Journal of Microbiology (2021) Vol. 59, No. 6, pp. 609–615
DOI 10.1007/s12275-021-1231-4
eISSN 1976-3794
pISSN 1225-8873

UBCG2: Up-to-date bacterial core genes and pipeline for phylogenomic analysis[§]

Jihyeon Kim[1,2†], Seong-In Na[1†], Dongwook Kim[1,2], and Jongsik Chun[1,2,3*]

[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*

(Received Apr 27, 2021 / Revised May 11, 2021 / Accepted May 11, 2021)

Phylogenomic tree reconstruction has recently become a routine and critical task to elucidate the evolutionary relationships among bacterial species. The most widely used method utilizes the concatenated core genes, universally present in a single-copy throughout the bacterial domain. In our previous study, a bioinformatics pipeline termed Up-to-date Bacterial Core Genes (UBCG) was developed with a set of bacterial core genes selected from 1,429 species covering 28 phyla. In this study, we revised a new bacterial core gene set, named UBCG2, that was selected from the more extensive genome sequence set belonging to 3,508 species spanning 43 phyla. UBCG2 comprises 81 genes with nine Clusters of Orthologous Groups of proteins (COGs) functional categories. The new gene set and complete pipeline are available at http://leb.snu.ac.kr/ubcg2.

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

Introduction

Phylogenomics has become an important routine task to infer evolutionary relationships among bacterial species (Chun and Rainey, 2014; Chun *et al.*, 2018; Na *et al.*, 2018). The most commonly adopted method uses concatenated core gene sequences (Wu and Scott, 2012; Darling *et al.*, 2014; Glaeser and Kämpfer, 2015; Parks *et al.*, 2017, 2018; Chun *et al.*, 2018; Zhu *et al.*, 2019; Asnicar *et al.*, 2020). This approach can infer a stable phylogeny with a higher resolution than the use of ribosomal RNA or a few protein-coding genes.

The genes selected for the core gene set vary according to taxonomic scope, from domain level to species level (Chun *et al.*, 2009; Na *et al.*, 2018; Lee, 2019). The genes suitable for phylogenomic analysis should be universally present as a single-copy in a target taxon. Domain level core gene sets have advantages over lower taxon-specific gene sets. They provide consistent and reproducible phylogenetic analysis across all species, as well as any taxonomic ranks, within that domain. One limitation of using domain level is that such gene sets can vary depending on the availability of complete genome sequences in public databases.

Our previously released software tool, Up-to-date Bacterial Core Genes (UBCG) (Na *et al.*, 2018), collected core genes used in several studies and screened only single-copy genes existing in most bacteria. This core gene set and accompanying software tool have been widely utilized for phylogenomic studies, especially for the classification of new bacterial taxa. This study aimed to update the bacterial core gene set utilizing significantly more bacterial genomes (3,508 species) compared with the previous version (1,429 species). The newly identified core gene set, named UBCG version 2 (UBCG2), comprises 81 bacterial core genes and a revised bioinformatic pipeline for building phylogenomic trees from genome assemblies. The software tools and manual are available for download at http://leb.snu.ac.kr/ubcg2.

Materials and Methods

Updating the bacterial core gene set

Firstly, we compiled potential single-copy core genes from previous studies, including our own set (UBCG) (Dupont *et al.*, 2012; Ankenbrand and Keller, 2016; Parks *et al.*, 2017; Na *et al.*, 2018). The resultant genes consisted of 148 candidate genes.

We then evaluated the presence of the 148 candidate core genes in the selected sequences using hidden Markov model (HMM) profiles obtained from the TIGRFAMs 15.0 (Selengut *et al.*, 2007) and Pfam 31.0 (El-Gebali *et al.*, 2019) databases. The genome sequences that were labeled as 'complete' in the NCBI database had been downloaded from the EzBioCloud database (Yoon *et al.*, 2017). A total of 3,508 sequences representing 3,508 different species were used to evaluate if a candidate gene is present as a single-copy and ubiquitous among the complete genomes. Coding sequences (CDSs) for each genome were predicted using Prodigal V2.6.3 (Hyatt *et al.*, 2010). The hmmscan program (HMMER 3.1b2; http://hmmer.org/) with a trusted cutoff (TC) option was used to detect the presence of the candidate genes. Genes with only one copy number in more than 95% of the complete genomes were selected for the UBCG2 gene set.

[†]These authors contributed equally to this work.
*For correspondence. E-mail: jchun@snu.ac.kr
[§]Supplemental material for this article may be found at http://www.springerlink.com/content/120956.

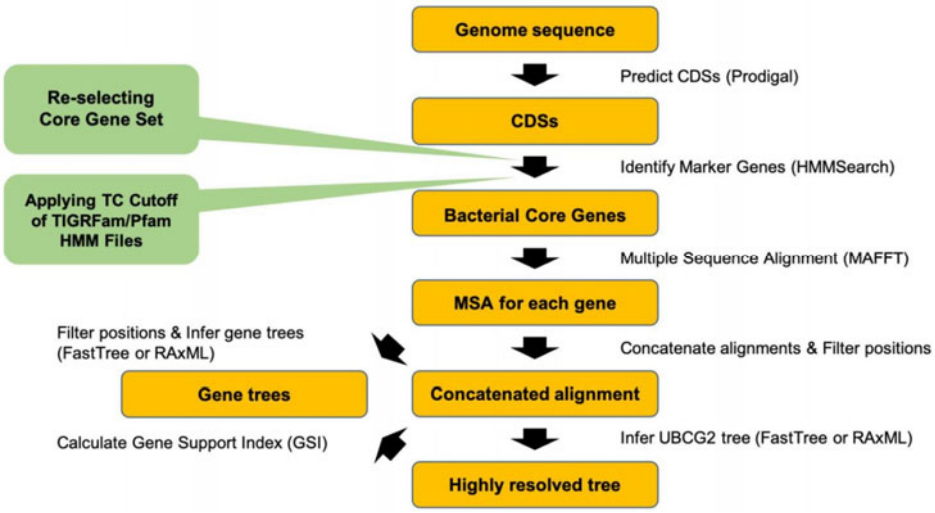

Fig. 1. The process of phylogenetic tree reconstruction using the UBCG2 pipeline. The pipeline generates 81 gene trees and concatenated UBCG2 tree that is labeled with gene support index (GSI) values when using UBCG2 gene set.

Software features

The UBCG2 phylogenomic pipeline was coded using Java language version 8 and is available at http://leb.snu.ac.kr/ubcg2. The overall workflow was identical to that of the previously released UBCG pipeline (Na *et al.*, 2018; Fig. 1). The pipeline infers a phylogenomic tree from a set of genomic sequences or CDSs by implementing external programs including Prodigal (Hyatt *et al.*, 2010), Hmmsearch (http://hmmer.org), Mafft (Katoh and Standley, 2013), RaxML (Stamatakis, 2014), and FastTree (Price *et al.*, 2010).

UBCG2 extracts the bacterial core genes and performs a multiple sequence alignment for each gene. Users can choose the alignment option out of the followings: (1) align nucleotide sequences, (2) align amino acid sequences, (3) align amino acid sequences, but use the nucleotide sequences, (4) align amino acid sequences, but use the first and second nucleotides in each codon (codon 12 option). Then, the pipeline removes gap-rich columns (by default, more than 50% of gap characters) in each alignment column and concatenates them into an extensive sequence alignment. UBCG2 infers the phylogenomic tree from this final alignment by applying RaxML or FastTree.

Results and Discussion

To identify the bacterial core genes among 3,508 bacterial species covering 43 phyla, we calculated the presence ratio (PR) and single-copy ratio (SR) of each candidate gene using the hmmscan program with the trusted cutoff (Table 1). We chose the trusted cutoffs, which vary for each gene instead of the fixed cutoff. For comparative purposes, when we also employed 10e-5 as a fixed cutoff for all genes in HMM-based search, the PR values of most genes increased, whereas the SR decreased (Table 1).

A core gene is defined as a gene with both PR and SR of 95% or higher with the trusted cutoffs. This stringent criterion resulted in 81 bacterial core genes, 11 fewer than the previous UBCG (92 genes; Na *et al.*, 2018).

Table 1 provides detailed information about 148 bacterial core gene candidates compiled from previous studies, including bcgTree (Dupont *et al.*, 2012; Ankenbrand and Keller, 2016), UBCG (Na *et al.*, 2018), and bac120 (Parks *et al.*, 2017). UBCG2 has 64 and 77 genes in common with bac120 and bcgTree, respectively. Of the genes used in bac120 and bcgTree sets, 56 and 30 genes were not included in UBCG2 gene set, respectively, as they did not meet the 95% SR criterion. In particular, *recG* and *rpsA* included in bac120 had an SR of 69.38% and 46.21%, respectively, and *proS*, *rpmH*, and *glyS* included in bcgTree had an SR of 78.76%, 75.03%, and 63.57%, respectively. The main reason for these discrepancies in the core gene sets between our study and previous studies is the combination of the use of trusted cutoff in HMM-based search and the larger number of reference genomes employed.

Fifteen genes included in UBCG version 1 were omitted in this version as they showed slightly lower SR values (93.16–94.93%) than the 95% cutoff (Supplementary data Table S1). Instead, four genes, namely, *trmD*, *era*, *ruvB*, and *rsmH*, were newly added to the UBCG2 gene set as they met our stringent criterion when 3,508 species were considered.

It is vital to detect target orthologs with appropriate cutoff criteria in order to identify single-copy core genes (Selengut *et al.*, 2007). If an applied cutoff is too loose, paralogous genes may be mistakenly identified as the correct sequence. Alternatively, if a cutoff is too strict, the corresponding gene sequence may not be detected, even though the gene is present in the genome sequence. In our study, we observed that PR and SR were significantly affected by the adopted cutoff criteria. Therefore, we employed the trusted cutoff defined by the curators of TIGRFAM and PFAM instead of the fixed cutoff (e.g., 10e-5) to ensure that only orthologs can be detected, which allows us to identify single-copy genes with more confidence.

In this study, we only used the complete genome sequences for calculating the bacterial core gene set as draft genome assemblies are often contaminated (Parks *et al.*, 2015; Lee *et al.*, 2017). To ensure a normalized representation for a broader taxonomic scope, each reference genome belongs to a separate species, which was validated by Average Nucleotide Identity-based identification (Ha *et al.*, 2019). As a result, the num-

Table 1. General information about bacterial core genes

A total of 148 genes collected from UBCG1, bac120, and bcgTree are shown. Their hidden Markov model (HMM) profiles, functions, presence, and single-copy ratio are listed. The ratios are derived from 3,508 complete genome sequences. There are two values for the ratios; when trusted cutoff is used and when 10E-5 (the expectation value; E-value) is used for the hmmscan. Whether each gene is included in the core gene sets is marked by O and X.

	Gene	HMM*	Length	COG†	COG category‡	Function	Presence ratio (%) Trusted cutoff is used (10E-5 E-value is used)	Single-copy ratio (%) Trusted cutoff is used (10E-5 E-value is used)	UBCG2	UBCG1	bac120 Parks *et al.* (2017)	bcgTree Dupont *et al.* (2012), Ankenbrand and Keller (2016)
1	*rplE*	PF00281.18	57	COG0094	J	50S ribosomal protein L5	99.91 (99.91)	99.71 (99.09)	○	○	×	○
2	*rpsH*	PF00410.18	127	COG0096	J	30S ribosomal protein S8	99.86 (99.86)	99.66 (97.63)	○	○	○	○
3	*rplB*	TIGR01171	275	COG0090	J	50S ribosomal protein L2	99.77 (99.94)	99.63 (99.03)	○	○	○	○
4	*rpsB*	TIGR01011	225	COG0052	J	30S ribosomal protein S2	99.83 (99.97)	99.49 (96.44)	○	○	○	○
5	*rpsI*	PF00380.18	120	COG0103	J	30S ribosomal protein S9	99.8 (99.8)	99.46 (98.03)	○	○	○	○
6	*rplC*	TIGR03625	202	COG0087	J	50S ribosomal protein L3	99.46 (99.97)	99.34 (98.92)	○	○	○	○
7	*rplN*	TIGR01067	122	COG0093	J	50S ribosomal protein L14	99.46 (99.63)	99.32 (98.77)	○	○	×	○
8	*rpsL*	TIGR00981	124	COG0048	J	30S ribosomal protein S12	99.4 (99.43)	99.17 (98.8)	○	○	×	○
9	*rplW*	PF00276.19	86	COG0089	J	50S ribosomal protein L23	99.32 (99.37)	99.14 (98.35)	○	○	×	○
10	*rpsP*	TIGR00002	78	COG0228	J	30S ribosomal protein S16	99.32 (99.63)	99.14 (99)	○	○	×	○
11	*rplJ*	PF00466.19	100	COG0244	J	50S ribosomal protein L10	99.26 (99.26)	99.09 (97.55)	○	○	○	○
12	*rpsJ*	TIGR01049	99	COG0051	J	30S ribosomal protein S10	99.29 (99.86)	99.09 (99.54)	○	○	×	○
13	*rpsC*	TIGR01009	212	COG0092	J	30S ribosomal protein S3	99.09 (99.91)	98.97 (93.9)	○	○	○	○
14	*rpsO*	TIGR00952	86	COG0184	J	30S ribosomal protein S15	99.2 (99.69)	98.95 (98.32)	○	○	×	○
15	*rplM*	TIGR01066	141	COG0102	J	50S ribosomal protein L13	99.06 (99.69)	98.89 (98.95)	○	○	○	○
16	*rplS*	TIGR01024	114	COG0335	J	50S ribosomal protein L19	98.95 (99.57)	98.77 (98.83)	○	○	×	○
17	*rpsE*	TIGR01021	156	COG0098	J	30S ribosomal protein S5	98.86 (99.83)	98.77 (98.86)	○	○	○	○
18	*tsf*	TIGR00116	293	COG0264	J	Elongation factor Ts	98.83 (99.52)	98.75 (95.44)	○	○	○	○
19	*rpmA*	TIGR00062	83	COG0211	J	50S ribosomal protein L27	98.83 (99.49)	98.72 (98.66)	○	○	×	○
20	*cgtA*	TIGR02729	329	COG0536	DL	GTPase ObgE/CgtA	98.77 (99.57)	98.57 (50.6)	○	○	○	○
21	*infB*	TIGR00487	587	COG0532	J	Translation initiation factor IF-2	98.72 (99.89)	98.57 (70.95)	○	○	○	○
22	*rplL*	TIGR00855	125	COG0222	J	50S ribosomal protein L7/L12	98.83 (99.34)	98.57 (91.65)	○	○	×	○
23	*rplQ*	TIGR00059	112	COG0203	J	50S ribosomal protein L17	98.6 (99.57)	98.52 (98.89)	○	○	○	○
24	*rplF*	TIGR03654	175	COG0097	J	50S ribosomal protein L6	98.6 (99.77)	98.52 (99.12)	○	○	○	○
25	*rplO*	TIGR01071	144	COG0200	J	50S ribosomal protein L15	98.55 (99.6)	98.46 (99.03)	○	○	○	○
26	*rpoA*	TIGR02027	298	COG0202	K	DNA-directed RNA polymerase subunit alpha	99.06 (100)	98.46 (91.79)	○	○	○	○
27	*rplV*	TIGR01044	103	COG0091	J	50S ribosomal protein L22	98.55 (99.26)	98.43 (98.43)	○	○	○	○
28	*rplU*	TIGR00061	101	COG0261	J	50S ribosomal protein L21	98.52 (98.8)	98.4 (97.98)	○	○	○	○
29	*rpsM*	TIGR03631	113	COG0099	J	30S ribosomal protein S13	98.49 (99.52)	98.38 (90.62)	○	○	×	○
30	*rpmC*	TIGR00012	56	COG0255	J	50S ribosomal protein L29	98.46 (98.69)	98.32 (96.29)	○	○	×	○
31	*rplI*	TIGR00158	148	COG0359	J	50S ribosomal protein L9	98.43 (98.86)	98.32 (98.03)	○	○	○	○
32	*rplX*	TIGR01079	104	COG0198	J	50S ribosomal protein L24	98.43 (99)	98.29 (97.58)	○	○	○	○
33	*ksgA*	TIGR00755	256	COG0030	J	Ribosomal RNA small subunit methyltransferase A	98.35 (99.37)	98.2 (4.59)	○	○	○	○
34	*rpsQ*	TIGR03635	72	COG0186	J	30S ribosomal protein S17	98.29 (99.69)	98.15 (95.15)	○	○	×	○
35	*rpsG*	TIGR01029	154	COG0049	J	30S ribosomal protein S7	98.26 (99.91)	98.06 (97.23)	○	○	○	○
36	*rpsS*	TIGR01050	92	COG0185	J	30S ribosomal protein S19	98.15 (99.63)	98.06 (98.95)	○	○	×	○
37	*rplK*	TIGR01632	140	COG0080	J	50S ribosomal protein L11	98.29 (99.71)	98.03 (98.46)	○	○	○	○
38	*rpoB*	TIGR02013	1238	COG0085	K	DNA-directed RNA polymerase subunit beta	98.35 (99.69)	98.03 (96.72)	○	○	○	○

Table 1. Continued

	Gene	HMM*	Length	COG†	COG category§	Function	Presence ratio (%) Trusted cutoff is used (10E-5 E-value is used)	Single-copy ratio (%) Trusted cutoff is used (10E-5 E-value is used)	UBCG2	UBCG1	bac120 Parks *et al.* (2017)	bcgTree Dupont *et al.* (2012), Ankenbrand and Keller (2016)
39	*rpsF*	TIGR00166	95	COG0360	J	30S ribosomal protein S6	98.15 (99.49)	98.03 (98.95)	○	○	○	○
40	*ybeY*	TIGR00043	111	COG0319	J	Endoribonuclease YbeY	98.23 (98.6)	98 (97.75)	○	○	○	○
41	*rplT*	TIGR01032	114	COG0292	J	50S ribosomal protein L20	98.15 (99.57)	97.98 (98.89)	○	○	○	○
42	*rplP*	TIGR01164	126	COG0197	J	50S ribosomal protein L16	98.06 (99.54)	97.98 (97.98)	○	○	○	○
43	*rsmH*	TIGR00006	310	COG0275	M	16S rRNA (cytosine[1402]-N[4])-methyltransferase	99.34 (99.63)	97.89 (83.44)	○	×	○	○
44	*trmD*	TIGR00088	233	COG0336	J	tRNA (guanine[37]-N[1])-methyltransferase	98 (99.03)	97.86 (97.32)	○	×	○	×
45	*dnaX*	TIGR02397	355	COG2812	L	DNA polymerase III subunit gamma	98.66 (99.46)	97.75 (0.51)	○	○	○	○
46	*rpmI*	TIGR00001	63	COG0291	J	50S ribosomal protein L35	98.03 (99.06)	97.81 (98.83)	○	○	×	○
47	*rplA*	TIGR01169	227	COG0081	J	50S ribosomal protein L1	97.92 (99.43)	97.81 (97.12)	○	○	○	○
48	*rplD*	TIGR03953	188	COG0088	J	50S ribosomal protein L4	97.83 (99.94)	97.69 (97.92)	○	○	○	○
49	*frr*	TIGR00496	176	COG0233	J	Ribosome-recycling factor	97.63 (99.52)	97.61 (97.55)	○	○	○	○
50	*engA*	TIGR03594	432	COG1160	R	GTPase Der	97.41 (99.86)	97.35 (0.74)	○	○	○	○
51	*rplR*	TIGR00060	114	COG0256	J	50S ribosomal protein L18	97.41 (98.23)	97.32 (97.06)	○	○	×	○
52	*ychF*	TIGR00092	368	COG0012	J	Ribosome-binding ATPase YchF	97.55 (98.89)	97.32 (96.15)	○	○	○	○
53	*nusA*	TIGR01953	340	COG0195	K	Transcription termination/antitermination protein NusA	97.26 (99.14)	97.21 (94.84)	○	○	○	○
54	*pheS*	TIGR00468	324	COG0016	J	Phenylalanine-tRNA ligase alpha subunit	97.38 (99.66)	97.18 (95.81)	○	○	○	○
55	*smpB*	TIGR00086	144	COG0691	O	SsrA-binding protein	97.78 (99.69)	97.15 (98.4)	○	○	○	○
56	*alaS*	TIGR00344	847	COG0013	J	Alanine-tRNA ligase	97.63 (99.46)	97.06 (65.34)	○	○	○	○
57	*tsaD*	TIGR03723	314	COG0533	J	tRNA N6-adenosine threonylcarbamoyltransferase	97.98 (99.29)	97.04 (68.7)	○	○	○	×
58	*prfA*	TIGR00019	361	COG0216	J	Peptide chain release factor 1	97.06 (99.83)	97.01 (66.13)	○	○	○	○
59	*leuS*	TIGR00396	843	COG0495	J	Leucine-tRNA ligase	97.72 (99.63)	96.98 (95.78)	○	○	○	○
60	*tilS*	TIGR02432	189	COG0037	J	tRNA(Ile)-lysidine synthase	96.92 (98.92)	96.81 (28.93)	○	○	○	○
61	*secY*	TIGR00967	414	COG0201	U	Protein translocase subunit SecY	97.69 (99.26)	96.81 (93.56)	○	○	○	○
62	*lepA*	TIGR01393	595	COG0481	J	Elongation factor 4	97.63 (99.4)	96.64 (38.88)	○	○	○	○
63	*ffh*	TIGR00959	428	COG0541	U	Signal recognition particle protein	96.69 (98.23)	96.61 (46.44)	○	○	○	○
64	*dnaG*	TIGR01391	414	COG0358	L	DNA primase	97.83 (99.26)	96.47 (66.56)	○	○	○	○
65	*infC*	TIGR00168	165	COG0290	J	Translation initiation factor IF-3	99.17 (99.69)	96.49 (95.61)	○	○	○	○
66	*ftsY*	TIGR00064	279	COG0552	U	Signal recognition particle receptor FtsY	96.29 (98.52)	96.18 (18.19)	○	○	○	○
67	*truB*	TIGR00431	210	COG0130	J	tRNA pseudouridine synthase B	96.41 (97.63)	96.12 (96.64)	○	○	○	×
68	*rpsD*	TIGR01017	200	COG0522	J	30S ribosomal protein S4	99 (100)	96.09 (30.87)	○	○	○	○
69	*nusG*	TIGR00922	172	COG0250	K	Transcription termination/antitermination protein NusG	95.9 (98.97)	95.81 (76.85)	○	○	○	○
70	*secA*	TIGR00963	787	COG0653	U	Protein translocase subunit SecA	97.86 (99.2)	95.81 (85.72)	○	○	○	○
71	*gmk*	TIGR03263	180	COG0194	F	Guanylate kinase	95.92 (98.52)	95.75 (20.92)	○	○	○	○
72	*fmt*	TIGR00460	315	COG0223	J	Methionyl-tRNA formyltransferase	95.81 (98.4)	95.67 (7.61)	○	○	○	○
73	*pheT*	TIGR00472	798	COG0072	J	Phenylalanine-tRNA ligase beta subunit	95.9 (99.2)	95.72 (78.19)	○	○	○	○
74	*serS*	TIGR00414	418	COG0172	J	Serine-tRNA ligase	98.23 (99.43)	95.58 (86.66)	○	○	○	○
75	*ileS*	TIGR00392	861	COG0060	J	Isoleucine-tRNA ligase 1	97.43 (99.66)	95.41 (94.21)	○	○	○	○
76	*hisS*	TIGR00442	406	COG0124	J	Histidine-tRNA ligase	98.03 (99.46)	95.35 (45.92)	○	○	○	○
77	*rpsR*	TIGR00165	70	COG0238	J	30S ribosomal protein S18	99.46 (99.66)	95.3 (94.87)	○	○	×	○

Table 1. Continued

	Gene	HMM*	Length	COG†	COG category§	Function	Presence ratio (%) Single-copy ratio (%) Trusted cutoff is used (10E-5 E-value is used)		UBCG2	UBCG1	bac120 Parks *et al.* (2017)	bcgTree Dupont *et al.* (2012), Ankenbrand and Keller (2016)
78	*recA*	TIGR02012	321	COG0468	L	DNA recombination and repair protein	97.01 (97.89)	95.21 (81.13)	○	○	○	○
79	*ruvB*	TIGR00635	305	COG2255	L	holliday junction DNA helicase RuvB	95.5 (97.66)	95.1 (6.78)	○	×	○	×
80	*era*	TIGR00436	270	COG1159	J	GTP-binding protein Era	95.58 (96.61)	95.13 (51.91)	○	×	○	○
81	*rpsT*	TIGR00029	87	COG0268	J	30S ribosomal protein S20	95.18 (97.83)	95.04 (96.89)	○	○	○	○
82	*aspS*	TIGR00459	586	COG0173	J	Aspartate-tRNA ligase	96.52 (99.69)	94.93 (8.04)	×	○	○	○
83	*rpoC*	TIGR02386	1147	COG0086	K	DNA-directed RNA polymerase subunit beta	95.01 (99.34)	94.84 (92.08)	×	○	○	○
84	*coaE*	TIGR00152	188	COG0237	H	Dephospho-CoA kinase	95.13 (98.6)	94.75 (18.24)	×	○	×	○
85	*pyrH*	TIGR02075	233	COG0528	F	UMP kinase	95.58 (98.86)	94.75 (4.9)	×	×	○	×
86	*rpsK*	TIGR03632	117	COG0100	J	30S ribosomal protein S11	94.84 (97.26)	94.75 (94.64)	×	○	○	○
87	*uvrC*	TIGR00194	574	COG0322	L	excinuclease ABC subunit C	95.3 (96.66)	94.67 (63.91)	×	×	○	×
88	*uvrB*	TIGR00631	658	COG0556	L	UvrABC system protein B	95.24 (96.81)	94.56 (13.2)	×	○	○	○
89	*yqgF*	TIGR00250	130	COG0816	L	putative transcription antitermination factor YqgF	94.73 (95.64)	94.61 (91.85)	×	×	○	×
90	*secG*	TIGR00810	73	COG1314	U	Protein-export membrane protein SecG	94.67 (95.32)	94.56 (92.73)	×	○	○	○
91	*metG*	TIGR00398	530	COG0143	J	methionine--tRNA ligase	97.41 (99.43)	94.38 (87)	×	×	○	×
92	*pgk*	PF00162.18	379	COG0126	F	Phosphoglycerate kinase	98.23 (98.2)	94.36 (93.53)	×	○	×	○
93	*trmU*	TIGR00420	351	COG0482	J	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	97.04 (99.4)	94.33 (8.24)	×	×	○	○
94	*rbfA*	TIGR00082	115	COG0858	J	30S ribosome-binding factor	94.27 (98.63)	94.21 (96.81)	×	○	○	○
95	*polA*	TIGR00593	890	COG0749	L	DNA polymerase I	94.16 (98.92)	94.04 (62.09)	×	×	○	×
96	*prfB*	TIGR00020	365	COG1186	J	peptide chain release factor 2	94.01 (96.86)	93.96 (74.29)	×	×	○	×
97	*murD*	TIGR01087	441	COG0771	M	UDP-N-acetylmuramoylalanine-D-glutamate ligase	94.53 (95.64)	93.93 (7.04)	×	×	○	×
98	*ligA*	TIGR00575	652	COG0272	L	DNA ligase	96.24 (98.55)	93.93 (65.11)	×	○	×	○
99	*dnaA*	TIGR00362	437	COG0593	L	Chromosomal replication initiator protein DnaA	94.64 (97.95)	93.9 (20.18)	×	○	○	○
100	*pyrG*	TIGR00337	526	COG0504	F	CTP synthase	94.13 (98.12)	93.79 (18.3)	×	○	○	○
101	*cysS*	TIGR00435	466	COG0215	J	Cysteine-tRNA ligase	97.15 (99.32)	93.79 (78.34)	×	○	○	○
102	*nusB*	TIGR01951	131	COG0781	J	transcription antitermination factor NusB	94.16 (98.66)	93.67 (30.07)	×	×	○	×
103	*argS*	TIGR00456	569	COG0018	J	Arginine-tRNA ligase	97.49 (99.09)	93.67 (88.63)	×	○	○	○
104	*coaD*	TIGR01510	155	COG0669	H	pantetheine-phosphate adenylyltransferase	94.93 (97.32)	93.56 (8.3)	×	×	○	×
105	*rnc*	TIGR02191	219	COG0571	K	Ribonuclease 3	95.18 (98.06)	93.59 (88.88)	×	○	○	○
106	*gyrB*	TIGR01059	639	COG0187	L	DNA gyrase, B subunit	94.53 (99.32)	93.56 (22.26)	×	×	○	○
107	*tig*	TIGR00115	410	COG1047	O	Trigger factor	94.3 (98.03)	93.16 (34.69)	×	○	○	○
108	*yeaZ*	TIGR03725	212	COG1214	O	tRNA threonylcarbamoyl adenosine modification protein YeaZ	93.27 (97.43)	93.13 (95.52)	×	×	○	×
109	*gyrA*	TIGR01063	800	COG0188	L	DNA gyrase, A subunit	97.66 (99.26)	93.1 (18.56)	×	×	○	○
110	*tyrS*	TIGR00234	406	COG0162	J	tyrosine--tRNA ligase	98.49 (99.63)	93.02 (15.45)	×	×	×	○
111	*murC*	TIGR01082	449	COG0773	M	UDP-N-acetylmuramate--L-alanine ligase	93.73 (96.04)	93.02 (6.7)	×	×	○	×
112	*mraY*	TIGR00445	321	COG0472	M	phospho-N-acetylmuramoyl-pentapeptide-transferase	94.1 (95.3)	92.99 (28.91)	×	×	○	×
113	*rpmF*	TIGR01031	56	COG0333	J	ribosomal protein bL32	96.81 (97.23)	92.99 (92.65)	×	×	×	○
114	*mfd*	TIGR00580	923	COG1197	L	transcription-repair coupling factor	92.96 (94.98)	92.93 (39.34)	×	×	○	×
115	*pnp*	PF03726.14	83	COG1185	J	Polyribonucleotide nucleotidyltransferase, RNA binding domain	92.99 (70.72)	92.9 (69.58)	×	×	○	×

Table 1. Continued

	Gene	HMM*	Length	COG†	COG category§	Function	Presence ratio (%) Single-copy ratio (%) Trusted cutoff is used (10E-5 E-value is used)		UBCG2	UBCG1	bac120 Parks *et al.* (2017)	bcgTree Dupont *et al.* (2012), Ankenbrand and Keller (2016)
116	*dnaN*	TIGR00663	367	COG0592	L	DNA polymerase III, beta subunit	97.55 (99.37)	92.9 (88.94)	×	×	○	○
117	*ftsZ*	TIGR00065	353	COG0206	D	cell division protein FtsZ	95.72 (96.98)	92.84 (86.15)	×	×	○	×
118	*rimM*	TIGR02273	166	COG0806	J	16S rRNA processing protein RimM	92.84 (94.56)	92.73 (88.54)	×	×	○	×
119	*recR*	TIGR00615	196	COG0353	L	recombination protein RecR	92.87 (95.41)	92.7 (90.48)	×	×	○	×
120	*grpE*	PF01025.19	166	COG0576	O	GrpE	99.71 (99.71)	92.67 (84.21)	×	×	○	○
121	*clpX*	TIGR00382	414	COG1219	O	ATP-dependent Clp protease, ATP-binding subunit ClpX	95.44 (99.8)	92.59 (2.96)	×	×	○	×
122	*typA*	TIGR01394	594	COG1217	T	GTP-binding protein TypA/BipA	92.16 (100)	92.02 (0.03)	×	×	○	×
123	*rsfS*	TIGR00090	99	COG0799	S	ribosome silencing factor	92.08 (92.56)	92.02 (92.1)	×	×	○	×
124	*rsmG*	TIGR00138	183	COG0357	J	16S rRNA (guanine(527)-N(7))-methyltransferase RsmG	92.08 (96.38)	91.79 (60.38)	×	×	○	×
125	*rpmB*	TIGR00009	58	COG0227	J	ribosomal protein bL28	97.38 (98.57)	91.68 (92.62)	×	×	×	○
126	*atpG*	TIGR01146	286	COG0224	C	ATP synthase F1, gamma subunit	92.93 (94.84)	91.68 (86.97)	×	×	○	×
127	*valS*	TIGR00422	863	COG0525	J	valine-tRNA ligase	91.9 (99.77)	91.42 (94.38)	×	×	×	○
128	*dnaK*	TIGR02350	596	COG0443	O	chaperone protein DnaK	97.75 (99.94)	91.19 (10.38)	×	×	○	○
129	*trmH*	TIGR00186	240	COG0566	J	RNA methyltransferase, TrmH family, group 3	91.62 (98.55)	91.08 (4.22)	×	×	○	×
130	*ribF*	TIGR00083	290	COG0196	H	riboflavin biosynthesis protein RibF	90.74 (96.49)	90.56 (90.96)	×	×	○	×
131	*thrS*	TIGR00418	565	COG0441	J	threonine--tRNA ligase	97.55 (99.17)	90.05 (67.9)	×	×	×	○
132	*secE*	TIGR00964	57	COG0690	U	preprotein translocase, SecE subunit	90.17 (90.85)	90.02 (89.37)	×	×	○	○
133	*radA*	TIGR00416	454	COG1066	O	DNA repair protein RadA	90.05 (99.6)	89.88 (6.1)	×	×	○	×
134	*hemN*	TIGR00539	361	COG0635	H	putative oxygen-independent coproporphyrinogen III oxidase	90.11 (93.73)	89.79 (19.98)	×	×	○	×
135	*rimP*	PF02576.17	73	COG0779	S	RimP N-terminal domain	89.51 (89.94)	89.37 (88.94)	×	×	○	×
136	*atpD*	TIGR01039	462	COG0055	C	ATP synthase F1, beta subunit	94.64 (99.34)	88.8 (0.43)	×	×	○	×
137	*ruvA*	TIGR00084	192	COG0632	L	Holliday junction DNA helicase RuvA	88.11 (97.01)	88.06 (71.72)	×	×	○	×
138	*purB*	TIGR00928	436	COG0015	F	adenylosuccinate lyase	89.42 (97.12)	86.46 (5.76)	×	×	○	×
139	*recN*	TIGR00634	563	COG0497	L	DNA repair protein RecN	85.03 (95.15)	84.95 (7.01)	×	×	○	×
140	*rsmD*	TIGR00095	194	COG0742	L	16S rRNA (guanine(966)-N(2))-methyltransferase RsmD	84.95 (97.95)	84.78 (14.25)	×	×	○	×
141	*guaB*	TIGR01302	450	COG0516	F	inosine-5′-monophosphate dehydrogenase	84.83 (98.8)	84.49 (1.57)	×	×	○	×
142	*rseP*	TIGR00054	421	COG0750	M	RIP metalloprotease RseP	83.32 (96.86)	81.9 (4.42)	×	×	○	×
143	*holA*	TIGR01128	314	COG1466	L	DNA polymerase III, delta subunit	81.41 (98.29)	80.93 (94.21)	×	×	○	×
144	*proS*	TIGR00409	568	COG0442	J	proline--tRNA ligase	79.08 (99.34)	78.76 (84.29)	×	×	×	○
145	*rpmH*	TIGR01030	44	COG0230	J	ribosomal protein bL34	75.14 (75.29)	75.03 (75.09)	×	×	×	○
146	*recG*	TIGR00643	629	COG1200	L	ATP-dependent DNA helicase RecG	69.56 (97.15)	69.38 (3.62)	×	×	○	×
147	*glyS*	TIGR00211	691	COG0751	J	glycine--tRNA ligase, beta subunit	63.71 (64)	63.57 (62.94)	×	×	×	○
148	*rpsA*	TIGR00717	516	COG0539	J	ribosomal protein bS1	46.38 (99.52)	46.21 (14.45)	×	×	○	×
									81	92	120	107

* Hidden Markov Model; † Clusters of Orthologous Groups of proteins;
§ Categories: C, energy production and conversion; D, cell cycle control and mitosis; F, nucleotide metabolism and transport; H, coenzyme metabolism; J, translation; K, transcription; L, replication and repair; M, cell wall/membrane/envelop biogenesis; O, post-translational modification, protein turnover, chaperone functions; R, general functional prediction only; S, function unknown; T, signal transduction; U, intracellular trafficking and secretion.

ber of species considered for determining the bacterial core gene set increased significantly from 1,429 (UBCG) to 3,508 (UBCG2).

Inferring phylogenomic trees using bacterial core genes has been widely used in taxonomy. It may become a standard method for the description of new taxa or genome-based phylogenetic studies, particularly for genus or higher-level taxa. We believe that our updated bacterial core gene set and accompanying easy-to-use bioinformatics pipeline should provide valuable means to researchers in the various fields of microbiology.

Acknowledgements

This research was supported by the Collaborative Genome Program for Fostering New Post-Genome Industry through the National Research Foundation of Korea (NRF) funded by the Ministry of Science ICT and Future Planning (NRF-2014M3C9A3063541), and the Korean Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA) of Korea (918013-04-4-SB010).

Conflict of Interest

The authors declare that they have no conflict of interest.

Open Access

References

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

Asnicar, F., Thomas, A.M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., May, U., *et al.* 2020. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 2500.

Chun, J., Grim, C.J., Hasan, N.A., Lee, J.H., Choi, S.Y., Haley, B.J., Taviani, E., Jeon, Y.S., Kim, D.W., Lee, J.H., *et al.* 2009. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc. Natl. Acad. Sci. USA* **106**, 15442–15447.

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.

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.

Darling, A.E., Jospin, G., Lowe, E., Matsen IV, F.A., 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.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., *et al.* 2019. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432.

Glaeser, S.P. and Kämpfer, P. 2015. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Syst. Appl. Microbiol.* **38**, 237–245.

Ha, S.M., Kim, C.K., Roh, J., Byun, J.H., Yang, S.J., Choi, S.B., Chun, J., and Yong, D. 2019. Application of the whole genome-based bacterial identification system, TrueBacID, using clinical isolates that were not identified with three matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) systems. *Ann. Lab. Med.* **39**, 530–536.

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.

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.

Lee, M.D. 2019. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* **35**, 4162–4164.

Lee, I., Chalita, M., Ha, S.M., Na, S.I., Yoon, S.H., and Chun, J. 2017. ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences. *Int. J. Syst. Evol. Microbiol.* **67**, 2053–2057.

Na, S.I., Kim, Y.O., Yoon, S.H., Ha, S.M., Baek, I., and Chun, J. 2018. UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J. Microbiol.* **56**, 280–285.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., and Hugenholtz, P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542.

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

Selengut, J.D., Haft, D.H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W.C., Richter, A.R., and White, O. 2007. TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.* **35**, D260–D264.

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

Wu, M. and Scott, A.J. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034.

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.

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J.G., Belda-Ferre, P., Al-Ghalith, G.A., Kopylova, E., McDonald, D., *et al.* 2019. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477.