Contamination Assessment

Martin Norling

Uppsala, November 15th 2016

However good your assembly looks, you should also take the time to figure out if this is the correct thing.

One simple measure that can sanity check an assembly is the GC content.

- GC-content varies between about 20% to 75% (*Buchnera aphidicola*, 20.1% - *Anaeromyxobacter dehalogenans, 74.4%*)

- GC content varies within genomes too, so fragments might show variations.

The go-to tool for all things classification is in many cases the **B**asic **L**ocal **A**lignment **S**earch **T**ool (BLAST).

The BLAST algorithm is based on smith-waterman local sequence alignment, optimized for speed by using heuristics. This reduces sensitivity though.

BLAST output

```
                                                              Score       E
Sequences producing significant alignments:                  (Bits)   Value

AP017377.1  Staphylococcus aureus DNA, complete genome, strain...  69.4     1e-09

ALIGNMENTS
>AP017377.1 Staphylococcus aureus DNA, complete genome, strain: OC8
Length=2897106

 Score = 69.4 bits (37),  Expect = 1e-09
 Identities = 37/37 (100%), Gaps = 0/37 (0%)
 Strand=Plus/Minus

Query  1          TGTGAAAGCACCAGTATTTAGTTTTAATAAATTGAAA    37
                  |||||||||||||||||||||||||||||||||||||
Sbjct  1641313   TGTGAAAGCACCAGTATTTAGTTTTAATAAATTGAAA    1641277
```
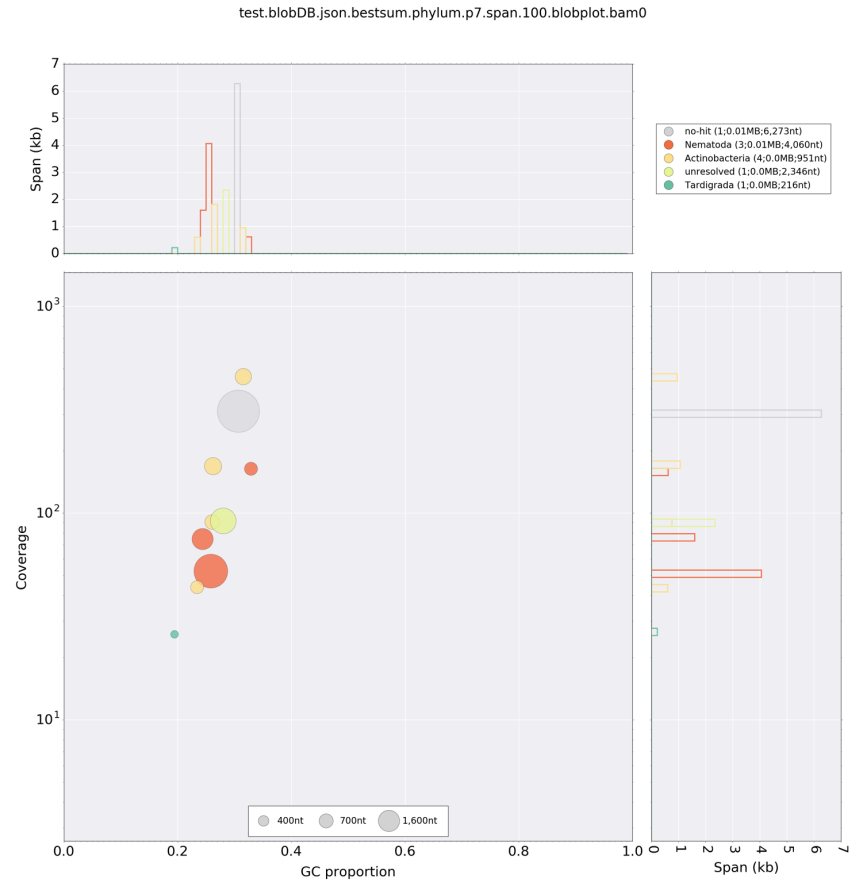
Blobtools

Creates a plot where contigs are taxonomically classified by BLAST, and plotted by Coverage and GC content.

Blobtools!

Blobtools work by creating a "blob database", which can then be used to produce tables and plots.

The blob database is created by parsing information from

- one *assembly* file   ← your assembly

- one (or more) *coverage* file(s) ← Your BAM-file

- one (or more) *taxonomy* file(s) ← BLAST output

- one *nodesDB* file ← Downloaded from NCBI

What to do with a BlobDB

- `blobtools view` can extract information from a blobDB

- `blobtools blobplot` can visualise information from a blobDB in a blobplot

- `blobtools covplot` can visualise information from a blobDB in a covplot

- `blobtools map2cov` can generate a COV file from one or more BAM/SAM/CAS file(s)

- `blobtools seqfilter` can filter FASTA files based on a list of sequence names

- `blobtools taxify` can annotate "taxid" field in Taxonomy files by using a TAXID file

Questions?