Using Core Genome Alignments to Assign Bacterial Species

Matthew Chung[a,b], James B. Munro[a], Julie C. Dunning Hotopp[a,b,c,]#

[a] Institute for Genome Sciences, University of Maryland Baltimore, Baltimore, MD 21201, USA

[b] Department of Microbiology and Immunology, University of Maryland Baltimore, Baltimore, MD 21201, USA

[c] Greenebaum Comprehensive Cancer Center, University of Maryland Baltimore, Baltimore, MD 21201, USA

#Address correspondence to Julie C. Dunning Hotopp, jdhotopp@som.umaryland.edu.

Running Title: Core Genome Alignments to Assign Bacterial Species

Word count Abstract: 371 words

Word count Text: 4,833 words

ABSTRACT

With the exponential increase in the number of bacterial taxa with genome sequence data, a new standardized method is needed to assign bacterial species designations using genomic data that is consistent with the classically-obtained taxonomy. This is particularly acute for unculturable obligate intracellular bacteria like those in the Rickettsiales, where classical methods like DNA-DNA hybridization cannot be used to define species. Within the Rickettsiales, species designations have been applied inconsistently, often obfuscating the relationship between organisms and the context for experimental results. In this study, we generated core genome alignments for a wide range of genera with classically defined species, including *Arcobacter, Caulobacter, Erwinia, Neisseria, Polaribacter, Ralstonia, Thermus*, as well as genera within the Rickettsiales including *Rickettsia*, *Orientia, Ehrlichia, Neoehrlichia, Anaplasma, Neorickettsia,* and *Wolbachia*. A core genome alignment sequence identity (CGASI) threshold of 96.8% was found to maximize the prediction of classically-defined species. Using the CGASI cutoff, the *Wolbachia* genus can be delineated into species that differ from the currently used supergroup designations, while the *Rickettsia* genus is delineated into nine species, as opposed to the current 27 species. Additionally, we find that core genome alignments cannot be constructed between genomes belonging to different genera, establishing a bacterial genus cutoff that suggests the need to create new genera from the *Anaplasma* and *Neorickettsia*. By using core genome alignments to assign taxonomic designations, we aim to provide a high-resolution, robust method for bacterial nomenclature that is aligned with classically-obtained results.

IMPORTANCE

With the increasing availability of genome sequences, we sought to develop and apply a robust, high-resolution method for the assignment of genera and species designations that can recapitulate classically-defined taxonomic designations. We developed genera and species cutoffs using both the length and sequence identity of core genome alignments as taxonomic criteria, respectively. These

41 criteria were then tested on diverse bacterial genera with an emphasis on the taxonomy of organisms
42 within the order Rickettsiales, where species designations have been applied inconsistently. Our results
43 indicate that the *Rickettsia* have an overabundance of species designations and that there are clear
44 demarcations of *Wolbachia* species that do not align precisely with the existing supergroup
45 designations. Lastly, we find that the current *Anaplasma* and *Neorickettsia* genus designations are both
46 too broad and need to be divided.

INTRODUCTION

While acknowledging the disdain some scientists have for taxonomy, Stephen Jay Gould frequently highlighted in his writings how the classifications arising from a good taxonomy both reflects and directs our thinking, stating, “the way we order reflects the way we think. Historical changes in classification are the fossilized indicators of conceptual revolutions” (1). Historically, bacterial species delimitation relied on the phenotypic, morphological, and chemotaxonomic characterization (2-4). The 1960s saw the introduction of molecular techniques in bacterial species delimitation through the use of GC-content (5), DNA-DNA hybridization (6), and 16S rRNA sequencing (7, 8). Currently, databases like SILVA (9) and Greengenes (10) use 16S rRNA sequencing to identify bacteria. However, 16S rRNA sequencing often fails to separate closely-related taxa, and its utility for species-level identification is questionable (10-12). Multilocus sequence analysis (MLSA) has also been used to determine species (13), as has phylogenetic analysis of both rRNA and protein-coding genes (3, 14, 15). Non-genomic mass-spectrometry-based approaches, in which expressed proteins and peptides are characterized, provide complementary data to phenotypic and genomic species delimitations (16, 17) and are used in clinical microbiology laboratories. However, DNA-DNA hybridization (DDH) remains the “gold standard” of defining bacterial species (18, 19), despite its inability to address non-culturable organisms and the intensive labor involved that limits its applicability. A new genome-based bacterial species definition is attractive given the increasing availability of bacterial genomes, rapid sequencing improvements with decreasing sequencing costs, and data standards and databases that enable data sharing.

Average nucleotide identity (ANI) and *in silico* DDH were developed as genomic era tools that allow for bacterial classification with a high correlation to results obtained using wet lab DDH, while bypassing the associated difficulties (20-22). For ANI calculations, the genome of the query organism is split into 1 kbp fragments, which are then searched against the whole genome of a reference organism. The average sequence identity of all matches having >60% overall sequence identity over >70% of their length is

71 defined as the ANI between the two organisms (18). *In silico* DDH, also referred to as digital DDH
72 (dDDH), uses the sequence similarity of conserved regions between the genomes of interests, such as
73 high scoring segment pairs (HSPs) or maximally unique matches (MUMs) (23), to calculate genome-to-
74 genome distances. These distances are converted to a dDDH value, which is intended to be analogous to
75 DDH values obtained using traditional laboratory methods. There are three formulas for calculating
76 dDDH values between two genomes using either (1) the length of all HSPs divided by the total genome
77 length, (2) the sum of all identities found in HSPs divided by the overall HSP length, and (3) the sum of all
78 identities found in HSPs divided by the total genome length, with the second formula being
79 recommended for assigning species designations for draft genomes (24, 25). However, these methods
80 do not report the total length of fragments that match the reference genome, and problems arise when
81 only a small number of fragments are unknowingly used.

82 Recent phylogenomic analyses have shifted towards using the core proteome, a concatenated alignment
83 constructed using the amino acid sequences of genes shared between the organisms of interest (26).
84 However, differences in annotation that affect gene calls can add an unnecessary variable when deriving
85 evolutionary relationships. We propose the use of a nucleotide core genome alignment, constructed
86 using all collinear genomic regions free of rearrangements (27), to infer phylogenomic relationships.
87 Using the length of the core genome alignment along with a sequence similarity matrix we aim to assign
88 taxonomic designations at the genus, species, and strain levels. A core genome alignment-based method
89 provides advantages to its protein-based counterpart in that it is of a higher resolution, independent of
90 annotation, transparent with respect to the data used in the calculations, and very amenable to data
91 sharing and deposition in data repositories.

92 The Rickettsiales are an order within the Alphaproteobacteria is composed of exclusively obligate,
93 intracellular bacteria where classic DNA-DNA hybridization is not possible and bacterial taxonomy is
94 uneven, with each of the genera having its own criteria for assigning genus and species designations.

Within the Rickettsiales, there are three major families: the Anaplasmataceae, Midichloriaceae, and Rickettsiaceae, with is an abundance of genomic data being available for genera within the Anaplasmataceae and Rickettsiaceae. However, species definitions in these families are inconsistent, and organisms in the *Wolbachia* genus lack community-supported species designations, instead relying on a system of supergroup designations.

The last reorganization of the Rickettsiaceae taxonomy occurred in 2001 (28), a time when there were <300 sequenced bacterial genomes (29). As of 2014, there have been more than 14,000 genomes sequenced and with this increase in available genomic information, more informed decisions can be made with regards to taxonomic classification (29). By using core genome alignments, we can condense whole genomes into positions shared between the input genomes and use sequence identity to infer phylogenomic relationships. In this study, we aim to both establish guidelines for delineating bacterial genus and species boundaries that are consistent with classically-defined genus and species designations and apply these guidelines to organisms in the Rickettsiales, including *Rickettsia, Orientia, Ehrlichia, Neoehrlichia, Anaplasma, Neorickettsia*, and *Wolbachia*.

RESULTS

Assessment of genus designations

Core genome alignments were generated using the whole genome aligner Mugsy with both complete and draft genome sequences assembled in ≤100 contigs (27). The Mugsy algorithm uses NUCmer to identify locally colinear blocks (LCBs) between the input genomes, where LCBs are shared colinear genomic regions free of rearrangements. LCBs found in all organisms are selected and filtered to include only positions present in all genomes. The resulting core genome alignment is used to generate a sequence similarity matrix.

Core genome alignments were constructed for the *Arcobacter*, *Caulobacter, Erwinia, Neisseria, Polaribacter, Ralstonia,* and *Thermus* genera. Each of the resultant core genome alignments were ≥0.33 Mbp in size and represented ≥13.6% of the average input genome sizes **(Table 1, Supplementary Table 1)**, suggesting that the technique is applicable to a wide range of classically defined bacterial taxa. Additionally, core genome alignments were successfully constructed from the genomes of four of the six Rickettsiales genera, including 69 *Rickettsia* genomes, 3 *Orientia* genomes, 16 *Ehrlichia* genomes, and 23 *Wolbachia* genomes **(Table 1)**. Each of these four core genome alignments are >0.87 Mbp in length and contain >10% of the average size of the input genomes.

We found our initial core genome alignments for *Anaplasma* and *Neorickettsia* to be considerably shorter, both being ~20 kbp and accounting for ≤2.3% of the average input genome sizes. We believe this reflects that the genomes within these two genera are too broad and need further refinement. Using input genomes from different genera to construct a core genome alignment yields an alignment of an insufficient size to accurately represent the evolutionary distances between the input genomes. For example, when the genome of *Neoehrlichia lotoris* is supplemented to the genomes used to create the 0.49 Mbp *Ehrlichia* core genome alignment, the resultant core genome alignment is 0.11 Mbp, representing only 8.9% of the average input genome size compared to the prior 39.8%. Therefore, we

used subsets of species to test whether the *Anaplasma* and *Neorickettsia* genera are too broadly defined. A core genome alignment generated using only the 20 *A. phagocytophilum* genomes in the 30 *Anaplasma* genome set is 1.25 Mbp and represents 83.3% of the average input genome size, while a core genome alignment of the remaining 10 *Anaplasma* genomes is 0.77 Mbp, 65.3% of the average genome input size. This result suggests that the *Anaplasma* genus should be split into two separate genera. Similarly, when the genome of *N. helminthoeca* Oregon is excluded from the *Neorickettsia* core genome alignment, a 0.77 Mbp *Neorickettsia* core genome alignment is generated which represents 87.4% of the average input genome size, suggesting *N. helminthoeca* Oregon is not of the same genus as the other three *Neorickettsia* genome. For the remainder of the manuscript, these genus reclassifications are used. Given these collective results we recommend that the genus classification level can be defined as a group of genomes that together will yield a core genome alignment that represents ≥10% of the average input genome sizes.

Advantages of nucleotide alignments over protein alignments for bacterial species analyses

While core protein alignments are increasingly used for phylogenetics, a core nucleotide alignment should have more phylogenetically informative positions in the absence of substitution saturation, yielding a greater potential for phylogenetic signal. Nucleotide-based analyses outperform amino acid-based analyses in terms of resolution, branch support, and congruence with independent evidence (30, 31) and outperform amino-acid based analyses at all time scales (32). A core protein alignment generated from 152 genes shared between the ten complete *Wolbachia* genomes contains 16,241 parsimony-informative positions while the core nucleotide alignment contains 124,074 such positions, indicating a 10-fold increase in potentially informative positions **(Figure 1)**. Substitution saturation can negatively impact nucleotide-based phylogenetic distance measurements relative to protein-based phylogenetic methods. However, for each core genome alignment, when the uncorrected pairwise genetic distances are plotted against the model-corrected distances, linear relationships are observed

for all alignments (r^2>0.995), indicating substitution saturation is not hindering the ability of the core genome alignments to represent evolutionary relationships (33) **(Supplementary Figure 1)**. Given our results for numerous taxa, we expect this result to be broadly applicable across bacterial species. Furthermore, maximum likelihood (ML) trees from the core protein alignment and the core nucleotide alignment are quite similar for branch length as well as topology, except for the wNo branch (**Figure 1**), which is indicative of a longstanding problem with resolution in the *Wolbachia* phylogeny (34-37).

Identifying a CGASI cutoff for species delineation

Within the *Rickettsia, Orientia, Ehrlichia, Anaplasma, Neorickettsia, Wolbachia, Arcobacter, Caulobacter, Erwinia, Neisseria, Polaribacter, Ralstonia,* and *Thermus* genome subsets, ANI, dDDH, and CGASI values were calculated for 7,264 pairwise genome comparisons **(Supplementary Table 2)**, of which 601 are between members of the same species. The ANI and CGASI follow a second-degree polynomial relationship (r^2 = 0.977) **(Figure 2A)**. Using this model, the ANI species cutoff of ≥95% is analogous to a CGASI cutoff of ≥96.8%. The dDDH and CGASI follow a third-degree polynomial model (r^2 = 0.978), with a dDDH of 70% being equivalent to a CGASI of 97.6%, indicating the dDDH species cutoff is generally more stringent than the ANI species cutoff **(Figure 2B)**.

The ideal CGASI threshold for species delineation would maximize the prediction of classically defined species, neither creating nor destroying the majority of the classically defined species. Therefore, all pairwise comparisons were classified as either intraspecies, between genomes with the same classically defined species designation, or interspecies, between genomes within the same genus but with different classically defined species designations. Every possible CGASI threshold value was then tested for the ability to recapitulate these classically-defined taxonomic classifications **(Figure 2C)**. In all cases, an abnormally high number of interspecies *Rickettsia* comparisons were found above both the established ANI and dDDH species thresholds consistent with previous observations that guidelines for establishing novel *Rickettsia* species are too lax (38), and as such they were excluded from this specific

analysis. Below a CGASI of 97%, classically defined species begin to be separated while organisms classically defined as different species begin to be collapsed. This coincides with the above calculated ANI-equivalent threshold but differs from the above calculated dDDH-equivalent **(Figure 2C)**. The dDDH-equivalent CGASI threshold of 97.6% failed to predict the classically defined taxa from 100 intraspecies comparisons **(Supplementary Table 3)** while the ANI-equivalent CGASI threshold of 96.8%failed to predict the classically defined taxa from 41 intraspecies comparisons **(Supplementary Table 4)**. Given these results, we selected the ANI-equivalent CGASI value of ≥96.8% to further analyze these taxa, which results in our recommendation of specific taxonomic changes within the Rickettsiaceae, including the *Rickettsia*, *Orientia*, *Anaplasma*, *Ehrlichia*, *Neorickettsia*, and *Wolbachia*.

Rickettsiaceae phylogenomic analyses

Rickettsia

The Rickettsiaceae family includes two genera, the *Rickettsia* and the *Orientia*, and while both genera are obligate intracellular bacteria, *Rickettsia* genomes have undergone more reductive evolution, having a genome size ranging from 1.1-1.5 Mbp (39) compared to the 2.0-2.2 Mbp size of the *Orientia* genome (40). Of the Rickettsiales, the *Rickettsia* genus contains the greatest number of sequenced genomes and named species, containing 69 genome assemblies in ≤100 contigs representing 27 unique species. *Rickettsia* genomes are currently classified based on the Fournier criteria, an MLST approach established in 2003 based on the sequence similarity of five conserved genes: the 16S rRNA, citrate synthase (*gltA*), and three surface-exposed protein antigens (*ompA*, *ompB*, and gene D) (41). To be considered a *Rickettsia* species, an isolate must have a sequence similarity of ≥98.1% 16S rRNA and ≥86.5% *gltA* to at least one preexisting *Rickettsia* species. Within the *Rickettsia*, using *ompA*, *ompB*, and gene D sequence similarities, the Fournier criteria also support the further classification of *Rickettsia* species into three groups: the typhus group, the spotted fever group (SFG), and the ancestral group (41). However, the

Fournier criteria has not yet been amended to classify the more recently established transitional *Rickettsia* group (42), indicating a need to update the *Rickettsia* taxonomic scheme.

A total of 69 *Rickettsia* genomes representative of 27 different established species were used for ANI, dDDH, and core genome alignment analyses, and regardless of the method used, a major reclassification is justified **(Figure 3, Supplementary Table 1)**. A core genome alignment constructed using the 69 *Rickettsia* species genomes yielded a core genome alignment size of ~0.56 Mbp, 42.4% of the lengths of the input *Rickettsia* genomes **(Table 1)**. Within the SFG *Rickettsia*, the CGASI between any two genomes is ≥98.2%, well within the proposed CGASI species cutoff of ≥96.8% (**Figure 3**), while the CGASI is ≤97.2% in the ancestral and typhus groups. If a CGASI cutoff of 96.8% is used to reclassify the *Rickettsia* species, all but two of the SFG *Rickettsia* genomes would be classified as the same species **(Figure 3)**, with the two remaining SFG *Rickettsia* genomes, *R. monacensis* IrR Munich and *Rickettsia* sp. Humboldt, being designated as the same species. This is consistent with ANI as well, while dDDH yields conflicting results (**Figure 3**). For the transitional group *Rickettsia*, *R. akari* and *R. australis* would be collapsed into a single species due to having CGASI values of 97.2% with one another. Similarly, *R. asembonensis, R. felis,* and *R. hoogstraalii,* all classified as transitional group *Rickettsia*, would be collapsed into another species, all having CGASI values of 97.2% with one another. This is consistent with a phylogenomic tree generated from the *Rickettsia* core genome alignments, where the SFG *Rickettsia* genomes have far less sequence divergence compared to the rest of the *Rickettsia* genomes **(Figure 3)**.

Orientia

The organisms within *Orientia* have no standardized criteria to define novel species, with *Orientia chuto* being defined as a novel species based on geographical location and the phylogenetic clustering of the 16S rRNA and two protein coding genes that encode for serine protease *htrA* (47 kDa gene) and an outer membrane protein (56-kDa gene), respectively (43). There are far fewer high-quality *Orientia* genomes, which is partly due to the large number of repeat elements found in *Orientia* genomes, with the genome

of *Orientia tsutsugamushi* being the most highly repetitive sequenced bacterial genome to date with ~42% of its genome being comprised of short repetitive sequences and transposable elements (44).

The *Orientia* core genome alignment was constructed using three *O.tsutusgamushi* genomes and is 0.97 Mbp in size, ~47.6% of the average input genome size, with CGASI values ranging from 96.3-97.2% **(Supplementary Figure 2).** Reclassifying the *Orientia* genomes using a CGASI species cutoff of 96.8% would result in *O. tsutsugamushi* Gillliam and *O.tsutsugamushi* Ikeda being classified as a separate species from *O. tsutsugamushi* Boryong (**Supplementary Figure 2**). In this case, this reclassification would not be consistent with recommendations from using ANI or dDDH, which is likely in part due to an imperfect correlation with the ANI and dDDH species cutoffs (18). In this case, we suspect that ANI is strongly influenced by the large number of repeats in the genome due to ANI calculations being based off the sequence identity of 1 kbp query genome fragments. In comparison, we do not anticipate that whole genome alignments would be confounded by the repeats. While the LCBs may be fragmented by the repeats, creating smaller syntenic blocks, the non-phylogenetically informative repeats are eliminated from an LCB-based analysis.

Anaplasmataceae phylogenomic analyses

Ehrlichia

Within the Anaplasmataceae, species designations are frequently assigned based on sequence similarity and clustering patterns from phylogenetic analyses generated using the sequences of specific genes, like the 16S rRNA, *groEL*, and *glt*A (45). As an example, the species designations for *Ehrlichia khabarensis* and *Ehrlichia ornithorhynchi* are justified based on having a lower sequence similarity for the 16S rRNA, *groEL*, and *glt*A below the maximum similarity that differentiates other *Ehrlichia* species (45-47).

A core genome alignment constructed using 16 *Ehrlichia* genomes, representative of four defined species, yields a 0.49 Mbp alignment, which equates to 39.8% of the average *Ehrlichia* genome size (1.25 Mbp) **(Table 1)**. Using a CGASI species cutoff of 96.8%, the *Ehrlichia chaffeensis* and *Ehrlichia ruminatium*

genomes were recovered as monophyletic and well supported species, which is consistent with dDDH and ANI **(Figure 4**). The two *Ehrlichia muris* and *Ehrlichia* sp. Wisconsin_h genomes have CGASI values >97.8%, indicating the three genomes represent one species, which is consistent with ANI, but not dDDH **(Figure 4**). The genomes of *Ehrlichia* sp. HF and *E. canis* Jake do not have CGASI values ≥96.8% with any other species, confirming their status as individual species, consistent with dDDH and ANI **(Figure 4**). This is consistent with an *Ehrlichia* phylogenomic tree **(Figure 4**).

Anaplasma

A novel *Anaplasma* species is currently defined based on phylogenetic analyses involving the 16S rRNA, *gltA,* and *groEL*, with a new species having a lower sequence identity and a divergent phylogenetic position relative to established *Anaplasma* species (48-51). A core genome alignment constructed using 30 *Anaplasma* genomes yielded a 20 kbp alignment, 1.4% of the average input genome size. As noted before, such low values are indicative of more than one genus represented in the taxa included in the analysis. Thus, CGASI analyses for the *Anaplasma* were done on two *Anaplasma* genome subsets, one containing the twenty *Anaplasma phagocytophilum* genomes and the other containing nine *Anaplasma marginale* and one *Anaplama centrale* genomes **(Table 1)**.

A 1.25 Mbp core genome alignment, consisting of 39.8% of the average input genome size, was constructed using twenty *A. phagocytophilum* genomes that all have CGASI values of ≥96.8%, supporting their designation as members of a single species **(Figure 5)**. This is supported by ANI, but dDDH again yields conflicting results **(Figure 5)**. The core genome alignment generated using the remaining ten *Anaplasma* genomes yields a 0.77 Mbp core genome alignment, 65.3% of the average input genome size. The *A. centrale* genome has CGASI values ranging from 90.7-91.0% when compared to the nine *A. marginale* genomes, supporting *A. centrale* as a separate species from *A. marginale*, consistent with existing taxonomy and with the ANI and dDDH species cutoffs **(Figure 5)**.

Neorickettsia

The *Neorickettsia* genus contains four genome assemblies: *Neorickettsia helminthoeca* Oregon, *Neorickettsia risticii* Illinois, *Neorickettsia sennetsu* Miyayama, and *Neorickettsia* sp. 179522. The genus was first established in 1954 with the discovery of *N. helminthoeca* (52) and in 2001, *N. risticii* and *N sennetsu*, both initially classified as *Ehrlichia* strains, were added to the *Neorickettsia* based on phylogenetic analyses of 16S rRNA and *groESL* (28). A core genome alignment constructed using all four *Neorickettsia* genomes yields a 20 kbp alignment, 2.3% of the average input genome size. When excluding the genome of *N. helminthoeca* Oregon, the three remaining *Neorickettsia* genomes form a core genome alignment of 0.76 Mbp in size, 87.4% of the average input genome size, indicating *N. risticii* Illinois, *N. sennetsu* Miyayama, and *Neorickettsia* sp. 179522 are three distinct species within the same genus while *N. helminthoeca* Oregon is of a separate genus **(Supplementary Figure 3)**. When assessing the *Neorickettsia* species designations using ANI and dDDH cutoffs, the four *Neorickettsia* genomes can only be determined to be different species, as the two techniques are unable to delineate phylogenomic relationships at the genus level.

Wolbachia

The current *Wolbachia* classification system currently lacks traditional species designations and instead groups organisms by supergroup designations using an MLST system consisting of 450-500 bp internal fragments of five genes, *gatB*, *coxA*, *hcpA*, *ftsZ*, and *fbpA* (53). A core genome alignment generated using 23 *Wolbachia* genomes yields a 0.18 Mbp alignment, amounting to 14.8% of the average input genome size. The CGASI values between the supergroup A *Wolbachia*, apart from *w*Inc SM, have CGASI values ≥96.8 **(Figure 6)**. The genome of *w*Inc SM has CGASI values ranging from 94.6%-95.9% when compared to other supergroup A *Wolbachia*, indicating that *w*Inc SM is a different species. This is also supported by ANI and, with a few exceptions, dDDH. However, a phylogenomic tree generated using the *Wolbachia* core genome alignment indicates *w*Inc SM is nested within the other supergroup A *Wolbachia* taxa, a clade with 100% bootstrap support **(Figure 6)**. The ANI, dDDH, and CGASI values of *w*Inc SM are likely

underestimated due to >10% of the genome having ambiguous nucleotide positions, leading to large penalties in sequence identity scores, indicating that these methods may not be suited for genomes with large numbers of ambiguous positions. Among the filarial *Wolbachia* supergroups C and D, the CGASI cutoff of 96.8% would split each of the traditionally recognized supergroups into two groups each, also supported by ANI and dDDH. While *w*Oo and *w*Ov would be the same species, *w*Di Pavia should be considered a different species. Similarly, *w*Bm and *w*Wb would be considered the same species, while *w*Ls should be designated as a separate species. The *w*Cle, *w*Fol, and *w*Ppe endosymbionts from supergroups E, F, and L, respectively, would all be considered distinct species using CGASI, ANI, or dDDH.

The results between CGASI, ANI, and dDDH in supergroup B are discordant. Five of the six supergroup B *Wolbachia* genomes, all except for *w*Tpre, have CGASI values ≥96.8% when compared to one of the other five, indicating the five genomes are of the same species. However, despite *w*Tpre being considered a different species if the CGASI cutoff of 96.8% is used, a phylogenomic tree constructed using the core genome alignment shows *w*Tpre to be nested within the supergroup B *Wolbachia* **(Figure 6).** If *w*Tpre is designated as a different species, this would create a paraphyletic clade in the supergroup B *Wolbachia*, suggesting that *w*Tpre is not a different species, but rather should be included with the other supergroup B *Wolbachia* as a single species. When assessing the ANI and dDDH values between the supergroup B *Wolbachia*, only *w*Pip JHB and *w*Pip Pel share both ANI values ≥95% and dDDH values ≥70% with one another **(Figure 6)**. The remaining three supergroup B *Wolbachia, w*No, *w*Stri, and *w*Aus have ANI values ≥95% with some of the other supergroup B *Wolbachia* while not having a dDDH ≥70% with any other genome, suggesting each supergroup B genome may be a separate species. Currently, the underlying basis for the differences in CGASI, ANI, and dDDH in supergroup B *Wolbachia* endosymbionts are not apparent. Given tradition, we would recommend caution and label all supergroup B *Wolbachia* as a single species with a potential re-evaluation in the future.

Other taxa

Overall, there were 41 pairwise comparisons of organisms across diverse taxa including *Arcobacter*, *Caulobacter*, *Neisseria*, *Orientia*, *Ralstonia*, and *Thermus* that are classically defined as the same species but would be classified as distinct species when assessed using our suggested CGASI threshold **(Supplementary Table 4)**. Additionally, there were 782 pairwise comparisons of organisms classically defined as different species that this threshold suggests should be the same species, of which only 10 were not in the genus *Rickettsia*, instead belonging to the *Arcobacter* and *Caulobacter* **(Supplementary Table 5)**. This suggests that using CGASI with this threshold yields robust results. However, as seen with the supergroup B *Wolbachia* results, using strictly a nucleotide identity-based threshold can lead to paraphyletic groups, such that examination with a phylogenomic tree constructed using the core genome alignment is recommended, as was also observed for *Neisseria*.

A 0.33 Mbp core genome alignment was generated from 66 *Neisseria* genomes, accounting for 14.7% of the average input genome size. The core genome alignment largely supports the classically defined species including *N. meningitidis*, *N. gonorrhoeae*, *N. weaveri,* and *N. lactamica,* although one *N. lactamica* isolate, *N. lactamica* 338.rep1_NLAC, appears to be inaccurately assigned **(Supplementary Figure 4)**. The CGASI values suggest each *N. elongata* isolate is a distinct species, as are the *N. flavescens* isolates. Like what was observed in the *Wolbachia*, a paraphyletic clade is observed when using a CGASI cutoff of 98.6% with the genome of *Neisseria* sp. HMSC061B04 being nested within a clade of two *N. mucosa* genomes, *N. lactamica* 338.rep1_NLAC, and several unnamed *Neisseria* taxa, while not having a CGASI ≥98.6% when compared to any other *Neisseria* genome **(Supplementary Figure 4)**. This highlights that while sequence identity cutoffs derived using nucleotide-based pairwise comparisons are important when delineating species, analyses using phylogenomic trees can aid in resolving instances near the CGASI species cutoff to ensure that paraphyletic clades do not occur. The two results are quite complimentary with the core genome alignment being amenable to ML-based phylogenetic approaches.

DISCUSSION

As noted with the development of MLST (54), sequence data has the advantages of being incredibly standardized and portable. While MLST methods allow for insight at the sub-species levels of taxonomic classifications, and are heavily relied upon during infectious disease outbreaks, they do not provide sufficient resolution at the species level. Increased resolution is required when inferring evolutionary relationships and can be obtained through the construction of whole genome alignments that maximize the number of evolutionarily informative positions.

By generating core genome alignments for different genera, we sought to identify a universal, high-resolution method for species classification. Using the core genome alignment, a set of genomes can be analyzed based on both the length and sequence identity of its core genome alignment. The length of the core genome alignment, which reflects the ability for the genomes to be aligned, is informative in delineating genera. Meanwhile, the pairwise comparisons of core genome alignment sequence identity can be used to delineate species, which can in turn be further validated with a phylogenetic analysis that can be used to resolve paraphyletic clades.

A current issue of the Mugsy aligner is the inability for the software to scale, being only able to computationally handle subsets of at most ~80 genomes. However, for heavily sequenced genera such as the *Rickettsia*, future species assignments for novel genomes do not require constructing a core genome alignment from every available genome to infer evolutionary relationships. Instead, we recommend that for each genus, the relevant experts in the community establish and curate a set of trusted genomes with at least one representative of each named species that should be used for constructing core genome alignments. After an initial assessment, refinement could then be made using a core genome alignment with many more genomes from closely related species.

Through this work we have identified criteria that largely reconstruct classical species definitions in a method that is transparent and portable. In the course of this work we have identified modifications

371 that need to be made to the species and genus designations of a number of organisms, particularly
372 within the Rickettsiales. While we have identified these instances, we recommend that any changes in
373 nomenclature be addressed by collaborative teams of experts in the respective communities.

374 # MATERIALS AND METHODS

375 ## Core genome alignments

376 Genomes used in the taxonomic analyses were downloaded from NCBI GenBank (55). OrthoANIu v1.2
377 (56) and USEARCH v6.1.544 (57) were used for the average nucleotide identity calculations. GGDC v2.1
378 (ggdc.dsmz.de) (25) paired with the recommended BLAST+ alignment tool (58) was used to calculate
379 dDDH values. For analyses in this paper, all dDDH calculations were performed using dDDH formula 2
380 due to the usage of draft genomes in taxonomic analyses (24). Core genome alignments were generated
381 for each of the genome subsets using Mugsy v1.2 (27) and MOTHUR v1.22 (59). Sequence identity
382 matrices for the core genome alignments were created using BioEdit v7.2.5
383 (brownlab.mbio.ncsu.edu/JWB/papers/1999Hall1.pdf). ML phylogenomic trees with 1000 bootstraps
384 were calculated for each core genome alignment using IQ-TREE v1.6.2 (60) paired with ModelFinder (61)
385 to select the best model of evolution and UFBoot2 (62) for fast bootstrap approximation. Trees were
386 visualized and annotated using iTOL v4.1.1 (63). Construction of neighbor-network trees were done
387 using the R packages ape (64) and phangorn (65).

388 ## Core protein alignments

389 Orthologs between complete genomes of the same species were determined using FastOrtho, a
390 reimplementation of OrthoMCL (66) that identifies orthologs using all by all BLAST searches. The amino
391 acid sequences of proteins present in all organisms in only one copy were aligned using MAAFT v7.313
392 (67). For every protein alignment, the best model of evolution was identified using ModelFinder (61) and
393 phylogenomic trees were constructed using an edge-proportional partition model (68) with IQ-TREE

v1.6.2 (60) and UFBoot2 (62) for fast bootstrap approximation. Comparative analyses of core protein and core nucleotide alignment trees were done using the R packages ape (64) and phangorn (65).

ACKNOWLEDGMENTS

We would like to thank Dr. Joseph J. Gillespie for helpful discussions and encouragement. This work was funded by the National Institute of Allergy and Infectious Diseases (U19AI110820).

REFERENCES

1. Gould SJ, William B. Provine Collection on Evolution and Genetics. 1984. Hen's teeth and horse's toes. Norton, New York.

2. Ramasamy D, Mishra AK, Lagier JC, Padhmanabhan R, Rossi M, Sentausa E, Raoult D, Fournier PE. 2014. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol 64:384-91.

3. Schleifer KH. 2009. Classification of Bacteria and Archaea: past, present and future. Syst Appl Microbiol 32:533-42.

4. Bisen PS, Debnath M, Prasad GB. 2012. Identification and Classification of Microbes. *In* P. S. Bisen MDaGBP (ed), Microbes doi:doi:10.1002/9781118311912.ch4.

5. Rossello-Mora R, Amann R. 2001. The species concept for prokaryotes. FEMS Microbiol Rev 25:39-67.

6. Brenner DJ, Fanning GR, Steigerwalt AG. 1972. Deoxyribonucleic acid relatedness among species of Erwinia and between Erwinia species and other enterobacteria. J Bacteriol 110:12-7.

7. Dubnau D, Smith I, Morell P, Marmur J. 1965. Gene conservation in Bacillus species. I. Conserved genetic and nucleic acid base sequence homologies. Proc Natl Acad Sci U S A 54:491-8.

8. Boone DR, Castenholz RW, SpringerLink (Online service). 2001. Bergey's Manual® of Systematic Bacteriology : Overview: A Phylogenetic Backbone and Taxonomic Framework for Procaryotic Systematics, Second Edition. ed. Springer New York : Imprint: Springer, New York, NY.

9. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590-6.

10. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610-8.

11. Palys T, Nakamura LK, Cohan FM. 1997. Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. Int J Syst Bacteriol 47:1145-56.

12. Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45:2761-4.

13. Rossello-Mora R, Amann R. 2015. Past and future species definitions for Bacteria and Archaea. Syst Appl Microbiol 38:209-16.

14. Yarza P, Yilmaz P, Pruesse E, Glockner FO, Ludwig W, Schleifer KH, Whitman WB, Euzeby J, Amann R, Rossello-Mora R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12:635-45.

15. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, Rooney AP, Yi H, Xu XW, De Meyer S, Trujillo ME. 2018. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int J Syst Evol Microbiol 68:461-466.

16. Karlsson R, Gonzales-Siles L, Boulund F, Svensson-Stadler L, Skovbjerg S, Karlsson A, Davidson M, Hulth S, Kristiansson E, Moore ER. 2015. Proteotyping: Proteomic

442 characterization, classification and identification of microorganisms--A prospectus. Syst
443 Appl Microbiol 38:246-57.
444 17. De Bruyne K, Slabbinck B, Waegeman W, Vauterin P, De Baets B, Vandamme P. 2011.
445 Bacterial species identification from MALDI-TOF mass spectra through data analysis and
446 machine learning. Syst Appl Microbiol 34:20-9.
447 18. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007.
448 DNA-DNA hybridization values and their relationship to whole-genome sequence
449 similarities. Int J Syst Evol Microbiol 57:81-91.
450 19. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de
451 Peer Y, Vandamme P, Thompson FL, Swings J. 2005. Opinion: Re-evaluating prokaryotic
452 species. Nat Rev Microbiol 3:733-9.
453 20. Auch AF, von Jan M, Klenk HP, Goker M. 2010. Digital DNA-DNA hybridization for
454 microbial species delineation by means of genome-to-genome sequence comparison.
455 Stand Genomic Sci 2:117-34.
456 21. Aanensen DM, Spratt BG. 2005. The multilocus sequence typing network: mlst.net.
457 Nucleic Acids Res 33:W728-33.
458 22. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-
459 Ponten T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus sequence typing of total-
460 genome-sequenced bacteria. J Clin Microbiol 50:1355-61.
461 23. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004.
462 Versatile and open software for comparing large genomes. Genome Biol 5:R12.

463 24. Auch AF, Klenk HP, Goker M. 2010. Standard operating procedure for calculating
464 genome-to-genome distances based on high-scoring segment pairs. Stand Genomic Sci
465 2:142-8.
466 25. Meier-Kolthoff JP, Auch AF, Klenk HP, Goker M. 2013. Genome sequence-based species
467 delimitation with confidence intervals and improved distance functions. BMC
468 Bioinformatics 14:60.
469 26. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV,
470 Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y
471 Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson
472 RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar
473 N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR,
474 White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli
475 R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus
476 agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A
477 102:13950-5.
478 27. Angiuoli SV, Salzberg SL. 2011. Mugsy: fast multiple alignment of closely related whole
479 genomes. Bioinformatics 27:334-42.
480 28. Dumler JS, Barbet AF, Bekker CP, Dasch GA, Palmer GH, Ray SC, Rikihisa Y, Rurangirwa
481 FR. 2001. Reorganization of genera in the families Rickettsiaceae and Anaplasmataceae
482 in the order Rickettsiales: unification of some species of Ehrlichia with Anaplasma,
483 Cowdria with Ehrlichia and Ehrlichia with Neorickettsia, descriptions of six new species

combinations and designation of Ehrlichia equi and 'HGE agent' as subjective synonyms of Ehrlichia phagocytophila. Int J Syst Evol Microbiol 51:2145-65.

29. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141-61.

30. Simmons MP, Ochoterena H, Freudenstein JV. 2002. Amino acid vs. nucleotide characters: challenging preconceived notions. Mol Phylogenet Evol 24:78-90.

31. Simmons MP, Carr TG, O'Neill K. 2004. Relative character-state space, amount of potential phylogenetic information, and heterogeneity of nucleotide and amino acid characters. Mol Phylogenet Evol 32:913-26.

32. Townsend JP, Lopez-Giraldez F, Friedman R. 2008. The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree. J Mol Evol 67:437-47.

33. Philippe H, Forterre P. 1999. The rooting of the universal tree of life is not reliable. J Mol Evol 49:509-23.

34. Bordenstein SR, Paraskevopoulos C, Dunning Hotopp JC, Sapountzis P, Lo N, Bandi C, Tettelin H, Werren JH, Bourtzis K. 2009. Parasitism and mutualism in *Wolbachia*: what the phylogenomic trees can and cannot say. Mol Biol Evol 26:231-241.

35. Casiraghi M, Bordenstein SR, Baldo L, Lo N, Beninati T, Wernegreen JJ, Werren JH, Bandi C. 2005. Phylogeny of *Wolbachia pipientis* based on *gltA*, *groEL* and *ftsZ* gene sequences: clustering of arthropod and nematode symbionts in the F supergroup, and evidence for further diversity in the *Wolbachia* tree. Microbiology 151:4015-22.

506 36. Lefoulon E, Bain O, Makepeace BL, d'Haese C, Uni S, Martin C, Gavotte L. 2016.
507 Breakdown of coevolution between symbiotic bacteria Wolbachia and their filarial
508 hosts. PeerJ 4:e1840.
509 37. Gerth M, Gansauge MT, Weigert A, Bleidorn C. 2014. Phylogenomic analyses uncover
510 origin and spread of the Wolbachia pandemic. Nat Commun 5:5117.
511 38. Walker DH, Ismail N. 2008. Emerging and re-emerging rickettsioses: endothelial cell
512 infection and early disease events. Nat Rev Microbiol 6:375-86.
513 39. Georgiades K, Merhej V, El Karkouri K, Raoult D, Pontarotti P. 2011. Gene gain and loss
514 events in Rickettsia and Orientia species. Biol Direct 6:6.
515 40. Liao HM, Chao CC, Lei H, Li B, Tsai S, Hung GC, Ching WM, Lo SC. 2017. Intraspecies
516 comparative genomics of three strains of Orientia tsutsugamushi with different
517 antibiotic sensitivity. Genom Data 12:84-88.
518 41. Fournier PE, Dumler JS, Greub G, Zhang J, Wu Y, Raoult D. 2003. Gene sequence-based
519 criteria for identification of new rickettsia isolates and description of Rickettsia
520 heilongjiangensis sp. nov. J Clin Microbiol 41:5456-65.
521 42. Gillespie JJ, Beier MS, Rahman MS, Ammerman NC, Shallom JM, Purkayastha A, Sobral
522 BS, Azad AF. 2007. Plasmids and rickettsial evolution: insight from Rickettsia felis. PLoS
523 One 2:e266.
524 43. Izzard L, Fuller A, Blacksell SD, Paris DH, Richards AL, Aukkanit N, Nguyen C, Jiang J,
525 Fenwick S, Day NP, Graves S, Stenos J. 2010. Isolation of a novel Orientia species (O.
526 chuto sp. nov.) from a patient infected in Dubai. J Clin Microbiol 48:4404-9.

527 44. Salje J. 2017. Orientia tsutsugamushi: A neglected but fascinating obligate intracellular
528 bacterial pathogen. PLoS Pathog 13:e1006657.
529 45. Rar VA, Pukhovskaya NM, Ryabchikova EI, Vysochina NP, Bakhmetyeva SV, Zdanovskaia
530 NI, Ivanov LI, Tikunova NV. 2015. Molecular-genetic and ultrastructural characteristics of
531 'Candidatus Ehrlichia khabarensis', a new member of the Ehrlichia genus. Ticks Tick
532 Borne Dis 6:658-67.
533 46. Inokuma H, Brouqui P, Drancourt M, Raoult D. 2001. Citrate synthase gene sequence: a
534 new tool for phylogenetic analysis and identification of Ehrlichia. J Clin Microbiol
535 39:3031-9.
536 47. Gofton AW, Loh SM, Barbosa AD, Paparini A, Gillett A, Macgregor J, Oskam CL, Ryan UM,
537 Irwin PJ. 2018. A novel Ehrlichia species in blood and Ixodes ornithorhynchi ticks from
538 platypuses (Ornithorhynchus anatinus) in Queensland and Tasmania, Australia. Ticks
539 Tick Borne Dis 9:435-442.
540 48. Ybanez AP, Matsumoto K, Kishimoto T, Inokuma H. 2012. Molecular analyses of a
541 potentially novel Anaplasma species closely related to Anaplasma phagocytophilum
542 detected in sika deer (Cervus nippon yesoensis) in Japan. Vet Microbiol 157:232-6.
543 49. Li H, Zheng YC, Ma L, Jia N, Jiang BG, Jiang RR, Huo QB, Wang YW, Liu HB, Chu YL, Song
544 YD, Yao NN, Sun T, Zeng FY, Dumler JS, Jiang JF, Cao WC. 2015. Human infection with a
545 novel tick-borne Anaplasma species in China: a surveillance study. Lancet Infect Dis
546 15:663-70.

547 50. Yang J, Liu Z, Niu Q, Liu J, Han R, Liu G, Shi Y, Luo J, Yin H. 2016. Molecular survey and
548 characterization of a novel Anaplasma species closely related to Anaplasma capra in
549 ticks, northwestern China. Parasit Vectors 9:603.
550 51. Battilani M, De Arcangeli S, Balboni A, Dondi F. 2017. Genetic diversity and molecular
551 epidemiology of Anaplasma. Infect Genet Evol 49:195-211.
552 52. Philip CB, Hadlow WJ, Hughes LE. 1954. Studies on salmon poisoning disease of canines.
553 I. The rickettsial relationships and pathogenicity of Neorickettsia helmintheca. Exp
554 Parasitol 3:336-50.
555 53. Baldo L, Dunning Hotopp JC, Jolley KA, Bordenstein SR, Biber SA, Choudhury RR, Hayashi
556 C, Maiden MC, Tettelin H, Werren JH. 2006. Multilocus sequence typing system for the
557 endosymbiont Wolbachia pipientis. Appl Environ Microbiol 72:7098-110.
558 54. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K,
559 Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: a
560 portable approach to the identification of clones within populations of pathogenic
561 microorganisms. Proc Natl Acad Sci U S A 95:3140-5.
562 55. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW.
563 2013. GenBank. Nucleic Acids Res 41:D36-42.
564 56. Yoon SH, Ha SM, Lim J, Kwon S, Chun J. 2017. A large-scale evaluation of algorithms to
565 calculate average nucleotide identity. Antonie Van Leeuwenhoek 110:1281-1286.
566 57. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST.
567 Bioinformatics 26:2460-1.

58. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

59. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537-41.

60. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268-74.

61. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587-589.

62. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol 35:518-522.

63. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242-5.

64. Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289-90.

65. Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592-3.

66. Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178-89.

590 67. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple
591 sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059-66.
592 68. Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for
593 Phylogenomic Inference from Supermatrices. Syst Biol 65:997-1008.

594

FIGURE LEGENDS

Figure 1: Comparison of phylogenomic trees generated using the core nucleotide and core protein alignments for ten complete *Wolbachia* genomes

Two phylogenomic trees were generated from the genomes of ten *Wolbachia* strains using **(A)** a core protein alignment (CPA) containing 152 genes present in only one copy in all ten genomes and **(B)** a core genome alignment (CNA) with members of the *Wolbachia* supergroups A (●), B (▲), C (■), D (⬟), and F (●) being represented. Shapes of the same color indicate that the multiple genomes are of the same species as determined using our determined CGASI cutoff of ≥96.8%. The single difference in topology is denoted in grey, otherwise the trees are largely similar in both topology and branch lengths, despite the core protein alignment being 77,868 bp long with 16,241 parsimony-informative positions while the core nucleotide alignment is 579,495 bp long with 124,074 such positions.

Figure 2: ANI, dDDH, and CGASI correlation analysis

CGASI, ANI and dDDH values were calculated for 7,264 pairwise comparisons of genomes within the same genus for *Rickettsia*, *Orientia, Ehrlichia, Anaplasma, Neorickettsia, Wolbachia, Caulobacter, Erwinia, Neisseria, Polaribacter, Ralstonia,* and *Thermus*. **(A)** The CGASI and ANI values for the interspecies comparisons follow a second-degree polynomial model (r^2 = 0.977) with the ANI species cutoff of ≥95% being equivalent to a CGASI of 96.8%, indicated by the blue dotted box. **(B)** The CGASI and dDDH values for all pairwise comparisons follow a third-degree polynomial model (r^2 = 0.978) with the dDDH species cutoff of ≥70% being equivalent to a CGASI of 97.6%, indicated by the red dotted box. **(C)** To identify the optimal CGASI cutoff to use when classifying species, for each increment of the CGASI species cutoff, plotted on the x-axis, the percentage of intraspecies and interspecies comparisons correctly assigned was determined based on classically defined species designations. The ideal cutoff should maximize the prediction of classically defined species for both interspecies and intraspecies comparisons. The ANI-equivalent CGASI species cutoff is represented by the blue dotted line while the dDDH-equivalent CGASI species cutoff is represented by the red dotted line.

Figure 3: Analysis of the ANI, dDDH, and CGASI values of 69 *Rickettsia* genomes

For 69 *Rickettsia* genomes, (**A**) the ANI and dDDH values and (**B**) CGASI values were calculated for each genome comparison and color-coded to illustrate the results with respect to cutoffs of ANI ≥95% and dDDH ≥70%. The color of the shape represents species designations as determined by a CGASI cutoff ≥96.8%. **(C)** A ML phylogenomic tree with 1000 bootstraps was generated using the core genome with bootstrap support values placed next to their corresponding nodes. **(D)** The relationships in the green box on Panel C cannot be adequately visualized at the necessary scale, so they are illustrated separately with a different scale. In all panels, the shape next to each *Rickettsia* genome represents whether the genome originates from an ancestral (●), transitional (⬟), typhus group (▲), or spotted fever group (■) *Rickettsia* species.

Figure 4: Analysis of the ANI, dDDH, and CGASI values of 16 *Ehrlichia* genomes

For 16 *Ehrlichia* genomes, (**A**) the ANI and dDDH values and (**B**) CGASI values were calculated for each genome comparison and color-coded to illustrate the results with respect to cutoffs of ANI ≥95% and dDDH ≥70%. The color of the shape represents species designations as determined by a CGASI cutoff ≥96.8%**. (C)** A ML phylogenomic tree with 1000 bootstraps was generated using the core genome with bootstrap support values placed next to their corresponding nodes.

Figure 5: Analysis of the ANI, dDDH, and CGASI values of 30 *Anaplasma* genomes

(**A**) For 30 *Anaplasma* genomes, the ANI and dDDH values were calculated for each genome comparison and color-coded to illustrate the results with respect to cutoffs of ANI ≥95% and dDDH ≥70%. When attempting to construct a core genome alignment using all 30 *Anaplasma* genomes, only a 20-kbp alignment was generated, accounting for <1% of the average *Anaplasma* genome size. Therefore, CGASI values were calculated after the *Anaplasma* genomes were split into two subsets containing (**B**) twenty *A. phagocytophilum* genomes, and (**C**) the remaining ten *Anaplasma* genomes. Two phylogenomic trees were generated using the core genome alignments from **(D)** 20 *A. phagocytophilum* genomes and **(E)**

644 the remaining ten *Anaplasma* genomes. In all panels, the shape next to each genome denotes genus
645 designations as determined by CGASI while the color of the shape denotes the species as defined by a
646 CGASI cutoff of ≥96.85%.

647 **Figure 6: Analysis of the ANI, dDDH values of 23 *Wolbachia* genomes and CGASI values of 22**
648 ***Wolbachia* genomes**

649 For 23 *Wolbachia* genomes, **(A)** The ANI and dDDH values and (**B**) CGASI values were calculated for each
650 genome comparison and color-coded to illustrate the results with respect to cutoffs of ANI ≥95% and
651 dDDH ≥70%. **(C)** A ML phylogenomic tree was generated using the *Wolbachia* core genome alignment
652 constructed using 22 *Wolbachia* genomes. The species designations in supergroup B show the CGASI-
653 designated species cluster of *w*Pip_Pel, *w*Pip_JHB, *w*Aus, and *w*Stri to be polyphyletic, indicating the
654 need for amendments to the CGASI species criteria. In all panels, the shape denotes the current
655 supergroups A (●), B (▲), C (■), D (⬟), E (⬢), F (⬣) and L (◆) and the color denotes the species as
656 defined by a CGASI cutoff of ≥96.85%.

TABLES

Table 1: Rickettsiales genome and core genome alignment statistics

Genus	Family	Class	Genomes Analyzed	Established Species Represented	Average Genome Size (Mbp)	Core Genome Alignment Size (Mbp)	Percentage Core Genome Composition	Minimum CGASI
Rickettsia	Rickettsiaceae	Alphaproteobacteria	69	27	1.32	0.56	42.4%	81.7%
Orientia	Rickettsiaceae	Alphaproteobacteria	3	1	2.04	0.97	47.5%	96.3%
Ehrlichia	Anaplasmataceae	Alphaproteobacteria	16	4	1.23	0.49	39.8%	81.6%
Neoehrlichia	Anaplasmataceae	Alphaproteobacteria	1	1	1.27	NA	NA	NA
With *Ehrlichia*			17	5	1.24	0.11	8.9%	80.1%
Anaplasma	Anaplasmataceae	Alphaproteobacteria	30	3	1.39	0.02	1.4%	84.0%
A. phagocytophilum			20	1	1.5	1.25	83.3%	98.9%
A. marginale and *A. centrale*			10	2	1.18	0.77	65.3%	90.6%
Neorickettsia	Anaplasmataceae	Alphaproteobacteria	4	3	0.87	0.02	2.3%	85.7%
Excluding *N. helminthoeca* Oregon			3	2	0.87	0.76	87.4%	85.7%
Wolbachia	Anaplasmataceae	Alphaproteobacteria	23	NA	1.22	0.18	14.8%	77.2%
Excluding *w*Ppe			22	NA	1.23	0.54	43.9%	80.1%
Arcobacter	Campylobacteraceae	Epsilonproteobacteria	44	12	2.27	0.43	18.9%	78.9%
Caulobacter	Caulobacteraceae	Alphaproteobacteria	26	4	4.97	1.61	32.4%	82.5%
Erwinia	Enterobacteriaceae	Gammaproteobacteria	22	10	4.42	0.60	13.6%	78.8%
Neisseria	Neisseriaceae	Betaproteobacteria	66	10	2.24	0.33	14.7%	80.5%
Polaribacter	Flavobacteriaceae	Flavobacteria	24	11	3.55	0.80	22.5%	77.9%
Ralstonia	Burkholderiaceae	Betaproteobacteria	21	3	5.44	1.24	22.8%	82.2%
Thermus	Thermaceae	Deinococci	19	11	2.29	1.10	48.0%	80.9%

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1: Assessing substitution saturation for core genome alignments

For each of the 14 core genome alignments that comprise ≥10% of the average input genome size, the uncorrected genetic distance between each of the members was plotted against the TN69-model corrected genetic distance. The red-line represents the best-fit line for each data set while the black, dotted-line represents the identity line (y=x). In all cases, the relationship between the two distances are linear (r^2 >0.995), indicating little substitution saturation in the core genome alignments.

Supplementary Figure 2: Analysis of the ANI, dDDH, and CGASI values of three *Orientia* genomes

For 3 *Orientia tsutsugamushi* genomes, **(A)** the ANI and dDDH values and (**B**) CGASI values were calculated for each genome comparison and color-coded to illustrate the results with respect to cutoffs of ANI ≥95% and dDDH ≥70%. Circles of the sample colors next to the names of each genome indicate members of the same species as defined by a CGASI ≥96.8%.

Supplementary Figure 3: Analysis of the ANI, dDDH, and CGASI of 4 *Neorickettsia* genomes

(a) For the 4 *Neorickettsia* genomes, the ANI and dDDH values were calculated for each genome comparison and color-coded to illustrate the results with respect to cutoffs of ANI ≥95% and dDDH ≥70%. **(b)** CGASI values were calculated and are illustrated using a core genome alignment that could only be constructed using 3 of the *Neorickettsia* genomes, excluding *N. helminthoeca* Oregon. Circles of the sample colors next to the names of each genome indicate members of the same species as defined by a CGASI ≥96.8%.

Supplementary Figure 4: Analysis of the ANI, dDDH, and CGASI of 66 *Neisseria* genomes

For the 66 *Neisseria* genomes, **(A)** the ANI and dDDH values and (**B**) CGASI values were calculated for each genome comparison and color-coded to illustrate the results with respect to cutoffs of ANI ≥95% and dDDH ≥70%. **(c)** A ML phylogenomic tree was generated using 1,000 bootstraps with the 0.33 Mbp

683 *Neisseria* core genome alignment. Circles of the sample colors next to the names of each genome

684 indicate members of the same species as defined by a CGASI ≥96.8

685 ## SUPPLEMENTARY TABLES

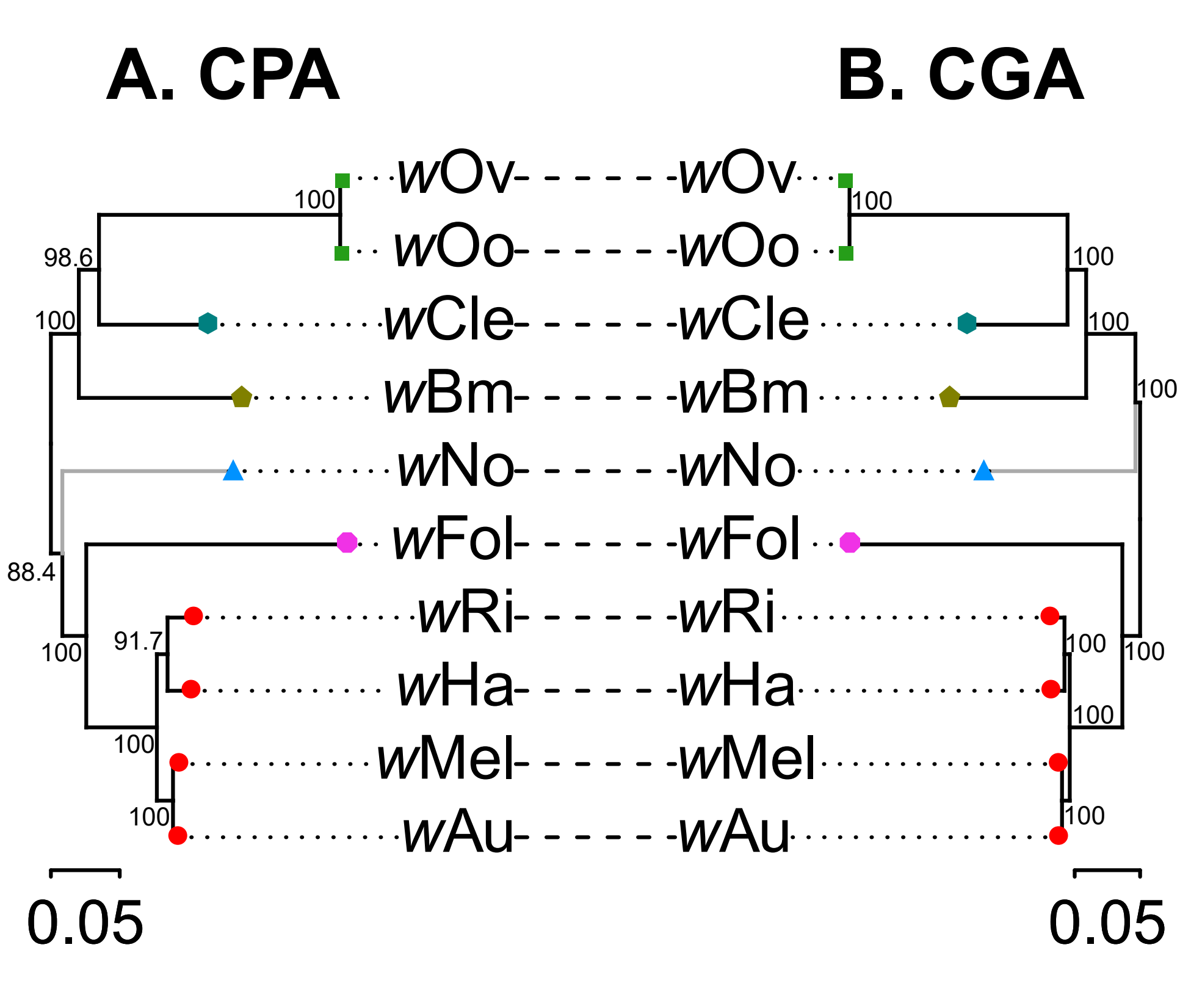
686 **Supplementary Table 1: Genomes used for ANI, dDDH, and CGASI analysis**

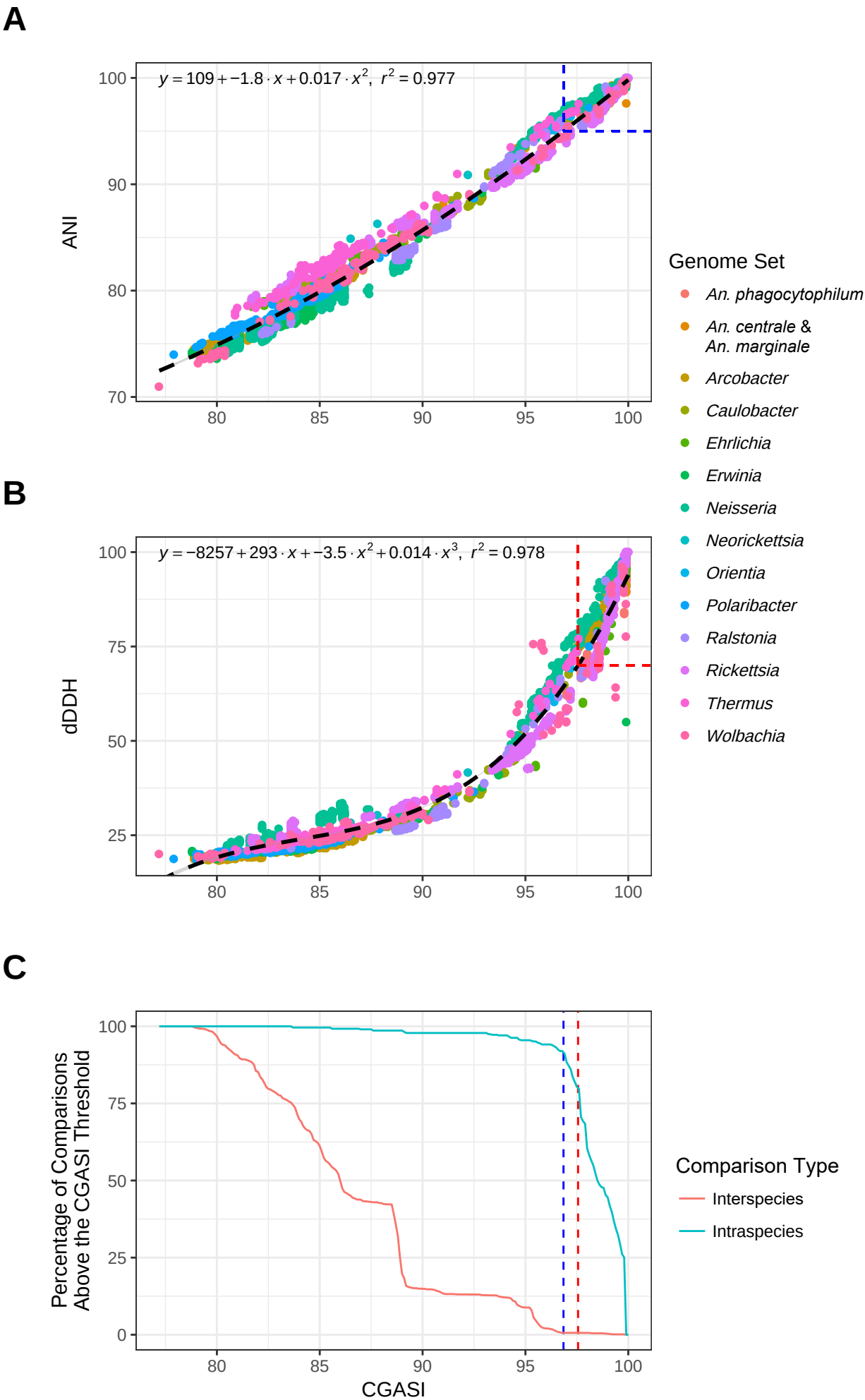
687 **Supplementary Table 2: ANI, dDDH, and CGASI values for 7,264 interspecies comparisons of**
688 **the *Rickettsia, Orientia, Ehrlichia, Anaplasma, Neorickettsia, Wolbachia, Arcobacter,***
689 ***Caulobacter, Erwinia, Neisseria, Polaribacter, Ralstonia,* and *Thermus* genera**

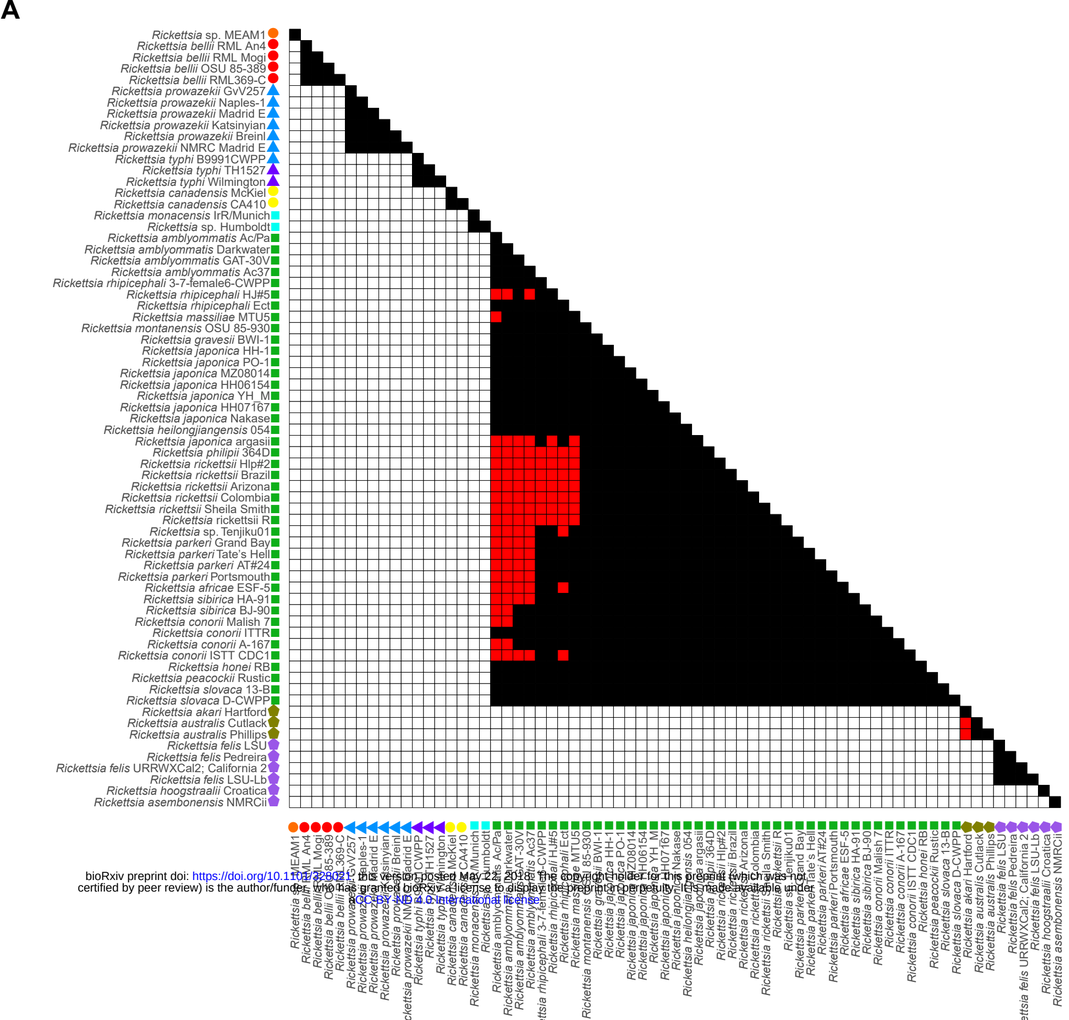
690 **Supplementary Table 3: ANI, dDDH, and CGASI values for 100 intraspecies comparisons with a**
691 **CGASI <97.6%**

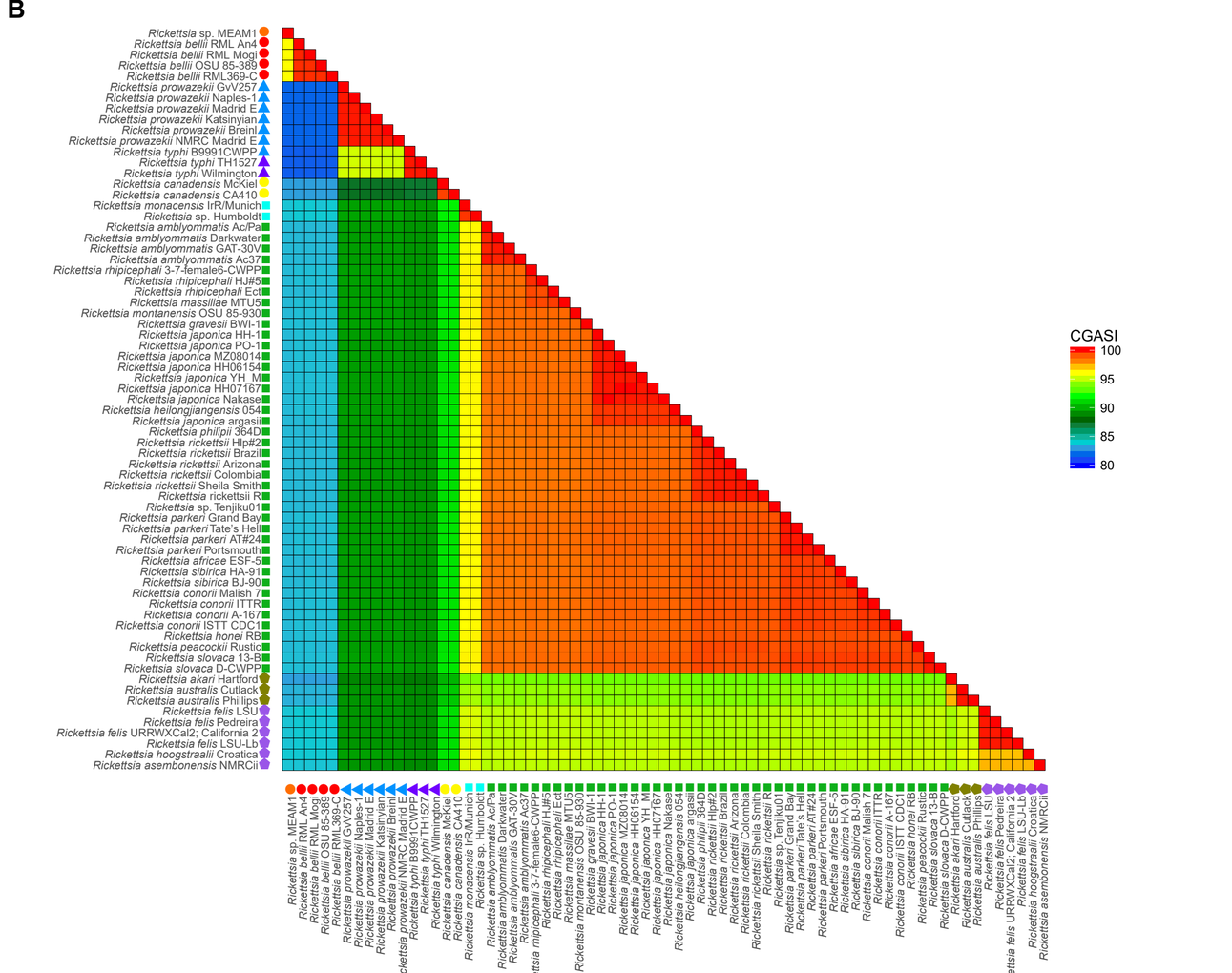
692 **Supplementary Table 4: ANI, dDDH, and CGASI values for 41 intraspecies comparisons with a**
693 **CGASI <96.8%**

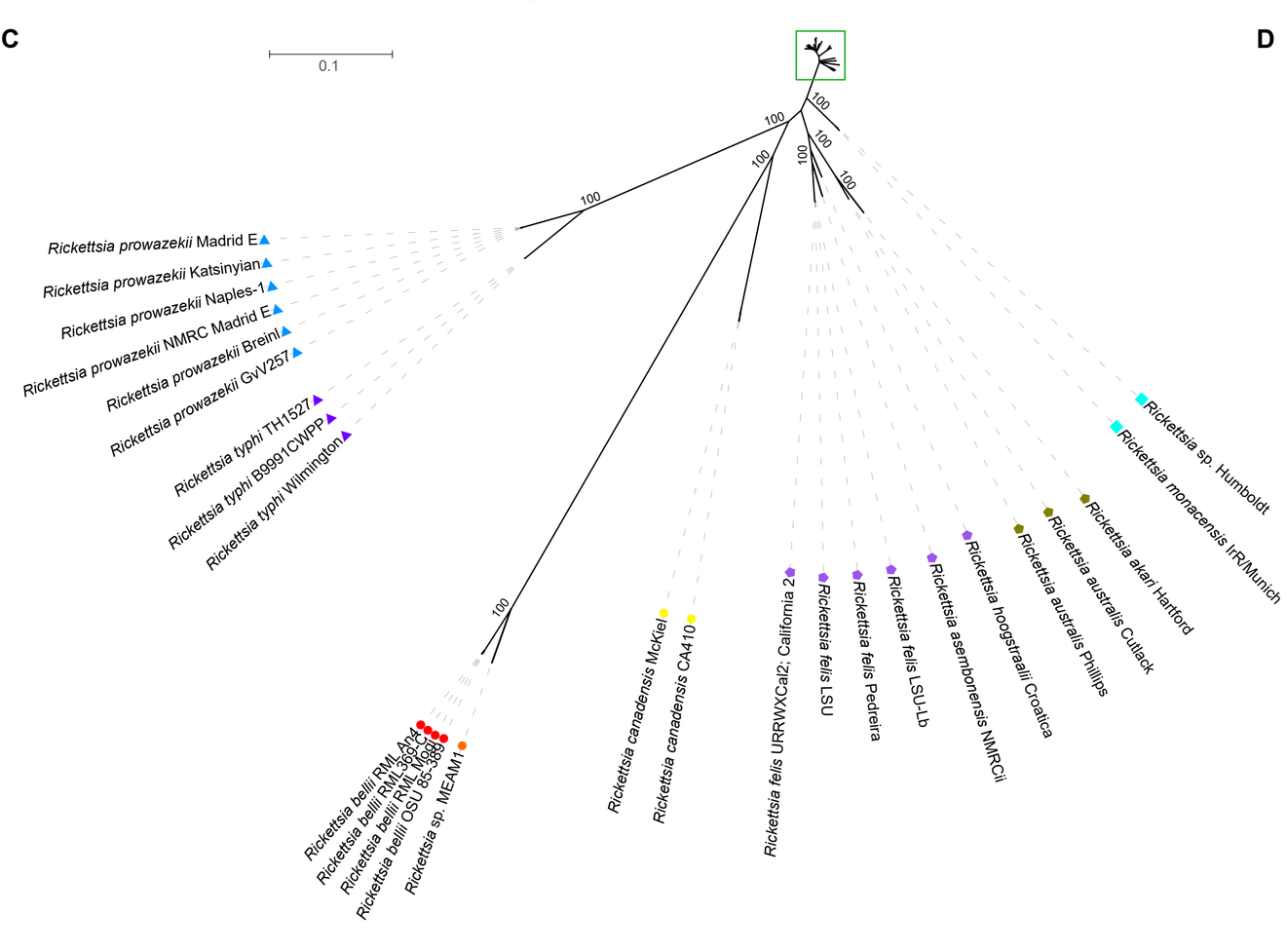
694 **Supplementary Table 5: ANI, dDDH, and CGASI values for 10 non-*Rickettsia* interspecies**
695 **comparisons determined to be intraspecies using the ANI-derived CGASI cutoff of 96.8%**

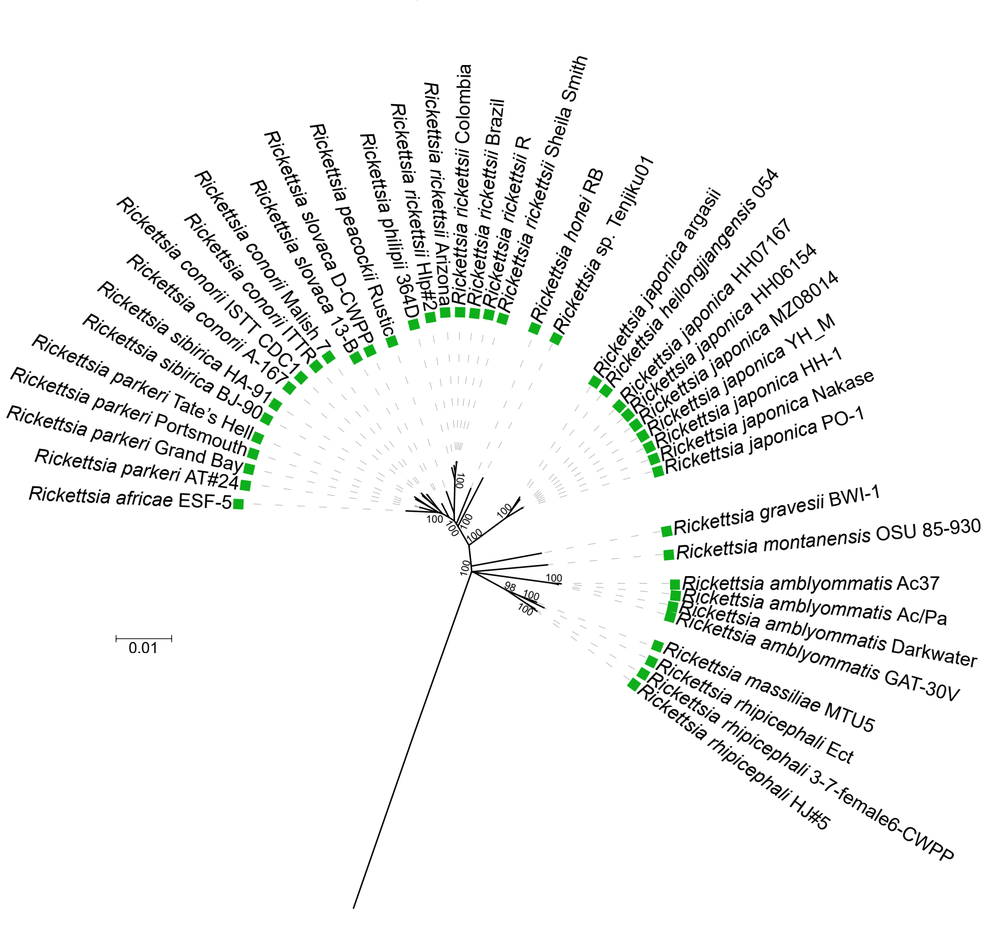

A. CPA
B. CGA
wOv
wOo
wCle
wBm
wNo
wFol
wRi
wHa
wMel
wAu
100
98.6
100
88.4
100
91.7
100
100
100
100
100
100
100
100
100
0.05
0.05


A
$y = 109 + -1.8 \cdot x + 0.017 \cdot x^2,\ r^2 = 0.977$
ANI
100
90
80
70
80
85
90
95
100
B
$y = -8257 + 293 \cdot x + -3.5 \cdot x^2 + 0.014 \cdot x^3,\ r^2 = 0.978$
dDDH
100
75
50
25
80
85
90
95
100
Genome Set
An. phagocytophilum
An. centrale & An. marginale
Arcobacter
Caulobacter
Ehrlichia
Erwinia
Neisseria
Neorickettsia
Orientia
Polaribacter
Ralstonia
Rickettsia
Thermus
Wolbachia
C
Percentage of Comparisons Above the CGASI Threshold
100
75
50
25
0
80
85
90
95
100
CGASI
Comparison Type
Interspecies
Intraspecies

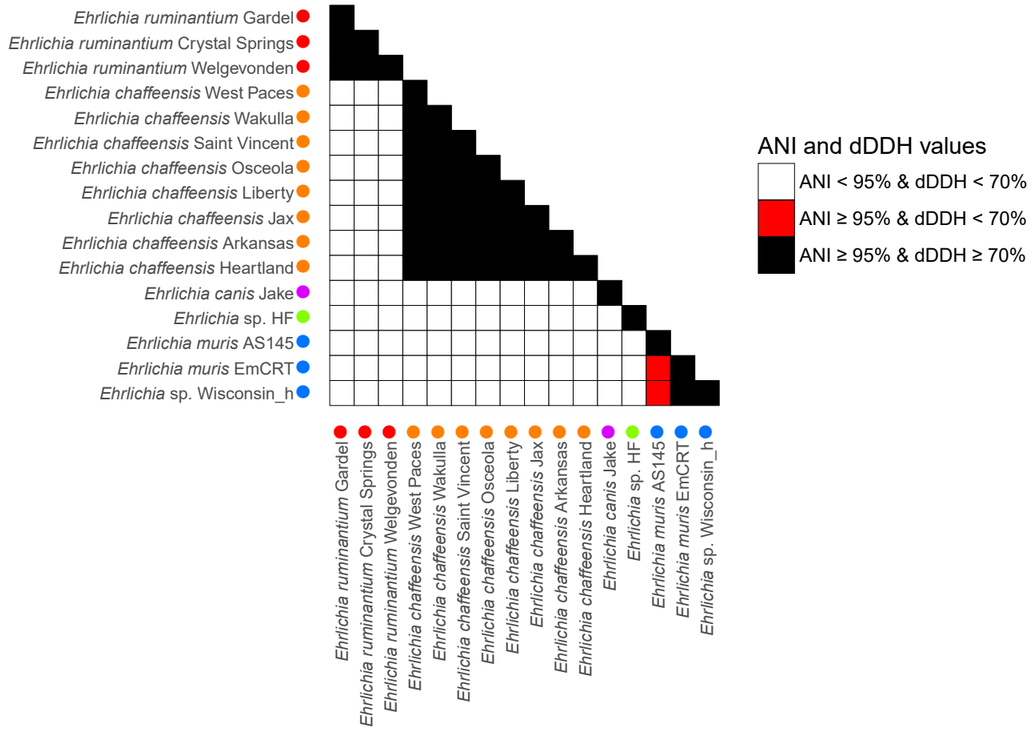




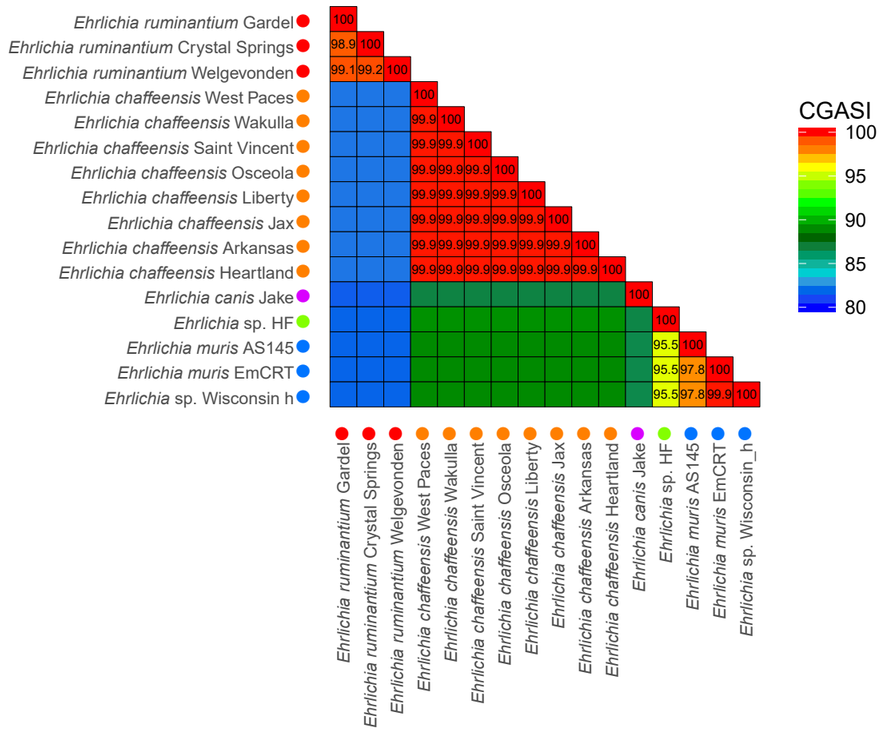




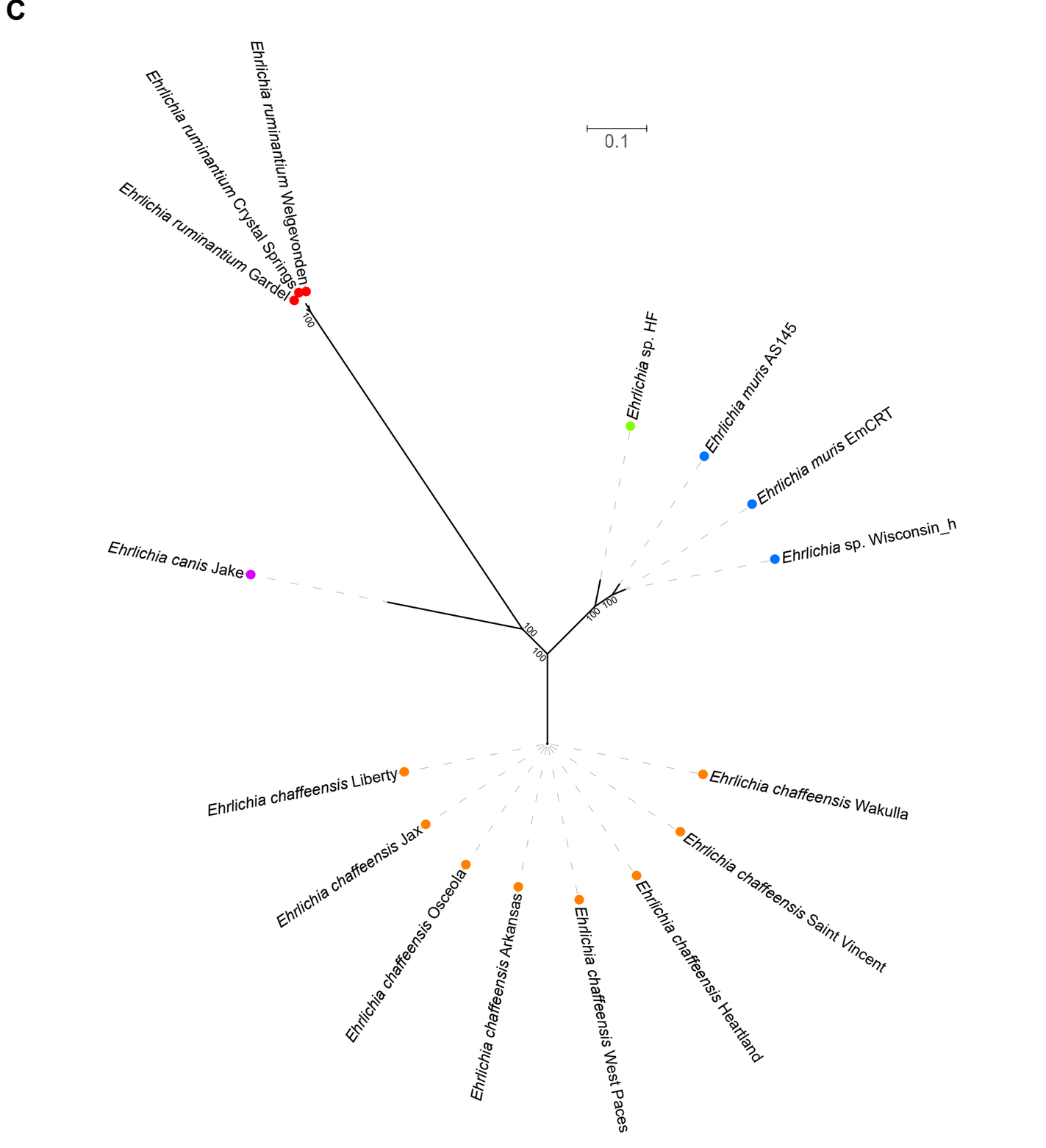
A

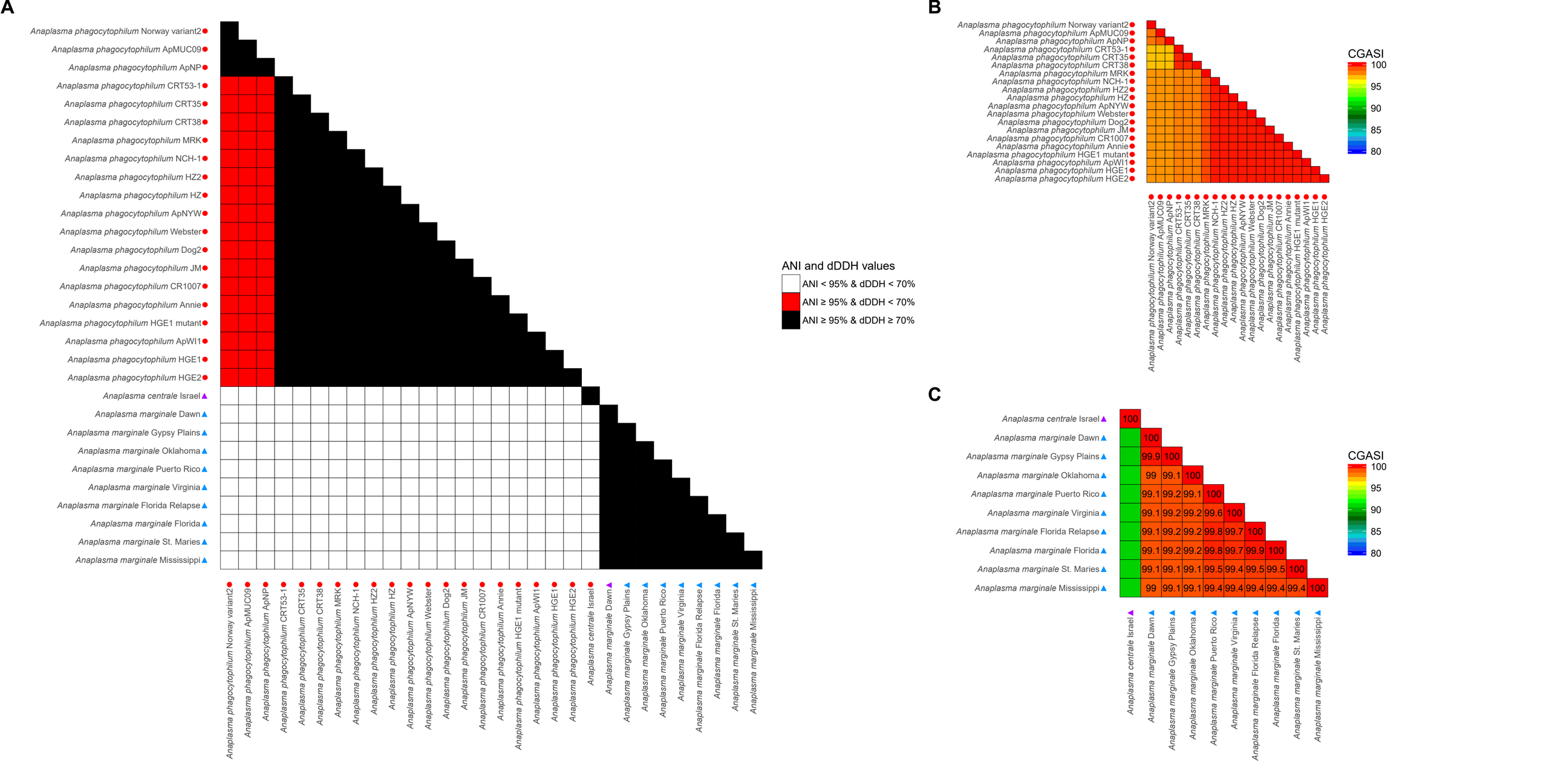

Ehrlichia ruminantium Gardel
Ehrlichia ruminantium Crystal Springs
Ehrlichia ruminantium Welgevonden
Ehrlichia chaffeensis West Paces
Ehrlichia chaffeensis Wakulla
Ehrlichia chaffeensis Saint Vincent
Ehrlichia chaffeensis Osceola
Ehrlichia chaffeensis Liberty
Ehrlichia chaffeensis Jax
Ehrlichia chaffeensis Arkansas
Ehrlichia chaffeensis Heartland
Ehrlichia canis Jake
Ehrlichia sp. HF
Ehrlichia muris AS145
Ehrlichia muris EmCRT
Ehrlichia sp. Wisconsin_h
ANI and dDDH values
ANI < 95% & dDDH < 70%
ANI ≥ 95% & dDDH < 70%
ANI ≥ 95% & dDDH ≥ 70%

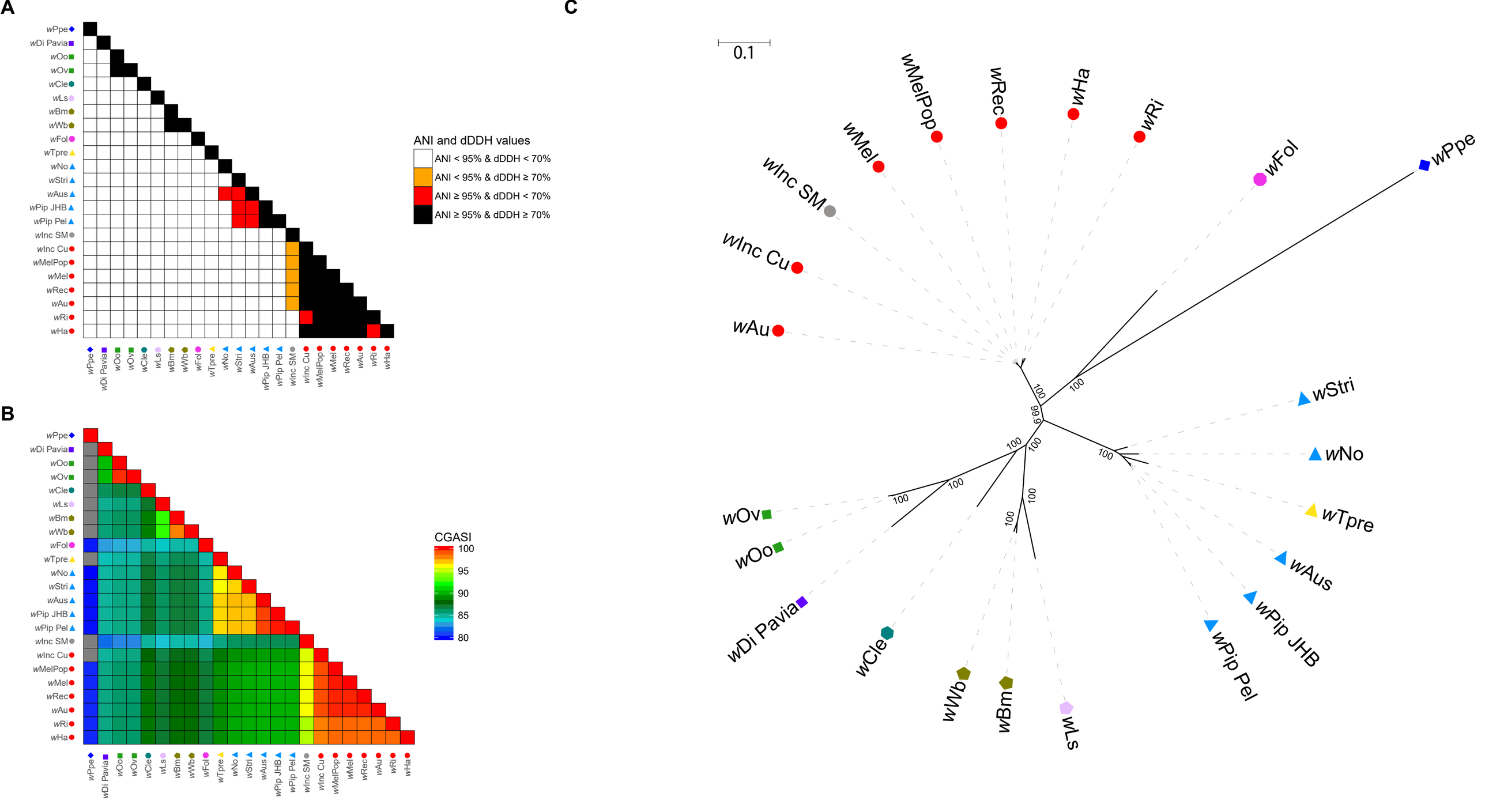
B


Ehrlichia ruminantium Gardel
Ehrlichia ruminantium Crystal Springs
Ehrlichia ruminantium Welgevonden
Ehrlichia chaffeensis West Paces
Ehrlichia chaffeensis Wakulla
Ehrlichia chaffeensis Saint Vincent
Ehrlichia chaffeensis Osceola
Ehrlichia chaffeensis Liberty
Ehrlichia chaffeensis Jax
Ehrlichia chaffeensis Arkansas
Ehrlichia chaffeensis Heartland
Ehrlichia canis Jake
Ehrlichia sp. HF
Ehrlichia muris AS145
Ehrlichia muris EmCRT
Ehrlichia sp. Wisconsin h
CGASI
100
95
90
85
80

C


0.1
Ehrlichia ruminantium Gardel
Ehrlichia ruminantium Crystal Springs
Ehrlichia ruminantium Welgevonden
Ehrlichia sp. HF
Ehrlichia muris AS145
Ehrlichia muris EmCRT
Ehrlichia sp. Wisconsin_h
Ehrlichia canis Jake
Ehrlichia chaffeensis Liberty
Ehrlichia chaffeensis Wakulla
Ehrlichia chaffeensis Jax
Ehrlichia chaffeensis Saint Vincent
Ehrlichia chaffeensis Osceola
Ehrlichia chaffeensis Arkansas
Ehrlichia chaffeensis West Paces
Ehrlichia chaffeensis Heartland
100
100
100
100
100


A
Anaplasma phagocytophilum Norway variant2
Anaplasma phagocytophilum ApMUC09
Anaplasma phagocytophilum ApNP
Anaplasma phagocytophilum CRT53-1
Anaplasma phagocytophilum CRT35
Anaplasma phagocytophilum CRT38
Anaplasma phagocytophilum MRK
Anaplasma phagocytophilum NCH-1
Anaplasma phagocytophilum HZ2
Anaplasma phagocytophilum HZ
Anaplasma phagocytophilum ApNYW
Anaplasma phagocytophilum Webster
Anaplasma phagocytophilum Dog2
Anaplasma phagocytophilum JM
Anaplasma phagocytophilum CR1007
Anaplasma phagocytophilum Annie
Anaplasma phagocytophilum HGE1 mutant
Anaplasma phagocytophilum ApWI1
Anaplasma phagocytophilum HGE1
Anaplasma phagocytophilum HGE2
Anaplasma centrale Israel
Anaplasma marginale Dawn
Anaplasma marginale Gypsy Plains
Anaplasma marginale Oklahoma
Anaplasma marginale Puerto Rico
Anaplasma marginale Virginia
Anaplasma marginale Florida Relapse
Anaplasma marginale Florida
Anaplasma marginale St. Maries
Anaplasma marginale Mississippi
ANI and dDDH values
ANI < 95% & dDDH < 70%
ANI ≥ 95% & dDDH < 70%
ANI ≥ 95% & dDDH ≥ 70%
B
CGASI
100
95
90
85
80
C
CGASI
100
95
90
85
80
100
99.9 100
99 99.1 100
99.1 99.2 99.1 100
99.1 99.2 99.2 99.6 100
99.1 99.2 99.2 99.8 99.7 100
99.1 99.2 99.2 99.8 99.7 99.9 100
99.1 99.1 99.1 99.5 99.4 99.5 99.5 100
99 99.1 99.1 99.4 99.4 99.4 99.4 99.4 100


A
wPpe
wDi Pavia
wOo
wOv
wCle
wLs
wBm
wWb
wFol
wTpre
wNo
wStri
wAus
wPip JHB
wPip Pel
wInc SM
wInc Cu
wMelPop
wMel
wRec
wAu
wRi
wHa
ANI and dDDH values
ANI < 95% & dDDH < 70%
ANI < 95% & dDDH ≥ 70%
ANI ≥ 95% & dDDH < 70%
ANI ≥ 95% & dDDH ≥ 70%
B
CGASI
100
95
90
85
80
C
0.1
wMelPop
wRec
wHa
wRi
wMel
wInc SM
wFol
wPpe
wInc Cu
wAu
wStri
wNo
wTpre
wAus
wPip JHB
wPip Pel
wLs
wBm
wWb
wCle
wDi Pavia
wOo
wOv
100
99.9