Phylogeny Inference of Closely Related Bacterial Genomes: Combining the Features of Both Overlapping Genes and Collinear Genomic Regions

Yan-Cong Zhang[1,2] and Kui Lin[1,2]

[1]State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing, China. [2]MOE Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing, China.

Supplementary Issue: Evolutionary Genomics

ABSTRACT: Overlapping genes (OGs) represent one type of widespread genomic feature in bacterial genomes and have been used as rare genomic markers in phylogeny inference of closely related bacterial species. However, the inference may experience a decrease in performance for phylogenomic analysis of too closely or too distantly related genomes. Another drawback of OGs as phylogenetic markers is that they usually take little account of the effects of genomic rearrangement on the similarity estimation, such as intra-chromosome/genome translocations, horizontal gene transfer, and gene losses. To explore such effects on the accuracy of phylogeny reconstruction, we combine phylogenetic signals of OGs with collinear genomic regions, here called locally collinear blocks (LCBs). By putting these together, we refine our previous metric of pairwise similarity between two closely related bacterial genomes. As a case study, we used this new method to reconstruct the phylogenies of 88 Enterobacteriale genomes of the class *Gammaproteobacteria*. Our results demonstrated that the topological accuracy of the inferred phylogeny was improved when both OGs and LCBs were simultaneously considered, suggesting that combining these two phylogenetic markers may reduce, to some extent, the influence of gene loss on phylogeny inference. Such phylogenomic studies, we believe, will help us to explore a more effective approach to increasing the robustness of phylogeny reconstruction of closely related bacterial organisms.

KEYWORDS: overlapping genes, locally collinear blocks, phylogeny inference, bacteria, genome evolution, pairwise similarity metric

SUPPLEMENT: Evolutionary Genomics

CITATION: Zhang and Lin. Phylogeny Inference of Closely Related Bacterial Genomes: Combining the Features of Both Overlapping Genes and Collinear Genomic Regions. *Evolutionary Bioinformatics* 2015:11(S2) 1–9 doi: 10.4137/EBO.S33491.

TYPE: Original Research

RECEIVED: September 23, 2015. **RESUBMITTED:** November, 10, 2015. **ACCEPTED FOR PUBLICATION:** November 16, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

PEER REVIEW: Six peer reviewers contributed to the peer review report. Reviewers' reports totaled 3171 words, excluding any confidential comments to the academic editor.

FUNDING: This work was supported by the National Natural Science Foundation of China (Grant No. 31171235) and the State Key Laboratory of Earth Surface Processes and Resource Ecology (Grant No. 2013-ZY-10). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: linkui@bnu.edu.cn

Introduction

It is the widespread nature of the arrangement of genome architecture that means two adjacent protein-coding genes have coding sequences that partially or entirely overlap.[1] This phenomenon has been observed in viruses,[2–4] prokaryotes[5–9] and eukaryotes.[10–13] For example, it has been reported that about one-third of all protein-coding genes across completely sequenced bacterial genomes are such overlapping genes (OGs).[14] As to the role of this configuration, OGs are usually assumed to be potentially involved in the regulation of gene expression[1,13,15,16] or improvement of genome compaction.[1,13,16,17]

In our previous works, we demonstrated that OGs could be used as rare genomic markers for bacterial phylogeny inference.[18,19] A previous study shows that gene content might change when gene order is altered too much.[20] As phylogenetic markers, it is evident that OGs do not evolve as slowly as gene content, because of their widespread arrangement in prokaryotic genomes and mutation at a universal rate. However, OGs evolve more conservatively than gene order, because the linkage between two OGs may be preserved for functional constraints.[11,14,21–23] Our idea is simple and intuitive that two closely related bacterial genomes sequenced completely are compared using a similarity metric, which measures the proportion of the shared OGs between the two genomes. Our results and the results of others show that there are indeed some phylogenetic signals within those orthologous OGs among closely related bacterial genomes.[24–28] However, it is obvious that OGs as phylogenetic markers may be inappropriate for comparing a set of too closely related bacterial genomes, such as complete strain genomes of one species, because it is possible that no evolutionary events of OGs would have occurred in a short time span. In addition, for too distantly related bacterial genomes, the OGs method used for phylogeny inference may also show poor performance, as a few orthologous OGs may be identified.

Another drawback of OGs as phylogenetic markers for bacterial phylogenomic analysis in our previous works is that current similarity metrics consider only the presence or absence of one pair of OGs and ignore the relationship with neighborhoods. Thus, the method is usually less sensitive to capturing inconsistent evolutionary events, such as intra-chromosome/genome translocations, gains of OGs from foreign genomes or species, and gene losses.[18,19] To solve the problem caused by gene losses, here we use another type of genomic feature called locally collinear blocks (LCBs). Each LCB, also known as a collinear genomic region, is a homologous region of sequence shared by two or more genomes.[29] Clearly, LCBs from different genomes can be used as orthologous regions, which are likely to contain multiple conserved genes even in their regulatory regions when they are sufficiently large. In some previous studies, LCBs have proved to be reliable genomic features for phylogenomic analysis among closely related genomes on the whole-genome scale.[30,31] By combining OGs with LCBs, additional constraint is built into computations of the similarity between two genomes. In this way, we aim to partly, if not completely, mitigate the above-mentioned drawback of OGs as phylogenetic markers and thus infer phylogeny more accurately.

To test this hypothesis, we studied the phylogenetic relationships of 88 Enterobacteriale genomes of the class *Gammaproteobacteria* with different genomic features. This dataset was selected as a case study because of our interest in the evolution of Enterobacteriale genomes.[18,19,30] First, Sibelia was used to delineate the potential collinear genomic regions between each pair of genomes among the studied species.[32,33] Next, the pairwise similarity of two genomes was measured using OGs as markers. Here, the similarity was computed conditional on the identified pairwise collinear genomic regions rather than on the whole genomes as in our previous works.[18,19] Then, we reconstructed the phylogeny, called OGs–LCBs phylogeny, from the distance matrix based on both OGs and LCBs. Compared with the phylogenies based on only OGs or only LCBs, called OGs phylogeny or LCBs phylogeny, respectively, our OGs–LCBs phylogeny was more similar to a standard 16S rRNA phylogeny and more consistent with taxonomy. This suggests that our OGs–LCBs phylogeny may be robust for phylogenomic analysis among the 88 closely related bacterial organisms. When dealing with inconsistent evolutionary events such as gene losses, combining these two types of genomic features on the whole-genome scale may capture more accurate phylogenetic signals regarding the evolutionary histories. The results demonstrated here show that the analysis of OGs together with LCBs should be useful in accurate phylogeny inference of closely related bacterial genomes.

Materials and Methods

Genome data. As a case study, the dataset of this study consists of 88 Enterobacteriales genomes of *Gammaproteobacteria* (Supplementary Table 1). On August 30, 2015, 112 results were retrieved from the National Center for Biotechnology Information (NCBI) genome dataset by searching with limits of selecting *Gammaproteobacteria* sequences located on the chromosome; these results stand for 112 groups or species. From the 112 groups, any group that had a reference genome to stand for the group, or any group comprising only one member, was chosen. With these selection criteria, 102 groups were selected. Then, the genomes of these selected 102 groups were downloaded from the NCBI database (ftp://ftp.ncbi.nlm.nih.gov/genomes/all). For computational accuracy, only genomes with NCBI assembly level of *complete genome* or *chromosome* were studied. Therefore, 90 complete or nearly complete genomes without plasmids were studied. Following our re-annotation of the 16S rRNAs, two genomes without annotated 16S rRNAs were excluded from the selected 90 genomes. Finally, 88 genomes belonging to the order Enterobacteriales, which cover most genera of the family Enterobacteriaceae, were used in further analysis.

Identification of LCBs among multiple genomes and reconstruction of phylogenies. Each LCB, also known as a collinear genomic region, is a homologous conserved block of sequence among different species.[29] As in our previous study,[30] the core LCBs, which are the set of collinear regions shared by all the species in the study, were used to reconstruct the evolutionary phylogeny. The whole-genome phylogeny inferred from collinear genomic segments was called LCBs phylogeny. First, LCBs were indentified using Sibelia version 3.0.6,[32,33] which can efficiently find LCBs among a large number of microbial genomes without alignment. Chromosome sequences were compared with parameters "-s loose -q -g -v -t tmp --gff -m 100." Second, for each LCB identified by Sibelia, multiple sequences alignment was performed using MAFFT (Multiple sequence Alignment based on Fast Fourier Transform) version 7.164 with parameters "--auto <data>,"[34] and ambiguously aligned regions were removed from the alignment using trimAl version 1.4 with default parameters.[35] Those trimmed alignments were converted to the Multiple Alignment Format (MAF) to get treated LCBs. Third, the core LCBs shared by all the studied genomes were assembled into a concatenated supermatrix. The maximum likelihood (ML) tree was inferred from the data matrices using FastTree version 2.1.8[36] with default Jukes-Cantor + CAT (classifying sites into CATegories) model.[19,36] Local Shimodaira–Hasegawa (SH)-like support was assessed using SH test with 1,000 bootstrap replicates, and the support values are given as names for the internal nodes.

Identification of OGs, orthologous OG pairs, and reconstruction of phylogenies. The phylogeny inferred from OGs was called OGs phylogeny.[18,19,27] OGs are defined as adjacent genes whose coding sequences are shared with each other. OGs were identified from each genome annotation using in-house Perl (a programming language) scripts (Supplementary File 5). All genes annotated as "unknown," "hypothetical" or "putative," which may be misannotated in the genomes downloaded from the NCBI,[27] were removed for more reliable analysis. Considering only protein-coding genes, the putative orthologous genes between two genomes were determined using the approach of bidirectional best hit (BBH). In accordance with previous studies,[18,27] we tested two types of parameter settings for the NCBI BLAST (the Basic Local Alignment Search Tool) program[37]: (1) *e*-value $<10^{-4}$ and identity $>40\%$, used by Luo et al[18] and (2) e-value $< 10^{-8}$, identity $> 45\%$, and coverage $> 85\%$ used by Cheng et al.[27] Orthologous OG pairs from two different genomes were defined as gene pairs that overlap in one genome and have respective orthologous counterparts that overlap in the other genome. Then, the distance matrix among the studied genomes was produced according to the definition of the distance between two genomes as shown in Equation 1. Finally, the neighbor-joining (NJ)[38] tree and unweighted pair-group method with arithmetic mean (UPGMA)[39] tree were inferred from the distance matrix using the Phylogeny Inference Package (PHYLIP) version 3.69[40,41] with default parameters.

The distance between two genomes is defined by Luo et al, which is as follows:

$$D_{ij} = 1 - \frac{x_{ij} + x_{ji}}{2 * \min(x_i, x_j)} (i, j = 1, 2, \ldots, N) \tag{1}$$

where x_i is the number of OG pairs in genome i, N is the number of studied genomes, and x_{ij} is the number of OG pairs in genome i with orthologs in genome j.[19] From this definition, an $N \times N$ distance matrix is produced for phylogeny inference.

Combining OGs and LCBs and reconstruction of phylogenies. By combining the features of OGs and collinear genome region, a phylogeny called OGs–LCBs phylogeny was inferred. First, pairwise comparison was performed to identify pairwise LCBs between any two genomes from the studied genomes using Sibelia version 3.0.6.[32,33] The parameters were set as "-s loose -q -g -v -t tmp --gff -m 100." Second, the orthologous OG pairs, identified in the process of reconstructing the OGs phylogeny described earlier, were selected as follows: if all genes of one orthologous OG pair were completely along one pairwise LCB between these two genomes, this OG pair was selected and called a collinear orthologous OG pair. Then, the distance between two genomes is defined as shown in Equation 2 using collinear orthologous OG pairs instead of the orthologous OG pairs used by Luo et al.[18,19] The distance matrix was generated according to this modified definition, and both the NJ and UPGMA phylogeny trees were inferred using the PHYLIP version 3.69.[40,41]

The distance between two genomes is defined as:

$$D'_{ij} = 1 - \frac{x'_{ij} + x'_{ji}}{2 * \min(x_i, x_j)} (i, j = 1, 2, \ldots, N) \tag{2}$$

where x_i is the number of OG pairs in genome i, N is the number of studied genomes, and x'_{ij} is the number of OG pairs in genome i with orthologs of genome j in the pairwise LCBs between these two genomes. From this definition, an $N \times N$ distance matrix is produced, which can be used to infer the phylogenetic relationships of the species being studied.

Reconstruction of phylogeny with 16S rRNA. To test the usefulness of our method, a standard 16S rRNA phylogeny was reconstructed. All the OGs phylogeny, LCBs phylogeny, and OGs–LCBs phylogeny were compared with this 16S rRNA phylogeny to quantitatively measure the similarity with the Robinson–Foulds topological distance (Table 1 and Supplementary Table 2).[42] The 16S rRNA genes of these 88 genomes downloaded from the NCBI may be annotated using different methods. In order to reduce the inference of differences caused by different annotation methods, rRNAs were re-annotated with our local annotation pipeline, in which Infernal release 1.1[43] was used to annotate possible noncoding RNAs accompanying the Rfam database (release 10.1).[44] Based on the re-annotated 16S rRNA gene, a standard 16S rRNA phylogeny for the 88 Enterobacteriale genomes was reconstructed as follows. First, multiple sequences alignment was performed on the 16S rRNA sequences using MAFFT version 7.164 with the parameters "--auto <data>."[34] Then, ambiguously aligned regions were trimmed using trimAl version 1.4 with default parameters.[35] FastTree version 2.1.8 was performed[36,45] using the default Jukes–Cantor + CAT model to construct an ML tree.[46] Using the SH test with 1,000 bootstrap replicates, local SH-like support was assessed, and the support values are given as names for the internal nodes. Most species belonging to one genus according to the taxonomy were well clustered in the 16S rRNA phylogeny (Fig. 1).

Results and Discussion

The LCBs phylogeny for 88 Enterobacteriales genomes. As a type of phylogenomic marker, LCBs have been proven useful for phylogenomic analysis among closely related genomes or intraspecific prokaryotic genomes on the whole-genome scale.[30,31] The phylogeny was constructed as described in the "Materials and methods" section to explore the performance of LCBs in phylogeny reconstruction among genomes in the Enterobacterials. For the 88 Enterobacteriales genomes, four-core LCBs with the total length of 1,392 bp were identified and comprised 0.02%–0.26% of the lengths of the studied genomes (Supplementary Tables 3 and 4). The phylogeny was constructed using the four-core LCBs and was called the LCBs phylogeny. To quantitatively measure the similarity of the LCBs phylogeny and the 16S rRNA phylogeny, the Robinson–Foulds topological distance was calculated (Table 1). By comparing these two types of phylogenies, we found that many species were not well clustered according to their genera in our LCBs phylogeny, but they were well grouped in the 16S rRNA phylogeny. Especially for the genera *Yersinia*, *Xenorhabdus*, and *Enterobacter*, species from each genus were scattered across different clades in the LCB phylogeny (Fig. 2). This observation indicates that more attention should be paid when using core LCBs as markers to analyze the phylogenetic relationship of species in one order. Although core LCBs are reliable phylogenetic markers of intraspecific relationship mentioned by Zhang et al,[30] minimal conservation of LCBs

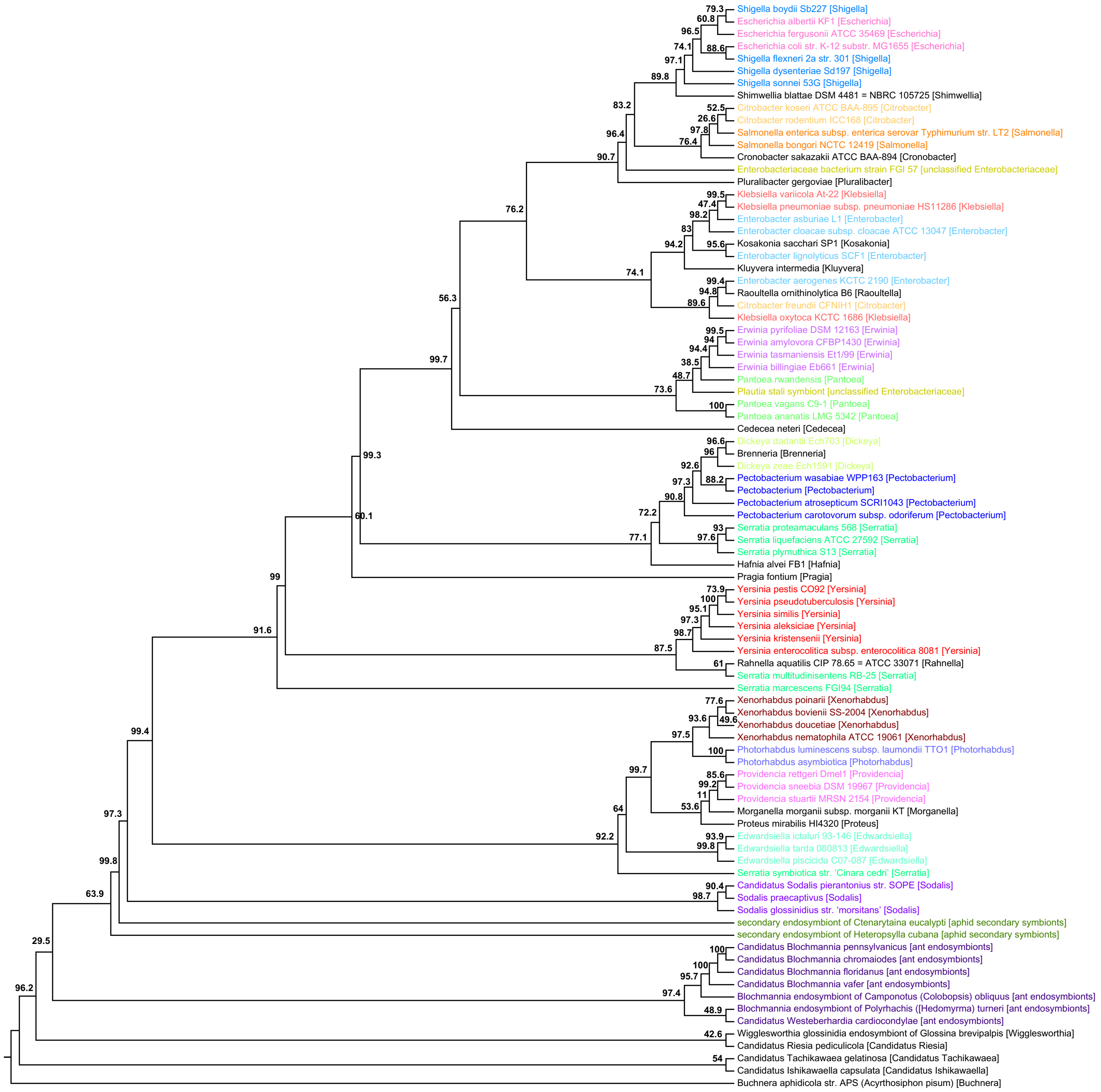
79.3
Shigella boydii Sb227 [Shigella]
60.8
Escherichia albertii KF1 [Escherichia]
96.5
Escherichia fergusonii ATCC 35469 [Escherichia]
74.1
88.6
Escherichia coli str. K-12 substr. MG1655 [Escherichia]
97.1
Shigella flexneri 2a str. 301 [Shigella]
Shigella dysenteriae Sd197 [Shigella]
89.8
Shigella sonnei 53G [Shigella]
Shimwellia blattae DSM 4481 = NBRC 105725 [Shimwellia]
83.2
52.5
Citrobacter koseri ATCC BAA-895 [Citrobacter]
26.6
Citrobacter rodentium ICC168 [Citrobacter]
97.8
96.4
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 [Salmonella]
76.4
Salmonella bongori NCTC 12419 [Salmonella]
90.7
Cronobacter sakazakii ATCC BAA-894 [Cronobacter]
Enterobacteriaceae bacterium strain FGI 57 [unclassified Enterobacteriaceae]
Pluralibacter gergoviae [Pluralibacter]
99.5
Klebsiella variicola At-22 [Klebsiella]
47.4
76.2
Klebsiella pneumoniae subsp. pneumoniae HS11286 [Klebsiella]
98.2
Enterobacter asburiae L1 [Enterobacter]
83
Enterobacter cloacae subsp. cloacae ATCC 13047 [Enterobacter]
94.2
95.6
Kosakonia sacchari SP1 [Kosakonia]
Enterobacter lignolyticus SCF1 [Enterobacter]
74.1
Kluyvera intermedia [Kluyvera]
99.4
Enterobacter aerogenes KCTC 2190 [Enterobacter]
94.8
Raoultella ornithinolytica B6 [Raoultella]
56.3
89.6
Citrobacter freundii CFNIH1 [Citrobacter]
Klebsiella oxytoca KCTC 1686 [Klebsiella]
99.5
Erwinia pyrifoliae DSM 12163 [Erwinia]
94
Erwinia amylovora CFBP1430 [Erwinia]
94.4
Erwinia tasmaniensis Et1/99 [Erwinia]
99.7
38.5
Erwinia billingiae Eb661 [Erwinia]
48.7
Pantoea rwandensis [Pantoea]
73.6
Plautia stali symbiont [unclassified Enterobacteriaceae]
100
Pantoea vagans C9-1 [Pantoea]
Pantoea ananatis LMG 5342 [Pantoea]
Cedecea neteri [Cedecea]
96.6
Dickeya dadantii Ech703 [Dickeya]
96
99.3
Brenneria [Brenneria]
92.6
Dickeya zeae Ech1591 [Dickeya]
97.3
88.2
Pectobacterium wasabiae WPP163 [Pectobacterium]
Pectobacterium [Pectobacterium]
90.8
Pectobacterium atrosepticum SCRI1043 [Pectobacterium]
72.2
60.1
Pectobacterium carotovorum subsp. odoriferum [Pectobacterium]
93
Serratia proteamaculans 568 [Serratia]
77.1
97.6
Serratia liquefaciens ATCC 27592 [Serratia]
Serratia plymuthica S13 [Serratia]
Hafnia alvei FB1 [Hafnia]
99
Pragia fontium [Pragia]
73.9
Yersinia pestis CO92 [Yersinia]
100
Yersinia pseudotuberculosis [Yersinia]
95.1
Yersinia similis [Yersinia]
97.3
Yersinia aleksiciae [Yersinia]
91.6
98.7
Yersinia kristensenii [Yersinia]
87.5
Yersinia enterocolitica subsp. enterocolitica 8081 [Yersinia]
61
Rahnella aquatilis CIP 78.65 = ATCC 33071 [Rahnella]
Serratia multitudinisentens RB-25 [Serratia]
Serratia marcescens FGI94 [Serratia]
77.6
Xenorhabdus poinarii [Xenorhabdus]
93.6
Xenorhabdus bovienii SS-2004 [Xenorhabdus]
49.6
Xenorhabdus doucetiae [Xenorhabdus]
99.4
97.5
Xenorhabdus nematophila ATCC 19061 [Xenorhabdus]
100
Photorhabdus luminescens subsp. laumondii TTO1 [Photorhabdus]
Photorhabdus asymbiotica [Photorhabdus]
99.7
85.6
Providencia rettgeri Dmel1 [Providencia]
99.2
Providencia sneebia DSM 19967 [Providencia]
11
Providencia stuartii MRSN 2154 [Providencia]
64
53.6
Morganella morganii subsp. morganii KT [Morganella]
97.3
Proteus mirabilis HI4320 [Proteus]
93.9
Edwardsiella ictaluri 93-146 [Edwardsiella]
92.2
99.8
Edwardsiella tarda 080813 [Edwardsiella]
99.8
Edwardsiella piscicida C07-087 [Edwardsiella]
Serratia symbiotica str. ‘Cinara cedri’ [Serratia]
90.4
Candidatus Sodalis pierantonius str. SOPE [Sodalis]
63.9
98.7
Sodalis praecaptivus [Sodalis]
Sodalis glossinidius str. ‘morsitans’ [Sodalis]
secondary endosymbiont of Ctenarytaina eucalypti [aphid secondary symbionts]
secondary endosymbiont of Heteropsylla cubana [aphid secondary symbionts]
29.5
100
Candidatus Blochmannia pennsylvanicus [ant endosymbionts]
Candidatus Blochmannia chromaiodes [ant endosymbionts]
100
95.7
Candidatus Blochmannia floridanus [ant endosymbionts]
96.2
Candidatus Blochmannia vafer [ant endosymbionts]
97.4
Blochmannia endosymbiont of Camponotus (Colobopsis) obliquus [ant endosymbionts]
48.9
Blochmannia endosymbiont of Polyrhachis ([Hedomyrma) turneri [ant endosymbionts]
Candidatus Westeberhardia cardiocondylae [ant endosymbionts]
42.6
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis [Wigglesworthia]
Candidatus Riesia pediculicola [Candidatus Riesia]
54
Candidatus Tachikawaea gelatinosa [Candidatus Tachikawaea]
Candidatus Ishikawaella capsulata [Candidatus Ishikawaella]
Buchnera aphidicola str. APS (Acyrthosiphon pisum) [Buchnera]

Figure 1. The phylogeny of 88 Enterobacteriales genomes inferred using 16S rRNA. The maximum likelihood tree was constructed, called 16S rRNA phylogeny, using the 16S rRNA gene. The 16S rRNA phylogeny was inferred from this single gene using FastTree version 2.1.8[36] with default Jukes–Cantor + CAT model. Local SH-like support was assessed using the SH test with 1,000 bootstrap replicates, and the support values are given as names for the internal nodes. Species are denoted with their taxa names in the NCBI, and the corresponding genera are indicated in the square brackets. Species in the same genus are colored with the same color. Those genera with only one member in the study were colored with black.

Table 1. The Robinson–Foulds topological distances between different types of phylogenies.

	LCBs PHYLOGENY[a]	OGs PHYLOGENY[b]	OGs-LCBs PHYLOGENY[b]
16S rRNA phylogeny[a]	144	128	118

Notes: [a]16S rRNA phylogeny and LCBs phylogeny were constructed with FastTree version 2.1.8.[36] [b]Orthologous genes were identified using the approach of BBH by setting the parameters with *e*-value $<10^{-8}$, identity >45%, and coverage >5% used by Cheng et al.[27] The OGs phylogeny and OGs–LCBs phylogeny were constructed with the NJ[38] method using the PHYLIP version 3.69.[41]

may not provide sufficient information and resolution for the analysis of phylogenetic relationship among bacterial genomes in the order Enterobacteriales.

The OGs phylogeny for 88 Enterobacteriales genomes. In addition, another type of phylogeny for the 88 Enterobacteriales genomes was inferred based on OGs, which was called OGs phylogeny. As described in the "Materials and methods" section, two types of parameter settings were tested to identify orthologous protein-coding genes. Therefore, two types of OGs phylogeny were reconstructed with the NJ method

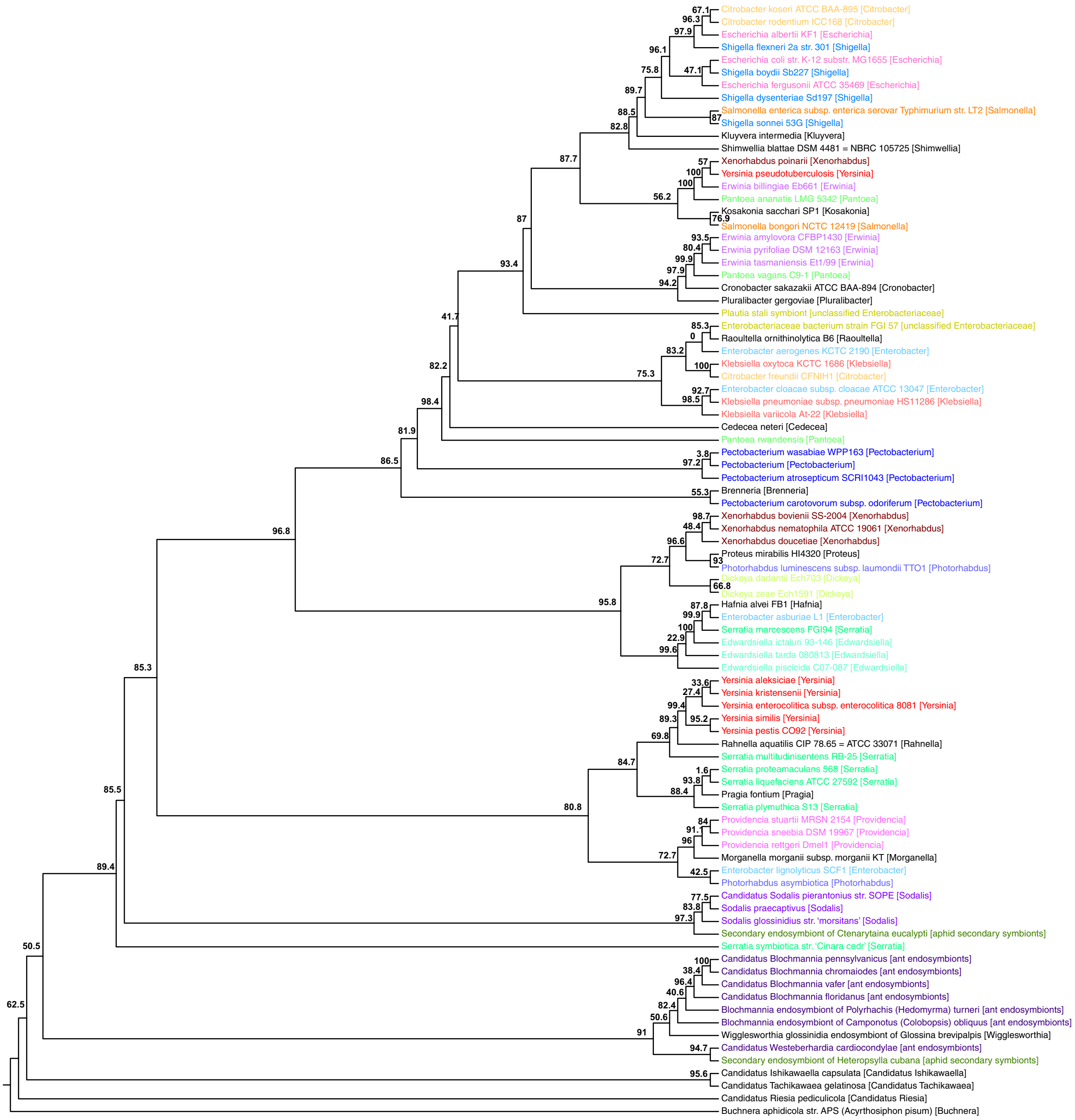
Citrobacter koseri ATCC BAA-895 [Citrobacter]
Citrobacter rodentium ICC168 [Citrobacter]
Escherichia albertii KF1 [Escherichia]
Shigella flexneri 2a str. 301 [Shigella]
Escherichia coli str. K-12 substr. MG1655 [Escherichia]
Shigella boydii Sb227 [Shigella]
Escherichia fergusonii ATCC 35469 [Escherichia]
Shigella dysenteriae Sd197 [Shigella]
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 [Salmonella]
Shigella sonnei 53G [Shigella]
Kluyvera intermedia [Kluyvera]
Shimwellia blattae DSM 4481 = NBRC 105725 [Shimwellia]
Xenorhabdus poinarii [Xenorhabdus]
Yersinia pseudotuberculosis [Yersinia]
Erwinia billingiae Eb661 [Erwinia]
Pantoea ananatis LMG 5342 [Pantoea]
Kosakonia sacchari SP1 [Kosakonia]
Salmonella bongori NCTC 12419 [Salmonella]
Erwinia amylovora CFBP1430 [Erwinia]
Erwinia pyrifoliae DSM 12163 [Erwinia]
Erwinia tasmaniensis Et1/99 [Erwinia]
Pantoea vagans C9-1 [Pantoea]
Cronobacter sakazakii ATCC BAA-894 [Cronobacter]
Pluralibacter gergoviae [Pluralibacter]
Plautia stali symbiont [unclassified Enterobacteriaceae]
Enterobacteriaceae bacterium strain FGI 57 [unclassified Enterobacteriaceae]
Raoultella ornithinolytica B6 [Raoultella]
Enterobacter aerogenes KCTC 2190 [Enterobacter]
Klebsiella oxytoca KCTC 1686 [Klebsiella]
Citrobacter freundii CFNIH1 [Citrobacter]
Enterobacter cloacae subsp. cloacae ATCC 13047 [Enterobacter]
Klebsiella pneumoniae subsp. pneumoniae HS11286 [Klebsiella]
Klebsiella variicola At-22 [Klebsiella]
Cedecea neteri [Cedecea]
Pantoea rwandensis [Pantoea]
Pectobacterium wasabiae WPP163 [Pectobacterium]
Pectobacterium [Pectobacterium]
Pectobacterium atrosepticum SCRI1043 [Pectobacterium]
Brenneria [Brenneria]
Pectobacterium carotovorum subsp. odoriferum [Pectobacterium]
Xenorhabdus bovienii SS-2004 [Xenorhabdus]
Xenorhabdus nematophila ATCC 19061 [Xenorhabdus]
Xenorhabdus doucetiae [Xenorhabdus]
Proteus mirabilis HI4320 [Proteus]
Photorhabdus luminescens subsp. laumondii TTO1 [Photorhabdus]
Dickeya dadantii Ech703 [Dickeya]
Dickeya zeae Ech1591 [Dickeya]
Hafnia alvei FB1 [Hafnia]
Enterobacter asburiae L1 [Enterobacter]
Serratia marcescens FGI94 [Serratia]
Edwardsiella ictaluri 93-146 [Edwardsiella]
Edwardsiella tarda 080813 [Edwardsiella]
Edwardsiella piscicida C07-087 [Edwardsiella]
Yersinia aleksiciae [Yersinia]
Yersinia kristensenii [Yersinia]
Yersinia enterocolitica subsp. enterocolitica 8081 [Yersinia]
Yersinia similis [Yersinia]
Yersinia pestis CO92 [Yersinia]
Rahnella aquatilis CIP 78.65 = ATCC 33071 [Rahnella]
Serratia multitudinisentens RB-25 [Serratia]
Serratia proteamaculans 568 [Serratia]
Serratia liquefaciens ATCC 27592 [Serratia]
Pragia fontium [Pragia]
Serratia plymuthica S13 [Serratia]
Providencia stuartii MRSN 2154 [Providencia]
Providencia sneebia DSM 19967 [Providencia]
Providencia rettgeri Dmel1 [Providencia]
Morganella morganii subsp. morganii KT [Morganella]
Enterobacter lignolyticus SCF1 [Enterobacter]
Photorhabdus asymbiotica [Photorhabdus]
Candidatus Sodalis pierantonius str. SOPE [Sodalis]
Sodalis praecaptivus [Sodalis]
Sodalis glossinidius str. ‘morsitans’ [Sodalis]
Secondary endosymbiont of Ctenarytaina eucalypti [aphid secondary symbionts]
Serratia symbiotica str. ‘Cinara cedr’ [Serratia]
Candidatus Blochmannia pennsylvanicus [ant endosymbionts]
Candidatus Blochmannia chromaiodes [ant endosymbionts]
Candidatus Blochmannia vafer [ant endosymbionts]
Candidatus Blochmannia floridanus [ant endosymbionts]
Blochmannia endosymbiont of Polyrhachis (Hedomyrma) turneri [ant endosymbionts]
Blochmannia endosymbiont of Camponotus (Colobopsis) obliquus [ant endosymbionts]
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis [Wigglesworthia]
Candidatus Westeberhardia cardiocondylae [ant endosymbionts]
Secondary endosymbiont of Heteropsylla cubana [aphid secondary symbionts]
Candidatus Ishikawaella capsulata [Candidatus Ishikawaella]
Candidatus Tachikawaea gelatinosa [Candidatus Tachikawaea]
Candidatus Riesia pediculicola [Candidatus Riesia]
Buchnera aphidicola str. APS (Acyrthosiphon pisum) [Buchnera]
67.1
96.3
97.9
96.1
47.1
75.8
89.7
88.5
87
82.8
87.7
57
100
100
56.2
76.9
87
93.5
80.4
99.9
97.9
94.2
93.4
41.7
85.3
0
83.2
100
82.2
75.3
92.7
98.5
98.4
81.9
3.8
86.5
97.2
55.3
96.8
98.7
48.4
96.6
72.7
93
66.8
95.8
87.8
99.9
100
22.9
99.6
85.3
33.6
27.4
99.4
89.3
95.2
69.8
84.7
1.6
93.8
88.4
85.5
80.8
84
91.1
96
89.4
72.7
42.5
77.5
83.8
97.3
50.5
100
38.4
96.4
40.6
62.5
82.4
50.6
91
94.7
95.6

Figure 2. Whole-genome phylogeny of 88 Enterobacteriales genomes inferred using LCBs. The maximum likelihood tree was constructed, called LCBs phylogeny, using the core LCBs shared by the 88 Enterobacteriales genomes. The LCBs phylogeny was inferred from the core LCBs using FastTree version 2.1.8[36] with default Jukes–Cantor + CAT model. Local SH-like support was assessed using the SH test with 1,000 bootstrap replicates, and the support values are given as names for the internal nodes. Species are denoted with their taxa names in the NCBI, and the corresponding genera are indicated in the square brackets. Species in the same genus are colored with the same color. Those genera with only one member in the study were colored with black.

using the PHYLIP software package (Supplementary Fig. 1 and Fig. 3). Comparing these two OGs phylogenies with the standard 16S rRNA phylogeny using Robinson–Foulds topological distance, we found that the OGs phylogeny with the second parameter setting was more similar to the 16S rRNA phylogeny (Supplementary Table 2). Furthermore, the second OGs phylogeny (Fig. 3) showed more consistency with taxonomy than the first one. Thus, we opted for the second parameter setting and the second OGs phylogeny was used in the remainder of our study. The OG information used for phylogeny inference with the second parameter setting was shown in Supplementary Table 5 (Workbooks 1 and 3). Similar to the 16S rRNA phylogeny, most of the genera in our OGs phylogeny were well grouped, except the ant endosymbionts. The seven species of ant endosymbionts were divided into three groups, which is inconsistent with the topology of the 16S rRNA phylogeny. We examined the genome sizes of the 88 genomes. Strikingly, the genome sizes of ant

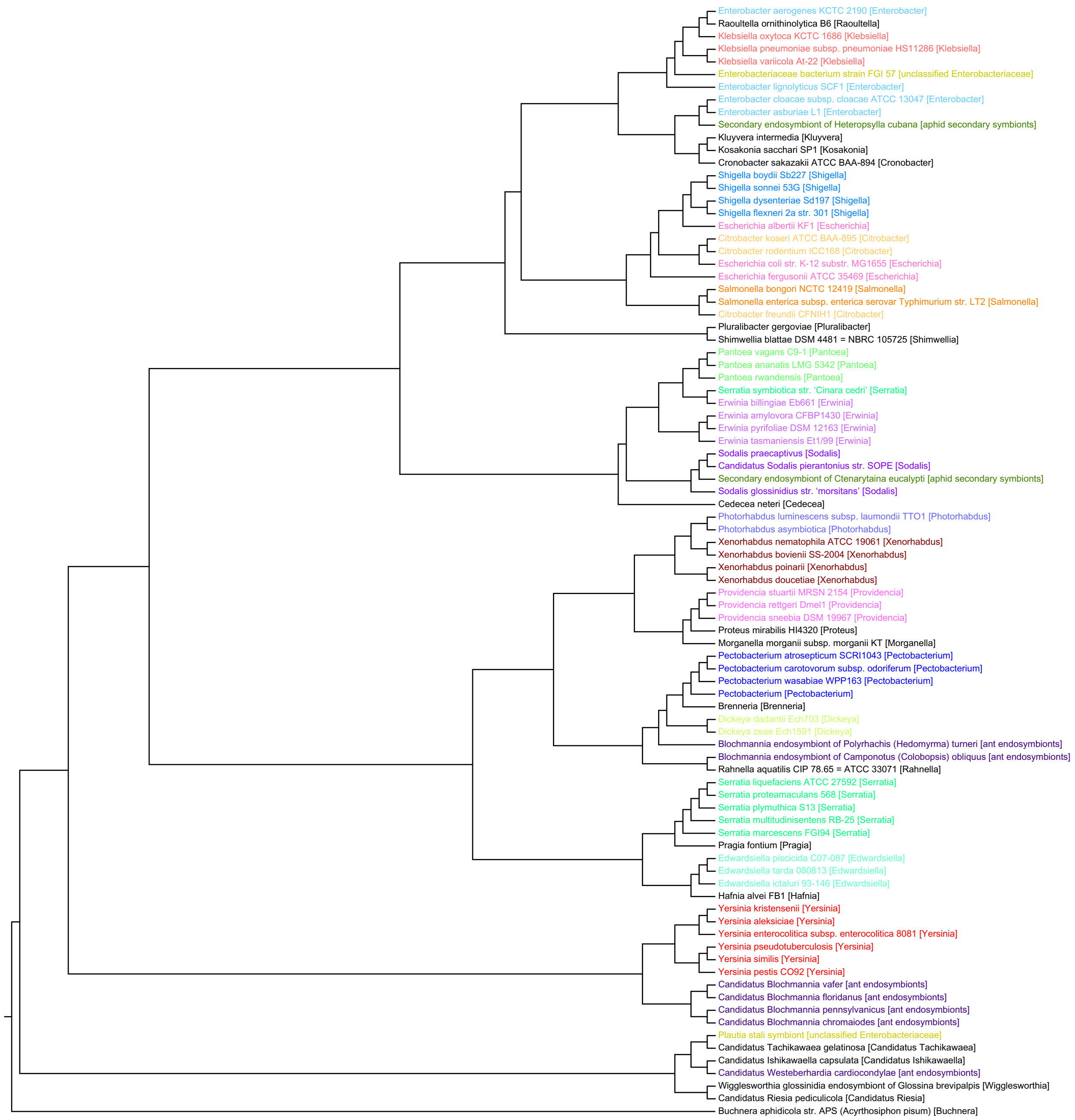
Enterobacter aerogenes KCTC 2190 [Enterobacter]
Raoultella ornithinolytica B6 [Raoultella]
Klebsiella oxytoca KCTC 1686 [Klebsiella]
Klebsiella pneumoniae subsp. pneumoniae HS11286 [Klebsiella]
Klebsiella variicola At-22 [Klebsiella]
Enterobacteriaceae bacterium strain FGI 57 [unclassified Enterobacteriaceae]
Enterobacter lignolyticus SCF1 [Enterobacter]
Enterobacter cloacae subsp. cloacae ATCC 13047 [Enterobacter]
Enterobacter asburiae L1 [Enterobacter]
Secondary endosymbiont of Heteropsylla cubana [aphid secondary symbionts]
Kluyvera intermedia [Kluyvera]
Kosakonia sacchari SP1 [Kosakonia]
Cronobacter sakazakii ATCC BAA-894 [Cronobacter]
Shigella boydii Sb227 [Shigella]
Shigella sonnei 53G [Shigella]
Shigella dysenteriae Sd197 [Shigella]
Shigella flexneri 2a str. 301 [Shigella]
Escherichia albertii KF1 [Escherichia]
Citrobacter koseri ATCC BAA-895 [Citrobacter]
Citrobacter rodentium ICC168 [Citrobacter]
Escherichia coli str. K-12 substr. MG1655 [Escherichia]
Escherichia fergusonii ATCC 35469 [Escherichia]
Salmonella bongori NCTC 12419 [Salmonella]
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 [Salmonella]
Citrobacter freundii CFNIH1 [Citrobacter]
Pluralibacter gergoviae [Pluralibacter]
Shimwellia blattae DSM 4481 = NBRC 105725 [Shimwellia]
Pantoea vagans C9-1 [Pantoea]
Pantoea ananatis LMG 5342 [Pantoea]
Pantoea rwandensis [Pantoea]
Serratia symbiotica str. ‘Cinara cedri’ [Serratia]
Erwinia billingiae Eb661 [Erwinia]
Erwinia amylovora CFBP1430 [Erwinia]
Erwinia pyrifoliae DSM 12163 [Erwinia]
Erwinia tasmaniensis Et1/99 [Erwinia]
Sodalis praecaptivus [Sodalis]
Candidatus Sodalis pierantonius str. SOPE [Sodalis]
Secondary endosymbiont of Ctenarytaina eucalypti [aphid secondary symbionts]
Sodalis glossinidius str. ‘morsitans’ [Sodalis]
Cedecea neteri [Cedecea]
Photorhabdus luminescens subsp. laumondii TTO1 [Photorhabdus]
Photorhabdus asymbiotica [Photorhabdus]
Xenorhabdus nematophila ATCC 19061 [Xenorhabdus]
Xenorhabdus bovienii SS-2004 [Xenorhabdus]
Xenorhabdus poinarii [Xenorhabdus]
Xenorhabdus doucetiae [Xenorhabdus]
Providencia stuartii MRSN 2154 [Providencia]
Providencia rettgeri Dmel1 [Providencia]
Providencia sneebia DSM 19967 [Providencia]
Proteus mirabilis HI4320 [Proteus]
Morganella morganii subsp. morganii KT [Morganella]
Pectobacterium atrosepticum SCRI1043 [Pectobacterium]
Pectobacterium carotovorum subsp. odoriferum [Pectobacterium]
Pectobacterium wasabiae WPP163 [Pectobacterium]
Pectobacterium [Pectobacterium]
Brenneria [Brenneria]
Dickeya dadantii Ech703 [Dickeya]
Dickeya zeae Ech1591 [Dickeya]
Blochmannia endosymbiont of Polyrhachis (Hedomyrma) turneri [ant endosymbionts]
Blochmannia endosymbiont of Camponotus (Colobopsis) obliquus [ant endosymbionts]
Rahnella aquatilis CIP 78.65 = ATCC 33071 [Rahnella]
Serratia liquefaciens ATCC 27592 [Serratia]
Serratia proteamaculans 568 [Serratia]
Serratia plymuthica S13 [Serratia]
Serratia multitudinisentens RB-25 [Serratia]
Serratia marcescens FGI94 [Serratia]
Pragia fontium [Pragia]
Edwardsiella piscicida C07-087 [Edwardsiella]
Edwardsiella tarda 080813 [Edwardsiella]
Edwardsiella ictaluri 93-146 [Edwardsiella]
Hafnia alvei FB1 [Hafnia]
Yersinia kristensenii [Yersinia]
Yersinia aleksiciae [Yersinia]
Yersinia enterocolitica subsp. enterocolitica 8081 [Yersinia]
Yersinia pseudotuberculosis [Yersinia]
Yersinia similis [Yersinia]
Yersinia pestis CO92 [Yersinia]
Candidatus Blochmannia vafer [ant endosymbionts]
Candidatus Blochmannia floridanus [ant endosymbionts]
Candidatus Blochmannia pennsylvanicus [ant endosymbionts]
Candidatus Blochmannia chromaiodes [ant endosymbionts]
Plautia stali symbiont [unclassified Enterobacteriaceae]
Candidatus Tachikawaea gelatinosa [Candidatus Tachikawaea]
Candidatus Ishikawaella capsulata [Candidatus Ishikawaella]
Candidatus Westeberhardia cardiocondylae [ant endosymbionts]
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis [Wigglesworthia]
Candidatus Riesia pediculicola [Candidatus Riesia]
Buchnera aphidicola str. APS (Acyrthosiphon pisum) [Buchnera]

Figure 3. Whole-genome phylogeny of 88 Enterobacteriales genomes based on OGs with the NJ method. The NJ tree was constructed, called OGs phylogeny, based on OGs for the 88 Enterobacteriales genomes. Orthologous genes were identified using the approach of BBH by setting the parameters with *e*-value $<10^{-8}$, identity $>45\%$, and coverage $>85\%$ used by Cheng et al.[27] The OGs phylogeny was constructed based on orthologous OG pairs with the NJ[38] method using the PHYLIP version 3.69.[41] Species are denoted with their taxa names in the NCBI, whose corresponding genera are indicated in the square brackets. Species in the same genus are colored with the same color. Those genera with only one member in the study were colored with black.

endosymbionts were much smaller than most other species (Supplementary Table 1). Previous studies have shown that species of ant endosymbionts have experienced substantial gene losses.[47–50] As Luo et al stated,[18] the OGs method might be less sensitive to inconsistent evolutionary events, such as a lot of reduction (gene losses). These results indicate that the genomic feature of OGs can contain useful phylogenomic signals for inferring the evolutionary histories of closely related genomes. However, inconsistent evolutionary events probably reduce the phylogenomic signals of OGs, leading to some conflicts or confusion of the evolutionary relationships. Based on the same distance matrix, UPGMA trees were also reconstructed with the UPGMA method using the PHYLIP. Similar results were also found in the UPGMA trees (Supplementary Figs. 2 and 3 and Supplementary Table 2).

As our results show, both OGs and LCBs used for the phylogeny inference of the 88 Enterobacteriales organisms revealed some merits and drawbacks. In seeking to maximize the extraction of genomic information on the whole-genome scale, we assume that the analysis of combining OGs and LCBs should reveal a comprehensive history of closely related bacterial organisms.

The OGs–LCBs phylogeny for 88 Enterobacteriales genomes. To test our hypothesis, pairwise comparison was performed between any two genomes from the 88 Enterobacteriales genomes to identify LCBs of pairwise genomes. Collinear orthologous OG pairs were identified according to the "Materials and methods" section. Combining the two types of genomic features (OGs and LCBs), we constructed the OGs–LCBs phylogeny using orthologous OGs within their collinear genomic regions. The NJ tree (Fig. 4) was inferred with the NJ method using the PHYLIP, based on the distance

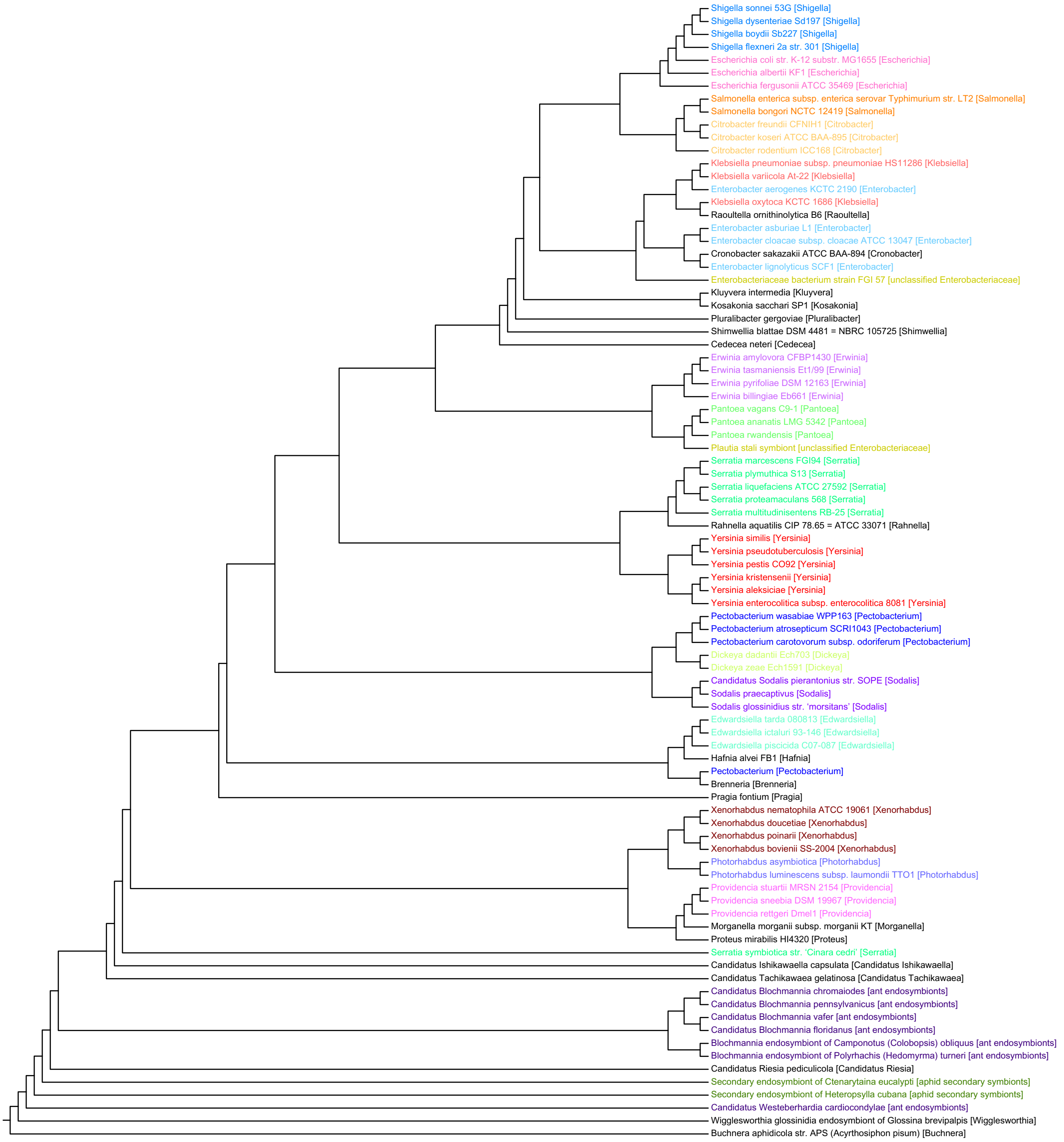
Shigella sonnei 53G [Shigella]
Shigella dysenteriae Sd197 [Shigella]
Shigella boydii Sb227 [Shigella]
Shigella flexneri 2a str. 301 [Shigella]
Escherichia coli str. K-12 substr. MG1655 [Escherichia]
Escherichia albertii KF1 [Escherichia]
Escherichia fergusonii ATCC 35469 [Escherichia]
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 [Salmonella]
Salmonella bongori NCTC 12419 [Salmonella]
Citrobacter freundii CFNIH1 [Citrobacter]
Citrobacter koseri ATCC BAA-895 [Citrobacter]
Citrobacter rodentium ICC168 [Citrobacter]
Klebsiella pneumoniae subsp. pneumoniae HS11286 [Klebsiella]
Klebsiella variicola At-22 [Klebsiella]
Enterobacter aerogenes KCTC 2190 [Enterobacter]
Klebsiella oxytoca KCTC 1686 [Klebsiella]
Raoultella ornithinolytica B6 [Raoultella]
Enterobacter asburiae L1 [Enterobacter]
Enterobacter cloacae subsp. cloacae ATCC 13047 [Enterobacter]
Cronobacter sakazakii ATCC BAA-894 [Cronobacter]
Enterobacter lignolyticus SCF1 [Enterobacter]
Enterobacteriaceae bacterium strain FGI 57 [unclassified Enterobacteriaceae]
Kluyvera intermedia [Kluyvera]
Kosakonia sacchari SP1 [Kosakonia]
Pluralibacter gergoviae [Pluralibacter]
Shimwellia blattae DSM 4481 = NBRC 105725 [Shimwellia]
Cedecea neteri [Cedecea]
Erwinia amylovora CFBP1430 [Erwinia]
Erwinia tasmaniensis Et1/99 [Erwinia]
Erwinia pyrifoliae DSM 12163 [Erwinia]
Erwinia billingiae Eb661 [Erwinia]
Pantoea vagans C9-1 [Pantoea]
Pantoea ananatis LMG 5342 [Pantoea]
Pantoea rwandensis [Pantoea]
Plautia stali symbiont [unclassified Enterobacteriaceae]
Serratia marcescens FGI94 [Serratia]
Serratia plymuthica S13 [Serratia]
Serratia liquefaciens ATCC 27592 [Serratia]
Serratia proteamaculans 568 [Serratia]
Serratia multitudinisentens RB-25 [Serratia]
Rahnella aquatilis CIP 78.65 = ATCC 33071 [Rahnella]
Yersinia similis [Yersinia]
Yersinia pseudotuberculosis [Yersinia]
Yersinia pestis CO92 [Yersinia]
Yersinia kristensenii [Yersinia]
Yersinia aleksiciae [Yersinia]
Yersinia enterocolitica subsp. enterocolitica 8081 [Yersinia]
Pectobacterium wasabiae WPP163 [Pectobacterium]
Pectobacterium atrosepticum SCRI1043 [Pectobacterium]
Pectobacterium carotovorum subsp. odoriferum [Pectobacterium]
Dickeya dadantii Ech703 [Dickeya]
Dickeya zeae Ech1591 [Dickeya]
Candidatus Sodalis pierantonius str. SOPE [Sodalis]
Sodalis praecaptivus [Sodalis]
Sodalis glossinidius str. 'morsitans' [Sodalis]
Edwardsiella tarda 080813 [Edwardsiella]
Edwardsiella ictaluri 93-146 [Edwardsiella]
Edwardsiella piscicida C07-087 [Edwardsiella]
Hafnia alvei FB1 [Hafnia]
Pectobacterium [Pectobacterium]
Brenneria [Brenneria]
Pragia fontium [Pragia]
Xenorhabdus nematophila ATCC 19061 [Xenorhabdus]
Xenorhabdus doucetiae [Xenorhabdus]
Xenorhabdus poinarii [Xenorhabdus]
Xenorhabdus bovienii SS-2004 [Xenorhabdus]
Photorhabdus asymbiotica [Photorhabdus]
Photorhabdus luminescens subsp. laumondii TTO1 [Photorhabdus]
Providencia stuartii MRSN 2154 [Providencia]
Providencia sneebia DSM 19967 [Providencia]
Providencia rettgeri Dmel1 [Providencia]
Morganella morganii subsp. morganii KT [Morganella]
Proteus mirabilis HI4320 [Proteus]
Serratia symbiotica str. 'Cinara cedri' [Serratia]
Candidatus Ishikawaella capsulata [Candidatus Ishikawaella]
Candidatus Tachikawaea gelatinosa [Candidatus Tachikawaea]
Candidatus Blochmannia chromaiodes [ant endosymbionts]
Candidatus Blochmannia pennsylvanicus [ant endosymbionts]
Candidatus Blochmannia vafer [ant endosymbionts]
Candidatus Blochmannia floridanus [ant endosymbionts]
Blochmannia endosymbiont of Camponotus (Colobopsis) obliquus [ant endosymbionts]
Blochmannia endosymbiont of Polyrhachis (Hedomyrma) turneri [ant endosymbionts]
Candidatus Riesia pediculicola [Candidatus Riesia]
Secondary endosymbiont of Ctenarytaina eucalypti [aphid secondary symbionts]
Secondary endosymbiont of Heteropsylla cubana [aphid secondary symbionts]
Candidatus Westeberhardia cardiocondylae [ant endosymbionts]
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis [Wigglesworthia]
Buchnera aphidicola str. APS (Acyrthosiphon pisum) [Buchnera]

Figure 4. Whole-genome phylogeny of 88 Enterobacteriales genomes based on both OGs and LCBs with the NJ method. The NJ tree was inferred for the 88 Enterobacteriales genomes called OGs–LCBs phylogeny, combing two types of genomic features OGs and LCBs. Orthologous genes were identified using the approach of BBH by setting the parameters with *e*-value $<10^{-8}$, identity $>5\%$, and coverage $>85\%$ used by Cheng et al.[27] Based on the orthologous OG pairs in collinear regions, the OGs–LCBs phylogeny was constructed with the NJ[38] method using PHYLIP version 3.69.[41] Species are denoted with their taxa names in the NCBI, and the corresponding genera are indicated in the square brackets. Species in the same genus are colored with the same color. Those genera with only one member in the study were colored with black.

matrix produced according to the methods (Supplementary Table 5: Workbook 2 and 3). Similarly, the UPGMA trees were also reconstructed with UPGMA method using the PHYLIP (Supplementary Fig. 4). According to the measurement of Robinson–Foulds topological distance, our OGs–LCBs phylogeny seems to be the most similar to the 16S rRNA phylogeny (Table 1). Consistent with the 16S rRNA phylogeny, almost all species in our OGs–LCBs phylogeny were clustered into groups according to their genera. In contrast to the OGs phylogeny, the species of ant endosymbionts, having experienced gene losses in their evolutionary histories,[47–50] were better grouped in our OGs–LCBs phylogeny (Fig. 4). This pattern suggests that combining both the OGs and LCBs features, to a certain extent, can reduce the effect of inconsistent evolutionary events and increase the robustness of the OGs methods for phylogenomic analysis. When the loss of genes had occurred within these genomes, smaller numbers of orthologous OGs pairs were identified, which might reduce the accuracy of their positions in the phylogenies when comparing with other genomes. However, if we use orthologous OGs pairs conditional on their collinear genomic regions rather than their whole genomes, the influence of gene loss on the accuracy of phylogeny inference may be mitigated. In addition, we also observed that *Serratia symbiotica* str. *Cinara cedri*, which belongs to *Serratia* genus, was clustered in a different group rather than the *Serrate* group in our OGs–LCBs phylogeny. Similarly, *S. symbiotica* str. *C. cedri* and other species of *Serratia* were clustered into different clades in our LCBs phylogeny, OGs phylogeny, and the 16S rRNA phylogeny. This observation may suggest that *S. symbiotica* str. *C. cedri* probably has undergone some inconsistent evolutionary events and evolved distantly with other species in the genus *Serratia*. We observed that the genome size of *S. symbiotica* str. *C. cedri* was much smaller than other *Serratia* species, indicating that genome reduction might have occurred in *S. symbiotica* str. *C. cedri*. Indeed, Lamelas et al showed that massive genomic decay had occurred in *S. symbiotica*.[51,52] Interestingly, for the clade of the genus *Escherichia* and *Shigella*, our OGs–LCBs phylogeny seems to be more in agreement with previous studies than the 16S rRNA phylogeny.[30,53,54]

There are many methods designed for the reconstruction of phylogenies among species.[30] A recent study by Facey et al introduced a comparative genomics approach to reconstruct a phylogeny of similar bacterial species, which compiled a dataset of loci shared by 9% of bacteria under study.[55] All these previous studies are generally separated into two categories: those based on the core genes or features, similar to the LCBs phylogeny, and those based on the variable gene content or other genomic features or metabolic pathways, such as the OGs phylogeny and OGs–LCBs phylogeny. In this study, we provide some evaluation of using only LCBs, only OGs, and the combination of OGs and LCBs in reconstructing phylogenies. All the results suggest that the reconstruction of phylogeny combining both OGs and LCBs should be reliable for phylogenomic analysis among closely related bacterial genomes. However, there are some limitations for our OGs–LCBs method. For example, currently, we have analyzed only the complete genomes, and further improvement may be needed to analyze the draft genomes.

Conclusions

In this work, we attempt to combine two types of genomic features (OGs and LCBs) for the phylogenomic analysis of closely related bacterial genomes. As a case study, we analyzed the phylogenetic relationship of 88 species in the order Enterobacteriales. Three different types of phylogenies were constructed based on OGs and LCBs. Our results demonstrated that by combining OGs with LCBs, more accurate phylogenetic signals were detected, which enabled us to construct more precise and robust phylogenies of the 88 genomes. The OGs method for phylogenomic analysis is usually less sensitive to some inconsistent evolutionary events, such as gene losses. Interestingly, combining OGs and LCBs as phylogenetic markers may reduce, to some extent, the influence of gene loss on phylogeny inference. In the OGs–LCBs phylogeny of 88 Enterobacteriales genomes, the species of ant endosymbionts, which have experienced substantial gene losses, were better clustered than in the OGs phylogeny. Thus, mining the phylogenetic signals of OGs, together with collinear genome regions, should be an effective approach to increasing the robustness of the OGs methods for phylogenomic analysis of closely related bacterial organisms.

Acknowledgments

The authors thank the six anonymous reviewers for their valuable comments and suggestions. The authors also thank Bing Liu and Xia Han for help in accessing the data from the NCBI and Bo-Wen Zhang for suggestions about the manuscript.

Author Contributions

Designed the study: YCZ, KL. Performed the bioinformatics analyses: YCZ. Wrote the first draft of the manuscript: YCZ, KL. Contributed to the writing of the manuscript: YCZ, KL. Agreed with manuscript results and conclusions: YCZ, KL. Jointly developed the structure and arguments for the paper: YCZ, KL. Made critical revisions and approved the final version: YCZ, KL. Both the authors reviewed and approved the final manuscript.

Supplementary Materials

Supplementary Table 1. List of 88 Enterobacteriales genomes of Gammaproteobacteria.

Supplementary Table 2. The Robinson-Foulds topological distances between different types of phylogenies.

Supplementary Table 3. The lengths and descriptions of the four core LCBs.

Supplementary Table 4. The annotation information of each genome for the four core LCBs.

Supplementary Table 5. The OG information used for inferring phylogenies of the 88 Enterobacteriales genomes.

Supplementary Figure 1. Whole-genome phylogeny of 88 Enterobacteriales genomes based on OGs with NJ method.

Supplementary Figure 2. Whole-genome phylogeny I of 88 Enterobacteriales genomes based on OGs with UPGMA method.

Supplementary Figure 3. Whole-genome phylogeny II of 88 Enterobacteriales genomes based on OGs with UPGMA method.

Supplementary Figure 4. Whole-genome phylogeny of 88 Enterobacteriales genomes based on both OGs and LCBs with UPGMA method.

Supplementary File 1. Perl scripts used to identify OGs from each genome annotation.

REFERENCES

1. Huvet M, Stumpf MPH. Overlapping genes: a window on gene evolvability. *BMC Genomics*. 2014;15:721.
2. Simon-Loriere E, Holmes EC, Pagan I. The effect of gene overlapping on the rate of RNA virus evolution. *Mol Biol Evol*. 2013;30(8):1916–28.
3. Pavesi A, Magiorkinis G, Karlin DG. Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of Deltaretroviruses. *PLoS Comput Biol*. 2013;9(8):e1003162.
4. Chirico N, Vianelli A, Belshaw R. Why genes overlap in viruses. *Proc Biol Sci*. 2010;277(1701):3809–17.
5. Cock PJ, Whitworth DE. Evolution of relative reading frame bias in unidirectional prokaryotic gene overlaps. *Mol Biol Evol*. 2010;27(4):753–6.
6. Sabath N, Graur D, Landan G. Same-strand overlapping genes in bacteria: compositional determinants of phase bias. *Biol Direct*. 2008;3:36.
7. Sakharkar KR, Chow VTK. Strategies for genome reduction in microbial genomes. *Genome Inform*. 2005;16(2):69–75.
8. Rogozin IB, Spiridonov AN, Sorokin AV, et al. Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet*. 2002;18(5):228–32.
9. Fukuda Y, Nakayama Y, Tomita M. On dynamics of overlapping genes in bacterial genomes. *Gene*. 2003;323:181–7.
10. Williams BAP, Slamovits CH, Patron NJ, et al. A high frequency of overlapping gene expression in compacted eukaryotic genomes. *Proc Natl Acad Sci U S A*. 2005;102(31):10936–41.
11. Sanna CR, Li W-H, Zhang L. Overlapping genes in the human and mouse genomes. *BMC Genomics*. 2008;9:169.
12. Makalowska I, Lin CF, Makalowski W. Overlapping genes in vertebrate genomes. *Comput Biol Chem*. 2005;29(1):1–12.
13. Normark S, Bergstrom S, Edlund T, et al. Overlapping genes. *Annu Rev Genet*. 1983;17(1):499–525.
14. Johnson ZI, Chisholm SW. Properties of overlapping genes are conserved across microbial genomes. *Genome Res*. 2004;14(11):2268–72.
15. Inokuchi Y, Hirashima A, Sekine Y, et al. Role of ribosome recycling factor (RRF) in translational coupling. *EMBO J*. 2000;19(14):3788–98.
16. Krakauer DC. Stability and evolution of overlapping genes. *Evolution*. 2000;54(3):731–9.
17. Clark MA, Baumann L, Thao MLL, et al. Degenerative minimalism in the genome of a psyllid endosymbiont. *J Bacteriol*. 2001;183(6):1853–61.
18. Luo Y, Fu C, Zhang D-Y, et al. Overlapping genes as rare genomic markers: the phylogeny of gamma-*Proteobacteria* as a case study. *Trends Genet*. 2006;22(11):593–6.
19. Luo Y, Fu C, Zhang D-Y, et al. BPhyOG: an interactive server for genome-wide inference of bacterial phylogenies based on overlapping genes. *BMC Bioinformatics*. 2007;8:266.
20. Wolf Y, Rogozin I, Grishin N, et al. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol*. 2001;1(1):8.
21. Fukuda Y, Washio T, Tomita M. Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Nucleic Acids Res*. 1999;27(8):1847–53.
22. Sakharkar KR, Sakharkar MK, Verma C, et al. Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*. *Int J Syst Evol Microbiol*. 2005;55:1205–1209.
23. Li W-H, Graur D. *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates; 1991.
24. Kim D-S, Cho C-Y, Huh J-W, et al. EVOG: a database for evolutionary analysis of overlapping genes. *Nucleic Acids Res*. 2009;37:D698–702.
25. Behura SK, Severson DW. Overlapping genes of *Aedes aegypti*: evolutionary implications from comparison with orthologs of *Anopheles gambiae* and other insects. *BMC Evol Biol*. 2013;13:124.
26. Jiang L-W, Lin K-L, Lu CL. OGtree: a tool for creating genome trees of prokaryotes based on overlapping genes. *Nucleic Acids Res*. 2008;36:W475–80.
27. Cheng C-H, Yang C-H, Chiu H-T, et al. Reconstructing genome trees of prokaryotes using overlapping genes. *BMC Bioinformatics*. 2010;11:102.
28. Lillo F, Krakauer DC. A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. *Biol Direct*. 2007;2:22.
29. Darling ACE, Mau B, Blattner FR, et al. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*. 2004;14(7):1394–403.
30. Zhang Y, Lin K. A phylogenomic analysis of *Escherichia coli/Shigella* group: implications of genomic features associated with pathogenicity and ecological adaptation. *BMC Evol Biol*. 2012;12:174.
31. Hazen TH, Sahl JW, Fraser CM, et al. Refining the pathovar paradigm via phylogenomics of the attaching and effacing *Escherichia coli*. *Proc Natl Acad Sci U S A*. 2013;110(31):12810–5.
32. Minkin I, Patel A, Kolmogorov M, Vyahhi N, Pham S. Sibelia: a scalable and comprehensive synteny block generation tool for closely related microbial genomes. In: Aaron D, Jens S, eds. *Algorithms in Bioinformatics*. Berlin: Springer; 2013:215–29.
33. Minkin I, Pham H, Starostina E, et al. C-Sibelia: an easy-to-use and highly accurate tool for bacterial genome comparison. *F1000Res*. 2013;2013(2):258–8.
34. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772–80.
35. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25(15):1972–3.
36. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010;5(3):e9490.
37. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10.
38. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4(4):406–25.
39. Nei M, Kumar S. *Molecular Evolution and Phylogenetics*. Oxford: Oxford University Press; 2000.
40. Plotree D, Plotgram D. PHYLIP-phylogeny inference package (version 3.2). *Cladistics*. 1989;5:163–6.
41. Felsenstein J. *PHYLIP: Phylogenetic Inference Program, Version 3.6*. Seattle: University of Washington; 2005.
42. Robinson D, Foulds LR. Comparison of phylogenetic trees. *Math Biosci*. 1981;53(1):131–47.
43. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 2013;29(22):2933–5.
44. Gardner PP, Daub J, Tate J, et al. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res*. 2011;39(suppl 1):D141–5.
45. Liu K, Linder CR, Warnow T. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS One*. 2011;6(11):e27731.
46. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 1981;17(6):368–76.
47. Gil R, Silva FJ, Zientz E, et al. The genome sequence of Blochmannia floridanus: comparative analysis of reduced genomes. *Proc Natl Acad Sci U S A*. 2003; 100(16):9388–93.
48. Gross R, Feldhaar H. Blochmannia in Camponotus: From the Genome of the Endosymbiont to Physiological Function for the Host. 2006.
49. Wernegreen JJ, Lazarus AB, Degnan PH. Small genome of Candidatus Blochmannia, the bacterial endosymbiont of Camponotus, implies irreversible specialization to an intracellular lifestyle. *Microbiology*. 2002;148(8):2551–6.
50. Degnan PH, Lazarus AB, Wernegreen JJ. Genome sequence of Blochmannia pennsylvanicus indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome Res*. 2005;15(8):1023–33.
51. Lamelas A, Gosalbes MJ, Manzano-Marín A, et al. *Serratia symbiotica* from the aphid Cinara cedri: a missing link from facultative to obligate insect endosymbiont. *PLoS Genet*. 2011;7(11):e1002357.
52. Burke GR, Moran NA. Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. *Genome Biol Evol*. 2011;3:195–208.
53. Sims GE, Kim S-H. Whole-genome phylogeny of *Escherichia coli*/Shigella group by feature frequency profiles (FFPs). *Proc Natl Acad Sci U S A*. 2011;108(20): 8329–34.
54. Meier-Kolthoff JP, Hahnke RL, Petersen J, et al. Complete genome sequence of DSM 30083 T, the type strain (U5/41 T) of *Escherichia coli*, and a proposal for delineating subspecies in microbial taxonomy. *Stand Genomic Sci*. 2014;9:2.
55. Facey PD, Méric G, Hitchings MD, et al. Draft genomes, phylogenetic reconstruction, and comparative genomics of two novel cohabiting bacterial symbionts isolated from *Frankliniella occidentalis*. *Genome Biol Evol*. 2015;7(8): 2188–202.