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. Pightling[1], Nicholas Petronella[2] 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

Table S7: Total numbers of false positive sites, true positive sites, ambiguous sites, and gaps detected in consensus sequences calculated from alignments of Illumina MiSeq reads to a non-identical reference with four reference-guided sequence assemblers before and after read-quality filtering and trimming. Total numbers of sites and gaps present in consensus sequences calculated from alignments of twelve sets of *Listeria monocytogenes* strain 08-5578 short-read sequence data with four reference-guided assemblers (BWA, MOSAIK, Novoalign, and SMALT) were counted. An *L. monocytogenes* strain EGD-e chromosome sequence obtained from the National Center for Biotechnology Information archive that is different at 25,347 nucleotide positions was used as a reference. The best values (Trim or No trim) for each aligner within each category are bolded.

	False Positive Sites		True Positive Sites		Ambiguous Sites		Gaps	
	Trim	No trim	Trim	No trim	Trim	No trim	Trim	No trim
BWA	**17726**	18379	**243145**	241542	**6402**	6591	**36049**	37120
MOSAIK	**9455**	9497	**266228**	265767	**3216**	3316	**21453**	22427
Novoalign	2626	**2610**	198216	**199378**	5268	**4956**	102590	**101571**
SMALT	**4280**	4476	**285851**	285815	**14241**	15012	**6411**	6573