WORKSHEET for SAGESA BIOINFORMATICS WORKSHOP 2022

Facilitator: Arun Decano

***De novo* Genome Assembly using Shovill**

(1) Select 2 *Escherichia coli* and 2 *Salmonella typhi* paired read libraries (from the ~/Raw_FQs/ directory) you want to work with.

(2) *De novo* assemble your chosen genomes by running the Shovill pipeline as below.

```
shovill --outdir [output_directory_name] --R1 [/path/to/R1.fq.gz/] --R2 [/path/to/R2.fq.gz/]
```

(3) Inspect the output files at the end of each Shovill run.

```
ls [output_directory_name]
```

Assessing the quality of assembled genomes using Quast

(1) Run Quast on your newly assembled genomes. The final assembled genomes from the Shovill run is be labelled as "contigs.fa" in the output directory. Rename the genome with the original sample name.

(2) From any of the summary report files, take note of the following parameters for each assembled genome:

 # contigs (>= 0 bp)
 # contigs (>= 1000 bp)
 # contigs (>= 5000 bp)
 # contigs (>= 10000 bp)
 # contigs (>= 25000 bp)
 # contigs (>= 50000 bp)
Total length (>= 0 bp)
Total length (>= 1000 bp)
Total length (>= 5000 bp)
Total length (>= 10000 bp)
Total length (>= 25000 bp)
Total length (>= 50000 bp)
 # contigs
Largest contig
Total length
GC (%)
N50
N75
L50
L75
 # N's per 100 kbp

(3) Which of the two *E. coli* genomes you assembled is more complete? Why so? Use and cite relevant assembly parameters to compare.

(4) Which of the two *S. typhi* genomes you assembled is more complete? Why so? Use and cite relevant assembly parameters to compare.

AMR Gene Annotation of *de novo* Assembled Genomes using Abricate

(1) Run the Abricate pipeline using CARD and Resfinder as reference databases on the 2 *E. coli* and 2 *S. typhi* genomes you've assembled using Shovill. Save the output to a .csv file as below.

```
abricate  --csv --db resfinder [sample_name.fa] > [sample_name]_resf.csv
abricate  --csv --db card [sample_name.fa] > [sample_name]_card.csv
```

(2) Combine reports aligned with CARD and Resfinder and save each set to a .csv file.

```
abricate  --csv --summary [sample-1_name.fa]  [sample-2_name.fa]  > summary_CARD.csv
abricate  --csv --summary [sample-1_name.fa]  [sample-2_name.fa]  > summary_RESF.csv
```

(3) Plot the number of AMR genes of the genomes you characterised. Only include results with 95-100% identity match with those in the two databases.