This document contains brief description of each file which is related to the construction of PlasGUN. The scripts and data listed in this document are not part of the PlasGUN software, but may be useful for researcher who want to reproduce or improve PlasGUN. If you have any question when using these scripts and data, please directly contact fangzc@pku.edu.cn.

before2013.tar.gz

This file contains plasmid genomes used to train PlasGUN (released before 2013).

---

after2013.tar.gz

This file contains plasmid genomes used to test PlasGUN (released after 2013).

CDSmat.tar.gz and **gbffmat.tar.gz**

This files contain genome annotation of each plasmid (saved as mat format), which are created by the scripts separateGB.m all_gb2table.m and gb2table.m.

---

up_std_num.m

Matlab script used to calculate the probability that which start codon in the longest coding ORF is the translation initiation site.

---

getSimParameter.m and **MetaSim.m**

Matlab scripts used to generate the short reads.

---

unknown_function_gene.m

Script used to calculate the proportion of genes of uncertain function.

---

train100to400.zip

After decompression, the folder contains data, related results and scripts of GroupS, which are described as follow:

trainParameter.mat and **testParameter.mat**

Generated by the script getSimParameter.m

All_Data1.mat to **All_Data6.mat**

Generated by the script MetaSim.m

simAnnotation.m, **adjust_uncertain_nt.m**, **findPosition.m**, **getORF.m**, **isbetweenMinMax.m**, **isCoding.m** and **nt2num.m**

Scripts used to extract annotate candidate ORFs.

All_Data1_annotation.mat to **All_Data6_annotation.mat**
Generated by the script simAnnotation.m

create_data.m
Script used to generate the mathematical model of candidate ORFs.

codon_onehot1.mat to **codon_onehot2.mat**, **TIS_onehot1.mat** to **TIS_onehot6.mat**, **length1.mat** to **length6.mat**, **orftype1.mat** to **orftype6.mat**, **label1.mat** to **label6.mat** and **test.fna**
Generated by the script create_data.m

create_num.m
Script used to number the sequences in test.fna.

test_num.m
Generated by the script create_num.m.

train.py
Script used to train the neural network.

predict.csv and **model_a.h5**
Generated by the script train.py. The file predict.csv is the prediction of the neural network.

best_HM.m
Script used to calculate the threshold in which the neural network achieves the highest Hm.

prodigal.txt, **metagun.metagun**, **MetaGeneMark.txt**, **FragGeneScane.txt** and **MetaGeneAnnotator.txt**
Output files of the comparative tools.

evaluate_PlasGUN.m, **evaluate.m**, **parse_gff.m**, **parse_MED.m**, and **parse_MetaGeneAnnotator.m**
Scripts used to evaluate each tool.

positive_negative_ration.m
Script used to calculate the ratio of coding and non-coding ORF.

---

train401to900.zip

After decompression, the folder contains data, related results and scripts of GroupL. The description of each file is the same as that in train100to400.zip as mentioned above.

---

cal_AUC.m

Calculate the AUC of PlasGUN.

---

draw_threshold.m and **evaluate_threshold.m**

Scripts used to draw Figure S2.

Szczepanowski_dataset.zip

After decompression, the folder contains data, related results and scripts related to Szczepanowski dataset, which are described as follow.

adjust_addnum.m

Script used to number the sequences in SzczepanowskiEtAl_Insight_JournalBiotech_2008.fna.

output_predicted_ORF.m, **parse_gff.m, parse_MED.m** and **parse_MetaGeneAnnotator.m**

Scripts used to create the faa file that contains the predicted proteins of the gene prediction tool.

FragGeneScane.faa, **MetaGeneAnnotator.txt**, **MetaGeneMark.faa**, **metagun.metagun**, **PlasGUN_05242.txt** and **prodigal.faa**

Prediction results of each tool.

metagun.faa, **PlasGUN_05242.faa** and **MetaGeneAnnotator.faa**

Generated by script output_predicted_ORF.m.

psi_blast.zip

Contains blast results of each tool.

evaluate_blast.m

Script used to evaluate the blast results of each tool.

SzczepanowskiEtAl_bowtie2.sam

Output of bowtie2, which mapped the short reads to RefSeq plasmid genomes database.
