A global resource for genomic predictions of antimicrobial resistance and surveillance of *Salmonella* Typhi at Pathogenwatch

Authors: Silvia Argimón[1], Corin A. Yeats[2], Richard J. Goater[1,11], Khalil Abudahab[1], Benjamin Taylor[2], Anthony Underwood[1], Leonor Sánchez-Busó[2], Vanessa K. Wong[3], Zoe A. Dyson[3,4,9], Satheesh Nair[5], Se Eun Park[6], Florian Marks[6], Andrew J. Page[7,11], Jacqueline A. Keane[7], Stephen Baker[8], Kathryn E. Holt[4,9], Gordon Dougan[3], and David M. Aanensen[1,2].

1 Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Hinxton, Cambridgeshire, United Kingdom

2 Centre for Genomic Pathogen Surveillance, Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, Oxfordshire, United Kingdom

3 Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, United Kingdom

4 London School of Hygiene and Tropical Medicine, London, United Kingdom

5 Gastrointestinal Bacterial Reference Unit, Public Health England, Colindale, London, United Kingdom

6 International Vaccine Institute, Seoul, Korea

7 Pathogen Informatics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, United Kingdom

8 Cambridge Institute of Therapeutic Immunology & Infectious Disease, Department of Medicine, University of Cambridge, Cambridge, United Kingdom

9 Department of Infectious Diseases, Monash University, Melbourne, Australia

10 Present address: European Molecular Biology Lab, Heidelberg, Baden-Wuerttemberg, Germany

11 Present address: Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, United Kingdom

Corresponding Authors: Silvia Argimón (silvia.argimon@cgps.group) and David M. Aanensen (david.aanensen@cgps.group)

Abstract

Background: Microbial whole-genome sequencing (WGS) is now increasingly used to inform public health investigations of infectious disease. This approach has transformed our understanding of the global population structure of *Salmonella enterica* serovar Typhi (*S.* Typhi), the causative agent of typhoid fever. WGS has been particularly informative for understanding the global spread of multi-drug resistant (MDR) typhoid. As WGS capacity becomes more decentralised, there is a growing opportunity for collaboration and sharing of surveillance data within and between countries to inform disease control policies. This requires freely available, community driven tools that reduce the barriers to access genomic data for public health surveillance and that deliver genomic data on a global scale.

Methods: Here we present the Pathogenwatch (https://pathogen.watch/styphi) scheme for *S.* Typhi, a web application enabling the rapid identification of genomic markers of antimicrobial resistance (AMR) and contextualization with public genomic data to identify high-risk clones at a population level. Data are delivered in single genome reports or in collections of genomes combined with geographic and other data using trees, maps and tables.

Results: We show that the clustering of *S.* Typhi genomes in Pathogenwatch is comparable to established bioinformatics methods, and that genomic predictions of AMR are largely concordant with phenotypic drug susceptibility data. We demonstrate the public health utility of Pathogenwatch with examples selected from over 4,300 public genomes available in the application.

Conclusions: Pathogenwatch democratises genomic epidemiology of *S.* Typhi by providing an intuitive entry point for the analysis of WGS and linked epidemiological data, enabling international public health monitoring of the emergence and spread of high risk clones.

Keywords: *Salmonella* Typhi, typhoid fever, enteric fever, antimicrobial resistance, genomics, whole-genome sequencing, surveillance, genomics, epidemiology, public health.

Introduction

Bacterial pathogens have the potential for rapid evolution and adaptation (1). The ability to rapidly sequence microbial genomes directly from the field is facilitating the tracking of pathogen evolution in real-time and in a geographical context. Genomic surveillance provides the opportunity to identify the emergence of genetic signatures indicating antimicrobial resistance (AMR), or adaptation to host, facilitating early intervention and minimising wider dissemination. Consequently, genomic data has the ability to transform the way in which we manage the emergence of microbes that pose a direct threat to human health in real time.

Although pathogen genomic data is being generated at a remarkable rate, we need to bridge the gap between genome sciences and public health with tools that make these data broadly and rapidly accessible to those who are not expert in genomics. To maximise the impact of ongoing surveillance programs, these tools need to quickly highlight high-risk clones by assigning isolates to distinct lineages and identifying genetic elements associated with clinically relevant features such as AMR or virulence. In this way, new isolates can be examined against the backdrop of a population framework that is continuously updated and that enables both the contextualisation of local outbreaks and the interpretation of global patterns.

Salmonella enterica subsp. *enterica* serovar Typhi (*S.* Typhi) causes typhoid (enteric) fever, a disease that affects approximately 20-30 million people every year (2, 3). The disease is predominant in low-income communities where public health infrastructure is poorly resourced. Similar to other infections, typhoid treatment is compromised by the emergence of *S.* Typhi with resistance to multiple antimicrobials, including those currently used for treatment (3). Until recently, epidemiological investigations and surveillance of typhoid fever have employed alternative molecular techniques such as pulse-field gel electrophoresis (PFGE (4)), multi-locus sequence typing (MLST (5)), multiple-locus variable-number tandem-repeat (VNTR) analysis (MLVA (6)), and phage-typing (4), which offer limited resolution for a bacterium that exhibits very limited genetic variability. Whole genome sequencing (WGS) has proven key to identify *S.* Typhi high-risk clones by linking the population structure to the presence of AMR elements. For example, the recent resurgence of multi-drug resistant (MDR) typhoid (defined as resistance to all the historical first-line agents chloramphenicol, ampicillin and co-trimoxazole) has been explained in part by the global spread of an MDR *S.* Typhi lineage known as haplotype H58 or subclade

4.3.1 (7, 8), which is associated with both acquired AMR genes (conferring MDR) and fluoroquinolone resistance mutations (7, 9).

WGS is increasingly being implemented in local and national public health laboratories, and web applications can provide rapid analysis and access to actionable information for infection control in the context of a global population framework. Online resources are available for *Salmonella enterica* species for the identification of acquired AMR mechanisms (10) and for *in silico* typing and visualisation of genome variation and relatedness based on WGS data (11, 12). Here, we describe Typhi Pathogenwatch, a web application to support genomic epidemiology and public health surveillance of *S.* Typhi through the rapid identification and linking of genetic lineages with AMR determinants. Typhi Pathogenwatch rapidly places new genomes within the population context, predicts their genotype according to established nomenclatures (5, 8, 11), and detects the presence of AMR determinants and plasmid replicon genes to assess public health risk. Typhi Pathogenwatch displays this information interactively, allowing users to link lineages, AMR patterns, geographical data and other metadata to quickly determine if similar strains have been previously identified, where and when. Furthermore, results can be downloaded or shared via a web address containing a unique collection identifier. This approach allows the rapid incremental addition of new data and can be used to underpin the international surveillance of typhoid, MDR and other public health threats.

Methods

The Pathogenwatch application

The Pathogenwatch user interface is a React (13) single-page application with styling based on Material Design Lite (14). Phylocanvas (15) is used for phylogenetic trees, Leaflet (16) is used for maps, and Sigma (17) is used for networks. The Pathogenwatch back-end, written in Node.js, consists of an API service for the user interface and four "Runner" services to perform analysis: species prediction, single-genome analyses, tree-building, and core genome multi-locus sequence typing (cgMLST) clustering. Runner services spawn Docker containers for queued tasks, streaming a FASTA file or prior analysis through standard input and storing JSON data from standard output. Data storage and task queuing/synchronisation are performed by a MongoDB cluster.

S. Typhi genome assemblies and data privacy

Genome assemblies can be uploaded by the user in FASTA format or assembled *de novo* from high-throughput short read data with the Pathogenwatch pipeline using SPAdes (18), as described in the Pathogenwatch documentation (19). Sequence data and metadata files uploaded by the user are kept private to the user account unless explicitly requested to be publicly shared. Genomes can be grouped into collections and kept private or set to be made available to collaborators through a web link. Users can also integrate private and potentially confidential metadata into the display without uploading it to the Pathogenwatch servers. This private metadata will not be shared even if the collection is set to be shared via web link (20).

Genomes from published studies with geographical localisation metadata and short read data on the European Nucleotide Archive (ENA) are available as public data and accessible to all users for browsing and for contextualisation of their own datasets. At the time of submission, 4389 public *S.* Typhi genomes from 26 studies were available (Additional File 1: Supplementary Table S1). The sequences of 2490 public genomes were generated at the Wellcome Sanger Institute with Illumina HiSeq technology and assembled as previously described (21). Briefly, FASTQ files were used to create multiple assemblies using VelvetOptimiser v2.2.5 and Velvet v1.2 (22). An assembly improvement step was applied to the assembly with the best N50, and contigs were scaffolded using

SSPACE (23) and sequence gaps filled using GapFiller (24). For the remaining 1899 public genomes the FASTQ files were downloaded from the ENA and assembled with Velvet as above, as well as with SPAdes v3.9.0 (18) and a range of *k*-mer sizes of 66-90% of the read length (in increments of 4). A total of 814 Velvet assemblies and 1068 SPAdes assemblies were included based on comparisons of the assembly stats and the Pathogenwatch core genome stats with both methods. Seventeen public genomes were excluded as neither assembly method produced a satisfactory draft genome based on the assembly stats and/or GC content. The public genomes metadata submitted to the ENA or made available as supplementary information in the corresponding publications was made available in the Metadata table (e.g., country/location, collection year/date, run and study accessions and PMID).

Characterisation and genotyping of *S.* Typhi genomes with Pathogenwatch

For both user-uploaded and public genomes, Pathogenwatch outputs a taxonomy assignment, a map of their locations, and assembly quality metrics. The taxonomy assignment is the best match to a microbial version of the RefSeq genome database release 78, as computed with Mash (25) (k=21, s=400). Details of the *speciator* tool can be found in the documentation (26).

Pathogenwatch also provides compatibility with *Salmonella* serotyping (SISTR (12)), multi-locus sequence typing (MLST (5)), core-genome MLST (cgMLST (11)) and *S.* Typhi single-nucleotide polymorphism (SNP)-based genotyping (GenoTyphi (8)). Detailed descriptions of the implementation of the typing tools can be found in the documentation (27).

The MLST and cgMLST schemes are periodically downloaded from Enterobase (28) and (29), respectively. Samples are typed as described in the documentation (https://cgps.gitbook.io/pathogenwatch/technical-descriptions/typing-methods/mlst and https://cgps.gitbook.io/pathogenwatch/technical-descriptions/typing-methods/cgmlst). Exact allele matches are reported using their allele ID. Multiple allele hits for a gene are reported if present. Inexact allele matches and novel STs are reported by hashing the matching allele sequence and the gene IDs, respectively.

Pathogenwatch implements SISTR (Salmonella In Silico Typing Resource (12)), which produces serovar predictions from WGS assemblies by determination of antigen gene and

cgMLST gene alleles using *blastn* v2.2.31+. Pathogenwatch uses the cgmlst_subspecies and serovar fields from the SISTR JSON output to specify the serovar.

GenoTyphi assigns *S.* Typhi genomes to a predefined set of clades and subclades based on a curated set of SNPs (8) that is regularly updated as novel lineages of epidemiological interest are identified (30). Pathogenwatch uses an in-house implementation designed to work with assembly output. The *blastn* v2.2.30 program is used to align the query loci and identify positions of diagnostic SNPs, which are then processed according to the rules of the GenoTyphi scheme (31). The genotype assignment and the number of diagnostic SNPs identified on the assemblies are reported.

The plasmid replicon marker sequences are detected in the user and public genome assemblies with *Inctyper*, which uses the PlasmidFinder Enterobacteriaceae database (32). Details of the *Inctyper* tool can be found in the documentation (33).

Generation of the *S.* Typhi core genome library

Pathogenwatch supports SNP-based neighbour joining trees of *S.* Typhi both for user genomes (collection trees) and public genomes (population tree and subtrees). The trees are inferred using a curated core gene library of 3284 *S.* Typhi genes (34) generated from a pan-genome analysis of 26 complete or high-quality draft genomes (Additional File 1: Supplementary Table S2) with Roary (35) and identity threshold of 95%. The core gene families were realigned using MAFFT v7.2.2.0 (36), and filtered or trimmed according to the quality of the alignments. The gene with the fewest average pairwise SNP differences to the other family members was selected as the representative for each family. We then selected 19 reference genomes (Additional File 1: Supplementary Table S2) belonging to different genotypes according to the population structure previously described (8). The gene families were then searched against each of the 19 reference genomes and filtered according to the following rules: a) only universal families with complete coverage of the representative were kept; b) all paralogues were removed; c) overlapping gene families were merged into a single, contiguous pseudo-sequence. A BLAST (37) core library was then built with the representative genes, and a profile of variant sites determined for the core genes present in each reference genome. Each of the 4389 public genomes was then clustered with its closest reference genome based on this profile of variant sites, thus

constituting each of the 19 population subtrees that Pathogenwatch employs to contextualise user-uploaded genomes.

Typhi Pathogenwatch genome clustering

The relationships between genomes are represented with trees (dendrograms) based on the genetic distance computed from substitution mutations in the core gene library, as described in detail in the documentation (38). User-provided assemblies are queried against the *S.* Typhi core gene library with *blastn* v2.2.30 (37) using an identity threshold of 90%. The core gene set of each query assembly is compared to the reference genome core that has the most variant sites in common. An overall relative substitution rate is determined, and loci that contain more variants than expected assuming a Poisson distribution are filtered out. Pairwise distances between assemblies (including user-provided and reference) are scored via a distance scoring algorithm that compares all variant positions from all pairs of core gene sets, SNPs are counted (generating a downloadable pairwise difference matrix) and normalised by the relative proportion of the core present (generating a downloadable pairwise score matrix). The pairwise score matrix is then used to infer a midpoint-rooted neighbour-joining tree using the Phangorn v2.4.0 (39) and Ape v5.1 (40) R packages. Trees are computed for the user assemblies only (collection tree), and for the user assemblies and public assemblies assigned to the same reference genome (public data subtrees), all of which are downloadable in Newick format.

We benchmarked the Pathogenwatch clustering method against other methods of SNP-based tree inference with three subsets of published genomes: Dataset I) 118 genomes spanning the population diversity of *S.* Typhi defined by GenoTyphi (Additional File 2: Supplementary Table S3); Dataset II) 138 closely related genomes, from a recent clonal expansion of the multidrug-resistant haplotype H58 within Africa (Additional File 2: Supplementary Table S4); and Dataset III) 43 strains from clade 3.2 including CT18, the first completed *S.* Typhi genome, which remains reference of choice for most population genomics studies (Additional File 2: Supplementary Table S5). For each subset a tree was generated with four different methods: 1) Pathogenwatch; 2) maximum likelihood (ML) with RAxML v8.2.8 (41) on SNPs extracted from an alignment of concatenated core genes generated using Roary (35); 3) neighbour joining (NJ) with FastTree (42) using the option –noml on the same alignment as 2); and 4) ML with RAxML v8.2.8 on SNPs extracted

from a previously published CT18-guided alignment (7). Five hundred bootstrap replicates were computed for the ML trees (methods 2 and 4). We compared the trees thus generated using the tree comparison software Treescape v1.10.18 (Kendall-Colijn distance, now available as Treespace (43)) and the Tanglegram algorithm of Dendroscope (44). The tree files used in the tree comparisons are provided in (45).

Genomes can also be clustered in Typhi Pathogenwatch based on their cgMLST profile using single linkage clustering. Distance scores are calculated between each pair of samples by identifying the genes which have been found in both samples and by counting the number of differences in the alleles. The SLINK algorithm (46) is used to quickly group genomes into clusters at a given threshold. For a given genome, users are able to see how many other genomes it is clustered with at a range of distance thresholds, view the structure of the cluster as a network graph, and view the metadata and analysis for sequences in that cluster.

Genomic predictions of antimicrobial resistance

Pathogenwatch predicts the presence of genes and single point mutations conferring AMR by querying genome assemblies using PAARSNP v2.4.9 with *blastn* v2.2.30 (37) and a curated *S.* Typhi database of genes and mutations (Additional File 1: Supplementary Table S6 (47)) known to confer resistance to ampicillin (AMP), chloramphenicol (CHL), broad-spectrum cephalosporins (CEP), ciprofloxacin (CIP), sulfamethoxazole (SMX), trimethoprim (TMP), the combination antibiotic co-trimoxazole (sulfamethoxazole-trimethoprim, SXT), tetracycline (TCY), azithromycin (AZM), colistin (CST) and meropenem (MEM). For details of the implementation see Pathogenwatch documentation (48)

PAARSNP also provides a prediction of AMR phenotype inferred from the combination of identified mechanisms. To benchmark the genotypic resistance predictions, we used a set of 1316 genomes from 16 published studies (Additional File 1: Supplementary Table S1) with drug susceptibility information available for at least one of the twelve antibiotics reported by Typhi Pathogenwatch. The drug susceptibility data reported was heterogeneous across the studies (minimum inhibitory concentration (MICs), disk diffusion diameters, and/or susceptible/intermediate/resistant (SIR)). We first compared the Typhi Pathogenwatch antibiotic resistance predictions to the drug susceptibility phenotype (SIR

interpretation provided by the studies) of 1316 genomes, grouping the Resistant and Intermediate classifications as non-susceptible. For each antibiotic, the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for detection of known resistance determinants, and their 95% confidence intervals (CI) were calculated with the epi.tests function of the epiR v1.0-14 package (49). False negative (FN) and false positive (FP) results were further investigated with alternative methods by querying the genome assemblies with Resfinder (10) and/or by mapping and local assembly of the sequence reads to the Bacterial Antimicrobial Resistance Reference Gene Database (Bioproject PRJNA313047) with ARIBA (50).

Seven studies reported ciprofloxacin MICs for a total of 889 *S.* Typhi strains, albeit interpreted with different breakpoint guidelines and versions (Additional File 2: Supplementary Table S1). We compared the Typhi Pathogenwatch ciprofloxacin resistance predictions (SIR) for each observed combination of genetic AMR determinants against the MIC values re-interpreted with the ciprofloxacin breakpoints for *Salmonella* spp. from CLSI M100 30th edition (susceptible MIC$\leq$0.06; intermediate MIC = 0.12 to 0.5; resistant MIC $\geq$1 (51)) with a script that is available at (45) .

Results

Overview of the Typhi Pathogenwatch application

We have developed a public health focused application for *S.* Typhi genomics that uses genome assemblies to perform three essential tasks for surveillance and epidemiological investigations, i.e., (i) placing isolates into lineages or clonal groups, (ii) identifying their closest relatives and linking to their geographic distribution, and (iii) detecting the presence of genes and mutations associated with AMR. These data can aid the local investigator to rapidly identify a potential source of transmission and to predict AMR phenotypes.

The Pathogenwatch application can be accessed at https://pathogen.watch/styphi, where users can create an account and upload and analyse their genomes (Figure 1 (52)). User data remains private and stored in their personal account. Pathogenwatch provides compatibility with typing information for MLST (5), cgMLST (11), *in silico* serotyping (SISTR (12)), a SNP genotyping scheme (GenoTyphi (8)), and plasmid replicon sequences (32). The results for a single genome are displayed in a genome report that can be downloaded as a PDF. The results for a collection of genomes can be viewed online and downloaded as trees and tables of genotypes, AMR predictions, assembly metrics, and genetic variation. Results can also be accessed at a later date and shared via a collection ID embedded in a unique weblink, thus facilitating collaborative surveillance.

Clustering genomes into lineages with Pathogenwatch

A fundamental process for interpreting large genomic datasets is to identify the nearest neighbours to the genome(s) under investigation. The pairwise genetic distance between isolates provides an operational unit for genomic surveillance, which we can combine with epidemiological metadata to make inferences during an investigation or for routine surveillance. Typhi Pathogenwatch clusters user genomes based on their genetic distance and displays their relationships in a collection tree.

We benchmarked the Pathogenwatch clustering method against established methods of SNP-based tree inference, i.e. maximum likelihood or neighbour-joining trees inferred from core genome SNPs or whole-genome SNPs. We used three sets of published genomes: I) 118 genomes spanning the population diversity of *S.* Typhi defined by GenoTyphi (8); II)

138 closely related genomes, from a clonal expansion of 4.3.1 within Africa (7); and III) 43 strains from clade 3.2 including CT18, the genome of choice for reference-guided population genomics studies (8). The Pathogenwatch trees clustered the diverse genomes from subset I according to genotype assignments (Additional File: Supplementary Figure S1a), and detected phylogeographic signal in the closely related genomes of subset II (Additional File: Supplementary Figure S1b), in agreement with previous studies. In addition, we found that the Typhi Pathogenwatch clustering algorithm produced trees comparable to the established methods based on visualisations of the tree space and tree topology (Additional File 3: Supplementary Figure S2).

Contextualisation with public data

Pathogenwatch contextualises the user-uploaded genomes with public genomes using a population tree of 19 diverse genome references (Additional File 3: Supplementary Figure S3) to guide the SNP-based clustering of user and public genomes into subsets of closely related genomes (population subtrees). Therefore, Pathogenwatch can display user genomes in the context of the most relevant public data, facilitating visualisation and interpretation. A previous investigation of a typhoid outbreak in Zambia exemplifies the value of this integrated contextualisation (53). This retrospective study identified clonal diversity and a two repertoires of AMR genes within outbreak organisms, which belonged to haplotype H58 (genotype 4.3.1). The study also identified an isolate from Central Africa as the nearest neighbour to the Zambian genomes among the only 5 genomes from 4.3.1 available for comparison at the time. Using Pathogenwatch, the clonal diversity of the outbreak strains can be rapidly contextualised with the 2500 H58 genomes available at the time of publication. This revealed two different clusters with close relationships to contemporary genomes from neighbouring countries Malawi and Tanzania (Figure 2a-b) that are also characterised by different *dfrA* genes (Figure 2c-d). The integration of genomic data and associated metadata from different studies in Pathogenwatch facilitates the contextualisation of a local outbreak via the web and without the need for bioinformatics expertise.

Users interested in exploring the public genomes without creating their own collections can browse the public data as a whole (54) or view by published study (55). At the time of submission, Typhi Pathogenwatch included 4389 public genomes from 26 published

articles (Additional File 1: Supplementary Table S1). The average length of the genome assemblies was 4,787,922 bp (sd = 85492, range = 4535,494 – 5,211,763), the average N50 was 196,001 bp (sd = 130667, range = 19,527 – 4,806,333), the average number of contigs was 59.3 (sd = 43.44, range = 1 – 633), the average number of non-ATCG characters was 1,841.1 (sd = 3,718, range = 1 – 48,002) and the GC content was 52.0% (sd = 0.1, range = 51.4 - 52.4). The genomes spanned the years 1905 to 2019, the majority of which were from 2000 onwards (N=3,795, 86.49%). Seventy-seven countries were reflected by the public genomes, with the largest representations from the Indian subcontinent (N=1,602, 36.50%), the United Kingdom (N=629, 14.33%) and Southeast Asia (N=570, 12.99%, Additional File 3: Supplementary Figure S4). Over 97% of the genomes were classified as either ST1 (68.2%) or ST2 (29.0%) using the 7-locus *Salmonella* MLST scheme, with the remaining 2.8% divided between 33 rare STs (Additional File 1: Supplementary Table S7). Similarly, over half of the genomes (N=2,500, 57.0%) belonged to the globally dominant MDR genotype 4.3.1, which is further discriminated into five genotypes with different temporal distributions and relative abundance (Additional File 3: Supplementary Figure S5).

Genomic predictions of antimicrobial resistance

Typhi Pathogenwatch provides resistance predictions for antimicrobials relevant to treatment of typhoid fever by querying genome assemblies with BLAST (37) and a curated library of known AMR genes and mutations (Additional File 1: Supplementary Table S6). To benchmark the Typhi Pathogenwatch predictions, we compared the genotypic resistance genotypes to the available drug susceptibility phenotypes (SIR interpretation) of 1316 genomes, grouping the Resistant and Intermediate classifications as insusceptible. The sensitivity of the Pathogenwatch genotypic predictions was at least 0.96 for all antibiotics with a computed value (Table 1); at the time of writing, there were no insusceptible isolates described for colistin or meropenem. The false negative (FN) calls for ampicillin (N=4), cephalosporins (N=2), chloramphenicol (N=6), and sulfamethoxazole-trimethoprim (N=7) corresponded to genomes for which no resistance gene was reported in the original genome studies (56-58), nor by an alternative bioinformatics method (50), in agreement with Pathogenwatch phenotype. For all of the 49 FN calls for ciprofloxacin, the Pathogenwatch genotypic predictions agree with the sequence analyses reported in the original genome studies (30, 58), in which no QRDR mutations or *qnr* genes were

detected. Only mutations outside of the quinolone-resistance determining region (QRDR) of *parE* (A364V, N=17) or *gyrA* (D538N, N=2) were detected in 20 genomes. These mutations have not as yet been shown to induce ciprofloxacin resistance and were therefore excluded from the Pathogenwatch AMR library.

The specificity of the Pathogenwatch genotypic predictions was at least 0.95 for most antimicrobials (Table 1), with the exception of ciprofloxacin, for which the specificity was 0.66 (95% confidence interval 0.58-0.73), showing that a third of the ciprofloxacin susceptible isolates were reported as insusceptible by Pathogenwatch. A closer inspection of the 57 false positive (FP) results showed that Pathogenwatch reported one (N=55), two (N=2) or three (N=1) mutations in the QRDR of *gyrA, gyrB* and/or *parC*, most frequently the single mutations *gyrA*_S83F (N=25) and *gyrB*_S464F (N=16). For 54 of these cases, the same mutations were reported in the original genome studies. For the remaining three genomes, no mutations were reported in the original studies, but we confirmed the presence of *gyrB*_S464F (N=2) or *gyrB*_S464Y (N=1) in the assemblies using Resfinder (10). Similarly, we confirmed the Pathogenwatch identification of $bla_{\text{TEM-1}}$, *catA1*, or *sul1-dfrA7* for all 47 of the FP calls for ampicillin (N=8), chloramphenicol (N=14), and sulfamethoxazole-trimethoprim (N=25), respectively, either from the original genome studies or with Resfinder.

The additive effect of QRDR mutations on ciprofloxacin susceptibility has been previously described (59). In addition, the presence of three non-synonymous mutations in the *gyrA* (S83F and D87N) and *parC* (S80I) genes was previously associated with ciprofloxacin resistance and fluoroquinolone treatment failure (59, 60) and was predictive of ciprofloxacin resistance in a study of reference laboratory isolates (61). Pathogenwatch thus reports this specific combination of mutations as resistant on the Antibiotics table with a red circle, while any other single, double or triple QRDR mutation is reported as decreased susceptibility (intermediate, yellow circle). We evaluated the ciprofloxacin MICs of 889 *S.* Typhi isolates from nine previous studies against the different combinations of resistance mechanisms identified by Pathogenwatch. Overall, the distribution of MIC values was consistent with the genomic predictions of AMR from Pathogenwatch (Figure 3). The MIC values linked to some of the mechanisms, however, straddled two or even all three SIR categories, explaining many of the differences observed between phenotype and genotypic predictions. The isolates with 1 or 2 QRDR mutations had intermediate MICs against ciprofloxacin, and support reporting as intermediate in Pathogenwatch. The

highest ciprofloxacin MIC values were observed for the combination of *gyrA*_S83F-*gyrA*_D87N-*parC*_S80I mutations, reported as resistant by Pathogenwatch. However, the triple combination *gyrA*_S83F-*gyrA*_D87G-*parC*_E84K was represented by 9 isolates with MICs in both the resistant (N=6) and the intermediate (N=3) ranges, and is reported by Pathogenwatch as intermediate. Further susceptibility testing of isolates with this combination of mutations is needed to refine genotypic predictions. Likewise, several other mechanisms potentially conferring insusceptibility to ciprofloxacin were found in the public genomes but had with no or little associated MIC data, including seven additional triple mutations (Additional File 1: Supplementary Table S8, Additional File 3: Supplementary Figure S6).

Genomic predictions of AMR are presented in three interactive and downloadable tables, Antibiotics, Genes, and SNPs, which display the predicted resistance profile, AMR genes and AMR-associated chromosomal SNPs found for each genome in the collection, respectively. The user can overlay the genotypic predictions on the tree and the map views for one or multiple antibiotics/genes/SNPs, thus intuitively linking resistance with genome clustering and geographic location. For example, the distribution of genomic predictions of ciprofloxacin resistant, MDR, or extremely drug resistant (XDR, defined as MDR + ciprofloxacin resistant) *S.* Typhi on the map and on the tree of 4389 public genomes highlight the lineages that represent a particular challenge to treatment and their geographical distribution (Additional File 3: Supplementary Figure S7). A summary of the genomic predictions of MDR and XDR *S.* Typhi highlights the differences in the distribution of high-risk clones by region, year and genotype, as inferred from the public genomes (Additional File 3: Supplementary Figure S8).

In addition, Pathogenwatch presents a granular picture of the different genetic mechanisms behind resistance to an antibiotic. For example, the distinct distribution of trimethoprim-resistance gene *dfrA15* in West Africa associated with genotype 3.1.1, and of *dfrA7* across Central and East Africa, associated with genotypes 2.5.1 and 4.3.1.1, respectively (62) (Additional File 3: Supplementary Figure S9). The most frequent AMR genes in the public collection of 4389 genomes associated with an MDR phenotype were $bla_{\text{TEM-1}}$ (ampicillin, N=1460), *catA1* (chloramphenicol, N=1406), *sul1* (sulfamethoxazole, N=1447), and *dfrA7* (trimethoprim, N=1232). Notably, $bla_{\text{CTX-M-15}}$ was the most frequent gene coding for an extended-spectrum beta-lactamase (N=92, Additional File 3: Supplementary Figure S10). The distribution of the sequence identity values of acquired

AMR genes found in the public genomes showed only minor deviations from being identical to the AMR library representatives (Additional File 3: Supplementary Figure S11).

Several plasmids have been implicated in the dissemination of drug-resistant *S.* Typhi. Notably, the MDR phenotype is linked to a composite transposon carrying multiple resistance genes, either located in IncH1 plasmids or integrated into the chromosome (7). An IncY plasmid that confers resistance to fluoroquinolones and third-generation cephalosporins was detected in XDR *S.* Typhi from an outbreak in Pakistan (56), while plasmids belonging to at least five different Inc types have been described in a recent pan-African study (62). Pathogenwatch identifies plasmid replicon marker sequences in the user genomes and reports them on the Typing table on the collection view for multiple genomes (Figure 1). A more detailed output is included in the single genome report, where any resistance genes found on the same assembled contig as the replicon marker sequence are also indicated. Pathogenwatch reported between one and four plasmid replicon marker sequences in a third of the public genomes (1,571/4,389, 35.79%, Additional File 3: Supplementary Figure S12a). Predictably, plasmid replicon markers were more frequent in genomes with predicted genotypic resistance, in particular those organisms that were resistant to multiple antimicrobials (Additional File 3: Supplementary Figure S12b). Notably, the cryptic plasmid pHCM2, which does not carry resistance genes (63), was the most common replicon detected amongst genomes in which acquired resistance genes were not detected. The distribution of replicon genes showed that the combination of IncH1A and IncH1B(R27) was prevalent in MDR genomes from Southeast Asia and East Africa belonging to clade 4.3.1, while the same combination with the addition of IncFIA(HI1) was more prevalent in West Africa, and associated with clade 3.1 (Additional File 3: Supplementary Figure 12b-d). The IncH1A and IncH1B(R27) detect fragments of the *repA2* and *repA* genes, respectively, of the IncHI1 conjugative plasmid which has historically been associated with the majority of MDR typhoid (7). IncFIA(HI1) detects fragments of the *repE* gene that is present in a subset of IncHI1 plasmids, including the plasmid sequence type PST2 variant common in *S*. Typhi 3.1 in West Africa, but is lacking from the PST6 variant that is widespread in *S*. Typhi 4.3.1 in East Africa and Asia (62).

Maximising the utility of genomic data

Pathogenwatch makes the public WGS data easily accessible and searchable, and also constitutes a growing resource to which new data can be added. While genomic predictions of AMR are based on known mechanisms, the predictions can easily be updated as new mechanisms are discovered. Azithromycin is one of the last oral treatment options for typhoid for which resistance is currently uncommon, of particular importance in endemic areas with high rates of fluoroquinolone-resistance and outbreaks of XDR *S.* Typhi. A non-synonymous point mutation in the gene encoding the efflux pump AcrB (R717Q) was recently singled out as a molecular mechanism of resistance to azithromycin in *S*. Typhi (64). Pathogenwatch detected the *acrB*_R717Q mutation in a collection of 12 Bangladeshi genomes of genotype 4.3.1.1 isolated between 2013 and 2016 in which this mutation was first described (Figure 4). Notably, Pathogenwatch also detected the *acrB*_R717Q mutation in three additional genomes, two from isolates recovered in England in 2014 (no travel history available (65)), and one from an isolate recovered in Samoa in 2007 (7). The Samoan genome 10349_1_30_Sam072830_2007 was typed as genotype 3.5.4, while the English genomes 65343 and 32480 (no travel information available) belonged to genotypes 4.3.1.1 and 4.3.2.1, respectively. Genome 65343 was closely related to the cluster of 12 genomes from Bangladesh where this mutation was first described, while genome 32480 belonged to a small cluster of five genomes from India or with travel history to India. Thus, reanalysis of public data with Pathogenwatch showed that the *acrB*_R717Q mutation has emerged in multiple genetic backgrounds, in multiple locations, and as early as 2007.

Pathogenwatch applied to rapid risk assessment

Typhoid fever is rare in countries with a good infrastructure for the provision of clean water and sanitation infrastructure, with most cases arising from travel to endemic areas (66). Ceftriaxone-resistant typhoid fever was recently reported in developed countries and associated with travel to Pakistan (67-69). These ceftriaxone resistant isolates were associated to the recent outbreak of XDR *S.* Typhi in the Sindh province of Pakistan by the epidemiological data, the antibiograms, and information derived from WGS of the clinical isolate, such as presence of resistance genes, and mobile genetic elements. In some cases the genomes were contextualised with retrospective genomes by building a phylogenetic tree using an existing bioinformatic pipeline (67, 68) .

We exemplify how Pathogenwatch facilitates this analysis with the genome from the isolate recovered in Canada (PHL5950, accession RHPM00000000 (69)). Pathogenwatch provides a printable genome report (Additional File 3: Supplementary Figure S13) including genotyping and *in silico* serotyping information, predicted resistance profile, and the presence of resistance genes and mutations. In addition, Pathogenwatch places the genome within the Pakistani XDR outbreak (Figure 5) and shows the close genetic relatedness (between 3 and 8 pairwise differences) of the isolates via the downloadable score matrix.

Pathogenwatch as a tool for international collaboration in typhoid surveillance

As WGS capacity becomes a reality in typhoid endemic countries, there is a growing opportunity for local genomic surveillance and for collaboration across borders. This is underscored by the growing number of genomes from the Indian Subcontinent (Additional File 3: Supplementary Figure S3), where epidemic clone 4.3.1 (H58) and the nested clade of fluoroquinolone-resistant triple mutants belonging to genotype 4.3.1.2 (H58 lineage II) have been shown to have originated (7, 60)). The triple mutants were first reported in Nepal (isolated in 2013-2014) and linked to isolates from India from 2008-2012 (60). More recent surveillance studies in India showed that this lineage was still prevalent in *S.* Typhi isolates collected in Nepal in 2016 and in India in 2016-2017 (30, 70). The public data integrated in Pathogenwatch showed that (at the time of writing) this lineage is represented by 195 public genomes from seven countries (India, Bangladesh, Nepal, Pakistan, Myanmar, Japan, and United Kingdom, Figure 6a, (7, 58, 59, 65, 70-73)) and from as early as 2006 (Japan, with travel history to India, Figure 6b (71)). Linking the tree and the map highlights distinct clusters of genomes that show evidence of transmission across borders, for example between India-Pakistan and India-Nepal (Figure 8c-d). In addition, three isolates recovered in 2016 in India were reported to be resistant to cephalosporins, linked to the presence of the bla_{SHV-12} gene (74); Pathogenwatch detected bla_{SHV-12}, *qnrB* and the IncX3 plasmid replicon in these genomes. Another previous study reported an IncN replicon in three genomes from the United Kingdom (two with travel history to India) that also carried resistance genes *dfrA15* (trimethoprim), *sul1* (sulfamethoxazole), and *tetA*(A) (tetracycline) (59). Pathogenwatch identified the same AMR genes and plasmid replicon in these genomes, and also in two closely related genomes from Japan with travel history to Nepal and India (Figure 6b). Altogether, these

observations suggest that this lineage circulating in South Asia and linked to treatment failure with fluoroquinolones, can acquire plasmids with additional AMR genes, with the concomitant risk of the clonal expansion of a lineage that poses additional challenges to treatment.

Discussion

Our understanding of the *S.* Typhi population structure, including MDR typhoid has improved dramatically since the introduction of WGS, which provides a level of discrimination much needed for a human-adapted pathogen that exhibits very limited genetic variability. Progress towards the widespread implementation of WGS for epidemiological investigations and integrated routine surveillance within public health settings needs to be accompanied by i) active surveillance programs in endemic regions; ii) implementation of WGS at laboratories in endemic regions; iii) analysis of WGS data with fast, robust and scalable tools that deliver information for public health action; iv) dissemination of WGS data through networks of collaborating reference laboratories at the national, international and global scales; and v) provision of WGS data and associated metadata through continuously growing databases that are amenable to interaction and interpretation (75). Here, we described Typhi Pathogenwatch, a web application for the genomic surveillance and epidemiology of *S.* Typhi, which enhances the utility of public WGS data and associated metadata by integration into an interactive resource that users can browse, or query with their own WGS data.

Rapid, timely access to information on local patterns of AMR may inform treatment regimens, which could ultimately lead to a reduction in morbidity and mortality associated with enteric fever as this is much greater in the absence of effective antimicrobial therapy (76) . Typhi Pathogenwatch provides a general framework for genomic predictions of AMR and of related strain clusters, and is accessible to users of all bioinformatics skills levels. This enables users with an understanding of genomics but no bioinformatics training to conduct surveillance and epidemiological investigations using WGS. Furthermore, it allows experienced bioinformaticians to rapidly perform the essential tasks listed in the results section, thus freeing up time for more advanced downstream analyses.

We demonstrated that genomic predictions of AMR are largely concordant with the resistance phenotype (overall concordance 96.34%, Table 1), but with the added value of immediate contextualisation with location, time and population structure in an interactive visualisation with which to easily explore these aspects. A previous study of 332 isolates analysed in a single reference laboratory reported only 0.03% discordant results (61) versus 3.66% from our data. Our results, however, amalgamated published susceptibility

data from thirteen different studies conducted in eight different countries. The availability of consistent laboratory antimicrobial susceptibility testing data is key for the periodic benchmarking and refinement of genomic predictions of AMR (77), in particular for ciprofloxacin due to the diverse combinations of mechanisms (Additional File 1: Supplementary Table S8). It should also be noted that Pathogenwatch was developed with a focus on surveillance, not for clinical decision making.

Novel mechanisms of AMR can easily be added to the curated Pathogenwatch AMR library, and the growing collection of public genomes can be retrospectively screened, potentially revealing the presence of a newly identified gene or mutation in genomes from isolates previously collected (Figure 4). This illustrates how the provision of public genomic data through Pathogenwatch maximises reusability from which new insights into novel AMR mechanisms can be derived.-The utility of maintaining a regularly updated archive of WGS data that can be rapidly 'mined' for the presence of newly discovered AMR gene was elegantly illustrated before by the retrospective discovery of the colistin resistance gene *mcr-1* in *S. enterica* and *Escherichia coli* genomes from Public Health England (78). Pathogenwatch extends this utility to the Typhi entire community, thus democratizing the reusability of the genomic data.

Contextualizing new genomes with existing data is now a routine part of genomic epidemiology, as it can complement epidemiological investigations to, among many applications, place the new genomes in or out of an outbreak, link to past outbreaks, and determine if the success of a resistant phenotype is the result of a single clonal expansion or multiple independent introductions (79). Using the publicly available genomes, we provided examples of the utility of Pathogenwatch to contextualise user-uploaded genomes for outbreak investigation in endemic areas (Figure 2) or for the management of patients in non-endemic countries with travel history to endemic areas (Figure 5). Analysing new genomes in the context of global genomes involves the retrieval, storage and bioinformatic analysis of large amounts of sequence data and linked metadata, which is time-consuming at the least, and largely unfeasible for hospitals or public-health agencies with limited computing infrastructure. Pathogenwatch bridges this gap and provides contextualisation with the closest-related genomes guided by the *S.* Typhi population tree (Additional File 3: Supplementary Figure S3**)** and subtrees.

The interpretation of the genomic context relies heavily on the completeness of the public collection used for contextualisation, and this in turn depends on the establishment of

local, national and international surveillance programs for the real-time management of emerging lineages that pose a direct threat to human health. The International Typhoid Consortium collected and sequenced around 40% of the global genomes available in Pathogenwatch for comparison (7, 8), but ongoing surveillance and WGS are needed to maintain a relevant, contemporary genome collection. Additionally, this requires retrieval and curation of the genome data and associated metadata, as Pathogenwatch does not currently support automated submissions.

Pathogenwatch can facilitate collaborative surveillance in endemic areas via data integration and shared collections for the early detection and containment of high-risk clones (Figure 6). Collections can be set to off-line mode to work while disconnected from the internet, which may be advantageous in settings with unreliable internet connections. Despite recent efforts to promote data openness in times of pandemics (80, 81), several challenges to sharing genomic data and linked metadata remain in both the academic and public-health settings (75). User-uploaded genomes and metadata remain in the Pathogenwatch user account, and collections also remain private unless the user specifically shares them via a collection URL. Moreover, Pathogenwatch offers a private metadata option to work with confidential information.

Recent improvements in our understanding of the disease burden and the dissemination of AMR, and the development of new typhoid conjugate vaccines have bolstered efforts to employ routine vaccination for the containment of typhoid fever (82). Routine surveillance coupled with WGS can inform decisions on suitable settings for the introduction of vaccination programs and on the evolution of pathogens in response to them (83, 84). Pathogenwatch could be linked to routine genomic surveillance around typhoid vaccination initiatives to monitor the population dynamics in response to the deployment of new vaccines.

While other tools have been developed for the analysis of WGS data of the *Salmonella enterica* species, such as the comprehensive database Enterobase (11) and the *in silico* typing resource SISTR (12), the *S.* Typhi analysis framework of Pathogenwatch has been developed with a focus on the epidemiology of this human-adapted serovar and AMR. The modular architecture allows new functionalities to be added to cater to the community needs.

Conclusions

Typhi Pathogenwatch combines accurate genomic predictions of AMR with genomic and geographic context within an easy to use interface for delivered for the community and to support ongoing typhoid surveillance programs.

List of abbreviations

AMR: antimicrobial resistance

cgMLST: core-genome multi-locus sequence typing

PFGE: pulse-field gel electrophoresis

MDR: multi-drug resistant

MLST: multi-locus sequence typing

MLVA: multiple-locus variable-number tandem-repeat analysis

QRDR: quinolone resistance determining region

VNTR: multiple-locus variable-number tandem-repeat

XDR: extremely-drug resistant

WGS: whole-genome sequencing

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The genome data and linked metadata presented are available from: https://pathogen.watch/collection/07lsscrbhu2x-public-genomes, https://pathogen.watch/collection/g5pbucot6e58-hendriksen-et-al-2015, and https://pathogen.watch/collection/11lsok8nrzts-wong-et-al-2018-idcases-15e00492

The tree comparison nexus files are available from https://gitlab.com/cgps/pathogenwatch/publications/styphi/benchmark_tree

The AMR benchmarking input files and script are available from https://gitlab.com/cgps/pathogenwatch/publications/styphi/benchmark_AMR

Competing interests

The authors declare no competing interests.

Funding

Pathogenwatch is developed with support from Li Ka Shing Foundation (Big Data Institute, University of Oxford) and Wellcome (grant number 099202). SA and DMA are supported by the National Institute for Health Research (UK) Global Health Research Unit on genomic Surveillance of AMR (16_136_111) and by the Centre for Genomic Pathogen Surveillance (http://pathogensurveillance.net).

ZAD received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement TyphiNET No 845681.

Authors' contributions

DMA conceived the Pathogenwatch application. CY, RJG, KA, BT, AU and DMA developed the Pathogenwatch application. SA drafted the manuscript. SA, DMA, KEH, SB, and GD contributed to the conception and design of the work, data interpretation, and substantially revised the manuscript. SA, CY, VKW, ZAD, SN, AJP, JAK, SEP and FM contributed to the acquisition and interpretation of data. SA, CY and LSB analysed the data. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Flora Stevens and Joanne Freedman from the Travel Health and IHR department at Public Health England for providing some of the travel information linked to isolates from the United Kingdom, and to Dr. Koji Yahara, Dr. Makoto Ohnishi and Dr. Masatomo Morita for providing the travel information linked to isolates from Japan.

References

1. Bryant J, Chewapreecha C, Bentley SD. Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. Future Microbiol. 2012;7(11):1283-96.
2. Crump JA, Mintz ED. Global Trends in Typhoid and Paratyphoid Fever. Clinical Infectious Diseases. 2010;50(2):241-6.
3. Wain J, Hendriksen RS, Mikoleit ML, Keddy KH, Ochiai RL. Typhoid fever. The Lancet. 2015;385(9973):1136-45.
4. Le TA, Fabre L, Roumagnac P, Grimont PA, Scavizzi MR, Weill FX. Clonal expansion and microevolution of quinolone-resistant Salmonella enterica serotype typhi in Vietnam from 1996 to 2004. J Clin Microbiol. 2007;45(11):3485-92.
5. Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, et al. Multilocus sequence typing as a replacement for serotyping in Salmonella enterica. PLoS Pathog. 2012;8(6):e1002776.
6. Das S, Samajpati S, Roy I, Sankar S, Gaind R, Deb M, et al. Molecular Subtyping of Salmonella enterica Serovar Typhi by Pulsed-Field Gel Electrophoresis and Multiple-Locus Variable-Number Tandem-Repeat Analysis in India: Their Association with Antimicrobial Resistance Profiles. Jpn J Infect Dis. 2017;70(5):536-43.
7. Wong VK, Baker S, Pickard DJ, Parkhill J, Page AJ, Feasey NA, et al. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of Salmonella Typhi identifies inter- and intracontinental transmission events. Nat Genet. 2015;47(6):632-9.
8. Wong VK, Baker S, Connor TR, Pickard D, Page AJ, Dave J, et al. An extended genotyping framework for Salmonella enterica serovar Typhi, the cause of human typhoid. Nat Commun. 2016;7:12827.
9. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, et al. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nat Genet. 2008;40(8):987-93.
10. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemoth. 2012;67(11):2640-4.
11. Alikhan NF, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of Salmonella. PLoS Genet. 2018;14(4):e1007261.
12. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VP, Nash JH, 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. 2016;11(1):e0147101.
13. React. https://reactjs.org/. Accessed 30 Jun 2020.
14. Material Design Lite. https://getmdl.io. Accessed 30 Jun 2020.
15. Phylocanvas. http://phylocanvas.org. Accessed 30 Jun 2020.
16. Leaflet. https://leafletjs.com/. Accessed 30 Jun 2020.
17. Sigma. http://sigmajs.org/. Accessed 30 Jun 2020.
18. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77.
19. Pathogenwatch short-read assembly. https://cgps.gitbook.io/pathogenwatch/technical-descriptions/short-read-assembly. Accessed 30 Jun 2020.

20. Pathogenwatch private metadata. https://cgps.gitbook.io/pathogenwatch/how-to-use-pathogenwatch/private-metadata. Accessed 30 Jun 2020.
21. Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J, Harris SR, et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. Microb Genom. 2016;2(8):e000083.
22. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821-9.
23. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27(4):578-9.
24. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol. 2012;13(6):R56.
25. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1):132.
26. Pathogenwatch speciator. https://cgps.gitbook.io/pathogenwatch/technical-descriptions/species-assignment/speciator. Accessed 30 Jun 2020.
27. Pathogenwatch typhing methods. https://cgps.gitbook.io/pathogenwatch/technical-descriptions/typing-methods. Accessed 30 Jun 2020.
28. Enterobase allele ST search. http://enterobase.warwick.ac.uk/species/senterica/allele_st_search. Accessed 30 Jun 2020.
29. Enterobase cgMLST database. http://enterobase.warwick.ac.uk/species/index/senterica. Accessed 30 Jun 2020.
30. Britto CD, Dyson ZA, Duchene S, Carter MJ, Gurung M, Kelly DF, et al. Laboratory and molecular surveillance of paediatric typhoidal Salmonella in Nepal: Antimicrobial resistance and implications for vaccine policy. PLoS Negl Trop Dis. 2018;12(4):e0006408.
31. GenoTyphi. https://github.com/katholt/genotyphi. Accessed 30 Jun 2020.
32. Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O, Villa L, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58(7):3895-903.
33. Pathogenwatch IncTyper. https://cgps.gitbook.io/pathogenwatch/technical-descriptions/inctyper. Accessed 30 Jun 2020.
34. Pathogenwatch S. Typhi core gene library. https://gitlab.com/cgps/cgps-core-fp/-/tree/master/schemes/90370. Accessed 30 Jun 2020.
35. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3.
36. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80.
37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.
38. Pathogenwatch core genome tree. https://cgps.gitbook.io/pathogenwatch/technical-descriptions/core-genome-tree. Accessed 30 Jun 2020.
39. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592-3.
40. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35(3):526-8.
41. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22(21):2688-90.
42. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490.

43. Jombart T, Kendall M, Almagro-Garcia J, Colijn C. treespace: Statistical exploration of landscapes of phylogenetic trees. Mol Ecol Resour. 2017.
44. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012;61(6):1061-7.
45. Pathogenwatch publication repository. https://gitlab.com/cgps/pathogenwatch/publications. Accessed 30 Jun 2020.
46. Sibson R. SLINK: An optimally efficient algorithm for the single-link cluster method. The Computer Journal. 1973;16(1):30-4.
47. Pathogenwatch S. Typhi AMR library. https://gitlab.com/cgps/pathogenwatch/amr-libraries/-/blob/master/90370.toml. Accessed 30 Jun 2020.
48. Pathogenwatch PAARSNP. https://cgps.gitbook.io/pathogenwatch/technical-descriptions/antimicrobial-resistance-prediction/paarsnp. Accessed 30 Jun 2020.
49. epiR package. https://www.rdocumentation.org/packages/epiR. Accessed 30 Jun 2020.
50. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microbial Genomics. 2017;3.
51. CLSI. Performance Standards for Antimicrobial Susceptibility Testing. Wayne, PA; 2020.
52. Pathogewatch S. Typhi video. https://youtu.be/faVZWU7xJII. Accessed 30 Jun 2020.
53. Hendriksen RS, Leekitcharoenphon P, Lukjancenko O, Lukwesa-Musyani C, Tambatamba B, Mwaba J, et al. Genomic signature of multidrug-resistant Salmonella enterica serovar typhi isolates related to a massive outbreak in Zambia between 2010 and 2012. J Clin Microbiol. 2015;53(1):262-72.
54. Pathogenwatch S. Typhi 4389 public genomes collection. https://pathogen.watch/collection/07lsscrbhu2x-public-genomes. Accessed 30 Jun 2020.
55. Pathogenwatch S. Tyhi public collections. https://pathogen.watch/collections/all?access=public&organismId=90370. Accessed 30 Jun 2020.
56. Klemm EJ, Shakoor S, Page AJ, Qamar FN, Judge K, Saeed DK, et al. Emergence of an Extensively Drug-Resistant Salmonella enterica Serovar Typhi Clone Harboring a Promiscuous Plasmid Encoding Resistance to Fluoroquinolones and Third-Generation Cephalosporins. mBio. 2018;9(1).
57. Pragasam AK, Pickard D, Wong V, Dougan G, Kang G, Thompson A, et al. Phylogenetic Analysis Indicates a Longer Term Presence of the Globally Distributed H58 Haplotype of Salmonella Typhi in Southern India. Clin Infect Dis. 2020.
58. Tanmoy AM, Westeel E, De Bruyne K, Goris J, Rajoharison A, Sajib MSI, et al. Salmonella enterica Serovar Typhi in Bangladesh: Exploration of Genomic Diversity and Antimicrobial Resistance. mBio. 2018;9(6).
59. Ingle DJ, Nair S, Hartman H, Ashton PM, Dyson ZA, Day M, et al. Informal genomic surveillance of regional distribution of Salmonella Typhi genotypes and antimicrobial resistance via returning travellers. PLoS Negl Trop Dis. 2019;13(9):e0007620.
60. Pham Thanh D, Karkey A, Dongol S, Ho Thi N, Thompson CN, Rabaa MA, et al. A novel ciprofloxacin-resistant subclade of H58 Salmonella Typhi is associated with fluoroquinolone treatment failure. Elife. 2016;5:e14003.
61. Day MR, Doumith M, Do Nascimento V, Nair S, Ashton PM, Jenkins C, et al. Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of Salmonella enterica serovars Typhi and Paratyphi. J Antimicrob Chemother. 2017.

62. Park SE, Pham DT, Boinett C, Wong VK, Pak GD, Panzner U, et al. The phylogeography and incidence of multi-drug resistant typhoid fever in sub-Saharan Africa. Nat Commun. 2018;9(1):5094.
63. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, et al. Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature. 2001;413(6858):848-52.
64. Hooda Y, Sajib MSI, Rahman H, Luby SP, Bondy-Denomy J, Santosham M, et al. Molecular mechanism of azithromycin resistance among typhoidal Salmonella strains in Bangladesh identified through passive pediatric surveillance. PLoS Negl Trop Dis. 2019;13(11):e0007868.
65. Ashton PM, Nair S, Peters TM, Bale JA, Powell DG, Painset A, et al. Identification of Salmonella for public health surveillance using whole genome sequencing. PeerJ. 2016;4:e1752.
66. Bhan MK, Bahl R, Bhatnagar S. Typhoid and paratyphoid fever. Lancet. 2005;366(9487):749-62.
67. Engsbro AL, Riis Jespersen HS, Goldschmidt MI, Mollerup S, Worning P, Pedersen MS, et al. Ceftriaxone-resistant Salmonella enterica serotype Typhi in a pregnant traveller returning from Karachi, Pakistan to Denmark, 2019. Euro Surveill. 2019;24(21).
68. Godbole GS, Day MR, Murthy S, Chattaway MA, Nair S. First Report of CTX-M-15 Salmonella Typhi From England. Clin Infect Dis. 2018;66(12):1976-7.
69. Wong W, Rawahi HA, Patel S, Yau Y, Eshaghi A, Zittermann S, et al. The first Canadian pediatric case of extensively drug-resistant Salmonella Typhi originating from an outbreak in Pakistan and its implication for empiric antimicrobial choices. IDCases. 2019;15:e00492.
70. Britto CD, Dyson ZA, Mathias S, Bosco A, Dougan G, Jose S, et al. Persistent circulation of a fluoroquinolone-resistant Salmonella enterica Typhi clone in the Indian subcontinent. J Antimicrob Chemother. 2020;75(2):337-41.
71. Matono T, Morita M, Yahara K, Lee KI, Izumiya H, Kaku M, et al. Emergence of Resistance Mutations in Salmonella enterica Serovar Typhi Against Fluoroquinolones. Open Forum Infect Dis. 2017;4(4):ofx230.
72. Oo KM, Myat TO, Htike WW, Biswas A, Hannaway RF, Murdoch DR, et al. Molecular mechanisms of antimicrobial resistance and phylogenetic relationships of Salmonella enterica isolates from febrile patients in Yangon, Myanmar. Trans R Soc Trop Med Hyg. 2019;113(10):641-8.
73. Pham Thanh D, Thompson CN, Rabaa MA, Sona S, Sopheary S, Kumar V, et al. The Molecular and Spatial Epidemiology of Typhoid Fever in Rural Cambodia. PLOS Neglected Tropical Diseases. 2016;10(6):e0004785.
74. Rodrigues C, Kapil A, Sharma A, Devanga Ragupathi NK, Inbanathan FY, Veeraraghavan B, et al. Whole-Genome Shotgun Sequencing of Cephalosporin-Resistant Salmonella enterica Serovar Typhi. Genome Announc. 2017;5(10).
75. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018;19(1):9-20.
76. Crump JA, Sjolund-Karlsson M, Gordon MA, Parry CM. Epidemiology, Clinical Presentation, Laboratory Diagnosis, Antimicrobial Resistance, and Antimicrobial Management of Invasive Salmonella Infections. Clin Microbiol Rev. 2015;28(4):901-37.
77. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clin Microbiol Infect. 2017;23(1):2-22.
78. Doumith M, Godbole G, Ashton P, Larkin L, Dallman T, Day M, et al. Detection of the plasmid-mediated mcr-1 gene conferring colistin resistance in human and food isolates

of Salmonella enterica and Escherichia coli in England and Wales. J Antimicrob Chemother. 2016;71(8):2300-5.
79. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012;13(9):601-12.
80. (COG-UK) TC-GU. An integrated national scale SARS-CoV-2 genomic surveillance network. The Lancet Microbe.
81. Perkel J. Democratic databases: science on GitHub. Nature. 2016;538(7623):127-8.
82. Bentsi-Enchill AD, Hombach J. Revised Global Typhoid Vaccination Policy. Clin Infect Dis. 2019;68(Suppl 1):S31-S3.
83. Colijn C, Corander J, Croucher NJ. Designing ecologically optimized pneumococcal vaccines using population genomics. Nat Microbiol. 2020;5(3):473-85.
84. Lo SW, Gladstone RA, van Tonder AJ, Lees JA, du Plessis M, Benisty R, et al. Pneumococcal lineages associated with serotype replacement and antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an international whole-genome sequencing study. Lancet Infect Dis. 2019;19(7):759-69.

Figures

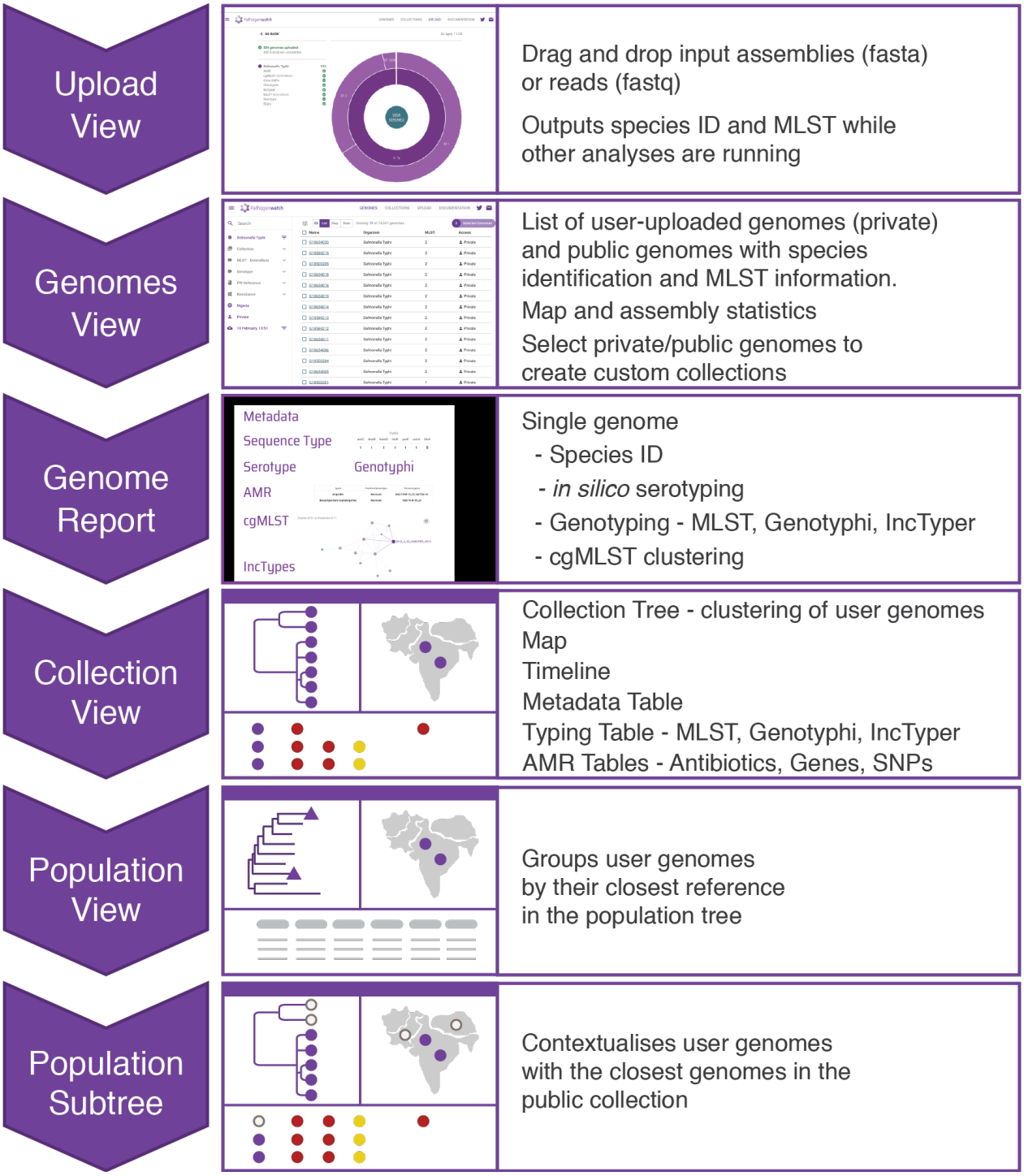

Figure 1. Workflow of the Typhi Pathogenwatch application. Input assemblies or sequence reads and metadata files can be uploaded via drag-and-drop onto the Upload page. Once the analyses completed, the genomes are listed on the Genomes page with Pathogenwatch outputs for speciation and MLST. Clicking on a genome name on the list pops up a Genome Report. The user can create collections of genomes. The Collection view displays the user genomes clustered by genetic similarity on a tree, their location on a map, a timeline, as well as tables for metadata, typing and AMR. The Population view displays the user genomes by their closest reference genome in the population tree. Clicking on one of the highlighted nodes (purple triangles) opens the Population subtree view, which contextualises the user genomes with the closest public genomes.

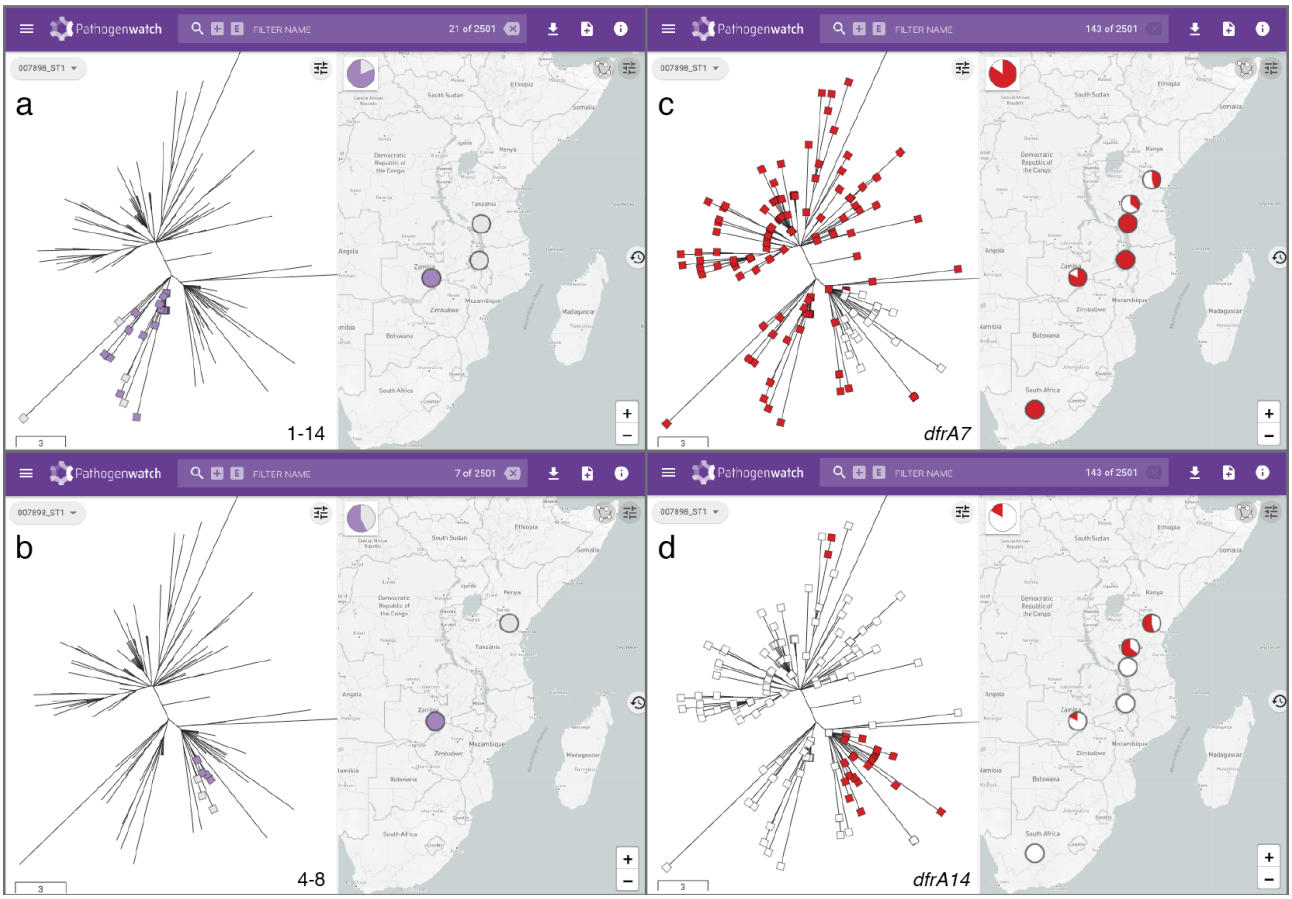

Figure 2. Pathogenwatch provides genomic context for outbreak investigations. a-b Genomes from an outbreak in Zambia (purple markers on tree and map) are linked by genetic relatedness to genomes from neighbouring countries Malawi and Tanzania (grey markers) forming 2 separate groups of 16 (**a**) and 4 (**b**) outbreak genomes, respectively. The number of pairwise differences (range) between outbreak and related genomes as downloaded from the Pathogenwatch score matrix are indicated on the bottom-right of the tree panel. **c-d** Differential distribution of trimethoprim resistance genes *dfrA7* (**c**) and *dfrA14* (**d**) across the two clades containing outbreak genomes. The data is available at https://pathogen.watch/collection/g5pbucot6e58-hendriksen-et-al-2015.

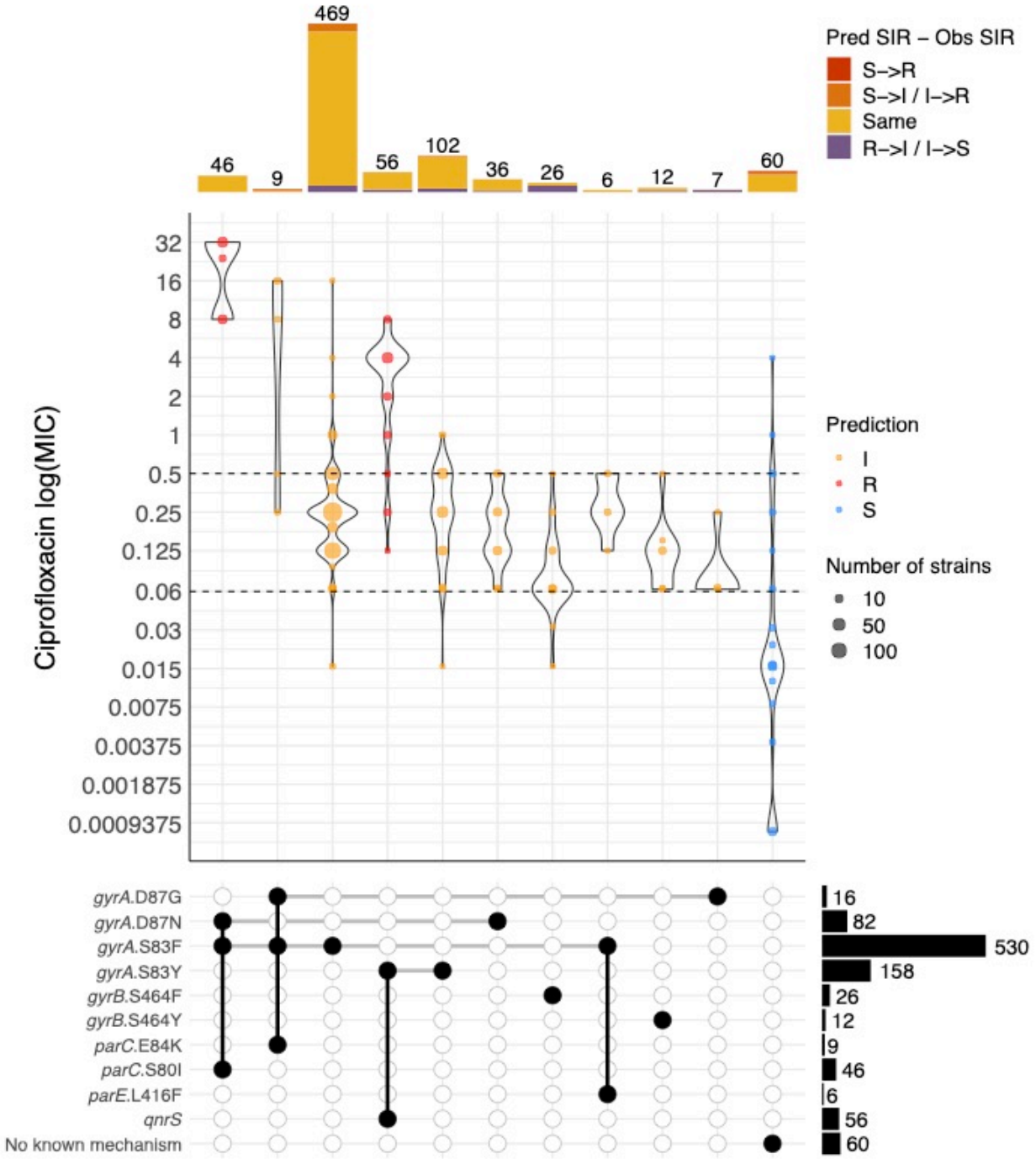

Figure 3. Distribution of minimum inhibitory concentration (MIC) values (mg/L) for ciprofloxacin in a collection of *S.* Typhi isolates with different combinations of genetic mechanisms that are known to confer resistance to this antibiotic. Only combinations observed in at least 5 genomes are shown. Dashed horizontal lines on the violin plots mark the CLSI clinical breakpoint for ciprofloxacin. Point colours inside violins represent the genotypic AMR prediction by Pathogenwatch on each combination of mechanisms. Barplots on the top show the abundance of genomes with each combination of mechanisms. Bar colours represent the differences between the predicted and the observed SIR (i.e. red for a predicted susceptible mechanism when the observed phenotype is resistant).

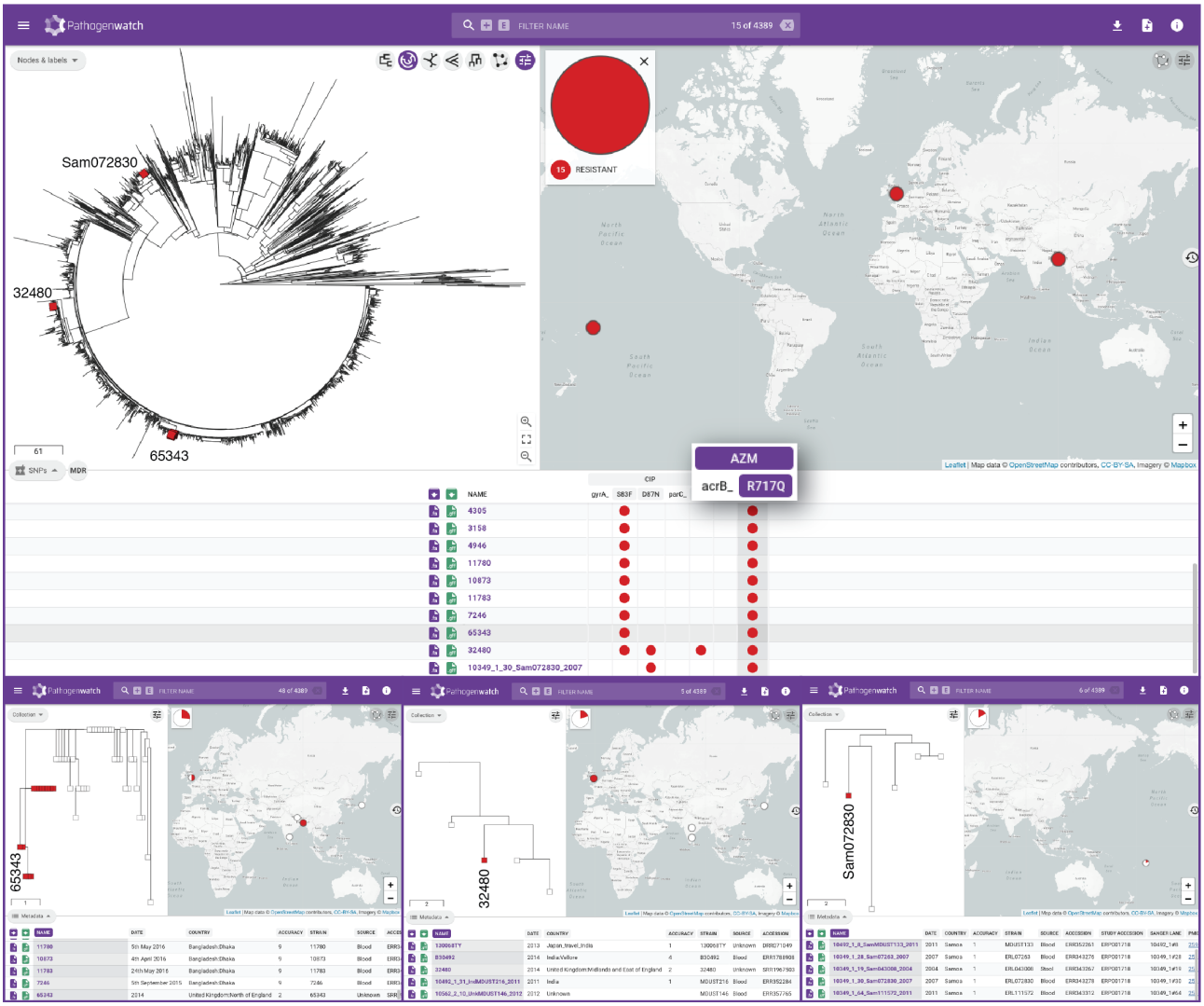

Figure 4. Pathogenwatch data reusability. Fifteen genomes carrying the *acrB*_R717Q mutation recently linked to azithromycin resistance in *S.* Typhi are shown in red on the tree of 4389 public genomes and on the map. The presence of the mutation is indicated by the red circles on the SNPs table. Three of these genomes (tree labels) belong to isolates collected before the mutation was first described and are shown in more detail in the bottom panels. The data is available at https://pathogen.watch/collection/07lsscrbhu2x-public-genomes

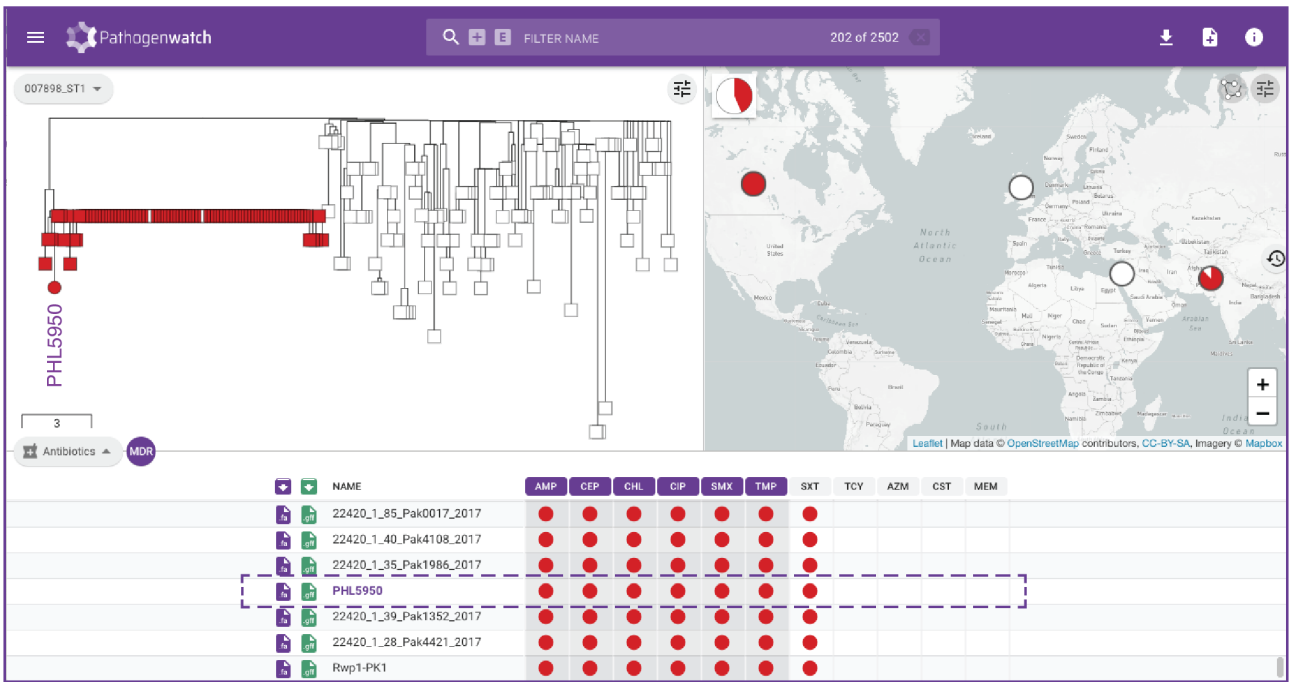

Figure 5. Rapid risk assessment of typhoid fever cases in non-endemic regions. Pathogenwatch places genome PHL5950 from an isolate recovered in Canada and with travel history to Pakistan within the XDR-outbreak in Pakistan (red markers). The data is available at https://pathogen.watch/collection/11lsok8nrzts-wong-et-al-2018-idcases-15e00492

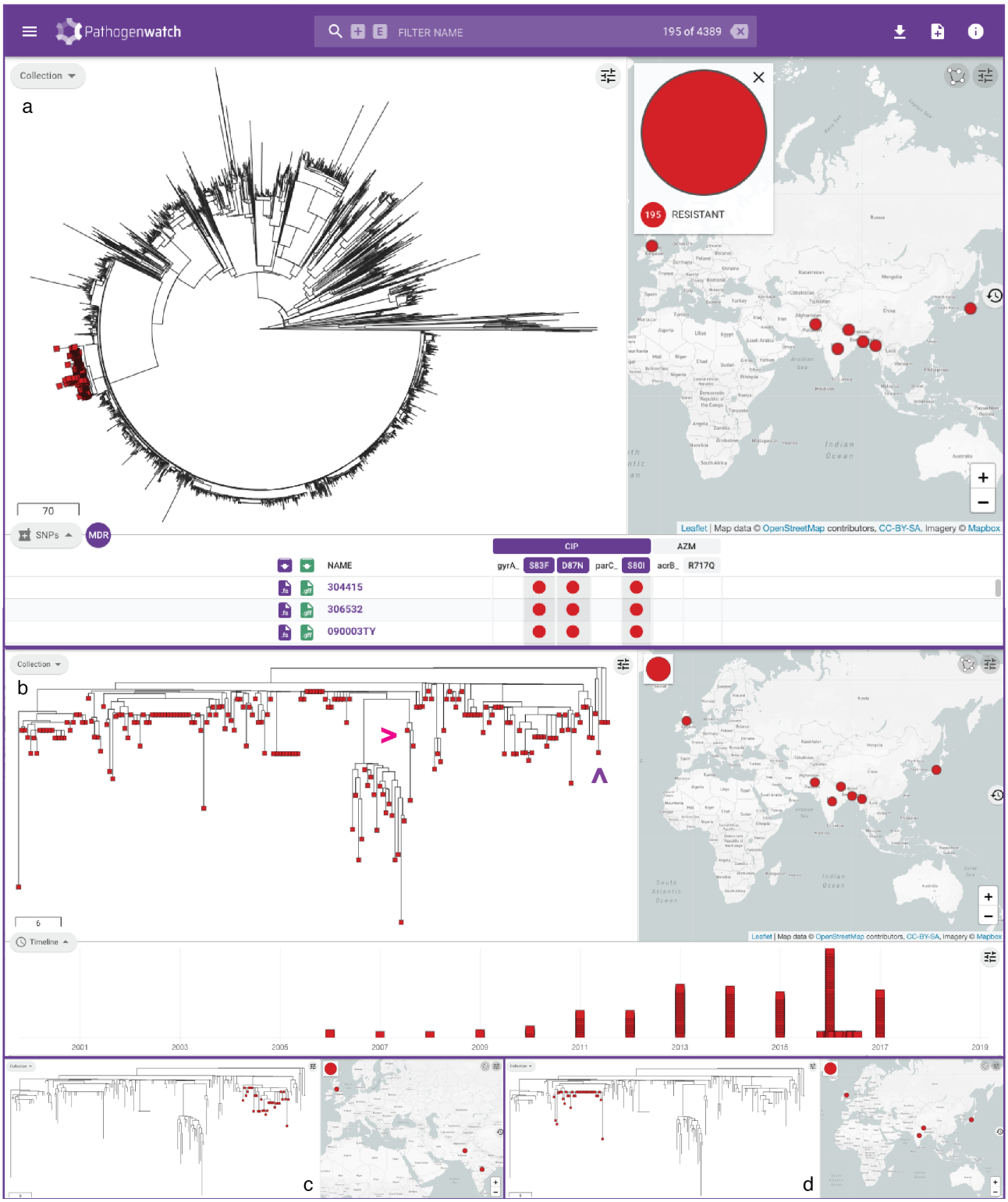

Figure 6. Pathogenwatch to for collaborative international surveillance of *S.* Typhi. a Pathogenwatch highlights 195 ciprofloxacin-resistant triple mutants on the public data tree and map by simultaneously selecting the mutations *gyrA*_S83F, *gyrA*_D87N, and *parC*_S80I on the SNPs table**. b** Detailed visualisation of the triple mutants showing the temporal distribution of the genomes on the timeline. Magenta arrowhead: 3 genomes from India with bla_{SHV-12}, *qnrB* and an IncX3 replicon. Purple arrowhead: 4 genomes with *sul1*, *dfrA15*, *tetA*(A) and an IncN replicon from the UK and Japan. Selecting individual clades on the tree shows distinct clades that span neighbouring countries India-Pakistan (**c**) and India-Nepal (**d**). The data is available at https://pathogen.watch/collection/07lsscrbhu2x-public-genomes

Tables

Antibiotic	Total	TN	TP	FN	FP	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	VME rate	ME rate	Concordance (%)
AMP	875	461	402	4	8	0.99 (0.97-1.00)	0.98 (0.97-0.99)	0.98 (0.96-0.99)	0.99 (0.98-1)	0.01	0.02	98.63
CEP	348	256	90	2	0	0.98 (0.92-1.00)	1.00 (0.99-1.00)	1.00 (0.96-1.00)	0.99 (0.97-1.00)	0.02	0	99.43
CHL	913	518	375	6	14	0.98 (0.97-0.99)	0.97 (0.96-0.99)	0.96 (0.94-0.98)	0.99 (0.98-1.00)	0.02	0.03	97.81
CIP	1282	111	1065	49	57	0.96 (0.94-0.97)	0.66 (0.58-0.73)	0.95 (0.93-0.96)	0.69 (0.62-0.76)	0.04	0.32	91.73
SXT	912	513	367	7	25	0.98 (0.96-0.99)	0.95 (0.93-0.97)	0.94 (0.91-0.96)	0.99 (0.97- 0.99)	0.02	0.05	96.49
TCY	44	40	4	0	0	1.00 (0.40-1.00)	1.00 (0.91-1.00)	1.00 (0.40-1.00)	1.00 (0.91-1.00)	0	0	100
AZM	156	144	12	0	0	1.00 (0.74-1.00)	1.00 (0.97-1.00)	1.00 (0.74-1.00)	1.00 (0.97-1.00)	0	0	100
CST	41	41	0	0	0	-	1.00 (0.91-1.00)	-	1.00 (0.91-1.00)	-	0	100
MEM	132	132	0	0	0	-	1.00 (0.97- 1.00)	-	1.00 (0.97- 1.00)	-	0	100

Table 1. Benchmark analysis of Typhi Pathogenwatch AMR predictions for ampicillin (AMP), chloramphenicol (CHL), broad-spectrum cephalosporins (CEP), ciprofloxacin (CIP), sulfamethoxazole-trimethoprim (SXT), tetracycline (TCY), azithromycin (AZM), colistin (CST) and meropenem (MEM). The total number of comparisons, true negatives (TN), true positives (TN), false negatives (FN), false positives (FN), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), very major error (VME) rate, major error (ME) rate, and concordance are shown. Confidence intervals (95%) are shown in parenthesis.