2019-02-14

The speciation and hybridization history of the genus *Salmonella*

Author names

Alexis Criscuolo[1], Sylvie Issenhuth-Jeanjean[2], Xavier Didelot[3], Kaisa Thorell[4], James Hale[5], Julian Parkhill[6], Nicholas R. Thomson[6], François-Xavier Weill[2], Daniel Falush[7,*], Sylvain Brisse[8,*]

Affiliation

[1] Bioinformatics and Biostatistics Hub, C3BI, USR 3756 IP CNRS, Paris, France

[2] Institut Pasteur, Enteric Bacterial Pathogens, World Health Organization Collaborating Centre for Reference and Research on *Salmonella*, Paris, France

[3] School of Life Sciences and Department of Statistics, University of Warwick, Coventry, United Kingdom

4 Department of Microbiology and Immunology, University of Gothenburg, Gothenburg, Sweden

[5] Environmental Research Institute, University College Cork, Cork, Ireland

[6] Wellcome Trust Sanger Institute, Hinxton, United Kingdom

7 The Milner Centre for Evolution, University of Bath, Bath, United Kingdom

[8] Institut Pasteur, Biodiversity and Epidemiology of Bacterial Pathogens, Paris, France

*Corresponding authors

S. Brisse: Biodiversity and Epidemiology of Bacterial Pathogens, Institut Pasteur, 28 rue du Dr Roux, F-75724 Paris, France. E-mail: sylvain.brisse@pasteur.fr; Phone +33 1 45 68 83 34.

D. Falush: The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom. E-mail: danielfalush@gmail.com

Keywords: *Salmonella*; speciation; hybridization; evolution; taxonomy; genomics

Abstract

Bacteria and archaea make up most of natural diversity but the mechanisms that underlie the origin and maintenance of prokaryotic species are poorly understood. We investigated the speciation history of the genus *Salmonella*, an ecologically diverse bacterial lineage, within which *S. enterica* subsp. *enterica* is responsible for important human food-borne infections. We performed a survey of diversity across a large reference collection using multilocus sequence typing, followed by genome sequencing of distinct lineages. We identified eleven distinct phylogroups, three of which were previously undescribed. Strains assigned to *S. enterica* subsp. *salamae* are polyphyletic, with two distinct lineages that we designate Salamae A and Salamae B. Strains of subspecies *houtenae* are subdivided into two groups, Houtenae A and B and are both related to Selander's group VII. A phylogroup we designate VIII was previously unknown. A simple binary fission model of speciation cannot explain observed patterns of sequence diversity. In the recent past, there have been large scale hybridization events involving an unsampled ancestral lineage and three distantly related lineages of the genus that have given rise to Houtenae A, Houtenae B and VII. We found no evidence for ongoing hybridization in the other eight lineages but detected more subtle signals of ancient recombination events. We are unable to fully resolve the speciation history of the genus, which might have involved additional speciation-by-hybridization or multi-way speciation events. Our results imply that traditional models of speciation by binary fission and divergence may not apply in *Salmonella*.

Data summary

Illumina sequence data were submitted to the European Nucleotide Archive under project number PRJEB2099 and are available from INSDC (NCBI/ENA/DDBJ) under accession numbers ERS011101 to ERS011146. The MLST sequence and profile data generated in this study have been publicly available on the *Salmonella* MLST web site between 2010 and the migration of the *Salmonella* MLST website to EnteroBase (https://enterobase.warwick.ac.uk/), and subsequently from there.

Introduction

Bacteria and archaea make up most of natural diversity, both in terms of species richness and biological functions [1,2]. However, the mechanisms that underlie the origin and maintenance of prokaryotic species are poorly understood. It is often assumed that there is a single phylogenetic tree representing the relationships amongst prokaryotic taxa, with the branch lengths reflecting divergence times between them. However, bacteria and archaea acquire foreign DNA by homologous and non-homologous recombination and can recombine frequently, including in the *Salmonella* genus [3–9]. High recombination rates can maintain genetic cohesion within a species, preventing divergence and speciation from occurring until barriers to gene flow develop. Recombination has been shown in laboratory experiments to be supressed by nucleotide mismatches between donor and recipient [10,11]. This property provides a potential mechanism for speciation. It has been shown by simulation that large effective population sizes and neutral genetic drift can precipitate speciation by increasing the average pairwise divergence between strains, leading to either binary or multi-way speciation events [5,12].

Conversely, distinct new lineages or species can potentially arise almost instantaneously by hybridization of existing distantly related ones. Such large-scale hybridization events can occur at once by recombination of large genomic regions (e.g., [13]), or through multiple exchanges of small chromosomal segments associated with ecological convergence [14]. Therefore, to describe relationships between prokaryotes and understand patterns of species richness and phenotypic diversity, it is important to characterise the process of speciation and gene flow between species, including large-scale hybridization events [15].

Salmonellae are a prominent speciation model, where experimental and genomic studies of recombination and hybridization have been pioneered [4-10,14]. The genus *Salmonella* is divided into a number of phylogroups, namely *bongori*, *enterica*, *salamae*, *arizonae*, *diarizonae*, *houtenae*, and *indica* [16–18]. *Salmonella bongori* has been classified a distinct species [18], while the other phylogroups are considered to be subspecies of a single species, *S. enterica*. These taxa are further subdivided into serovars based on antigenic variation of flagellins and O-antigen.

Members of the genus *Salmonella* are major pathogens of humans and other warm-blooded animals. Human infections mostly involve *S. enterica* subspecies *enterica*, which can cause gastroenteritis, enteric fever and other infections [19,20]. Other *S. enterica* subspecies, as well

as the species *S. bongori*, are more typically isolated from cold blooded animals or the environment, and are rarely reported from human infections [21].

Here we are concerned with evolutionary relationships rather than taxonomy and we designate phylogroups by names that derived from these subspecies' labels, e.g. Bongori, Arizonae, Diarizonae, etc., with Enterica representing subspecies *enterica*. We use italicised names such as *houtenae* to refer to previous subspecies designations, which sometimes differ from our phylogroup assignments. A seventh *S. enterica* subgroup (group VII) was distinguished based on multilocus enzyme electrophoresis and gene sequencing [22–24]. Note that phylogenetic re-evaluation [25] of the proposed species *Salmonella subterranea* [26] shows that it does not belong to the *Salmonella* genus.

Phylogenetic analyses of the evolutionary relationships amongst the different *Salmonella* lineages have led to contradictory conclusions with several proposed phylogenetic trees [9,23,24,27–37]. This lack of consensus might reflect technical issues with phylogenetic reconstruction but a more biologically interesting possibility is that the history of *Salmonella* is not well-characterized by a simple model in which speciation proceeds stepwise by irreversible binary fissions.

To test this hypothesis, we sampled the genetic diversity within the little studied groups from cold-blooded hosts and used whole genome sequences from representative isolates of phylogroups to characterize the genetic relationships between them and to infer historical populations splits and gene flow. We show that while a binary fission model of speciation works for some of the *Salmonella* lineages, there are several important historical events that cannot be characterized in this way.

Methods

Taxonomic sampling and MLST analyses

A total of 367 strains (**Table S1**) from outside the subspecies enterica were selected from the collection of the World Health Organization Collaborative Centre for Reference and Research on *Salmonella*, Institut Pasteur, Paris, France. This center contains the reference strains of all *Salmonella* serovars and their variants. The 367 strains represented approximately one third of currently described serovars outside enterica and were selected to maximize the diversity of antigenic formulae. MLST was performed on these strains using updated primers adapted from those of Kidgell et al. [38] to amplify DNA from *S. bongori* and all subspecies of *S. enterica*. The novel primers are described in **Table S3**; note that they have been publicly available on the MLST web site between 2008 and the migration of the *Salmonella* MLST website to EnteroBase, and subsequently from there.

A phylogenetic tree was inferred from the median distance matrix of the seven genes with the algorithm BioNJ* [39]. A supermatrix of characters was built by concatenating the seven MSAs with the program Concatenate (www.supertriplets.univ-montp2.fr/PhyloTools.php), and the nucleotide diversity of groups was defined using the index π [40] with the program DnaSP [41]. Minimum spanning trees were built using the software tool BioNumerics (Applied-Maths, Belgium).

Strain selection and genome sequencing

A set of 46 strains were selected for whole genome sequencing (**Table S2**). Genome sequencing was achieved by Illumina 2 x 50 nt paired-end sequencing for all strains. The characteristics of the obtained *de novo* assemblies are summarised in **Table S2**. This set was completed with genome sequences gathered from the GenBank repository (*i.e.*, 16 *S. enterica subsp. enterica*, 1 *S. enterica subsp. arizonae*, and 1 *S. bongori* strains), as well as 9 *S. bongori* genome sequences from Fookes et al. [34]. This led to a total of 73 genomes (**Table S2**): *S. enterica* subsp. *enterica*, 16; subsp. *salamae*, 13; subsp. *arizonae*, 9; subsp. *diarizonae*, 10; subsp. *houtenae*, 6; subsp. *indica*, 4; *S. bongori*, 10; and VII, 2.

Core gene construction

Each of the 4,423 protein sequences from *S. enterica* strain LT2 [42] was used as query to perform BLAST similarity searches [43] against the genome sequence of each of the other 72 strains. Clusters of homologous sequences were built by considering only the first tblastn hit (E-value < 10^{-5}), and every cluster that did not contain 73 sequences (*i.e.*, one per strain) was discarded. Next, orthology was assessed within each cluster by performing reciprocal tblastn, leading to 2,328 clusters of putative orthologous coding sequences from the core gene set of the 73 strains. For each of these clusters, sequences were translated, and a multiple sequence alignment (MSA) was performed with ProbCons [41] and next back-translated to obtain a codon-level MSA. The 2,328 MSAs were concatenated into a supermatrix of 2,137,446 nucleotide characters that was used to infer a balanced minimum-evolution phylogenetic tree using FastME [44] with pairwise p-distances. Branch support was assessed for each internal branch with a MSA-based bootstrap procedure based on 1,000 replicates. This procedure samples the MSA with replacement according to the same procedure as the standard bootstrap with nucleotide characters.

Recombination analysis

We applied four separate and complementary methods to analyse the ancestral recombination events that occurred during the evolution of the genus *Salmonella*. Firstly, we applied chromosome painting on the 73 genomes, using CHROMOPAINTER [45] to reconstruct each genome as a mosaic of all the others. The results were summarized as a heatmap of coancestry values, where each coancestry value corresponds to the number of fragments copied from one genome to another (**Figure 2**)**.** Secondly, we performed pairwise comparisons of genomes using a gene-by-gene approach. For each pair of genomes, we computed the genetic distances of all shared genes, and the distribution of these distances was plotted as a cumulative curve (**Figure 3**). Thirdly, the CHROMOPAINTER analysis was repeated using only nine unrelated genomes: one for each of the 12 phylogroups but excluding VII and Houtenae B due to recent shared ancestry with Houtenae A, and considering Enterica A and B as a single group. Each genome was therefore reconstructed as a mosaic of the other eight unrelated genomes. This allowed us to explore deeper relationships between phylogroups, since when all genomes are included each genome from a phylogroup copies mostly from other genomes of the same phylogroup (**Figure 2**). The resulting coancestry matrix was plotted as a heatmap (**Figure 4**). Fourthly, we applied the Treemix algorithm with parameter K=3 [46] to one genome from

each of the 12 phylogroups in order to reconstruct their relationships as a vertical phylogenetic tree augmented with horizontal transfer events (**Figure 5**).

Pan-genome analyses

Analysis of accessory genome was performed using the Roary pan-genome pipeline version 3.6.2 [47]. Since the draft genomes were very unequally fragmented and synteny information therefore was of variable reliability we used the “don't-split-paralogs” option. The analysis was performed with a protein identity cut-off of 85% and the core genome was defined as genes present in > 99% of the genomes studied. The Pearson correlation between accessory gene content of the genomes were visualised using the R software CORRPLOT package [48].

Results and Discussion

In order to survey the diversity of *Salmonella* outside *S. enterica* subsp. *enterica*, a total of 367 strains, comprising about a third of the known non-Enterica serovars, were selected from the World Health Organization Collaborating Centre for Reference and Research on *Salmonella* (Institut Pasteur, Paris, France) reference collection and subjected to multilocus sequence typing (MLST) (**Tables S1**, **S2**). A phylogenetic tree was built (**Figure S1**), revealing a novel group (labelled VIII) and suggesting a polyphyletic origin of *S. enterica* subsp. *salamae* (Salamae A and B) and of *S. enterica* subsp. *houtenae* (Houtenae A and B). Within-phylogroup nucleotide diversity (**Figure S1 inset**) was the highest in Arizonae (π = 1.6%), lowest in Houtenae groups, Bongori, Salamae B and Diarizonae (π ranging from 0.35% to 0.42%), whereas it was intermediate in Salamae A, Indica and Enterica. Minimum spanning tree analysis of MLST profiles illustrates the genotypic diversity within each group (**Figure S2**).

Based on MLST diversity, 46 genomes were chosen for genome sequencing and compared to 27 previously published genome sequences of Enterica, Arizonae, and Bongori (**Table S2**). A phylogenetic tree based on the genome sequences is shown in **Figure 1**. This tree implies that *S. enterica* subsp. *salamae* is not a monophyletic group but instead forms two lineages with distinct evolutionary histories that we designate Salamae A and Salamae B. Whereas Salamae A contained 138 (88%) of the *salamae* strains, Salamae B comprised 18 isolates collected from a human (one isolate), a bat (one isolate) or reptiles (16 isolates, including 6 from chameleons). In contrast, 49 (41.5%) of Salamae A isolates were from humans, and only 34 (28.8%) were from cold-blooded animals, suggesting important ecological and pathogenic differences between the two Salamae groups. *S. enterica* subsp. *houtenae* was also subdivided into two distinct phylogroups, which we have designated Houtenae A and Houtenae B, and which clustered together with group VII on the tree. The genome-wide phylogenetic analysis also uncovers a hitherto unknown phylogroup, labelled VIII, made of strains previously identified as either salamae, diarizonae or of the former Hisingen serotype of *S. enterica* subsp. *enterica* [25]. The description of Salamae B, Houtenae B and VIII represent the first novel *Salmonella* phylogroups described since the identification of group VII by Selander and colleagues more than 25 years ago [24,27]. Our analysis therefore defines 11 phylogroups within *Salmonella*. The phylogenetic tree also shows further subdivisions at shallower levels, including the division of *S. enterica* subsp. *enterica* into Enterica A and Enterica B as previously described [5,9]. Note that the genomes of the present study have been publicly

available from International Nucleotide Sequence Database Collaboration (INSDC) since 2011, and were used in a genome-based phylogenetic analysis of *Salmonella* by Alikhan *et al*. [49]; the three novel *Salmonella* groups were labelled as novel subspecies A (Houtenae B), B (VIII) and C (Salamae B) in [49].

Recent recombination between phylogroups.

We used chromosome painting of the above set of 73 strains in order to investigate shared ancestry and recombination events between different phylogroups. Specifically, the CHROMOPAINTER algorithm uses a Hidden Markov Model to reconstruct each isolate as a mosaic of stretches of DNA of the other isolates in the sample [45]. The results can be summarized as a heatmap indicating how many stretches from each other sample are used in the reconstruction. The organism used in the reconstruction is assumed to be the most closely related for each stretch of DNA. **Figure 2** shows a heatmap illustrating the proportion of DNA used to paint each isolate across the genome, with dark blue corresponding to 0% and dark red corresponding to 10%. We call this proportion the coancestry value. Each phylogroup shows higher coancestry within the same phylogroup than with others. The highest coancestry between strains in different phylogroups is between Houtenae A, Houtenae B and VII. However, Houtenae B shows higher Enterica ancestry (particularly with Enterica B) than do Houtenae A or VII. The two deepest branching Salamae A strains show high levels of coancestry with several other groups including Salamae B, Diarizonae, Indica and VIII. One strain of Enterica A (SL483) is exceptional in showing higher coancestry levels with Enterica B.

In order to test whether high coancestry between groups might be explained by recent recombination between them, we looked for evidence of sharing of very similar stretches of DNA between pairs of lineages [14] by plotting, for each pairwise comparison, the proportion of genes with divergence below a threshold increasing from 0% to 25% (**Figure 3**). Consistent with recent recombination between them, Enterica B and Houtenae B showed many more genes with very similar sequences than expected based on their position in the phylogenetic tree, with 20% of the genes of an Enterica B strain having less than 1% divergence to Houtenae B, compared to only 5% between Enterica B and Houtenae A (**Figure 3A**). These divergence curves are also consistent with recent recombination between Houtenae A, Houtenae B and VII. For example, approximately 5% of the VII genome and 6% of Houtenae A has less than 0.1% divergence with Houtenae B (**Figure 3B**), suggesting that

there has been very recent recombination between these three phylogroups. There is no analogous signal of recent recombination between any of the strains of Salamae A or Salamae B with each other or with other phylogroups based on cumulative divergence curves (*e.g.*, **Figure 3C**). The smudged pattern of coancestry of the deeper branching Salamae A and Salamae B strains in **Figure 2** can potentially be explained by them retaining ancestral variants that have been lost by the rest of the phylogroup and therefore does not necessarily indicate recent recombination between lineages. **Figure 3D** illustrates the absence of any signal of recent recombination with Arizonae.

Evidence for hybridization in the origin of the phylogroups

We next examined the origins of the phylogroups themselves. Recombination events which predate the generation of the diversity observed *within* each phylogroup are unlikely to be picked up in the chromosome painting analysis in **Figure 2**: members of a phylogroup that have inherited the same ancestrally imported stretch will be painted by each other for those stretches. Therefore, we selected a single strain from each phylogroup and performed a distinct chromosome painting analysis. We excluded VII and Houtenae B due to the recent shared ancestry with Houtenae A, and also included only a single representative for both Enterica A and Enterica B. The chromosome painting results (**Figure 4**) show high coancestry between Bongori and Arizonae and between Indica and Enterica. These relationships can be interpreted using a vertical phylogenetic model, as they agree with a large number of different analyses including ours (**Figure 1)** that Arizonae is the deepest branching lineage after Bongori and that Indica is a sister group of Enterica [9,24,33,34]. Conversely, the chromosome painting analysis reveals a large number of intransitive relationships (*i.e.*, in which A has elevated coancestry with B and B has high coancestry with C but C does not have high coancestry with A). First, Diarizonae and Arizonae have high coancestry, as do Diarizonae and Salamae B but Salamae B and Arizonae do not (**Figure 4**). Second, Houtenae A and Salamae A have high coancestry with each other and the phylogenetic tree suggests they are sister taxa. However, they have different relationships to other phylogroups. Houtenae A, but not Salamae A, shows high coancestry with Arizonae, while Salamae A shows higher shared ancestry with Indica and Enterica. Intransitive patterns of coancestry are also evident for VIII, Salamae B and Diarizonae and for VIII, Salamae B and Bongori. An intransitive pattern is not predicted by any phylogenetic model and is indicative of mixture in

the history. These observations suggest a complex pattern of homologous recombination events that predate diversification within phylogroups.

A scenario involving three recent hybridization events

To complement the results above, we used Treemix to infer a history that allows for recombination events in the origins of the phylogroups. Treemix attempts to model the covariance matrix reflecting SNP sharing between strains by assuming a phylogenetic model of divergence via genetic drift, but with a limited number *K* of mixing events in the history. Our application of Treemix to *Salmonella* gave results which varied in important details depending on the value of K. Each of the events that were identified at a given value of K had counterparts in the inference performed for higher values, but details of the inferred phylogenetic tree and the location and direction of the hybridization events were not consistent. For example, for K=1 and K=2 Houtenae A and Houtenae B are sister taxa whose common ancestor received genetic material from VII, while for K=3, VII and Houtenae B share a common ancestor, which contributed genetic material to Houtenae A. We present the Treemix results for K=3 (**Figure 5**) because all of the events inferred are supported by signals identified by chromosome painting and cumulative divergence (**Figures 2, 3 and 4**). The Treemix results with K=3 imply that Houtenae A, Houtenae B and VII all have hybrid origins. All three of them received DNA from a shared lineage which branched between Arizonae and Diarizonae (black arrowhead, **Figure 5**), but differ in the remaining source of their ancestry (red arrows, **Figure 5**), which, according to the Treemix estimates, account for about half of their genome in all three cases (1: ancestor of Arizonae to VII: 0.461; 2: Enterica B to Houtenae B: 0.42; 3: ancestor of VII to houtenae A: 0.49). Note that according to this reconstruction, no pure, or nearly pure, representative of this shared ancestral lineage is present in the sample, a feature which is likely to have contributed largely to the instability of the Treemix analysis and makes all types of evolutionary reconstruction considerably more challenging.

The second source for Houtenae B is inferred to be Enterica B (red arrow 2, **Figure 5**), which is consistent with the results from chromosome painting and of the pairwise distances, as discussed above, and is consistent with recent genetic exchange having taken place. The second source for VII is inferred to branch at the same point as Arizonae does in the tree. The deep position of this ancestry source is supported by the distribution of pairwise distances VII has to shallower branching lineages such as Diarizonae or Salamae A, which are more similar

to the distribution found for Arizonae than to that of either Houtenae A or Houtenae B (*e.g.*, **Figure 3C**). The distribution of distances to Arizonae is similar to that of other shallow-branching lineages, suggesting that the recombination was not with Arizonae itself. Finally, the second source for Houtenae A branches next to Salamae A, which is consistent with the reconstructed position of Houtenae A in the phylogenetic tree in Figure 1 and the high coancestry of Houtenae A and Salamae A in Figure 4. However, unlike for Houtenae B, there is no signal of recent recombination of Houtenae A with other lineages in Figure 2. Furthermore, the pairwise distance curves of Salamae A to Houtenae A and Houtenae B are comparable (**Figure 3C**). These features imply that there has not been recent recombination between Houtenae A and Salamae A. Instead, they are consistent with the second source that contributed to Houtenae A being an unsampled sister taxa to Salamae A.

Unequal evolutionary rates of the different taxa

One important feature of the phylogenetic tree (**Figure 1**) is the different branch lengths leading to each phylogroup. This feature might be caused by either unequal substitution rates between lineages or by recombination, which can cause hybrid lineages to branch closer to the root. Evidence for unequal substitution rates comes for example from comparisons with Bongori or Arizonae, which can tentatively be treated as outgroups. Salamae A and Salamae B have smaller genetic distances than other lineages to either (**Figure 3D**), despite the chromosome painting indicating no evidence of elevated recombination between them. Furthermore, Salamae A and Salamae B show low genetic distances compared to potential sister lineages to all taxa, suggesting that they have substantially lower substitution rates than other groups. Because our reconstruction of *Salmonella* evolutionary history is incomplete and uncertain, we do not attempt to formally model all of these processes occurring together.

Accessory genome relationships

Accessory genes contribute most to ecological specialization and the pattern of horizontal gene transfer among phylogroups might provide important complementary information regarding functional and ecological correlates of the recombination history that we inferred in this work [9]. We therefore analysed the pan-genome of the dataset, which with a protein identity cut-off of 85% rendered a core genome of 1818 gene clusters and a total pan-genome of 21973 genes. Unfortunately, estimations of the strain relationships based on gene

presence/absence and analysis of the shared ancestry revealed that the analyses were strongly affected by the fragmentation of the genomic assemblies (**Table S2**), as was particularly visible for the highly fragmented Diarizonae genomes (**Figure S3)**. Analysis of the horizontal gene transfer pattern among phylogroups therefore requires higher quality assemblies and will be the subject of future studies.

Conclusions

We investigated the diversification and hybridization history within *Salmonella*, a group of prominent public health importance and an early model for microbial speciation and evolutionary studies. By sampling largely in the non-enterica subspecies, we uncovered three novel phylogenetic groups that had not been recognized since the last group, VII, was described in 1991. Our snapshot of diversity within phylogroups of *Salmonella* implies that recombination among phylogroups is relatively rare at any point in time but that when it happens it can be with distantly related lineages rather than sister taxa and can involve large fractions of the core genome. These events are likely to provide substantial potential for phenotypic innovation but may also entail a great deal of hybrid disgenesis.

The three hybridization events that we have been able to elucidate with any degree of certainty are ongoing or took place in the recent past and all involved a lineage that is not present in unhybridized form in the dataset. This circumstance makes it challenging to estimate simple properties of the events such as the direction of hybridization and the proportion of genome acquired from each source. We can nevertheless robustly conclude that the hybridization has involved at least three entirely different branches of the *Salmonella* tree and has led to the formation of three phylogroups, namely Houtenae A, Houtenae B and VII. Interestingly the latter group was inferred to be a ‘hybrid’ in early MLEE studies [24]. This suggests an interesting question that is likely to be informative about the general nature of species boundaries in bacteria, namely what has happened to make one lineage particularly prone to hybridization in the recent past?

We see less conclusive but nevertheless still strong evidence for hybridization events in the more distant past. Phylogenetic trees of *Salmonella* phylogroups are notoriously unstable, including in different analyses we have performed (data not shown). In particular,

relationships amongst Salamae A, Salamae B, Diarizonae, Enterica and VIII are difficult to elucidate. The coancestry relationships between these lineages are highly intransitive (**Figure 4**). One possibility is that this intransitivity is due to a complex multi-way speciation event [5], such that there is no true splitting order to infer. However, it may also represent hybridization events after stepwise speciation. The two lineages that branch deeply (**Figure 1**), namely VIII and Diarizonae, both show evidence of shared ancestry with basal lineages, Bongori and Arizonae, respectively (**Figure 4**), which is likely to have substantially affected their phylogenetic position.

The events of recombination inferred in this work explain the difficulties to reconstruct the phylogeny of the genus that have led to multiple distinct hypotheses on the phylogenetic relationships among subspecies. The phylogenetic relationships which do appear to be reasonably certain are that Bongori split from the other phylogroups first, followed by Arizonae and that Indica is a sister group to Enterica. Houtenae A seems to have been a sister taxon of Salamae A, prior to its mixture event. These examples demonstrate that in the right circumstances, phylogenetic signal can be preserved over long evolutionary time periods despite recombination between phylogroups. The problem of reconstructing ancestral hybridization events is a hard one and we do not have the tools or genomes available to reconstruct an entire history with any degree of confidence.

Our results demonstrate that bacterial species histories are complex. There is considerable phylogenetic signal in the data, consistent with the evolution and long-term persistence barriers to gene flow between lineages but also examples for hybridization events that may reverse species boundaries, sometimes between taxa separated by large genetic distances, rather than between sister taxa. These results mean that phylogenetic trees displaying relationships between species will often represent considerable simplifications of evolutionary history and in the worst case can be entirely misleading. Further work in multiple taxa will elucidate the evolutionary and ecological factors that precipitate speciation and hybridization events.

Author contributions

Conceptualization: SB, DF. Supervision: **SB, DF, NRT, FXW.** ***Performed the experiments:*** **SIJ.** ***Data curation:*** **JH, SB, SIJ, FXW.** ***Data analysis:*** **AC, SB, XD, KT.** ***Writing – original draft:*** **AC, SB, DF.** ***Writing – review & editing:*** **all.**

Conflicts of interest

The author(s) declare that there are no conflicts of interest.

Funding information

This work was supported financially by a grant from Region Ile De France to AC and by a Walton Visiting Scientist grant from the Science Foundation of Ireland to SB.

Acknowledgements

We acknowledge the expert help of Laure Diancourt, Coralie Tran and Virginie Passet (Institut Pasteur) for MLST data production, and of Mark Achtman for his support at the start of the project and for bioinformatics assistance in MLST data curation.

References

1. Hugenholtz P, Skarshewski A, Parks DH. Genome-Based Microbial Taxonomy Coming of Age. Cold Spring Harb Perspect Biol. **2016**; 8(6).
2. Adam PS, Borrel G, Brochier-Armanet C, Gribaldo S. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. ISME J. **2017**; 11(11):2407–2425.
3. Brown EW, Mammel MK, LeClerc JE, Cebula TA. Limited boundaries for extensive horizontal gene transfer among Salmonella pathogens. Proc Natl Acad Sci USA. **2003**; 100(26):15676–15681.
4. Octavia S, Lan R. Frequent recombination and low level of clonality within Salmonella enterica subspecies I. Microbiology (Reading, Engl). **2006**; 152(Pt 4):1099–1108.
5. Falush D, Torpdahl M, Didelot X, Conrad DF, Wilson DJ, Achtman M. Mismatch induced speciation in Salmonella: model and data. Philosophical transactions of the Royal Society of London. **2006**; 361(1475):2045–53.
6. Sangal V, Harbottle H, Mazzoni CJ, et al. Evolution and population structure of Salmonella enterica serovar Newport. Journal of bacteriology. **2010**; 192(24):6465–76.
7. Didelot X, Bowden R, Street T, et al. Recombination and population structure in Salmonella enterica. PLoS Genet. **2011**; 7(7):e1002191.

8. Achtman M, Wain J, Weill F-X, et al. Multilocus Sequence Typing as a Replacement for Serotyping in Salmonella enterica. PLoS Pathog. **2012**; 8(6):e1002776.
9. Desai PT, Porwollik S, Long F, et al. Evolutionary Genomics of Salmonella enterica Subspecies. MBio. **2013**; 4(2).
10. Zahrt TC, Maloy S. Barriers to recombination between closely related bacteria: MutS and RecBCD inhibit recombination between Salmonella typhimurium and Salmonella typhi. Proc Natl Acad Sci USA. **1997**; 94(18):9786–9791.
11. Vulić M, Dionisio F, Taddei F, Radman M. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci USA. **1997**; 94(18):9763–9767.
12. Hanage WP, Fraser C, Spratt BG. The impact of homologous recombination on the generation of diversity in bacteria. Journal of theoretical biology. **2006**; 239(2):210–9.
13. Chen L, Mathema B, Pitout JDD, DeLeo FR, Kreiswirth BN. Epidemic Klebsiella pneumoniae ST258 is a hybrid strain. MBio. **2014**; 5(3):e01355-01314.
14. Didelot X, Achtman M, Parkhill J, Thomson NR, Falush D. A bimodal pattern of relatedness between the Salmonella Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome research. **2007**; 17(1):61–8.
15. Sheppard SK, McCarthy ND, Falush D, Maiden MCJ. Convergence of Campylobacter species: implications for bacterial evolution. Science. **2008**; 320(5873):237–239.
16. Brenner FW, Villar RG, Angulo FJ, Tauxe R, Swaminathan B. Salmonella nomenclature. J Clin Microbiol. **2000**; 38(7):2465–2467.
17. Tindall BJ, Grimont PA, Garrity GM, Euzeby JP. Nomenclature and taxonomy of the genus Salmonella. Int J Syst Evol Microbiol. **2005**; 55(Pt 1):521–4.
18. Reeves MW, Evins GM, Heiba AA, Plikaytis BD, Farmer JJ. Clonal nature of Salmonella typhi and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis, and proposal of Salmonella bongori comb. nov. J Clin Microbiol. **1989**; 27(2):313–320.
19. Parry CM, Hien TT, Dougan G, White NJ, Farrar JJ. Typhoid fever. N Engl J Med. **2002**; 347(22):1770–1782.
20. Sánchez-Vargas FM, Abu-El-Haija MA, Gómez-Duarte OG. Salmonella infections: an update on epidemiology, management, and prevention. Travel Med Infect Dis. **2011**; 9(6):263–277.
21. Lamas A, Miranda JM, Regal P, Vázquez B, Franco CM, Cepeda A. A comprehensive review of non-enterica subspecies of Salmonella enterica. Microbiol Res. **2018**; 206:60–73.
22. Selander R, Beltran P, Smith N. Evolutionary genetics of Salmonella. Evolution at the molecular level. Selander RK, Clark AG, Whittam TS, editors; 1991. p. 25–57.
23. Nelson K, Selander RK. Evolutionary genetics of the proline permease gene (putP) and the control region of the proline utilization operon in populations of Salmonella and Escherichia coli. J Bacteriol. **1992**; 174(21):6886–6895.
24. Boyd EF, Wang FS, Whittam TS, Selander RK. Molecular genetic relationships of the salmonellae. Applied and environmental microbiology. **1996**; 62(3):804–8.
25. Guibourdenche M, Roggentin P, Mikoleit M, et al. Supplement 2003-2007 (No. 47) to the White-Kauffmann-Le Minor scheme. Research in microbiology. **2010**; 161(1):26–9.
26. Shelobolina ES, Sullivan SA, O'Neill KR, Nevin KP, Lovley DR. Isolation, characterization, and U(VI)-reducing potential of a facultatively anaerobic, acid-resistant Bacterium from Low-pH, nitrate- and U(VI)-contaminated subsurface sediment and description of Salmonella subterranea sp. nov. Appl Environ Microbiol. **2004**; 70(5):2959–2965.

27. Nelson K, Whittam TS, Selander RK. Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (gapA) in natural populations of Salmonella and Escherichia coli. Proc Natl Acad Sci U S A. **1991**; 88(15):6667–71.
28. Thampapillai G, Lan R, Reeves PR. Molecular evolution in the gnd locus of Salmonella enterica. Mol Biol Evol. **1994**; 11(6):813–28.
29. Christensen H, Nordentoft S, Olsen JE. Phylogenetic relationships of Salmonella based on rRNA sequences. Int J Syst Bacteriol. **1998**; 48 Pt 2:605–610.
30. Brown EW, Kotewicz ML, Cebula TA. Detection of recombination among Salmonella enterica strains using the incongruence length difference test. Mol Phylogenet Evol. **2002**; 24(1):102–120.
31. Whittam TS, Bumbaugh AC. Inferences from whole-genome sequences of bacterial pathogens. Curr Opin Genet Dev. **2002**; 12(6):719–725.
32. Porwollik S, Wong RM-Y, McClelland M. Evolutionary genomics of Salmonella: gene acquisitions revealed by microarray analysis. Proc Natl Acad Sci USA. **2002**; 99(13):8956–8961.
33. McQuiston JR, Herrera-Leon S, Wertheim BC, et al. Molecular phylogeny of the salmonellae: relationships among Salmonella species and subspecies determined from four housekeeping genes and evidence of lateral gene transfer events. Journal of bacteriology. **2008**; 190(21):7060–7.
34. Fookes M, Schroeder GN, Langridge GC, et al. Salmonella bongori provides insights into the evolution of the Salmonellae. PLoS Pathog. **2011**; 7(8):e1002191.
35. Trujillo S, Keys CE, Brown EW. Evaluation of the taxonomic utility of six-enzyme pulsed-field gel electrophoresis in reconstructing Salmonella subspecies phylogeny. Infect Genet Evol. **2011**; 11(1):92–102.
36. Pettengill JB, Timme RE, Barrangou R, et al. The evolutionary history and diagnostic utility of the CRISPR-Cas system within Salmonella enterica ssp. enterica. PeerJ. **2014**; 2:e340.
37. Kisiela DI, Chattopadhyay S, Libby SJ, et al. Evolution of Salmonella enterica virulence via point mutations in the fimbrial adhesin. PLoS Pathog. **2012**; 8(6):e1002733.
38. Kidgell C, Reichard U, Wain J, et al. Salmonella typhi, the causative agent of typhoid fever, is approximately 50,000 years old. Infect Genet Evol. **2002**; 2(1):39–45.
39. Criscuolo A, Gascuel O. Fast NJ-like algorithms to deal with incomplete distance matrices. BMC Bioinformatics. **2008**; 9:166.
40. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA. **1979**; 76(10):5269–5273.
41. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. **2009**; 25(11):1451–2.
42. McClelland M, Sanderson KE, Spieth J, et al. Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. Nature. **2001**; 413(6858):852–6.
43. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. **1997**; 25(17):3389–402.
44. Lefort V, Desper R, Gascuel O. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Mol Biol Evol. **2015**; 32(10):2798–2800.
45. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. **2012**; 8(1):e1002453.
46. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. **2012**; 8(11):e1002967.
47. Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. **2015**; 31(22):3691–3693.

48. Wei, T, Simko V. R package "corrplot": Visualization of a Correlation Matrix (Version 0.84) [Internet]. 2017. Available from: https://github.com/taiyun/corrplot

49. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of Salmonella. PLoS Genet. **2018**; 14(4):e1007261.

Figures

Figure 1. Phylogenetic tree of 73 *Salmonella* strains based on all shared core genes.

The balanced minimum-evolution phylogenetic tree was constructed using FastME (see Methods). The 11 phylogroups are indicated above their ancestral branch; Enterica groups A and B are also indicated. Bootstrap/branch support values are indicated at the nodes.

Figure 2. Coancestry matrix of 73 *Salmonella* genomes, computed using CHROMOPAINTER.

Figure 3. Cumulative curves of gene-by-gene distances between selected pairs of genomes. A: Comparisons with Enterica (group B, serovar Schwartzengrund CVM19633). The arrowhead shows that 20% (0.20, Y-axis) of the genes of an Enterica B strain have less than 1% (0.01, x-axis) divergence to Houtenae B. **B**: Comparisons with Houtenae B (2193/78). The arrowhead shows that 5% of the VII genome and 6% of Houtenae A has less than 0.1% divergence with Houtenae B. **C:** Comparisons with Salamae A (1268/72). **D:** Comparisons with Arizonae (CDC 129-73).

Figure 4. Coancestry matrix between 9 unrelated genomes, computed using CHROMOPAINTER.

Figure 5. Treemix analysis of 12 genomes representative of phylogroups diversity.

The arrowhead indicates the position of the ancestor contributing to extant Houtenae A, Houtenae B and VII lineages. The red arrows indicate gene fluxes inferred by Treemix.

Supporting Information

Table S1. Strains studied by MLST.

Table S2. Genomic sequence data.

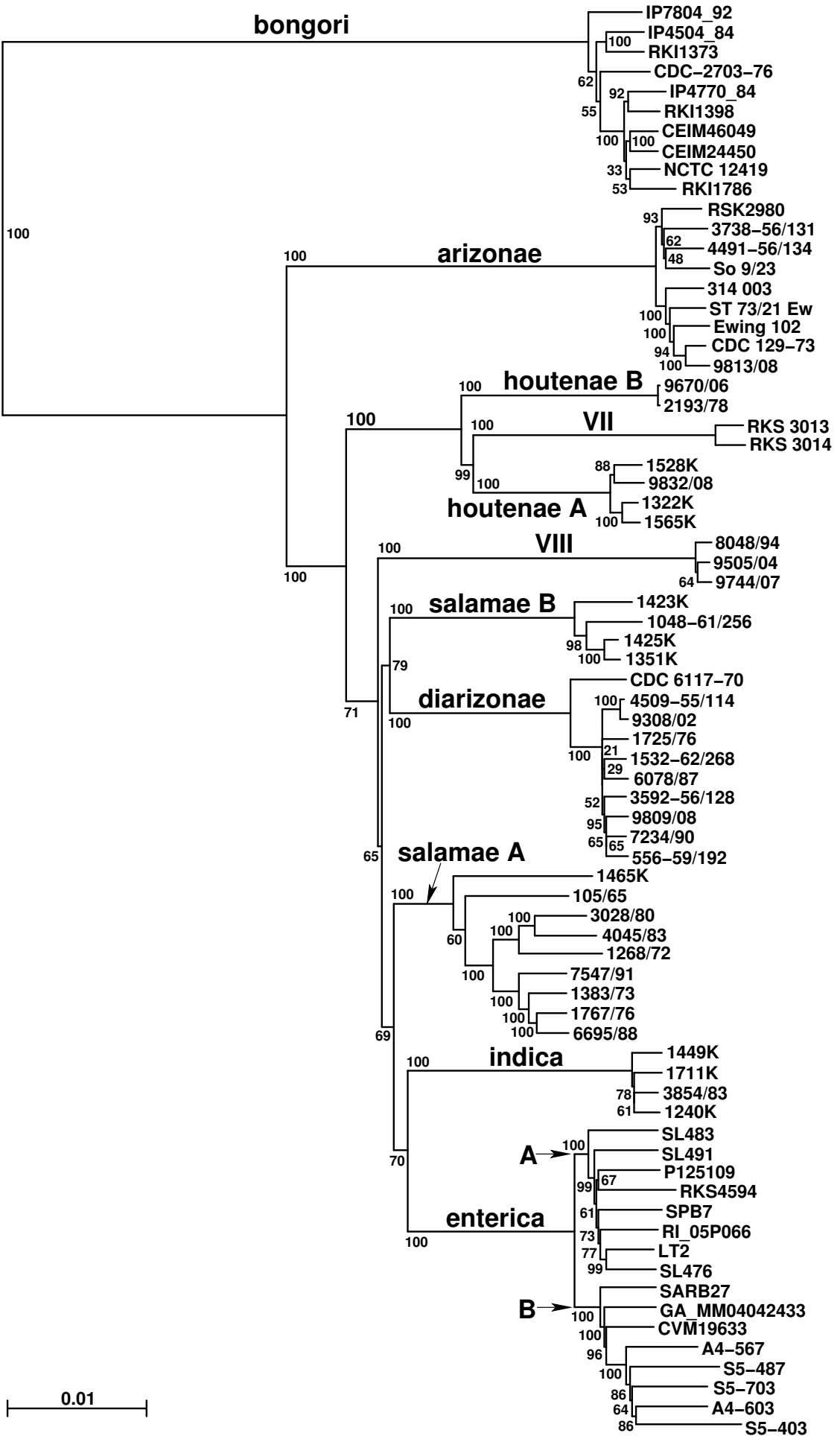
Table S3. Primers and conditions used for MLST gene amplification and sequencing for non-Enterica isolates.

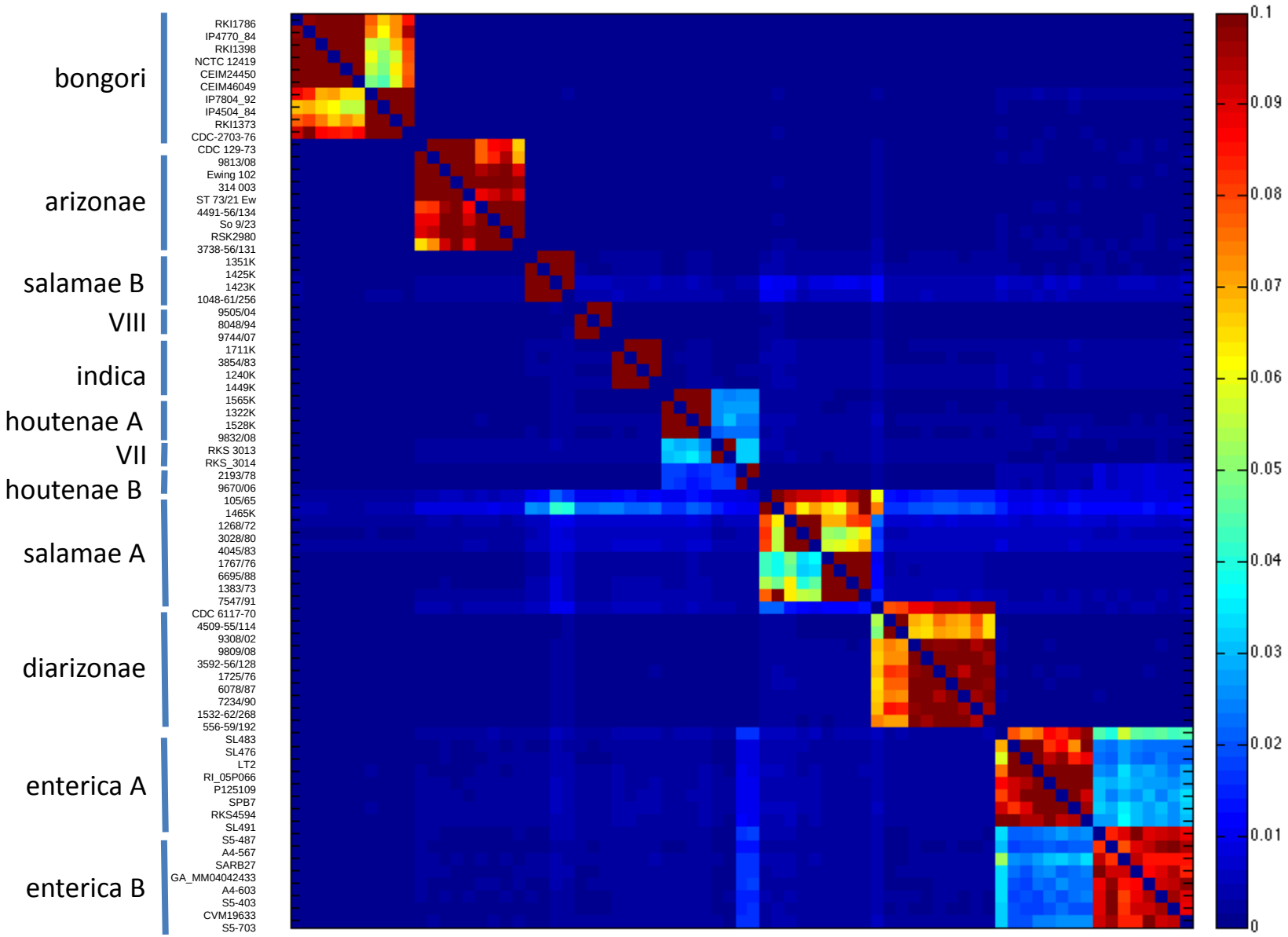
Figure S1. BioNJ* tree of 382 *Salmonella* strains based on seven housekeeping gene sequences. The inset shows the average nucleotide diversity of each phylogroup (houtenae comprises Houtenae A and Houtenae B) at the seven MLST genes.

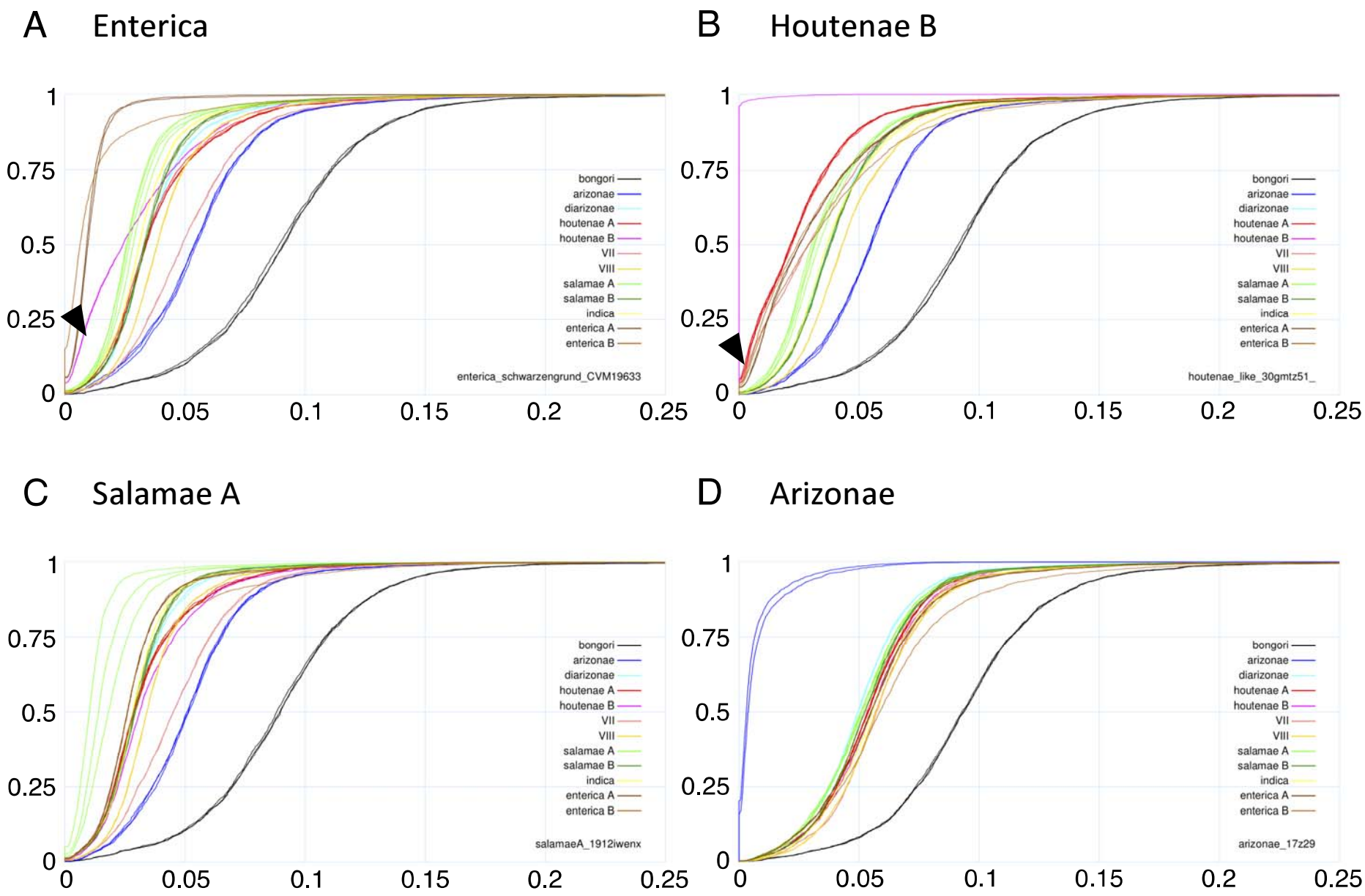
Figure S2. Minimum spanning tree representations of the genotypic diversity within *Salmonella* groups. The minimum spanning trees were constructed for each group based on number of mismatches among MLST allelic profiles. Strains selected for genome sequencing are represented by blue sectors (or blue circles when only one strain shared the corresponding genotype). Grey zones surround groups of sequence types that are connected successively by single allelic mismatches and are equivalent to clonal complexes or 'eBURST' groups (Achtman et al., 2012).

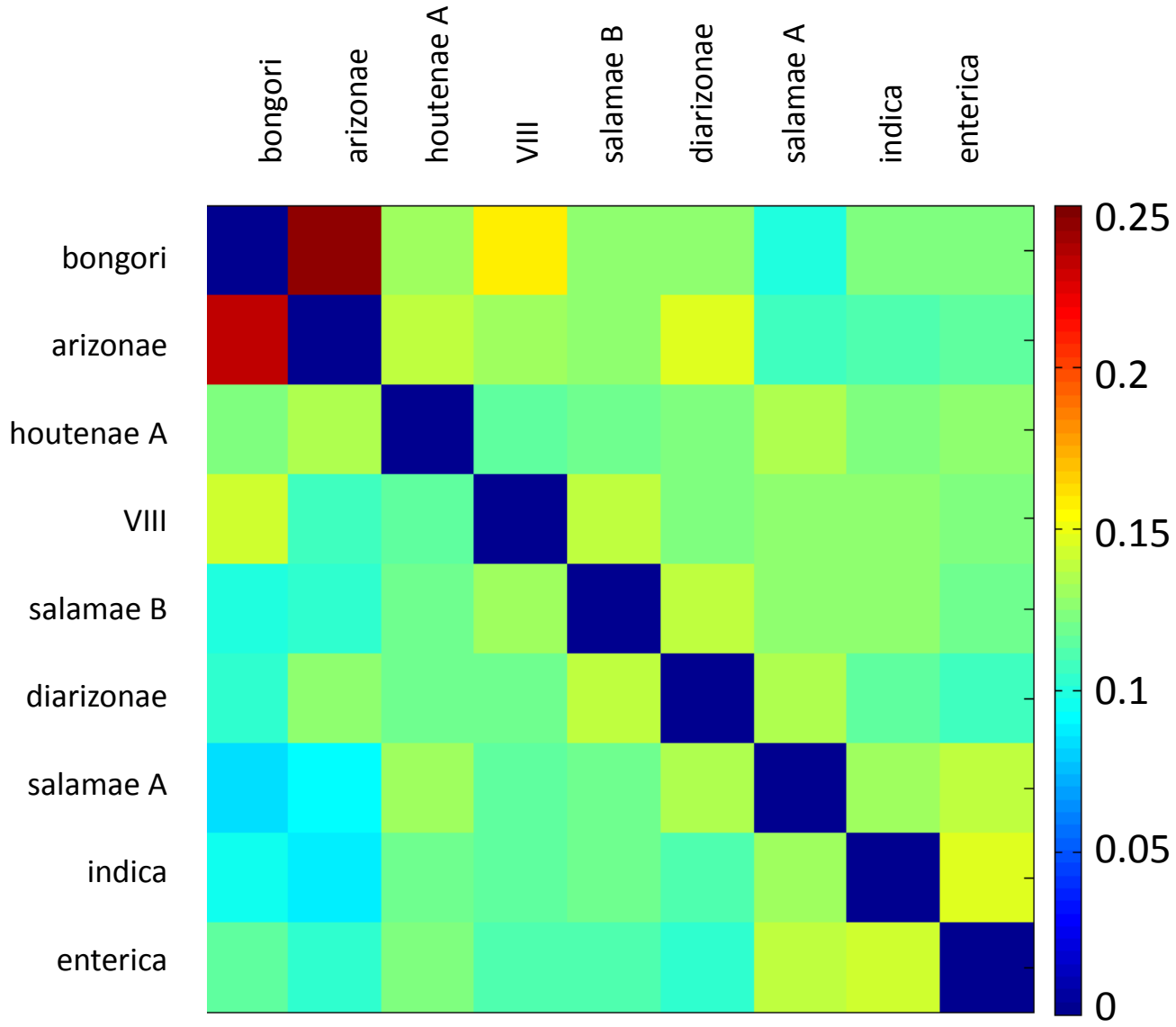
Figure S3. Heatmap of the proportion of shared genes

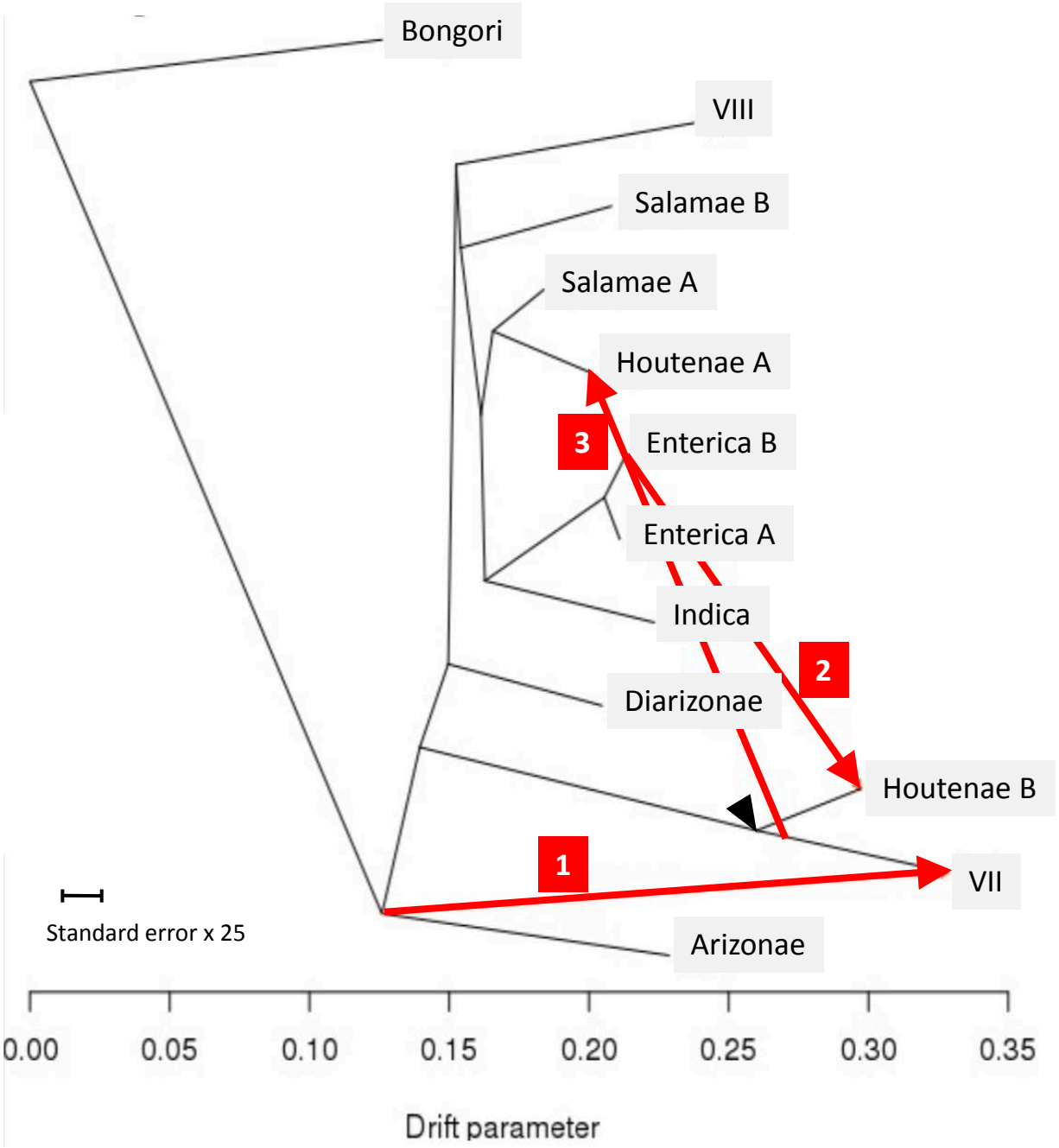
Strains are ordered according to the phylogeny in Figure 1 (left). The proportion of shared genes was computed from the ROARY output with a protein identity cut-off of 85% and the "don't split paralogs" option.

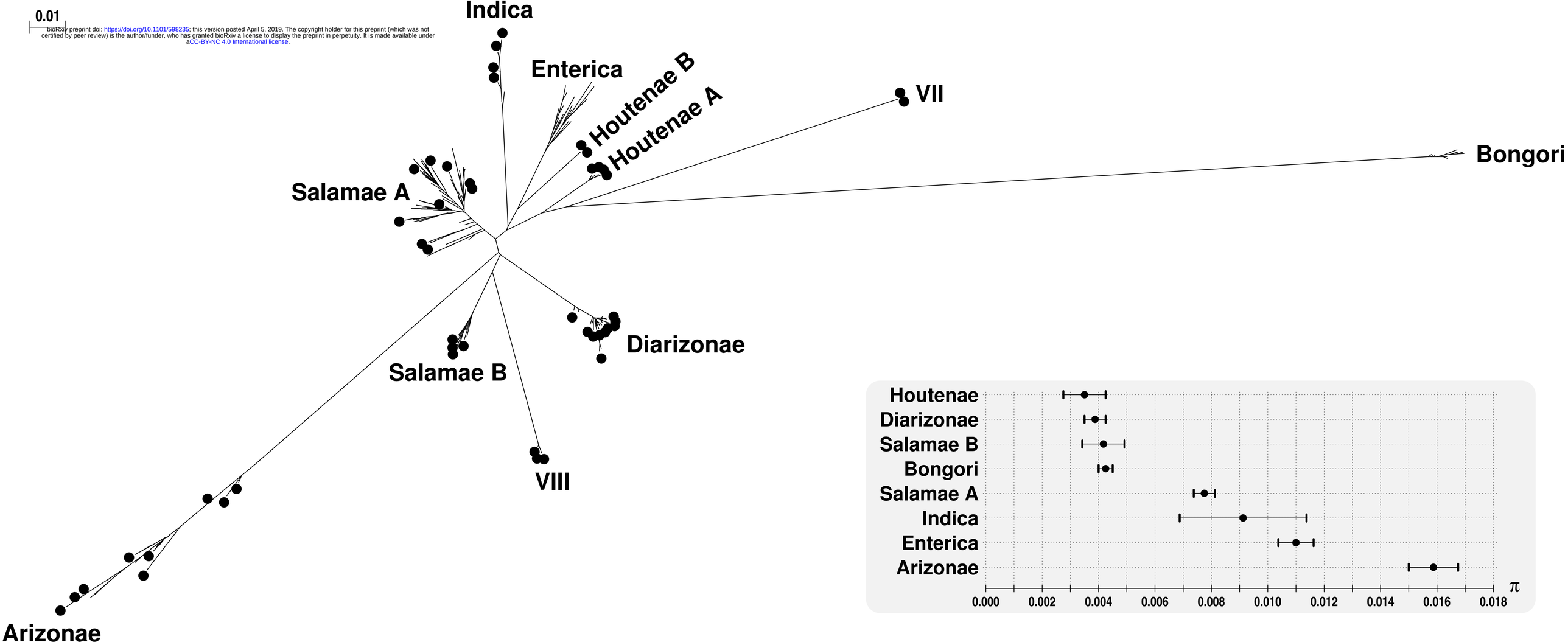

bongori
IP7804_92
IP4504_84
RKI1373
CDC−2703−76
IP4770_84
RKI1398
CEIM46049
CEIM24450
NCTC 12419
RKI1786
arizonae
RSK2980
3738−56/131
4491−56/134
So 9/23
314 003
ST 73/21 Ew
Ewing 102
CDC 129−73
9813/08
houtenae B
9670/06
2193/78
VII
RKS 3013
RKS 3014
houtenae A
1528K
9832/08
1322K
1565K
VIII
8048/94
9505/04
9744/07
salamae B
1423K
1048−61/256
1425K
1351K
diarizonae
CDC 6117−70
4509−55/114
9308/02
1725/76
1532−62/268
6078/87
3592−56/128
9809/08
7234/90
556−59/192
salamae A
1465K
105/65
3028/80
4045/83
1268/72
7547/91
1383/73
1767/76
6695/88
indica
1449K
1711K
3854/83
1240K
A
SL483
SL491
P125109
RKS4594
enterica
SPB7
RI_05P066
LT2
SL476
B
SARB27
GA_MM04042433
CVM19633
A4−567
S5−487
S5−703
A4−603
S5−403
0.01


bongori
arizonae
salamae B
VIII
indica
houtenae A
VII
houtenae B
salamae A
diarizonae
enterica A
enterica B
RKI1786
IP4770_84
RKI1398
NCTC 12419
CEIM24450
CEIM46049
IP7804_92
IP4504_84
RKI1373
CDC-2703-76
CDC 129-73
9813/08
Ewing 102
314 003
ST 73/21 Ew
4491-56/134
So 9/23
RSK2980
3738-56/131
1351K
1425K
1423K
1048-61/256
9505/04
8048/94
9744/07
1711K
3854/83
1240K
1449K
1565K
1322K
1528K
9832/08
RKS 3013
RKS_3014
2193/78
9670/06
105/65
1465K
1268/72
3028/80
4045/83
1767/76
6695/88
1383/73
7547/91
CDC 6117-70
4509-55/114
9308/02
9809/08
3592-56/128
1725/76
6078/87
7234/90
1532-62/268
556-59/192
SL483
SL476
LT2
RI_05P066
P125109
SPB7
RKS4594
SL491
S5-487
A4-567
SARB27
GA_MM04042433
A4-603
S5-403
CVM19633
S5-703
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0

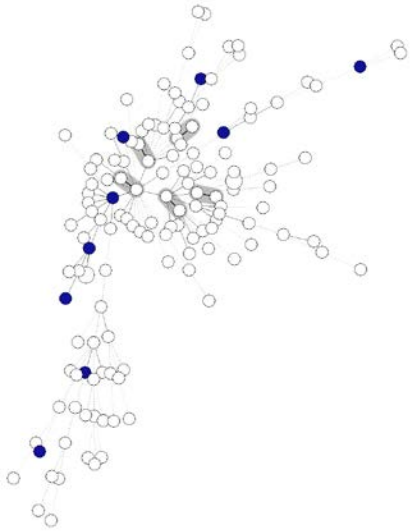



bongori
arizonae
houtenae A
VIII
salamae B
diarizonae
salamae A
indica
enterica
0.25
0.2
0.15
0.1
0.05
0

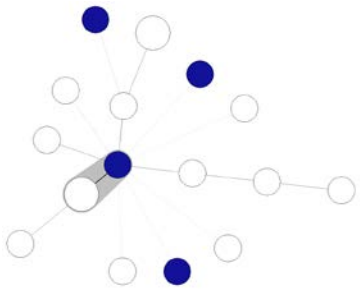

Bongori
VIII
Salamae B
Salamae A
Houtenae A
3
Enterica B
Enterica A
Indica
2
Diarizonae
Houtenae B
1
VII
Arizonae
Standard error x 25
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Drift parameter



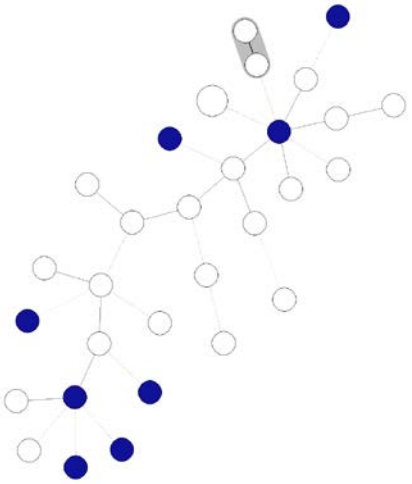
Salamae A



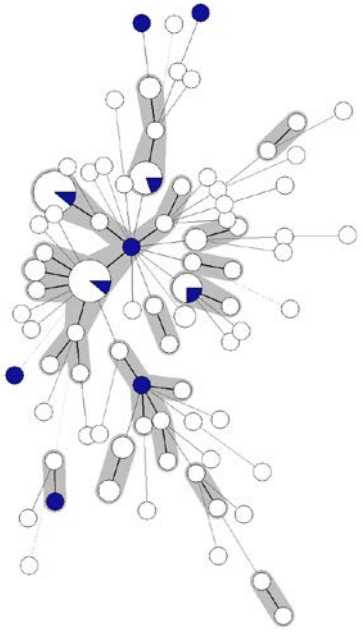
Salamae B



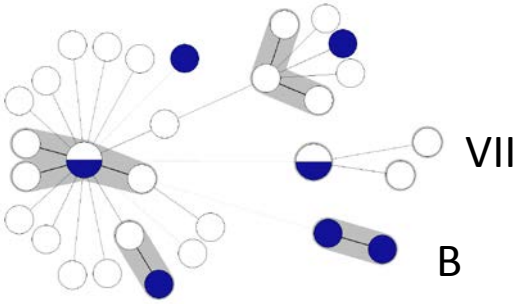
Arizonae



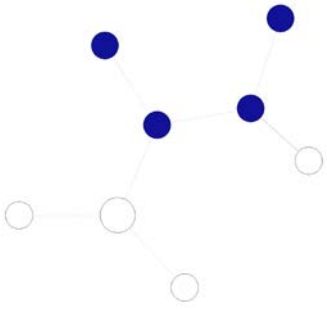
Diarizonae



Houtenae A, B and VII

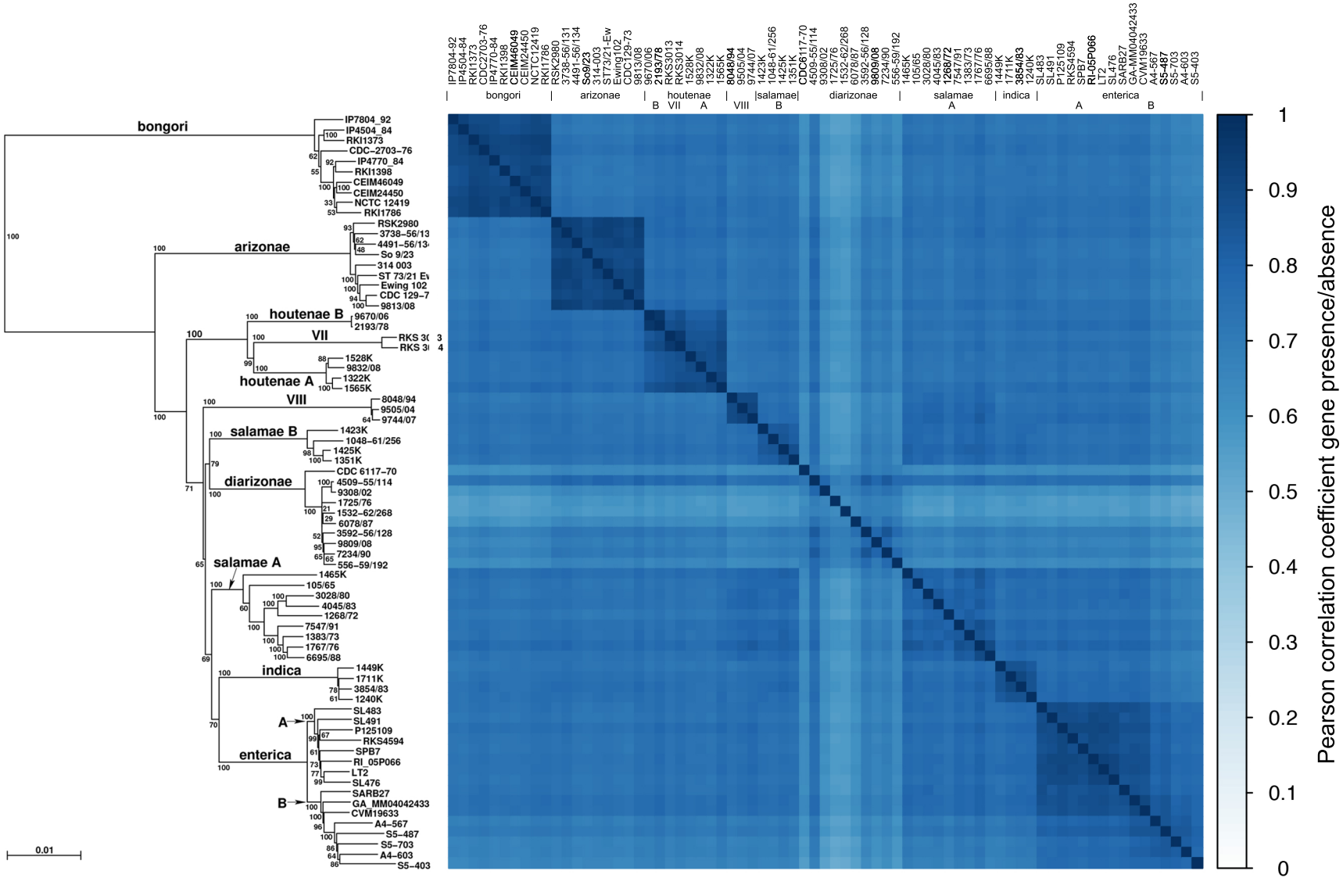


Indica



VIII

S. bongori


bongori
arizonae
houtenae B
VII
houtenae A
VIII
salamae B
diarizonae
salamae A
indica
A
enterica
B
IP7804_92
IP4504_84
RKI1373
CDC–2703–76
IP4770_84
RKI1398
CEIM46049
CEIM24450
NCTC 12419
RKI1786
RSK2980
3738–56/131
4491–56/134
So 9/23
314 003
ST 73/21 Ev
Ewing 102
CDC 129–7
9813/08
9670/06
2193/78
1528K
9832/08
1322K
1565K
8048/94
9505/04
9744/07
1423K
1048–61/256
1425K
1351K
CDC 6117–70
4509–55/114
9308/02
1725/76
1532–62/268
6078/87
3592–56/128
9809/08
7234/90
556–59/192
1465K
105/65
3028/80
4045/83
1268/72
7547/91
1383/73
1767/76
6695/88
1449K
1711K
3854/83
1240K
SL483
SL491
P125109
RKS4594
SPB7
RI_05P066
LT2
SL476
SARB27
GA_MM04042433
CVM19633
A4–567
S5–487
S5–703
A4–603
S5–403
0.01
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Pearson correlation coefficient gene presence/absence