How to create a pangenome of isolated genome sequences using Roary and Prokka?

Author : Tariq Abdullah

Categories : Softwares, Tools

Date : February 23, 2019

Roary is a pangenome genome pipeline, which calculates pangenome of a set of related prokaryotic isolates [1]. It takes annotated assemblies in the gff3 format generated by Prokka [2] and provides the pangenome. The working methodology has been explained in our previous article. In this article, we will learn how to create the pangenome of a few isolated genome sequences using Roary [1] and Prokka [2].

Input for Roary

1. Genome sequences in the form of gff3 files.

Downloading the genome sequences

At first, you need to download genome sequences as per your need, which you can easily download yourself or by using ncbi-genome-download package. It provides several scripts to download genome sequences from NCBI FTP servers. To install this package, open a terminal (Ctrl + T) and type the following commands:

```
$ pip install ncbi-genome-download
```

After downloading this package, you can download the genome assemblies as per your requirements such as fasta sequences of all bacteria, viral genome, RefSeq genome sequences in GenBank format, fungal genomes and so on. (Remember, while downloading gff3 files, you need to download Genbank files with the nucleotide sequence because gff3 files on the NCBI website contain annotation only). I will download all bacterial sequences in fasta format using the following command (showing this example with only a few sequences only):

```
$ ncbi-genome-download --format fasta bacteria
```

Annotating the genome sequences

Go into the directory of Roary, create a new folder, let's name it as *example*, and save download these sequences. After downloading, you will see many fasta files in the same folder. Now start annotating them to determine the attributes and location of the genes present in them, and also to obtain gff3 files which are used as an input in roary. This can be easily done with Prokka [2]. Open the terminal and type the following commands:

```
$ cd Downloads/Roary/example/
```

```
$ prokka --kingdom Bacteria --outdir prokka_GCA_000006285 --genus Salmonella --locustag GCA_000006285 GCA_000006285.2_ASM828v3_genomic.fna
```

You can further add other description such as organism details (genus, species, etc.). Make sure you annotate all the genome sequences you are dealing with and remember to change the output directory name, locus tag, and assembly version accordingly. After running this command, a new directory will be created in the name of each sequence and it will be consisting of 12 files with different extensions including the gff3 file.

Creating pangenome/Running Roary

We have got gff3 files of the genome sequences in the directories, now we need to copy the gff3 file from each directory into another directory (let's say, *gff_all*). After that, open the terminal again and type the following command to run roary:

```
$ roary -f ./tutorial -e -n -v ./gff_all/*.gff
```

At this stage, Roary will get all the coding sequences, translate them into protein sequences, and generate pre-clusters. After that, roary will look for the paralogs by using blastp [3] and will create clusters using MCL [4]. Finally, it will take every isolate and order them according to the presence/absence of orthologs. This will take time depending upon the number of sequences (or gff3 files) you are using.

If you want to create a pangenome without the core alignment, then use the following command:

```
$ roary -f ./tutorial -v ./gff_all/*.gff
```

If you want to change the percentage identity of blastp (not advised to go below 90%), then use the following command:

```
$ roary -f ./tutorial -i 90 -v ./gff_all/*.gff
```

These commands will result in a new directory called *tutorial* (as given name in the command), where all result files will be found. You can see the summary statistics in the file named '*summary_statistics.txt*', it will look like this:

summary_statistics.txt