EasyCGTree

An Easy Tool for Constructing Core-Gene Tree

Version 3

Dao-Feng Zhang

Update on August 23rd, 2021

Any suggestion or questions, please contact: *zdf1987@163.com*

Contents

1. What is EasyCGTree?

EasyCGTree is a Perl script, developed to construct **genome-based Maximum-Likehood (ML) phylogenetic tree,** by taking microbial genomes or proteomes in fasta format and reference amino acid sequences of a set of core genes as input data. EasyCGTree includes two approaches to infer phylogeny: **supermatrix** and **supertree**. An alternative approach attached to supermatrix is also helpful and memory-saving when performing intraspecific analysis with large size of input data, that is using the single nucleotide polymorphisms (SNPs) within the concatenation core gene sequence to infer phylogeny (simply termed **cgSNP approach**). It has integrated all the steps needed between the input data and the resulted tree file into one Perl script, which would make it easier to infer a core-gene tree. Furthermore, intermediate data of an EasyCGTree run can be directly used as input data of many other applications.

2. How to install EasyCGTree?

There is no necessary to compile **EasyCGTree** after uncompressing the downloaded package (https://github.com/zdf1987/EasyCGTree), rather than install Perl language environment. Several extra programs would be invoked in **EasyCGTree**. Users need to install and configure the extra programs only in case that they want to use the latest version. Here is a list of the extra programs:

Windows OS users (Testes under Windows 10):

1) blast+ (not blast) package: https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/. After installation, move blastp, tblastn, and makeblastdb into the folder "bin", or set the full path in command line (see ***Section 6***).

2) ActivePerl: https://www.perl.org/get.html

3*) FastTree: http://www.microbesonline.org/fasttree/#Usage

4*) muscle: http://www.drive5.com/muscle/

5*) consense: https://evolution.genetics.washington.edu/phylip.html

*: a version of muscle, consense and FastTree had been included in the **EasyCGTree** package.

Linux OS users (Tested under Ubuntu):

1) blast+ package: https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/. After

installation, move blastp, tblastn, and makeblastdb into the folder "bin", or set the full path in command line (see ***Section 6***).

2) Perl: https://www.perl.org/get.html

3*) FastTree: http://www.microbesonline.org/fasttree/#Usage

4*) Clustal Omega. http://www.clustal.org/

5*) consense: https://evolution.genetics.washington.edu/phylip.html

*: a version of FastTree, consense and clustalo had been included in the **EasyCGTree** package for Linux version. ***Users need to use "chmod +x filename" within folder 'bin' to make them executable.***

EasyCGTree *DO NOT* support MAC OS currently.

3. Input data

As mentioned above, genome/proteome data (**Genome data**; see ***Section 3.1***) in fasta format and reference amino acid sequences of a set of core genes (**Query data/Reference data;** see ***Section 3.2*** and ***3.3***, respectively) are required to run **EasyCGTree**. **Query data** are used in BLAST search against **Genome data** (gene calling) by using blast+ software; related homologs are extracted from the Genome data based on the BLAST results; and extracted data, together with **Reference data**, are subsequently aligned (by using muscle or clustalo), trimmed and concatenated to generated input data for FastTree.

3.1 Genome data (Genomes/proteomes in fasta format)

The fasta files are required to be uncompressed and a specialized name (unique, no spacer, no "_", and ending with ".fas"). ***Don't worry when you have dozens of genomes. There is a Perl script "formatGenomes.pl" to do this laborious job*** (see ***Section 4.1.1***). The user just needs to gather the genome sequences.

3.2 Core-gene sets for gene calling (Query data)

The set of genes determines that which genes will be extracted from the genomes/proteomes to infer a core-gene tree. How to define the core genes of the data set of interest? The best way is to define the pan-genome first. Nevertheless, if a researcher just wants a core-gene tree, pan-genome analysis is not the optimal way. Maybe, retrieving a core gene set from previous studies or public databases is a better choice.

An optional way to retrieve a core gene set is to download from the Genome

Taxonomy Database (GTDB) server (https://gtdb.ecogenomic.org/). GTDB defined a gene set named bac120 that includes 120 ubiquitous single-copy genes across the domain *Bacteria*, and a gene set named ar122 that includes 122 ubiquitous single-copy genes across the domain *Archaea*. Users just need to download all the marker genes of *Bacteria*, *Archaea*, or both as wishes, and related taxonomy lists. ***There is a Perl script "GetReferencFromGTDB.pl" in our package to help extract reference sequences*** (see ***Section 4.1.2.1***).

Gene sets can be also selected from those mentioned in the following ***section 3.3*** of this manual. Gene sets mentioned in this part and ***section 3.3*** should ***have the same gene numbers and contain the same set of homologous genes.***

**Because of lower accuracy in gene calling by using nucleotide sequence, EasyCGTree requests protein sequence used for homologous gene searching.

3.3 Pre-prepared core gene sets (Reference data) *(optional)*

Some users may already have a pre-defined core-gene set of some genomes and probably a tree based on these genomes and this gene set or you have finished an EasyCGTree run previously. Now, you want to include some new genomes into the core-gene tree. In this case, it will save much time by commanding a type "A3" run (see ***Section 4.2***). These gene sets should have the same gene numbers and contain the same set of homologous genes as those mentioned in ***Section 3.2***. The files of gene sets should be fasta-formated and have a specialized name (unique, no spacer, no "_", and ending with ".fas"). **You can also get the 'Reference data' directly from the 'TEM4_GeneSeqs' folder of a previous run** (see ***Section 5***).

**Only gene sets of protein sequence is allowed.

4. Run EasyCGTree

4.1 Preparations

4.1.1 Genome data (Genomes/proteomes)

Gather the genomes/proteomes of fasta format into a folder, name them in English style. ***Copy the folder into the directory of EasyCGTree (e.g. D:/EasyCGTree).***

Name the folder whatever you like, it will be referred as "MyGenomes" as an example in the following steps. Users must ensure the data is in fasta format.

****Formatted names excluding extension (should be '.fas') will be the labels**

present on the tree tips.

4.1.2 Query data (used for gene calling)

***** Notably, protein sequence is requested.***

4.1.2.1 bac120/ar122 core-gene set

1) Create a folder (e.g. GTDBdata) in directory wherever you like, and name it.
2) Download data from GTDB (https://gtdb.ecogenomic.org/) and put them into the folder "GTDBdata".

   Download the marker genes of *Bacteria*, *Archaea*, or both as you wish. Taking release95.0 as an example: ar122_taxonomy_r95.tsv (https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/ar122_taxonomy_r95.tsv) and ar122_marker_genes_reps_r95.tar.gz (https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/genomic_files_reps/ar122_marker_genes_reps_r95.tar.gz) for *Archaea*; or bac120_taxonomy_r95.tsv (https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/bac120_taxonomy_r95.tsv) and bac120_marker_genes_reps_r95.tar.gz (https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/genomic_files_reps/bac120_marker_genes_reps_r95.tar.gz) for *Bacteria*.

   Marker genes of all genomes included in this release is available at: https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/genomic_files_all/

   Uncompress the files and move the "faa" folder into "GTDBdata".
3) Copy GetReferencFromGTDB.pl into the folder "GTDBdata".
4) Assume that the bac120 gene sets of the genus *Bacillus* was wanted.

   Change the working directory to the parent folder "GTDBdata" and type:

```
> perl GetReferencFromGTDB.pl bac120_taxonomy_r95.tsv faa g__Bacillus
```

   The bac120 genes of each genomes, belonging to *Bacillus* and included in GTDB, will be extracted and written into a file named with the species label and in the format as descript in ***Section 4.1.2.2***. A log file (.csv) named with the taxon specified will be created to report the included genomes and related details (not all the genes of bac120/ar122 could be found in the genomes included in GTDB). The "g" means genus, and the users can specify any taxa ranging from genus to phylum with a label of "s, g, f, o, c, p", respectively. Please note, there are two "_", not one.

If you specify a higher taxon or a large genus, you will get a lot of bac120 gene sets. In most instances, one gene set for a genus is feasible, because the divergence within a genus is limited for bac120/ar122 genes (https://gtdb.ecogenomic.org/). However, more gene sets used in **EasyCGTree** will increase the accuracy of the gene-calling. ***In contract, too much gene sets will slow down the running speed. We recommend to use gene sets that meets the following criteria:*** **a)** each gene set contain all the genes (120 or 122) of bac120/ar122; **b)** the number of gene sets between M/60 and M/15 be used in following analysis (M means the number of members of a genus); **c)** and the species of these gene sets should be widespread in the genus tree based on 16S rRNA gene.

Under Linux OS, there is a perl script "GetRepRef.pl" that can help select representatives of the gene sets from DTDB.

5) Gather the selected gene sets in a folder (e.g. query), and ***move it into the directory of EasyCGTree.***

4.1.2.2 Personal core-gene set

The users can use personal core-gene sets, but ensure that: **a)** ***they contain no paralogs;*** **b)** ***each gene is labeled in a format of XXX_XXX_XXX_geneSymbol and is unique within a gene set and among gene sets;*** **c)** ***they have the same gene numbers; they contain the same set of homologous genes.*** The latter two are recommended but not imperative if the user was well-known about how **EasyCGTree** works. For gene labels, we recommend to use "generic name"_"specific epithet"_"strain number"_"geneSymbol". The separator "_" is not allowed in each of the four divisions. For running **EasyCGTree** correctly, ***the position of gene symbol is coercive, and gene symbol should be kept consistent among homologs in all gene sets.*** The "generic name", "specific epithet" and "strain number" are not coercive, and can be set waywardly (e.g. 1_1_1_ geneSymbol).

4.1.3 Reference data (used for tree inference) (*optional*)

***** Notably, protein sequence is requested.***

This data should be prepared for a type "A3" run (see ***Section 4.2***). These gene sets should have the same gene numbers and contain the same set of homologous genes as those mentioned in ***Section 3.2*** **and** ***4.1.2***. ***Notably, these gene sets should be gathered into a folder (e.g. Reference)*** and be prepared ***in the format as descript in Section 4.1.2.2. File names excluding extension (should be '.fas') will be the labels present on the tree tips.***

4.1.4 Format the names of Genome and Reference data files (*recommended*)

The file names of Genome and Reference data should have a specialized name

(unique, no spacer, no "_", and ending with ".fas"), but in most time, they do not meet the requirements. So, they need formatting. There is a perl script "formatGenomes.pl" that can help do this laborious job.

From now on, users need to run "cmd.exe" for Windows or "Terminal" for Linux before execute the Perl scripts in **EasyCGTree**.

Change the working directory to the parent folder of the folder including the Genome and Reference data (e.g. MyGenomes and Reference).

For Windows users: (press the key "Enter" when finishing a line)

```
> d:
> cd ./EasyCGTree
> perl formatGenomes.pl MyGenomes
> perl formatGenomes.pl Reference
```

4.2 Ready to Run EasyCGTree

With the input data prepared correctly, it is very easy to run **EasyCGTree**. ***Ensure that the two (or three) folders (Genome, Query and even Reference data) are in the home directory*** of **EasyCGTree**, change the working directory to that of **EasyCGTree** and type:

Perl EasyCGTree.pl [options]

The options do not need to be ordered. For example:

```
> perl EasyCGTree.pl -input myGenomes -query query -seq_type nucl
> perl EasyCGTree.pl -input myGenomes -query query -seq_type prot -mode A3 -reference Reference
```

Then, you will get a tree (named after the name of the folder containing the genomes, e.g. myGenome.tree) in Newick format and many files/folders generated during running the script.

4.2.1 Essential setting

-input <String>

Input data (genome/proteome) directory

-query <String>

Directory of protein sequence data for gene calling

-seq_type <String, 'nucl', 'prot'>

nucl/prot determines the sequence type (protein or nucleotide) used for tree inference. Setting nucl means the input data mentioned in ***Section 3.1*** **and** ***4.1.1*** should be nucleotide sequence, while setting prot means the input data should be protein sequence.

4.2.2 Optional setting

-mode <String, 'A', 'A1', 'A2', 'A3'>

A/A1/A2/A3 means running mode:

Table 1 Allowed combinations of command line settings and input data

Input data			Other settings		
-input	-query	-reference	-mode	-seq_type	-tree#
nucl*	prot	not requested	A	nucl	sm, snp
prot				prot	sm, st
nucl*			A1	nucl	not requested
prot				prot	
nucl*			A2	nucl	sm, snp
prot				prot	sm, st
prot		prot	A3	prot	sm, st

sm, supermatrix; st, supertree; snp, single nucleotide polymorphisms.
*, nucl means either genome sequence or proteome in nucleotide sequence.

A, command complete run without input data mentioned in ***Section 3.3*** and ***4.1.3***, and yield all output files (see ***Section 5***);

A1, first part run (yield TEM1-3 folders, query sequences and a log file);

A2, second part run (yield TEM4-9 folders, a log file, concatenated sequences and a tree);

A3, complete run with input data mentioned in ***Section 3.3*** and ***4.1.3***, conflicting to '-seq_type nucl'.

The division of A1 and A2 mode is aimed to save running time when users need to optimize parameters (see ***Section 6*), because BLAST search (Fig. 1, step 2) is time-consuming when the gene sets used in gene calling and/or genomes increase (see ***Section 4.1.2.1***).

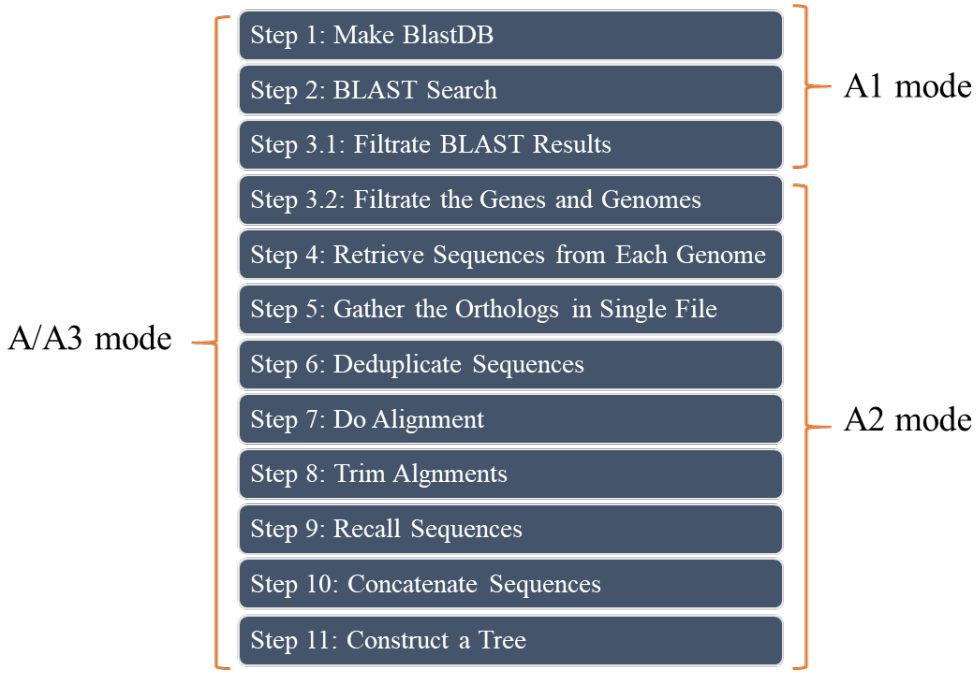

Fig. 1 Program flowchart of EasyCGTree for a complete run

-tree <String, 'sm', 'st', 'snp'>

Specify the approach (sm, supermatrices; st, supertree; or snp, cgSNP) used for tree inference. The default is 'sm'. The 'st' and 'snp' will determine that the '-gene_cutoff' will be set as '1.0'.

-reference <String>

Directory of protein sequence data used in tree inference.

-thread <Int>

Specify the number of threads used by the blast+ programs under Windows OS, or by blast+ and clustalo programs under Linux OS. FastTreeMP will use all the threads available under Linux OS. Please set it depending on your computer and it should be integer. The default is 2.

-blast_dir <String>

Specify the location of the program tblastn, e.g. /share/bin/. The default is '.\\bin\\' for Windows version and './bin/' for Linux version. If users want to use blast+ applications of which the path was added to the environment variable, just set '-blast_dir' with no value.

-iden_cutoff <Int, 50...100>

Specify the cutoff (50-100) for filtering the BLAST results. The bac120 and ar122 consist of highly conserved house-keeping genes. However, lower cutoff might introduce wrong signals when inferring ML trees if some core genes were not well

assembled or sequenced in some genomes. The default setting is 50.

-gene_cutoff <Decimal, 0.5...1>

Specify the cutoff (0.5-1) for omitting low-prevalence gene. It means that only genes present in more than the proportion (will be 80% when use the default 0.8) of the genomes determined by setting '-genome_cutoff (see next paragraph) will be used in following analysis. Lower cutoff will keep more genes used for inference of supermatrix approach, and missing genes in some genomes will be treat as gaps. The 'st' and 'snp' of '-tree' option will set the '-gene_cutoff' as '1.0' automatically.

-genome_cutoff <Decimal, 0.5...1>

Specify the cutoff (0.5-1) for omitting low-quality genomes. It means that only genomes harboring more than the proportion (will be 80% when use the default 0.8) of all the core genes will be used to infer a ML tree. Lower cutoff will keep more genomes used to infer ML tree, but the genes used will be fewer.

-help (Optional)

Display help message

#More comments on '-genome_cutoff' and '-gene_cutoff'

The two parameters may be confusing to some users. In most situations, leave them alone and the default setting will be OK for a run involving limited number of genomes (<30), especially when the genomes are all from type strains. However, if much more genomes are used, it is inevitable that some genes are not detected in some genomes, and the number of common genes decrease. In fact, many genomes collected in the representative database of GTDB contain an incomplete bac120/ar122 gene set, and users will find it when gather reference gene sets of bac120/ar122 (see ***Section 3.2***). It is competing between increasing common genes and increasing genomes of interest when inferring a core-gene tree.

These two parameters are expected to compromise this contradiction by excluding low-prevalence genes and low-quality genomes. Users should be informed that these two parameters determine the lower limit on the quality of input genomes (the selected gene set is determined by the selected genomes). Lower cutoffs make no sense if the genomes are all of high quality. For example: users will always get 120/122 genes to infer a tree if all the genomes contain a complete bac120/ar122 gene set, no matter what is specified (0.5-1).

5. Output files

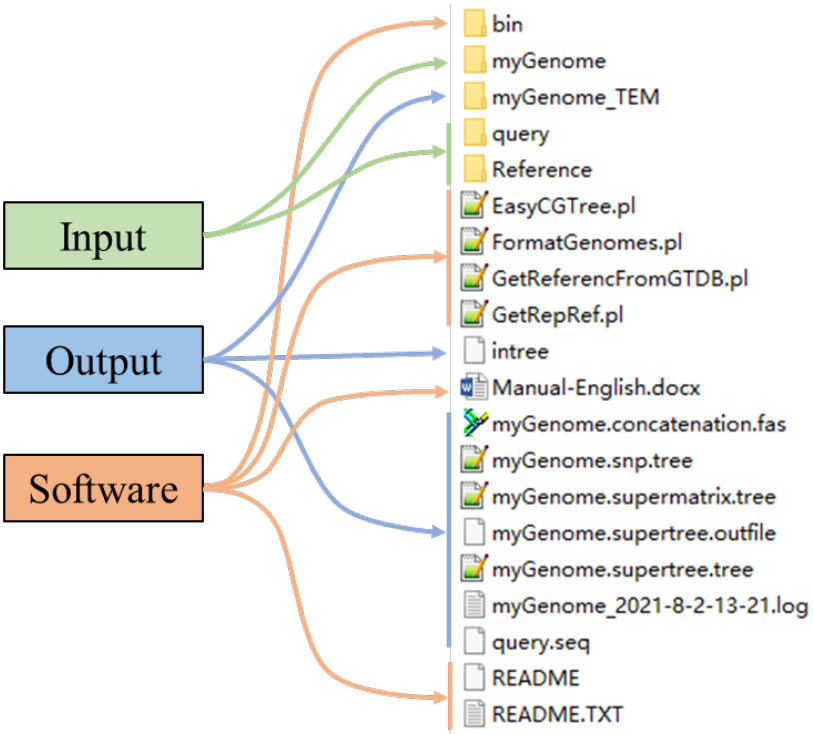

Fig. 2 Content of the EasyCGTree home directory after running

1) Phylogenetic trees of Newick format. Users could display it by using FigTree, MEGA, iTOL or other tree viewers. Take the example in Fig. 2:

myGenome.supermatrix.tree: generated by supermatrix approach (set '-tree sm ' or default).

myGenome.supertree.tree: generated by supertree approach using consense program in phylip software package (set '-tree st'). **iTOL is recommended to display this tree correctly.**

myGenome.snp.tree: generated by supermatrix approach (set '-tree snp ').

myGenome.supertree.outfile: generated by supermatrix approach using consense program in phylip software package (set '-tree st'). It records the details to generate '**myGenome.supertree.tree**' from' **intree**' by consense program.

intree: includes the trees constructed from each selected gene. It was generated from the tree files in 'myGenome/TEM11_geneTrees' (see below).

2) A fasta-format file (myGenome.concatenation.fas) **contains concatenated sequences of wanted genes extracted from each genome, which is used by FastTree software to infer a ML tree.** This file is named after the genome folder with addition of ".concatenation.fas". This file is generated from the folder TEM9_ReCallAln (see below). This file is also accessible to other phylogeny-inferring software, e.g. IQ-Tree and RaxML.

3) A log file (myGenome_2021-8-2-13-21.log) **named after the genome folder**

and the starting time. It records the detailed information about the run, such as: how many/which genes was included or excluded; how many/which genomes was included or excluded.

4) A file (query.seq) **contains the query sequences used for BLAST search.** It is named after the folder containing reference gene sets and ends with ".seq".

5) A folder (myGenome_TEM) **named after the genome folder and ending with "_TEM".** Subfolders' names in this folder are universal to all the runs of EasyCGTree. It contains the following folders.

a. TEM1_blastDB: containing the databases created by makeblastdb program for each genome.

b. TEM2_blastOUT: containing BLAST results of each genome in format 6 by tblastn program.

c. TEM3_blastOUT_S: containing screened BLAST results of TEM2_blastOUT. If two or more references gene sets were used, the best result for each gene will be selected based on the bitscore. Results below the identity cutoff will be dropped and the cutoff (default 50%) could be set in parameters.txt (**see *Section 6***).

d. TEM4_GeneSeqs: containing gene sequences retrieve from each genome based on the results of TEM3_blastOUT_S. **Files of protein sequence could be directly used as Query (see *Section 3.2* and *4.1.2*) or Reference data (see *Section 3.3* and *4.1.3*) in another EasyCGTree run.**

e. TEM5_GeneCluster: containing files gathering homologs from different genomes. One gene (cluster) was gathered in a single file.

f. TEM6_DedupGeneCluster: containing files including non-redundant sequences. These files are descendant of those in TEM5_GeneCluster. Genes with the same sequence are deduplicated, and the correspondence between representatives and analogues are recorded in files of TEM60_DedupList.

g. TEM60_DedupList: see TEM6_DedupGeneCluster.

h. TEM7_Alignment: containing fasta-format files of alignments created based on files in TEM6_DedupGeneCluster.

i. TEM8_AlignTrimmed: containing fasta-format files of trimmed alignments created based on files in TEM7_Alignment.

j. TEM9_ReCallAln: containing fasta-format files of trimmed alignments of all the genomes used to infer ML tree. These files are generated based on the files in TEM8_AlignTrimmed and TEM60_DedupList.

k. TEM11_geneTree: containing tree files constructed from the aligned gene clusters in TEM9_ReCallAln.

6. Usefull scripts

6.1 formatGenomes.pl

Used to format names of files in a directory. The name will be revised to be with no spacer, no "_", and ending with ".fas".

Usage:

```
> perl formatGenomes.pl Dir_name
```

Find more information in ***Section 3.1*** and ***4.1.4*****.**

6.2 GetReferencFromGTDB.pl

Used to retrieve bac120/ar122 gene sets within a taxon (from genus to phylum) from a local GTDB database. **Taking the genus *Bacillus* as example** (option 'g__Bacillus'), the bac120 genes of each genomes, belonging to *Bacillus* and included in GTDB, will be extracted and write into a file named with the species label and in the format as descript in ***Section 4.1.2.2***. A log file (.csv) named with the taxon specified will be created to report the included genomes and related details (absence/presence of bac120/ar122 genes).

Usage example:

```
> perl GetReferencFromGTDB.pl bac120_taxonomy_r95.tsv faa g__Bacillus
```

'bac120_taxonomy_r95.tsv': table of GTDB taxonomy for all bacterial genomes assigned to a GTDB species cluster (replace it according to your GTDB release).

'faa': name of the folder including protein sequences of marker genes of genomes collected by the GTDB release.

'g__Bacillus': "g" means genus, and the users can specify any taxa ranging from genus to phylum with a label of "g, f, o, c, p", respectively; "Bacillus" is the taxon name of interest, which is **case sensitive**. Please note, there are two "_", not one.

Find more information in ***Section 4.1.2.1***.

6.3 GetRepRef.pl (for Linux OS)

Used to decrease the number of the gene sets by selecting representatives. It needs to use clustalo program, so users should **keep 'GetRepRef.pl' and 'bin' folder together to use it**.

Usage:

```
> perl GetRepRef.pl -input myGeneSets
```

Options:

-input <String>

Input data directory (essential option)

-gene_number <Int>

Number of genes randomly selected for distance calculation (6-24, default: 12; smaller number, less time cost).

-diverg_cutoff <Decimal, 0…0.5>

Minimum distance allowed to screen representative data sets. [default: 0.05]

-thread <Int>

Number of threads to be used by 'clustalo'. [default: 2]

-local_dist

Distance matrix file generated in previous run. It will be used if it exists, so there is no need to specify it in the command line if it is still name after the input directory. This option can save time of doing alignment, and the file needs to be named after the input directory, for example: input.fas.dist).

-help

Display help message.

Find more information in ***Section 4.1.2.1***.

7. Hardware Requirement

A normal PC is good enough to run EasyCGTree, because clustalo, FastTree, and muscle are fast and approachable. However, the speed depends on the size of the input data. When Genome data <100, Query data < 10, and gene number in Query data < 200, a PC will finish the analysis within several hours. If bigger size input data (especially Genome data) was used, the version of Linux OS and powerful PC/server are recommended, because FastTree and clustalo under Linux support multi-threads (muscle and FastTree under Windows only support single thread).

8. Performance

Table 2 Running performance with test data 1

Test data information		Taxonomy: Erythrobacteraceae; Proteomes: 2; Query data sets: 2; Genes in each query set: 288; Reference data sets: 75			
Machine configuration		OS: Windows 10; Memory: 16G; CPU: i7-9700		OS: Ubuntu 18.04 LTS; Memory: 128G; CPU: 2x Xeon E5-2680 v4	
Options	-input	E2K3			
	-query	query			
	-seq_type	prot			
	-mode	A3			
	-reference	Reference			
	-thread	2(default)		50	
	-tree	sm (default)	st	sm (default)	st
Command line*		1	2	3	4
Running time		21 m 26 s	29 m 21 s	15 m 43 s	19 m 36 s

* 1, perl EasyCGTree.pl -input E2K3 -query query -seq_type prot -mode A3 -reference Reference

2, perl EasyCGTree.pl -input E2K3 -query query -seq_type prot -mode A3 -reference Reference -tree st

3, perl EasyCGTree.pl -input E2K3 -query query -seq_type prot -mode A3 -reference Reference -thread 50

4, perl EasyCGTree.pl -input E2K3 -query query -seq_type prot -mode A3 -reference Reference -tree st -thread 50

Table 3 Running performance with test data 2

Test data information		Taxonomy: Staphylococcus aureus; Genomes: 601; Query data sets: 2; Genes in each query set: 1413			
Machine configuration		OS: Windows 10; Memory: 16G; CPU: i7-9700		OS: Ubuntu 18.04 LTS; Memory: 128G; CPU: 2x Xeon E5-2680 v4	
Options	-input	S.aureus			
	-query	querySA			
	-seq_type	nucl			
	-mode	A			
	-thread	4(default)		50	
	-tree	snp	sm (default)	snp	sm (default)
Command line*		1	2	3	4
Running time		**FastTree Error: out of memory**	**FastTree Error: out of memory**	9 h 46 m 36 s	15 h 16 m 55 s

* 1, perl EasyCGTree.pl -input S.aureus -query querySA -seq_type nucl -mode A -thread 4 -tree snp

2, perl EasyCGTree.pl -input S.aureus -query querySA -seq_type nucl -mode A -thread 4

3, perl EasyCGTree.pl -input S.aureus -query querySA -seq_type nucl -mode A -thread 50 -tree snp

4, perl EasyCGTree.pl -input S.aureus -query querySA -seq_type nucl -mode A -thread 50

9. FAQ

1. Q: I have a genome with no proteome/CDS sequence released. How to use EasyCGTree to infer a tree from amino sequecnes?

A: You must get the proteome/CDS sequence before start. There is an easy way to get CDS quickly, that is using the online tool CDSeasy (https://bioinfo-mml.sjtu.edu.cn/STEP/STEP_CDSeasy.php). You will get the CDS in a few minutes.

2. Q: Can EasyCGTree build trees of two or three approaches (supertree, supermatrix, and cgSNP) simultaneously?

A: No, it can't. That is because supermatrix approach allow a gene using in tree inferring absent in some genomes (depends on '-gene_cutoff') to maintain more genetic information, while supertree and cgSNP approaches automatically set '-gene_cutoff 1. In another word, different approaches use different genes, so they can't be analyzed simultaneously. If there is no reference data, users can use '-mode A2' to infer a tree of other approaches, which will save some time.

3. Q: Can I use EasyCGTree to define common genes/core genes of a set of genomes by using CDS of one genome as Query data?

A: In theory, it seems to be feasible if '-iden_cutoff', '-genome_cutoff' and '-gene_cutoff' were well used. Nevertheless, as mentioned above, users should make sure that the query data contains no paralogs. Almost certainly, all the bacterial genomes contain paralogs. Our suggestion is that users should further screen the common genes according their function annotation and select those associated with basal metabolism.

4. Why not support Mac OS?

A: Taking "easy to install" and "easy to use" as the philosophy, EasyCGTree is arbitrary in selecting third-party applications that can be used directly by EasyCGTree. Unfortunately, we failed to find valid applications to realize its function.

10. References

Zhang, D.F.; Cui, X.W.; Zhao, Z.; Zhang, A.H.; Huang, J.K.; Li, W.J. *Sphingomonas hominis* sp. nov., isolated from hair of a 21-year-old girl. Antonie Van Leeuwenhoek 2020, 113: 1523-1530, doi:10.1007/s10482-020-

01460-z. (**If you use EasyCGTree, please cite this paper befor the paper describing EasyCGTree is published.**)

Parks, D.H.; Chuvochina, M.; Waite, D.W.; Rinke, C.; Skarshewski, A.; Chaumeil, P.A.; Hugenholtz, P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **2018**, *36*, 996-1004, doi:10.1038/nbt.4229.

Felsenstein, J. 2009. PHYLIP (Phylogeny Inference Package) version 3.7a. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 1989, *5*: 164-166.

Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997**, *25*, 3389-3402, doi:10.1093/nar/25.17.3389.

Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.Z.; Lopez, R.; McWilliam, H.; Remmert, M.; Soding, J., et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **2011**, *7*, doi:10.1038/msb.2011.75.

Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **2009**, *26*, 1641-1650, doi:10.1093/molbev/msp077.

Edgar, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **2004**, 32(5), 1792-1797, doi: 10.1186/1471-2105-5-113.

Zhang, X.M.; Zhang, D.F.; Zhang, Y.L.. Altererythrobacter flava sp. nov., a new member of the family Erythrobacteraceae, isolated from a surface seawater sample. *Antonie Van Leeuwenhoek*, 2021, 114: 497-506, doi:10.1007/s10482-021-01531-9.

Zhang, D.F.; Cui, X.W.; Li W.J.; Zhang, X.M.; Xue, H.P.; Huang, J.K.; Zhang, A.H.. Description of Salinimonas profundi sp. nov. a deep-sea bacterium harboring a transposon Tn6333. *Antonie Van Leeuwenhoek*, 2021, 114:69–81, doi:10.1007/s10482-020-01501-7.