scientific reports

OPEN

The integrated genomic surveillance system of Andalusia (SIEGA) provides a One Health regional resource connected with the clinic

Carlos S. Casimiro-Soriguer[1,2], Javier Pérez-Florido[1,2], Enrique A. Robles[1], María Lara[1], Andrea Aguado[1], Manuel A. Rodríguez Iglesias[3], José A. Lepe[2,4,5], Federico García[5,6,7], Mónica Pérez-Alegre[8], Eloísa Andújar[8], Victoria E. Jiménez[8], Lola P. Camino[8], Nicola Loruso[9], Ulises Ameyugo[9], Isabel María Vazquez[9], Carlota M. Lozano[9], J. Alberto Chaves[9] & Joaquin Dopazo[1,2✉]

The One Health approach, recognizing the interconnectedness of human, animal, and environmental health, has gained significance amid emerging zoonotic diseases and antibiotic resistance concerns. This paper aims to demonstrate the utility of a collaborative tool, the SIEGA, for monitoring infectious diseases across domains, fostering a comprehensive understanding of disease dynamics and risk factors, highlighting the pivotal role of One Health surveillance systems. Raw whole-genome sequencing is processed through different species-specific open software that additionally reports the presence of genes associated to anti-microbial resistances and virulence. The SIEGA application is a Laboratory Information Management System, that allows customizing reports, detect transmission chains, and promptly alert on alarming genetic similarities. The SIEGA initiative has successfully accumulated a comprehensive collection of more than 1900 bacterial genomes, including *Salmonella enterica*, *Listeria monocytogenes*, *Campylobacter jejuni*, *Escherichia coli*, *Yersinia enterocolitica* and *Legionella pneumophila*, showcasing its potential in monitoring pathogen transmission, resistance patterns, and virulence factors. SIEGA enables customizable reports and prompt detection of transmission chains, highlighting its contribution to enhancing vigilance and response capabilities. Here we show the potential of genomics in One Health surveillance when supported by an appropriate bioinformatic tool. By facilitating precise disease control strategies and antimicrobial resistance management, SIEGA enhances global health security and reduces the burden of infectious diseases. The integration of health data from humans, animals, and the environment, coupled with advanced genomics, underscores the importance of a holistic One Health approach in mitigating health threats.

Keywords One health, Surveillance, Whole genome sequencing, Resistances, AMR, Epidemiology

[1]Andalusian Platform for Computational Medicine, Andalusian Public Foundation Progress and Health-FPS, Seville, Spain. [2]Institute of Biomedicine of Seville, IBiS, University Hospital Virgen del Rocío/CSIC/University of Seville, 41013 Seville, Spain. [3]Servicio de Microbiología. Hospital Universitario Puerta del Mar, 11009 Cádiz, Spain. [4]Servicio de Microbiología, Unidad Clínica Enfermedades Infecciosas, Microbiología y Medicina Preventiva, Hospital Universitario Virgen del Rocío, 41013 Sevilla, Spain. [5]Centro de Investigación Biomédica en Red en Enfermedades Infecciosas (CIBERINFEC), ISCIII, Madrid, Spain. [6]Servicio de Microbiología. Hospital Universitario San Cecilio, 18016 Granada, Spain. [7]Instituto de Investigación Biosanitaria, Ibs.GRANADA, 18012 Granada, Spain. [8]Genomic Unit, Andalusian Molecular Biology and Regenerative Medicine Center (CABIMER), CSIC University of Seville University Pablo de Olavide, Seville, Spain. [9]Dirección General de Salud Pública y Ordenación Farmacéutica, Consejería de Salud y Consumo- Junta de Andalucía, Seville, Spain. ✉email: joaquin.dopazo@ juntadeandalucia.es

Abbreviations

AMR	Anti-microbial resistance
CC	Clonal complexes
cgMLST	Core genome multi-locus sequence typing
CSV	Comma-separated values
ECDC	European Centre for Disease Prevention and Control
EFSA	European Food Safety Authority
LIMS	Laboratory Information Management System
MLST	Multi-locus sequence-based typing
RASFF	Rapid alert system for food and feed
SIEGA	Sistema Integrado de Epidemiologia Genomica de Andalucia
SNP	Single nucleotide polymorphism
ST	Sequence types
SVEA	Sistema de Vigilancia Epidemiológica de Andalucía
WGS	Whole genome sequencing

The One Health approach emphasizes the interconnectedness of human, animal, and environmental health, and has gained relevance in recent years due to the emergence of zoonotic diseases and increasing antibiotic resistance[1]. The importance of One Health in epidemiological surveillance lies in its capacity to monitor and control infectious diseases across different domains, enabling a comprehensive understanding of disease dynamics and risk factors[2]. By integrating human, animal, and environmental health data, One Health surveillance systems facilitate early detection and response to health threats, thereby enhancing global health security and reducing the burden of disease[3]. Furthermore, this integrated approach fosters multidisciplinary collaborations among stakeholders to develop and implement coordinated strategies aimed at disease prevention and control[4]. Recently, the conventional process of serotyping through serology has been undergoing a gradual transformation, with molecular typing methods such as multi-locus sequence-based typing (MLST)[5] increasingly complementing or replacing traditional methods. However, these techniques do not possess the necessary discriminatory power to differentiate between closely related strains[6], which limits its application in many epidemiological studies. In recent years, High-throughput sequencing (HTS) technologies have revolutionized the field, enabling rapid and cost-effective analyses of complete genomes[7]. Whole-genome sequences (WGS) offer an unparalleled level of discrimination among genetically related isolates, allowing for the exploration of compelling questions such as accurate phylogenetic and phylogenomic analyses[8], as well as the examination of serotype- or subtype-determining genes[9]. Thus, the use of sequence-derived typing is gaining acceptance and is being employed for source attribution and epidemiological surveillance[10]. Accordingly, different analytical strategies have been developed[11–13], and numerous studies have demonstrated the potential of WGS in epidemiological investigations[14–18]. This advancement in genomic sequencing is revolutionizing the way in which the epidemiology of various pathogens is approached and assessed and, actually, the European Centre for Disease Prevention and Control (ECDC)[19] and the World Health Organization[20] have recommended the use of WGS as the gold standard methodology for surveillance of bacterial pathogens. One interesting aspect of WGS is its ability to provide not only precise typing but also the opportunity to conduct additional analyses beyond routine surveillance, such as assessing the existence of genetic factors related to antimicrobial resistance and virulence, which holds great significance in the context of the One Health approach[21]. This approach is rapidly emerging as the primary framework for monitoring and managing antimicrobial resistance, establishing a compelling connection between genomics and comprehensive control strategies.

Andalusia, with a population of 8.5 million inhabitants, is the third largest region in Europe and has the size of a medium-sized European country, like Switzerland or Austria.

On May 15, 2020, the Andalusian Local Ministry of Health entrusted the Progress and Health Foundation to initiate a program aimed at monitoring pathogens through whole genome sequencing and bioinformatic analysis of a specific set of bacteria with significant implications in public health[22]. This marked the inception of the SIEGA initiative.

Subsequently, the Regional Ministry of Health and Consumer Affairs designed a genomic sequencing circuit aimed at facilitating this integration, which was included in Instruction 130/2019[23] on the treatment and sequencing of biological agents isolates in Andalusia with a focus on health protection, inspired in The Transformation of Reference Microbiology Methods and Surveillance for Salmonella With the Use of Whole Genome Sequencing in England and Wales[24]. Figure 1 sketches the general operating layout of the circuit.

Any laboratory that isolates any strain of the pathogens of interest can send it, under the conditions set out in that instruction and accompanied by basic metadata, to the Public Health Laboratory of Málaga, under the Regional Ministry of Health and Consumer Affairs, where DNA extraction is carried out. Once the DNA is extracted, it is preserved in freezing conditions and sent to the sequencing laboratory at the Andalusian Molecular Biology and Regenerative Medicine Center (CABIMER).

Additionally, other collaborating centers have been incorporated, which have either provided DNA extracts (University Hospital Puerta del Mar in Cádiz) or directly provided sequences (University Hospital Virgen del Rocío or University Hospital San Cecilio).

Non-clinical origin samples primarily stem from official veterinary control activities in primary production, such as the implementation of the Annual National Program for the Control of Certain Serotypes of Salmonella in meat chickens of the species *Gallus gallus*. They also originate from Official Control Services in stages subsequent to primary production, which are included in the National Official Control Plan for the Food Chain. This was agreed upon in coordination meetings between the competent authorities. Additionally, other samples
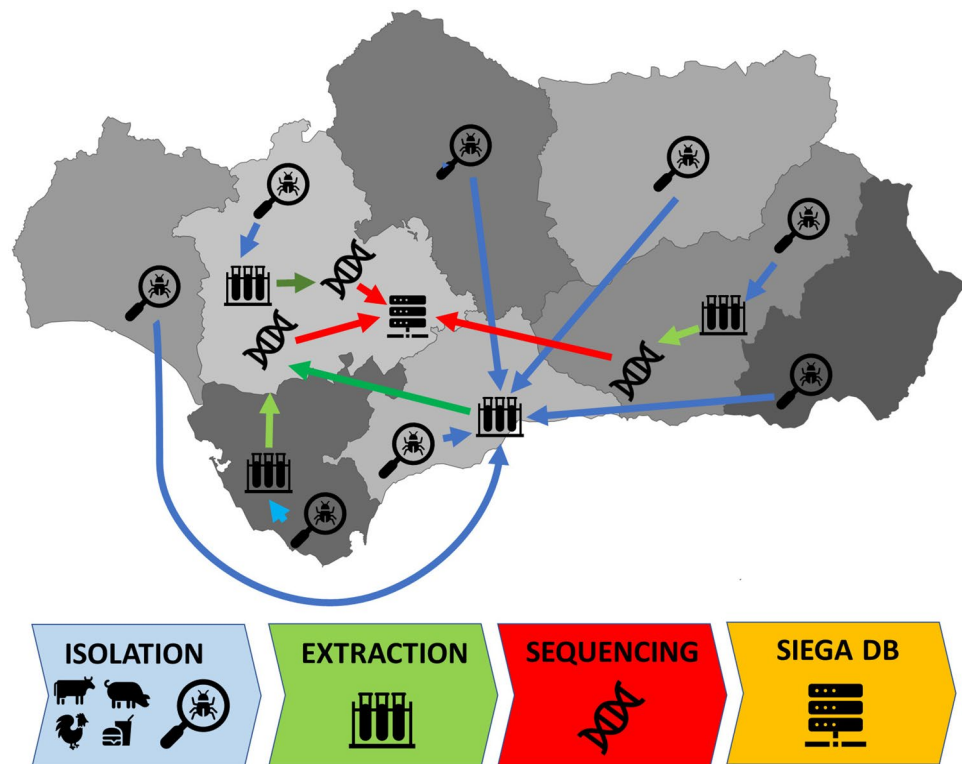
Figure 1. The SIEGA circuit. The different provinces that collect samples send them for extraction of DNA to the different reference laboratories, which is subsequently sent to the sequencing facilities and finally the resulting genomic data is uploaded in the central SIEGA data management system. The map of Andalusia was generated using mapSpain software from: https://ropenspain.github.io/mapSpain/articles/x02_mapasesp.html. The figure was generated with PowerPoint.

from investigations conducted within the framework of the management of foodborne outbreaks are included. All these samples are analyzed in designated official laboratories in accordance with the provisions of Article 37 of Regulation 2017/625, which, in the case of Andalusia, follow the aforementioned Instruction 130/2019.

Clinical origin samples are obtained in the exercise of the healthcare function of the referring centers on a voluntary basis, also applying the aforementioned Instruction 130/2019.

Here we present an application, SIEGA (acronym for "Integrated system of genomic epidemiology in Andalusia" in Spanish), that supports and facilitates the region-wide genomic surveillance system. As of July 2023, SIEGA contains a total of 1906 bacterial genomes corresponding to 670 *Salmonella enterica*, 688 *Listeria monocytogenes*, 276 *Campylobacter jejuni*, 191 *Escherichia coli,* 23 *Yersinia enterocolitica* and 58 *Legionella pneumophila*, collected at the different provinces of Andalusia in food products, factories, farms, water systems and human clinical samples (although these numbers change rapidly due to the continuous increase of samples). SIEGA contains information on the origin of the samples, their typing and resistance and virulence genes. It consist of two modules, one of them is public and contains descriptive information on the samples in the conventional NEXSTRAIN representation[25], and the other one, the SIEGA management data system, is a private Laboratory Information Management System (LIMS) for the use of the personnel of public health and the clinicians involved in the surveillance program. The LIMS allows users to build customized reports on the isolates, provides tools for detecting transmission chains and implements an automated alert system that promptly signals whenever a newly detected bacterium exhibits a predetermined genetic similarity to any existing database entry, enhancing vigilance and response capabilities. Some statistics on the isolates as well as a more detailed study on the relationships between them and the distribution of resistances across samples is provided.

Results
The public SIEGA webpage
SIEGA has a public website[25] where a detailed description of the circuit and updated information on the species under surveillance is available for the general public. Having a public webpage for a project of genomic surveillance of pathogens is vital for promoting transparency, disseminating knowledge, fostering collaboration, and bridging the gap between the scientific community and the general public. By providing open access to information on the surveillance achievements, SIEGA encourages participation from diverse stakeholders, and creates a more knowledgeable and engaged society in the ongoing fight against infectious diseases. SIEGA offers access to Nextstrain Auspice[26] viewers for the different species under surveillance: *Salmonella enterica, Listeria monocytogenes, Campylobacter jejuni, Escherichia coli, Legionella pneumophila* and *Yersinia enterocolitica*

(see Fig. 2). In the viewers, the public can explore the relationships among the samples sequenced in the region and other international samples of reference. It is also possible to locate in a map the geographical origin of the different isolates, including the animated options available in the interface, which emulate in a very visual way the transmission of the samples over time.

The SIEGA data management system

The SIEGA data management system serves as a private LIMS designed for utilization by personnel in public health and participating clinicians of the surveillance program. SIEGA facilitates the seamless uploading of raw sequencing data and orchestrates automated processing, including quality control assessments, considering the Guidelines for reporting Whole Genome Sequencing-based typing data through the EFSA One Health WGS System of the European Food Safety Authority (EFSA)[27]. Through this platform, users can generate tailored reports concerning the isolates, explore potential transmission chains, and deploy an automated alert mechanism that promptly signals any genetic similarity between newly identified bacteria and entries within the existing database. This system bolsters vigilance and response capabilities. Furthermore, SIEGA furnishes statistical insights into the isolates, along with a comprehensive exploration of their interrelationships and the distribution of resistances across samples.

Within the SIEGA interface, each organism has five distinct subsections to facilitate comprehensive analysis that include sample status, metadata, control results and Flexible Table, a wizard to combine metadata for complex representations (see Table 1).

The entire SIEGA data management system has been designed to free users of the need to have in-house expertise in genomic data management and resources to store such data. Simultaneously, the system offers a centralized database of all genomes sampled, allowing optimal exploitation of the results and the correct implementation of one health surveillance. It provides a convenient user permission structure to allows data sharing and collaborative work (if desired), detailed quality control for the standard pipelines used for data processing and accurate data traceability. Beyond facilitating collaborative work, user permission also helps with data privacy in samples of human origin. In any case, metadata does not include any field with personal identification and, ultimately, it is user responsibility not including any information that would reveal the origin of the sample. In general, data is treated in an open data philosophy where possible, according to FAIR principles (findable, accessible, interoperable and reusable)[28].

Detailed reports on each sample, that include MLST, cgMLST, serotyping outcomes, antimicrobial resistance and virulence genes, plasmids and a phylogenetic tree with the most related samples are provided. Additionally, customized phylogenetic analysis can be carried out in an easy and intuitive way. Finally, one of the most interesting features is the automatic alert system. SIEGA can be configured to automatically send a warning when a new sample is introduced that meets some criteria defined by the user, based on genetic distance, serotype, presence of antimicrobial or virulence genes, etc. (see Table 1 and Supplementary results for details).
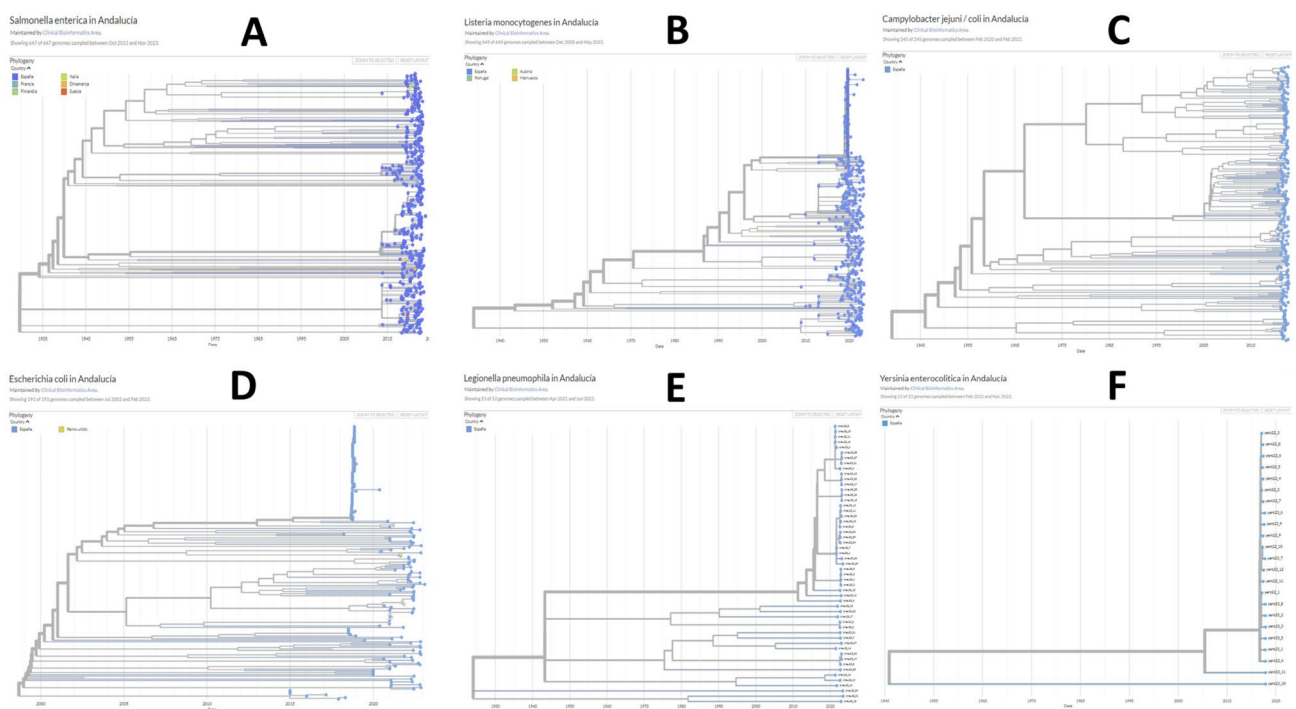

Figure 2. Phylogenies from Nextstrain viewers for: (**A**) *Salmonella enterica*, (**B**) *Listeria monocytogenes*, (**C**) *Campylobacter jejuni*, (**D**) *Escherichia coli*, (**E**) *Legionella pneumophila* and (**F**) *Yersinia enterocolitica*.

Feature	Description
Interface	With different subsections to facilitate the analysis: *Sample Status* information on each sample, including the quality control status *Metadata* information needed for sample and study traceability *Application control* information on the status and results of the pipeline of analysis *Results* detailed overview of the results obtained for any sample *Flexible Table* allows merging of any variety of result tables, metadata, or statuses for complex representation of results
User management	Access permission structure allowing users and group with many options to share data across users and groups
Data upload	Comma-separated values (CSV) file for metadata Fastq R1 and R2 files for each sample
Data processing and quality control	Quality control information on the result of each application is reportes (green: good, yellow: fair, red: poor)
Data traceability	Each sequenced sample has a distinctive identifier, used for associating relevant metadata
Automatic reporting	A report is generated for each sample including: MLST cgMLST Serotyping outcomes Antimicrobial resistance genes Virulence genes Plasmids A phylogenetic tree with the most related samples
Customized phylogenetic analysis	All existing tables can be combined using a wizard (flexible table) that enables the creation of combined tables with relevant data for each case, from which phylogenetic trees can be generated
Alert system	Automatic alerts based on various criteria can be established: Allelic distances Specific sequence type match Serotype predicted Presence of antimicrobial resistance genes Presence of virulence genes

Table 1. SIEGA features and description.

Advantages of SIEGA

The overarching goal of the SIEGA initiative, and particularly its SIEGA data management system, is to streamline the implementation of epidemiological surveillance, especially at the regional level or within large communities engaged in the One Health approach. It offers a continuously upgraded environment for managing genomic data, eliminating the need for end-users to establish their own bioinformatics teams for data processing and interpretation, as well as to invest in computer infrastructure for data storage and analysis. Furthermore, because data undergo uniform processing through a regularly updated pipelines adhering to international analysis standards, the results are consistent and can be readily compared. This democratizes genomic surveillance by involving all necessary stakeholders, as the SIEGA platform provides the essential resources for both data processing and interpretation.

Current users

Microbiology laboratories across different hospitals in Andalusia are actively utilizing SIEGA. In addition, the "Sistema de Vigilancia Epidemiológica de Andalucía" (SVEA) from the "Junta de Andalucía" and professionals in the field of Health Protection are also users of SIEGA. In adherence to the One Health principle, individuals involved in both animal health management and laboratory aspects have been incorporated as users of SIEGA.

Prospective new users must contact SIEGA through the contact address in the public SIEGA web page.

Genomic surveillance of isolates

Salmonella enterica

The SIEGA encompasses a dataset comprising 670 whole genome sequences of *Salmonella enterica*, which were sequenced from June 2020 to July 2023 using samples collected between 2013 and 2023. Within this dataset, 42.54% (285) of the samples were sourced from clinical origins, 34.63% from the food-related sector, and 21.34% from livestock sources. A total of 448 distinct Sequence Types (STs) were identified and categorized into clonal complexes (CCs). The prevailing ST, ST 309694 (corresponding to clonal complex ST-71), was encountered on 26 occasions. Additionally, 83 strains exhibited concurrence with more than one ST. 6 STs (ST-67337, ST-138467, ST-197094, ST-207307, ST-247937, and ST-320298) were found cross-wide clinical, food-related, and livestock-origin samples. Similarly, 18 STs were identified in both clinical and food samples, and 7 STs were identified in clinical and livestock-origin samples.

Listeria monocytogenes

The SIEGA includes a dataset comprising 678 whole genome sequences of *Listeria monocytogenes*, which were sequenced from June 2019 to July 2023. Within this dataset, 69.61% (472) of the samples were sourced from clinical origins, including all the samples from Andalusia sequenced by the Neisseria, Listeria and Bordetella Unit of the National Centre for Microbiology in Spain while investigating the Listeriosis outbreak caused by

contaminated stuffed pork in Spain in 2019[29], and 30.38% (206) from food origin. A total of 248 distinct STs were identified and categorized into CCs. The prevailing ST, ST 29,514 (corresponding to clonal complex ST-388), was encountered on 210 occasions.

Campylobacter spp
The SIEGA contains 276 whole genome sequences of *Campylobacter*, received between December 2020 and June 2023, corresponding to both *C. jejuni* and *C. coli*. Most of the sequences have been obtained from human clinical strains from two reference hospitals in Cádiz and Seville. There are some STs, grouped into CCs, which have been detected with greater frequency in clinical samples. From ST-16294 (corresponding to the clonal complex ST-206) 12 sequences have been obtained, with the interest of being detected from 2020 to 2023 and in a scattered way, with 8 isolates in Cádiz and 4 in Seville. Their identical virulome and resistome profiles have been recovered from these sequences, using the tools described, particularly ABRicate[30] on VFDB[31] and CARD[32], databases. Another frequent STs have been ST-12550 (ST-573CC) and ST-18855 (ST-52CC).

Escherichia coli
The SIEGA includes 121 whole genome sequences of *Escherichia coli*, 44 of them downloaded from the EnteroBase website[33] for reference and 77 of them collected between November 2021 and May 2023. To date, most of the sequences (72) have been obtained from food samples taken at retail level on behalf of the monitoring programme of anti-microbial resistance (AMR) according to the provisions of the Commission Implementing Decision (EU) 2020/1729[34] implemented in Andalusia. The human clinical strains come from two reference hospitals in Cádiz and Seville. There are STs, grouped into CCs, which have been detected with greater frequency in food samples. The prevailing ST, ST 169652 (corresponding to clonal complex ST-10) was encountered in 6 occasion all from food samples. Additionally, other 4 strains grouped in ST142026 (CC 155) and 3 strains exhibited concurrence with ST 191979 (CC 162) or 60064 (CC 93). To the date, no shared CCs have been detected in the food and human clinical origin samples.

Yersinia enterocolitica
The SIEGA encompasses 23 whole genome sequences of *Yersinia enterocolitica*, received in the year 2022, sampled between February and November. To date 21 (91.3%) of the sequences have been obtained from clinical samples from one of the reference hospitals, the Hospital Virgen del Rocio in Seville, and 2 were obtained from food samples. There are some STs, grouped into CCs, which have been detected with greater frequency in these clinical samples. The prevailing ST, ST-1574 (corresponding to clonal complex ST-135), was encountered on 4 occasions. Additionally, 3 strains exhibited concurrence with ST-52 and 2 strains grouped into ST-1716 (corresponding also to clonal complex ST-135). Figure 3 depicts the genetic relationships between all the *Yersinia enterocolitica* samples.

Legionella pneumophila
Legionella data stored in SIEGA, includes the comparative analyses of 58 *Legionella pneumophila* isolates during 2021–2023. Of these, 12 isolates corresponded to clinical isolates and 46 to environmental isolates. Clonal relation between the Isolates was determined by cgMLST. This scheme classified the isolates into 18 ST (sequence type). The most abundant being ST 293 (20 isolates, 34.5%) and 180F (11 isolates, 18.9%). Four ST (293, 427, 489 and 180F) were present in both clinical and environmental isolates. In addition, we identified 3 STs (95, 98, and 524F) in clinical isolates that are not associated with environmental origin, suggesting that they derived from unrecognized sources.

Analysis of antimicrobial resistance
In the EU, the new legislation related to the harmonized monitoring and reporting of AMR from 2021[35] authorized whole genome sequencing as an alternative method to supplementary phenotypic testing of *Salmonella* and *E. coli* in certain conditions. The SIEGA allows the monitoring of the presence of resistance genes in the different microorganisms facilitating the tracking of the dissemination or emergence of AMR throughout the food chain under a One Health approach. For example, the presence of AMR genes in the population of Salmonella included in the SIEGA can be analyzed, categorized by antimicrobial classes (Fig. 4). This analysis reveals that 52.7% (347) exhibit resistance genes to only 1 group of antimicrobials, while 15% (99) carry resistance genes to two distinct classes of antimicrobials. In contrast, 32.2% (212) demonstrate resistance genes to 3 or more classes of antimicrobials. In a similar manner, this could be carried out with the other microorganisms hosted in the database, or further analysis could be conducted by delving into the multiple variables, for instance, this could involve monitoring the emergence of *Salmonella* strains harboring colistin resistance genes, the occurrence of *Salmonella* strains harboring resistance genes to fluoroquinolones and third-generation cephalosporins or monitoring the presence of resistance genes to Critically Important Antibiotics (CIAs) in the database. The Fig. 4 represents a summary of the observed frequency of potential multi-resistances, represented as the number of different AMR genes corresponding to different antimicrobial classes harbored by each individual sample.

Another perspective for monitoring antimicrobial resistance is through the surveillance of plasmids harboring these resistance genes. SIEGA flexible tables allow the integration of data to both antimicrobial resistance and plasmid tracking, facilitating a comprehensive analysis. This data can be graphically represented on a phylogenetic tree, which illustrates the STs that have acquired resistance-bearing plasmids. Moreover, this representation can highlight whether such acquisitions have occurred within the same time, geographical region, livestock farm, food processing plant, grocery store or healthcare facility, thus providing critical insights into the patterns and pathways of resistance spread. Figure 5 illustrates a case is the plasmid NZ_AJ437107, which harbors a
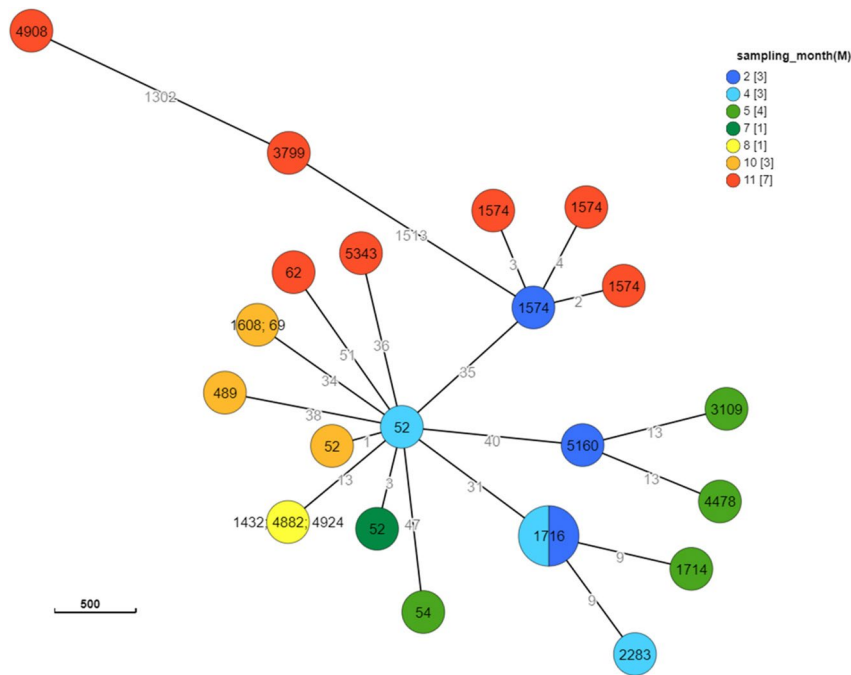
Figure 3. *Y. enterocolitica* GrapeTree representation, generated within the SIEGA application. Node labels represent ST (in some cases an ambiguous ST assignation occurred and more than one number is displayed) and node color correspond to the sampling month (a warm gradient has been used to better display the time scale). Numbers in the branches correspond to the allelic distances among nodes. The GrapeTree representation provides an intuitive visualization of the temporal scale of sampling and the genetic similarities among the samples. Using different labels from the metadata and the results tables, it is possible to obtain visual representations of many aspects of the epidemiology of the selected samples.
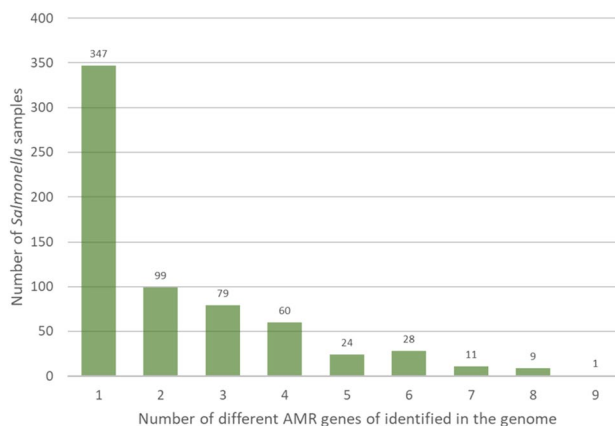
Figure 4. Observed frequency of potential multi-resistance cases found among the *Salmonella* samples sequenced. Number of samples in which from only one to up to 9 different AMR genes have been found.

beta-lactam resistance gene. This plasmid has been acquired in both livestock and food samples with the same sequence type, similar date and from the same geographical location, suggesting a potential selection pressure in certain environments favoring the acquisition of beta-lactam resistance genes.

Some successful SIEGA use cases

Despite its incipient use, SIEGA has already proven its usefulness in several cases. Among them it is worth mentioning the investigations carried out in connection with an outbreak of Salmonella Agona, as declared by Norwegian authorities[36], entailed, irrespective of field investigations, a comprehensive review of the SIEGA database in pursuit of genomic congruences. This study resulted in the absence of any coincident strains. Another case was the investigation regarding the food alert notification issued under the identifier 2020.5961 within the European Commission Rapid Alert System for Food and Feed (RASFF)[37], involving actions that extended
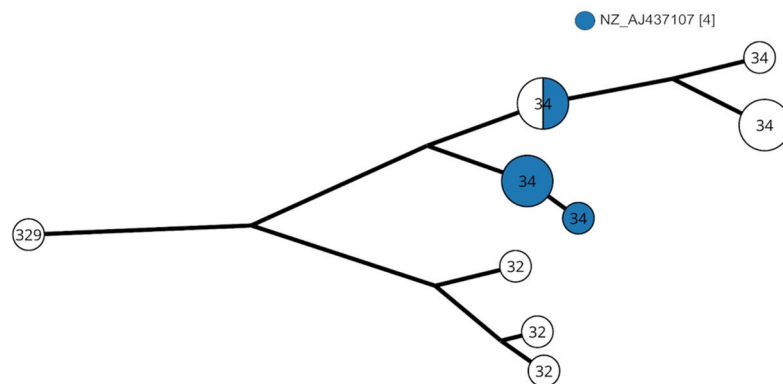
Figure 5. Phylogenetic tree, based on allelic differences (log scale), of Salmonella enterica isolated from the same locality, grouped by allelic profile (the circle size correlates with the number of samples). Blue dots represent strains harboring the plasmid NZ_AJ437107, and the number inside each dot indicates the corresponding MLST for each sample.

beyond on-site measures. These actions included obtaining genomic sequences from food samples supplied by the Finnish food safety authorities. These sequences were then integrated into the SIEGA database. However, no matches were identified both at the time of integration and among subsequent samples added to the system. Further investigations undertaken in relation to the "Joint ECDC-EFSA rapid outbreak assessment" released on July 27, 2023[38], in which Spain was identified as one of the conceivable sources of the suspected transmission vehicle, involved, regardless of field actions, a reassessment of the SIEGA database in search of genomic coincidences, resulting in the non-existence of any matching strain.

These cases clearly illustrate how SIEGA is valuable in ruling out the existence of coincident strains in our database with strains from evaluations of other outbreaks detected at the national or European level, thus facilitating decision-making. On the other hand, alerts are generated that point out possible connections between the stored strains and the ongoing outbreaks, accelerating the possible identification of the source of origin in an outbreak, an aspect that is very difficult to identify with previous methods.

Discussion

In order to effectively apply the One Health approach to the surveillance and control of antimicrobial resistance, it is crucial to adopt methodologies that ensure consistent results across diverse settings and various sample sources, facilitating comprehensive and reliable data analysis[39]. Traditionally, surveillance efforts have predominantly centered around specific organisms, prioritizing genotypic and/or phenotypic characteristics that are deemed significant for individual pathogens. Consequently, a multitude of distinct methodologies emerged, giving rise to challenges of consistency, applicability, standardization, and scalability across laboratories and pathogens. Addressing these issues necessitated the implementation of rigorous protocols, regular quality and harmonization controls, as well as the adoption of shared platforms and common controls for result dissemination, thereby mitigating some of the aforementioned concerns[40]. Nonetheless, the advent of high-throughput sequencing and the accessibility of obtaining comprehensive or nearly comprehensive genome sequences from bacterial isolates within a reasonable timeframe and cost have brought about a profound transformation in the pursuit of surveillance objectives, addressing a significant portion of the previously mentioned challenges. While complete realization remains a work in progress, the trajectory appears unambiguous, with the majority of agencies transitioning toward genome-centric surveillance. This approach presents numerous advantages, notably the capacity to establish connections among surveillance outcomes across diverse levels, locations, and species, which holds particular relevance for the One Health framework[41,42].

Currently, the genomic surveillance in Andalusia comprises 6 main pathogens, although new pathogens will be included in brief, like Vibrio or others, pending further decisions. Specifically, *Campylobacter* is one of the major foodborne pathogens of concern in its growing trend of antimicrobial resistance. *C. jejuni* and *E. coli* are the major causative agents, with *C. jejuni* contributing to most of the cases in approximately 90% in the world. Infection is transmitted to humans due to consumption of contaminated food and water[43]. It is very necessary to establish the chains of transmission between reservoir animals, food and humans. There are few published studies on the genomic characterization of human clinical strains[44,45] and their relationship with outbreaks[46]. It would be very useful to establish sentinel surveillance, such as the one implemented in Ireland recently[47]. The difficulty in maintaining the viability of *Campylobacter* in environmental samples is a challenge to increase the data that provide comparative information with clinical strains.

Although SIEGA represents a regional solution, Andalusia due to its large, country-like size, can be considered an illustrative case of implementation for a medium-sized region in Europe, serving as an exemplary guide for other authorities wishing to adopt a similar model. Actually, within an international context, SIEGA has yielded results in ruling out the involvement of products originating from Andalusia in outbreaks occurring in other European countries. Moreover, upon reaching a national-level management agreement, it will facilitate the inclusion of non-human origin strains in the European Food Safety Authority (EFSA) sequencing database,

as the quality controls and bioinformatics analysis tools described in [reference] have been implemented. This will enable both EFSA and the ECDC at the European level to interact with genomic information originating from Andalusia. It is also important to remark that SIEGA has been designed with the capability to incorporate new domains, such as strains originating from other territories, countries or entities with its multi-level user management system and flexible permissions. Obviously, expanding the scope and the user's base comes with the challenge of scalability and sustainability. While sustainability is assured by the support of the Andalusian Local Ministry of Health and technical limitations due to scalability to a large number of samples are not expected in a near future, a limitation could arise from the need of implementing many diverse pathogens (viruses, bacteria, and maybe fungi) in the near future, to implement new sequencing technologies and other types of samples, like metagenomic surveys, wastewater surveillance, etc.

There are some publicly available tools that can be used for genomic surveillance. Pathogenwatch[48] is perhaps the most popular one, that share some functionalities with SIEGA. It can upload and process about 20 of pathogen species and provide some functionalities like AMR prediction and Core SNP-based trees, although data sharing facilities are quite limited and it does not implement anything like an alert system. The EFSA One Health WGS system[27] is another tool with restricted access, only for official data providers, designed by country officers, also with limited functionalities and managing only three pathogen species. Another recent initiative, the Swiss Pathogen Surveillance Platform, developed a tool for genomic surveillance, but mainly oriented to facilitate visualization of the results from an epidemiologic point of view, like maps, time evolution of samples, etc.[49]. Other specific tool are Microreact[50] or Nexstrain[26], with functionalities limited to data visualization and sharing for genomic epidemiology.

Summarizing, this work presents SIEGA, an advanced integrated One Health system that allows precise surveillance of environmental pathogens, permitting a comprehensive characterization of new isolates with data quality control and data traceability included. SIEGA facilitates automated generation of customized reports that include similarities with other samples, represented as a dendrogram, potential AMR or virulence genes detected, among other details. The alerting functionality of SIEGA, that alerts as immediately as one sample is introduced that meets a predefined similarity conditions, is a revolutionary tool for detecting transmission chains at an early stage. In addition, the possibility of crossing metadata and representing them over sample dendrograms allows performing detailed retrospective studies of different sample features (e.g., emergence or transmission of AMR, etc.) Moreover, since SIEGA connects environmental with clinical samples it allows tracing clinical occurrences back to their environmental origins, permitting rapid interventions targeted to the source of the outbreak. In addition, the success in discarding relationships in the use cases underscore the intricate challenges in elucidating the etiology and dynamics of microorganism-related incidents. The pursuit of genomic concordance within the SIEGA database, although yielding no direct matches, serves to illuminate the genetic diversification that these pathogens can exhibit. These episodes emphasize the need for continued vigilance, inter-agency cooperation, and cutting-edge molecular methodologies to fortify our comprehension and management of such microbial phenomena. It is worth noting that the modular structure of SIEGA allows the incorporation of new pathogens as the surveillance policies consider them relevant.

Thus, by leveraging high-throughput sequencing technologies, advanced bioinformatics tools, and robust data sharing platform, SIEGA offers a comprehensive view of the genetic landscape of pathogens across different geographical regions and host populations.

Conclusions

Centralized circuits of genomic surveillance based on whole genome sequencing (WGS) as SIEGA provide a convenient and efficient way to monitor infectious diseases and detect outbreaks in real-time[51]. Such facilities enable the characterization and relatedness determination of bacterial isolates, aiding in tracking transmission patterns and implementing effective infection control measures[4]. Moreover, WGS has been incorporated into public health surveillance systems, offering significant contributions to outbreak investigations, infection prevention, and control[51]. The potential integration of WGS into epidemiological investigations has been highlighted, emphasizing the need to establish optimal models for data integration and evaluate public health impacts resulting from genomic surveillance[52]. By harnessing the power of whole-genome sequencing, centralized circuits of genomic surveillance offer immense potential for improving disease surveillance and response, ultimately contributing to the overarching goals of the One Health framework[1,4,39].

Methods

DNA extraction protocol

For the majority of the samples, DNA extraction was conducted using the PureLink Genomic DNA Mini Kit (Invitrogen). Subsequently, quantification of the extracted DNA was performed using the QUBIT FLEX fluorometer, and the quality of the eluate was assessed through electrophoresis using the Egel Power Snap Electrophoresis Device.

Sequencing

The majority of the samples have been sequenced at the CABIMER Genomic Unit, with a significant contribution from Listeria sequences obtained from the National Center of Microbiology (469 samples). These sequences were incorporated under a mutual agreement for sequence exchange established between the Andalusian Public Foundation Progress and Health-FPS and the National Center of Microbiology in 2020.

Whole genome sequencing of the isolates has been carried out at the CABIMER genomic facility. Library construction and sequencing was performed at the Genomics Core Facility of CABIMER. DNA libraries were prepared using Nextera DNA Flex Library prep kit (Illumina) following the manufacturer's instructions. Currently,

high-throughput sequencing was performed on the NextSeq 500 Sequencing System (Illumina), although other sequencing technologies, such as nanopore will be included in a near future.

Sequencing data processing

The raw reads are filtered using the fastp[53] application (v0.23.4), followed by a search for potential contamination from other organisms using the kraken2[54] and bracken[55] applications (v2.1.2). Coverage quality control is carried out using qualimap2[56] (v2.2.2). Subsequently, with quality-controlled reads, a de novo assembly is performed using SPAdes[57] (v3.15.4) and the quality of the assembly is assessed using QUAST[58] (v5.0.2).

Sample typing

MLST (Multi-Locus Sequence Typing) is acquired using the MentaLiST[59] application (v1.0.0) for those organisms whose databases are updated within this tool. CgMLST (core genome Multi-Locus Sequence Typing) profiles are generated using Chewbbaca[60] (v3.0.0) with the schema available in the Chewie Nomenclature Server (chewie-NS)[61] for *Listeria monocytogenes. Salmonella enterica* and *Escherichia coli* schemas are filtered based on the EFSA gene list[62] CgMLST profiles. Moreover cgMLST profiles based in other databases are obtained using a custom script that utilizes the BLAST + [63] tool (v2.12.0 +) and allelic profiles from the following public databases as references: Pasteur Institute[64] for *Listeria monocytogenes*, EnteroBase[33] for *Yersinia enterocolitica, Salmonella enterica* and *Escherichia coli* and PubMLST[65] for *Campylobacter jejuni/coli.* Additionally, the applications LisSero[66,67] (v0.4.1), SeqSero2[68] (v1.2.1), and SerotypeFinder[69] (v2.0.1) are employed to determine the serotype of *Listeria monocytogenes*, *Salmonella enterica*, and *Escherichia coli*, respectively.

New species can be easily incorporated either by using its own typing schema, if the corresponding application is available, or with a customized script, as mentioned above.

AMR genes

To identify antimicrobial resistances, two approaches were employed. Firstly, the ResFinder[70] application (v4.5.0) was utilized, along with its dedicated database. Secondly, the ABRicate[30] tool (v1.0.1) was employed in conjunction with various databases, namely CARD[32], MegaRES[71], and ARG-ANNOT[72].

Virulence genes

To identify virulence genes, two applications were employed: VirulenceFinder[73] (v2.0.4), with its dedicated database and ABRicate using the Virulence Factor Database (VFDB)[31] as a reference.

Plasmids

To identify plasmids present in the samples two approaches are employed: the PlasmidFinder[74] application (v2.1.6) and the Mash Screen[75] application (v2.3) combined with the PlasDB database[76].

Core genome determination and phylogenetic analysis

Using the parSNP[77] application (v1.7.4), the core genome is analyzed, SNP selection is performed, and subsequent phylogenetic trees are constructed based on SNP differences for each of the organisms present in SIEGA. On the other hand, allelic differences are obtained using a custom script that compares the cgMLST of each sample for each organism, generating a matrix for each organism. This matrix is then utilized by GrapeTree[78] (v2.1) to construct the phylogenetic tree based on allelic distances, which could be visualized with GrapeTree or Taxonium[79]. The ETE3[80] application (v3.1.2) is used for the selection and generation of sub-trees included in the reports. The reference sequences (NCBI database identifiers) used for the different species were for *Listeria monocytogenes* NC_003210.1, *Salmonella enterica* NZ_SRHS01000001.1, *Escherichia coli* NC_000913.3, *Campylobacter jejuni / coli* NZ_CYQA01000001.1, *Yersinia enterocolitica* NZ_CQAE01000001.1 and *Legionella pneumophila* NZ_CP015941.1.

Data availability

The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request. Data are located in controlled access data storage at SIEGA application.

References

1. Mackenzie, J. S. & Jeggo, M. The one health approach—Why is it so important?. *Trop. Med. Infect. Dis.* **4**, 88 (2019).
2. Rüegg, S. R. *et al.* A blueprint to evaluate One Health. *Front. Public Health* **5**, 20 (2017).
3. Machalaba, C. C. *et al.* Global avian influenza surveillance in wild birds: A strategy to capture viral diversity. *Emerg. Infect. Dis.* **21**, 14145 (2015).
4. Queenan, K. *et al.* Roadmap to a One Health agenda 2030. *CABI Rev.* https://doi.org/10.1079/PAVSNNR201712014 (2017).
5. Mather, A. *et al.* Distinguishable epidemics of multidrug-resistant Salmonella Typhimurium DT104 in different hosts. *Science* **341**, 1514–1517 (2013).
6. Scaltriti, E. *et al.* Differential single nucleotide polymorphism-based analysis of an outbreak caused by Salmonella enterica serovar Manhattan reveals epidemiological details missed by standard pulsed-field gel electrophoresis. *J. Clin. Microbiol.* **53**, 1227–1238 (2015).
7. Pattabhiramaiah, M. & Mallikarjunaiah, S. *Sequencing Technologies in Microbial Food Safety and Quality* 393–424 (CRC Press, 2021).

8. Yachison, C. A. *et al.* The validation and implications of using whole genome sequencing as a replacement for traditional serotyping for a national Salmonella reference laboratory. *Front. Microbiol.* **8**, 1044 (2017).
9. Banerji, S., Simon, S., Tille, A., Fruth, A. & Flieger, A. Genome-based Salmonella serotyping as the new gold standard. *Sci. Rep.* **10**, 1–10 (2020).
10. EFSA Panel on Biological Hazards. Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms. *EFSA J.* **17**, e05898 (2019).
11. Tewolde, R. *et al.* MOST: A modified MLST typing tool based on short read sequencing. *PeerJ* **4**, e2308 (2016).
12. Zhang, S. *et al.* SeqSero2: Rapid and improved Salmonella serotype determination using whole-genome sequencing data. *Appl. Environ. Microbiol.* **85**, e01746-e1719 (2019).
13. Yoshida, C. E. *et al.* The Salmonella in silico typing resource (SISTR): An open web-accessible tool for rapidly typing and subtyping draft Salmonella genome assemblies. *PLoS ONE* **11**, e0147101 (2016).
14. Li, Z. *et al.* Whole genome sequencing analyses of Listeria monocytogenes that persisted in a milkshake machine for a year and caused illnesses in Washington State. *BMC Microbiol.* **17**, 134. https://doi.org/10.1186/s12866-017-1043-1 (2017).
15. Rounds, J. *et al.* Prospective Salmonella Enteritidis surveillance and outbreak detection using whole genome sequencing, Minnesota 2015–2017. *Epidemiol. Infect.* **148**, e254 (2020).
16. Pendleton, S., Hanning, I., Biswas, D. & Ricke, S. C. Evaluation of whole-genome sequencing as a genotyping tool for Campylobacter jejuni in comparison with pulsed-field gel electrophoresis and flaA typing1 1Presented as part of the Next-Generation Sequencing Tools: Applications for Poultry Production and Food Safety Symposium at the Poultry Science Association's annual meeting in Athens, Georgia, July 10, 2012. *Poultry Sci.* **92**, 573–580. https://doi.org/10.3382/ps.2012-02695 (2013).
17. Catoiu, E. A., Phaneuf, P., Monk, J. & Palsson, B. O. Whole-genome sequences from wild-type and laboratory-evolved strains define the alleleome and establish its hallmarks. *Proc. Natl. Acad. Sci. USA* **120**, e2218835120. https://doi.org/10.1073/pnas.2218835120 (2023).
18. Petzold, M., Prior, K., Moran-Gilad, J., Harmsen, D. & Lück, C. Epidemiological information is key when interpreting whole genome sequence data - lessons learned from a large Legionella pneumophila outbreak in Warstein, Germany, 2013. *Euro Surveill.* https://doi.org/10.2807/1560-7917.Es.2017.22.45.17-00137 (2017).
19. ECDC. *ECDC Strategic Framework for the Integration of Molecular and Genomic Typing into European Surveillance and Multi-country Outbreak Investigations.* https://www.ecdc.europa.eu/sites/default/files/documents/framework-for-genomic-surveillance.pdf (2019).
20. WHO. *Whole-Genome Sequencing for Surveillance of Antimicrobial Resistance.* https://apps.who.int/iris/bitstream/handle/10665/334354/9789240011007-eng.pdf (2020).
21. Mancilla-Becerra, L. M., Lías-Macías, T., Ramírez-Jiménez, C. L. & León, J. B. Multidrug-resistant bacterial foodborne pathogens: Impact on human health and economy. *Pathog. Bact.* **2019**, 1–18 (2019).
22. *Commitment to the Andalusian Public Foundation of Progress and Health to Carry out the Maintenance, Updating and Improvement of SIEGA.* https://www.juntadeandalucia.es/haciendayadministracionpublica/apl/pdc_sirec_documentacion/rest/descargar/documento/73341 (2022).
23. Andalucía, J. D. *Instrucción 130/2019 sobre Tratamiento y Secuenciación de Aislados de Agentes Biológicos en Andalucía.* https://juntadeandalucia.es/export/drupaljda/Instrucci%C3%B3n.130-2019%20Secuenciaci%C3%B3n%20y%20Aislamientos%20rev1.pdf (2019).
24. Chattaway, M. A. *et al.* The transformation of reference microbiology methods and surveillance for Salmonella with the use of whole genome sequencing in England and Wales. *Front. Public Health* **7**, 317 (2019).
25. *SIEGA (Integrated System for Genomic Epidemiology in Andalusia).* http://clinbioinfosspa.es/projects/siega/ (2020).
26. Hadfield, J. *et al.* Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
27. Authority, E. F. S. *et al. Guidelines for Reporting Whole Genome Sequencing-Based Typing Data Through the EFSA One Health WGS System.* Report No. 2397–8325 (Wiley Online Library, 2022).
28. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
29. Fernández-Martínez, N. F. *et al.* Listeriosis outbreak caused by contaminated stuffed pork, Andalusia, Spain, July to October 2019. *Eurosurveillance* **27**, 2200279 (2022).
30. Seemann, T. *Abricate.* https://github.com/tseemann/abricate (2020).
31. Liu, B., Zheng, D., Jin, Q., Chen, L. & Yang, J. VFDB 2019: A comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* **47**, D687–D692. https://doi.org/10.1093/nar/gky1080 (2018).
32. Alcock, B. P. *et al.* CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525. https://doi.org/10.1093/nar/gkz935 (2019).
33. Zhou, Z. *et al.* The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Res.* **30**, 138–152 (2020).
34. *Commission Implementing Decision (EU) 2020/1729 of 17 November 2020 on the Monitoring and Reporting of Antimicrobial Resistance in Zoonotic and Commensal Bacteria and Repealing Implementing Decision 2013/652/EU.* https://eur-lex.europa.eu/eli/dec_impl/2020/1729/oj (2020).
35. Authoity, E. F. S. & Prevention Control, E. C. F. D. The European Union Summary Report on Antimicrobial Resistance in zoonotic and indicator bacteria from humans, animals and food in 2020/2021. *EFSA J.* **21**, e07867 (2023).
36. *Agurk fra Spania er mistenkt smittekilde i utbrudd av salmonella.* https://www.fhi.no/nyheter/2022/agurk-fra-spania-er-mistenkt-smittekilde-i-utbrudd-av-salmonella/ (2022).
37. *NOTIFICATION 2020.5961. Foodborne outbreak suspected to be caused by Salmonella enterica ser. Kedougou in courgettes from NL and ES.* https://webgate.ec.europa.eu/rasff-window/screen/notification/457471 (2020).
38. Prevention, E. C. F. D. & Control, E. F. S. A. Multi-country outbreak of Salmonella Senftenberg ST14 infections, possibly linked to cherry-like tomatoes. *EFSA Support. Publ.* **20**, 8211 (2023).
39. Gerner-Smidt, P. *et al.* Whole genome sequencing: Bridging one-health surveillance of foodborne diseases. *Front. Public Health* **7**, 172 (2019).
40. Nadon, C. *et al.* PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Eurosurveillance* **22**, 30544 (2017).
41. Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* **19**, 9–20 (2018).
42. Gwinn, M., MacCannell, D. & Armstrong, G. L. Next-generation sequencing of infectious pathogens. *JAMA* **321**, 893–894 (2019).
43. EFSA, J. European Centre for Disease Prevention and Control (ECDC); European Food Safety Authority (EFSA); European Medicines Agency (EMA). *Joint Interagency Antimicrobial Consumption and Resistance Analysis (JIACRA) Report* **15**, e04872 (2017).
44. Redondo, N., Carroll, A. & McNamara, E. Molecular characterization of Campylobacter causing human clinical infection using whole-genome sequencing: Virulence, antimicrobial resistance and phylogeny in Ireland. *PLoS ONE* **14**, e0219088 (2019).
45. Bravo, V. *et al.* Genomic analysis of the diversity, antimicrobial resistance and virulence potential of clinical Campylobacter jejuni and Campylobacter coli strains from Chile. *PLoS Negl. Trop. Dis.* **15**, e0009207 (2021).
46. Joseph, L. A. *et al.* Evaluation of core genome and whole genome multilocus sequence typing schemes for Campylobacter jejuni and Campylobacter coli outbreak detection in the USA. *Microb. Genom.* **9**, 1012 (2023).

47. Brehony, C., Lanigan, D., Carroll, A. & McNamara, E. Establishment of sentinel surveillance of human clinical campylobacteriosis in Ireland. *Zoonoses Public Health* **68**, 121–130 (2021).
48. *Pathogenwatch*. https://pathogen.watch/ (2018).
49. Neves, A. *et al.* The Swiss Pathogen Surveillance Platform–towards a nation-wide One Health data exchange platform for bacterial, viral and fungal genomics and associated metadata. *Microb. Genom.* **9**, 001001 (2023).
50. *Microreact*. https://microreact.org/ (2024).
51. Lo, S. W. & Jamrozy, D. Genomics and epidemiological surveillance. *Nat. Rev. Microbiol.* **18**, 478–478. https://doi.org/10.1038/s41579-020-0421-0 (2020).
52. Ferdinand, A. S. *et al.* An implementation science approach to evaluating pathogen whole genome sequencing in public health. *Genome Med.* **13**, 121. https://doi.org/10.1186/s13073-021-00934-7 (2021).
53. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
54. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257. https://doi.org/10.1186/s13059-019-1891-0 (2019).
55. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
56. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294. https://doi.org/10.1093/bioinformatics/btv566 (2015).
57. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes de novo assembler. *Curr. Protoc. Bioinform.* **70**, e102 (2020).
58. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075. https://doi.org/10.1093/bioinformatics/btt086 (2013).
59. Feijao, P. *et al.* MentaLiST: A fast MLST caller for large MLST schemes. *Microb. Genom.* **4**, 146 (2018).
60. Silva, M. *et al.* chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microb. Genom.* **4**, 166 (2018).
61. Mamede, R., Vila-Cerqueira, P., Silva, M., Carriço, J. A. & Ramirez, M. Chewie Nomenclature Server (chewie-NS): A deployable nomenclature server for easy sharing of core and whole genome MLST schemas. *Nucleic Acids Res.* **49**, D660–D666. https://doi.org/10.1093/nar/gkaa889 (2020).
62. Mirko, R. *EFSA cgMLST Gene Lists for Escherichia coli and Salmonella enterica chewieNS Schema*. https://zenodo.org/record/6655441 (2022).
63. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinform.* **10**, 421. https://doi.org/10.1186/1471-2105-10-421 (2009).
64. Moura, A. *et al.* Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. *Nature Microbiol.* **2**, 1–10 (2016).
65. Jolley, K. A., Bray, J. E. & Maiden, M. C. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* **3**, 124 (2018).
66. Doumith, M., Buchrieser, C., Glaser, P., Jacquet, C. & Martin, P. Differentiation of the major Listeria monocytogenes serovars by multiplex PCR. *J. Clin. Microbiol.* **42**, 3819–3822 (2004).
67. Kwong, J., Zhang, J. & Seemann, T. *LisSero. In Silico Serogroup Typing Prediction for Listeria Monocytogenes*. https://github.com/MDU-PHL/LisSero (2021).
68. Zhang, S. *et al.* SeqSero2: Rapid and improved serotype determination using whole-genome sequencing data. *Appl. Environ. Microbiol.* **85**, e01746-01719. https://doi.org/10.1128/AEM.01746-19 (2019).
69. Joensen, K. G., Tetzschner, A. M. M., Iguchi, A., Aarestrup, F. M. & Scheutz, F. Rapid and easy *In Silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J. Clin. Microbiol.* **53**, 2410–2426. https://doi.org/10.1128/jcm.00008-15 (2015).
70. Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **75**, 3491–3500. https://doi.org/10.1093/jac/dkaa345 (2020).
71. Lakin, S. M. *et al.* MEGARes: An antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Res.* **45**, D574–D580. https://doi.org/10.1093/nar/gkw1009 (2016).
72. Gupta, S. K. *et al.* ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.* **58**, 212–220 (2014).
73. Joensen, K. G. *et al.* Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *J. Clin. Microbiol.* **52**, 1501–1510 (2014).
74. Carattoli, A. & Hasman, H. PlasmidFinder and in silico pMLST: Identification and typing of plasmid replicons in whole-genome sequencing (WGS). *Horizontal Gene Transfer: Methods and Protocols* 285–294 (2020).
75. Ondov, B. D. *et al.* Mash Screen: High-throughput sequence containment estimation for genome discovery. *Genome Biol.* **20**, 232. https://doi.org/10.1186/s13059-019-1841-x (2019).
76. Galata, V., Fehlmann, T., Backes, C. & Keller, A. PLSDB: A resource of complete bacterial plasmids. *Nucleic Acids Res.* **47**, D195–D202. https://doi.org/10.1093/nar/gky1050 (2018).
77. Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* **15**, 1–15 (2014).
78. Zhou, Z. *et al.* GrapeTree: Visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res.* **28**, 1395–1404 (2018).
79. Sanderson, T. Taxonium, a web-based tool for exploring large phylogenetic trees. *eLife* **11**, e82392. https://doi.org/10.7554/eLife.82392 (2022).
80. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638. https://doi.org/10.1093/molbev/msw046 (2016).

Acknowledgements

Author contributions

Conceptualization, JD. and J.A.C.; methodology, C.S.C.S., J.P.F., M.L., A.A., M.P.A., E.A., V.E.J. and L.P.C.; software, E.A.R.; formal analysis, C.S.C.S. and J.P.F.; investigation, M.A.R.I., J.A.L., F.G., N.L., U.A., I.M.V.; data

curation, C.S.C.S., J.P.F., M.L., A.A.; writing—original draft preparation, J.D., J.A.C., C.S.C.S., J.P.F.; writing—review and editing, J.D., J.A.C., C.S.C.S., J.P.F., J.A.L., M.A.R.I., C.M.L.; supervision, J.D., J.A.C.; funding acquisition, J.D., J.A.L., F.G., C.S.C.S. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-70107-0.

Correspondence and requests for materials should be addressed to J.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.