Tutorial of DAFGA
- explained with pmoA dataset (included in test folder) as an example

Requirement:
(1) refseq.gp: reference sequences of a gene of interest retrieved from NCBI non-redundant protein database in genpeptide format (.gp). (given here as *pmoA_refseqs.gp*)

(2) amplicon.fna: If you have multiple libraries, make mapping file and assign samples to multiplex reads using split_libraries.py of QIIME. (given here as *mapping.txt* and *pmoA_amplicons.fna*)

Workflow :
(1) dafga_refDB.py -gp pmoA_refseqs.gp -o refDB -- email *personal_email_address*
(2) dafga_correlation.py -r refDB -o corr
(3) dafga_otus.py -i pmoA_amplicons.fna -o otus -t +1 -c corr
(4) dafga_taxonomy.py -g otus -r refDB -c corr -o taxa

Final outputs to be used for diversity measurments (QIIME):
(1) OTU_mapping.txt in otus folder
(2) taxa_at_*rank*_iden.txt or taxa_at_*rank*_siml.txt in taxa folder
(3) rep_phylo.tre in taxa folder

First two files (OTU_mapping.txt and taxa_at_rank_iden/smil.txt) are converted into OTU table in biom format by make_otu_tables.py of QIIME.
It is subjected to alpha- and beta- diversiy measurement of QIIME along with rep_phylo.tre.

dafga_refDB.py

Description:
DAFGA parses the publicly available sequences of functional genes that are retrieved from the NCBI protein database in genpept format. It excludes all the environmental sequences that are not taxonomically affiliated and subsequently constructs the reference database to be used for taxonomic assignment (saved as "*_ref_seqs.fasta"). DAFGA retrieves the full taxonomic lineage of each reference sequence from the NCBI taxon ID (saved as "*_ID_to_taxonomy.txt"). In addition, it separately collect reference sequences that are described at strain level (saved as "*_strain.fasta") and fetches near full-length 16S rRNA gene sequences (1200-1600 bp in length) of the source organisms from the NCBI nucleotide database (saved as *_strain_16S_rRNAs.fasta").

*Note: **Internet connection is required to execute this command**

[REQUIRED]	
-gb	Reference sequences of a functional gene that are retrieved from NCBI protein database in a .gp format
--email	Email address necessary to use the Entrez Utilities Web Service
-o	Folder where output files will be saved

[OPTIONAL]	
-l	Minimum length of reference sequences (default: 50)

[OUTPUT]	
processed.gp	The processed.gp doesn't include unrecognizable lines by biopython which start with CONTIG, SecStr, and Het.
*_ref_seqs.fasta	A set of functional gene sequences with taxonomy information, which will be used for taxonomic assignment
*_ID_to_taxonomy.txt	Taxonomy information of reference functional gene sequences, which include the NCBI taxon ID and the complete taxonomic lineage
*_strain.fasta	A subset of reference gene sequences containing explicit source information at strain level.
*_strain_16S_rRNAs.fasta	The 16S rRNA gene sequences of strain-level source organisms. They pair with functional gene sequences in the *_source_strains.fasta

dafga_correlation.py

Description:
The pairs of functional and 16S rRNA gene sequences ("*_strain.fasta" and "*_strain_16S_rRNAs.fasta") are used to compute the evolutionary rate(ER) of a functional gene in relation to that of its 16S rRNA gene by pairwise alignments between all the strain level source organisms using the EMBOSS Smith-Waterman alignment tool (saved in folder “fg_split” and “ssu_split”). Outputs of pairwise alignments are plotted on scatter plot comparing functional and 16S rRNA gene sequence divergence between source organisms (correlation_plot.pdf). Taxonomic thresholds of a functional gene are deduced from linear regression analysis of those of 16S rRNA gene sequences corresponding to different taxonomic ranks (phylum: 80%, class: 85%, order: 90%, family: 93%, genus: 95%, species: 97%, and strain: 99%), resulting in "corresponding_similarity_to_16S.txt".

[REQUIRED]	
-r	The output directory of **dafga_refDB.py** that contains 16S rRNA and functional gene sequences of source organsms (*strain_16S_rRNA.fasta and *strain.fasta)
-o	Folder where output files will be saved

[OUTPUT]	
fg_split	Pairwise alignment of functional gene sequences
ssu_split	Pairwise alignment of 16S rRNA gene sequences
correlation_plot.pdf	Correlation plot between functional and 16S rRNA gene sequence divergence
corresponding_similarity_to_16S.txt	Similarity thresholds of a functional gene corresponding to different taxonomic ranks. The values are extrapolated from the linear regression curve in the correlation plot.

dafga_otus.py

Description:
Insertion and deletion errors in next-generation sequencing can often cause a shift in reading frame during translation, thereby inflating diversity of functional gene species. To overcome this problem and to reduce overestimation of diversity, DAFGA uses a two-step procedure for OTU clustering. Amplicon reads, which passed user-defined quality-filtering strategy, are pre-clustered into OTUs defined by high nucleotide sequence identity (option: --nt_id, defalult >97%), using USEARCH (Edgar, 2010). A consensus sequence of each OTU is selected to represent pre-clustered OTUs (saved in folder "preclustering"). The translated consensus sequences are subjected to final OTU clustering using the gene-dependent sequence identity threshold that corresponds to the desired taxonomic rank (user can define with --taxa option). Centroid sequences are selected to represent each final OTU and used for both taxonomic assignment and phylogenetic tree construction. The OTU mapping file is generated by merging pre-clustering and clustering output.

[REQUIRED]	
-i	Amplicon sequences resulting from **split_library.py (QIIME)**
-c	The output directory of **dafga_correlation.py** that contains correlation_plot.pdf and corresponding_similarity_to_16S.txt.
-o	The output directory where the output files will be saved
-t	Starting position for translation in amplicon sequences(+1,+2, or +3) If not sure, do BLASTX with several query sequences to NCBI nr protein database and take the frame value of alignments

[OPTIONAL]	
--gen_code	NCBI codon table for translation [1-6 and 9-16] (Default: 11)
--nt_id	Sequence identity threshold for preclustering using nucleotide sequence [0-1] (default:0.97)
--taxa	Taxonomic rank to be used for OTU clustering. Identity threshold will be taken from corresponding_similarity_to_16S.txt (Default: species)

[OUTPUT]	
preclustering	The output folder of preclustring including out mapping file and representative sequences
clustering	The output folder of clustring including out mapping file and representative sequences
OTU_mapping.txt	The file contains OTU identifiers and amplicon identifiers belonging to each OTU identifier. It is generated from merging preclustering and clustering output.
rep_seqs.fasta	Translated representative sequences of OTUs

dafga_taxonomy.py

Description:
The number of functional gene sequences that are available for the construction of reference databases is limited and biased towards taxa characterized by sequenced genomes. For instance, dependent on the functional marker gene, the number of known sequences deposited in FunGene ranges from 100 to maximum 77,876 (including environmental sequences), while more than three million 16S rRNA sequences are available. In similarity-based taxonomic classification of functional gene sequences, lack of close homologs at low taxonomic ranks, such as genus or species, can often lead to incorrect classification due to high sequence divergence. Therefore, DAFGA takes into account the absolute identity or similarity score in alignment with the best homolog, and performs assignment at the most reliable taxonomic rank by referring to the correlation plot between the functional and 16S rRNA gene sequences.

This script takes "*_ref_seqs.fasta" as an input to construct BLAST database and compare "rep_seqs.fasta" to find the most homologous sequences in the database. It generates sequence alignment profiles of the representative sequences, containing histograms of query coverage and idenity/similarity in alignments (align_prof.pdf). Users can filter poor quality alignment with respect to alignment length of query sequences using --cov option, and they can also specify the lowest taxonomic level of assignment based on idenity/similarity value in alignments (option: --taxa) by referring taxonomic threshold given in "corresponding_similarity_to_16S.txt". Taxonomic assignment and phylogenetic tree of the representative sequences are respectively saved as "*_taxa_assignment.txt" and constructed as rep_phylo.tre using FastTree method.

[REQUIRED]	
-g	The output directory of the **OTU_clustering.py** containing
-r	The output directory of the **reference_db.py** containing
-c	The output directory of the **correlation_plot.py** containing
-o	The output directory where the output files will be saved

[OPTIONAL]	
--evalue	E-value threshold to use in the BLAST search to the reference database (default: 10)
--num_threads	Number of threads(CPUs) to use in the BLAST search (default: 1)
--taxa	Selected taxonomic level for reliable assignment (default: species)
--cov	Minimum query coverage(%) in sequence alignments to reference sequences (default: 30)

--phylo	FastTree methods for tree building (default: FastTree)

[OUTPUT]	
*taxa_assignment.txt	The taxonomy of the most homologous sequences of representative sequence and the alignment scores including query coverage, similarity, identity and reference sequence ID. Query coverage greater than 100% is due to deletion in query sequences.
align_prof.pdf	The alignment profile of representative sequences to the most homologous sequence. It contains histograms of (a) query coverage (%) and (b) absolute identity and similarity value in alignments. You can refer to this profile to determine reliable taxonomic level for taxonomic assignment.
.tre	Phylogenetic tree of representative sequence is constructed