SUPPLEMENTARY MATERIALS

1 DLTree algorithm

The algorithm used in DLTree has been described previously (Yu, et al., 2010; Yu, et al., 2009; Yu, et al., 2005). We only give a brief summary of the essentials here. We collect all protein sequences or protein-coding DNA sequences in a genome and counts the number of (overlapping) *k*-mers to form a raw vector with N^k (=20^k or 4^k) components, depending on whether protein or protein-coding DNA sequences are used. Furthermore, the numbers of *k*-mers are predicted from that of (*k*-1)-mers and 1-mers by the DL theory. The differences between the prediction and the actual counts are taken as new components of a renormalized vector. So an organism can be represented by a renormalized vector and the distance between species can be represented by the distance between their renormalized vectors. If we denote $p(a_1a_2\cdots a_k)$ the actual count of *k*-mers $a_1a_2\cdots a_k$, and define

$$\overrightarrow{A_i} = (A_1, A_2, \cdots, A_{N_k}) \tag{1}$$

$$D_{chord}(\vec{A}_i, \vec{A}_j) = \sqrt{2(1 - C(\vec{A}_i, \vec{A}_j))} \tag{2}$$

where $D_{chord}(\vec{A}_i, \vec{A}_j)$ represents the chord distance between genomes $\vec{A}_i$ and $\vec{A}_j$, $C(\vec{A}_i, \vec{A}_j) = \dfrac{\vec{A}_i \bullet \vec{A}_j}{\left|\vec{A}_i\right| \times \left|\vec{A}_j\right|}$,

$$A_i(a_1a_2\cdots a_k) = \begin{cases} \dfrac{p(a_1a_2\cdots a_k) - q(a_1a_2\cdots a_k)}{q(a_1a_2\cdots a_k)} & , for \quad q \neq 0 \\ 0 & , for \quad q = 0 \end{cases}, \tag{3}$$

and in Eq. (3), $q(a_1a_2\cdots a_k)$ is the background for *k*-mer $a_1a_2\cdots a_k$ and it is defined as

$$q(a_1a_2\cdots a_k) = \frac{p(a_1)p(a_2a_3\cdots a_k) + p(a_1a_2\cdots a_{k-1})p(a_k)}{2}. \tag{4}$$

There are some long-standing challenges in alignment-free algorithms. (**Challenge 1**) Requirement of huge memory. In the protein sequences case (generally there are 20 kinds of amino acids), the memory required is almost 10GB (20^7 *8B, each element in the renormalized vector is stored as a double) when *k* is equal to 7. So, how to represent it for computers with limited memory? (**Challenge 2**) Duplicate computation. In fact, it is possible that the distance between two species may be computed more than once. So, how to avoid the duplicate computation for reducing the load of the entire system? (**Challenge 3**) Time-consuming computation. Actually, the CPU time is mostly spent on computing the distances between species (or genomes). Not all renormalized vectors are stored in memory all the time in a machine. So, how to schedule those renormalized vectors to reduce the running time?

The following solutions were proposed to address the above challenges.

(**Solution 1**) Compressed vector. We found that the number of non-zero elements in a renormalized vector is very small. For example, when *k* is equal to 7, the maximum and minimum ratio of the number of non-zero elements to that of all elements are 7.7653% and 0.0079% respectively in all 253 species. Therefore it is necessary to compress the renormalized vectors.

Table S1. Data structures of three compressed vectors. For the SCV, the key and value represent the index and value of an element in a vector, respectively. For the PCV, we divide a vector of length of 20^7 into 20^2 sub-vectors of length 20^5 by the first two characters of a *k*-mer. We only show *k*-mers with prefix AA and AC as examples. For the MCV, on the basis of the PCV, we take out the index of an element of which the value is -1 as key2 and discard its value.

SCV		PCV				MCV					
		AA		AC		AA			AC		
key	value	key	value	key	value	Key	value	key2	Key	value	key2
0	1	0	1	0	1	0	1	4	0	1	4
10	2	10	2	10	2	10	2	5	10	2	5
100	3	100	3	100	3	100	3	110	100	3	110
200	4	200	4	200	4	200	4	300	200	4	300

Description of data structures. We design four data structures to store a renormalized vector. Full Vector (FV) is a vector without compression, only in memory and not written to disc. Single Compressed Vector (SCV) is a vector, only storing the non-zero values and its indexes. Partitioned Compressed Vector (PCV) is a vector with partition, only storing the non-zero values and its indexes in each partition. Minimal Compressed Vector (MCV) is a vector with partition, storing the values and its indexes for non-zero values except for -1 and only the indexes for -1 values in each partition. Notice that most of non-zero values are -1. The data structures are depicted in Table S1.

Experimental Settings. The following settings were used for the experiments: Machine Model: HP Z800 Workstation; Platform: Linux (Cent OS 7.0); Processor: Intel(R) Xeon(R) CPU X5660; CUP Single Core Speed: 2.80GHz; Number of Cores: 2; Cache: 12MB; FSB Speed: 1333MHz; RAM: 35GB; Hyper Threading™: Enabled ; HDD Average Read Speed: 117.85 MB/s; Values of *k* for *k*-mer: 6 and 7; Programming Language Used: Java.

Performance with different data structures. For all protein sequences, we tested the performance of *DLTree* with the above data structures on the 253 species under different computation modes. Table S2 shows FV is the best under all computation modes when *k*=6, because the average non-zero ratio is 30.0978% and it is too high. Table S3 shows PCV is the best when *k*=7, because the average non-zero ratio is 2.7085% and it is too low. So we adopt the FV for *k*=6 and the PCV for *k*=7 as the data structures in the *DLTree* based on the trade-off between the storage and computation time. As for the DNA sequences, we also do similar performance tests (the results were listed in Tables S4-S6.) and get similar conclusion.

Table S2. The running time (in seconds) of *DLTree* (for all protein sequences) at *k*=6 with different modes and data structures. We adopt a binary search algorithm when computing the distance between vectors. The speed in the serial-compare mode is lower than that in parallel-compare mode, so it is not necessary to do a full test in serial-compare mode.

Mode	FV	SCV	PCV	MCV
parallel-load + parallel-compare	1860	7321	5463	6716
serial-load + parallel-compare	2063	5983	5190	6421

Table S3. The running time (in seconds) of *DLTree* (for all protein sequences) at *k*=7 with different modes and data structures. The value with FV is an estimated value because it could not run through due to memory limitation. The estimated value 574408 was obtained based the running time for 1898/31878 of the entire task (34200 seconds).

Mode	FV	SCV	PCV	MCV
serial-load + parallel-compare	574408	13938	9581	11192

Table S4. The running time (in seconds) of *DLTree* (for all coding DNA sequences) at *k*=13 with different modes and data structures. We adopt a binary search algorithm when computing the distance between vectors. The speed in the serial-compare mode is lower than that in parallel-compare mode, so it is not necessary to do a full test in serial-compare mode.

Mode	FV	SCV	PCV	MCV
parallel-load + parallel-compare	2404	3670	3409	4907
serial-load + parallel-compare	2647	3728	3333	5041

Table S5. The running time (in seconds) of *DLTree* (for all coding DNA sequences) at *k*=14 with different modes and data structures. The value with FV is an estimated value. The estimated value 17875 and 15970 were obtained based the running time for 29211/31878 and 12096/31878 of the entire task (16380 and 6060 seconds respectively).

Mode	FV	SCV	PCV	MCV
parallel-load + parallel-compare	17875	5118	4854	6896
serial-load + parallel-compare	15970	5075	4626	7061

Table S6. The running time (in seconds) of *DLTree* (for all coding DNA sequences) at *k*=15 with different modes and data structures. The value with FV is an estimated value. The estimated value 255535 was obtained based the running time for 1497/31878 of the entire task (12000 seconds).

Mode	FV	SCV	PCV	MCV
serial-load + parallel-compare	255535	7794	7308	9658

(Solution 2) Fingerprint. We define the fingerprint of a file as the combination of its MD5-value computed by MD5 algorithm (Rivest, 1992) and its size, so we can identify two identical files by comparing their fingerprints. The fingerprint has two advantages: (i) If the fingerprints of a new file in a new job and some existing file in the *DLTree* server are identical, then we can view the renormalized vector of the existing file as that of the new file, avoiding duplicate computation;(ii) If the fingerprints of two new files in a new job and two existing files in the *DLTree* server are identical respectively, then we can view the distance between those two existing files as that of those two new files, avoiding duplicate compare.

(**Solution3**) **Scalable memory management**. The original algorithm for scalable memory management was presented in (Anaththa, et al., 2014). Here, we modify this algorithm by (i) cutting off the duplicate computation and avoiding loading the corresponding files at the same time; (ii) changing the serial mode to parallel mode in all procedures of reading files, computing vectors, writing compressed vectors and computing distances if it can improve the performance to run in parallel mode in those procedures.

2 DLTree web server

Input. DLTree allows two kinds of input data: selected genomes from the inbuilt database and user's uploaded data. Users may upload their own sequences to the DLTree web server. All protein sequences of one genome should be included in a single FASTA file (Figure S1). The file name (without extension) will be displayed as the species name in the trees. Due to the limitation of storage and computation power, up to 100MB uncompressed data can be uploaded once a time and up to 400 files can be included in a job once a time. For the inbuilt genomes, users are allowed to use keywords to select species of interest. For example, for the time being entering 'Archaea' as a keyword would bring up all the 30 Archaea names.

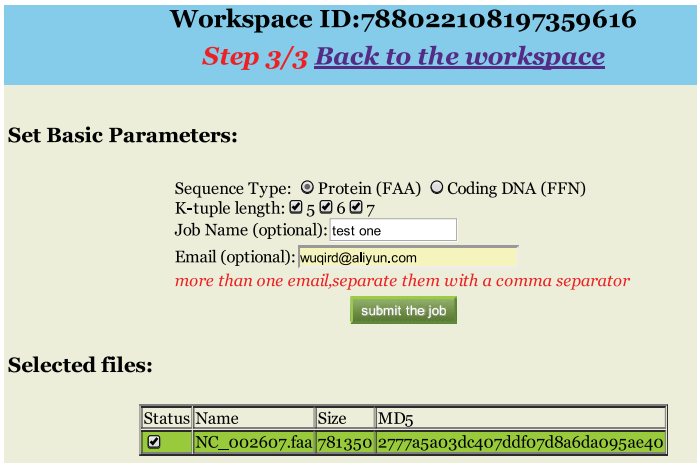

Figure S1. The parameter setting and job submission page.

Output. For each submission, users are able to track its status (queued, running or finished) by clicking on the "checking status" button. If an email is provided, the user will be notified by email when the job has been done, and the resulting data are reported on a webpage at the URL assigned. An example output page is available at: http://dltree.xtu.edu.cn/example. The output data include: (i) two distance matrix files *.meg and *.phylip (which can be used to construct the phylogenetic trees); (ii) a Newick tree format file *.tree (which can be viewed in MEGA); and (iii) six color-coded files (which will not be generated if none of species in the inbuilt database are selected).Those files are directly uploaded to a common iTOL (Letunic and Bork, 2016) personal account in order to be displayed in a different manner. The results are kept on the server for 90 days and will be deleted after that to save disc space in our system. All the results listed on the result page are collected together in a compressed file, which is provided for download on the

same page. Users are encouraged to download this file to their computer to store the results permanently. The output results are introduced in Figures S2-S3.

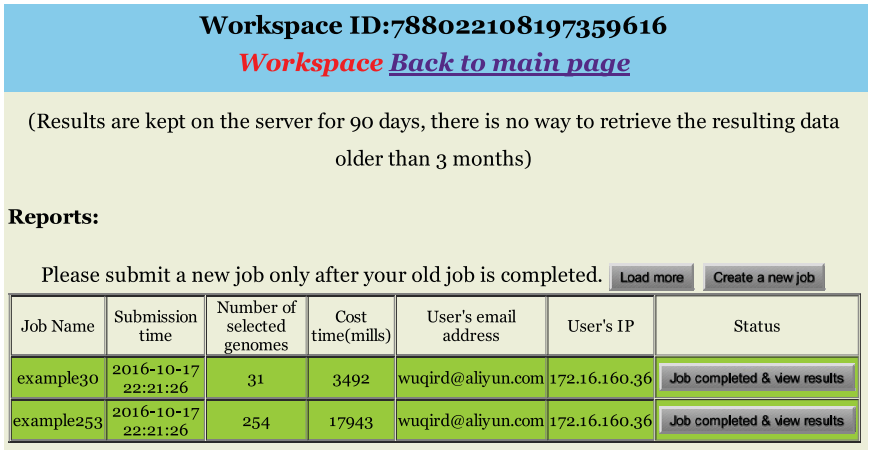

Figure S2. The workspace page displays the status of users' jobs. Once the job was labelled as 'Job completed & view results', users can access the result by clicking on its status button.

Figure S3. An example phylogenetic tree. The tree consists of 30 species (labelled by the Taxonomy ID) in the inbuilt database and one species (labelled by 'NC 002607') uploaded by the user was displayed using iTOL (Letunic and Bork, 2016). Each species was associated to six color strips from inner circle to outer circle according to the domain, kingdom, phylum, class, order and family. However species uploaded by the user were associated to six black strips to distinguish them from others. Species with the same color in a circle belong to the same class.

3 The inbuilt database

Table S7. The information of the 253 species in the inbuilt database, which were downloaded from the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). Scientific names are truncated. Only the first 30 characters of each are reserved. All protein or DNA coding sequences of one species (including Eukaryotes) are included in one FASTA file and just list one access record number in this table. Abbreviation *GI* : GenBank identifier; *TI*: Taxonomy identifier; *ARN*: Access record number.

ID	*GI*	*TI*	Scientific name	*ARN*	Kingdom
1	15789341	64091	Halobacterium sp. NRC-1	NC_002607	Archaea
2	20088900	188937	Methanosarcina acetivorans C2A	NC_003552	Archaea

3	20093441	190192	Methanopyrus kandleri AV19	NC_003551	Archaea
4	11497622	224325	Archaeoglobus fulgidus DSM 430	NC_000917	Archaea
5	15668173	243232	Methanocaldococcus jannaschii	NC_000909	Archaea
6	48477073	263820	Picrophilus torridus DSM 9790	NC_005877	Archaea
7	14521983	272844	Pyrococcus abyssi GE5	NC_000868	Archaea
8	13540832	273116	Thermoplasma volcanium GSS1	NC_002689	Archaea
9	88601323	323259	Methanospirillum hungatei JF-1	NC_007796	Archaea
10	518651361	333146	Ferroplasma acidarmanus fer1	NC_021592	Archaea
11	116753326	349307	Methanosaeta thermophila PT	NC_008553	Archaea
12	147918683	351160	Methanocella arvoryzae MRE50	NC_009464	Archaea
13	110666977	362976	Haloquadratum walsbyi DSM 1679	NC_008212	Archaea
14	126177953	368407	Methanoculleus marisnigri JR1	NC_009051	Archaea
15	119718919	368408	Thermofilum pendens Hrk 5	NC_008698	Archaea
16	119871521	384616	Pyrobaculum islandicum DSM 418	NC_008701	Archaea
17	146302786	399549	Metallosphaera sedula DSM 5348	NC_009440	Archaea
18	126464914	399550	Staphylothermus marinus F1	NC_009033	Archaea
19	134045047	402880	Methanococcus maripaludis C5	NC_009135	Archaea
20	124484830	410358	Methanocorpusculum labreanum Z	NC_008942	Archaea
21	124026907	415426	Hyperthermus butylicus DSM 545	NC_008818	Archaea
22	148642061	420247	Methanobrevibacter smithii ATC	NC_009515	Archaea
23	154149550	456442	Methanoregula boonei 6A8	NC_009712	Archaea
24	312136231	523846	Methanothermus fervidus DSM 20	NC_014658	Archaea
25	284163296	543526	Haloterrigena turkmenica DSM 5	NC_013743	Archaea
26	302347803	666510	Acidilobus saccharovorans 345-	NC_014374	Archaea
27	429216256	1056495	Caldisphaera lagunensis DSM 15	NC_019791	Archaea
28	385805052	1163730	Fervidicoccus fontis Kam940	NC_017461	Archaea
29	478482032	1236689	Candidatus Methanomethylophilu	NC_020913	Archaea
30	408402625	1237085	Candidatus Nitrososphaera garg	NC_018719	Archaea
31	261414415	59374	Fibrobacter succinogenes subsp	NC_013410	Bacteria
32	154243959	78245	Xanthobacter autotrophicus Py2	NC_009720	Bacteria
33	158520018	96561	Desulfococcus oleovorans Hxd3	NC_009943	Bacteria
34	111017023	101510	Rhodococcus jostii RHA1	NC_008268	Bacteria
35	117923319	156889	Magnetococcus marinus MC-1	NC_008576	Bacteria
36	15836606	160492	Xylella fastidiosa 9a5c	NC_002488	Bacteria
37	71278268	167879	Colwellia psychrerythraea 34H	NC_003910	Bacteria
38	114319167	187272	Alkalilimnicola ehrlichii MLHE	NC_008340	Bacteria
39	16124257	190650	Caulobacter crescentus CB15	NC_002696	Bacteria

40	110799791	195103	Clostridium perfringens ATCC 1	NC_008261	Bacteria
41	125972526	203119	Ruminiclostridium thermocellum	NC_009012	Bacteria
42	116490127	203123	Oenococcus oeni PSU-1	NC_008528	Bacteria
43	225847841	204536	Sulfurihydrogenibium azorense	NC_012438	Bacteria
44	113461928	205914	Haemophilus somnus 129PT	NC_008309	Bacteria
45	147668653	216389	Dehalococcoides mccartyi BAV1	NC_009455	Bacteria
46	116254466	216596	Rhizobium leguminosarum bv. vi	NC_008380	Bacteria
47	15605613	224324	Aquifex aeolicus VF5	NC_000918	Bacteria
48	29374662	226185	Enterococcus faecalis V583	NC_004668	Bacteria
49	225871700	240015	Acidobacterium capsulatum ATCC	NC_012483	Bacteria
50	15805043	243230	Deinococcus radiodurans R1	NC_001263	Bacteria
51	118466558	243243	Mycobacterium avium 104	NC_008595	Bacteria
52	108885075	243273	Mycoplasma genitalium G37	NC_000908	Bacteria
53	15638996	243276	Treponema pallidum subsp. pall	NC_000919	Bacteria
54	146328959	246195	Dichelobacter nodosus VCS1703A	NC_009446	Bacteria
55	108761581	246197	Myxococcus xanthus DK 1622	NC_008095	Bacteria
56	42521651	264462	Bdellovibrio bacteriovorus HD1	NC_005363	Bacteria
57	294672794	264731	Prevotella ruminicola 23	NC_014033	Bacteria
58	50364816	265311	Mesoplasma florum L1	NC_006055	Bacteria
59	108802857	266117	Rubrobacter xylanophilus DSM 9	NC_008148	Bacteria
60	13470331	266835	Mesorhizobium loti MAFF303099	NC_002678	Bacteria
61	152966184	266940	Kineococcus radiotolerans SRS3	NC_009664	Bacteria
62	110636428	269798	Cytophaga hutchinsonii ATCC 33	NC_008255	Bacteria
63	15604718	272561	Chlamydia trachomatis D/UW-3/C	NC_000117	Bacteria
64	126697567	272563	Peptoclostridium difficile 630	NC_009089	Bacteria
65	116510844	272622	Lactococcus lactis subsp. crem	NC_008527	Bacteria
66	121633902	272831	Neisseria meningitidis FAM18	NC_008767	Bacteria
67	206889903	289376	Thermodesulfovibrio yellowston	NC_011296	Bacteria
68	119355858	290317	Chlorobium phaeobacteroides DS	NC_008639	Bacteria
69	119964126	290340	Arthrobacter aurescens TC1	NC_008711	Bacteria
70	74316019	292415	Thiobacillus denitrificans ATC	NC_007404	Bacteria
71	51891139	292459	Symbiobacterium thermophilum I	NC_006177	Bacteria
72	206895191	309798	Coprothermobacter proteolyticu	NC_011295	Bacteria
73	206901546	309799	Dictyoglomus thermophilum H-6-	NC_011297	Bacteria
74	221632038	309801	Thermomicrobium roseum DSM 515	NC_011959	Bacteria
75	85375839	314225	Erythrobacter litoralis HTCC25	NC_007722	Bacteria
76	304319678	314260	Parvularcula bermudensis HTCC2	NC_014414	Bacteria

77	115522031	316055	Rhodopseudomonas palustris Bis	NC_008435	Bacteria
78	159896534	316274	Herpetosiphon aurantiacus DSM	NC_009972	Bacteria
79	103485499	317655	Sphingopyxis alaskensis RB2256	NC_008048	Bacteria
80	114561189	318167	Shewanella frigidimarina NCIMB	NC_008345	Bacteria
81	163845604	324602	Chloroflexus aurantiacus J-10-	NC_010175	Bacteria
82	111219506	326424	Frankia alni ACN14a	NC_008278	Bacteria
83	338731897	331113	Simkania negevensis Z	NC_015713	Bacteria
84	107021563	331271	Burkholderia cenocepacia AU 10	NC_008060	Bacteria
85	169823698	334413	Finegoldia magna ATCC 29328	NC_010376	Bacteria
86	114330037	335283	Nitrosomonas eutropha C91	NC_008344	Bacteria
87	114565577	335541	Syntrophomonas wolfei subsp. w	NC_008346	Bacteria
88	116747453	335543	Syntrophobacter fumaroxidans M	NC_008554	Bacteria
89	118578450	338966	Pelobacter propionicus DSM 237	NC_008609	Bacteria
90	167036432	340099	Thermoanaerobacter pseudethano	NC_010321	Bacteria
91	163854305	340100	Bordetella petrii DSM 12804	NC_010170	Bacteria
92	145294043	340322	Corynebacterium glutamicum R	NC_009342	Bacteria
93	109896333	342610	Pseudoalteromonas atlantica T6	NC_008228	Bacteria
94	147674590	345073	Vibrio cholerae O395	NC_009457	Bacteria
95	134297882	349161	Desulfotomaculum reducens MI-1	NC_009253	Bacteria
96	83642914	349521	Hahella chejuensis KCTC 2396	NC_007645	Bacteria
97	187734517	349741	Akkermansia muciniphila ATCC B	NC_010655	Bacteria
98	120552945	351348	Marinobacter hydrocarbonoclast	NC_008740	Bacteria
99	148262086	351605	Geobacter uraniireducens Rf4	NC_009483	Bacteria
100	117927212	351607	Acidothermus cellulolyticus 11	NC_008578	Bacteria
101	146295086	351627	Caldicellulosiruptor saccharol	NC_009437	Bacteria
102	148283998	357244	Orientia tsutsugamushi str. Bo	NC_009488	Bacteria
103	108562425	357544	Helicobacter pylori HPAG1	NC_008086	Bacteria
104	119943795	357804	Psychromonas ingrahamii 37	NC_008709	Bacteria
105	160878163	357809	Lachnoclostridium phytoferment	NC_010001	Bacteria
106	226309588	358681	Brevibacillus brevis NBRC 1005	NC_012491	Bacteria
107	148266448	359786	Staphylococcus aureus subsp. a	NC_009487	Bacteria
108	121602380	360095	Bartonella bacilliformis KC583	NC_008783	Bacteria
109	108805999	360102	Yersinia pestis Antiqua	NC_008150	Bacteria
110	118474813	360106	Campylobacter fetus subsp. fet	NC_008599	Bacteria
111	119025019	367928	Bifidobacterium adolescentis A	NC_008618	Bacteria
112	145592567	369723	Salinispora tropica CNB-440	NC_009380	Bacteria
113	220930851	373903	Halothermothrix orenii H 168	NC_011899	Bacteria

114	110677422	375451	Roseobacter denitrificans OCh	NC_008209	Bacteria
115	310817279	378806	Stigmatella aurantiaca DW4/3-1	NC_014623	Bacteria
116	226225407	379066	Gemmatimonas aurantiaca T-27	NC_012489	Bacteria
117	198282149	380394	Acidithiobacillus ferrooxidans	NC_011206	Bacteria
118	117617947	380703	Aeromonas hydrophila subsp. hy	NC_008570	Bacteria
119	156740029	383372	Roseiflexus castenholzii DSM 1	NC_009767	Bacteria
120	104779317	384676	Pseudomonas entomophila L48	NC_008027	Bacteria
121	116871423	386043	Listeria welshimeri serovar 6b	NC_008555	Bacteria
122	111114824	390236	Borrelia afzelii PKo	NC_008277	Bacteria
123	104773258	390333	Lactobacillus delbrueckii subs	NC_008054	Bacteria
124	148269146	390874	Thermotoga petrophila RKU-1	NC_009486	Bacteria
125	150019914	391009	Thermosipho melanesiensis BI42	NC_009616	Bacteria
126	114326665	391165	Granulibacter bethesdensis CGD	NC_008343	Bacteria
127	120601052	391774	Desulfovibrio vulgaris DP4	NC_008751	Bacteria
128	110669658	393115	Francisella tularensis subsp.	NC_008245	Bacteria
129	148557804	393480	Fusobacterium nucleatum subsp.	NC_009506	Bacteria
130	110832862	393595	Alcanivorax borkumensis SK2	NC_008260	Bacteria
131	114568555	394221	Maricaulis maris MCS10	NC_008347	Bacteria
132	182677003	395963	Beijerinckia indica subsp. ind	NC_010581	Bacteria
133	120608715	397945	Acidovorax citrulli AAC00-1	NC_008752	Bacteria
134	126640116	400667	Acinetobacter baumannii ATCC 1	NC_009085	Bacteria
135	148358140	400673	Legionella pneumophila str. Co	NC_009494	Bacteria
136	154250457	402881	Parvibaculum lavamentivorans D	NC_009719	Bacteria
137	134096621	405948	Saccharopolyspora erythraea NR	NC_009142	Bacteria
138	120434373	411154	Gramella forsetii KT0803	NC_008571	Bacteria
139	118475779	412694	Bacillus thuringiensis str. Al	NC_008600	Bacteria
140	209963360	414684	Rhodospirillum centenum SW	NC_011420	Bacteria
141	163849458	419610	Methylobacterium extorquens PA	NC_010172	Bacteria
142	154706441	434922	Coxiella burnetii Dugway 5J108	NC_009727	Bacteria
143	150002609	435590	Bacteroides vulgatus ATCC 8482	NC_009614	Bacteria
144	150006675	435591	Parabacteroides distasonis ATC	NC_009615	Bacteria
145	162446889	441768	Acholeplasma laidlawii PG-8A	NC_010163	Bacteria
146	148271179	443906	Clavibacter michiganensis subs	NC_009480	Bacteria
147	148559145	444178	Brucella ovis ATCC 25840	NC_009505	Bacteria
148	187250424	445932	Elusimicrobium minutum Pei191	NC_010644	Bacteria
149	257067224	446465	Brachybacterium faecium DSM 48	NC_013172	Bacteria
150	296127870	446466	Cellulomonas flavigena DSM 201	NC_014151	Bacteria

151	297558986	446468	Nocardiopsis dassonvillei subs	NC_014210	Bacteria
152	269793359	446469	Sanguibacter keddieii DSM 1054	NC_013521	Bacteria
153	291297539	446470	Stackebrandtia nassauensis DSM	NC_013947	Bacteria
154	269954811	446471	Xylanimonas cellulosilytica DS	NC_013530	Bacteria
155	162448270	448385	Sorangium cellulosum So ce56	NC_010162	Bacteria
156	182411827	452637	Opitutus terrae PB90-1	NC_010571	Bacteria
157	220915124	455488	Anaeromyxobacter dehalogenans	NC_011891	Bacteria
158	182433794	455632	Streptomyces griseus subsp. gr	NC_010572	Bacteria
159	188584643	457570	Natranaerobius thermophilus JW	NC_010718	Bacteria
160	256826461	469378	Cryptobacterium curtum DSM 156	NC_013170	Bacteria
161	284041472	469383	Conexibacter woesei DSM 14684	NC_013739	Bacteria
162	269124278	471852	Thermomonospora curvata DSM 43	NC_013510	Bacteria
163	229818503	471853	Beutenbergia cavernae DSM 1233	NC_012669	Bacteria
164	256831256	471856	Jonesia denitrificans DSM 2060	NC_013174	Bacteria
165	256823906	478801	Kytococcus sedentarius DSM 205	NC_013169	Bacteria
166	258650272	479431	Nakamurella multipartita DSM 4	NC_013235	Bacteria
167	271961610	479432	Streptosporangium roseum DSM 4	NC_013595	Bacteria
168	256389233	479433	Catenulispora acidiphila DSM 4	NC_013131	Bacteria
169	269836034	479434	Sphaerobacter thermophilus DSM	NC_013523	Bacteria
170	284028000	479435	Kribbella flavida DSM 17836	NC_013729	Bacteria
171	269797070	479436	Veillonella parvula DSM 2008	NC_013520	Bacteria
172	258404139	485915	Desulfohalobium retbaense DSM	NC_013223	Bacteria
173	255529917	485917	Pedobacter heparinus DSM 2366	NC_013061	Bacteria
174	256419058	485918	Chitinophaga pinensis DSM 2588	NC_013132	Bacteria
175	192359033	498211	Cellvibrio japonicus Ueda107	NC_010995	Bacteria
176	167628138	498761	Heliobacterium modesticaldum I	NC_010337	Bacteria
177	262193327	502025	Haliangium ochraceum DSM 14365	NC_013440	Bacteria
178	291294405	504728	Meiothermus ruber DSM 1279	NC_013946	Bacteria
179	383787662	511051	Caldisericum exile AZM16c01	NC_017096	Bacteria
180	238915977	515620	[Eubacterium] eligens ATCC 277	NC_012778	Bacteria
181	239616412	521045	Kosmotoga olearia TBF 19.5.1	NC_012785	Bacteria
182	257783815	521095	Atopobium parvulum DSM 20469	NC_013203	Bacteria
183	296137751	521096	Tsukamurella paurometabola DSM	NC_014158	Bacteria
184	258510021	521098	Alicyclobacillus acidocaldariu	NC_013205	Bacteria
185	257124815	523794	Leptotrichia buccalis C-1013-b	NC_013192	Bacteria
186	256827819	525897	Desulfomicrobium baculatum DSM	NC_013173	Bacteria
187	269791620	525903	Thermanaerovibrio acidaminovor	NC_013522	Bacteria

188	256370825	525909	Acidimicrobium ferrooxidans DS	NC_013124	Bacteria
189	284988630	526225	Geodermatophilus obscurus DSM	NC_013757	Bacteria
190	262200047	526226	Gordonia bronchialis DSM 43247	NC_013441	Bacteria
191	283777803	530564	Pirellula staleyi DSM 6068	NC_013720	Bacteria
192	308047723	550540	Ferrimonas balearica DSM 9799	NC_014541	Bacteria
193	295129530	553199	Propionibacterium acnes SK137	NC_014039	Bacteria
194	261854631	555778	Halothiobacillus neapolitanus	NC_013422	Bacteria
195	226938935	557598	Laribacter hongkongensis HLHK9	NC_012559	Bacteria
196	225618951	565034	Brachyspira hyodysenteriae WA1	NC_012225	Bacteria
197	190570479	570417	Wolbachia endosymbiont of Cule	NC_010981	Bacteria
198	288939765	572477	Allochromatium vinosum DSM 180	NC_013851	Bacteria
199	302390798	574087	Acetohalobium arabaticum DSM 5	NC_014378	Bacteria
200	291612473	580332	Sideroxydans lithotrophicus ES	NC_013959	Bacteria
201	300021539	582899	Hyphomicrobium denitrificans A	NC_014313	Bacteria
202	253995375	583345	Methylotenera mobilis JLW8	NC_012968	Bacteria
203	294053542	583355	Coraliomargarita akajimensis D	NC_014008	Bacteria
204	297567993	589865	Desulfurivibrio alkaliphilus A	NC_014216	Bacteria
205	284047387	591001	Acidaminococcus fermentans DSM	NC_013740	Bacteria
206	224372071	598659	Nautilia profundicola AmH	NC_012115	Bacteria
207	261749099	600809	Blattabacterium sp. (Periplane	NC_013418	Bacteria
208	291278434	639282	Deferribacter desulfuricans SS	NC_013939	Bacteria
209	372486702	640081	Dechlorosoma suillum PS	NC_016616	Bacteria
210	296392441	640132	Segniliparus rotundus DSM 4498	NC_014168	Bacteria
211	313674130	643867	Marivirga tractuosa DSM 4126	NC_014759	Bacteria
212	302341445	644282	Desulfarculus baarsii DSM 2075	NC_014365	Bacteria
213	297570614	644284	Arcanobacterium haemolyticum D	NC_014218	Bacteria
214	317120850	644966	Thermaerobacter marianensis DS	NC_014831	Bacteria
215	319788983	648996	Thermovibrio ammonificans HB-1	NC_014926	Bacteria
216	297622253	649638	Truepera radiovictrix DSM 1709	NC_014221	Bacteria
217	336065243	650150	Erysipelothrix rhusiopathiae s	NC_015601	Bacteria
218	317050200	653733	Desulfurispirillum indicum S5	NC_014836	Bacteria
219	337285378	667014	Thermodesulfatator indicus DSM	NC_015681	Bacteria
220	390945348	679935	Alistipes finegoldii DSM 17242	NC_018011	Bacteria
221	350268399	693746	Oscillibacter valericigenes Sj	NC_016048	Bacteria
222	332980606	697281	Mahella australiensis 50-1 BON	NC_015520	Bacteria
223	328954614	700015	Coriobacterium glomerans PW2	NC_015389	Bacteria
224	317123178	710696	Intrasporangium calvum DSM 430	NC_014830	Bacteria

225	297620247	716544	Waddlia chondrophila WSU 86-10	NC_014225	Bacteria
226	334143100	717773	Thioalkalimicrobium cyclicum A	NC_015581	Bacteria
227	326793323	717774	Marinomonas mediterranea MMB-1	NC_015276	Bacteria
228	327402012	755732	Fluviicola taffensis DSM 16823	NC_015321	Bacteria
229	300309347	757424	Herbaspirillum seropedicae SmR	NC_014323	Bacteria
230	327398174	760142	Hippea maritima DSM 10411	NC_015318	Bacteria
231	332662000	760192	Haliscomenobacter hydrossis DS	NC_015510	Bacteria
232	338173995	765952	Parachlamydia acanthamoebae UV	NC_015702	Bacteria
233	307543590	768066	Halomonas elongata DSM 2581	NC_014532	Bacteria
234	333981748	857087	Methylomonas methanica MC09	NC_015572	Bacteria
235	374286831	862908	Halobacteriovorax marinus SJ	NC_016620	Bacteria
236	326802615	866775	Aerococcus urinae ACS-120-V-Co	NC_015278	Bacteria
237	343082722	880070	Cyclobacterium marinum DSM 745	NC_015914	Bacteria
238	328951747	880072	Desulfobacca acetoxidans DSM 1	NC_015388	Bacteria
239	383760956	926550	Caldilinea aerophila DSM 14535	NC_017079	Bacteria
240	320159411	926569	Anaerolinea thermophila UNI-1	NC_014960	Bacteria
241	385808587	945713	Ignavibacterium album JCM 1651	NC_017464	Bacteria
242	393198728	1002809	Solibacillus silvestris StLB04	NC_018065	Bacteria
243	383765076	1142394	Phycisphaera mikurensis NBRC 1	NC_017080	Bacteria
244	543961848	1163617	Sulfuricella denitrificans skB	NC_022357	Bacteria
245	397689004	1191523	Melioribacter roseus P3M-2	NC_018178	Bacteria
246	409913918	1231626	Cardinium endosymbiont cEper1	NC_018605	Bacteria
247	508605262	1234679	Carnobacterium maltaromaticum	NC_019425	Bacteria
248	507379057	1276227	Spiroplasma chrysopicola DF-1	NC_021280	Bacteria
249	512550082	1303518	Chthonomonas calidirosea T49	NC_021487	Bacteria
250	507384423	1321370	Idiomarina loihiensis GSL 199	NC_021286	Bacteria
251	258597608	36329	Plasmodium falciparum 3D7	NC_004317	Eukaryota
252	19115952	284812	Schizosaccharomyces pombe 972h	NC_003424	Fungi
253	363748723	931890	Eremothecium cymbalariae DBVPG	NC_016450	Fungi

References

Anaththa, P.D.K., Kelly, W. and Tian, Y.C. (2014) Optimizing I/O Cost and Managing Memory for Composition Vector Method Based on Correlation Matrix Calculation in Bioinformatics, *Current Bioinformatics*, **9**, 234-245(212).

Letunic, I. and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees, *Nucleic Acids Research*, **44**, W242-W245.

Rivest, R.L. (1992) The MD5 message-digest algorithm. RFC 1321.

Yu, Z.G., *et al.* (2010) Whole-proteome phylogeny of large dsDNA viruses and parvoviruses through a composition vector method related to dynamical language model, *Bmc Evolutionary Biology*, **10**, : 192.

Yu, Z.G., *et al.* (2009) Proper distance metrics for phylogenetic analysis using complete genomes without sequence alignment, *International Journal of Molecular Sciences*, **11**, 1141-1154.

Yu, Z.G., *et al.* (2005) Phylogeny of Prokaryotes and Chloroplasts Revealed by a Simple Composition Approach on All Protein Sequences from Complete Genomes Without Sequence Alignment, *Journal of Molecular Evolution*, **60**, 538-545.