**MIP 280A4: Microbial Sequence Analysis**

**Assembly, part 2, In-class exercise questions**

**Mapping trimmed reads to the entire boa constrictor genome**

1. What percentage of reads mapped to the nuclear boa constrictor genome [the "overall alignment rate"] (1 pt)
2. How does this compare to the percentage of reads that mapped to just the mitochondrial genome? (1 pt)
3. Does it make sense that about this percentage of reads mapped to the entire genome given the source of this library? Explain your answer (1 pt)
4. Describe two possible sources of non-mapping reads? (2 pts)
5. The non-mapping reads should be in new files named SRR1984309\_not\_boa\_mapped.1.fastq and SRR1984309\_not\_boa\_mapped.2.fastq. How many reads are in each of these files? What bash command did you use to answer this question? (2 pts)

**Mapping trimmed reads to the entire boa constrictor genome (ignoring read pairing this time)**

1. What percentage of reads mapped *uniquely* to the boa constrictor genome? (1 pt)
2. What percentage of reads mapped non-uniquely (>1 time) to the boa constrictor genome? (1 pt)
3. What can you say about the regions of the boa constrictor genome to which these reads mapped non-uniquely? (1 pt)
4. Spades produced an output file named scaffolds.fasta. How is this file different from contigs.fasta? (1 pt)
5. In general, what is needed to go beyond contigs to produce a scaffolded assembly? (1 pt)
6. What kmer sizes did Spades use during this assembly? (1 pt)
7. What are the 5 largest contigs produced by the Spades assembler? (1 pt each)

|  |  |  |
| --- | --- | --- |
| **Contig** | **Closest existing Genbank sequence** | **Your interpretation of what this contig is.** |
| NODE\_1… |  |  |
| NODE\_2… |  |  |
| NODE\_3… |  |  |
| NODE\_4… |  |  |
| NODE\_5… |  |  |