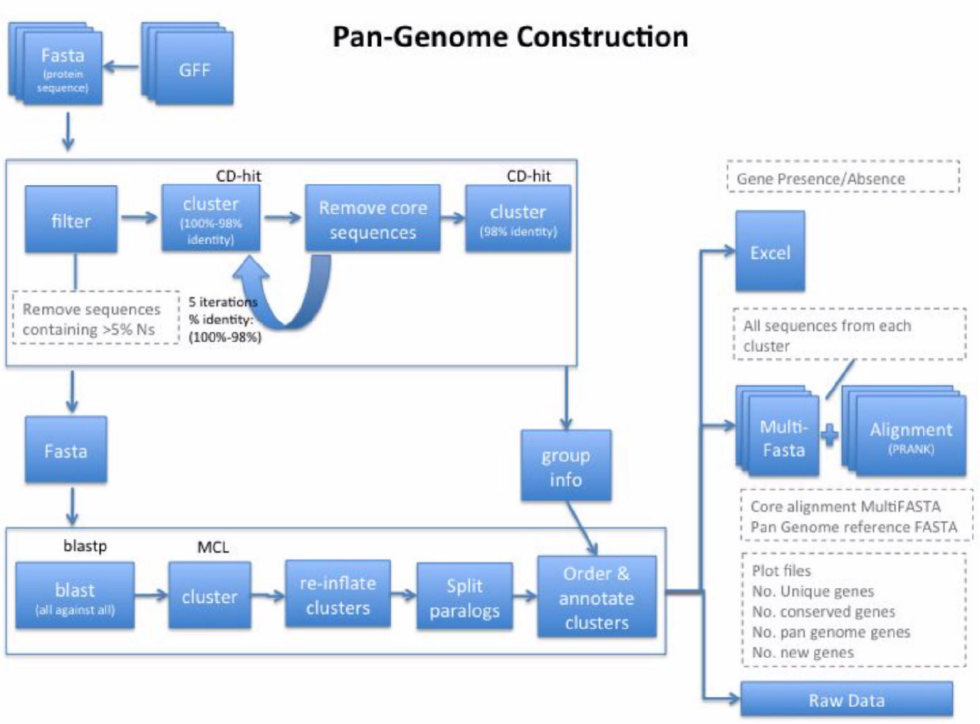
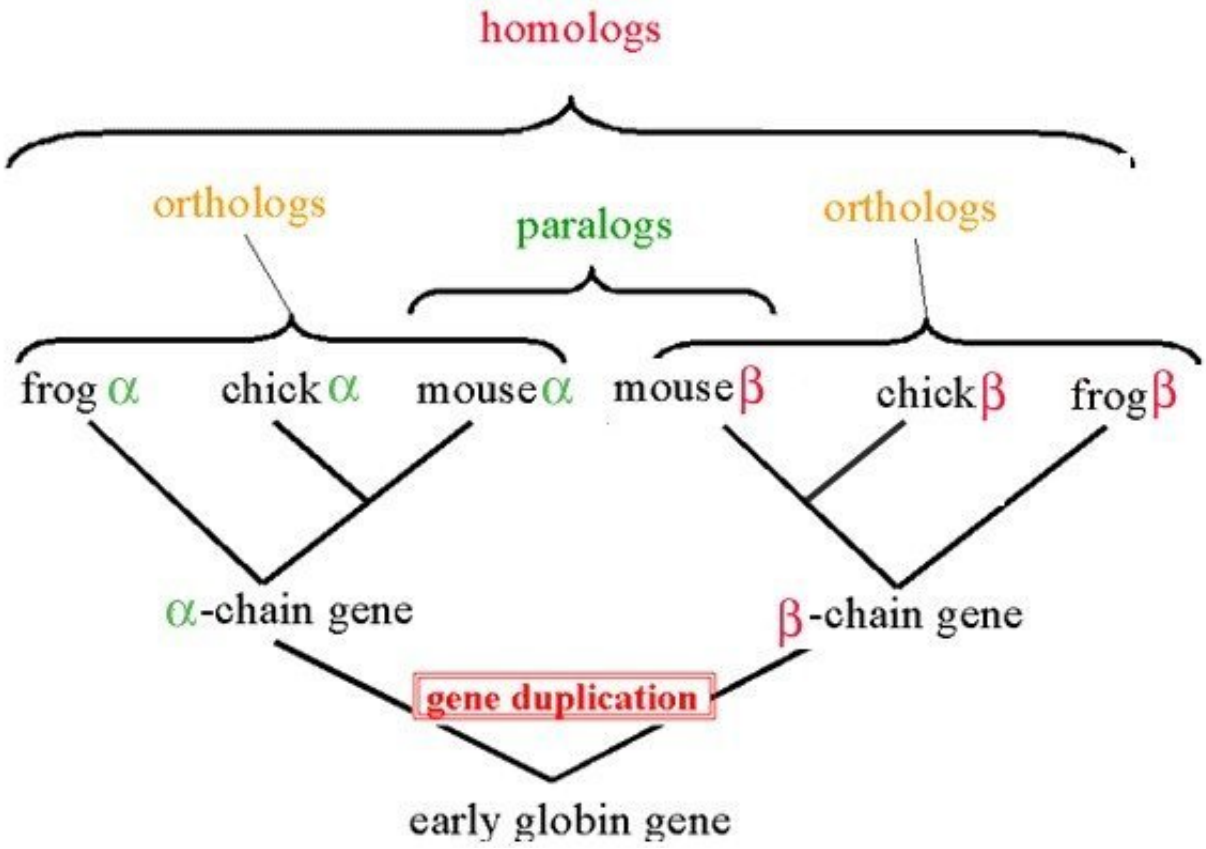
[Roary][https://github.com/sanger-pathogens/Roary]是一个快速计算pan genome软件，将输入gff3格式文件(prokka输出的gff文件)根据序列相似性分类，计算pan genome。首先将输入文件的编码区域提取出来，然后翻译成对应的蛋白序列文件，过滤掉部分的序列，然后使用CD-HIT对其进行迭代聚类，这将在很大程度上减少蛋白序列内容；然后根据设定的序列一致性(默认为95%)对所有序列使用BLASTP进行比对；接着使用MCL进行聚类；最后将CD-HIT聚类后的结果和MCL聚类后的结果汇总，得到pan genome蛋白序列。



Sup. Fig. 13: A flowchart of the steps in the application.

Paralogs



Usage

```
roary *.gff
```

使用8线程生成核心基因比对

```
roary -e --mafft -p 8 *.gff
```

检测软件是否正确安装

```
roary -a
```

-o clusters输出文件名[clustered_proteins]

-f 输出文件路径[.]

-e 如果不使用--mafft，则使用PRANK针对使用codon aware alignment构建core genes的multiFASTA比对，速度慢但是准确

-n 和-e一起使用MAFFT执行核酸的比对快速构建core gene，快速但是准确性不高

以上core_gene_alignment.aln(不能排除重组)可用于输入构建系统发育树，可使用snp_sites先过滤，以减少运行时间和内存

-i blastp比对的最小一致性[95]

-cd 基因存在于该比例的isloates中时判定为core[99]

-r 创建R图，需求R和ggplot2

-s 不进行paralogs split

-t 翻译蛋白密码表[11]

-ap 允许paralogs存在于core aligment

Output

gene_presence_absence.csv 每个isolate中基因在每个group内存在分布

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Gene	Non	Annotation	No. isc	No. sequen	Avg sequen	Genome Fra	Order withir	Access	Access	QC	Min group s	Max group s	Avg group s	36170	38218	38377	38588	39401	42395	
2	pgrR_3		HTH-type tr	6	6	1	1	1427				554	887	720	ENJFHDCD_	IALAIEND_0	LLFCEEMA_	MJNCPELE_	GHLKHLGD_	EDNNINJN_00106	
3	group_103		hypothetica	6	6	1	1	1455				125	125	125	ENJFHDCD_	IALAIEND_0	LLFCEEMA_	MJNCPELE_	GHLKHLGD_	EDNNINJN_00134	
4	rsxC		Electron trar	6	6	1	1	1490				2054	2261	2192	ENJFHDCD_	IALAIEND_0	LLFCEEMA_	MJNCPELE_	GHLKHLGD_	EDNNINJN_00169	
5	asr		Acid shock p	6	6	1	1	1527				293	407	359	ENJFHDCD_	IALAIEND_0	LLFCEEMA_	MJNCPELE_	GHLKHLGD_	EDNNINJN_00206	
6	group_106		hypothetica	6	6	1	1	1533				149	149	149	ENJFHDCD_	IALAIEND_0	LLFCEEMA_	MJNCPELE_	GHLKHLGD_	EDNNINJN_00214	
7	crnA		Creatinine a	6	6	1	1	1581				656	881	768	ENJFHDCD_	IALAIEND_0	LLFCEEMA_	MJNCPELE_	GHLKHLGD_	EDNNINJN_00258	
8	group_111		putative FAI	6	6	1	1	1585				1418	1421	1418	ENJFHDCD_	IALAIEND_0	LLFCEEMA_	MJNCPELE_	GHLKHLGD_	EDNNINJN_00262	
9	thlA_2		Acetyl-CoA	6	6	1	1	1598				749	1181	965	ENJFHDCD_	IALAIEND_0	LLFCEEMA_	MJNCPELE_	GHLKHLGD_	EDNNINJN_00269	
10	group_114		hypothetica	6	6	1	1	3114				302	302	302	ENJFHDCD_	IALAIEND_0	LLFCEEMA_	MJNCPELE_	GHLKHLGD_	EDNNINJN_02467	

core_gene_alignment.aln 核心保守基因的multi-FASTA比对

```
$bioawk -c fastx '{print $name,length($seq)}' core_gene_alignment.aln
36170   4114273
38218   4114273
38377   4114273
38588   4114273
39401   4114273
42395   4114273
```

`fasttree -nt -gtr core_gene_alignment.aln > my_core_gene_alignmnt.newick`

clustered_proteins 一个cluster一行，包含序列ID

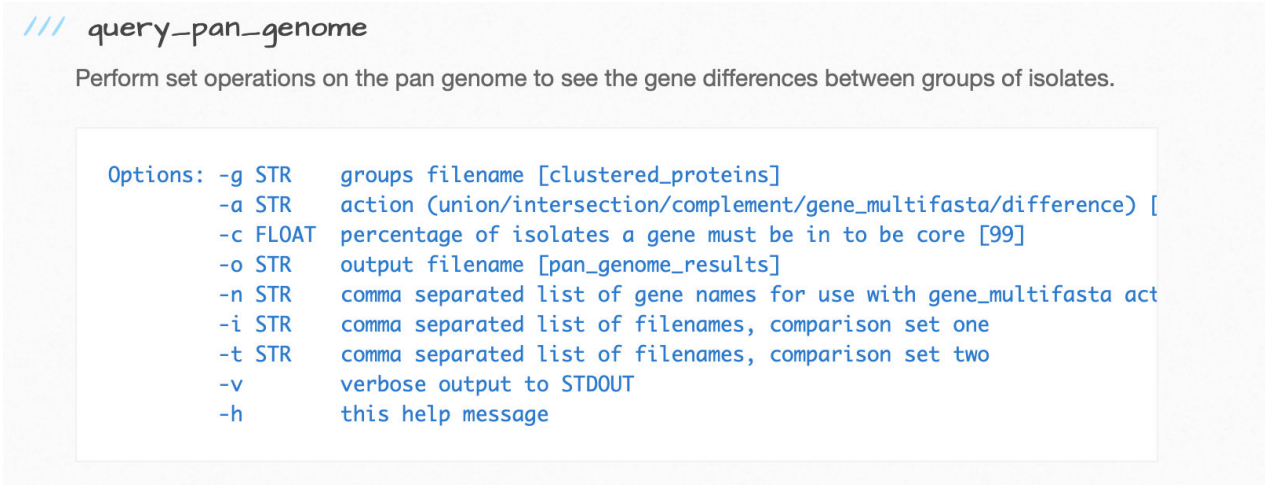
```
1 adh1_1: ENJFHDCD_02879  IALAIEND_03209  LLFCEEMA_02761  MJNCPELE_03749  GHLKHLGD_02911  EDNNINJN_02655
2 group_255: ENJFHDCD_00992       IALAIEND_01872  LLFCEEMA_04436  GHLKHLGD_00634  EDNNINJN_03775  MJNCPELE_02613
3 ompR_2: ENJFHDCD_02786  IALAIEND_04541  LLFCEEMA_02852  MJNCPELE_03842  GHLKHLGD_03004  EDNNINJN_02561
4 mdtM_2: ENJFHDCD_05060  IALAIEND_05023  LLFCEEMA_04500  MJNCPELE_02545  GHLKHLGD_05265  EDNNINJN_03708
5 pgrR_11: ENJFHDCD_05085 IALAIEND_05062  LLFCEEMA_01724  MJNCPELE_01076  GHLKHLGD_05289  EDNNINJN_01884
6 group_6374: ENJFHDCD_04571      IALAIEND_04569  LLFCEEMA_02517  MJNCPELE_04801  GHLKHLGD_04775  EDNNINJN_02906
7 group_385: ENJFHDCD_03633       IALAIEND_03367  LLFCEEMA_01349  GHLKHLGD_03776  EDNNINJN_01438  MJNCPELE_02946
8 yqjC: ENJFHDCD_00224    IALAIEND_00169  LLFCEEMA_03836  MJNCPELE_01946  GHLKHLGD_00169  EDNNINJN_04134
```

Accessory_binary_genes.fa.newick accessory genome内基因分布关系的newick tree，可使用FigTree打开，查看acessory genes的对应关系图，该关系图较为粗糙

First of all we construct a FASTA file with the binary presence and absence of genes, where 'A' means a gene is present and 'C' means it is absent. Only the first 4000 genes in the accessory genome are considered to limit the running time and memory usage of FastTree. FastTree is then run with the fastest possible settings to produce a Newick tree.

```
1 36170   AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
2 38218   AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
3 38377   CAAAAACCCCCCCCCCCCCCCCACCCCCCCCCACCCCCCCCCCCCCCCCAAAAACCCCCACCCCCCCCCCCACCCCCCCCCCC
4 38588   CAAAAACCCCAAAACAAAAAACACCCCCCCCCCAAAAAAAAAAAAAAAAAAACCAACCCCCCCCCCCCCCCCCCCCCCCCCCCA
5 39401   AAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
6 42395   CCCCCCCAAACCCCCAAAAAACACCCCCCCCCCAAAAAAAAAAAAAAACCCCCAAAACAAACAAAACCCCACCCCCACCACCCCCCA
```

query_pan_genome



you need all Roary output within the same folder as the .gff files so query_pan_genome works

查看isloates中所有基因

```
query_pan_geonme -a union *.gff
```

查看isolates中基因交集

```
query_pan_genome -a intersection *.gff
```

查看isloates中的acessory 基因

```
query_pan_genome -a complement *.gff
```

提取基因的序列并构建multi-FASTA文件

```
query_pan_genome -a gene_multifasta -n gryA,mecA,abc *.gff
```

存在于两组isolates中的基因分布差异

```
query_pan_genome -a difference --input_set_one 1.gff,2.gff --input_set_two
3.gff,4.gff,5.gff
```

Receipe for using roary

1. Annotate FASTA files with PROKKA
2. Roary -e --mafft *.gff
3. FastTree -nt -gtr core_gene_alignment.aln > my_tree.newick

---