Infection, Genetics and Evolution 16 (2013) 314–325

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Population structures in the SARA and SARB reference collections of *Salmonella enterica* according to MLST, MLEE and microarray hybridization

Mark Achtman [a,*], James Hale [a,*], Ronan A. Murphy [a], E. Fidelma Boyd [a,1], Steffen Porwollik [b,1]

[a] Department of Microbiology and Environmental Research Institute, University College Cork, Lee Road, Cork, Ireland
[b] Vaccine Research Institute of San Diego, San Diego, CA, USA

ARTICLE INFO

Article history:
Received 18 January 2013
Received in revised form 27 February 2013
Accepted 5 March 2013
Available online 16 March 2013

Keywords:
Population genetics
Reference collection
Salmonella enterica

ABSTRACT

In the 1980's and 1990's, population genetic analyses based on Multilocus Enzyme Electrophoresis (MLEE) provided an initial overview of the genetic diversity of multiple bacterial species, including *Salmonella enterica*. The genetic diversity within *S. enterica* subspecies *enterica* according to MLEE is represented by the SARA and SARB reference collections, each consisting of 72 isolates, which have been extensively used for comparative analyses. MLEE has subsequently been replaced by Multilocus Sequence Typing (MLST). Our initial MLST results indicated that some strains within the SARB collection differed from their published descriptions. We therefore performed MLST on four versions of the SARB collection from different sources and one collection of SARA, and found that multiple isolates in SARB and SARA differ in serovar from their original description, and other SARB isolates differed between different sources. Comparisons with a global MLST database allowed a plausible reconstruction of the serovars of the original collection. MLEE, MLST and microarrays were largely concordant at recognizing closely related strains. MLST was particularly effective at recognizing discrete population genetic groupings while the two other methods provided hints of higher order relationships. However, quantitative pair-wise phylogenetic distances differed considerably between all three methods. Our results provide a translation dictionary from MLEE to MLST for the extant SARA and SARB collections which can facilitate genomic comparisons based on archival insights from MLEE.

1. Introduction

Salmonellosis in humans and other warm-blooded animals is usually caused by *Salmonella enterica* subspecies *enterica*, henceforth designated *enterica*. (Achtman et al., 2012) *enterica* is traditionally subdivided into >1500 serovars by the classical Kauffmann–White scheme on the basis of serological properties of lipopolysaccharide and two alternatively expressed phases of the flagellar antigens (Grimont and Weill, 2007). Some serologically indistinguishable groupings are subdivided on the basis of nutritional properties (biotyping). During the 1980's and early 1990's, the population genetic structure of *enterica* was investigated by MultiLocus Enzyme Electrophoresis (MLEE), which is based on the assignment of allelic designations to electrophoretic variants of protein enzymes. MLEE was a powerful tool for population genetic analyses because multiple enzymes were tested, whose combined allelic designations defined Electrophoretic Types (ETs) (Selander et al., 1986). Phylogenetic analyses of MLEE were based on the genetic distances between ETs, calculated as the fraction of allelic differences, which facilitated seminal insights about the population genetic structure of *enterica* by Selander and his colleagues (Selander et al., 1987; Beltran et al., 1988, 1991; Boyd et al., 1993, 1996). Analyses of MLEE data from thousands of isolates indicated that some serovar designations correspond to relatively uniform, monophyletic lineages derived from a single ancestor whereas other serovars confounded polyphyletic lineages derived from multiple ancestors (Boyd et al., 1993; Selander et al., 1990). Unfortunately, the strains that had been subjected to these detailed ET assignments have been lost, as have the details of the MLEE assignments, except for two *enterica* reference collections designated SARA (Beltran et al., 1991) and SARB (Boyd et al., 1993), each composed of 72 isolates.

SARA contains representatives of the genetic diversity within the so-called 'Typhimurium complex', which included the apparently related *enterica* serovars Typhimurium, Saintpaul, Heidelberg, Paratyphi B and Muenchen. SARA includes one representative of each of the 48 ETs identified among 916 isolates from these serovars, supplemented by up to five additional isolates from the most common ETs (Beltran et al., 1991). SARB represents the broader MLEE diversity found among pathogenic *enterica*, and encompasses

* Corresponding authors. Current Address: Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK (M. Achtman).
E-mail addresses: achtmanmark@gmail.com (M. Achtman), james.hale@me.com (J. Hale), fboyd@UDel.Edu (E.F. Boyd), sporwoll@uci.edu (S. Porwollik).
[1] Current Addresses: Department of Biological Sciences, University of Delaware, Newark, DE 19716, USA (E.F. Boyd); Deparment of Pathology and Laboratory Medicine, University of California, Irvine, CA 92697, USA (S. Porwollik).

1567-1348
http://dx.doi.org/10.1016/j.meegid.2013.03.003

37 serovars that are commonly isolated from humans and domesticated animals (Boyd et al., 1993). Fifty-six SARB isolates are distinct from SARA, but 16 others are present in both collections (Supplementary Table S1). Nineteen serovars in SARB are represented by at least two ETs, and 11 of these serovars were polyphyletic. Since their initial description, the SARA and SARB collections have been used as representatives of *enterica* for a variety of studies, as witnessed by 70 citations for SARA (Beltran et al., 1991) and 139 for SARB (Boyd et al., 1993) as of December 2012. However, some of the original serovar designations may have been inaccurate and/or some SARB strains were contaminated according to published results (Torpdahl and Ahrens, 2004; Porwollik et al., 2004; Achtman et al., 2012; Uzzau et al., 1999) and personal communications from Helmut Tschäpe, Francois-Xavier Weill, Ken Sanderson, Mia Torpdahl and Howard Ochman. An analysis of the genomic contents of a selection of these isolates by microarrays revealed multiple discrepancies between MLEE and genomic content, even after some contaminated strains had been excluded (Porwollik et al., 2004).

MLEE has been replaced by MultiLocus Sequence Typing (MLST) (Maiden et al., 1998), which has now been applied to 75 bacterial species (http://pubmlst.org/databases.shtml) and is widely regarded as the Gold Standard for revealing bacterial population genetic patterns (Feil et al., 2004; Aanensen and Spratt, 2005; Maiden, 2006; Didelot and Maiden, 2010). Similar to MLEE, MLST screens polymorphisms in multiple housekeeping genes, based on gene fragments of 400–600 bp which can be readily sequenced in both directions. Allelic variants are assigned unique, numerical designations, and combinations of these alleles are referred to as Sequence Types (STs). MLST differs from MLEE in that it scores all sequence polymorphisms whereas MLEE only detects polymorphisms that result in changes in the electrical charge of proteins. MLST is usually only applied to few genes, typically six to eight, whereas MLEE was used with up to three to five times as many proteins. Finally, MLST tends to identify clusters of closely related STs, referred to as eBGs (eBurst groups) in *enterica*, whereas MLEE was largely used to reconstruct genealogical trees. More than 4000 isolates from >500 serovars of *enterica* fell into 1092 STs, most of which were grouped within 138 eBGs (Achtman et al., 2012). Many eBGs contained only isolates from a single serovar but numerous exceptions were identified.

Phylogenetic trees according to MLST and MLEE were roughly comparable for *Neisseria meningitidis* in an analysis that was performed when MLST was first described (Maiden et al., 1998). The phylogenetic relationships based on MLEE between subspecies of *S. enterica* were roughly comparable with the sequence of the malate dehydrogenase gene, and MLEE detected 57% of the amino acid changes detected by sequencing (Boyd et al., 1994). However, subsequent comparisons of MLEE with MLST have been rare, and their relative discriminatory abilities for eBurst groupings has not been extensively tested. The continued use of the SARA and SARB collections as reference strains for *enterica* might benefit from a retrospective analysis of the discrimination between these methods, especially because MLST has been claimed to have the potential to replace serotyping in these organisms (Achtman et al., 2012) and is increasingly being used to identify genetically related groups of isolates (Fabre et al., 2012; Haase et al., 2012; Hauser et al., 2012; Toboldt et al., 2012). Here we describe the properties of the SARA and SARB reference collections according to MLST, and relate those data to results obtained by MLEE and microarray hybridization.

2. Materials and methods

2.1. Sources of isolates in the SARA and SARB collections

The SARB collection was assembled in Robert K. Selander's laboratory (Pennsylvania State University, Pennsylvania, USA) by E. Fidelma Boyd (Boyd et al., 1993). We examined two sets of SARB isolates that were obtained independently by Mia Torpdahl (Statens Serum Institut, Copenhagen, Denmark) and Steffen Porwollik. Both MT and SP had obtained their copies of the SARB collection from Kenneth E. Sanderson (University of Calgary, Calgary, Alberta, Canada), who in turn had received them from Selander. We also investigated a third copy obtained by Howard Ochman (Yale University, West Haven, Connecticut, USA) from Selander as well as a fourth copy maintained by Fidelma Boyd. Our copy of the SARA collection (Beltran et al., 1991) was also from Fidelma Boyd, who obtained it from Selander during her stay in his laboratory.

2.2. MLST

After streaking to single colonies, and storage of strains in robotically friendly format, DNA was extracted and Multilocus Sequence Typing was performed semi-automatically as previously described (O'Farrell et al., 2012), using the seven gene fragments and oligonucleotides primers that are published on the *S. enterica* MLST website (http://mlst.ucc.ie). The MLST data for the reconstructed SARA and SARB collections have also been deposited at the *S. enterica* MLST website.

2.3. Electrophoretic Types

We extracted the Electrophoretic Types for the SARB collection from the data for 24 loci presented in the original analysis by Boyd et al. (1993). Of those 24 loci, ADK was assigned allelic variant 3 for all isolates because it was monomorphic for all SARA and SARB strains (Boyd et al., 1993). A 25th locus, THD, was excluded in order to allow comparisons with the SARA collection for which that locus had not been tested. Electrophoretic Types based on 20 polymorphic loci for the SARA collection were presented by Beltran et al. (1991). Table 2 in Beltran et al. indicates that the SARA collection had also been tested for ADK, LG2, IPO and MPI, and was monomorphic for those enzymes. The allele for those loci in SARA was assigned as variant 3 because Boyd et al. listed that allele for strains that are present in both SARA and SARB, except for the corrections that ET Sp4 contained LG2 variant 2, Sp4 contained CAT variant 2.5 and He3 possessed PGI variant 3 (footnotes to Table 2 in Boyd et al., 1993). ET designations with additional letters, such as Pb1a and Pb5a, refer to ETs with the same allelic assignments as the ET without the additional letter, except for some enzymes which were inactive and were scored as 0 for a missing allele.

2.4. Microarray analysis

Microarray data for 39 of the isolates analyzed here has already been published (Porwollik et al., 2004) but we also report unpublished data with 35 additional isolates that were specifically tested for this report, all of which is publicly available (http://www.sdibr.org). Microarray analysis, data acquisition and initial data analysis were performed as described by Porwollik et al. using PCR probes that corresponded to the 4936 annotated open reading frames (ORF) of serovar Typhimurium strain LT2, including plasmids pSLT and R46, as well as additional chromosomal CDSs of serovar Typhi strain CT18. These data were scored as present (1), absent (0) and uncertain (2). In order to concentrate on reliable scores from the core genome, we excluded CDSs that were annotated as belonging to mobile genetic elements or prophages (STM0893-0929; STM1005-1056; STM2584-2636; STM2694-2740; STM4417-4436; pSLT plasmid; R46 plasmid; STY1014-1073; STY1591-1643; STY2016-2039), or where the signal intensity did not allow a reliable distinction between presence and absence, leaving a total of 4,483 test scores per strain.

After excluding isolates of incorrect serovar and duplicates, we compared MLST and microarray data from 68 strains where both assays had been performed, and from a subset of 42 isolates from SARA and SARB where ET allelic scores were also available. Pairwise similarity matrices for the microarray data were calculated as the relative frequencies of identical scores over all hybridizations that were scored present or absent. On average, 3864 ± 208 ORFs were scored for each pair of isolates.

2.5. Population genetic analyses

Minimal spanning trees were calculated using the MST algorithm implemented in Bionumerics 6.5 (Applied Maths, Belgium). Distance matrices were generated with pair-wise deletion of missing data with python scripts, which were also used for reformatting data. Linear regression and mantel comparisons were performed with R (R Development Core Team, 2004, http://www.r-project.org/).

3. Results

This analysis was triggered because initial investigations of the SARB collection by MLST (Falush et al., 2006) indicated that some strains that had been received by Mia Torpdahl from K. R. Sanderson did not correspond to their original descriptions. Similar observations have been reported on the basis of serotyping and fingerprinting (Uzzau et al., 1999) or microarray analyses (Porwollik et al., 2002). We also learned of additional unpublished serotyping discrepancies from our colleagues. In order to resolve these issues, we performed MLST on a second copy of the SARB collection from Sanderson's laboratory that had been used by Steffen Porwollik for microarray analyses. In addition, Howard Ochman performed MLST on a third copy of the SARB collection that he had received directly from R. K. Selander. The results revealed additional discrepancies between the three collections, and occasional mixed cultures within supposedly pure stocks. We then performed MLST on the SARA and SARB collections that had been maintained by E. Fidelma Boyd since her stay in Selander's lab. Here we summarize the composition of these four collections of SARB and one collection of SARA, and reconstruct the original composition of SARA and SARB.

3.1. Strain discrepancies

For ten isolates, the multilocus sequence type (ST) was consistent between multiple collections but the originally reported serotype differed from that of multiple other isolates with the same ST (Table 1). Reserotyping of those 10 isolates confirmed that the serotype predicted by ST was correct, and differed from the published serotypes of those strains, usually due to a single epitope difference in the antigenic formula or a different biovar. Four of these ten isolates existed as two pairs of the same strain within SARA and SARB, and yielded uniform serotypes and STs for each pair. For a fifth strain, SARB50 (ET Pc4, supposedly a serovar Paratyphi C isolate), the strain in the Torpdahl collection was serovar Oranienburg and the strain in the Boyd collection was rough: m, t:– which is compatible with the loss of smooth LPS by Oranienburg.

We conclude that these ten isolates probably reflect serotyping mistakes. Some of these discrepancies have been previously described, stimulated by unexpected microarray results that led to re-serotyping (Porwollik et al., 2004) or after re-evaluation of biotyping results (Uzzau et al., 2000; Achtman et al., 2012).

Five MLST discrepancies were observed between individual collections (Table 2). Discrepant serovars between individual single colonies have also been observed by others (W. Rabsch, pers. comm.), as have differing alleles between single colonies for individual MLST gene fragments (H. Ochman, pers. comm.). Two of these collection-specific discrepancies concern a known inversion of SARB19 (ET En7, supposedly a serovar Enteritidis strain; ST77) and SARB20 (Em1, supposedly a serovar Emek isolate; ST76) in the Sanderson collection (Porwollik et al., 2004). ET En7 was only isolated once (Switzerland), and no other strain has been assigned to ST77. ET Em1 was also only isolated once (Israel) but a second ST76 isolate is known, strain 297K, which is the Kauffmann reference strain for serovar Emek and had been isolated in Israel in 1949. The three other collection-specific discrepancies consisted of SARB49, SARB58 and SARB69, which differed in serovar and ST between different collections (Table 2). For each of these, we assigned the correct serotype on the basis of the serovars and STs of additional SARB isolates from other collections, and these three isolates probably correspond to cultures that were contaminated during storage of the SARB collection in individual laboratories.

We detected two further cases where the ST of SARA or SARB strains from different collections differed by a single allelic difference. SARA34 (He1, a serovar Heidelberg isolate) was ST15 in the Ochman collection and ST1615 in the Boyd collection; these STs differ at nucleotide 208 in *purE*. Three other He1 SARA strains are also ST15 (SARA30, SARA31, SARA32) and one other He1 SARA strain (SARA33) is ST1615 (Table 3), suggesting an additional strain mixup. We arbitrarily assigned ST1615 to the original SARA34 strain because the strain in our collection has that ST. Similarly, SARB61 (Sv2, a serovar Stanleyville strain) was ST1630 in F. Boyd's collection and ST97 in M. Torpdahl's collection, which differ by eight nucleotides in *dnaN*. Each of those STs is only present once within the MLST database, and they are not closely related to any

Table 1
Erroneous serotype assignments in the SARA and SARB collections.

Strain	ET	Serovar [antigenic formula]		Concordant ST
		Supposed	Observed (ST)	T P B
SARB5	Cs6	Choleraesuis [6,7:c:1,5]	Decatur (67) [6,7:c:1,5]	■ □ ■
SARB7	Cs13	Choleraesuis [6,7:c:1,5]	Decatur (69) [6,7:c:1,5]	■ □ ■
SARB35/SARA71	Mu4	Muenchen [6,8:d:1,2]	Manhattan (18) [6,8:d:1,5]	■ ■ ■
SARB40	Pn2	Panama [9,12:l,v:1,5]	Javiana (24) [9,12:l,z_{28}:1,5]	■ ■ ■
SARB41	Pn12	Panama [9,12:l,v:1,5]	Javiana (24) [9,12:l,z_{28}:1,5]	■ ■ ■
SARB47/SARA62	Pb7	Paratyphi B [4,12:b:1,2]	Limete (89) [4,12:b:1,5]	■ ■ ■
SARB50	Pc4	Paratyphi C [6,7:c:1,5]	Oranienburg (91) [6,7,14:m,t]	■ □ ■
SARA61	Pb6	Paratyphi B [4,12:b:1,2]	Agona (13) [4,12:f,g,s:-]	□ □ ■
SARB70	Ts3	Typhisuis [6,7:c:1,5]	Decatur (70) [6,7:c:1,5]	■ □ ■
SARA72	Mu4a	Muenchen[6,8:d:1,2]	Manhattan (113) [6,8:d:1,5]	□ □ ■

Note: Serovar assignments except for Decatur are based on reserotyping of still another copy of the Sanderson SARB collection by Helmuth Tschäpe, RKI, Wernigerode, and of the Boyd collections for SARB47, SARB50, SARA61 and SARA72 by F.-X. Weill. The serotyping of SARB5, SARB7 and SARB70 as Decatur is based on the data reported by Achtman et al. (2012). STs that were concordant between the T (Mia Torpdahl version of Sanderson collection), P (Steffen Porwollik version of Sanderson collection) and B (Fidelma Boyd) collections are indicated by a filled box and strains that were not MLST typed are indicated by an open box. SARB50 is rough in the B collection.

Table 2
Contaminated cultures in the SARB collections.

		Serovar (ST) [antigenic formula]		Concordant ST
Strain	ET	True	False	T P O B
SARB19	En7	Enteritidis (77) [9,12:g,m:-]	Emek (76) [8:g.m.s:-]	x x ■ ■
SARB20	Em1	Emek (76) [8:g.m.s:-]	Enteritidis (77) [9,12:g,m:-]	x x ■ ■
SARB49	Pc2	Paratyphi C (114) [6,7:c:1,5]	Limete (89) [4,12:b:1,5]	x ■ □ ■
SARB58	Se1	Sendai (85) [9,12:a:1,5]	Saintpaul (95) [4,12:e,h:1,2]	x x ■ ■
SARB69	Ts1	Typhisuis (147) [6,7:c:1,5]	Thompson (26) [6,7:k:1,5]	■ □ □ x

Note: The true serovar assignments are based on reserotyping of still other copies of the Sanderson SARB collection by Wolfgang Rabsch, RKI, Wernigerode, and of the Boyd collections by F.-X. Weill, Institut Pasteur, Paris. Collections are designated by T (Mia Torpdahl version of Sanderson collection), P (Steffen Porwollik version of Sanderson collection), O (Howard Ochman) and B (Fidelma Boyd). The true STs (corresponding to the serovar assignment in column 3) are indicated by a filled box, false STs (corresponding to the false serovar assignments in column 4) by an x and strains that were not MLST typed are indicated by an open box.

other ST. We arbitrarily assigned ST1630 to the original SARB61 culture (Table 4) because the serovar had been independently confirmed for that strain and we possess it in our collection. After correcting these problems, we were able to make a plausible reconstruction of the serovar, ET and ST properties of the original SARA and SARB collections (Tables 3 and 4).

After excluding the 16 strains that are common to the SARA and SARB collections (Supplementary Table S1), SARA plus SARB contain a total of 128 unique strains from 103 ETs and 46 serovars or serovar variants. A summary of the relationship between ETs, STs and eBGs is presented in Supplementary Tables S2 and S3.

3.2. MLST versus MLEE

How similarly do MLEE ETs and MLST STs differentiate individual strains of *enterica*? Because the phylogenies of *N. meningitidis* were very similar by MLEE and MLST (Maiden et al., 1998), we expected that the resolution between the two methods within *enterica* would also be comparable. However, a detailed comparison of the results with both methods showed that their resolutions differed dramatically. The MLEE data analyzed here consist of the alleles identified at 24 enzyme loci, of which four were invariant in all isolates. The MLST data are based on sequences of seven gene fragments, all of which are polymorphic. Pair-wise comparisons of genetic distances (fraction of different alleles) for all 128 unique strains in SARA plus SARB showed that most pairs of isolates do not share any identical MLST alleles (Fig. 1A) whereas all pairs do share some MLEE alleles (Fig. 1B). The genetic similarities of pair-wise comparisons according to the two methods were significantly correlated (Mantel test, $R^2 = 0.27$, $p < e^{-4}$) (Fig. 2), but this level of correlation is relatively weak, and the 95% confidence limits of a predicted straight line excluded the few pairs of related strains with high genetic similarities according to both methods. Based on these results, MLEE would be expected to recognize broader groupings of strains whereas MLST would be more applicable for differentiating between groups of closely related isolates.

We tested these predictions by a greedy network approach (minimal spanning tree of allelic differences; MSTree) based on the eBurst algorithm (Feil et al., 2004; Feil, 2010). eBurst groupings are commonly used for the analysis of MLST data because they focus on close relationships, and do not attempt to reconstruct deep phylogenies, which can be difficult to elucidate when homologous recombination is frequent (Hanage et al., 2006a,b). Homologous recombination is common in *S. enterica* (Brown et al., 2002; Falush et al., 2006), resulting in little deep phylogenetic signal even when 300 kb of sequences from multiple gene fragments other than the seven loci chosen for MLST are sequenced per strain (Didelot et al., 2011). In contrast, MLST eBurst groupings within *enterica* (designated eBGs) represent natural groupings that are broadly consistent, independent of the clustering method (Achtman et al., 2012). eBGs also correlated well with serovar except where the serovar was polyphyletic (Achtman et al., 2012). However, eBurst-based clustering has not been previously applied to MLEE data, or to the microarray data analyzed below.

Figs. 3 and 4 present a comparison of the eBG clustering of the same isolates using either MLEE data (Fig 3) or MLST data (Fig 4) as input. The MLEE-based MSTree revealed that ETs from SARA plus SARB fell into eBurst clusters of single locus variants (SLVs), several of which were uniformly associated with a cluster-specific serovar. However, multiple eBurst clusters were not uniquely associated with a single serovar. One large cluster contained predominantly Typhimurium (Tm) but also Saintpaul (Sp) and a second cluster conflated Enteritidis (En) with Dublin (Du). These assignments are supported by the analyses of 300 kb of housekeeping gene sequences, which grouped Tm and Sp within Lineage 2 and En and Du in Lineage 4 (Didelot et al., 2011). However, a third MLEE SLV cluster conflated Paratyphi B (Pb) and Muenchen (Mu), and a fourth conflated Javiana, (Pn2), Rubislaw (Ru1) and Panama (Pn), which was not supported by the 300 kb sequence analyses. More relaxed clustering of both SLVs and double locus variants (DLVs; dashed lines) of MLEE ETs identified a larger 'supercluster', which in large part corresponds to the entire original SARA collection, the so-called 'Typhimurium Complex'. This supercluster included many of the ETs found in serovars Tm, Sp, Heidelberg (He), Pb, and Mu (Fig. 3), which is again not supported by the 300 kb analyses (Didelot et al., 2011). The other clusters that arose when links to DLVs were included were also largely inconsistent with both serovar and sequence based analyses.

The MSTree of the same isolates based on MLST SLVs provided a contrasting view of genetic clustering in *enterica* (Fig. 4). Each MLST cluster corresponds to a unique eBG. Several pairs of DLVs in Fig. 4 are indicated as belonging to a common eBG because data from additional strains on the *enterica* MLST website supplied missing SLV links, e.g. Cs1 (Choleraesuis) and Cs11 are DLVs within eBG6, which contains other Choleraesuis isolates, and Pc1 (Paratyphi C), Pc2 and Ts1 (Typhisuis) are DLVs within eBG20, which contains other Paratyphi C and Typhisuis isolates (Achtman et al., 2012). (eBGs 6 and 20 are discrete because their central STs are DLVs and intermediate SLV links have not been identified.) Similarly, En1 (Enteritidis) and En3 are DLVs of Ga2 (Gallinarum), which is itself a DLV of Pu3 and Pu4 (Gallinarum var. Pullorum); all of these STs belong to eBG4 which includes multiple other isolates of those serovars. The MLST results are in accordance with genomic sequences, which have shown close similarities between Paratyphi C and Choleraesuis (Liu et al., 2009) and between Enteritidis and Gallinarum (Thomson et al., 2008). With rare exceptions, such as eBG6 and eBG20 described above, the fine-grained structure of ST clusters is not informative about higher order genetic relationships. For example, unlike MLEE, MLST did not identify a relationship between Typhimurium and Saintpaul under the standard conditions used during eBURST analysis.

Table 3
Summary of the SARA collection.

Strain	RKS No.	Original Designation	Correct Serovar	Prior Serovar	Host	Country	Year	ET	ST	eBG
SARA1	284	INSP24	Typhimurium		Human	Mexico		Tm1	19	1
SARA2	4939	LT2	Typhimurium					Tm1	19	1
SARA3	145	NVSL7095	Typhimurium		Horse	USA	1987	Tm1	19	1
SARA4	183	NVSL5820	Typhimurium		Rabbit	USA	1986	Tm1	19	1
SARA5	810	IVB232	Typhimurium			Mongolia		Tm1	19	1
SARA6	345	CDCB1213	Typhimurium		Human	USA		Tm2	19	1
SARA7	821	IVB665/81	Typhimurium			Norway		Tm3	36	138
SARA8	811	IVB5560	Typhimurium			Finland		Tm5	36	138
SARA9	203	NVSL2816	Typhimurium		Parrot	USA	1987	Tm7	98	1
SARA10	154	NVSL6814	Typhimurium		Opossum	USA	1987	Tm9	19	1
SARA11	829	IVB276/25	Typhimurium			Thailand		Tm10	19	1
SARA12	147	NVSL6993	Typhimurium		Horse	USA	1987	Tm11	19	1
SARA13	837	IVB1430	Typhimurium			France		Tm12	19	1
SARA14	842	IVB76/67	Typhimurium			Panama		Tm13	19	1
SARA15	149	NVSL6968	Typhimurium		Dog	USA	1987	Tm14	19	1
SARA16	350	CDCB1236	Typhimurium		Human	USA		Tm15	19	1
SARA17	1164	IVB48/81	Typhimurium			Yugoslavia		Tm16	19	1
SARA18	151	NVSL6938	Typhimurium		Horse	USA	1987	Tm17	19	1
SARA19	93	INSP85	Typhimurium		Human	Mexico		Tm21	19	1
SARA20	839	IVB1544	Typhimurium			France		Tm22	19	1
SARA21	4535	USFW318	Typhimurium		Heron	USA		Tm23	99	1
SARA22	1688	CDCB1605	Saintpaul		Human	USA		Sp1	50	14
SARA23	1689	CDCB1722	Saintpaul		Human	USA		Sp2	49	14
SARA24	1690	CDCB2076	Saintpaul		Human	USA		Sp3	50	14
SARA25	1380	IVB516	Saintpaul			France		Sp3	27	14
SARA26	3748	IP67/88	Saintpaul		Human	France	1988	Sp3	27	14
SARA27	3755	IP78/88	Saintpaul		Human	France	1988	Sp3	27	14
SARA28	3763	IP86/88	Saintpaul		Human	France	1988	Sp3	27	14
SARA29	1686	CDCB1400	Saintpaul		Human	USA		Sp4	95	209
SARA30	539	NVSL7039	Heidelberg		Chicken	USA	1987	He1	15	26
SARA31	560	NVSL5876	Heidelberg		Pig	USA	1987	He1	15	26
SARA32	562	NVSL5145	Heidelberg		Dog	USA	1986	He1	15	26
SARA33	576	INSP94	Heidelberg		Human	Mexico		He1	1615	26
SARA34	1364	IVB7135/1990	Heidelberg			Israel		He1	1615	26
SARA35	1389	IVB126/82	Heidelberg			Brazil		He2	15	26
SARA36	1391	IVB588/24	Heidelberg			Thailand		He3	15	26
SARA37	543	NVSL5208	Heidelberg		Turkey	USA	1987	He4	15	26
SARA38	540	NVSL4960	Heidelberg		Turkey	USA	1987	He5	15	26
SARA39	646	CDCB2487	Heidelberg		Human	USA		He7	15	26
SARA40	1347	IVB218/82	Heidelberg			USA		He8	15	26
SARA41	3222	IP155/76	Paratyphi B		Human	France	1976	Pb1	86	5
SARA42	3279	DMS724/74	Paratyphi B		Human	UK	1974	Pb1	86	5
SARA43	3305	DMS220/82	Paratyphi B		Human		1982	Pb1	86	5
SARA44	3265	DMS2434	Paratyphi B		Human		1965	Pb1	86	5
SARA45	3596	IP7/88	Paratyphi B		Cow	France	1988	Pb1	86	5
SARA46	3294	DMS3254/7/811	Paratyphi B		Human		1981	Pb1a	86	5
SARA47	3249	DMS3205/83	Paratyphi B var. Java		Sewage	UK	1983	Pb2	43	5
SARA48	3237	DMS843/82	Paratyphi B var. Java		Human	UK	1982	Pb2a	149	5
SARA49	3267	DMS2442	Paratyphi B var. Java		Sewage	UK	1982	Pb2b	43	5
SARA50	3202	DMS106/76	Paratyphi B var. Java		Food		1976	Pb3	110	5
SARA51	3193	IP53/76	Paratyphi B var. Java		Human	France	1976	Pb3	110	5
SARA52	3614	IP87/87	Paratyphi B var. Java		Cow	France	1987	Pb3	110	5
SARA53	3605	IP16/88	Paratyphi B var. Java		Human	France	1988	Pb3	110	5
SARA54	3597	IP8/88	Paratyphi B var. Java		Human	France	1988	Pb3	110	5
SARA55	3211	IP47/81	Paratyphi B var. Java		Human	France	1981	Pb3a	110	5
SARA56	3201	IP83/76	Paratyphi B var. Java		Human	France	1976	Pb4	88	19
SARA57	3274	DMS2471	Paratyphi B monophasic		Water	UK	1965	Pb5	42	32
SARA58	3218	IP59/81	Paratyphi B var. Java monophasic		Human	France	1981	Pb5a	42	32
SARA59	3219	IP61/81	Paratyphi B var. Java monophasic		Human	France	1981	Pb5b	42	32
SARA60	3192	IP52/76	Paratyphi B var. Java monophasic		Food	France	1976	Pb5c	734	32
SARA61	3277	DMS203/74	Agona	Paratyphi B	Water	UK	1974	Pb6	13	54
SARA62	3215	DMS53/81	Limete	Paratyphi B	Human		1981	Pb7	89	
SARA63	4283	IP6/88	Muenchen		Human	France	1988	Mu1	111	8
SARA64	4129	NVSL519	Muenchen		Cow	USA	1986	Mu1	82	8
SARA65	4135	NVSL2817	Muenchen		Chicken	USA	1987	Mu1	82	8
SARA66	4277	CDCB2026	Muenchen		Human	USA		Mu1	112	8
SARA67	4317	INSP46	Muenchen		Human	Mexico		Mu1	112	8
SARA68	4292	IP15/88	Muenchen		Human	France	1988	Mu1a	112	8
SARA69	4288	IP11/88	Muenchen		Human	France	1988	Mu2	83	
SARA70	4300	IP25/88	Muenchen		Human	France	1988	Mu3	84	
SARA71	4272	CDCB1293	Manhattan	Muenchen	Human	USA		Mu4	18	27
SARA72	4306	IP31/88	Manhattan	Muenchen	Human	France	1988	Mu4a	113	27

Table 4
Summary of the SARB collection.

Strain	RKS No.	Original Designation.	Correct Serovar	Prior Serovar	Host	Country	Year	ET	ST	eBG
SARB1	1701	IVB36/79	Agona			Peru		Ag1	13	54
SARB2	2403	CDCB1487	Anatum		Human	USA		An1	64	65
SARB3	4231	DMS2819	Brandenburg			UK	1988	Ba2	65	12
SARB4	1280	NVLS6321	Choleraesuis var. Kunzendorf		Pig	USA	1986	Cs1	66	6
SARB5	1239	IVB651/79	Decatur	Choleraesuis		USA		Cs6	67	141
SARB6	3169	CDC3327/54	Choleraesuis var. senso stricto			Thailand	1954	Cs11	68	6
SARB7	4640	IP6562/88	Decatur	Choleraesuis		Australia	1988	Cs13	69	142
SARB8	4647	631 K	Decatur			France		Dt1	70	
SARB9	246	NVSL4111	Derby		Avian	USA	1986	De1	71	
SARB10	241	NVSL5558	Derby		Cow	USA	1986	De13	40	57
SARB11	243	NVSL5283	Derby		Turkey	USA	1986	De31	72	
SARB12	1518	NVSL5618	Dublin		Cow	USA	1986	Du1	10	53
SARB13	4717	IP82/3144	Dublin		Cow	France	1982	Du3	73	53
SARB14	1550	IVB3540/24	Dublin/Enteritidis	Dublin	Human	Thailand		Du2	74	32
SARB15	4239	DMS3618	Duisburg			UK	1988	Di1	75	
SARB16	53	CDCSSU7998	Enteritidis			USA		En1	11	4
SARB17	761	IVB176/82	Enteritidis		Human	Brazil	1982	En2	6	
SARB18	69	CDCSSU8074	Enteritidis			USA	1988	En3	11	4
SARB19	1208	IVB470/82	Enteritidis			Switzerland	1988	En7	77	
SARB20	1216	IVB4793/3366	Emek			Israel		Em1	76	
SARB21	2962	CDC4801/72	Gallinarum		Human	USA	1972	Ga2	78	4
SARB22	4241	DMS3005	Haifa			Scotland	1988	Ha1	49	14
SARB23	539	NVSL7039	Heidelberg		Chicken	USA	1987	He1	15	26
SARB24	1391	IVB588/24	Heidelberg			Thailand		He3	15	26
SARB25	4250	DMS3702	Indiana			UK	1988	Id1	17	
SARB26	1490	CDCB3460	Infantis		Human	USA		In1	32	31
SARB27	1452	IVB385/72	Infantis			Senegal		In3	79	
SARB28	2833	CDC4648/53	Miami		Human	USA	1953	Mi1	80	111
SARB29	4381	IP2/79	Miami		Human	Guiana	1979	Mi5	48	42
SARB30	1762	CDCB2131	Montevideo		Human	USA		Mo1	4	40
SARB31	1740	CDCB2604	Montevideo		Human	USA		Mo6	81	40
SARB32	3121	ATCC8388	Muenchen					Mu1	82	8
SARB33	4288	IP11/88	Muenchen		Human	France	1988	Mu2	83	
SARB34	4300	IP25/88	Muenchen		Human	France	1988	Mu3	84	
SARB35	4272	CDCB1293	Manhattan	Muenchen	Human	USA		Mu4	18	27
SARB36	2016	CDCB3465	Newport		Human	USA		Np8	5	2
SARB37	1915	INSP15	Newport		Human	Mexico		Np11	31	7
SARB38	1956	NVSL3882	Newport		Snake	USA	1987	Np15	46	3
SARB39	1793	IVBBendia	Panama			Italy		Pn1	48	42
SARB40	1776	CDCB1171	Javiana	Panama	Human	USA		Pn2	24	17
SARB41	1779	CDCB1433	Javiana	Panama	Human	USA		Pn12	24	17
SARB42	4993	ATCC9150	Paratyphi A					Pa1	85	11
SARB43	3222	DMS155/76	Paratyphi B		Human	France	1976	Pb1	86	5
SARB44	3202	DMS106/76	Paratyphi B var. Java		Food	Middle East	1976	Pb3	110	5
SARB45	3201	IP83/76	Paratyphi B var. Java		Human	France	1976	Pb4	88	19
SARB46	3274	DMS2471	Paratyphi B monophasic		Water	UK	1965	Pb5	42	32
SARB47	3215	DMS53/81	Limete	Paratyphi B	Human	Africa	1981	Pb7	89	
SARB48	4587	33 K	Paratyphi C			France		Pc1	90	20
SARB49	4594	IP2/88	Paratyphi C		Human	France	1988	Pc2	114	20
SARB50	4620	IP4/77	Oranienburg	Paratyphi C	Human	France	1977	Pc4	91	50
SARB51	2266	IVB978/87	Gallinarum var. Pullorum			Germany		Pu3	92	4
SARB52	2246	IVBItalianStandard	Gallinarum var. Pullorum			Germany		Pu4	92	4
SARB53	4256	DMS3853	Reading			UK	1988	Re1	93	43
SARB54	4938	ATCC10717	Rubislaw					Ru1	94	133
SARB55	1690	CDCB2076	Saintpaul			USA		Sp3	50	14
SARB56	1686	CDCB1400	Saintpaul			USA		Sp4	95	209
SARB57	4261	DMS1253	Schwarzengrund			UK	1988	Sw1	96	33
SARB58	2866	CDC1035/74	Sendai			UK	1974	Se1	85	11
SARB59	2358	NVSL6673	Senftenberg		Chicken	USA	1987	Sf1	14	55
SARB60	4264	DMS1112	Stanley			UK	1988	St1	51	29
SARB61	4267	DMS3705	Stanleyville			UK	1988	Sv2	1630	
SARB62	1767	CDCB2637	Thompson		Human	USA		Th1	26	28
SARB63	3333	IPE.88.374	Typhi		Human	Senegal	1988	Tp1	2	13
SARB64	3320	IPE.88.353	Typhi		Human	Senegal	1988	Tp2	3	13
SARB65	284	INSP24	Typhimurium		Human	Mexico		Tm1	19	1
SARB66	203	NVSL2816	Typhimurium		Parrot	USA	1987	Tm7	98	1
SARB67	837	IVB1430	Typhimurium			France		Tm12	19	1
SARB68	4535	USFW318	Typhimurium		Heron	USA		Tm23	99	1
SARB69	3134	CDC277/68	Typhisuis		Pig	USA	1968	Ts1	147	20
SARB70	3133	CDC1426/67	Decatur	Typhisuis	Pig	USA	1967	Ts3	70	
SARB71	4000	IP5/88	Wien			France	1988	Wi1	101	
SARB72	3998	IP3/88	Wien			France	1988	Wi2	102	

NOTE: SARB50 was serotyped as Oranienburg in the collection from Mia Torpdahl and as rough:m, t: – in the collection from Fidelma Boyd.

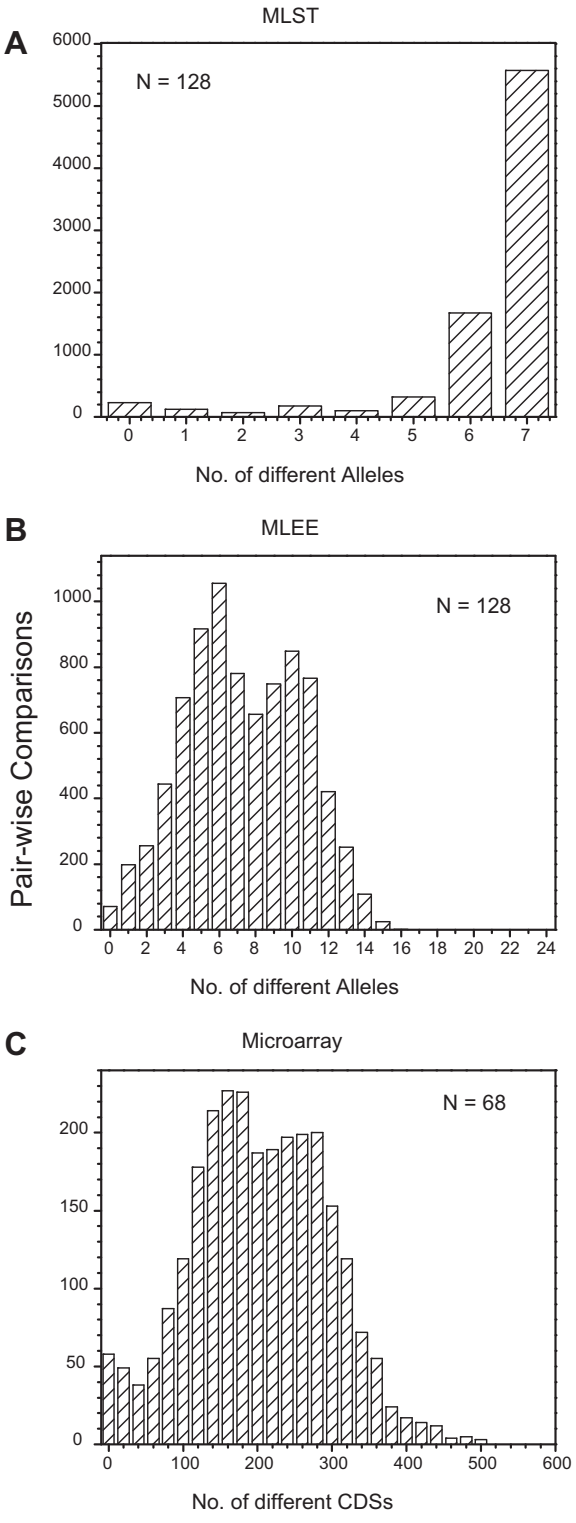

Fig. 1. Histograms of pair-wise genetic distances between unique isolates within SARA and SARB. (A) MLST data for 128 strains for seven gene fragments. (B) MLEE data for 128 strains for 24 enzyme loci. (C) Microarray hybridization data from 68 strains.

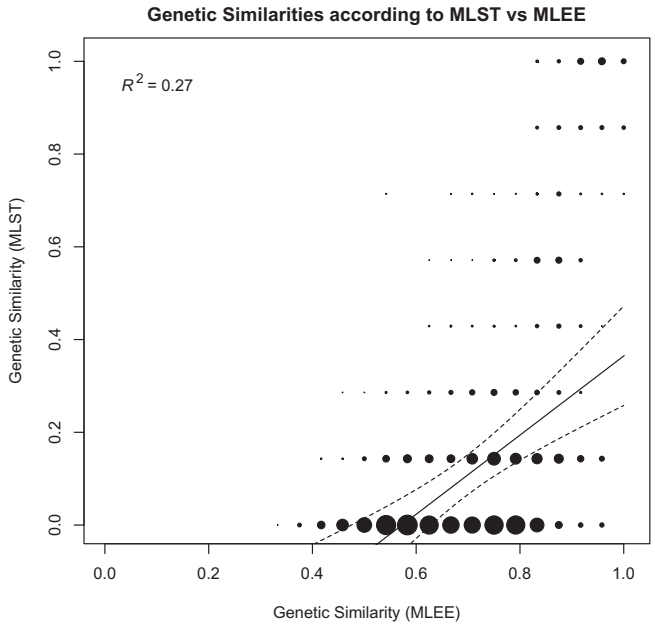

Fig. 2. Genetic similarities between pairs of strains according to MLST *versus* MLEE. Genetic similarities were calculated from the data in Fig. 1A and B, except that those values were normalized as a fraction of the number of loci tested and then subtracted from 1.0. Circle sizes reflect the number of values. The solid line indicates a linear regression of MLST similarity on MLEE similarity with 95% confidence intervals indicated by dashed lines. Calculations were performed in *R*.

Almost all ST clusters in Fig. 4 were uniform in serovar, and also consistent with the serovars of additional isolates in those eBGs on the MLST website. For example, Sp1 (Saintpaul), Sp2 and Sp3 belong to eBG14, as does Ha1 (Haifa). According to the MLST website, eBG14 is largely composed of Saintpaul isolates but also contains a second Haifa isolate. Similarly, Mu4 and Mu4a were found to be Manhattan rather than Muenchen. They belong to eBG27, which contains multiple other Manhattan isolates. In contrast, the Muenchen isolates of Mu1 and Mu1a belong to eBG8, which contains multiple other Muenchen isolates. Indeed, almost all of the misserotyped isolates belong to eBGs or STs in which additional isolates are consistent in serovar with our reconstructions of SARA and SARB.

SARA and SARB represent the maximal diversity obtained by MLEE from a much larger group of isolates. Our comparison of MLEE with MLST is thus a comparison between the maximal MLEE diversity with random MLST diversity. It is therefore not surprising that MLST did not distinguish between some closely related ET SLVs. For example, 12 Typhimurium ETs all belong to ST19 (Table S2). However, in some cases, MLST had higher resolution than MLEE, such as Mu1, which was sub-divided into ST82, ST111 and ST112.

3.3. Comparison of MLEE and MLST to microarray hybridization

Porwollik et al. (2004) performed microarray hybridization analysis of strains from SARA and SARB, as well as from other *enterica*. We performed MLST on 74 of these isolates to compare the results between microarrays, MLEE and MLST. The microarrays were based on PCR products of annotated CDSs from genomes of serovars Typhimurium (strain LT2) and Typhi (CT18) (see Materials and Methods). We excluded six erroneous or superfluous strains from further analysis (Supplementary Table S4): Three strains were excluded because they belonged to pairs of isolates from the SARB collections of Boyd and Porwollik which had been separately stored in distinct laboratories for over two decades. Each pair differed by ⩽23 CDSs. In contrast, two versions of SARB52 differed by 271 CDS calls although the strains had identical STs and serovars, suggesting that one of the DNA preparations used for the microarray experiments was from an unrelated bacterial strain. We also excluded two other strains which probably also reflect DNA mix-ups, where the resulting microarray data suggested this

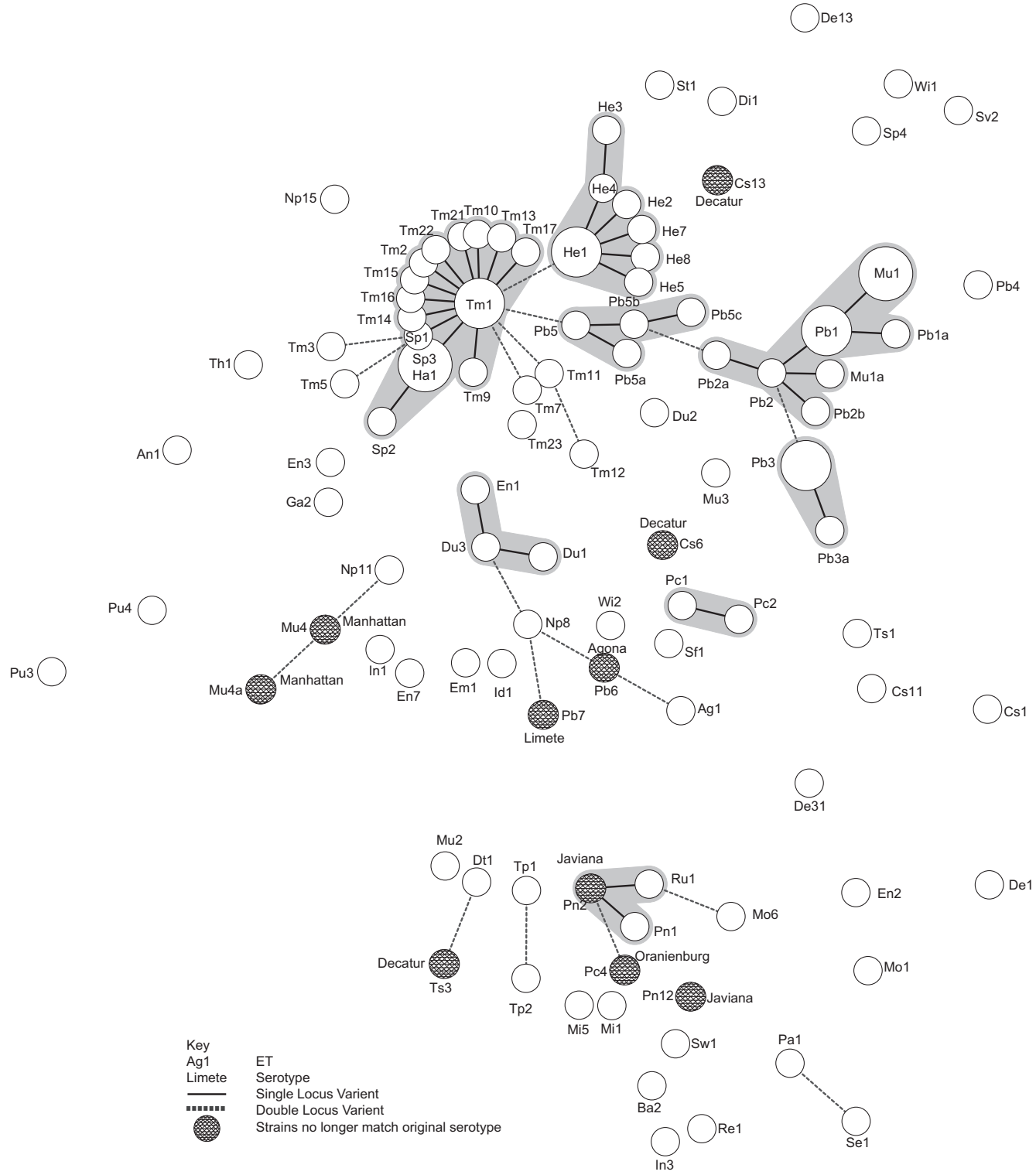

Fig. 3. Minimal spanning tree of MLEE relationships among 128 unique strains in SARA and SARB. eBurst-like groupings based on chains of single locus variants (SLVs) are indicated by grey shading. Each circle indicates a distinct ET whose ET designation is indicated within the circle, or directly next to it. Circle sizes reflect numbers of strains. Cross-hatched circles indicate strains with the indicated serovars, which do not match the original published information. Short black lines joining ET circles indicate SLVs whereas double locus variants (DLVs) are indicated by dashed lines.

strain to be nearly identical to an unrelated isolate with different ET and ST assignments (Table S4). After excluding these six sets of microarray data, 68 isolates remained which had been tested by both MLST and microarray, including 42 that had also been tested by MLEE.

The maximum number of CDSs that differed between each pair was less than 500 (~13% of CDSs tested), and most pairs of strains differed by many fewer CDSs (Fig. 1C). The boundary between identical and distinct strains was somewhat fuzzy, and there is no clear trough between closely related and more distant

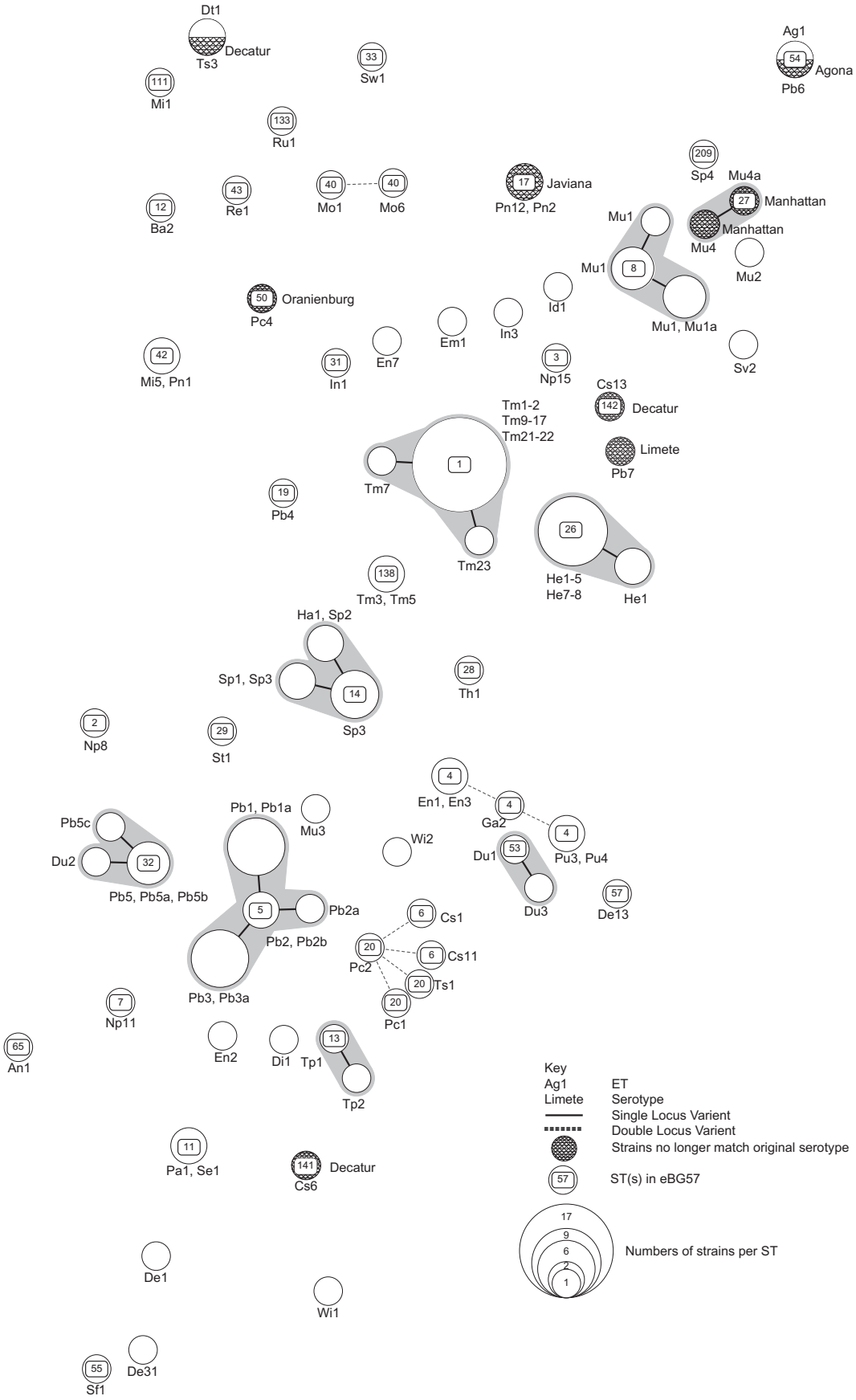

Fig. 4. Minimal spanning tree of MLST relationships among 128 unique strains in SARA and SARB. Data are depicted as in Fig. 3 with the following modifications. Each circle indicates an ST. eBG assignments are indicated by numbers within rounded rectangles. For eBGs containing multiple SLVs, the eBG designation is only indicated once within a central ST. Other singleton STs have not been assigned to an eBG.

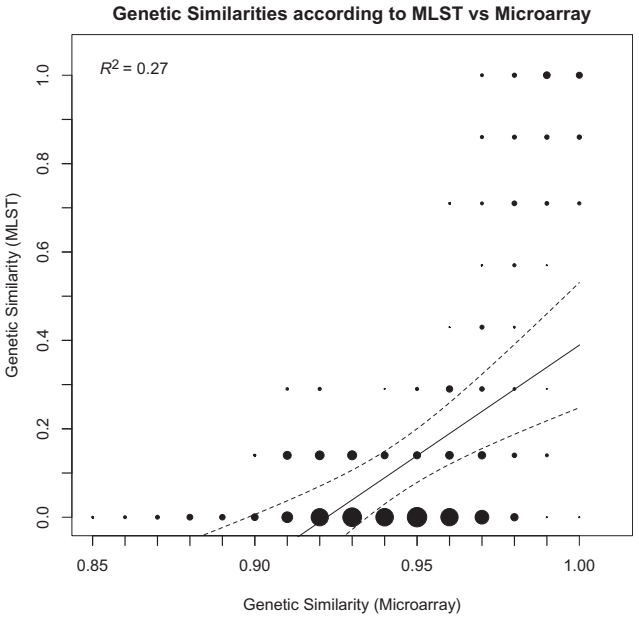

Fig. 5. Genetic similarities between pairs of strains according to MLST *versus* microarray hybridization. These comparisons are based on 68 strains which had been tested by both methods. Other details are as in Fig. 2.

relationships in the frequency distribution of genetic distance according to microarray hybridization. At least 85% of CDSs are present or absent in both strains in pair-wise comparisons among pairs of strains that are distinct by MLST (Fig. 5). As in the comparison between MLEE and MLST, the genetic similarities obtained by both MLST and microarrays are significantly correlated but the quantitative level of correlation is only intermediate (Mantel test, $R^2 = 0.27$, $p < e^{-4}$).

An MSTree of the microarray results showed clusters of related isolates within the serovars Enteritidis, Typhimurium, Montevideo, Muenchen, Paratyphi A and Paratyphi B (Fig. 6). It also indicated moderately close relationships between Choleraesuis (eBG6), Typhisuis and Paratyphi C (eBG20); between Enteritidis, Pullorum and Gallinarum (eBG4); between Paratyphi A and Sendai (eBG11); and between Typhimurium (eBG1), Saintpaul (eBG14) and Heidelberg (eBG26). However, these relationships were intermingled with other closely related pair-wise distances of strains that are thought be unrelated by other methods. Thus, the microarray calls supported the assignments to eBGs by MLST but did not distinguish as strongly between the different serovars as MLST.

The similarities between MLST, MLEE and microarrays are supported by pair-wise comparisons of the genetic similarities between these methods for all 42 strains which had been tested by all three methods (Supporting Figure S1). The strongest correlation in the three-way comparison was for MLEE *versus* microarrays (Table S5), but even that correlation was still quite low ($R^2 = 0.18$), and the pattern corresponded more to a broad cloud of points rather than fitting tightly to a straight line. Thus, the quantitative similarity measurements differ dramatically between all three methods. We note that higher correlations were observed with larger sample sizes, possibly due to the inclusion of more closely related isolates, but none of the comparisons yielded a very strong correlation.

4. Discussion

At the time MLEE was implemented for population genetic analyses of bacteria, populations were distinguished on the basis of somewhat arbitrary cut-offs based on branch lengths within phylogenetic trees. Phylogenetic trees are still being used in a similar fashion to detect relationships between *enterica* genomes (Leekitcharoenphon et al., 2012; Jacobsen et al., 2011; den Bakker et al., 2011). However, branch lengths within phylogenies are distorted when recombination is frequent, and frequent recombination also blurs signals of deep genealogical relationships. Indeed, recombination is so frequent in *enterica* that it retains few signs of deeper phylogenetic signals even when extensive sequence datasets are compared (Didelot et al., 2011). We therefore used eBurst grouping by a minimal spanning tree to compare MLEE, microarray hybridization and MLST for their relative abilities to assign individual isolates to populations. For clonal lineages with fully parsimonious genealogies, similar topologies are indicated by MSTrees and maximum parsimony or maximum likelihood analyses (Cui et al., 2012), except that eBurst groupings are more intuitive for visual interpretations. eBGs continue to recognize populations even when phylogenetic methods fail due to frequent recombination (Wirth et al., 2006; Turner et al., 2007). Furthermore, eBGs are particularly effective when 1000's of bacteria are analyzed (Achtman et al., 2012), a task that is difficult to handle with many phylogenetic methods, or at least very difficult to visualize. Based on the eBurst approach, overlapping groupings were obtained by all three methods, confirming that these represent natural evolutionary groupings, independent of the source of data. MLST was superior to MLEE and microarray hybridization for its consistency in recognizing the serovar-specific eBGs within *enterica* but is considerably poorer for recognizing higher-order relationships such as those that exist between serovars Typhimurium and Saintpaul. However, although both MLEE and microarray hybridization clustered Typhimurium and Saintpaul, they also spuriously clustered other groups of organisms, and differed from each other in the clusters that they recognized. We conclude that none of these methods is suitable for reliably detecting higher order relationships, and suggest that a better understanding of the genealogy of *enterica* will only be possible after the reconstruction of recombination events in analyses of full genomes. Such methods have recently been developed for deducing the details of speciation events (Abby et al., 2012; Szöllösi et al., 2012), but have not yet been applied for genealogical investigations within species.

Our results also show that modern copies of the SARA and SARB collections do not fully reflect the published descriptions of their serovars and relationships (Beltran et al., 1991; Boyd et al., 1993), and that a few strains also differ between individual collections. These problems are not surprising. A few percent of the *enterica* strains supplied to us by other laboratories have been incorrect (Achtman et al., 2012), and we were unable to ensure uncontaminated cultures in our own hands until we implemented robotically assisted microbiology within a safety cabinet (O'Farrell et al., 2012). The reconstructed compositions of SARA and SARB (Tables 3 and 4), and the information provided here about their MLST STs, can be used for quality control purposes by laboratories who use SARA and/or SARB as representatives of *enterica*. Our detailed listing of mixed strains presented here will also support the reconstruction of potentially confusing results in the literature that resulted from such strain mix-ups. Finally, the MLST data of SARA and SARB now allow a comparison of those collections within the context of the 1000's of strains that have been MLST typed (Achtman et al., 2012), and may help with the choice of strains for future genomic analyses.

Acknowledgements

We gratefully acknowledge the receipt of strains from Mia Torpdahl, information from Kenneth E. Sanderson, MLST data from Howard Ochman, serotyping by and information from Wolfgang

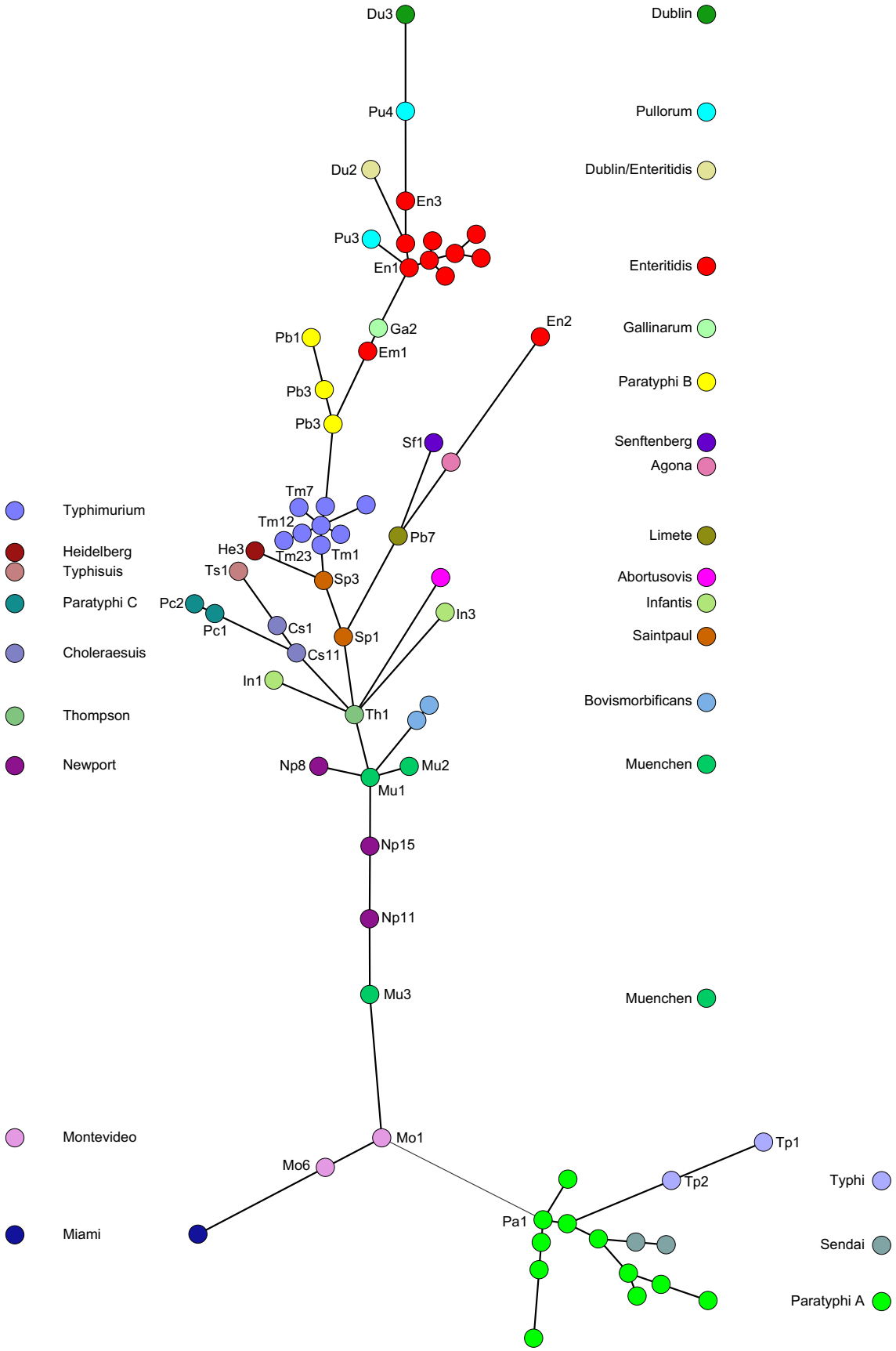

Fig. 6. Minimal spanning tree of distances according to numbers of hybridizing CDSs in microarray analyses of 68 isolates. Circles correspond to individual strains of the ET types indicated, color coded by serovar as indicated at the right and left. Circles without ET types have not been subjected to MLEE. The lengths of the lines are roughly proportional to the numbers of differences in presence/absence scores for CDSs between pairs of strains. Note that the designation Em1 for strain SARB20 is used even though microarrays were performed with strain SARB19 which is serovar Enteritidis (Table 2).

Rabsch and Francois-Xavier Weill, and assistance in the preparation of figures in R by Jana K. Haase, Angela McCann and Thibaut Jombard. This work was supported by grants from the Science Foundation of Ireland (05/FE1/B882) to MA and NIH grants AI034829, AI034829 and AI075093 to SP.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2013.03.003.

References

Aanensen, D.M., Spratt, B.G., 2005. The multilocus sequence typing network: mlst.net. Nucleic Acids Res. 33, W728–W733.

Abby, S.S., Tannier, E., Gouy, M., Daubin, V., 2012. Lateral gene transfer as a support for the tree of life. Proc. Natl. Acad. Sci. USA 109, 4962–4967.

Achtman, M., Wain, J., Weill, F.-X., Nair, S., Zhou, Z., Sangal, V., Krauland, M.G., Hale, J.L., Harbottle, H., Uesbeck, A., Dougan, G., Harrison, L.H., Brisse, S., 2012. *S. enterica* MLST study group, multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. PLoS Pathog. 8, e1002776.

Beltran, P., Musser, J.M., Helmuth, R., Farmer III, J.J., Frerichs, W.M., Wachsmuth, I.K., Ferris, K., McWhorter, A.C., Wells, J.G., Cravioto, A., Selander, R.K., 1988. Toward a population genetic analysis of *Salmonella*: Genetic diversity and relationships among strain of serotypes *S. choleraesuis, S. derby, S. dublin, S. enteritidis, S. heidelberg, S. infantis, S. newport, and S. typhimurium*. Proc. Natl. Acad. Sci. USA 85, 7753–7757.

Beltran, P., Plock, S.A., Smith, N.H., Whittam, T.S., Old, D.C., Selander, R.K., 1991. Reference collection of strains of the *Salmonella typhimurium* complex from natural populations. J. Gen. Microbiol. 137, 601–606.

Boyd, E.F., Nelson, K., Wang, F.-S., Whittam, T.S., Selander, R.K., 1994. Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica*. Proc. Natl. Acad. Sci. USA 91, 1280–1284.

Boyd, E.F., Wang, F.S., Beltran, P., Plock, S.A., Nelson, K., Selander, R.K., 1993. *Salmonella* reference collection B (SARB): strains of 37 serovars of subspecies I. J. Gen. Microbiol. 139, 1125–1132.

Boyd, E.F., Wang, F.S., Whittam, T.S., Selander, R.K., 1996. Molecular genetic relationships of the Salmonellae. Appl. Environ. Microbiol. 62, 804–808.

Brown, E.W., Kotewicz, M.L., Cebula, T.A., 2002. Detection of recombination among *Salmonella enterica* strains using the incongruence length difference test. Mol. Phylogenet. Evol. 24, 102–120.

Cui, Y., Yu, C., Yan, Y., Li, D., Li, Y., Jombart, T., Weinert, L.A., Wang, Z., Guo, Z., Xu, L., Zhang, Y., Zheng, H., Qin, N., Xiao, X., Wu, M., Wang, X., Zhou, D., Qi, Z., Du, Z., Wu, H., Yang, X., Cao, H., Wang, H., Wang, J., Yao, S., Rakin, A., Li, Y., Falush, D., Balloux, F., Achtman, M., Song, Y., Wang, J., Yang, R., 2012. Historical variations in mutation rate in an epidemic pathogen, *Yersinia pestis*. Proc. Natl. Acad. Sci. USA, 110.

den Bakker, H.C., Moreno Switt, A.I., Cummings, C.A., Hoelzer, K., Degoricija, L., Rodriguez-Rivera, L.D., Wright, E.M., Fang, R., Davis, M., Root, T., Schoonmaker-Bopp, D., Musser, K.A., Villamil, E., Waechter, H., Kornstein, L., Furtado, M.R., Wiedmann, M., 2011. A whole genome SNP based approach to trace and identify outbreaks linked to a common Salmonella enterica subsp. enterica serovar Montevideo Pulsed Field Gel Electrophoresis type. Appl. Environ. Microbiol. 77, 8648–8655.

Didelot, X., Bowden, R., Street, T., Golubchik, T., Spencer, C., McVean, G., Sangal, V., Anjum, M.F., Achtman, M., Falush, D., Donnelly, P., 2011. Recombination and population structure in *Salmonella enterica*. PLoS Pathog. 7, e1002191.

Didelot, X., Maiden, M.C., 2010. Impact of recombination on bacterial evolution. Trends Microbiol. 18, 315–322.

Fabre, L., Zhang, J., Guigon, G., Le Hello, S., Guibert, V., Accou-Demartin, M., de Romans, S., Lim, C., Roux, C., Passet, V., Dianvourt, L., Guibourdenche, M., Issenhuth-Jeanjean, S., Achtman, M., Brisse, S., Sola, C., Weill, F.-X., 2012. CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. PLoS ONE 7, e36995.

Falush, D., Torpdahl, M., Didelot, X., Conrad, D.F., Wilson, D.J., Achtman, M., 2006. Mismatch induced speciation in *Salmonella*: model and data. Philos. Trans. R. Soc. Lond B Biol. Sci. 361, 2045–2053.

Feil, E.J., 2010. Linkage, selection and clonal complex. In: Robinson, D.A., Falush, D., Feil, E.J. (Eds.), Bacterial Population Genetics in Infectious Disease. John Wiley & Sons, Inc., Hoboken, NJ, pp. 19–35.

Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P., Spratt, B.G., 2004. EBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from Multilocus Sequence Typing data. J. Bacteriol. 186, 1518–1530.

Grimont, P.A., Weill, F.-X., 2007. Antigenic formulae of the Salmonella serovars, nineth ed. WHO Collaborating Centre for Reference and Research on *Salmonella*, Paris, France.

Haase, J.K., Brown, D.J., Weill, F.X., Mather, H., Foster, G., Brisse, S., Wain, J., Achtman, M., 2012. Population genetic structure of 4,12:a: – *Salmonella enterica* from harbour porpoises. Appl. Environ. Microbiol. 78, 8829–8833.

Hanage, W.P., Fraser, C., Spratt, B.G., 2006a. Sequences, sequence clusters and bacterial species. Philos. Trans. R. Soc. Lond B Biol. Sci. 361, 1917–1927.

Hanage, W.P., Fraser, C., Spratt, B.G., 2006b. The impact of homologous recombination on the generation of diversity in bacteria. J. Theor. Biol. 239, 210–219.

Hauser, E., Tietze, E., Helmuth, R., Junker, E., Prager, R., Schroeter, A., Rabsch, W., Fruth, A., Toboldt, A., Malorny, B. 2012. Clonal Dissemination of Salmonella enterica serovar Infantis in Germany. Foodborne. Pathog. Dis. 9.

Jacobsen, A., Hendriksen, R.S., Aaresturp, F.M., Ussery, D.W., Friis, C., 2011. The *Salmonella enterica* Pan-genome. Microb. Ecol. 62, 487–504.

Leekitcharoenphon, P., Lukjancenko, O., Friis, C., Aarestrup, F.M., Ussery, D.W., 2012. Genomic variation in *Salmonella enterica* core genes for epidemiological typing. BMC Genomics 13, 88.

Liu, W.Q., Feng, Y., Wang, Y., Zou, Q.H., Chen, F., Guo, J.T., Peng, Y.H., Jin, Y., Li, Y.G., Hu, S.N., Johnston, R.N., Liu, G.R., Liu, S.L., 2009. *Salmonella paratyphi* C: Genetic divergence from *Salmonella choleraesuis* and pathogenic convergence with *Salmonella typhi*. PLoS ONE 4, e4510.

Maiden, M.C., 2006. Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. 60, 561–588.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA 95, 3140–3145.

O'Farrell, B., Haase, J.K., Velayudhan, V., Murphy, R.A., Achtman, M., 2012. Transforming microbial genotyping: A robotic pipeline for genotyping bacterial strains. PLoS ONE 7, e48022.

Porwollik, S., Boyd, E.F., Choy, C., Cheng, P., Florea, L., Proctor, E., McClelland, M., 2004. Characterization of *Salmonella enterica* subspecies I genovars by use of microarrays. J. Bacteriol. 186, 5883–5898.

Porwollik, S., Mei-Yi, W.R., McClelland, M., 2002. Evolutionary genomics of *Salmonella*: Gene acquisitions revealed by microarray analysis. Proc. Natl. Acad. Sci. USA 99, 8956–8961.

R Development Core Team. 2004. R: a language and environment for statistical computing.

Selander, R.K., Beltran, P., Smith, N.H., Helmuth, R., Rubin, F.A., Kopecko, D.J., Ferris, K., Tall, B.T., Cravioto, A., Musser, J.M., 1990. Evolutionary genetic relationships of clones of *Salmonella* serovars that cause human typhoid and other enteric fevers. Infect. Immun. 58, 2262–2275.

Selander, R.K., Caugant, D.A., Ochman, H., Musser, J.M., Gilmour, M.N., Whittam, T.S., 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl. Environ. Microbiol. 51, 873–884.

Selander, R.K., Musser, J.M., Caugant, D.A., Gilmour, M.N., Whittam, T.S., 1987. Population genetics of pathogenic bacteria. Microb. Pathog. 3, 1–7.

Szöllösi, G.J., Boussau, B., Abby, S.S., Tannier, E., Daubin, V., 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc. Natl. Acad. Sci. USA 109, 17513–17518.

Thomson, N.R., Clayton, D.J., Windhorst, D., Vernikos, G., Davidson, S., Churcher, C., Quail, M.A., Stevens, M., Jones, M.A., Watson, M., Barron, A., Layton, A., Pickard, D., Kingsley, R.A., Bignell, A., Clark, L., Harris, B., Ormond, D., Abdellah, Z., Brooks, K., Cherevach, I., Chillingworth, T., Woodward, J., Norberczak, H., Lord, A., Arrowsmith, C., Jagels, K., Moule, S., Mungall, K., Sanders, M., Whitehead, S., Chabalgoity, J.A., Maskell, D., Humphrey, T., Roberts, M., Barrow, P.A., Dougan, G., Parkhill, J., 2008. Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome. Res. 18, 1624–1637.

Toboldt, A., Tietze, E., Helmuth, R., Fruth, A., Junker, E., Malorny, B., 2012. Human *Salmonella enterica* serovar Paratyphi B (dT+) infections in Germany originate from reptiles and rarely from poultry. Appl. Environ. Microbiol. 78, 7347–7357.

Torpdahl, M., Ahrens, P., 2004. Population structure of *Salmonella* investigated by amplified fragment length polymorphism. J. Appl. Microbiol. 97, 566–573.

Turner, K.M., Hanage, W.P., Fraser, C., Connor, T.R., Spratt, B.G., 2007. Assessing the reliability of eBURST using simulated populations with known ancestry. BMC Microbiol. 7, 30.

Uzzau, S., Brown, D.J., Wallis, T., Rubino, S., Leori, G., Bernard, S., Casadesus, J., Platt, D.J., Olsen, J.E., 2000. Host adapted serotypes of *Salmonella enterica*. Epidemiol. Infect. 125, 229–255.

Uzzau, S., Hovi, M., Stocker, B.A., 1999. Application of ribotyping and IS*200* fingerprinting to distinguish the five *Salmonella* serotype O6,7:c:1,5 groups: Choleraesuis *sensu stricto*, Choleraesuis var. Kunzendorf, Choleraesuis var. Decatur, Paratyphi C, and Typhisuis. Epidemiol. Infect 123, 37–46.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., Karch, H., Reeves, P.R., Maiden, M.C., Ochman, H., Achtman, M., 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. 60, 1136–1151.