Paratype: A genotyping framework and an open-source tool for *Salmonella* Paratyphi A

Arif M. Tanmoy[1,2*], Yogesh Hooda[1,3*], Mohammad S. I. Sajib[1,4], Kesia E. da Silva[5], Junaid Iqbal[6], Farah N. Qamar[6], Stephen P. Luby[5], Gordon Dougan[7], Zoe A. Dyson[7,8,9,12], Stephen Baker[8], Denise O. Garrett[10], Jason R. Andrews[5], Samir K. Saha[1,11,$], Senjuti Saha[1,$,#]

[1] Child Health Research Foundation, Dhaka, Bangladesh. arif.tanmoy@chrfbd.org, yhooda@chrfbd.org, saiful.sajib@chrfbd.org, samir@chrfbd.org, senjutisaha@chrfbd.org

[2] Department of Medical Microbiology and Infectious Diseases, Erasmus University Medical Center, Rotterdam, the Netherlands.

[3] MRC-Laboratory Molecular Biology, Cambridge, UK.

[4] Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow, UK.

[5] Division of Infectious Diseases and Geographic Medicine, Stanford University School of Medicine, Stanford, California, USA. kesiaeds@stanford.edu, sluby@stanford.edu, jandr@stanford.edu

[6] Department of Paediatrics and Child Health, Aga Khan University, Karachi, Pakistan. junaid.iqbal@aku.edu, farah.qamar@aku.edu

[7] Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. gd312@medschl.cam.ac.uk

[8] Cambridge Institute of Therapeutic Immunology and Infectious Disease, Department of Medicine, University of Cambridge, Cambridge, United Kingdom. sgb47@medschl.cam.ac.uk

[9] Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, UK. Zoe.Dyson@lshtm.ac.uk

[10] Applied Epidemiology Team, Sabin Vaccine Institute, Washington, DC, USA. Denise.Garrett@Sabin.org

[11] Department of Microbiology, Bangladesh Institute of Child Health, Dhaka Shishu Hospital, Dhaka, Bangladesh.

[12] Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria 3004, Australia.

*, [$] Equal contribution;

[#] Corresponding author

Abstract

Background: *Salmonella enterica* serovar Paratyphi A (*Salmonella* Paratyphi A) is the primary causative agent of paratyphoid fever, which is responsible for an estimated 3.4 million infections annually. However, little genomic information is available on population structure, antimicrobial resistance (AMR), and spatiotemporal distribution of the pathogen. With rising antimicrobial resistance and no licensed vaccines, genomic surveillance is important to track the evolution of this pathogen and monitor transmission.

Results: We performed whole-genome sequencing of 817 *Salmonella* Paratyphi A isolates collected from Bangladesh, Nepal, and Pakistan and added publicly available 562 genomes to build a global database representing 37 countries, covering 1917-2019. To track the evolution of *Salmonella* Paratyphi A, we used the existing lineage scheme, developed earlier based on a small dataset, but certain sub-lineages were not homologous, and many isolates could not be assigned a lineage. Therefore, we developed a single nucleotide polymorphism based genotyping scheme, Paratype, a tool that segregates *Salmonella* Paratyphi A into three primary and nine secondary clades, and 18 genotypes. Each genotype has been assigned a unique allele definition located on a conserved gene. Using Paratype, we identified genomic variation between different sampling locations and specific AMR markers, and mutations in the O2-polysaccharide synthesis locus, a candidate for vaccine development.

Conclusions: This large-scale global analysis proposes the first genotyping tool for *Salmonella* Paratyphi A. Paratype has already been released (https://github.com/CHRF-Genomics/Paratype) as an open-access, command-line tool and is being adopted for large scale genomic analysis. This tool will assist future genomic surveillance and help inform prevention and treatment strategies.

Keywords: *Salmonella* Paratyphi A; Paratyphoid fever; Paratyphi A genotyping; Genomics; Antimicrobial resistance; Global analysis; Epidemiology; Enteric fever; Neglected tropical disease.

Background

Paratyphoid fever, caused by *Salmonella enterica* subspecies *enterica* serovar Paratyphi A (*Salmonella* Paratyphi A) is a systemic febrile illness that affects an estimated 3.4 million people each year, and causes 19,100 deaths globally [1]. The disease is clinically indistinguishable from typhoid fever, caused by *Salmonella enterica* subspecies *enterica* serovar Typhi (*Salmonella* Typhi). Much like typhoid, paratyphoid fever is also endemic in many low- and middle-income countries of South Asia and Sub-Saharan Africa, due to fecal contamination of water, food and the environment. However, barring a few countries (e.g., China, Myanmar), paratyphoid fever is usually less prevalent than typhoid fever [2,3]. *Salmonella* Paratyphi A continues to be an inadequately studied pathogen [4] hampering the implementation of evidence-based policies for the treatment and prevention of paratyphoid fever.

Relative to *Salmonella* Typhi, little genomic information is available on population structure, antimicrobial resistance (AMR), and spatiotemporal distribution of *Salmonella* Paratyphi A. The first *Salmonella* Paratyphi A genome was published in 2004 and had a size of 4.5 Mb, with ~4,200 genes. To determine the global diversity of *Salmonella* Paratyphi A isolates, Bayesian analysis was conducted on a set of 149 *Salmonella* Paratyphi A genomes, which identified that the last common ancestor of all *Salmonella* Paratyphi A existed for at least 450 years prior to differentiating into at least seven distinct lineages (A to G) which have circulated globally [5]. Whole genome sequencing was also used to characterize clonal paratyphoid outbreaks in Cambodia [6] and China [7] and further extend the lineage scheme to include sub-lineages within Lineage A and C. However, very few studies have characterized isolates from countries in South Asia, which contributes over 80% of all paratyphoid infections [8,9]. Available studies are sporadic, and either focused on genomes

121 from a specific geographical location or provide no information on antimicrobial resistance
122 markers, potential vaccine targets, and other virulence factors.

123

124 To address this data gap, we performed whole-genome sequencing of 817 *Salmonella*
125 Paratyphi A isolates collected from Bangladesh, Nepal, and Pakistan and combined them
126 with whole-genome sequence data of another 562 isolates reported in the literature to build a
127 global database of 1,379 *Salmonella* Paratyphi A isolates. To track the evolution of
128 *Salmonella* Paratyphi A over a century, we used the existing lineage scheme and found that
129 certain sub-lineages were not homologous, and many isolates could not be assigned a specific
130 lineage. This motivated us to develop a single nucleotide polymorphism (SNP) based
131 genotyping scheme, called Paratype. The scheme is phylogenetically informative and
132 successfully segregates the global population structure into three primary, seven secondary,
133 and 18 distinct subclades/genotypes. We also identified the specific antimicrobial resistance
134 genes, mutations, and plasmids present in *Salmonella* Paratyphi A genomes and correlated
135 these with the different genotypes.

136

137 **Results**

138 Whole-genome sequencing and compilation of global *Salmonella* Paratyphi A genomes

139 The Child Health Research Foundation (CHRF) has been conducting typhoid and
140 paratyphoid fever surveillance in Bangladesh since 1999 and has generated a biobank of
141 1,123 *Salmonella* Paratyphi A isolates from 1999-2018 [10–12]. We selected 528 of these
142 isolates, covering all age groups, years of isolation, and hospitalization status
143 (hospitalized/out-patient), and performed whole-genome sequencing on these isolates
144 (Additional file 1: Table S1). Of these, 180 *Salmonella* Paratyphi A isolates were collected as

145 part of the Surveillance of Enteric Fever in Asia Project (SEAP, 2014 - 2019) study, a multi-
146 country international effort to better understand the epidemiology and impact of enteric fever
147 in South Asia [13]. In addition to Bangladesh, 133 isolates were sequenced from the SEAP
148 study conducted in Pakistan, and 156 from Nepal.

149

150 To contextualize these genomes, we conducted a literature search to compile all publicly
151 available *Salmonella* Paratyphi A genomes (for which raw reads were available) to build a
152 database of 560 additional isolates from 10 studies (Additional file 1: Table S2). Two
153 reference genomes (ATCC 9150 and AKU_12601) were also included. The largest dataset
154 consisted of 254 isolates, published by Public Health England as part of their *Salmonella*
155 surveillance [8,14]; 164 of these isolates were linked to travel, most commonly to South Asia.
156 In our study, we assigned these isolates to the countries where the patient acquired the
157 infection. Our final data, including the genomes we sequenced, consisted of a total of 1,379
158 isolates from 37 different countries, spanning over 103 years - 1917 to 2019. Most of the
159 isolates (1,112/ 1,379; 81%) were from countries in South Asia (541 from Bangladesh, 268
160 from Nepal, 187 from Pakistan and 115 from India). South Asian countries also bear a
161 disproportionately high burden of paratyphoid fever; of the estimated 3.4 million global
162 paratyphoid infections in 2019, 2.8 (82%) million are estimated to have occurred in South
163 Asia [1].

164

165 Following assembly from raw reads, the pan-genome analysis identified 6,983 genes, of
166 which 4,114 (59% of all genes) were conserved in more than 95% of isolates (Additional file
167 1: Figure S1). The average genome size was 4.5 Mb with ~4,300 genes, and the pan genome
168 does not appear to be closed (decay parameter, alpha = 0.67). Overall, 2,550 genes were

found to be present in less than 15% of isolates, and these included genes often found in prophages and other mobile regions, and genes encoding adhesins, antimicrobial resistance markers, and several hypothetical proteins.

Genotyping scheme for *Salmonella* Paratyphi A

To investigate the genomic diversity of *Salmonella* Paratyphi A, we identified 8,346 single nucleotide polymorphisms (SNPs) in the 1,379 isolates. These were used in RAxML [15] to generate a Maximum-likelihood phylogenetic tree of the global collection of *Salmonella* Paratyphi A isolates (Figure 1). A previously reported lineage scheme, proposed for *Salmonella* Paratyphi A by Zhou et al. [5] and extended by subsequent studies [6,7,9,16,17] was overlaid on the RAxML tree. This highlighted the insufficiency of the current lineage scheme to fully capture the diversity of *Salmonella* Paratyphi A present. First, while the isolates from lineages B & D - G clustered together, several isolates previously assigned to lineages A and C in the scheme did not. Second, some sequences belonged to clades that diverged from isolates before the exitance of the most recent common ancestor for lineages A and B, indicating that these isolates should be considered to be in a different lineage. This was not surprising considering that when this scheme was devised, there were limited number of sequenced *Salmonella* Paratyphi A genomes available, particularly from South Asia.

To build a genotyping scheme based on a larger number of representative samples, first, we used fastBAPS [18] to generate a potential list of clusters in the RAxML tree (Additional file 1: Figure S2). Next, we selected a set of 315 isolates that included two isolates per year for all fastBAPS clusters selected randomly and performed phylodynamic analysis using the Bayesian Evolutionary Analysis by Sampling Trees (BEAST) software [19] (Figure 2). Based

on these analyses, we devised a genotyping scheme with three primary clades, nine secondary clades, and 18 genotypes that have circulated globally in the last 100 years.

To aid further genomic epidemiological studies, we identified 18 additional alleles (Additional file 1: Table S3) that are unique to each of the 18 *Salmonella* Paratyphi A genotypes. These alleles were present in conserved genes involved in essential cellular functions such as protein synthesis, DNA replication, or metabolism. Identification of these genotype-specific alleles allowed us to write a Python script – "Paratype" – that assigns genotypes to *Salmonella* Paratyphi A genomes using fastq, bam, or vcf files obtained during whole-genome sequencing and variant calling. The Paratype software tool (available at: https://github.com/CHRF-Genomics/Paratype/) has 100% sensitivity and specificity and was able to assign the correct genotype to all the 1,379 genomes that were present in our database.

Temporal and geographic distribution of different genotypes

Upon the establishment of the "Paratype" scheme, we considered the geographical distribution of the different genotypes (Figure 3). Genotype 0.1 under primary clade 0 was phylogenetically unique (matches with lineage H of Zhou et al [5]); there was only one isolate belonging to this genotype/primary clade that was isolated in Hong Kong in 1971. The genome of this isolate was distinct from all other genomes obtained thus far, contained 1288 unique SNPs, and may represent a lineage that is now extinct, or present at very low numbers in areas that have not been sampled. The other two primary clades, clades 1 and 2, emerged between 1700-1800 and contain genomes that have been collected in the last two decades. Clade 1 contains strains largely from lineage F, and fastBAPS predicted two sub-clusters within this clade. One of these clusters was largely found in Bangladesh and has been

assigned secondary clade 1.2, then sub-divided into genotypes 1.2.1 and 1.2.2 which appear to have diverged in the 1950s. Both these genotypes are currently present in Bangladesh and other South Asian countries (Figure 2). The other cluster with13 genomes from Bangladesh that were first isolated in 1999 have been assigned to genotype 1.1. The remaining 10 genomes were obtained between 1917 to 1963 and have been assigned genotype 1.0.

Most *Salmonella* Paratyphi A genomes (1254/1379; 91%) have been assigned to primary clade 2, which contains genomes belonging to the lineage A-E of the previous scheme. Genomes that belonged to lineages B, D, and E have now been assigned to genotypes 2.4, 2.2, and 2.0, respectively. Within genotype 2.0, 13 unique and recent isolates from Pakistan were identified and have been assigned as genotype 2.0.1. Genotype 2.1 contains isolates from Nepal that were sampled during the SEAP study, yet the genotype emerged in the 1800s and is distinct from all other isolates in clade 2. Two clusters in fastBAPS, comprising of strains largely from what was formerly C lineage are now assigned to genotype 2.3. Genotype 2.3 has been subdivided into genotypes 2.3.1 to 2.3.3, each of which belongs to a distinct geographical location: 2.3.1 is found predominantly in Cambodia and South-East Asia; 2.3.2 and 2.3.3 are found largely in South Asia. An outbreak of paratyphoid fever in China during 2010 – 2011 [7] was caused by isolates of genotype 2.3.3, and these likely originated in South Asia. The former lineages A and B have been assigned genotype 2.4, which is further divided into 2.4.1 to 2.4.4. While genotypes 2.4.1 and 2.4.2 have been observed in different countries in South Asia, genotype 2.4.4 is predominantly found in Bangladesh, and 2.4.3 is largely present in Nepal.

Different countries in South Asia had unique genotype distributions. Predominant genotypes present in Bangladesh were 2.4.4 (56%) followed by 1.2.2 (14%) and 2.3.3 (13%). In Nepal, 2.4.3 (47%), 2.3.3 (16%) and 2.4.1 (14%) were three most common genotypes. Pakistan had genotypes 2.3.3 (25%), 2.3 (16%) and 2.4 (15%). In India, genotypes 2.4.2 (22%), 2.4 (20%), 2.4.1 (19%), 2.3.3 (17%), and 2.3 (16%) were commonly identified.

Antimicrobial resistance markers in *Salmonella* Paratyphi A

To characterize genomic determinants of antimicrobial resistance in *Salmonella* Paratyphi A, we screened the 1,379 genomes for the presence of antimicrobial genes and markers using ResFinder [20] (Figure 4a) and plasmids using PlasmidFinder [21] (Figure 4b). Of the 1,379 isolates, 1,015 (74%) isolates showed no predicted plasmids and 1356/1379 had no predicted antimicrobial resistance genes. Five genomes with the IncHI1 plasmid were identified, two genomes (both from India) contained resistance genes for trimethoprim and chloramphenicol, and the other three genomes contained genes for trimethoprim, chloramphenicol and ampicillin designated as MDR isolates (one each from India, Pakistan, and Thailand). All five genomes belonged to genotype 2.3 and the strains were isolated between 1999-2004. We also identified a genome belonging to genotype 2.4.4 containing *bla*CTX-M-15 and *bla*TEM-1B on an IncI1-I plasmid; the originating strain was isolated from a patient who contracted the infection in Bangladesh in 2017 [22]. There were 14 isolates from the genotype 2.3.1 that contain *bla*TEM-116, which can lead to resistance to ampicillin; all 14 were reported from Cambodia[6]. Another isolate from genotype 2.3.3 (from Pakistan, 2015) contained a *qnrB19* gene on a Col(pHAD28) plasmid, which has been shown to lead to quinolone resistance in other *Salmonella* species [23].

In addition to antimicrobial resistance genes, we also identified chromosomal mutations in the *acrB* gene and the quinolone resistance determining region (QRDR) to identify isolates resistant to azithromycin and ciprofloxacin respectively. Six of 1,379 genomes contained an AcrB R717 mutation, all from Bangladesh and these belonged to genotypes 2.3.3 (1/6) and 2.4.4 (5/6) [24,25]. The first azithromycin resistant *Salmonella* Paratyphi A isolate was identified in 2014, and this resistance has emerged independently at least twice in two different genotypes. On the other hand, a majority (1177/1397; 84%) of genomes had mutations in the QRDR region. The most commonly found single mutation was gyrA-S83F (941/1379), followed by gyrA-S83Y (205/1379). Two isolates contained double mutations in the QRDR region; one of them belonged to genotype 2.0.1 (gyrA-S83F & D87N, Pakistan, 2017) and another belonged to genotype 2.3.3 (gyrA-S83F & D87G, UK, 2016). Barring genotype 0.1, 1.0 and 2.2, all other genotypes had at least one genome with a QRDR mutation (Figure 4c). The first QRDR mutation was identified in 1997 in India in genotype 2.4 and their prevalence have increased over time. In 2012 and 2013, there was an outbreak in Cambodia caused by a strain from genotype 2.3.1 that did not have any QRDR mutation leading to a temporary increase in proportion of *Salmonella* Paratyphi A with no QRDR mutations during these two years (Additional file 1: Figure S3).

Characterization of mutations in the O2-antigen biosynthetic gene cluster

The majority of the vaccines being developed for *Salmonella* Paratyphi A use the O2-antigen that is unique to this serovar conjugated to a carrier protein [26]. Recently, through in-silico metabolic reconstruction, an 18.9 kb region containing genes involved in O-antigen biosynthesis was identified as important for determining the specific molecular features of the O2-antigen found in *Salmonella* Paratyphi A [27]. We identified the SNPs in the O2-antigen biosynthesis genes found in the 1,379 genomes to investigate the conservation of this

289 genomic loci. In total, 84 SNPs were found, of which 13 were present in more than 10
290 genomes. The most common SNP was at genomic location 8,68,444 (G> C; synonymous
291 mutation in *prt* gene encoding paratose synthase), which was found in 17% (239/1,379) of all
292 isolates. Out of those 13 common SNPs (n≥10), seven led to non-synonymous mutations
293 (Additional file 1: Figure S4) that could potentially change the O2-antigen structure and
294 chemistry.

295

296 **Discussion**

297 *Salmonella* Paratyphi A is the causative agent of paratyphoid fever, a neglected tropical
298 disease with a high burden and mortality in low-and-middle-income countries. Limited
299 information is available regarding its genomic diversity, especially from South Asian
300 countries that collectively are responsible for over 80% of all paratyphoid cases. As genomic
301 surveillance becomes more prominent, there is a need for a coherent and easy-to-use scheme
302 that can be deployed by public health researchers that do not require extensive compute
303 resources.

304

305 We sequenced 817 isolates from Bangladesh, Pakistan and Nepal collected over the last 20
306 years and compiled a collection of all genomes of *Salmonella* Paratyphi A publicly available
307 thus far. We describe a genotyping framework for *Salmonella* Paratyphi A using 1,379
308 isolates obtained from 1917 through 2019. Rather than being guided by a single approach, we
309 combined ML-based phylogenetics with BAPS and Bayesian analysis via BEAST to design a
310 genotyping scheme for *Salmonella* Paratyphi A. The scheme divided the *Salmonella*
311 Paratyphi A population into 18 different genotypes, and each can be identified by the
312 presence of an allele that is located on the coding sequence of a conserved gene, involved in

313 housekeeping functions. We only found 8,346 SNPs from all 1,379 isolates, with minimal
314 recombination, and thus, this genotyping scheme based on SNP alleles can support robust
315 genotyping and accommodate future evolution of *Salmonella* Paratyphi A. And to assist with
316 that, we have developed Paratype, an open-source Python script for genotyping of *Salmonella*
317 Paratyphi A genomes. Paratype can detect the genotype of *Salmonella* Paratyphi A genomes
318 directly from raw fastq read data. It can also detect mutations in the acrB efflux pump
319 (determinant of macrolide resistance) and in the QRDR region (determinant of ciprofloxacin
320 non-susceptibility).

321

322 In this genotyping scheme, we propose three primary clades 0, 1, and 2, which diverged
323 before the 1800s (Figure 2). While only a single isolate of primary clade 0 was obtained in
324 1971, isolates belonging to clade 1 and 2 have been routinely identified over the past two
325 decades. Clade 2 is the most abundant and has been subdivided into four secondary clades:
326 2.1 - 2.4, which probably emerged in the 1800s or the early 1900s. Clade 2.3 could be
327 subdivided into 2.3.1 - 2.3.3, each with distinct geographic distribution. Clade 2.4 was also
328 sub-divided into genotypes 2.4.1 - 2.4.4. Genotype 2.4.4 was the most abundant and was
329 predominantly present in Bangladesh. This genotype emerged in the late 1990s to early 2000s
330 and possesses high rates of ciprofloxacin resistance (Figure 2). Five of the isolates from this
331 genotype also contained AcrB-R717Q mutation that leads to azithromycin resistance, while
332 one was found to harbor a plasmid containing extended-spectrum beta-lactamase gene
333 (*bla*CTX-M-15) [22].

334

335 In line with findings of previous studies, the rates of acquisition of antimicrobial resistance
336 markers in *Salmonella* Paratyphi A is lower relative to *Salmonella* Typhi (Figure 4) [6,9].

Although a few isolates did acquire the IncHI1 plasmid in the late 1990s to early 2000s (Figure 4a), no massive spread across the globe was noted; this unlike *Salmonella* Typhi lineage H58 (genotype 4.3.1) carrying the IncHI1 plasmid spread and became the dominant lineage in the last 30 years[28]. This is also true for chromosomal mutations such as QRDR and AcrB mutations, which are overall less prevalent in *Salmonella* Paratyphi A than in *Salmonella* Typhi [28,29]. Considering the genetic similarities between *Salmonella* Typhi and Paratyphi A, and the fact that they occupy the same environmental niche, the differences in the presence of AMR genes between these typhoidal *Salmonella* serovars warrants further investigation.

The specific O-antigen in the *Salmonella* Paratyphi A is thought to be conserved (assigned to serogroup O2) and several vaccine candidates are currently under development, utilizing the O2 antigen conjugated to a carrier protein as the main vaccine antigen. We compared the 18.9 kbp region responsible for the synthesis of the O2 antigen in this serovar [27] and found 83 SNPs in this region, of which 7 non-synonymous mutations were present in >10 isolates. While it is not clear if these mutations affect the O2-antigen chemistry, the low mutation rate and no observed recombination events in the cluster suggests that the O2 antigen vaccine will have a broadly protective response against all the *Salmonella* Paratyphi A genotypes sampled thus far. However, any variations in this region should be carefully monitored through genomic surveillance.

The conclusions that we can draw from this analysis are subject to certain limitations. First, the available genomes are an incomplete sample; *Salmonella* Paratyphi A is a neglected pathogen, and hence the available genomes and might not have broad representativeness

across geographies or time. Specifically, a small proportion of genomes were available from countries in sub-Saharan Africa and India. Second, while the tool has high sensitivity and specificity on our dataset, as more genomes become available over time and novel mechanism of AMR emerge, this tool will require updates from the bigger scientific community. Like all genotyping tools, Paratype is a living tool that will require updates. Our diverse group of authors plans to continually monitor the library of publicly available genomes, accept update requests via GitHub, and incorporate any required updates in the Paratype scheme accordingly.

Conclusions

This study reports the first large-scale global analysis of *Salmonella* Paratyphi A genomes and proposes the first genotyping tool for this pathogen. Paratype, which has already been released (https://github.com/CHRF-Genomics/Paratype) as an open-access, easy-to-use, command-line tool, is being tested and adopted by researchers for large scale genomic analysis (https://doi.org/10.5281/zenodo.5520408). This tool will assist future genomic surveillance studies and will help inform prevention and treatment strategies for this neglected pathogen.

Methods

Study site and isolate selection

Child Health Research Foundation in Bangladesh has been preserving invasive *Salmonella* isolates since 1999 and maintains a biobank of >9000 typhoidal *Salmonella* isolates, largely from children (<18 years of age) that were isolated from the blood of the patients in two different settings: in-patient (hospitalized), and out-patient (community) facility [30]. Clinical

and epidemiological data were collected for all hospitalized patients. From this biobank, among 640 *Salmonella* Paratyphi A isolates collected till December 2016, 348 were randomly selected for whole-genome sequencing (WGS) (Additional file 1: Table S1). A set of 469 *Salmonella* Paratyphi A isolates were also added to this collection, isolated under the Surveillance for Enteric Fever in Asia (SEAP) project from three different typhoid-endemic countries, Bangladesh (n= 180), Nepal (n= 156), and Pakistan (n=133). The SEAP-Bangladesh isolates (n=180) were selected using randomization to represent 483 isolates collected between 2016 and 2018. In contrast, SEAP-Nepal isolates included all pre-SEAP isolates (2014 – 2016) and randomly selected SEAP isolates (2017 – 2019). The SEAP-Pakistan isolates were selected prioritizing the availability of geographic information and susceptibility profile during 2016 – 2018.

To add to all the isolates sequenced in this study, we also collected raw fastq data of 560 *Salmonella* Paratyphi A isolates from 37 different countries and 10 published articles (Additional file 1: Table S2). Complete chromosomal sequences of *Salmonella* Paratyphi A ATCC 9150 (NC_006511) and AKU_12601 (NC_011147) were also included [31,32]. For travel-related paratyphoid cases, the country of "traveling from" was considered as the country of origin. If no travel data is available, the country of "reported from" was considered as the country. Overall, for globally distributed 562 *Salmonella* Paratyphi A, year and country data were available for 507 and 536 respectively (Additional file 1: Table S2). In total, we obtained a global collection of 1,379 *Salmonella* Paratyphi A covering a timeline of 1917 – 2019 and 37 countries [see Additional file 2 for more details].

Whole-genome sequencing

Salmonella Paratyphi A isolates from 1999-2016 (before the start of the SEAP project) from Bangladesh (n=348) were subcultured on MacConkey agar media and kept overnight at 37°C. In case of any visible contamination, a single colony was picked and subcultured again. Later, all colonies were swapped and resuspended into 1 ml of water. From this suspension, 400 µL was used for DNA extraction using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) and sent to Novogene (NovogeneAIT, Singapore) for WGS on Novaseq 6000 platform (PE150). All SEAP isolates were extracted using the same protocol and were sequenced on Illumina HiSeq 4000 platform (PE150) at the Welcome Sanger Institute, Cambridge, UK.

Systematic literature review of existing *Salmonella* Paratyphi A genomes

To contextualize the genomes sequenced in this study, we conducted a systematic search to compile all publicly available *Salmonella* Paratyphi A genomes (for which raw reads and metadata were available) to build a database of 560 additional isolates from 10 studies (Additional file 1: Table S2). First, the search terms "(Salmonella Paratyphi A) AND (Molecular Epidemiology)" "Salmonella Paratyphi A genome" and "(Salmonella Paratyphi A) AND (Genomic Epidemiology)" were used in PubMed advanced search builder. Next, the hits were filtered by selecting dates between 1900 and 2019 and the total number of publications remaining were 231. After screening the abstracts and titles manually and eliminating duplicated, only 7 studies were found to have any kind of genome/metadata available for further analysis. In addition, three studies [8,9,22] that meets our criteria (published and both metadata and raw reads available) but missed/not published during the initial PubMed search were incorporated from European Nucleotide Archive (ENA) database, taking the final number of incorporated publications to 10.

433

434 Quality check, genome assembly, annotation, and pan-genome analysis

435 Raw fastq reads of all *Salmonella* Paratyphi A were quality-checked using FastQC and
436 trimmed using Trimmomatic if necessary[33]. All 1,377 sets of raw fastq reads were
437 assembled using Unicycler v0.4.8 (*default with --min_fasta_length 200*)[34]. The assembled
438 contigs (n = 1,377) and downloaded complete chromosomes (n = 2) were annotated using
439 Prokka (*--gcode 11 --mincontiglen 200*) [35]. The annotated GFF files of all 1,379 isolates
440 were used to build a pan- and core-genome of *Salmonella* Paratyphi A using Roary v3.3 (
441 *options: -t 11 -e --mafft -n*)[36]. The gene_presence_absence matrix output was used to
442 perform the Heap's law analysis to understand the open/closedness of the pan-genome (*heaps*
443 function of *micropan* library on R; 1000 permutations).

444

445 SNP-based phylogenetic analyses

446 For the complete "global+SEAP" raw data collection, fastq reads of 1,377 *Salmonella*
447 Paratyphi A and fasta of two RefSeq chromosomes (NC_006511 and NC_011147) were
448 mapped against the *Salmonella* Paratyphi A AKU_12601 (FM200053.1) using Bowtie2
449 v2.3.5.1 [37]. Candidate SNPs were identified using SAMtools (v1.10) and BCFtools
450 (v1.10.2) [38]. Only the homozygous, unambiguous SNPs with a Phred-quality score of >20
451 were selected using a customized Python script. SNPs were discarded if they had strand bias
452 p <0.001, mapping bias p <0.001 or tail bias p <0.001 (using vcfutils.pl script, from
453 SAMtools). SNPs located in phage or repeat regions (118.9 kb for *Salmonella* Paratyphi A
454 AKU_12601 as described in Sajib et al. [25]) were also excluded using a customized python
455 script. Gubbins v2.3.4 was used to detect the recombinant regions [39] and SNPs in those
456 regions were excluded as well using the same python script, resulting in a set of 8,346

457 chromosomal SNPs positions for the "global+SEAP" collection (n= 1,379). All SNP alleles
458 were extracted (fasta) using a customized python script and merged to produce SNP
459 alignment.

460

461 Maximum likelihood trees (MLT) were built from the chromosomal SNP alignments using
462 RAxML v8.2.12 (with the Generalized Time-Reversible model and a Gamma distribution to
463 model site-specific rate variation; GTRGAMMA in RAxML) [15]. Support for the MLT was
464 calculated using 100 bootstrap pseudo-analyses of the alignment. The MLT was outgroup-
465 rooted by including the pseudo-alleles from *Salmonella* Typhi CT18 (NC_003198.1) in the
466 alignment. Tree visualization was done using iTol v5.5 [40], including the previous Paratyphi
467 A lineages proposed by Zhou *et al* [5].

468

469 Bayesian analysis and identifying phylogenetically informative clades and subclades

470 In addition to SNP-based MLT, we investigated the population structure of the global
471 *Salmonella* Paratyphi A collection using a Bayesian approach, implemented with the SNP
472 alignment using fastBaps[40]. To maintain compatibility with the phylogeny, some minor
473 modifications were made to the clustering pattern proposed by the least conservative
474 Dirichlet prior hyperparameters on fastbaps, *optimise.baps*. This eventually resulted in a total
475 of 16 different clusters. A customized python script was used to randomly select two
476 isolates/year/cluster to represent this global collection of *Salmonella* Paratyphi A, leading to
477 two independent sample sets of 315 isolates each. The alignment of SNP-alleles for this
478 representative sample set was used to understand the evolutionary diverging pattern of
479 different *Salmonella* Paratyphi A clusters over time using BEAST v1.10.4 [19]. The
480 GTR+Γ(4) substitution model was selected for this analysis with the exponential unrelated

relaxed clock as clock type and Bayesian skyline coalescent model as tree prior. The analysis considered the year of isolation as tip dates and continued for 500 million steps with sampling every 50,000 iterations. The BEAST analysis was run twice each on the two independently generated sets of isolates. The resulting log files and model parameters were analyzed on Tracer v1.7.1. TreeAnnotator v1.10 was used to generate the maximum–clade-credibility (MCC) tree [41]. The tree was visualized on FigTree v1.4.4 with a time scale. For the model with the highest posterior values (joint effective sample size (ESS) of 544) used for further analysis, time to last common ancestor (MRCA) was calculated to be 1407 AD (95% highest posterior density (HPD) interval [721.0, 1637.3]). Based on the diverging patterns suggested by the MCC tree, we assigned the clusters (defined as described above) into primary clades, secondary clades, and subclades on the MLT. However, a few visible clusters on the MLT could not be assigned to specific subclades due to a lack of clustering information from fastBaps, likely due to the low number of SNPs unique to these clusters.

SNP-based genotyping scheme and paratype

We further divided the 16 clusters obtained from fastBAPS into 18 genotypes and identified a set of 18 SNP alleles, located in a coding sequence for conserved genes to define each assigned secondary clade and sub-clades. Each SNP allele was unique to only one subclade or, to one secondary clade and its corresponding subclades (if any). Therefore, we assigned the term “genotype” to each of the 18 secondary clades or subclades. Sorted read alignment (BAM) files generated during the SNP analysis were used to assign the genotypes for each isolate using a customized Python script, named Paratype (available at https://github.com/CHRF-Genomics/Paratype). Briefly, under BAM mode *(--mode bam*), Paratype uses *samtools index* (if bam file is not indexed), *samtools mpileup*, and *bcftools call* to extract the consensus base calls at those 18 SNP loci from the BAM file. The resulting

variant call format (VCF) file is then processed to identify the presence of the defining SNP alleles and follow cladistic logic to assign the genotype of the isolate, as well as the primary clade, secondary clades, and subclade information. Paratype only considers high-quality SNP alleles (Phred score >20 and 75% read_ratio for the allele) to assign genotypes. Read_ratio is calculated by the number of high-quality alternative-allele reads on both strands, divided by the total number of high-quality reads. In addition, Paratype also has fastq mode (*--mode fastq*) where a user can provide a set of paired-end raw fastq data file (can be gzipped) and Paratype performs reference mapping (against the *Salmonella* Paratyphi AKU_12601 genome) using Bowtie and SAMtools and follows the same steps described above to detect the genotype of the isolates. Although the bam mode is the default for the tool, the fastq mode is more accurate and should be user-friendly to non-coding specializing researchers; however, it is more time-consuming. Paratype also runs on vcf mode (*--mode vcf*) which is faster, but also the least accurate if the provided SNPs are not highly trusted.

Plasmid, resistance gene, and mutation analysis

All assembled contigs were screened with PlasmidFinder v2.1 [21] and ResFinder v3.2 [20] to detect plasmid amplicons and acquired AMR genes respectively. Both results were parsed using customized python scripts. To detect mutations in *gyrA* and *acrB* genes, we used the same Paratype script. It uses the same files used for genotyping and produces gene- and position-specific non-silent and silent mutation results.

We also explored the genomic region where the genes related to O2-antigen biosynthesis are located (860,008 – 878,865 of AKU_12601 genome). We detected all SNPs in that region with the number of isolates having those and their corresponding amino-acid changes using

the Paratype. Two additional python scripts were used to count position-specific SNPs and mutations for the 18.9 kbp region.

Data visualization and statistical analysis

R (v4.0.4) base function and several packages including dplyr, ggplot2, micropan and scatterpie were used for data visualization and statistical analysis.

Declarations

Ethics approval and consent to participate

Ethical approval for the parent studies were obtained from the Bangladesh Institute of Child Health Ethical Review Committee, Nepal Health Research Council, Aga Khan University Hospital Ethics Committee and Pakistan National Ethics Committee, Stanford University Institutional Review Board, and U.S. Centers for Disease Control and Prevention. Informed written consent and clinical information were taken from adult participants and legal guardians of child participants.

Consent for publication

Not applicable (No data from individual person was used for analysis).

Availability of data and materials

The genome dataset supporting the conclusions of this article are available in the European Nucleotide Archive (ENA) under study accession ERP132884. The genotyping tool for

552 *Salmonella* Paratyphi A, Paratype is available at https://github.com/CHRF-
553 Genomics/Paratype (https://doi.org/10.5281/zenodo.5520408). Customized Python scripts
554 and color scheme used in the manuscript are available at https://github.com/CHRF-
555 Genomics/CHRF_Paratyphi_scripts. The metadata supporting the conclusions of this article
556 is included in Additional file 2.

557

558 Competing interests

559 The authors declare no competing interests.

560

561 Funding

562 This study was supported by the Bill and Melinda Gates Foundation (grant numbers INV-
563 023821 and INV-008335). The funding body did not have any role in the design of the study,
564 analysis, and interpretation of data or, in writing the manuscript.

565

566 Authors' contributions

567 AMT, YH, MSIS, SKS and SS were involved in conceptualization and design of the study.
568 MSIS performed the DNA extraction for sequencing in Bangladesh and the literature review
569 for the global database construction. AMT, YH and MSIS performed bioinformatic analysis
570 under supervision of SS and SKS. JRA provided continuous guidance during bioinformatic
571 analysis. AMT and YH designed the genotyping scheme and AMT wrote the Paratype script.
572 YH and MSIS conducted the statistical analyses and visualization. KES, JI, ZAD, SB and
573 JRA reviewed the results. ZAD and SB reviewed genotyping scheme and the Paratype tool.
574 AMT, YH, MSIS and SS wrote the first draft of the manuscript. KES, JI, FNQ, SPL, GD,

ZAD, SB, DOG, JRA and SKS reviewed the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

We are thankful to Mr. Hafizur Rahman, Mr. Dipu Chandra Das and Ms. Nusrat Alam of the Child Health Research Foundation for their help with the wet-lab procedures. We are also thankful to the entire SEAP team for their unwavering support and coordination between the teams.

References

1. Stanaway JD, Reiner RC, Blacker BF, Goldberg EM, Khalil IA, Troeger CE, et al. The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet Infectious Diseases [Internet]. [cited 2019 Feb 28];0. Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(18)30685-6/abstract

2. Crump JA, Mintz ED. Global trends in typhoid and paratyphoid fever. Clin Infect Dis. 2010;50:241–6.

3. Lu X, Li Z, Yan M, Pang B, Xu J, Kan B. Regional Transmission of *Salmonella* Paratyphi A, China, 1998–2012. Emerg Infect Dis. 2017;23:833–6.

4. Furuse Y. Analysis of research intensity on infectious disease by disease burden reveals which infectious diseases are neglected by researchers. PNAS. National Academy of Sciences; 2019;116:478–83.

597 5. Zhou Z, McCann A, Weill F-X, Blin C, Nair S, Wain J, et al. Transient Darwinian
598 selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of
599 enteric fever. Proc Natl Acad Sci U S A. 2014;111:12199–204.

600 6. Kuijpers LMF, Le Hello S, Fawal N, Fabre L, Tourdjman M, Dufour M, et al. Genomic
601 analysis of *Salmonella enterica* serotype Paratyphi A during an outbreak in Cambodia, 2013–
602 2015. Microb Genom [Internet]. 2016 [cited 2019 Aug 25];2. Available from:
603 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5320704/

604 7. Yan M, Yang B, Wang Z, Wang S, Zhang X, Zhou Y, et al. A Large-Scale Community-
605 Based Outbreak of Paratyphoid Fever Caused by Hospital-Derived Transmission in Southern
606 China. PLOS Neglected Tropical Diseases. Public Library of Science; 2015;9:e0003859.

607 8. Ashton PM, Nair S, Peters TM, Bale JA, Powell DG, Painset A, et al. Identification of
608 *Salmonella* for public health surveillance using whole genome sequencing. PeerJ. PeerJ Inc.;
609 2016;4:e1752.

610 9. Britto CD, Dyson ZA, Duchene S, Carter MJ, Gurung M, Kelly DF, et al. Laboratory and
611 molecular surveillance of paediatric typhoidal Salmonella in Nepal: Antimicrobial resistance
612 and implications for vaccine policy. PLOS Neglected Tropical Diseases. 2018;12:e0006408.

613 10. Saha S, Islam M, Uddin MJ, Saha S, Das RC, Baqui AH, et al. Integration of enteric fever
614 surveillance into the WHO-coordinated Invasive Bacterial-Vaccine Preventable Diseases (IB-
615 VPD) platform: A low cost approach to track an increasingly important disease. PLOS
616 Neglected Tropical Diseases. 2017;11:e0005999.

617 11. Saha S, Islam M, Saha S, Uddin MJ, Rahman H, Das RC, et al. Designing
618 Comprehensive Public Health Surveillance for Enteric Fever in Endemic Countries:
619 Importance of Including Different Healthcare Facilities. J Infect Dis. 2018;218:S227–31.

620 12. Saha S, Islam MS, Sajib MSI, Saha S, Uddin MJ, Hooda Y, et al. Epidemiology of
621 Typhoid and Paratyphoid: Implications for Vaccine Policy. Clin Infect Dis. 2019;68:S117–
622 23.

623 13. Barkume C, Date K, Saha SK, Qamar FN, Sur D, Andrews JR, et al. Phase I of the
624 Surveillance for Enteric Fever in Asia Project (SEAP): An Overview and Lessons Learned. J
625 Infect Dis. 2018;218:S188–94.

626 14. Day MR, Doumith M, Do Nascimento V, Nair S, Ashton PM, Jenkins C, et al.
627 Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of *Salmonella*
628 *enterica* serovars Typhi and Paratyphi. Journal of Antimicrobial Chemotherapy.
629 2018;73:365–72.

630 15. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
631 large phylogenies. Bioinformatics. 2014;30:1312–3.

632 16. Britto CD, Dyson ZA, Mathias S, Bosco A, Dougan G, Jose S, et al. Persistent circulation
633 of a fluoroquinolone-resistant *Salmonella enterica* Typhi clone in the Indian subcontinent.
634 Journal of Antimicrobial Chemotherapy. 2020;75:337–41.

635 17. Sherchan JB, Morita M, Matono T, Izumiya H, Ohnishi M, Sherchand JB, et al.
636 Molecular and Clinical Epidemiology of *Salmonella* Paratyphi A Isolated from Patients with
637 Bacteremia in Nepal. The American Journal of Tropical Medicine and Hygiene. The
638 American Society of Tropical Medicine and Hygiene; 2017;97:1706–9.

639 18. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. Fast hierarchical Bayesian
640 analysis of population structure. Nucleic Acids Res. 2019;47:5539–49.

641 19. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees.
642 BMC Evolutionary Biology. 2007;7:214.

643 20. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, et al. ResFinder 4.0
644 for predictions of phenotypes from genotypes. Journal of Antimicrobial Chemotherapy.
645 2020;75:3491–500.

646 21. Carattoli A, Hasman H. PlasmidFinder and In Silico pMLST: Identification and Typing
647 of Plasmid Replicons in Whole-Genome Sequencing (WGS). Methods Mol Biol.
648 2020;2075:285–94.

649 22. Nair S, Day M, Godbole G, Saluja T, Langridge GC, Dallman TJ, et al. Genomic
650 surveillance detects *Salmonella enterica* serovar Paratyphi A harbouring blaCTX-M-15 from
651 a traveller returning from Bangladesh. PLOS ONE. Public Library of Science;
652 2020;15:e0228250.

653 23. Jibril AH, Okeke IN, Dalsgaard A, Menéndez VG, Olsen JE. Genomic Analysis of
654 Antimicrobial Resistance and Resistance Plasmids in *Salmonella* Serovars from Poultry in
655 Nigeria. Antibiotics. Multidisciplinary Digital Publishing Institute; 2021;10:99.

656 24. Hooda Y, Sajib MSI, Rahman H, Luby SP, Bondy-Denomy J, Santosham M, et al.
657 Molecular mechanism of azithromycin resistance among typhoidal Salmonella stains in
658 Bangladesh identified through passive pediatric surveillance. PLOS Neglected Tropical
659 Diseases. 2019;13:e0007868.

660 25. Sajib MSI, Tanmoy AM, Hooda Y, Rahman H, Andrews JR, Garrett DO, et al. Tracking
661 the Emergence of Azithromycin Resistance in Multiple Genotypes of Typhoidal Salmonella.
662 mBio [Internet]. American Society for Microbiology; 2021 [cited 2021 Feb 16];12. Available
663 from: https://mbio.asm.org/content/12/1/e03481-20

664 26. Sahastrabuddhe S, Carbis R, Wierzba TF, Ochiai RL. Increasing rates of *Salmonella*
665 Paratyphi A and the current status of its vaccine development. Expert Review of Vaccines.
666 Taylor & Francis; 2013;12:1021–31.

667 27. Seif Y, Monk JM, Machado H, Kavvas E, Palsson BO. Systems Biology and Pangenome
668 of *Salmonella* O-Antigens. mBio [Internet]. American Society for Microbiology; 2019 [cited
669 2020 Aug 11];10. Available from: https://mbio.asm.org/content/10/4/e01247-19

670 28. Wong VK, Baker S, Pickard DJ, Parkhill J, Page AJ, Feasey NA, et al. Phylogeographical
671 analysis of the dominant multidrug-resistant H58 clade of *Salmonella* Typhi identifies inter-
672 and intracontinental transmission events. Nat Genet. 2015;47:632–9.

673 29. Wong VK, Baker S, Connor TR, Pickard D, Page AJ, Dave J, et al. An extended
674 genotyping framework for *Salmonella enterica* serovar Typhi, the cause of human typhoid.
675 Nature Communications. 2016;7:12827.

676 30. Saha SK, Baqui AH, Hanif M, Darmstadt GL, Ruhulamin M, Nagatake T, et al. Typhoid
677 fever in Bangladesh: implications for vaccination policy. The Pediatric Infectious Disease
678 Journal. 2001;20:521–4.

679 31. McClelland M, Sanderson KE, Clifton SW, Latreille P, Porwollik S, Sabo A, et al.
680 Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of
681 *Salmonella enterica* that cause typhoid. Nature Genetics. Nature Publishing Group;
682 2004;36:1268–74.

683 32. Holt KE, Thomson NR, Wain J, Langridge GC, Hasan R, Bhutta ZA, et al. Pseudogene
684 accumulation in the evolutionary histories of *Salmonella enterica* serovars Paratyphi A and
685 Typhi. BMC Genomics. 2009;10:36.

686 33. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence
687 data. Bioinformatics. 2014;30:2114–20.

688 34. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome
689 assemblies from short and long sequencing reads. PLOS Computational Biology. Public
690 Library of Science; 2017;13:e1005595.

691 35. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics.
692 2014;30:2068–9.

693 36. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid
694 large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3.

695 37. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods.
696 Nature Publishing Group; 2012;9:357–9.

697 38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
698 Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

699 39. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid
700 phylogenetic analysis of large samples of recombinant bacterial whole genome sequences
701 using Gubbins. Nucleic Acids Research. 2015;43:e15–e15.

702 40. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new
703 developments. Nucleic Acids Research. 2019;47:W256–9.

704 41. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian
705 phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution
706 [Internet]. 2018 [cited 2021 Mar 28];4. Available from: https://doi.org/10.1093/ve/vey016

707

Figure legends

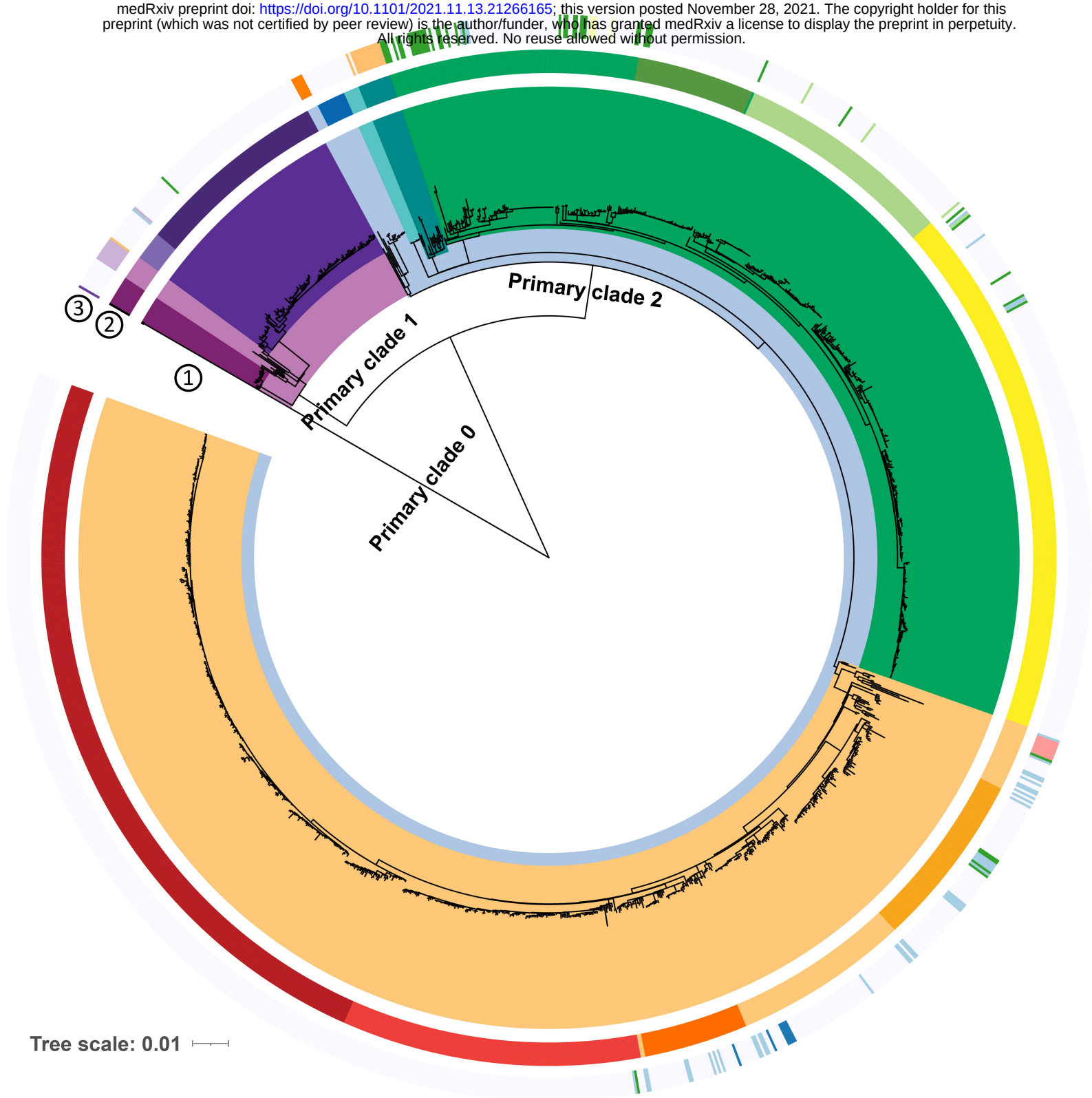
Figure 1: Genotyping scheme for *Salmonella* Paratyphi A. The scheme is composed of three primary, nine secondary and 18 genotypes on a phylogenetic tree of 1,379 isolates. The 9 secondary clades as highlighted by the coloring of the inner ring. 18 genotypes identified and are shown in the colored middle ring of the figure. The previously proposed lineage system is shown in the outer ring.

Figure 2: Maximum clade credibility tree of 315 representative *Salmonella* Paratyphi A isolates. The tree shows the last common ancestor of all *Salmonella* Paratyphi A existed at least 600 years ago (tMRCA - 1407 AD). The different genotypes are temporally resolved. Countries with greater than or equal to 5 isolates are also included.

Figure 3: Geographical distribution of *Salmonella* Paratyphi A genotypes. The country of isolation for 1378 sequenced *Salmonella* Paratyphi A isolates is shown. The distribution of genotypes per country is shown as scattered pie charts. The size of the each pie chart represents the number of sequences available. A difference in circulating genotypes is observed indicating local populations differ in several endemic countries. Further details are provided in Additional file 2.

Figure 4: Presence of antimicrobial resistance genes, plasmids, and chromosomal mutations linked to quinolone resistance across different *Salmonella* Paratyphi A genotypes. The diversity of **a)** Antimicrobial resistance genes **b)** Plasmids and **c)** quinolone resistance determining region (QRDR) mutations present *Salmonella* Paratyphi A is shown.

732



1 Secondary Clades

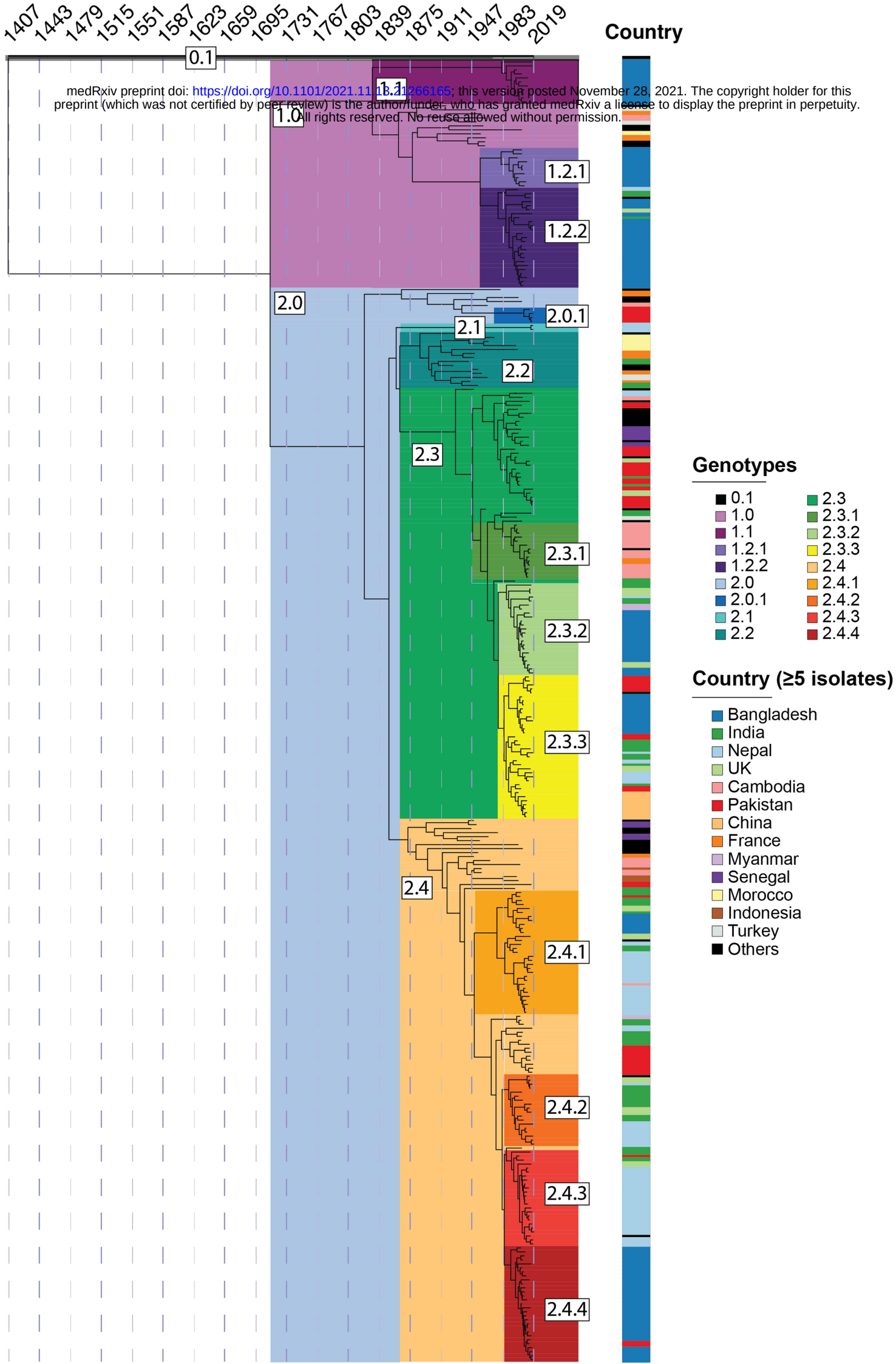
0.1	2.1	
1.0	2.2	
1.1	2.3	
1.2	2.4	
2.0		

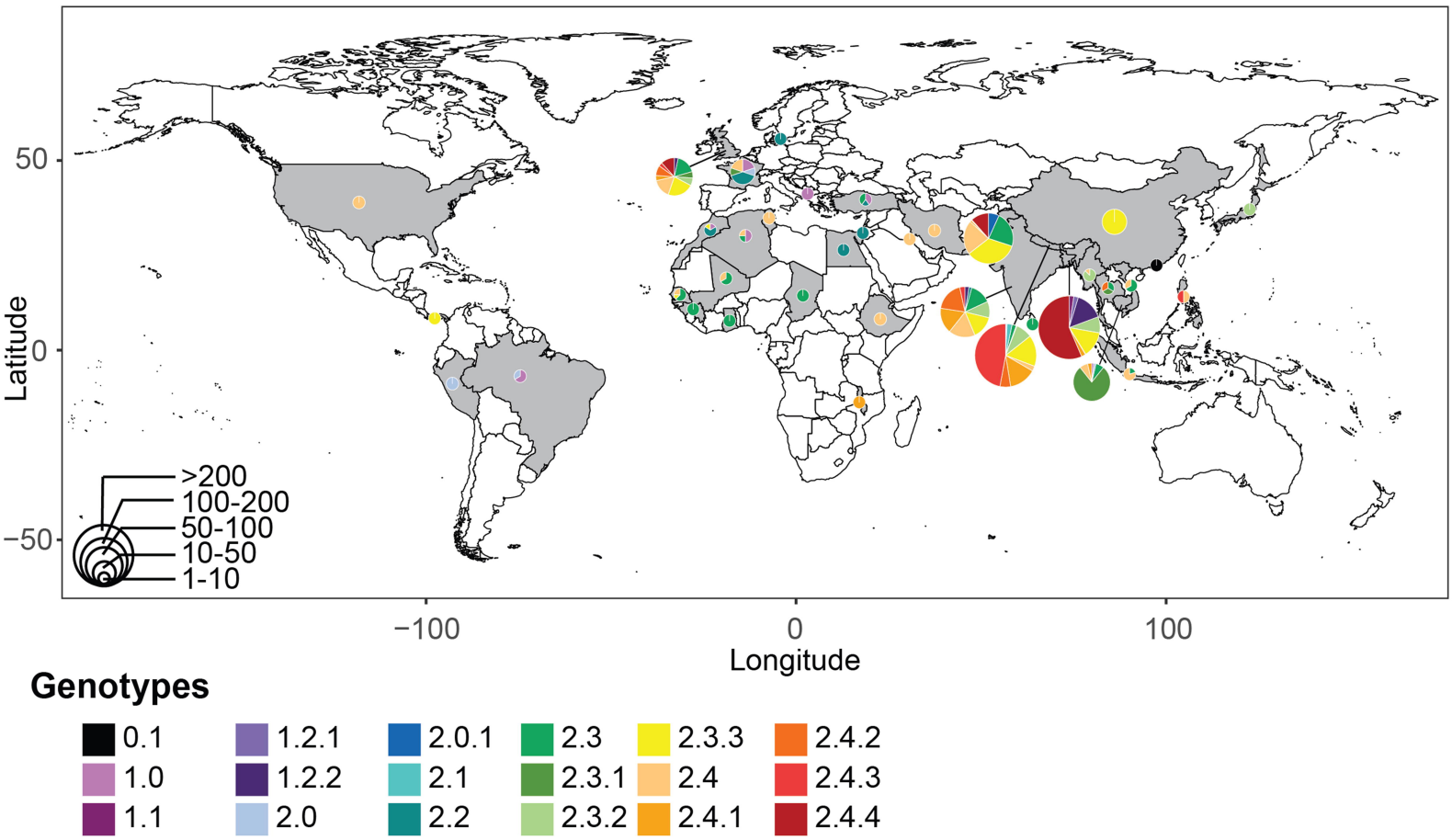
2 Genotypes

0.1	2.0	2.3.1	2.4.2
1.0	2.0.1	2.3.2	2.4.3
1.1	2.1	2.3.3	2.4.4
1.2.1	2.2	2.4	
1.2.2	2.3	2.4.1	

3 Lineages

A	C5	NA
A1	D	
B	E	
C	F	
C4	G	





Genotypes

0.1	1.2.1	2.0.1	2.3	2.3.3	2.4.2
1.0	1.2.2	2.1	2.3.1	2.4	2.4.3
1.1	2.0	2.2	2.3.2	2.4.1	2.4.4