Lab Report 1

Zachary Foster

April 15, 2014

In order to test the effectiveness of *de novo* assembly, we simulated short read data from a 100,000bp section of the Haemophilus influenzae Rd KW20 genome [2] provided by our instructor. We explored the attributes of the sequence using `infoseq` from the EMBOSS tools [3]. The total genome was 1,830,138bp in length and has a %GC of 38.15. The custom perl script `sample_suquence.pl` was used to extract the 100,000bp section between locations 100,000 and 200,000 of the genome. We used `grinder` to simulate 100bp paired-end reads with a 500bp insert size and an average coverage of 10. To assemble the simulated reads we used Velvet with a k-mer length of 33bp. The assembly resulted in 37 contigs with an average length of 2722bp and an N50 of 7731bp, determined using `python`. We then used `nucmer` from the MUMMER software package [1] to make a dotplot of the contigs against the genome they were derived from (Figure 1). The contigs were not sorted relative to their corresponding sequence in the part of the genome they were simulated from, so the dotplot is disorganized. However, since straight lines span the entire length of the reference sequence, we likely have assembled most of the original sequence. We determined the average coverage of the contigs using a custom R script `velvet_cp.R`. Most of the contigs had coverage around 8bp and a few had significantly more or less (Figure 2).

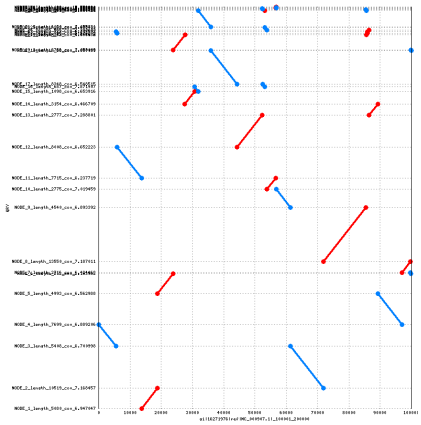

Figure 1: Dot plot of contigs vs. the sequence they were derived from.

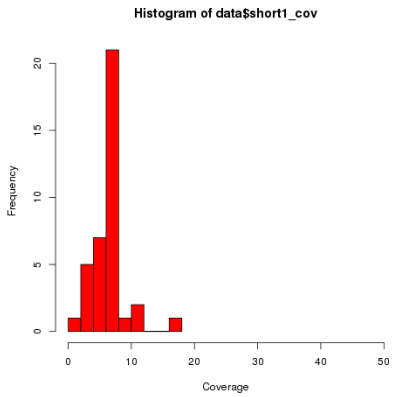

Figure 2: Distribution of contig average coverage.

References

[1] Arthur L. Delcher, Steven L. Salzberg, and Adam M. Phillippy. "Using MUMmer to Identify Similar Regions in Large Sequence Sets". en. In: *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., 2002. ISBN: 9780471250951. URL: `http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1003s00/abstract` (visited on 04/15/2014).

[2] R D Fleischmann et al. "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd". eng. In: *Science (New York, N.Y.)* 269.5223 (July 1995). PMID: 7542800, pp. 496–512. ISSN: 0036-8075.

[3] Sue A. Olson. "Emboss opens up sequence analysis". en. In: *Briefings in Bioinformatics* 3.1 (Mar. 2002). PMID: 12002227, pp. 87–91. ISSN: 1467-5463, 1477-4054. DOI: `10.1093/bib/3.1.87`. URL: `http://bib.oxfordjournals.org/content/3/1/87` (visited on 04/15/2014).