



Using SNP addresses for *Salmonella Typhimurium* DT104 in routine veterinary outbreak detection




Original Paper

Cite this article: Bettridge JM, Snow LC, Tang Y, Petrovska L, Lawes J and Smith RP (2023). Using SNP addresses for *Salmonella Typhimurium* DT104 in routine veterinary outbreak detection. *Epidemiology and Infection*, **151**, e187, 1–14
<https://doi.org/10.1017/S0950268823001723>

Received: 27 May 2023
Revised: 12 September 2023
Accepted: 17 September 2023

Keywords:
genomic typing; molecular epidemiology; *Salmonella (Typhimurium)*; surveillance; veterinary epidemiology

Corresponding author:
R. P. Smith;
Email: Richard.P.Smith@apha.gov.uk

J. M. Bettridge^{1,2} , L. C. Snow¹, Y. Tang³, L. Petrovska³ , J. Lawes¹ and R. P. Smith¹ 

¹Department of Epidemiological Sciences, Animal and Plant Health Agency, Weybridge, UK; ²Natural Resources Institute, University of Greenwich, Chatham, UK and ³Department of Bacteriology, Animal and Plant Health Agency, Weybridge, UK

Abstract

SNP addresses are a pathogen typing method based on whole-genome sequences (WGSs), assigning groups at seven different levels of genetic similarity. Public health surveillance uses it for several gastro-intestinal infections; this work trialled its use in veterinary surveillance for salmonella outbreak detection. Comparisons were made between temporal and spatio-temporal cluster detection models that either defined cases by their SNP address or by phage type, using historical data sets. Clusters of SNP incidents were effectively detected by both methods, but spatio-temporal models consistently detected these clusters earlier than the corresponding temporal models. Unlike phage type, SNP addresses appeared spatially and temporally limited, which facilitated the differentiation of novel, stable, or expanding clusters in spatio-temporal models. Furthermore, these models flagged spatio-temporal clusters containing only two to three cases at first detection, compared with a median of seven cases in phage-type models. The large number of SNP addresses will require automated methods to implement these detection models routinely. Further work is required to explore how temporal changes and different host species may impact the sensitivity and specificity of cluster detection. In conclusion, given validation with more sequencing data, SNP addresses are likely to be a valuable addition to early warning systems in veterinary surveillance.

Introduction

Non-typhoidal salmonella (NTS) remains one of the most significant causes of foodborne disease worldwide. In Europe, it is second only to *Campylobacter* as a cause of gastro-intestinal infection and an important cause of foodborne outbreaks [1]. Whilst some of the approximately 2,600 serovars are host-adapted and cause extra-intestinal disseminated infections, most have a broad host range and circulate in multiple vertebrate species, causing localised gastroenteritis in hosts, but rarely invasive disease [2]. *Salmonella enterica* serovar Typhimurium (*S. Typhimurium*) has traditionally been considered the archetypal broad host range serovar, although a growing body of evidence, initially from sero-, bio-, and phage typing, and now from molecular methods, indicates that the Typhimurium serovar could more correctly be considered as a collection of pathovariants that differ significantly in their degree of host adaptation [3]. Some pathovariants appear to have become host-adapted to wild avian species by a convergent evolutionary process akin to that seen in other host-adapted serovars, such as *S. Typhi* and *S. Choleraesuis* [4]. However, these appear to have evolved as a distinct phylogenetic clade from a common ancestor, and the majority of variants associated with known livestock epidemics in the last 30 years have evolved as separate lineages from a basal ancestral broad host range variant. These include the definitive phage types (DT)104, U288 and the most recent monophasic *S. Typhimurium* (*S.* 4, [5],12:i:-) sequence type (ST) 34, representing successive waves of dominant clones that have accounted for up to 60% of human infections for several years before a new strain arises [4]. *S. Typhimurium* DT104 is estimated to have first arisen around the middle of the twentieth century, acquiring multiple drug resistance genes in the 1970s and disseminating widely throughout Europe and subsequently Asia and the Americas in the 1980s and 1990s [5]. The epidemic of human infections with DT104 appeared to peak in the late 1990s in England, Wales, and Ireland, although it was somewhat later in other countries, including Scotland [6]. Although now overtaken by the monophasic *S. Typhimurium* ST34, which emerged in pig populations and spread globally [7], likely due to enhanced resistance to heavy metals over other variants [8, 9], DT104 continues to circulate in animal populations, causing sporadic outbreaks in both animals and humans. A zoonotic outbreak in 2016, first identified in sheep, cattle, and horses in Anglesey, North Wales [10], presented the basis for this study.

© Crown Copyright - Crown copyright, 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (<http://creativecommons.org/licenses/by-nc-nd/4.0>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use and/or adaptation of the article.

Advances in typing technologies are transforming surveillance and outbreak investigations in animal and human health. In addition to providing valuable information about the population structure of salmonella, whole-genome sequence (WGS) data are becoming more widely used in epidemiology, providing high levels of sensitivity and specificity for detecting clusters of gastro-intestinal infections [11–13] and source attribution for foodborne pathogens [14–17]. The UK Health Security Agency (UKHSA; formerly Public Health England) first adopted WGS typing of salmonella in April 2014, and routinely from 2015, using multi-locus sequence typing (MLST) to assign isolates to a sequence type [18]. This was complemented by a method that introduced a strain-level nomenclature, known as the SNP address, based on single-nucleotide polymorphisms (SNPs) across the whole genome, providing fine-level typing [19]. The method uses hierarchical single-linkage clustering on a matrix of pairwise SNP differences relative to a reference genome. Clustering is performed at seven descending levels of SNP difference (250, 100, 50, 25, 10, 5, and 0 SNPs) to generate a seven-integer code, where each integer identifies the group membership at the corresponding level [20]. UKHSA now routinely uses the SNP address as the primary method to prospectively monitor for clusters of cases of gastro-intestinal disease that are microbiologically linked, with automated methods that extract 5-SNP-level single-linkage clusters to assess for outbreak investigation [19]. The 5-SNP threshold has been demonstrated to give a high likelihood that cases relate to a common source, whilst analysis at the 10-SNP threshold may be useful to uncover wider epidemiological links [21].

The Animal and Plant Health Agency (APHA) began trialling the UKHSA SnapperDB software [20, 22] to assign SNP addresses to strains in 2018, as part of joint One Health activities, and now regularly implements it as an additional tool for characterising salmonella for surveillance and outbreak response. Salmonella surveillance in animals differs by several important aspects from that in human health, including active surveillance through the National Control Programmes; dealing with multiple animal species with different host-adapted salmonellae; and the fact that salmonella infections may persist for prolonged periods on livestock holdings, resulting in diagnostic samples from single premises being received at intervals of potentially months or years. Bearing this in mind, this study was undertaken with a view to exploring the use of the SNP address in outbreak detection in UK veterinary surveillance. In 2016, there was a human outbreak of *S. Typhimurium* DT104 (designated t5:459), where veterinary salmonella isolates identified through normal surveillance were sequenced and had SNP addresses generated at UKHSA and confirmed to cluster genetically with each other and with human isolates. Veterinary investigations verified epidemiological links between several premises from which isolates originated. APHA was subsequently able to define its own SNP address for the outbreak and generate a phylogeny from further current and historic sequenced isolates to place the outbreak strain within the context of the population of UK *S. Typhimurium* isolates from animal species. This study examines how two readily available early detection models functioned when applied to incidents defined by SNP address, to help adapt or develop early detection systems for the near future. The t5:459 outbreak was used to examine the timeliness of cluster detection using the SNP address in comparison with the phage-type definition.

Methods

Sample selection for sequencing and SNP address generation

Under the Zoonoses (Amendment) (England) Order 2021, laboratory isolations of salmonella from British livestock must be reported to APHA, generally followed by the submission of samples to APHA laboratories. However, only a small subset was sequenced each year on a risk basis; additional isolates were sequenced for specific research projects or during outbreaks. Three previous projects generated sequencing data that were suitable for this research, with all sequencing carried out between 2015 and 2019. Project RDOZO347, described by Mellor [23], contributed 406 sequences, investigated the historical context of DT104, and sequenced isolates from 1992 to 2016. Isolate sources included livestock, companion animal, and environmental samples and were either DT104 or a related phage type (including DT104b, U302, DT120, and DT12 isolates). Most isolates with a known sequence type were ST19, plus five ST34 isolates.

Sequence data for a further 39 isolates (collected from 2014 to 2017) were generated under APHA's SE553 Project [24], and a further 55 sequences were taken from the CR2000F project (2017 to 2019). These projects focused on generating sequencing data for isolates known or suspected to have a link to the t5:459 outbreak; the internal project reports contain sensitive information and are not in the public domain. Ninety isolates were DT104/ST19; two isolates were U302/ST19; and of two further DT104 isolates, one was ST34 and one was not typable. Isolates encompassed all sources as above, but with a heavy bias towards cattle and sheep, the main species implicated in the outbreak (Supplementary Figure S1).

DNA extraction and sequencing

Genomic DNA of all isolates was extracted with the KingFisher MagMAX™ CORE instrument and the MagMAX™ CORE Nucleic Acid Purification Kit (Thermo Fisher Scientific, UK) from 270 µL of overnight cultures in LB broth, following the manufacturer's instruction. DNA concentrations were normalised before sequencing library generation with the Nextera XT Kit following the manufacturer's instructions (Illumina UK). Libraries were normalised and pooled before running on an Illumina MiSeq or NextSeq instrument to generate 150 base pair paired-end reads. Sequencing data were demultiplexed, and sequencing adapters were removed with bcl2fastq software (Illumina UK) to generate per-sample raw data fastq files.

Sequencing data analyses

Sequenced isolates were analysed with an APHA in-house *in silico* serotyping pipeline [25] to confirm the serovar. Essentially, the Salmonella bioinformatics pipeline (publication in preparation) runs the following tools: fastp, FASTQC, KmerID, and Quast (quality control software); Shovill (genome assembly); SeqSero, MOST, and SISTR (serotype determination); and SRST2 for each sample. The combined outputs from these tools generated sequencing quality metrics, a genome assembly (and the associated metrics), 7-gene MLST profile, and an assigned serotype according to three different tools. Finally, a consensus serotype Typhimurium was assigned based on the output from the three approaches. SNP address generation was performed at the clonal/eBURST group level using SnapperDB [20]. A pairwise distance matrix of SNP distances was calculated and used to generate an isolate-level

hierarchical clustering nomenclature – the SNP Address – using *S. Typhimurium* AE006468 as the reference genome. The seven levels of the SNP address were then used to create a simplified pairwise matrix and dendrogram (Figure 1).

Data processing

To obtain location data and produce the data set of phage-type isolates, all records of *S. Typhimurium* serovars identified between 1 January 2002 and 30 September 2019 were extracted from the APHA salmonella database (25,335 records). Before 2002, location data were not routinely collected, and thus, sequenced isolates from 1992 to 2001 were excluded from most analyses (exceptions are noted below.) The data fields queried from the database were the date of sample collection, the location where the sample was obtained, the species or source from which the sample was collected, and the reason for sample submission and phage type. Location data were one or more of the following: business name, address, postcode, grid reference, or, for agricultural premises only, the holding reference number. Records were sequentially removed

from the data set if they were missing phage type ($n = 9,845$), date ($n = 32$), were duplicate records ($n = 3,080$), or did not contain sufficient location data to establish the postcode or the postcode was not within the mainland UK ($n = 2,534$). Among sequenced isolates ($n = 500$), all had complete data for phage type and date, but 82 were removed at the location data step; of which, all but two were collected before 2003.

An online tool (<https://gridreferencefinder.com/postcodeBatchConverter/>) generated an approximate latitude and longitude for each record using the postcode centroid. Unique identifiers were created for each premises to differentiate those that shared postcodes. Where postcode sharing was due to the simultaneous presence of multiple premises within a small area, coordinates were manually adjusted to the true location of each premise to avoid overlap. Premises that sequentially shared a location due to a change in ownership or use were counted as separate premises. Overall, 3,269 postcodes and 3,349 premises were identified.

Data were next aggregated to the case level. Sixty per cent of premises had only one isolate (range 1–409). Our definition of a ‘case’ was similar to the definition of ‘incident’ used in salmonella

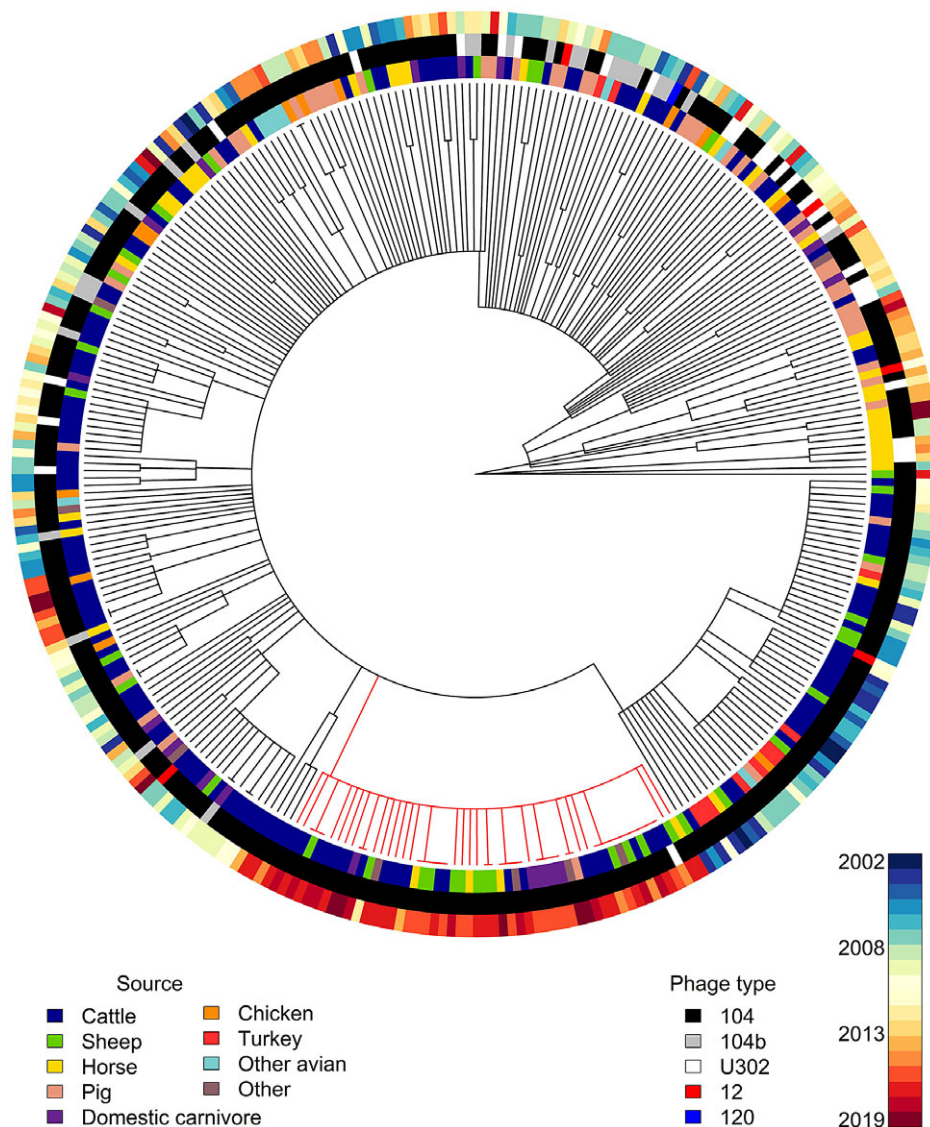


Figure 1. Dendrogram of SNP addresses from 329 incidents. The outbreak clade, which shares SNP address with the t5:459 human outbreak strain, is shown in red.

surveillance [26] as either the first and subsequent isolations of a particular definitive phage type from an animal, group of animal, or environment on a premises (DT cases; $n = 4,165$), or as the first and subsequent isolations of a particular 5-SNP address (SNP cases; $n = 329$). However, unlike the usual definition of an incident, no constraints were placed on the time elapsed between isolates belonging to the 5-SNP group or phage type to count them as the same case, due to the sporadic nature of available genetic data. For all subsequent related isolates, the time since the first detection of the DT or SNP case on the premises was calculated. At this stage, premises where animals were not permanently kept (e.g. abattoirs, butchers, laboratories, and referral veterinary clinics) were identified with multiple submissions. These could not be classified as single cases, as the animals tested would likely have originated from a variety of sources, but could potentially trigger apparent spatial clusters; therefore, it was decided to remove these records from the data set entirely ($n = 56$, including two sequenced isolates). Similar premises with only a single submission were retained, as removal of all such records would remove a larger proportion of companion animal isolates compared with the tiny fraction of livestock samples collected from such premises, and companion animal isolates were already relatively low. Although this means that the spatial data for these samples are not necessarily the same as that of the animal source, nevertheless it can give an indication of epidemiologically linked premises.

Sources were grouped into nine categories: cattle, sheep, horses, pigs, and domestic carnivores (incorporating cats, dogs, and ferrets); chickens, turkeys, and other avian species (incorporating ducks, quails, psittacines, and groups of mixed avian species); and Other (incorporating environmental samples, reptiles and amphibians, and samples where the source was defined as 'Other' in the original database).

Defining outcomes of interest

Based on UKHSA's previous published work [21] and preliminary examination of the data set, it was decided to test outcomes defined by both the 5-SNP and 10-SNP clustering thresholds on the data set of sequenced isolates. A 12-month rolling average was calculated, for each threshold level, of the proportion of cases contributed by each SNP group. As a comparator, analyses run using the whole cleaned data set of phage types also used two levels of outcome. Firstly, cases were defined as belonging to the DT104 group if they comprised any DT104, DT104 variant, or U302 phage type; secondly, analyses were run separately for DT104 and U302 phage types only.

Statistical methods

The Farrington method [27] is a log-linear model that calculates an expected value of cases for the current time period based on historical data, and a threshold above which an observed count is declared to be unusual. It can adjust for overdispersion, seasonality, secular trends, and past outbreaks and is currently used in the APHA Early Detection System (EDS) [28] to detect potential salmonella outbreaks using serotype, phage type, and antimicrobial resistance patterns. Models were run in R using the `farringtonFlexible` function from the package 'surveillance' [29], using 5 years of historical data, with a half-window of 1 month. A two-thirds power transformation was included to adjust for overdispersion where there were low counts of cases. The model has a default threshold of no alarm if there were fewer than five cases in 4 weeks. Given the

very low numbers of any given SNP address overall, the model was run both using the default threshold and a lower threshold of no alarm if there were fewer than three cases in 12 weeks. Farrington models for the SNP cases included pre-2002 cases, due to the overall low number of isolates with sequence data. For the phage-type models, only the cleaned post-2002 data were used, as it was not possible to determine isolates originating from the same case without accompanying location data.

SaTScan™ [30] is free software that includes a prospective space-time permutation scan statistic, designed to scan a defined geographical area to detect outbreaks of any scale or location that are still in existence at the end of the scan period [31]. We elected to use a Bernoulli model, where cases and controls are defined, and which searched for circular spatial clusters where the expected risk of being a case was significantly greater within the cluster than outside of it. Although the software can conduct a prospective scan using only cases, it was felt that it was important to consider the underlying population, as the sample collection was geographically biased due to the outbreak investigation contributing isolates over and above those arising from routine passive surveillance. For models using the sequenced data, cases were defined as the SNP group of interest and controls as any other SNP group. Only sequenced isolates from 2002 onwards were used, which could be matched with location data in the cleaned data set. For the phage-type models, controls were any phage type other than DT104 (including variants) and U302. To simulate the kind of regular scanning that would be undertaken as part of routine surveillance, sequential data sets were created for each quarter (3 months) from March 2003 to September 2019 containing data on all cases and controls up to that time point, and the models were applied to each data set in turn. Space-time clusters identified by each model were scrutinised with respect to the cases allocated to that cluster, to determine whether they were novel or had been detected by previous models. To be defined as a new space-time cluster, at least 50% of the cluster cases had to have not been allocated to any space-time cluster in any earlier model. Space-time clusters were allocated an identifying letter: Those with spatial overlap, but no temporal continuity with an earlier cluster – that is the space-time cluster was not detected by sequential models, but had one or more intervening quarters where it was absent – were allocated the same letter with a subsequent number to differentiate them (such as β and $\beta:1$). Clusters were also mapped and tabulated, to cross-check their distributions in space and time and to compare the detection and p-values across the different models. To test for associations between sample source and cluster inclusion at any point, chi-squared tests were used, or Fisher's exact test, if categories contained five or fewer expected observations.

Results

Descriptive analysis of sequenced isolates

SNP isolates came from 310 premises, 247 with a single isolate and 63 premises with between two and nine isolates. On 39 premises, the additional isolates were deemed to belong to the same case, regardless of whether cases were classified by phage type or SNP address. Eight premises with sequenced isolates were identified as having two or more DT types (Supplementary Table S1), although on seven of these, the isolates of different phage types were allocated to the same SNP-5 group. There were 17 premises with additional isolates that clustered only above the 5-SNP threshold and were thus defined as having multiple SNP cases (Table 1). On only one of

Table 1. Premises where two or more isolates were sequenced and more than one SNP case identified

Host	Phage type	Number of isolates	SNP address ^a	Days between sample collection	SNP clustering threshold
		1	1.2.2.2.1117.1261.2336		
Cattle	104	1	1.2.2.2.1117.1912.2244	806	≤10
		1	1.2.2.2.1117.1912.2307	83	≤5
		1	1.2.2.2.450.1195.2294		
Cattle	104	1	1.2.2.2.450.1195.2294	71	0
		1	1.2.2.2.7.7.29	1,839	≤25
Cattle	104	1	1.2.2.2.450.1195.2286		
		1	1.2.2.2.450.1204.1252	668	≤10
Cattle	104	1	1.2.2.2.571.610.1331		
		1	1.2.2.2.398.424.440	2	≤25
		1	1.2.2.649.1132.1282.1367		
Pig	104	2	1.2.2.649.1132.1282.x	119	≤5
		1	1.2.2.649.1132.1231.1285	0	≤10
Pig	104	1	1.2.2.2.1128.1226.1281		
		1	1.2.2.2.1128.1226.1373	523	≤5
		2	1.2.2.2.1739.2038.x	1,192	≤25
		3	1.2.2.2.1726.2020.x	806	≤25
Pig	104	1	1.2.2.2.1111.1205.1253		
		1	1.2.2.2.5.976.1299	61	≤25
Pig	104	1	1.4.44.84.1121.2019.2445		
		1	1.4.44.84.1121.2032.2464	0	≤10
		2	1.2.2.649.1735.2033.x		
Pig	104	1	1.2.2.932.1736.2035.2485	0	≤50
		1	1.2.2.932.1736.2035.2467	226	≤5
Pig	104	1	1.2.496.928.1729.2049.2495		
		1	1.2.496.928.1729.2023.2453	0	≤10
Horse	104	1	1.2.2.2.1692.1972.2361		
		1	1.2.2.2.1683.1961.2346	405	≤25
		1	1.7.28.55.73.254.2506		
Horse	104	1	1.2.2.2.1111.1205.1394	1	≤250
		1	1.7.28.55.73.254.1396	6 ^b	≤5
Chicken	104	1	1.2.2.676.1186.1300.1375		
		1	1.2.2.2.1204.1324.1421	0	≤50
Chicken	104	1	1.2.2.2.1141.1242.1297		
		1	1.2.2.2.1128.1226.1281	7	≤25
Turkey	U302	1	1.2.2.635.1090.1174.1209		
	104b	1	1.2.2.635.1219.1346.1449	0	≤25
		1	1.2.2.2.77.2034.2466		
Quail	104	1	1.2.2.2.77.81.82	16	≤10
		1	1.2.2.2.77.334.2474	49	≤10
		2	1.2.2.2.77.334.x	231	≤5
Non-livestock premises	104	1	1.2.2.668.1174.1285.1357		
		1	1.2.2.2.5.20.1412	7	≤50

^aMultiple isolates collected on the same day with ≤5 SNP differences have the 0 SNP groups represented as 'x'.^bThe number of days elapsed is between this and the most closely related previous sample, that is the first one from this case. An identical isolate was found on linked premises 329 days later.

these premises were the isolates also differentiated by phage typing. A comparison of genetic distance *versus* time interval between all intra-premises isolate pairs confirmed a tendency for the median time difference to increase with higher thresholds of SNP clustering, up to the 25-SNP level (Supplementary Table S2). However, on six livestock and one non-livestock-keeping premises, isolates that clustered only above the 10-SNP threshold were collected within a week of each other (Table 1). Conversely, although 85% of isolate pairs ($n = 132/156$) falling within the same 5-SNP or 0-SNP cluster were collected less than three months apart, there were seven incidents, where isolates in the same 0- or 5-SNP cluster could be identified more than six months after the original sample was collected, and two incidents, in cattle and sheep, where isolates in the same 5-SNP cluster were identified on the same estate over two years later. Also, in cattle, two isolates in the same 5-SNP cluster were found on two neighbouring farms over three years apart.

Selection of strains to trial detection methods

Figure 1 shows a dendrogram of the first detected isolates from all cases with SNP data, clustered by the seven levels of SNP address. The t5:459 outbreak cluster (SNP-5 group 7) is highlighted and falls within the largest 25-SNP cluster, which comprised 76% ($n = 249$) of incidents. Within this group, there were 74 different 10-SNP clusters and 107 5-SNP clusters. The t5:459 outbreak cluster comprised 72 isolates from 53 SNP cases. The field epidemiological investigations at the time of the outbreak linked the first reported case in humans to a case of salmonellosis in cattle in April 2014. However, this data set shows an isolate from a sheep in Anglesey, collected in March 2012, within the same 5-SNP cluster, suggesting that this strain had been circulating in animals for at least two years.

Using a twelve-month rolling average, four 10-SNP clusters were identified that contained at least five cases and had contributed at least 20% of cases in the previous 12 months (Figure 2). Six 5-SNP clusters, including the 2016 outbreak cluster, also met the above criteria and were all subgroups of the identified 10-SNP clusters. Three were located within the same 10-SNP cluster (SNP-10 group 5), along with twelve other 5-SNP groups each of one or two isolates. SNP-5 group 1,195 contained more than 50% of the cases within its SNP-10 group 450 supergroup, as did SNP-5 group 221 with respect to SNP-10 group 88. In contrast, the t5:459 outbreak group did not cluster with any other incidents within the 10-SNP threshold.

A total of nine SNP groups were selected to use as case definitions for all cluster detection models, six at the 5-SNP level and three at the 10-SNP level. For the SNP-5 group 7 (t5:459 outbreak) strain, models at the 10-SNP level included no additional cases, they were identical to 5-SNP level models and are not shown. Three additional case definitions were used for SaTScan models only, as they either had fewer than five incidents and therefore would not meet the default alarm threshold for Farrington models, or the 5-SNP and 10-SNP groups contained identical incidents. The full list of case definitions is shown in Table 2.

Farrington models

Figure 3 shows the months where the number of cases was in exceedance of the threshold calculated by the Farrington models. Except for SNP-10 group 88 and its subgroup SNP-5 group 221, the other ten models raised at least one alarm during the study period. All alarms raised at the default threshold (five cases in four weeks) were also included at the low threshold of three cases in twelve

weeks, and whilst this threshold decrease resulted in only one or two additional alarms in the phage-type models, there were between 1 and 8 additional alarms (median 3) in the SNP models, excluding those where no alarms were raised.

There was some overlap between alarms raised by the 10-SNP models and those run on their 5-SNP sub-cluster(s), and between the DT104 group and phage types DT104 and U302, but in no case was there complete agreement. SNP-5 group 20 and SNP-5 group 976 raised one and two additional alarms, respectively, that did not appear in their supergroup, SNP-10 group 5. Both SNP-10 group 5 and group 450 raised alarms that did not correspond with any seen in their modelled subgroups. Phage-type models at the subgroup level also raised alarms that were not observed at the supergroup level and *vice versa*.

For the t5:459 outbreak, the earliest alarm was raised in January 2016 by the low threshold SNP model. The default threshold U302 phage-type model raised an alarm two months later, and the default threshold SNP model, DT104 phage type, and the DT104 group models each raised an alarm five months after that. No other concordance between the phage type and SNP group models could be confidently identified. The closest alarms were in November and December 2006 in the DT104 group model, followed by an alarm in January 2007 by the SNP-10 group 5 (low threshold) model. The 23 phage-type incidents in November were predominantly in pigs, cattle, and turkeys and widely dispersed across the whole of Great Britain; likewise, the three incidents in the SNP model were in different species and widely dispersed. A second pair of alerts close in time was raised by the SNP-5 group 976 model in July 2006 followed by one in the U302 phage-type model in September of the same year. However, it seems unlikely that these are related, as no sequenced U302 isolates ($n = 35$) were in the SNP-5 group 976 or its supergroup, SNP-10 group 5.

SaTScan

A total of twelve SNP addresses were used as the case definition in the Bernoulli models, with controls defined as any other SNP addresses. Space-time clusters were detected at both the 5-SNP ($n = 16$) and 10-SNP ($n = 10$) level models. All SaTScan models run on SNP cases identified statistical case clusters earlier than their corresponding Farrington models. In addition, the 5-SNP level addresses within the SNP10 group 450 supergroup, which could not be used in Farrington models, as there were fewer than three isolates, recorded detectable clusters in SaTScan models. When first detected, seven of the clusters identified in the 5-SNP models were deemed statistically significant (p -value < 0.05), whereas none of the SNP-10 clusters were, though one had a p -value of 0.051 at first detection (Table 3 and Supplementary Table S3). Both the minimum number and median number of cases in a cluster at first detection were 2 for the SNP-5 models and 3 for the SNP-10 models, although there was more variation observed between the different SNP sets than there was between SNP-5 and SNP-10 levels belonging to the same set. Space-time clusters tended to remain relatively consistent over subsequent sequential models, with some shifts in radius and location, and the occasional appearance of a clearly distinct secondary cluster (Figure 4).

Models run on the DT104 group data detected 35 distinct clusters, with between two and eight detected at every time interval (Supplementary Tables S4 and S5). The minimum number of cases in a cluster at first detection was 5, and the median was 7. Only one cluster had a p -value of < 0.05 at first detection – although this cluster was, more accurately, a continuation of an earlier cluster

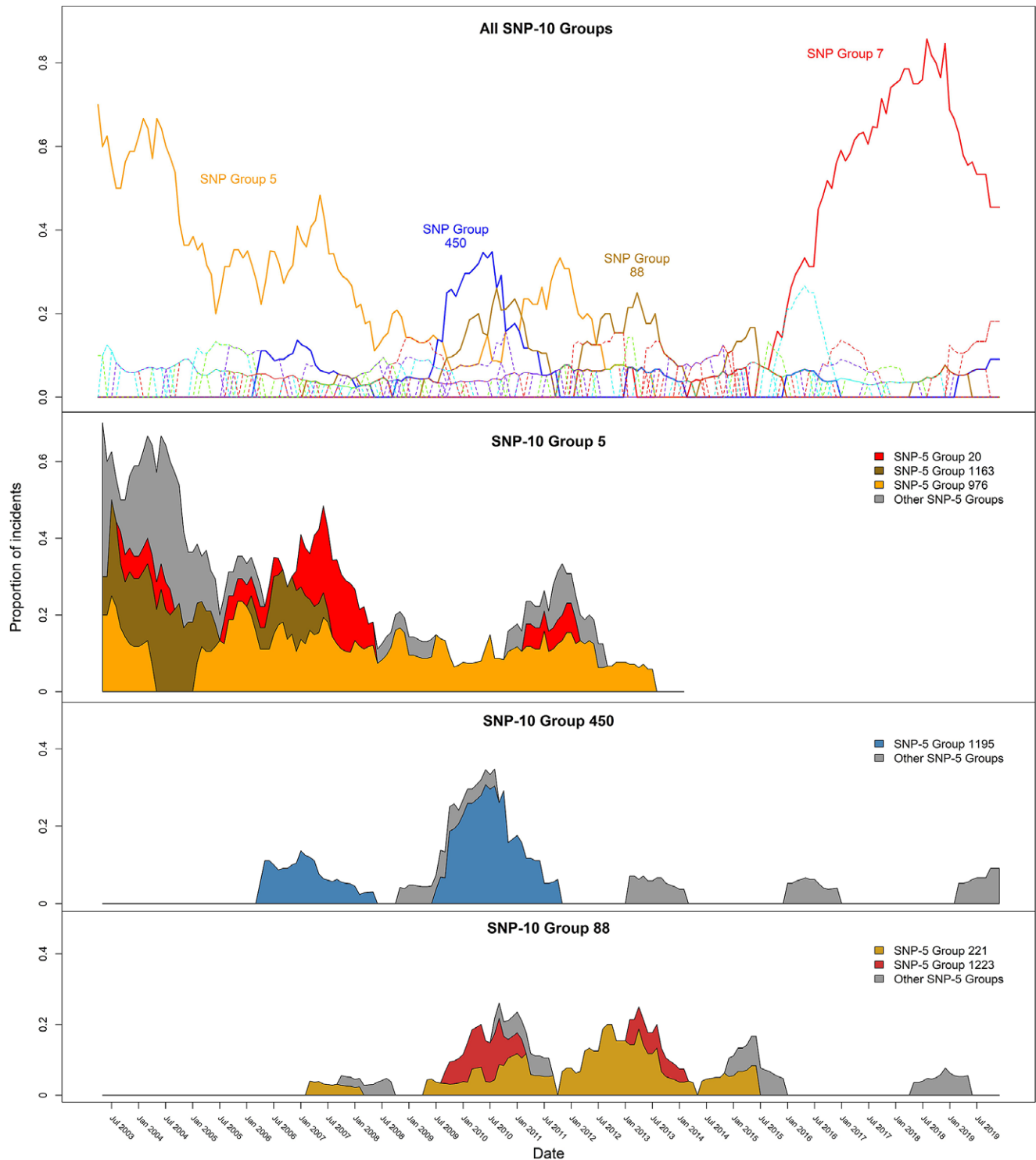


Figure 2. Twelve-month rolling averages of the proportion of cases contributed by different SNP groups, with breakdowns of the 10-SNP clusters contributing at least five incidents overall, and 20% or more of all cases in any 12-month period. All cases within SNP-10 group 7 belonged to the same subgroup (SNP-5 group 7).

that split into two distinct foci. Longitudinal tracking of clusters showed that some remained stable over time, whereas others appeared only transiently, shifted in size and location, merged, or even split and recombined (Figure 5). Six clusters presented a p -value of <0.05 at some point in their lifespan, though this could be up to two years after the first detection.

The t5:459 outbreak was initially detected, after just two cases in Anglesey, in June 2014 by the SNP-5 level model (SNP-5 cluster A). The earliest detection by the DT104 group models was in March 2016, although a transient cluster centred on the Llŷn peninsula was detected in June 2012, which included the earliest known case in a sheep on Anglesey with the outbreak SNP

Table 2. Case definitions for models

Grouping method	Supergroup	Subgroup(s)	Number of cases between 2002 and 2019	Maximum proportion of 12-month rolling average	Farrington models	SaTScan models
Phage type	DT104 group		1,261	0.49	✓	✓
		DT104	832	0.34	✓	
		U302	278	0.2	✓	
SNP	SNP-10 group 7		53	0.86	Identical to subgroup	
		SNP-5 group 7	53	0.86	✓	✓
	SNP-10 group 5		54	0.7	✓	✓
		SNP-5 group 20	10	0.23	✓	✓
		SNP-5 group 976	23	0.25	✓	✓
		SNP-5 group 1,163	8	0.27	✓	✓
	SNP-10 group 450		18	0.35	✓	✓
		SNP-5 group 1,195	13	0.31	✓	✓
	SNP-10 group 88		17	0.26	✓	✓
		SNP-5 group 221	9	0.2	✓	✓
		SNP-5 group 1,223	4	0.13		✓
	SNP-10 group 1,088		3	0.13	Identical to subgroup	
		SNP-5 group 1,170	3	0.13		✓
	SNP-10 group 398		6	0.18	Identical to subgroup	
		SNP-5 group 424	6	0.18		✓

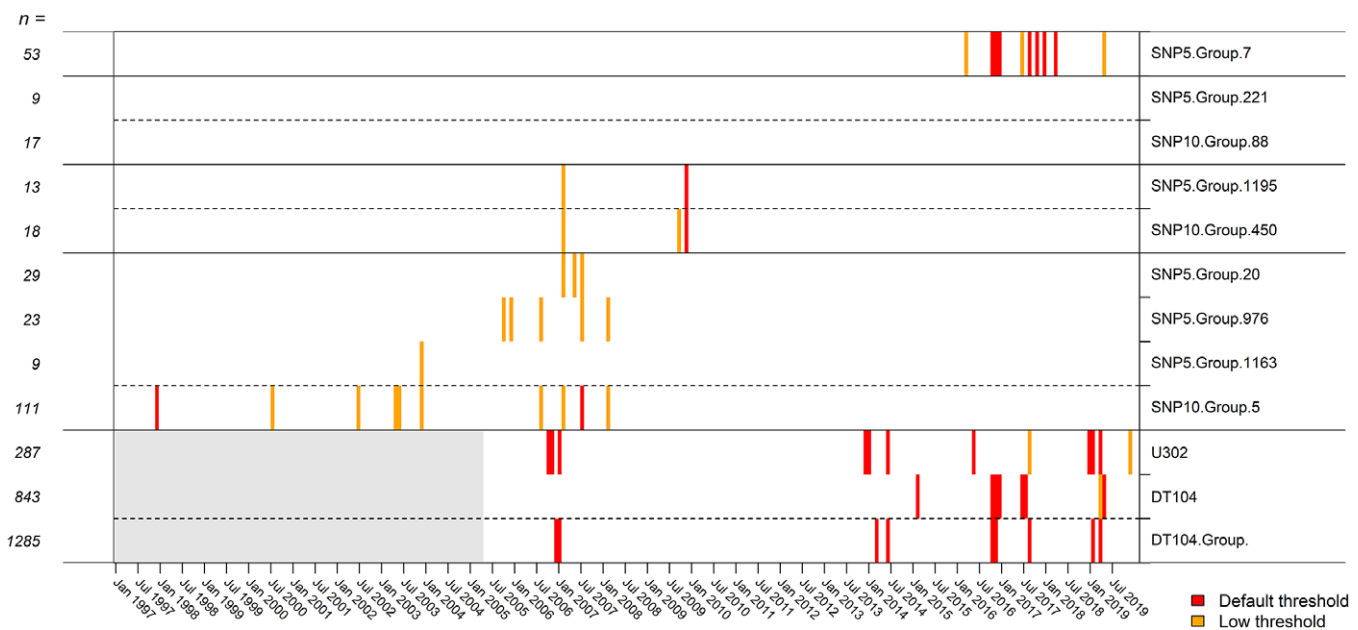


Figure 3. Monthly alarms raised by Farrington models with a five-year run-in period for different case definitions. The default model threshold suppressed alarms if there were fewer than five cases in 4 weeks. The low threshold models suppressed alarms if there were fewer than three cases in 12 weeks. The total number of cases for each model is shown on the left (n). Greyed-out areas show where there are no predictions (run-in period or earlier).

address. None of the other cases in this cluster had any sequence information.

The animal source of the sample appeared to influence its likelihood of being included in a cluster for the SNP-5 level (Fisher's exact test: $p < 0.0005$), the SNP-10 level (Fisher's exact

test: $p < 0.0005$), and the phage type (chi-squared statistic: 38.7, $df = 8$, $p < 0.0001$) models. For all models, cattle and pigs contributed strongly to the test statistic, with cattle incidents being more likely and pig incidents less likely to be included in clusters (Supplementary Table S6). Sheep and turkeys were also more likely

Table 3. SaTScan detection of space–time clusters of cases with sequenced isolates between March 2003 and September 2019

Case definition	Cluster ID	Status at first detection				Relative risk ^a (P-value)	Observations
		Date	Number of cases	Radius (km)	Species (n)		
SNP-5 group 7	A	June 2014	2	13.73	Cattle (1) Sheep (1)	Infinity (0.005)	The cluster persisted until the end of the study period and eventually contained 43 cases. The cluster increased in size and the centre drifted eastwards and then southwards as new cases occurred
	B	Sept 2016	5	19.92	Dog (4) Cat (1)	26.8 (>0.001)	The first five cases occurred within a 6-week period. Two further cases (one in sheep and one in cattle) were included in this cluster one year later, but subsequently reassigned to cluster A
SNP-10 group 5	C	Mar 2003	3	122.01	Cattle (2) Horse (1)	Infinity (0.198)	The predominant cluster starts in the southwest of England and south Wales, then moves to mid-Wales, and eventually into the Welsh borders and west of England. There is broad overlap with clusters identified by the SNP-5 group 976 models, but many other SNP-5 groups are included in this cluster. Cluster D largely corresponds with cluster M (SNP-5 group 1,163) but also incorporates a SNP-5 group 20 case. Clusters E to I only appear transiently. Clusters E and H incorporate a mixture of SNP-5 groups, cluster G is dominated by SNP-5 group 976 with a single SNP-5 group 20 case, and clusters F and I only contain SNP-5 group 976 cases
	D	Sept 2003	3	84.94	Cattle (3)	2.17 (0.944)	
	E	Dec 2006	3	35.79	Cattle (1) Turkey (2)	2.46 (0.998)	
	F	Sep 2007	3	14.20	Cattle (3)	2.58 (0.995)	
	G	June 2010	6	77.81	Cattle (5) Turkey (1)	2.50 (0.989)	
	H	Sept 2011	5	164.08	Cattle (2) Sheep (1) Chicken (1) Turkey (1)	3.09 (0.908)	
SNP-5 group 20	J	Dec 2006	2	183.75	Turkey (2)	20.33 (0.365)	One very large cluster (J) at first detection, subsequently split into two smaller clusters (K and L)
	K	Sept 2008	3	56.74	Sheep (1) Turkey (2)	11.70 (0.518)	
	L	Sept 2008	4	79.76	Pig (1) Turkey (3)	13.14 (0.102)	
SNP-5 group 1,163	M	Sept 2003	2	3.53	Cattle (2)	Infinity (0.103)	A small cluster, persisting for three years, with a maximum radius of 27.4 km, and a shift in location of around 20 km
SNP-5 group 976	N	Mar 2005	2	6.83	Sheep (2)	18.0 (0.072)	Cluster N begins in mid-Wales but rapidly expands when new cases appear in the southwest of England. A cluster persists in Wales until 2012, but periodically splits and recombines as some cases are temporarily reassigned into secondary smaller clusters (O, P, and Q). Cluster R first appears in the southeast of England and moves northwards with sudden changes in size, eventually focusing on a small area in East Anglia
	O	Mar 2008	4	23.13	Cattle (4)	7.13 (0.287)	
	P	Dec 2010	4	36.58	Cattle (4)	4.72 (0.996)	
	Q	Sept 2012	2	31.21	Cattle (2)	9.38 (0.924)	
	R	June 2007	2	162.4	Horse (1) Turkey (1)	2.77 (0.995)	
SNP-10 group 450	S	June 2006	2	66.57	Cattle (1) Cat (1)	29.5 (0.051)	Cluster S corresponds with cluster U, but persists for longer and incorporates two other related SNP-5 groups in cattle in the same area and a case from retail premises in West Yorkshire. Cluster T incorporates a SNP-5 1195 case from Hampshire and two cases with a different SNP-5 group in Essex
	T	Mar 2013	2	80.19	Pig (1) Dog (1)	9.57 (0.885)	
SNP-5 group 1,195	U	June 2006	2	66.57	Cattle (1) Cat (1)	29.5 (0.049)	This cluster is dominated by cases in cattle in the Cheshire and Staffordshire area, with sporadic cases elsewhere in other species. There is a transient split into two clusters in June 2010
SNP-10 group 88	V	Dec 2007	2	185.65	Pig (1) Dog (1)	8.75 (0.189)	This cluster was initially large, but subsequently focused on a more limited area around the

(Continued)

Table 3. (Continued)

Case definition	Cluster ID	Status at first detection			Species (n)	Relative risk ^a (P-value)	Observations
		Date	Number of cases	Radius (km)			
							Devon/Cornwall border. Most cases comprised the SNP-5 groups 221 and 1,223, but three other related SNP-5 groups were detected in cattle in the same area
SNP-5 group 221	W	June 2009	2	6.52	Cattle (1) Pig (1)	Infinity (0.017)	The cluster was confined to west and south Devon, with the majority of cases occurring in cattle
SNP-5 group 1,223	X	Dec 2009	2	2.29	Cattle (2)	Infinity (0.010)	This was a very localised cluster in north Cornwall, with a maximum radius of less than 5 km
SNP-5 group 424	Y	June 2016	2	9.78	Cattle (2)	Infinity (0.018)	This was initially very localised in the Oxfordshire/Warwickshire border, but the cluster rapidly expanded to incorporate three cases in Pembrokeshire
SNP-5 group 1,170	Z	June 2014	2	112.94	Horses (2)	Infinity (0.0047)	This cluster contained only three cases, all in horses, but from Staffordshire, Essex, and North Yorkshire

Note: For each model, controls consisted of all other sequenced isolates.

^aThe risk ratios are often infinite because all cases are included in the circle.

to be included in SaTScan clusters across model types, whilst chickens, equids, and other avian species were less likely to be included. Domestic carnivores did not show an association either way, although they were included in 20% of the SNP clusters at the cluster's first appearance, despite forming only 6% of the sequenced samples.

Discussion

In this study, we have applied, to *S. Typhimurium* isolates, a form of strain differentiation that is relatively new to the field of veterinary surveillance. Early detection of serovars, phage types, and AMR patterns in veterinary isolates has been routinely carried out for salmonella in Great Britain as part of passive surveillance activities, but the transition towards molecular methods for differentiating salmonellae in the laboratory now demands an adaptation of epidemiological methods to incorporate the increased typing resolution that WGS brings. The SNP address has proved its ready accessibility to existing surveillance methods in the human sphere [19, 21]; this study provides an initial investigation of its applicability to the veterinary field using historical data sets.

Two methods of outbreak detection have been explored here, and comparisons have been drawn between using single-linkage SNP thresholds at the 5- and 10-SNP level and the traditional phage-type method to define cases. Although we used only one verified outbreak to explore the sensitivity and timeliness of the methods prospectively, the exploration of further SNP addresses has provided insights into how the methods are likely to behave going forward, and potential problems and pitfalls that may be encountered.

The Farrington models provided a robust, easy-to-implement method of detecting a higher-than-expected number of cases and, for the SNP addresses, detected an exceedance after a relatively small number of cases. The weakness of exceedance methods is when serovars or phage types are very common, as the models lack

the sensitivity to detect small increases in case numbers, but WGS may provide a way of narrowing the resolution of the data and thus making it easier to pick up a rapid increase in the detection of a particular strain or group of strains. The level of resolution was easy to adjust using different levels of the SNP address, although it was not clear whether using either the 5- or 10-SNP level gave a particular advantage. As single-linkage clustering is impacted by population coverage, additional sampling could lead to the subsequent merging of SNP-5 groups, and therefore, surveillance at the SNP-10 level may be easier to implement on a purely practical level. The alarms raised by the different SNP addresses were temporally limited, compared with the phage-type models that raised alarms across the whole time period analysed. Alarms raised by the groups with the largest number of incidents (SNP-10 group 5 and sub-groups and SNP-5 group 7) appeared approximately correlated with temporal clusters of the phage-type alarms, although the missing data from the early years in the phage-type models precluded any formal analysis of this. The SNP models with fewer incidents (SNP-10 groups 88 and 450 and corresponding sub-groups) did not demonstrate corresponding alarms in the phage-type models.

With respect to the t5:459 outbreak, the SNP address did not raise alerts any earlier than the phage-type case definition. Both models first identified an exceedance in August 2016 using the default threshold, although at the lower threshold of three cases in 12 weeks, the SNP address detected the first exceedance in January. However, for routine use, lowering the threshold would need to be carefully trialled, as the decreased specificity may result in unnecessary investigations. Lowering the specificity of the model by adjusting the default exceedance threshold identified more clusters, and was necessary in this data set, due to the relatively low proportion of sequenced isolates compared with overall salmonella submissions. Where total case numbers were low, as with many of the SNP group models, the time interval between cases was generally too large for the models to raise any alarms using the default threshold. The

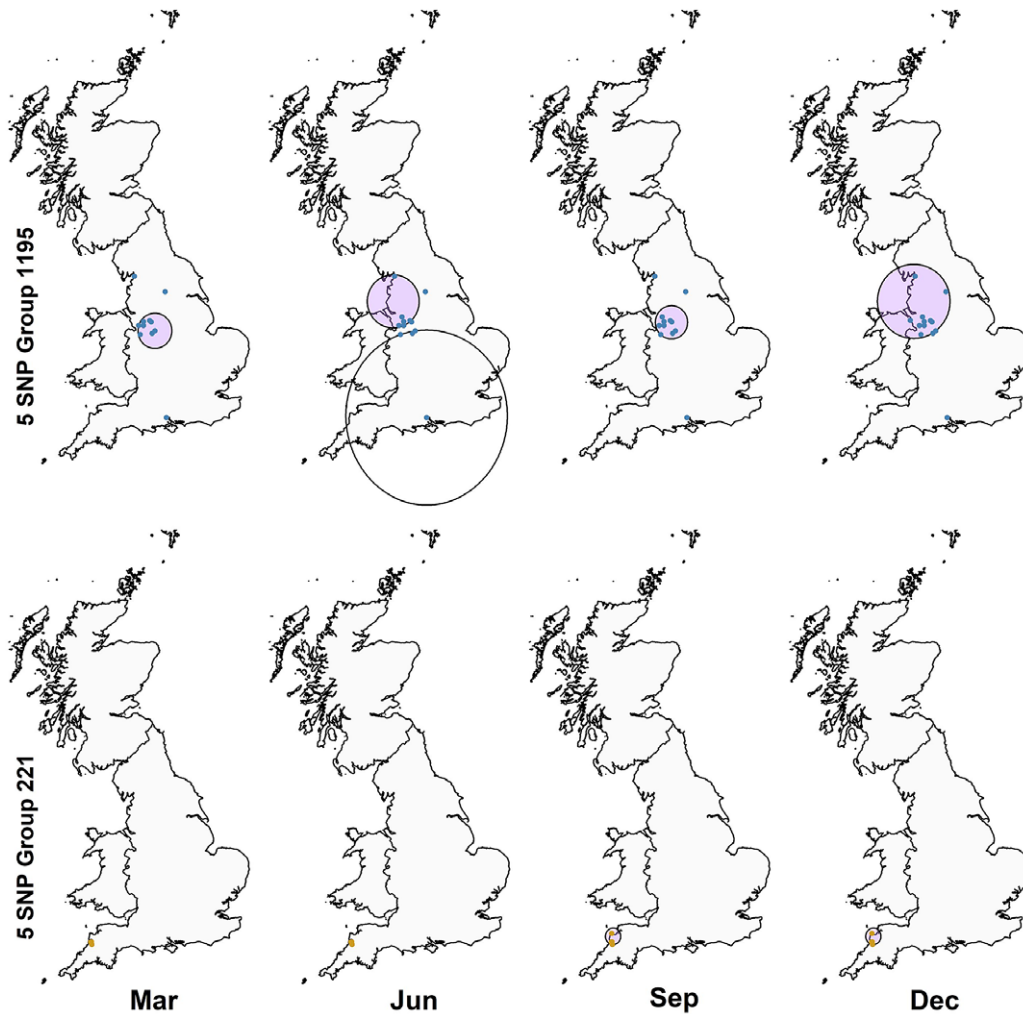


Figure 4. Examples of SaTScan space–time clusters detected for different SNP-5 groups for four quarters of 2010. Cases are shown as dots, and circles show the model clusters. The colour intensity of the circles is related to the *P*-value returned by the model, with darker circles having lower *P*-values.

effect of changing the threshold was much less in the phage-type models as the number of phage-type cases was orders of magnitude greater than the number of SNP cases. As these models do appear to be relatively sensitive to the number of cases, their future potential for routine surveillance is likely to depend on the number of sequenced isolates available. A further possibility that could be explored using sequence typing would be cgMLST, which may give sufficient numbers of related cases, but offer a better resolution than phage typing.

The space–time models created using SaTScan appeared to give a distinct advantage over the Farrington exceedance models for more timely detection of clusters, in all cases detecting clusters earlier than their Farrington counterparts, often after only two or three cases. Interestingly, this was not only true for cases that were very localised, such as the t5:459 (SNP-5 group 7) and SNP-5 group 1,223, but also when the cases were more spatially distant, such as SNP-5 group 1,170, where the first three cases occurred in East Yorkshire, Staffordshire, and Essex. For the t5:459 outbreak strain, the SNP address models first detected a space–time cluster in June 2014, whereas the phage-type models' first detection of a space–time cluster in the same location did not occur until March 2016 (with the first *p*-value <0.05 not occurring until December 2016). The SNP address models were also considerably easier to interpret,

with only one to three clusters appearing at any time point. Mapping the space–time clusters at sequential timepoints clearly indicated which were persistent and which were new. Identification of persistent space–time clusters, especially those spreading, either locally or into new areas, may be the flag that would trigger an investigation, as an indicator that a strain is successful, in the evolutionary sense, in the veterinary sector and becoming prevalent enough to pose risks to human health. The phage-type models, on the other hand, were much more laborious to interpret, with many more transient clusters, cluster splitting, recombining, or suddenly shifting in size and location, making tracking of clusters over time a labour-intensive process. There is a potential risk that, as more isolates are routinely sequenced and the data sets get larger, this could also apply to SNP address models. However, from the data for SNP-10 group 5, which was already in existence at the start of the study period and had a more widespread distribution of incidents around the country, this did not seem to be the case. A small number of models were also trialled using SNP address as cases and any other sequenced or non-sequenced incidents as controls, to see whether the number of controls in the background population affected the detection of the SNP clusters (data not shown). Neither the identification of the clusters nor their relative risks or *p*-values appeared dramatically altered, which suggests that the method

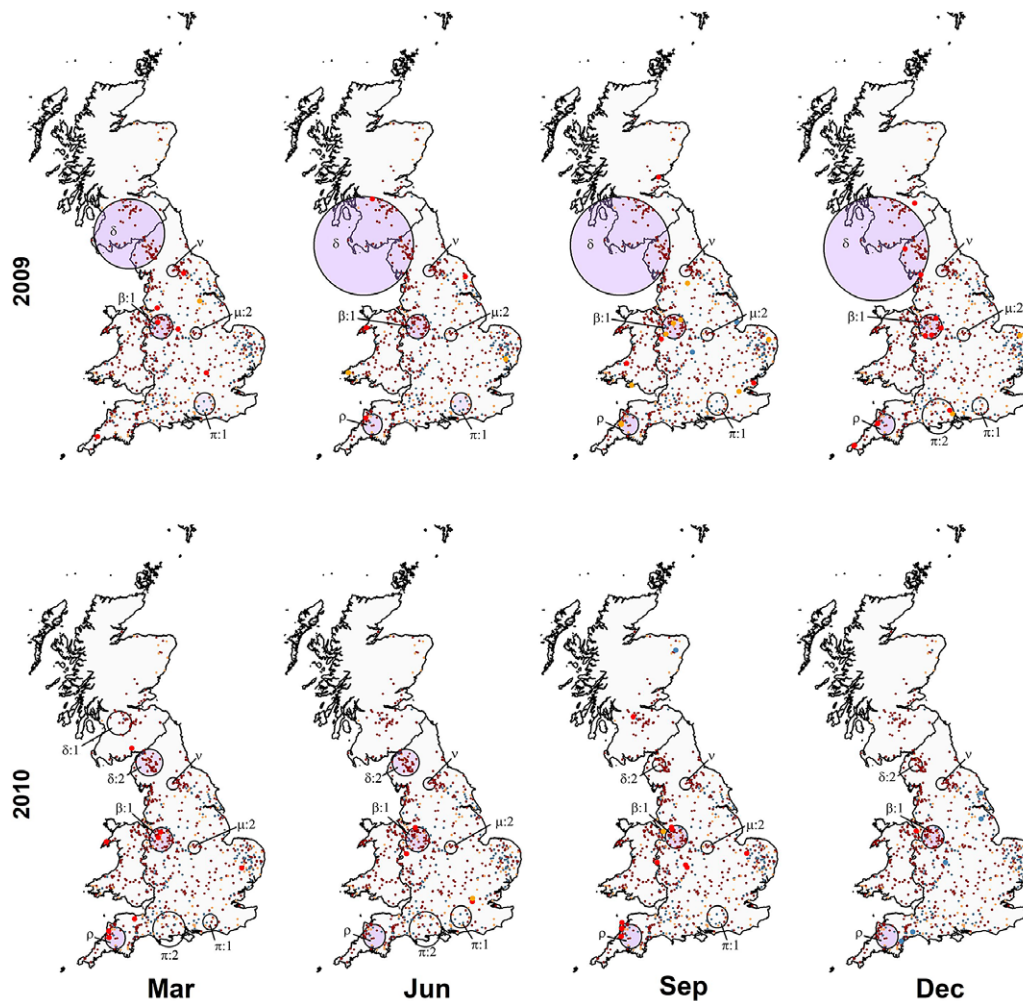


Figure 5. SaTScan clusters detected for DT104 group models for 2009 and 2010, with an example of a splitting cluster (δ becomes $\delta:1$ and $\delta:2$). The $\delta:1$ cluster appears only transiently in the first quarter of 2010.

could work equally well with larger data sets of sequenced isolates going forward. Furthermore, the data here indicate that the SNP clusters were time-limited, in a way that phage-type data were not, with strains arising and being superseded, making it easier to distinguish between new and existing clusters. Some of the explanations for the sudden shifts in the phage-type clusters may be due to them incorporating more than one strain of unrelated salmonella within the DT104 phage type that overlapped in time and space.

Although the SNP address appears to perform satisfactorily in both methods for detecting an exceedance of cases retrospectively, a fundamental issue for both methods is how to decide which SNP addresses to run any models on. In a routine surveillance context, the 'outbreak' strain in the currently used EDS model would have been unlikely to have been run until there were at least five sequenced isolates and thus an exceedance may not have been detected as early as suggested by the models shown here. Deciding what outcome to use for the EDS is fundamental and identifying new strains for the EDS to look for requires much time and effort. Using the 12-month rolling average to identify SNP groups with a large or rapidly increasing share of the caseload was a straightforward way to identify SNP addresses to investigate here and would be amenable to automation. One of the main drawbacks of SaTScan was the requirement for manual upload of data, which would have

made it unfeasible to run many models for different SNP addresses in SaTScan on a regular basis. A potential solution tried was the multinomial SaTScan model, which avoids the issue of having *a priori* knowledge of which strain to look for. However, this performed poorly in early trials on our data set, which was thought to be due to the large number of categories. Recently developed packages allowing SaTScan to interface with R software may make routine use of this programme more feasible, and from the fact that almost all the SNP clusters tended to show spatial, as well as temporal clustering, it seems likely that incorporating a spatial element into routine detection models would give an advantage over relying on temporal clustering alone. This is perhaps more important in veterinary surveillance, due to the expected substantially lower coverage of animal diagnostic samples tested compared with human samples.

The 2016 DT104 outbreak, on which this work was based, defined livestock incidents as cases if their SNP address was within the same single-linkage cluster at the 5-SNP level as the human outbreak strain. In human cases, the 5-SNP threshold is used for outbreak investigations due to the high likelihood that cases relate to a common source, whilst analysis at the 10-SNP threshold is undertaken to uncover deeper epidemiological links [21]. The same SNP address levels appeared to work well here, and the fact that

>75% of isolates fell within the same cluster at the next highest level of SNP address makes it seem probable that analysing at any higher level may lose the advantage of the greater resolution provided by the SNP address over the phage type. This was born out by some of our earlier trials with detection models run on different levels of the SNP address (not shown here). However, previous studies have found that whilst the number of SNP differences between isolates within outbreaks is usually small (2–12 SNP differences) larger differences of up to 249 SNPs may exist and may be dependent on the serovar [5]. The SNP differences between strains may also depend on the SNP-based sub-typing workflows used [32]. Batch effects, that is technical sources of variation in subsets of sequencing data arising from DNA extraction, technician skills, library preparation, sequencing lane differences affecting coverage, or bioinformatics tools used for trimming and assemblage, are a potential issue even where all work is carried out within the same laboratory. These have been noted as a particular source of bias in multi-year projects [33] where samples are added incrementally and allocation to different libraries and lanes cannot be fully randomised. Whilst this work used three different sets of sequencing data generated at APHA, and there was relatively little temporal overlap between the samples selected for inclusion, we did note that there were eleven 5-SNP clusters that contained samples sequenced under different projects. Although we cannot rule out false SNP differences in this data set, it appears that the SNP address may be relatively robust to these types of information bias, even at the 5-SNP threshold. However, we would stress the need for thresholds at which to include or exclude cases to be considered on an outbreak-by-outbreak basis, especially if multiple laboratories with different pipelines begin to contribute to sample testing. Where the single-linkage threshold chosen is too discriminatory, cases may be wrongly excluded; however, it is too inclusive and much time is spent investigating cases that are not related.

Thus, for the definition of a strain using SNP address to be most useful, studies looking at background strain diversity are required, both between and within farms, in order to inform how many samples need to be collected and sequenced to identify outbreak incidents. If there is high diversity within a phage type, many more samples might be needed per farm than if using serotype or phage type as the classifier. The temporal aspect also needs to be considered, as a rapidly changing organism may be misclassified by the single-linkage clustering method as a new strain, especially if the coverage of samples being sequenced is low, as intervening isolates are more likely to be missed. This could potentially be countered by progressively relaxing the cut-off chosen for classifying cases as part of the same outbreak as time intervals between samples increase. However, we also demonstrated isolates within the same SNP-5 cluster that were collected over two years apart. Leekitcharoenphon et al. [5] were unable to find an association between time of isolation and the number of SNP differences and suggested the existence of groups of isolates that comprise single clonal haplotypes with virtually no genetic change over time. Knowing more about changes in strains over time will be particularly important, as it is apparent that clusters and indeed outbreaks can span many years.

Animal host species would also appear to be an important consideration for the application of the SNP address to early detection. It was unsurprising that cattle and sheep had an above-average chance of appearing in SaTScan SNP clusters for the sequenced data, given that this data set was neither random nor representative, but augmented by the outbreak investigation. It was more surprising that cattle and sheep were considerably more likely to be included in phage-type clusters, along with turkeys, whilst

pigs, chickens, horses, and other avian species were much less likely to be included. The reasons for this could be multi-factorial, including things such as uneven host population densities; different movement and mixing patterns in different livestock species; national salmonella control plans applied to poultry, thus lowering *S. Typhimurium* risk in these, but not other species; or bacterial factors, such as the predominance of monophasic *S. Typhimurium* ST34 in pigs, and more competition or cross-protection from this or other serovars. The effect of the host species on the bacterial diversity will thus also need to be determined to develop effective ways to use the SNP address.

The use of the SNP address in the case definition will also depend on the strain in question, as a new or rare strain would be valuable in defining cases, a more common strain that is present over much of the population would be less so. The changes over time could be most clearly seen for the SNP clusters that arose during the study period, whilst it was more difficult to interpret the clusters for SNP-10 group 5 and subgroups, which were already the dominant strains in 2003, and more widely dispersed over the country. This same caveat can be applied to serotype or phage-type classification, as was evident from our models using the phage type as the case definition. Prospectively, SNP clusters detected from passive surveillance data will need to be verified using epidemiological information and proven transmission links. Using surveillance data at UKHSA for the seven most common salmonella serovars (Enteritidis, Typhimurium, Typhi, Paratyphi A, Java, Agona, and Newport), Waldram et al. [21] only found a significant epidemiological link for 17 of 32 clusters of isolates. However, Ågren et al. [34] have shown that cattle herds with known epidemiological contacts generally showed smaller SNP differences between *S. Dublin* isolates than where no known links were found.

In conclusion, this work demonstrated that SNP addresses could perform well for detecting outbreaks in a timely manner, although a large number of different strains will pose challenges and will only be feasible if systems can be automated. As the SNP address becomes integrated into routine typing methods for all salmonella submitted to APHA laboratories, larger data sets are likely to soon be available to start answering some of the questions about background diversity, although more targeted and systematic sampling will be needed to answer some of the other questions posed about the influence of time, space, and species on cluster detection. It is also worth noting that there is an urgent need for international consensus about how WGS data are used for typing. Currently, an SNP address is run per institution using the same reference strain, but each institutional database generates different values for SNP addresses dependent on their unique content, making data comparisons between institutions more complex. A shared database holding SNP information for sequenced human, animal, food, and environmental isolates would enhance joint outbreak investigations by shortening the time it currently takes to share the data. Collaboration between laboratories should be a future priority, to agree on a method that can be used on a large scale, and is internationally recognised in the same way as sero- and phage typing, as WGS becomes a more routine part of surveillance and epidemiological investigation of livestock disease outbreaks.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/S0950268823001723>.

Data availability statement. The authors have provided anonymised data within the article and in the supplementary materials. Some of the isolates were

published in two previous papers, as detailed in the manuscript. However, due to the sensitive nature of the outbreak, the authors are currently unable to publish the full list of isolates. Data can be provided on request to APHA.

Acknowledgements. The authors thank Margarida Carvalho Abecasis for preliminary scoping work on the project and Lesley Larkin for her insightful comments on the manuscript.

Author contribution. Conceptualization: R.P.S.; Funding acquisition: R.P.S., L.P.; Methodology: R.P.S., J.L., L.S.; Supervision: R.P.S., J.L., L.P.; Writing – review & editing: R.P.S., J.L., L.P., Y.T., L.S., J.M.B.; Data curation: L.P., Y.T.; Writing – original draft: L.P., L.S., J.M.B.; Formal analysis: L.S., J.M.B.

Financial support. Funding was received from the European Union’s Horizon 2020 research and innovation programme, as part of the COMPARE project, under grant agreement no. 643676. J.B. is supported by Research England’s ‘Expanding Excellence in England (E3) Fund’. The generation of WGS data for all isolates in this study was supported by Defra RDOZO347, with APHA SE0355 and CR2000F R&D funding to LP.

Competing interest. The authors declare none.

References

- [1] **European Food Safety Authority, European Centre for Disease Prevention and Control** (2021). The European Union one health 2019 Zoonoses report. *EFSA Journal* **19**, e06406. <https://doi.org/10.2903/j.efsa.2021.6406>
- [2] **Kingsley RA and Bauml AJ** (2000). Host adaptation and the emergence of infectious disease: the Salmonella paradigm. *Molecular Microbiology* **36**, 1006–1014.
- [3] **Rabsch W, et al.** (2002). Salmonella enterica serotype Typhimurium and its host-adapted variants. *Infection and Immunity* **70**, 2249–2255.
- [4] **Bawn M, et al.** (2020). Evolution of Salmonella enterica serotype Typhimurium driven by anthropogenic selection and niche adaptation. *PLoS Genetics* **16**, e1008850.
- [5] **Leekitcharoenphon P, et al.** (2016). Global genomic epidemiology of Salmonella enterica Serovar Typhimurium DT104. *Applied and Environmental Microbiology* **82**, 2516–2526.
- [6] **Helms M, et al.** (2005) International Salmonella Typhimurium DT104 infections, 1992–2001. *Emerging Infectious Diseases* **11**, 859–867.
- [7] **Hauser E, et al.** (2010). Pork contaminated with *Salmonella enterica* Serovar 4,(5),12:I:–, an emerging health risk for humans. *Applied and Environmental Microbiology* **76**, 4601–4610.
- [8] **Branchu P, et al.** (2019). SGI-4 in monophasic Salmonella Typhimurium ST34 is a novel ICE that enhances resistance to copper. *Frontiers in Microbiology* **10**, 1118.
- [9] **Petrovska L, et al.** (2016). Microevolution of monophasic *Salmonella* Typhimurium during epidemic, United Kingdom, 2005–2010. *Emerging Infectious Diseases* **22**, 617–624.
- [10] **APHA** (2017). Disease surveillance in England and Wales, December 2016. *Veterinary Record* **180**, 39–42.
- [11] **Allard MW, et al.** (2012) High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach. *BMC Genomics* **13**, 32.
- [12] **McDonnell J, et al.** (2013) Retrospective analysis of whole genome sequencing compared to prospective typing data in further informing the epidemiological investigation of an outbreak of *Shigella sonnei* in the UK. *Epidemiology and Infection* **141**, 2568–2575.
- [13] **Dallman TJ, et al.** (2015). Whole-genome sequencing for National Surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clinical Infectious Diseases* **61**, 305–312.
- [14] **Dallman T, et al.** (2016). Phylogenetic structure of European Salmonella Enteritidis outbreak correlates with national and international egg distribution network. *Microbial Genomics* **2**, e000070. <https://doi.org/10.1099/mgen.0.000070>
- [15] **Jenkins C, Dallman TJ and Grant KA** (2019). Impact of whole genome sequencing on the investigation of food-borne outbreaks of Shiga toxin-producing *Escherichia coli* serogroup O157:H7, England, 2013 to 2017. *Eurosurveillance* **24**, 1800346. <https://doi.org/10.2807/1560-7917.ES.2019.24.4.1800346>
- [16] **Mughini-Gras L, et al.** (2019). Critical orientation in the jungle of currently available methods and types of data for source attribution of foodborne diseases. *Frontiers in Microbiology* **10**, 2578.
- [17] **Arnold M, et al.** (2021) Bayesian source attribution of Salmonella Typhimurium isolates from human patients and farm animals in England and Wales. *Frontiers in Microbiology* **12**, 579888.
- [18] **Ashton PM, et al.** (2016). Identification of *Salmonella* for public health surveillance using whole genome sequencing. *PeerJ* **4**, e1752.
- [19] **Chattaway MA, et al.** (2019). The transformation of reference microbiology methods and surveillance for Salmonella with the use of whole genome sequencing in England and Wales. *Frontiers in Public Health* **7**, 317.
- [20] **Dallman T, et al.** (2018). SnapperDB: A database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* **34**, 3028–3029.
- [21] **Waldram A, et al.** (2018). Epidemiological analysis of Salmonella clusters identified by whole genome sequencing, England and Wales 2014. *Food Microbiology* **71**, 39–45.
- [22] **Dallman T** (2018). *SnapperDB*. UKHSA. Available at <https://github.com/ukhsa-collaboration/snapperdb> (accessed 6 November 2023).
- [23] **Mellor KC, et al.** (2022). Contrasting long-term dynamics of antimicrobial resistance and virulence plasmids in Salmonella Typhimurium from animals. *Microbial Genomics* **8**, 000826. <https://doi.org/10.1099/mgen.0.000826>
- [24] **Science Search database.** Available at <https://sciencesearch.defra.gov.uk> (accessed 12 September 2023).
- [25] **APHA** (2023). *NextflowSerotypingPipeline*. APHA-BAC. Available at <https://github.com/APHA-BAC/NextflowSerotypingPipeline> (accessed 6 November 2023).
- [26] **APHA** (2022). *Salmonella in animals and feed in Great Britain 2021*. (ISBN 1 8995 1534 8).
- [27] **Farrington CP, et al.** (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **159**, 547.
- [28] **Kosmider R, et al.** (2006). A statistical system for detecting *Salmonella* outbreaks in British livestock. *Epidemiology and Infection* **134**, 952–960.
- [29] **Salmon M, Schumacher D and Höhle M** (2016). Monitoring count time series in R: aberration detection in public health surveillance. *Journal of Statistical Software* **70**, 1–35. <https://doi.org/10.18637/jss.v070.i10>
- [30] **Kulldorff M** (2021). *Information Management Services, Inc. SaTScan: Software for the spatial and space-time scan statistics*. Available at <https://www.satscan.org/> (accessed 6 November 2023).
- [31] **Kulldorff M, et al.** (2005). A space–time permutation scan statistic for disease outbreak detection. *PLoS Medicine* **2**, e59.
- [32] **Saltykova A, et al.** (2018). Comparison of SNP-based subtyping workflows for bacterial isolates using WGS data, applied to Salmonella enterica serotype Typhimurium and serotype 1,4,(5),12:I:–. *PLoS One* **13**, e0192504.
- [33] **Leigh DM, et al.** (2018). Batch effects in a multiyear sequencing study: false biological trends due to changes in read lengths. *Molecular Ecology Resources* **18**, 778–788.
- [34] **Ågren ECC, et al.** (2016). Comparison of whole genome sequencing typing results and epidemiological contact information from outbreaks of *Salmonella* Dublin in Swedish cattle herds. *Infection Ecology & Epidemiology* **6**, 31782.