Assembly analysis exercises

### Blast exercies

Use environment BTG\_env\_blast

Have a look at the help file for blastn

blastn -h

1.

Use the 16s v3-v4 sequence found in BTG\_2024/precomputed\_data/16s\_phylogeny/16s\_v3\_v4\_reference.fasta as query to identify the 16s v3-v4 region in one of the Listeria assemblies in BTG\_2024/precomputed\_data/spades\_assemblies/spades\_filtered.

Use standard output format, this will give you a visually pleasing overview of the alignment.

2.

Do the same as above, but output a more programmatic friendly format using -outfmt 6

3.

Do it again, but change the output to include data that you think may be relevant. Have a look at <https://www.metagenomics.wiki/tools/blast/blastn-output-format-6> to see output options

4.

The fasta file BTG\_2024/data/databases/Listeria\_monocytogenes\_ariba\_mlst /02.cdhit.all.fa contains allele variants of the 7 mlst genes in the Listeria monocytogenes scheme. Try to identify which alleles are found in SRR27240806

5.

Set up a blast database of all the Listeria assemblies found in BTG\_2024/precomputed\_data/spades\_assemblies/spades\_filtered and identify the 16s v3-v4 region in all of them with a single blast command using there reference found in BTG\_2024/precomputed\_data/16s\_phylogeny/16s\_v3\_v4\_reference.fasta as query.

### Blast solutions

Activate environment

source activate BTG\_env\_blast

To check input/output options for blastn

blastn -h

1. To blast 16s v3-v4 region against an assambly

blastn -query BTG\_2024/precomputed\_data/16s\_phylogeny/16s\_v3\_v4\_reference.fasta -subject BTG\_2024/precomputed\_data/spades\_assemblies/spades\_filtered/SRR27240806.fasta -out blast\_out.txt

1. To do the same, but with a tsv separated output

blastn -query BTG\_2024/precomputed\_data/16s\_phylogeny/16s\_v3\_v4\_reference.fasta -subject BTG\_2024/precomputed\_data/spades\_assemblies/spades\_filtered/SRR27240806.fasta -out blast\_out.txt -outfmt 6

1. Indicate what to include in tsv-output (here we include ID of subject sequence match, percent identity, length of alignment, length of query sequence, biscore and evalue)

blastn -query BTG\_2024/precomputed\_data/16s\_phylogeny/16s\_v3\_v4\_reference.fasta -subject BTG\_2024/precomputed\_data/spades\_assemblies/spades\_filtered/SRR27240806.fasta -out blast\_out.txt -outfmt “6 sseqid pident length qlen bitscore evalue”

1. Blast against all the alleles from the mlst reference fasta at once. But only output perfect matches:

blastn -query BTG\_2024/data/databases/Listeria\_monocytogenes\_ariba\_mlst /02.cdhit.all.fa -subject BTG\_2024/precomputed\_data/spades\_assemblies/spades\_filtered/SRR27240806.fasta -outfmt 6 -perc\_identity 100 -qcov\_hsp\_perc 100 | less

Cat BTG\_2024/precomputed\_data/spades\_assemblies/spades\_filtered/\*.fasta > Listeria \_assemblies\_combined.fasta

Makeblastdb -in Listeria\_assemblies\_combined.fasta -out Listeria\_blast\_DB -dbtype nucl

blastn -query BTG\_2024/precomputed\_data/16s\_phylogeny/16s\_v3\_v4\_reference.fasta -db Listeria\_blast\_DB -out blast\_out.txt

### MLST exercise

Use enviroment BTG\_env\_mlst

Run mlst on all assemblies found in BTG\_2024/precomputed\_data/spades\_assemblies/spades\_filtered

### MLST solution

Source activate BTG\_env\_mlst

mlst BTG\_2024/precomputed\_data/spades\_assemblies/spades\_filtered/\*.fasta > mlst.txt

### Ariba MLST exercise

Use environment BTG\_env\_ariba

Check how to run ariba:

ariba run -h

The database you need to use is located in /home/gebt/BTG\_2024/data/databases/Listeria\_monocytogenes\_ariba\_mlst/

This is the one you need to give as “prepareref\_dir” input

Run ariba on two sets of reads.

A nice clean sample:
/home/gebt/BTG\_2024/data/listeria\_cluster/illumina\_reads/SRR27240806\_R1.fastq.gz /home/gebt/BTG\_2024/data/listeria\_cluster/illumina\_reads/SRR27240806\_R2.fastq.gz

And a special contaminated sample:

/home/gebt/BTG\_2024/data/listeria\_cluster/illumina\_reads\_Qcissues/SRR27240832\_R1.fastq.gz

/home/gebt/BTG\_2024/data/listeria\_cluster/illumina\_reads\_Qcissues/SRR27240832\_R2.fastq.gz

### Ariba MLST solution

Sample 1:
ariba run BTG\_2024/data/databases/Listeria\_monocytogenes\_ariba\_mlst/ BTG\_2024/data/listeria\_cluster/illumina\_reads/SRR27240806\_R1.fastq.gz BTG\_2024/data/listeria\_cluster/illumina\_reads/SRR27240806\_R1.fastq.gz ariba\_out\_ SRR27240806

Sample 2:

ariba run BTG\_2024/data/databases/Listeria\_monocytogenes\_ariba\_mlst/ BTG\_2024/data/listeria\_cluster/illumina\_reads/ SRR27240832\_R1.fastq.gz BTG\_2024/data/listeria\_cluster/illumina\_reads/ SRR27240832\_R1.fastq.gz ariba\_out\_ SRR27240832

## Bakta exercise

Use enviroment BTG\_env\_bakta\_1.9.2

Run bakta on one of the Listeria assemblies. Database is located in BTG\_2024/data/databases/db-light/

## Bakta solution

Activate environment

source activate BTG\_env\_bakta\_1.9.2

Input/output options

bakta -h

Run bakta gene annotation on a genome assembly

Bakta --compliant -d /home/gebt/BTG\_2024/data/databases/db-light/ -o bakta\_output\_folder /home/gebt/BTG\_2024/precomputed\_data/spades\_assemblies/spades\_filtered/SRR27240806.fasta

A little case usage of blastx:

blastx -query /home/gebt/BTG\_2024/data/databases/Listeria\_monocytogenes\_ariba\_mlst/02.cdhit.all.fa -subject /home/gebt/BTG\_2024/precomputed\_data/bakta\_spades/SRR27240806/SRR27240806.faa | less

IN CASE BAKTA DB ISN’T WORKING

Have a look at the bakta github repository at <https://github.com/oschwengers/bakta>

We already have a bakta pre-installed for you, so you don’t have to do that.

Activate the environment env\_bakta\_1.9.2

Download the light database from bakta into a location of your choosing

Also download the