1 Pangenome Construction with Roary

No exercises in this section.

2 The Pangenome explained

2.1 Check your understanding

Q1: The pangenome contains:
a) Only genes present in one isolate in a population
b) All genes from all isolates in a population
c) Only genes present in all isolates in a population

A1: b) All genes from all isolates in a population

Q2: Core genes are:
a) Often important for basic cell functions
b) Present in only a subset of the isolates of a population
c) Often related to drug resistance

A2: a) Often important for basic cell functions

3 Preparing the input data for Roary

3.1

3.2 Check your understanding

Q3: Why do we need to run Prokka?
a) It will perform QC on our data
b) It will annotate our data
c) We don't, Roary can handle fasta files as input

A3: b) It will annotate our data

Q4: Why do we use the –locustag option when we run Prokka?
a) To make it easier to keep track of the output files
b) Because Roary won't work without it
c) To make the Roary results easier to interpret

A4: c) To make the Roary results easier to interpret

4 Performing QC on your data

4.1

4.2

4.3

4.4

4.5 Check your understanding

Q5: Why is it important to QC your data?
A5: If the data is bad going in, the results will be bad coming out.

Q6: You're not getting any core genes when you run Roary. What could be the reason?
A6: Most commonly there is contamination, or the geomes are too fragmented.

Q7: What is the size of the assembly for sample1?
A7: 2096319

Q8: How many contigs are in the assembly of sample1?
A8: 38

5 Constructing a Pangenome with Roary

5.1

5.2

5.3 Check your understanding

Q9: Why do we want to run Roary with MAFFT?
a) Because it's quicker than to run Roary without the -e option
b) To get more accurate results
c) To generate a core gene alignment

A9: c) To generate a core gene alignment

Q10: Why do we use the -p otion?
a) We have to when we use MAFFT
b) To speed up the run
c) To get a nice tree

A10: b) To speed up the run

6 Exploring the results

6.1

6.2

6.3

6.4 Check your understanding

Q11: Approximately how many genes would you expect to see in the summary_statistics.txt file if you are working with a species with a genome size of 5,000,000 bases?
a) 500
b) 5000
c) 50,000

A11: b) 5000

Q12: What does the accessory_binary_genes.fa.newick file provide?
a) A pylogenetic tree ready for publishing
b) Nothing, it is useless
c) A quick insight to the data

A12: c) A quick insight to the data

Q13: For query_pan_genome, what option should you use to get the accessory genome?
a) union
b) intersection
c) complement

A13: c) complement

7 Visualising the results with phandango

7.1 Check your understanding

Q14: What is the name of this gene cluster?
A14: lytN

Q15: Is this a core gene?
A15: No (it is only present in sample 1 and 2)

8 Creating genome assemblies

8.1

8.2

8.3 Check your understanding

Note these metrics may differ slightly.

Q16: What is the size of the assembly?
A16: 2100403

Q17: How many contigs did it assemble into?
A17: 164

Q18: What is the largest contig?
A18: 205299

Q19: What is the N50?
A19: 73737

Q20: Is this a good assembly?
A20: Yes, a reasonable quality assembly.