OrthoPhyl – Streamlining large scale, orthology-based phylogenomic studies of bacteria at broad evolutionary scales

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

There are a staggering number of publicly available bacterial genome sequences (at writing, 2.0 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. Here we present OrthoPhyl, which takes bacterial genome assemblies and reconstructs trees from whole genome codon alignments. The analysis pipeline 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 infer orthologs in the full dataset. 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 OrthoPhyl to trees generated with kSNP3 and GToTree along with published trees using 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/blob/OrthoPhyl_1.0/ . Installation and execution instructions are provided in the associated github README.md file. Third party software versions and OrthoPhyl execution files will remain static in the *OrthoPhyl_1.0* branch, with the *main* branch housing current development. For versions of software dependencies see **Supplemental Table 1**. To aid in usability, a Singularity container is available at *https://cloud.sylabs.io/library/earlyevol/default/orthophyl* or with the command *singularity pull library://earlyevol/default/orthophyl:1.0_ms*. Also see the GitHub page for Singularity usage guide. Assemblies used within this manuscript are available from https://www.ncbi.nlm.nih.gov/assembly/ with accessions in **Supplemental tables 2-4**.

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 2,018,201 bacterial genome assemblies available through NCBI alone (ncbi.nlm.nih.gov/assembly/?term=bacteria). 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).

47 Unfortunately, most whole genome phylogenetic analysis methods require specialized
48 bioinformatic data processing and expertise (Smith, 2013; Lozano-Fernandez, 2022). There
49 are few very large scale trees to place these sequences into an evolutionary framework. The
50 available large-scale trees are under-resolved and build by reconciling many disparate
51 phylogenetic studies, such as NCBI taxonomy (Schoch *et al.*, 2020) or are generated with
52 methods unsuited for narrow and broad evolutionary distances (Hördt *et al.*, 2020) reducing
53 their utility at multiple evolutionary scales. This flood of genomic data necessitates an easy-
54 to-use phylogenetic analysis pipeline to help reveal the evolutionary context for the myriad
55 of sequenced bacteria.

56 **1.1 Analysis options**

57 **1.1.1 Single and Multi-locus**

58 A classic way to generate bacterial trees is to compare single homologous loci (e.g., 16S
59 ribosomal, recA, gyrA, rpoB, or dnaK genes). While largely replaced by other methods, single
60 gene trees have some advantages, which include the ability to use conserved primer sites to
61 amplify sequences, very low computational intensity, and the availability of curated
62 databases (SILVA, RDP, NCBI GenBank, etc.). Using genes other than rDNA allows for greater
63 resolution at certain evolutionary distances but still lacks broad resolution (i.e., Near
64 resolution for swiftly evolving genes, far for conserved) (Yang, 1998). However, these trees
65 are prone to misleading topologies due to horizontal gene transfers, incomplete lineage
66 sorting, paralogous gene conversions, and limited phylogenetic information (Huerta-Cepas *et*
67 *al.*, 2007). Using multi-locus alignments to build phylogenetic trees increases the total
68 amount of phylogenetic information (more sequence used), reduces the effects of

69 incongruent gene tree topologies due to recombination or horizontal gene transfer, and
70 widens the breadth of evolutionary resolution (Gontcharov *et al.*, 2004). However, in multi-
71 locus analyses of taxa without well vetted multi-locus methods, locus selection is non-trivial
72 and the effort and cost of primer design and amplification optimization scales with the
73 number of sequences used, which can rapidly become prohibitive for many labs.
74 Additionally, strains of many taxa, like Brucella, are not differentiated well by available multi-
75 locus methods, leading to unknown evolutionary histories (Sankarasubramanian *et al.*,
76 2019).

77 **1.1.2 Whole genome**

78 Classically, whole genome alignments were required to identify Single Nucleotide
79 Polymorphisms (SNPs) of assemblies (see Shakya *et al.*, 2020; Treangen *et al.*, 2014 for
80 method details). These methods add considerable phylogenetic data for tree building and
81 can produce different topologies than multi-locus alignments, illustrating their potential
82 utility (Baltrus *et al.*, 2014). A major drawback of whole genome alignments is that they tend
83 produce short alignments, or fail outright, for more distantly related genomes (Shakya *et al.*,
84 2020; Chung *et al.*, 2018; Darling *et al.*, 2004; Angiuoli and Salzberg, 2011). Thus,
85 phylogenetic information drops quickly as a function of evolutionary distance. The practical
86 result is a lack of resolution at even moderate evolutionary distances. For alignments to a
87 reference, having the "missing" data enriched for the more divergent sites will possibly lead
88 to an underestimation of the sequence to reference evolutionary distances for more
89 divergent sequences (Bertels *et al.*, 2014; Spencer *et al.*, 2007; Shavit Grievink *et al.*, 2013).
90 Alternatively, one can use all-vs-all whole genome alignments, however they scale poorly

91 and are limited in the number of input taxa that can be aligned (Angiuoli and Salzberg,
92 2011).

93 **1.1.3 K-mer overlap**

94 An emergent method to quickly estimate evolutionary histories from genome
95 assemblies is to use *k*-mer content. This involves identifying subsequences of length *k* within
96 each assembly and calculating the shared *k*-mer content (for details see Bussi *et al.*, 2021).
97 The advantage of this class of methods is that it is extremely fast relative to alignment-based
98 methods. The comparison of *k*-mer overlap between sequences naturally leads to pairwise
99 distance matrices, thus neighbor-joining trees can be generated directly from the results
100 (Saitou and Nei, 1987). Recent work has integrated more complicated distance statistics with
101 success (Tang *et al.*, 2023). Crucially, *k*-mer distances become non-linearly associated with
102 true percent identity, leading to erroneous estimated branch lengths (Jain *et al.*, 2018).

103 An alternative *k*-mer based approach is kSNP (Gardner and Hall, 2013; Gardner *et al.*,
104 2015; Hall and Nisbet, 2023), which identifies SNPs by flanking conserved *k*-mers. This
105 approach allows the use of tree building methods including parsimony, maximum likelihood,
106 and Bayesian. Like whole genome alignments and *k*-mer overlap methods, kSNP suffers from
107 a rapid decline in phylogenetic information as sequences diverge. Additionally, the
108 signal/noise ratio degrades quickly because non-homologous *k*-mers are increasingly
109 common as sequences diverge (Gardner and Hall, 2013).

110 **1.1.4 Predicted gene alignments**

111 To use the wealth of data generated by NGS methods and simultaneously simplify the
112 problem of whole genome alignments, CDS or protein sequences derived from annotations

113 or transcriptome sequencing can be compared. This method breaks the problem of
114 identifying homologous loci into two tractable parts: gene identification, then identification
115 of homologous sequences. This approach identifies many phylogenetically informative sites
116 (like whole genome), but alignments are more tractable (like single loci). However, gene-
117 based alignments are negatively affected by inferred paralogs, i.e., duplicated genes or
118 contamination. Thus, most methods filter out gene families with paralogs. However, see
119 ASTRAL-Pro for a method that is "paralog aware" (Zhang *et al.*, 2020). A widely used pipeline
120 that accomplishes this task is GToTree, which annotates assemblies, uses precomputed
121 hidden Marcov models to identify single copy orthologs, aligns predicted amino acid
122 sequences, then generates trees with FastTree2 or IQTree2 from concatenated alignments
123 (Lee, 2019). While focused on protein alignments, GToTree can also be run in nucleotide
124 mode to use CDS alignments for phylogenetic inference instead.
125 A trade-off arises while using predicted coding or protein sequences to build species
126 trees. Coding sequences (CDS) have the most phylogenetic information due to codon
127 degeneracy but are hard to align at high divergencies (States *et al.*, 1991). Proteins remain
128 alignable at high divergences but lack the information content of CDS alignments. This can
129 be addressed by converting protein alignments to corresponding codon alignments,
130 leveraging nucleotide phylogenetic information and protein alignment accuracy
131 (Wernersson and Pedersen, 2003; Bininda-Emonds, 2005). However, at increasing
132 evolutionary distances, even with accurate alignments, mutational saturation in nucleotide
133 data can negatively affect phylogenetic estimates. Evolutionary model selection and dense
134 taxon sampling can largely alleviate these issues, see Kapli *et al.*, 2023 for details.

135 **1.2 Benefits of an automated workflow**

136 The skills required to generate bacterial phylogenetic trees from whole genomes are
137 extensive. Some of the many steps include gathering and annotating assemblies, identifying
138 and filtering orthologs, sequence alignment, trimming and concatenation, and finally, tree
139 inference (Ashford *et al.*, 2020; Lozano-Fernandez, 2022). Each one of these steps, especially
140 for large data sets, requires expertise in picking parameters, file management, and data
141 format manipulation and filtering. Many of these steps also require familiarity with UNIX
142 command-line.

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

151 To meet this demand, we present OrthoPhyl, a phylogenomic pipeline that takes bacterial
152 genomes as input, annotates them, identifies orthologs, converts protein to nucleotide
153 alignments, and builds species trees with both concatenated alignments and gene tree to
154 species tree reconciliation. The workflow accepts an arbitrarily large number of input
155 genomes (tested up to 1200 here). To accelerate analysis, a subset of samples representing
156 the whole dataset's diversity are identified, their proteomes are used to identify orthologs

157 and build hidden Markov models, which are then expanded to the full input dataset by

158 iterative searches. This strategy allows the generation of trees for 689 Brucella assemblies in

159 ~48hrs using 30 cpus and 58.3GB memory with no hands-on time required. This pipeline is

160 designed to be an easy to install, scalable, turn-key solution for generating high-resolution

161 bacterial trees from diverse clades.

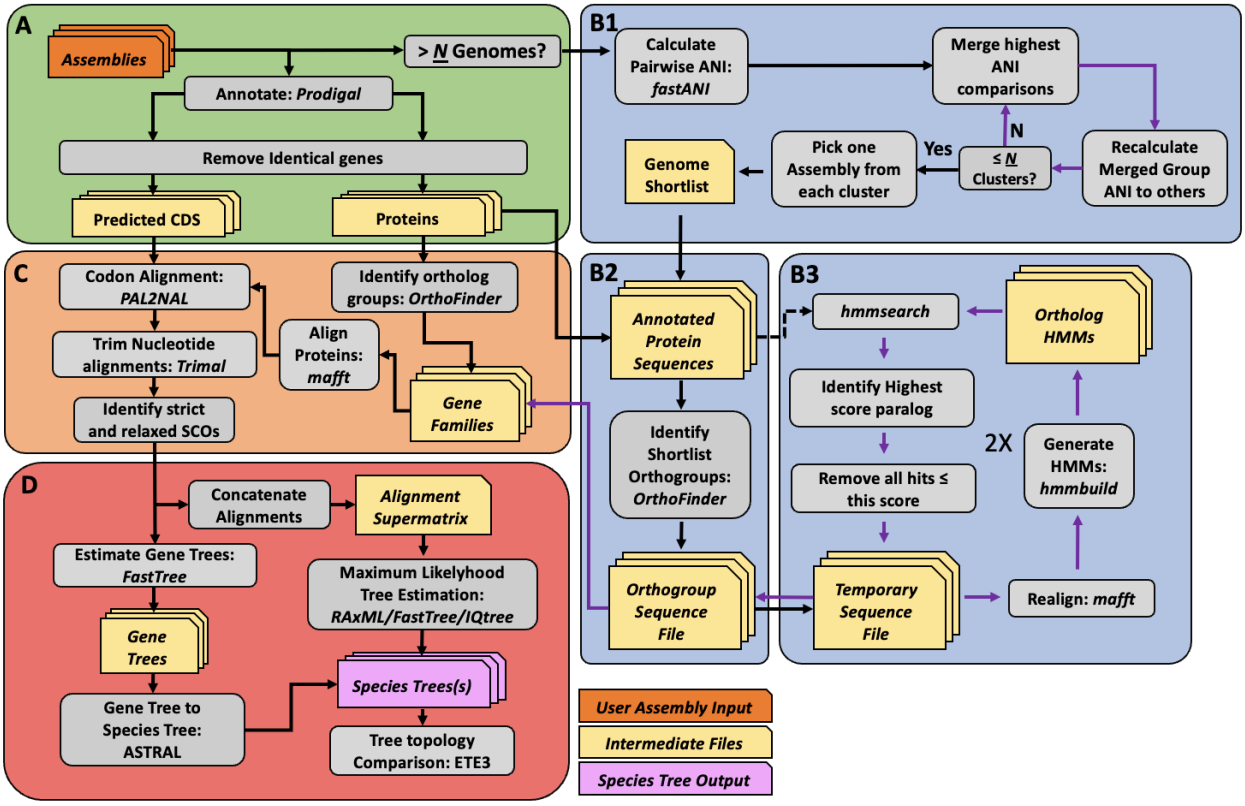

Figure 1 - Workflow diagram of OrthoPhyl. Grey boxes indicate processes. Programs used in each step are listed unless a custom script was used. Orange, tan, and purple boxes represent user input, intermediate files, and species tree outputs, respectively. Purple arrows show iterative approaches. The workflow is divided into four main tasks A) annotate assemblies and remove identical CDSs. If more than "N" assemblies are being analyzed, B1) identify a subset of diversity-spanning assemblies, B2) pass them through OrthoFinder to generate orthogroups, and B3) expand the OrthoFinder-identified orthogroups to the full dataset of assemblies through iterative HMM searches. C) Align full orthogroup protein sets, generate and trim matching codon alignments, then filter orthogroups by taxon representation. Finally, D) estimate species tree topologies with concatenated codon alignment supermatrices along with a gene tree to species tree consensus method.

162

163 **2 Methods**

164 **2.1 OrthoPhyl's Workflow**

165 The structure of OrthoPhyl can be broken down into four main steps: Genome
166 assemblies are annotated (**Figure 1A**), resulting proteins are assigned to orthogroups (**Figure**
167 **1B**), orthogroup proteins are aligned and converted to codon alignments (**Figure 1.C**), and
168 finally, concatenated codon alignments are used to infer phylogenetic trees (**Figure 1.D**). All
169 processes bellow use default parameters unless specified otherwise.

170 **2.1.1 Structural Annotation**

171 OrthoPhyl starts by generating gene calls from input assemblies (Figure 1A) with
172 Prodigal (Hyatt *et al.*, 2010). Redundant CDSs (> 99.9% nucleotide identity) found in a
173 genome are removed using bbmap's *dedupe* (BBMap, 2022) with the logic they are likely the
174 result of recent gene duplications and provide no or little phylogenetic information. This
175 removes unnecessary paralogs from the dataset to preserve the usefulness of the containing
176 orthogroup for downstream analysis.

177 **2.1.2 Orthogroup assignment**

178 To infer orthogroups to use for tree generation, *OrthoFinder* is used with default
179 parameters except for using multiple sequence alignments based gene tree inference (Emms
180 and Kelly, 2015, 2019). When species trees are being generated for large numbers (default: >
181 30) of genome assemblies, finding orthogroups directly becomes intractable due to the
182 exponential increase in computational time of the all-vs-all homology search. To overcome
183 this, OrthoPhyl identifies a diversity spanning subset of assemblies to identify orthogroups.

184 Briefly, average nucleotide identity (ANI) is estimated for all assembly pairs with *FastANI*
185 (Jain *et al.*, 2018). A custom algorithm is used to cluster assemblies based on pairwise ANI
186 values. Briefly, through successive rounds, samples with the highest ANI are merged into
187 clusters. The merged samples' ANI to other samples and clusters are averaged. Rounds of
188 merging are performed until N number (default: 30) of clusters are formed. Single
189 representative samples for each of the N groups are chosen as input for OrthoFinder to infer
190 a reduced set of orthogroups. *FastANI* fails to give an ANI percent for genome comparisons
191 with ANI less than ~75%. For these instances, an arbitrary value of 50% is used. The
192 representative genome picking will be unaffected if there are less than 30 clusters at this or
193 greater distances to all other clusters.
194 To expand orthogroups to the full assembly dataset, proteins from each orthogroup
195 are realigned with *Mafft* (Katoh and Standley, 2013), then *hmmer* (Eddy, 2008, 2009, 2011)
196 generates hidden Markov models and searches against all predicted proteins using default
197 parameters. OrthoPhyl then finds the minimum HMM hit score cutoff which removes all
198 paralogs, the remaining hits above this score (if any) are designated as the final orthogroup,
199 ensuring all potential paralogs are removed and the greatest number of taxa are
200 represented in the orthogroup. Each orthogroup is then realigned and HMMs are again
201 generated and searched against all proteins to capture additional, more divergent
202 orthologues.
203 **2.1.3 Codon Alignment generation**
204 Once orthogroups are identified for the full set of assemblies, protein sequences are
205 realigned with *Mafft* (Katoh and Standley, 2013). These alignments are then used as the

206 basis for codon alignment using *PAL2NAL* with codon table 11 (Suyama *et al.*, 2006). Codon
207 alignments are then trimmed with *trimAL* using arguments "-resoverlap .5 -seqoverlap 50 -gt
208 .80 -cons 60 -w 3" (Capella-Gutiérrez *et al.*, 2009). Users can change trimming options with
209 the "control_files.user" if required. *Alignment_Assessment* (Portik *et al.*, 2016), is used to
210 visualize the quality of codon alignments, including their phylogenetic signals, taxa per
211 alignment, and alignments per taxa **(Figure 2)**. This is critical for ensuring high-quality data
212 are used for generating trees and can aid in troubleshooting phylogenies with low branch
213 support.

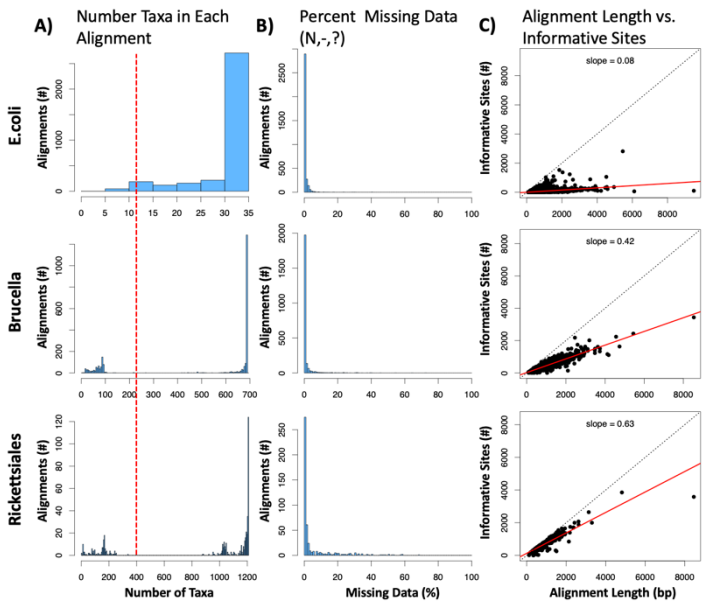

Figure 2 – Alignment metric visualization generated by OrthoPhyl. A) Histograms of the number of taxa in each alignment is shown. Dotted red line indicates the >30% taxa cutoff for alignments used to generate the "SCO relaxed" trees. B) Amount of missing data (including gaps and ambiguous bases) per alignment is shown with a historgram. This is exclusive of genomes without detected orthologs. C) The number of phylogenetically informative sites (sites with at least 2 different states in at least 2 taxa each) 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.

214 Our tool then classifies orthogroups as relaxed or strict single-copy orthologs (SCOs).
215 Strict SCOs are genes found in every taxon with no paralogs, while relaxed SCOs are found in
216 a subset of assemblies still with no paralogs. The percent of assemblies with the SCO to
217 count in this set is tunable with the default set as 30% (shown in Figure 2 A: red dashed line).
218 See (Wiens and Morrill, 2011) for analysis on the effects of missing data. These two sets of
219 orthogroups are used in parallel moving forward to generate species trees.

220 **2.1.4 Species Tree Estimation**

221 Two general methods are used to infer species trees for the input assemblies: gene-tree
222 to species-tree estimation with *ASTRAL-III* (Zhang *et al.*, 2018) and maximum likelihood
223 methods based on codon alignment supermatrices generated with catfasta2phyml
224 (https://github.com/nylander/catfasta2phyml). First, to use *ASTRAL*, ML gene-trees are
225 generated with either "*FastTree2 -gtr -gamma*" (Price *et al.*, 2010) or *IQtree2* with default
226 settings (Nguyen *et al.*, 2015; Kalyaanamoorthy *et al.*, 2017; Hoang *et al.*, 2018) per user
227 input. Then ASTRAL will generate species trees with the strict and relaxed SCO gene tree
228 datasets. For the maximum likelihood species tree methods, the user can specify any
229 combination of *IQtree*, *FastTree2 and/*or *RAxML* (Stamatakis, 2014) to infer tree
230 structure(s). Default parameters are used for each tree method except for RAxML and
231 FastTree2 being set to use the GTR + gamma model and 100 and 1000 bootstraps (with
232 Shimodaira-Hasegawa test), respectively, and IQtree2 set to run ModelFinder to identify a
233 best fit evolutionary model and run 1000 ultrafast bootstraps.

234 Finally, the trees generated during the OrthoPhyl run are compared with generalized
235 Robinson-Foulds metrics provided by ETEtoolkit (Huerta-Cepas *et al.*, 2007) indicating the
236 stability of the resultant tree structure when different inference methods and matrix
237 completenesses are used. All default parameters used with these tools can be found in the
238 OrthoPhyl repository at OrthoPhyl/control_file.defaults.

239 **2.2 Datasets**

240 As a proof of concept, OrthoPhyl was used to build trees for three bacterial clades. We
241 analyzed well characterized *E. coli*/*Shigella* strains. These included a list of 34 complete
242 genomes analyzed in (Shakya *et al.*, 2020), obtained from NCBI.
243 The genus *Brucella* was used to show OrthoPhyl's ability to resolve very closely related
244 sequences while dealing with very long relative branch lengths in the same tree
245 (*Ochrobactrum* group). Assemblies were acquired using NCBI Taxon number 234 with the
246 utility script "*gather_genomes.sh*" (Supplemental Methods - Gathering and Filtering
247 Assemblies and **Supplemental Figure 1**) which is packaged with OrthoPhyl
248 (github.com/eamiddlebrook/OrthoPhyl/utils/). This resulted in 689 assemblies after filtering
249 for >98% completion, < 0.1% duplication, and < 1% contamination with CheckM (Parks *et al.*,
250 2015). Here, duplication is the total number of loci identified as marker genes divided by the
251 number of identified marker genes (see Supplemental methods "**Gathering and Filtering**
252 **Assemblies"** and "utils/checkm_assemblies.slurm" within the github repository). It is
253 important to note that *Ochrobactrum* species were recently moved to the genus *Brucella*
254 (Hördt *et al.*, 2020). However, for clarity and due to controversy within the field, we chose to
255 keep the *Ochrobactrum* labeling for these species. An assembly of *Mycoplana dimorpha*
256 (GCA_003046475.1) was added to the dataset as an outgroup.

Table 1 - OrthoPhyl output, runtime, and ram usage

	Assembly Data Set		
Metric	***E. coli* + *Shigella***	***Brucella***	**Rickettsiales**
Number of assemblies	34	689+1	1201+7
Number SCOs - strict	2217	897	42
Total Length (bp)	2,105,000	855,924	29,248
% Informative SNPs	6.1	41.4	81.5
% missing data	0.4	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	-	1762	348
Total Length (bp)	-	1,627,004	295,820
% Informative SNPs	-	43.3	78.9
% missing data	-	2.7	10.6
ANI % (avg/med/min)*	-	95.0/99.8/63.2	82.6/55.7/48.0
CPU time (hr:min)	16:20	373:09:00	152:00:00
CPU efficiency	47%	24%	41%
Total Runtime (hr:min)	3:11	48:34:00	12:10
Max Mem Used (GB)	13.07	58.34	17.3

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

Finally, a tree for all NCBI Rickettsiales assemblies was generated to illustrate OrthoPhyl's utility in dealing with bacterial order level divergences. For Rickettsiales assemblies, we again used *gather_genomes.sh,* this time with NCBI taxon number 766. To illustrate a more challenging scenario, filtering of Rickettsiales assemblies was less stringent with cut-offs at > 95% completion, < 0.2% duplication, and < 1.5% contamination, again using CheckM (Parks *et al.*, 2015) and our independent calculation of duplication (Supplemental methods "**Gathering and Filtering Assemblies"**. This resulted in 1201 assemblies. Seven *Pelagibacter* assemblies were added to this dataset to serve as an outgroup. Accessions and stats for all genome assemblies used in this manuscript are available in **Supplemental Tables 2-4**.

3 Implementation

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

276 To illustrate Orthophyl's ability to build trees from moderate numbers of closely related
277 samples on a local Linux computer, we ran our *E. coli* analysis (34 genomes) on a desktop
278 with a 12-core Intel processor with 16 GB RAM running RHEL8 (centOS8). Because the
279 assemblies are of high quality and the stains are closely related, OrthoPhyl was used to
280 generate trees from only the strict SCO dataset. Run statistics are provided in **Table 1**. The
281 full pipeline took 3.25 hours to build trees with FastTree2 and ASTRAL. OrthoPhyl identified
282 2142 strict SCOs and alignment of these sequences constitutes 2.105 MB with 6%
283 phylogenetically informative sites for the SCO (**Table 1**).

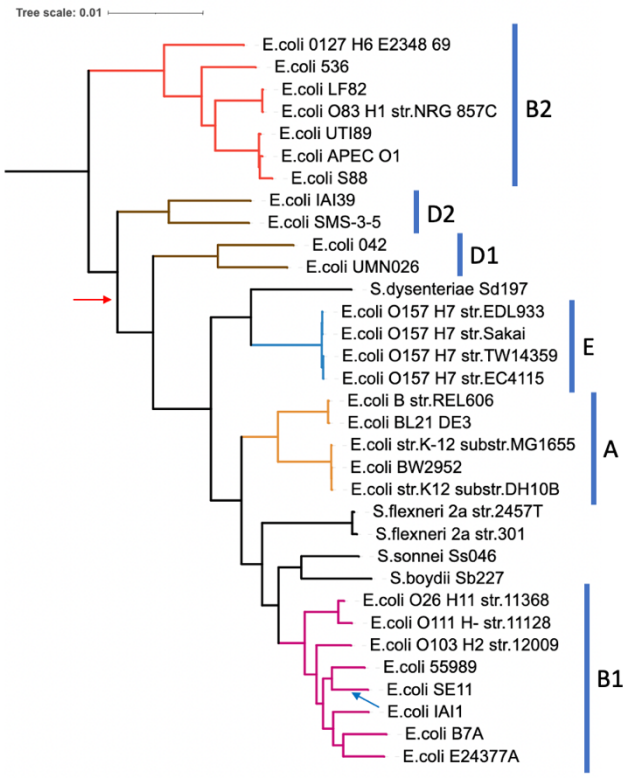

Figure 3 - Strict single copy ortholog based maximum likelihood phylogeny of 34 Escherichia and Shigella assemblies inferred by FastTree2 using the GTR gamma model with the default 1000 bootstraps. The tree was constructed with 2217 genes totaling 2.04MB of sequence. Tips are labeled with the given species and strain from NCBI's BioSample database (details in Supplemental Table 2). All bootstrap support values are 100% except for the split leading to *E. coli* strains APEC O1 and S88 (black arrow), which is 98.7%. Bars next to tree indicate *E.coli* phylotypes [Shakya 2017]. The tree was rooted at *E. fergusonii* then root node was removed. The blue and red arrows indicate the alternative placement of E. coli IAI1 and the root from Shakya 2017, respectively.

284 The resulting tree built by FastTree2 from the SCO-strict dataset shows successful
285 differentiation of the *E. coli* phylotypes **(Figure 3)**. This tree shows B1 and A as sister
286 phylotypes, with *S. sonnei, S. boydii* and *S. flexneri* between them. That clade is in turn next
287 to phylotype E and *S. dysenteriae*, then D1 and D2. Finally, B2 is sister to all of them. This

288 topology is identical to the reported tree from (Shakya *et al.*, 2020) using whole genome
289 alignments, except for the placement of the root (*E. fergusonii)*, where they have the root
290 placed between D2/B2 (red arrow) and the rest while OrthoPhyl's tree indicates it should be
291 placed between B2 and the other phylotypes.

292 **3.2 Genus level Species trees of 689 assemblies: *Brucella/Ochrobactrum***

293 For the 689 Brucella/Ochrobactrum assemblies analyzed (plus Mycoplana dimorpha
294 outgroup), OrthoPhyl was run on a RedHat8 compute node with 30 CPUs and 500 GB of ram.
295 The full analysis (generating strict and relaxed SCO trees with *FastTree2* and *ASTRAL*) took
296 just over 2 days (48hrs 31 min). Total ram usage was moderate at 58GB. CPU usage
297 efficiency was 24%, with a total CPU time of approx. 373 hours **(Table 1)**.

298 OrthoPhyl identified 785 strict and 1635 relaxed SCOs in the full dataset of assemblies.
299 The distribution of taxon number represented in each alignment shows that most
300 orthogroup alignments have all taxa represented, with approximately 1400 orthogroups
301 having 680-689 taxa (**Figure 2A**). The orthogroup alignments have every little missing data
302 (gaps or ambiguous bases) with the vast majority showing less than 10% (**Figure 2B**). **Figure**
303 **2C** shows the number of informative sites vs. each alignment's length. A slope of 0.42,
304 indicates the alignments are far from the value that would constitute noisy, erroneous
305 alignments.

306 The maximum likelihood (ML) *Brucella* phylogeny generated by FastTree2 based on the

307 1635 relaxed SCOs **(Figure 4)** shows very close agreement with trees generated by (Ashford

308 *et al.*, 2020) for the *Ochrobactrum* clades (see **Supplemental Figure 2** for high resolution tree

309 with accessions, bootstrap values, and species labeled). For instance, it supports two major

310 *Ochrobactrum* clades, Group A - *O. anthropi, O. lupini, O. tritici. O. pecoris, O. oryzae, O,*

311 *ciceri, O. intermedium, O. pseudointermedium,* and *O. daejeonense,* and Group B - *O.*

Figure 4 – A) Maximum likelihood tree of Brucella/Ochrobactrum assemblies generated by FastTree2 within OrthoPhyl. If provided by NCBI, species are labeled. 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,* which 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 Supplemental figure 2.

312 *grignonenense, O. pituitosa(pituitosum), O. quorumnocens, O.rhizosphaerae, O. pseudo*
313 *grignonense, O. thiophenivorans,* and *O. gallinifaecis* **(Figure 4 orange bars)**. Like Ashford et
314 al., our tree supports *O. endophytica* as basal to the other *Ochrobactrum*. However, our tree
315 shows the clade with *O. haematophila, O. soli* and *O. teleogrylli* as sister to Group B
316 *Ochrobactrum* with >90% support instead of Group A, like in Ashford et al. (Figure 4A and
317 Supplemental Figure 2, orange arrows). A major distinction between our tree topology and
318 the one presented in Ashford et al. is that our tree shows the monophyletic Brucella clade
319 splitting the *Ochrobactrum* clade between *O. endophytica* and all other *Ochrobactrum* with
320 >90% support (**Supplemental Figure 2, black arrow**). In contrast, the Brucellae in Ashford et
321 al. split *Ochrobactrum* between Group A and B, as in (Hördt *et al.*, 2020). This alternative
322 placement of Brucellae is shown in **Figure 4A** and **Supplemental Figure 2** with blue arrow.
323 The difference could be a product of having different outgroups (*Mycoplana ramose, M.*
324 *dimorpha,* and *Rhizobium etli* for Ashford et al. and Mycoplana dimorpha for OrthoPhyl), and
325 our tree being subject to long branch attraction between *Brucella* and the unbroken long
326 Mycoplana branch. Additionally, our tree indicates that *O. thiophenivorans, O. pituitosa,* and
327 *O. rhizosphaerae* samples are para/polyphyletic with >90% bootstrap support for all
328 associated splits (**Figure 4A** and **Supplemental Figure 2, astrisks**), indicating they were
329 possibly mislabeled when uploaded to NCBI.
330 *Ochrobactrum lupini* and *O. ciceri* are closely associated with *O. anthropi* and *O.*
331 *intermedia* (respectively). Like Ashford et al. comparing field isolates to type strains, we had
332 many *O. anthropi* and *O. intermedia* assemblies which show that *O. lupini* and *O. ciceri* arose
333 from within their respective clades and are likely misclassified as species, but instead

334 possibly should be classified as *O. anthropi* and *O. intermedia*, respectively. For *O. lupini* and
335 *O. anthropi*, this is supported by additional work (Gazolla Volpiano *et al.*, 2019).
336 The clade comprising the classical *Brucella* agrees with published phylogenies (Figure
337 4B) (Suárez-Esquivel *et al.*, 2020; Ashford *et al.*, 2020). *Brucella vulpis* at the base*, B.*
338 *pinnipedialis* arising from within *B. ceti* (rendering *B. ceti* paraphyletic) (Orsini *et al.*, 2022), *B.*
339 *canis* arising from within *B. suis* (rendering *B. suis* paraphyletic), and *B. melitensis* and *B.*
340 *abortus* being sister groups. Interestingly, several Brucella assemblies seem to be mislabeled
341 (**Supplemental Figure 2, grey tip labels**). Whether these placements are accurate or not
342 remains to be determined, however, B. melitensis and B. abortus are very closely related
343 (precluding saturation) and do not show signs of significant interspecies recombination
344 (Vishnu *et al.*, 2015; Suárez-Esquivel *et al.*, 2020). The classical *Brucella* have very short
345 branches at the crown (*B. microti*, *B.neotomae*, *B.ovis*, *B. suis*/*canis*, *B. abortus*/*melitensis*
346 and *B. pinnipedialis*/*ceti*), indicating very rapid divergence times for these clades, as seen
347 elsewhere (Ashford *et al.*, 2020; Suárez-Esquivel *et al.*, 2020; Orsini *et al.*, 2022). While the
348 topology of this tree is congruent with several published trees, more analyses are required
349 to create the most robust tree possible (see Discussion).

350 **3.2.1 GToTree vs. OrthoPhyl: *Brucella/Ochrobactrum***

351 OrthoPhyl compares favorably to the phylogenomics pipeline GToTree (Lee, 2019). We
352 ran GToTree v1.8.4 in nucleotide mode on the same genome dataset as OrthoPhyl allowing
353 the use of 28 cores while searching for orthologs with the Alphaproteobacteria HMM gene
354 set containing 117 models. All other options were left as default. The full analysis took
355 slightly more time than OrthoPhyl, using 50 hours of real time with a total CPU time of 1057

356 hours. The alignment length produced was 74.9kb, with 51 million total characters and
357 approximately 1 million being missing/ambiguous.
358 The GToTree topology for deep splits is very similar to the tree generated by OrthoPhyl
359 above. Specifically, all of the species level splits of the Ochrobactrum group are identical
360 between the trees (**Supplemental Figures 2** and **3**). Additionally, the Brucella clade splits the
361 main Orchrobactrum clade and O. endophytica in both trees. Notable disagreements
362 between the trees include many strain level differences in the Ochrobactrum species
363 subtrees and the basal Brucella. Additionally, the classical Brucella clade shows much lower
364 resolution in the GToTree phylogeny, with many splits supported in less than 50% of
365 bootstrap trees (**Supplemental Figure 3, red dots**). These results are perhaps expected due
366 to their relatively recent radiations and GToTree using a highly conserved gene set from
367 Alphaproteobacteria, likely missing Brucella/Ochrobactrum specific phylogenetic signal
368 which could resolve these relationships.
369 **3.2.2 kSNP4 vs. OrthoPhyl: *Brucella/Ochrobactrum***
370 To compare performance of OrthoPhyl with an alternative assembly-to-tree method,
371 kSNP4 (Gardner *et al.*, 2015; Hall and Nisbet, 2023) was run with the same hardware as
372 OrthoPhyl. See Supplemental methods for details. The resulting kSNP4 matrix for the
373 Brucella/Ochrobactrum dataset has 7.15 million SNPs, which is considerably longer than the
374 maximum genome length. Many of these are likely artifacts of the sequences being
375 divergent enough to have mutations within the flanking k-mers used to identify SNPs. Thus,
376 much of the matrix consists of missing data (~78%) indicating *k*-mers not being shared
377 between assemblies.

378 FastTree2 was used externally to generate a tree with the *k*-mer based SNP alignment
379 matrix from kSNP4. Like the OrthoPhyl tree, sites found in less than 30% of assemblies were
380 removed up to the point of removing 80% of total sites. This left 1,569,787 SNPs (exactly
381 20% of original SNPs, with ~13% total missing data). The trimmed SNP alignment file was
382 then used to generate a tree with FastTree2 using the GTR+gamma model and default
383 parameters. For a high-resolution tree with species, strain, bootstap support and assembly
384 accessions labeled, see **Supplemental Figure 4**.
385 For deep branches outside of the classic Brucella, this tree shows largely the same
386 topology as the tree generated by OrthoPhyl. Some notable exceptions within the
387 Ochrobactrum include: 1) the placement of O. gallinifaecis and one of the *O. thiophenivorans*
388 being sister to *O. pseudogrignonensis* in the OrthoPhyl tree while it is sister to the clade of *O.*
389 *pseudogrignonensis, O. quorumnocens, O, pituitosa* (among others) in the kSNP4 based tree
390 **(Figure 4A, orange arrow)**, and 2) *O. daejeonensis* is placed one node more basal in the
391 kSNP4 tree **(Figure 4A, red arrow)**. For the *Brucella* clade, there are 2 major differences. The
392 first is the placement of the root, with kSNP4 showing it coming in-between B. ovis and the
393 rest of Brucella **(Figure 4B, black arrow)**. The second is (given the OrthoPhyl root) *B.*
394 *neotomae* being placed, along with two B. suis samples **(Figure 4B, asterisk)** on the short
395 branch leading to the *B. melitensis, B. abortus, B. ovis, B. pinnipedialis, B. ceti* clade, instead
396 of sister to all *Brucella* internal to *B. microti* **(Figure 4B, purple arrow)**. Compared to kSNP4,
397 OrthoPhyl had favorable runtime and resource usage requirements. Major differences
398 include max memory usage (58 vs 204 GB), max storage (53 vs 480GB) and runtime (48 vs >
399 216 hours) for OrthoPhyl and kSNP4, respectively.

400 **3.3 Order Level Divergence Tree with 1200 samples: Rickettsiales**

401 The order Rickettsiales was chosen as a challenging taxon to test OrthoPhyl's ability to
402 deal with wide phylogenetic distances and extensive genome reduction. Additionally, this set
403 of assemblies was not filtered stringently, leaving a total of 1208 assemblies (7 outgroup) of
404 varying quality (see Datasets section above). As with the *Brucella* analysis, OrthoPhyl was run
405 on a 30 cpu compute node with 500 GB of available ram. The program took just 12 hours
406 and 10 minutes to complete using a total of 17.3 GB memory. The CPU usage efficiency was
407 41%, with a total CPU time of approx. 152 hours **(Table 1)**.

408 Even with this challenging genomic dataset, OrthoPhyl identifies > 29,000 homologous
409 base pairs from strict SCOs and 295,000 bp from relaxed SCOs (found in 366 or more
410 assemblies). These homologous sequences come from 42 and 348 genes respectively (**Table**
411 **1**). The relaxed SCO FastTree2 ML tree reconstructed from the 1201 Rickettsiales assemblies
412 recovers all valid genera as monophyletic (**Figure 5A**). The sole paraphyletic genus recovered
413 is *Candidatus Jidaibacter*. Other than *Jidaibacter*, the recovered topology of genera is
414 consistent with previously published analyses (Salje, 2021; Schön *et al.*, 2022), with
415 two major clades 1) *Wolbachia* sister to *Ehrlichia* and *Anaplasma*, with *Neoricketsia* basal to
416 them and 2) *Rickettsia* and *Orientia* separated from the rest by the root. A high-resolution
417 tree is provided in **Supplemental Figure 5,** which provides species, assembly accession,
418 bootstrap support and number of SCOs per sample.

419 To illustrate OrthoPhyl's ability to resolve narrow evolutionary windows along with

420 order level divergences in the same tree, the species rich *Rickettsia* genus subtree is shown

421 **(Figure 5B)**. Only three *Rickettsia* with species level identification (provided by NCBI

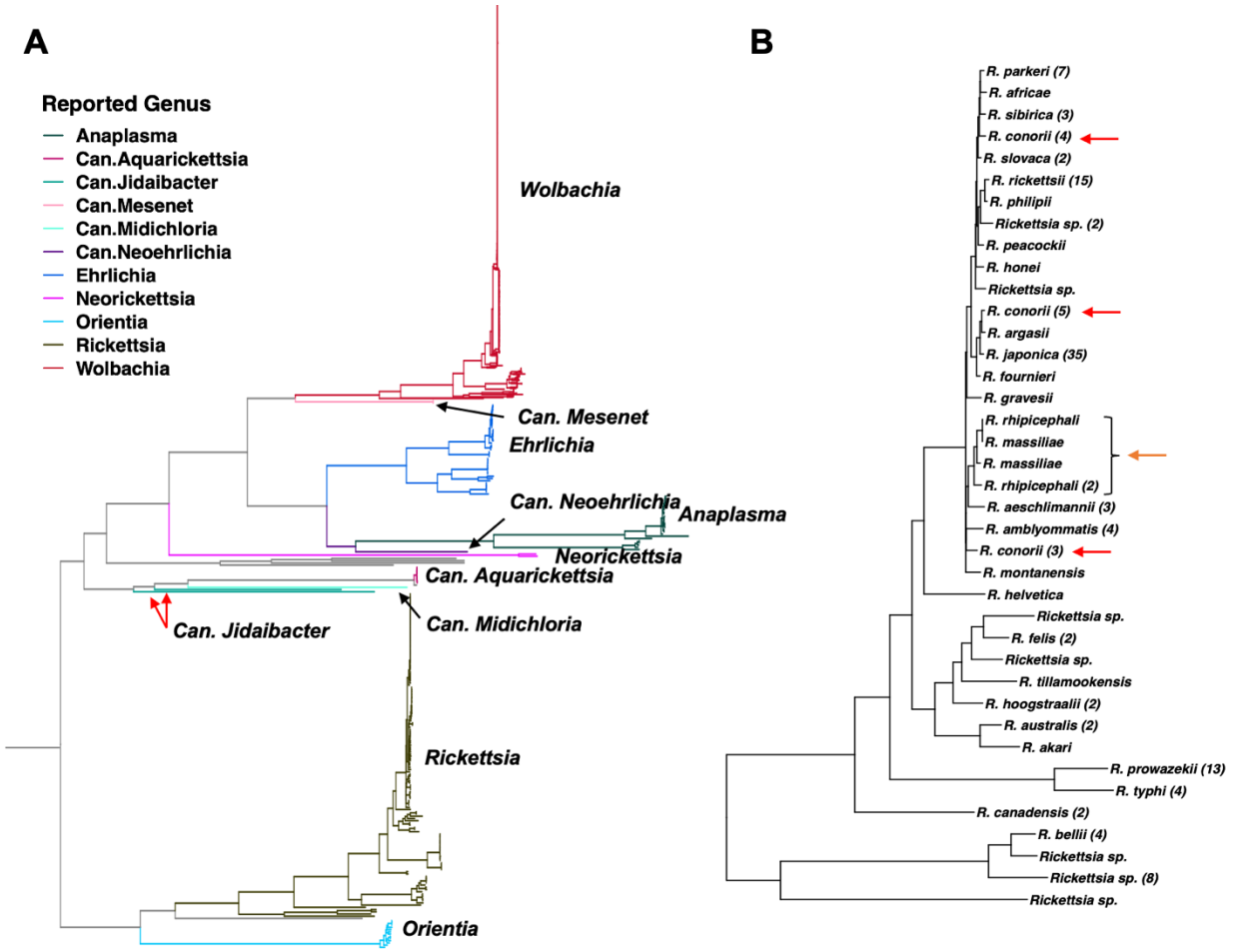

Figure 5 - A) Maximum likelihood tree for Rickettsiales generated by FastTree2 within OrthoPhyl. Branch colors show reported genera for each sample according to NCBI metadata. To aid in identification, labels are also provided. Samples with no genus level identification are labeled in grey. Internal 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 (*Canidatus-Jidibacter*) is labeled with red arrows. **B)** The *Rickettsia* subtree is shown with monophyletic species identifiers collapsed. Following species tip labels, in parentheses is the number of collapsed tips. The two most basal branches are omitted, as they are identified only 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).*

422 metadata) show polytomy. First, samples labeled *R. conorii* are placed in 3 very different
423 clades across the tree **(Figure 5B, red arrows)**, indicating that they are possibly mislabeled in
424 NCBI's genome database. Analysis in Hördt et al, using digital DNA:DNA hybridization, also
425 indicates sequences labeled as *R. conorii* and related species are phylogenenetically
426 problematic. Additionally, the placement of *R. massiliae* and *R. rhipicephali* samples render
427 each other polyphyletic (**Figure 5B, orange arrow**). Other studies mirror our results (Diop *et*
428 *al.*, 2018, 2020). The four R. conorii that are analyzed by Diop et al. are found to be
429 monophyletic in both their papers and this one, with ours having additional R. conorii
430 samples which appear polyphyletic. Diop et al. like us also recovered *R. massiliae* and *R.*
431 *rhipicephali* as problematic.
432 It should be noted that, although we recover identical topologies to others, deeper splits
433 in OrthoPhyl's tree could be affected by saturation at informative sites (**Supplemental figure**
434 **6**). Thus, a robust Rickettsiales analysis would involve also inferring trees from alignments
435 excluding saturated genes and/or sites and potentially from corresponding amino acid
436 matrices.
437
438 **<u>4 Key features</u>**
439 **4.1 Codon alignments**
440 Using codon alignments to infer phylogenetic trees leverages the favorable signal-to-
441 noise ratio of protein alignments (States *et al.*, 1991) while gaining phylogenetic signal
442 through codon degeneracy, allowing greater amounts of information per gene (Wernersson
443 and Pedersen, 2003; Bininda-Emonds, 2005; Kapli *et al.*, 2023).

444 **4.2 Many input assemblies**

445 Whole genome alignment tools, such as Mugsy and Mauve do not scale well for large
446 datasets, with them being limited to generating multiple alignments for about 80 input
447 assemblies (Angiuoli and Salzberg, 2011; Darling *et al.*, 2004)**.** OrthoPhyl can generate
448 alignments for >1000 assemblies, allowing large trees to be inferred. It does this by
449 identifying orthogroups in a diversity-spanning subset of assemblies (see Methods), then
450 assigning proteins to these orthogroups for the full set of predicted proteins by iterative
451 HMM searches. This reduces the computationally demanding all-vs-all protein searching that
452 would make these large analyses intractable for many researchers.

453 **4.3 User friendly**

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

461

462 **5 Discussion**

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

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

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

480 The *Brucella/Ochrobactrum* and Rickettsiales trees are, to our knowledge, the densest
481 phylogenies published for these clades, let alone using whole genome methods. While more
482 work is necessary to validate the topologies presented here, they are a starting point for
483 assessing metadata-based species labels and expected clade monophyly. Future work will
484 test the robustness of topologies to different evolutionary models, gene sets, tree inference
485 methods, and sample incorporation.

486 **5.1 Workflow considerations**

487 There are many parameters to tune during the various steps of phylogenomic inference.
488 Since we focus on usability there is an inevitable tradeoff with optimization, mainly through
489 automated or hard coded parameter choice. OrthoPhyl it does not produce publication
490 ready trees without user inspection of analysis metrics such as alignment quality, single copy
491 ortholog numbers, percent phylogenetically informative sites, mutational saturation and of
492 course branching support values. However, caution should be exercised when interpreting
493 bootstrap support from phylogenomic methods as support values can coalesce on erroneous
494 topologies because of mutational saturation and compositional bias and/or model
495 misspecification. Ultimately, trees should be compared to existing data to ensure major
496 clades are consistent and evidence for novel branching should be closely scrutinized.
497 Users should take different approaches to inferring trees depending on their final goals.
498 If one wishes to infer an initial phylogenetic tree to inform experimental design or taxon
499 sampling, running OrthoPhyl with default parameters to generate a tree with FastTree2 will
500 likely be sufficient. This will produce a tree using the widely applicable GTR+gamma model.
501 To infer a robust tree for publication, researchers should consider using IQtree2, which runs
502 ModelFinder to choose the evolutionary model which best fits their data. They should also
503 ensure even taxon sampling, with a focus on breaking up long branches. For trees aimed at
504 revising the taxonomy of a studied group, users should consider screening out genes with
505 high saturation and inspecting gene trees for signals of horizontal gene transfer.
506 Additionally, building trees using multiple “best” models, tree software, and gene sets then
507 comparing results will add robustness to the inferred topologies. We refer readers to
508 (Lozano-Fernandez, 2022) for an in-depth discussion on the subject.

509 **5.2 Future**

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

519 In the current iteration of OrthoPhyl, very ancient paralogs (originating before the most
520 recent common ancestor of the dataset species) might be clustered into the same
521 orthogroup, leading to the orthogroup's removal during filtering even though 2 sub-trees of
522 the orthogroup could be consistent with the species tree. Future iterations of this software
523 will filter the orthogroups in a deep paralog aware manner, using a rapidly generated
524 neighbor-joining species tree to identify and rescue such orthogroups.

525 Although this pipeline is geared towards prokaryotes, only two changes are needed to
526 adapt it to eukaryotic phylogenetics: 1) allowing users to input precomputed annotations
527 directly, avoiding the complex task of Eukaryotic annotation within the workflow itself and 2)
528 taking into consideration paralog subtrees which are consistent with a consensus species
529 tree. The next version of OrthoPhyl will incorporate these changes to expand the use cases
530 to Eukaryotic phylogenetics.

OrthoPhyl identifies gene family groups for the dataset provided. While this allows the workflow to adapt to any set of genome assemblies, it requires compute resources that could be saved if predefined gene family models could be used as a starting point. Un-clustered protein sequences would then be fed into the standard OrthoPhyl ortholog inference pipeline to capture additional gene families.

5.4 Conclusion

Phylogenetic reconstruction is critical to understanding the evolution of pathogenicity, novel traits, horizontal gene transfer, and metabolic potentials of bacteria. Unfortunately, large-scale phylogenomic analyses of diverse bacterial groups require deep understanding of bioinformatic methods and great care in dealing with the myriad intermediate files and formats used during the analysis. To our knowledge, OrthoPhyl is the only software to take >1000 bacterial assemblies with order level divergence, generate ortholog codon alignments, then reconstruct accurate 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 and knowledge, to leverage the wealth of publicly available genomic data to enrich their ongoing analyses with robust phylogenomic inferences across a broad swath of bacterial diversity.

6 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

Angiuoli,S.V. and Salzberg,S.L. (2011) Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, **27**, 334–342.

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.

Bussi,Y. *et al.* (2021) Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PLOS ONE*, **16**, e0258693.

Capella-Gutiérrez,S. *et al.* (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Chung,M. *et al.* (2018) Using Core Genome Alignments To Assign Bacterial Species. *mSystems*, **3**, e00236-18.

Darling,A.C.E. *et al.* (2004) Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Res.*, **14**, 1394–1403.

Diop,A. *et al.* (2020) Genome sequence-based criteria for demarcation and definition of species in the genus Rickettsia. *Int. J. Syst. Evol. Microbiol.*, **70**, 1738–1750.

Diop,A. *et al.* (2018) Rickettsial genomics and the paradigm of genome reduction associated with increased virulence. *Microbes Infect.*, **20**, 401–409.

Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform. Int. Conf. Genome Inform.*, **23**, 205–211.

Eddy,S.R. (2008) A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. *PLOS Comput. Biol.*, **4**, e1000069.

591 Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLOS Comput. Biol.*, **7**, e1002195.
592 Emms,D.M. and Kelly,S. (2019) OrthoFinder: phylogenetic orthology inference for comparative
593 genomics. *Genome Biol.*, **20**, 238.
594 Emms,D.M. and Kelly,S. (2015) OrthoFinder: solving fundamental biases in whole genome
595 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, **16**,
596 157.
597 Gardner,S.N. *et al.* (2015) kSNP3.0: SNP detection and phylogenetic analysis of genomes
598 without genome alignment or reference genome. *Bioinformatics*, **31**, 2877–2878.
599 Gardner,S.N. and Hall,B.G. (2013) When Whole-Genome Alignments Just Won't Work: kSNP v2
600 Software for Alignment-Free SNP Discovery and Phylogenetics of Hundreds of Microbial
601 Genomes. *PLoS ONE*, **8**, e81760.
602 Gazolla Volpiano,C. *et al.* (2019) Reclassification of Ochrobactrum lupini as a later heterotypic
603 synonym of Ochrobactrum anthropi based on whole-genome sequence analysis. *Int. J.*
604 *Syst. Evol. Microbiol.*, **69**, 2312–2314.
605 Gontcharov,A.A. *et al.* (2004) Are Combined Analyses Better Than Single Gene Phylogenies? A
606 Case Study Using SSU rDNA and rbcL Sequence Comparisons in the Zygnematophyceae
607 (Streptophyta). *Mol. Biol. Evol.*, **21**, 612–624.
608 Hall,B.G. and Nisbet,J. (2023) Building Phylogenetic Trees From Genome Sequences With
609 kSNP4. *Mol. Biol. Evol.*, **40**, msad235.
610 Hoang,D.T. *et al.* (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol.*
611 *Evol.*, **35**, 518–522.
612 Hördt,A. *et al.* (2020) Analysis of 1,000+ Type-Strain Genomes Substantially Improves
613 Taxonomic Classification of Alphaproteobacteria. *Front. Microbiol.*, **11**.
614 Huerta-Cepas,J. *et al.* (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic
615 Data. *Mol. Biol. Evol.*, **33**, 1635–1638.
616 Huerta-Cepas,J. *et al.* (2007) The human phylome. *Genome Biol.*, **8**, R109.
617 Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site
618 identification. *BMC Bioinformatics*, **11**, 119.
619 Jain,C. *et al.* (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear
620 species boundaries. *Nat. Commun.*, **9**, 5114.
621 Kalyaanamoorthy,S. *et al.* (2017) ModelFinder: fast model selection for accurate phylogenetic
622 estimates. *Nat. Methods*, **14**, 587–589.
623 Kapli,P. *et al.* (2023) DNA Sequences Are as Useful as Protein Sequences for Inferring Deep
624 Phylogenies. *Syst. Biol.*, **72**, 1119–1135.
625 Katoh,K. and Standley,D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7:
626 Improvements in Performance and Usability. *Mol. Biol. Evol.*, **30**, 772–780.
627 Konstantinidis,K.T. and Tiedje,J.M. (2007) Prokaryotic taxonomy and phylogeny in the genomic
628 era: advancements and challenges ahead. *Curr. Opin. Microbiol.*, **10**, 504–509.
629 Lee,M.D. (2019) GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics*, **35**,
630 4162–4164.
631 Lozano-Fernandez,J. (2022) A Practical Guide to Design and Assess a Phylogenomic Study.
632 *Genome Biol. Evol.*, **14**, evac129.
633 Nguyen,L.-T. *et al.* (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating
634 Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

635 Orsini,M. *et al.* (2022) Brucella ceti and Brucella pinnipedialis genome characterization unveils
636 genetic features that highlight their zoonotic potential. *MicrobiologyOpen*, **11**, e1329.
637 Parks,D.H. *et al.* (2015) CheckM: assessing the quality of microbial genomes recovered from
638 isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.
639 Portik,D.M. *et al.* (2016) An evaluation of transcriptome-based exon capture for frog
640 phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura).
641 *Mol. Ecol. Resour.*, **16**, 1069–1083.
642 Price,M.N. *et al.* (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large
643 Alignments. *PLOS ONE*, **5**, e9490.
644 Puigbò,P. *et al.* (2019) Genome-Wide Comparative Analysis of Phylogenetic Trees: The
645 Prokaryotic Forest of Life. *Methods Mol. Biol. Clifton NJ*, **1910**, 241–269.
646 Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing
647 phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
648 Salje,J. (2021) Cells within cells: Rickettsiales and the obligate intracellular bacterial lifestyle.
649 *Nat. Rev. Microbiol.*, **19**, 375–390.
650 Sankarasubramanian,J. *et al.* (2019) Development and evaluation of a core genome multilocus
651 sequence typing (cgMLST) scheme for Brucella spp. *Infect. Genet. Evol.*, **67**, 38–43.
652 Schön,M.E. *et al.* (2022) The evolutionary origin of host association in the Rickettsiales. *Nat.*
653 *Microbiol.*, **7**, 1189–1199.
654 Shakya,M. *et al.* (2020) Standardized phylogenetic and molecular evolutionary analysis applied
655 to species across the microbial tree of life. *Sci. Rep.*, **10**, 1723.
656 Shavit Grievink,L. *et al.* (2013) Missing Data and Influential Sites: Choice of Sites for
657 Phylogenetic Analysis Can Be As Important As Taxon Sampling and Model Choice.
658 *Genome Biol. Evol.*, **5**, 681–687.
659 Smith,D.R. (2013) The battle for user-friendly bioinformatics. *Front. Genet.*, **4**, 187.
660 Spencer,M. *et al.* (2007) Conditioned Genome Reconstruction: How to Avoid Choosing the
661 Conditioning Genome. *Syst. Biol.*, **56**, 25–43.
662 Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of
663 large phylogenies. *Bioinforma. Oxf. Engl.*, **30**, 1312–1313.
664 States,D.J. *et al.* (1991) Improved sensitivity of nucleic acid database searches using application-
665 specific scoring matrices. *Methods*, **3**, 66–70.
666 Suárez-Esquivel,M. *et al.* (2020) Brucella Genomics: Macro and Micro Evolution. *Int. J. Mol. Sci.*,
667 **21**, 7749.
668 Suyama,M. *et al.* (2006) PAL2NAL: robust conversion of protein sequence alignments into the
669 corresponding codon alignments. *Nucleic Acids Res.*, **34**, W609–W612.
670 Tang,R. *et al.* (2023) KINN: An alignment-free accurate phylogeny reconstruction method based
671 on inner distance distributions of *k*-mer pairs in biological sequences. *Mol. Phylogenet.*
672 *Evol.*, **179**, 107662.
673 Treangen,T.J. *et al.* (2014) The Harvest suite for rapid core-genome alignment and visualization
674 of thousands of intraspecific microbial genomes.
675 Varghese,N.J. *et al.* (2015) Microbial species delineation using whole genome sequences.
676 *Nucleic Acids Res.*, **43**, 6761–6771.
677 Vishnu,U.S. *et al.* (2015) Identification of Recombination and Positively Selected Genes in
678 Brucella. *Indian J. Microbiol.*, **55**, 384–391.

679 Wernersson,R. and Pedersen,A.G. (2003) RevTrans: multiple alignment of coding DNA from
680 aligned amino acid sequences. *Nucleic Acids Res.*, **31**, 3537–3539.
681 Wiens,J.J. and Morrill,M.C. (2011) Missing Data in Phylogenetic Analysis: Reconciling Results
682 from Simulations and Empirical Data. *Syst. Biol.*, **60**, 719–731.
683 Yang,Z. (1998) On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.*, **47**, 125–133.
684 Zhang,C. *et al.* (2018) ASTRAL-III: polynomial time species tree reconstruction from partially
685 resolved gene trees. *BMC Bioinformatics*, **19**, 153.
686 Zhang,C. *et al.* (2020) ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Mol.*
687 *Biol. Evol.*, **37**, 3292–3307.
688
689
690