OrthoPhyl – A turn-key solution for large scale whole genome bacterial phylogenomics

Earl A Middlebrook[1], Robab Katani[2], Jeanne M Fair[1]
[1] Genomics and Bioanalytics Group, Los Alamos National Lab
[2] Huck Institutes of Life Sciences, Pennsylvania State University

Abstract

Motivation – There are a staggering number of publicly available bacterial genome sequences (at writing, 1.5 million assemblies in NCBI's GenBank alone), and the deposition rate continues to increase. This wealth of data begs for phylogenetic analyses to place these sequences within an evolutionary context. A phylogenetic placement not only aids in taxonomic classification, but informs the evolution of novel phenotypes, targets of selection, and horizontal gene transfer. Building trees from multi-gene codon alignments is a laborious task that requires bioinformatic expertise, rigorous curation of orthologs, and heavy computation. Compounding the problem is the lack of tools that can streamline these processes for building trees from large scale genomic data.

Results - Here we present OrthoPhyl, which takes as input bacterial genome assemblies and reconstructs trees from whole genome codon alignments. The workflow can analyze an arbitrarily large number of input genomes (>1200 tested here) by identifying a diversity spanning subset of assemblies and using these genomes to build gene models to identify orthologs in the full dataset. OrthoPhyl is an easy-to-run, turn-key solution for generating high-resolution bacterial trees from datasets with a wide range of divergences. To illustrate the versatility of OrthoPhyl, we show three use-cases: E. coli/Shigella, Brucella/Ochrobactrum, and the order Rickettsiales. We compare trees generated with our

software to published trees from alternative methods. We show that OrthoPhyl trees are consistent with other methods while incorporating more data, allowing for greater numbers of input genomes, and more flexibility of analysis.

Availability and Implementation - Code used in this manuscript is available at *https://github.com/eamiddlebrook/OrthoPhyl.git*. To aid in usability, a Singularity container is available at *https://cloud.sylabs.io/library/earlyevol/default/orthophyl*. Installation and execution instructions are provided in the associated github readme file. Assemblies used with this manuscript are available from https://www.ncbi.nlm.nih.gov/assembly/ with accessions in **Supplementary table 1-3**.

Contact – Earl A. Middlebrook – earlm@lanl.gov

1 Introduction

Underpinning every aspect of an organism's biology is its evolutionary history. Phylogenetic methods strive to reconstruct these histories using genetic data. With the advent of "Next-" and "Third-gen" sequencing methodologies, the amount of genetic data in public databases has increased dramatically. At writing there are currently 1,502,852 bacterial genome assemblies available through NCBI alone (ncbi.nlm.nih.gov/assembly). The wealth of available genetic data has transformed the fields of bacterial phylogenetics and taxonomy (Konstantinidis and Tiedje, 2007; Jain *et al.*, 2018; Varghese *et al.*, 2015). Unfortunately, most whole genome phylogenetic analysis methods require specialized bioinformatic data processing and expertise (Smith, 2013). To compound the issue, many available large-scale trees are under-resolved (NCBI taxonomy) or are generated with methods unsuited for

48 narrow and broad evolutionary distances (Hördt *et al.*, 2020) reducing their utility at multiple
49 evolutionary scales. This flood of genetic data necessitates an easy-to-use phylogenetic
50 analysis pipeline to help reveal the evolutionary context for the myriad of sequenced
51 bacteria.

52 **1.1 Analysis options**

53 **1.1.1 Single and Multi-locus**

54 A common way to generate bacterial trees is to compare single homologous loci (e.g.,
55 16S ribosomal, recA, gyrA, rpoB, or dnaK genes). This method has advantages which include
56 the ability to use conserved primer sites to amplify sequences, very low computational
57 intensity, and the availability of curated databases (SILVA, NCBI GenBank, etc.). Using genes
58 other than rDNA allows for greater resolution at certain evolutionary distances but still lacks
59 broad resolution (i.e., Near resolution for swiftly evolving genes, far for conserved) (Yang,
60 1998). These trees are prone to misleading topologies due to horizontal gene transfers,
61 incomplete lineage sorting, paralogous gene conversions, and limited phylogenetic
62 information (Huerta-Cepas *et al.*, 2007). Using multi-locus alignments to build phylogenetic
63 trees increases the total number of phylogenetically informative sites (more sequence used),
64 reduces the effects of incongruent topologies, and widens the breadth of resolvable
65 evolutionary resolution (Gontcharov *et al.*, 2004). However, in multi-locus analyses, the
66 effort and cost of primer design and amplification optimization scales with the number of
67 sequences used, which can rapidly become prohibitive.

68 **1.1.3 Whole genome**

69 Classically, whole genome alignments were required to identify Single Nucleotide
70 Polymorphisms (SNPs) of assemblies (see Shakya *et al.*, 2020; Treangen *et al.*, 2014 for
71 method details). These methods add considerable phylogenetic data for tree building and
72 can produce different topologies than multi-locus alignments, illustrating their utility (Baltrus
73 *et al.*, 2014). A major drawback of whole genome alignments is that they tend to fail for
74 genomes with <85% identity depending on the aligner used (Shakya *et al.*, 2020). Thus,
75 phylogenetic information drops quickly as a function of evolutionary distance. The practical
76 result is a lack of resolution at even moderate evolutionary distances. Additionally, having
77 the "missing" data enriched for the more divergent sites will possibly lead to an
78 underestimation of the sequence to reference evolutionary distances for more divergent
79 sequences (Bertels *et al.*, 2014; Spencer *et al.*, 2007; Shavit Grievink *et al.*, 2013).

80 **1.1.4 K-mer overlap**

81 An emergent method to quickly estimate evolutionary histories from genome
82 assemblies is to use *k*-mer content. This involves identifying subsequences of length *k* within
83 each assembly and calculating the shared *k*-mer content (for details see Bussi *et al.*, 2021).
84 The advantage of this class of methods is that they are extremely fast relative to alignment-
85 based methods. The comparison of *k*-mer overlap between sequences naturally leads to
86 pairwise distance matrices, thus, neighbor-joining trees can be generated directly from the
87 results (Saitou and Nei, 1987). Crucially, *k*-mer distances quickly become non-linearly
88 associated with percent identity, leading to erroneous estimated branch lengths (Jain *et al.*,
89 2018).

90 An interesting advance in *k*-mer based approaches is kSNP (Gardner and Hall, 2013;
91 Gardner *et al.*, 2015), which identifies SNPs by flanking conserved *k*-mers. This approach
92 allows the use of tree building methods including parsimony, maximum likelihood, and
93 Baysian. Like whole genome alignments and *k*-mer overlap methods, kSNP suffers from a
94 rapid decline in phylogenetically informative data as sequences diverge. Additionally, the
95 signal/noise ratio degrades quickly because non-homologous *k*-mers are increasingly
96 common as sequences diverge (Gardner and Hall, 2013).

97 **1.1.5 Whole transcriptome/proteome alignments**

98 To use the wealth of data generated by NGS methods and simultaneously alleviate the
99 problem of whole genome alignments, whole transcriptome/proteome alignment can be
100 used. This method breaks the problem of identifying homologous loci into two tractable
101 parts: gene identification, then identification of homologous sequences. This approach
102 identifies many phylogenetically informative sites (like whole genome), but alignments are
103 more tractable (like single loci). However, gene-based alignments are negatively affected by
104 paralogs, i.e., duplicated genes or contamination. Thus, most methods filter out gene
105 families with paralogs. However, see ASTRAL-Pro for a method that is "paralog aware"
106 (Zhang *et al.*, 2020).

107 A trade-off arises while using predicted transcript/protein sequences to build species
108 trees. Transcripts have the most phylogenetic information (silent mutation) but are hard to
109 align at high divergencies (States *et al.*, 1991). Proteins remain alignable at high divergences
110 but lack the information content of transcripts. This can be addressed by converting protein
111 alignments to corresponding codon alignments, leveraging nucleotide phylogenetic

information and protein alignment accuracy (Wernersson and Pedersen, 2003; Bininda-Emonds, 2005).

1.2 Benefits of an automated workflow

The skills required to generate bacterial phylogenetic trees from whole genomes are extensive. Some of the many steps in building robust species trees from whole genome assemblies include gathering and annotating assemblies, identifying and filtering orthologs, sequence alignment, trimming and concatenation, and finally, tree inference (Ashford *et al.*, 2020). Each one of these steps, especially for large data sets, requires expertise in picking parameters, file management, and data format manipulation and filtering. Most of these steps also require familiarity with UNIX command-line.

The phylogenetic context of organisms is becoming increasingly important for standard molecular and evolutionary studies. Even with the available data and a clear need for the phylogenetic placement of focal study species, technical barriers preclude many researchers from inferring their evolutionary histories. An easy-to-use phylogenomic pipeline will encourage wide adoption of phylogenomic methods to complement ongoing environmental and clinical microbiological research, and perhaps help standardize the estimation of phylogenetic trees within and between research labs.

To meet this demand, we present OrthoPhyl, a phylogenomic pipeline that takes bacterial genomes as input, annotates them, identifies orthologs, converts protein to nucleotide alignments, and builds species trees with both concatenated alignments and gene tree to species tree reconciliation. The workflow accepts an arbitrarily large number of input genomes (tested up to 1200). To speed up analysis, a subset of samples representing the

134 whole dataset's diversity are identified, their proteomes are used to build hidden Markov
135 models and then identify orthologs in the full dataset. This strategy allows the generation of
136 trees for 689 Brucella assemblies in ~48hrs using 30 cpus and 58.3GB memory with no
137 hands-on time required. This pipeline is designed to be an easy to install, scalable, turn-key
138 solution for generating high-resolution bacterial trees from diverse clades.
139

140 **2 Methods**

141 **2.1 Main Pipeline**

142 OrthoPhyl's workflow can be broken down into four main steps. Genome assemblies are
143 annotated (**Figure 2.A**), resulting proteins are assigned to orthogroups (**Figure 2.B**),
144 orthogroup proteins are aligned and converted to codon alignments (**Figure 2.C**), and finally,
145 concatenated codon alignments are used to infer phylogenetic trees (**Figure 2.D**).

146 **2.1.1 Annotation**

147 OrthoPhyl starts by annotating assemblies with Prodigal (Hyatt *et al.*, 2010). This
148 provides both protein and transcript sequences that are used in subsequent steps.
149 Redundant transcripts (> 99.9% nucleotide identity) found in a genome are removed using
150 bbmap's *dedupe* (BBMap, 2022) with the logic they are likely the result of recent gene
151 duplications and provide no or little phylogenetic information. This removes unnecessary

152 paralogs from the dataset to preserve the usefulness of the containing orthogroup for

153 downstream analysis.

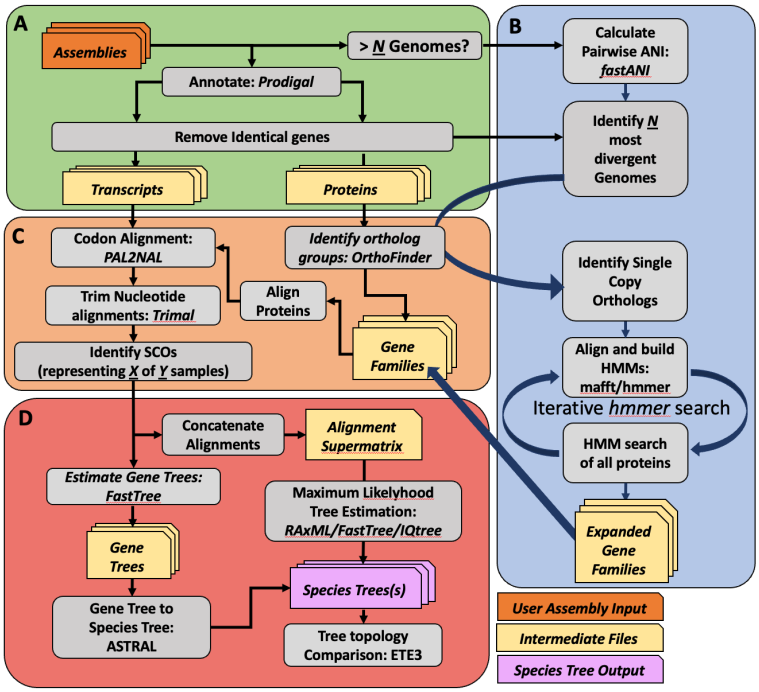

Figure 1 - Visual workflow diagram of OrthoPhyl. Grey boxes indicate processes. Programs used in each step are listed unless a custom script was used. Orange, tan, and purple boxes represent user input, intermediate files, and species tree outputs, respectively. The workflow is divided into four main tasks **A)** annotate assemblies and remove identical transcripts, **B)** if more than "N" assemblies are to being analyzed, identify orthologs in a diversity spanning subset of assemblies, then expand to the full dataset with iterative HMM searches, **C)** align protein sequences and generate codon alignments, finally **D)** estimate species tree topologies with concatenated codon alignment supermatrix along with a gene tree to species tree consensus method.

154 **2.1.2 Orthogroup assignment**

155 To identify orthogroups (OGs) to use for tree generation, *OrthoFinder* is used (Emms and

156 Kelly, 2015, 2019). When trees are being generated for large numbers (default: > 30) of

157 genome assemblies, finding orthogroups directly becomes intractable. Thus, OrthoPhyl

158 identifies a diversity spanning subset of assemblies to identify OGs. Briefly, average

159 nucleotide identity (ANI) is estimated for all assembly pairs with *FastANI* (Jain *et al.*, 2018). A

160 custom hierarchical clustering-like algorithm is used to cluster assemblies based on pairwise

161 ANI values, then single representatives for each of the N groups (default: 30) are chosen as

162 input for OrthoFinder. *FastANI* fails to give an ANI percent for genome comparisons with ANI

163 less than ~75%. For these instances, an arbitrary value of 50% is used. The representative

164 genome picking will be unaffected if there are less than 30 clades at this or greater distances

165 to all other clades. Even with several clades at greater than 75% ANI, the ultimate list of OGs
166 is unlikely to be affected.

167 To expand identified OGs to the full assembly dataset, OG proteins are realigned with
168 *Mafft* (Katoh and Standley, 2013), then *hmmer* (Eddy, 2008, 2009, 2011) generates hidden
169 Markov models and searches against all predicted proteins. OrthoPhyl optimizes the HMM
170 hit score cutoffs to maximize the number of taxa with proteins in the OG, while removing
171 paralogous hits. The expanded protein set for each OG is then realigned and HMMs are
172 again generated and searched against all proteins to capture additional, more divergent
173 orthologues.

174 **2.1.3 Codon Alignment generation**

175 Once OGs are identified for the full set of assemblies, protein sequences are realigned
176 with *Mafft* (Katoh and Standley, 2013). These alignments are then used as the basis for
177 codon alignment using *PAL2NAL* (Suyama *et al.*, 2006). Codon alignments are then trimmed
178 with *TRIMAL* (Capella-Gutiérrez *et al.*, 2009) to reduce noise associated with alignment
179 edges. *Alignment_Assessment* (Portik *et al.*, 2016), is used to assess the quality of codon
180 alignments, including their phylogenetic signals, taxa per alignment, and alignments per taxa
181 **(Figure 2)**. This is critical for ensuring high-quality data are used for generating trees and can
182 aid in troubleshooting phylogenies with low branch support.

183 Our tool then classifies OGs as relaxed or strict single-copy orthologs (SCOs). Strict SCOs
184 are genes found in every taxon with no paralogs, while relaxed SCOs are found in X% of

assemblies still with no paralogs, with X being a tunable parameter (default:30%). Moving forward, these two sets of OGs are used in parallel to generate species trees.

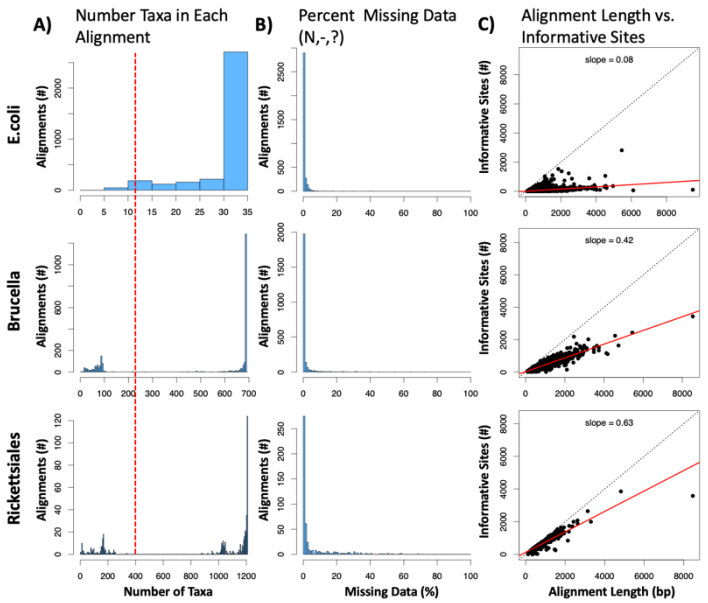

Figure 2 – Alignment metric visualization generated by OrthoPhyl. A) Histograms of the number of taxa in each alignment is shown. Dotted red line indicates the >30% taxa cutoff for alignments used to generate the "SCO relaxed" trees. B) Amount of missing data (including gaps and ambiguous bases) per alignment is shown with a historgram. This is exclusive of genomes without detected orthologs. C) The number of phylogenetically informative sites (variation in > 1 taxa) compared to alignment length for each orthogroup are shown. A linear regression is shown with a solid red line with associated slopes. A one-to-one relationship is shown as a black dotted line.

2.1.4 Species Tree Estimation

Two methods are used to generate species trees for the input assemblies, gene-tree to species-tree estimation with *ASTRAL-III* (Zhang *et al.*, 2018) and gene alignment supermatrix to maximum likelihood (ML). First, to use *ASTRAL*, ML gene-trees are generated with either *FastTree* (Price *et al.*, 2010) or *IQtree* (Nguyen *et al.*, 2015; Kalyaanamoorthy *et al.*, 2017; Hoang *et al.*, 2018) per user input. Then strict and relaxed SCO trees are used as input to *ASTRAL* to generate species trees. Trees are also generated from concatenated codon alignments and maximum likelihood methods. The user can specify any combination of *IQtree*, *FastTree* or *RAxML* (Stamatakis, 2014) to infer tree structure. Finally, these trees are compared with generalized Robinson-Foulds metrics provided by ETEtoolkit (Huerta-Cepas *et al.*, 2007).

2.2 Datasets

199 As a proof of concept, OrthoPhyl was used to build trees for three bacterial clades. We
200 analyzed well characterized *E. coli*/*Shigella* strains. These included a list of 34 complete
201 genomes analyzed in (Shakya *et al.*, 2020), obtained from NCBI.
202 The genus *Brucella* was used to show OrthoPhyl's ability to resolve very closely related
203 sequences while dealing with very long relative branch lengths in the same tree
204 (*Ochrobactrum* group). Assemblies were acquired using NCBI Taxon number 234 with the
205 utility script "*gather_genomes.sh*" (supp. Methods - Gathering and Filtering Assemblies and
206 **Supplementary Figure 1**) which is packaged with OrthoPhyl
207 (github.com/eamiddlebrook/OrthoPhyl/utils/). This resulted in 689 assemblies after filtering
208 for >98% completion, < 0.1% duplication, and < 1% contamination with CheckM (Parks *et al.*,
209 2015). It is important to note that *Ochrobactrum* species were recently moved to the genus
210 *Brucella* (Hördt *et al.*, 2020). However, for clarity and due to controversy within the field, we
211 chose to keep the *Ochrobactrum* labeling for these species. An assembly of *Mycoplana*
212 *dimorpha* (GCA_003046475.1) was added to the dataset as an outgroup.
213 Finally, a tree for all NCBI Rickettsiales assemblies was generated to illustrate
214 OrthoPhyl's utility in dealing with bacterial order level divergences. For Rickettsiales
215 assemblies, we again used *gather_genomes.sh,* this time with NCBI taxon number 766. To
216 illustrate a more challenging scenario, filtering of Rickettsiales assemblies was less stringent
217 with cut-offs at > 95% completion, < 0.2% duplication, and < 1.5% contamination, again
218 using CheckM (Parks *et al.*, 2015). This resulted in 1202 assemblies. Seven *Pelagibacter*
219 assemblies were added to this dataset to serve as an outgroup. Accessions and stats for all
220 genome assemblies used in this manuscript are available in **Supplementary Table 1-3**.

221

222 **3 Implementation**

223 **3.1 Small Tree of Closely Related Assemblies: *E. coli and Shigella***

224 To illustrate Orthophyl's ability to build trees from moderate numbers of samples on a
225 local Linux computer, we ran our *E. coli* analysis (34 genomes) on a desktop with a 12-core
226 Intel processor with 16gb ram running RHEL8 (centOS8). Run statistics are provided in **Table**
227 **2**. The full pipeline took less than 4 hours to build 2 maximum likelihood trees with *RAxML*
228 and 2 ASTRAL trees, one each from the strict and relaxed SCO gene sets.

229 OrthoPhyl identified 2142 strict and 3426 relaxed SCOs (found in 10 genomes) in the
230 analyzed *E. coli* and *Shigella* assemblies. Alignment of these sequences constitutes 2.04 and
231 3.19 MB with 6 and 7.2% phylogenetically informative sites for the SCO and relaxed SCO
232 sets, respectively (**Table 1**).

233 The resulting tree build by FastTree2 from the SCO-relaxed dataset shows successful
234 differentiation of the *E. coli* phylotypes **(Figure 3)**. This tree shows B1 and A as sister
235 phylotypes, with *S. sonnel, S. boydii* and *S. flexneri* between them. That clade is in turn next
236 to phylotype E and *S. dysenteriae*, then D1 and D2. Finally, B2 is sister to all of them. This
237 topology is identical to the reported tree from (Shakya *et al.*, 2020) using whole genome
238 alignments, except for the placement of the root (*E. fergusonii)*, where they have the root
239 placed between D2/B2 and the rest while OrthoPhyl's tree indicates it should be placed
240 between B2 and the other phylotypes.

241

Table 1 - OrthoPhylo output, runtime, ram usage, storage

	Assembly Data Set		
Metric	*E. coli + Shigella*	*Brucella*	Rickettsiales
Number of assemblies	34	689	1208
Number SCOs - strict	2142	897	42
Total Length (bp)	2,040,000	855,924	29,248
% Informative SNPs	6	41.4	81.5
% missing data	0.3	0.5	2.3
ANI % (avg/med/min)*	97.6/99.2/89.9	95.4/99.8/63.7	83.2/57.3/48.1
Number SCOs – relaxed	3426	1762	348
Total Length (bp)	3,190,000	1,627,004	295,820
% Informative SNPs	7.2	43.3	78.9
% missing data	9.3	2.7	10.6
ANI % (avg/med/min)*	97.3/99.0/89.1	95.0/99.8/63.2	82.6/55.7/48.0
CPU time (hr:min)	50:25:00	373:09:00	152:00:00
CPU efficiency	19%	24%	41%
Total Runtime (hr:min)	9:47	48:34:00	12:10
Max Mem Used (GB)	>16	58.34	17.3

*ANI percentage is calculated with ExANIFA (github.com/eamiddlebrook/ExANIFA) on the final concatenated codon alignments excluding gapped positions. See Supplemental Methods for details.

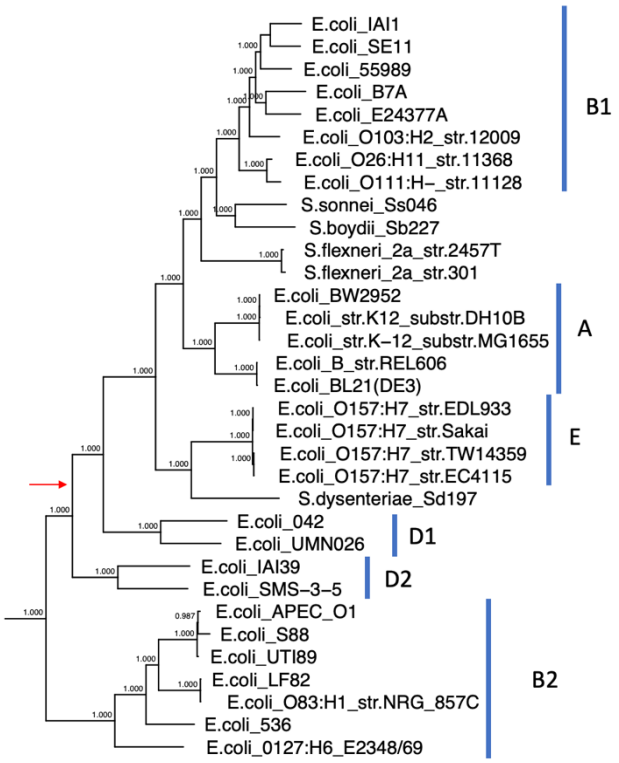

Figure 3 - Strict single copy ortholog based maximum likelihood phylogeny of 34 Escherichia and Shigella assemblies. The tree was constructed with 2142 genes totaling 2.04MB of sequence. Tips are labeled with the given species and strain from NCBI's BioSample database (details in supplementary table 1). Bars next to tree indicate *E.coli* phylotypes [Shakya 2017]. The tree was rooted at *E. fergusonii* then root node was removed. The red arrow indicates the alternative root placement from Shakya 2017.

242

243 **3.2 Genus level Species trees of 689 assemblies: *Brucella/Ochrobactrum***

244 For the 689 Brucella/Ochrobactrum assemblies analyzed (plus *Mycoplana dimorpha*
245 outgroup), OrthoPhyl was run on a RedHat8 compute node with 30 CPUS and 500 GB of ram.
246 The full analysis (generating strict and relaxed SCO with *FastTree2* and *ASTRAL* trees) took
247 just over 2 days (48hrs 31 min). Total ram usage was moderate at 58GB. CPU usage
248 efficiency was 25%, with a total CPU time of approx. 373 hours **(Table 1)**.

249 OrthoPhyl identified 785 strict and 1635 relaxed SCOs in the full dataset of assemblies.
250 The distribution of taxon number represented in each alignment shows that most
251 orthogroup alignments have all taxa represented, with approximately 1400 OGs having 680-
252 689 taxa (**Figure 2 A**). The OG alignments have every little missing data (gaps or ambiguous
253 bases) with the vast majority showing less than 10% (**Figure 2 B**). **Figure 2 C** shows the
254 number of informative sites vs. each alignment's length. A slope of 0.42, indicates the
255 alignments are far from the value that would constitute noisy alignments or over-saturation
256 of mutations.

257 The maximum likelihood (ML) *Brucella* phylogeny generated by FastTree based on the
258 1635 relaxed SCOs **(Figure 4)** shows very close agreement with trees generated by (Ashford
259 *et al.*, 2020) for the *Ochrobactrum* clades (see **Supp. Figure 2** for high resolution tree with all
260 accessions and species labeled). For instance, it supports two major *Ochrobactrum* clades,
261 Group A - *O. anthropi, O. lupini, O. tritici. O. pecoris, O. oryzae, O, cicero, O. intermediaum,*
262 *O. pseudointermedium,* and *O.daejeonense,* and Group B - *O. grignonenense, O.*
263 *pituitosa(pituitosum), O. quorumnocens, O.rhizosphaerae, O. pseudo grignonenense, O.*
264 *thiophenivorans,* and *O. gallinifaecis* **(Figure 4 orange bars)**. Like Ashford et al., our tree

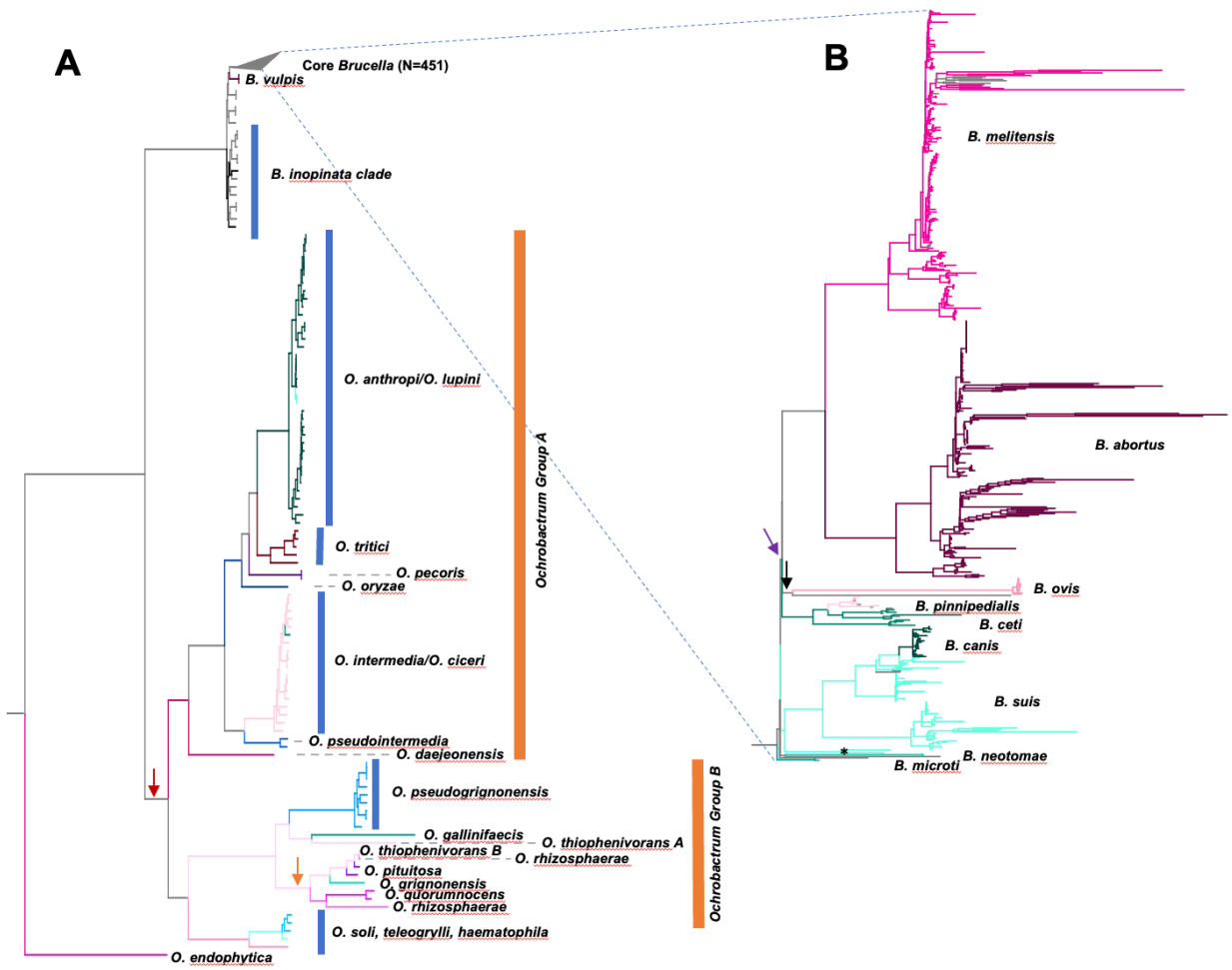

Figure 4 – A) Maximum likelihood tree of Brucella/Ochrobactrum assemblies generated by FastTree within OrthoPhyl. If provided by NCBI, species are labeled. Branch colors. To aid in visualization of mono/polyphyly, branches are colored by reported species. Internal branch colors are given based on majority rule of offspring. Branch lengths represent relative number of mutations per site. The tree was rooted with *Mycoplana dimorpha* that was subsequently removed, and the classic Brucella clade was collapsed. **B)** Subtree of the collapsed classic Brucella clade. A high-resolution version of the full tree showing Species, strain, accession, and bootstrap support is available in **Supplementary figure 2**.

supports *O. endophytica* as basal to the other *Ochrobactrum*. However, the clade with *O. haematophila, O. soli* and *O. teleogrylli* is shown as sister to Group B *Ochrobactrum* instead of Group A. A major distinction between our tree topology and the one presented in Ashford et al. is that our tree shows the monophyletic Brucella clade splitting the *Ochrobactrum* clade between *O. endophytica* and all other *Ochrobactrum*. In contrast, the Brucellae in Ashford et al. split *Ochrobactrum* between Group A and B, as in (Hördt *et al.*, 2020). This

271 difference could be a product of having different outgroups (*Mycoplana ramose, M.*
272 *dimorpha,* and *Rhizobium etli* for Ashford et al. and just *M. dimorpha* for OrthoPhyl), and our
273 tree being subject for long branch attraction between *Brucella* and the root. Additionally,
274 our tree indicates that one *O. thiophenivorans* and *O. rhizosphaerae* sample were likely
275 mislabeled when uploaded, as both have examples rendered polyphyletic by assemblies of
276 many other species with high support.

277 *Ochrobactrum lupini* and *O. ciceri* are closely associated with *O. anthropi* and *O.*
278 *intermedia* (respectively). Like Ashford et al. comparing field isolates to type strains, we had
279 many *O. anthropi* and *O. intermedia* assemblies which show that *O. lupini* and *O. ciceri* arose
280 from within their respective clades and are likely misclassified as species, but instead should
281 be classified as *O. anthropi* and *O. intermedia*, respectively. For *O. lupini* and *O. anthropi*, this
282 is supported by previous work (Gazolla Volpiano *et al.*, 2019).

283 The clade comprising the classical *Brucella* strongly agrees with published phylogenies
284 (Suárez-Esquivel *et al.*, 2020; Ashford *et al.*, 2020). *Brucella vulpis* at the base*, B.*
285 *pinnipedialis* arising from within *B. ceti* (rendering *B. ceti* paraphyletic) (Orsini *et al.*, 2022), *B.*
286 *canis* arising from within *B. suis* (rendering *B. suis* paraphyletic), and *B. melitensis* and *B.*
287 *abortus* being sister groups. The classical *Brucella* have very short branches at the crown (*B.*
288 *microti*, *B.neotomae*, *B.ovis*, *B. suis*/*canis*, *B. abortus*/*melitensis* and *B. pinnipedialis*/*ceti*),
289 indicating very rapid divergence times for these clades, as seen elsewhere (Ashford *et al.*,
290 2020; Suárez-Esquivel *et al.*, 2020; Orsini *et al.*, 2022).

291 **3.2.1 kSNP vs. OrthoPhyl**

292 To compare performance of OrthoPhyl with an alternative assembly-to-tree method,
293 kSNP4 (Gardner *et al.*, 2015) was run with the same hardware as OrthoPhyl. See
294 supplementary methods for details. The resulting kSNP4 alignment shows 7.15 million SNPs
295 were identified, considerably longer than the maximum genome length of the dataset. As
296 expected, much of this SNP alignment consisted of missing data (~78%) indicating *k*-mers not
297 being shared between assemblies.
298 FastTree2 was used externally to generate a tree with the *k*-mer based SNPs. To reduce
299 runtime, memory usage, and gap-based noise, SNPs found in less than 30% of assemblies
300 were removed, with a minimum of 20% being preserved, this left 1,569,787 SNPs (exactly
301 20% of original SNPs, with ~13% total missing data). This SNP file was then used to generate
302 a tree with FastTree2 using the GTRGAMMA model. For a high-resolution tree with species,
303 strain, and assembly accessions labeled, see **(Supp. Figure 3)**. For deep branches outside of
304 the classic Brucella, this tree shows largely the same topology as the tree generated by
305 OrthoPhyl. Some notable exceptions within the Ochrobactrum include: 1) the placement of
306 O. gallinifaecis and one of the *O. thiophenivorans* being sister to *O. pseudogrignonensis* in
307 the OrthoPhyl tree while it is sister to the clade of *O. pseudogrignonensis, O. quorumnocens,*
308 *O, pituitosa* (among others) in the kSNP based tree **(Figure 4A, orange arrow)**, and 2) *O.*
309 *daejeonensis* is placed one node more basal in the kSNP tree **(Figure 4A, red arrow)**. For the
310 *Brucella* clade, there are 2 major differences. The first is the placement of the root, with
311 kSNP showing it coming in-between B. ovis and the rest of Brucella **(Figure 4B, black arrow)**.
312 The second is (given the OrthoPhyl root) *B. neotomae* being place, along with two B. suis
313 samples **(Figure 4B, asterisk)** on the short branch leading to the *B. melitensis, B. abortus, B.*

314 *ovis, B. pinnipedialis, B. ceti* clade, instead of sister to all *Brucella* internal to *B. microti*
315 **(Figure 4B, purple arrow)**. Compared to kSNP4, OrthoPhyl had favorable runtime and
316 resource usage requirements. Major differences include max memory usage (58 vs 204 GB),
317 max storage (53 vs 480GB) and runtime (48 vs > 216 hours) for OrthoPhyl and kSNP4,
318 respectively.
319
320 **3.2.3 Order Level Divergence Tree with 1200 samples: Rickesiales**
321 Rickettsiales was chosen as a challenging taxon to test OrthoPhyl's ability to deal with
322 wide phylogenetic distances and extensive genome reduction. Additionally, this set of
323 assemblies was not filtered stringently, leaving a total of 1209 assemblies (7 outgroup) of
324 varying quality (see Datasets section above). As with the *Brucella* analysis, OrthoPhyl was run
325 on a 30 cpu compute node with 500 GB of available ram. The program took just 12 hours
326 and 10 minutes to complete using a total of 17.3 GB memory. The CPU usage efficiency was
327 41%, with a total CPU time of approx. 152 hours **(Table 1)**.
328 Even with this challenging genomic dataset, OrthoPhyl identifies > 29,000 homologous
329 base pairs from strict SCOs and 295,000 bp from relaxed SCOs (found in 366 or more
330 assemblies). These homologous sequences come from for 42 and 348 genes respectively
331 (**Table 1**). The relaxed SCO FastTree ML tree reconstructed from the 1202 Rickettsiales
332 assemblies recovers all valid genera as monophyletic (**Figure 5A**). The sole paraphyletic
333 genus recovered is *Candidatus Jidaibacter*. Other than *Jidaibacter*, the recovered topology of
334 genera is consistent with previously published analyses (Salje, 2021; Schön *et al.*, 2022), with
335

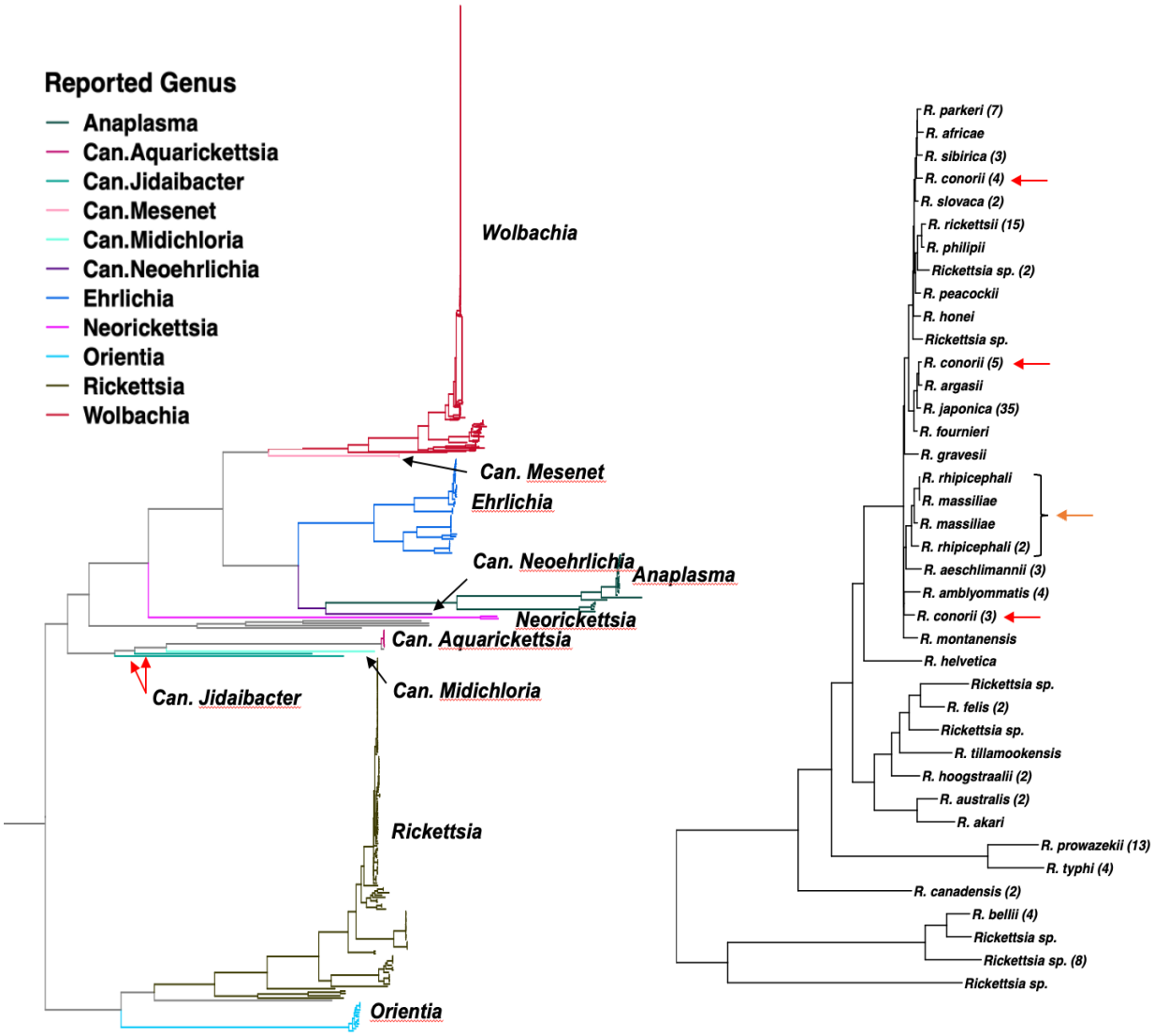

Figure 6 - A) Maximum likelihood tree for Rickettsiales generated by FastTree within OrthoPhyl. Branch colors show reported genera for each sample according to NCBI metadata. To aid in identification, labels are also provided. Samples with no genus level identification are labeled in grey. Internal branch colors are given based on majority rule of offspring. Note the red Wolbachia clade has been compressed vertically (1/5 ratio) thus clade widths are not proportional to number of samples across the tree. Branch lengths are relative probability of a mutation per site. The tree was rooted with a *Pelagibacter* outgroup that was subsequently removed. Candidate genera (Candidatus) are labeled Can. The single, genus level polytomy is labeled in red (*Canidatus-Jidibacter*). **B)** The *Rickettsia* subtree is shown with monophyletic species identifiers collapsed. Following tip labels in parentheses is the number of collapsed tips. The two most basal branches are omitted, as they are only identified to genus. All assemblies with species level classification show monophyly with the exception of *R. conorii (red arrows)* and *R. massiliae/R. rhipicephali (orange arrow).*

337 two major clades 1) *Wolbachia* sister to *Ehrlichia* and *Anaplasma*, with *Neoricketsia* basal to
338 them and 2) *Rickettsia* and *Orientia* separated from the rest by the root. A high-resolution
339 tree is provided in **Supplementary Figure 4,** which provides species, assembly accession and
340 number of SCOs per sample.

341 To illustrate OrthoPhyl's ability to resolve narrow evolutionary windows along with
342 order level divergences in the same tree, the species rich *Rickettsia* genus subtree is shown
343 **(Figure 5B)**. Only three *Rickettsia* with species level identification (provided by NCBI
344 metadata) show polytomy. First, samples labeled *R. conorii* are placed in 3 very different
345 clades across the tree **(Figure 5B, red arrows)**, indicating that they are likely misidentified.
346 Analysis in Hördt et al, using digital DNA:DNA hybridization, also indicates *R. conorii* and
347 related species need revision. Additionally, the placement of *R. massiliae* and *R. rhipicephali*
348 samples render each other polyphyletic (**Figure 5B**, orange arrow). Excluding these samples,
349 the estimated *Rickettsia* tree is identical to the one presented in [Diop 2017].

350

351 **<u>4 Key features</u>**

352 **4.1 Codon alignment based**

353 Most ortho-gene methods use protein sequences, which are easy to align. However, our
354 software creates codon alignments of orthogroups, giving it two major advantages. Using
355 codon alignments to infer phylogenetic trees leverages protein alignment accuracy. Because
356 proteins have a 20 characters while nucleotides have only 4 in their "alphabet", protein
357 alignments have a much better signal-to-noise ratio (States *et al.*, 1991). Using nucleotide

358 alignments adds phylogenetic signal through codon degeneracy, allowing greater amounts of
359 information per gene (Wernersson and Pedersen, 2003; Bininda-Emonds, 2005).

360 **4.2 Many input assemblies**

361 This tool can generate trees with >1000 assemblies. It does this by identifying
362 orthogroups in a diversity-spanning subset of assemblies, then assigning proteins to these
363 orthogroups for the full set of predicted proteins by iterative HMM searches. This reduces
364 the computationally demanding all-vs-all protein searching that would otherwise be
365 required while retaining the ability to identify the majority of orthogroups spanning the
366 dataset.

367 **4.3 User friendly**

368 OrthoPhyl manages the formatting and organization of intermediate files, which is a
369 non-trivial task in a phylogenetic workflow such as this. As with any phylogenetic method,
370 alignment and tree estimation parameters will likely need to be tuned to obtain robust
371 results, thus we exposed these variables for altering at runtime. If OrthoPhyl fails due to
372 input error or lack of memory recourses, it can restart at the last completed step. Being
373 robust to restarting also allows users to tune parameters without rerunning the entire
374 pipeline.

375

376 **<u>5 Discussion</u>**

377 OrthoPhyl successfully reconstructs phylogenetic trees for clades with broad
378 evolutionary divergences. This is achieved with little user input; only requiring one command
379 line input pointing the pipeline to the assembly and output directories. OrthoPhyl can

380 generate trees for >30 bacterial assemblies on a laptop or >1000 assemblies on a
381 workstation or single compute node with moderate resources (30 cpus and 100GB RAM).
382 Many useful options are provided by OrthoPhyl, including alignment trimming parameters,
383 number of assemblies from which to identify initial SCOs, tree building software to use, and
384 evolutionary models for tree estimation.

385 With few exceptions, the trees estimated for the *E. coli/Shigella,*
386 *Brucella/Ochrobactrum,* and Rickettsiales datasets are consistent with published topologies.
387 This illustrates OrthoPhyl's ability to resolve broad evolutionary relationships: from highly
388 similar genomes with ANIs of >99% for *E.coli/Shigella* and *Brucella/Ochrobactrum* to highly
389 dissimilar assemblies with ANIs of 57% for the Rickettsiales. With OrthoPhyl being able to
390 generate whole genome-based phylogenies for more than 1000 assemblies, it fills a large
391 gap in phylogenomic analysis software, providing an easy to use, assembly-to-tree tool that
392 will help many researchers incorporate evolutionary analysis into their ongoing bacterial
393 studies.

394 The *Brucella/Ochrobactrum* and Rickettsiales trees are, to our knowledge, the largest
395 phylogenies published for these clades, let alone using whole genome methods. While more
396 work is necessary to validate the topologies presented here, they are a starting point for
397 assessing metadata-based species labels and expected clade monophyly. Future work will
398 test the robustness of topologies to different evolutionary models and maximum-likelihood
399 methods.

400 **5.1 Pipeline limitations**

There are many parameters to tune during the various steps within this pipeline. Since we focus on usability there is an inevitable tradeoff with optimization, mainly through automated parameter choice. Therefore, this pipeline is useful for "democratizing" the generation of large bacterial phylogenetic trees but does not produce publication ready trees without user inspection of intermediate file such as alignment quality, single copy ortholog numbers, percent phylogenetically informative sites, and of course bootstrap values. Ultimately, trees should be compared to existing data to ensure major clades are consistent, and if not, total evidence for novel branching should be investigated.

5.2 Future

Future versions of this software package will include several additional evolutionary analyses. Bayesian tree inference methods will be incorporated into the pipeline to allow more flexibility in tree estimation methods. Since codon alignments of single copy orthologs are generated while running this pipeline, we also plan on integrating a software package, e.g., ETE3 (Huerta-Cepas *et al.*, 2016), to test evolutionary models such as positive/negative selection and neutrality. A "forest of life" based horizontal gene transfer analysis will be added using gene tree comparisons. This analysis will look at k-means clustering of gene trees to build multiple consensus trees to identify if blocks of genes do not agree with a single consensus species tree hypothesis (Puigbò *et al.*, 2019).

In the current iteration of OrthoPhyl, very ancient paralogs (originating before the most recent common ancestor of the dataset species) might be clustered into the same orthogroup, leading to the orthogroup’s removal during filtering even though the 2 sub-trees of the orthogroup could be consistent with the species tree. Future iterations of this

423 software will filter the orthogroups in a deep paralog aware manner, using a rapidly
424 generated neighbor-joining species tree to identify and rescue such orthogroups.
425 Although this pipeline is geared towards prokaryotes, only two changes are needed to
426 adapt it to eukaryotic phylogenetics: 1) allowing users to input precomputed annotations
427 directly, avoiding the complex task of Eukaryotic annotation within the workflow itself and 2)
428 taking into consideration paralog subtrees which are consistent with a consensus species
429 tree. The next version of OrthoPhyl will incorporate these changes to expand the use cases
430 to Eukaryotic phylogenetics.
431 OrthoPhyl identifies gene family groups for the dataset provided. While this allows the
432 workflow to adapt to any set of genome assemblies, it requires considerable compute
433 resources to complete. Alternatively, predefined gene family models could be used as a
434 starting point, then the un-clustered protein sequences would be fed into the standard
435 OrthoPhyl orthogene identification pipeline to capture additional gene families with
436 phylogenetic signal.
437 **5.3 Conclusion**
438 Phylogenetic reconstruction is critical to understanding the evolution of pathogenicity,
439 novel traits, horizontal gene transfer, and metabolic potentials of bacteria. Unfortunately,
440 large-scale phylogenomic analyses of diverse bacterial groups require deep understanding of
441 bioinformatic methods and great care in dealing with the myriad intermediate files and
442 formats used during the analysis. To our knowledge, OrthoPhyl is the only software to take
443 >1000 bacterial assemblies with order level divergence, generate ortholog codon
444 alignments, then reconstruct phylogenetic trees without user input or management of

intermediate files being necessary. Thus, OrthoPhyl will allow many research groups, including those with modest computing resources, to leverage of the wealth of publicly available genomic data to enrich their ongoing analyses with phylogenomic inferences across a broad swath of bacterial diversity.

7 Acknowledgments

The authors would like to thank Migun Shakya, Taehyung Kwon and John Gillece for their valuable insight and comments and the whole Disease Surveillance and Molecular Epidemiology of Brucella project team members in Tanzania and Rwanda, especially Prof. Joram Buza for comments and suggestions.

This research was funded by the Defense Threat Reduction Agency through Triad National Security, LLC, operator of the Los Alamos National Laboratory under Contract No. 89233218CNA000001 with the U.S. Department of Energy.

8 Conflict of interest

The authors declare no conflict of interest.

9 Bibliography

Ashford,R.T. *et al.* (2020) Application of Whole Genome Sequencing and Pan-Family Multi-Locus Sequence Analysis to Characterize Relationships Within the Family Brucellaceae. *Front. Microbiol.*, **11**.

Baltrus,D.A. *et al.* (2014) Incongruence between multi-locus sequence analysis (MLSA) and whole-genome-based phylogenies: Pseudomonas syringae pathovar pisi as a cautionary tale. *Mol. Plant Pathol.*, **15**, 461–465.

BBMap (2022) *SourceForge*.

Bertels,F. *et al.* (2014) Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads. *Mol. Biol. Evol.*, **31**, 1077–1088.

Bininda-Emonds,O.R. (2005) transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics*, **6**, 156.

475 Bussi,Y. *et al.* (2021) Large-scale k-mer-based analysis of the informational properties of
476 genomes, comparative genomics and taxonomy. *PLOS ONE*, **16**, e0258693.
477 Capella-Gutiérrez,S. *et al.* (2009) trimAl: a tool for automated alignment trimming in large-scale
478 phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
479 Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference.
480 *Genome Inform. Int. Conf. Genome Inform.*, **23**, 205–211.
481 Eddy,S.R. (2008) A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical
482 Significance Estimation. *PLOS Comput. Biol.*, **4**, e1000069.
483 Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLOS Comput. Biol.*, **7**, e1002195.
484 Emms,D.M. and Kelly,S. (2019) OrthoFinder: phylogenetic orthology inference for comparative
485 genomics. *Genome Biol.*, **20**, 238.
486 Emms,D.M. and Kelly,S. (2015) OrthoFinder: solving fundamental biases in whole genome
487 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, **16**,
488 157.
489 Gardner,S.N. *et al.* (2015) kSNP3.0: SNP detection and phylogenetic analysis of genomes
490 without genome alignment or reference genome. *Bioinformatics*, **31**, 2877–2878.
491 Gardner,S.N. and Hall,B.G. (2013) When Whole-Genome Alignments Just Won't Work: kSNP v2
492 Software for Alignment-Free SNP Discovery and Phylogenetics of Hundreds of Microbial
493 Genomes. *PLoS ONE*, **8**, e81760.
494 Gazolla Volpiano,C. *et al.* (2019) Reclassification of Ochrobactrum lupini as a later heterotypic
495 synonym of Ochrobactrum anthropi based on whole-genome sequence analysis. *Int. J.*
496 *Syst. Evol. Microbiol.*, **69**, 2312–2314.
497 Gontcharov,A.A. *et al.* (2004) Are Combined Analyses Better Than Single Gene Phylogenies? A
498 Case Study Using SSU rDNA and rbcL Sequence Comparisons in the Zygnematophyceae
499 (Streptophyta). *Mol. Biol. Evol.*, **21**, 612–624.
500 Hoang,D.T. *et al.* (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol.*
501 *Evol.*, **35**, 518–522.
502 Hördt,A. *et al.* (2020) Analysis of 1,000+ Type-Strain Genomes Substantially Improves
503 Taxonomic Classification of Alphaproteobacteria. *Front. Microbiol.*, **11**.
504 Huerta-Cepas,J. *et al.* (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic
505 Data. *Mol. Biol. Evol.*, **33**, 1635–1638.
506 Huerta-Cepas,J. *et al.* (2007) The human phylome. *Genome Biol.*, **8**, R109.
507 Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site
508 identification. *BMC Bioinformatics*, **11**, 119.
509 Jain,C. *et al.* (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear
510 species boundaries. *Nat. Commun.*, **9**, 5114.
511 Kalyaanamoorthy,S. *et al.* (2017) ModelFinder: fast model selection for accurate phylogenetic
512 estimates. *Nat. Methods*, **14**, 587–589.
513 Katoh,K. and Standley,D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7:
514 Improvements in Performance and Usability. *Mol. Biol. Evol.*, **30**, 772–780.
515 Konstantinidis,K.T. and Tiedje,J.M. (2007) Prokaryotic taxonomy and phylogeny in the genomic
516 era: advancements and challenges ahead. *Curr. Opin. Microbiol.*, **10**, 504–509.
517 Nguyen,L.-T. *et al.* (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating
518 Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

519 Orsini,M. *et al.* (2022) Brucella ceti and Brucella pinnipedialis genome characterization unveils
520 genetic features that highlight their zoonotic potential. *MicrobiologyOpen*, **11**, e1329.
521 Parks,D.H. *et al.* (2015) CheckM: assessing the quality of microbial genomes recovered from
522 isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.
523 Portik,D.M. *et al.* (2016) An evaluation of transcriptome-based exon capture for frog
524 phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura).
525 *Mol. Ecol. Resour.*, **16**, 1069–1083.
526 Price,M.N. *et al.* (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large
527 Alignments. *PLOS ONE*, **5**, e9490.
528 Puigbò,P. *et al.* (2019) Genome-Wide Comparative Analysis of Phylogenetic Trees: The
529 Prokaryotic Forest of Life. *Methods Mol. Biol. Clifton NJ*, **1910**, 241–269.
530 Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing
531 phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
532 Salje,J. (2021) Cells within cells: Rickettsiales and the obligate intracellular bacterial lifestyle.
533 *Nat. Rev. Microbiol.*, **19**, 375–390.
534 Schön,M.E. *et al.* (2022) The evolutionary origin of host association in the Rickettsiales. *Nat.*
535 *Microbiol.*, **7**, 1189–1199.
536 Shakya,M. *et al.* (2020) Standardized phylogenetic and molecular evolutionary analysis applied
537 to species across the microbial tree of life. *Sci. Rep.*, **10**, 1723.
538 Shavit Grievink,L. *et al.* (2013) Missing Data and Influential Sites: Choice of Sites for
539 Phylogenetic Analysis Can Be As Important As Taxon Sampling and Model Choice.
540 *Genome Biol. Evol.*, **5**, 681–687.
541 Smith,D.R. (2013) The battle for user-friendly bioinformatics. *Front. Genet.*, **4**, 187.
542 Spencer,M. *et al.* (2007) Conditioned Genome Reconstruction: How to Avoid Choosing the
543 Conditioning Genome. *Syst. Biol.*, **56**, 25–43.
544 Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of
545 large phylogenies. *Bioinforma. Oxf. Engl.*, **30**, 1312–1313.
546 States,D.J. *et al.* (1991) Improved sensitivity of nucleic acid database searches using application-
547 specific scoring matrices. *Methods*, **3**, 66–70.
548 Suárez-Esquivel,M. *et al.* (2020) Brucella Genomics: Macro and Micro Evolution. *Int. J. Mol. Sci.*,
549 **21**, 7749.
550 Suyama,M. *et al.* (2006) PAL2NAL: robust conversion of protein sequence alignments into the
551 corresponding codon alignments. *Nucleic Acids Res.*, **34**, W609–W612.
552 Treangen,T.J. *et al.* (2014) The Harvest suite for rapid core-genome alignment and visualization
553 of thousands of intraspecific microbial genomes.
554 Varghese,N.J. *et al.* (2015) Microbial species delineation using whole genome sequences.
555 *Nucleic Acids Res.*, **43**, 6761–6771.
556 Wernersson,R. and Pedersen,A.G. (2003) RevTrans: multiple alignment of coding DNA from
557 aligned amino acid sequences. *Nucleic Acids Res.*, **31**, 3537–3539.
558 Yang,Z. (1998) On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.*, **47**, 125–133.
559 Zhang,C. *et al.* (2018) ASTRAL-III: polynomial time species tree reconstruction from partially
560 resolved gene trees. *BMC Bioinformatics*, **19**, 153.
561 Zhang,C. *et al.* (2020) ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Mol.*
562 *Biol. Evol.*, **37**, 3292–3307.