Journal of Microbiology (2018) Vol. 56, No. 4, pp. 280–285
DOI 10.1007/s12275-018-8014-6
eISSN 1976-3794
pISSN 1225-8873

UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction[§]

Seong-In Na[1,2], Yeong Ouk Kim[1,2], Seok-Hwan Yoon[4], Sung-min Ha[3,4], Inwoo Baek[2,3], and Jongsik Chun[1,2,3,4]*

[1]*Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 00826, Republic of Korea*
[2]*Institute of Molecular Biology & Genetics, Seoul National University, Seoul 00826, Republic of Korea*
[3]*School of Biological Sciences, Seoul National University, Seoul 00826, Republic of Korea*
[4]*ChunLab, Inc., Seoul 06725, Republic of Korea*

(Received Jan 8, 2018 / Revised Jan 27, 2018 / Accepted Jan 28, 2018)

Genome-based phylogeny plays a central role in the future taxonomy and phylogenetics of *Bacteria* and *Archaea* by replacing 16S rRNA gene phylogeny. The concatenated core gene alignments are frequently used for such a purpose. The bacterial core genes are defined as single-copy, homologous genes that are present in most of the known bacterial species. There have been several studies describing such a gene set, but the number of species considered was rather small. Here we present the up-to-date bacterial core gene set, named UBCG, and software suites to accommodate necessary steps to generate and evaluate phylogenetic trees. The method was successfully used to infer phylogenomic relationship of *Escherichia* and related taxa and can be used for the set of genomes at any taxonomic ranks of *Bacteria*. The UBCG pipeline and file viewer are freely available at https://www.ezbiocloud.net/tools/ubcg and https://www.ezbiocloud.net/tools/ubcg_viewer, respectively.

Keywords: phylogeny, phylogenetic analysis, phylogenomics, bacterial core gene

Introduction

Advancement of DNA sequencing technologies allows microbiologists to employ genome-based methods routinely in various disciplines (Radford *et al.*, 2012; Tagini and Greub, 2017). Among those, taxonomy is a science that heavily relies upon molecular phylogeny. A single gene, notably 16S rRNA gene (16S), has been widely used for the taxonomy of prokaryotes and served as the general framework (Rosselló-Mora and Amann, 2001). However, 16S is well known for its limited phylogenetic resolution that hampers the utility at the species or subspecies level (Fox *et al.*, 1992). Recently, the use of genome sequences is recommended for taxonomic purposes instead of conventional DNA-DNA hybridization and 16S rRNA phylogeny (Chun and Rainey, 2014; Chun *et al.*, 2018).

Genome-based phylogeny, also called phylogenomics, is inferred using a set of core genes rather than a single gene (Eisen and Fraser, 2003). The core genes are defined as single-copy, homologous that are universally present in the target group which can be any taxonomic ranks. The core gene sets for the domain *Bacteria* have been proposed in several times, which varied due to the availability of genome sequences at the time of analysis. Creevey *et al.* (2011) identified 40 bacterial core genes when they considered 191 species. In the latter studies, 37 genes were identified from 666 genomes (note that these were not from 666 species; Wu *et al.*, 2013), and 107 genes were suggested using the Comprehensive Microbial Resource genome database at the time of analysis (Dupont *et al.*, 2012). More recently, two phylogenomic software tools, namely Phylosift (Darling *et al.*, 2014) and bcgTree (Ankenbrand and Keller, 2016), used these gene sets respectively. Because the numbers of genomes and species used in the previous studies are rather limited, there is an urgent need to update the bacterial core gene set using the up-to-date version of public genome databases. Here, we identified the up-to-date bacterial core gene (UBCG) set from the complete genome sequences representing 1,429 species. Also, user-friendly bioinformatic tools for inferring phylogenomic trees using this gene set were provided.

Materials and Methods

Identification of bacterial core gene set

The UBCG set was identified using the complete genome sequences available from the EzBioCloud database (https://www.ezbiocloud.net/; Yoon *et al.*, 2017). To normalize the bias in the number of complete genome sequences among species, we chose single complete genome per a species (1,429 species representing 28 phyla).

A candidate set of bacterial core genes was compiled from the previous studies (Creevey *et al.*, 2011; Dupont *et al.*, 2012; Darling *et al.*, 2014). In addition, we carried out clustering of protein sequences from representative genomes that were chosen for each family using the UCLUST (Edgar, 2010) with 50% identity and 50% query_cov parameters. A total of 34 gene clusters existing in more than 50% of the genomes were identified and included to our candidate gene set for the further analysis. This process ensures our candidate gene set

*For correspondence. E-mail: jchun@snu.ac.kr; Tel.: +82-2-880-8153
[§]Supplemental material for this article may be found at http://www.springerlink.com/content/120956

represents all potentially core genes for the domain *Bacteria*. The final set contains 133 genes (Supplementary data Table S1).

The hidden Markov model (HMM) profiles (Supplementary data Table S1) for each candidate genes were downloaded from either Pfam (Finn *et al.*, 2016) or TIGRFAMs (Haft *et al.*, 2013) databases. The 1,429 complete genome sequences representing 1,429 species were screened for the candidate gene set using the hmmsearch program in the HMMER package (Eddy, 2011) with trusted-cutoff values recommended by the corresponding databases. The genes that are present as single-copy in at least 95% of 1,429 species were selected as UBCGs.

Software implementation

We developed a phylogenomics pipeline using JAVA programming language and external bioinformatics software tools (Fig. 1). The first step is to extract UBCGs using Prodigal (for gene-finding; Hyatt *et al.*, 2010) and hmmsearch (for identification of the genes using HMM; Eddy, 2011) from a whole genome assembly. The HMM profiles and cutoff values are the same as those described in the previous section. Our software tool saves the resulting UBCG sequences (both DNA and protein) in a JSON format file that can be used for the next step. In the second step, a set of JSON files containing UBCG sequences and metadata of the genome assemblies are selected for multiple alignments of each gene using MAFFT (Katoh and Standley, 2013). Lastly, the phylogenetic trees are inferred for each gene as well as a concatenated sequence of the 92 UBCGs. The phylogenetic tree generated from a concatenated alignment is named a UBCG tree. FastTree (Price *et al.*, 2010) and RAxML (Stamatakis, 2014) can be run under the UBCG pipeline automatically. Other programs for phylogenetic treeing, such as MEGA (http://www.megasoftware.net/), can also be used for tree reconstruction with the FASTA-formatted alignment files generated by the pipeline.

We assumed that a UBCG tree is one representing the true evolutionary history of whole genomes. However, it may be different from those inferred by the individual gene trees. Therefore, we devised a method to estimate the robustness of each branch in a UBCG tree using individual gene trees. If a bipartition in the UBCG tree is also present in a given single gene tree, this gene is considered to support that branch. The number of single gene trees supporting a branch in a UBCG tree is calculated and designated the Gene Support Index (GSI); the GSI value of 92 means that the branch is supported by all UBCGs. The higher the GSI is, the more robustly the branch is supported. If a gene is not present in some genomes resulting in a partial gene tree, only the existing leaves are considered. When the number of genomes is large, there is more chance in that gene trees do not support the branches in a UBCG tree. Therefore, we designed our pipeline to accept the threshold value that is used to decide the portion of genomes support the branch in a UBCG tree. The default is 95%, meaning that a gene tree supports the given branch in UBCG tree if 95% of genomes agree.

Inference of phylogenies for the *Escherichia* and related taxa using UBCG

We tested the UBCG pipeline to infer the phylogenomic relationship among *Escherichia* and related taxa including the genera *Citrobacter*, *Klebsiella*, *Salmonella*, and *Shigella*. A total of 29 genomes and corresponding 16S sequences were retrieved from the EzBioCloud database (Supplementary data Table S2). The 16S sequences were aligned using the EzEditor2 software with secondary structure information (https://www.ezbiocloud.net/tools/ezeditor2; Jeon *et al.*, 2014). The UBCG trees were generated using both nucleotide and amino acid sequences. The individual UBCGs were aligned, concatenated, and the alignment positions that had gap characters more than 50% were excluded. The final nucleotide and protein alignments were used to infer the phylogenetic trees with the GTR + CAT (for nucleotide) and the JTT + CAT (for protein) models, respectively. All phylogenetic trees were built using RAxML tool.

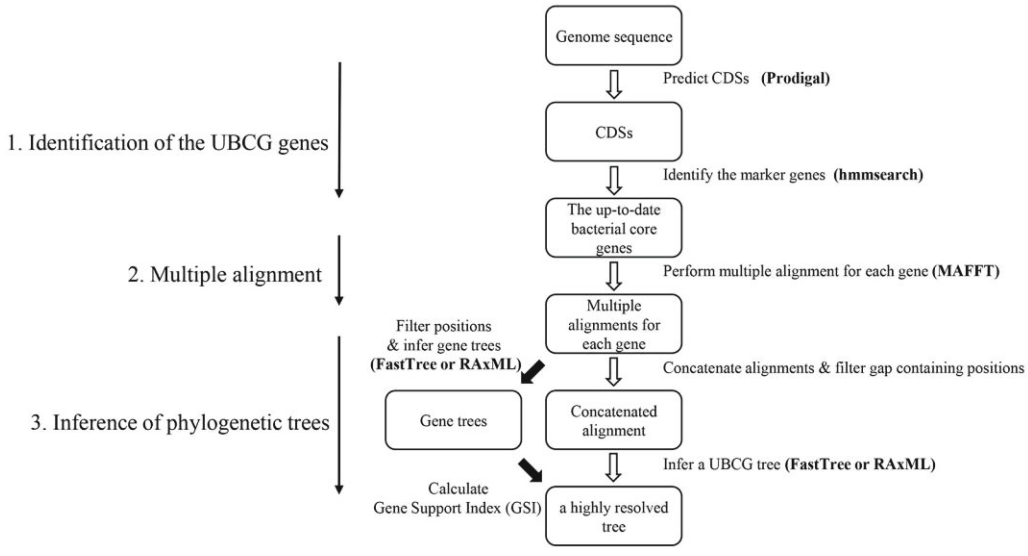

Fig. 1. The process of phylogenetic tree reconstruction using the UBCG pipeline. Each of UBCG genes is aligned separately before being concatenated into a single alignment. The pipeline generates 92 gene trees and one UBCG tree that is labeled with Gene Support Index (GSI) values. Externally executed software tools are indicated in the parentheses.

Table 1. General information of the UBCGs (Up-to-date Bacterial Core Genes)

Gene	Functional category (COG)*		HMM profile	Function
alaS	J	COG0013	TIGR00344	Alanine-tRNA ligase
argS	J	COG0018	TIGR00456	Arginine-tRNA ligase
aspS	J	COG0173	TIGR00459	Aspartate-tRNA ligase
cgtA	DL	COG0536	TIGR02729	GTPase ObgE/CgtA
coaE	H	COG0237	TIGR00152	Dephospho-CoA kinase
cysS	J	COG0215	TIGR00435	Cysteine-tRNA ligase
dnaA	L	COG0593	TIGR00362	Chromosomal replication initiator protein DnaA
dnaG	L	COG0358	TIGR01391	DNA primase
dnaX	L	COG2812	TIGR02397	DNA polymerase III subunit gamma
engA	R	COG1160	TIGR03594	GTPase Der
ffh	U	COG0541	TIGR00959	Signal recognition particle protein
fmt	J	COG0223	TIGR00460	Methionyl-tRNA formyltransferase
frr	J	COG0233	TIGR00496	Ribosome-recycling factor
ftsY	U	COG0552	TIGR00064	Signal recognition particle receptor FtsY
gmk	F	COG0194	TIGR03263	Guanylate kinase
hisS	J	COG0124	TIGR00442	Histidine-tRNA ligase
ileS	J	COG0060	TIGR00392	Isoleucine-tRNA ligase 1
infB	J	COG0532	TIGR00487	Translation initiation factor IF-2
infC	J	COG0290	TIGR00168	Translation initiation factor IF-3
ksgA	J	COG0030	TIGR00755	Ribosomal RNA small subunit methyltransferase A
lepA	J	COG0481	TIGR01393	Elongation factor 4
leuS	J	COG0495	TIGR00396	Leucine-tRNA ligase
ligA	L	COG0272	TIGR00575	DNA ligase
nusA	K	COG0195	TIGR01953	Transcription termination/antitermination protein NusA
nusG	K	COG0250	TIGR00922	Transcription termination/antitermination protein NusG
pgk	G	COG0126	PF00162	Phosphoglycerate kinase
pheS	J	COG0016	TIGR00468	Phenylalanine-tRNA ligase alpha subunit
pheT	J	COG0073	TIGR00472	Phenylalanine-tRNA ligase beta subunit
prfA	J	COG0216	TIGR00019	Peptide chain release factor 1
pyrG	F	COG0504	TIGR00337	CTP synthase
recA	L	COG0468	TIGR02012	DNA recombination and repair protein
rbfA	J	COG0858	TIGR00082	30S ribosome-binding factor
rnc	K	COG0571	TIGR02191	Ribonuclease 3
rplA	J	COG0081	TIGR01169	50S ribosomal protein L1
rplB	J	COG0090	TIGR01171	50S ribosomal protein L2
rplC	J	COG0087	TIGR03625	50S ribosomal protein L3
rplD	J	COG0088	TIGR03953	50S ribosomal protein L4
rplE	J	COG0094	PF00281	50S ribosomal protein L5
rplF	J	COG0097	TIGR03654	50S ribosomal protein L6
rplI	J	COG0359	TIGR00158	50S ribosomal protein L9
rplJ	J	COG0244	PF00466	50S ribosomal protein L10
rplK	J	COG0080	TIGR01632	50S ribosomal protein L11
rplL	J	COG0222	TIGR00855	50S ribosomal protein L7/L12
rplM	J	COG0102	TIGR01066	50S ribosomal protein L13
rplN	J	COG0093	TIGR01067	50S ribosomal protein L14
rplO	J	COG0200	TIGR01071	50S ribosomal protein L15
rplP	J	COG0197	TIGR01164	50S ribosomal protein L16
rplQ	J	COG0203	TIGR00059	50S ribosomal protein L17
rplR	J	COG0256	TIGR00060	50S ribosomal protein L18
rplS	J	COG0335	TIGR01024	50S ribosomal protein L19
rplT	J	COG0292	TIGR01032	50S ribosomal protein L20
rplU	J	COG0261	TIGR00061	50S ribosomal protein L21
rplV	J	COG0091	TIGR01044	50S ribosomal protein L22
rplW	J	COG0089	PF00276	50S ribosomal protein L23
rplX	J	COG0198	TIGR01079	50S ribosomal protein L24
rpmA	J	COG0211	TIGR00062	50S ribosomal protein L27

Table 1. Continued

Gene	Functional category (COG)*		HMM profile	Function
rpmC	J	COG0255	TIGR00012	50S ribosomal protein L29
rpmI	J	COG0291	TIGR00001	50S ribosomal protein L35
rpoA	K	COG0202	TIGR02027	DNA-directed RNA polymerase subunit alpha
rpoB	K	COG0085	TIGR02013	DNA-directed RNA polymerase subunit beta
rpoC	K	COG0086	TIGR02386	DNA-directed RNA polymerase subunit beta'
rpsB	J	COG0052	TIGR01011	30S ribosomal protein S2
rpsC	J	COG0092	TIGR01009	30S ribosomal protein S3
rpsD	J	COG0522	TIGR01017	30S ribosomal protein S4
rpsE	J	COG0098	TIGR01021	30S ribosomal protein S5
rpsF	J	COG0360	TIGR00166	30S ribosomal protein S6
rpsG	J	COG0049	TIGR01029	30S ribosomal protein S7
rpsH	J	COG0096	PF00410	30S ribosomal protein S8
rpsI	J	COG0103	PF00380	30S ribosomal protein S9
rpsJ	J	COG0051	TIGR01049	30S ribosomal protein S10
rpsK	J	COG0100	TIGR03632	30S ribosomal protein S11
rpsL	J	COG0048	TIGR00981	30S ribosomal protein S12
rpsM	J	COG0099	TIGR03631	30S ribosomal protein S13
rpsO	J	COG0184	TIGR00952	30S ribosomal protein S15
rpsP	J	COG0228	TIGR00002	30S ribosomal protein S16
rpsQ	J	COG0186	TIGR03635	30S ribosomal protein S17
rpsR	J	COG0238	TIGR00165	30S ribosomal protein S18
rpsS	J	COG0185	TIGR01050	30S ribosomal protein S19
rpsT	J	COG0268	TIGR00029	30S ribosomal protein S20
secA	U	COG0653	TIGR00963	Protein translocase subunit SecA
secG	U	COG1314	TIGR00810	Protein-export membrane protein SecG
secY	U	COG0201	TIGR00967	Protein translocase subunit SecY
serS	J	COG0172	TIGR00414	Serine-tRNA ligase
smpB	O	COG0691	TIGR00086	SsrA-binding protein
tig	O	COG0544	TIGR00115	Trigger factor
tilS	J	COG0037	TIGR02432	tRNA(Ile)-lysidine synthase
truB	J	COG0130	TIGR00431	tRNA pseudouridine synthase B
tsaD	J	COG0533	TIGR03723	tRNA N6-adenosine threonylcarbamoyltransferase
tsf	J	COG0264	TIGR00116	Elongation factor Ts
uvrB	L	COG0556	TIGR00631	UvrABC system protein B
ybeY	J	COG0319	TIGR00043	Endoribonuclease YbeY
ychF	J	COG0012	TIGR00092	Ribosome-binding ATPase YchF

*COG, clusters of orthologous group; D, cell cycle control and mitosis; F, nucleotide metabolism and transport; G, carbohydrate transport and metabolism; H, coenzyme metabolism; J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination and repair; O, post-translational modification, protein turnover, and chaperones; R, general function prediction only; U, intracellular trafficking, secretion, and vesicular transport.

Language and software availability

The UBCG pipeline was written in Java language and is run under Linux and Mac OS X. It can be run under Windows on a Linux virtual machine. The executable files and manual are available at https://www.ezbiocloud.net/tools/ubcg. A webpage that can be used to visualize and access the sequences of the extracted UBCGs from a JSON file is written in JavaScript and available at https://www.ezbiocloud.net/tools/ubcg_viewer.

Results and Discussion

The bacterial core genes are generally defined as the genes that are present in most of the bacterial species (Wu *et al.*, 2009; Rinke *et al.*, 2013; Shih *et al.*, 2013). It is well known that our efforts for genome sequencing are heavily skewed towards pathogenic species. To reduce this sampling bias, we selected a single complete genome for each species. The resulting reference data set consisted of 1,429 genome sequences covering 28 phyla; the taxonomic coverage used in this study is the largest to date. A total of 133 candidate genes compiled from the previous studies and our *de novo* clustering were used to screen this genome dataset. Using HMM-based search, 92 genes were found to exist as a single-copy in more than 95% of the complete genome sequences considered. Therefore, the final UBCG set consists of 92 genes covering 10 functional categories (Table 1). Out of 92, 67 UBCGs belong to the COG J category (translation, ribosomal structure, and biogenesis).

To test the utility of our method, we applied the UBCG pipeline to the set of genomes belonging to *Escherichia coli*

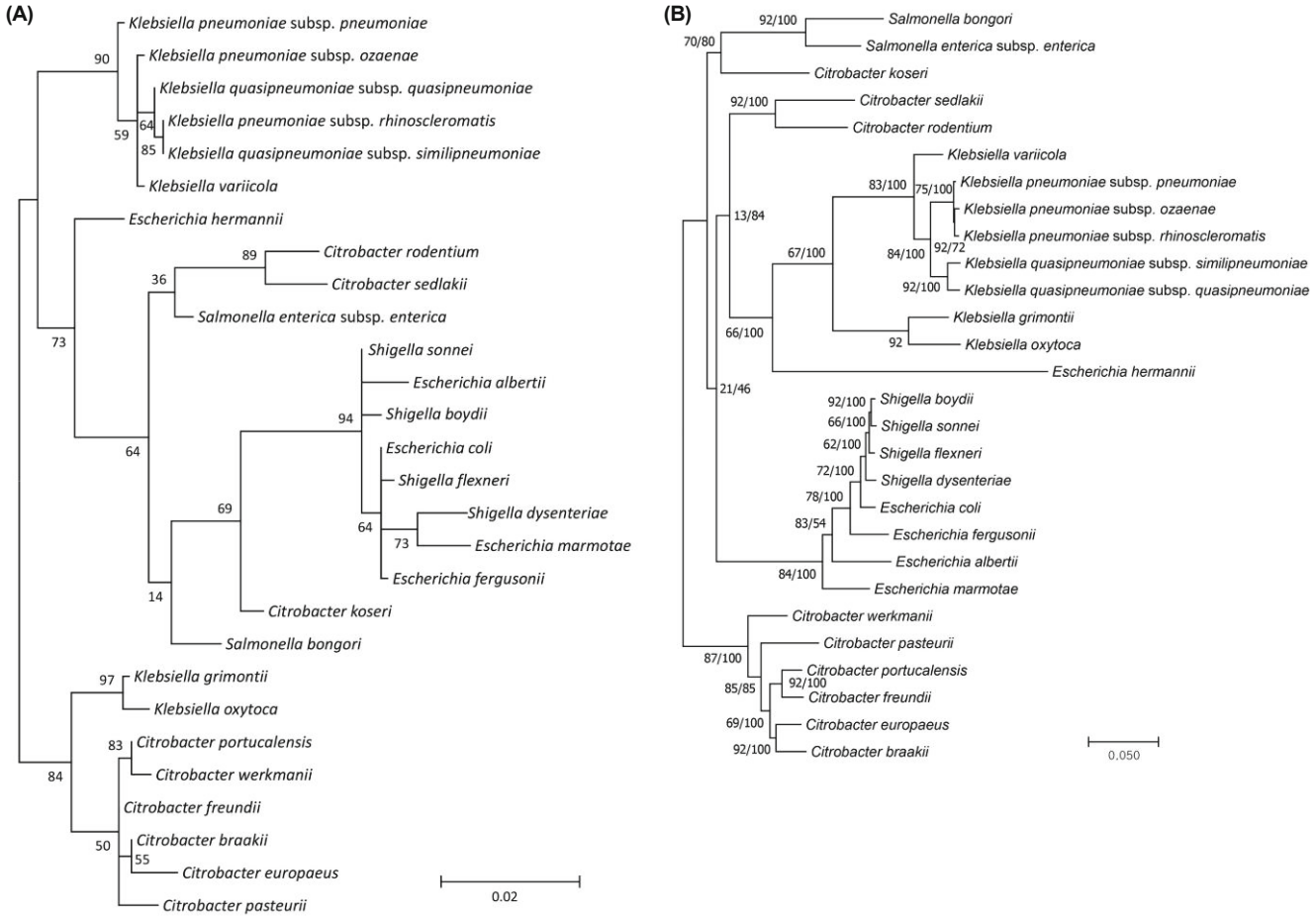

Fig. 2. Phylogenetic trees of *Escherichia coli* and related taxa. Unrooted maximum likelihood phylogenetic trees were inferred using RAxML ver. 8.2.11 using GTR + CAT model. Bootstrap analysis was carried out using 100 replications. (A) Phylogenetic tree inferred using 16S sequences. A total of 1,406 nucleotide positions were used. Percentage bootstrap values are given at branching points. Bar, 0.02 substitution per position. (B) Phylogenetic tree inferred using UBCGs (concatenated alignment of 92 core genes). A total of 88,911 nucleotide positions were used. Gene support indices (GSIs) and percentage bootstrap values are given at branching points. Bar, 0.05 substitution per position.

and related taxa. This group shows high level or 16S similarities while genome sequences are readily available for type strains. Maximum likelihood phylogenetic trees inferred using 16S and UBCG sequences were generated and compared (Fig. 2). Differences in tree topologies were evident in two phylogenetic trees. In the 16S tree (Fig. 2A), the members of the genera *Escherichi*a and *Shigella* were not differentiated whereas our UBCG method (Fig. 2B) clearly separated *E. coli*/*Shigella* spp. from the other species (*Escherichia albertii*, *E. fergusonii*, *E. marmotae*). It is noteworthy that *E. coli* and *Shigella* spp. belong to the same genomic species on the basis of high average nucleotide identities (> 95%). The clade containing *Escherichia/Shigella* was supported by 100% bootstrap and 83 GSI supports. The latter means that 83 out of 92 UBCGs supported this concatenated the clade, implying that possible events of lateral gene transfer occurred in nine genes. For example, in the *sec*Y gene-phylogenetic tree (Supplementary data Fig. S2), *E. albertii* and *E. marmotae* were not included in the *Escherichi/Shigella* clade, contradicting the UBCG tree topology (Fig. 2B).

Eight species of the genus *Klebsiella* were recovered as a monophyletic clade only in the UBCG tree, but not in the 16S tree (Fig. 2). Moreover, the closely related group containing *Klebsiella pneumoniae*, *K. quasipneumoniae*, and *K. variicola* was not differentiated in the 16S tree whereas our UBCG tree evidently confirmed at least their current species-level classification. The utility of our method for the subspecies-level analysis requires the further investigation. We also inferred phylogenetic trees from concatenated amino acid sequences of UBCGs (Supplementary data Fig. S1) whose tree topology was very similar to the corresponding nucleotide sequence-derived tree.

The GSI values indicate the reliability of branches in the genome-based phylogenetic trees, complementing other statistical measures such as the bootstrap (Felsenstein, 1985). The UBCG pipeline automatically generates the maximum likelihood trees with GSI values, making this method readily available for the users who are not skillful in bioinformatics. Additionally, trees and multiple sequence alignments are provided for all 92 core genes, which can then be used for gene-based phylogenetic analysis. The nature of GSI at the various levels of taxonomic ranks is a subject for future study.

Here, using an example set, we showed that our method provides better resolution than 16S gene at the species and genus levels. Because it is based on the domain-level core genes, the UBCG pipeline can be applied to any taxonomic ranks.

Recently, the phylogenomic treeing approach using core gene set has been proposed as a minimal standard in describing new genus or higher taxa for the domain *Bacteria* (Chun *et al.*, 2018). In this study, we introduce a new phylogenomic method that is universally applicable to any phyla of the domain *Bacteria*. The significance of the branches in the resulting phylogenomic tree is readily evaluated by the number of supporting single-gene trees. The UBCG set and accompanying bioinformatic pipelines should provide accurate and easy-to-use means of generating phylogenomic trees for not only taxonomic purposes but also other microbiological disciplines.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Acknowledgements

This research was supported, in part, by the National Research Foundation of Korea (Grants NRF-2015R1A2A2A01008404 and NRF-2014M3C9A3063541).

References

Ankenbrand, M.J. and Keller, A. 2016. bcgTree: automatized phylogenetic tree building from bacterial core genomes. *Genome* **59**, 783–791.

Chun, J. and Rainey, F.A. 2014. Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea*. *Int. J. Syst. Evol. Microbiol.* **64**, 316–324.

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D.R., da Costa, M.S., Rooney, A.P., Yi, H., Xu, X.W., De Meyer, S., *et al.* 2018. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **68**, 461–466.

Creevey, C.J., Doerks, T., Fitzpatrick, D.A., Raes, J., and Bork, P. 2011. Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PLoS One* **6**, e22099.

Darling, A.E., Jospin, G., Lowe, E., Matsen, F.I., Bik, H.M., and Eisen, J.A. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243.

Dupont, C.L., Rusch, D.B., Yooseph, S., Lombardo, M.J., Richter, R.A., Valas, R., Novotny, M., Yee-Greenbaum, J., Selengut, J.D., Haft, D.H., *et al.* 2012. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6**, 1186–1199.

Eddy, S.R. 2011. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195.

Edgar, R.C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461.

Eisen, J.A. and Fraser, C.M. 2003. Phylogenomics: intersection of evolution and genomics. *Science* **300**, 1706–1707.

Felsenstein, J. 1985. Confidence-limits on phylogenies – an approach using the bootstrap. *Evolution* **39**, 783–791.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., *et al.* 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285.

Fox, G.E., Wisotzkey, J.D., and Jurtshuk, P.J. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int. J. Syst. Bacteriol.* **42**, 166–170.

Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K., and Beck, E. 2013. TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.* **41**, D387–D395.

Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119.

Jeon, Y.S., Lee, K., Park, S.C., Kim, B.S., Cho, Y.J., Ha, S.M., and Chun, J. 2014. EzEditor: a versatile sequence alignment editor for both rRNA- and protein-coding genes. *Int. J. Syst. Evol. Microbiol.* **64**, 689–691.

Katoh, K. and Standley, D.M. 2013. MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780.

Price, M.N., Dehal, P.S., and Arkin, A.P. 2010. FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490.

Radford, A.D., Chapman, D., Dixon, L., Chantrey, J., Darby, A.C., and Hall, N. 2012. Application of next-generation sequencing technologies in virology. *J. Gen. Virol.* **93**, 1853–1868.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., *et al.* 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437.

Rosselló-Mora, R. and Amann, R. 2001. The species concept for prokaryotes. *FEMS Microbiol. Rev.* **25**, 39–67.

Shih, P.M., Wu, D.Y., Latifi, A., Axen, S.D., Fewer, D.P., Talla, E., Calteau, A., Cai, F., de Marsac, N.T., Rippka, R., *et al.* 2013. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc. Natl. Acad. Sci. USA* **110**, 1053–1058.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

Tagini, F. and Greub, G. 2017. Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 2007–2020.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., Kunin, V., Goodwin, L., Wu, M., Tindall, B.J., *et al.* 2009. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature* **462**, 1056–1060.

Wu, D.Y., Jospin, G., and Eisen, J.A. 2013. 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* **8**, e77033.

Yoon, S.H., Ha, S.M., Kwon, S., Lim, J., Kim, Y., Seo, H., and Chun, J. 2017. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.* **67**, 1613–1617.