Evaluation of whole genome sequencing for outbreak detection of Salmonella enterica

Leekitcharoenphon, Pimlapas; Nielsen, Eva M.; Kaas, Rolf Sommer; Lund, Ole; Aarestrup, Frank Møller

Published in:
P L o S One

Link to article, DOI:
10.1371/journal.pone.0087991

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Evaluation of Whole Genome Sequencing for Outbreak Detection of *Salmonella enterica*

Pimlapas Leekitcharoenphon¹,²*, Eva M. Nielsen³, Rolf S. Kaas¹,², Ole Lund², Frank M. Aarestrup¹

¹ Division for Epidemiology and Microbial Genomics, National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark, ² Department of System Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark, ³ Department of Microbiology and Infection Control, Statens Serum Institut, Copenhagen, Denmark

Abstract

*Salmonella enterica* is a common cause of minor and large food borne outbreaks. To achieve successful and nearly ‘real-time’ monitoring and identification of outbreaks, reliable sub-typing is essential. Whole genome sequencing (WGS) shows great promises for using as a routine epidemiological typing tool. Here we evaluate WGS for typing of *S. Typhimurium* including different approaches for analyzing and comparing the data. A collection of 34 *S. Typhimurium* isolates was sequenced. This consisted of 18 isolates from six outbreaks and 16 epidemiologically unrelated background strains. In addition, 8 *S. Enteritidis* and 5 *S. Derby* were also sequenced and used for comparison. A number of different bioinformatics approaches were applied on the data; including pan-genome tree, k-mer tree, nucleotide difference tree and SNP tree. The outcome of each approach was evaluated in relation to the association of the isolates to specific outbreaks. The pan-genome tree clustered 65% of the *S. Typhimurium* isolates according to the pre-defined epidemiology, the k-mer tree 88%, the nucleotide difference tree 100% and the SNP tree 100% of the strains within *S. Typhimurium*. The resulting outcome of the four phylogenetic analyses were also compared to PFGE revealing that WGS typing achieved the greater performance than the traditional method. In conclusion, for *S. Typhimurium*, SNP analysis and nucleotide difference approach of WGS data seem to be the superior methods for epidemiological typing compared to other phylogenetic analytic approaches that may be used on WGS. These approaches were also superior to the more classical typing method, PFGE. Our study also indicates that WGS alone is insufficient to determine whether strains are related or un-related to outbreaks. This still requires the combination of epidemiological data and whole genome sequencing results.

Introduction

*Salmonella* is a common cause of infectious disease in human and animals. *Salmonella* is classically divided into species *Shigoni* and *S.enterica*; the latter further divided into more than 2,500 different serotypes [1,2]. It is, however, only a limited number of serovars that are responsible for most infections and in Europe, the most prevalent *S.enterica* serovars isolated from humans are Enteritidis and Typhimurium, responsible for over 75% of the human cases of salmonellosis [3]. *Salmonella* infections can occur as minor and major foodborne outbreaks (major outbreak - an outbreak that attracts intensive publicity). In order to elucidate the epidemiology and implement the control programs, reliable and rapid sub-typing is essential [4,5]. Today, different typing methods are commonly used as a central part of the detection and investigation of *Salmonella* outbreaks, for instance, serotyping, phage typing, pulse-field gel electrophoresis (PFGE) and multilocus variable number of tandem repeat analysis (MLVA) [6–8]. PFGE has been the gold standard for epidemiological investigations of foodborne bacterial pathogens including *Salmonella* [9]. A drawback of PFGE is that it is unable to separate very closely related strains because the low rate of genetic variation does not significantly impact the electrophoretic mobility of a restriction fragment [6]. MLVA has major benefits in epidemiological surveillance of some *Salmonella* [10], but serotype specific protocols are needed for high discrimination.

During recent years the cost of whole genome sequencing (WGS) has decreased dramatically and the technology becomes increasingly available for routine use around the world [4,11]. Moreover, the speed of sequencing is decreasing from several days or weeks to perhaps hours for a bacterial genome in the near future [12]. The combination of low cost and high speed of WGS, opens an opportunity for WGS to become very useful and practical in various bacterial infectious studies [13–15] including the routine use in diagnostic and public health microbiology [12,16]. WGS has also been successfully used for elucidating the evolution of some *Salmonella* sub-types [15,17]. Nevertheless, prior to implementing WGS in routine surveillance, it is essential to evaluate it compared to traditional method and to determine which analytic approaches that might be most useful for a given bacterial species and sub-type.

This study was conducted to evaluate WGS for outbreak typing of *S.enterica*. A collection of presumed epidemiologically related and un-related *S.enterica* strains were sequenced and analyzed using four different bioinformatics approaches. The outcome was evaluated according to the pre-defined expected epidemiological...
whole genome sequencing

Methods

Bacterial Isolates and Molecular Typing

Salmonella strains were derived from the Danish laboratory-based surveillance system of human gastrointestinal infections in 2000–2010. The procedures for isolation, identification, serotyping, antimicrobial susceptibility testing, PFGE and MLVA of the isolates included in this study have been described previously [9,18]. The S. Typhimurium collection consisted of 18 isolates from 6 previously described outbreaks or clusters, primarily defined by MLVA [9,10] and 16 strains that were expected to be epidemiologically un-related to the outbreaks. The outbreaks were selected to cover outbreaks that were restricted in time and location [10] as well as some epidemiologically challenging outbreaks (outbreak 1–3) that lasted several months [9]. The isolates from each outbreak/cluster were selected to include some of the known diversity within these (e.g. based on phage type, MLVA, PFGE as well as the time span of the outbreak). The 16 background strains were selected, so at least two isolates belonged to the same phage type as that of each of the 6 outbreaks. The set of S. Enteritidis consisted of 5 isolates from a couple of outbreaks and 3 background strains. The S. Derby collection comprised 3 isolates from a single outbreak and 2 background strains. Isolate information was included in Table 1.

Whole Genome Sequencing

The total set of forty-seven Salmonella enterica genomes was selected for multiplexed, paired-end sequencing on the Illumina GAIIx genome analyzer (Illumina, Inc., San Diego, CA). The procedures for DNA and library preparation including sequencing in this study have been described previously and according to Hendriksen et al [13]. The paired-end reads had read length at 101 bp. The genomic data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. ERP002633. The raw reads can be accessed online at http://www.ebi.ac.uk/ena/data/view/ERP002633. De novo short read assembly was performed on the set of raw reads using Velvet [19], which is a part of the pipeline available on the Center for Genomic Epidemiology (www.genomicepidemiology.org) [20,21]. The de novo assembly produced contigs with average N50 = 232,749.

A number of publicly available Salmonella genomic data were integrated to this study to make total setting of analyzed data rose to 271 genomes. A set of 39 S. Montevideo genomes was retrieved via Bioproject 61937 with the accession numbers AESR00000000-101 bp. The genomic data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. ERP002633. The raw reads can be accessed online at http://www.ebi.ac.uk/ena/data/view/ERP002633. De novo short read assembly was performed on the set of raw reads using Velvet [19], which is a part of the pipeline available on the Center for Genomic Epidemiology (www.genomicepidemiology.org) [20,21]. The de novo assembly produced contigs with average N50 = 232,749.

A number of publicly available Salmonella genomic data were integrated to this study to make total setting of analyzed data rose to 271 genomes. A set of 39 S. Montevideo genomes was retrieved via Bioproject 61937 with the accession numbers AESR00000000-101 bp. The genomic data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. ERP002633. The raw reads can be accessed online at http://www.ebi.ac.uk/ena/data/view/ERP002633. De novo short read assembly was performed on the set of raw reads using Velvet [19], which is a part of the pipeline available on the Center for Genomic Epidemiology (www.genomiepidemiology.org) [20,21]. The de novo assembly produced contigs with average N50 = 232,749.

A number of publicly available Salmonella genomic data was integrated to this study to make total setting of analyzed data rose to 271 genomes. A set of 39 S. Montevideo genomes was retrieved via Bioproject 61937 with the accession numbers AESR00000000-101 bp. The genomic data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. ERP002633. The raw reads can be accessed online at http://www.ebi.ac.uk/ena/data/view/ERP002633. De novo short read assembly was performed on the set of raw reads using Velvet [19], which is a part of the pipeline available on the Center for Genomic Epidemiology (www.genomiepidemiology.org) [20,21]. The de novo assembly produced contigs with average N50 = 232,749.

A number of publicly available Salmonella genomic data was integrated to this study to make total setting of analyzed data rose to 271 genomes. A set of 39 S. Montevideo genomes was retrieved via Bioproject 61937 with the accession numbers AESR00000000-101 bp. The genomic data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. ERP002633. The raw reads can be accessed online at http://www.ebi.ac.uk/ena/data/view/ERP002633. De novo short read assembly was performed on the set of raw reads using Velvet [19], which is a part of the pipeline available on the Center for Genomic Epidemiology (www.genomiepidemiology.org) [20,21]. The de novo assembly produced contigs with average N50 = 232,749.

A number of publicly available Salmonella genomic data was integrated to this study to make total setting of analyzed data rose to 271 genomes. A set of 39 S. Montevideo genomes was retrieved via Bioproject 61937 with the accession numbers AESR00000000-101 bp. The genomic data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. ERP002633. The raw reads can be accessed online at http://www.ebi.ac.uk/ena/data/view/ERP002633. De novo short read assembly was performed on the set of raw reads using Velvet [19], which is a part of the pipeline available on the Center for Genomic Epidemiology (www.genomiepidemiology.org) [20,21]. The de novo assembly produced contigs with average N50 = 232,749.

A number of publicly available Salmonella genomic data was integrated to this study to make total setting of analyzed data rose to 271 genomes. A set of 39 S. Montevideo genomes was retrieved via Bioproject 61937 with the accession numbers AESR00000000-101 bp. The genomic data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. ERP002633. The raw reads can be accessed online at http://www.ebi.ac.uk/ena/data/view/ERP002633. De novo short read assembly was performed on the set of raw reads using Velvet [19], which is a part of the pipeline available on the Center for Genomic Epidemiology (www.genomiepidemiology.org) [20,21]. The de novo assembly produced contigs with average N50 = 232,749.

A number of publicly available Salmonella genomic data was integrated to this study to make total setting of analyzed data rose to 271 genomes. A set of 39 S. Montevideo genomes was retrieved via Bioproject 61937 with the accession numbers AESR00000000-101 bp. The genomic data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. ERP002633. The raw reads can be accessed online at http://www.ebi.ac.uk/ena/data/view/ERP002633. De novo short read assembly was performed on the set of raw reads using Velvet [19], which is a part of the pipeline available on the Center for Genomic Epidemiology (www.genomiepidemiology.org) [20,21]. The de novo assembly produced contigs with average N50 = 232,749.

A number of publicly available Salmonella genomic data was integrated to this study to make total setting of analyzed data rose to 271 genomes. A set of 39 S. Montevideo genomes was retrieved via Bioproject 61937 with the accession numbers AESR00000000-101 bp. The genomic data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. ERP002633. The raw reads can be accessed online at http://www.ebi.ac.uk/ena/data/view/ERP002633. De novo short read assembly was performed on the set of raw reads using Velvet [19], which is a part of the pipeline available on the Center for Genomic Epidemiology (www.genomiepidemiology.org) [20,21]. The de novo assembly produced contigs with average N50 = 232,749.

A number of publicly available Salmonella genomic data was integrated to this study to make total setting of analyzed data rose to 271 genomes. A set of 39 S. Montevideo genomes was retrieved via Bioproject 61937 with the accession numbers AESR00000000-101 bp. The genomic data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. ERP002633. The raw reads can be accessed online at http://www.ebi.ac.uk/ena/data/view/ERP002633. De novo short read assembly was performed on the set of raw reads using Velvet [19], which is a part of the pipeline available on the Center for Genomic Epidemiology (www.genomiepidemiology.org) [20,21]. The de novo assembly produced contigs with average N50 = 232,749.
Table 1. Epidemiological information for the 47 *Salmonella* genomes used in this study (source: human).

<table>
<thead>
<tr>
<th>ID</th>
<th>Serotype</th>
<th>Received date</th>
<th>Outbreak/Background</th>
<th>Phage type</th>
<th>STTR9</th>
<th>STTR5</th>
<th>STTR6</th>
<th>STTR10</th>
<th>STTR3</th>
<th>MLVA pattern</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>0803T57157</td>
<td>Typhimurium</td>
<td>3/11/08</td>
<td>&gt;1600 cases (Outbreak)</td>
<td>DT135</td>
<td>2</td>
<td>15</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>ERR277220</td>
</tr>
<tr>
<td>0808S61603</td>
<td>Typhimurium</td>
<td>8/6/08</td>
<td>&gt;1600 cases (Outbreak)</td>
<td>DT135</td>
<td>2</td>
<td>11</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td>ERR277226</td>
</tr>
<tr>
<td>0902R11254</td>
<td>Typhimurium</td>
<td>2/10/09</td>
<td>&gt;1600 cases (Outbreak)</td>
<td>DT135</td>
<td>2</td>
<td>11</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td>ERR277229</td>
</tr>
<tr>
<td>000419417</td>
<td>Typhimurium</td>
<td>4/7/00</td>
<td>Background –</td>
<td>U292</td>
<td>2</td>
<td>11</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td>ERR274480</td>
</tr>
<tr>
<td>0207T641</td>
<td>Typhimurium</td>
<td>7/16/02</td>
<td>Background –</td>
<td>U292</td>
<td>2</td>
<td>10</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>ERR277205</td>
</tr>
<tr>
<td>0808F31478</td>
<td>Typhimurium</td>
<td>8/27/08</td>
<td>&gt;200 cases (Outbreak)</td>
<td>DT135</td>
<td>2</td>
<td>15</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>ERR277223</td>
</tr>
<tr>
<td>0903R11327</td>
<td>Typhimurium</td>
<td>3/10/09</td>
<td>&gt;200 cases (Outbreak)</td>
<td>DT135</td>
<td>2</td>
<td>15</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>ERR277222</td>
</tr>
<tr>
<td>0508R6811</td>
<td>Typhimurium</td>
<td>8/24/05</td>
<td>Background –</td>
<td>DT135</td>
<td>2</td>
<td>11</td>
<td>5</td>
<td>10</td>
<td></td>
<td></td>
<td>ERR277218</td>
</tr>
<tr>
<td>0811R10987</td>
<td>Typhimurium</td>
<td>11/28/08</td>
<td>Background –</td>
<td>DT135</td>
<td>3</td>
<td>18</td>
<td>NA</td>
<td></td>
<td>20</td>
<td></td>
<td>ERR277224</td>
</tr>
<tr>
<td>0808R40331</td>
<td>Typhimurium</td>
<td>8/7/08</td>
<td>Background –</td>
<td>DT135</td>
<td>2</td>
<td>11</td>
<td>11</td>
<td></td>
<td>9</td>
<td></td>
<td>ERR277225</td>
</tr>
<tr>
<td>0804R9234</td>
<td>Typhimurium</td>
<td>4/4/08</td>
<td>~100 cases (Outbreak)</td>
<td>DT3</td>
<td>3</td>
<td>20</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td>ERR277222</td>
</tr>
<tr>
<td>0810R10649</td>
<td>Typhimurium</td>
<td>10/2/08</td>
<td>~100 cases (Outbreak)</td>
<td>DT3</td>
<td>3</td>
<td>20</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td>ERR277227</td>
</tr>
<tr>
<td>0901M16079</td>
<td>Typhimurium</td>
<td>1/27/09</td>
<td>~100 cases (Outbreak)</td>
<td>U292</td>
<td>3</td>
<td>20</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td>ERR277228</td>
</tr>
<tr>
<td>0905W16624</td>
<td>Typhimurium</td>
<td>5/15/09</td>
<td>~100 cases (Outbreak)</td>
<td>DT3</td>
<td>3</td>
<td>14</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td>ERR277230</td>
</tr>
<tr>
<td>01110T17035</td>
<td>Typhimurium</td>
<td>10/30/01</td>
<td>Background –</td>
<td>DT3</td>
<td>2</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td></td>
<td></td>
<td>ERR277203</td>
</tr>
<tr>
<td>0505F34633</td>
<td>Typhimurium</td>
<td>5/13/05</td>
<td>Background –</td>
<td>DT3</td>
<td>4</td>
<td>15</td>
<td>8</td>
<td>–2</td>
<td></td>
<td></td>
<td>ERR277213</td>
</tr>
<tr>
<td>0508R6701</td>
<td>Typhimurium</td>
<td>8/10/05</td>
<td>50 cases. Source: restaurant</td>
<td>DT104</td>
<td>3</td>
<td>11</td>
<td>18</td>
<td>17</td>
<td></td>
<td></td>
<td>ERR277214</td>
</tr>
<tr>
<td>0508R6707</td>
<td>Typhimurium</td>
<td>8/5/05</td>
<td>50 cases. Source: restaurant</td>
<td>DT104</td>
<td>3</td>
<td>11</td>
<td>18</td>
<td>17</td>
<td></td>
<td></td>
<td>ERR277216</td>
</tr>
<tr>
<td>0508R6762</td>
<td>Typhimurium</td>
<td>8/23/05</td>
<td>50 cases. Source: restaurant</td>
<td>DT104</td>
<td>3</td>
<td>11</td>
<td>18</td>
<td>17</td>
<td></td>
<td></td>
<td>ERR277217</td>
</tr>
<tr>
<td>0210H31581</td>
<td>Typhimurium</td>
<td>10/24/02</td>
<td>Background –</td>
<td>DT104</td>
<td>3</td>
<td>14</td>
<td>19</td>
<td>21</td>
<td></td>
<td></td>
<td>ERR277206</td>
</tr>
<tr>
<td>0510R6956</td>
<td>Typhimurium</td>
<td>10/19/05</td>
<td>Background –</td>
<td>DT104</td>
<td>3</td>
<td>12</td>
<td>9</td>
<td>25</td>
<td></td>
<td></td>
<td>ERR277219</td>
</tr>
<tr>
<td>0408R5930</td>
<td>Typhimurium</td>
<td>8/26/04</td>
<td>Outbreak</td>
<td>DT12</td>
<td>4</td>
<td>4</td>
<td>14</td>
<td>7</td>
<td></td>
<td></td>
<td>ERR277210</td>
</tr>
<tr>
<td>0408R5960</td>
<td>Typhimurium</td>
<td>8/24/04</td>
<td>Outbreak</td>
<td>DT12</td>
<td>4</td>
<td>4</td>
<td>14</td>
<td>7</td>
<td></td>
<td></td>
<td>ERR277211</td>
</tr>
<tr>
<td>0409R5985</td>
<td>Typhimurium</td>
<td>9/8/04</td>
<td>Outbreak</td>
<td>DT12</td>
<td>4</td>
<td>4</td>
<td>14</td>
<td>7</td>
<td></td>
<td></td>
<td>ERR277212</td>
</tr>
<tr>
<td>0112F33212</td>
<td>Typhimurium</td>
<td>12/21/01</td>
<td>Background –</td>
<td>DT12</td>
<td>4</td>
<td>13</td>
<td>13</td>
<td>8</td>
<td></td>
<td></td>
<td>ERR277204</td>
</tr>
<tr>
<td>0406R5753</td>
<td>Typhimurium</td>
<td>6/30/04</td>
<td>Background –</td>
<td>DT12</td>
<td>4</td>
<td>17</td>
<td>12</td>
<td>7</td>
<td></td>
<td></td>
<td>ERR277207</td>
</tr>
<tr>
<td>0407M287</td>
<td>Typhimurium</td>
<td>7/5/04</td>
<td>Background –</td>
<td>DT12</td>
<td>4</td>
<td>17</td>
<td>12</td>
<td>7</td>
<td></td>
<td></td>
<td>ERR277208</td>
</tr>
<tr>
<td>0407W47858</td>
<td>Typhimurium</td>
<td>7/7/04</td>
<td>Background –</td>
<td>DT12</td>
<td>4</td>
<td>17</td>
<td>12</td>
<td>7</td>
<td></td>
<td></td>
<td>ERR277209</td>
</tr>
<tr>
<td>0508R6706</td>
<td>Typhimurium</td>
<td>8/3/05</td>
<td>Background –</td>
<td>DT12</td>
<td>4</td>
<td>14</td>
<td>9</td>
<td>10</td>
<td></td>
<td></td>
<td>ERR277215</td>
</tr>
<tr>
<td>1004F19825</td>
<td>O4,12; H1: –</td>
<td>4/18/10</td>
<td>Outbreak</td>
<td>DT120</td>
<td>3</td>
<td>12</td>
<td>10</td>
<td>NA</td>
<td></td>
<td></td>
<td>ERR277232</td>
</tr>
<tr>
<td>1005R12913</td>
<td>Typhimurium</td>
<td>5/31/10</td>
<td>Outbreak</td>
<td>DT120</td>
<td>3</td>
<td>12</td>
<td>10</td>
<td>NA</td>
<td></td>
<td></td>
<td>ERR277233</td>
</tr>
<tr>
<td>1006R12965</td>
<td>Typhimurium</td>
<td>6/16/10</td>
<td>Outbreak</td>
<td>DT120</td>
<td>3</td>
<td>12</td>
<td>10</td>
<td>NA</td>
<td></td>
<td></td>
<td>ERR277234</td>
</tr>
<tr>
<td>0909R12120</td>
<td>Typhimurium</td>
<td>9/15/09</td>
<td>Background –</td>
<td>DT120</td>
<td>3</td>
<td>12</td>
<td>9</td>
<td>NA</td>
<td></td>
<td></td>
<td>ERR277231</td>
</tr>
<tr>
<td>1007T82809</td>
<td>O4,5,12; H: –</td>
<td>7/12/10</td>
<td>Background –</td>
<td>DT120</td>
<td>3</td>
<td>14</td>
<td>7</td>
<td>NA</td>
<td></td>
<td></td>
<td>ERR277235</td>
</tr>
<tr>
<td>0905R11565</td>
<td>Enteritidis</td>
<td>5/18/09</td>
<td>Outbreak</td>
<td>PT8</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
<td>ERR277236</td>
</tr>
<tr>
<td>0905R11609</td>
<td>Enteritidis</td>
<td>5/26/09</td>
<td>Outbreak</td>
<td>PT8</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
<td>ERR277237</td>
</tr>
<tr>
<td>0909R12091</td>
<td>Enteritidis</td>
<td>9/4/09</td>
<td>Outbreak</td>
<td>PT8</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
<td>ERR277238</td>
</tr>
<tr>
<td>0910R12287</td>
<td>Enteritidis</td>
<td>10/23/09</td>
<td>Background –</td>
<td>PT8</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
<td>ERR248795</td>
</tr>
</tbody>
</table>
Single nucleotide polymorphisms (SNPs) were identified using a genobox pipeline available on the Center for Genomic Epidemiology (www.genomicepidemiology.org) [34]. The pipeline consists of various freely available programs. Basically, the paired-end reads from each isolates were aligned against the reference genome, S. Typhimurium str. LT2, using Burrows-Wheeler Aligner (BWA) [35]. The average depth coverage was 74. SAMtools [36] ‘mpileup’ command and bedtools [37] were used to determine and filter SNPs. The qualified SNPs were selected once they met the following criteria: (1) a minimum coverage (number of reads mapped to reference positions) of 20; (2) a minimum distance of 20 bps between each SNP; (3) a minimum quality score for each SNP at 30; and (4) all indels were excluded. The qualified SNPs found within Salmonella core genes were ultimately used to make SNP tree because SNPs within the non-core reflect the high proportion of mobile or extra-chromosomal elements, including prophage and genomic islands [14,38].

SNP tree was not only constructed from raw reads but also from contigs or assembled genomes. We used the software package called MUMmer version 3.23 [39]. An application named Nucmer (which is a part of MUMmer) was introduced to align each of contigs to the reference genome. SNPs were determined from the resulting alignments with another MUMmer application called “show-snps” (with options “-CIlrT”). The final set of SNPs was filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.

For each genome, the final qualified SNPs for each genome were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script. If SNP is not found in the reference genome or the base coverage is less than a minimum of 20, it is not considered. The qualified SNPs were filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.

For each genome, the final qualified SNPs for each genome were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script. If SNP is not found in the reference genome or the base coverage is less than a minimum of 20, it is not considered. The qualified SNPs were filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.

For each genome, the final qualified SNPs for each genome were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script. If SNP is not found in the reference genome or the base coverage is less than a minimum of 20, it is not considered. The qualified SNPs were filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.

For each genome, the final qualified SNPs for each genome were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script. If SNP is not found in the reference genome or the base coverage is less than a minimum of 20, it is not considered. The qualified SNPs were filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.

For each genome, the final qualified SNPs for each genome were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script. If SNP is not found in the reference genome or the base coverage is less than a minimum of 20, it is not considered. The qualified SNPs were filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.

For each genome, the final qualified SNPs for each genome were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script. If SNP is not found in the reference genome or the base coverage is less than a minimum of 20, it is not considered. The qualified SNPs were filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.

For each genome, the final qualified SNPs for each genome were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script. If SNP is not found in the reference genome or the base coverage is less than a minimum of 20, it is not considered. The qualified SNPs were filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.

For each genome, the final qualified SNPs for each genome were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script. If SNP is not found in the reference genome or the base coverage is less than a minimum of 20, it is not considered. The qualified SNPs were filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.

For each genome, the final qualified SNPs for each genome were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script. If SNP is not found in the reference genome or the base coverage is less than a minimum of 20, it is not considered. The qualified SNPs were filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.

For each genome, the final qualified SNPs for each genome were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script. If SNP is not found in the reference genome or the base coverage is less than a minimum of 20, it is not considered. The qualified SNPs were filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.
the performance from other approaches (Table 2). However, the pan-genome tree revealed high performance for clustering strains according to their phage type (Figure S2).

**K-mer tree.** K-mer tree was constructed from the frequency profile of k-mers across the selected genomes. The size of k is a sensitive factor for the performance of k-mer tree. A number of various k were evaluated on the set of 34 S. Typhimurium. Figure 3 showed an increase in the percentage of concordance with increasing k value. There was a rise in the concordance to a level of 88% concordance at k = 30. The percentage remained at this level when k > 30 suggesting that this range of k achieved the highest performance of k-mer tree. Therefore, we chose k = 35 to build the final k-mer tree.

Figure 1B showed that k-mer tree gave higher resolution and more reliable tree than the pan-genome tree. However, some outbreak-related isolates were mixed up with the background strains (Figure 1B). Interestingly, the expanded tree in Figure 2B was capable to place the S. Enteritidis outbreak strains into two distinct clusters according to their outbreak groups. The tree also succeeded with clustering S. Derby outbreak strains. Nevertheless, the k-mer tree exhibited 88% and 89% concordance for the set of 34 and 47 isolates respectively (Table 2). The time consuming of k-mer tree was only 5.2 minutes per genome (including the time for assemble process). This is the fastest method compared to the others.

**Nucleotide difference tree.** As a baseline, we implemented a simple approach, the nucleotide difference tree (ND tree), which based on nucleotide difference between a pair of read mapped reference genomes. For the set of 34 S. Typhimurium, the ND tree classified outbreak-related strains into six obvious clusters (Figure 1C) with 100% concordance (Table 2). Thus, the typing ability of the ND tree was superior to the pan-genome tree and the k-mer tree. For the set of 47 genomes, the performance of the ND tree was slightly reduced (Figure 2C). The percentage of concordance decreased from 100 to 91% (Table 2).

**SNP tree.** SNP tree was computed from concatenated qualified SNPs identified from mapping raw reads to core genes of the reference genome [14,38]. From figure 1D, the SNP tree clustered S. Typhimurium outbreak-related strains into six clusters with 100% concordance (Table 2) and furthermore differentiated them accurately from the background isolates. For the set of 47 genomes, SNP tree was able to categorized S. Derby isolates but unable to ultimately classify the S. Enteritidis strains (Figure 2D). The percentage of concordance was dropped from 100 to 91% (Table 2). This is due to the choice of reference genome, SNP tree and ND tree were able to cluster S. Enteritidis outbreak strains.
concordantly by applying publicly available S. Enteritis str. P125109 as a reference genome (data not shown). On average, 4.69 Mb of reference genome was covered by S. Typhimurium genomes meanwhile the reference genome was mapped with 4.63 Mb and 4.60 Mb when adding S. Enteritis and S. Derby.

The performance of SNP tree from raw reads was slightly higher than the one from contigs but constructing the SNP tree from contigs was faster (Table 2). In addition, the identified SNPs were distributed thoroughly across core genes of the reference genome (Figure 4) suggesting that the mutation occurred randomly through the core genes.

Figure 5 revealed that minimum and maximum number of SNP difference within the outbreak strains were significantly less than those numbers between outbreak-related isolates and background isolates. The number of SNP difference between isolates within outbreaks ranged from 2 to 12 except the outbreak 5 (DT12) where the maximum number was relatively high (3-30 SNPs). Besides, the number of days within outbreak strains was unrelated

Table 2. Evaluation results.

<table>
<thead>
<tr>
<th>WGS typing methods</th>
<th>Percentage of concordance</th>
<th>Time (Minutes per genome)</th>
<th>Reference based method</th>
<th>Type of input</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>34 isolates</td>
<td>47 isolates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pan-genome tree</td>
<td>65</td>
<td>64</td>
<td>13</td>
<td>Reference free</td>
</tr>
<tr>
<td>K-mer tree</td>
<td>88</td>
<td>89</td>
<td>5.2</td>
<td>Reference free</td>
</tr>
<tr>
<td>Nucleotide difference tree</td>
<td>100</td>
<td>91</td>
<td>15</td>
<td>Reference-based</td>
</tr>
<tr>
<td>SNP tree (raw reads)</td>
<td>100</td>
<td>91</td>
<td>20</td>
<td>Reference-based</td>
</tr>
<tr>
<td>SNP tree (contigs)</td>
<td>100</td>
<td>89</td>
<td>5.5</td>
<td>Reference-based</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0087991.t002

Figure 2. WGS typing results for the set of 47 genomes. (A) pan-genome tree, (B) K-mer tree, (C) nucleotide difference tree and (D) SNP tree. The labeled color was displayed the same as Figure 1.
doi:10.1371/journal.pone.0087991.g002
Figure 3. Percentage of concordance of k-mer tree on various size of k. This evaluation was conducted on the set of 34 S. Typhimurium. doi:10.1371/journal.pone.0087991.g003

Figure 4. Distribution of SNPs across Salmonella core genes. Black bars represent number of SNPs at each core gene. Red and green small circles are core genes in the form of DNA and protein sequences respectively. The seven black dots represent house-keeping genes for MLST analysis of Salmonella. doi:10.1371/journal.pone.0087991.g004
to the number of SNP difference (Figure S3) and this relation seems to be random.

Comparison with Published Studies

Four publicly available *Salmonella* outbreak dataset were integrated and analyzed by SNP approach. These data comprised of background and outbreak-related strains except *S*. Heidelberg that contained only outbreak strains. An average number of SNP difference or pairwise SNP distance between strains within outbreaks and between outbreak-related strains and background strains were summarized in Figure 6. *S*. Montevideo and *S*. Enteritidis supported our finding that a SNP distance within outbreak strains was less than that between outbreak and background strains. Interestingly, *S*. Agona showed the higher number of SNP difference within outbreak strains and these numbers from two sub-outbreak clusters were higher than the SNP distance between background and outbreak strains. The number of SNP differences between strains within an outbreak is likely to vary for each serotype making it difficult to find the threshold for the case definition of an outbreak.

We reproduced SNP tree and k-mer tree based on 271 genomes from publicly available *Salmonella* genomes together with the genomes under study (Figure S4A and S4B). It was not possible to reproduce the tree by ND tree because most of the published data are assembled genomes and the ND tree was invented primarily for raw reads. The reproduced trees from SNP and k-mer formed distinct clusters according to serotypes. However, combining different serovar strains, k-mer and SNP trees illustrated the similar tree topology of *S*. Typhimurium cluster as they showed in Figure 1B and 1D respectively. Nonetheless, the reproduced SNP tree exhibited less resolution than the tree constructed from the strains with identical serovar as in Figure 1D.

Discussions

The objective of this study was to determine the strengths and drawbacks of WGS using different analytic approaches compared to traditional typing method, PFGE, for retrospectively outbreak typing of *Salmonella*. A set of thirty-four human *S*. Typhimurium strains from six different outbreaks together with background strains plus eight *S*. Enteritidis isolates from two outbreaks and five *S*. Derby strains from a single outbreak were used as test sets. A number of recent studies have already used WGS for epidemiological typing of single outbreaks [13,14,17]. However, these studies have only used SNP analysis and not other analytic procedures. We evaluated different of analytical approaches on the WGS data set and compared to PFGE typing - the gold standard

---

**Figure 5. Minimum and maximum number of SNP difference.** Green shaded bars show the minimum and maximum number of SNP difference between isolates within outbreaks and red shaded bars represent the number of SNP difference between outbreak-related isolates and background isolates. doi:10.1371/journal.pone.0087991.g005
method for epidemiological studies. In our study, WGS based typing using SNP tree and ND tree was able to compete with PFGE for outbreak clustering.

The performance of the four selected WGS based typing methods was validated based on the outbreak related *Salmonella enterica* strains. Pan-genome tree failed to perform accurate clusters as the variation in protein level among the outbreak strains was not appropriate for outbreak typing, although the pan-genome tree showed meaningful clusters corresponding to phage types. This could be due to the content of prophages. The k-mer tree gave the expected clustering but was still unable to employ the complete outbreak typing. Interestingly, the k-mer tree revealed a better clustering when combining *Salmonella* strains from different serovars. This is most likely because the k-mer tree is independent from the reference genome. Another advantage of k-mer analysis is that the frequencies-based approach is much faster. Thus, it is expected to be applicable for both closely and more distantly related strains with very short time consumption for analysis. On the other hand, a deficiency is the loss of information as the huge amount of DNA sequence data is condensed into a vector of k-mer counts. Furthermore, the order of k-mers in compared sequences is neglected [30]. The nucleotide difference tree (ND tree) identified the number of nucleotide difference between a pair of raw read mapped reference genomes rather than identify the difference as SNP. This method gave the results similarly to the SNP tree. Additionally, it is important to note that SNP not being found in the reference genome is considered as not being a variation and the corresponding nucleotide from the reference is expected. This might not always be the right choice. The ND tree does not face this problem, as it does not require the concatenated sequence for alignment. ND tree was found to be somewhat sensitive to its setting. In initial calculations the mismatch score was set to $-1$, and in this tree all *S. Enteritidis* and *S. Derby* strains became identical (data not showed). The final results used a mismatch score as $-3$, which is also the default in the short read alignment program, BWA.

Ultimately, SNP and ND trees were equally superior methods for clustering outbreak related isolates of *S. Typhimurium* (Figure 1C and 1D). As mentioned above, ND tree was sensitive to the parameter settings, while SNP tree failed to categorize strains with different serovars because this method depends heavily on the reference genome and this has to be closely related to the strains investigated for example the reference genome should be at least the same serovar as the strains under study. Using an inappropriate reference genome will cause exceed number of SNPs which affects the final SNP tree for instance the decreasing of the percentage concordance when adding strains with different serovars from the reference genome (Table 2, SNP tree with a set of 47 genomes). In addition, SNP tree constructed from contigs exhibited slightly less concordance than the one from the raw reads. In term of speed, the SNP tree from contigs can be achieved very fast (almost as fast as k-mer tree). It might be an alternative choice of using SNP tree for real-time typing.

We found that the numbers of SNP difference between isolates within outbreaks were very small and ranged from 2 to 12 with an exception for the outbreak 5 (DT12) where the number ranged from 3 to 30 SNP differences. Comparing to publicly available *Salmonella* genomes, the SNP distance between strains within outbreaks was possibly ranged from 4 to 249 depending on serotype suggesting that finding a general threshold to define an outbreak for all *Salmonella* might not be possible. However, these numbers may be useful as an indicator of expected SNP distance in a particular serovar or a sub-outbreak cluster within serovar. Nevertheless, by using a small number of isolates from specific

---

**Figure 6. The pairwise SNPs distance.** This is the average number of SNP difference between strains within outbreaks and between outbreak-related strains and background strains from the four published dataset. doi:10.1371/journal.pone.0087991.g006
outbreaks, this reduced sampling may be introduce some of other variables affecting the predictions. It may take dozens of isolates to determine the actual scope or threshold of an outbreak.

Recent studies support SNP tree as an outbreak surveillance tool such as S. Montevideo outbreak in United States [17,42], S. Enteritidis shell egg outbreak in US in 2010 [25], S. Agona [24] and a 2011 multistate outbreak in the US of S. Heidelberg [22,23]. Nonetheless, the SNP detection and validation need to be improved, and this method needs to be further evaluated in other bacterial pathogens to elucidate the usefulness of using SNP tree. Perhaps, for further pathogens, other approaches might be the most superior beside SNP analysis. In addition, it is especially a need to determine the importance of using different sequencing platforms, different analytic procedures and different reference strains for creating the SNP trees. Moreover, the robustness of this analytical approach for cluster detection in a routine setting has to be evaluated. The fact that the tree topology may give less resolution when new strains are added might cause some problems in the interpretation in a routine setting and over time.

In our study, we were unable to find an association between time (days) of isolation and number of SNP difference between isolates belonging to the same outbreak. This contrasts studies of methicillin-resistant Staphylococcus aureus (MRSA) spreading between humans in hospital community, where the time and number of SNPs are correlated [14]. This might be due to the dissimilarities in the epidemiology of these bacterial pathogens. MRSA transfers from human to human within a hospital, whereas Salmonella has its natural reservoir in various sources, animals and human. Thus, the transmission route of Salmonella to human is indirect and even though two strains are isolated with a given time interval this might not entirely reflect the number of generations that they differ. Nonetheless, this observation is in agreement with that was reported by Okoro et al [43]. They show that the number of days (25–486 days) between isolation of index and recurrent isolates of S. Typhimurium from infected patients had no obvious impact on the numbers of SNP differences accumulated, and suggest the existence of groups of isolates that comprise single clonal haplotypes with virtually no genetic change over time.

The strains included in this study were selected based on detailed epidemiological information as estimated to belong or not belonging to the same outbreak. Since the true epidemiology is not known, it cannot be excluded that strains not being part of an outbreak have been falsely included or that true outbreak strains have been falsely categorized as non-outbreak related. Based on the detailed epidemiological information available and carefully selection of isolates, we do believe that the reference material reflects the true epidemiology and that the methods SNP and ND are superior to the currently used methods for epidemiological typing such as PFGE. However, only time and routine implementation of the new WGS technologies in routine investigations will provide the value of WGS as supporting outbreak detection and control.

It is also important to note that WGS is as all other typing tools to support for decision making and should always be used in combination with epidemiological and/or clinical information. For example, the different phylogenetic trees shown in this study were not meaningful without any support from epidemiological information (the color dots in the trees). Thus, it is essential to combine epidemiological data and whole genome sequencing results.

In conclusion, this study suggests that WGS and analysis using SNP and/or nucleotide difference approaches are superior methodologies for epidemiological typing of S. Typhimurium isolates and might be very successfully applied for outbreak detection. For the very fast but rough result, k-mer tree might meet this requirement with constructing the tree in high speed and giving high accuracy in clade level.

Supporting Information

Figure S1 An UPGMA band based comparison of pulsed-field gel electrophoresis (PFGE) XbaI profiles.
(PDF)

Figure S2 Pan-genome tree with phage typing labels.
(PDF)

Figure S3 The relation between number of days and number of SNP difference among the outbreak strains.
(PDF)

Figure S4A SNP tree constructed from 271 genomes from published data and Salmonella genomes under this study.
(PDF)

Figure S4B K-mer tree constructed from 271 genomes from published data and Salmonella genomes under this study.
(PDF)

Author Contributions

Conceived and designed the experiments: PL, FMA. Performed the experiments: PL, EMN RSK OL. Analyzed the data: PL. Contributed reagents/materials/analysis tools: PL, EMN RSK OL. Wrote the paper: PL, EMN OL FMA.

References

Outbreak Detection of S. enterica