# Software

* SNVPhyl - <http://snvphyl.readthedocs.org/en/latest/>
* Contact: Aaron Petkau <aaron.petkau@phac-aspc.gc.ca>

# Files

* **results.xlsx**: The identified matches/non-matches.
* **listeriaDistanceMatrix.xlsx**: A SNV/SNP distance matrix for the listeria dataset.
* **Listeria\_supplemental/input**: Input files (assembled genome + masking/invalid positions) for re-running analysis.
* **Listeria\_supplemental/results/\***: Resulting output files from SNVPhyl for separate runs to include/exclude different genomes.
* **Listeria\_supplemental/results/\*/phylogeneticTree.txt**: Phylogenetic tree in Newick format.
* **Listeria\_supplemental/results/\*/snpTable.tsv**: A table of all identified variants, include those used for further analysis (status set to valid) or those excluded from further analysis.
* **Listeria\_supplemental/results/\*/vcf2core.tsv**: A table of the number of positions excluded (phage or repeats), those evaluated, and those part of the core for each analysis.
* **salmonellaDistanceMatrix.xlsx**: A SNV/SNP distance matrix for the salmonella dataset.
* **Salmonella\_supplemental/input**: Input files (assembled genome + masking/invalid positions) for re-running analysis.
* **Salmonella\_supplemental/results/\***: Resulting output files from SNVPhyl for separate runs to include/exclude different genomes.
* **Salmonella\_supplemental/results/\*/phylogeneticTree.txt**: Phylogenetic tree in Newick format.
* **Salmonella\_supplemental/results/\*/snpTable.tsv**: A table of all identified variants, include those used for further analysis (status set to valid) or those excluded from further analysis.
* **Salmonella\_supplemental/results/\*/vcf2core.tsv**: A table of the number of positions excluded (phage or repeats), those evaluated, and those part of the core for each analysis.

# Methods

## Listeria

1. Genome ASM\_68 assembled using SPAdes. Contigs filtered to remove short contigs (< 1000 bp), low coverage (< 0.33 \* mean coverage). Repeats identified as contigs with > 1.75 \* mean coverage and removed (to avoid mapping to repeat regions when constructing phylogeny).

2. Phages identified on ASM\_68 using PHAST (http://phast.wishartlab.com/) and entered into an invalid-positions file.

3. Reads were run through an instance of SNVPhyl (version 0.2-beta-1) along with the ASM\_68 reference and the invalid-positions file. Parameters set as "min\_coverage=10", "min\_mean\_mapping=30", "alternative\_allele\_proporition=0.75". Reads for ASM\_68 were kept in analysis to check for any differences compared to reference genome.

4. Output files (phylogenetic tree, vcf2core, snpTable) were saved from pipeline.

5. Some post-processing (re-labeling sample names and re-arranging distance matrix rows/columns to match tree) was done.

6. Above steps were repeated for different combinations of samples to include within a phylogenetic tree.

## Salmonella

1. Genome ASM\_31 assembled using SPAdes. Contigs filtered to remove short contigs (< 1000 bp), low coverage (< 0.33 \* mean coverage). Repeats identified as contigs with > 1.75 \* mean coverage and removed (to avoid mapping to repeat regions when constructing phylogeny).

2. Phages identified on ASM\_31 using PHAST (http://phast.wishartlab.com/) and entered into an invalid-positions file.

3. Reads were run through an instance of SNVPhyl (version 0.2-beta-1) along with the ASM\_31 reference and the invalid-positions file. Parameters set as "min\_coverage=10", "min\_mean\_mapping=30", "alternative\_allele\_proporition=0.75". Reads for ASM\_31 were kept in analysis to check for any differences compared to reference genome.

4. Output files (phylogenetic tree, vcf2core, snpTable) were saved from pipeline.

5. There were 3 sites identified where reads for ASM\_31 differed from the assembled reference for ASM\_31. These sites were removed from analysis by labeling as "filtered-reference" in the "snpTable.tsv" file and re-generating a SNV alignment from the remaining sites.

6. The re-generated SNV alignment was used to construct a SNV distance matrix and phylogenetic tree (using phyml).

7. Some post-processing (re-labeling sample names and re-arranging distance matrix rows/columns to match tree) was done.

8. Above steps were repeated for different combinations of samples to include within a phylogenetic tree.