1) and a metabolic kinase "Phosphatidylinositol-4-phosphate 3-kinase" (PI3KC2A) in autophagy. Using KinOrtho-defined orthologs, Gene Ontology (GO) annotations, and machine learning models, we prioritize understudied kinases for functional studies by developing a Novel Inferred Annotation Score (NIAS). The KinOrtho pipeline and ortholog datasets are available at the GitHub repository (https://github.com/esbgkannan/KinOrtho), and the patterns of conservation in aligned orthologs sequences are visualized in both KinView [57] (https://prokino.uga.edu/kinview/) and the IDG resource Pharos [58] (https://pharos.nih.gov/).

Results

Overview of KinOrtho algorithm

KinOrtho is a query- and graph-based orthology inference method that combines full-length and domain-based orthology inference approaches. It consists of two pipelines (full-length and domain-based) and six main steps: (i) homology search, (ii) building kinome databases, (iii) all-vs-all homology search, (iv) orthology inference, (v) cluster analysis, and (vi) combining the results from two pipelines (Fig. 1).

Because KinOrtho is query-based, it omits a large portion of unnecessary sequence comparisons unrelated to the query sequence(s). This characteristic makes KinOrtho a more efficient tool to identify orthologs of interest across the tree of life. We applied KinOrtho to identify the orthologs of 545 human kinases across some 17,000 species found in UniProt reference proteomes [59]. Without target genes, traditional orthology inference methods start from an all-vs-all homology search, which would require orders of magnitude (more than two quadrillion) pairwise sequence comparisons for this sample of reference proteomes. Because our query sequences were human kinases, only about eight thousand species were found to have human kinase homologs, which resulted in a nearly 2000-fold reduction in the number of comparisons to be made (Additional file 1: Table S1). This makes KinOrtho one of the most computationally efficient orthology inference methods currently available for the identification of kinases.

Fig. 1 Overview of the KinOrtho algorithm. KinOrtho is an orthology inference method combining full-length and domain-based approaches. It consists of six main steps: (i) homology search against reference proteomes, (ii) building BLAST databases, (iii) all-vs-all homology search, (iv) orthology inference, (v) cluster analysis, and (vi) combining the results of full-length and domain-based approaches

When performing orthology inference, we adopted the definition of orthologous relationship used by OrthoMCL [35] (see Methods), which resulted in twice as many orthologous relationships using similarity in full-length sequences compared to similarities within the kinase domain alone (302 million for full-length vs. 148 million for kinase domain; Additional file 1: Table S1). However, the application of graph-based clustering and further refinement of the clusters resulted in a comparable number of orthologous relationships in the full-length (97 million) and domain-based (100 million) pipelines (referred to as full-length set and domain-based set, respectively). Finally, the combination of both pipelines resulted in 75 million overlapping orthologous relationships (termed "overlapping set" from here on), including ~133,000 relationships between human kinases and kinases from other species (Additional file 1: Table S1). Since this is the most refined set of relationships obtained from KinOrtho, this overlapping set will be referred to as KinOrtho throughout this manuscript. In contrast, the full-length and domain-based results will be explicitly stated when mentioned.

Benchmarking and comparison of KinOrtho with other orthology inference methods

To evaluate and compare the performance of KinOrtho with other orthology inference methods in identifying kinase orthologs, we applied KinOrtho to the well-curated Quest for Orthologs (QfO) reference proteomes 2018 [60]. As shown in Fig. 2a, the overall comparison metrics for KinOrtho are better (in terms of the overall precision and recall) than existing methods in the benchmarking datasets based on the enzyme classification conservation test, agreement with reference gene phylogenies, and species tree discordance benchmarks. The remaining metrics are shown in Additional file 1: Figure S1. It is also important to note that the selection of orthologs from KinOrtho's full-length pipeline, domain-based pipeline, and the overlapping results all yielded similar performance (compared to other methods), suggesting robustness and agreement across these methods.

Next, a direct comparison between the pairs of orthologs identified by KinOrtho was performed alongside other methods to gain insights into overlapping predictions. In general, 35–60% of the orthologs identified by KinOrtho were also identified by other methods (blue bars in Fig. 2b). Besides, KinOrtho (overlapping set) consistently found orthologs not identified by other methods (orange bars in Fig. 2b). Several unique KinOrtho-defined human kinase orthologs, such as cyclin-dependent protein kinase (CDK) orthologs, are described in Supplementary Results and shown in Additional file 1: Figure S2–S4. On the other hand, KinOrtho consistently omitted at least 10% of the orthologs (average: 23.2%; gray bars in Fig. 2b) identified by other methods. This number is significantly reduced when considering KinOrtho full-length or domain-based sets alone (average: 14.3% and 17.2%, respectively; gray bars in Additional file 1: Figure S5), suggesting that KinOrtho eliminates putative ortholog sequences that lack the well-defined bilobal kinase domain associated with protein kinases. The ability of KinOrtho to delineate the orthologs based on the protein kinase domain against the orthologs based on other conserved domains is described in Supplementary Results and shown in Additional file 1: Figure S6. Additional details about the utility and benchmarking of the domain-based approach are discussed below. Finally, we generated a similarity heat map to quantify orthology predictions by KinOrtho and other methods (Fig. 2c). Similarities
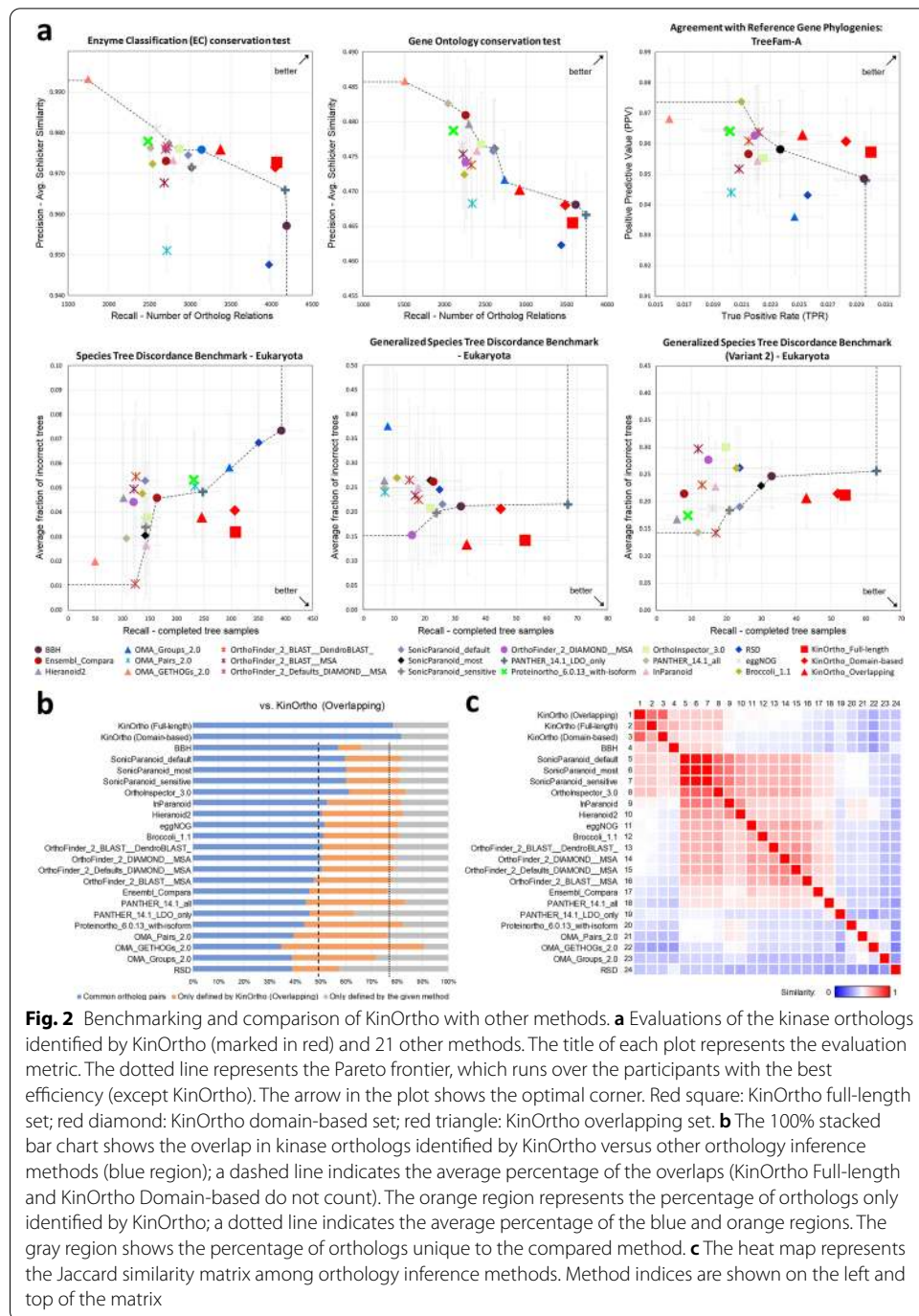
Fig. 2 Benchmarking and comparison of KinOrtho with other methods. **a** Evaluations of the kinase orthologs identified by KinOrtho (marked in red) and 21 other methods. The title of each plot represents the evaluation metric. The dotted line represents the Pareto frontier, which runs over the participants with the best efficiency (except KinOrtho). The arrow in the plot shows the optimal corner. Red square: KinOrtho full-length set; red diamond: KinOrtho domain-based set; red triangle: KinOrtho overlapping set. **b** The 100% stacked bar chart shows the overlap in kinase orthologs identified by KinOrtho versus other orthology inference methods (blue region); a dashed line indicates the average percentage of the overlaps (KinOrtho Full-length and KinOrtho Domain-based do not count). The orange region represents the percentage of orthologs only identified by KinOrtho; a dotted line indicates the average percentage of the blue and orange regions. The gray region shows the percentage of orthologs unique to the compared method. **c** The heat map represents the Jaccard similarity matrix among orthology inference methods. Method indices are shown on the left and top of the matrix

measured by the Jaccard similarity coefficient between two ortholog sets ranged from 25.3% (Reciprocal Smallest Distance (RSD) [61] vs. Orthologous Matrix (OMA) [39]) to 81.5% (SonicParanoid [40] vs. OrthoInspector [41]. The average similarity among all methods was 50.4%. Orthology results from Bidirectional Best Hits (BBH) [62] and two BBH- and graph-based methods, OrthoInspector and SonicParanoid, were found to have the most agreement with KinOrtho results (average similarities: 59.9%, 59.8%, and 58.7%, respectively).

Inferring functional associations using KinOrtho-based identification of kinase domain fusion and fission events

KinOrtho's ability to find orthologs for individual domains allows identifying domain fusion and fission events for kinases with multiple kinase domains. In the human kinome, there are 13 kinases with two tandem kinase domains within the same polypeptide, many of which are functionally annotated phosphorylation targets of Mitogen Activated Protein Kinase (MAPK) signaling pathways. Figure 3a illustrates the four scenarios of finding domain-based orthologs for these 13 kinases: (1) tandem domains in one kinase match tandem domains in another kinase, (2) tandem domains in one kinase match tandem domains in another kinase in reverse order, (3) two domains from two human kinases match tandem domains in one kinase from another species, and (4) tandem domains in one human kinase match two domains from two kinases in different species. Traditional full-length BBH-based orthology inference methods do not have the resolution to distinguish between these
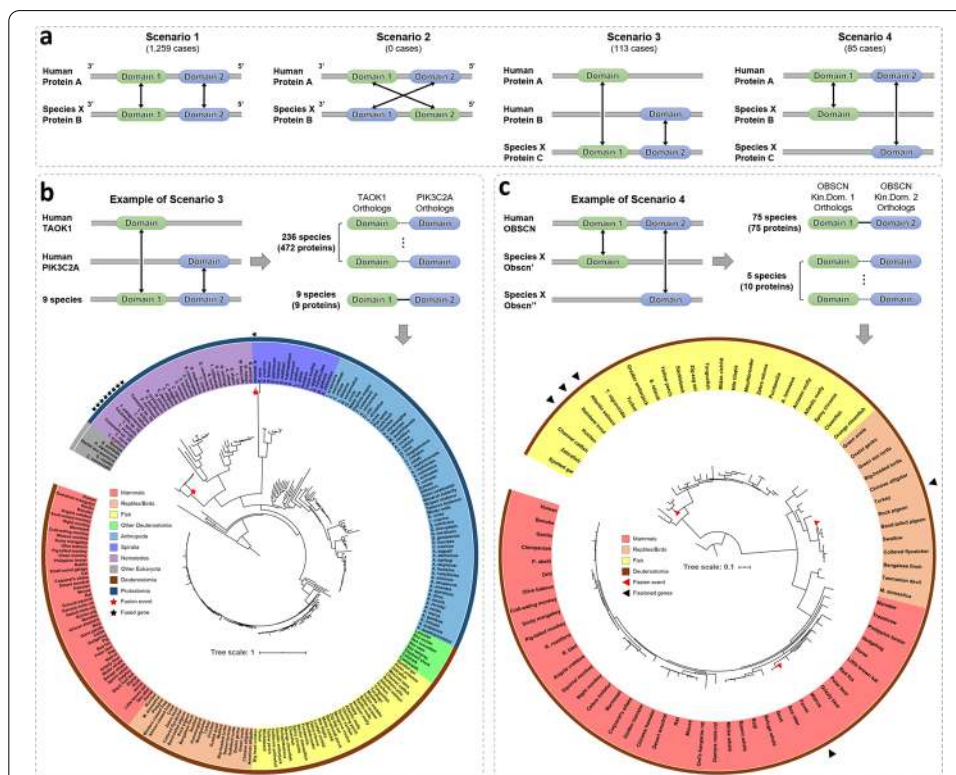

Fig. 3 Scenarios of a single protein with tandem kinase domains and examples of potential gene fusion and fission events. **a** Four scenarios of a single protein with tandem kinase domains and their domain-based orthologs. Arrows represent orthologous pairs. The number of cases for each scenario is shown in parentheses. **b** Example of Scenario 3 and phylogenetic tree analysis on the fusion event of TAOK1 orthologs and PIK3C2A orthologs. Human TAOK1's and PIK3C2A's domain-based orthologs in 245 species were aligned, concatenated (represented by a dotted line if these two domains are from different species), and used to build a phylogenetic tree. Species names are labeled at the leaves of the circular-mode phylogenetic tree. The leaves are colored according to the clade of the species (refer to legend). Black stars mark the species with a TAOK1-PIK3C2A fused gene. The time when potential fusion events occurred is indicated by red stars on the tree. **c** Example of Scenario 4 and analysis of the fission event of OBSCN orthologs. The domain-based orthologs of human OBSCN's two kinase domains in 80 species were used to build a phylogenetic tree. Black triangles mark the species with fission events. Red triangles indicate the time when potential fission events occurred

scenarios. However, KinOrtho's domain-based approach allows the definition of orthologs from all scenarios, thus identifying fusion and fission events in orthologous sequences.

Scenario 3 in Fig. 3a reflects potential domain fusion events. Although we identified 113 potential fusion events, we use TAOK1 and PIK3C2A as an example to illustrate how integrating evolutionary data with other contextual data (protein-protein interaction, co-expressions, and co-occurrence) can reveal potential functions for understudied kinases (Additional file 2). TAOK1, an understudied kinase, belongs to the STE20 family, while PIK3C2A belongs to the PI3K family. We found nine instances of potential domain fusion events between these two kinase domains (Fig. 3b). In addition, we identified 236 species with both TAOK1 and PIK3C2A domain-based orthologs in different kinases. We concatenated the sequences of TAOK1 orthologs and PIK3C2A orthologs for each species and then built a phylogenetic tree. We found that eight kinases with TAOK1 and PIK3C2A domains reside in the same clade of Nematodes. Based on this, we postulate two potential fusion events (indicated by red stars on the phylogenetic tree, Fig. 3b). Proteins involved in a fusion event usually belong to the same functional category [63]. As an example, TAOK1, an understudied kinase, shows high co-expression with PIK3C2A in 17,382 normal samples and 1376 cancer samples in the Genotype-Tissue Expression project [64] (GTEx, version 8) and the Cancer Dependency Portal [65] (DepMap, 20Q4), respectively. The correlations (Pearson correlation coefficient = 0.856 in normal samples and 0.612 in cancer samples) are among the top 0.15% of all kinase pairs (Additional file 1: Figure S7). The co-expressed patterns are conserved in *A.aegypti*, *B.taurus*, *D.melanogaster*, and *S.mansoni* (STRING [66], version 11.0), suggesting a possible physical interaction. Moreover, TAOK1 and PIK3C2A have been reported to be involved in the autophagy response [67]. Based on these observations, we predict a functional association between the understudied kinase TAOK1 and PIK3C2A in human cellular biology, perhaps involving communication between the membrane, where phospholipids are sensed, and the cytosol, where TAOK1 has known functions in relaying information to MAPK pathways.
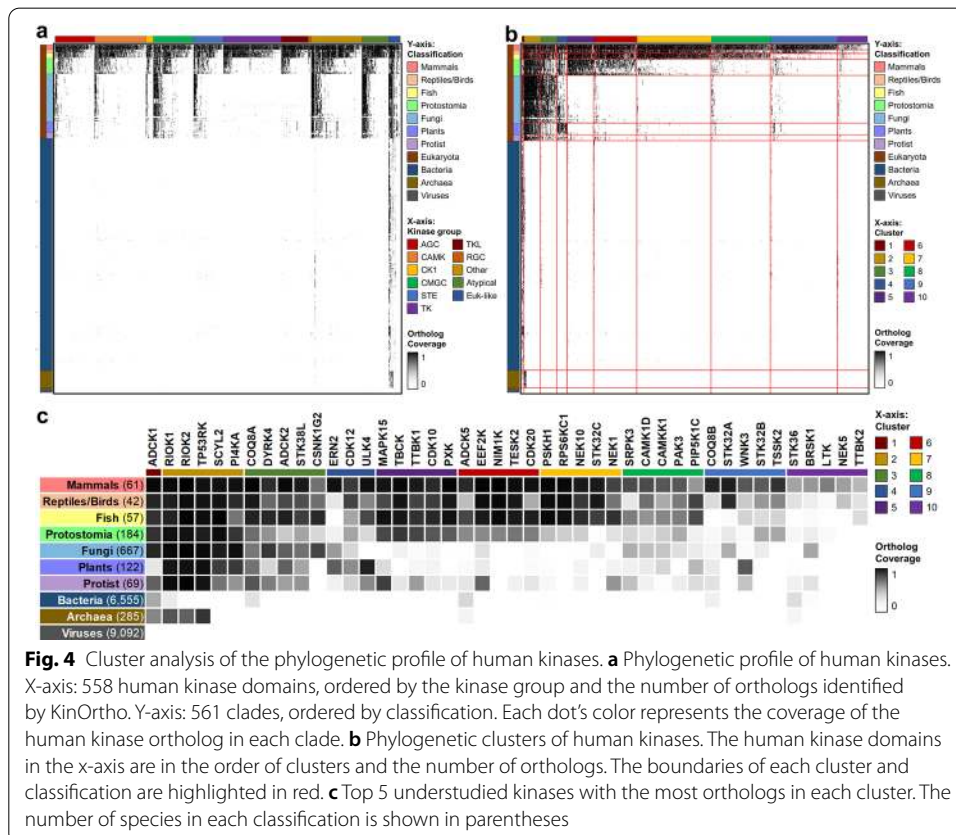
We also analyzed cases in Scenario 4 for potential domain fission events. We found ten kinases in five species (ferrets, turkeys, Atlantic salmon, rainbow trout, and huchen) matching the tandem kinase domains in human Obscurin (OBSCN) kinase (Fig. 3c). The tandem kinase domain arrangement in OBSCN is conserved in 75 species (Scenario 1). In species where the tandem domains are encoded in two different proteins, we concatenated the domains and performed phylogenetic comparisons with species where the two domains are naturally fused. The concatenated sequences of Atlantic salmon, rainbow trout, and huchen occur in the same clade. Based on the phylogenetic tree, we estimate three kinase domain fission events (marked by red triangles in Fig. 3c). Although the functional significance of these fission events is unclear, the established role of OBSCN in eye development [68, 69] suggests a role for these events in the evolution of vision in these species [70–73].

Phylogenetic profile analysis reveals the evolutionary depth of human protein kinase conservation and enriched molecular functions across species

We next sought to classify human kinases based on conservation depth across species by building a phylogenetic profile of KinOrtho-defined orthologs. Figure 4a highlights a human kinase phylogenetic profile consisting of 558 human kinase domains and their orthologs across 561 clades. As expected, human kinase orthologs are barely present

Huang *et al. BMC Bioinformatics*   (2021) 22:446

Page 9 of 25

in bacteria, archaea, and viruses, except for the orthologs of eukaryotic-like protein kinases. Consistent with previous findings, four eukaryotic-like kinases, Protein adenylyltransferase SelO, mitochondrial (SELENOO; 3936 orthologs), AarF domain-containing protein kinase 1 (ADCK1; 3258 orthologs), Ketosamine-3-kinase (FN3KRP; 2234 orthologs), and Serine/threonine-protein kinase RIO1 (RIOK1; 1849 orthologs) have the most orthologs. In contrast, Casein kinase II subunit alpha 3 (CSNK2A3), Rhodopsin kinase GRK1 (GRK1), Putative serine/threonine-protein kinase PR-KY (PRKY), and Probable serine/threonine-protein kinase SIK1B (SIK1B) are "orphan" kinases with no orthologs based on KinOrtho's stringent criteria. The phylogenetic profile also shows that the kinases in tyrosine kinase (TK), tyrosine kinase-like (TKL), and receptor guanylate cyclase (RGC) groups are mainly conserved in Metazoa (including mammals, reptiles, birds, fish, and protostomes), which is consistent with the findings of a previous study [74].

Many understudied kinases have escaped analysis due to weak conservation in model organisms. Based on the distribution of orthologs across different species, we organized human kinases into 10 clusters (Fig. 4b). The top 5 understudied kinases with the most orthologs in each cluster are highlighted in Fig. 4c. Kinases within each cluster are closely related (small Euclidean distance) with high co-occurrence with each other. Using this clustered phylogenetic profile, we sought to identify potentially conserved kinase-regulated biological functions across species. For example, because the kinases in Cluster 5 are highly conserved in Metazoa, we can hypothesize a role for these kinases in

Fig. 4 Cluster analysis of the phylogenetic profile of human kinases. **a** Phylogenetic profile of human kinases. X-axis: 558 human kinase domains, ordered by the kinase group and the number of orthologs identified by KinOrtho. Y-axis: 561 clades, ordered by classification. Each dot's color represents the coverage of the human kinase ortholog in each clade. **b** Phylogenetic clusters of human kinases. The human kinase domains in the x-axis are in the order of clusters and the number of orthologs. The boundaries of each cluster and classification are highlighted in red. **c** Top 5 understudied kinases with the most orthologs in each cluster. The number of species in each classification is shown in parentheses

metazoan-specific biological functions. To this end, we performed Gene Ontology (GO) enrichment analyses using the GO annotations of all kinases as background. We identified 802 significantly enriched GO terms (false discovery rate (FDR) < 0.05), and the GO term is annotated for at least five human kinases in the cluster. The top three enriched GO terms for each cluster are shown in Table 1, and the entire list is shown in Additional file 3.

The orthologs of human kinases in Cluster 2 are present in most eukaryotic species. The most enriched GO term in Cluster 2 is a biological process term "cell cycle checkpoint" (GO:0000075), which encompasses a variety of DNA and spindle-assembly checkpoints, well-established control mechanisms that control progression through the eukaryotic cell cycle [75]. EKC/KEOPS complex subunit TP53RK (TP53RK), an understudied kinase with 1367 orthologs, plays a vital role in the cell cycle and G1 checkpoint control [76, 77]. However, this GO term is currently absent in both human TP53RK annotation and TP53RK ortholog annotation. Kinases in Cluster 9 are mostly present in mammals. Consistently, the cellular component term "acrosomal vesicle" (GO:0001669) is the most enriched GO term in Cluster 9. Acrosomal vesicles, components in the sperm's head, contain enzymes essential for fertilization [78]. All members of testis-specific serine/threonine-kinases (TSSK) belong to Cluster 9, and they are all understudied kinases: TSSK1B, TSSK2, TSSK3, TSSK4, and TSSK6. Currently, TSSK1B, TSSK2, and TSSK4 are annotated with this GO term. Although TSSK3, TSSK6, and their orthologs lack this annotation, both TSSK3 and TSSK6 are highly expressed in testis (median Transcripts Per Million = 60.86 and 424.6, respectively) [64]. TSSK6 is also reported to be involved in the acrosome reaction and egg fertilization [79]. Therefore, based on the cluster analysis of the phylogenetic profile, we predict TSSK3 and TSSK6 function in acrosomal biology and vesicle localization.

Machine learning model to prioritize understudied kinases using KinOrtho and GO annotations

The human kinome contains several understudied kinases of unknown function. We next wanted to investigate if KinOrtho-defined orthologs, along with sequence similarities and GO annotations from different species, can be used to infer the functions of understudied kinases using "guilt-by-association" [80] and machine learning methods. To this end, we trained machine learning classifiers using orthology and functional annotations of well-studied human kinases to predict whether functional annotation could be transferred from orthologs in other species to human kinases. In brief, we built training sets using the manually curated GO annotations of well-studied kinases (Fig. 5). The input features of each training instance represent the GO annotation status of human kinase orthologs; the output shows whether the human kinase has this GO term annotation or not. Input features were weighted based on the sequence similarities between the human kinase and orthologs. The training sets contained 0.3 million instances with 730 GO terms for 393 well-studied kinases and their orthologs across 176 species. After class balancing and 10-fold cross-validation, random forest displayed high prediction accuracies among various machine learning methods attempted (90.9%, 92.1%, and 95.5% for the training sets of biological process, cellular component, and molecular function, respectively; Table 2).

Table 1 GO term enrichment analysis on the phylogenetic clusters of human kinases

Cluster	#Kinases	Conserved in	Enriched GO term	K	k	Fold	FDR
2	27128	Eukaryota	GO:0000075 (cell cycle checkpoint)	4765	3619	3.66	< 1.0E−320
			GO:0098805 (Whole membrane)	1289	890	3.33	2.3E−306
			GO:0046488 (Phosphatidylinositol metabolic process)	2634	1772	3.24	< 1.0E−320
3	16899	Vertebrata Protostomia (partial) Fungi	GO:0000165 (MAPK cascade)	1943	1025	4.08	< 1.0E−320
			GO:0005694 (Chromosome)	284	120	3.27	< 1.0E−320
			GO:1902749 (Regulation of cell cycle G2/M phase transition)	983	380	2.99	2.2E−10
4	6933	Vertebrata Protostomia (partial) Plants	GO:0010468 (Regulation of gene expression)	9690	796	1.55	< 1.0E−320
			GO:0006807 (Nitrogen compound metabolic process)	20390	1470	1.36	< 1.0E−320
			GO:0044238 (Primary metabolic process)	23630	1663	1.33	2.1E−09
5	12496	Metazoa	GO:0008047 (Enzyme activator activity)	778	341	4.59	1.9E−10
			GO:0090287 (Regulation of cellular response to growth factor stimulus)	627	249	4.15	3.7E−11
			GO:0016055 (Wnt signaling pathway)	415	161	4.06	< 1.0E−320
6	16115	Vertebrata Protostomia (partial)	GO:0045669 (Positive regulation of osteoblast differentiation)	437	243	4.51	1.1E−102
			GO:0030500 (Regulation of bone mineralization)	508	244	3.90	< 1.0E−320
			GO:0034645 (Cellular macromolecule biosynthetic process)	797	330	3.36	1.0E−10
7	17842	Vertebrata	GO:0042629 (Mast cell granule)	54	54	7.33	5.5E−46
			GO:0030522 (Intracellular receptor signaling pathway)	152	146	7.04	8.6E−116
			GO:0042102 (Positive regulation of T cell proliferation)	211	202	7.01	1.6E−159
8	10994	Mammals Reptiles/Birds (partial) Fish	GO:0051965 (Positive regulation of synapse assembly)	250	177	8.42	2.2E−128
			GO:0005004 (GPI-linked ephrin receptor activity)	215	130	7.19	5.4E−81
			GO:0005005 (Transmembrane-ephrin receptor activity)	267	143	6.37	3.9E−79
9	10202	Mammals Reptiles/Birds (partial)	GO:0001669 (Acrosomal vesicle)	103	75	9.33	3.2E−58
			GO:0031253 (Cell projection membrane)	283	161	7.29	1.8E−99
			GO:0014068 (Positive regulation of phosphatidylinositol 3-kinase signaling)	413	164	5.09	4.8E−11
10	3043	Mammals (partial)	GO:0035173 (Histone kinase activity)	665	96	6.20	1.1E−11
			GO:0050321 (Tau-protein kinase activity)	318	43	5.81	8.7E−12
			GO:0051051 (Negative regulation of transport)	818	98	5.15	< 1.0E−320

Based on fold enrichment, only the top three enriched GO terms in each cluster are shown. Cluster: the cluster ID shown in Fig. 4b; #Kinases: the total number of human kinases and their orthologs in the cluster; K: the total number of kinases associated with the GO term; k: the number of kinases associated with the GO term in the cluster; Fold: fold enrichment; FDR: false discovery rate
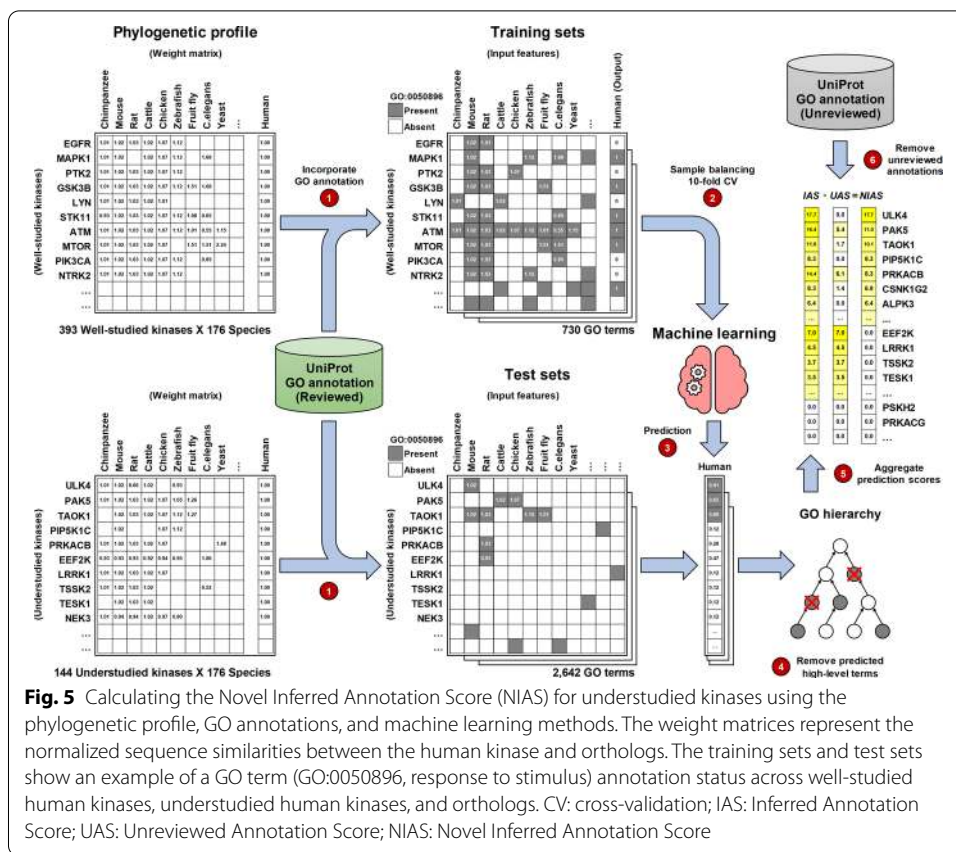
Fig. 5 Calculating the Novel Inferred Annotation Score (NIAS) for understudied kinases using the phylogenetic profile, GO annotations, and machine learning methods. The weight matrices represent the normalized sequence similarities between the human kinase and orthologs. The training sets and test sets show an example of a GO term (GO:0050896, response to stimulus) annotation status across well-studied human kinases, understudied human kinases, and orthologs. CV: cross-validation; IAS: Inferred Annotation Score; UAS: Unreviewed Annotation Score; NIAS: Novel Inferred Annotation Score

Table 2 Performance of GO annotation prediction in each training set

GO domain	Model	Accuracy	Precision	Recall	F-measure	AUC
Biological process	Logistic regression	0.896	0.942	0.844	0.891	0.914
	SVM	0.896	0.943	0.843	0.890	0.896
	Random forest	<u>0.909</u>	<u>0.944</u>	<u>0.869</u>	<u>0.905</u>	<u>0.923</u>
Cellular component	Logistic regression	0.908	0.955	0.856	0.903	0.924
	SVM	0.908	0.955	0.858	0.903	0.908
	Random forest	<u>0.921</u>	<u>0.959</u>	<u>0.880</u>	<u>0.918</u>	<u>0.932</u>
Molecular function	Logistic regression	0.940	0.959	0.919	0.939	0.961
	SVM	0.942	0.959	0.925	0.941	0.942
	Random forest	<u>0.955</u>	<u>0.964</u>	<u>0.945</u>	<u>0.955</u>	<u>0.965</u>

The best performance in each measurement and each GO domain is highlighted in underlined

Next, we used the trained models (random forests) to predict the functions of understudied kinases. To this end, we constructed test sets in which understudied human kinases could be annotated based on the GO terms available for one or more of their orthologs. This resulted in 16 thousand instances with 2642 GO terms for 144 understudied kinases. Application of the pre-trained random forest models on the test sets resulted in 11,573 predictions of kinase-GO term pairs as present (Additional file 1: Table S2). Among these predicted annotations, 8933 predictions (77.2%) already existed in the UniProt as manually reviewed annotations, while 2640 predictions (22.8%) did not have UniProt annotations. Instead of referring to these 2640 predictions as false

positives, we considered them as missing annotations. In fact, 1452 of them (55%) were found to be unreviewed electronic annotations from Ensembl [81], InterPro [82], the UniProt Consortium, or the GO Consortium. The remaining 1188 annotations, including 236 lowest-level GO term annotations, were novel inferred annotations (available in Additional file 4). By aggregating the prediction score of each novel inferred annotation, we calculated a Novel Inferred Annotation Score (NIAS) for each understudied human kinase (the last step in Fig. 5).

Our analysis reveals that Serine/threonine-protein kinase ULK4 (ULK4) has the highest NIAS among all understudied human kinases. It has 22 novel inferred annotations. Fifteen of them with prediction scores higher than 0.9 are inferred from mouse/rat Ulk4. Twelve of these inferred annotations with a high score are associated with neuronal function and brain development, such as "ventricular system development" (GO:0021591), "corpus callosum development" (GO:0022038), "neuronal stem cell division" (GO:0036445), and "GABAergic neuron differentiation" (GO:0097154). A role for human ULK4 in neuronal function and brain development has been suggested [83–85], and it is an unusual pseudokinase that binds to nucleotides in the absence of cations [86, 87]. Serine/threonine-protein kinase PAK 5 (PAK5), an understudied kinase with the second highest NIAS, has 12 novel inferred annotations. The association between PAK5 and "activation of MAPK activity" (GO:0000187), the GO term with the highest prediction score for PAK5, is also known from the literature, where these kinases act as upstream regulators of MAPK modules [88]. We also identified 80 understudied kinases with a NIAS of 0. We can still prioritize these proteins by Inferred Annotation Score (IAS, which aggregates all prediction scores regardless of existing unreviewed annotations; see Methods) for further manual curation or experimental validation. For example, the NIAS of Eukaryotic elongation factor 2 kinase (EEF2K) is 0. Kinases in this list include pseudokinases such as Serine/threonine-protein kinase H2 (PSKH2), which represent the "darkest" of kinases with little or no information across species and no functional biology currently reported [28].

Discussion

Here we map human kinase orthologs across diverse species by developing a kinase orthology inference method called KinOrtho. We demonstrate that KinOrtho performs better than existing orthology inference methods based on comparisons across standard benchmarking datasets and metrics. KinOrtho utilizes domain-based orthology inference to eliminate orthologs with no kinase domains, allowing researchers to focus on the functional domains of interest. KinOrtho's query-based characteristic enables users to identify orthologs of specific kinases across thousands of species within a reasonable time. In contrast to orthologous groups provided by other methods, this approach provides one-to-one ortholog, in-paralog, and co-ortholog relationships, thereby revealing functional relationships and separating even the most closely related paralogous sequences.

While KinOrtho's performance is better than existing methods based on metrics in the benchmarking dataset (Fig. 2a and Additional file 1: Figure S1), overlap in orthologous relationships defined by various methods in the benchmarking dataset is significantly low (only 29.9% similarity; Additional file 1: Figure S8), presumably because of

the variability in orthology definition, methods used, or even potential genome assembly errors in the UniProt reference database. Thus, the interpretation of ortholog sets should be made with some caution. The Alliance of Genome Resources (AGR) has recently established orthologous relationships among humans and six model organisms: *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Mus musculus*, *Rattus norvegicus*, and *Saccharomyces cerevisiae* [89]. The orthologous relationships in AGR are based on the consensus of seven orthology inference methods [38, 39, 41, 42, 81, 90, 91] and five databases [92–96]. Comparison of KinOrtho-defined human kinase orthologs with AGR-defined orthologs reveals nearly 70.5% similarity. The greatest difference in kinase orthology sets occurs in the CMGC and CAMK groups, presumably because of the deeper conservation of these kinases across taxa (Additional file 1: Figure S2).

In our previous study [97], we developed an annotation score (AS) for prioritizing understudied kinases based on existing knowledge stored in curated databases, such as mutations, pathways, expressions, and post-translational modifications (updated AS are available in Additional file 4). In this study, we propose a complementary NIAS for prioritizing understudied kinases based on missing knowledge (mainly their biological functions) inferred by machine learning methods. Because these two scores reflect different aspects of kinase annotation status, they should be used in conjunction when prioritizing understudied kinases. For example, the NIAS score can be informative when prioritizing understudied kinases based on information available from other organisms, while the AS can be helpful when prioritizing kinases based on curation status. Although we attempted to generate an aggregate score (AS') by introducing NIAS into the original AS calculation (Additional file 4), the difference between AS and AS' was not significant. Therefore, we recommend using AS and NIAS independently when prioritizing understudied kinases for experimental studies.

For illuminating understudied kinases, ion channels, G-protein-coupled receptors, or other protein families, a broader collection of manually curated biological functions from various species would be immensely helpful. Although we propose KinOrtho as a tool that can be generalized to a broad range of protein families, its query- and domain-based characteristics may result in lower sensitivity of small protein families with fewer orthologs. Moreover, a fair comparison between KinOrtho and other orthology inference methods cannot be made if the results are evaluated based on a tiny subset of benchmarking datasets. Therefore, to identify the orthologs of those protein families with few members, users are recommended to use KinOrtho in conjunction with other orthology inference methods.

Conclusions

In this study, we have developed an efficient query-based orthology inference method that combines full-length and domain-based orthology inference methods to comprehensively map human kinase orthologs across the tree of life. KinOrtho performed better than existing methods in a benchmarking dataset and identified putative domain fusion and fission events. We confirmed kinase-associated molecular functions enriched across species using phylogenetic profiles after identifying overlapping orthologous relationships from full-length and domain-based pipelines. Finally, we prioritized and inferred functions of understudied human kinases using KinOrtho-defined orthology

and GO annotations as features in machine learning. Our studies serve as a conceptual starting point for investigating understudied human kinase biology by leveraging evolutionary information. This is exemplified, but by no means limited to, pharmacologically tractable protein families such as the protein kinases.

Methods

KinOrtho workflow

KinOrtho is a query-based, graph-based, and combinatorial orthology inference method. It consists of six main steps (Fig. 1):

1 Homology search for the query sequences of interest against reference proteomes
2 Building Basic Local Alignment Search Tool [98] (BLAST) databases, containing full-length and domain-based databases
3 All-vs-all homology search for the rebuilt databases
4 Orthology inference and determining orthologs, paralogs, and co-orthologs
5 Cluster analysis and filtering out the orthologous relationships between two proteins in different clusters or the clusters without query sequences
6 Combining the results of full-length and domain-based methods

Query sequences

The query sequences used in this study were based on a broader mapping of human kinome composition performed recently [19]. We collected 545 human kinases, containing 483 eukaryotic protein kinases (ePKs), 19 eukaryotic-like protein kinases (PKLs), and 43 atypical protein kinases (aPKs). Based on a manually curated eukaryotic protein kinase sequence profile [27], Pfam [99], and Conserved Domain Database [100], we manually annotated and collected 558 kinase domain sequences from the 545 human kinases. More information about the domain name, domain boundary, and kinase group are available in Additional file 5.

Reference proteomes

We applied KinOrtho to the UniProt reference proteomes (release 2019_11), which are chosen to broadly represent the taxonomic diversity [59]. It is also the most well-curated and extensive collection of entire proteomes across the tree of life. The reference proteomes contain 18,870,318 protein sequences spanning the tree of life (Additional file 1: Table S1). To benchmark the performance of KinOrtho, we applied KinOrtho to the Quest for Orthologs (QfO) reference proteomes 2018 [60], which contains 885,338 protein sequences from 48 eukaryotic species, 82,507 sequences from 23 bacterial species, and 17,317 sequences from 7 archaea species (Additional file 1: Table S1). To benchmark KinOrtho based on a domain-based kinase classification, we also applied KinOrtho to the model organisms in KinBase [101], which includes 15 species and 7597 kinase sequences (Additional file 1: Table S3).

Homology search and building BLAST databases

Before performing a time-consuming all-vs-all homology search for all reference proteomes, KinOrtho looks for potential homologs of query sequences by screening the reference proteomes using NCBI BLAST+ [102] (version 2.7.1) with default settings, except for the E-value threshold. Referring to other orthology inference methods, such as OrthoMCL-DB [103] and PANTHER [90], KinOrtho uses 10^{-5} as a default E-value threshold for BLAST search. This threshold has been demonstrated to balance between false-positive and false-negative rates [104]. An additional experiment showed that choosing the default E-value threshold of BLAST+ (10^1) yielded similar sequence comparisons with choosing 10^{-5} in the benchmarking dataset (Additional file 1: Table S4) but not significantly increased the performance based on the six benchmarking metrics shown in Additional file 1: Figure S9. Then, KinOrtho builds two sets of BLAST databases as new reference proteomes ("kinomes" hereafter) based on full-length and domain-based query sequences. To build a full-length kinome for each proteome, KinOrtho keeps the sequences in the BLAST result, generates a new sequence file, and then applies the "makeblastdb" function provided by NCBI BLAST+. To build a domain-based kinome for each proteome, KinOrtho generates a new sequence file and builds a BLAST database based on the BLAST hit region (between "sstart" and "send") of the sequences in the BLAST result. After building a set of full-length kinomes and a set of domain-based kinomes, KinOrtho performs an all-vs-all homology search for each set using the E-value threshold (10^{-5}) mentioned above.

Orthology inference

The orthologous relationships identified by KinOrtho include orthologs, in-paralogs, and co-orthologs. KinOrtho defines a pair of one-to-one orthologs using the Bidirectional Best Hits (BBH) method [62]. A pair of in-paralogs is defined as two protein sequences with a higher similarity score (BLAST bit score) in the same species than the homologous sequences in other species. A pair of co-orthologs is defined based on the following two criteria: (i) ortholog of one sequence is the in-paralog of the other, or (ii) in-paralog of each sequence are a pair of orthologs. Using all orthologous relationships as edges, KinOrtho builds two graphs by connecting the kinases in the full-length and domain-based kinomes, respectively.

Cluster analysis

To identify orthologous groups, KinOrtho performs the Markov Cluster (MCL) Algorithm [44] (version 14.137) for the two graphs. MCL is a fast, unsupervised clustering method using a simulation of flow in graphs. It has been utilized in other graph-based orthology inference methods [35, 105] and detecting protein families [106]. In the orthologous relationship graphs, KinOrtho assigns the negative logarithm of the E-value as a weight for each edge. If an E-value is reported 0 by the BLAST program, KinOrtho assigns an arbitrary E-value of 10-200. Considering the systematic differences among species, such as nucleotide composition bias, KinOrtho normalizes the weights based on the method used by OrthoMCL [35]. For the orthologs or co-orthologs between any two species, KinOrtho normalizes the weights by dividing them by the average weight of

all the orthologs or co-orthologs between the two species. For in-paralogs, the weights are divided by the average weight of all in-paralogs in each kinome. When performing MCL after setting a normalized weight for each edge, KinOrtho chooses 1.5 as a default inflation value to control the cluster tightness. This value is the best inflation value to balance the sensitivity and selectivity for functional classification [35]. Each protein is assigned to a cluster, after which KinOrtho refines orthologous relationships by filtering out the relationships between two proteins in different clusters or the clusters without query sequences.

Combining results

In the last step, KinOrtho combines the orthologous relationships from full-length and domain-based results. We define an "overlapping orthologous relationship" as a relationship present in both full-length and domain-based results. For example, in Scenario 1 of Fig. 3a, if A_1B_1 (meaning the pair of A's 1st kinase domain and B's 1st kinase domain) and A_2B_2 are domain-based orthologs and A–B is a full-length ortholog pair, then both A_1B_1 and A_2B_2 are defined as overlapping orthologous relationships. However, in Scenario 4 of Fig. 3a, if A_1B_1 and A_2C_1 are domain-based orthologs and A-B is a full-length ortholog pair, only A_1B_1 is an overlapping orthologous relationship, but A_2C_1 is not. Because non-overlapping relationships are also informative in domain-based orthology analyses, KinOrtho keeps all the results from full-length and domain-based methods.

Comparison of orthology inference methods

There are 21 public orthology inference results available at Ortholog Benchmarking Webservice [60] (Additional file 1: Table S5; similarity matrices are shown in Additional file 1: Figure S8). These datasets generated by full-length orthology inference methods contain the kinase relationships and all other proteins' relationships in the QfO reference proteomes 2018. To make the orthologs identified by KinOrtho and those identified by the 21 methods comparable, we performed the following preprocessing for the compared datasets. First, because KinOrtho defines orthologs based on the BBH method, we only kept one-to-one relationships in the compared datasets. Second, to identify the kinase orthologs in the compared datasets, we only kept the relationships with at least one protein found in the ortholog relationships identified by KinOrtho (either full-length or domain-based approach). Finally, to identify human kinase orthologs in the compared datasets, we only kept the relationships involving human kinases. The numbers of remaining proteins and ortholog relationships are shown in Additional file 1: Table S5. We submitted these 21 preprocessed one-to-one kinase ortholog datasets to Ortholog Benchmarking Webservice for performance evaluation.

Protein domain annotation

This study employed the annotations in Pfam [99] (version 32.0) as known protein domain annotations. There are 305,472 proteins with at least one orthologous relationship identified by KinOrtho from the UniProt reference proteomes. In these proteins, 197,327 of them have 398,313 domain annotations, and 149,080 have at least one of the two major protein kinase domains: "Pkinase" and "Pkinase_Tyr".

Phylogenetic analysis

The phylogenetic tree analyses in this study were utilized to investigate the gene fusion and fission events of proteins with tandem kinase domains (Fig. 3). First, we obtained the domain-based orthologs of the kinase domains of interest. To identify gene fusion events, we used TAOK1 and PIK3C2A as an example. There are 245 species (including humans) having both TAOK1 and PIK3C2A orthologs. To identify gene fission events, we used OBSCN as an example. There are 80 species with orthologs for each of the (tandem) kinase domains in OBSCN. Second, we aligned those two sets of kinase domain orthologs separately by Multiple Alignment using Fast Fourier Transform [107] (MAFFT, version 7.407). We used options "L-INS-i", "−localpair", and "−maxiterate 10000" to generate more accurate alignments. Third, the two kinase domains' alignments were concatenated as a single alignment file: the first kinase domain's orthologs followed by the second kinase domain's orthologs for each species. Fourth, we used IQ-TREE [108] with options "-m TEST" (standard model), "-bb 1000" (bootstrap replicates), and "-alrt 1000" (approximate likelihood ratio test) to build consensus trees. Finally, phylogenetic trees were visualized using Interactive Tree Of Life [109] (iTOL, version 4).

Cluster and enrichment analyses on phylogenetic profile

The phylogenetic profile of human kinases in this study was built upon 558 human kinase domains and their orthologs identified by KinOrtho's both full-length and domain-based approaches across the 17,134 species in the UniProt reference proteomes (Fig. 4). We manually grouped these species into 561 clades based on the NCBI Taxonomy database [110]. Each clade contains at least five species; each clade in eukaryotes, bacteria, archaea, or viruses contains at most 41, 240, 35, or 2287 species, respectively. Then we calculated an ortholog coverage for each kinase-clade pair by dividing the number of orthologs by the total number of species in each clade. Based on this phylogenetic profile (a kinase-clade matrix), in addition to ordering the human kinase domains by their groups defined by KinBase (Fig. 4a), we clustered them using k-means clustering [111] (Fig. 4b). We used an R package "factoextra" [112] (version 1.0.7) with options "kmeans" (clustering function), "nstart = 50" (initial random centroids), "nboot = 500" (number of bootstrap samples), and "gap_stat" (compute gap statistic [113]) to determine the optimal number of clusters. We found that the optimal number of clusters was 10 (Additional file 1: Figure S10).

We performed Gene Ontology (GO) enrichment analyses using the GO annotations of all human kinases and their orthologs in each cluster. We extracted all three GO domains (biological process, cellular component, and molecular function) annotations from UniProt [114] (release 2019_11). We then expanded the lowest-level GO terms to all-levels GO terms for every kinase based on the hierarchical controlled vocabulary defined by the GO Consortium [115, 116]. Because the GO terms are annotated at the protein level instead of the domain level, we removed duplicate annotations if a protein's tandem kinase domains are in the same cluster. When performing enrichment analyses, we chose all kinase orthologs' annotations as background, used Fisher's exact test, and then controlled the FDR by the Benjamini-Hochberg procedure [117]. A significantly enriched GO term is defined based on whether the FDR < 0.05, and at least five human kinases in the cluster have an

annotation. If multiple GO terms in the same lineage are enriched in a cluster, we only keep the lowest-level term.

Novel Inferred Annotation Score

The Novel Inferred Annotation Score (NIAS) proposed in this study is used to estimate the number of potential annotations we can infer from orthologous relationships to annotate understudied human kinases, which are defined by the NIH Illuminating the Druggable Genome program (IDG) [13] (Additional file 5, last updated on June 11, 2019). The scoring system was built upon machine learning-based annotation inference models using overlapping orthologous relationships and GO annotations (Fig. 5).

First, to prevent prediction models from being biased by unreviewed data, we only used the manually reviewed (non-electronic) annotations of all well-studied kinases and their orthologs to build training sets for the three GO domains. The GO terms annotated for less than 100 kinases were excluded from the training sets. Each instance of a kinase-GO term pair in the training sets showed the values of output and input features based on the GO annotation status (1 for present and 0 for absent) of a well-studied human kinase and its orthologs, respectively. For example, in Fig. 5, an instance shows that mouse's and rat's Egfr genes have a GO term "response to stimulus" (GO:0050896) annotation, but human EGFR has not. The training sets consisted of 0.3 million instances with 730 GO terms for 393 well-studied kinases and their orthologs across 176 species. To prioritize the annotations from different species, we further introduced a sequence similarity for each ortholog. Sequence similarities are defined by the average of the normalized weights generated when we built orthologous relationship graphs.

After building the three training sets, we built annotation inference models for each training set using logistic regression, support vector machine (SVM), and random forest implemented by WEKA [118]. All models were trained with class balancing (using instance reweighting) and 10-fold cross-validation to prevent overfitting. The three machine learning methods' prediction performances for each training set are shown in Table 2. Because random forest showed the best performance among the three training sets, we used the annotation inference models built by random forest to predict missing GO annotations for understudied human kinases. When building test sets, we used the GO terms annotated for at least one understudied kinase ortholog. There were 11,503 instances in the biological process test set, 1810 instances in the cellular component test set, and 2507 instances in the molecular function test set. After applying the random forest models, the confusion matrices built upon the annotation inference result and existing annotations are shown in Additional file 1: Table S2. We collected those GO annotations currently absent but predicted as present for each understudied kinase and then only kept the lowest-level term in each GO term lineage. The summation of each prediction score calculated by random forest is defined as an Inferred Annotation Score (IAS):

$$S(k_g) = \begin{cases} P(k_g) & \text{if } P(k_g) > 0.5 \text{ and} \\ & \quad \text{for all } P(k_{g'}) <= 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$IAS(k) = \sum_{g=1}^{G} S(k_g),$$

where $P(k_g)$ is the prediction score of gth GO term annotation for understudied kinase k, $g = \{1, 2, \ldots, G\}$, and g' represents any descendant of gth GO term. Because the predicted annotations may include existing unreviewed electronic annotations, we defined the NIAS of an understudied kinase by subtracting the unreviewed annotation score (UAS) from IAS:

$$U(k_g) = \begin{cases} 1 & \text{if } k_g \text{ is an unreviewed annotation} \\ 0 & \text{otherwise} \end{cases}$$

$$UAS(k) = \sum_{g=1}^{G} S(k_g)U(k_g)$$

$$NIAS(k) = IAS(k) - UAS(k)$$

Abbreviations
ANOVA: Analysis of variance; AS: Annotation score; BLAST: Basic Local Alignment Search Tool; FDR: False discovery rate; GO: Gene Ontology; HMM: Hidden Markov models; IAS: Inferred Annotation Score; MCL: Markov Cluster; NIAS: Novel Inferred Annotation Score; QfO: Quest for Orthologs; SVM: Support vector machine; UAS: Unreviewed Annotation Score.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04358-3.

Additional file 1. Supplementary results, figures, and tables.

Additional file 2. Fusion event.

Additional file 3. GO enrichment.

Additional file 4. Novel inferred annotation.

Additional file 5. Human protein kinase list.

Availability of data and materials
All source codes are available at https://github.com/esbgkannan/KinOrtho.

Declarations

Ethics approval and consent to participate
Not applicable.

Huang *et al. BMC Bioinformatics* (2021) 22:446

Page 21 of 25

Author details
[1]Institute of Bioinformatics, University of Georgia, 120 Green St., Athens, GA 30602, USA. [2]PREP@UGA, University of Georgia, 500 D.W. Brooks Drive, Athens, GA 30602, USA. [3]Department of Biochemistry and Molecular Biology, University of Georgia, 120 Green St., Athens, GA 30602, USA. [4]Department of Biochemistry and Systems Biology, University of Liverpool, Crown St, Liverpool, UK.

References

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991;252(5013):1651–6.
2. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, Goldstein MM, Grigoriev IV, Hackett KJ, Haussler D, Jarvis ED, Johnson WE, Patrinos A, Richards S, Castilla-Rubio JC, van Sluys MA, Soltis PS, Xu X, Yang H, Zhang G. Earth BioGenome Project: sequencing life for the future of life. Proc Natl Acad Sci USA. 2018;115(17):4325–33.
3. Weiss KM. Genetic variation and human disease: principles and evolutionary approaches, vol. 11. Cambridge: Cambridge University Press; 1993.
4. Hardman G, Perkins S, Brownridge PJ, Clarke CJ, Byrne DP, Campbell AE, Kalyuzhnyy A, Myall A, Eyers PA, Jones AR, Eyers CE. Strong anion exchange-mediated phosphoproteomics reveals extensive human non-canonical phosphorylation. EMBO J. 2019;38(21):100847.
5. Shrestha S, Katiyar S, Sanz-Rodriguez CE, Kemppinen NR, Kim HW, Kadirvelraj R, Panagos C, Keyhaninejad N, Colonna M, Chopra P, Byrne DP, Boons GJ, van der Knaap E, Eyers PA, Edison AS, Wood ZA, Kannan N. A redox-active switch in fructosamine-3-kinases expands the regulatory repertoire of the protein kinase superfamily. Sci Signal. 2020;13(639):eaax6313.
6. Hunter T. The proteins of oncogenes. Sci Am. 1984;251(2):70–9.
7. Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov. 2002;1(9):727–30.
8. Arslan MA, Kutuk O, Basaga H. Protein kinases as drug targets in cancer. Curr Cancer Drug Targets. 2006;6(7):623–34.
9. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 2002;298(5600):1912–34.
10. Wilson LJ, Linley A, Hammond DE, Hood FE, Coulson JM, MacEwan DJ, Ross SJ, Slupsky JR, Smith PD, Eyers PA, Prior IA. New perspectives, opportunities, and challenges in exploring the human protein kinome. Cancer Res. 2018;78(1):15–29.
11. Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH. Too many roads not taken. Nature. 2011;470(7333):163–5.
12. Knapp S, Arruda P, Blagg J, Burley S, Drewry DH, Edwards A, Fabbro D, Gillespie P, Gray NS, Kuster B, Lackey KE, Mazzafera P, Tomkinson NC, Willson TM, Workman P, Zuercher WJ. A public-private partnership to unlock the untargeted kinome. Nat Chem Biol. 2013;9(1):3–6.
13. Illuminating the Druggable Genome. Understudied proteins. https://commonfund.nih.gov/idg/understudiedpro teins. Accessed 11 June 2019 (2019)
14. Nguyen DT, Mathias S, Bologa C, Brunak S, Fernandez N, Gaulton A, Hersey A, Holmes J, Jensen LJ, Karlsson A, Liu G, Ma'ayan A, Mandava G, Mani S, Mehta S, Overington J, Patel J, Rouillard AD, Schürer S, Sheils T, Simeonov A, Sklar LA, Southall N, Ursu O, Vidovic D, Waller A, Yang J, Jadhav A, Oprea TI, Guha R. Pharos: collating protein information to shed light on the druggable genome. Nucleic Acids Res. 2017;45(D1):995–1002.
15. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013;41(Web Server issue):518–22.
16. Vanrobays E, Gelugne JP, Gleizes PE, Caizergues-Ferrer M. Late cytoplasmic maturation of the small ribosomal subunit requires RIO proteins in Saccharomyces cerevisiae. Mol Cell Biol. 2003;23(6):2083–95.
17. Fry AM, O'Regan L, Sabir SR, Bayliss R. Cell cycle regulation by the NEK family of protein kinases. J Cell Sci. 2012;125(Pt 19):4423–33.
18. Luo H, Lin Y, Liu T, Lai FL, Zhang CT, Gao F, Zhang R. DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. Nucleic Acids Res. 2021;49(D1):677–86.
19. Moret N, Liu C, Gyori BM, Bachman JA, Steppi A, Taujale R, Huang L-C, Hug C, Berginski M, Gomez S, Kannan N, Sorger PK. Exploring the understudied human kinome for research and therapeutic opportunities. bioRxiv. 2020. https://doi.org/10.1101/2020.04.02.022277.
20. Wei L, Liu Y, Dubchak I, Shon J, Park J. Comparative genomics approaches to study organism similarities and differences. J Biomed Inform. 2002;35(2):142–50.
21. Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool. 1970;19(2):99–113.
22. Fitch WM. Homology a personal view on some of the problems. Trends Genet. 2000;16(5):227–31.
23. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309–38.
24. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997;278(5338):631–7.

Huang *et al. BMC Bioinformatics*   (2021) 22:446

Page 22 of 25

25. O'Brien KP, Remm M, Sonnhammer EL. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 2005;33(Database issue):476–80.

26. Ribeiro AJM, Das S, Dawson N, Zaru R, Orchard S, Thornton JM, Orengo C, Zeqiraj E, Murphy JM, Eyers PA. Emerging concepts in pseudoenzyme classification, evolution, and signaling. Sci Signal. 2019;12(eaat594):9797.

27. Kwon A, Scott S, Taujale R, Yeung W, Kochut KJ, Eyers PA, Kannan N. Tracing the origin and evolution of pseudokinases across the tree of life. Sci Signal. 2019;12(578):eaav3810.

28. Shrestha S, Byrne DP, Harris JA, Kannan N, Eyers PA. Cataloguing the dead: breathing new life into pseudokinase research. FEBS J. 2020;287(19):4150–69.

29. Eyers PA, Keeshan K, Kannan N. Tribbles in the 21st century: the evolving roles of tribbles pseudokinases in biology and disease. Trends Cell Biol. 2017;27(4):284–98.

30. Bailey FP, Byrne DP, McSkimming D, Kannan N, Eyers PA. Going for broke: targeting the human cancer pseudokinome. Biochem J. 2015;465(2):195–211.

31. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2009;19(2):327–35.

32. Kaduk M, Sonnhammer E. Improved orthology inference with Hieranoid 2. Bioinformatics. 2017;33(8):1154–9.

33. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019;47(D1):419–26.

34. Huerta-Cepas J, Szklarczyk D, Heller D, Hernndez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47(D1):309–14.

35. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.

36. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinform. 2011;12:124.

37. Altenhoff AM, Gil M, Gonnet GH, Dessimoz C. Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS ONE. 2013;8(1):53786.

38. Sonnhammer EL, Zstlund G. In Paranoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res. 2015;43(Database issue):234–9.

39. Train CM, Glover NM, Gonnet GH, Altenhoff AM, Dessimoz C. Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. Bioinformatics. 2017;33(14):75–82.

40. Cosentino S, Iwasaki W. SonicParanoid: fast, accurate and easy orthology inference. Bioinformatics. 2019;35(1):149–51.

41. Nevers Y, Kress A, Defosset A, Ripp R, Linard B, Thompson JD, Poch O, Lecompte O. OrthoInspector 3.0: open portal for comparative genomics. Nucleic Acids Res. 2019;47(D1):411–8.

42. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.

43. Derelle R, Philippe H, Colbourne JK. Broccoli: combining phylogenetic and network analyses for orthology assignment. Mol Biol Evol. 2020;37:3389–96.

44. vanDongen S. A cluster algorithm for graphs. Information Systems [INS] (R 0010) (2000)

45. Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. Computational methods for Gene Orthology inference. Brief Bioinform. 2011;12(5):379–91.

46. Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, Bork P. Orthology prediction methods: a quality assessment using curated protein families. BioEssays. 2011;33(10):769–80.

47. Gabaldn T. Large-scale assignment of orthology: back to phylogenetics? Genome Biol. 2008;9(10):235.

48. Hulsen T, Huynen MA, de Vlieg J, Groenen PM. Benchmarking ortholog identification methods using functional genomics data. Genome Biol. 2006;7(4):31.

49. Storm CE, Sonnhammer EL. Comprehensive analysis of orthologous protein domains using the HOPS database. Genome Res. 2003;13(10):2353–62.

50. Krishnamurthy N, Brown D, Sjlander K. FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. BMC Evol Biol. 2007;7 Suppl 1:12.

51. Chen TW, Wu TH, Ng WV, Lin WC. DODO: an efficient orthologous genes assignment tool based on domain architectures. Domain based ortholog detection. BMC Bioinform. 2010;11 Suppl 7:6.

52. Uchiyama I, Mihara M, Nishide H, Chiba H. MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucleic Acids Res. 2015;43(Database issue):270–6.

53. Persson E, Kaduk M, Forslund SK, Sonnhammer ELL. Domainoid: domain-oriented orthology inference. BMC Bioinform. 2019;20(1):523.

54. Sreelatha A, Yee SS, Lopez VA, Park BC, Kinch LN, Pilch S, Servage KA, Zhang J, Jiou J, Karasiewicz-Urbańska M, Łobocka M, Grishin NV, Orth K, Kucharczyk R, Pawłowski K, Tomchick DR, Tagliabracci VS. Protein AMPylation by an evolutionarily conserved pseudokinase. Cell. 2018;175(3):809–21.

55. Krupa A, Abhinandan KR, Srinivasan N. KinG: a database of protein kinases in genomes. Nucleic Acids Res. 2004;32(Database issue):153–5.

56. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science. 1999;285(5428):751–3.

57. McSkimming DI, Dastgheib S, Baffi TR, Byrne DP, Ferries S, Scott ST, Newton AC, Eyers CE, Kochut KJ, Eyers PA, Kannan N. KinView: a visual comparative sequence analysis tool for integrated kinome research. Mol BioSyst. 2016;12(12):3651–65.

58. Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen DT, Bologa CG, Jensen LJ, Vidović D, Koleti A, Schürer SC, Waller A, Yang JJ, Holmes J, Bocci G, Southall N, Dharkar P, Mathé E, Simeonov A, Oprea TI. TCRD and Pharos 2021: mining the human proteome for disease biology. Nucleic Acids Res. 2021;49(D1):1334–46.

Huang *et al. BMC Bioinformatics*   (2021) 22:446

Page 23 of 25

59. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23(10):1282–8.

60. ...Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Pryszcz LP, Schreiber F, da Silva AS, Szklarczyk D, Train CM, Bork P, Lecompte O, von Mering C, Xenarios I, Sjlander K, Jensen LJ, Martin MJ, Muffato M, Gabaldn T, Lewis SE, Thomas PD, Sonnhammer E, Dessimoz C, Altenhoff AM, Boeckmann B, Capella-Gutierrez S, DeLuca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Pryszcz LP, Schreiber F, da Silva AS, Szklarczyk D, Train CM, Lecompte O, Xenarios I, Sjlander K, Martin MJ, Muffato M, Gabaldn T, Lewis SE, Thomas PD, Sonnhammer E, Dessimoz C. Standardized benchmarking in the quest for orthologs. Nat Methods. 2016;13(5):425–30.

61. Wall DP, Fraser HB, Hirsh AE. Detecting putative orthologs. Bioinformatics. 2003;19(13):1710–1.

62. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA. 1999;96(6):2896–901.

63. Yanai I, Derti A, DeLisi C. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc Natl Acad Sci USA. 2001;98(14):7940–5.

64. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, Berghuis B, Turner L, Hudson E, Feenstra K, Sobin L, Robb J, Branton P, Korzeniewski G, Shive C, Tabor D, Qi L, Groch K, Nampally S, Buia S, Zimmerman A, Smith A, Burges R, Robinson K, Valentino K, Bradbury D, Cosentino M, Diaz-Mayoral N, Kennedy M, Engel T, Williams P, Erickson K, Ardlie K, Winckler W, Getz G, DeLuca D, MacArthur D, Kellis M, Thomson A, Young T, Gelfand E, Donovan M, Grant G, Mash D, Marcus Y, Basile M, Liu J, Zhu J, Tu Z, Cox NJ, Nicolae DL, Gamazon ER, Kyung H, Konkashbaev A, Pritchard J, Stevens M, Flutre T, Wen X, Dermitzakis T, Lappalainen T, Guigo R, Monlong J, Sammeth M, Koller D, Battle A, Mostafavi S, McCarthy M, Rivas M, Maller J, Rusyn I, Nobel A, Wright F, Shabalin A, Feolo M, Sharopova N, Sturcke A, Paschal J, Anderson JM, Wilder EL, Derr LK, Green ED, Struewing JP, Temple G, Volpi S, Boyer JT, Thomson EJ, Guyer MS, Ng C, Abdallah A, Colantuoni D, Insel TR, Koester SE, Little AR, Bender PK, Lehner T, Yao Y, Compton CC, Vaught JB, Sawyer S, Lockhart NC, Demchok J, Moore HF. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580–5.

65. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, Meyers RM, Ali L, Goodale A, Lee Y, Jiang G, Hsiao J, Gerath WFJ, Howell S, Merkel E, Ghandi M, Garraway LA, Root DE, Golub TR, Boehm JS, Hahn WC. Defining a cancer dependency map. Cell. 2017;170(3):564–76.

66. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):607–13.

67. Bennetzen MV, Mario G, Pultz D, Morselli E, Frgeman NJ, Kroemer G, Andersen JS. Phosphoproteomic analysis of cells treated with longevity-related autophagy inducers. Cell Cycle. 2012;11(9):1827–40.

68. Kontrogianni-Konstantopoulos A, Bloch RJ. Obscurin: a multitasking muscle giant. J Muscle Res Cell Motil. 2005;26(6–8):419–26.

69. Perry NA, Ackermann MA, Shriver M, Hu LY, Kontrogianni-Konstantopoulos A. Obscurins: unassuming giants enter the spotlight. IUBMB Life. 2013;65(6):479–86.

70. Fraser NH, Metcalfe NB, Thorpe JE. Temperature-dependent switch between diurnal and nocturnal foraging in salmon. Proc R Soc Lond B. 1993;252(1334):135–9.

71. Kadri S, Metcalfe NB, Huntingford FA, Thorpe JE. Daily feeding rhythms in Atlantic salmon I: feeding and aggression in parr under ambient environmental conditions. J Fish Biol. 1997;50(2):267–72. https://doi.org/10.1111/j.1095-8649.1997.tb01357.x.

72. Dickson JG. The wild Turkey: biology and management. Mechanicsburg: Stackpole Books; 1992.

73. Fox JG, Marini RP. Biology and diseases of the ferret. New York: Wiley; 2014.

74. Suga H, Dacre M, de Mendoza A, Shalchian-Tabrizi K, Manning G, Ruiz-Trillo I. Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. Sci Signal. 2012;5(222):35.

75. Hartwell LH, Weinert TA. Checkpoints: controls that ensure the order of cell cycle events. Science. 1989;246(4930):629–34.

76. Abe Y, Matsumoto S, Wei S, Nezu K, Miyoshi A, Kito K, Ueda N, Shigemoto K, Hitsumoto Y, Nikawa J, Enomoto Y. Cloning and characterization of a p53-related protein kinase expressed in interleukin-2-activated cytotoxic T-cells, epithelial tumor cell lines, and the testes. J Biol Chem. 2001;276(47):44003–11.

77. Richardson E, Stockwell SR, Li H, Aherne W, Cuomo ME, Mittnacht S. Mechanism-based screen establishes signalling framework for DNA damage-associated G1 checkpoint response. PLoS ONE. 2012;7(2):31627.

78. Moreno RD, Ramalho-Santos J, Chan EK, Wessel GM, Schatten G. The Golgi apparatus segregates from the lysosomal/acrosomal vesicle during rhesus spermiogenesis: structural alterations. Dev Biol. 2000;219(2):334–49.

79. Sosnik J, Miranda PV, Spiridonov NA, Yoon SY, Fissore RA, Johnson GR, Visconti PE. Tssk6 is required for Izumo relocalization and gamete fusion in the mouse. J Cell Sci. 2009;122(Pt 15):2741–9.

80. Oliver S. Guilt-by-association goes global. Nature. 2000;403(6770):601–3.

81. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Marugn JC, Cummins C, Davidson C, Dodiya K, Fatima R, Gall A, Giron CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Maurel T, McDowall M, McMahon A, Mohanan S, Moore B, Nuhn M, Oheh DN, Parker A, Parton A, Patricio M, Sakthivel MP, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sycheva M, Szuba M, Taylor K, Thormann A, Threadgold G, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Flint B, Frankish A, Hunt SE, Iisley G, Kostadima M, Langridge N, Loveland JE, Martin FJ, Morales J, Mudge JM, Muffato M, Perry E, Ruffier M, Trevanion SJ, Cunningham F, Howe KL, Zerbino DR, Flicek P. Ensembl 2020. Nucleic Acids Res. 2020;48(D1):682–8.

82. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira F, Marchler-Bauer A, Mi H, Natale DA, Necci M, Nuka G,

Huang *et al. BMC Bioinformatics*   (2021) 22:446

Page 24 of 25

Orengo C, Pandurangan AP, Paysan-Lafosse T, Pesseat S, Potter SC, Qureshi MA, Rawlings ND, Redaschi N, Richardson LJ, Rivoire C, Salazar GA, Sangrador-Vegas A, Sigrist CJA, Sillitoe I, Sutton GG, Thanki N, Thomas PD, Tosatto SCE, Yong SY, Finn RD. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 2019;47(D1):351–60.

83. Lang B, Pu J, Hunter I, Liu M, Martin-Granados C, Reilly TJ, Gao GD, Guan ZL, Li WD, Shi YY, He G, He L, Stefnsson H, St Clair D, Blackwood DH, McCaig CD, Shen S. Recurrent deletions of ULK4 in schizophrenia: a gene crucial for neuritogenesis and neuronal motility. J Cell Sci. 2014;127(Pt 3):630–40.

84. Lang B, Zhang L, Jiang G, Hu L, Lan W, Zhao L, Hunter I, Pruski M, Song NN, Huang Y, Zhang L, St Clair D, McCaig CD, Ding YQ. Control of cortex development by ULK4, a rare risk gene for mental disorders including schizophrenia. Sci Rep. 2016;6:31126.

85. Khamrui S, Ung PMU, Secor C, Schlessinger A, Lazarus MB. High-resolution structure and inhibition of the schizophrenia-linked pseudokinase ULK4. J Am Chem Soc. 2020;142(1):33–7.

86. Murphy JM, Zhang Q, Young SN, Reese ML, Bailey FP, Eyers PA, Ungureanu D, Hammaren H, Silvennoinen O, Varghese LN, Chen K, Tripaydonis A, Jura N, Fukuda K, Qin J, Nimchuk Z, Mudgett MB, Elowe S, Gee CL, Liu L, Daly RJ, Manning G, Babon JJ, Lucet IS. A robust methodology to subclassify pseudokinases based on their nucleotide-binding properties. Biochem J. 2014;457(2):323–34.

87. Eyers PA. Marveling at the Incredible ULK4. Structure. 2020;28(11):1181–3.

88. Pandey A, Dan I, Kristiansen TZ, Watanabe NM, Voldby J, Kajikawa E, Khosravi-Far R, Blagoev B, Mann M. Cloning and characterization of PAK5, a novel member of mammalian p21-activated kinase-II subfamily that is predominantly expressed in brain. Oncogene. 2002;21(24):3939–48.

89. Agapite J, Albou LP, Aleksander S, Argasinska J, Arnaboldi V, Attrill H, Bello SM, Blake JA, Blodgett O, Bradford YM, Bult CJ, Cain S, Calvi BR, Carbon S, Chan J, Chen WJ, Cherry JM, Cho J, Christie KR, Crosby MA, Pons J, Dolan ME, Santos GD, Dunn B, Dunn N, Eagle A, Ebert D, Engel SR, Fashena D, Frazer K, Gao S, Gondwe F, Goodman J, Gramates LS, Grove CA, Harris T, Harrison MC, Howe DG, Howe KL, Jha S, Kadin JA, Kaufman TC, Kalita P, Karra K, Kishore R, Laulederkind S, Lee R, MacPherson KA, Marygold SJ, Matthews B, Millburn G, Miyasato S, Moxon S, Mueller HM, Mungall C, Muruganujan A, Mushayahama T, Nash RS, Ng P, Paulini M, Perrimon N, Pich C, Raciti D, Richardson JE, Russell M, Gelbart SR, Ruzicka L, Schaper K, Shimoyama M, Simison M, Smith C, Shaw DR, Shrivatsav A, Skrzypek M, Smith JR, Sternberg PW, Tabone CJ, Thomas PD, Thota J, Toro S, Tomczuk M, Tutaj M, Tutaj M, Urbano JM, Auken KV, Slyke CEV, Wang SJ, Weng S, Westerfield M, Williams G, Wong ED, Wright A, Yook K. Alliance of Genome Resources Portal: unified model organism research platform. Nucleic Acids Res. 2020;48(D1):650–8.

90. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129–41.

91. Kaduk M, Riegler C, Lemp O, Sonnhammer EL. HieranoiDB: a database of orthologs inferred by Hieranoid. Nucleic Acids Res. 2017;45(D1):687–90.

92. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO Gene Nomenclature Committee (HGNC). Hum Genet. 2001;109(6):678–80.

93. Li H, Coghlan A, Ruan J, Coin LJ, Hériché JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK, Zheng W, Dehal P, Wang J, Durbin R. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. 2006;34(Database issue):572–80.

94. DeLuca TF, Cui J, Jung JY, St Gabriel KC, Wall DP. Roundup 2.0: enabling comparative genomics for over 1800 genomes. Bioinformatics. 2012;28(5):715–6.

95. Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. 2014;42(Database issue):897–902.

96. Ruzicka L, Howe DG, Ramachandran S, Toro S, Van Slyke CE, Bradford YM, Eagle A, Fashena D, Frazer K, Kalita P, Mani P, Martin R, Moxon ST, Paddock H, Pich C, Schaper K, Shao X, Singer A, Westerfield M. The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources. Nucleic Acids Res. 2019;47(D1):867–73.

97. Huang LC, Ross KE, Baffi TR, Drabkin H, Kochut KJ, Ruan Z, D'Eustachio P, McSkimming D, Arighi C, Chen C, Natale DA, Smith C, Gaudet P, Newton AC, Wu C, Kannan N. Integrative annotation and knowledge discovery of kinase post-translational modifications and cancer-associated mutations through federated protein ontologies and resources. Sci Rep. 2018;8(1):6518.

98. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

99. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):427–32.

100. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020;48(D1):265–8.

101. Manning G. KinBase: genomics, evolution and function of protein kinases (2002)

102. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinform. 2009;10:421.

103. Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics Chapter. 2011;6:1–19.

104. Chen F, Mackey AJ, Vermunt JK, Roos DS. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE. 2007;2(4):383.

105. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.

106. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30(7):1575–84.

Huang *et al. BMC Bioinformatics*   (2021) 22:446

Page 25 of 25

107. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.

108. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.

109. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):256–9.

110. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012;40(Database issue):136–43.

111. MacQueen J, et al. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. Oakland, CA, USA (1967)

112. Kassambara A, Mundt F. Package 'factoextra'. Extract and visualize the results of multivariate data analyses. 2017;76.

113. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B (Stat Methodol). 2001;63(2):411–23.

114. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):506–15.

115. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.

116. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47(D1):330–8.

117. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol). 1995;57(1):289–300.

118. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The weka data mining software: an update. ACM SIGKDD Explor Newsl. 2009;11(1):10–8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.