

# From bacterial ASVs to a two-dimensional matrix of gene presence or absence

## Step 1

首先按照配置好vsearch软件。

从GTDB数据库下载配置好细菌基因组数据库bac120\_ssu\_reps\_r207.fna。

创建一个vsearch比对用的数据库（bac120\_ssu\_reps\_r207.udb我已经放在/public/home/2022122/chenhuilong/ph\_preference/data目录下）

vsearch --makeudb\_usearch bac120\_ssu\_reps\_r207.fna -output bac120\_ssu\_reps\_r207.udb

比对ASV序列到参考基因组数据库。

vsearch --usearch\_global refseq.fasta --db bac120\_ssu\_reps\_r207.udb --strand both --notrunc --id 0.99 --maxrejects 100 --maxaccepts 100 --blast6out PAN\_607\_aligned\_ssu.tsv --threads 16

aligned\_ssu.tsv结果如下，即blast output format 6 file：

1	ASV_29951	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_29951	GR_GCA_002414185.1	100	257	0	0	257	1	1	1210	-1	0
2	ASV_29964	Bacteria	NA	NA	NA	NA	NA	ASV_29964	GR_GCA_001821055.1	99.2	261	2	0	261	1	1	1714	-1	0
3	ASV_31077	NA	NA	NA	NA	NA	NA	ASV_31077	RS_GCF_007115105.1	99.6	251	1	0	251	1	1	1591	-1	0
4	ASV_33204	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_33204	GR_GCA_003487785.1	100	221	0	0	221	1	1	1564	-1	0
5	ASV_30952	Eukaryota	NA	NA	NA	NA	NA	ASV_30952	GR_GCA_002293685.1	100	257	0	0	257	1	1	1631	-1	0
6	ASV_26826	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_26826	GR_GCA_001821055.1	100	255	0	0	255	1	1	1714	-1	0
7	ASV_11091	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_11091	GR_GCA_001821055.1	100	253	0	0	253	1	1	1714	-1	0
8	ASV_30305	Bacteria	NA	NA	NA	NA	NA	ASV_30305	GR_GCA_001821055.1	100	250	0	0	250	1	1	1714	-1	0
9	ASV_29338	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_29338	GR_GCA_001821055.1	100	252	0	0	252	1	1	1714	-1	0
10	ASV_29053	Eukaryota	NA	NA	NA	NA	NA	ASV_29053	GR_GCA_002483765.1	99.6	250	1	0	250	1	1	1837	-1	0
11	ASV_22970	Bacteria	NA	NA	NA	NA	NA	ASV_22970	GR_GCA_002483765.1	99.6	251	1	0	251	1	1	1837	-1	0
12	ASV_10874	Bacteria	NA	NA	NA	NA	NA	ASV_10874	GR_GCA_002483765.1	99.6	251	1	0	251	1	1	1837	-1	0
13	ASV_26310	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_26310	GR_GCA_002483765.1	99.6	251	1	0	251	1	1	1837	-1	0
14	ASV_19308	Bacteria	NA	NA	NA	NA	NA	ASV_19308	GR_GCA_002294185.1	99.6	253	1	0	253	1	1	1579	-1	0
15	ASV_18315	NA	NA	NA	NA	NA	NA	ASV_18315	GR_GCA_002294185.1	99.6	253	1	0	253	1	1	1579	-1	0
16	ASV_17098	Bacteria	NA	NA	NA	NA	NA	ASV_17098	GR_GCA_002294185.1	99.6	253	1	0	253	1	1	1579	-1	0
17	ASV_27407	Eukaryota	NA	NA	NA	NA	NA	ASV_27407	GR_GCA_002293685.1	100	256	0	0	256	1	1	1631	-1	0
18	ASV_27641	Eukaryota	NA	NA	NA	NA	NA	ASV_27641	RS_GCF_007115105.1	99.6	253	1	0	253	1	1	1591	-1	0
19	ASV_29144	Eukaryota	NA	NA	NA	NA	NA	ASV_29144	GR_GCA_002483765.1	99.6	259	1	0	259	1	1	1837	-1	0
20	ASV_17510	NA	NA	NA	NA	NA	NA	ASV_17510	GR_GCA_002307715.1	100	265	0	0	265	1	1	1758	-1	0
21	ASV_29539	Bacteria	Proteobacteria	NA	NA	NA	NA	ASV_29539	GR_GCA_003455565.1	100	255	0	0	255	1	1	1227	-1	0
22	ASV_28759	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_28759	GR_GCA_003455565.1	100	255	0	0	255	1	1	1227	-1	0
23	ASV_22091	Bacteria	Proteobacteria	Alphaproteobacteria	NA	NA	NA	ASV_22091	GR_GCA_003455565.1	100	255	0	0	255	1	1	1227	-1	0
24	ASV_26448	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_26448	GR_GCA_001821055.1	100	252	0	0	252	1	1	1714	-1	0
25	ASV_26448	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_26448	GR_GCA_002483765.1	99.6	252	1	0	252	1	1	1837	-1	0
26	ASV_27487	NA	NA	NA	NA	NA	NA	ASV_27487	GR_GCA_002483765.1	100	250	0	0	250	1	1	1835	-1	0
27	ASV_32860	NA	NA	NA	NA	NA	NA	ASV_32860	RS_GCF_007115105.1	99.6	255	1	0	255	1	1	1591	-1	0
28	ASV_240	Bacteria	Acidobacteria	Acidobacteriia	Subgroup_2	NA	NA	ASV_240	GR_GCA_013286465.1	99.2	253	2	0	253	1	1	1426	-1	0
29	ASV_29512	Eukaryota	NA	NA	NA	NA	NA	ASV_29512	GR_GCA_003487785.1	100	234	0	0	234	1	1	1564	-1	0
30	ASV_33313	NA	NA	NA	NA	NA	NA	ASV_33313	GR_GCA_002293685.1	100	251	0	0	251	1	1	1631	-1	0
31	ASV_27105	Bacteria	Proteobacteria	NA	NA	NA	NA	ASV_27105	RS_GCF_007115105.1	99.6	229	1	0	229	1	1	1591	-1	0
32	ASV_10401	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_10401	RS_GCF_007115105.1	99.6	229	1	0	229	1	1	1591	-1	0
33	ASV_14617	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_14617	RS_GCF_007115105.1	99.6	232	1	0	232	1	1	1591	-1	0
34	ASV_24237	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_24237	RS_GCF_007115105.1	99.1	225	2	0	225	1	1	1591	-1	0
35	ASV_31022	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_31022	RS_GCF_007115105.1	99.1	227	2	0	227	1	1	1591	-1	0
36	ASV_28195	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA	ASV_28195	RS_GCF_007115105.1	99.1	225	2	0	225	1	1	1591	-1	0
37	ASV_31447	NA	NA	NA	NA	NA	NA	ASV_31447	GR_GCA_002483765.1	100	254	0	0	254	1	1	1837	-1	0
38	ASV_30889	Eukaryota	NA	NA	NA	NA	NA	ASV_30889	RS_GCF_007115105.1	99.6	262	1	0	262	1	1	1591	-1	0
39	ASV_8526	Eukaryota	NA	NA	NA	NA	NA	ASV_8526	GR_GCA_003452965.1	99	202	2	0	202	1	1	1968	-1	0
40	ASV_11480	NA	NA	NA	NA	NA	NA	ASV_11480	GR_GCA_003452965.1	99	202	2	0	202	1	1	1968	-1	0
41	ASV_10115	Eukaryota	NA	NA	NA	NA	NA	ASV_10115	GR_GCA_003452965.1	99	202	2	0	202	1	1	1968	-1	0
42	ASV_23807	Eukaryota	NA	NA	NA	NA	NA	ASV_23807	GR_GCA_012959805.1	100	203	0	0	203	1	1	1569	-1	0
43	ASV_26334	NA	NA	NA	NA	NA	NA	ASV_26334	GR_GCA_003452965.1	99	202	2	0	202	1	1	1968	-1	0
44	ASV_27507	Eukaryota	NA	NA	NA	NA	NA	ASV_27507	GR_GCA_003452965.1	99	202	2	0	202	1	1	1968	-1	0
45	ASV_14284	Eukaryota	NA	NA	NA	NA	NA	ASV_14284	GR_GCA_003452965.1	99	202	2	0	202	1	1	1968	-1	0
46	ASV_32449	Archaea	Euryarchaeota	Methanobacteriia	Methanobacteriales	Methanobacteriaceae	Methanobrevibacter	ASV_32449	GR_GCA_017381935.1	99.6	254	1	0	254	1	1	1258	-1	0
47	ASV_31154	Bacteria	NA	NA	NA	NA	NA	ASV_31154	GR_GCA_002483765.1	100	256	0	0	256	1	1	1837	-1	0

## Step 2

使用“python according\_to\_vsearch\_blast6out\_result\_extract\_the\_single\_match\_genome.py”脚本（optional—）根据PAN\_607\_aligned\_ssu.tsv得到大于等于99.6%的结果并提出ASV与细菌基因组的唯一对应ID（asv\_vs\_genome\_id.txt）

```
usage: according_to_vsearch_blast6out_result_extract_the_single_match_genome.py [-h] -i blast6Format -o1 asvToGenome -o2 outputFolder

Matching ASVs to unique genomes

optional arguments:
  -h, --help            show this help message and exit
  -i blast6Format, --inputFilePath blast6Format
                        Input blast6out format file path
  -o1 asvToGenome, --outputFilePath asvToGenome
                        Output asv to genome id file path
  -o2 outputFolder, --outputFolderPath outputFolder
                        Output folder path
```

python according\_to\_vsearch\_blast6out\_result\_extract\_the\_single\_match\_genome.py -i  
PAN\_607\_aligned\_ssu.tsv -o1 asv\_genome\_id\_unique\_match\_ID.txt -o2 result

	A	B	C	D	E	F	G	H	
1	ASV_113	GB_GCA_003222395.1							
2	ASV_1222	RS_GCF_000239795.1							
3	ASV_1254	RS_GCF_012641395.1	RS_GCF_007829155.1	RS_GCF_014207095.1	GB_GCA_900465675.1	RS_GCF_001692325.1	RS_GCF_017876795.1	RS_GCF_900094585.1	RS_GCF_017876795.1
4	ASV_1280	GB_GCA_003165335.1							
5	ASV_129	RS_GCF_001905365.1	RS_GCF_900116325.1	RS_GCF_000745675.1	RS_GCF_004795105.1	RS_GCF_900091955.1	GB_GCA_000470535.1	RS_GCF_008042275.1	RS_GCF_008042275.1
6	ASV_1337	RS_GCF_001956915.1							
7	ASV_137	RS_GCF_000426105.1							
8	ASV_144	RS_GCF_010731775.1	RS_GCF_018278905.1	RS_GCF_900566085.1	RS_GCF_001672975.1	RS_GCF_001667775.1	RS_GCF_001665875.1	RS_GCF_902652685.1	RS_GCF_001665875.1
9	ASV_1441	RS_GCF_000620785.1	RS_GCF_000423625.1						

基因组ID表示只有此行的ASV可以匹配到，所以取哪个都行。

39	ASV_384	RS_GCF_008369935.1							
40	ASV_386								
41	ASV_390	RS_GCF_900455735.1	RS_GCF_017068395.1	RS_GCF_001692325.1	GB_GCA_900468285.1	RS_GCF_001692325.1	RS_GCF_017876795.1	RS_GCF_900094585.1	RS_GCF_017876795.1
42	ASV_421	GB_GCA_003220615.1							
43	ASV_430								
44	ASV_435	GB_GCA_017881025.1							
45	ASV_450	GB_GCA_016711085.1							

如果运气不好，有些ASV没有单独它匹配到的细菌基因组，就从原始的blast6out结果表格中找一个ASV的匹配结果复制并粘贴到B列，因为ASV可能重复，把B列中重复的后面的列的值替换被重复的那个细菌基因组ID即可。

反之，如果B列都有唯一匹配的基因组序列：

	A	B	C	D	E	F	G	H	I	J	K
1	OTU1	RS_GCF_003641185.1	RS_GCF_003641145.1	RS_GCF_000463075.2	RS_GCF_014131735.1						
2	OTU10	RS_GCF_002706425.1									
3	OTU12	RS_GCF_001441165.1									
4	OTU13	GB_GCA_900468015.1	GB_GCA_900473205.1	GB_GCA_900468285.1							
5	OTU17	RS_GCF_001433855.1									
6	OTU18	RS_GCF_001436415.1	RS_GCF_004101845.1								
7	OTU19	RS_GCF_000160575.1									
8	OTU21	RS_GCF_000193635.1	RS_GCF_004521965.1								
9	OTU24	RS_GCF_005405305.1	RS_GCF_014058425.1								
10	OTU26	RS_GCF_001543285.1	RS_GCF_002083135.2	RS_GCF_001543205.1							
11	OTU27	RS_GCF_000739975.1	RS_GCF_002154895.1	GB_GCA_009911365.1							
12	OTU28	RS_GCF_001020205.1	RS_GCF_001276985.1	RS_GCF_001889325.1	RS_GCF_001276215.1	RS_GCF_002008345.1					
13	OTU29	RS_GCF_006151785.1									
14	OTU3	RS_GCF_000392875.1									
15	OTU31	RS_GCF_002607735.1	RS_GCF_002777965.1	RS_GCF_011045815.1	RS_GCF_003144395.1	RS_GCF_003144405.1	RS_GCF_016239305.1				
16	OTU32	RS_GCF_001436555.1									
17	OTU33	RS_GCF_005405265.1									
18	OTU35	RS_GCF_001437055.1									
19	OTU36	RS_GCF_001936335.1									
20	OTU37	RS_GCF_001423405.1	RS_GCF_900234795.1	RS_GCF_014845115.1	RS_GCF_001542815.1						
21	OTU39	RS_GCF_018314255.1									
22	OTU4	RS_GCF_013394305.1	RS_GCF_002813755.1	RS_GCF_017316305.1	RS_GCF_003957375.1						
23	OTU41	RS_GCF_003946675.1	RS_GCF_003946645.1								
24	OTU43	RS_GCF_001423425.1									
25	OTU45	RS_GCF_002583405.1	RS_GCF_001311075.1	RS_GCF_001434895.1							
26	OTU46	RS_GCF_000159315.1									
27	OTU47	RS_GCF_001436295.1									
28	OTU48	RS_GCF_000359585.1									
29	OTU49	RS_GCF_014522265.1	RS_GCF_001512295.1	RS_GCF_001650425.1	RS_GCF_000730625.1						
30	OTU5	RS_GCF_011078175.1	RS_GCF_001875655.1								
31	OTU50	RS_GCF_001434705.1									
32	OTU51	RS_GCF_003946465.1									
33	OTU52	RS_GCF_001691565.1	RS_GCF_000383475.1	RS_GCF_001552475.1	RS_GCF_900102005.1	RS_GCF_001644225.1					
34	OTU53	RS_GCF_019218635.1									
35	OTU54	RS_GCF_900095695.1	RS_GCF_000735375.1								
36	OTU57	RS_GCF_005405245.1									
37	OTU59	RS_GCF_001437425.1									
38	OTU62	RS_GCF_002907375.1									
39	OTU63	RS_GCF_000829035.1									
40	OTU67	GB_GCA_000751555.1									
41	OTU68	RS_GCF_000407605.1									
42	OTU7	RS_GCF_014634805.1	RS_GCF_000014445.1	RS_GCF_014634745.1							
43	OTU71	RS_GCF_015159595.1									
44	OTU74	GB_GCA_003072625.1									
45	OTU76	RS_GCF_002095385.1									
46	OTU77	RS_GCF_009789535.1									
47	OTU78	RS_GCF_001434785.1									
48	OTU8	RS_GCF_001437285.1									
49											

就直接复制A和B列即可。

asv\_vs.\_genome\_id.txt文件内容如下：

```
ASV_113 GB_GCA_003222395.1
ASV_1222 RS_GCF_000239795.1
ASV_1254 RS_GCF_012641395.1
... ..
ASV_998 RS_GCF_001542915.1
```

## Step 3

使用“batch\_according\_to\_match\_id\_file\_extract\_genome\_multisequence.py”脚本根据上一步得到的asv\_vs.\_genome\_id.txt文件，从已处理好的集群中的细菌基因组数据库中，批量萃取到ASV对应的细菌基因组的全蛋白组序列，并自动生成在genome\_extract\_result文件夹中。如下：

## 细菌基因组数据库目

录: /public/home/2022122/chenhuilong/ph\_preference/data/protein\_faa\_reps/bacteria\_processed

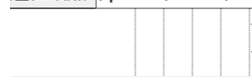
```
usage: batch_according_to_match_id_file_extract_genome_multisequence.py [-h] -i asvToGenome -d proteomeDatabase -o outputFolder
Batch extraction of target bacterial proteome

optional arguments:
  -h, --help            show this help message and exit
  -i asvToGenome, --inputFilePath asvToGenome
                        Input asv to genome id file path
  -d proteomeDatabase, --inputDatabaseFolderPath proteomeDatabase
                        Path to the folder holding the bacterial proteomes
  -o outputFolder, --outputFolderPath outputFolder
                        Output folder path
```

```
python batch_according_to_match_id_file_extract_genome_multisequence.py -i
asv_vs._genome_id.txt -d
/public/home/2022122/chenhuilong/ph_preference/data/protein_faa_reps/bacteria_processed -o
genome_extract_result
```

genome\_extract\_result文件夹中结果如下:

远程站点: /public/home/2022122/chenhuilong/ph\_preference/data/protein\_faa\_reps/genome\_extract\_result

					
文件名	文件大小	文件类型			
..					
GB_GCA_000421065.1_protein.fasta	1,912,659	FASTA File			
GB_GCA_000620725.1_protein.fasta	2,053,964	FASTA File			
GB_GCA_000688615.1_protein.fasta	1,596,741	FASTA File			
GB_GCA_002101905.1_protein.fasta	1,574,974	FASTA File			
GB_GCA_003165335.1_protein.fasta	1,658,673	FASTA File			
GB_GCA_003168335.1_protein.fasta	1,609,598	FASTA File			
GB_GCA_003219415.1_protein.fasta	1,155,787	FASTA File			
GB_GCA_003220265.1_protein.fasta	2,020,157	FASTA File			
GB_GCA_003220615.1_protein.fasta	1,946,288	FASTA File			

## Step 4

对提取出来的目标蛋白组进行基因类型注释

**conda activate pfam\_annotation** (我在集群中建了一个conda镜像 (里面安装配置好HMMER软件并配置好Pfam数据库) )

使用“batch\_run\_hmmscan.py”脚本根据上一步得到的全蛋白组序列, 批量自动进行hmmscan的程序调用和全蛋白组的基因注释, 并自动将基因注释结果生成在目标文件夹中。

**#hmmscan注释一个细菌基因组的时长大约1个小时, 所以这步会很耗时。**

```
usage: batch_run_hmmscan.py [-h] -i targetProteome -o outputFolder
Batch annotation of genotypes via HMMER

optional arguments:
  -h, --help            show this help message and exit
  -i targetProteome, --inputFolderPath targetProteome
                        Input Path to the folder holding the target bacterial proteomes
  -o outputFolder, --outputFolderPath outputFolder
                        Output folder path
```

```
python batch_run_hmmscan.py -i genome_extract_result -o hmmscanResult
```

hmmscanResult文件夹内容如下:

远程站点: /public/home/2022122/chenhuilong/ph\_preference/data/protein\_faa\_reps/hmmscanResult

	genome_extract_result	
	hmmscanResult	
	presence_absence_table_result	

文件名	文件大小	文件类型
..		
GB_GCA_000421065.1_hmmscan.tbl	3,329,022	TBL 文件
GB_GCA_000620725.1_hmmscan.tbl	3,580,917	TBL 文件
GB_GCA_000688615.1_hmmscan.tbl	2,844,513	TBL 文件
GB_GCA_002101905.1_hmmscan.tbl	2,981,860	TBL 文件
GB_GCA_003165335.1_hmmscan.tbl	3,133,205	TBL 文件
GB_GCA_003168335.1_hmmscan.tbl	3,226,902	TBL 文件
GB_GCA_003219415.1_hmmscan.tbl	2,280,846	TBL 文件
GB_GCA_003220265.1_hmmscan.tbl	4,154,911	TBL 文件
GB_GCA_003220615.1_hmmscan.tbl	4,043,950	TBL 文件

conda deactivate

## Step 5

根据hmmscanResult文件夹中的结果，使用

“batch\_according\_to\_hmmResult\_extract\_all\_gene\_presence\_or\_absence\_table-command\_format.py”脚本批量提取细菌基因组全部基因类型的存在或缺失矩阵表。

```
usage: batch_according_to_hmmResult_extract_all_gene_presence_or_absence_table-command_format.py [-h] -i asvToGenome -m hmmscanResult -o outputFolder
Obtaining full gene presence or absence matrices based on HMMER annotation results
optional arguments:
  -h, --help            show this help message and exit
  -i asvToGenome, --inputFilePath asvToGenome
                        Input asv to genome id file path
  -m hmmscanResult, --inputFolderPath hmmscanResult
                        Input Path to the folder holding the target HMMER annotation results
  -o outputFolder, --outputFolderPath outputFolder
                        Output folder path
```

```
python batch_according_to_hmmResult_extract_all_gene_presence_or_absence_table-
command_format.py -i asv_vs_genome_id.txt -m hmmscanResult -o
all_gene_presence_absence_table_result
```

#终端打印出来的为所有基因类型的数量，如，18935个。

所有基因类型存在或缺失矩阵存在生成的all\_gene\_presence\_absence\_table\_result文件夹中。如下：

作者更新之后的 > 实操 > hmmscan > all\_gene\_presence\_absence\_table\_result

名称	修改日期	类型
allGenePresenceAbsenceBinarizeTable.xls	2023/11/26 20:17	Microsoft Excel 97-...

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	ASV/OTU	X5.FTHF_cyc.lig	AAA_25	AAA_assoc.C	AcetylCoA_hyd.C	ASH	Big_3_5	CitMHS	CpsB_CapC	CpXC	CsbD	Cys_rich_CPXG	Cytidylate_kin2	CytoC_RC	DHquinase_I	Exo_endo_pl
2	ASV_113	0	1	0	1	1	1	1	0	1	1	1	1	1	1	1
3	ASV_1222	0	1	1	1	0	1	1	1	0	1	1	1	0	0	1
4	ASV_1254	0	1	1	1	1	0	1	0	0	1	0	1	1	0	1
5	ASV_1280	0	1	1	1	0	0	1	0	1	1	1	1	0	0	1
6	ASV_129	0	1	0	1	1	1	1	0	1	1	0	1	1	0	1
7	ASV_1337	0	1	1	1	1	1	1	0	0	1	0	1	0	1	1
8	ASV_137	0	1	1	1	0	0	1	0	0	1	0	1	0	0	1
9	ASV_144	0	1	0	1	0	0	1	0	0	1	1	1	0	0	1
10	ASV_1441	0	1	0	1	0	0	1	0	1	1	0	1	0	0	1
11	ASV_151	0	1	0	1	0	1	1	1	1	1	0	1	0	0	1
12	ASV_167	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	ASV_168	0	1	0	0	0	0	1	0	1	1	0	1	0	0	1
14	ASV_180	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1