D210–D213 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 11 November 2010*
doi:10.1093/nar/gkq1007

tRNADB-CE 2011: tRNA gene database curated manually by experts

Takashi Abe[1,*], Toshimichi Ikemura[1], Junichi Sugahara[2], Akio Kanai[2], Yasuo Ohara[1], Hiroshi Uehara[1], Makoto Kinouchi[3], Shigehiko Kanaya[4], Yuko Yamada[1], Akira Muto[5] and Hachiro Inokuchi[1]

[1]Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga, 526-0829, [2]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, 997-0035, [3]Department of Bio-System Engineering, Graduate School of Science and Engineering, Yamagata University, Yonezawa, Yamagata, 992-0038, [4]Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, 630-0192 and [5]Faculty of Agriculture and Life Science, Hirosaki University, Hirosaki, Aomori, 036-8561, Japan

Received September 10, 2010; Accepted October 7, 2010

ABSTRACT

We updated the tRNADB-CE by analyzing 939 complete and 1301 draft genomes of prokaryotes and eukaryotes, 171 complete virus genomes, 121 complete chloroplast genomes and approximately 230 million sequences obtained by metagenome analyses of 210 environmental samples. The 287 102 tRNA genes in total, and thus two times of the tRNA genes compiled previously, are compiled, in which sequence information, clover-leaf structure and results of sequence similarity and oligonucleotide-pattern search can be browsed. In order to pool collective knowledge with help from any experts in the tRNA research field, we included a column to which comments can be added on each tRNA gene. By compiling tRNAs of known prokaryotes with identical sequences, we found high phylogenetic preservation of tRNA sequences, especially at a phylum level. Furthermore, a large number of tRNAs obtained by metagenome analyses of environmental samples had sequences identical to those found in known prokaryotes. The identical sequence group, therefore, can be used as phylogenetic markers to clarify the microbial community structure of an ecosystem. The updated tRNADB-CE provided functions, with which users can obtain the phylotype-specific markers (e.g. genus-specific markers) by themselves and clarify microbial community structures of ecosystems in detail. tRNADB-CE can be accessed freely at http://trna.nagahama-i-bio.ac.jp.

INTRODUCTION

More than 99% of microorganisms that inhabit natural environments are difficult to culture under laboratory conditions. Metagenomic analyses of mixed genome samples have been developed (1–3) to explore the enormous number of novel genome resources. In accord with the remarkable progress of DNA-sequencing technology, a vast quantity of metagenomic sequences obtained from a wide variety of environmental samples have been decoded and released from DDBJ/EMBL/GenBank. Because a significant portion of environmental DNA sequences is derived from unculturable microbes, we can acquire new knowledge of tRNA sequences from novel genomes. The 154 455 tRNA genes found in the metagenomic sequences were included in the tRNADB-CE. When we focused on a group of tRNAs with an identical sequence, we found tRNAs found only in a particular lineage of phylogenetic groups. Notably, such phylotype-specific tRNA sequences were found in many species-unknown genomic fragments obtained by metagenome analyses of environmental samples. This shows that tRNA is a good phylogenetic marker for discovering phylotype composition and microbial community structure in an environmental ecosystem.

MATERIALS AND METHODS

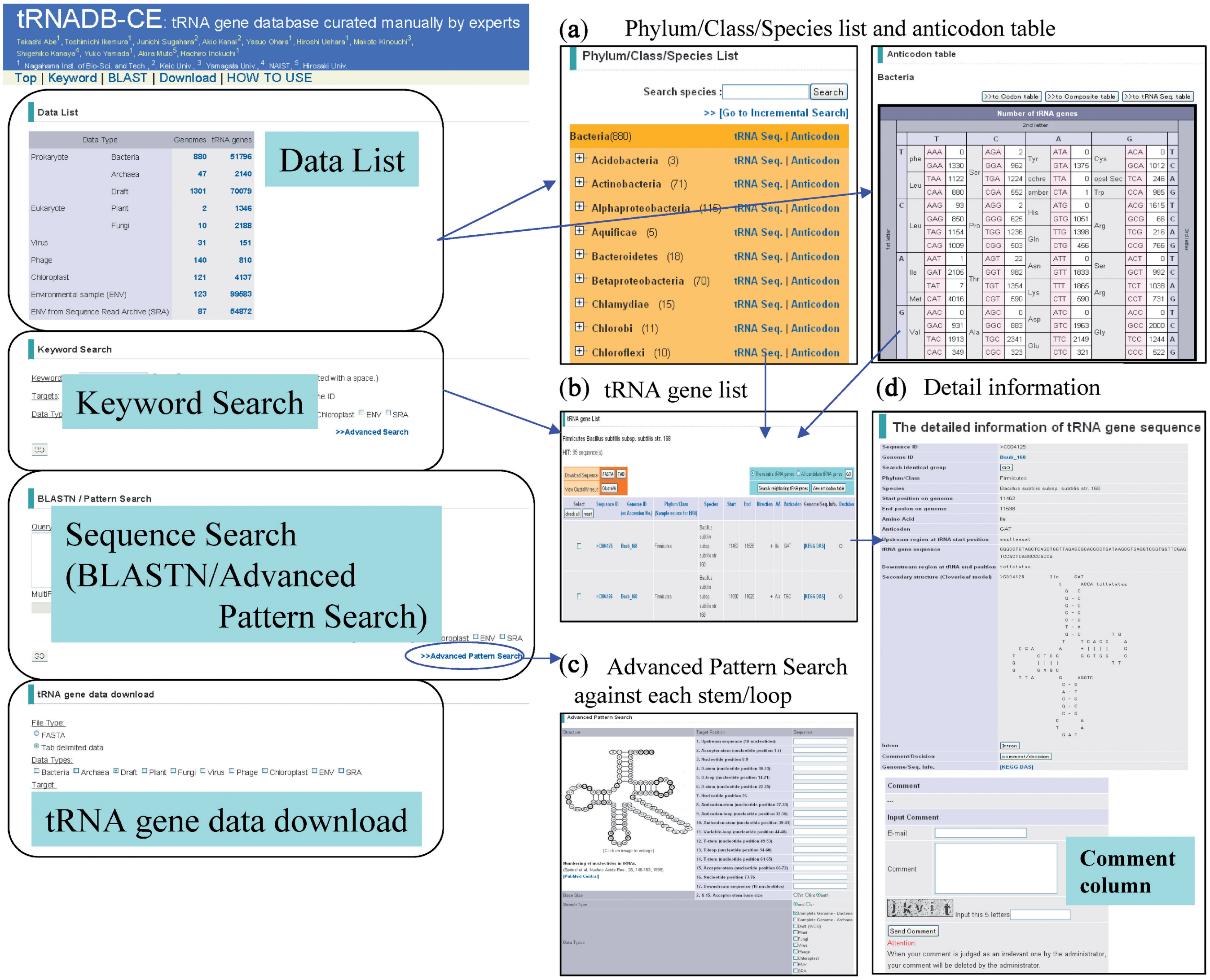
The following sources of DNA sequences were used: the complete genomes of 927 prokaryotes and of 171 viruses released by Genome Information Broker (GIB, http://gib.genes.nig.ac.jp/) and Genome Information Broker for Viruses (GIB-V, http://gib-v.genes.nig.ac.jp/) of DDBJ up to March 2009; the complete genomes of 121 chloroplasts released by Organelle Genome Database (GOBASE, http://gobase.bcm.umontreal.ca/) up to March 2009; the

*To whom correspondence should be addressed. Tel: +81 749 64 8100; Fax: +81 749 64 8140; Email: takaabe@nagahama-i-bio.ac.jp draft genome sequences of 1301 prokaryotes released by WGS division of DDBJ/EMBL/GenBank up to August 2009; the complete genomes of 12 eukaryotes and the 17 million metagenomic sequences released by DDBJ/EMBL/GenBank up to March 2010; and the 217 million metagenomic sequences obtained using a next-generation sequencer and released by Sequence Read Achieve (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra/) in NCBI up to March 2010.

RESULTS AND DISCUSSION

Update of registered tRNA genes and a new function for organizing collected knowledge

The 287 102 tRNA genes in total (53 936, 70 079, 961, 3534, 4137 and 154 455 genes from 927 complete prokaryote genomes, 1301 draft prokaryote genomes, 171 complete virus, 12 complete eukaryote genomes, 121 complete chloroplast genomes and 210 metagenomic samples, respectively) were registered in the updated tRNADB-CE. This was two times as many tRNA genes as registered in the previous version (4). This exhaustive search for tRNA genes was performed by running three computer programs used for tRNA gene search, tRNAscan-SE (5), ARAGORN (6) and tRNAfinder (7) in combination. This method should enhance the completeness and accuracy of prediction since the programs' algorithms are partially different and render somewhat different results. The tRNA genes found accordantly by all three programs were stored in tRNADB-CE without further checks. Then, the residual, discordant cases (~4% of the total of tRNA gene candidates), except for those found in metagenomic sequences, were checked, manually by three experts (Y.Y., A.M. and H.I.) in the tRNA experimental field and were classified into three categories: (i) reliable tRNA genes, (ii) not tRNA genes and (iii) ambiguous cases. Users can browse or download



tRNADB-CE: tRNA gene database curated manually by experts

Takashi Abe[1], Toshimichi Ikemura[1], Junichi Sugahara[2], Akio Kanai[2], Yasuo Ohara[1], Hiroshi Uehara[1], Makoto Kinouchi[3], Shigehiko Kanaya[4], Yuko Yamada[1], Akira Muto[5], Hachiro Inokuchi[1]

[1]. Nagahama Inst. of Bio-Sci. and Tech., [2]. Keio Univ., [3]. Yamagata Univ., [4]. NAIST, [5]. Hirosaki Univ.

Top | Keyword | BLAST | Download | HOW TO USE

Data List

Data Type		Genomes	tRNA genes
Prokaryote	Bacteria	880	51796
	Archaea	47	2140
	Draft	1301	70079
Eukaryote	Plant	2	1346
	Fungi	10	2188
Virus		31	151
Phage		140	810
Chloroplast		121	4137
Environmental sample (ENV)		123	99583
ENV from Sequence Read Archive (SRA)		87	54872

Keyword Search

Keyword ... ted with a space.)
Targets: ... ne ID
Data Typ ... Chloroplast □ ENV □ SRA
>>Advanced Search

BLASTN / Pattern Search

Sequence Search (BLASTN/Advanced Pattern Search)

>>Advanced Pattern Search

tRNA gene data download

File Type:
○ FASTA
● Tab delimited data
Data Types:
□ Bacteria □ Archaea ☑ Draft □ Plant □ Fungi □ Virus □ Phage □ Chloroplast □ ENV □ SRA
Target:

(a) Phylum/Class/Species list and anticodon table

Phylum/Class/Species List

Search species: Search
>> [Go to Incremental Search]

Bacteria(880)	tRNA Seq. | Anticodon
⊞ Acidobacteria (3)	tRNA Seq. | Anticodon
⊞ Actinobacteria (71)	tRNA Seq. | Anticodon
⊞ Alphaproteobacteria (115)	tRNA Seq. | Anticodon
⊞ Aquificae (5)	tRNA Seq. | Anticodon
⊞ Bacteroidetes (18)	tRNA Seq. | Anticodon
⊞ Betaproteobacteria (70)	tRNA Seq. | Anticodon
⊞ Chlamydiae (15)	tRNA Seq. | Anticodon
⊞ Chlorobi (11)	tRNA Seq. | Anticodon
⊞ Chloroflexi (10)	tRNA Seq. | Anticodon

Anticodon table

Bacteria

>>to Codon table | >>to Composite table | >>to tRNA Seq. table

Number of tRNA genes

1st letter	T			C			A			G			3rd letter
T	phe	AAA	0	Ser	AGA	2	Tyr	ATA	0	Cys	ACA	0	T
		GAA	1330		GGA	962		GTA	1375		GCA	1012	C
	Leu	TAA	1122		TGA	1224	ochre	TTA	0	opal Sec	TCA	246	A
		CAA	880		CGA	552	amber	CTA	1	Trp	CCA	985	G
C	Leu	AAG	93	Pro	AGG	2	His	ATG	0	Arg	ACG	1615	T
		GAG	850		GGG	625		GTG	1051		GCG	66	C
		TAG	1154		TGG	1236	Gln	TTG	1398		TCG	216	A
		CAG	1009		CGG	503		CTG	456		CCG	766	G
A	Ile	AAT	1	Thr	AGT	22	Asn	ATT	0	Ser	ACT	0	T
		GAT	2105		GGT	982		GTT	1833		GCT	992	C
		TAT	7		TGT	1354	Lys	TTT	1865	Arg	TCT	1038	A
	Met	CAT	4016		CGT	590		CTT	590		CCT	731	G
G	Val	AAC	0	Ala	AGC	0	Asp	ATC	0	Gly	ACC	0	T
		GAC	931		GGC	883		GTC	1963		GCC	2000	C
		TAC	1913		TGC	2341	Glu	TTC	2149		TCC	1244	A
		CAC	349		CGC	323		CTC	321		CCC	522	G

(b) tRNA gene list

tRNA gene List

Firmicutes Bacillus subtilis subsp. subtilis str. 168

HIT: 85 sequence(s).

(c) Advanced Pattern Search against each stem/loop

Advanced Pattern Search

(d) Detail information

The detailed information of tRNA gene sequence

Sequence ID	>C004125
Genome ID	Bsub_168
Search identical group	GO
Phylum/Class	Firmicutes
Species	Bacillus subtilis subsp. subtilis str. 168
Start position on genome	11462
End posion on genome	11538
Amino Acid	Ile
Anticodon	GAT
Upstream region at tRNA start position	aagttaagat
tRNA gene sequence	GGGCCTGTAGCTCAGCTGGTTAGAGCGCACGCCTGATAAGCGTGAGGTCGGTGGTTCGAGTCCACTCAGGCCCACCA
Downstream region at tRNA end position	tcttatataa
Secondary structure (Cloverleaf model)	>C004125 Ile GAT
Intron	Intron
Comment/Decision	comment/decision
Genome/Seq. Info.	[KEGG DAS]

Comment

Input Comment

E-mail

Comment

Comment column

Input this 5 letters

Send Comment

Attention:
When your comment is judged as an irrelevant one by the administrator, your comment will be deleted by the administrator.

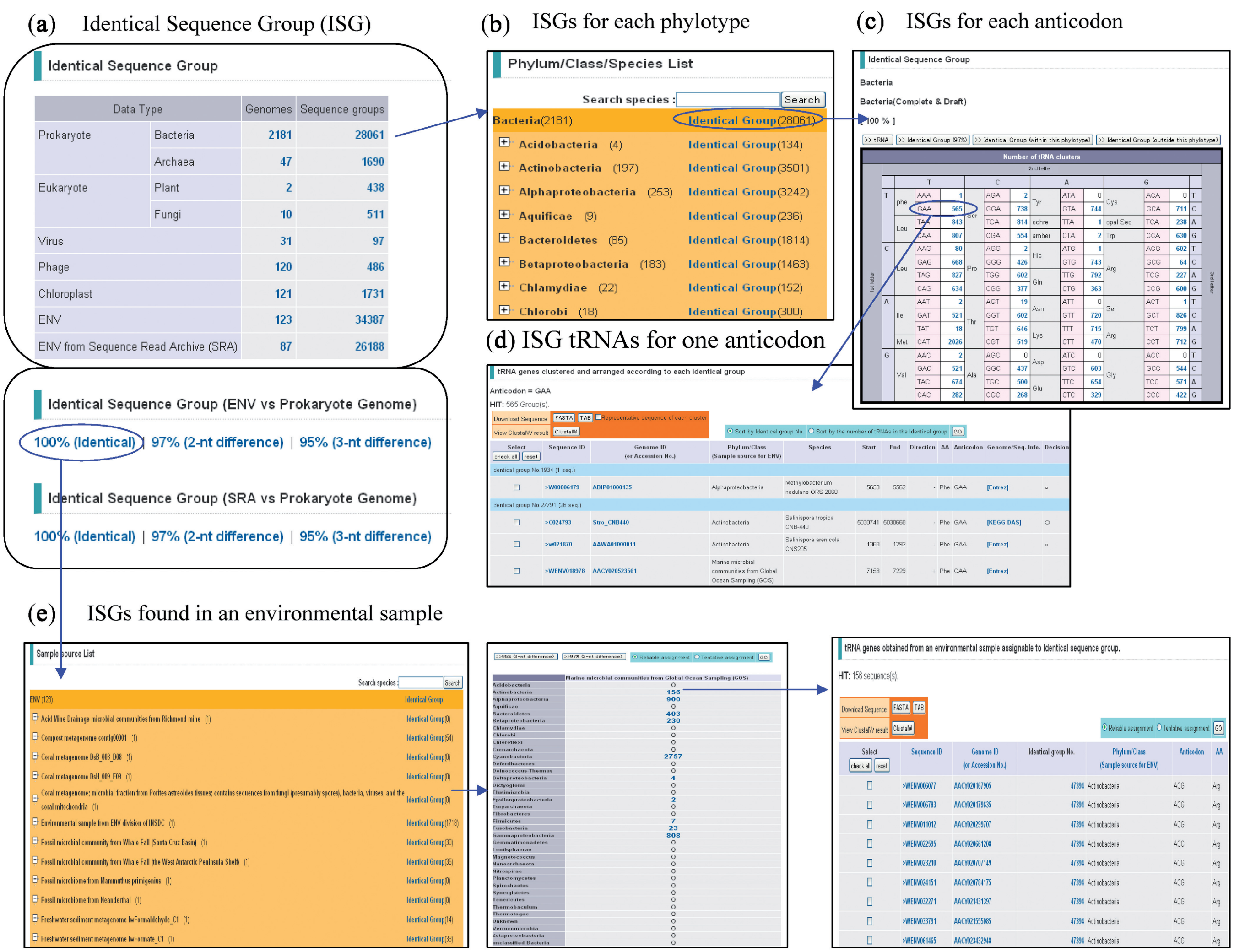
Figure 1. Basic functions of tRNADB-CE. (**a**) Phylum/class/species list and anticodon table, (**b**) tRNA gene list, (**c**) Advanced pattern search against each stem/loop and (**d**) Detail information.

either the reliable tRNA genes or all candidate genes by choosing 'The reliable tRNA genes' or 'All candidate genes'. Criteria used for this manual evaluation were described in detail previously (4). The tRNA genes of Archaea obtained from SPLITSdb (8) were included in this new version. Basic functions of the database were described previously in detail and briefly in Figure 1.

To aim at establishing a very reliable database utilizing collective knowledge in the various experimental fields of tRNA research, we developed a new function for including comments on each of the registered tRNA genes on 'the detailed information of tRNA gene sequence page' (Figure1d). User can add comments by entering e-mail address and password, while we reserve the right to remove irrelevant comments. We hope that the accumulation of user comments will provide annotation of higher quality and that this database will become an information sharing system in the tRNA gene community.

Identical sequence groups and their use as phylogenetic markers for environmental metagenomic sequences

When we conducted the clustering of 124 015 sequences of tRNA genes, except for the 3′-CCA terminal sequence, from prokaryotic genomes by sequence alignment using the CD-HIT (9), we found high phylogenetic preservation of tRNA genes: i.e. a particular tRNA sequence was found only in a particular lineage of phylogenetic groups. We designated here the tRNA group with an identical sequence as 'Identical Sequence Group: ISG' (Figure 2a) and listed the numbers of ISGs for each phylotype (Figure 2b) and for each anticodon (Figure 2c). The tRNAs with one anticodon type were classified and listed according to ISG along with the phylotype information of each tRNA (Figure 2d), and therefore, the range of phylotypes found for each ISG could be examined. If we focused on ISGs composed of more than five sequences, 97.1% of ISGs were conserved at a phylum level, showing most tRNAs to be good phylogenetic markers at least at a phylum level. The ISGs could provide a strategy


(a) Identical Sequence Group (ISG)
(b) ISGs for each phylotype
(c) ISGs for each anticodon
(d) ISG tRNAs for one anticodon
(e) ISGs found in an environmental sample

Identical Sequence Group

Data Type		Genomes	Sequence groups
Prokaryote	Bacteria	2181	28061
	Archaea	47	1690
Eukaryote	Plant	2	438
	Fungi	10	511
Virus		31	97
Phage		120	486
Chloroplast		121	1731
ENV		123	34387
ENV from Sequence Read Archive (SRA)		87	26188

Identical Sequence Group (ENV vs Prokaryote Genome)

100% (Identical) | 97% (2-nt difference) | 95% (3-nt difference)

Identical Sequence Group (SRA vs Prokaryote Genome)

100% (Identical) | 97% (2-nt difference) | 95% (3-nt difference)

Phylum/Class/Species List

Search species : [ ] Search

Bacteria(2181)	Identical Group(28061)
⊞ Acidobacteria (4)	Identical Group(134)
⊞ Actinobacteria (197)	Identical Group(3501)
⊞ Alphaproteobacteria (253)	Identical Group(3242)
⊞ Aquificae (9)	Identical Group(236)
⊞ Bacteroidetes (85)	Identical Group(1814)
⊞ Betaproteobacteria (183)	Identical Group(1463)
⊞ Chlamydiae (22)	Identical Group(152)
⊞ Chlorobi (18)	Identical Group(300)

Identical Sequence Group

Bacteria

Bacteria(Complete & Draft)

[100 %]

>> tRNA | >> Identical Group (97%) | >> Identical Group (within this phylotype) | >> Identical Group (outside this phylotype)

Number of tRNA clusters

1st letter		T			C			A			G			3rd letter
T	phe	AAA	1	Ser	AGA	2	Tyr	ATA	0	Cys	ACA	0	T
		GAA	565		GGA	738		GTA	744		GCA	711	C
	Leu	TAA	843		TGA	814	ochre	TTA	1	opal Sec	TCA	238	A
		CAA	807		CGA	554	amber	CTA	2	Trp	CCA	630	G
C	Leu	AAG	80	Pro	AGG	2	His	ATG	1	Arg	ACG	602	T
		GAG	668		GGG	426		GTG	743		GCG	64	C
		TAG	827		TGG	602	Gln	TTG	792		TCG	227	A
		CAG	634		CGG	377		CTG	363		CCG	600	G
A	Ile	AAT	2	Thr	AGT	19	Asn	ATT	0	Ser	ACT	1	T
		GAT	521		GGT	602		GTT	720		GCT	826	C
		TAT	18		TGT	646	Lys	TTT	715	Arg	TCT	799	A
	Met	CAT	2026		CGT	519		CTT	470		CCT	712	G
G	Val	AAC	2	Ala	AGC	0	Asp	ATC	0	Gly	ACC	0	T
		GAC	521		GGC	437		GTC	603		GCC	544	C
		TAC	674		TGC	500	Glu	TTC	654		TCC	571	A
		CAC	282		CGC	268		CTC	329		CCC	422	G

tRNA genes clustered and arranged according to each identical group

Anticodon = GAA

HIT: 565 Group(s).

Download Sequence: FASTA | TAB | □ Representative sequence of each cluster

View ClustalW result: ClustalW

● Sort by Identical group No. ○ Sort by the number of tRNAs in the Identical group GO

Select (check all / reset)	Sequence ID	Genome ID (or Accession No.)	Phylum/Class (Sample source for ENV)	Species	Start	End	Direction	AA	Anticodon	Genome/Seq. Info.	Decision
Identical group No.1934 (1 seq.)											
□	>W08006179	ABIP01000135	Alphaproteobacteria	Methylobacterium nodulans ORS 2060	5653	5562	-	Phe	GAA	[Entrez]	○
Identical group No.27791 (26 seq.)											
□	>C024793	Stro_CNB440	Actinobacteria	Salinispora tropica CNB-440	5030741	5030668	-	Phe	GAA	[KEGG DAS]	○
□	>w021870	AAWA01000011	Actinobacteria	Salinispora arenicola CNS205	1368	1292	-	Phe	GAA	[Entrez]	○
□	>WENV018978	AACY020523561	Marine microbial communities from Global Ocean Sampling (GOS)		7153	7229	+	Phe	GAA	[Entrez]	

Sample source List

Search species : [ ] Search

ENV (123)	Identical Group
⊟ Acid Mine Drainage microbial communities from Richmond mine (1)	Identical Group(0)
⊟ Compost metagenome contig00001 (1)	Identical Group(54)
⊟ Coral metagenome DsB_003_D08 (1)	Identical Group(0)
⊟ Coral metagenome DsH_009_E09 (1)	Identical Group(0)
⊟ Coral metagenome; microbial fraction from Porites astreoides tissues; contains sequences from fungi (presumably spores), bacteria, viruses, and the coral mitochondria (1)	Identical Group(0)
⊟ Environmental sample from ENV division of INSDC (1)	Identical Group(1718)
⊟ Fossil microbial community from Whale Fall (Santa Cruz Basin) (1)	Identical Group(30)
⊟ Fossil microbial community from Whale Fall (the West Antarctic Peninsula Shelf) (1)	Identical Group(35)
⊟ Fossil microbiome from Mammuthus primigenius (1)	Identical Group(0)
⊟ Fossil microbiome from Neanderthal (1)	Identical Group(0)
⊟ Freshwater sediment metagenome IwFormaldehyde_C1 (1)	Identical Group(14)
⊟ Freshwater sediment metagenome IwFormate_C1 (1)	Identical Group(33)

>>95% (3-nt difference) | >>97% (2-nt difference) | ● Reliable assignment ○ Tentative assignment GO

	Marine microbial communities from Global Ocean Sampling (GOS)
Acidobacteria	0
Actinobacteria	156
Alphaproteobacteria	900
Aquificae	0
Bacteroidetes	403
Betaproteobacteria	230
Chlamydiae	0
Chlorobi	0
Chloroflexi	0
Crenarchaeota	0
Cyanobacteria	2757
Deferribacteres	0
Deinococcus-Thermus	0
Deltaproteobacteria	4
Dictyoglomi	0
Elusimicrobia	0
Epsilonproteobacteria	2
Euryarchaeota	0
Fibrobacteres	0
Firmicutes	7
Fusobacteria	23
Gammaproteobacteria	808
Gemmatimonadetes	0
Lentisphaerae	0
Magnetococcus	0
Nanoarchaeota	0
Nitrospirae	0
Planctomycetes	0
Spirochaetes	0
Synergistetes	0
Tenericutes	0
Thermobaculum	0
Thermotogae	0
Unknown	0
Verrucomicrobia	0
Zetaproteobacteria	0
unclassified Bacteria	0

tRNA genes obtained from an environmental sample assignable to Identical sequence group.

HIT: 156 sequence(s).

Download Sequence: FASTA | TAB

View ClustalW result: ClustalW

● Reliable assignment ○ Tentative assignment GO

Select (check all / reset)	Sequence ID	Genome ID (or Accession No.)	Identical group No.	Phylum/Class (Sample source for ENV)	Anticodon	AA
□	>WENV006077	AACY020167905	47394	Actinobacteria	ACG	Arg
□	>WENV006783	AACY020179635	47394	Actinobacteria	ACG	Arg
□	>WENV011012	AACY020299707	47394	Actinobacteria	ACG	Arg
□	>WENV022595	AACY020661208	47394	Actinobacteria	ACG	Arg
□	>WENV023210	AACY020707149	47394	Actinobacteria	ACG	Arg
□	>WENV024151	AACY020784175	47394	Actinobacteria	ACG	Arg
□	>WENV032271	AACY021431397	47394	Actinobacteria	ACG	Arg
□	>WENV033791	AACY021555085	47394	Actinobacteria	ACG	Arg
□	>WENV061465	AACY023432948	47394	Actinobacteria	ACG	Arg

Figure 2. List and search for (**a**) ISG, (**b**) ISGs for each phylotype, (**c**) ISGs for each anticodon, (**d**) ISG tRNAs for one anticodon and (**e**) ISGs found in an environmental sample.

to select reliable phylogenetic markers. At the genus level, ~60% of ISGs were conserved, showing there may exist good genus-specific markers. By combining the data provided by this database with other detail knowledge of a particular tRNA obtained by experiments or from literature, users may get useful phylogenetic markers (e.g. genus-specific markers) by themselves.

Interestingly, among 154 455 tRNA genes found in metagenomic sequences derived from environmental samples, 35 739 tRNA genes (23%) were identical in sequence to genes from known prokaryotes. Using tRNAs found in an environment sample that were assigned to ISGs, we could predict microbial community structures in an environmental ecosystem at least at a phylum level (Figure 2e). The database has also a function to search for sequences with 97 or 95% sequence identify (2- or 3-nt difference, respectively) (Figure 2a). By using tools in the database and the specific markers found by users (e.g. genus-specific markers), users can independently clarify microbial populations in an ecosystem. This strategy can be applied even to data of short sequences obtained using next-generation sequencers, such as SRA in NCBI. In metagenomic analysis by a next-generation sequencer where the length of sequences obtained is short, phylogenetic characterization of the short sequences was particularly difficult using existing methods, except for sequences derived from dominant species or sequences unambiguously mapped on a known sequenced genome. Because tRNA genes are searchable even from short genomic fragments of around 100 bases, tRNA genes should become one of the most effective means for identifying the microbial populations in an ecosystem when analyzing a vast number of metagenomic sequences obtained by next-generation sequencers.

ACCESS TO THE DATABASE

tRNADB-CE can be accessed freely from http://trna.nagahama-i-bio.ac.jp.

FUNDING

Integrated Database Project and a Grant-in-Aid for Publication of Scientific Research Results from the Ministry of Education, Culture, Sports, Science and Technology, Japan. Funding for open access charge: Integrated Database Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Amann,R.I., Ludwig,W. and Schleifer,K.H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.*, **59**, 143–169.
2. Delong,E.F. and Karl,D.M. (2005) Genomic perspectives in microbial oceanography. *Nature*, **437**, 336–342.
3. Denef,V.J., Mueller,R.S. and Banfield,J.F. (2010) AMD biofilms: using model commuities to study evolution and ecological complexity in nature. *ISME J.*, **4**, 599–610.
4. Abe,T., Ikemura,T., Ohara,Y., Uehara,H., Kinouchi,M., Kanaya,S., Yamada,Y., Muto,A. and Inokuchi,H. (2009) tRNADB-CE: tRNA gene database curated manually by experts. *Nucleic Acids Res.*, **37**, D163–D168.
5. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
6. Laslett,D. and Canback,B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.*, **32**, 11–16.
7. Kinouchi,M. and Kurokawa,K. (2006) tRNAfinder: A software system to find all tRNA genes in the DNA sequence based on the cloverleaf secondary structure. *J. Comput. Aided Chem.*, **7**, 116–126.
8. Sugahara,J., Kikuta,K., Fujishima,K., Yachie,N., Tomita,M. and Kanai,A. (2008) Comprehensive analysis of archaeal tRNA genes reveals rapid increase of tRNA introns in the order thermoproteales. *Mol. Biol. Evol.*, **25**, 2709–2716.
9. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.