Food Microbiology xxx (2017) 1–7

Contents lists available at ScienceDirect

Food Microbiology

journal homepage: www.elsevier.com/locate/fm

Epidemiological analysis of *Salmonella* clusters identified by whole genome sequencing, England and Wales 2014

Alison Waldram [a, b, c, 1], Gayle Dolan [a, b, c, 1], Philip M. Ashton [d], Claire Jenkins [d, *], Timothy J. Dallman [d]

[a] Field Epidemiology Services, National Infection Service, Public Health England, UK
[b] Field Epidemiology Training Programme, Public Health England, London, UK
[c] European Programme for Intervention Epidemiology Training, European Centre for Disease Prevention and Control, Stockholm, Sweden
[d] Gastrointestinal Bacteria Reference Unit, National Infection Service, Public Health England, London, UK

ARTICLE INFO

Article history:
Received 26 October 2016
Available online xxx

Keywords:
Salmonella
Whole genome sequencing
Epidemiology
Outbreak
Surveillance

ABSTRACT

The unprecedented level of bacterial strain discrimination provided by whole genome sequencing (WGS) presents new challenges with respect to the utility and interpretation of the data. Whole genome sequences from 1445 isolates of *Salmonella* belonging to the most commonly identified serotypes in England and Wales isolated between April and August 2014 were analysed. Single linkage single nucleotide polymorphism thresholds at the 10, 5 and 0 level were explored for evidence of epidemiological links between clustered cases. Analysis of the WGS data organised 566 of the 1445 isolates into 32 clusters of five or more. A statistically significant epidemiological link was identified for 17 clusters. The clusters were associated with foreign travel (n = 8), consumption of Chinese takeaways (n = 4), chicken eaten at home (n = 2), and one each of the following; eating out, contact with another case in the home and contact with reptiles. In the same time frame, one cluster was detected using traditional outbreak detection methods. WGS can be used for the highly specific and highly sensitive detection of biologically related isolates when epidemiological links are obscured. Improvements in the collection of detailed, standardised exposure information would enhance cluster investigations.

1. Introduction

An increasing number of laboratories are adopting whole genome sequencing (WGS) as the method of choice for the molecular typing of infectious organisms of clinical and public health significance. To date, existing literature largely reports the use of WGS to retrospectively delineate previously identified outbreaks or to compare WGS derived typing data with traditional typing methods, such as pulsed field gel electrophoresis (PFGE) and multi-locus variable number tandem repeat analysis (MLVA) to detect linked cases (Deng et al., 2015; Inns et al., 2013; Leekitcharoenphon et al., 2014; Taylor et al., 2015). As the technology matures and becomes increasingly financially viable, the implementation of this approach as a routine public health tool for prospective surveillance seems inevitable. In the United States and Denmark, routine WGS sequencing in both the food and clinical sector for *Listeria monocytogenes* has been implemented (Jackson et al., 2016; Kvistholm Jensen et al., 2016). In England and Wales, WGS has become the first line public health microbiological method for a number of organisms including, *Mycobacterium tuberculosis*, *Salmonella* species, Shiga toxin-producing *Escherichia coli* and *Shigella* species (Ashton et al., 2015; Dallman et al., 2013; Pankhurst et al., 2016).

Since April 2014, all *Salmonella* isolates received by the Gastrointestinal Bacteria Reference Unit (GBRU), Public Health England (PHE) have been sequenced and traditional phenotypic serotyping has been greatly reduced (Ashton et al., 2015). Phage typing and molecular typing by PFGE and MLVA have been withdrawn. For the WGS analysis, a three-step bioinformatics workflow first assigns a sub-species designation using a kmer comparison approach (Ashton et al., 2015). This is followed by clustering into clonal e-BURST Groups (EBGs) as defined by their multi-locus sequence type (MLST), which correlates well with serotype level classification (Achtman et al., 2012). Finally, single nucleotide polymorphisms (SNPs) are identified in the *Salmonella* genome with respect to a

* Corresponding author.
E-mail address: claire.jenkins@phe.gov.uk (C. Jenkins).
[1] These authors contributed equally to the work.

http://dx.doi.org/10.1016/j.fm.2017.02.012
0740-0020/

reference genome representative of each EBG and used to derive phylogenetic relationships between isolates (Ashton et al., 2015).

Prior to 2014, outbreaks of *Salmonella* were detected either by (i) identifying common exposures between cases, (ii) temporally or geographically linked cases of a particular serotype or phage type or (iii) identifying a higher than expected number of cases of a particular serotype or phage type when compared to preceding weeks and to the same time period in preceding years (exceedance algorithm) (Noufaily et al., 2013). Compared to serotype and phage type, WGS data provides a higher level of strain discrimination and this unprecedented level of discrimination presents new challenges with respect to the utility and interpretation of the data. For example, it is important to clarify the criteria used to determine whether or not a case belongs to an outbreak and how the WGS analysis contributes to the case definition. The aim of the study was to determine whether WGS could be used to detect previously unidentified clusters of salmonellosis and if epidemiological links could be demonstrated between these clustered cases. The analysis was used to inform the development of a routine system to identify *Salmonella* clusters in a robust and timely manner to ensure a directed and prompt public health response.

2. Methods

2.1. Bacterial isolates

All *Salmonella* isolates belonging to the seven most commonly identified EBGs in England and Wales submitted from local hospital laboratories to the GBRU at PHE and received between April and August 2014 were included in this study (n = 1510). Duplicate isolates from the same patient and non-human isolates were excluded from the dataset (n = 65). The remaining 1445 isolates included EBG4 *S.* Enteritidis (n = 789), EBG1 *S.* Typhimurium (n = 474), EBG13 *S.* Typhi (n = 60), EBG11 *S.* Paratyphi A (n = 41), EBG5 *S.* Paratyphi B/Java (n = 23), EBG54 *S.* Agona (n = 35) and EBG3 *S.* Newport (n = 23).

2.2. Whole genome sequencing

DNA was extracted from cultures of *Salmonella* species for sequencing on the Illumina HiSeq 2500 instrument as described previously (Ashton et al., 2014). High quality Illumina reads from isolates were mapped to the relevant reference genome defined by EBG (Genbank accession EBG4:AM933172, EBG1:AE006468, EBG13:AE014613, EBG11:CP000026, EBG5:CP000886, EBG54: CP001138, EBG3:CP001113) using BWA-MEM(Li and Durbin, 2010). SNPs were identified using GATK2 (McKenna et al., 2010) in unified genotyper mode. Core genome positions that had a high quality SNP (>90% consensus, minimum depth 10x, GQ $\geq$ 30) in at least one isolate were extracted. The pairwise SNP distance was calculated for each pair of isolates for each EBG. For each EBG, the distance matrix was subjected to single linkage clustering at 250, 100, 50, 25, 10, 5 and 0 SNPs. This hierarchical approach facilitates the generation of a SNP profile or "SNP address" that groups isolates together into clusters of increasing levels of similarity.

2.3. Cluster definitions

In this study, single linkage SNP thresholds at the 10, 5 and 0 level were explored for evidence of epidemiological links between clustered cases. Cases that clustered at the 0 or 5 SNP level also appeared in clusters at the 5 and 10 SNP levels, i.e. isolates belonging to closely related clusters (SNP difference = 0 or 5) were nested within clusters comprising more diverse isolates (SNP difference = 5 or 10). Clusters that nested at different genetic thresholds were grouped together and designated a group identification number (Table 1). Where identified clusters at different SNP thresholds contained the same cases (±1 case), the cluster at the lowest (i.e. 0 SNP threshold) was investigated. Clusters were defined as five or more isolates with a matching SNP address at the 0, 5 or 10 SNP level and received by GBRU within consecutive weeks (i.e. isolates that had a receipt date within one week of at least one other isolate in the same cluster). These criteria were developed with the aim of identifying clusters that were more likely to be linked to the same point source of infection. Isolates within clusters were defined as cases. Controls were defined as isolates with any exposure information and more than 50 SNP differences from any other isolate in the study population.

2.4. Epidemiological investigations

Exposure information was collected for each case using routinely conducted case investigation questionnaires, where available. Exposures were categorised into eight high level categories; travel (foreign and domestic), animal (direct contact and indirect contact via their environment), food (history and eating out) and contact with others that were ill (in and outside of the household). Each high level category was further categorised into lower level categories if appropriate information was available e.g. country visited or species of animal. Known outbreaks detected using traditional surveillance methods were identified by contacting local health protection areas. Clusters were described by size, duration and the case demographics (age, sex and geographical distribution) of the cases within them. Clusters were analytically investigated if they contained five or more isolates with available exposure information. Controls were randomly selected for the analysis of each cluster. The sample size calculation for analytical investigation was based on a cluster of 10 cases, returning an odds ratio (OR) of 5 or more with an exposure between 20 and 80% (based on a literature review of previous Salmonella outbreaks) (Centers for Disease Control and Prevention (CDC), 2014, 2013; Chironna et al., 2014; Gicquelais et al., 2014; Kunwar et al., 2013; Zenner et al., 2013). To ensure an efficient design, i.e. minimising sample size whilst maximising power, two controls per case were used.

2.5. Statistical analysis

Univariable analysis was undertaken to generate hypotheses about potential exposures. ORs, associated 95% confidence intervals and p-values (using Fisher's exact test) were calculated to estimate the strength of any association between exposure and illness. Clusters in which 75% or more of cases had travelled abroad were classified as foreign travel associated.

3. Results

3.1. General description of all the clusters

Analysis of the WGS data organised 566 of the 1445 isolates into 32 clusters of five or more isolates (Fig. 1). Of these 32 clusters, 18 were within the 10 SNP threshold, 16 were within the 5 SNP threshold and 11 were within the 0 SNP threshold, and were categorised into 17 groups of nested clusters. Clusters arose within three serotypes; *S.* Typhimurium (n = 6), *S.* Enteritidis (n = 24) and *S.* Typhi (n = 2) (Table 1). No clusters were identified during the study period for serotypes *S.* Agona, *S.* Newport, *S.* Paratpyhi A or *S.* Java. The median number of isolates per cluster was 7 (IQR: 5–18

Please cite this article in press as: Waldram, A., et al., Epidemiological analysis of *Salmonella* clusters identified by whole genome sequencing, England and Wales 2014, Food Microbiology (2017), http://dx.doi.org/10.1016/j.fm.2017.02.012