Wu *BMC Genomics* 2018, **19**(Suppl 1):921
DOI 10.1186/s12864-017-4327-9

BMC Genomics

RESEARCH Open Access

ezTree: an automated pipeline for identifying phylogenetic marker genes and inferring evolutionary relationships among uncultivated prokaryotic draft genomes

Yu-Wei Wu

From 16th International Conference on Bioinformatics (InCoB 2017)
Shenzhen, China. 20-22 September 2017

Abstract

Background: Inferring phylogenetic trees for newly recovered genomes from metagenomic samples is very useful in determining the identities of uncultivated microorganisms. Even though 16S ribosomal RNA small subunit genes have been established as "gold standard" markers for inferring phylogenetic trees, they usually cannot be assembled very well in metagenomes due to shared regions among 16S genes. Using single-copy marker genes to build genome trees has become increasingly popular for uncultivated species. Predefined marker gene sets were discovered and have been applied in various genomic studies; however these gene sets might not be adequate for novel, uncultivated, draft, or incomplete genomes. The automatic identification of marker gene sets among a set of genomes with different assembly qualities has thus become a very important task for inferring reliable phylogenetic relationships for microbial populations.

Results: A computational pipeline, ezTree, was developed to automatically identify single-copy marker genes for a group of genomes and build phylogenetic trees from the marker genes. Testing ezTree on a group of proteobacteria species revealed that ezTree was highly effective in pinpointing marker genes and constructing reliable trees for different groups of bacterial genomes. Applying ezTree to genomes that were recently recovered from metagenomes also showed that ezTree can help elucidate taxonomic relationships among newly recovered genomes and existing ones.

Conclusions: The development of ezTree can help scientists build reliable phylogenetic trees for uncultivated species retrieved from environmental samples. The uncovered single-copy marker genes may also provide crucial hints for understanding shared features of a group of microbes. The ezTree pipeline is freely available at https://github.com/yuwwu/ezTree under a GNU GPLv3 license.

Keywords: Marker gene, Phylogenetic tree, Uncultivated species

Background

Metagenomics and single-cell genomics have been established as promising methods for mining and investigating novel organisms from a wide variety of environments. The term "microbial dark matter" was proposed to describe uncultivated organisms that can only be sequenced and studied from microbial communities [1], and a new view of the tree of life was proposed to plug more than 1000 newly recovered uncultivated genomes into existing phylogenetic trees [2]. Increasing numbers of studies have focused on analyzing novel genomes extracted from a huge variety of microbial communities [3–12], thus expanding and pushing our knowledge toward understanding these organisms and the roles they play in the environments.

One of the most popular techniques for investigating microbial communities is metagenomics, which seeks to directly obtain genomic sequences from the environments. Computational binning techniques [13–22] were developed

Correspondence: yuwei.wu@tmu.edu.tw
Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, No. 250, Wuxing St., Xinyi District, 110 Taipei, Taiwan

to extract individual organisms directly from metagenomes. To understand the microbial diversity of the recovered genomes and place them in the tree of life, phylogenetic marker genes have been used to build trees for the newly identified species. 16S ribosomal RNA small-subunit genes, one of the most widely adopted phylogenetic markers, have been established as "gold standard" for probing the taxonomy of newly recovered organisms and constructing phylogenetic trees [23, 24]. However, due to shared regions of 16S rRNA genes, it is still a very challenging task for de Bruijn graph-based metagenomic assemblers, such as Meta-IDBA [25], SPAdes [26], Ray Meta [27], and MEGAHIT [28], to assemble intact 16S rRNA genes from metagenomes [29]. As a result, genomes recovered from metagenomes usually lack 16S genes (or consist of only very short gene fragments), making it impossible or very difficult to build phylogenetic trees using 16S sequences.

Whole-genome information was proposed for refining phylogenetic relationships between or among individual species [30–33]. Concatenated protein trees (trees based on combined protein data alignments) were proposed to compensate for 16S gene-based trees and are potentially more robust and informative [34]. In order to build concatenated protein trees, one needs to identify phylogenetic marker genes, defined as genes that appear once and only once in every organism considered in the study [35]. Genes satisfying this criterion have been used as markers for reliably reconstructing phylogenetic relationships for prokaryotic species, as demonstrated in previous studies [36, 37]. Various attempts have been made to discover such marker gene sets. For example, Ciccarelli et al. identified 31 marker genes in 191 bacterial species and built a highly resolved tree-of-life [36]. Different marker gene sets were also reported by other people [35, 37, 38]. The checkM software also discovered lineage-specific marker gene sets and used them to check the completeness and contamination ratios of prokaryotic genomes recovered from metagenomes [39].

Since individual genomes recovered from metagenomes are rarely complete, some of the genes from the predefined marker gene sets may be missing from the recovered genomes. Moreover, since constructing phylogenetic trees usually involves dozens or even hundreds of genomes, one may need to laboriously check the copy number of each gene in every genome in order to identify the marker gene set for building phylogenetic trees. Even though reliable gene prediction tools such as Prodigal [40] and FragGeneScan [41] were developed to alleviate efforts to predict genes from newly recovered prokaryotic genomes, tools to automatically identify marker genes in a group of genomes are still needed to infer taxonomic relationships for a set of genomes.

Herein, I introduce a computational pipeline for inferring marker genes and phylogenetic trees from a set of prokaryotic genomes. The pipeline takes a set of genomes, including newly recovered, fragmented, or incomplete ones, and is able to predict protein-coding genes from the input genomes, identify marker genes shared by all genomes, and produce concatenated protein alignments of marker genes along with a maximum-likelihood (ML) phylogenetic tree. Users with newly recovered genomes of any quality can very easily and effortlessly employ this pipeline to build a tree and infer the taxonomy of recovered species.

Methods

The pipeline was designed to take in a set of prokaryotic genomic sequences in fasta format. Genomic sequences can be complete, fragmented, or even incomplete. If users prefer, they may also input protein sequences instead of entire genomes. The workflow of the pipeline includes predicting protein-coding genes from the genomes, assigning functional profiles to the genes, identifying single-copy marker genes for the set of genomes, and aligning sequences to produce a phylogenetic tree, as depicted in Fig. 1. The implementation details of the pipeline are described below.

Gene prediction and functional annotation

Gene prediction was performed using Prodigal [40] with parameter "-p meta" to accommodate novel or newly recovered fragmented genomes. The pipeline is able to check whether inputs are genomic sequences or proteins and skip the gene-prediction step for protein inputs. After extracting protein-coding genes from the genomes, the amino acid sequences were compared to PFAM hidden Markov models [42] using HMMER3 [43] with e-value cutoff set to 1e-10 (which was chosen to achieve a balance between sensitivity and specificity, as illustrated in Fig. 2, in which the greatest number of marker genes was identified at 1e-10 and 1e-15). Only the top hit for each gene was retained in order to preserve only the most likely mapping results and facilitate the search for single-copy marker genes.

Single-copy marker gene identification

Gene annotations across all genomes were compared within and between genomes to look for marker genes. PFAM profiles that appeared more than once in each genome were discarded; the remaining profiles were further compared among all genomes. Only single-copy PFAM profiles that were found in all genomes were kept for further processing.

Sequence alignment and phylogenetic tree construction

Once single-copy marker genes were identified for the set of input genomes, amino acid sequences of the genes

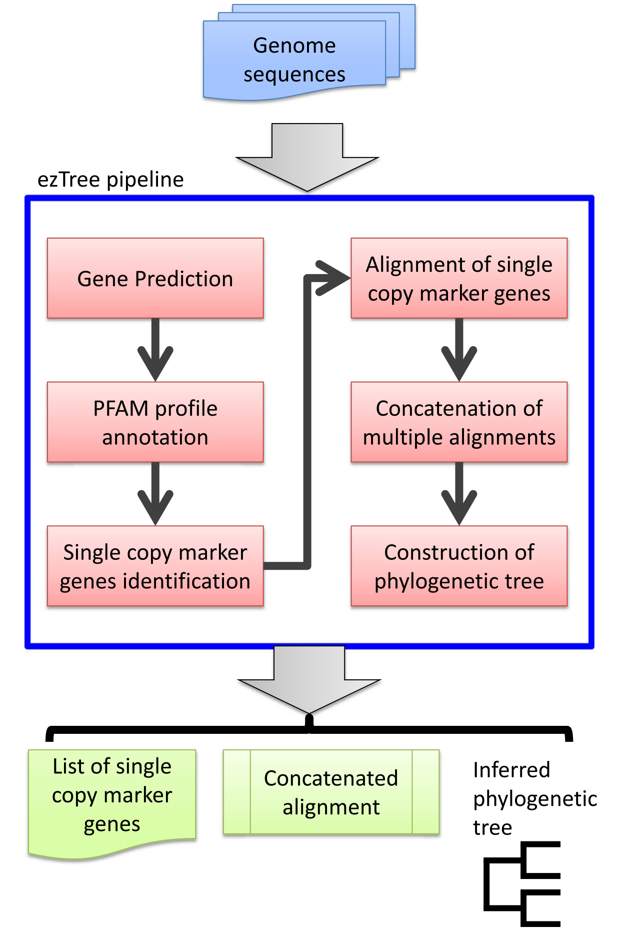
Genome sequences
ezTree pipeline
Gene Prediction
PFAM profile annotation
Single copy marker genes identification
Alignment of single copy marker genes
Concatenation of multiple alignments
Construction of phylogenetic tree
List of single copy marker genes
Concatenated alignment
Inferred phylogenetic tree

Fig. 1 Workflow of ezTree to look for single-copy marker genes and use them to construct a phylogenetic tree for a set of input genomes

were collected from all genomes and separately aligned using MUSCLE [44]. Alignments were then concatenated, one-by-one, to form a single alignment file. Gblocks [45] was further employed to remove highly variable or gapped positions in order to generate more-reliable trees. Finally FastTree [46] was used to generate an ML tree from the concatenated alignment with default options (JTT model, 1000 bootstraps).

Pipeline output

Given a set of genomes, the pipeline was designed to identify 1) a list of marker genes; 2) a concatenated alignment file; and 3) the tree in the Newick format built by FastTree. The tree can be viewed using tools such as MEGA7 [47], TreeView [48], and FigTree [49]. If users wish, they can also take the alignment and use other tree-reconstruction software such as RAxML [50], Mr. Bayes [51], MEGA7 [47], PhyML [52], and IQ-TREE [53] to produce their own trees.

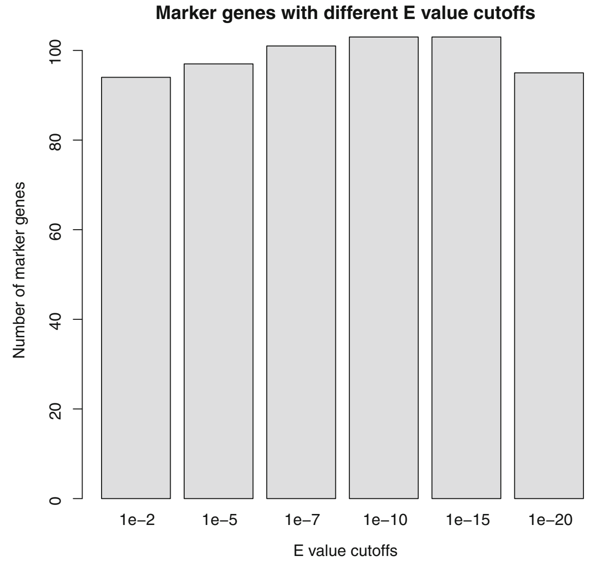

Fig. 2 Number of single-copy marker genes extracted using different e-value cutoffs. The test dataset is the proteobacteria dataset listed in Additional file 1: Table S1

Mapping PFAM profiles and cluster of orthologous groups (COG) categories

PFAM profiles and COG categories were mapped through the gene ontology (GO) website, which consists of COG-to-GO and PFAM-to-GO mapping results [54]. The mapping was done in two steps: 1) “cog2go” and “pfam2go” files were downloaded; and 2) COGs and PFAMs that could be mapped to the same GO terms were extracted. Definitions of COG categories were downloaded from the NCBI COG website [55].

Availability

The pipeline along with a README and a tutorial PDF file is publicly available at https://github.com/yuwwu/ezTree under the GNU GPLv3 license. The set of Proteobacteria genomes used in the evaluation can also be downloaded from the github website.

Results

The ezTree pipeline was first evaluated using 23 Proteobacteria genomes, among which six were draft genomes (i.e., genomes with more than one scaffold; detailed genome information is listed in Additional file 1: Table S1). ezTree successfully identified marker genes and built phylogenetic trees for genomes that shared the same species, genus, family, order, class, and phylum ranks, as shown in Fig. 3. Tree structures were consistent with known topologies, suggesting that ezTree was able to reconstruct phylogenetic relationships among the species. The high bootstrap support values for all branches of

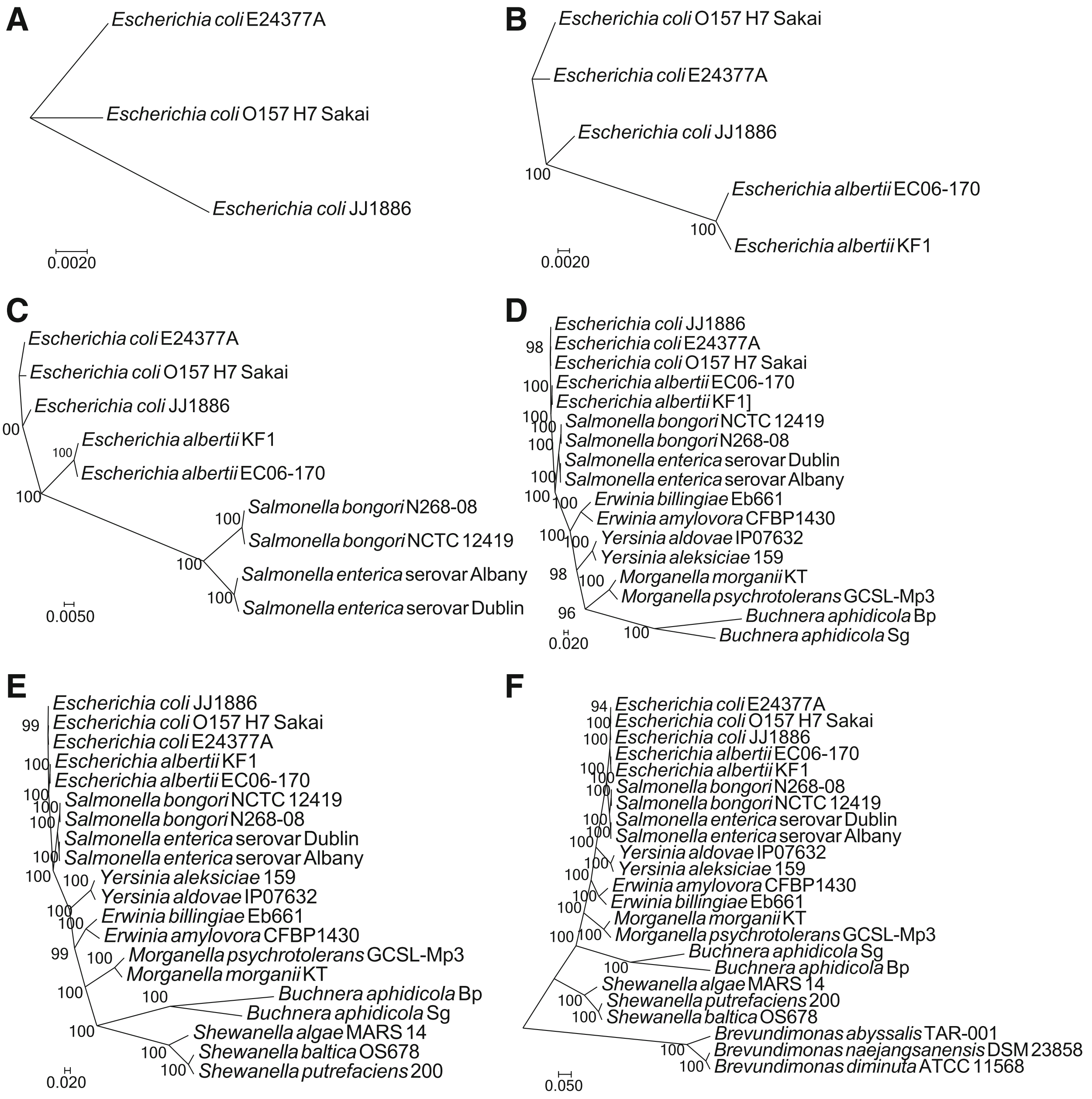
A
Escherichia coli E24377A
Escherichia coli O157 H7 Sakai
Escherichia coli JJ1886
0.0020
B
Escherichia coli O157 H7 Sakai
Escherichia coli E24377A
Escherichia coli JJ1886
100
Escherichia albertii EC06-170
100
Escherichia albertii KF1
0.0020
C
Escherichia coli E24377A
Escherichia coli O157 H7 Sakai
Escherichia coli JJ1886
100
Escherichia albertii KF1
100
Escherichia albertii EC06-170
100
Salmonella bongori N268-08
100
Salmonella bongori NCTC 12419
100
Salmonella enterica serovar Albany
100
Salmonella enterica serovar Dublin
0.0050
D
Escherichia coli JJ1886
98
Escherichia coli E24377A
Escherichia coli O157 H7 Sakai
100
Escherichia albertii EC06-170
Escherichia albertii KF1]
100
100
Salmonella bongori NCTC 12419
100
Salmonella bongori N268-08
Salmonella enterica serovar Dublin
100
Salmonella enterica serovar Albany
100
100
Erwinia billingiae Eb661
Erwinia amylovora CFBP1430
100
100
Yersinia aldovae IP07632
Yersinia aleksiciae 159
98
100
Morganella morganii KT
Morganella psychrotolerans GCSL-Mp3
96
Buchnera aphidicola Bp
100
Buchnera aphidicola Sg
0.020
E
Escherichia coli JJ1886
99
Escherichia coli O157 H7 Sakai
Escherichia coli E24377A
100
Escherichia albertii KF1
Escherichia albertii EC06-170
100
100
Salmonella bongori NCTC 12419
100
Salmonella bongori N268-08
Salmonella enterica serovar Dublin
100
Salmonella enterica serovar Albany
100
100
Yersinia aleksiciae 159
Yersinia aldovae IP07632
100
100
Erwinia billingiae Eb661
Erwinia amylovora CFBP1430
99
100
Morganella psychrotolerans GCSL-Mp3
Morganella morganii KT
100
100
Buchnera aphidicola Bp
Buchnera aphidicola Sg
100
Shewanella algae MARS 14
100
Shewanella baltica OS678
100
Shewanella putrefaciens 200
0.020
F
94
Escherichia coli E24377A
100
Escherichia coli O157 H7 Sakai
100
Escherichia coli JJ1886
Escherichia albertii EC06-170
100
100
Escherichia albertii KF1
100
Salmonella bongori N268-08
Salmonella bongori NCTC 12419
100
Salmonella enterica serovar Dublin
100
100
Salmonella enterica serovar Albany
Yersinia aldovae IP07632
100
Yersinia aleksiciae 159
100
Erwinia amylovora CFBP1430
100
Erwinia billingiae Eb661
100
Morganella morganii KT
100
100
Morganella psychrotolerans GCSL-Mp3
Buchnera aphidicola Sg
100
Buchnera aphidicola Bp
Shewanella algae MARS 14
100
Shewanella putrefaciens 200
100
Shewanella baltica OS678
Brevundimonas abyssalis TAR-001
100
Brevundimonas naejangsanensis DSM 23858
100
Brevundimonas diminuta ATCC 11568
0.050

Fig. 3 Reconstructed phylogenetic trees of a set of Proteobacteria genomes. Shared taxonomic ranks for each subset of genomes are (**a**) species; (**b**) genus; (**c**) family; (**d**) order; (**e**) class; and (**f**) phylum

the trees indicate that trees built from the identified marker genes were very reliable.

Besides building trees, ezTree was also able to identify shared single-copy marker genes from the examined genomes. Numbers of identified marker genes are shown in the upper part of Fig. 4. As expected, genomes with the same species-, genus-, or family-level taxonomy shared many more marker genes (1161, 1051, and 917 marker genes, respectively, for species, genus, and family levels) than those with the same order-, class-, or phylum-level taxonomy (167, 149, and 103 marker genes, respectively, for order, class, and phylum levels). The COG categories of marker genes were identified by mapping PFAM profiles against COGs (see Implementation for details). One of the COG categories, "[J] Translation, ribosomal structure and biogenesis," clearly stood out as the most abundant gene category for marker genes, as shown in the lower part of Fig. 4. This is consistent with other marker gene-related analyses, in which ribosomal proteins accounted for the majority of marker genes. For example, Huson et al. reported using 41 marker genes to guide gene-centric assembly of orthologous gene families, in which 30 of

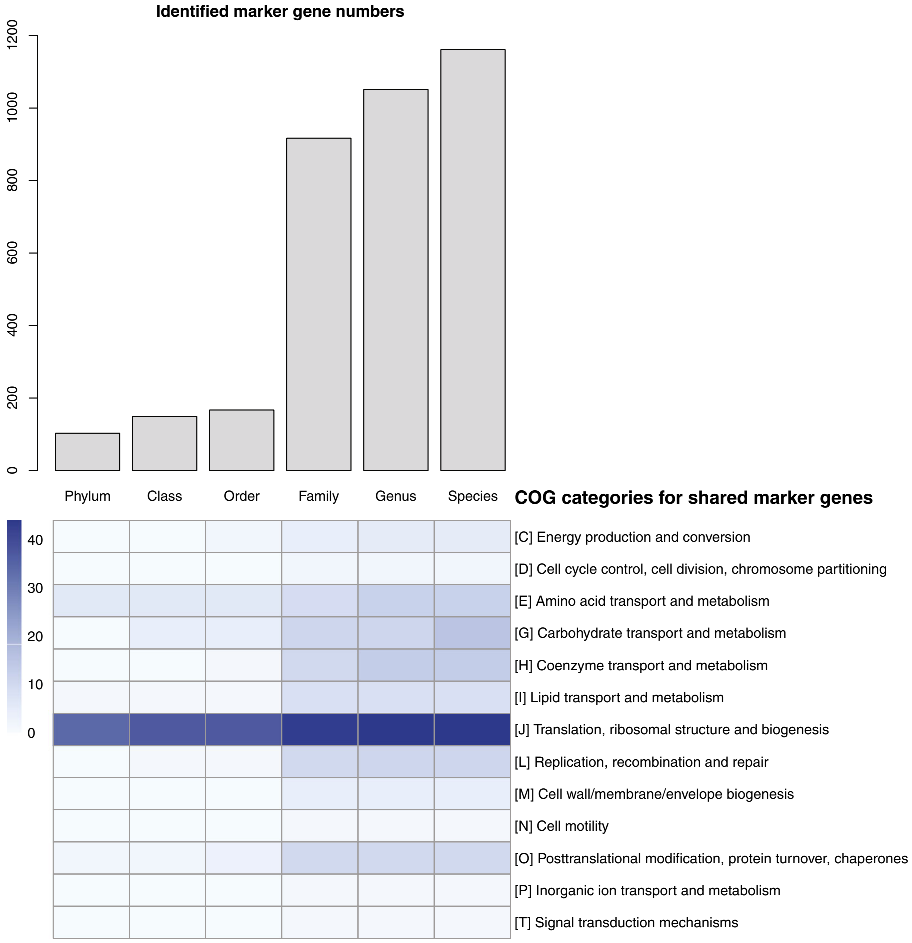

Fig. 4 Numbers and COG annotations of the marker genes predicted from Proteobacteria genomes. Upper: The number of single-copy marker genes identified for Proteobacteria genomes shown in Fig. 3. Lower: Heatmap of predicted cluster of orthologous group (COG) categories of marker genes

41 (73%) marker genes were ribosomal proteins [56]. A new tree-of-life was also built on a set of 16 ribosomal protein sequences of organisms [2]. Note that only a fraction of genes can be mapped to COG categories due to the mapping between PFAM and COG; however, the consistency between this and past works cannot be overlooked.

The ezTree pipeline was also used to identify marker genes and phylogenetic relationships for several newly recovered genomes from metagenomes. In 2016, Wawrik et al. reported that the bacterial species *Smithella* sp. SDB coupled with hydrogenotrophic methanogens could degrade water-insoluble paraffins [57]. The draft genomes of *Smithella* sp. SDB, *Methanosaeta* sp. SDB, *Methanolinea* sp. SDB, and *Methanoculleus* sp. SDB were downloaded and applied ezTree to them along with other genomes downloaded from NCBI. For *Smithella* sp. SDB, ezTree successfully identified 31 marker genes (Additional file 1: Table S3) from a group of Syntrophobacterales, and the tree (Fig. 5; genome information is available in Additional file 1: Table S2) for the involved genomes was consistent with the 16S tree (Fig. 2 of Wawrik et al.'s paper [57]). Note that among the 17 Syntrophobacterales genomes, only four were complete genomes; the numbers of scaffolds of the draft genomes ranged from as low as 22 to as high as 1037. This clearly demonstrates the ability of ezTree to identify marker genes and build trees from draft genomes of any assembly quality.

ezTree also identified 75 single-copy marker genes for the three Methanomicrobia SDB genomes (genome information can be found in Additional file 1: Table S4; marker genes are listed in Additional file 1: Table S5). The tree built from concatenated proteins also clearly placed the three SDB genomes in their corresponding places, as shown in Fig. 6. Bootstrap values were very significant for most branches, lending support to the reliability of the constructed tree. It was interesting to observe that species of *Methanolinea* and *Methanosaeta* recovered by Wawrik et al. [57] were more closely

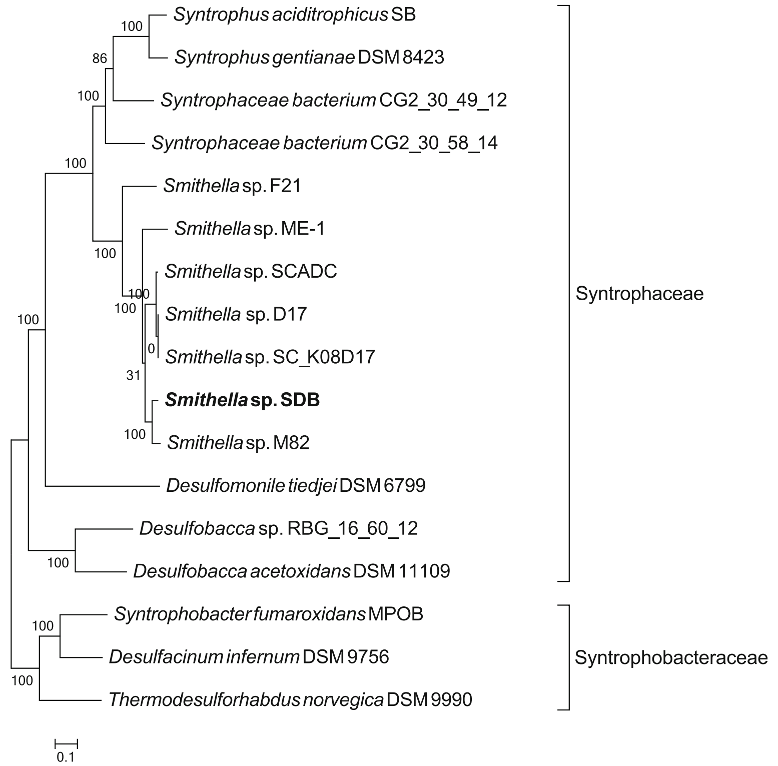
Syntrophus aciditrophicus SB
Syntrophus gentianae DSM 8423
Syntrophaceae bacterium CG2_30_49_12
Syntrophaceae bacterium CG2_30_58_14
Smithella sp. F21
Smithella sp. ME-1
Smithella sp. SCADC
Smithella sp. D17
Smithella sp. SC_K08D17
Smithella sp. SDB
Smithella sp. M82
Desulfomonile tiedjei DSM 6799
Desulfobacca sp. RBG_16_60_12
Desulfobacca acetoxidans DSM 11109
Syntrophobacter fumaroxidans MPOB
Desulfacinum infernum DSM 9756
Thermodesulforhabdus norvegica DSM 9990
Syntrophaceae
Syntrophobacteraceae
100
86
100
100
100
100
100
100
0
31
100
100
100
100
100
0.1

Fig. 5 Reconstructed phylogenetic tree of a set of Syntrophobacterales genomes. The SDB genome described in Wawrik et al. [57] is highlighted in bold for clarity

related to known species, while the *Methanoculleus* sp. SDB was more distantly related to other *Methanoculleus*, hinting that the recovered *Methanoculleus* sp. SDB species may have the potential to become a new genus under the Methanomicrobiaceae family.

Last, ezTree was applied to a Myxococcales species recovered from enriched cellulolytic microbial consortia derived from green waste compost [22]. This recovered genome was the most abundant species in one of the two microbial communities and was found in 2014 to be distantly related to *Sorangium cellulosum*, as shown in Fig. 5 of Wu et al.'s MaxBin paper [22]. Applying ezTree to a set of Myxococcales genomes yielded 56 marker genes (Additional file 1: Table S7), and the resulting tree built from the marker genes (shown in Fig. 7) indicated that the recovered Myxococcales species was more closely related to *Labilithrix luteola* DSM 27648 and *Sandaracinus amylolyticus* DSM 53668, which were deposited in NCBI on August and May 2015, respectively. Detailed information of the involved genomes can be found in Additional file 1: Table S6. In other words, with more genomes deposited in NCBI, the Myxococcales species can now be pinpointed to the Sorangiineae suborder. The tree also hinted that the Myxococcales species probably does not belong to either *Labilithrix* or *Sandaracinus* genera, as the three species formed distinct branches on the tree. More genomes are still needed to fully uncover the identity of this uncultivated species.

Discussion

A computational pipeline, ezTree, was developed to automatically infer single-copy marker genes and build reliable phylogenetic trees for a set of genomes. ezTree accepts both complete and draft genomes, including those with hundreds or even thousands of contigs or scaffolds, and is capable of automatically predicting and identifying phylogenetic marker genes. This functionality is very useful since increasing numbers of genomes are being recovered from metagenomes, and the first question we often ask is "what is it" when we are facing a new genome. ezTree thus provides an easy yet useful way to build trees and infer phylogenetic relationships with other species for newly recovered genomes.

One aspect worth noting is that ezTree needs no genome annotation information; it automatically infers annotations through the PFAM hidden Markov models. This feature relieves scientists of the burden of annotating genomes by themselves. In other words, after scientists obtain new prokaryotic genomes, they can put them—whether they are complete or are merely draft genomes—into the ezTree

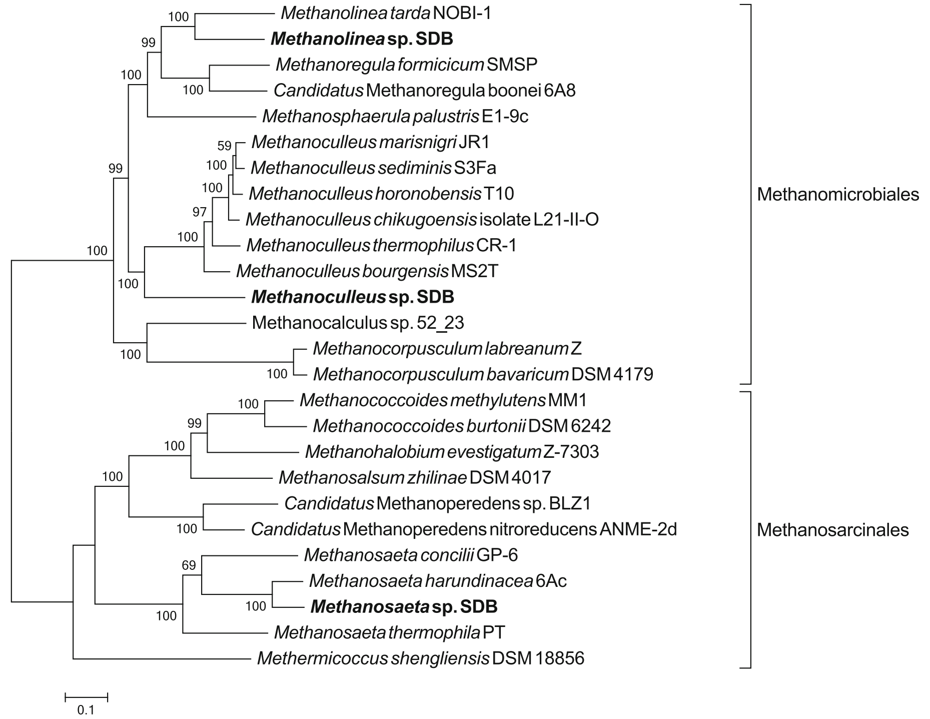
Methanolinea tarda NOBI-1
Methanolinea sp. SDB
Methanoregula formicicum SMSP
Candidatus Methanoregula boonei 6A8
Methanosphaerula palustris E1-9c
Methanoculleus marisnigri JR1
Methanoculleus sediminis S3Fa
Methanoculleus horonobensis T10
Methanoculleus chikugoensis isolate L21-II-O
Methanoculleus thermophilus CR-1
Methanoculleus bourgensis MS2T
Methanoculleus sp. SDB
Methanocalculus sp. 52_23
Methanocorpusculum labreanum Z
Methanocorpusculum bavaricum DSM 4179
Methanococcoides methylutens MM1
Methanococcoides burtonii DSM 6242
Methanohalobium evestigatum Z-7303
Methanosalsum zhilinae DSM 4017
Candidatus Methanoperedens sp. BLZ1
Candidatus Methanoperedens nitroreducens ANME-2d
Methanosaeta concilii GP-6
Methanosaeta harundinacea 6Ac
Methanosaeta sp. SDB
Methanosaeta thermophila PT
Methermicoccus shengliensis DSM 18856
Methanomicrobiales
Methanosarcinales
100
99
100
100
59
100
99
100
97
100
100
100
100
100
100
99
100
100
100
69
100
100
0.1

Fig. 6 Reconstructed phylogenetic tree of a set of Methanomicrobia genomes. Three SDB genomes identified by Wawrik et al. [57] are highlighted in bold

pipeline to infer the most likely taxonomic assignments of the novel species.

Identifying marker genes is a very important task in defining a taxonomic lineage; they can also be used to detect the completeness and contamination levels of genomes recovered from environmental samples. The ability of ezTree to identify marker genes provides scientists an easy route to investigate such information. For example, testing ezTree on collections of proteobacteria genomes and several newly identified genomes yielded highly reliable species trees, and different numbers of marker genes were also inferred by this process. These marker genes may be very important in defining distinct taxonomic ranks for a certain species, genus, family, order, class, or phylum.

Another issue related to ezTree is the selection of evolutionary models to build phylogenetic trees. Trees for the proteobacteria genomes and the Myxococcales genomes were built using different amino acid substitution models, including JTT, WAG, and LG, to test whether the selection of evolutionary models affects the tree topologies. Another option, Gamma20 model, which rescales the branch lengths and computes a Gamma20-based likelihood, was also included in the test. As shown in Additional file 1: Figure S1 and Figure S2, the trees using different amino acid substitution models are almost identical to each other, suggesting that issues related to model selection may be minor for common cases. The ezTree pipeline also provides an option to select models so that users may flexibly choose different evolutionary models or compare one model against the other.

To further validate marker gene sets, PFAM profiles of marker genes were mapped to COG categories. The greatest amount of mapped marker genes belonged to the category "[J] Translation, ribosomal structure, and biogenesis." This result is consistent with other marker gene sets discovered by other groups, in which ribosomal proteins were indispensable in marker gene sets, and lent support to the robustness of the marker gene sets identified by ezTree.

With the help of ezTree, we can now infer more-accurate taxonomic assignments for newly recovered genomes. An example can be seen in the inferred tree of Myxococcales species recovered from adapted compost microbial communities. Without the availability of *Labilithrix luteola* DSM 27648 and *Sandaracinus amylolyticus* DSM 53668, we would only know that this species is distantly related to *Sorangium cellulosum* but have no idea about its actual taxonomy. Now we can safely put it in the Sorangiineae suborder since it closely grouped together with other genomes from this taxonomic lineage. Perhaps after more genomes are extracted either from pure cultures or from environmental samples and are deposited in NCBI, we can

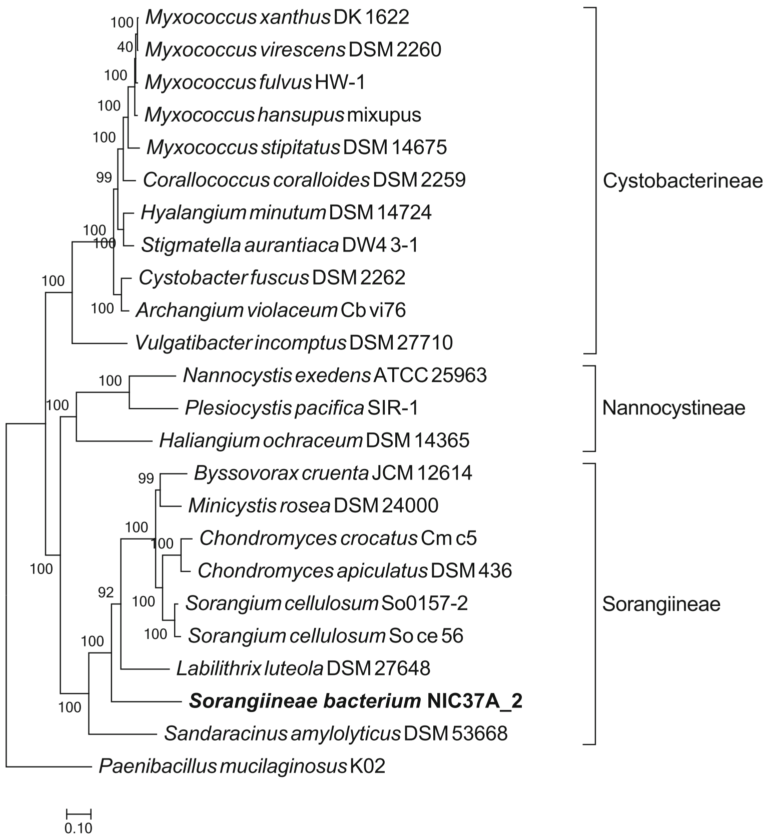
100
Myxococcus xanthus DK 1622
40
Myxococcus virescens DSM 2260
100
Myxococcus fulvus HW-1
100
Myxococcus hansupus mixupus
100
Myxococcus stipitatus DSM 14675
99
Corallococcus coralloides DSM 2259
Hyalangium minutum DSM 14724
100
100
Stigmatella aurantiaca DW4 3-1
100
Cystobacter fuscus DSM 2262
100
Archangium violaceum Cb vi76
Vulgatibacter incomptus DSM 27710
Cystobacterineae
100
Nannocystis exedens ATCC 25963
100
Plesiocystis pacifica SIR-1
Haliangium ochraceum DSM 14365
Nannocystineae
99
Byssovorax cruenta JCM 12614
Minicystis rosea DSM 24000
100
100
Chondromyces crocatus Cm c5
100
Chondromyces apiculatus DSM 436
92
100
Sorangium cellulosum So0157-2
100
Sorangium cellulosum So ce 56
100
Labilithrix luteola DSM 27648
***Sorangiineae bacterium* NIC37A_2**
100
Sandaracinus amylolyticus DSM 53668
Sorangiineae
Paenibacillus mucilaginosus K02
0.10

Fig. 7 Reconstructed phylogenetic tree of a set of Myxococcales genomes. The recovered species, *Sorangiineae bacterium* NIC37A_2, is highlighted in bold for clarity

eventually designate a more-accurate taxonomy for this and other novel species.

Conclusions

The ezTree pipeline can be used to extract marker genes and build concatenated-protein trees given a set of complete or draft genomes. Without prior knowledge except the genomic sequences, ezTree can infer single-copy marker genes for genomes and use the genes to build phylogenetic trees. Testing ezTree on multiple genome sets indicated that ezTree can be used to build highly reliable trees, providing crucial hints into defining the taxonomic lineages of the newly recovered prokaryotic genomes.

Availability and requirements

Project name: ezTree v0.1.
Project Home Page: https://github.com/yuwwu/ezTree
Operating Systems: Linux.
Programming Language: Perl.
Other requirements: None.
License: GNU GPLv3.
Any Restrictions to Use By Non-Academics: None.

Additional file

Additional file 1: Figure S1. The comparison of trees built for the set of Proteobacteria genomes provided by FastTree. **Figure S2.** The comparison of trees built for the set of Myxococcales genomes using different models provided by FastTree. **Table S1.** List of Proteobacteria genomes and their NCBI accession numbers used in the evaluation of ezTree. **Table S2.** List of Syntrophobacterales genomes and NCBI accession numbers used in inferring the tree for *Smithella* sp. SDB. **Table S3.** Single-copy marker genes identified for Syntrophobacterales genomes. **Table S4.** List of Methanomicrobia genomes and NCBI accession numbers used in inferring the tree for *Methanoculleus* sp. SDB, *Methanolinea* sp. SDB, and *Methanosaeta* sp. SDB. **Table S5.** Single-copy marker genes identified for Methanomicrobia genomes. **Table S6.** List of Myxococcales genomes and NCBI accession numbers used in inferring the tree for Sorangiineae bacterium NIC37A_2. **Table S7.** Single-copy marker genes identified for Myxococcales. (PDF 827 kb)

Abbreviations

COG: Cluster of orthologous groups; GO: Gene ontology; NCBI: National Center for Biotechnology Information; PFAM: the protein family database

Acknowledgements

I thank Lauren Tom at the Joint BioEnergy Institute, Emeryville CA, U.S.A. and Kshitij Tandon at Biodiversity Research Center, Academia Sinica, Taiwan for testing and commenting on the pipeline. I also thank the three anonymous reviewers for their valuable suggestions to improve this manuscript.

Funding
This work was supported by Taiwan Ministry of Science and Technology grant MOST106–2218-E-038-002-MY2 and Taipei Medical University grant TMU105-AE1-B19. The publication cost is covered by the Taiwan Ministry of Science and Technology grant MOST106–2218-E-038-002-MY2.

Availability of data and materials
The NCBI accession numbers of genomes mentioned in this manuscript are enumerated in Supplementary Information.

Author's contributions
YWW conceived the project, developed and tested the pipeline, and wrote the manuscript.

About this supplement
This article has been published as part of BMC Genomics Volume 19 Supplement 1, 2018: 16th International Conference on Bioinformatics (InCoB 2017): Genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-19-supplement-1.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The author declares that he has no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published: 19 January 2018

References

1. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499(7459):431–7.
2. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. A new view of the tree of life. Nat Microbiol. 2016;1(5):16048.
3. Becraft ED, Dodsworth JA, Murugapiran SK, Ohlsson JI, Briggs BR, Kanbar J, De Vlaminck I, Quake SR, Dong HL, Hedlund BP, et al. Single-cell-genomics-facilitated read binning of candidate phylum EM19 genomes from geothermal spring metagenomes. Appl Environ Microb. 2016;82(4):992–1003.
4. Denef VJ, Mueller RS, Chiang EN, Liebig JR, Vanderploeg HA. *Chloroflexi* CL500–11 Populations That Predominate Deep-Lake Hypolimnion Bacterioplankton Rely on Nitrogen-Rich Dissolved Organic Matter Metabolism and C-1 Compound Oxidation. Appl Environ Microb. 2016;82(5):1423–32.
5. Evans PN, Parks DH, Chadwick GL, Robbins SJ, Orphan VJ, Golding SD, Tyson GW. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. Science. 2015;350(6259):434–8.
6. Garcia SL, Buck M, McMahon KD, Grossart HP, Eiler A, Warnecke F. Auxotrophy and intrapopulation complementary in the "interactome" of a cultivated freshwater model community. Mol Ecol. 2015;24(17):4449–59.
7. Hiras J, YW W, Eichorst SA, Simmons BA, Singer SW. Refining the phylum Chlorobi by resolving the phylogeny and metabolic potential of the representative of a deeply branching, uncultivated lineage. Isme J. 2016; 10(4):833–45.
8. Hultman J, Waldrop MP, Mackelprang R, David MM, McFarland J, Blazewicz SJ, Harden J, Turetsky MR, McGuire AD, Shah MB, et al. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature. 2015;521(7551):208–12.
9. Nelson WC, Maezato Y, YW W, Romine MF, Lindemann SR. Identification and resolution of microdiversity through metagenomic sequencing of parallel consortia. Appl Environ Microb. 2016;82(1):255–67.
10. Nobu MK, Narihiro T, Kuroda K, Mei R, Liu WT. Chasing the elusive Euryarchaeota class WSA2: genomes reveal a uniquely fastidious methylreducing methanogen. Isme J. 2016;10(10):2478–87.
11. Pinto AJ, Marcus DN, Ijaz UZ, Santos QMBD, Dick GJ, Raskin L. Metagenomic evidence for the presence of Comammox Nitrospira-like bacteria in a drinking water system. Msphere. 2016;1(1):e00054–15.
12. Vavourakis CD, Ghai R, Rodriguez-Valera F, Sorokin DY, Tringe SG, Hugenholtz P, Muyzer G. Metagenomic insights into the uncultured diversity and physiology of microbes in four hypersaline soda Lake brines. Front Microbiol. 2016;7:211.
13. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31(6):533–8.
14. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144–6.
15. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. Peerj. 2014;2:e603.
16. Kang DWD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. Peerj. 2015;3:e1165.
17. Laczny CC, Sternal T, Plugaru V, Gawron P, Atashpendar A, Margossian HH, Coronado S, van der Maaten L, Vlassis N, Wilmes P. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome. 2015;3:1.
18. Lin HH, Liao YC. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. Sci Rep-Uk. 2016;6:24175.
19. YY L, Chen T, Fuhrman JA, Sun FZ. COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. Bioinformatics. 2017;33(6):791–8.
20. Strous M, Kraft B, Bisdorf R, Tegetmeyer HE. The binning of metagenomic contigs for microbial physiology of mixed cultures. Front Microbiol. 2012;3:410.
21. YW W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32(4):605–7.
22. YW W, Tang YH, Tringe SG, Simmons BA, Singer SW. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome. 2014;2:26.
23. Rossello-Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev. 2001;25(1):39–67.
24. Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kampfer P, Maiden MCJ, Nesme X, Rossello-Mora R, Swings J, Truper HG, et al. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Micr. 2002;52:1043–7.
25. Peng Y, Leung HCM, Yiu SM, Chin FYL. Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics. 2011;27(13):I94–I101.
26. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
27. Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. Ray meta: scalable de novo metagenome assembly and profiling. Genome Biol. 2012;13(12):R122.
28. Li DH, Liu CM, Luo RB, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674–6.
29. Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol. 2011;12(5):R44.
30. Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Micr. 2014;64:346–51.
31. Meier-Kolthoff JP, Auch AF, Klenk HP, Goker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. Bmc Bioinformatics. 2013;14:60.
32. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. PNAS. 2009;106(45):19126–31.
33. DY W, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, et al. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. Nature. 2009;462(7276):1056–60.
34. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. Universal trees based on large combined protein sequence data sets. Nat Genet. 2001; 28(3):281–5.

35. DY W, Jospin G, Eisen JA. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PLoS One. 2013;8(10):e77033.
36. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science. 2006; 311(5765):1283–7.
37. Dupont CL, Rusch DB, Yooseph S, Lombardo MJ, Richter RA, Valas R, Novotny M, Yee-Greenbaum J, Selengut JD, Haft DH, et al. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. Isme J. 2012;6(6):1186–99.
38. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P. Prediction of effective genome size in metagenomic samples. Genome Biol. 2007;8(1)
39. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7):1043–55.
40. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. Bmc Bioinformatics. 2010;11:119.
41. Rho MN, Tang HX, Ye YZ. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38(20):e191.
42. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1): D279–85.
43. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23(1):205–11.
44. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
45. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17(4):540–52.
46. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26(7): 1641–50.
47. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33(7):1870–4.
48. Page RDM. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci. 1996;12(4):357–8.
49. FigTree [http://tree.bio.ed.ac.uk/software/figtree/] Last access date: May 10, 2017.
50. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
51. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17(8):754–5.
52. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307–21.
53. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
54. Gene oncology download mappings [http://www.geneontology.org/page/download-mappings] Last access date: April 3, 2017.
55. NCBI COGs [https://www.ncbi.nlm.nih.gov/COG/index.html] Last access date: March 13, 2017.
56. Huson DH, Tappu R, Bazinet AL, Xie C, Cummings MP, Nieselt K, Williams R. Fast and simple protein-alignment-guided assembly of orthologous gene families from microbiome sequencing reads. Microbiome. 2017;5:11.
57. Wawrik B, Marks CR, Davidova IA, McInerney MJ, Pruitt S, Duncan KE, Suflita JM, Callaghan AV. Methanogenic paraffin degradation proceeds via alkane addition to fumarate by '*Smithella*' spp. mediated by a syntrophic coupling with hydrogenotrophic methanogens. Environ Microbiol. 2016;18(8):2604–19.