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]The Ethics and Society Research Centre, 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. Therefore, 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 too 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 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. To our knowledge, this is the first program that allows the identification of orphan genes 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.

48 **1. Introduction**

49 Following the introduction of large-scale, high-throughput automated DNA sequencing in the
50 mid1990s, the comparative analysis of whole genomes has revealed that a significant number of
51 protein-coding open reading frames (ORFs) occur only in one species or only in taxonomically
52 restricted groups (TRG's). These TRG's occur at systematic ranks from genus upwards. Species-
53 specific ORFs have been designated as "orphan" (sometimes spelled "ORFan") genes, whereas more
54 widely distributed ORFs, not present universally but only Linnaean taxa, have been designated as
55 "taxonomically restricted genes" (TRGs). By definition, orphan genes have no counter orthologous
56 (i.e., homologous) sequences in any other species. Similarly, TRGs are not present in any genomes
57 outside their respective taxa. For example, a TRG may be found only in the genus *Drosophila*, but not
58 in any other Dipteran or in any larger systematic category: Insecta, Arthropoda, and so forth.

59 Historically, most genetic analyses have omitted the study of orphan and taxonomically
60 restricted genes and favored focusing on conserved genes, common either to all organisms (e.g., a few
61 tens of ribosomal sequences) or to broader systematic categories (e.g., the Wnt signaling pathway in
62 Metazoa), due to the difficulty of studying taxonomically unique ORFs. This is because functional
63 roles are generally assigned to newly sequenced genes via homology criteria (e.g., existing
64 annotations). Since conserved genes were straightforward to analyze in evolutionary theory,
65 systematics, translational and medical research, this focus is easily understood. However, we believe
66 that understanding orphan genes, which evidence shows (Prabh and Rödelsperger 2016; Wissler et al.
67 2013; Yu and Stoltzfus 2012), are uniquely involved in making one species distinct from another
68 phenotypically, is equally significant.

69 The last decade has seen a rapidly increasing interest in studying orphan genes (Gao et al. 2020;
70 Weisman et al. 2020; and Johnson 2018). Genomic sequencing has revealed that significant fractions
71 of the genes in the complete genome of a given species do not possess orthologous sequences in other

species. Therefore, TRGs represent evolutionarily important mediators of phenotypic novelty (Johnson and Tsutsui 2011; Khalturin et al. 2009; Tautz 2011; Tautz and Domazet-Lošo 2011). Previous dogma implied that all genes have descended by modification from the set of coding sequences present in the Last Universal Common Ancestor (LUCA). Thus, the prevailing theory had posited that the duplication of existing genes, followed by a divergence into new functions, brought forth all extant genes (Lupas et al. 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 (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 one of the most intriguing aspects 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).

95 However, web-based, easy-to-use intuitive tools are further necessary and useful for identifying and
96 discovering orphan genes across taxonomic lineage phylostratigraphy - the process of determining the
97 phylogenetic origin of every gene in a genome (Domazet-Lošo et al. 2007).

98 The tool options a researcher can find today to study orphan genes are very limited, simply
99 because most software solutions focus on identifying orthologs or inferencing ortho groups and because
100 their scope is generally limited to proteins. For instance, ORFanFinder (Ekstrom and Yin 2016)
101 functionality is limited to plants, bacteria, and fungi. While the URL provided in the original
102 publication (http://cys.bios.niu.edu/orfanfinder) does not work, it appears that ORFanFinder is active
103 at http://bcb.unl.edu/orfanfinder/ although it does not seem to have been updated. SequenceServer
104 (Priyam et al. 2019) performs blast without classifying the proteins/DNA sequences to taxonomic
105 levels. The software Geneious (Kearse et al. 2012) performs alignment and can build a phylogenetic
106 tree, but identifying orphans using this software may be arduous. Similarly, OrthoFinder (Emms and
107 Kelly 2019) provides the option to use DIAMOND (Buchfink et al. 2015) or it's recommended
108 MMseq2 (Mirdita et al. 2019) is for aligning sequences. OMA orthology (Altenhoff et al. 2018) or the
109 series of analytical resources developed by the Bioinformatics Resource Centers (BRCs) for Infectious
110 Diseases program [bioinformatics tools, workspaces and services for bioinformatics data analysis like
111 AmoebaDB, FungiDB, OrthoMCL (Li et al. 2003)] only show orthologous genes/proteins and do not
112 identify orphan genes. In short, these tools perform alignment or identify conserved genes from the
113 genomes but are not uniquely designed to identify orphans as ORFanID is designed to do.

114 ORFanID's distinctiveness is threefold: (1) Its scope includes processing not just the
115 protein/amino acid sequences but also DNA/nucleotide sequences. (2) Using its built-in homology
116 interpreter and classifier, the search engine provides the taxonomic rank of a gene either as an orphan
117 gene or as a gene restricted to a taxonomic level in the tree of life; (3) As Orphans and TRGs are

118 identified, ORFanID builds its own database with the results of the analysis and provides the researcher
119 with the possibility to mine the data further.

120 **2. Methods**

121 ***2.1 Algorithm and Implementation***

122 ORFanID identifies orphan genes and TRGs from a given list of DNA or protein sequences
123 (Figure 1) mainly by accession number. Based on the protein or DNA sequence, detectable homologous
124 sequences are found in the NCBI non-redundant databases using the BLAST alignment tool. All the
125 tools mentioned above recognize NCBI databases are among the most trusted sources. BLAST is a
126 well-established tool, but users acknowledge it is slow in processing as the databases grow rapidly.

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

136 After finding the complete lineage for the individual homology for each input DNA or protein
137 sequence, the ORFanID Classifier begins searching from the Superkingdom level working towards the
138 Species level to find common ancestors (Figure 3). The taxonomic rank at which homologs are located
139 then yields the designation "Taxonomically Restricted Gene" (TRG) for the cluster of orthologs
140 (homologs) captured by the classifier. If no orthologous sequences are found, and the ORF remains a
141 species-unique sequence, it will be classified as an "Orphan Gene." If the ORF is unique at the
142 subspecies level, the gene is classified as a "Strict Orphan" by ORFanID. Using this algorithm,

143 ORFanID identifies and displays orphan genes or TRGs. Results can be viewed and analyzed
144 graphically. Further details on the ORFanID algorithm are provided in the Supplementary Information
145 section of this publication.

146 ***2.2 Operation***

147 ORFanID accepts either the protein or amino acid sequences of a single or multiple gene
148 sequences in the FASTA format. Users can easily retrieve numerous genes or protein sequences by
149 providing multiple accessions to the sequence search engine. Currently, ORFanID supports both NCBI
150 as well as Uniprot accessions. Optionally, the user can upload a FASTA file or directly copy the
151 sequence into the ORFanID engine according to specifications provided on the ORFanID website.
152 Secondly, users can select the species from the dropdown menu, containing species scientific names,
153 the NCBI taxonomy ID, and an image of the species for convenient visual recognition. Finally, by
154 using the advanced parameters, the accuracy of the results can be fine-tuned based on the E-value and
155 a maximum number of target sequences for each BLAST search. Default values for the advanced
156 parameters are as follows: max target sequences, 500; identity 60% (minimum percent similarity of
157 alignment); expected value, 10-3; since most cited papers use this value.

158 As Figure 4 illustrates, four distinct examples are provided to new users for demonstrative
159 purposes. Once the user submits their sequence and the species, along with the advanced parameters,
160 ORFanID will execute the steps discussed in the Algorithm section above and display results in Figure
161 5A. Figure 5A, the top-left table, summarizes the metadata of the analysis. The graph on the top right
162 depicts the taxonomy of the TRG or orphan gene for convenient visualization. The table at the bottom
163 shows the categorization levels of the genes discovered, along with the actual taxonomy level for each
164 gene.

165 The homology results can be viewed graphically by selecting any of the graph icons at the
166 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. 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 both as FASTA file sequences and as 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 prokaryotic and eukaryotic organisms, such as *E. coli*, *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.

192 Similarly, we tested several proteins in *H. sapiens*, including N-cym (NP_001316897.1) and
193 SPANX (NP_073152.2). N-cym protein regulates the stability of the proto-oncogene MYCN in
194 neuroblastoma cells. In contrast, the SPANX family proteins are expressed in human spermatozoa and
195 are extensively studied in Down syndrome patients (Salemi et al. 2009). A previous study predicted
196 that these two proteins are restricted to the order Primates (Kouprina et al. 2007). Surprisingly,
197 ORFanID showed that only N-cym is order restricted, while SPANX was classified as a family
198 restricted protein. To understand this disparity, we searched OrthoDB (Kriventseva et al. 2019) and the
199 non-redundant database of NCBI using *H. sapiens* SPANX protein (NP_073152.2) as the query.
200 Interestingly, both databases show that SPANX proteins are restricted to the Hominidae family. These
201 results support that the ORFanID algorithm is accurate in classifying proteins to their respective
202 taxonomic group based upon the functionality of BLAST and the choice of parameter settings.

203 Next, to accurately identify species-specific orphan genes, we tested genes from various species
204 previously shown to be orphans by using ORFanID. The *A. thaliana QQS* gene is one of the first plant
205 genes shown to be a species-level orphan (Li et al. 2009). It is involved in the regulation of starch
206 biosynthesis in leaves (Li et al. 2009), increasing of seed protein, and enhancing the resistance to
207 pathogens (Qi et al. 2018). Our results show that ORFanID accurately classifies QQS protein
208 (NP_189695.1) as an *A. thaliana* specific protein. Similarly, we tested the species-specific orphan
209 genes of some model organisms, such as *D. melanogaster*, *C. elegans*, and *S. cerevisiae*. ORFanID
210 precisely identified these published genes as species-specific orphans. The *D. melanogaster* group-
211 specific orphan genes, jeanbaptiste, and karr were predominantly expressed in the male germline and
212 are functionally very important because RNAi-mediated knock-down of these genes lead to male-
213 specific developmental defects and partial-lethality (Reinhardt and Jones 2013). As we expected,
214 ORFanID grouped these genes as strict species-level orphans. In *S. cerevisiae*, the de novo genes bsc4
215 and fyv5, which regulate DNA repair and vegetative growth, respectively, were accurately predicted
216 as species-specific by ORFanID. Finally, we analyzed the *C. elegans*-specific gene ify-1 required for

217 proper chromosome segregation during cell division. As expected, ORFanID rightly categorized ify-1
218 as a species-specific protein. This result was further confirmed by the database "wormbase.org" (Harris
219 et al. 2020), which did not show any orthologous of the ify-1 gene. Taken together, these results suggest
220 that ORFanID works accurately and reliably in classifying and identifying species-specific orphan
221 genes. Refer to Supplemental Information for a complete list of taxonomically restricted or species-
222 specific orphans we tested using ORFanID.

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

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

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

239 **4. Discussion**

240 Here we present the implementation of a web-based, easy-to-use intuitive tool helpful in
241 identifying and discovering orphan genes across any taxonomic lineage by geneticists/biologists,

242 bioinformaticians, and computational biologists. The search engine is designed to require minimal
243 computational skills to operate when compared to other bioinformatics programs. This web-based
244 interactive tool is based on the BLAST database and uses the genomic sequences for various species
245 made available through NCBI (Fischer and Eisenberg 1999) and (Yin and Fischer 2006). While
246 previous stand-alone programs such as ORFanFinder have existed in the past, their scope was limited
247 mainly to protein databases. To our knowledge, this is the first program that allows users to identify
248 orphan genes and TRGs using both nucleotide and protein sequences; and from ORFanID, any taxon
249 of interest found in the NCBI databases can be classified.

250 Taken together with its interactive user interface, ORFanID allows users to seamlessly use the
251 BLAST database to investigate and categorically identify orphan genes. These findings can lead to the
252 creation of databases of these clandestine genes pre-identified at the various taxonomic levels as
253 ascertained by ORFanID, leading to a much deeper understanding of the purpose of orphan genes, their
254 function in genomes, and their potential impact on life.

255

256

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 at: (please insert the link when ready)

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, A., Glover, N., Train, C.M., Kaleb, K., Vesztrocy, A.W., Dylus, D., Farias, T., Zile, K., Stevenson, C., Long, J., Redestig, H., Gonnet, G.H., Dessimoz, C. 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

Buchfink, B., Xie, C., Huson, D. 2015. Fast and Sensitive Protein Alignment using DIAMOND. Nature Methods 12: 59–60. doi:10.1038/nmeth.3176

Domazet-Lošo, T., Brajković, J., Tautz, D. 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

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

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

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

Gao, Q., Xiu, J., Xia, E., Wu, X., Gu, L., Yan, H., Xia, Y., Li, S. 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., Arnaboldi, T., Cain, S., Chan, J., Chen, W., Cho, J., Davis, P., Gao, S., Grove, C.A., Kishore, R., Lee, R.Y.N., Muller, H.M., Nakamura, C., Nuin, P., Paulini, M., Raciti, D., Rodgers, F.H., Russell,

307 M., Schindelman, G., Auken, K.V., Wang, Q., Williams, G., Wright, A.J., Yook, K., Howe, K.L.,
308 Schedl, T., Stein, L., Sternberg, P.W. 2020. WormBase: a modern Model Organism Information
309 Resource. Nucleic Acids Research 48 (D1): D762–D767. doi:10.1093/nar/gkz920
310 Ibrahim, A., Colson, P., Merhej, V., Zgheib, R., Maatouk, M., Naud, S., Bittar, F., Raoult, D. 2021.
311 Rhizomal Reclassification of Living Organisms. International Journal of Molecular Sciences 22 (11):
312 5643. doi:10.3390/ijms22115643

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

315 Johnson, B. 2018. Taxonomically Restricted Genes Are Fundamental to Biology and
316 Evolution. Frontiers in Genetics (Evolutionary and Population Genetics) 9: 407.
317 doi:10.3389/fgene.2018.00407
318 Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper,
319 A., Markowitz, S., Duran, C., Thierer, T., Ashton, B, Meintjes, P., Drummond, A. 2012. Geneious
320 Basic: An integrated and extendable desktop software platform for the organization and analysis of
321 sequence data. Bioinformatics 28 (12): 1647– 1649. doi:10.1093/bioinformatics/bts199

322 Khalturin, K., Georg, H., Sebastian, F., René, A., Bosch, T.C.G. 2009. More than just orphans: are
323 taxonomically-restricted genes important in evolution? Trends in Genetics 25 (9): 404-413.
324 doi:10.1016/j.tig.2009.07.006

325 Kouprina, N., Noskov, V., Pavlicek, A., Collins, K., Bortz, P., Ottolenghi, C., Loukinov, D.,
326 Goldsmith, P., Risinger, J.I., Kim, J.H., Westbrook, V.A., Solomon, G., Sounders, H., Herr, J.C., Jurka,
327 J., Lobanenkov, V., Schlessinger, D., Larionov, V. 2007. Evolutionary Diversification of SPANX-N
328 Sperm Protein Gene Structure and Expression. PLOS ONE. doi:10.1371/journal.pone.0000359

329 Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A., Zdobnov, E.M.
330 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral
331 genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Research 8 (47):
332 D807D811. doi:10.1093/nar/gky1053

333 Li, L., Foster, C.M., Gan, Q., Nettleton, D., James, M.G., Myers, A.M., Wurtele, E.S. 2009.
334 Identification of the novel protein QQS as a component of the starch metabolic network in
335 Arabidopsis leaves. Plant Journal 58 (3): 485-498. doi:10.1111/j.1365-313X.2009.03793.x

336 Li, L., Stoeckert, C.J. Jr, Roos, D.S. 2003. OrthoMCL: Identification of Ortholog Groups for
337 Eukaryotic Genomes. Genome Research (Cold Spring Harbor Laboratory Press) (13):
338 2178-2189. doi:10.1101/gr.1224503

339 Lupas, A.N., Ponting, C.P., Russell, R.B. 2001. On the evolution of protein folds: are similar motifs
340 in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?
341 Journal of Structural Biology 134 (2-3): 191-203. doi:10.1006/jsbi.2001.4393

342 Mirdita, M., Steinegger, M., Söding, J. 2019. MMseqs2 desktop and local web server app for fast,
343 interactive sequence searches. Bioinformatics 35 (16): 2856–2858.
344 doi:10.1093/bioinformatics/bty1057

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

347 Priyam, A., Woodcroft, B.J., Rai, V., Moghul, I., Munagala, A., Ter, F.,
348 Chowdhary, H., Pieniak, I., Maynard, L.J., Gibbins, M.A., Moon, H.K., Davis-Richardson, A., Uludag,
349 M., Watson-Haigh, N.S., Challis, R., Nakamura, H., Favreau, E., Gómez, E.A., Pluskal, T., Leonard,
350 G., Rumpf, W., Wurm, Y. 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, M., Wenguang, Z., Zhao, X., Hohenstein, J., Kandel, Y., O'Conner, S., Wang, Y., Du, C., Nettleton, D., MacIntosh, G.C., Tylka, G.L., Wurtele, E.S., Whitham, S.A., Li, L. 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, J.A., Jones, C.D. 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. 2009. 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., Domazet-Lošo, T. 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, C.M., Murray, A.W., Eddy, S.R. 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., Gadau, J., Simola, D.F., Helmkampf, M., Bornberg-Bauer, E. 2013. Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes. Genome biology and evolution 5 (2): 439–455. doi:10.1093/gbe/evt009

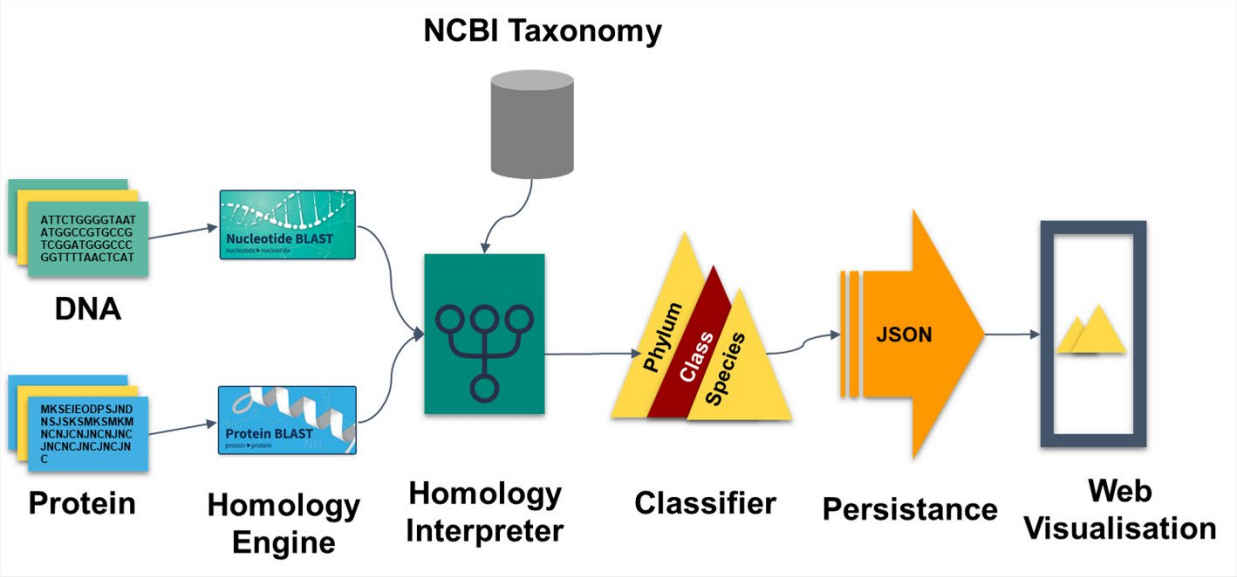
Yin, Y., Fischer, D. 2006. On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. BMC Evolutionary Biology 6 (63). doi:10.1186/1471-2148-6-63

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

374

375

376 **Figures**



378 **Figure 1:** Core engine of ORFanID

379 ORFanID is a web-based application developed on Java Spring Framework. The Web UI was

380 developed in Thymeleaf and MaterializedCSS, which follows Google Material design concepts.

381

A

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	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.

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

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	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 sequences.

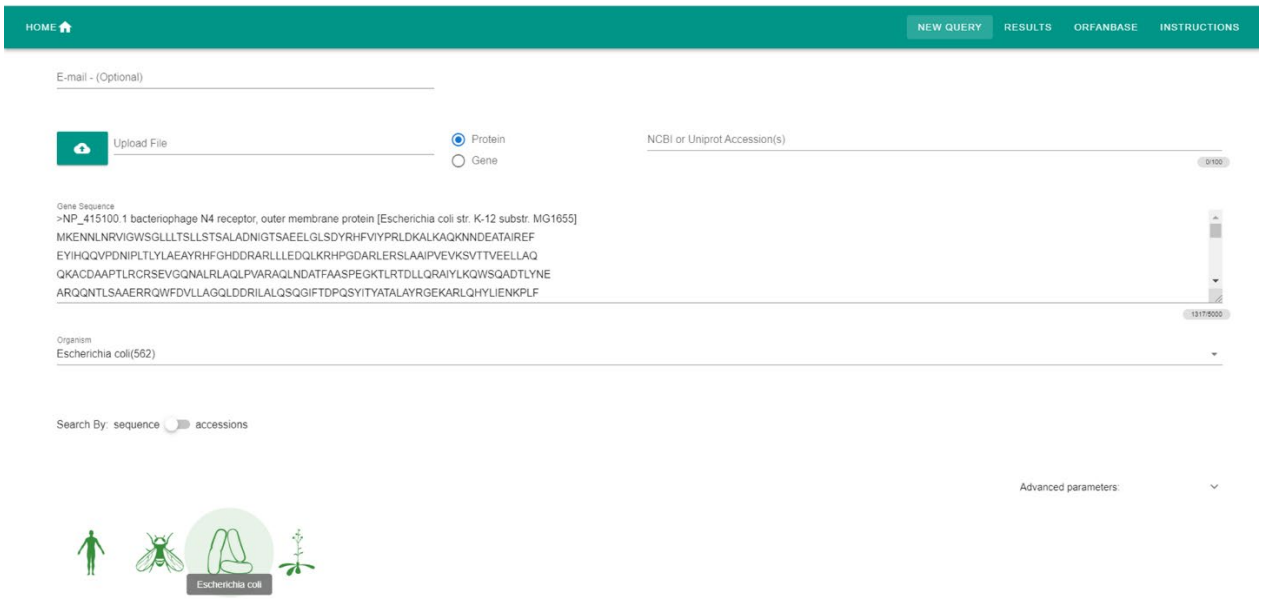


Figure 4: ORFanID Input Page.

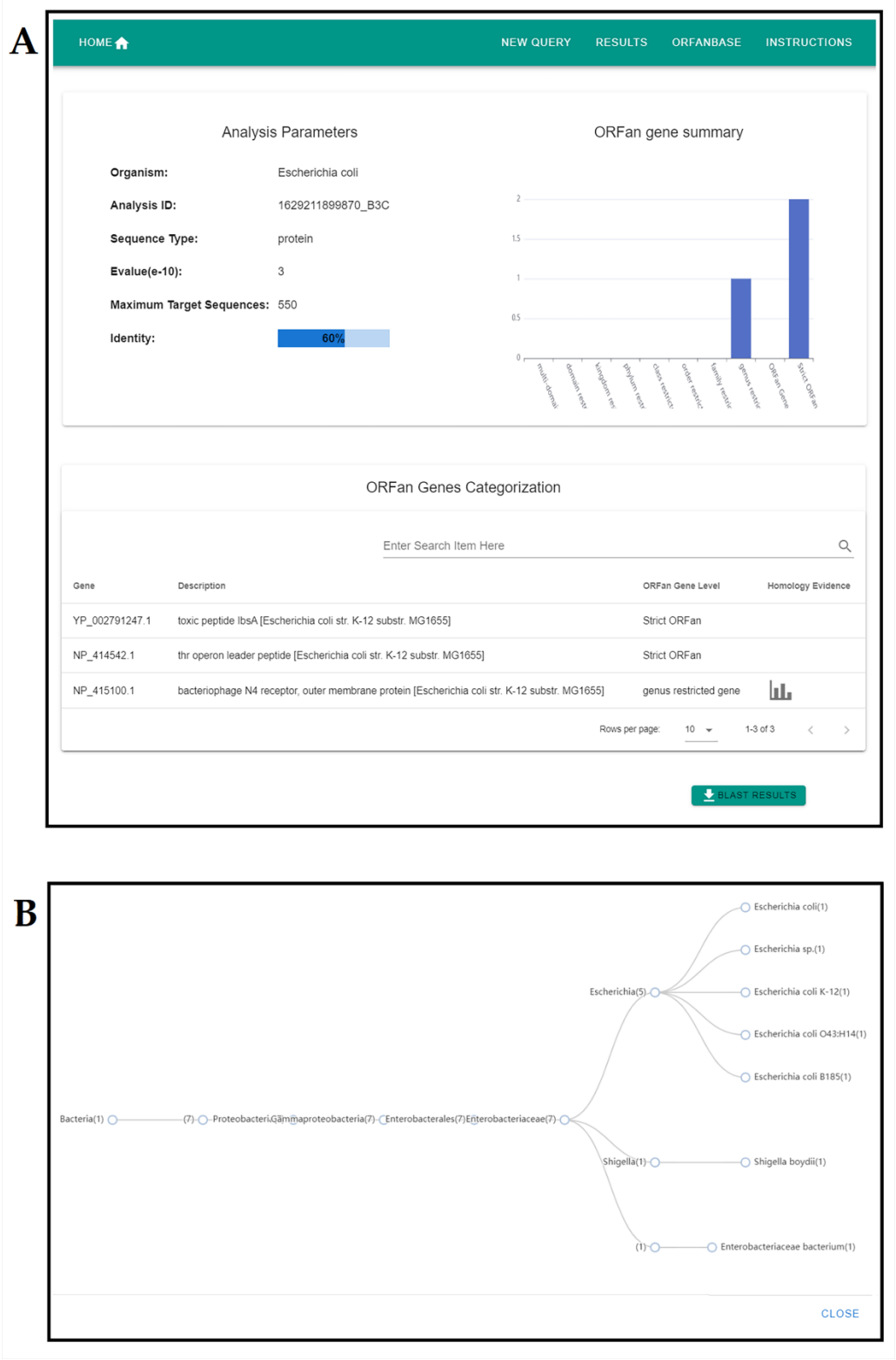


Figure 5: (A) ORFanID Results Page. (B) ORFanID BLAST Results Graphical Display.