ORTHOREFINE: IMPROVING IDENTIFICATION OF ORTHOLOGOUS GENES

by

JOHN LUDWIG

(Under the Direction of Jan Mrázek)

ABSTRACT

Identifying orthologous genes continues to be an early and imperative step in genome analysis but remains a challenging problem. While synteny (conservation of gene order) has previously been used independently and in combination with other methods to identify orthologs, applying synteny in ortholog identification has yet to be automated in a user-friendly manner. This desire for automation and ease-of-use led me to develop OrthoRefine, a standalone program that uses synteny to improve ortholog identification. OrthoRefine implements a look-around window approach to detect synteny, which is used to distinguish orthologs from paralogs in situations where other methods cannot separate paralogs from orthologs reliably. OrthoRefine, applied as a postprocessing step to results obtained with other methods, was tested in tandem with OrthoFinder, one of the most used software for identification of orthologs in recent years, and OMA, an online database of orthologous genes. I evaluated improvements provided by OrthoRefine in several datasets comprised of bacterial, eukaryotic, and archaeal genomes. OrthoRefine efficiently eliminates paralogs from orthologous groups detected by OrthoFinder and those obtained from OMA. Using synteny increased specificity and functional ortholog identification; additionally, analysis of BLAST e-values, phylogenetics, and operon occurrence further supported using synteny for ortholog identification. A comparison of several window

sizes suggested that smaller window sizes (eight genes) were generally the most suitable for identifying orthologs via synteny. However, larger windows (30 genes) performed better in datasets containing less closely related genomes. A typical run of OrthoRefine with ~10 bacterial genomes can be completed in a few minutes on a regular desktop PC.

INDEX WORDS: Ortholog, synteny, homolog, paralog, orthogroup, bioinformatics

ORTHOREFINE: IMPROVING IDENTIFICATION OF ORTHOLOGOUS GENES

by

JOHN LUDWIG

BS, Georgia Southern University, 2013

MS, Georgia Southern University, 2015

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSPHY

ATHENS, GEORGIA

2024

ORTHOREFINE: IMPROVING IDENTIFICATION OF ORTHOLOGOUS GENES

by

JOHN LUDWIG

Major Professor:	Jan Mrázek
Committee:	Ellen Neidle
	Anna Karls
	Liang Liu

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2024

DEDICATION

This one, with love, is for Claire.

ACKNOWLEDGEMENTS

To Jan, we thank for his eternal patience and understanding while mentoring. To the committee members, Liang, Anna, and Ellen, your resilience and support has been appreciated. To Lorenza, thank you for starting our scientific careers (the food was good too).

To Father and Mother, we would not have had the means to finish our academic years as we pleased without your support. To Sister, we've made you proud.

To our friends we've made along the way – Paula, Cynthia, Eduardo, Dakota & Alyssa, Gant, and Brandan – Thnks fr th Mmrs ♪. To the remaining members of Doritos Dust Dungeon, we will forever cherish our time spent together in the war of the BlackRock and the tranquility we forged Between Two Rocks. To the remaining members of the Drowned Fish, may the odds be ever in your favor when Gambit doesn't ask questions. To Cynthia, Paula, Gant, and Brandan – a special remembrance for your assistance on our finest day, as there were Two Rings to bring us all, and in the august of Dublin's September bind us.

To my best friend, I still draw authority from the love felt that night as was proclaimed in the shadow of the Moon's totality when dusk joined us that midday.

To those that will inherit from Syd, you are heralded in the stars as the impending inspiration – for The Dragon has many heads.

To the unenumerated, a sincere thanks for the perfect moments had, but not preserved.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Discerning the evolutionary relationship between genes and genomes remains a fundamental step in the search for answers to many biological questions. The need for accurate assessment of evolutionary relationships among genes is inherent to many different tasks: whether the goal is the construction of phylogenetic trees [1], using orthologs to infer the unknown function of a hypothetical gene [2-5], constructing databases for functional and comparative genomics [6, 7], verifying genome assemblies [8], or understanding the principles of genome organization and evolution [9-11]. However, the accurate identification of evolutionary relationships between genes can be confounded by evolutionary events, such as gene or genome duplication, gene loss, and gene acquisition via horizontal gene transfer [12]. Furthermore, automating the ortholog identification process is challenging, although not intractable [13].

In the context of gene history and comparisons, genes that evolved by divergence from a shared ancestor are classified as homologs [1, 14]. (There is a story about the changing definition of homolog starting with its introduction to scientific literature in 1848 by Owen; from “same organ in different animals” to the context of genetic inheritance and whether the definition of homolog relies on the definition of analogue [14-18] .) Gene homology can be further divided into three types based on the events that allowed the genes to take different evolutionary paths: 1) orthologs diverged because of a speciation event, 2) paralogs diverged following a gene duplication event [1], and 3) xenologs arose due to a horizontal gene transfer event [19]. To

relate gene duplication events with speciation events, paralogs can be further classified as inparalogs, those that arose from a duplication event after the speciation event, and outparalogs, when the duplication event preceded the speciation event [20].

Paralogs arise from a gene duplication event that creates a second copy of a gene in the genome, which can lead to several different outcomes. The most common result, non-functionalization, is that one of the copies is lost via direct deletion or deleterious mutation [21-23]. Alternatively, both copies of the gene can be retained resulting in either subfunctionalization, neofunctionalization, or superfunctionalization. Subfunctionalization occurs when the duplicated gene(s) becomes more specialized in function [24]. This functional divergence is facilitated by a reduced selective constraint on the duplicate copies of the gene and may take different forms [24, 25]. For example, promiscuous enzymes (those that catalyze the same chemical reaction but on different substrates) can become specialized to one particular substrate, or the genes' regulation is differentiated by placing the copies under the control of different regulatory elements [26]. In neofunctionalization, the reduced selective constraints allow one of the post-duplication genes to evolve a new function while the other gene maintains the original function [27]. In some instances, the organism might benefit from maintaining both copies of the gene without further divergence. This is sometimes referred to as superfunctionalization [28], and the fitness benefit comes from the increased synthesis rate of the gene's product; a typical example is multiple copies of rRNA genes commonly present in genomes of fast-growing bacteria [29-31].

While all genes experience genetic changes, orthologs tend to retain the function of the shared ancestral gene more often than paralogs [25, 32]; this property of orthologs is routinely used to predict functions of genes that are yet to be experimentally characterized. In the early

days of genome sequencing, a gene with an unknown function would be compared to a database of known genes (via pairwise alignments of the protein amino acid sequences) and assigned the function of the closest match [3, 4]. However, this led to a propagation of errors, and modern methods replaced pairwise gene-to-gene comparisons with an ensemble approach, typically utilizing hidden Markov models [33-35].

An additional application of gene orthology is in phylogenetics where, in general, the goal is to elucidate the evolutionary relationship or history between genetic units (genes, proteins, or genomes). Under a strict interpretation of an alignment, the aligned data are expected to have an orthologous relationship [1, 19, 36-39]. Including paralogs in the alignment is thought to reduce the accuracy of a phylogenetic tree derived from the alignments [19, 37, 40]. However, recent research suggests that the negative effect of including paralogs may be less significant than previously thought [41-43]. Another application of gene orthology is in judging the quality of *de novo* genome assemblies by verifying that they have the expected single-copy orthologs present in other related species [8]. Regardless of the application, the core assumption is that orthologous genes tend to retain their function after divergence, which is less likely for paralogs or xenologs.

Current informatics-based methods for the identification of orthologs are rooted in either phylogenetics, as is the case with Ortholuge which compares ratios of phylogenetic distances between ingroups and an outgroup to improve the accuracy of prior ortholog identification [44], or reciprocal best hit (RBH) via BLAST (or another sequence alignment tool) combined with a clustering algorithm – e.g., OrthoMCL performs Markov clustering on RBH results [45]. Other ortholog identification programs include OrthoLoger [46], TreeFam [47], and InParanoid [48]. However, the original software that implemented these methods is generally no longer functional

on present-day computers and operating systems due to dependencies on obsolete versions of software components or the software is challenging to set up and use [49].

An alternative but often not an ideal solution for the identification of orthologs is to use one of several available databases of orthologous genes, such as OrthoDB [50], eggNOG [51], PANTHER [52], or OMA [53]. This solution requires that the database contains information on the genome of interest and that the data from the database is in an accessible format that can be used for analysis. In addition, the databases generally do not allow changing the parameters of the algorithm used to identify the orthologs (e.g., sequence similarity cutoff, clustering parameters, phylogenetic models and their parameters), and the default parameters may not be ideal for different types of studies [54].

OrthoFinder combines reciprocal best-hits with phylogenetics to identify orthologs. The major advantages over other software include that it is user-friendly, easy to install (it does not rely on additional software not included in the installation), and offers increased accuracy for ortholog identification compared to many earlier methods [13, 49]. In OrthoFinder's 2015 paper, the authors noted the possibility of using synteny (conservation of gene order) to refine ortholog identification. However, they chose not to use synteny because reliable syntenic information breaks down over long evolutionary distances, and syntenic information is not immediately available for *de novo* assemblies. I note that for *de novo* assemblies, a genome annotation may be obtained by submitting the data for automated annotation at NCBI, or it can be directly generated by the user using the same pipeline [55].

The term synteny was initially conceived to describe genes linked together during inheritance (chromosome mapping, see [56]) and referred to two or more genes located on the same chromosome [57]. More recently, particularly in the context of prokaryotic genomes, the

term “synteny” has been used to refer to conserved gene order in comparative genomics [5, 58, 59]. This is how I use the term “synteny” in this work.

OrthoFinder is an effective tool that provides results suitable for many tasks. Nonetheless, incorporating synteny into the criteria for identifying orthologs as an additional postprocessing step can enhance the program’s ability to distinguish orthologs from paralogs and further refine some of the HOGs (hierarchical orthogroups) reported by OrthoFinder. Here I present a new program, OrthoRefine (https://github.com/jl02142/OrthoRefine), which automates the task of using synteny information to refine the HOGs identified by OrthoFinder into groups of syntenic orthologs, orthologs grouped based on evidence of synteny. My results below, and other recent work [60], show that ortholog identification via OrthoFinder can be enhanced by using synteny information.

In chapter 2 of this dissertation, I cover a brief overview of OrthoFinder’s methods, OrthoRefine’s algorithm, and runtime parameters. I benchmark OrthoFinder vs. OrthoFinder + OrthoRefine, and I discuss several specific examples of orthogroups that are improved via syntenic information. In chapter 3, I extend OrthoRefine to use the OrthoXML standardized input format and I use OrthoRefine to compare an ortholog database that utilizes the OrthoXML format with OrthoFinder.

CHAPTER 2

OrthoRefine: automated enhancement of prior ortholog identification via synteny.

In this chapter, I implement and test OrthoRefine in combination with OrthoFinder, currently one of the most frequently used tools for ortholog identification. I provide a brief overview of OrthoFinder's methods, describe OrthoRefine's main algorithm for identifying orthologs via synteny, analyze and discuss the effects of OrthoRefine's parameters and make recommendations on their values, benchmark OrthoRefine, and finally investigate and offer discourse on specific examples where synteny improved prior ortholog identification.

Implementation

OrthoFinder Summary

OrthoFinder was initially described in 2015 [49], and updated software and manuscript were released four years later [13]. We used version 2.5.2, downloaded from the GitHub repository on April 6th, 2021 [61]. OrthoFinder begins with pairwise all-against-all alignments of genes in all compared genomes and records the protein pairwise sequence similarity. It then uses the BLAST bit score [62] and normalizes the data so that sequence length does not influence the bit score. Normalization also makes the scores comparable for genomes at various levels of phylogenetic distance. Orthogroups, defined as the set of all genes predicted to be descendants of a single gene of the last common ancestor (this can include orthologs and paralogs) [49], are constructed by first grouping any genes that have a greater normalized similarity score than a cutoff score automatically determined by OrthoFinder. The second step for forming orthogroups,

and the final step of the 2015 version, is clustering the pairwise hits into orthogroups using the MCL algorithm [63].

The 2019 software begins where the 2015 version ended; the 2015 output (the clustering of the orthogroups) is used to create an unrooted gene tree for each orthogroup using DendroBLAST [64]. The unrooted gene trees are forwarded to STAG (Species Tree Inference from All Genes) [65] to infer an unrooted species tree, and the unrooted species tree is rooted by STRIDE (Species Tree Root Inference from Gene Duplication Events) [66]. The rooted species tree is then used to root each unrooted gene tree (orthogroup tree) [13]. To delineate the orthogroups into orthologs and identify duplication events (paralogs), the final step of the 2019 software (pre-version 2.4.0) is a hybrid algorithm that was designed to merge the accuracy of DLCpar [67] with the speed of the species overlap algorithm [68]. The update in version 2.4.0 introduced a new final output in the form of HOGs obtained from the analysis of the rooted gene trees [61].

In summary, the 2019 version uses the orthogroups from the 2015 algorithm and applies phylogenetics to identify orthologs and gene duplication events. However, since OrthoFinder version 2.4.0 of July 2020, the authors have replaced the now deprecated orthogroups with hierarchical orthogroups (HOGs), which are more accurate orthogroups inferred at each level in the species tree. Version 2.5.2 continues to supply the end user with the deprecated orthogroups delineated into orthologs; however, these data are not from the improved HOGs. Additionally, the end user must manually analyze many different data files or write their script(s) to process the data [60]. I desired to 1. use the more accurate HOGs, 2. keep the analysis automated, 3. and enhance the ortholog identification with synteny by refining the HOGs into orthologs and paralogs, which led me to create OrthoRefine.

While there are options and parameter adjustments that may be used when running OrthoFinder (e-value of BLAST, MCL inflation parameter, and phylogenetic parameters, etc.), my focus was on running OrthoFinder with default settings because, in my review of literature citing OrthoFinder, OrthoFinder was generally used with default parameters.

OrthoRefine Methods

By default, OrthoRefine is only applied to HOGs with genes from at least two genomes and at least two genes from the same genome (paralogs), with an option to verify synteny for HOGs that have only a single gene from each genome. The latter may still include paralogs if genes were duplicated and the original copy was subsequently lost, which can be revealed by synteny. The analysis begins by constructing a window centered at each gene of the HOG. OrthoRefine evaluates the synteny by counting matching pairs of genes inside the window; matching pairs consist of genes assigned to the same HOG in the initial OrthoFinder output (Figure 1). I note that genes only need to be within the window and are not required to be in the same order, and genes that do not have a homolog in the other genome are not included in the window (see Figure 7). The synteny ratio, sr, is calculated by taking the number of matching pairs and dividing it by the window size, w (Eqs. 1). If the ratio is greater than a cutoff (default 0.5), the genes at the center of the window are considered syntenic. After a pairwise comparison between all genes of different genomes in the original HOG, any subset of genes linked by synteny is referred to as a syntenic ortholog group (SOG). A HOG can thus be refined into one SOG (by removing paralogs which do not exhibit synteny with any other genes from the original HOG) or into more than one SOG if the original HOG contained multiple distinct subgroups of genes linked by synteny within each subgroup but not between the subgroups.

$$x_i = \begin{cases} 1\ if match \\ 0\ if no\ match \end{cases}, \quad sr = \frac{\sum_1^w x_i}{w} \qquad (1)$$

where sr is the synteny ratio, w is the window size, and i is the serial number of the gene within the window. The gene being evaluated for synteny (at the center of the window) is not counted.

I expect OrthoRefine to be used primarily in tasks that would benefit from resolving orthologous relationships to no more than a single ortholog from each compared genome in each orthologous group. OrthoRefine was designed to emulate several of the qualities that make OrthoFinder a desirable tool for the end user: speed, ease-of-use, and self-containment (no dependencies); OrthoRefine requires only the output from OrthoFinder (or any orthogroup file formatted to match OrthoFinder's format; see Chapter 3) and genome annotations (in the RefSeq features table format used by NCBI) and does not depend on any other software or data that could complicate its use. The only input required to be created by the end user is a text file where each line specifies the RefSeq accession for each genome used as input for OrthoFinder.

OrthoRefine Parameters

Two runtime parameters control OrthoRefine, window size and synteny ratio. There is no consensus on the required amount of synteny - how many surrounding genes in a window must be orthologs to conclude that the gene of interest is an ortholog – or the size of the window. I recommend a smaller window size and larger synteny ratio when analyzing datasets containing closely related genomes, e.g., window size eight and synteny ratio 0.5. In contrast, a larger window size and or a lower synteny ratio may be more appropriate as the evolutionary distance increases, e.g., window size 30 and synteny ratio 0.2. After testing several combinations of window size and synteny ratio, I selected a window size of eight and synteny ratio of 0.5 as default parameters (see Results and Discussion for details).

OrthoRefine Example

I demonstrate the use of OrthoRefine on HOG 19 from the OrthoFinder output for representative genomes of four different species of the *Escherichia* genus: *Escherichia coli* strain K12 substrain MG1655 (NCBI genome assembly GCF_000005845.2), *Escherichia fergusonii* (GCF_013892435.1), *Escherichia albertii* (GCF_016904755.1), and *Escherichia marmotae* (GCF_902709585.1). OrthoFinder, with default parameters, was used to identify HOGs, and OrthoRefine was subsequently applied with window size eight and synteny ratio cutoff 0.5.

The HOG included four *E. coli* genes (b0652, b3271, b4106, & b4096), fives genes from *E. fergusonii* (HVX45_RS09410, HVX45_RS02390, HVX45_RS04025, HVX45_RS07420, & HVX45_RS11505), two genes from *E. marmotae* (GV529_RS14465 & GV529_RS05870), and one gene from *E. albertii* (JRC41_RS15115). Most of the genes were annotated as encoding ATP-binding cassette (ABC) transporters. Figure 1 shows how OrthoRefine determined which of b3271 or b0652 of *E. coli* is the ortholog of RS11505 of *E. fergusonii.* As eight out of eight genes surrounding RS11505 had a match in the window centered at b3271, I concluded that there is a syntenic relationship between *E. coli's* b3271 and *E. fergusonii's* RS11505, and they are orthologs while b0652 is presumed to be a paralog of RS11505; none of the genes surrounding b0562 had a match within the window around RS11505. This HOG was ultimately refined into two SOGs: the first included a single syntenous ortholog from each genome and the second SOG contained a single syntenous ortholog from *E. coli*, *E. fergusonii*, and *E. marmotae*. The remaining genes initially placed in HOG 19 by OrthoFinder were excluded by OrthoRefine as putative paralogs (see Results and Discussion for details).

Results and Discussion

Datasets Used to Evaluate OrthoRefine's Performance

I analyzed several datasets including taxa of different levels of divergence. The first dataset, Quest for Orthologs [69], included 23 diverse bacterial genomes used to test OrthoRefine using the community standard benchmarking tool [70]. The second dataset was the four *Escherichia* species detailed above. The third dataset comprising four Gammaproteobacteria – *E. coli, Klebsiella pneumoniae, Salmonella enterica,* and *Pseudomonas aeruginosa* – was adopted from a prior study [71]. The fourth dataset was a collection of sixteen members of the phylum Actinomycetota. The fifth dataset was used to test OrthoRefine's performance with eukaryotic genomes; three *Saccharomyces* genomes were selected: *Saccharomyces mikatae, Saccharomyces cerevisiae,* and *Saccharomyces kudriavzevii*.

Benchmarking OrthoRefine

Orthology Benchmarking [70], a web-based benchmarking tool, was used to evaluate OrthoRefine's ability to improve functional ortholog identification (gene ontology conservation (GO) & enzyme classification (EC); [72]) and specificity (Robinson-Foulds (RF) distance; [72]). Because the RefSeq annotations and sequences corresponding to the date when the benchmarking data were generated (2020) are no longer available on the NCBI website (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/), I was unable to use precisely the same collection of proteins that was used for the benchmarking (the annotations are required to determine gene location to assess synteny) and my results are not directly comparable to the data on the benchmarking server. However, I utilized the benchmarking tool to compare OrthoFinder and OrthoRefine results using a dataset that was composed of the same 23 bacterial genomes (Table 1) as the benchmarking dataset but with current annotations and protein sequences

downloaded in August 2023 (See Appendix A for details on generating the dataset). This dataset includes genomes spanning diverse bacterial phyla and the large evolutionary distances among the genomes makes synteny less effective, providing a stringent test for OrthoRefine.

As expected, OrthoRefine increased functional ortholog identification and specificity accuracy compared to OrthoFinder alone. Of the combinations for window size and synteny ratio I tested, window size ten and synteny ratio 0.5 resulted in the lowest RF distance (conceptually defined as a normalized sum of differences between the benchmark orthogroups and the user proposed orthogroups; a lower score indicates higher specificity). The highest average Schlicker score [73] (a measure of the overall mutual similarity of the function classifications of the orthologs) for the GO terms was observed at window size 40 and synteny ratio 0.5, while the highest average Schlicker score for EC was recorded for window sizes 20-40 at a synteny ratio of 0.5 (Figure 2).

Evaluating OrthoRefine's Runtime Parameters

In the absence of a gold standard dataset of true orthologs to assess OrthoRefine's accuracy, I consider that, for most applications, the most desirable set of orthologs would include genes from the maximum number of the genomes analyzed (high sensitivity) while containing no paralogs (i.e., a single gene from each genome; high specificity). I, therefore, evaluated OrthoRefine's output for the average maximum number of orthologous genes (AMNOG) defined as the dataset's average maximum number of orthologs present in SOGs without paralogs. For HOGs that were refined into multiple SOGs, the SOG with the most genomes represented is included in the AMNOG calculation. I propose to use the AMNOG as a relative measure of sensitivity while specificity is, in theory, at or near 100% (SOGs containing paralogs are excluded). In general, changing parameters did not dramatically change the AMNOG measure. I

observed that larger windows and lower synteny ratio performed slightly better in datasets consisting of diverse genomes. For datasets of closely related genomes, smaller window sizes performed at least equally as well as larger window sizes and within those smaller window sizes, a larger synteny ratio tended to result in a higher AMNOG (Table 2).

Testing OrthoRefine on Datasets of Varying Phylogenetic Diversity

Escherichia Dataset

I evaluated OrthoRefine on four species of the *Escherichia* genus. The close relationship among the genomes was expected to make accurate identification of orthologs easier for OrthoFinder and OrthoRefine due to the low divergence between orthologous gene sequences and the high conservation of gene order. Indeed, 64% of OrthoFinder's HOGs were comprised of precisely one gene from each genome (1-to-1 HOGs), while 25% of the HOGs were missing an ortholog in at least one genome but included no more than one gene per genome (0-or-1 HOGs). 11% of HOGs combined orthologs and paralogs (at least one genome contributed more than one gene to the same HOG). OrthoRefine modified 87% of these paralogous HOGs (synteny eliminated at least one paralog and or divided the HOG into at least one SOG); the remaining 13% were split between either confirmed (all genes assigned by OrthoFinder to a HOG were supported by synteny) (3%) or unconfirmed (insufficient synteny support for the original HOG or any SOG subgroup) (10%). Additionally, OrthoRefine confirmed 97% of the 1-to-1 HOGs and 88% of the 0-or-1 HOGs (Figure 3).

Case study 1: HOG19 – (ABC transporters)

HOG19, identified by OrthoFinder, contained a group of genes encoding ATP-binding cassette (ABC) transporters. ABC transporters are easily identifiable by the presence of the distinctive ATP-binding domain, but their further classification remains challenging, in part, due

to the vast diversity of substrates they can transfer and the often subtle differences that can affect substrate specificity [74].

OrthoRefine divided this HOG into two subgroups: SOG19.0 which consisted of b0652, JRC41_RS15115, HVX45_RS07420, & GV529_RS05870 and SOG19.1 which consisted of b3271, HVX45_RS11505, & GV529_RS14465, while excluding b4106, b4096, HVX45_RS02390, HVX45_RS04205, & HVX45_RS09410 as presumed paralogs (Figure 4).

The BLAST e-values and percent identity from OrthoFinder's alignment supported the presence of these two natural subgroups in HOG19 (Table 3), as did the phylogenetic tree made independently of OrthoFinder (Figure 5) using Muscle, version 5 [75], RAxML, version 8.2.12 [76], R, version 4.1.2 [77], and the R library ape, version 5.6-2 [78] (See Appendix B for commands used). Previously reported operon structures further supported the division of HOG19 by OrthoRefine (Figure 6). b0652, annotated as encoding glutamate/aspartate ABC transporter ATP subunit *gltl*, is a member of the *gltIJKL* operon [79]. All members of SOG19.0 (b0652, JRC41_RS15115, HVX45_RS07420, & GV529_RS05870) were in the same operon in their respective genomes. Similarly, all members of SOG19.1 (b3271, HVX45_RS11505, & GV529_RS14465) were in the previously reported *yhdWXYZ* operon in their respective genomes [80, 81]. Members of both operons encode products of the same length and the second and third sub-units (J/X & K/Y) followed the previously reported pattern of the first three of the more specific substrate-binding sub-units being about 30% smaller than their non-specific counterparts [81]. The exceptions, *gltI* & *yhdW*, were probably due to a frameshift in the *yhd* operon [80, 81].

Case study 2: HOG21 – (*rpnA*/*rpnE* homologs)

HOG 21 comprised eight genes from the four genomes; the genes were annotated as encoding recombination-promoting nuclease (*rpnA*), *rpnE*, insertion sequence family not

classified yet (ISNCY), or hypothetical protein. The recombination-promoting nucleases are thought to be involved with horizontal gene transfer, though RpnE was inactivein recombination-determining assays [82].

Similar to HOG19 above, OrthoRefine divided HOG21 into two subgroups: SOG21.0 – the *rpnE* and ISNCY group, which consisted of b2244, HVX_RS21485, GV529_RS12150, & JRC41_RS07400 and SOG21.1 – the *rpnA* group, which consisted of b3411, HVX45_RS12120, & GV529_RS10930 while excluding HVX45_RS22925 as a presumed paralog (Figure 7). The BLAST e-values and percent identity supported dividing HOG21 into these two natural subgroups (Table 4); additionally, the phylogenetic tree agreed with the two subgroups but included HVX45_RS22925 in the *rpnA* group (Figure 8). The HVX45_RS22925 gene encodes a short (68 amino acids) hypothetical protein similar to the C-terminal segment of RpnA, which is much larger (292 amino acids in *E. coli*). The similarity probably leads to this hypothetical protein being included in HOG21 by OrthoFinder, but its short length suggests that it is not a true ortholog of RpnA - if it is a functional protein at all.

Gammaproteobacteria Dataset (Lim et al. 2022)

In their publication, the authors analyzed the dataset (Table 5) with OrthoFinder and highlighted a specific HOG which contained paralogs; I analyzed the same genomes with OrthoFinder and OrthoRefine to resolve the paralog HOG to a 1-to-1 relationship. I observed a lower percentage of 1-to-1 HOGs (27%) than in the *Escherichia* dataset, presumably due to the larger evolutionary distance among the genomes. The percentage of 0-or-1 HOGs (43%) and HOGs with paralogs (30%) increased. When using OrthoRefine to process OrthoFinder results, I observed a lower percentage of 1-to-1 HOGs (34%) and 0-or-1 HOGs (63%) confirmed by synteny and generally higher number of HOGs that were modified by OrthoRefine (Figure 9).

Case study 3: HOG346 - (*sdiA*)

sdiA encodes a LuxR family transcription factor and is thought to regulate transcription of cell division genes [83] and genes involved in acid tolerance [84]. OrthoFinder included potential paralogs in this HOG in the original analysis by Lim et al. ([71] Figure 2 C & D) and the same genes were included in this HOG in my results when we used OrthoFinder without OrthoRefine. *E. coli*, *S. enterica*, and *K. pneumoniae* each contributed one gene to HOG 346, while *P. aeruginosa* contributed four genes. OrthoRefine, with window size eight and synteny ratio 0.5, did not resolve HOG346; none of the four *P. aeruginosa* homologs could be classified as syntenous due to a lack of matches within the window. However, this HOG was resolved with the parameters I recommend for more distantly related genomes - window size 30 and synteny ratio 0.2, which identified the *P. aeruginosa* gene AFI95_RS32400 (transcription regulator *luxR* family) as the ortholog of *sdiA* in *E. coli*, *S. enterica*, and, by proxy, *K. pneumoniae* (Table 6).

I speculate that the lack of synteny between the *K. pneumoniae* gene and any of the four genes of *P. aeruginosa* could stem from the fact that the syntenous genes between *E. coli*, *S. enterica*, and *P. aeruginosa* were motility genes, whereas *K. pneumoniae* is non-motile [85], and therefore not expected to contain these genes. Nevertheless, this conclusion is apparent only from the synteny analysis, whereas neither sequence similarity (Table 7) nor the phylogenetic tree (Figure 10) could differentiate an ortholog from the paralogs in the *P. aeruginosa* genome.

Actinomycetota Dataset

This dataset of sixteen arbitrarily selected genomes from the phylum Actinomycetota (Table 8) further increased the evolutionary distances among the analyzed genomes. The Actinomycetota are classified based on 16S [86], 23S rRNA, and distinct indels [87] and are the "high" G+C division of gram-positive bacteria while Bacillota, formerly Firmicutes, are the

"low" G+C gram-positive bacteria [88]; rarely, Bacillota may become gram-negative later in the life cycle [89]. Of the HOGs identified by OrthoFinder, only 2% were 1-to-1, 60% were 0-or-1, and 38% were paralogous. OrthoRefine modified 65% of the HOGs with paralogs; additionally, OrthoRefine confirmed 49% and 32% of the 1-to-1 and 0-or-1 HOGs (Figure 11).

Case study 4: HOG 402 – (PknB)

HOG 402 is comprised of 23 genes from the sixteen species, which are all annotated as encoding proteins of the kinase B (PknB) family – which contains penicillin-binding proteins (PBP) and serine/threonine kinases (STKP) characterized by the presence of a serine/threonine kinase-associated domain (PASTA). The *pknB* gene is essential in *Mycobacterium tuberculosis* [90, 91], where it controls cell division and cell wall synthesis [92]; however, *pknB* was found to be not essential in *Streptomyces coelicolor,* where it is thought to be involved in the development cycle and antibiotic production [93]. In PBPs, the function of the PASTA domain appears to be species specific [94], and there is a lack of consensus on its exact function [95]. In STKPs, the PASTA domain is thought to bind peptidoglycan and β lactam (penicillin group antibiotics) [96].

OrthoRefine split the HOG into four SOGs (Figure 12). SOG 402.0 contained the genes from *O. timonensis*, *O. uli*, *D. detoxificans*, *C. curtum*, and *E. lenta*. SOG 402.1 included two genes from *A. ferrooxidans* and one member each from *E. rhizosphaerae*, *E. halophilus*, *A. cellulolyticus*, *S. fradiae*, *S. avermitilis*, and *S. griseus;* the two genes from *A. ferrooxidans* are in tandem next to each other, which prevents them from being differentiated by synteny. SOG 402.2 contained similar pairs of tandem paralogs from the *Streptomyces* genera. SOG 402.3 included genes from *R. tropicus* and *R. marinus*, whereas the genes from *R. xylanophilus* and *A. oris* lacked the required synteny to be assigned to any SOG. The phylogenetic tree built

independently of OrthoFinder could not delineate the *Streptomyces* orthologs and paralogs (Figure 13).

The members of SOG 402.0 and SOG 402.1 were identified as members of a previously identified operon from *Mycobacterium* [91, 92, 97] and *Streptomyces* [94]. The members of SOG 402.2 are not organized in the same operon, which provided further evidence for placing these genes into their own SOG. The operons for members of SOG 402.1 were consistently found to have a gene encoding a STPK adjacent to a gene encoding a PBP, *ftsW*, *stp1*, and a gene encoding a forkhead-associated (FHA) domain; members of SOG 402.0 had a similar arrangement except the gene encoding the PBP and *ftsW* were fused (Figure 12). I could not detect an analogous operon containing the genes included in HOG402 from the *Rubrobacter* genomes. However, a manual inspection of the annotation for *R. xylanophilus, R. tropicus*, & *R. marinus* reveals the operon not with the *Rubrobacter* genes assigned to HOG402 but rather with those assigned to HOG401 (RXYL_RS00115, GBA63_RS00140, & GBA65_RS00120). I also detected a gene fusion or split - which has previously been shown to reduce the accuracy of ortholog identification [98] - between members of SOG 402.0 and SOG 402.1, which would explain why OrthoRefine split these groups into their own SOGs instead of combining them into a single SOG. Additionally, *A. oris* has an additional gene in its operon that was not present in SOG 402.0 or 402.1, which would explain why its gene failed to be grouped with any SOG.

In most *Streptomyces,* a gene encoding a STPK with four PASTA domains is positioned next to a gene encoding a PBP without the PASTA domain [95]. The *Streptomyces* proteins encoded by the genes assigned to SOG402.1 (SAVERM_RS22430, SGR_RS18440, & CP974_RS14705) have four PASTA domains, while the tandem pairs of *Streptomyces* genes in SOG402.2 encode proteins that have one PASTA domain. The genes from *Rubrobacter* revealed

from the manual inspection (HOG 401) encode a STPK protein with four PASTA domains and those genes are adjacent to a gene that encodes a PBP, providing further confidence that these genes are the orthologs of SOG 402.1 and not the original *Rubrobacter* genes assigned to HOG 402.

***Saccharomyces* Dataset**

I evaluated three Saccharomyces genomes for orthologs to test OrthoRefine's performance on eukaryotic genomes: *S. mikatae* (GCF_947241705.1), *S. cerevisiae* (GCF_000146045.2), and *S. kudriavzevii* (GCF_947243775.1). As expected, due to the small evolutionary distance between the three *Saccharomyces* genomes, 95% of OrthoFinder's HOGs were 1-to-1, 2% were 0-or-1, and 3% were paralogous. OrthoRefine modified 82% of the paralogous HOGs and confirmed 99% of 1-to-1 HOGS and 85% of the 0-or-1 HOGs (Figure 14).

Case Study 5: HOG 55 - (glyceraldehyde-3-phosphate dehydrogenase)

HOG 55 is composed of two genes from each of the three genomes - annotated as encoding glyceraldehyde-3-phosphate dehydrogenase, either *tdh2* or *tdh3* - which are known paralogs [99]. OrthoRefine split the HOG into two SOGs, correctly separating the members of the two groups: SOG55.0, the *tdh2* group, was comprised of SMKI_10G2100, YJR009C, & SKDI_10G2170, whereas SOG55.1, the *tdh3* group, was comprised of SMKI_16G0680, YGR192C, & SKDI_07G4440 (Figure 15). The BLAST e-value and percent identity mostly agreed with these groupings (Table 9); however, SKDI_07G4440 was the best match for both *S. cerevisiae* genes, which led to a failure in correct ortholog assignment for the two paralogs in *S. cerevisiae* based on sequence similarity alone. It has previously been reported that orthologs sometimes have a lower percent identity than their paralogs [97, 100]. This result shows that synteny can, at least in some instances, resolve such discrepancies in sequence divergence. The

phylogenetic analysis mostly supported the synteny groupings; however, similar to the BLAST e-values, there was a lack of support to tell where to group the genes from *S. kudriavzevii* (Figure 16).

Table 1. Names and RefSeq accessions for the 23 genomes from the Quest for Orthologs.

Genius species	RefSeq accession
Mycobacterium tuberculosis	GCF_000195955.2
Pseudomonas aeruginosa	GCF_000006765.1
Thermotoga maritima	GCF_000008545.1
Chlamydia trachomatis	GCF_000008725.1
Streptomyces coelicolor	GCF_000203835.1
Leptospira interrogans	GCF_000092565.1
Escherichia coli	GCF_000005845.2
Neisseria meningitidis	GCF_000008805.1
Deinococcus radiodurans	GCF_000008565.1
Bradyrhizobium diazoefficiens	GCF_000011365.1
Synechocystis	GCF_000009725.1
Chloroflexus aurantiacus	GCF_000018865.1
Bacillus subtilis	GCF_000009045.1
Gloeobacter violaceus	GCF_000011385.1
Aquifex aeolicus	GCF_000008625.1
Helicobacter pylori	GCF_000008525.1
Fusobacterium nucleatum	GCF_000007325.1
Rhodopirellula baltica	GCF_000196115.1
Geobacter sulfurreducens	GCF_000007985.2
Mycoplasma genitalium	GCF_000027325.1
Dictyoglomus turgidum	GCF_000021645.1

Bacteroides thetaiotaomicron	GCF_000011065.1
Thermodesulfovibrio yellowstonii	GCF_000020985.1

Table *2*. Combinations of window size and synteny ratio on the AMNOG.

		4 *Escherichia*	4 Gammaproteo-bacteria	16 Actino-mycetota	3 *Saccharomyces*	
Window size	Synteny ratio	Average max number orthologous genes (AMNOG)				Average AMNOG
2	0.5	3.16	2.52	3.2	2.3	2.8
4	0.25	3.15	2.54	3.21	2.33	2.81
4	0.5	3.2	2.57	3.41	2.44	2.91
6	0.2	3.17	2.6	3.41	2.32	2.88
6	0.5	3.19	2.59	3.38	2.55	2.93
8	0.2	3.19	2.57	3.42	2.41	2.9
8	0.3	3.2	2.57	3.44	2.54	2.94
8	0.5	3.24	2.57	3.39	2.59	2.95
10	0.2	3.22	2.58	3.4	2.39	2.9
10	0.3	3.22	2.57	3.45	2.6	2.96
10	0.5	3.23	2.56	3.34	2.67	2.95
30	0.2	3.22	2.6	3.4	2.72	2.99
30	0.3	3.23	2.56	3.31	2.73	2.96
30	0.5	3.17	2.54	3.02	2.68	2.85
40	0.2	3.2	2.59	3.44	2.73	2.99
40	0.3	3.21	2.57	3.29	2.69	2.94
40	0.5	3.18	2.53	2.9	2.7	2.83

Table 3. BLAST e-values and percent identity for HOG 19.
BLAST e-values and percent identity, reported by OrthoFinder, for two genes from *E. coli*, b0652 & b3271, as BLASTed against the other members of HOG19 from *E. albertii*, *E. fergusonii*, and *E. marmotae*. Bolded values are the lowest e-value and highest percent identity from b0652 or b327. b4106 & b4096 were ommitted due to high e-value (1.6e-35 & 2.6e-18) and low percent identity (39.1 & 31.2). HVX45_RS02390, HVX45_RS04025, & HVX45_RS09410 were omitted for the same reason (best e-value with E. coli. = 5.9e-56, best percent identity = 51.08).

HOG 19				
E. coli b0652			*E. coli* b3271	
e-value	% identity	Gene to be BLAST against	e-value	% identity
1.0e-131	**98.8**	JRC41_RS15115 (*E. albertii*)	2.1e-82	60.6
4.4e-133	**100.0**	HVX45_RS07420 (*E. fergusonii*)	2.8e-82	60.2
4.2e-133	**100.0**	GV529_RS05870 (*E. marmotae*)	2.7e-82	59.8
3.7e-79	60.2	HVX45_RS11505 (*E. fergusonii*)	**1.1e-142**	**98.0**
1.3e-78	60.2	GV529_RS14465 (*E. marmotae*)	**2.3e-142**	**97.6**

Table 4. BLAST e-values and percent identity for HOG 21.
BLAST e-values and percent identity, reported by OrthoFinder, for two genes from *E. coli*, b2244 & b3411, as BLASTed against the other members of HOG21 from *E. albertii*, *E. fergusonii*, and *E. marmotae*. Bolded values are the lowest e-value and highest percent identity. HVX45_RS22925 had no reported best match within the HOG and had a very different length of 68 amino acids vs. the other member's average length of 305 amino acids.

HOG 21				
Escherichia coli b2244 (*rpnE*)			*Escherichia coli* b3411 (*rpnA*)	
e-value	% identity	Gene to be BLAST against	e-value	% identity
2.2e-155	**89.3**	JRC41_RS07400 (*E. albertii*)	3.3e-100	57.9
3.9e-163	**90.9**	HVX45_RS21485 (*E. fergusonii*)	1.5e-98	65.2
5.9e-161	**92.2**	GV529_RS12150 (*E. marmotae*)	4.4e-100	58.1
4.7e-100	60.7	HVX45_RS12120 (*E. fergusonii*)	**9.0e-157**	**93.5**
5.1e-96	57.1	GV529_RS10930 (*E. marmotae*)	**3.9e-149**	**86.3**
N/A	N/A	HVX45_RS22925 (*E. fergusonii*)	N/A	N/A

Table 5. Names and RefSeq accessions for genomes used in Lim et al. 2022.
Gene annotations are in parenthesis: Suppressor of cell division A (*sdiA*), transcription regulator (tr) which is further noted as part of the luminescence (*luxR*) family, regulator of elastase *lasB* (*lasR*), rhamnolipid regulator (*rhlR*), and quorum-sensing transcription repressor (*qscR*).

Genius species	RefSeq accession	Gene locus tag (gene annotation)
Escherichia coli	GCF_000005845.2	b1916 (*sdiA*)
Salmonella enterica	GCF_000006945.2	STM1950 (*sdiA*)
Klebsiella pneumoniae	GCF_000445405.1	N559_RS09495 (*sdiA*)
Pseudomonas aeruginosa	GCF_001181725.1	AFI95_RS32400 (tr, *luxR* family) AFI95_RS29375 (*lasR*) AFI95_RS07465 (*rhlR*) AFI95_RS28195 (*qscR*)

Table 6. Synteny ratio between genes for HOG 346.
The four Gammaproteobacteria genomes were analyzed with OrthoRefine (window size = 30; synteny ratio = 0.2).

		Synteny ratio					
Genius species		*E. coli*	*S. enterica*	*Pseudomonas aeruginosa*			
	Locus tag	b1916	STM1950	AFI95_RS32400	AFI95_RS29375	AFI95_RS07465	AFI95_RS28195
E. coli	b1916	-	-	0.22	0.00	0.00	0.00
S. enterica	STM1950	0.9	-	0.33	0.03	0.00	0.00
K. pneumoniae	N559_RS09495	0.86	0.83	0.00	0.00	0.00	0.00

Table 7. BLAST percent identity for HOG 346.
BLAST percent identity, reported by OrthoFinder, between genes for HOG 346 of the four Gammaproteobacteria genomes (*E. coli*, *S. enterica*, *K. pneumoniae*, & *P. aeruginosa*).

BLAST percent identity						
			P. aeruginosa			
	b1916	STM1950	AFI95_RS32400	AFI95_RS29375	AFI95_RS07465	AFI95_RS28195
b1916	-	-	37.3	30.1	40.8	33.2
STM1950	71.3	-	34.7	30.2	45.0	34.0
N559_RS09495	65.8	66.7	25.7	Not reported	43.7	31.0

Table 8. Species names and RefSeq accession for the sixteen Actinomycetota genomes.

Genus species	RefSeq accession
Cryptobacterium curtum	GCF_000023845.1
Olsenella timonensis	GCF_900119915.1
Olsenella uli	GCF_000143845.1
Eggerthella lenta	GCF_021378605.1
Egibacter rhizosphaerae	GCF_004322855.1
Egicoccus halophilus	GCF_004300825.1
Denitrobacterium detoxificans	GCF_001643775.1
Rubrobacter xylanophilus	GCF_000014185.1
Rubrobacter tropicus	GCF_011492945.1
Rubrobacter marinus	GCF_011492965.1
Acidimicrobium ferrooxidans	GCF_000023265.1
Streptomyces fradiae	GCF_008704425.1
Streptomyces griseus	GCF_000010605.1
Streptomyces avermitilis	GCF_000009765.2
Acidothermus cellulolyticus	GCF_000015025.1
Actinomyces oris	GCF_016127955.1

Table 9. BLAST e-values and percent identity for HOG 55.
BLAST e-values and percent identity, reported by OrthoFinder, for two genes from *S. cerevisiae*, YJR009C & YGR192C, as BLASTed against the other members of HOG55 from *S. mikatae* & *S. kudriavzevii*. Bolded values are the lowest e-value and highest percent identity.

HOG 55				
S. cerevisiae YJR009C			*S. cerevisiae* YGR192C	
e-value	% identity	Gene to be BLAST against	e-value	% identity
1.5e-186	**99.4**	SMKI_10G2100	6.4e-182	96.4
1.6e-183	97.3	SKDI_10G2170	3.2e-181	95.8
8.7e-182	96.4	SMKI_16G0680	**2.4e-187**	**99.4**
1.4e-184	**97.9**	SKDI_07G4440	**3.0e-182**	**96.4**

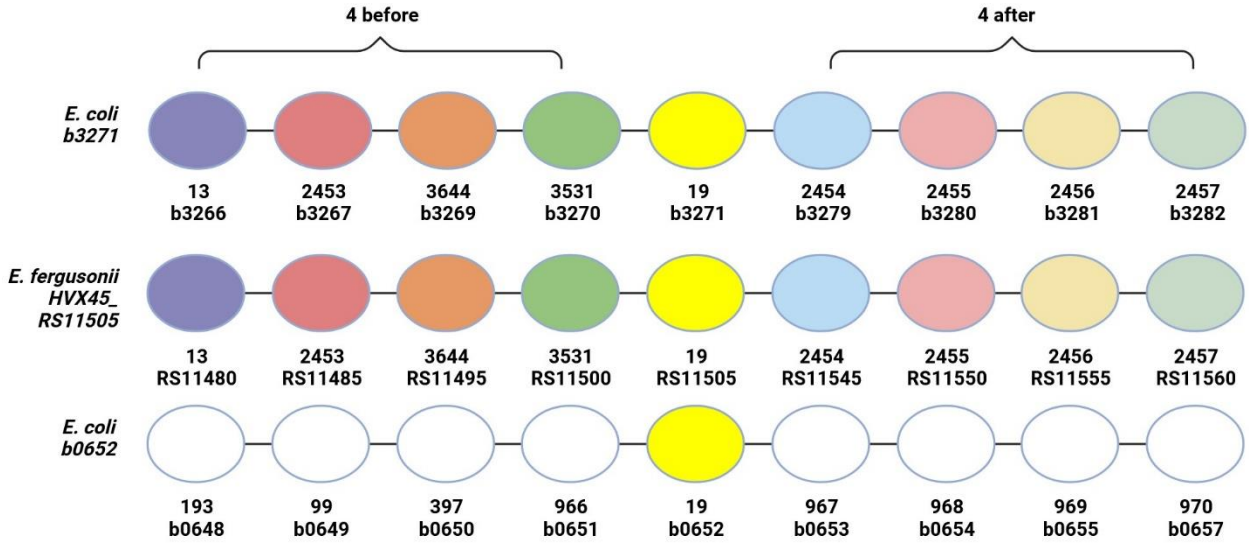

Figure 1. Example of OrthoRefine's synteny analysis.
The window around three genes assigned to HOG19 by OrthoFinder demonstrates how OrthoRefine determines which of the *E. coli* genes is an ortholog of *E. fergusonii's* HVX45_RS11505. The HOG19 genes are shown with yellow fill, other genes assigned to the same HOG are shown in matching colors, and genes that have orthologs in other genomes outside the displayed window are shown in white. The first number below each circle denotes the HOG assigned by OrthoFinder, while the second entry shows the locus tag.

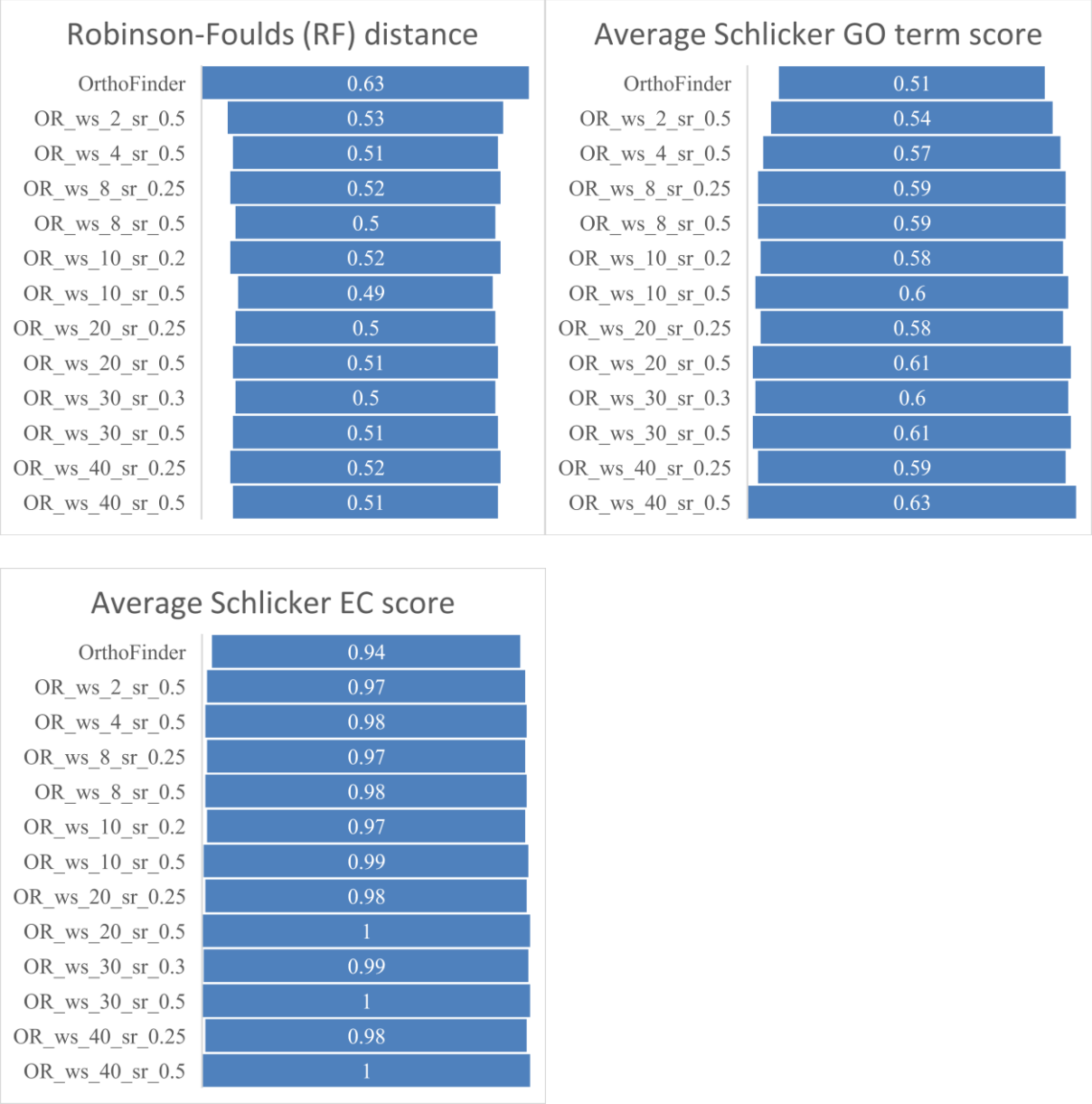

Figure 2. Benchmarking results for OrthoFinder and OrthoRefine on the Quest for Orthologs bacterial dataset.
OrthoRefine was run with different parameters for window size (ws) and synteny ratio (sr) (A) Robinson-Foulds (RF) distance as a measure of specificity (lower values indicate higher specificity). (B) Average Schlicker scores for gene ontology (GO) and (C) enzyme classification (EC) as a measure of functional ortholog identification (higher scores indicate improvement).

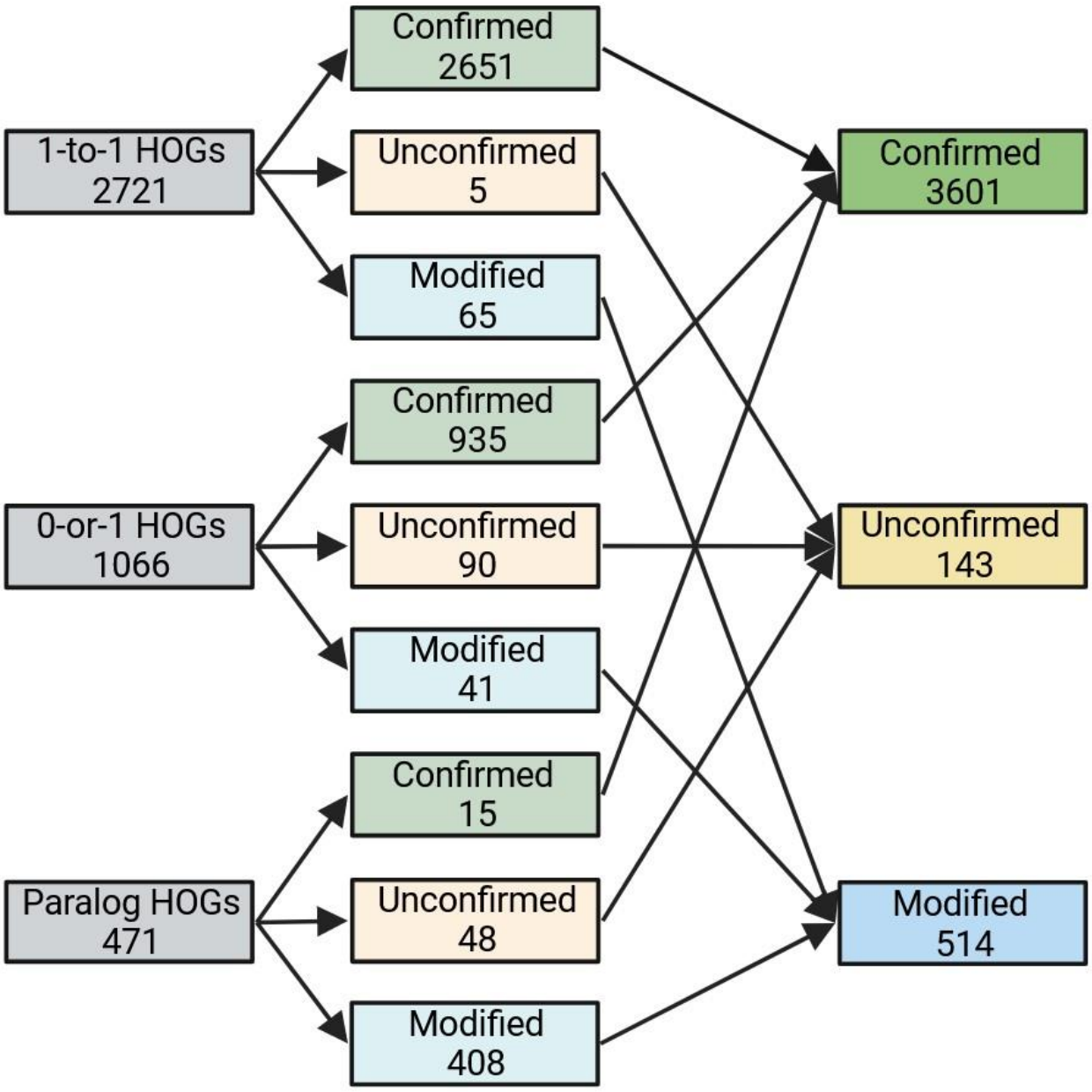

Figure 3. Summary statistics for the four *Escherichia* genomes.
The genomes were analyzed with OrthoRefine (window size = 8; synteny ratio = 0.5). 1-to-1 HOGs contained precisely one gene per genome. 0-or-1 HOGs were missing an ortholog in at least one genome and none of the genomes contributed more than one gene. Paralog HOGs are those where at least one genome contributed more than one gene. Confirmed HOGs are those where all genes assigned by OrthoFinder to a HOG were supported by synteny. Unconfirmed HOGs lacked synteny support for all genes assigned to a HOG by OrthoFinder or any SOG subgroup. HOGs where synteny eliminated at least one paralog and/or divided the HOG into at least one SOG are designated Modified HOGs. HOGs comprised of genes of only one genome could not be analyzed by OrthoRefine and were excluded.

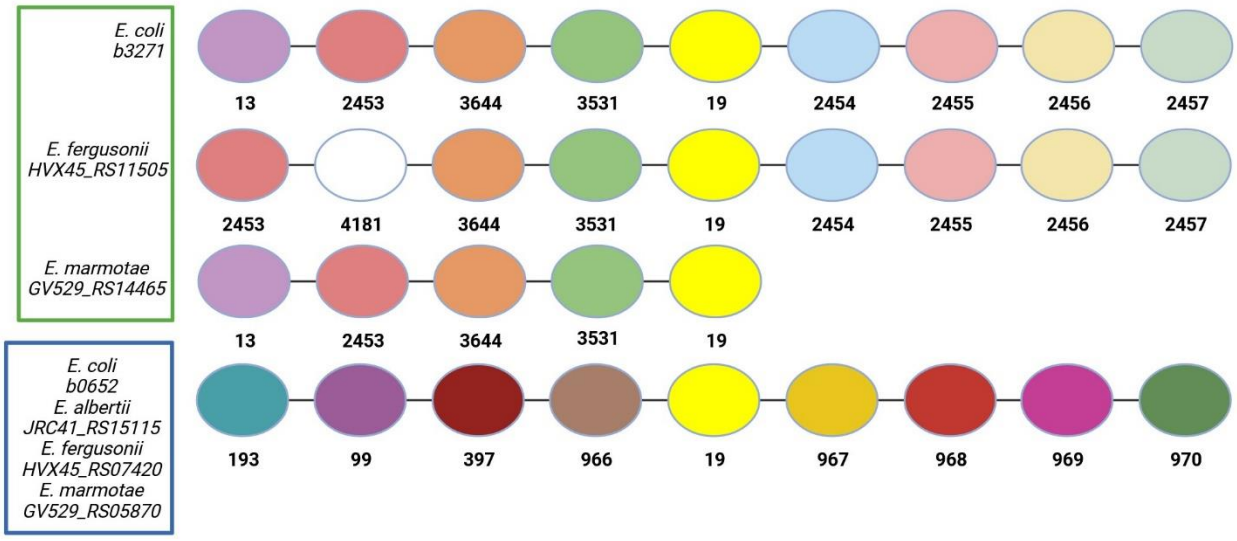

Figure 4. Synteny analysis of HOG 19.
Matched colored circles represent genes assigned to the same HOG by OrthoFinder, with the HOG numbers shown below each circle; white circles denote genes with orthologs in other genomes located outside the displayed window. Green and blue boxes mark the two SOGs delineated by OrthoRefine. The missing data from *E. marmotae* to the right of the HOG19 member is due to a scaffold boundary in the assembly. The neighborhoods for genes marked by the blue box are identical and are collapsed into a single line. The other members of HOG19 (b4106, b4096, HVX45_RS02390, HVX45_RS04025, & HVX45_RS09410) are omitted because they have no syntenic matches to any other member of HOG19.

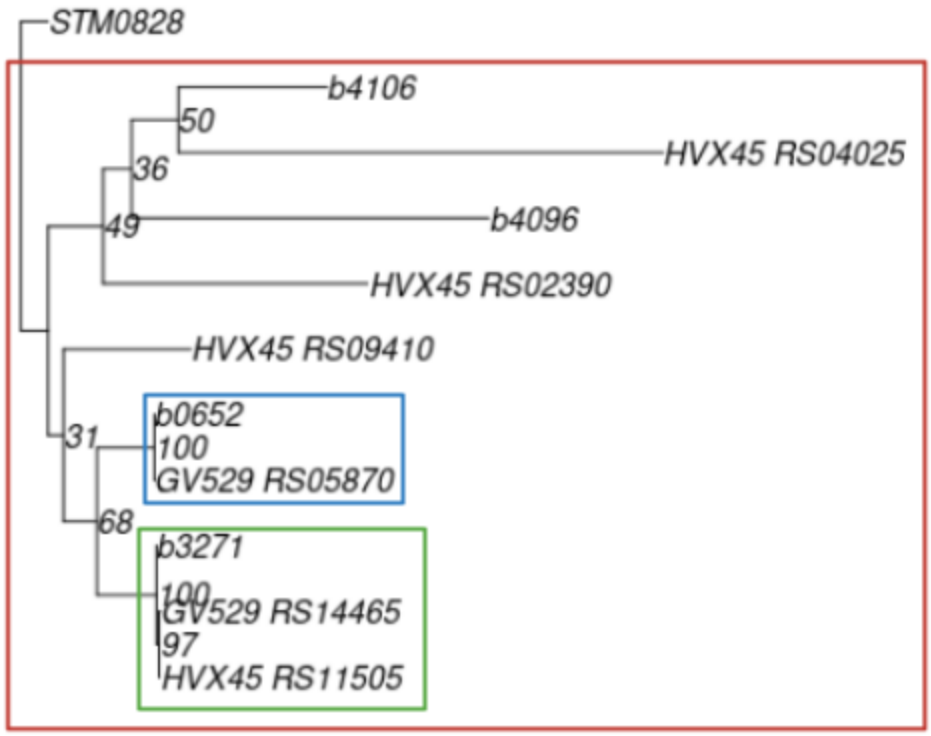

Figure 5. Phylogenetic tree of HOG 19.
The tree was generated by RAxML comprised of sequences from *E. coli* (prefix b), *E. fergusonii* (HVX45), *E. marmotae* (GV529), and *E. albertii* (JRC41). b0652 of *E. coli* is a representative of HVX45_RS07420 of *E. fergusonii* and JRC41_RS15115 of *E. albertii*. The best BLAST hit from *S. enterica*, STM0828, was used to root the tree. Boxes have been placed around Orthofinder's grouping (red) and OrthoRefine's groupings (blue or green). Node values are bootstrap support with n = 1000. There was bootstrap support (>70 [101]) for the blue and green groupings but not for additional groups.

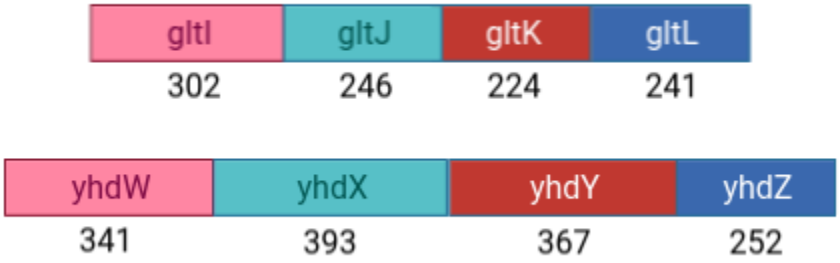

Figure 6. The operons of HOG 19.
The *glt*IJKL and *yhd*WXYZ ABC transporter operons of the *Escherichia* genera. The product length is below each gene. The RNA gene, *sroC*, has been omitted from the *glt*IJKL operon for simplicity.

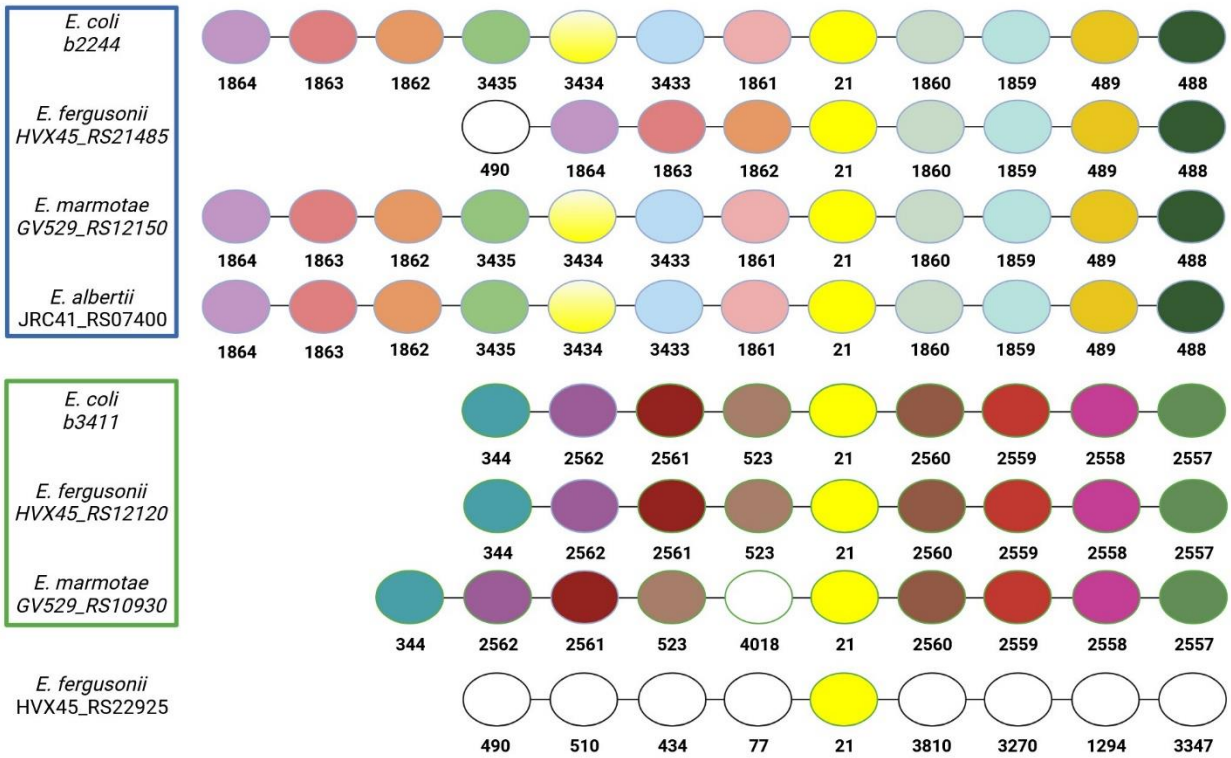

Figure 7. Synteny analysis of HOG 21.
Matched colored circles represent genes assigned to the same HOG by OrthoFinder, which is shown below each circle; white circles denote genes with orthologs in other genomes located outside the displayed window. Green and blue boxes mark the two SOGs identified from HOG21. The analysis was performed with window size 8 (four genes on each side of HOG21). However, because the synteny is evaluated separately for each pair of genomes and the *E. fergusonii* genome contains no representative of HOGs 3433, 3434, and 3435, these genes are excluded from comparisons with *E. fergusonii* (OrthoRefine ignores genes that do not have a counterpart in the other genome) and the window instead includes an additional three genes (HOGs 1862, 1863, and 1864), which allows HVX45_RS21485 to be identified as the syntenous ortholog of b2244, GV529_RS12150, and JRC41_RS07400.

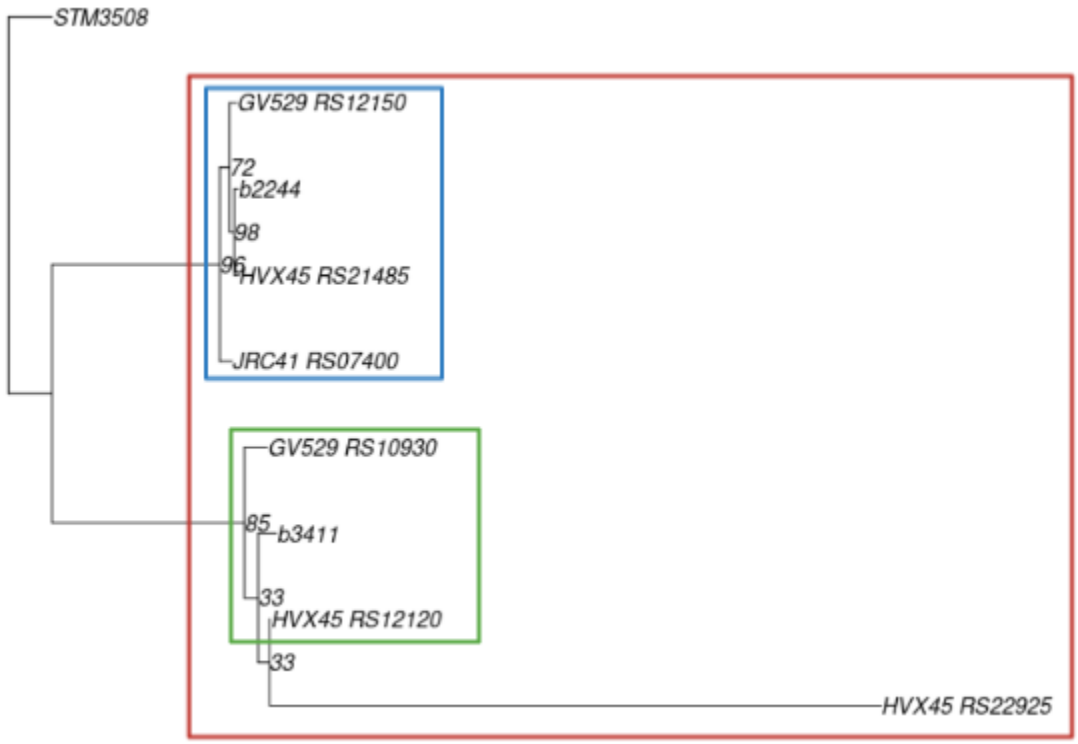

Figure 8. Phylogenetic tree of HOG 21.
The tree was comprised of sequences from *E. coli* (prefix b), *E. fergusonii* (HVX45), *E. marmotae* (GV529), and *E. albertii* (JRC41). The best BLAST hit from *S. enterica*, STM3508, was used to root the tree. Boxes have been placed around Orthofinder's grouping (red) and OrthoRefine's groupings (blue or green). Node values are bootstrap support with n = 1000. There was bootstrap support (>70) for the blue and green groupings. HVX45_RS22925 had bootstrap support to be included in the green group; however, such a grouping would be non-monophyletic [102] and thus violated the species overlap method used to tell ortholog from paralog [68].

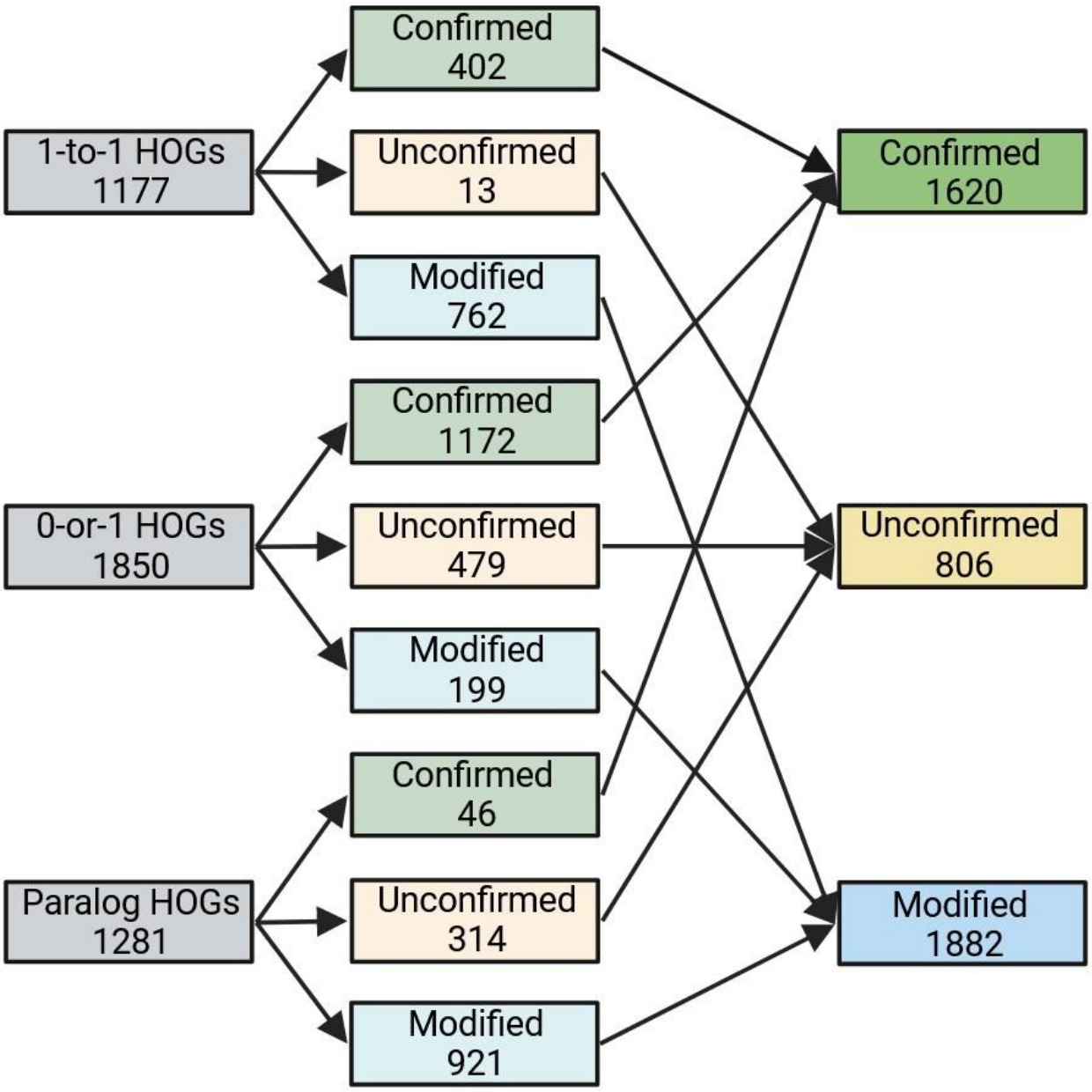

Figure 9. Summary statistics for the four Gammaproteobacteria genomes.
The genomes were analyzed with OrthoRefine (window size = 8; synteny ratio = 0.5). See legend to Figure 3.

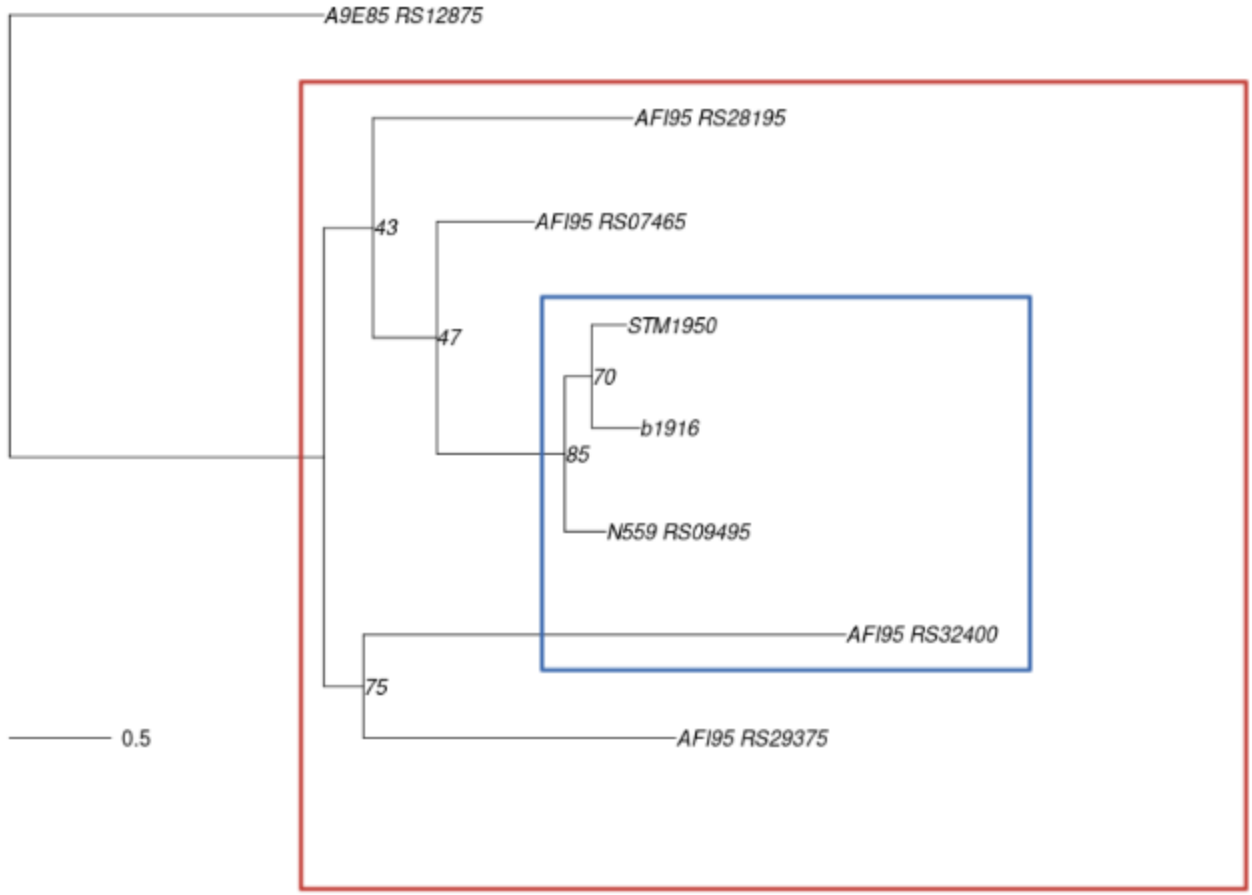

Figure 10. Phylogenetic tree of HOG 346.
The tree was comprised of sequences from *E. coli* (b1916), *K. pneumoniae* (N559_RS09495), *S. enterica* (STM1950), and *P. aeruginosa* (prefix AFI95). The best BLAST hit from *Legionella pneumophila*, A9E85_RS12875, was used to root the tree. Boxes have been placed around OrthoFinder's grouping (red) and OrthoRefine's grouping (blue). There was a lack of bootstrap support (>70) for any of the four *P. aeruginosa* genes to be grouped with the genes from the other species. Node values are bootstrap support with n = 1000.

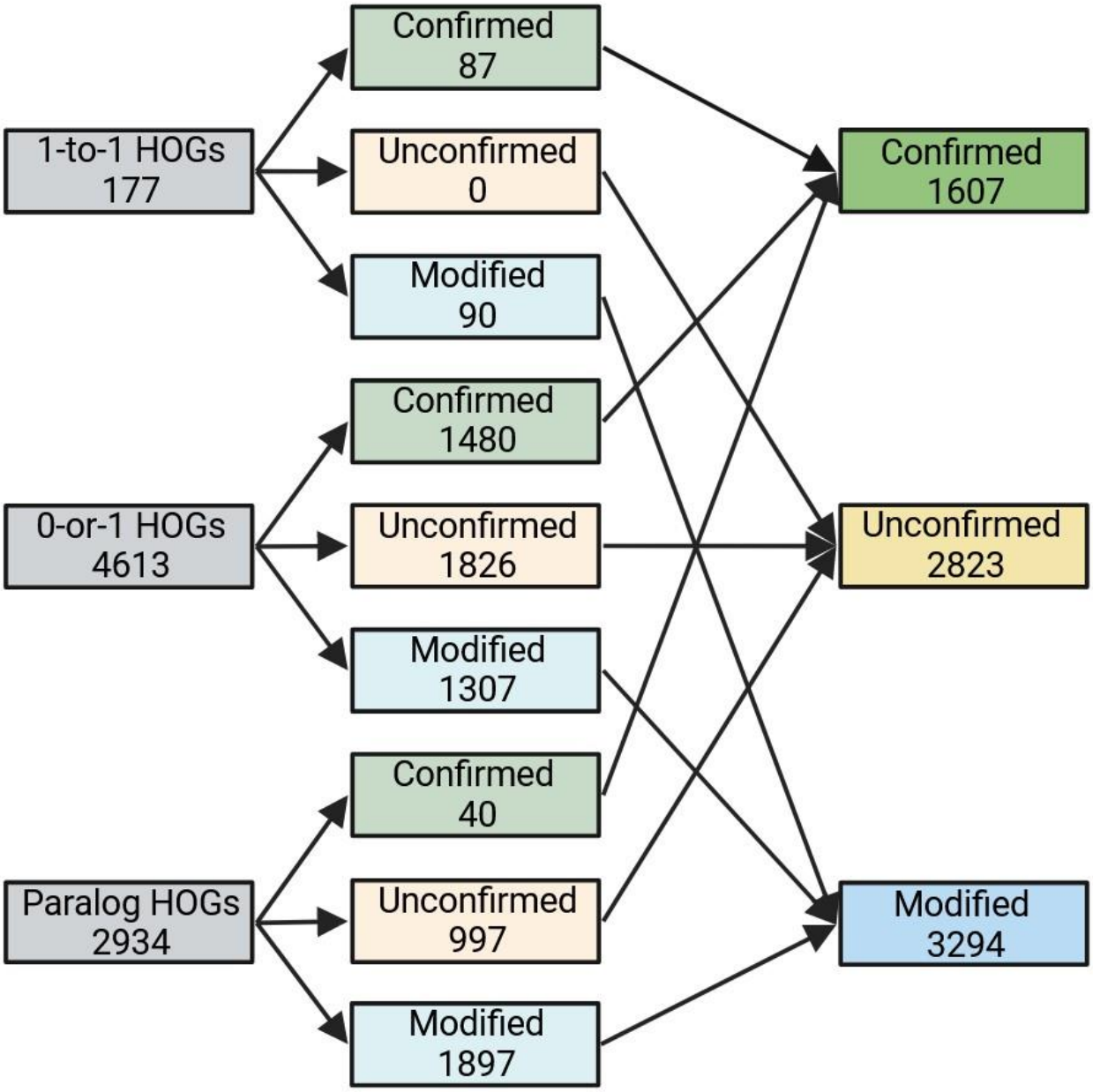

Figure 11. Summary statistics for the sixteen Actinomycetota genomes.
The genomes were analyzed with OrthoRefine (window size = 8; synteny ratio = 0.5). See legend to Figure 3.

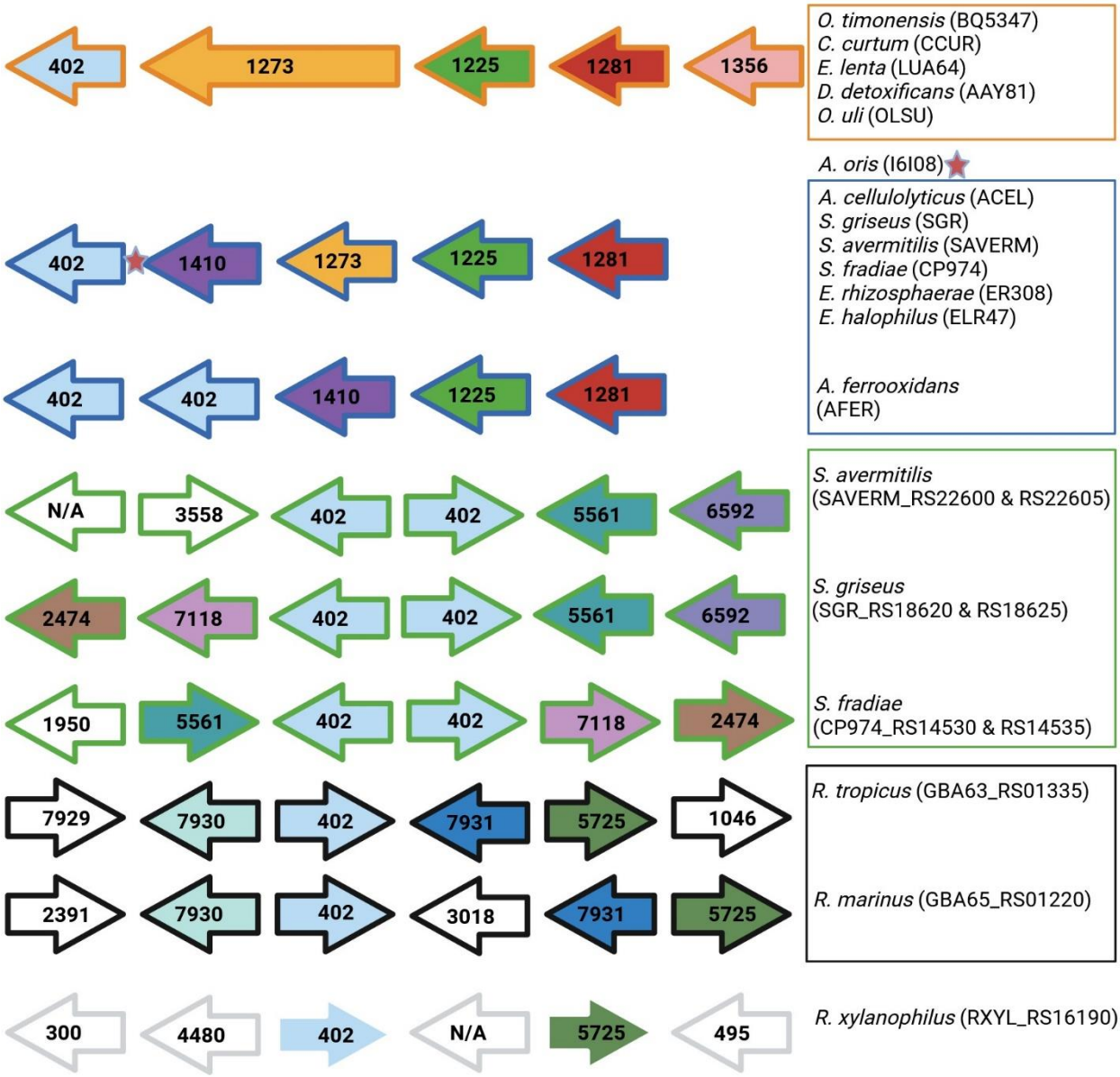

Figure 12. Synteny analysis of HOG 402.
Matched colored arrows represent the same HOG number, which have been placed inside the arrows; white arrows denote genes with no match from the same HOG within the window, arrows containing N/A were not assigned to a HOG. The Actinomycetota operons of *pknB* have been divided based on SOG assignment and their edges color coded (SOG 402.0 orange, 402.1 blue, 402.2 green, or 402.3 black). Additional matches within the window have been omitted from the figure as the focus was on the operon. Due to an additional STPK (red star), assigned to HOG 400, between the STPK and the PBP, *A. oris* was not assigned to any SOG but otherwise has the same operon as SOG 402.1.

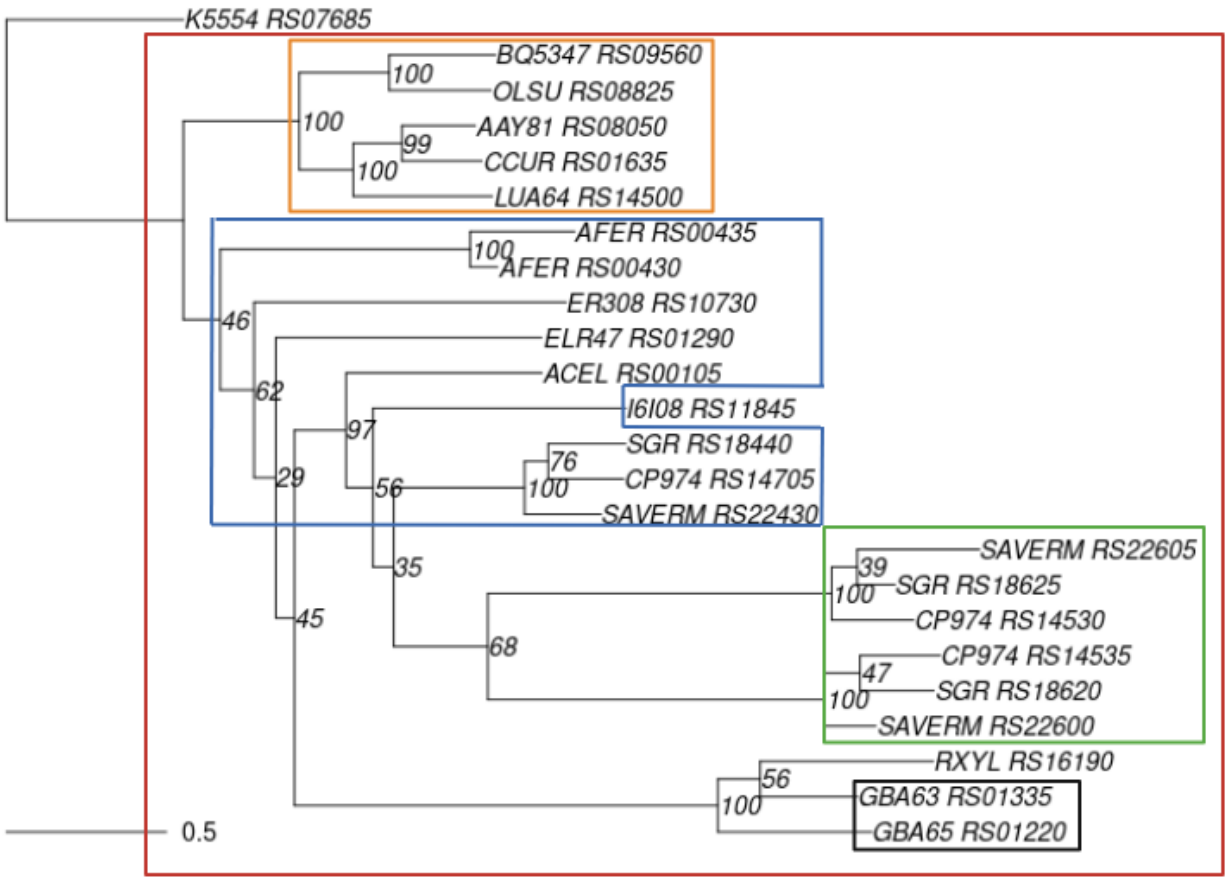

Figure 13. Phylogenetic tree of HOG 402.
The tree was comprised of sequences from the 16 Actinomycetota; the *Gelria* gene with the best BLAST hit was used to root the tree. OrthoFinder grouped all genes into a single HOG (red box), while OrthoRefine split the group into four SOGs (orange, blue, black, and green boxes). Node values are bootstrap support values with n = 1,000. There was a lack of bootstrap support (>70) to delineate the *Streptomyces* orthologs from paralogs using the species overlap method.

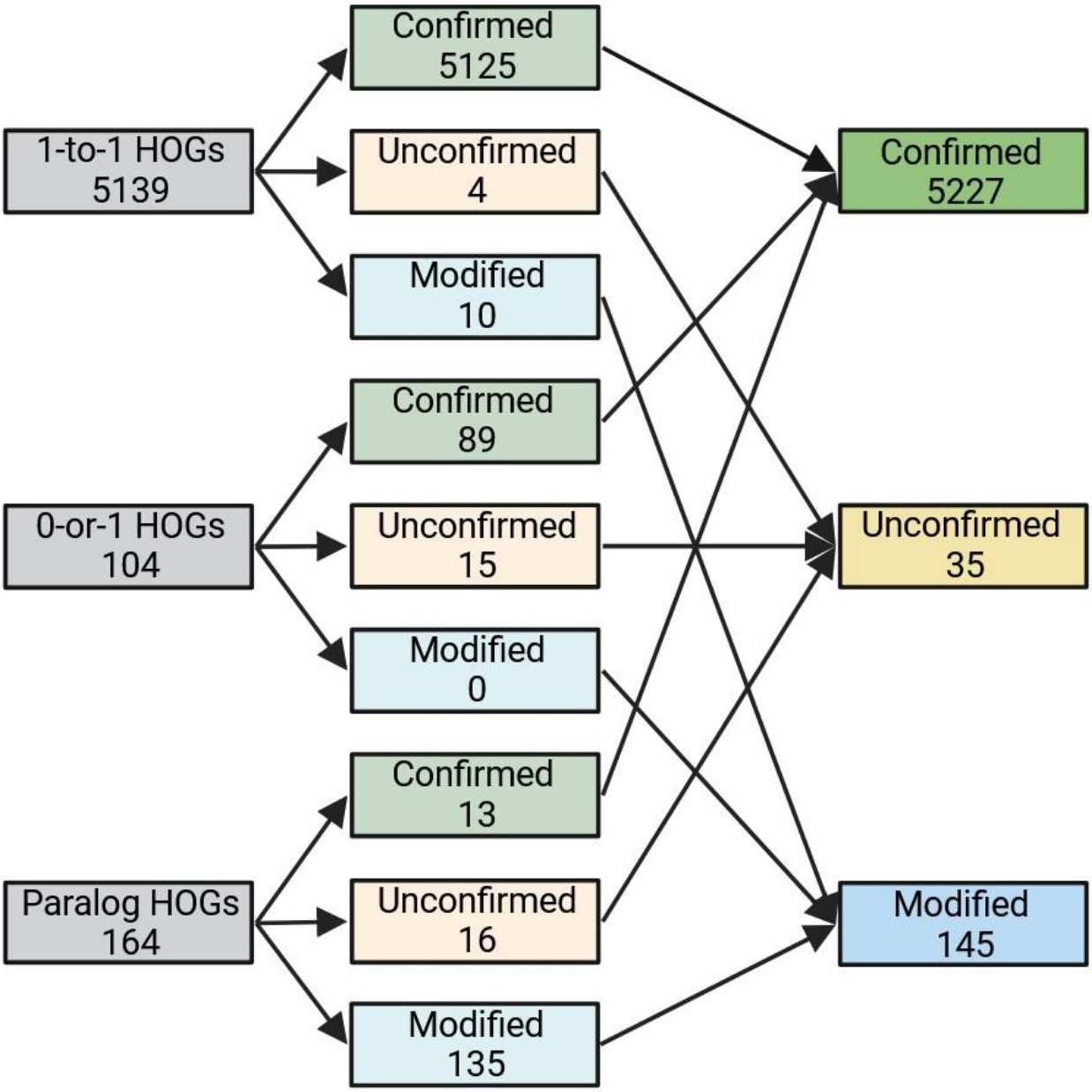

Figure 14. Summary statistics for the sixteen *Saccharomyces* genomes.
The genomes were analyzed with OrthoRefine (window size = 8; synteny ratio = 0.5). See legend to Figure 3.

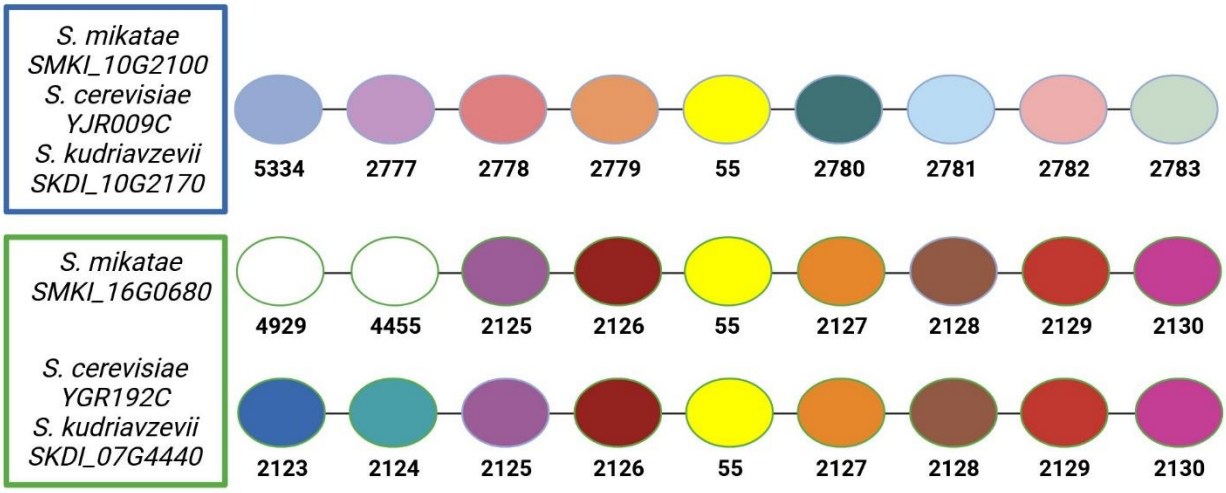

Figure 15. Synteny analysis of HOG 55.
Matched colored circles represent genes assigned to the same HOG (shown below each circle); white circles denote genes with orthologs in other genomes located outside the displayed window. Blue and green boxes mark the two SOGs derived from HOG55.

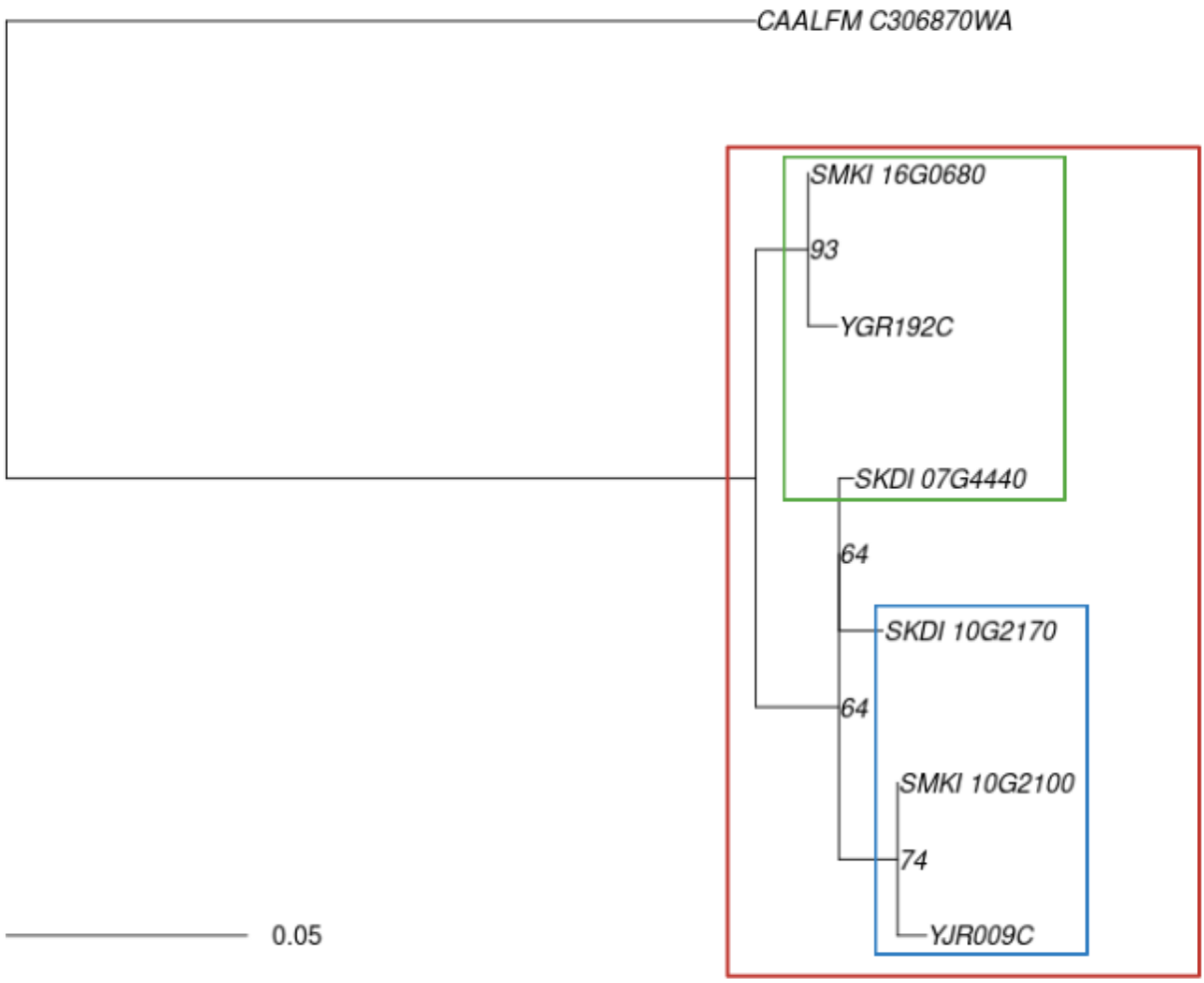

Figure 16. Phylogenetic tree of HOG 55.
The tree was comprised of sequence from *S. cerevisiae* (YJR), *S. mikatae* (SMKI), and *S. kudriavzevii* (SKDI). The best BLAST hit from *Candida albicans*, CAALFM_C306870WA, was used to root the tree. Boxes have been placed around OrthoFinder's grouping (red) and OrthoRefine's groupings (blue or green). Node values are bootstrap support with n = 1000.

CHAPTER 3

Expanding OrthoRefine's Functionality

In this chapter, I extend OrthoRefine's functionality to accept OrthoXML input files and apply OrthoRefine to data from one commonly used ortholog database, OMA (orthologous matrix).

Introduction

A recurring issue in ortholog analysis is lack of adherence to a uniformly accepted standard data format. Most programs and databases use custom formats that are often difficult to compare and complicate further downstream processing of data. The OrthoXML file format [103, 104] was designed to standardize how ortholog data are communicated; currently, there are eight databases that provide data in OrthoXML format [105]. This work aimed to expand OrthoRefine's utility by accepting input provided in the OrthoXML format. Additionally, I utilized OrthoRefine to compare the results obtained with OrthoFinder [13, 49] and the OMA database [53].

Methods

I used a combination of the OrthoXML documentation [106] and the OMA OrthoXML files [107] to design a script to convert from OrthoXML to OrthoFinder's output format, the native input format for OrthoRefine. However, the provided documentation was insufficient to unambiguously interconvert the formats, which was further complicated by the need to match unique OMA protein identifiers to gene identifiers in the genome annotations (See Appendix C for example files and details). Moreover, although OrthoXML is listed as an optional output

format in OrthoFinder's documentation, this feature appears nonfunctional (the program crashes if the option is selected). Overcoming these difficulties required extensive testing and validation. Unfortunately, due to specificities of OMA's data and the use of nonstandard protein identifiers, the script may require modifications to work with OrthoXML files from other sources.

The dataset to test the conversion script and compare the OMA database with OrthoFinder was comprised of three Archaea genomes from within the same genus: *Archaeoglobus fulgidus* (GCF_000008665.1), *Archaeoglobus profundus* (GCF_000025285.1), and *Archaeoglobus veneficus* (GCF_000194625.1). OrthoFinder was run with default settings and OMA's HOG database file was converted for use with OrthoRefine; OrthoRefine was applied to OrthoFinder's and OMA's data using the recommended settings (window size = 8, synteny ratio = 0.5). The average maximum number of orthologous genes (AMNOG, see Chapter 2), summary statistics generated as part of OrthoRefine's analysis, and specific examples of HOGs were utilized in comparing OrthoFinder and the OMA-hog database results.

Results and discussion

The data from OrthoRefine's analysis of OMA had a higher AMNOG (a surrogate measure of sensitivity) than the OrthoFinder results at lower window sizes but experienced a notable fall-off as the window size increased beyond 10 (Table 10); no trend was noted on smaller vs larger synteny ratio. I speculate that OMA's fall-off was due to how the conversion script interpreted the orthogroups into HOGs (at the lowest orthogroup level); a planned analysis of OrthoFinder's OrthoXML conversion, which was intended to be used as guide for which orthogroup level(s) to consider as a HOG when converting from OrthoXML format, was abandoned due to a bug resulting in OrthoFinder's non-functionality of printing to the OrthoXML format. Nevertheless, an analysis of OrthoFinder's and OMA's raw data revealed

that OrthoFinder produced more HOGs (OrthoFinder 1721, OMA 1550), more gene paralogs (OrthoFinder 807, OMA 287), and overall, more genes grouped into HOGs (OrthoFinder 5235, OMA 4379) – suggesting higher sensitivity in the OrthoFinder data. Additionally, the OMA data contained fewer paralogous HOGs as a percentage of total HOGs (OrthoFinder 25%, OMA 13%) and more 1-to-1 (OrthoFinder 54%, OMA 61%) or 0-or-1 HOGs (OrthoFinder 21%, OMA 26%) (Figure 17,18) – suggesting higher specificity in the OMA data. (A prior study identified 1,001 1-to-1 orthologs between *A. fulgidus*, *A. profundus*, *A. veneficus*, and *A. sulfaticallidus* [108]; OrthoFinder identified 929 while OMA identified 951; *A. sulfaticallidus* was not included in OMA's database.) The Orthology Benchmarking results (see chapter 2) also concluded that OrthoFinder had a higher sensitivity than OMA and that OMA had higher specificity than OrthoFinder [109].

OrthoFinder's and OMA's OrthoRefine outputs were compared for proportions of different outcomes when a HOG was refined by synteny. The most common outcome was that OrthoFinder's result could be further refined with synteny, but OMA already had the correct refinement according to OrthoRefine (42%). For 16% of HOGs, OrthoFinder had the correct refinement originally whereas OMA's HOG could be further refined. OrthoFinder and OMA identified the same orthologs that were further refined by OrthoRefine in 30% of HOGs, and in 12% of HOGs OrthoFinder results required fewer refinements (contained fewer paralogs) than OMA. When both OMA and OrthoFinder results could be refined, OMA never produced HOGs that would require fewer refinements than OrthoFinder, probably because OMA's dataset contained fewer paralogous HOGs than OrthoFinder's (Figure 17, 18).

Case Study 6: OrthoFinder HOG 186, OMA HOG 1435 - (hydrogenase iron-sulfur subunit)

OrthoFinder HOG 186 and OMA HOG 1435 were comprised of the same genes: AF_RS06935 of *A. fulgius*, ARCPR_RS07755 of *A. profundus*, and ARCVE_RS10480, ARCVE_RS07785, & ARCVE_RS02635 of *A. veneficus*; all were annotated as encoding a hydrogenase iron-sulfur subunit. The synteny analysis conducted on OrthoFinder's output identified ARCVE_RS07785 as the ortholog of AF_RS06935 and ARCPR_RS07755. The synteny analysis on OMA's output failed to group any gene of *A. veneficus* with *A. fulgius* and *A. profundus* as there were insufficient genes that matched in the window, which resulted in the synteny ratio failing to be at least 0.5 (Figure 19). Notably, ARCVE_RS07785 is listed as the ortholog of AF_RS06935 & ARCPR_RS07755 in the OMA web interface whereas ARCVE_RS10480 & ARCVE_RS02635 were listed as paralogs [110]; this points to a possible discrepancy between the OMA database files provided for download and the information displayed online. Inspection of OMA's data revealed that one gene that was matched within the window of the OrthoFinder analysis but not the OMA analysis, ARCVE_RS07800, was in OMA's list of genes but was not included as a member of any OMA HOG; regardless of how the OrthoXML script interpreted orthogroup levels when converting to HOGs, ARCVE_RS07800 would have never been included in a HOG. Another gene that matched within the window of the OrthoFinder analysis but not the OMA, ARCVE_RS07780, was assigned to an OMA HOG that only had two of the three genomes (the HOG was missing an *A. fulgius* gene); thus, it could not be included in the window for the pairwise analysis between *A. fulgius* and *A.* veneficus. This demonstrates how OrthoRefine will work better with a program that prioritizes sensitivity over

specificity as OrthoRefine can improve specificity via eliminating false positives (paralogs) but it cannot improve sensitivity.

ARCPR_RS07755 is known as *mvhD*, which has been reported to form an operon with *mvhG* and *mvhA* in *Methanobacterium thermoautotrophicum* [111]. The three *Archaeoglobus* genomes were inspected and *mvhG* and *mvhA* homologs were located immediately downstream of AF_RS06935, ARCPR_RS07755, and ARCVE_RS07785. However, ARCVE_RS10480 and ARCVE_RS02635 were not in proximal locations to the *mvhG* and *mvhA* homologs (Figure 20). OrthoRefine's conclusion that ARCVE_RS07785 is the correct *mvhD* ortholog in *A. veneficus* is consistent with the presence of the *mvh*DGA operon, whereas BLAST e-values and percent identity (Table 11), and the phylogenetic tree (Figure 21) on their own were insufficient to differentiate the ortholog from paralogs.

As a caveat of the OrthoFinder and OMA comparison, the data used to generate the genome annotations required by OrthoRefine and the data incorporated into OMA's database were mismatched in date; for some genomes, this was a difference of several years. For the *Archaeoglobus* dataset, this led to 17 genes in the OMA data that could not be matched to a gene in the annotation. The discrepancies between the OMA database and current genome annotations vary among different genomes and can be much larger.

Ultimately, I found the OMA database implementation of OrthoXML to be problematic. Particularly limiting was the use of protId as a nonstandard OMA-specific identifier, which prevented conversion of OMA's HOG OrthoXML format to include standard and up-to-date identifiers used in major sequence databases such as NCBI or EMBL.

Table 10. Combinations of window size and synteny ratio on the AMNOG for three *Archaeoglobus* genomes.

		OrthoFinder	OMA
Window Size	Synteny ratio	Average max number of orthologous genes (AMNOG)	
2	0.5	2.04	2.24
4	0.25	2.08	2.35
4	0.5	2.18	2.4
6	0.2	2.21	2.37
6	0.5	2.19	2.33
8	0.2	2.16	2.27
8	0.3	2.22	2.36
8	0.5	2.08	2.21
10	0.2	2.13	2.27
10	0.3	2.1	2.25
10	0.5	1.95	1.94
30	0.2	1.86	1.67
30	0.3	1.81	1.33
30	0.5	2	1.33
40	0.2	1.85	1.53
40	0.3	1.83	0.5
40	0.5	0	0

Table 11. BLAST e-values and percent identity for OrthoFinder HOG 186.
Bolded values are the lowest e-value and highest percent identity. Three genomes from *Archaeoglobus* were analyzed: *A. fulgius* (AF), *A. profundus* (ARCPR), and *A. veneficus* (ARCVE).

	ARCPR_RS07755		ARCVE_RS10480		ARCVE_RS07785		ARCVE_RS02635	
	e-value	% identity	e-value	% identity	e-value	% identity	e-value	% identity
AF_RS06935	2.1e-36	52	1.5e-28	43.4	**2.0e-33**	**46.6**	3.9e-29	45.1
ARCPR_RS07755	-	-	6.5e-32	45.1	**1.5e-44**	**58.5**	8.8e-37	50.4

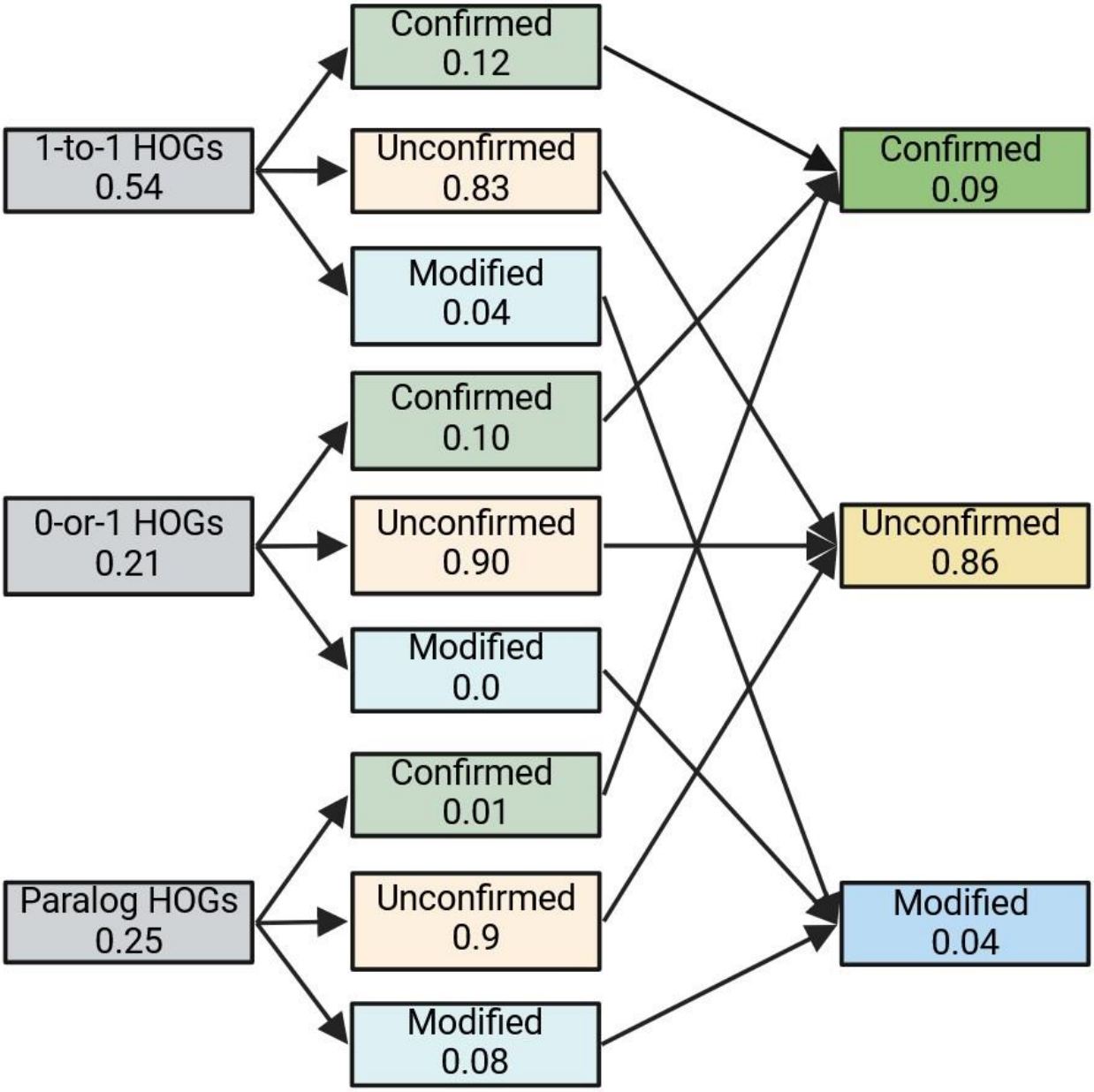

Figure 17. Summary statistics of three *Archaeoglobus* genomes analyzed with OrthoFinder and then OrthoRefine (window size 8, synteny ratio 0.5); data are percentages. See legend to Figure 3.

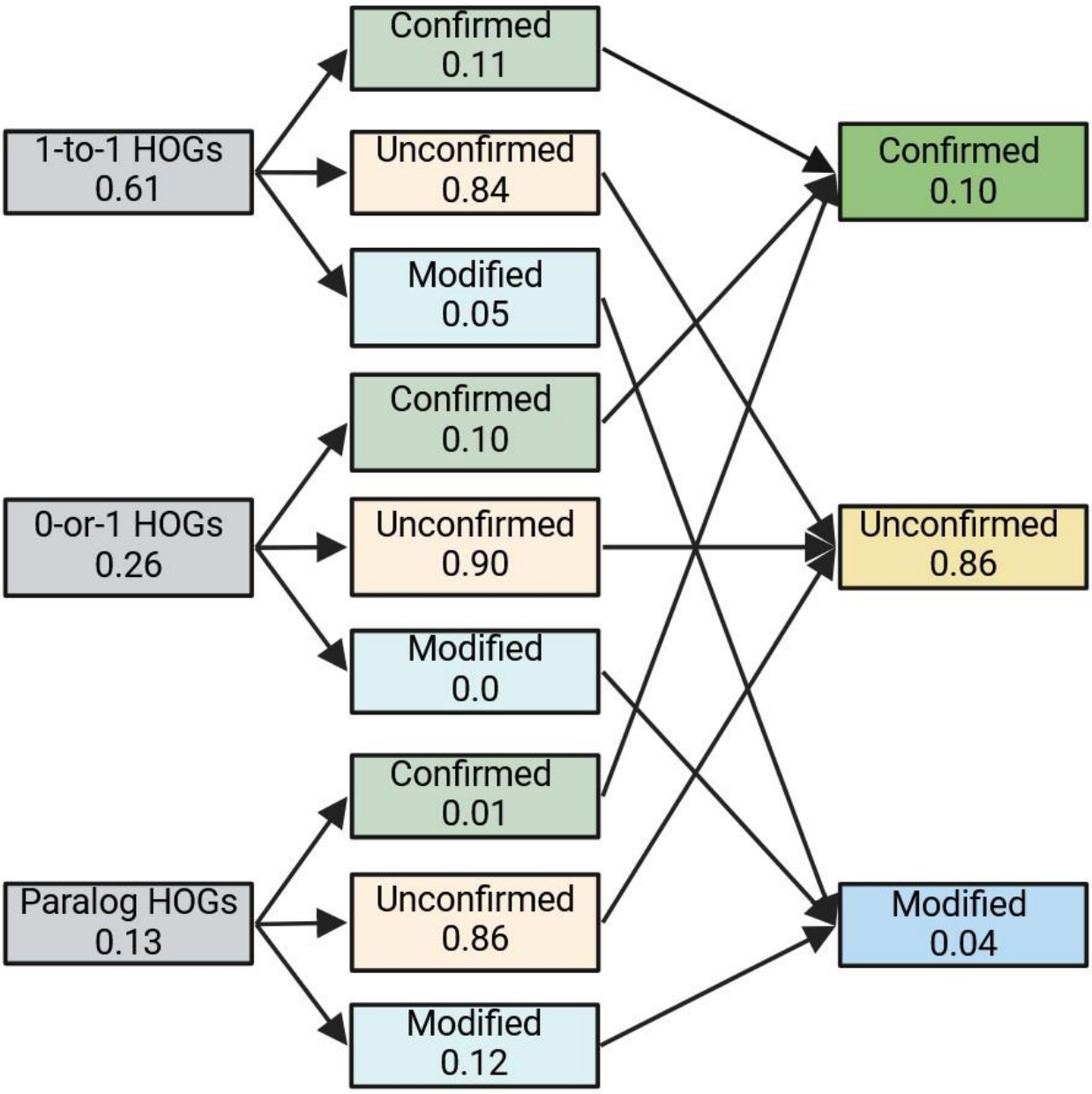

Figure 18. Summary statistics of three *Archaeoglobus* genomes from the OMA database analyzed with OrthoRefine (window size 8, synteny ratio 0.5); data are percentages. See legend to Figure 3.

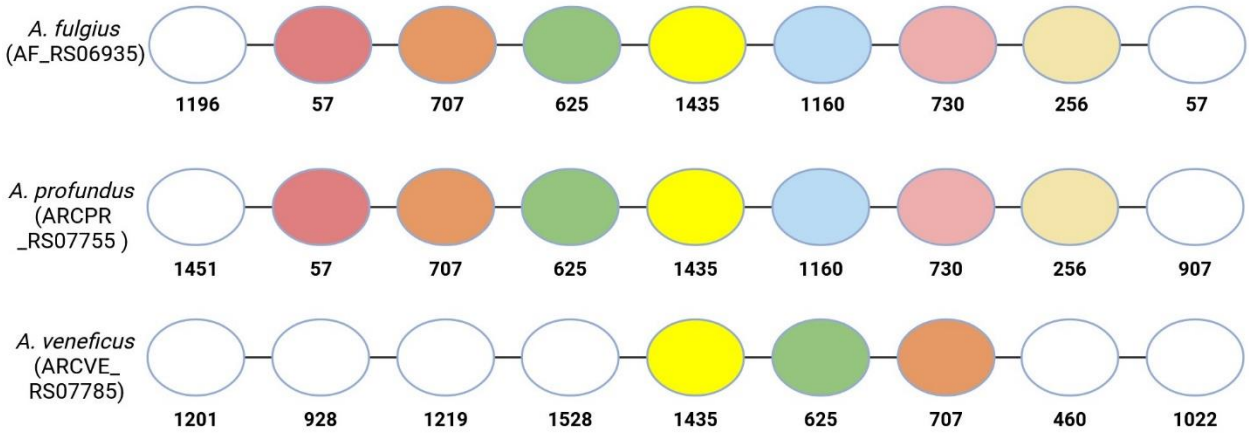

Figure 19. Synteny analysis of OMA HOG 1435.
Matched colored circles represent genes assigned to the same HOG by OMA, with the HOG numbers shown below each circle; white circles denote genes with orthologs in other genomes located outside the displayed window.

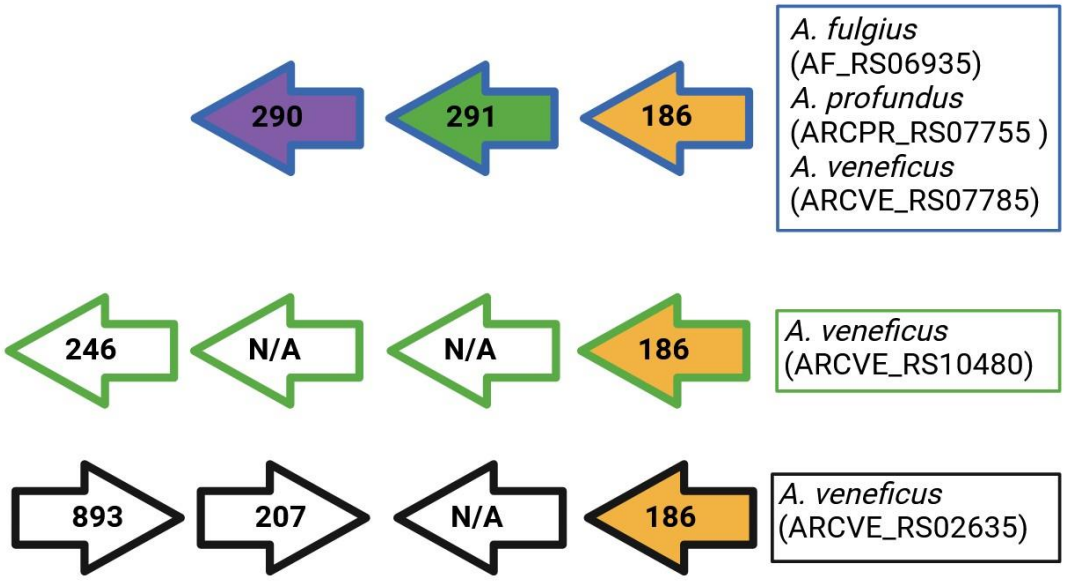

Figure 20. Operon analysis of HOG 186.
Matched colored arrows represent the same HOG number, which have been placed inside the arrows; white arrows denote genes with no match from the same HOG surrounding the operon, arrows containing N/A were not assigned to a HOG.

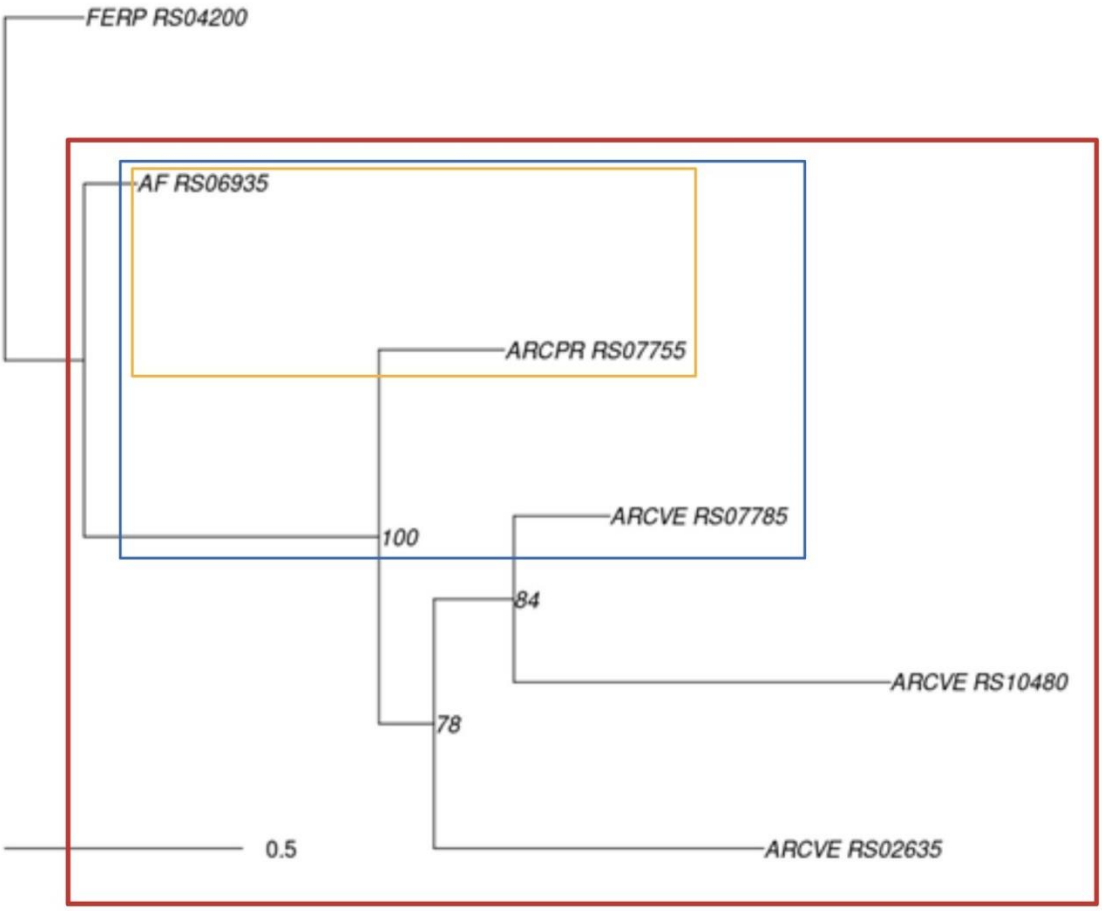

Figure 21. Phylogenetic tree of HOG 186.
The tree was comprised of sequences from three genomes of *Archaeoglobus*: *A. fulgius* (AF), *A. profundus* (ARCPR), and *A. veneficus* (ARCVE). The best BLAST hit from *Ferroglobus placidus*, FERP_RS04200, was used to root the tree. Boxes have been placed around OrthoFinder's & OMA's grouping (red), OrthoRefine's analysis of OrthoFinder's output (blue), and OrthoRefine's analysis of OMA's database file (gold).

CHAPTER 4

Conclusion

I developed OrthoRefine, a standalone program that automates refinements of ortholog identification by evaluating gene synteny. OrthoRefine is designed to mimic the desirable properties of OrthoFinder, namely ease-of-use (no dependencies on additional software, a simple input, and support scripts to download data or create summary statistics), automation, and speed. I expect OrthoRefine to be most beneficial when the desired orthologous relationship is 1-to-1 (i.e., a single ortholog from each genome with no paralogs). The value of synteny for ortholog identification has been demonstrated in previous studies [60, 100, 112-114], but in the absence of easy-to-use tools to identify syntenous orthologs automatically, such studies have been time-intensive and generally limited in their scope. This work further demonstrates how the use of synteny, automated in OrthoRefine, can enhance ortholog identification by analyzing different data sets and groups separated by different evolutionary distances. In addition to confirmation of OrthoRefine's ability to increase specificity of functional ortholog identification via the community benchmarking tool, detailed investigation of several cases by manual inspection of sequence alignments, phylogenetic trees, and operon structures provided additional independent support for OrthoRefine's results.

To further expand OrthoRefine's utility, I developed a script to convert the standard ortholog file format, OrthoXML, to OrthoRefine's input format to facilitate use of OrthoRefine in combination with other software for ortholog identification (in addition to OrthoFinder). I used this script to apply OrthoRefine to data from the OMA database and compared the

performance of OrthoRefine when applied to OMA data and OrthoFinder results. One outcome of adapting OrthoRefine for use with the database is that it demonstrated a need for compliance with standard data formats and data identifiers adopted by major bioinformatics resources (e.g. NCBI, EBI, KEGG). OMA's use of nonstandard protein identifiers in combination with dated genome annotations makes it difficult to relate the OMA database information to information from other databases.

REFERNCES

1. Fitch WM: **Distinguishing homologous from analogous proteins**. *Syst Zool* 1970, **19**(2):99-113.
2. Koonin EV, Bork P, Sander C: **Yeast chromosome III: new gene functions**. *EMBO J* 1994, **13**(3):493-503.
3. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling**. *Proc Natl Acad Sci U S A* 1999, **96**(6):2896-2901.
4. Goltsman DS, Denef VJ, Singer SW, VerBerkmoes NC, Lefsrud M, Mueller RS, Dick GJ, Sun CL, Wheeler KE, Zemla A *et al*: **Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing "Leptospirillum rubarum" (Group II) and "Leptospirillum ferrodiazotrophum" (Group III) bacteria in acid mine drainage biofilms**. *Appl Environ Microbiol* 2009, **75**(13):4599-4615.
5. Yelton AP, Thomas BC, Simmons SL, Wilmes P, Zemla A, Thelen MP, Justice N, Banfield JF: **A semi-quantitative, synteny-based method to improve functional predictions for hypothetical and poorly annotated bacterial and archaeal genes**. *PLoS Comput Biol* 2011, **7**(10):e1002230.
6. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families**. *Science* 1997, **278**(5338):631-637.
7. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
8. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**(19):3210-3212.
9. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV: **Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli**. *Curr Biol* 1996, **6**(3):279-291.
10. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S *et al*: **The complete genome sequence of the gram-positive bacterium Bacillus subtilis**. *Nature* 1997, **390**(6657):249-256.
11. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O *et al*: **Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths**. *PLoS Genet* 2009, **5**(1):e1000344.
12. Maddison WP: **Gene Trees in Species Trees**. *Systematic Biology* 1997, **46**(3):523-536.
13. Emms DM, Kelly S: **OrthoFinder: phylogenetic orthology inference for comparative genomics**. *Genome Biol* 2019, **20**(1):238.
14. Reeck GR, de Haen C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH, et al.: **"Homology" in proteins and nucleic acids: a terminology muddle and a way out of it**. *Cell* 1987, **50**(5):667.
15. Owen R: **On the archetype and homologies of the vertebrate skeleton**. London: John Van Voorst; 1848.

16. Boyden A: **Genetics and Homology**. *The Quarterly Review of Biology* 1935, **10**(4):448-451.
17. Boyden A: **Homology and Analogy: A Century After the Definitions of "Homologue" and "Analogue" of Richard Owen**. *The Quarterly Review of Biology* 1943, **18**(3):228-241.
18. Fitch WM: **Homology a personal view on some of the problems**. *Trends Genet* 2000, **16**(5):227-231.
19. Gray GS, Fitch WM: **Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus**. *Mol Biol Evol* 1983, **1**(1):57-66.
20. Sonnhammer EL, Koonin EV: **Orthology, paralogy and proposed classification for paralog subtypes**. *Trends Genet* 2002, **18**(12):619-620.
21. Haldane JBS: **The Part Played by Recurrent Mutation in Evolution**. *The American Naturalist* 1933, **67**(708):5-19.
22. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes**. *Science* 2000, **290**(5494):1151-1155.
23. Conrad B, Antonarakis SE: **Gene duplication: a drive for phenotypic diversity and cause of human disease**. *Annu Rev Genomics Hum Genet* 2007, **8**:17-35.
24. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations**. *Genetics* 1999, **151**(4):1531-1545.
25. Gout JF, Lynch M: **Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization**. *Mol Biol Evol* 2015, **32**(8):2141-2148.
26. Kleinjan DA, Bancewicz RM, Gautier P, Dahm R, Schonthaler HB, Damante G, Seawright A, Hever AM, Yeyati PL, van Heyningen V, Coutinho P: **Subfunctionalization of duplicated zebrafish pax6 genes by cis-regulatory divergence**. *PLoS Genet* 2008, **4**(2):e29.
27. Ohno S: **Evolution by Gene Duplication.** . New York: SpringerVerlag; 1970.
28. Dvornyk V, Vinogradova O, Nevo E: **Long-term microclimatic stress causes rapid adaptive radiation of kaiABC clock gene family in a cyanobacterium, Nostoc linckia, from "Evolution Canyons" I and II, Israel**. *Proc Natl Acad Sci U S A* 2002, **99**(4):2082-2087.
29. Sawada M, Osawa S, Kobayashi H, Hori H, Muto A: **The number of ribosomal RNA genes in Mycoplasma capricolum**. *Mol Gen Genet* 1981, **182**(3):502-504.
30. Ulbrich N, Kumagai I, Erdmann VA: **The number of ribosomal RNA genes in Thermus thermophilus HB8**. *Nucleic Acids Res* 1984, **12**(4):2055-2060.
31. Espejo RT, Plaza N: **Multiple Ribosomal RNA Operons in Bacteria; Their Concerted Evolution and Potential Consequences on the Rate of Evolution of Their 16S rRNA**. *Front Microbiol* 2018, **9**:1232.
32. Koonin EV: **Orthologs, paralogs, and evolutionary genomics**. *Annu Rev Genet* 2005, **39**:309-338.
33. Karplus K, Barrett C, Hughey R: **Hidden Markov models for detecting remote protein homologies**. *Bioinformatics* 1998, **14**(10):846-856.
34. Eddy SR: **A new generation of homology search tools based on probabilistic inference**. *Genome Inform* 2009, **23**(1):205-211.

35. Prestat E, David MM, Hultman J, Tas N, Lamendella R, Dvornik J, Mackelprang R, Myrold DD, Jumpponen A, Tringe SG *et al*: **FOAM (Functional Ontology Assignments for Metagenomes): a Hidden Markov Model (HMM) database with environmental focus**. *Nucleic Acids Res* 2014, **42**(19):e145.
36. Fitch WM: **Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology**. *Systematic Biology* 1971, **20**(4):406-416.
37. Doolittle WF: **Phylogenetic classification and the universal tree**. *Science* 1999, **284**(5423):2124-2129.
38. Kapli P, Yang Z, Telford MJ: **Phylogenetic tree building in the genomic age**. *Nat Rev Genet* 2020, **21**(7):428-444.
39. Young AD, Gillung JP: **Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics**. *Systematic Entomology* 2020, **45**(2):225-247.
40. Page RD, Charleston MA: **From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem**. *Mol Phylogenet Evol* 1997, **7**(2):231-240.
41. Struck TH: **The impact of paralogy on phylogenomic studies - a case study on annelid relationships**. *PLoS One* 2013, **8**(5):e62892.
42. Hellmuth M, Wieseke N, Lechner M, Lenhof HP, Middendorf M, Stadler PF: **Phylogenomics with paralogs**. *Proc Natl Acad Sci U S A* 2015, **112**(7):2058-2063.
43. Smith ML, Hahn MW: **New Approaches for Inferring Phylogenies in the Presence of Paralogs**. *Trends Genet* 2021, **37**(2):174-187.
44. Fulton DL, Li YY, Laird MR, Horsman BG, Roche FM, Brinkman FS: **Improving the specificity of high-throughput ortholog prediction**. *BMC Bioinformatics* 2006, **7**:270.
45. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome Res* 2003, **13**(9):2178-2189.
46. Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, Zdobnov EM: **OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity**. *Nucleic Acids Res* 2023, **51**(D1):D445-D451.
47. Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A: **TreeFam v9: a new website, more species and orthology-on-the-fly**. *Nucleic Acids Res* 2014, **42**(Database issue):D922-925.
48. Sonnhammer EL, Östlund G: **InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic**. *Nucleic Acids Res* 2015, **43**(Database issue):D234-239.
49. Emms DM, Kelly S: **OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy**. *Genome Biol* 2015, **16**(1):157.
50. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simao FA, Zdobnov EM: **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-D811.
51. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ *et al*: **eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses**. *Nucleic Acids Res* 2019, **47**(D1):D309-D314.
52. Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, Thomas PD: **PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API**. *Nucleic Acids Res* 2021, **49**(D1):D394-D403.

53. Altenhoff AM, Train CM, Gilbert KJ, Mediratta I, Mendes de Farias T, Moi D, Nevers Y, Radoykova HS, Rossier V, Warwick Vesztrocy A *et al*: **OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more**. *Nucleic Acids Res* 2021, **49**(D1):D373-D379.
54. Chen F, Mackey AJ, Vermunt JK, Roos DS: **Assessing performance of orthology detection strategies applied to eukaryotic genomes**. *PLoS One* 2007, **2**(4):e383.
55. **NCBI Prokaryotic Genome Annotation Pipeline** [https://github.com/ncbi/pgap]
56. Renwick JH: **The mapping of human chromosomes**. *Annu Rev Genet* 1971, **5**:81-120.
57. Renwick JH: **Progress in mapping human autosomes**. *Br Med Bull* 1969, **25**(1):65-73.
58. Kilian A, Kudrna DA, Kleinhofs A, Yano M, Kurata N, Steffenson B, Sasaki T: **Rice-barley synteny and its application to saturation mapping of the barley Rpg1 region**. *Nucleic Acids Res* 1995, **23**(14):2729-2733.
59. Passarge E, Horsthemke B, Farber RA: **Incorrect use of the term synteny**. *Nat Genet* 1999, **23**(4):387.
60. Walden N, Schranz ME: **Synteny Identifies Reliable Orthologs for Phylogenomics and Comparative Genomics of the Brassicaceae**. *Genome Biol Evol* 2023, **15**(3).
61. **OrthoFinder Readme** [https://github.com/davidemms/OrthoFinder]
62. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC Bioinformatics* 2009, **10**:421.
63. Dongen S: **Graph clustering by flow simulation**. Utrecht University; 2000.
64. Kelly S, Maini PK: **DendroBLAST: approximate phylogenetic trees in the absence of multiple sequence alignments**. *PLoS One* 2013, **8**(3):e58537.
65. Emms DM KS: **STAG: Species Tree Inference from All Genes**. In. bioRxiv; 2018.
66. Emms DM, Kelly S: **STRIDE: Species Tree Root Inference from Gene Duplication Events**. *Mol Biol Evol* 2017, **34**(12):3267-3278.
67. Wu YC, Rasmussen MD, Bansal MS, Kellis M: **Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees**. *Genome Res* 2014, **24**(3):475-486.
68. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T: **The human phylome**. *Genome Biol* 2007, **8**(6):R109.
69. Altenhoff AM, Garrayo-Ventas J, Cosentino S, Emms D, Glover NM, Hernandez-Plaza A, Nevers Y, Sundesha V, Szklarczyk D, Fernandez JM *et al*: **The Quest for Orthologs benchmark service and consensus calls in 2020**. *Nucleic Acids Res* 2020, **48**(W1):W538-W545.
70. Nevers Y, Jones TEM, Jyothi D, Yates B, Ferret M, Portell-Silva L, Codo L, Cosentino S, Marcet-Houben M, Vlasova A *et al*: **The Quest for Orthologs orthology benchmark service in 2022**. *Nucleic Acids Res* 2022, **50**(W1):W623-W632.
71. Lim PK, Davey EE, Wee S, Seetoh WS, Goh JC, Zheng X, Phang SKA, Seah ESK, Ng JWZ, Wee XJH *et al*: **Bacteria.guru: Comparative Transcriptomics and Co-Expression Database for Bacterial Pathogens**. *J Mol Biol* 2022, **434**(11):167380.
72. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs inference projects and methods**. *PLoS Comput Biol* 2009, **5**(1):e1000262.
73. Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T: **A new measure for functional similarity of gene products based on Gene Ontology**. *BMC Bioinformatics* 2006, **7**:302.

74. Saier MH, Jr.: **A functional-phylogenetic classification system for transmembrane solute transporters**. *Microbiol Mol Biol Rev* 2000, **64**(2):354-411.
75. Edgar RC: **Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny**. *Nature Communications* 2022, **13**(1):6968.
76. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014, **30**(9):1312-1313.
77. Team RC: **R: A language and environment for statistical computing.** . In. Vienna, Austria.: R Foundation for Statistical Computing; 2021.
78. Paradis E, Schliep K: **ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R**. *Bioinformatics* 2019, **35**(3):526-528.
79. Miyakoshi M, Chao Y, Vogel J: **Cross talk between ABC transporter mRNAs via a target mRNA-derived sponge of the GcvB small RNA**. *EMBO J* 2015, **34**(11):1478-1492.
80. Walshaw DL, Poole PS: **The general L-amino acid permease of Rhizobium leguminosarum is an ABC uptake system that also influences efflux of solutes**. *Mol Microbiol* 1996, **21**(6):1239-1252.
81. Walshaw DL, Lowthorpe S, East A, Poole PS: **Distribution of a sub-class of bacterial ABC polar amino acid transporter and identification of an N-terminal region involved in solute specificity**. *FEBS Lett* 1997, **414**(2):397-401.
82. Kingston AW, Ponkratz C, Raleigh EA: **Rpn (YhgA-Like) Proteins of Escherichia coli K-12 and Their Contribution to RecA-Independent Horizontal Transfer**. *J Bacteriol* 2017, **199**(7).
83. Yamamoto K, Yata K, Fujita N, Ishihama A: **Novel mode of transcription regulation by SdiA, an Escherichia coli homologue of the quorum-sensing regulator**. *Mol Microbiol* 2001, **41**(5):1187-1198.
84. Ma X, Zhang S, Xu Z, Li H, Xiao Q, Qiu F, Zhang W, Long Y, Zheng D, Huang B *et al*: **SdiA Improves the Acid Tolerance of E. coli by Regulating GadW and GadY Expression**. *Front Microbiol* 2020, **11**:1078.
85. Fife MA, Davis BR, Ewing WH: **The Biochemical reactions of the tribe Klebsielleae**. 1965.
86. Ludwig W KH: **Bergey's manual® of systematic bacteriology**, vol. 2: Springer Science & Business Media; 2001.
87. Gao B, Gupta RS: **Conserved indels in protein sequences that are characteristic of the phylum Actinobacteria**. *International journal of systematic and evolutionary microbiology* 2005, **55**(6):2401-2412.
88. Woese CR: **Bacterial evolution**. *Microbiol Rev* 1987, **51**(2):221-271.
89. Turnbull PCB: **Bacillus**. In: *Medical Microbiology.* Edited by Baron S. Galveston (TX): University of Texas Medical Branch at Galveston
Copyright © 1996, The University of Texas Medical Branch at Galveston.; 1996.
90. Drews SJ, Hung F, Av-Gay Y: **A protein kinase inhibitor as an antimycobacterial agent**. *FEMS Microbiol Lett* 2001, **205**(2):369-374.
91. Fernandez P, Saint-Joanis B, Barilone N, Jackson M, Gicquel B, Cole ST, Alzari PM: **The Ser/Thr protein kinase PknB is essential for sustaining mycobacterial growth**. *J Bacteriol* 2006, **188**(22):7778-7784.

92. Kang CM, Abbott DW, Park ST, Dascher CC, Cantley LC, Husson RN: **The Mycobacterium tuberculosis serine/threonine kinases PknA and PknB: substrate identification and regulation of cell shape**. *Genes Dev* 2005, **19**(14):1692-1704.
93. Jones G, Del Sol R, Dudley E, Dyson P: **Forkhead-associated proteins genetically linked to the serine/threonine kinase PknB regulate carbon flux towards antibiotic biosynthesis in Streptomyces coelicolor**. *Microb Biotechnol* 2011, **4**(2):263-274.
94. Ogawara H: **Self-resistance in Streptomyces, with Special Reference to beta-Lactam Antibiotics**. *Molecules* 2016, **21**(5).
95. Ogawara H: **Distribution of PASTA domains in penicillin-binding proteins and serine/threonine kinases of Actinobacteria**. *J Antibiot (Tokyo)* 2016, **69**(9):660-685.
96. Yeats C, Finn RD, Bateman A: **The PASTA domain: a beta-lactam-binding domain**. *Trends Biochem Sci* 2002, **27**(9):438.
97. Narayan A, Sachdeva P, Sharma K, Saini AK, Tyagi AK, Singh Y: **Serine threonine protein kinases of mycobacterial genus: phylogeny to function**. *Physiol Genomics* 2007, **29**(1):66-75.
98. Li G, Ma Q, Mao X, Yin Y, Zhu X, Xu Y: **Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes**. *Nucleic Acids Res* 2011, **39**(22):e150.
99. Solis-Escalante D, Kuijpers NG, Barrajon-Simancas N, van den Broek M, Pronk JT, Daran JM, Daran-Lapujade P: **A Minimal Set of Glycolytic Genes Reveals Strong Redundancies in Saccharomyces cerevisiae Central Metabolism**. *Eukaryot Cell* 2015, **14**(8):804-816.
100. Notebaart RA, Huynen MA, Teusink B, Siezen RJ, Snel B: **Correlation between sequence conservation and the genomic context after gene duplication**. *Nucleic Acids Res* 2005, **33**(19):6164-6171.
101. Hillis DM, Bull JJ: **An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis**. *Systematic Biology* 1993, **42**(2):182-192.
102. Hennig W: **Phylogenetic Systematics.** , 1st edn: University of Illinois Press; 1966.
103. Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer EL: **InParanoid 7: new algorithms and tools for eukaryotic orthology analysis**. *Nucleic Acids Res* 2010, **38**(Database issue):D196-203.
104. Schmitt T, Messina DN, Schreiber F, Sonnhammer EL: **Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information**. *Brief Bioinform* 2011, **12**(5):485-488.
105. **OrthoXML & SeqXML** [https://www.orthoxml.org/xml/Main.html]
106. **OrthoXML Documentation** [https://orthoxml.org/0.4/orthoxml_doc_v0.4.html]
107. **OMA Database** [https://omabrowser.org/oma/current/]
108. Birkeland NK, Schonheit P, Poghosyan L, Fiebig A, Klenk HP: **Complete genome sequence analysis of Archaeoglobus fulgidus strain 7324 (DSM 8774), a hyperthermophilic archaeal sulfate reducer from a North Sea oil field**. *Stand Genomic Sci* 2017, **12**:79.
109. **Orthology Benchmarking** [https://orthology.benchmarkservice.org]
110. **OMA browser** [https://omabrowser.org/oma/pps/23663/]
111. Reeve JN, Beckler GS, Cram DS, Hamilton PT, Brown JW, Krzycki JA, Kolodziej AF, Alex L, Orme-Johnson WH, Walsh CT: **A hydrogenase-linked gene in**

Methanobacterium thermoautotrophicum strain delta H encodes a polyferredoxin. *Proc Natl Acad Sci U S A* 1989, **86**(9):3031-3035.
112. Catchen JM, Conery JS, Postlethwait JH: **Automated identification of conserved synteny after whole-genome duplication**. *Genome Res* 2009, **19**(8):1497-1505.
113. Jun J, Mandoiu, II, Nelson CE: **Identification of mammalian orthologs using local synteny**. *BMC Genomics* 2009, **10**:630.
114. Georgescu CH, Manson AL, Griggs AD, Desjardins CA, Pironti A, Wapinski I, Abeel T, Haas BJ, Earl AM: **SynerClust: a highly scalable, synteny-aware orthologue clustering tool**. *Microb Genom* 2018, **4**(11).

APPENDIX A

Ortholog Benchmarking Commands

This appendix contains the commands used to generate the data to benchmark OrthoFinder and OrthoRefine using the community standard benchmark (see Chapter 2).

```
# Download data files from Benchmark website
ftp://ftp.ebi.ac.uk/pub/databases/reference_proteomes/previous_releases/qfo_release-2020_04_with_updated_UP000008143/QfO_release_2020_04_with_updated_UP000008143.tar.gz
# move benchmark files to own directory, note you have to change file path where you want to move files
ls *.fasta | grep -v "DNA" | grep -v "addit" | xargs -I {} mv {} ../../benchmark_to_submit/bacteria/
# need to change fasta header line OrthoRefine's pattern matching
for thing in *.fasta; do sed 's/^>..|/>/' $thing | sed 's/|/ |/' > temp.fasta; mv temp.fasta $thing; done
# Use README file to look up GCF accession for benchmarking dataset
# Included at the end of this file is a copy/paste file with input for using download script and input with GCF, GCA, UP, and species name
# Download GCF using OrthoRefine's download script
./download_ft_fafiles "$file_name"
# Download GCA files manually
```

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/

Rename feature table file so they don’t interfere with OrthoRefine when running benchmark on their data

for thing in *feature_table.txt; do mv $thing "${thing%%????????????????}ft.txt"; done

Compile support script convert_locus_tag_to_prot_id.cpp as convert_locus_tag_to_prot_id.exe

g++ convert_locus_tag_to_prot_id.cpp -o convert_locus_tag_to_prot.exe

Perform conversion on REFSEQ feature table to change to uniport ID

for thing in GCF*_ft.txt; do thing2="GCA${thing##???}"; thing3="${thing2%_*}";

thing4="${thing3%_*}".txt; ./convert_locus_tag_to_prot.exe $thing $thing4; done

rename fasta file to match OrthFinder output to OrthoRefine input, note detailed input file which is included at the end of this file

cat detailed_input_all_bacteria_benchmark.txt | awk '{print $3"* "$1".fasta"}' > move_list

while read line; do mv $line; done < move_list

Move .faa files to different directory so OrthoFinder doesn’t use them

mv *.faa /path/to/dir

Run OrthoFinder

/path/to/OrthoFinder/./orthofinder -f ./

use dos2unix on OrthoFinder output

dos2unix N0.tsv

Generate OrthoRefine results, also get lines that are extra to remove from OrthoFinder output to match OrthoRefine’s dataset which has missing data (difference between NCBI and uniprot as NCBI is 2023 data and uniprot is 2020, can’t find NCBI 2020 data backup)

./OrthoRefine.exe -input input_all_bacteria_benchmark.txt -OF_file N0.tsv -window_size 8 -synteny_ratio 0.5 -benchmark 1 -run_all 1 -print_all_orthofinder 1 | grep "prod" | cut -d " " -f3 > list_to_remove_from_orthofinder_as_no_ft_match.txt

Convert OrthoFinder output file to 2 columns for submission

./convert_orthofinder_out_to_benchmark_submit.exe N0.tsv > 16_bac_orthofinder.tsv

remove extra lines of data from OrthoFinder's output that are not in the feature table to submit to benchmark. Keeps dataset the same between OrthoFinder and OrthoRefine. Slow but works

while read line; do sed -i "/$line/d" ./16_bac_orthofinder.tsv; done < list_to_remove_from_orthofinder_as_no_ft_match.txt

Uniq to remove duplicates from OrthoRefine output

cat OrthoRefine_outfile | sort | uniq > sorted_OrthoRefine_outfile

Submit to Ortholog benchmarking service

input file

GCF_000195955.2

GCF_000006765.1

GCF_000008545.1

GCF_000008725.1

GCF_000203835.1

GCF_000092565.1

GCF_000005845.2

GCF_000008805.1

GCF_000008565.1

GCF_000011365.1

GCF_000009725.1

GCF_000018865.1

GCF_000009045.1

GCF_000011385.1

GCF_000008625.1

GCF_000008525.1

GCF_000007325.1

GCF_000196115.1

GCF_000007985.2

GCF_000027325.1

GCF_000021645.1

GCF_000011065.1

GCF_000020985.1

#detailed input file

GCF_000195955.2 GCA_000195955.2 UP000001584 Mycobacterium tuberculosis strain H37RV

GCF_000006765.1 GCA_000006765.1 UP000002438 Pseudomonas aeruginosa PAO1

GCF_000008545.1 GCA_000008545.1 UP000008183 Thermotoga maritima MSB8

GCF_000008725.1 GCA_000008725.1 UP000000431 Chlamydia trachomatis (strain D/UW-3/Cx)

GCF_000203835.1 GCA_000203835.1 UP000001973 Streptomyces coelicolor (strain ATCC BAA-471 / A3(2) / M145) surpressed by REFSEQ

GCF_000092565.1 GCA_000092565.1 UP000001408 Leptospira interrogans serogroup Icterohaemorrhagiae serovar Lai (strain 56601)

GCF_000005845.2 GCA_000005845.2 UP000000625 Escherichia coli (strain K12)

GCF_000008805.1 GCA_000008805.1 UP000000425 Neisseria meningitidis serogroup B (strain MC58)

GCF_000008565.1 GCA_000008565.1 UP000002524 Deinococcus radiodurans (strain ATCC 13939 / DSM 20539 / JCM 16871 / LMG 4051 / NBRC 15346 / NCIMB 9279 / R1 / VKM B-1422)

GCF_000011365.1 GCA_000011365.1 UP000002526 Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)

GCF_000009725.1 GCA_000009725.1 UP000001425 Synechocystis sp. (strain PCC 6803 / Kazusa)

GCF_000018865.1 GCA_000018865.1 UP000002008 Chloroflexus aurantiacus (strain ATCC 29366 / DSM 635 / J-10-fl)

GCF_000009045.1 GCA_000009045.1 UP000001570 Bacillus subtilis (strain 168)

GCF_000011385.1 GCA_000011385.1 UP000000557 Gloeobacter violaceus (strain ATCC 29082 / PCC 7421)

GCF_000008625.1 GCA_000008625.1 UP000000798 Aquifex aeolicus (strain VF5)

GCF_000008525.1 GCA_000008525.1 UP000000429 Helicobacter pylori (strain ATCC 700392 / 26695) (Campylobacter pylori) surpressed by REFSEQ

GCF_000007325.1 GCA_000007325.1 UP000002521 Fusobacterium nucleatum subsp. nucleatum (strain ATCC 25586 / CIP 101130 / JCM 8532 / LMG 13131) surpressed by REFSEQ

GCF_000196115.1 GCA_000196115.1 UP000001025 Rhodopirellula baltica (strain DSM 10527 / NCIMB 13988 / SH1)

GCF_000007985.2 GCA_000007985.2 UP000000577 Geobacter sulfurreducens (strain ATCC 51573 / DSM 12127 / PCA)

GCF_000027325.1 GCA_000027325.1 UP000000807 Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195)

GCF_000021645.1 GCA_000021645.1 UP000007719 Dictyoglomus turgidum (strain Z-1310 / DSM 6724)

GCF_000011065.1 GCA_000011065.1 UP000001414 Bacteroides thetaiotaomicron (strain ATCC 29148 / DSM 2079 / NCTC 10582 / E50 / VPI-5482)

GCF_000020985.1 GCA_000020985.1 UP000000718 Thermodesulfovibrio yellowstonii (strain ATCC 51303 / DSM 11347 / YP87)

APPENDIX B

Generating Phylogenetic Trees Commands

This appendix contains the commands used to align the fasta files, build the phylogenetic trees, and plot the phylogenetic tree figures (see Chapter 2).

```
#Bash
# align with muscle 5
./muscle5.1.linux_intel64 -align "$sequence.fasta" -output "$aligned.fasta"
# build tree with raxml
raxml/standard-RAxML-master/./raxmlHPC-PTHREADS-AVX -T 10 -f a -m
PROTGAMMAAUTO -p 12345 -x 12345 -o STM0828 -# 1000 -s "$aligned.fasta" -n
"$hog_number_tree"
```

```
#R
library(ape)
setwd() # set working directory
data1 <- read.tree("RAxML_bipartitions.hog_number_tree")
plot.phylo(data1, show.node.label = TRUE)
add.scale.bar()
```

Tree figures have been rotated about nodes to place certain groups next to other groups.

APPENDIX C

Example Feature Table File and OMA Database File

This appendix contains an example RefSeq feature table file (genome annotation) and two example files from the OMA database, oma-groups.orthoXML, which does not contain paralogs, and oma-hogs.orthoXML.xml, which does contain paralogs. The tab-delimited feature table file displays a single gene and protein product (CDS) with the remaining genes and proteins omitted. The product accession is bolded, the gene name is bold with italics, the locus tag is bold with solid underline, and the attributes are bold with dashed underline; the other columns, denoted by #...#, have been omitted. A single gene and ortholog group are displayed for the oma-groups and oma-hogs files, with the remaining genes and ortholog groups omitted. The geneId is bolded with a dashed underline, the protId (a unique OMA identifier that cannot be referenced to other databases) is unbolded with a solid underline, and the gene id & the geneRef id is unbolded with italics. The OrthoXML conversion script (orthoxml_convert.cpp, See Chapter 3) requires all three files. The conversion is achieved by matching the protId information between the two OMA files, storing the geneId from the OMA-group file for matching with the feature table later, and matching the gene id from the start of the oma-group file to the geneRef id in the orthologGroup at the end of the oma-group file – all genes contained within the orthologGroup that were from the genomes specified by the user were grouped into a hog. The geneId from the oma-group file is then compared to the attributes, locus_tag, and name column of the feature table to match one of the columns; if no match can be found, the geneId from OMA is printed in the output. If a match can be found, the locus_tag column between the gene and CDS lines is

verified as matching, and then the information from the product_accession column is printed. The conversion output is a tab-delimited HOG file that matches OrthoFinder's formatting, which is the format OrthoRefine requires as input.

GCF_000008665.1_ASM866v1_feature_table.txt

# feature	2…10	**product_accession**	12…13	***name***	15…16	**locus_tag**	18…19	**attributes**
gene	2…17					**AF_RS00030**	18…19	**AF0008,AF_0008**
CDS	2…10	**WP_010877522.1**	12…13	***MFS transporter***	15…16	**AF_RS00030**	18…20	

…

oma-groups.orthoXML.xml

<species name="Archaeoglobus fulgidus (strain ATCC 49558 / DSM 4304 / JCM 9628 / NBRC 100126 / VC-16)" NCBITaxId="-627288153">

<database name="Genome Reviews" version="01-SEP-2009 (Rel. 110, Last updated, Version 111)">

<genes>

…

<gene id="*17953*" geneId="**AF_0008**" protId="ARCFU00008"/>

…

</genes>

</database>

</species>

oma-hogs.orthoXML

<species name="Archaeoglobus fulgidus (strain ATCC 49558 / DSM 4304 / JCM 9628 / NBRC 100126 / VC-16)" NCBITaxId="-627288153">

<database name="Archaeoglobus fulgidus (strain JCM 9628 / DSM 4304 / VC-16 / ATCC 49558 / NBRC 100126) chromosome, complete sequence." version="01-SEP-2009 (Rel. 110, Last updated, Version 111)">

<genes>

…

<gene id="*17953*" protId="ARCFU00008"/>

…

</genes>

</database>

</species>

…

<groups>

<orthologGroup id="####">

```
            <geneRef id="17953”/>
            <geneRef id="#####”/>
            <geneRef id="#####”/>
            <geneRef id="#####”/>
        </orthologGroup>
        …
</groups>
```

APPENDIX D

OrthoXML Conversion and Additional Summary Statistic Commands

```
# To perform conversion from OrthoXML to OrthoRefine input
# taxid.txt is a single-column file that contains the taxid from the OMA file for the genome
(NOT NCBI TaxId)
./orthoxml_convert.exe --input input.txt --ncbiTaxId taxid.txt --orthoxmlGroup oma-
groups.orthoXML.xml --orthoxmlHog oma-hogs.orthoXML --Output output.txt

# To obtain AMNOG
./OrthoRefine.exe --input input.txt --OF_file OF_file.tsv --window_size 8 --synteny_ratio 0.5 --
run_combo

# To obtain summary stats
./summary_stats.sh --exe ./OrthoRefine.exe --window_size 8 --synteny_ratio 0.5 --input input.txt
--OF_file OF_file.tsv

# To count number of paralogs in OrthoFinder format
less N0.tsv | grep -o ',' | wc -l

# To count number of genes in OrthoFinder format
less N0.tsv | cut -f 4- | tail -n +2 | grep -o '_' | wc -l
```