The Effect of Whole Genome Sequencing and Assembly Upon In Silico Serotyping of Salmonella enterica Using the SISTR Platform

John HE Nasha, Catherine Yoshidab, Deyuan Miaoa, Peter Kruczkiewiczb & Eduardo N. Taboadaa

The Laboratory for Foodborne Zoonoses, Public Health Agency of Canada, aGuelph, Ontario, Canada; bLethbridge, Alberta, Canada.

Background
Whole genome sequencing (WGS) has become ubiquitous, inexpensive and widely accepted. It can replace laboratory-based testing in applications such as the typing of pathogens for outbreaks, as well as in surveillance and diagnostics. This applies in particular to serotyping where the reagents are often expensive or difficult to procure. To this end, the Laboratory for Foodborne Zoonoses of the Public Health Agency of Canada has developed the Salmonella In Silico Typing Resource (SISTR) which is in open beta status, in order to rapidly classify serotypes of Salmonella enterica using rapidly assembled WGS data. It is available at [https://fs.corefacility.ca/sistr-app/](https://fs.corefacility.ca/sistr-app/).

SISTR incorporates molecular serotyping based upon specific probes derived from O-chain biosynthesis genes and H-antigen genes [1], combined with a core-genome MLST (cgMLST) analysis module, as an in silico typing assay [2]. In theory, SISTR has the ability to detect over 2,000 serotypes of S. enterica. It contains a database of over 4,000 genomes of S. enterica obtained from public data repositories.

Objectives
WGS occurs under a variety of platforms, sequence coverage, and read types, and with many different assembly methods. Our objective was to examine the effect of these parameters on the accuracy of SISTR’s serotype predictions.

Methods

- 4,331 genomes of serotypes of S. enterica were downloaded from public data repositories into SISTR. The vast majority of these were unfinished sets of contigs. These represented 246 serotypes of S. enterica.
- Available metadata associated with these genomes (e.g. serotype, assembly metrics, assembler, WGS technology & platform, location of isolation, source, host, etc.) are available for the vast majority of these genomes.
- Each of these genomes was subjected to serotype prediction by SISTR. Where the predicted serotype of a genome did not match the serotype in the metadata, the reason for this was investigated further.
- Assembly metrics were entered into SISTR and used to investigate the relationships between genome assembly parameters (N50, contigs >1000 bp, etc.) and various prediction mismatch parameters (see legend to Figures 1 & 2).
- For efficient computation of phylogenies we used a method derived from whole genome MLST (wgMLST) previously described by Sheppard et al. [3] but focusing on core genes (i.e. cgMLST). We used 584 high-quality genomes of S. enterica to derive a set of 330 core genes for the scheme, which we term “cgMLST330”.

Results and Discussion

- Of the 4,189 WGS assemblies of S. enterica that passed QC, the serotype of 3,962 (94.58%) was predicted accurately by SISTR using a combination of O- and H-antigen identification or cgMLST (Figure 1). This included 108 cases in which a serotype variant was identified, and 147 cases where evidence from cgMLST suggests that the serotype indicated in the metadata is likely to be incorrect.
- We examined the effect of genome sequencing parameters (N50 and number of contigs >1000 bp) upon accuracy of serotype prediction (Figure 2). We also examined the effect of presence of full copies of the 330 loci used in cgMLST330 upon accuracy of SISTR to predict the serotype of the Salmonella WGS assemblies (Figure 3).
- Our analyses showed that several factors influenced the accurate prediction of Salmonella serotype, including quality of genome assembly and the number of complete cgMLST loci. In particular:
  - A high proportion of cgMLST330 loci (typically >315) combined with good assembly metrics (high N50 or few numbers of contigs greater than 1000 bp – assuming good sequence coverage) yielded a high accuracy of prediction. However, accurate serotype prediction was also observed for assemblies of lesser quality (i.e. poor N50 or numbers of contigs >1000 bp).
  - Where the number of cgMLST330 loci was low and genome assembly qualities were high, serotype could not be predicted; several genomes with these parameters were subsequently shown to be Enterobacteriaceae.
  - Where the number of cgMLST330 loci were low, and genome assemblies were of low quality, serotype prediction was rarely possible.
  - The most significant source of incorrect serovar prediction was due to genome assemblies with incorrect or absent serotype information provided in the metadata.

Figure 1 – A breakdown of serovar prediction accuracy

98.4% 3.2% 1.8% 0.2% 0.0% 0.0%

Serovar matches (n=4,270)
All mismatches (n=266)
Serovar mismatches (n=401)
Serovar evidence (n=299)
Partial match (n=149)
No match (n=26)

Table 1: Serovar Category Breakdown

Figure 2 – Effect of genome assembly parameters upon serotype prediction

<table>
<thead>
<tr>
<th>Complete cgMLST330 loci</th>
<th>Contigs &gt; 1,000 bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=4,270</td>
<td>100%</td>
</tr>
<tr>
<td>n=243</td>
<td>95%</td>
</tr>
<tr>
<td>n=91</td>
<td>78%</td>
</tr>
<tr>
<td>n=12</td>
<td>50%</td>
</tr>
</tbody>
</table>

Table 2: Serovar category by genome assembly metrics

Figure 3 – Effect of complete cgMLST330 loci and contigs >1000 bp upon serotype prediction

<table>
<thead>
<tr>
<th>Complete cgMLST330 loci</th>
<th>Contigs &gt; 1,000 bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=4,270</td>
<td>100%</td>
</tr>
<tr>
<td>n=243</td>
<td>95%</td>
</tr>
<tr>
<td>n=91</td>
<td>78%</td>
</tr>
<tr>
<td>n=12</td>
<td>50%</td>
</tr>
</tbody>
</table>

Table 3: Serovar category by genome assembly metrics

References