2-2022

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

Richard S. Gunasekera
Biola University

Komal K.B. Raja

Suresh Hewapathirana

Thushara Galbadage
Biola University

Emanuel Tundrea

See next page for additional authors

Recommended Citation

Gunasekera, Richard S.; Raja, Komal K.B.; Hewapathirana, Suresh; Galbadage, Thushara; Tundrea, Emanuel; Gunasekera, Vinodh; and Nelson, Paul A., "ORFanID: A Web-Based Search Engine for the Discovery and Identification of Orphan and Taxonomically Restricted Genes" (2022). *Faculty Articles & Research*. 668.
https://digitalcommons.biola.edu/faculty-articles/668

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], Thushara Galbadage[4], Emanuel Tundrea[5], Vinodh Gunasekera[6], and Paul A. Nelson[1*]

[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, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK,

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

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

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

*** Correspondence:** Corresponding
Authors
richard.gunasekera@biola.edu
paul.alfredp@gmail.com

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

Abstract

With the multiplicity of genomes sequenced today, it has been shown that significant percentages of genes in any given taxon do not possess orthologous sequences in other taxa. These sequences are typically designated as orphans/ORFans when found as singletons in one species only or taxonomically restricted genes (TRGs) when found at higher taxonomic ranks. Quantitative and collective studies of these genes are necessary for understanding their biological origins. Currently, orphan gene identifying software is limited, and those previously available are either not functional, are limited in their database search range, or are very complex algorithmically. Thus, an interested researcher studying orphan genes must harvest their data from many disparate sources. ORFanID is a graphical web-based search engine that efficiently finds both orphan genes and TRGs at all taxonomic levels, from DNA or amino acid sequences in the entire NCBI database cluster and other large bioinformatics repositories. This algorithm allows the easy identification of both orphan genes and TRGs using both nucleotide and protein sequences in any species of interest. ORFanID identifies genes unique to any taxonomic rank, from species to a domain, using standard NCBI systematic classifiers. The software allows for user control of the NCBI database search parameters. The results of the search are provided in a spreadsheet as well as a graphical display. All the tables in the software are sortable by column, and results can be easily filtered with fuzzy search functionality. In addition, the visual presentation is expandable and collapsible by taxonomy. Availability and Implementation**:** The ORFanID Web Application and Database are available at http://www.orfangenes.com/ 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 mid1990s, the comparative analysis of whole genomes has revealed that a large number of protein-coding open reading frames (ORFs) occur only in one species or only in taxonomically restricted genes (TRGs). These TRG's occur at systematic ranks from genus upwards. Species-specific ORFs have been designated as "orphan" (sometimes spelled "ORFan") genes, whereas more widely distributed ORFs, not present universally but only at lower taxonomic ranks, have been designated as TRGs. By definition, orphan genes have no orthologous (i.e., homologous) sequences in any other species. Similarly, 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 in any larger systematic category: Insecta, Arthropoda, and so forth.

Historically, genetic studies and analyses have favored focusing on conserved genes, common either to all organisms (e.g., ribosomal genes) or to broader systematic categories (e.g., the Wnt signaling pathway in Metazoa), due to likely a lack of interest in studying taxonomically unique ORFs and orphan genes. This is because functional roles are generally assigned to newly sequenced genes via homology criteria (e.g., existing annotations). However, it is shown that orphan genes are uniquely involved in making one species distinct from another phenotypically (Prabh and Rödelsperger 2016; Wissler et al. 2013; Yu and Stoltzfus 2012) which is also quite significant.

The last decade has seen a rapidly increasing interest in studying orphan genes (Gao, et al. 2020; Weisman, Murray and Eddy 2020; and B. Johnson 2018). Genomic sequencing has revealed that large fractions of the genes in the complete genome of a given species do not possess orthologous sequences in other species. Therefore, TRGs represent important mediators of phenotypic novelty (Johnson and Tsutsui 2011; Khalturin et al. 2009; D. Tautz 2011; Tautz and Domazet-Lošo2011). Previous theory held that all genes have descended by modification from the set of coding sequences present in the

Last Universal Common Ancestor (LUCA) (Merhej and Raoult, 2012; Doolittle and Bapteste, 2007). Thus, the prevailing theory had posited that the duplication of existing genes, followed by a divergence into new functions, brought forth all extant genes (Ohno, S. 1970, Lupas, Ponting and Russell 2001). 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 finding of orphan genes has led to a changing picture of gene evolution and formation. The novel emergence of genes suggests de novo formation may be a predominant mechanism for gene emergence (Tautz and Domazet-Lošo 2011). This radical shift in thinking has significant ramifications for understanding the nature of genes, the genome's non-coding regions, and fully functional sequences (D. Tautz 2011). This calls for the re-analysis of existing data and comparative genomic analysis, which can provide new and accurate understandings leading to theories in genetics, genomics, and the tree of life (Ibrahim et al. 2021). In time, the biological understanding of the origins and the functions of orphan genes will have applications in both medicine and evolutionary biology across the tree of life. Therefore, it can be asserted that 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 calls for easy-to-use bioinformatics tools that can be utilized by scientists from the life sciences and those with computational backgrounds. Tools have been made available through BLAST (Fischer and Eisenberg 1999; Yin and Fischer 2006) and related machine language programs (Ekstrom and Yin 2016).
However, tools are further necessary and useful for discovering orphan genes and assigning TRGs to their taxonomic levels (sometimes known as phylostratigraphy) (Domazet-Lošo, Brajković and Tautz 2007).

The tool options researchers can find today to study orphan genes are very limited, simply because most software solutions focus on identifying orthologs or inferencing ortho groups and because their scope is generally limited to proteins. For instance, ORFanFinder (Ekstrom and Yin 2016) functionality is limited to plants, bacteria, and fungi. While the URL provided in the original publication (DOI: 10.1093/bioinformatics/btw122) does not work, it appears that ORFanFinder is active at http://bcb.unl.edu/orfanfinder/ although it does not seem to have been updated. SequenceServer (Priyam et al. 2019) performs blast without classifying the proteins/DNA sequences to taxonomic levels. The software Geneious (Kearse et al. 2012) performs alignment and can build a phylogenetic tree but identifying orphans using this software may be arduous. Similarly, OrthoFinder (Emms and Kelly 2019) provides the option to use DIAMOND (Buchfink, Xie and Huson 2015) or it's recommended MMseq2 (Mirdita, Steinegger and Söding 2019) for aligning sequences. OMA orthology (Altenhoff et al. 2018) 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 (Li, Stoeckert and Roos 2003)] only show orthologous genes/proteins and do not identify orphan genes. A new gene classification platform, www.shoot.bio, may also be helpful to align and compare gene origins but using protein (amino acid) sequences only. In short, these tools perform alignment or identify conserved genes from the genomes but are not uniquely designed to identify orphans as ORFanID is designed to do.

ORFanID's distinctiveness is threefold: (1) Its scope includes processing not just the protein/amino acid sequences but also DNA/nucleotide sequences. (2) Using 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 mine the data further.

2. Methods

2.1 Algorithm and Implementation

ORFanID identifies orphan genes and TRGs from a given list of DNA or protein sequences (Figure 1) mainly by NCBI accession number. Based on the protein or DNA sequence, detectable homologous sequences are found in the NCBI non-redundant databases using the BLAST alignment tool. All the tools mentioned above recognize NCBI databases are among the most trusted sources. BLAST is a well-established tool, but users acknowledge it is slow in processing as the databases grow rapidly. Therefore, the ORFanID search engine can also be slower at times based on BLAST and NCBI server performance.

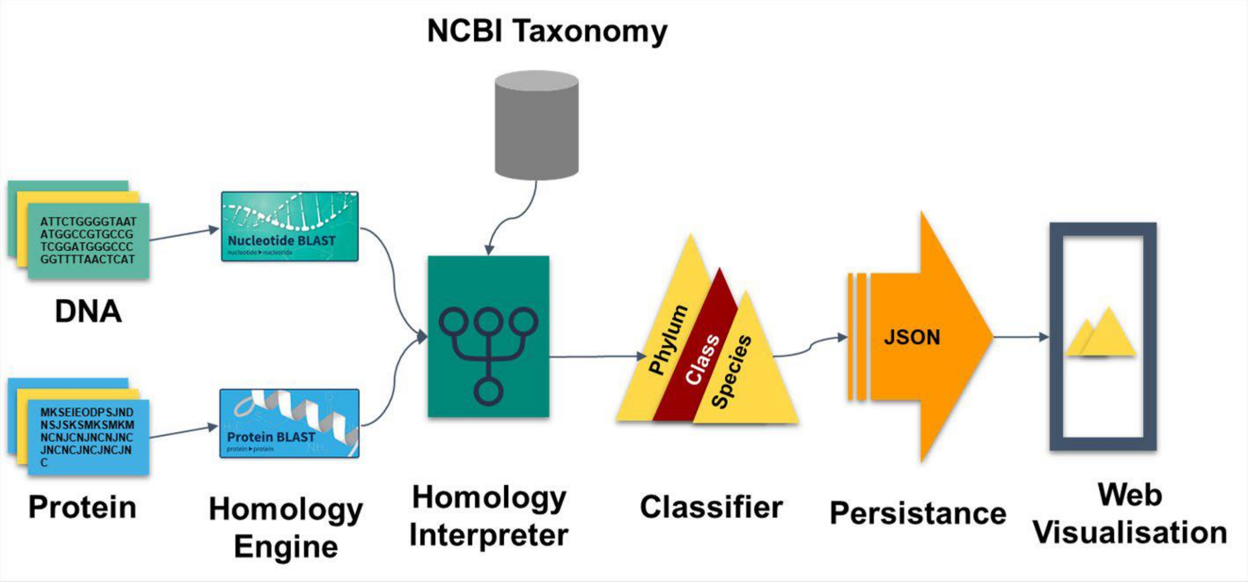
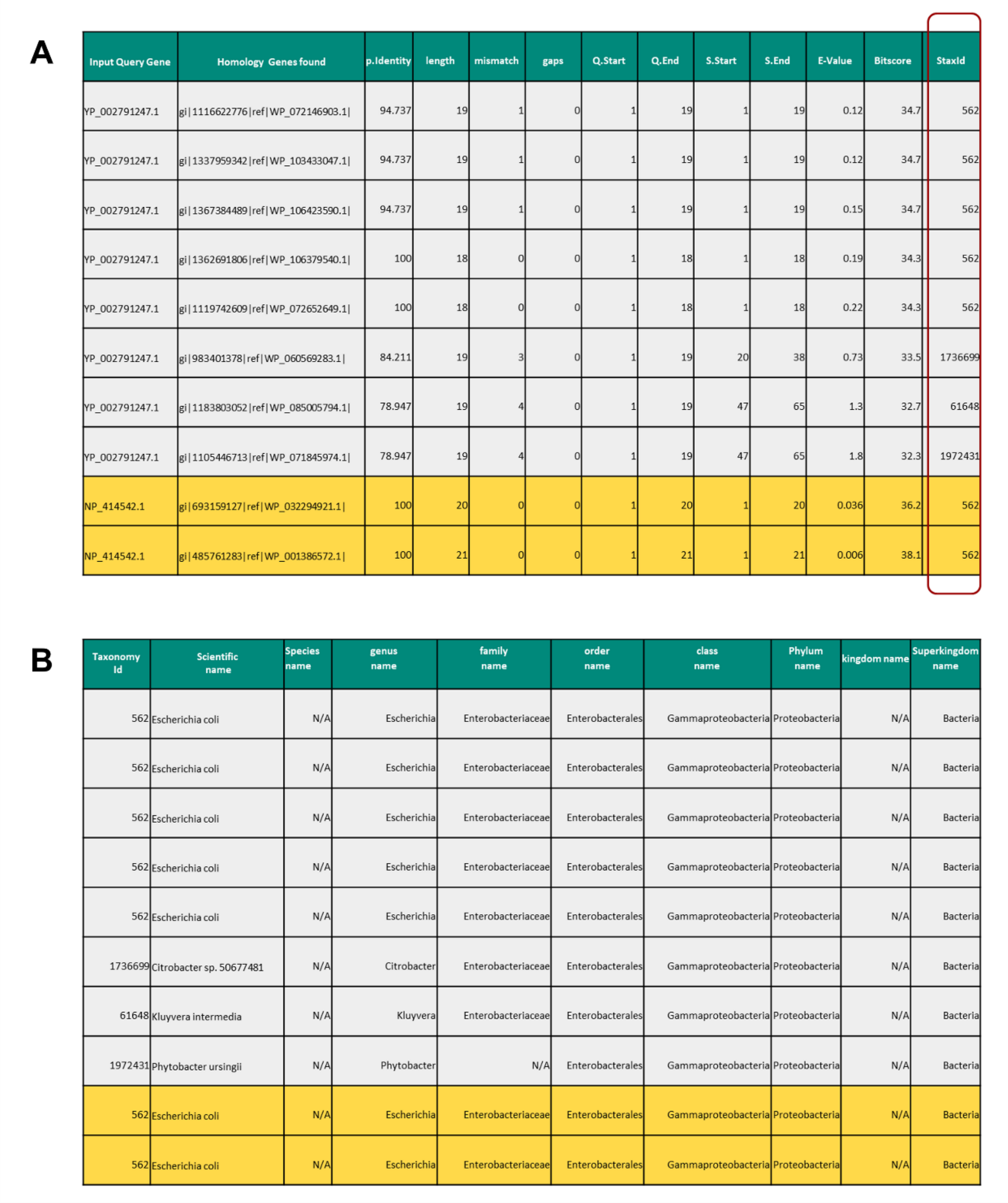


Figure 1: Core engine of ORFanID

ORFanID is a web-based application developed on Java Spring Framework. The Web UI was developed in Thymeleaf and MaterializedCSS, which follows Google Material design concepts.

If the query sequence is a protein, the BLASTP program is used, while the BLASTN program is used for the nucleotide sequences. ORFanID allows the user to submit multiple sequences in FASTA format so that the queries can be queued for processing using software message brokering techniques. The results will be combined into a single blast report in tab-delimited format, downloadable by the user. ORFanID sends a customized BLAST command to retrieve taxonomy IDs of each hit (the "staxid" column of Figure 2A). Using the rank lineage information file provided by the NCBI Taxonomy Database, the ORFanID Homology Interpreter finds the defined taxonomy level (species, genus, family, order, class, phylum, kingdom, and superkingdom) for each sequence found in the BLAST results. (Figure 2B).

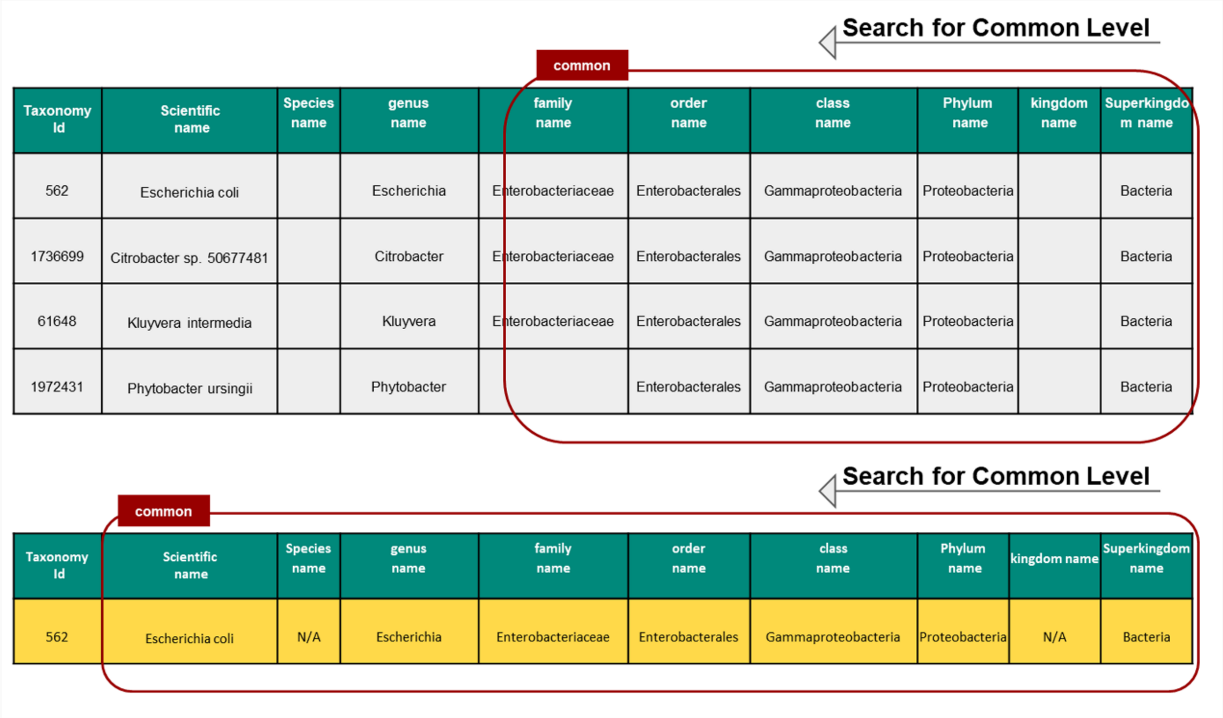


Input Query Gene	Homology Genes found	p.Identity	length	mismatch	gaps	Q.Start	Q.End	S.Start	S.End	E-Value	Bitscore	StaxId
YP_002791247.1	gi|1116622776|ref|WP_072146903.1|	94.737	19	1	0	1	19	1	19	0.12	34.7	562
YP_002791247.1	gi|1337959342|ref|WP_103433047.1|	94.737	19	1	0	1	19	1	19	0.12	34.7	562
YP_002791247.1	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

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	N/A	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: (A) Finding homologous genes by sequence similarity.
(B) Finding taxonomy levels for each homology.

After finding the complete lineage (Linnaeus taxonomy) for the individual homology for each input DNA or protein sequence, the ORFanID Classifier begins searching from the Superkingdom level working towards the Species level to find common ancestors (Figure 3). The taxonomic rank at which homologs are located then yields the designation “Taxonomically Restricted Gene” (TRG) for the cluster of orthologs (homologs) captured by the classifier. Here ORFanID is using the single ortholog de-classfier rule -meaning ORFanID is sufficient for the finding of only one ortholog at a certain level in the tree of life- in order to classify the gene of interest as a TRG at that level (or as an orphan at the final level). ORFanID does not need a significant number of orthologs (as the SMOTE technique would employ) to find that a sequence is not an orphan (Blagus et al 2022). 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” by ORFanID. Using this algorithm, ORFanID identifies and displays orphan genes or TRGs. Results can be viewed and analyzed graphically.


Search for Common Level
common
Search for Common Level
common

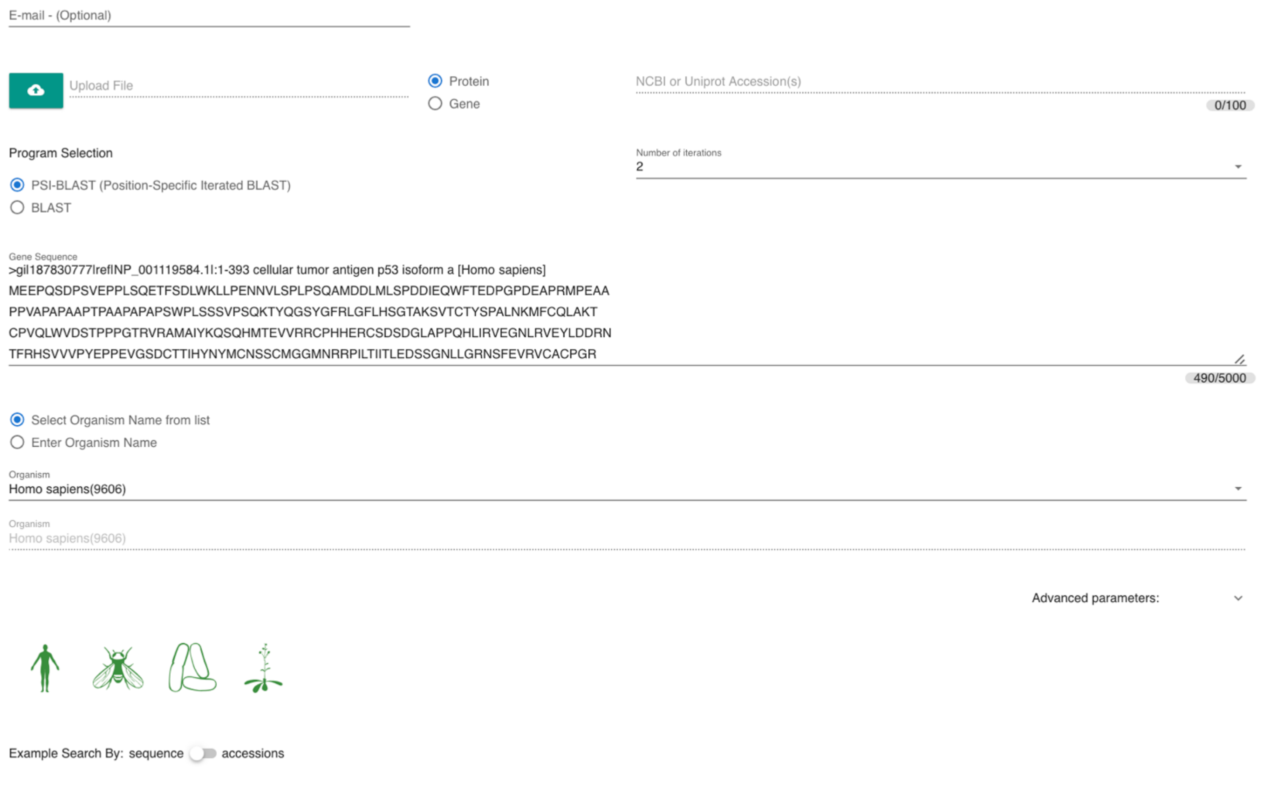
Taxonomy Id	Scientific name	Species name	genus name	family name	order name	class name	Phylum name	kingdom name	Superkingdom name
562	Escherichia coli		Escherichia	Enterobacteriaceae	Enterobacterales	Gammaproteobacteria	Proteobacteria		Bacteria
1736699	Citrobacter sp. 50677481		Citrobacter	Enterobacteriaceae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

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

Figure 3: Gene Classification. The search starts from the superkingdom to species level until it finds homologous species.

2.2 Operation

ORFanID accepts either the protein or nucleotide (gene) sequences of a single or multiple gene sequences in the FASTA format. Users can easily retrieve numerous protein or gene sequences by providing multiple accessions to the sequence search engine. Currently, 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 according to specifications provided on the ORFanID website. Secondly, users can select the species from the dropdown menu, containing species scientific names, the NCBI taxonomy ID, and an image of the species for convenient visual recognition. Finally, by using the advanced parameters, the accuracy of the results can be fine-tuned based on the E-value and a maximum number of target sequences for each BLAST search. Default values for the advanced parameters are as follows: max target sequences, 500; identity ranging from 40 - 100%; expected value, 10^{-3}; since most cited papers use this value.



E-mail - (Optional)

Upload File

● Protein
○ Gene

NCBI or Uniprot Accession(s)

0/100

Program Selection

● PSI-BLAST (Position-Specific Iterated BLAST)
○ BLAST

Number of iterations
2

Gene Sequence

```
>gil187830777lreflNP_001119584.1l:1-393 cellular tumor antigen p53 isoform a [Homo sapiens]
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAA
PPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKT
CPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRN
TFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGR
```

490/5000

● Select Organism Name from list
○ Enter Organism Name

Organism
Homo sapiens(9606)

Organism
Homo sapiens(9606)

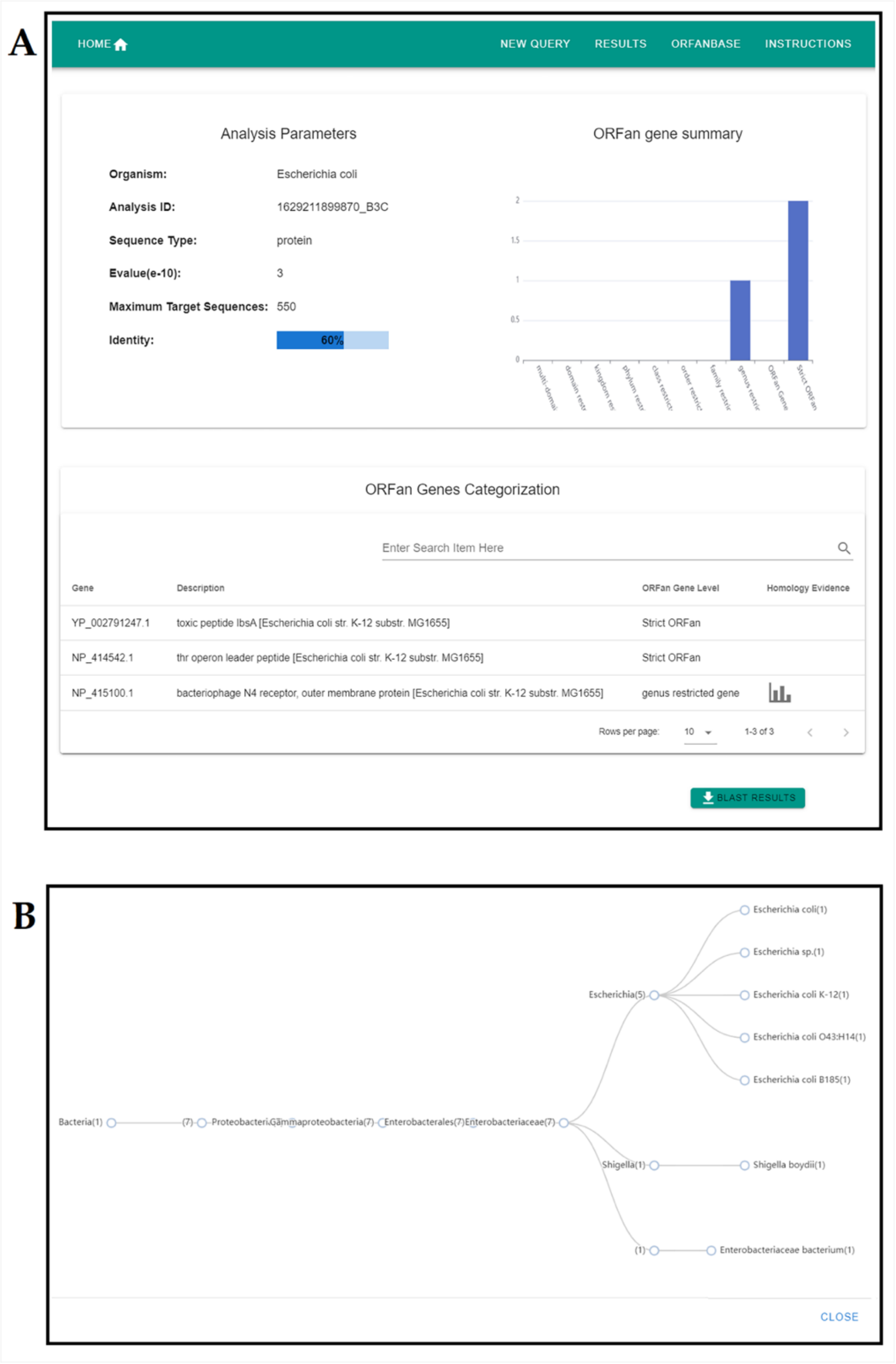
Advanced parameters:

Example Search By: sequence accessions

SUBMIT ➤

Figure 4: ORFanID Input Page.

As Figure 4 illustrates, four distinct examples are provided to new users for demonstrative purposes. Once the user submits their sequence and the species, along with the advanced parameters, ORFanID will execute the steps discussed in the Algorithm section above and display results in Figure 5A. Figure 5A, the top-left table, summarizes the metadata of the analysis. The graph on the top right depicts the taxonomy of the TRG or orphan gene for convenient visualization. The table at the bottom shows the categorization levels of the genes discovered, along with the actual taxonomy level for each gene.



Analysis Parameters

Organism:	Escherichia coli
Analysis ID:	1629211899870_B3C
Sequence Type:	protein
Evalue(e-10):	3
Maximum Target Sequences:	550
Identity:	60%

ORFan Genes Categorization

Gene	Description	ORFan Gene Level	Homology Evidence
YP_002791247.1	toxic peptide IbsA [Escherichia coli str. K-12 substr. MG1655]	Strict ORFan	
NP_414542.1	thr operon leader peptide [Escherichia coli str. K-12 substr. MG1655]	Strict ORFan	
NP_415100.1	bacteriophage N4 receptor, outer membrane protein [Escherichia coli str. K-12 substr. MG1655]	genus restricted gene	📊

Figure 5:
(A) ORFanID Results Page.
(B) ORFanID BLAST Results Graphical Display.

The homology results can be viewed graphically by selecting any of the graph icons at the bottom of the results page (Figure 5B). The number of BLAST matches for each taxonomic rank is visualized in this interactive chart. Each node in the tree shows the related orthologs found at each taxonomic rank. The graphical display is expandable and collapsible by taxonomy level. All the tables on the results page are sortable by column, and the results can be easily filtered by fuzzy search functionality. Besides its web service, ORFanID can be downloaded and installed on a local server. Furthermore, the local BLAST software needs to be downloaded from NCBI and run on the server. The Source-code and installation instructions are freely available online.

At the user experience level, ORFanID is equipped with a dedicated web page that outlines the installation and operating instructions. The site also includes video tutorials that will help users understand its web interface. As noted previously, the sequence submissions page also provides four example sets of input gene data for demonstrational purposes. They are available to be inputted both as FASTA file sequences or by 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 operation of the ORFanID application is architected for ease of use, rich graphics, and reasonable speed for providing a research instrument to identify 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* (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 the ORFanID effectively assigns the protein accessions to their respective taxonomic levels

(Supplementary Table 1). For instance, paired box-containing pax-6 proteins are tissue-specific transcriptional factors highly conserved across most animal phyla. As expected, testing the *Drosophila* melanogaster Pax-6 proteins, Eyeless (NP_726607.1) and Twin of eyeless (NP_001259080.1) using ORFanID, showed that these proteins are restricted to phylum and class levels.

Similarly, we tested several proteins in *H. sapiens*, including N-cym (NP_001316897.1) and SPANX (NP_073152.2). N-cym protein regulates the stability of the proto-oncogene MYCN in neuroblastoma cells. In contrast, the SPANX family proteins are expressed in human spermatozoa and are extensively studied in Down syndrome patients (Salemi et al., 2009). A previous study predicted that these two proteins are restricted to the order Primates (Kouprina et al. 2007). 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 (Kriventseva et al. 2019) and the non-redundant database of NCBI using *H. sapiens* SPANX protein (NP_073152.2) as the query. 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 upon the functionality of BLAST and the choice of parameter settings.

Next, to accurately identify species-specific orphan genes, we tested genes from various species previously shown to be orphans by using ORFanID. The *A. thaliana QQS* gene is one of the first plant genes shown to be a species-level orphan (Li et al., 2009). It is involved in the regulation of starch biosynthesis in leaves (Li et al., 2009), increasing of seed protein, and enhancing the resistance to pathogens (Qi et al., 2018). 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, were predominantly expressed in the male germline and are functionally very important because RNAi-mediated knock-down of these genes lead to male-specific developmental defects and partial-lethality (Reinhardt and Jones 2013). As we expected, ORFanID grouped these genes as strict species-level orphans. In *S. 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. As expected, ORFanID rightly categorized ify-1 as a species-specific protein. This result was further confirmed by the database "wormbase.org" (Harris et al. 2020), which did not show any orthologous of the ify-1 gene. Taken together, these results suggest that ORFanID works accurately and reliably in classifying and identifying species-specific orphan genes. Refer to Supplemental Information for a complete list of taxonomically restricted or species-specific orphans we tested using ORFanID.

Nonetheless, as the NCBI and other such databases grow, what is currently classified as an orphan may be re-classified at another level in their taxonomic group. Furthermore, the database of orphans and TRGs is a dynamic list. As such, ORFans may have a provisional status and thus change as sampling grows. However, we noticed over the years that the percentage of orphans in every species tends to stabilize with the growing number of genomes synthesized. Hence, ORFanID regularly updates the databases and files obtained from NCBI.

We compared ORFanFinder to ORFanID, resulting from genes from five different organisms (*Caenorhabditis elegans*, *Escherichia coli*, *Homo sapiens*, *Oryza sativa*, and *Zea mays*). Results show that ORFanID is more sensitive to give accurate results in classifying orphan and strict-orphan genes (ORFanFinder is no longer updated, and the last version of its database is from 2018). Refer to Supplemental Table 2 for more details.

We also compared ORFanID with abSENSE (Weisman, Murray, and Eddy 2020) and found that ORFanID classifies only 68% of the Fungal genes as orphans and 90% of the Insect. This attests that ORFanID is accurate in classifying orphans and TRGs. This experiment may also attest to the reality in which a percentage of orphans may prove not to be orphans when more organisms are sequenced and available in the DNA public databases. Refer to Supplemental Table 3 for more detailed results.

4. Discussion

Here we present the implementation of a web-based, easy-to-use intuitive tool helpful in identifying and discovering orphan genes across any taxonomic lineage by geneticists/biologists, bioinformaticians, and computational biologists. The search engine is designed to require minimal computational skills to operate when compared to other bioinformatics programs. This web-based interactive tool is based on the BLAST database and uses the genomic sequences for various species made available through NCBI (Fischer and Eisenberg 1999) and (Yin and Fischer 2006). While previous stand-alone programs such as ORFanFinder have existed in the past, their scope was limited mainly to protein databases. This algorithm 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 from ORFanID, any taxon of interest found in the NCBI databases can be classified.

There are other practical uses of this search engine that can help reliable species identification. For example, if one wanted to detect the presence or absence of human DNA in a sample, PCR for a human ORFans would be an effective way of doing this. Or 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, PCR for a rhino orfans could be used for this kind of identification. The use of orfans could be more reliable than DNA “barcoding,” where a widespread gene is sequenced, and a species is identified based on a few SNPs that are thought to be species specific.

Taken together with 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 pre-identified at the various taxonomic levels as ascertained by ORFanID, leading to a much 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. The authors declare that there is no conflict of interest.

6. Funding

This work was supported in part by the Peter & Carla 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 Peter & Carla 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.

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 a database 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

References

Altenhoff, Adrian, Natasha Glover, Clément-Marie Train, Klara Kaleb, Alex Warwick Vesztrocy, David Dylus, Tarcisio de Farias, et al. 2018. “The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces.” Nucleic Acids Research 46 (D1): D477–D485. doi:10.1093/nar/gkx1019.

Blagus, R., Lusa, L. 2013. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* **14,** 106. doi:10.1186/1471-2105-14-106

Buchfink, Benjamin, Chao Xie, and Daniel Huson. 2015. “Fast and Sensitive Protein Alignment using DIAMOND.” Nature Methods 12: 59–60. doi:10.1038/nmeth.3176.

Carvunis, AR., Rolland, T., Wapinski, I. et al. 2012. “Proto-genes and de novo gene birth.” Nature 487: 370–374. doi:10.1038/nature11184.

Clamp, Michele and Fry, Ben and Kamal, Mike and Xie, Xiaohui and Cuff, James and Lin, Michael F. and Kellis, Manolis and Lindblad-Toh, Kerstin and Lander, Eric S. 2007. “Distinguishing proteincoding and non-coding genes in the human genome.” Proceedings of the National Academy of Sciences (National Academy of Sciences) 104 (49): 19428--19433. doi:10.1073/pnas.0709013104.

Daubin, Vincent,Ochman, Howard. 2004. “Bacterial Genomes as New Gene Homes: The Genealogy of ORFans in *E. coli*.” Genome Research (Cold Spring Harbor Laboratory Press) 14: 1036-1042. doi:10.1101/gr.2231904.

Domazet-Lošo, Tomislav, Josip Brajković, and Diethard Tautz. 2007. “A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages.” Trends in Genetics 23 (11): 533-539. doi:10.1016/j.tig.2007.08.014.

Donoghue, M.T., Keshavaiah, C., Swamidatta, S.H. et al. 2011. "Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*." BMC Ecology and Evolution 11 (47). doi:10.1186/1471-2148-11-47.

Doolittle, W.F. and Bapteste, E., 2007. Pattern pluralism and the Tree of Life hypothesis. Proceedings of the National Academy of Sciences, 104(7), pp.2043-2049.

Ekstrom, Alex, and Yanbin Yin. 2016. "ORFanFinder: automated identification of taxonomically restricted orphan genes." Bioinformatics 32 (13): 2053–2055. doi:10.1093/bioinformatics/btw122.

Emms, David M, and Steven Kelly. 2019. "OrthoFinder: phylogenetic orthology inference for comparative genomics." Genome Biology 20 (238). doi:10.1186/s13059-019-1832-y.

Fischer, D, and D Eisenberg. 1999. "Finding families for genomic ORFans." Bioinformatics 15 (9): 759–762. doi:10.1093/bioinformatics/15.9.759.

Gao, Qijuan, Xiu Jin, Enhua Xia, Xiangwei Wu, Lichuan Gu, Hanwei Yan, Yingchun Xia, and Shaowen Li. 2020. "Identification of Orphan Genes in Unbalanced Datasets Based on Ensemble Learning." Frontiers in Genetics (Systems Biology Archive) 11: 820. doi:10.3389/fgene.2020.00820.

Harris, T, T Arnaboldi, S Cain, J Chan, W Chen, J Cho, P Davis, et al. 2020. "WormBase: a modern Model Organism Information Resource." Nucleic Acids Research 48 (D1): D762–D767. doi:10.1093/nar/gkz920.

Ibrahim, Ahmad, Philippe Colson, Vicky Merhej, Rita Zgheib, Mohamad Maatouk, Sabrina Naud, Fadi Bittar, and Didier Raoult. 2021. "Rhizomal Reclassification of Living Organisms." International Journal of Molecular Sciences 22 (11): 5643. doi:10.3390/ijms22115643.

Johnson, B.R., and N.D. Tsutsui. 2011. “Taxonomically restricted genes are associated with the evolution of sociality in the honey bee.” BMC Genomics 12 (164). doi:10.1186/1471-2164-12-164.

Johnson, Brian. 2018. “Taxonomically Restricted Genes Are Fundamental to Biology and

Evolution.” Frontiers in Genetics (Evolutionary and Population Genetics) 9: 407. doi:10.3389/fgene.2018.00407.

Kearse, Matthew, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, et al. 2012. “Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data.” Bioinformatics 28 (12): 1647–1649. doi:10.1093/bioinformatics/bts199.

Khalturin, Konstantin, Hemmrich Georg, Fraune Sebastian, Augustin René, and Thomas C.G. Bosch. 2009. “More than just orphans: are taxonomically-restricted genes important in evolution?” {Trends in Genetics 25 (9): 404-413. doi:10.1016/j.tig.2009.07.006.

Klasberg, Steffen, Tristan Bitard-Feildel, Ludovic Mallet. 2016. “Computational Identification of Novel Genes: Current and Future Perspectives.” Bioinformatics and Biology Insights 10: 121-131. doi:10.4137/BBI.S39950.

Kouprina, Natalay, Vladimir Noskov, Adam Pavlicek, Keith Collins, Pamela Bortz, Chris Ottolenghi,

Dmitri Loukinov, et al. 2007. “Evolutionary Diversification of SPANX-N Sperm Protein Gene Structure and Expression.” PLOS ONE. doi:10.1371/journal.pone.0000359.

Kriventseva, EV, D Kuznetsov, F Tegenfeldt, M Manni, R Dias, FA Simão, and EM Zdobnov. 2019. “OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs." Nucleic Acids Research 8 (47): D807D811. doi:10.1093/nar/gky1053.

Kuhn, RM, D Haussler, and WJ Kent. 2013. "The UCSC genome browser and associated tools." Briefings in Bioinformatics 144-161. doi:10.1093/bib/bbs038.

Lander, Eric S. 2019. "2018 William Allan Award: Discovering the Genes for Common Disease: From Families to Populations." The American Journal of Human Genetics 104 (3): 375-383. doi:10.1016/j.ajhg.2019.01.016.

Li, L, CM Foster, Q Gan, D Nettleton, MG James, AM Myers, and ES Wurtele. 2009. "Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves." Plant Journal 58 (3): 485-498. doi:10.1111/j.1365-313X.2009.03793.x.

Li, Li, Christian J. Jr Stoeckert, and David S Roos. 2003. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." Genome Research (Cold Spring Harbor Laboratory Press) (13): 2178-2189. doi:10.1101/gr.1224503.

Lin, H., Moghe, G., Ouyang, S. et al. 2010. "Comparative analyses reveal distinct sets of lineagespecific genes within *Arabidopsis thaliana*." BMC Evolutionary Biology 10 (41). doi:10.1186/14712148-10-41.

Lupas, AN, CP Ponting, and RB Russell. 2001. "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?" Journal of Structural Biology 134 (2-3): 191-203. doi:10.1006/jsbi.2001.4393.

Merhej, V. and Raoult, D., 2012. Rhizome of life, catastrophes, sequence exchanges, gene creations, and giant viruses: how microbial genomics challenges Darwin. Frontiers in cellular and infection microbiology, 2, p.113.

Mirdita, Milot, Martin Steinegger, and Johannes Söding. 2019. “MMseqs2 desktop and local web server app for fast, interactive sequence searches.” Bioinformatics 35 (16): 2856–2858. doi:10.1093/bioinformatics/bty1057.

Ohno S. Evolution by gene duplication. Berlin: Springer; 1970.

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

Priyam, Anurag, Ben J Woodcroft, Vivek Rai, Ismail Moghul, Alekhy Munagala, Fili Ter, Hite Chowdhary, et al. 2019. “Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases.” Molecular Biology and Evolution 36 (12): 2922–2924. doi:10.1093/molbev/msz185.

Qi, Mingsheng, Zheng Wenguang, Xuefeng Zhao, Jessica Hohenstein, Yuba Kandel, Seth O'Conner, Yifan Wang, et al. 2018. “*QQS* orphan gene and its interactor NF-YC4 reduce susceptibility to pathogens and pests.” Plant Biotechnology Journal 17 (1): 252-263. doi:10.1111/pbi.12961.

Reinhardt, JA, and CD Jones. 2013. “Two rapidly evolving genes contribute to male fitness in *Drosophila*.” Journal of Molecular Evolution 77 (5-6): 246-59. doi:10.1007/s00239-013-9594-8.

Salemi, M., Romano, C., Barone, C. et al. 2008. “SPANX-B and SPANX-C (Xq27 region) gene dosage analysis in Down’s syndrome subjects with undescended testes.” Journal of Genetics 88: 93-97. doi:10.1007/s12041-009-0013-2.

Tautz, D. 2014. “One size does not fit all.” eLIFE Evolutionary Morphology. doi:10.7554/eLife.02088.

Tautz, D, and T Domazet-Lošo. 2011. “The evolutionary origin of orphan genes.” Nature Reviews Genetics 12 (10): 692-702. doi:10.1038/nrg3053.

Tautz, D. 2011. “Not just another genome.” BMC Biology 9 (8). doi:10.1186/1741-7007-9-8.

Weisman, Caroline M., Andrew W. Murray, and Sean R. Eddy. 2020. “Many, but not all, lineagespecific genes can be explained by homology detection failure.” PLoS Biology 18 (11). doi:10.1371/journal.pbio.3000862.

Wissler, L, J Gadau, Simola, D. F, M Helmkampf, and E Bornberg-Bauer. 2013. “Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes.” Genome biology and evolution 5 (2): 439–455. doi:10.1093/gbe/evt009.

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

Yu, G, and A Stoltzfus. 2012. “Population diversity of ORFan genes in *Escherichia coli*.” Genome Biology and Evolution 4 (11): 1176-87. doi:10.1093/gbe/evs081.