ORFanID: A Web-Based Search Engine for the Discovery and Identification of Orphan and Taxonomically Restricted Genes

Richard S. Gunasekera[1*], Komal K. B. Raja[2], Suresh Hewapathirana[3], Emanuel Tundrea[4], Vinodh Gunasekera[5], Thushara Galbadage[6] and Paul A. Nelson[7]

[1] Department of Chemistry, Physics and Engineering, School of Science, Technology & Health, Biola University, La Mirada, CA

[2] Department of Pathology & Immunology, Baylor College of Medicine, Houston, TX

[3] European Bioinformatics Institute, Welcome Genome Campus, Hinxton, Cambridgeshire, UK

[4] Griffiths School of Management and IT, Emanuel University of Oradea, Romania

[5] Bioinformatics, Chesalon USA, Inc., Houston TX

[6] Department of Kinesiology and Public Health, School of Science, Technology & Health, Biola University, La Mirada, CA

[7] Biola University, La Mirada, CA

*** Correspondence:**

Richard S. Gunasekera

richard.gunasekera@biola.edu

Keywords: ORFanID, orphan gene, ortholog, Taxonomically Restricted Genes (TRG), DNA, protein, bioinformatics

Abstract

With the numerous genomes sequenced today, it has been revealed that a noteworthy percentage of genes in a given taxon of organisms in the phylogenetic tree of life do not have orthologous sequences in other taxa. These sequences are commonly referred to as "orphans" or "ORFans" if found as single occurrences in a single species or as "taxonomically restricted genes" (TRGs) when found at higher taxonomic levels. Quantitative and collective studies of these genes are necessary for understanding their biological origins. However, the current software for identifying orphan genes is limited in its functionality, database search range, and very complex algorithmically. Thus, researchers studying orphan genes must harvest their data from many disparate sources. ORFanID is a graphical web-based search engine that facilitates the efficient identification of both orphan genes and TRGs at all taxonomic levels, from DNA or amino acid sequences in the NCBI database cluster and other large bioinformatics repositories. The software allows users to identify genes that are unique to any taxonomic rank, from species to domain, using NCBI systematic classifiers. It provides control over NCBI database search parameters, and the results are presented in a spreadsheet as well as a graphical display. The tables in the software are sortable, and results can be filtered using the fuzzy search functionality. The visual presentation can be expanded and collapsed by the taxonomic tree to its various branches. Example results from searches on five species and gene expression data from specific orphan genes are provided in the Supplementary Information.

Availability and Implementation: The ORFanID web application and database can be accessed at http://www.orfangenes.com/. The website provides textual and video instructions on how to use the tool. The Source Code for ORFanID is available at https://github.com/orfanid.

1. Introduction

Following the introduction of large-scale, high-throughput automated DNA sequencing in the mid-1990s, the comparative analysis of whole genomes revealed a large number of protein-coding open reading frames (ORFs) that occur only in one species or as taxonomically restricted genes (TRGs). These TRGs occur at systematic ranks from the genus upwards. ORFs that are specific to one species are designated as "orphan" (sometimes spelled "ORFan") genes, while more widely distributed ORFs that are not present universally but only at lower taxonomic ranks, are designated as TRGs [1]. By definition, orphan genes have no orthologous sequences in any other species, and TRGs are not present in any genomes outside their respective taxa. For example, a TRG may be found only in the genus *Drosophila*, but not in any other Dipteran or any larger systematic category such as Insecta, Arthropoda, and so on.

Early molecular genetic studies focused on conserved genes, either common to all organisms (e.g., ribosomal genes) or broader systematic categories (e.g., the Wnt signaling pathway in Metazoa), as it was difficult to study TRGs or orphans when relying on the amplification and sequencing of sub-kilobase regions. The neglect of orphans has continued to some extent in the genomic era, as functional roles are mainly assigned to newly sequenced genes mainly via homology criteria (e.g., existing annotations in other species). However, it has been shown in some cases that orphan genes are uniquely involved in making one species distinct from another phenotypically [2-4]. The last decade has seen a rapidly increasing interest in studying orphan genes [5-7]. Genomic sequencing has revealed that large fractions of genes in a given species' complete genome do not possess orthologous sequences in other species, so TRGs represent important mediators of phenotypic novelty [8-10].

The previous theory held that all genes descended by coding sequences present in the Last Universal Common Ancestor (LUCA) and that the duplication of existing genes, followed by a divergence into new functions, brought forth all extant genes [11-13]. This duplication-and-divergence process model predicted that any extant ORF should reveal its history compared to the known universe of other ORFs, given that all ORFs stand on branching lineages descended from their predecessors in LUCA.

However, the discovery of orphan genes has led to a changing picture of gene evolution and formation suggesting that de novo formation may be a predominant mechanism for gene emergence [10]. This shift in thinking has significant ramifications for understanding genes, the genome's non-coding regions, and fully functional sequences and calls for the re-analysis of existing data and comparative genomic analysis to provide new and accurate understandings [14, 15]. The biological understanding of the origins and the functions of orphan genes will have applications in medicine and evolutionary biology across the tree of life. Therefore, orphan genes (and TRGs) represent an intriguing aspect of biology, lying at the intersection of genomics, genetics, comparative and structural biology, phylogenetics, and evolution.

To study growing numbers of these novel genes across genomes, easy-to-use bioinformatics tools are needed that can be utilized by scientists from the life sciences and those with computational backgrounds. Tools such as BLAST and related software have been made available [1, 16, 17]. However, further tools are necessary for discovering orphan genes and assigning TRGs to their taxonomic levels (sometimes known as phylostratigraphy) [18].

Currently, the tool options researchers can find today for studying orphan genes are limited, as most software solutions focus on identifying orthologs or inferencing ortho groups and are generally limited to proteins. For instance, ORFanFinder functionality is limited to plants, bacteria,

and fungi, and the URL in the original publication (DOI: 10.1093/bioinformatics/btw122) is not active [17]. However, a web search has revealed that ORFanFinder is available at http://bcb.unl.edu/orfanfinder/ although this software has not been updated since 2016. SequenceServer performs BLAST without classifying the proteins/DNA sequences into taxonomic levels [19]. The software Geneious can perform alignment and build a phylogenetic tree but identifying orphans using this software is challenging [20]. Similarly, OrthoFinder provides the option to use DIAMOND or its recommended MMseq2 for sequence alignment [21-24]. OMA orthology or the series of analytical resources developed by the Bioinformatics Resource Centers (BRCs) for Infectious Diseases program [bioinformatics tools, workspaces, and services for bioinformatics data analysis like AmoebaDB, FungiDB, OrthoMCL] only show orthologous genes/proteins and do not identify orphan genes [25, 26]. A newer gene classification platform, www.shoot.bio, may also be helpful to align and compare gene origins but it only uses protein (amino acid) sequences. In short, these tools perform alignment or identify conserved genes from genomes, but they are not suited to identify orphans as ORFanID is designed to do.

ORFanID's distinctiveness is in three aspects: (1) It processes not just protein/amino acid sequences but also DNA/nucleotide sequences. (2) With its built-in homology interpreter and classifier, this search engine provides the taxonomic rank of a gene either as an orphan gene or as a gene restricted to a taxonomic level in the tree of life; (3) As ORFans and TRGs are identified, ORFanID builds its own database with the results of the analysis and provides the researcher with the possibility to further explore the data.

2. Methods

2.1 Algorithm and Implementation

ORFanID identifies orphan genes and TRGs from a given list of DNA or protein sequences mainly using NCBI accession numbers (Figure 1). It detects homologous sequences in the NCBI non-redundant databases by using the BLAST alignment tool. All the previously mentioned tools acknowledge that NCBI databases are among the most trusted sources. BLAST is a well-established tool, but its slowness in processing is acknowledged as the databases continue to grow rapidly. Since the ORFanID search engine utilizes local BLASting and a 24-core 48-thread server which significantly improves the execution speed of BLAST as compared with the web-based BLAST searches.

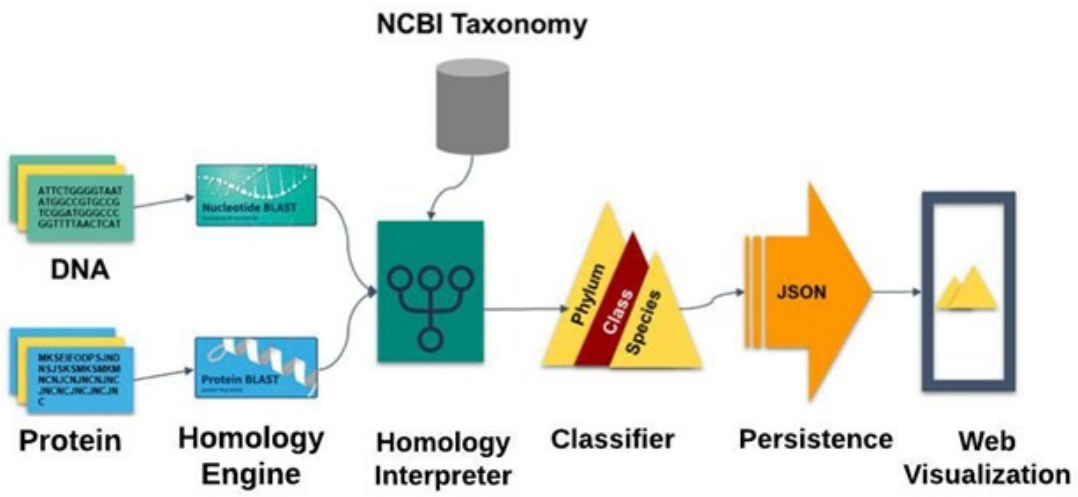


Figure 1: A graphical representation of the core engine of ORFanID. DNA and protein sequences are fed through a homology interpreter and classified. ORFanID is a web-based application developed on the Java Spring Framework.

If the query sequence is a protein, the BLASTP algorithm is used, while the BLASTN algorithm is used for the nucleotide sequences. In addition, ORFanID also supports PSI-BLAST which uses position-specific scoring matrices to detect distant evolutionary relationships between

sequences. ORFanID allows the user to submit multiple sequences in FASTA format, which are then processed using software message brokering techniques. The results are combined into a single blast report in a tab-delimited format that can be downloaded by the user. ORFanID sends a customized BLAST command to retrieve the taxonomy IDs of each hit. Using the rank lineage information file provided by the NCBI Taxonomy Database, the ORFanID Homology Interpreter determines the defined taxonomy level (species, genus, family, order, class, phylum, kingdom, and superkingdom) for each sequence found in the BLAST results. (Figure 2).

A

Homologous Species Taxonomy IDs
562, 1736699, 61648, 1972431

Input Query Gene	Homology Genes found	p.Identity	length	mismatch	gaps	Q.Start	Q.End	S.Start	S.End	E-Value	Bitscore	StaxId
YP_002791247.1	gi|1116622776|ref|WP_072146903.1|	94.737	19	1	0	1	19	1	19	0.12	34.7	562
YP_002791247.1	gi|1337959342|ref|WP_103433047.1|	94.737	19	1	0	1	19	1	19	0.12	34.7	562
YP_002791247.1	gi|1367384489|ref|WP_106423590.1|	94.737	19	1	0	1	19	1	19	0.15	34.7	562
YP_002791247.1	gi|1362691806|ref|WP_106379540.1|	100	18	0	0	1	18	1	18	0.19	34.3	562
YP_002791247.1	gi|1119742609|ref|WP_072652649.1|	100	18	0	0	1	18	1	18	0.22	34.3	562
YP_002791247.1	gi|983401378|ref|WP_060569283.1|	84.211	19	3	0	1	19	20	38	0.73	33.5	1736699
YP_002791247.1	gi|1183803052|ref|WP_085005794.1|	78.947	19	4	0	1	19	47	65	1.3	32.7	61648
YP_002791247.1	gi|1105446713|ref|WP_071845974.1|	78.947	19	4	0	1	19	47	65	1.8	32.3	1972431
NP_414542.1	gi|693159127|ref|WP_032294921.1|	100	20	0	0	1	20	1	20	0.036	36.2	562
NP_414542.1	gi|485761283|ref|WP_001386572.1|	100	21	0	0	1	21	1	21	0.006	38.1	562

B

Taxonomy Id	Scientific name	Species name	genus name	family name	order name	class name	Phylum name	kingdom name	Superkingdom name
562	Escherichia coli	N/A	Escherichia	Enterobacteriaceae	Enterobacterales	Gammaproteobacteria	Proteobacteria	N/A	Bacteria
562	Escherichia coli	N/A	Escherichia	Enterobacteriaceae	Enterobacterales	Gammaproteobacteria	Proteobacteria	N/A	Bacteria
562	Escherichia coli	N/A	Escherichia	Enterobacteriaceae	Enterobacterales	Gammaproteobacteria	Proteobacteria	N/A	Bacteria
562	Escherichia coli	N/A	Escherichia	Enterobacteriaceae	Enterobacterales	Gammaproteobacteria	Proteobacteria	N/A	Bacteria
562	Escherichia coli	N/A	Escherichia	Enterobacteriaceae	Enterobacterales	Gammaproteobacteria	Proteobacteria	N/A	Bacteria
1736699	Citrobacter sp. 50677481	N/A	Citrobacter	Enterobacteriaceae	Enterobacterales	Gammaproteobacteria	Proteobacteria	N/A	Bacteria
61648	Kluyvera intermedia	N/A	Kluyvera	Enterobacteriaceae	Enterobacterales	Gammaproteobacteria	Proteobacteria	N/A	Bacteria
1972431	Phytobacter ursingii	N/A	Phytobacter	Enterobacteriaceae	Enterobacterales	Gammaproteobacteria	Proteobacteria	N/A	Bacteria
562	Escherichia coli	N/A	Escherichia	Enterobacteriaceae	Enterobacterales	Gammaproteobacteria	Proteobacteria	N/A	Bacteria
562	Escherichia coli	N/A	Escherichia	Enterobacteriaceae	Enterobacterales	Gammaproteobacteria	Proteobacteria	N/A	Bacteria

Figure 2: Homologies and taxid found by ORFanID for the input sequence. (A) Displays the homologies and taxid found by ORFanID for the input sequence. In this example, the homologies and taxids of two sequences, YP_002791247.1(Q1) and NP_414542.1(Q2), [i.e., genes from Escherichia Coli species (NCBI Taxonomy ID: 562)]. Eight homologous hits were found from four species consisting of the taxonomy ids: 562, 1736699, 61648, and 1972431. (B) Shows the traversal of the algorithm based on the taxid of the sequence homologies that are found by ORFanID. In this example, eight hits that are homologous to the YP_002791247.1 gene and two hits that are homologous to NP_414542.1 are shown. Then the rank lineages are extracted from the taxonomy lineage file in NCBI for each hit.

After finding the complete lineage (Linnaeus taxonomic category) for each individual homolog of the input DNA or protein sequence, the ORFanID Classifier starts searching from the Superkingdom level and works its way toward the Species level to find additional homologs (Figure 3). The taxonomic rank where the homologs are located designates them as a "Taxonomically Restricted Gene" (TRG) cluster of orthologs (homologs) by the use of the classifier. ORFanID uses the single ortholog de-classifier rule, meaning that it is sufficient for finding only one ortholog at a certain level in the tree of life, to classify the gene of interest as a TRG at that level or as an orphan at the final level. ORFanID does not require a significant number of orthologs (as the SMOTE technique would) to determine that a sequence is not an orphan [27]. If no orthologous sequences are found, and the ORF remains a species-unique sequence, it will be classified as an "Orphan Gene." If the ORF is unique at the subspecies level, the gene is classified as a "Strict Orphan". Using this algorithm, ORFanID identifies and displays both orphan genes and TRGs. The results

can be viewed and analyzed graphically. Further details on the ORFanID algorithm are shown in Figure 4.

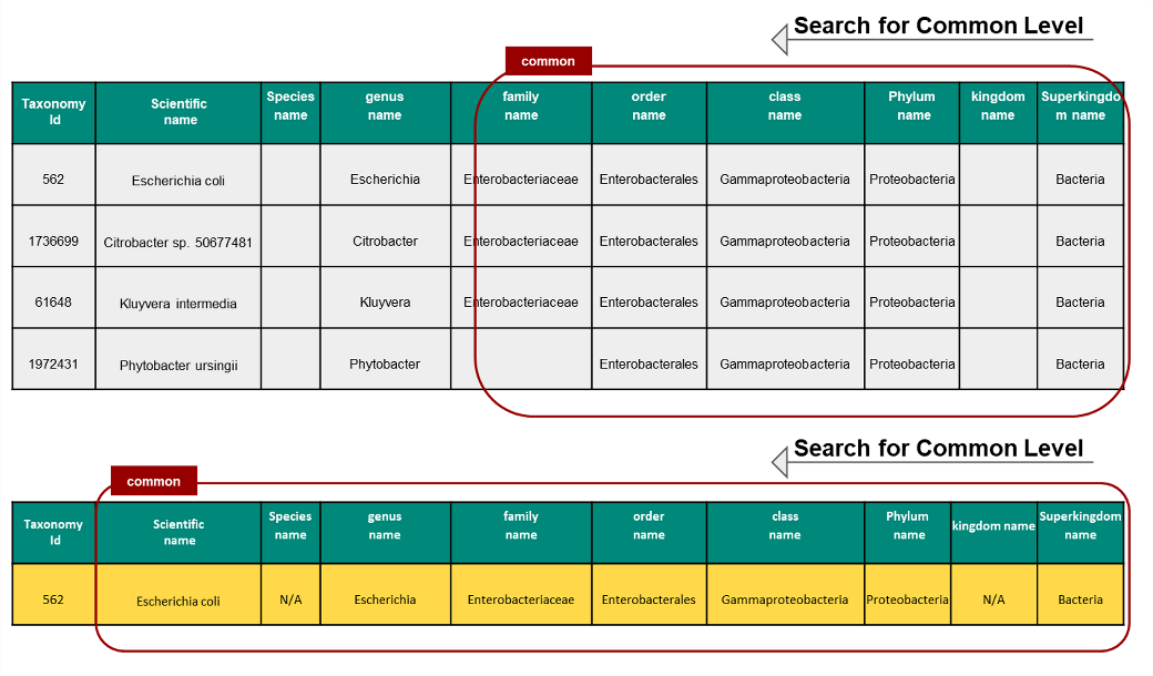


Figure 3: Gene Classification on ORFanID. The search starts from the superkingdom and steps through each subsequent taxonomy to the species level until it finds homologous sequences. If no homologous sequences are found, the sequence is classified as an ORFan gene.

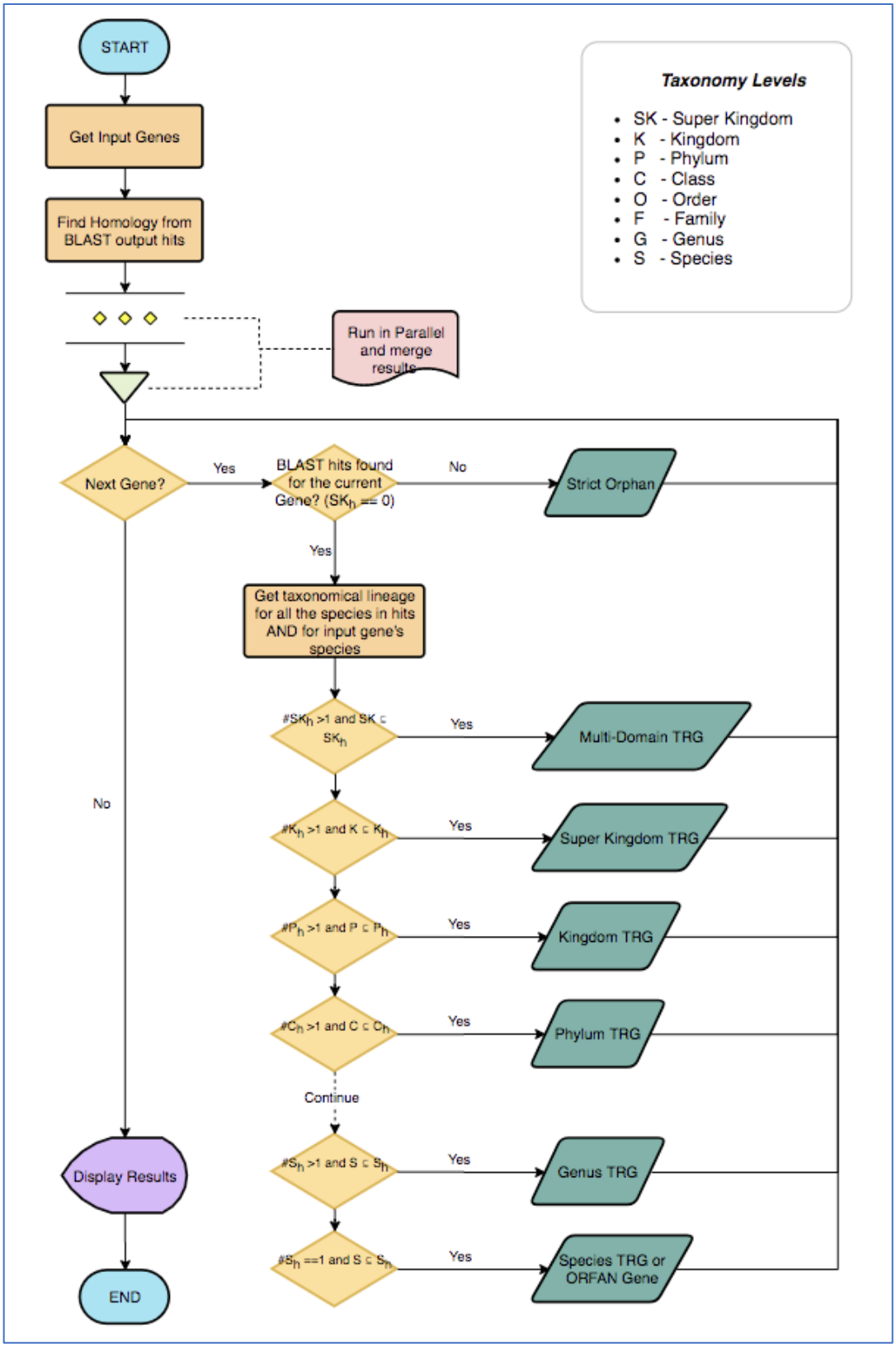

START
Get Input Genes
Find Homology from BLAST output hits
Run in Parallel and merge results
Taxonomy Levels
SK - Super Kingdom
K - Kingdom
P - Phylum
C - Class
O - Order
F - Family
G - Genus
S - Species
Next Gene?
Yes
BLAST hits found for the current Gene? (SKh == 0)
No
Strict Orphan
Yes
Get taxonomical lineage for all the species in hits AND for input gene's species
#SKh >1 and SK ⊂ SKh
Yes
Multi-Domain TRG
No
#Kh >1 and K ⊂ Kh
Yes
Super Kingdom TRG
#Ph >1 and P ⊂ Ph
Yes
Kingdom TRG
#Ch >1 and C ⊂ Ch
Yes
Phylum TRG
Continue
#Sh >1 and S ⊂ Sh
Yes
Genus TRG
#Sh ==1 and S ⊂ Sh
Yes
Species TRG or ORFAN Gene
Display Results
END

Figure 4: An example of the algorithm of the ORFanID engine is explained in a UML Flow Chart diagram. To find the homolog, our algorithm begins its scan at the Super Kingdom rank. Let SK_h be the set of Super Kingdom taxonomies of the subject genome. If $\#SK_h > 1$, there are no common homologs to be found. This gene (G) will be categorized as a "*Multi-domain*" Taxonomically Restricted Gene (TRG). However, if all of the subject genomes belong to the same Super Kingdom as the input genome ($\#SK_h = 1$), the algorithm moves to the next level which is the Kingdom (K) rank. Let K_h be the set of Kingdom taxonomies of the homologous genomes. If $\#K_h > 1$ the subject sequences belong to multiple Kingdoms. This means that while the Kingdoms are not common, their Super Kingdom is common to the subject genomes. Therefore, if the homolog can be found at the Super Kingdom rank, the gene (G) will be classified as "*Super Kingdom*" TRG. If all the subject genomes belong to the same Kingdom ($\#K_h=1$), the algorithm moves to the Phylum level. Let P_h be the set of Phylum taxonomies of the subject genomes. ORFanID then checks to see how many distinct Phyla ($\#P_h$) are found in the subject sequences. If $\#P_h > 1$ Phyla are not common, but their Kingdom is common. If the homolog can be found at Kingdom ranks, therefore the gene (G) will be categorized as a "*Kingdom*" TRG. In this way, this search will continue to the Class(C), Order(O), Family(F), Genus(G), and Species(S) levels. If a homolog is found, the ORFanID algorithm will identify the sequence as a TRG at the appropriate taxonomical rank. If the homologs are found only at the Species level the input sequence is identified as an "ORFan Gene". If there are no homologs even at the Species level, the sequence is named a "Strict Orfan".

2.2 Operation

ORFanID accepts either protein or nucleotide (gene) sequences singly or as multiple gene sequences in the FASTA format (Figure 5). Users can easily retrieve multiple protein or gene

sequences by providing multiple accessions to the sequence search engine. ORFanID supports both NCBI as well as Uniprot accessions. Optionally, the user can upload a FASTA file or directly copy the sequence into the ORFanID engine as specified on the ORFanID website. Next, users can select a species from the dropdown menu, which includes species' scientific names, NCBI taxonomy IDs, and images of the species for easy visual recognition. Finally, the accuracy of the results can be fine-tuned using advanced parameters such as E-value, percent identity, and a maximum number of target sequences for each BLAST and PSI-BLAST search (Figure 6). The default values for the advanced parameters are; max target sequences, 500; percent identity ranging from 40 - 100%; expected value, 10^{-3}; as used in most cited papers.

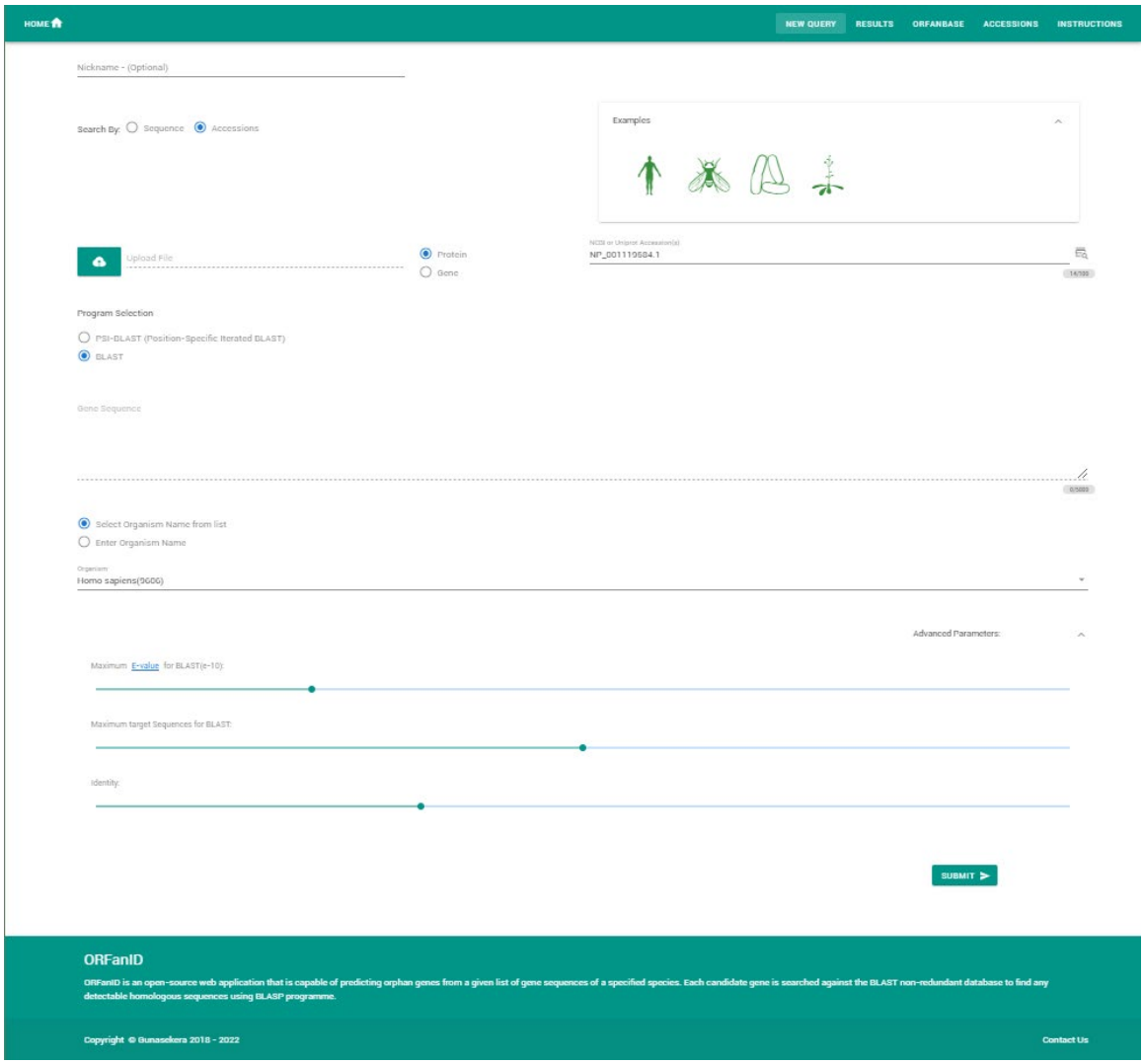
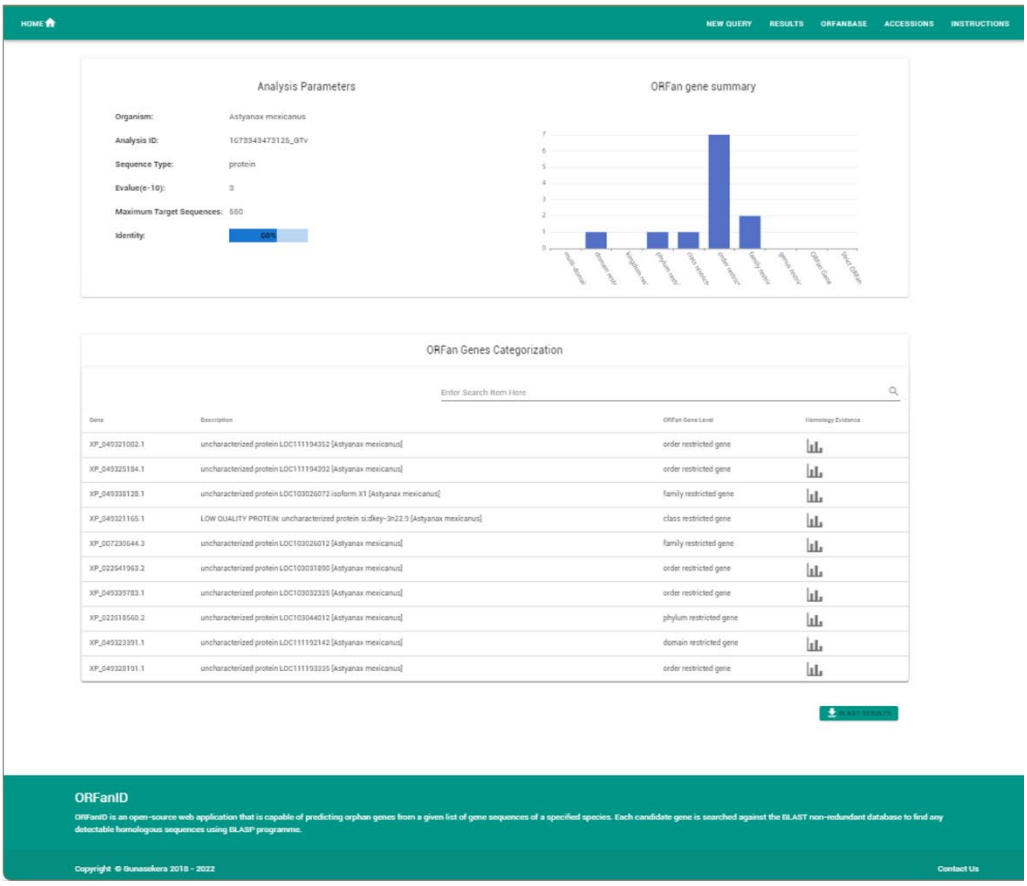


Figure 5: ORFanID Input Page. Four distinct examples are provided to new users for demonstrative purposes. After the user submits their sequence, species, and advanced parameters, ORFanID will perform the steps discussed in the Algorithm section above (Figure 4) and display results as in Figure 6A.

A


HOME
NEW QUERY
RESULTS
ORFANBASE
ACCESSIONS
INSTRUCTIONS
Analysis Parameters
Organism:
Astyanax mexicanus
Analysis ID:
1673343473125_0Tv
Sequence Type:
protein
Evalue(e-10):
3
Maximum Target Sequences:
550
Identity:
60%
ORFan gene summary
ORFan Genes Categorization
Enter Search Item Here
Gene
Description
ORFan Gene Level
Homology Evidence
XP_049321002.1
uncharacterized protein LOC111194352 [Astyanax mexicanus]
order restricted gene
XP_049325184.1
uncharacterized protein LOC111194392 [Astyanax mexicanus]
order restricted gene
XP_049338128.1
uncharacterized protein LOC103026072 isoform X1 [Astyanax mexicanus]
family restricted gene
XP_049321165.1
LOW QUALITY PROTEIN: uncharacterized protein si:dkey-3n22.5 [Astyanax mexicanus]
class restricted gene
XP_007230644.3
uncharacterized protein LOC103026012 [Astyanax mexicanus]
family restricted gene
XP_022541963.2
uncharacterized protein LOC103031890 [Astyanax mexicanus]
order restricted gene
XP_049339783.1
uncharacterized protein LOC103032325 [Astyanax mexicanus]
order restricted gene
XP_022518560.2
uncharacterized protein LOC103044012 [Astyanax mexicanus]
phylum restricted gene
XP_049323391.1
uncharacterized protein LOC111192142 [Astyanax mexicanus]
domain restricted gene
XP_049328191.1
uncharacterized protein LOC111193335 [Astyanax mexicanus]
order restricted gene
ORFanID
ORFanID is an open-source web application that is capable of predicting orphan genes from a given list of gene sequences of a specified species. Each candidate gene is searched against the BLAST non-redundant database to find any detectable homologous sequences using BLASTP programme.
Copyright © Gunasekera 2018 - 2022
Contact Us

B

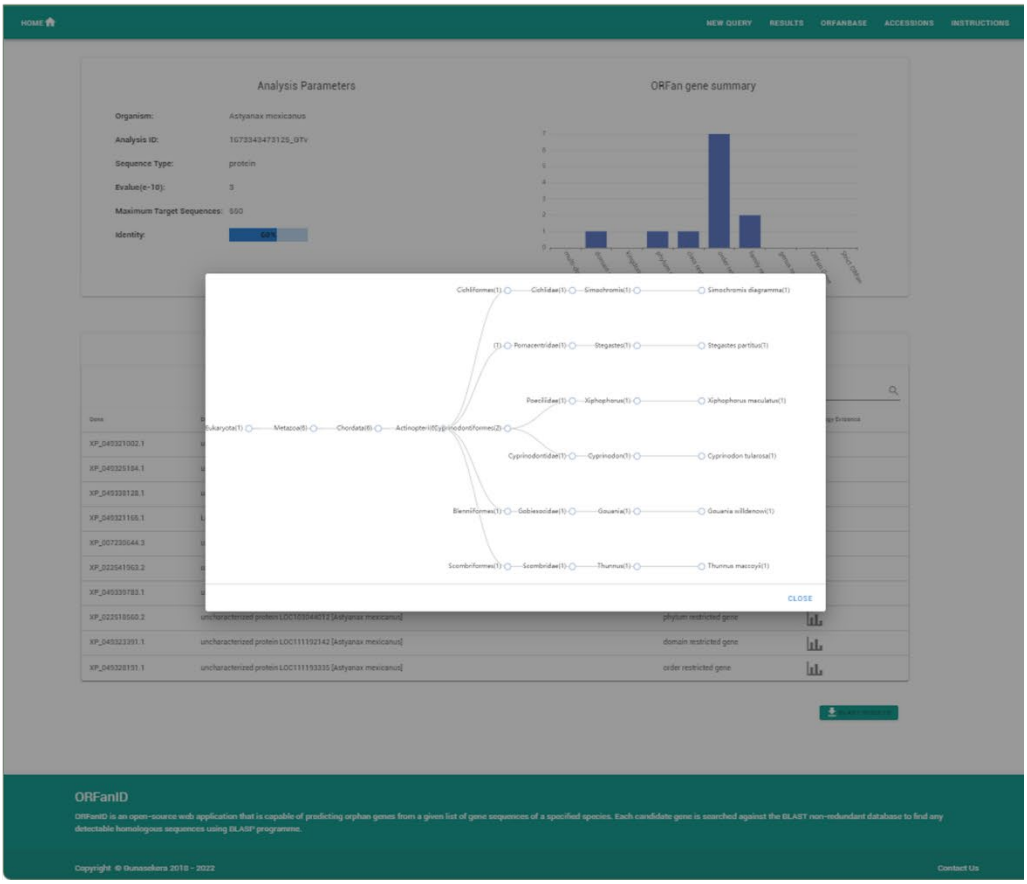

HOME
NEW QUERY
RESULTS
ORFANBASE
ACCESSIONS
INSTRUCTIONS
Analysis Parameters
Organism:
Astyanax mexicanus
Analysis ID:
1673343473125_0Tv
Sequence Type:
protein
Evalue(e-10):
3
Maximum Target Sequences:
550
Identity:
60%
ORFan gene summary
Cichliformes(1)
Cichlidae(1)
Simochromis(1)
Simochromis diagramma(1)
Pomacentridae(1)
Stegastes(1)
Stegastes partitus(1)
Poeciliidae(1)
Xiphophorus(1)
Xiphophorus maculatus(1)
Eukaryota(1)
Metazoa(8)
Chordata(8)
Actinopterii(Cyprinodontformes)(2)
Cyprinodontidae(1)
Cyprinodon(1)
Cyprinodon tularosa(1)
Blenniformes(1)
Gobiesocidae(1)
Gouania(1)
Gouania willdenowi(1)
Scombriformes(1)
Scombridae(1)
Thunnus(1)
Thunnus maccoyii(1)
CLOSE
ORFanID
ORFanID is an open-source web application that is capable of predicting orphan genes from a given list of gene sequences of a specified species. Each candidate gene is searched against the BLAST non-redundant database to find any detectable homologous sequences using BLASTP programme.
Copyright © Gunasekera 2018 - 2022
Contact Us

Figure 6: ORFanID Results Page. (A) This page consists of a top left table that summarizes the metadata of the analysis, a graph on the top right that depicts the taxonomy of the TRGs or orphan genes for easy visualization, and a table at the bottom that displays the categorization levels and description of the analyzed genes. (B) ORFanID BLAST Results in Graphical Display. The homology results can be viewed graphically by selecting any of the graph icons at the bottom of the results page (Figure 6A). This interactive chart visualizes the number of BLAST matches for each taxonomic rank. Each node in the tree represents the related orthologs found at each taxonomic rank, and. the graphical display is expandable and collapsible by taxonomy level. In addition, all the tables on the results page are sortable by column and can be easily filtered using the fuzzy search functionality. In addition to its web service, ORFanID can be downloaded and installed on a local server, but the local BLAST software must be downloaded from NCBI and run on the server. The Source-code and installation instructions are freely available online.

ORFanID has a dedicated web page that outlines the installation and operating instructions and includes video tutorials to help users understand its web interface. As previously noted, the sequence submissions page also provides four example sets of input gene data for demonstrational purposes, which can be input either as FASTA file sequences or NCBI accession numbers. These four examples analyze sequences of the following species: *Escherichia coli* (562), *Drosophila* melanogaster (7227), *Homo sapiens* (9606), and *Arabidopsis thaliana* (3702). The ORFanID application is designed for ease of use, rich graphics, and reasonable speed to provide a research tool for identifying orphan genes at all taxonomic levels.

3. Results and Evaluations

We tested the functionality of ORFanID by analyzing the protein and DNA sequences (or NCBI accessions) of various organisms, such as *C. elegans*, *S. cerevisiae*, *D. melanogaster*, and *H. sapiens* (Tabel 1 and Supplementary Table 1). We adjusted the parameters (e-value <=10^3) to filter low-quality results (transposable elements, low-complexity protein regions, etc.) from our analyses. Our results show that ORFanID effectively assigns the protein accessions to their respective taxonomic levels (Tabel 1 and Supplementary Table 1). For instance, paired box-containing Pax-6 proteins are tissue-specific transcriptional factors highly conserved across most animal phyla. As expected, ORFanID classified *Drosophila* melanogaster Pax-6 proteins, Eyeless (NP_726607.1), and Twin of eyeless (NP_001259080.1) phylum and class-level proteins, respectively.

Similarly, we tested several proteins in *Homo sapiens*, including N-cym (NP_001316897.1) and SPANX (NP_073152.2). The N-cym protein regulates the stability of the proto-oncogene MYCN in neuroblastoma cells, while the SPANX family proteins are expressed in human spermatozoa and are extensively studied in Down syndrome patients [28]. A previous study predicted that these two proteins were restricted to the order Primates [29]. Surprisingly, ORFanID showed that only N-cym is order restricted, while SPANX was classified as a family-restricted protein. To understand this disparity, we searched OrthoDB and the non-redundant database of NCBI using *H. sapiens* SPANX protein (NP_073152.2) as the query [30]. Interestingly, both databases show that SPANX proteins are restricted to the Hominidae family. These results support that the ORFanID algorithm is accurate in classifying proteins to their respective taxonomic group based on the functionality of BLAST and the choice of parameter settings.

Next, to accurately identify species-specific orphan genes, we tested genes from various species that were previously shown to be orphans by using ORFanID. The *Arabidopsis thaliana* (*A. thaliana*) *QQS* gene was one of the first plant genes shown to be a species-level orphan [31]. It

regulates starch biosynthesis in leaves, increases seed protein, and enhances resistance to pathogens [31, 32]. Our results show that ORFanID accurately classifies QQS protein (NP_189695.1) as an *A. thaliana*-specific protein. Similarly, we tested the species-specific orphan genes of some model organisms, such as *D. melanogaster*, *C. elegans*, and *S. cerevisiae*. ORFanID precisely identified these published genes as species-specific orphans. The *D. melanogaster* group-specific orphan genes, *jeanbaptiste*, and *karr* are predominantly expressed in the male germline and are functionally important [33]. RNAi-mediated knock-down of these genes leads to male-specific developmental defects and partial lethality. These genes were accurately grouped as strict species-level orphans by ORFanID. In *Saccharomyces cerevisiae*, the *de novo* genes *bsc4* and *fyv5*, which regulate DNA repair and vegetative growth, respectively, were accurately predicted as species-specific by ORFanID. Finally, we analyzed the *C. elegans*-specific gene *ify-1* required for proper chromosome segregation during cell division. ORFanID correctly categorized *ify-1* as a species-specific protein, a result confirmed by the "wormbase.org" database, which did not show any orthologous of the *ify-1* gene [34]. Taken together, these results suggest that ORFanID works accurately and reliably in classifying and identifying species-specific orphan genes. A complete list of taxonomically restricted or species-specific orphans we tested using ORFanID is available in the Supplemental Information.

It is important to note that as the NCBI and other sequence-based databases grow, what is currently classified as an orphan may not be an orphan with the availability of new sequence information. Furthermore, the database of orphans and TRGs are dynamic and can change as sampling grows. However, we have noticed that the percentage of orphans in every species tends to stabilize with the increasing number of genomes synthesized. The ORFanID is updated regularly so that the databases and files obtained from NCBI are current.

ORFanID yields accurate results in comparison with other classifiers

We compared ORFanFinder to ORFanID, from five different organisms (*Caenorhabditis elegans*, *Escherichia coli*, *Homo sapiens*, *Oryza sativa*, and *Zea mays*). The results showed that ORFanID is more sensitive and gives more accurate results in classifying orphan and strict-orphan genes. It should be noted that ORFanFinder is no longer updated, and the last version of its database is from 2018. We also compared ORFanID with abSENSE and found that ORFanID accurately classifies 29% of the fungal genes as orphans and 37% of the Insect genes (Supplemental Table 2) [6]. This supports the accuracy of ORFanID in classifying orphans and TRGs. However, it should be noted that as more organisms are sequenced and their DNA becomes available in public databases, a percentage of these classified orphans may prove not to be true orphans.

4. Discussion

Here we present the implementation of a web-based, easy-to-use, intuitive tool designed to help geneticists, biologists, bioinformaticians, and computational biologists, identify and discover orphan genes across any taxonomic lineage. The search engine is designed to be user-friendly and requires minimal computational skills when compared to other bioinformatics programs. The tool is based on the BLAST algorithm and uses the genomic sequences of various species made available through NCBI [1, 16]. Although previous stand-alone programs such as ORFanFinder existed in the past, their scope was limited to protein databases. ORFanID, on the other hand, is the first web-based, easy-to-use tool that allows users to identify orphan genes and TRGs using both nucleotide and protein sequences and classify any taxon of interest found in the NCBI databases.

With the increase in interest in orphan genes, researchers are using various approaches to identify and understand these unique genes with specific biological significance. Recently a new deep learning model (CNN + Transformer) was used to identify orphan genes in moso bamboo [35].

Another recently published website, eggNOG 6.0 database, provides a large number of species with functional annotations that allows the comparison of genes and potentially further our understanding of orphan genes [36]. One potential limitation of using BLAST as the approach to identify ORFan genes is that BLAST may not be able to identify gene lineages lacking detectable homologs. This could lead to the potential erroneous identification of ORFan genes. A recent study showed that many lineage-specific ORFan genes could be explained by homology detection failure [6]. However, these findings will need to be explored further, and ORFanID would allow the comparison and evaluation of many such genes.

There are other practical applications of this search engine that can provide reliable species identification. For example, if someone wanted to detect the presence or absence of human DNA in a sample, using PCR for human ORFans would be an effective way of doing this. If customs officers wanted to check the identity of unknown goods for CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora) listed species, such as black rhinoceros, using PCR for a rhino ORFans could be used for this type of identification. The use of orfans could be more reliable than DNA "barcoding," where a common gene is sequenced, and a species is identified based on a few SNPs that are considered species-specific.

In addition to its interactive user interface, ORFanID allows users to seamlessly use the BLAST database to investigate and categorically identify orphan genes. These findings can lead to the creation of databases of these clandestine genes that have been pre-identified at the various taxonomic levels by ORFanID, resulting in a deeper understanding of the purpose of orphan genes, their function in genomes, and their potential impact on life.

5. Author Contributions

RG conceptualized and led the project along with PN. SH and VG architected and developed the software. KR tested the program and helped with writing. TG and ET helped with writing and editing. VG managed the software team that brought this software to life. The authors declare that there is no conflict of interest.

6. Funding

This work was supported in part by the Roth Family.

7. Acknowledgments

We acknowledge the programming work of intern Savidu Dias, video creations by intern Megan Rupp, and research fellow Jesse Gentile, including technical support of engineering interns Kenneth Chen and Daniel Dirksen. We thank Marisa Arthur for her expert editorial assistance. The support of the Roth Family for the interns and the expert technical work of Chesalon USA, Inc. is greatly appreciated.

8. Supplementary Information

Supplementary data are available in the online repository of the journal. The Supplementary Material for this article can be found online at (Link to supplemental Information)

9. Data Availability and Implementation Statement

The ORFanID Web Application and Database are freely available to non-commercial users at http://www.orfangenes.com/. The Development Stack of ORFanID comprises a VueJs front-end, a

back-end in the Spring Boot framework, and databases in PostgreSQL. The software modules of ORFanID are containerized using Docker and communicate using RESTful APIs. ORFanID supports all major browsers. The Source Code for ORFanID is freely available at https://github.com/orfanid

Tables

Table 1. Table showing the ORFanID classification of different genes. Five known sequences are accurately classified by ORFanID.

Species	Taxonomy ID	Accession	ORFanID Classification
D. melanogaster	7227	NP_723266.1	Strict ORFan [1]
D. melanogaster	7227	NP_726607.1	Phylum
C.elegans	6239	NP_500848.2	Order [2]
C. elegans	6239	NP_494931.1	Orfan gene [3]
S.cervisiae	4932	NP_001335743.1	Strict ORFan [4]
S. cervisiae	4932	NP_014130.1	Strict ORFan [5]
H. sapiens	9606	NP_073152.2	Genus [6]
H. sapiens	9606	NP_001316897.1	Order [7]
D. rerio	7955	XP_002665008.4	ORFan gene [8]
D. rerio	7955	NP_571379.1	Class [9]

[1] Orphan gene based on Levine et al., 2006. No orthologs in orthodb.
[2] TRG based on Verster et al., 2017. WormBase & orthodb show Nematode orthologs
[3] TRG based on Verster et al., 2017. WormBase & orthodb show orthologs in *Caenorhabditis genus*
[4] *de novo* gene. No orthologs in other databases
[5] Strict Orphan based on publication (Cai J et al., 2008).BSC4 may be involved in the DNA repair pathway during the stationary phase of S. cerevisiae and contribute to the robustness of *S. cerevisiae* when shifted to a nutrient-poor environment
[6] Order restricted based on a publication and family based on ortho database
[7] Order restricted based on a publication. Orthodb shows conservation in eutheria clade
[8] It was shown to be genus restricted. According to orthodb, this protein has orthologs in the Actinopterygii class
[9] Should be present in all seeing animals. should be multi-domain

References

1. Fischer D, Eisenberg D. Finding families for genomic ORFans. Bioinformatics. 1999;15(9):759-62. Epub 1999/09/28. doi: 10.1093/bioinformatics/15.9.759. PubMed PMID: 10498776.

2. Prabh N, Rödelsperger C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? BMC Bioinformatics. 2016;17(1):226. doi: 10.1186/s12859-016-1102-x.

3. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol. 2013;5(2):439-55. Epub 2013/01/26. doi: 10.1093/gbe/evt009. PubMed PMID: 23348040; PubMed Central PMCID: PMCPMC3590893.

4. Yu G, Stoltzfus A. Population diversity of ORFan genes in Escherichia coli. Genome Biol Evol. 2012;4(11):1176-87. Epub 2012/10/05. doi: 10.1093/gbe/evs081. PubMed PMID: 23034216; PubMed Central PMCID: PMCPMC3514957.

5. Gao Q, Jin X, Xia E, Wu X, Gu L, Yan H, et al. Identification of Orphan Genes in Unbalanced Datasets Based on Ensemble Learning. Front Genet. 2020;11:820. Epub 2020/11/03. doi: 10.3389/fgene.2020.00820. PubMed PMID: 33133122; PubMed Central PMCID: PMCPMC7567012.

6. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol. 2020;18(11):e3000862. Epub 2020/11/03. doi: 10.1371/journal.pbio.3000862. PubMed PMID: 33137085; PubMed Central PMCID: PMCPMC7660931.

7. Johnson BR. Taxonomically Restricted Genes Are Fundamental to Biology and Evolution. Front Genet. 2018;9:407. Epub 2018/10/09. doi: 10.3389/fgene.2018.00407. PubMed PMID: 30294344; PubMed Central PMCID: PMCPMC6158316.

8. Johnson BR, Tsutsui ND. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. BMC Genomics. 2011;12:164. Epub 2011/03/31. doi: 10.1186/1471-2164-12-164. PubMed PMID: 21447185; PubMed Central PMCID: PMCPMC3072959.

9. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009;25(9):404-13. Epub 2009/09/01. doi: 10.1016/j.tig.2009.07.006. PubMed PMID: 19716618.

10. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011;12(10):692-702. Epub 2011/09/01. doi: 10.1038/nrg3053. PubMed PMID: 21878963.

11. Merhej V, Raoult D. Rhizome of life, catastrophes, sequence exchanges, gene creations, and giant viruses: how microbial genomics challenges Darwin. Front Cell Infect Microbiol. 2012;2:113. Epub 2012/09/14. doi: 10.3389/fcimb.2012.00113. PubMed PMID: 22973559; PubMed Central PMCID: PMCPMC3428605.

12. Doolittle WF, Bapteste E. Pattern pluralism and the Tree of Life hypothesis. Proceedings of the National Academy of Sciences. 2007;104(7):2043-9. doi: doi:10.1073/pnas.0610699104.

13. Lupas AN, Ponting CP, Russell RB. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol. 2001;134(2-3):191-203. Epub 2001/09/12. doi: 10.1006/jsbi.2001.4393. PubMed PMID: 11551179.

14. Tautz D. Not just another genome. BMC Biology. 2011;9(1):8. doi: 10.1186/1741-7007-9-8.

15. Ibrahim A, Colson P, Merhej V, Zgheib R, Maatouk M, Naud S, et al. Rhizomal Reclassification of Living Organisms. International Journal of Molecular Sciences. 2021;22(11):5643. PubMed PMID: doi:10.3390/ijms22115643.

16. Yin Y, Fischer D. On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. BMC Evolutionary Biology. 2006;6(1):63. doi: 10.1186/1471-2148-6-63.

17. Ekstrom A, Yin Y. ORFanFinder: automated identification of taxonomically restricted orphan genes. Bioinformatics. 2016;32(13):2053-5. doi: 10.1093/bioinformatics/btw122.

18. Domazet-Loso T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 2007;23(11):533-9. Epub 2007/11/22. doi: 10.1016/j.tig.2007.08.014. PubMed PMID: 18029048.

19. Priyam A, Woodcroft BJ, Rai V, Moghul I, Munagala A, Ter F, et al. Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. Mol Biol Evol. 2019;36(12):2922-4. Epub 2019/08/15. doi: 10.1093/molbev/msz185. PubMed PMID: 31411700; PubMed Central PMCID: PMCPMC6878946.

20. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647-9. Epub 2012/05/01. doi: 10.1093/bioinformatics/bts199. PubMed PMID: 22543367; PubMed Central PMCID: PMCPMC3371832.

21. Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics. 2021;37(18):3029-31. doi: 10.1093/bioinformatics/btab184.

22. Mirdita M, Steinegger M, Söding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics. 2019;35(16):2856-8. Epub 2019/01/08. doi: 10.1093/bioinformatics/bty1057. PubMed PMID: 30615063; PubMed Central PMCID: PMCPMC6691333.

23. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238. Epub 2019/11/16. doi: 10.1186/s13059-019-1832-y. PubMed PMID: 31727128; PubMed Central PMCID: PMCPMC6857279.

24. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59-60. Epub 2014/11/18. doi: 10.1038/nmeth.3176. PubMed PMID: 25402007.

25. Altenhoff AM, Glover NM, Train CM, Kaleb K, Warwick Vesztrocy A, Dylus D, et al. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res. 2018;46(D1):D477-d85. Epub 2017/11/07. doi: 10.1093/nar/gkx1019. PubMed PMID: 29106550; PubMed Central PMCID: PMCPMC5753216.

26. Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178-89. Epub 2003/09/04. doi: 10.1101/gr.1224503. PubMed PMID: 12952885; PubMed Central PMCID: PMCPMC403725.

27. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013;14:106. Epub 2013/03/26. doi: 10.1186/1471-2105-14-106. PubMed PMID: 23522326; PubMed Central PMCID: PMCPMC3648438.

28. Salemi M, Romano C, Barone C, Calí F, Caraci F, Romano C, et al. SPANX-B and SPANX-C (Xq27 region) gene dosage analysis in Down's syndrome subjects with undescended testes. J

Genet. 2009;88(1):93-7. Epub 2009/05/07. doi: 10.1007/s12041-009-0013-2. PubMed PMID: 19417550.

29. Kouprina N, Noskov VN, Pavlicek A, Collins NK, Schoppee Bortz PD, Ottolenghi C, et al. Evolutionary diversification of SPANX-N sperm protein gene structure and expression. PLoS One. 2007;2(4):e359. Epub 2007/04/05. doi: 10.1371/journal.pone.0000359. PubMed PMID: 17406683; PubMed Central PMCID: PMCPMC1831492.

30. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019;47(D1):D807-d11. Epub 2018/11/06. doi: 10.1093/nar/gky1053. PubMed PMID: 30395283; PubMed Central PMCID: PMCPMC6323947.

31. Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, et al. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. Plant J. 2009;58(3):485-98. Epub 2009/01/22. doi: 10.1111/j.1365-313X.2009.03793.x. PubMed PMID: 19154206.

32. Qi M, Zheng W, Zhao X, Hohenstein JD, Kandel Y, O'Conner S, et al. QQS orphan gene and its interactor NF-YC4 reduce susceptibility to pathogens and pests. Plant Biotechnol J. 2019;17(1):252-63. Epub 2018/06/08. doi: 10.1111/pbi.12961. PubMed PMID: 29878511; PubMed Central PMCID: PMCPMC6330549.

33. Reinhardt JA, Jones CD. Two rapidly evolving genes contribute to male fitness in Drosophila. J Mol Evol. 2013;77(5-6):246-59. Epub 2013/11/14. doi: 10.1007/s00239-013-9594-8. PubMed PMID: 24221639; PubMed Central PMCID: PMCPMC3880551.

34. Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Cho J, et al. WormBase: a modern Model Organism Information Resource. Nucleic Acids Res. 2020;48(D1):D762-d7. Epub 2019/10/24. doi: 10.1093/nar/gkz920. PubMed PMID: 31642470; PubMed Central PMCID: PMCPMC7145598.

35. Zhang X, Xuan J, Yao C, Gao Q, Wang L, Jin X, et al. A deep learning approach for orphan gene identification in moso bamboo (Phyllostachys edulis) based on the CNN + Transformer model. BMC Bioinformatics. 2022;23(1):162. doi: 10.1186/s12859-022-04702-1.

36. Hernández-Plaza A, Szklarczyk D, Botas J, Cantalapiedra Carlos P, Giner-Lamia J, Mende DR, et al. eggNOG 6.0: enabling comparative genomics across 12 535 organisms. Nucleic Acids Research. 2022;51(D1):D389-D94. doi: 10.1093/nar/gkac1022.