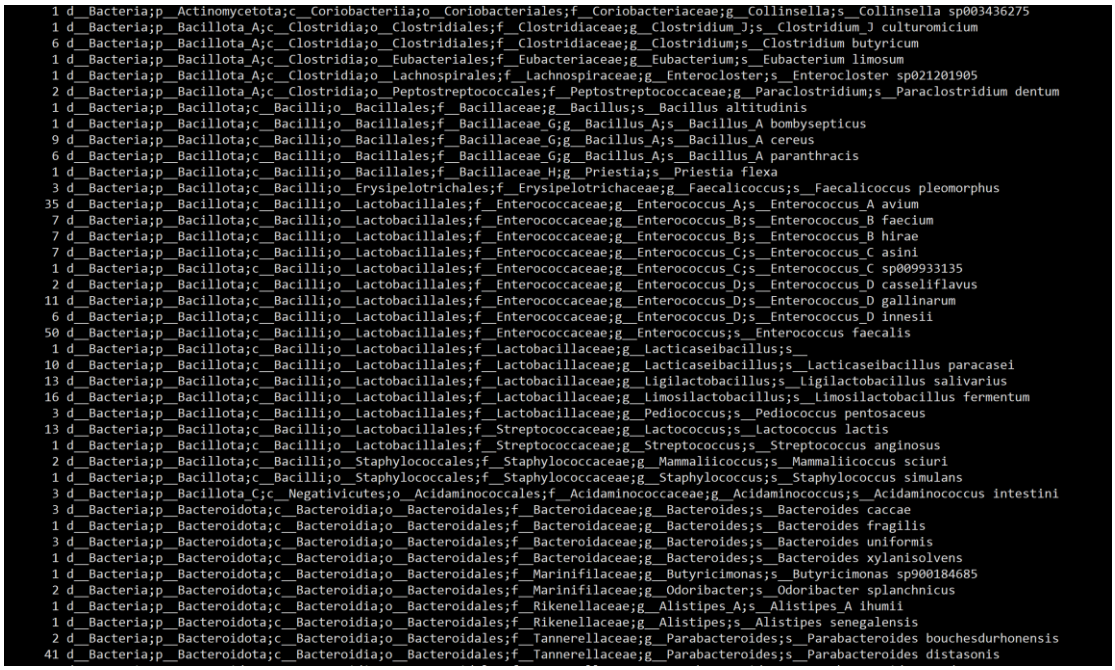
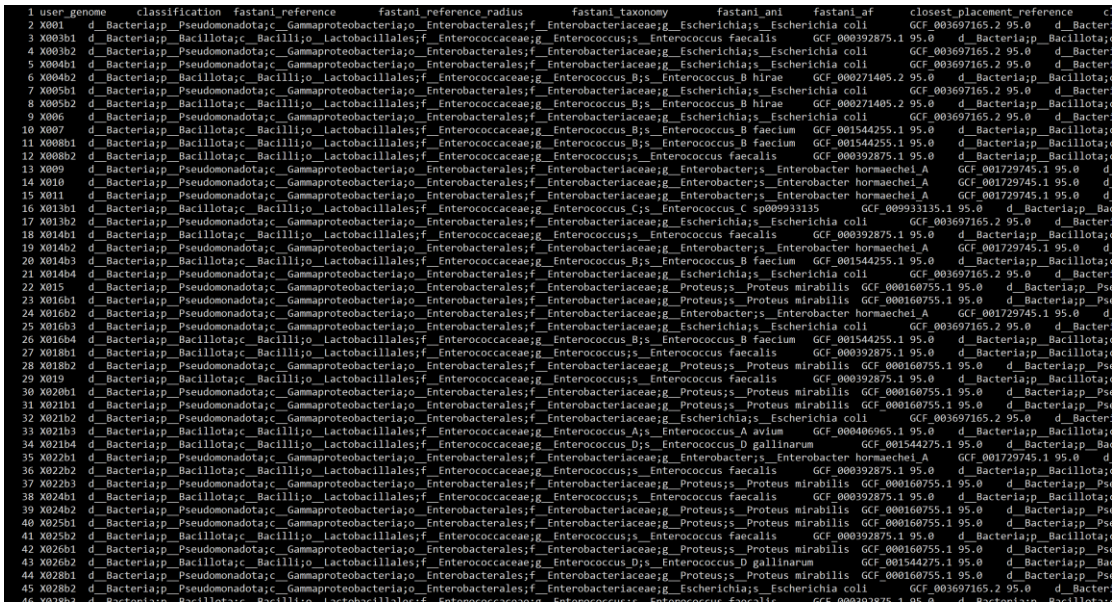
roary_split.py 的目标是根据数量超过特定阈值的每个物种，分别创建文件夹。在每个文件夹内，生成一个 txt 文件，记录该物种的所有 MAGs，并在物种文件夹下创建指向 Prokka 注释的 gff 文件的软链接。输入文件包括 species.txt 和 tax.bac120.summary.tsv 两个文件。

其中，species.txt 记录了每个组装的 MAG 的数量：



```
  1 d__Bacteria;p__Actinomycetota;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;g__Collinsella;s__Collinsella sp003436275
  1 d__Bacteria;p__Bacillota_A;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium_J;s__Clostridium_J culturomicium
  6 d__Bacteria;p__Bacillota_A;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium;s__Clostridium butyricum
  1 d__Bacteria;p__Bacillota_A;c__Clostridia;o__Eubacteriales;f__Eubacteriaceae;g__Eubacterium;s__Eubacterium limosum
  1 d__Bacteria;p__Bacillota_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Enterocloster;s__Enterocloster sp021201905
  2 d__Bacteria;p__Bacillota_A;c__Clostridia;o__Peptostreptococcales;f__Peptostreptococcaceae;g__Paraclostridium;s__Paraclostridium dentum
  1 d__Bacteria;p__Bacillota;c__Bacilli;o__Bacillales;f__Bacillaceae;g__Bacillus;s__Bacillus altitudinis
  1 d__Bacteria;p__Bacillota;c__Bacilli;o__Bacillales;f__Bacillaceae_G;g__Bacillus_A;s__Bacillus_A bombysepticus
  9 d__Bacteria;p__Bacillota;c__Bacilli;o__Bacillales;f__Bacillaceae_G;g__Bacillus_A;s__Bacillus_A cereus
  6 d__Bacteria;p__Bacillota;c__Bacilli;o__Bacillales;f__Bacillaceae_G;g__Bacillus_A;s__Bacillus_A paranthracis
  1 d__Bacteria;p__Bacillota;c__Bacilli;o__Bacillales;f__Bacillaceae_H;g__Priestia;s__Priestia flexa
  3 d__Bacteria;p__Bacillota;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Faecalicoccus;s__Faecalicoccus pleomorphus
 35 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_A;s__Enterococcus_A avium
  7 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_B;s__Enterococcus_B faecium
  7 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_B;s__Enterococcus_B hirae
  7 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_C;s__Enterococcus_C asini
  1 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_C;s__Enterococcus_C sp009933135
  2 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_D;s__Enterococcus_D casseliflavus
 11 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_D;s__Enterococcus_D gallinarum
  6 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_D;s__Enterococcus_D innesii
 50 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__Enterococcus faecalis
  1 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lacticaseibacillus;s__
 10 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lacticaseibacillus;s__Lacticaseibacillus paracasei
 13 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Ligilactobacillus;s__Ligilactobacillus salivarius
 16 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Limosilactobacillus;s__Limosilactobacillus fermentum
  3 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Pediococcus;s__Pediococcus pentosaceus
 13 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Lactococcus;s__Lactococcus lactis
  1 d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus anginosus
  2 d__Bacteria;p__Bacillota;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Mammaliicoccus;s__Mammaliicoccus sciuri
  1 d__Bacteria;p__Bacillota;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus simulans
  3 d__Bacteria;p__Bacillota_C;c__Negativicutes;o__Acidaminococcales;f__Acidaminococcaceae;g__Acidaminococcus;s__Acidaminococcus intestini
  3 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides caccae
  1 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides fragilis
  3 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides uniformis
  1 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides xylanisolvens
  1 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Marinifilaceae;g__Butyricimonas;s__Butyricimonas sp900184685
  2 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Marinifilaceae;g__Odoribacter;s__Odoribacter splanchnicus
  1 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes_A;s__Alistipes_A ihumii
  1 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes;s__Alistipes senegalensis
  2 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Tannerellaceae;g__Parabacteroides;s__Parabacteroides bouchesdurhonensis
 41 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Tannerellaceae;g__Parabacteroides;s__Parabacteroides distasonis
```

tax.bac120.summary.tsv，其记载了每个 MAG 所属的 species：



```
  1 user_genome     classification  fastani_reference       fastani_reference_radius        fastani_taxonomy        fastani_ani     fastani_af      closest_placement_reference     c
  2 X001    d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli       GCF_003697165.2 95.0    d__Bacter
  3 X003b1  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__Enterococcus faecalis      GCF_000392875.1 95.0    d__Bacteria;p__Bacillota;
  4 X003b2  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli       GCF_003697165.2 95.0    d__Bacter
  5 X004b1  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli       GCF_003697165.2 95.0    d__Bacter
  6 X004b2  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_B;s__Enterococcus_B hirae     GCF_000271405.2 95.0    d__Bacteria;p__Bacillota;
  7 X005b1  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli       GCF_003697165.2 95.0    d__Bacter
  8 X005b2  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_B;s__Enterococcus_B hirae     GCF_000271405.2 95.0    d__Bacteria;p__Bacillota;
  9 X006    d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli       GCF_003697165.2 95.0    d__Bacter
 10 X007    d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_B;s__Enterococcus_B faecium    GCF_001544255.1 95.0    d__Bacteria;p__Bacillota;
 11 X008b1  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_B;s__Enterococcus_B faecium    GCF_001544255.1 95.0    d__Bacteria;p__Bacillota;
 12 X008b2  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__Enterococcus faecalis      GCF_000392875.1 95.0    d__Bacteria;p__Bacillota;
 13 X009    d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Enterobacter;s__Enterobacter hormaechei_A     GCF_001729745.1 95.0    d
 14 X010    d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Enterobacter;s__Enterobacter hormaechei_A     GCF_001729745.1 95.0    d
 15 X011    d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Enterobacter;s__Enterobacter hormaechei_A     GCF_001729745.1 95.0    d
 16 X013b1  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_C;s__Enterococcus_C sp009933135      GCF_009933135.1 95.0    d__Bacteria;p__Ba
 17 X013b2  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli       GCF_003697165.2 95.0    d__Bacter
 18 X014b1  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__Enterococcus faecalis      GCF_000392875.1 95.0    d__Bacteria;p__Bacillota;
 19 X014b2  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Enterobacter;s__Enterobacter hormaechei_A     GCF_001729745.1 95.0    d
 20 X014b3  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_B;s__Enterococcus_B faecium    GCF_001544255.1 95.0    d__Bacteria;p__Bacillota;
 21 X014b4  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli       GCF_003697165.2 95.0    d__Bacter
 22 X015    d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Proteus;s__Proteus mirabilis  GCF_000160755.1 95.0    d__Bacteria;p__Pse
 23 X016b1  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Proteus;s__Proteus mirabilis  GCF_000160755.1 95.0    d__Bacteria;p__Pse
 24 X016b2  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Enterobacter;s__Enterobacter hormaechei_A     GCF_001729745.1 95.0    d
 25 X016b3  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli       GCF_003697165.2 95.0    d__Bacter
 26 X016b4  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_B;s__Enterococcus_B faecium    GCF_001544255.1 95.0    d__Bacteria;p__Bacillota;
 27 X018b1  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__Enterococcus faecalis      GCF_000392875.1 95.0    d__Bacteria;p__Bacillota;
 28 X018b2  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Proteus;s__Proteus mirabilis  GCF_000160755.1 95.0    d__Bacteria;p__Pse
 29 X019    d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__Enterococcus faecalis      GCF_000392875.1 95.0    d__Bacteria;p__Bacillota;
 30 X020b1  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Proteus;s__Proteus mirabilis  GCF_000160755.1 95.0    d__Bacteria;p__Pse
 31 X021b1  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Proteus;s__Proteus mirabilis  GCF_000160755.1 95.0    d__Bacteria;p__Pse
 32 X021b2  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli       GCF_003697165.2 95.0    d__Bacter
 33 X021b3  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_A;s__Enterococcus_A avium     GCF_000406965.1 95.0    d__Bacteria;p__Bacillota;
 34 X021b4  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_D;s__Enterococcus_D gallinarum       GCF_001544275.1 95.0    d__Bacteria;p__Ba
 35 X022b1  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Enterobacter;s__Enterobacter hormaechei_A     GCF_001729745.1 95.0    d
 36 X022b2  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__Enterococcus faecalis      GCF_000392875.1 95.0    d__Bacteria;p__Bacillota;
 37 X022b3  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Proteus;s__Proteus mirabilis  GCF_000160755.1 95.0    d__Bacteria;p__Pse
 38 X024b1  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__Enterococcus faecalis      GCF_000392875.1 95.0    d__Bacteria;p__Bacillota;
 39 X024b2  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Proteus;s__Proteus mirabilis  GCF_000160755.1 95.0    d__Bacteria;p__Pse
 40 X025b1  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Proteus;s__Proteus mirabilis  GCF_000160755.1 95.0    d__Bacteria;p__Pse
 41 X025b2  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__Enterococcus faecalis      GCF_000392875.1 95.0    d__Bacteria;p__Bacillota;
 42 X026b1  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Proteus;s__Proteus mirabilis  GCF_000160755.1 95.0    d__Bacteria;p__Pse
 43 X026b2  d__Bacteria;p__Bacillota;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_D;s__Enterococcus_D gallinarum       GCF_001544275.1 95.0    d__Bacteria;p__Ba
 44 X028b1  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Proteus;s__Proteus mirabilis  GCF_000160755.1 95.0    d__Bacteria;p__Pse
 45 X028b2  d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli       GCF_003697165.2 95.0    d__Bacter
```

使用方法：

python roary_split.py –help

-i：输入文件 1， species.txt 文件

--tax：输入文件 2，tax.bac120.summary.tsv 文件

-t：genome 数量的阈值，只有超过该数值的 genome 才会被创建文件夹，默认：10

--thread：roary 程序调用的线程数

--donotalign：是否使用 MAFFT 对基因比对（roary 内置参数）

-p：Prokka 文件夹。注意必须为绝对路径，不能使用相对路径
-o：输出文件，输出每个种及其对应的 species 数目用于后期作图。

```
$python roary_split.py --help
usage: roary_split.py [-h] -i <file> --tax <file> -t <file> --thread <file>
                      --donotalign <file> -p <file> -o <file> [--version]

Roary results

optional arguments:
  -h, --help            show this help message and exit
  -i <file>, --input <file>
                        Species file: species.txt.
  --tax <file>          Tax file: tax.bac120.summary.tsv.
  -t <file>, --threshold <file>
                        Number of genome, Default: 10.
  --thread <file>       Number of threads, Default: 10.
  --donotalign <file>   Do-not-align genes, Default: T.
  -p <file>, --prokka <file>
                        Prokka directory [ABS]. this directory is usde to link
                        gff file into sub-directory and process roary.
  -o <file>, --out <file>
                        Output file.
  --version             Display version
```