046 – CFSAN SNP Pipeline v2 (CSP2): Accurate estimation of genetic distance among pathogens via rapid assembly alignment

Robert Literman, James Pettengill, and Hugh Rand
Center for Food Safety and Applied Nutrition, 5001 Campus Dr, College Park MD 20742

Introduction

Accurate estimation of genetic distances is crucial for tracking pathogen movement throughout the supply chain and responding effectively to outbreaks. FDA uses the CFSAN SNP Pipeline (CSP) for this purpose, which relies on mapping whole genome sequence (WGS) data against a reference genome. These analyses can be time-consuming, sometimes taking hours to complete depending on the number of isolates analyzed, their sequencing depth, and available computational resources. As WGS becomes more commonplace, isolate clusters will continue to expand and these analytical runtimes will grow in kind.

We present CFSAN SNP Pipeline Version 2 (CSP2), a bioinformatics pipeline for genetic distance estimation that replaces read mapping with MUMmer fast whole-genome alignment, capable of analyzing 100 isolates in around 10 minutes with comparable results to CSP. CSP2 estimates distances from genome assemblies which facilitates the inclusion of isolates lacking WGS data, which CSP cannot analyze. CSP2 is coded in Nextflow, allowing the user to easily adjust parameters to best suit their biological and computational infrastructure needs.

Here we compare the analytical runtimes and genetic distance estimations from CSP and CSP2 for clusters of *Salmonella*, *E. coli*, *L. monocytogenes*, and *Cronobacter*. We find that CSP2 results are strongly correlated with those from CSP but are generated in a fraction of the time (Figure 1).

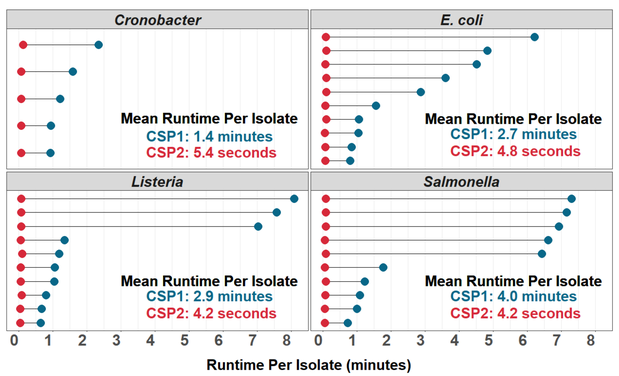

Figure 1. Analytical runtimes for clusters of *Cronobacter* (n=5), *E. coli, Listeria,* and *Salmonella* (n=10). Mean runtimes per isolate were 4.5 seconds for **CSP2** (**red**), compared to 3 minutes for **CSP** (**blue**), corresponding to a ~97% reduction in runtime.

Materials and Methods

A total of 35 clusters of *Cronobacter* (n = 5), *E. coli*, *L. monocytogenes*, and *S. enterica* (n = 10) were randomly selected from the NCBI Pathogen Detection database (median cluster size: 55 isolates). For each cluster, we estimated the number of genetic differences (i.e., single-nucleotide polymorphisms or SNPs) among isolates using CSP and CSP2 and compared results to distances from the NCBI Pathogen Detection database (Figure 2).

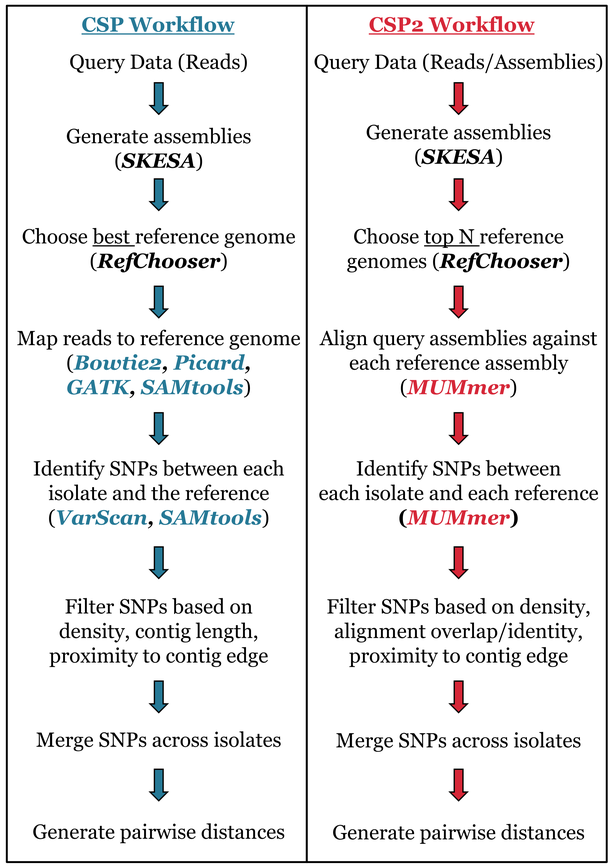

Figure 2. Analytical flow charts for **CSP** (left) and **CSP2** (right). While the pipelines perform similar tasks, in CSP2 the rapid genome aligner MUMmer (**red text**) takes the place of 5 different software packages required for CSP (**blue text**), including the time-intensive read mapping and SNP calling steps.

Results

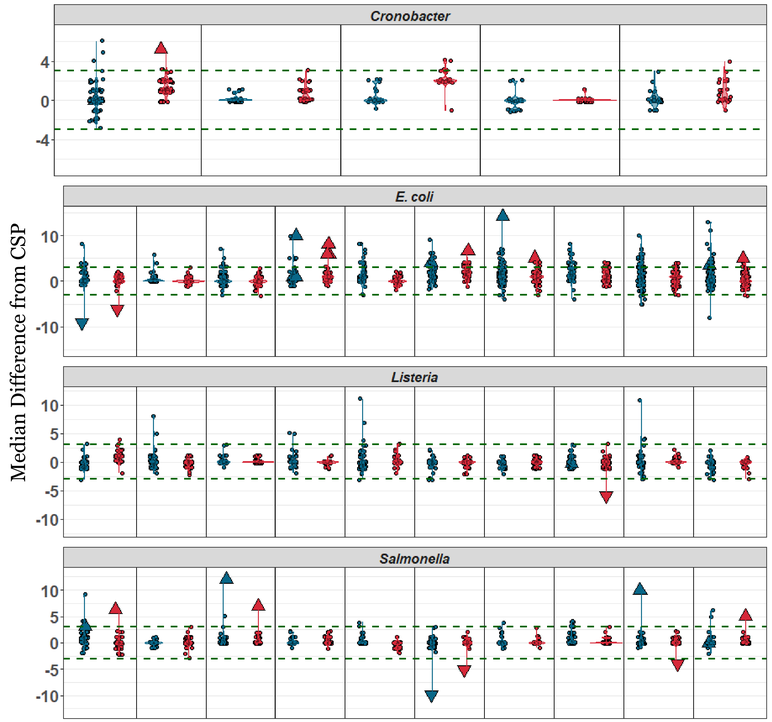

Figure 3: The median deviation between the genetic distances estimated by CSP for each isolate and all other cluster isolates was calculated for distances estimated by **NCBI** (**blue**) and **CSP2** (**red**). Triangles denote isolates where, relative to other cluster isolates, CSP2 estimated SNP distances that were systematically smaller (▽) or larger (△) than CSP. CSP2 genetic distances for the vast majority of isolates fall within 3 SNPs of CSP estimates (green dashed lines), and outlier isolates provide opportunities to fine-tune filtering criteria or to define flagging conditions (e.g., poor assemblies, too many contigs, etc.).

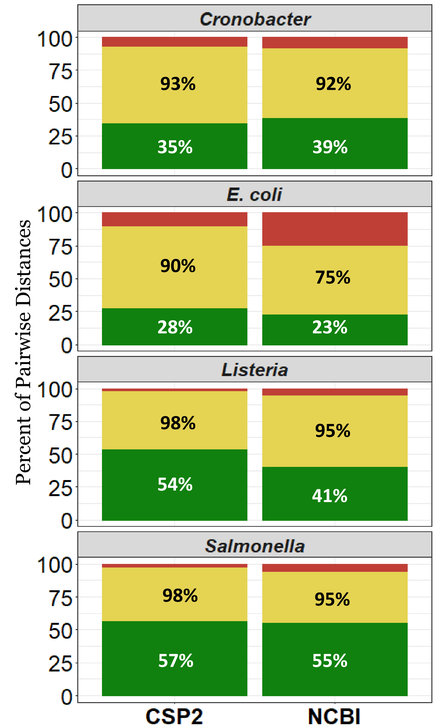

Figure 4: Relative to CSP, pairwise distances estimated by CSP2 or NCBI were categorized into three groups: **Identical**, **Within 3 SNPs**, **4+ SNPs different**. For each species, 90% - 98% of CSP2 estimates were within 3 SNPs of CSP estimates, compared to 75% - 95% of NCBI estimates.

Conclusions and Future Directions

CSP2 genetic distance estimates are strongly correlated with genetic distances generated by CSP (Figures 3 + 4) but are generated in a fraction of the time (Figure 1). In most cases, genetic distance estimates from CSP2 are closer to CSP distances than those from the NCBI Pathogen Detection database (Figure 4). Development of CSP2 is ongoing, with a special focus on:

1. Further refining CSP2 filtering criteria via:
   - Testing of more clusters and species, and incorporation of simulated data
   - Focused investigation of clusters and isolates with the largest deviations from CSP estimates
2. Configuring CSP2 for inclusion on GalaxyTrakr

GitHub