Assembly, assembly stats, virulotyping, serotyping

Valeria Michelacci

Bioinformatics training,
July 2019

Assembly (e.g.: SPAdes)

Short sequencing reads

.fastq file

```
@HWI-ST700693:238:B0224ACXX:1:1101:1218:1982
NACACTTGCTTTGGTGACAGCGGGGCATCCTCAAGC
+
#1=DDDDDHAFF?GEFGIIIIIIIIIIIIIIIIIFI
@HWI-ST700693:238:B0224ACXX:1:1101:1161:1986
NGATTTTGACCTCTCCAGTTTCCTCTTAACACTTTG
+
#1:BDFFFGHHHGJJJIIJHIJJJJJJJJJJJJJJJ
@HWI-ST700693:238:B0224ACXX:1:1101:1193:1989
NTATCCAGCCTGCGGTGCTACTTGGTGGAAGAGGAT
+
#1=DDFFFHGHGGJJFGHJJIJJJIEGECHDFHCC?
@HWI-ST700693:238:B0224ACXX:1:1101:1440:1981
NTCAAGAATCCAAGTGGGGCCAGCATAATGTACGCT
+
#1=DDFFFHGHDFDAEGIIFGIICGGHGBFGEFDHI
@HWI-ST700693:238:B0224ACXX:1:1101:1367:1983
NATTAGAACAGATCGCTACTTCGCCCGAAGATACAT
+
#4BDFFFFHHHHHJGIJIJJJIJJJJJJJJJJJJIJ
@HWI-ST700693:238:B0224ACXX:1:1101:1395:1988
NTGGAAACGTTTTTAAACGCGGAGACAGCGTGGAGT
+
#1=DDFFFHCFFHJJJIJJIJJIJJJJGGIFHIGI7
@HWI-ST700693:238:B0224ACXX:1:1101:1285:1994
NCTTTGCTGTATTGACCGTTTGTAGATTTGAATCTT
+
#4=DDFFFHBHHHHIGIJFHIJFGGGIGIHIJIJII
@HWI-ST700693:238:B0224ACXX:1:1101:1632:1989
NTCTATGAATGTTCAAGCGGTAGCTGAGGAGAGTCC
+
```

FastqSize ≈ GenomeSize x Coverage x 2

At least 0.5 GB per genome

Partially assembled genome (contigs)

.fasta file

```
>NODE_1_length_449_cov_4.835189
ATCTTTCGCGCCTTCCAGCTCCAGCCATTCGGAACCGTTCGCCAGAAAACGGGCGTAATC
GGGTAAGACATAGCGCGGTTTGTACGGCGCATGACCTTCAAACATATCGCAGATTACACC
TTCATCCAGCGCGCGGCGGGCTTCGGCAGGAAGCTGTGGGTAAGGCAGATTGTTTTCTGC
TTCCAGTGCCAGAAAATGGCGCTTCTGCTCCGGGCTAAGCACTGGGCTGGTGACAATTTG
CTGGCAACGTTGTTGCAGTGCATTTTCATGAGAAGTGGGCATCTTCTTTTCCTTTTATGC
CGAAGGTGATGCGCCATTGTAAGAAGTTTCGTGATGTTCACTTTGATCCTGATGCGTTTG
CCACCACTGACGCATTCATTTGAAAGTGAATTATTTGAACCAGATCGCATTACAGTGATG
CAAACTTGTAAGTAGATTTCCTTAATTGTGATGTGTATCGAAGTGTGTTGCGG
>NODE_2_length_309_cov_4.686084
ACTGGTCAGTGCGGGTATCCTTGGACAATGGCCGATTGGACGTCTGGCGGATAAGTTTGG
TCGACTGCTGGTGTTGCGTGTTCAGGTCTTTGTCGTCATTCTCGGCAGTATCGCGATGCT
TAGCCAGGCGGCGATGGCCCCAGCGTTATTCATCCTCGGTGCCGCTGGCTTTACGCTATA
TCCGGTGGCGATGGCATGGGCTTGCGAGAAAGTTGAACATCATCAACTGGTGGCGATGAA
CCAGGCCTTACTGTTGAGCTATACTGTGGGAAGTCTGCTTGGCCCGTCATTTACCGCTAT
GCTAATGCAGAATTTCTCCGATAATTTATTGTT
>NODE_3_length_101_cov_3.346535
AGCGCATGAGCGCGCAGCGCCGCCGTTACGTGGTGCATCAGCATGATGTTGGCCGGAGAG
TACAGAGACTCCCCTTCATCCATGATGCCCTCTTTCACCAGCAGTTCTTCAATCATCACC
AGACC
>NODE_4_length_311_cov_3.610933
CATCAACGCTAAAAGCCAGATGACGCAGACCGCAAGCTTCCGGTCGGCTGGGTCGTTCCG
GCGGGAACGGAAATGAGAAAAGCTCAATCACATATTGCCCATTAAGCGCCAAATCCCCTT
TCCATGAGTCGCGCGCTTCGCGATAGACTTCGCTTTGCAGCGTGAAACCAAGAATATCGC
AGTAGAAAGCTTTGCTCACCGCATAATCCGTCGCAATAATCGCAATATGGTGAACCTGTT
TTAAACCCAGCATAACGTCTCCTTTATTTGTTAACAGCACGTTACTCGCCCGGAAGCCGC
TCTGGCAAGTTATCCCGCCATTTTTAGGACTCGTA
>NODE_5_length_186_cov_4.973118
CGAAGATATAAGAAAGCGAACCAGAAAGAATGCCGGAGAACTTCATCAATTCATCACCTG
CATTGAGCAGATTTTGCAGGTTCTCAATAACCGGTAATCCAGCCCCAACGTTGGTGTCAT
AGAGGAATTTACGCCGCGATTTTTCCGCCGCATAACGCAACTGATGGTAGTAATCCATCG
ACGAGGTGTTGGCCTTTTTGTTCGGCGTGA
```

FastaSize for *E. coli* contigs

~5.5 MB

Filter SPAdes repeats/1

Input: contigs file & file with contigs stats

Coverage cut-off ratio:
This is the average coverage ratio cutoff. For example: if the average coverage is 100 and a coverage cut-off ratio of 0.5 is used, then any contigs with coverage lower than 50 will be eliminated.

Repeat cut-off ratio:
This is the coverage ratio cutoff to determine repeats in contigs. For exmaple: if the average coverage is 100 and a repeat cut-off ratio of 1.75 is used, then any contigs with coverage more than or equal to 175 will be marked as repeats.

Length cut-off

Length for average coverage calculation (default = 5000)

Filter SPAdes repeats/2

What does it do?

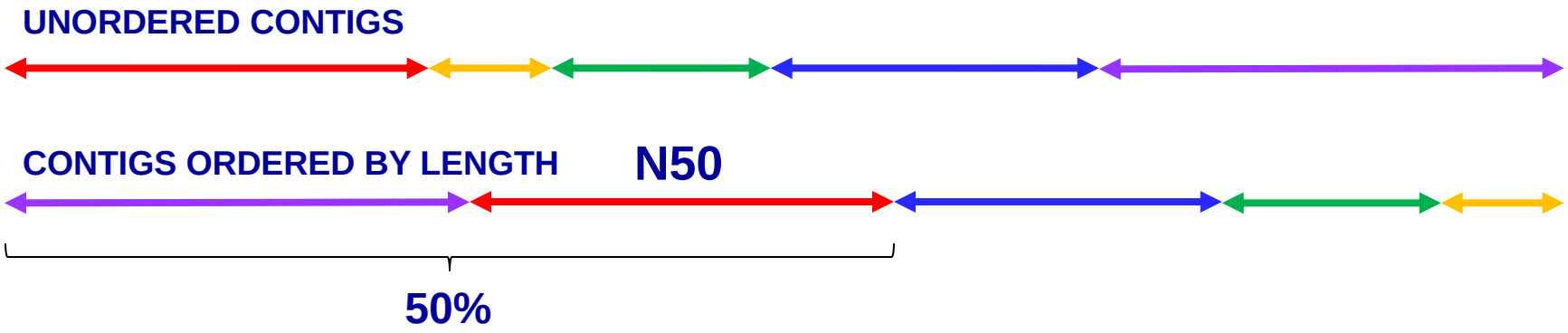
Using the output of SPAdes (a fasta and a stats file, either from contigs or scaffolds), it filters the fasta files, discarding all sequences that are under a given length or under a calculated coverage. Repeated contigs are detected based on coverage.

Output

- **Filtered sequences (with repeats)**
- Will contain the filtered contigs/scaffolds including the repeats. These are the sequences that passed the length and minumum coverage cutoffs.
- For workflows, this output is named **output_with_repeats**
- **Filtered sequences (no repeats)**
- Will contain the filtered contigs/scaffolds excluding the repeats. These are the sequences that passed the length, minimum coverage and repeat cutoffs.
- For workflows, this output is named **output_without_repeats**
- **Repeat sequences**
- Will contain the repeated contigs/scaffolds only. These are the sequences that were exluded for having high coverage (determined by the repeat cutoff).
- For workflows, this output is named **repeat_sequences_only**
- **Discarded sequences**
- If selected, will contain the discarded sequences. These are the sequences that fell below the length and minumum coverage cutoffs, and got discarded.
- For workflows, this output is named **discarded_sequences**
- **Results summary** : If selected, will contain a summary of all the results.

Assembly stats

N50 the **length** of the smallest contig among the set of the largest contigs that together cover at least 50% of the assembly



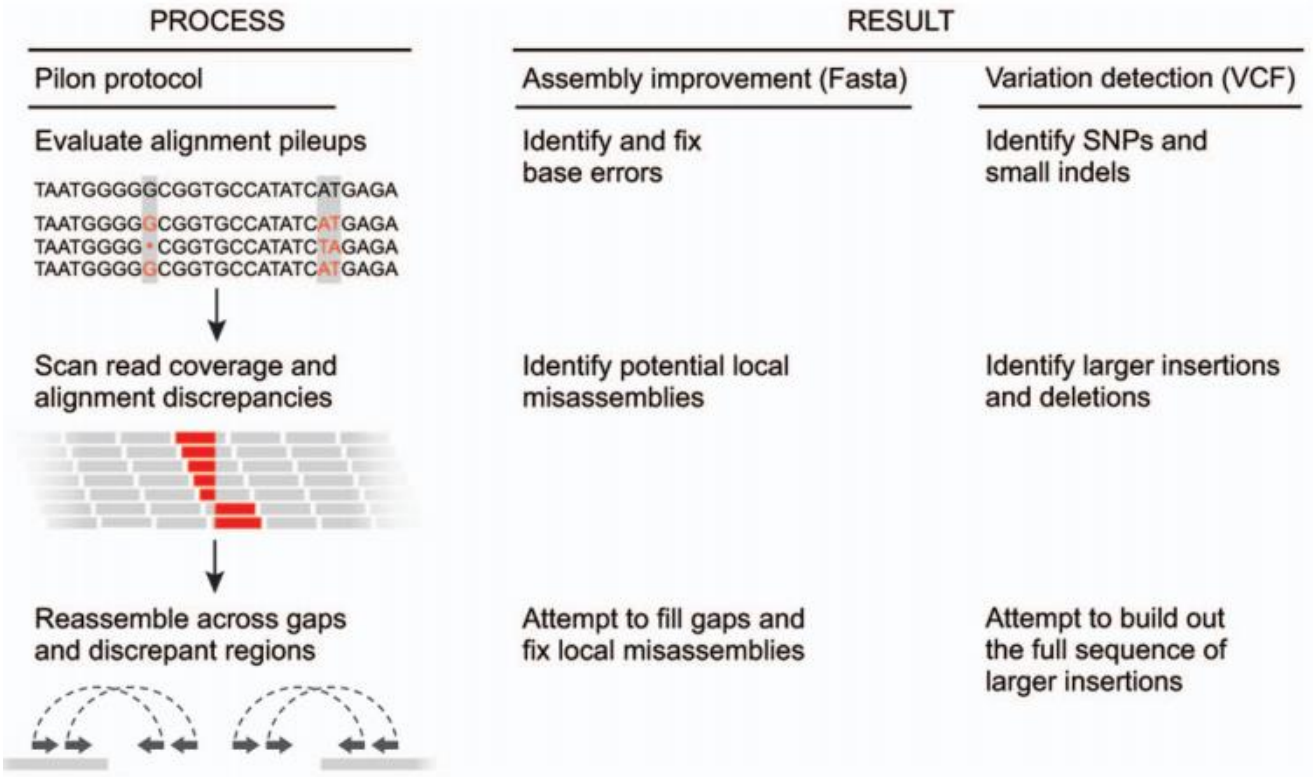
Other intuitive parameters to check:

Maximum contig length

Coverage of the contigs

Consensus length

Pilon – contigs refinement



Realignment of the reads on a «reference sequence»: we use Bowtie2 as alignment tool and the contigs as ref seq

Pilon uses the result of the alignment to improve the assembly: it outputs better assembled contigs

Assembly stats: check bacterial contigs

Contigs Evaluator v1.0 on file dataset_126093.dat

Estimated genome size:	5000000 bp
Assembled nucleotides:	5754440 bp
Estimated coverage:	1.15 x
N. contigs:	818
Average contig length:	7035
Median contig length:	457
Maximum contig length:	165129
N. contigs >= 200 bp:	572 (69.9 %)
N. contigs >= 2,000 bp:	204 (24.9 %)
N50:	58429
NG50:	72678

Assembly stats: Quast