

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_Corionobacteria;o_Corionobacteriales;f_Corionobacteriaceae;g_Collinsella;s_Collinsella sp003436275
1 d_Bacteria;p_Bacillota A;c_Clostridia;o_Clostridiales;f_Clostridiaceae;g_Clostridium J;s_Clostridium J culturomicum
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_Bacillus A;s_Bacillus A bombysepticus
9 d_Bacteria;p_Bacillota;c_Bacilli;o_Bacillales;f_Bacillaceae;g_Bacillus A;s_Bacillus A cereus
6 d_Bacteria;p_Bacillota;c_Bacilli;o_Bacillales;f_Bacillaceae;g_Bacillus A;s_Bacillus A paranthracis
1 d_Bacteria;p_Bacillota;c_Bacilli;o_Bacillales;f_Bacillaceae;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_Lactocaseibacillus;s_
10 d_Bacteria;p_Bacillota;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae;g_Lactocaseibacillus;s_Lactocaseibacillus 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_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 humii
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 bouchesdurhoniensis
41 d_Bacteria;p_Bacteroidota;c_Bacteroidia;o_Bacteroidales;f_Tannerellaceae;g_Parabacteroides;s_Parabacteroides distans
```

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_anl fastani_af closest_placement reference C
2 X0001 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Escherichia coli GCF_003697165.2 95.0 d_Bacter
3 X00302 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 X00302 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Escherichia coli GCF_003697165.2 95.0 d_Bacter
5 X00401 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Escherichia coli GCF_003697165.2 95.0 d_Bacter
6 X00402 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 X00501 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Escherichia coli GCF_003697165.2 95.0 d_Bacter
8 X00502 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 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 X00801 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 X00802 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 X01301 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 X01302 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Escherichia coli GCF_003697165.2 95.0 d_Bacter
18 X01401 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 X01402 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Enterobacter;s_Enterobacter hormaechei A GCF_001729745.1 95.0 d
20 X01403 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 X01404 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_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_Ps
23 X01601 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Proteus;s_Proteus mirabilis GCF_000160755.1 95.0 d_Bacteria;p_Ps
24 X01602 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Enterobacter;s_Enterobacter hormaechei A GCF_001729745.1 95.0 d
25 X01603 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Escherichia coli GCF_003697165.2 95.0 d_Bacter
26 X01604 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 X01801 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 X01802 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Proteus;s_Proteus mirabilis GCF_000160755.1 95.0 d_Bacteria;p_Ps
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 X02001 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Proteus;s_Proteus mirabilis GCF_000160755.1 95.0 d_Bacteria;p_Ps
31 X02101 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Proteus;s_Proteus mirabilis GCF_000160755.1 95.0 d_Bacteria;p_Ps
32 X02102 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Escherichia coli GCF_003697165.2 95.0 d_Bacter
33 X02103 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 X02104 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 X02201 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Enterobacter;s_Enterobacter hormaechei A GCF_001729745.1 95.0 d
36 X02202 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 X02203 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Proteus;s_Proteus mirabilis GCF_000160755.1 95.0 d_Bacteria;p_Ps
38 X02401 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 X02402 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Proteus;s_Proteus mirabilis GCF_000160755.1 95.0 d_Bacteria;p_Ps
40 X02501 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Proteus;s_Proteus mirabilis GCF_000160755.1 95.0 d_Bacteria;p_Ps
41 X02502 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 X02601 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Proteus;s_Proteus mirabilis GCF_000160755.1 95.0 d_Bacteria;p_Ps
43 X02602 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 X02801 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Proteus;s_Proteus mirabilis GCF_000160755.1 95.0 d_Bacteria;p_Ps
45 X02802 d_Bacteria;p_Pseudomonadota;c_Gammaproteobacteria;o_Enterobacterales;f_Enterobacteriaceae;g_Escherichia coli GCF_003697165.2 95.0 d_Bacter
46 X02803 d_Bacteria;p_Bacillota;c_Bacilli;o_Lactobacillales;f_Enterococcaceae;g_Enterococcus;s_Enterococcus faecalis GCF_000392875.1 95.0 d_Bacteria;p_Bacillota;
```

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

-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。

```
usage: roary_split.py [-h] -i <file> --tax <file> -t <file> -p <file> -o <file> [--version]
```

Roary results

options:

```
-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>
                        Threshold of geome number, default: 10.
-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
```