

Supporting Information for:

Choice of reference sequence and assembler for alignment of *Listeria monocytogenes* short-read sequence data greatly influences rates of error in SNP analyses

Arthur W. Pightling¹, Nicholas Petronella² and Franco Pagotto^{1*}

* Corresponding author Franco.Pagotto@hc-sc.gc.ca

¹ Listeriosis Reference Service for Canada, Research Division, Bureau of Microbial Hazards, Food Directorate, Health Products and Food Branch, Health Canada, Ottawa, Ontario, Canada

² Biostatistics and Modelling Division, Bureau of Food Surveillance and Science Integration, Food Directorate, Health Products and Food Branch, Health Canada, Ottawa, Ontario, Canada

Figure S1: Comparison of consensus sequences calculated from alignments of Illumina MiSeq reads to a nearly identical reference with four reference-guided sequence assemblers both before and after read-quality filtering and trimming.

Listeria monocytogenes strain 08-5578 genomic DNA was sequenced twelve times with an Illumina MiSeq benchtop sequencer and the resulting reads were assembled before and after read-quality filtering and trimming with four reference-guided assemblers (BWA, MOSAIK, Novoalign, and SMALT). An *L. monocytogenes* strain 08-5578 chromosome sequence obtained from the National Center for Biotechnology Information archive that differs at three nucleotide positions was used as a reference. The total numbers of false positive sites (a), true positive sites (b), ambiguous sites (c), and gaps (d) present in all consensus sequences were counted. Error bars were calculated as the square root of the standard deviation of each dataset.

Figure S1

