JOURNAL OF BACTERIOLOGY, Sept. 2004, p. 5883–5898 Vol. 186, No. 17
0021-9193/04/$08.00+0 DOI: 10.1128/JB.186.17.5883–5898.2004

Characterization of *Salmonella enterica* Subspecies I Genovars by Use of Microarrays

S. Porwollik, E. F. Boyd, C. Choy, P. Cheng, L. Florea, E. Proctor, and M. McClelland*

Sidney Kimmel Cancer Center, San Diego, California

Received 4 March 2004/Accepted 26 May 2004

Subspecies 1 of *Salmonella enterica* is responsible for almost all *Salmonella* infections of warm-blooded animals. Within subspecies 1 there are over 2,300 known serovars that differ in their prevalence and the diseases that they cause in different hosts. Only a few of these serovars are responsible for most *Salmonella* infections in humans and domestic animals. The gene contents of 79 strains from the most prevalent serovars were profiled by microarray analysis. Strains within the same serovar often differed by the presence and absence of hundreds of genes. Gene contents sometimes differed more within a serovar than between serovars. Groups of strains that share a distinct profile of gene content can be referred to as "genovars" to distinguish them from serovars. Several misassignments within the *Salmonella* reference B collection were detected by genovar typing and were subsequently confirmed serologically. Just as serology has proved useful for understanding the host range and pathogenic manifestations of *Salmonella*, genovars are likely to further define previously unrecognized specific features of *Salmonella* infections.

The bacterial genus *Salmonella* is divided into two species, *Salmonella bongori* and *S. enterica*. *S. enterica* itself is comprised of six subspecies: they are *S. enterica* subsp. *enterica*, *S. enterica* subsp. *salamae*, *S. enterica* subsp. *arizonae*, *S. enterica* subsp. *diarizonae*, *S. enterica* subsp. *indica*, and *S. enterica* subsp. *houtenae*, or I, II, IIIa, IIIb, IV, and VI, respectively (21). Of these six subspecies, only subspecies I is associated with disease in warm-blooded animals. To date, there are over 2,300 serovars identified within subspecies I. However, only a small fraction of the thousands of described subspecies I serovars frequently cause disease in humans and domestic animals. For example, the annual report of the Centers for Disease Control and Prevention (CDC) for the year 2001 registered 360 different serovars in human infections in the U.S. Approximately 50% of these infections were caused by only three *Salmonella* serovars, specifically Typhimurium, Enteritidis, and Newport. The 12 most prevalent *Salmonella* serovars were responsible for >70% of all human *Salmonella* infections (http://www.cdc.gov/ncidod/dbmd/phlisdata/salmonella.htm). Similarly, 41.8% of all veterinary infections were attributed to only two *Salmonella* serovars, namely, Typhimurium and Newport. The 10 most prevalent veterinary serovars caused 70% of all infections.

The *Salmonella* reference B (SARB) collection of *Salmonella* subspecies I strains represents 72 protein electrophoretic types (ETs) within 37 medically important serovars selected to embody the maximum diversity within subspecies I (4). These ETs were determined by multilocus enzyme electrophoresis (MLEE), a technique that reveals the presence and anodal mobility of enzymes (26). A total of 24 enzymes were surveyed in several thousand strains to establish the SARB set. For 19 of the 37 serovars included in the SARB collection, more than one MLEE type was found. In these cases, the most prevalent type was included in the SARB set together with less prevalent types that were the most different from the common type. A genetic distance tree constructed with these data showed that several serovars (including Dublin, Enteritidis, Infantis, Muenchen, Newport, and Saint Paul) were apparently of polyphyletic origin, while others, such as Heidelberg, Montevideo, Typhi, and Typhimurium, clustered together and were therefore monophyletic. ETs occurred at different frequencies. For example, serovar Typhimurium is represented by four different MLEE types in the SARB collection, termed Tm1, Tm7, Tm12, and Tm23. Whereas Tm1 was the most prevalent type and was detected in 258 isolates, Tm7 and Tm23 were only detected in two strains and Tm12 was detected in 27 isolates (4).

Comparative genomic hybridization using microarray technology has been extensively employed to monitor the gene contents of closely related bacterial species (reviewed in references 10 and 13). Differences in the genetic repertoire within the different *Salmonella* subspecies, including the divergence of the different subspecies of the salmonellae, have been investigated in two recent studies by use of a *Salmonella* microarray chip (7, 23). This report now concentrates on the differences between *Salmonella* isolates that belong to subspecies I and supplements these previous studies to include all of the most medically relevant serovars of *S. enterica*. Besides representing an overview of the extensive genetic variations found between these isolates, we confirm that *Salmonella* strains of the same serovar are not always genotypically closely related, and those differences are characterized at single-gene resolution. While several isolates of subspecies I serovars have been previously genotyped, we can now describe the sometimes remarkable diversity between isolates in the same serovar. We propose that *Salmonella* genovars may be a useful description for certain strain characteristics within a serovar. Genovars, which classify strains within a species on the basis of gene

* Corresponding author. Mailing address: Sidney Kimmel Cancer Center, 10835 Altman Row, San Diego, CA 92103. Phone: (858) 450-5990, ext. 280. Fax: (858) 550-3998. E-mail: mmcclelland@skcc.org.

content, are different from genomovars; "genomovar" is a term that has been used to describe similarities among species that are phylogenetically distinguishable from each other but which are phenotypically indistinguishable.

MATERIALS AND METHODS

Strain and microarray specifications. Details about the strains employed in this study are shown in Table 1. We used a *Salmonella*-specific microarray that represented PCR-amplified sequences from the annotated open reading frames (ORFs) of *S. enterica* serovar Typhimurium LT2 supplemented with annotated chromosomal ORFs from the serovar Typhi CT18 strain that were >10% divergent from those of serovar Typhimurium (22). The overall *S. enterica* serovar Typhimurium genome coverage for the array was 96.6% (4,338 genes), and the overall coverage of the *S. enterica* serovar Typhi genome was 94.5% (4,348 genes), excluding plasmids. The array also contained PCR products representing the genes found on the LT2 virulence plasmid pSLT and the ORFs of R46, a resistance plasmid present in various enterobacteria. The DNAs were spotted onto Ultra-GAPS glass slides (Corning Inc., Corning, N.Y.) in 50% dimethyl sulfoxide.

DNA labeling. Genomic DNAs of serovar Typhimurium LT2, serovar Typhi CT18 and TY2, and the SARB *S. enterica* strains were prepared from fresh overnight cultures by the use of either GenElute bacterial genomic DNA kits (Sigma, St. Louis, Mo.) or DNEasy kits (Qiagen, Valencia, Calif.) according to the manufacturer's instructions. Cells were grown in Luria broth at 37°C. The harvested nucleic acids were labeled according to P. Brown's protocol (http://cmgm.stanford.edu/pbrown/protocols/4_genomic.html) with 12 μg of random hexamers (Sigma Genosys, The Woodlands, Tex.), 10 U of Klenow enzyme (New England Biolabs, Beverly, Mass.), and 2 nmol of Cy3-dCTP (Amersham, Piscataway, N.J.) for 16 h at 37°C. Serovar Typhimurium LT2 genomic DNA was labeled with Cy5-dCTP. Probes were purified with a Qiaquick PCR purification kit (Qiagen) as suggested by the manufacturer, eluted in 1 mM Tris-HCl, pH 8.0, dried, and resuspended in 20 μl of sterile water.

DNAs from recent clinical *S. enterica* isolates were embedded in plugs of 2% LMP agarose and stored in 1 mM Tris-HCl, pH 8.0. For labeling, a modification of the Brown procedure was employed. Briefly, a plug was separated from the storage buffer and solubilized at 62°C for 10 min. Then 21 μl of the solubilized plug containing the genomic DNA was used directly in the labeling reaction, without any further modifications. Probes were subsequently purified by the addition of 175 μl of buffer QG from a Qiaquick gel extraction kit (solubilization buffer) and 10 μl of 3 M sodium acetate, pH 5.2. After mixing, 55 μl of isopropanol was added and the suspensions were loaded onto the standard Qiaquick PCR purification columns from which they were retrieved as described above.

Hybridization and data acquisition. Immediately before use, the labeled probes for serovar Typhimurium LT2 (control sample) and one of the query *S. enterica* strains (experimental sample) were unified, mixed with 40 μl of 2× hybridization buffer (50% formamide, 10× SSC [1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate], 0.2% sodium dodecyl sulfate), and boiled for 5 min. Standard protocols for hybridization in formamide buffer (Corning instruction manual, Corning Inc.) were applied for prehybridization, hybridization, and posthybridization wash processes. A ScanArray 5000 laser scanner (Packard BioChip Technologies, Billerica, Mass.) was employed for image acquisition with ScanArray 2.1 software. Signal intensities were quantified with the QuantArray 3.0 software package (Packard BioChip Technologies).

Data analysis. Spot signal intensities were measured by adaptive quantitation. The local background was subtracted from the recorded spot intensities, and data were normalized by determination of the contribution of every spot to the total signal in that channel. Ratios of the contributions were calculated. Negative values (i.e., local background intensities higher than the spot signal) were considered no data. Since the array was spotted in triplicate, a single hybridization resulted in three data points per gene, and the median of the three ratios per gene was reported.

The presence or absence of the *S. enterica* serovar Typhimurium LT2 genes in the other *S. enterica* genomes was evaluated based on a comparison of normalized hybridization signal ratios of the query strain to serovar Typhimurium LT2 for the respective gene spot. Genes that displayed a ratio of >0.67 and which in addition were neighbored on the LT2 genome by elements that also displayed ratios of >0.67 were included in the calculation of the presence baseline *P*. *P* was set to be the median of the ratios for this set of genes. The standard deviation (SD_P) of these ratios was calculated for each query strain. Similarly, medians and SDs for genes with ratios of <0.5 which were neighbored by elements with ratios of <0.5 were also determined (absence baseline *A* and SD_A, respectively). Ratios which were higher than the presence threshold, set at 2 SD_P below the baseline *P*, were scored as "present," whereas genes with ratios lower than the absence threshold, set at 2 SD_A above the baseline *A*, were scored as "absent." Genes that were outside of these thresholds and those that displayed ratios between 0.5 and 0.67 were scored as "uncertain." Genes with signals that were among the lowest 5% of all LT2 genes for the control sample (serovar Typhimurium LT2) were considered missing data.

The array also contained several plasmid genes, based on the LT2 virulence plasmid pSLT, which is present in some serovar Typhimurium strains, and the resistance plasmid R46. In addition, 471 serovar Typhi-specific elements that are not present in serovar Typhimurium LT2 were also represented on the array. For these elements, presence in the query strain was assumed if the median signal strength of the respective spot was among the top 70% of all DNA spots on the chip (including LT2 genes). The lowest 20% of all signals were assigned to the "absent" category. If signal strengths were ranked between these thresholds, the genes were scored as "uncertain."

Presence and absence predictions for genes were also performed for genome sequence data obtained for different *Salmonella* serovars. This predictor calculated the highest percent similarity over a 100-bp window of all chip sequences representing LT2 chromosomal genes to the sequenced genome and calculated the 75th percentile of these values (P_{75}, which was usually 100, except for *S. bongori*, for which P_{75} equaled 99). Array elements that displayed similarities equal to or higher than $P_{75} - 5$ were considered to be present and those with similarities that were lower than $P_{75} - 15$ were considered to be absent. The remaining values were attributed to the "uncertain" category.

Phylogenetic trees. The predictions obtained for every gene for each strain investigated (0 = absent, 1 = uncertain, "?" = missing data, 2 = present) were incorporated into the PAUP software program (http://paup.csit.fsu.edu) as previously described (23). For tree building, the highly mobile prophage regions of both the serovar Typhimurium LT2 and serovar Typhi CT18 genomes were excluded from the data set. A more condensed matrix was also employed in which regions rather than single genes were used in order to better approximate the number of insertion-deletion events that caused the observed diversity. In these cases, predictions included five different categories, as follows: 0, absent; 1, primarily absent; 2, uncertain; 3, primarily present; 4, present.

Array data accession number. The data presented here have been deposited at the GEO database of the National Center for Biotechnology at http://www.ncbi.nlm.nih.gov/geo under series number GSE1035.

RESULTS

We characterized the genetic contents of recent clinical isolates of the most prevalent *S. enterica* serovars by comparative genomic hybridization to a microarray representing almost all annotated ORFs of both the serovar Typhimurium LT2 and the serovar Typhi CT18 isolates. Genomic DNAs from recent clinical isolates were obtained for every serovar representing the 12 most common clinical and the 10 most common veterinary isolates in the U.S. in 2001 according to the 2001 annual report of the CDC (http://www.cdc.gov/ncidod/dbmd/phlisdata/salmonella.htm). The collection included serovars Typhimurium, Enteritidis, Newport, Heidelberg, Javiana, Montevideo, Oranienburg, Muenchen, Thompson, Saint Paul, Java, and Infantis as well as serovars Agona, Cholerasuis, Senftenberg, Muenster, and Dublin. Profiling was also performed on representatives of the same serovars from the SARB collection, a set of *S. enterica* isolates collected more than 10 years ago (4), as well as on two serovar Abortusovis strains and SARB isolates for the rarer serovars Paratyphi B, Paratyphi C, Sendai, Gallinarum, and Typhisuis (Table 1).

Overall, 867 Typhimurium LT2 chromosomal genes (21% of all annotated LT2 ORFs, excluding those for tRNAs and rRNAs) were found to be absent (or to have no close homologue) from at least one isolate of this representative set of subspecies I strains and reliably present in other strains. Figure 1 depicts the status of these polymorphic genes in order of their positions on the LT2 genome in the investigated subspecies I

TABLE 1. *Salmonella* strains used for this study

Salmonella subspecies and serovar (serogroup)	% Human infections[a]	% Veterinary infections[a]	Isolate no.	SARA/SARB/SARC characterization		Name or no. for clinical isolates	Isolate designation
				Name	Occurrence[b]		
S. enterica subspecies I							
Abortusovis (B)			1			15-5	AbA1
			2			SS44	AbA2
Agona (B)	1.2	5	1	SARB1	114		Ag1
			2			022481	AgA1
Choleraesuis (C1)		3.5	1	SARB4	131		Cs1
			2	SARB6	3		Cs11
			3			S1380	CsA1
Dublin (D1)		2	1	SARB12	128		Du1
			2	SARB13	36		Du3
			3	SARB14	5		Du2
			4			011277	DuA1
Enteritidis (D1)	17.7	2.2	1	SARB16	357		En1
			2	SARB18	3		En3
			3	SARB20[c]	1		En7
			4			021834	EnA1
Gallinarum (D1)			1	SARB21	13		Ga2
Heidelberg (B)	5.9	6	1	SARB24[c]	173		He1
			2	SARB23[c]	3		He3
			3	SARA32	173		He1b
			4			024509	HeA1
			5			022477	HeA2
Infantis (C1)	1.4	1.9	1	SARB26	109		In1
			2	SARB27	1		In3
			3			022226	InA1
Java[d] (B)	1.5		1			022007	JaA1
			2			022382	JaA2
Javiana (D1)	3.4		1			024358q	JvA1
Muenster (E1)		2.8	1			021785	MeA1
			2			001186	MeA2
Montevideo (C1)	2	2.6	1	SARB30	38		Mo1
			2	SARB31	3		Mo6
			3			011650	MoA1
			4			002693	MoA2
Muenchen (C2)	1.8	0.9	1	SARB32	46		Mu1
			2	SARB33	19		Mu2
			3	SARB34	4		Mu3
			4			011795	MuA2
Newport (C2)	10	13.6	1	SARB37	228		Np11
			2	SARB36	111		Np8
			3	SARB38	1		Np15
			4			995115	NpA1
			5			994730	NpA2
Oranienburg (C1)	1.9		1			020420	OrA1
			2			020150	OrA2
Paratyphi A (A)			1	SARB42	117		Pa1
Paratyphi B (B)			1	SARB43	139		Pb1
			2	SARB44			Pb3
			3	SARB47			Pb7
			4			PbA1	PbA1
			5			PbA3	PbA3
			6			PbA7	PbA7
Paratyphi C (C1)			1	SARB48	60		Pc1
			2	SARB49	27		Pc2
Sendai (D1)			1	SARB58	1		Se1
Senftenberg (E4)		3.9	1	SARB59	67		Sf1
			2			021998	SfA1
Saint Paul (B)	1.5		1	SARB55	27		Sp3
			2	SARB56	1		Sp4
			3	SARA25	27		Sp3b
			4	SARA27	27		Sp3c
			5	SARA22			Sp1
			6	SARA23			Sp2
			7			021173	SpA1
			8			021964	SpA2
Thompson (C1)	1.6		1	SARB62	8		Th1
			2			024724	ThA1
Typhimurium (B)	22.1	28.2	1	SARB65	258		Tm1
			2	SARB67	27		Tm12
			3	SARB66	2		Tm7

Continued on following page

TABLE 1—*Continued*

Salmonella subspecies and serovar (serogroup)	% Human infections[a]	% Veterinary infections[a]	Isolate no.	SARA/SARB/SARC characterization		Name or no. for clinical isolates	Isolate designation
				Name	Occurrence[b]		
			4	SARB68	2		Tm23
			5			996933	TmA1
			6			000175	TmA2
Typhi (D1)	1.1		1	SARB63	276		Tp1
			2	SARB64	53		Tp2
			3			024513	TpA1
			4			022621	TpA2
			5				CT18
			6				Ty2
Typhisuis (C1)			1	SARB69	4		Ts1
S. enterica subspecies VI			1	SARC14			VI
S. bongori			1	SARC11			Bo

[a] Contribution to all *Salmonella* infections in 2001 in the U.S., according to the CDC 2001 annual report.
[b] Number of isolates of this electrophoretic type found during the establishment of the SARB collection.
[c] SARB misassignments that were corrected.
[d] Now called *S. enterica* serovar Paratyphi B var. L-tartrate(+).

strains. The distribution of homologues is also shown for genes that are present in serovar Typhi CT18 but absent from serovar Typhimurium LT2 and for the genes of the virulence plasmid pSLT and of pKM101, an R46 derivative (16). For plasmid genes, only those that were detected as present in at least one isolate of a subspecies I serovar other than Typhimurium are shown, and for CT18 genes, only those that were detected as present in at least one isolate of a subspecies I serovar other than Typhi are shown.

Polymorphic genes generally occurred in clusters. In total, we noted 85 regions of polymorphic LT2 chromosomal genes. Table 2 lists these regions of two or more continuous genes that were found to be absent from at least one of the *Salmonella* subspecies I strains examined. Groups of genes in these clusters often displayed heterogeneous patterns of presence and absence. For example, the first four genes of the region STM4483-STM4498 were reported to be absent exclusively for the serovar Typhi isolates, whereas the remaining genes of this region were absent from >80% of all strains tested. Therefore, the number of insertion and deletion events that are responsible for the polymorphy of these clusters is probably much higher than the number of clusters itself.

Overall, there were fewer than 60 singular LT2 chromosomal genes that were not part of a cluster of genes that were absent from at least one subspecies I isolate. Among these genes were *ratB*, *envR*, *rfc*, *fhuA*, *avrA*, and *malX*. The distribution pattern of these six genes is also listed in Table 2.

Table 2 also includes gene clusters from the serovar Typhi CT18 genome that were detected in at least one isolate of another serovar and summarizes the presence and absence of the genes of plasmids pSLT and R46. All absence and presence predictions, at single-gene resolution, can be found as supplementary information (supplement A) at http://bioinformatics.skcc.org/mcclelland/salmonella/subspecies1/.

In order to base any conclusions on these results, we needed to be confident of the assignments of the strains that we used for these studies and also confident of the microarray results.

Confidence in strain assignments. Since the establishment of the SARB collection, some strains have been reassigned to different serovars. SARB70, for example, was originally typed as serovar Typhisuis but was later classified as serovar Decatur. Similarly, two serovar Choleraesuis strains (SARB5, Cs6; SARB7, Cs13) were later excluded from this serovar (29). After obtaining the genovars for some SARB isolates, we encouraged the retyping of some isolates by Ken Sanderson (University of Calgary, Calgary, Alberta, Canada). These efforts revealed a few additional strains that have either been misclassified or swapped at an early stage in the dissemination of the collection. These misassignments included strain SARB50 (Pc4), which is not serovar Paratyphi C, and strains SARB19 and SARB20 (En7 and Em1), which were swapped with each other in our SARB collection. In addition, SARB35 (Mu4) is not serovar Munchen but the very closely related serovar Manhattan. All of the other serovar assignments were confirmed by this process. Thus, the serotypes of the SARB strains for which we are presenting results can be viewed with a high level of confidence.

The recent clinical isolates investigated in this study have each been subjected to pulsed-field gel electrophoretic (PFGE) analysis and have been assigned to serovars based on the observed patterns. For serovar classification, these patterns were compared to an extensive database comprised of PFGE patterns of thousands of clinical isolates cross-referenced with serological assignments (28). While errors cannot be 100% excluded, their probability is very low.

Comparison of microarray data with sequence data. In order to assess the quality of the microarray data, we made a comparison between the predictions of gene status for serovar Typhi CT18 as well as serovar Typhi Ty2 based on the array data and predictions based on available genome sequences (8, 18). Of all the chromosomal genes of the LT2 genome present on the array (in total, 4,338 spots), the microarray undercalled (i.e., called absent when the gene was in fact present) only 12 ORFs each for CT18 and Ty2 and overcalled (i.e., called present when the gene was in fact absent) 22 and 11 ORFs, respectively. The error rate for microarray predictions can therefore be estimated to be below 1%. The serovar Paratyphi A genome sequence of SARB42 (ATCC 9150), which was

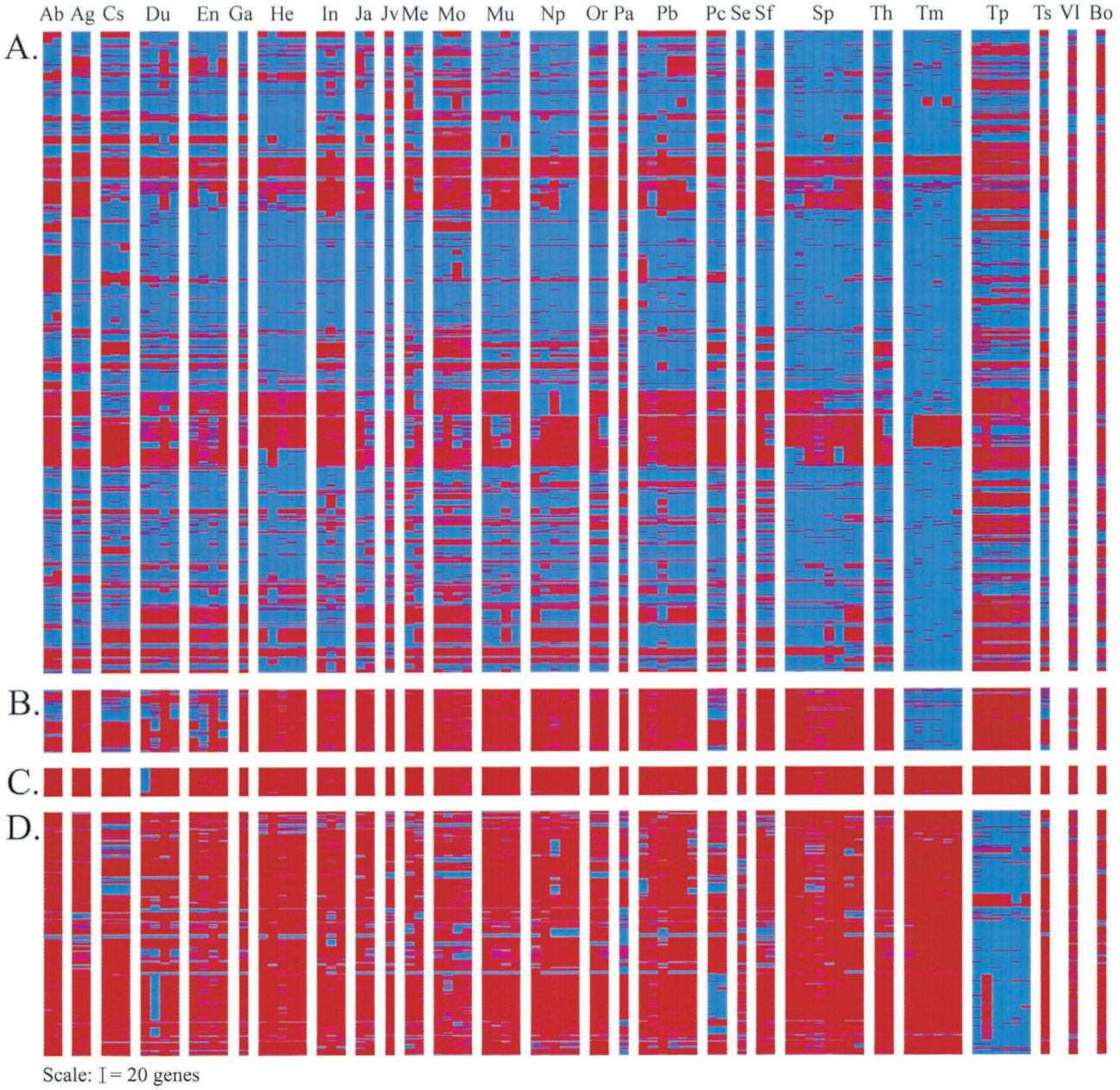
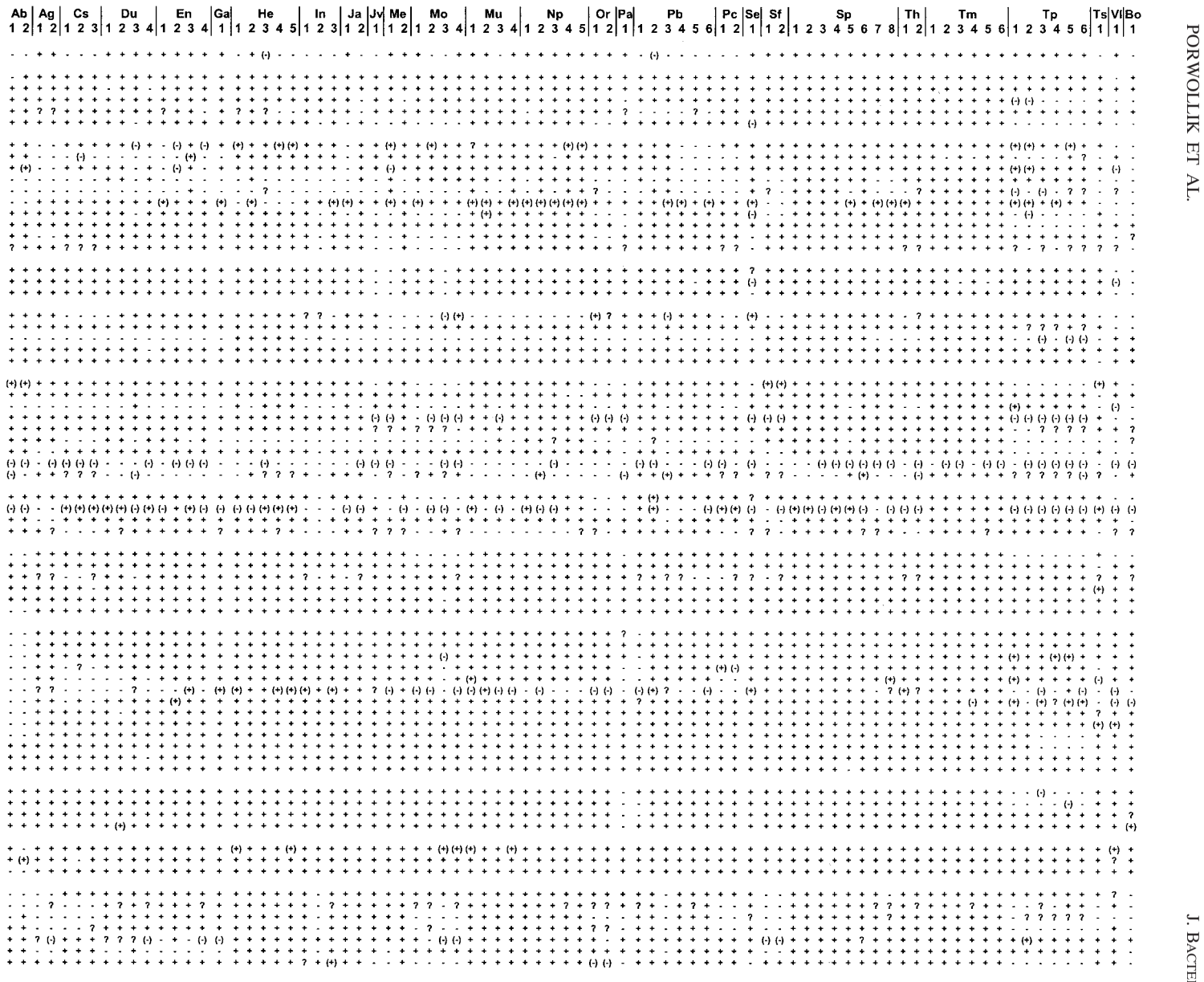


FIG. 1. *S. enterica* serovar Typhimurium LT2 and Typhi CT18 gene homologues with heterogeneous distribution patterns in *S. enterica* subspecies I serovars. Gene status is color-coded as follows: blue, present; purple, uncertain; red, absent. The strains are depicted, from left to right, in order of appearance in Table 1. (A) Serovar Typhimurium LT2 chromosomal genes. Only ORFs that are absent from at least one subspecies I strain are shown. (B) Plasmid pSLT. (C) Plasmid R46. (D) Genes present in serovar Typhi CT18, but absent from serovar Typhimurium LT2. In panels B and C, only genes that were predicted to be present in at least one subspecies I isolate outside serovar Typhimurium are shown. In panel D, only genes detected in at least one subspecies I isolate outside serovar Typhi are depicted.

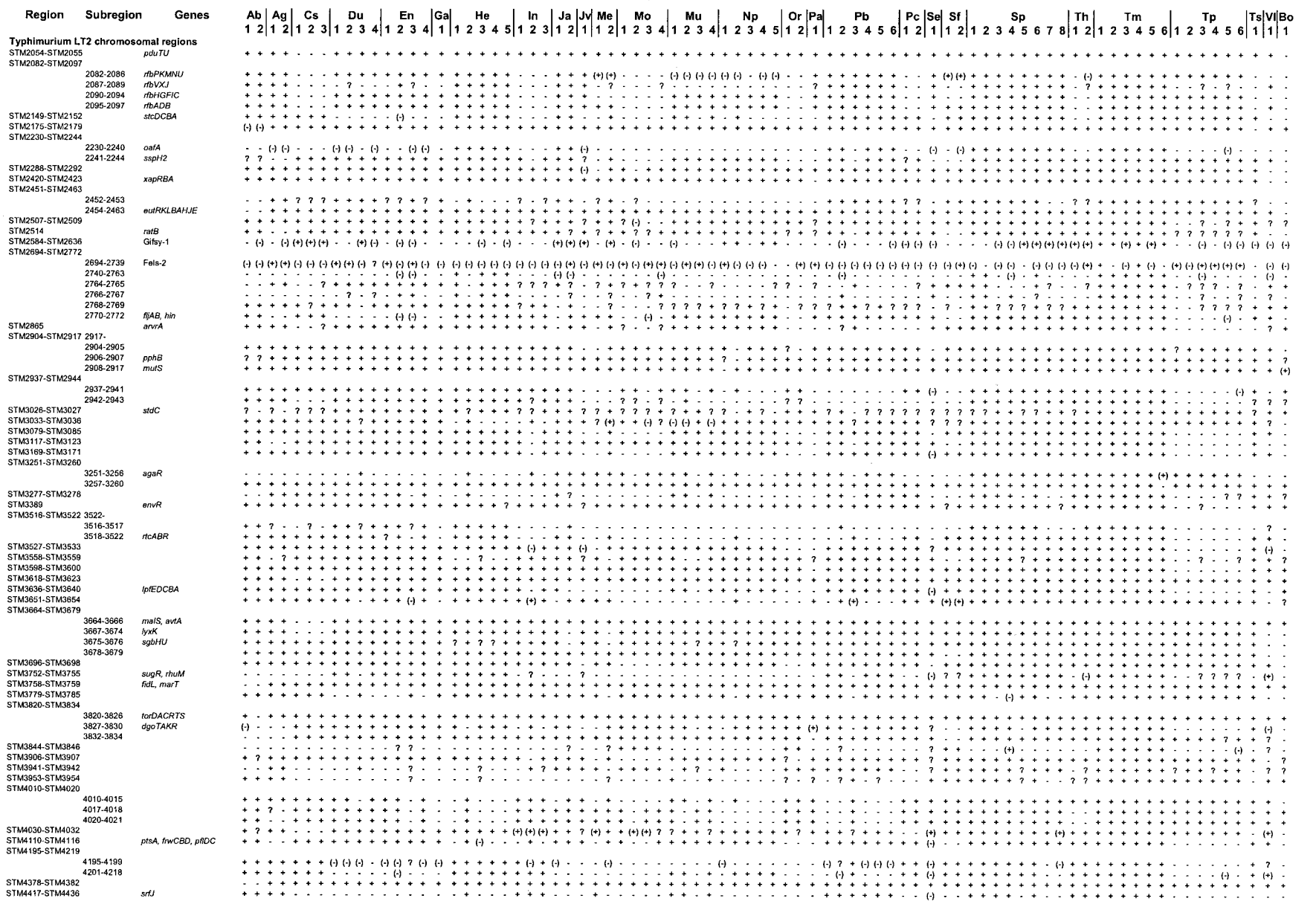
generated by the Genome Sequencing Center, St. Louis, Mo. (http://genome.wustl.edu/pub/seqmgr/bacterial/salmonella/S.paratyphiA), was also compared to the SARB42 microarray predictions, and we found only three overcalls and three undercalls, confirming the excellent concordance of microarray predictions with sequence data. In further agreement, the available sequence of *S. bongori* 12419, generated by the Sanger Center, Hinxton, Cambridge, United Kingdom (ftp://ftp.sanger.ac.uk/pub/pathogens/Salmonella/), displayed only 4 undercalls and 21 overcalls compared to the *S. bongori* SARC11 microarray assignments, despite being a different strain.

When the microarray data are compared to genome sequences that are not yet complete, genes present in the sequence data but not detected in the microarray (undercalls) should be rare and genes present in the microarray data that have not yet been sequenced in the incomplete genome sequence (overcalls) should be common. The partial sequence of

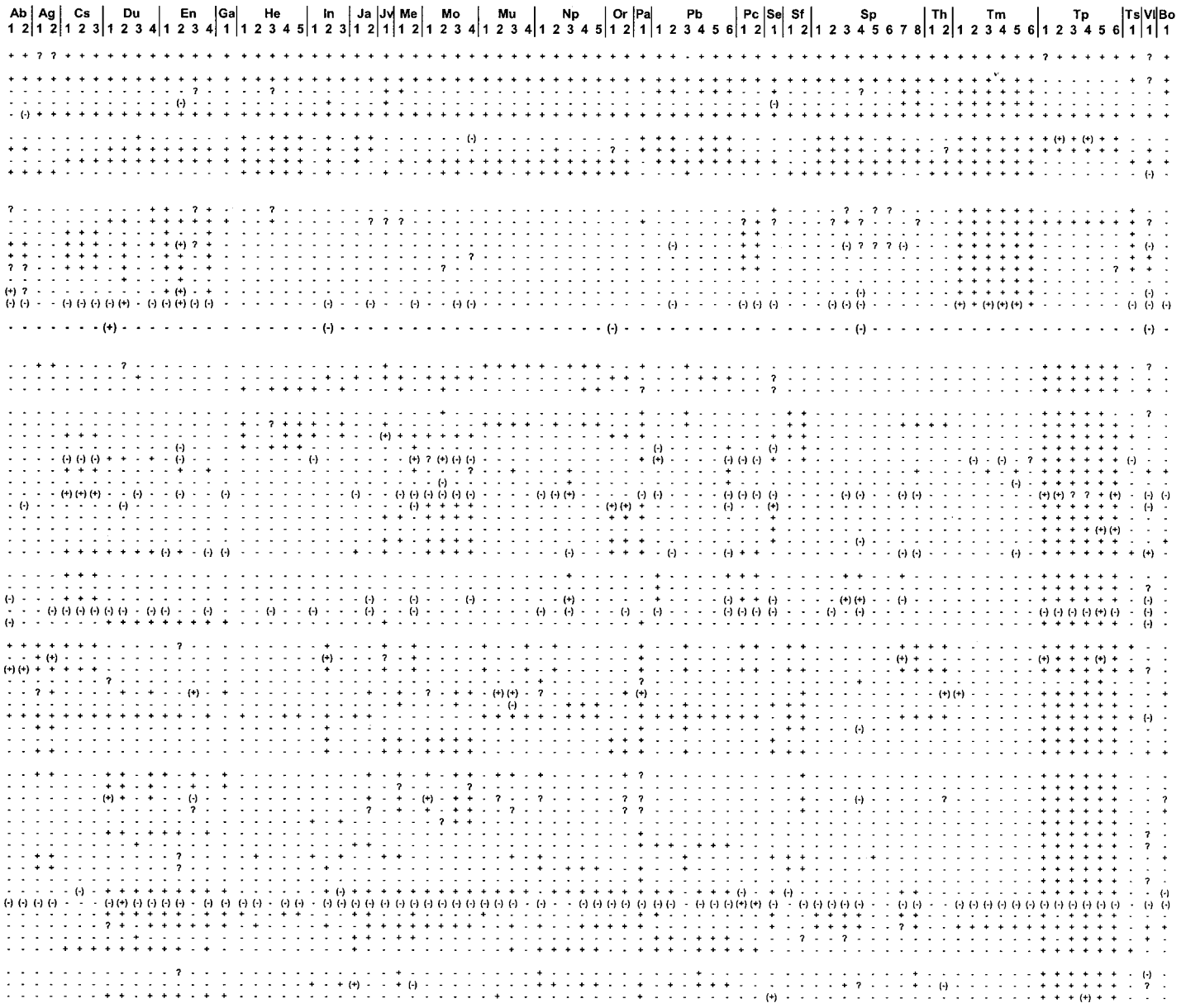
TABLE 2. Regions of two or more genes that exhibit inconsistent distribution within *S. enterica* subspecies I isolates[a]



Continued on following page



Continued on following page



[a] The isolates are sorted left to right in the same order presented in Table 1. Symbols: −, region absent; (−), region primarily absent; ?, status uncertain or equal numbers of genes in the region are predicted to be absent and present; (+), region primarily absent; +, region present. Some named singular genes with interesting distribution patterns are also depicted (see text for details).

an *S. enterica* serovar Gallinarum strain yielded 31 undercalls and 208 overcalls when it was compared to the microarray predictions for a different strain in MLEE type Ga2. The small number of undercalls suggests a close relationship between these two strains and the large number of overcalls indicates that there are several gaps that still need to be closed in the sequence. The same scenario was found when we compared sequence data for serovars Paratyphi C and Dublin with microarray predictions for all serovar Paratyphi C and Dublin strains investigated. While Du1, Du3, and DuA1 all yielded <10 undercalls, Du2 displayed 76 undercalls, excluding the Du2 MLEE type as the closest relative to the sequenced isolate. For serovar Paratyphi C, Pc1 and Pc2 yielded similarly small numbers of undercalls (10 and 8, respectively). The average number of apparent overcalls for serovar Paratyphi C was 343 and that for serovar Dublin was 644, indicating the degree of sequence completion in each genome sequencing project.

Serovar Typhimurium LT2 chromosomal genes. Of the four temperate prophage genomes present in serovar Typhimurium LT2, Fels-1 cannot be found in any other bacterial isolate to date. The other three phages are predominantly absent from many subspecies I serovars. However, several gene clusters within phage are frequently detected in other isolates, presumably due to the mosaic structure of phage genomes that leads to cross-hybridization of portions of related phages.

Genetic elements that were frequently missing or divergent were the entire *rfb* locus, responsible for the lipopolysaccharide side chain structure, *rfc* (the O antigen polymerase), and the fimbrial operons *saf*, *stc*, *sti*, *stj*, and *lpf*. The major flagellar filament protein FliC and its cap, FliD, were divergent in or absent from almost exactly the same isolates as the phase 2 flagellin FljB protein, together with the FliC repressor FljA and the Hin invertase, a system that enables the expression of FljB. The allantoin/glyoxylate cluster (STM0514-0532) has previously been observed to frequently be deleted from *Salmonella* genomes (11, 23). The reason for its instability is unknown to date. Some sugar transport operons (*dgo* and *frw*) or operons involved in sugar metabolism (*sgb*) were also absent quite frequently, suggesting a redundancy of these systems in the life cycle of *S. enterica* subspecies I isolates. Prominent individual genes that were absent from several isolates included *fhuA*, encoding an outer membrane receptor for ferrichrome and phages, the gene for the outer membrane protein RatB, which is involved in fecal shedding (14), *envR*, encoding a transcriptional repressor of the multidrug transport protein AcrF, and the *malX* pseudogene.

A total of 149 genes were absent from just one of the isolates investigated. Among these were the *suf* operon, encoding selenocysteine lyase and transport components (absent from Typhisuis Ts1), the *cai/fix* operon involved in carnithine metabolism (absent from Abortusovis AbA1), the *tor* operon, encoding the regulation and function of trimethylamine-*N* oxide reductase (absent from Abortusovis AbA2), and the *xap* operon, which is necessary for xanthosine transport (absent from Paratyphi A Pa1). The gene for the outer membrane protein BigA and the gene encoding topoisomerase IV, *parE*, were missing from Typhi Tp2. As part of a nine-gene cluster (STM2907 to STM2917), *mutS*, which is involved in DNA mismatch repair, was not present in Newport Np11. The gene encoding the murein lipoprotein Lpp (STM1377), a protein that connects the inner and outer membranes in the bacterium, and part of the *cit* operon (STM0618 to STM0621) involved in citrate lyase function were not detected in Typhisuis Ts1.

A subset of 74 *S. enterica* serovar Typhimurium LT2 genes were previously identified as subspecies I signature genes, as they are present in strains belonging to subspecies I but not in strains from the other subspecies investigated (23). In the present study, which extends the number of subspecies I strains examined, 31 genes of the original 74 were still not detected as being absent from any subspecies I isolate. Only four of these genes have annotated gene names (the acid phosphatase *phoN*, the gene *sinI*, encoding an outer membrane protein, and two *cit* genes, *citC2* and *citX2*). One candidate, STM0305, which in this study was detected in SARC14, a member of subspecies VI, had to be excluded from the signature set. Nineteen of the remaining signature genes are organized into six operons, all of which encode at least one gene product that is predicted to span the inner membrane: they are STM0041 to STM0042, STM0649 to STM0652, STM2132 to STM2135, STM2273 to STM2275, STM2573 to STM2575, and STM3547 to STM3550. All six operons are probably involved in transport processes, and four of these operons also include a predicted transcriptional regulator. Few of the genes in these six operons had borderline scores (less than 6 of 79), and therefore these operons may be suitable candidates for the easy detection and distinction of subspecies I *Salmonella* strains from all others. This is particularly true for the operons STM0041-0042 (a hydrolase and a putative galactoside symporter) and STM2573-2575 (containing a permease, a putative ketopanthoate reductase, and a putative regulatory element), in which no genes had a borderline score.

Serovar Typhi-specific chromosomal genes. Several of the Typhi CT18 genes present on the array were also detected in many other subspecies I isolates (Table 2). Among the operons frequently detected were the *rfbVXES* cluster involved in O antigen biosynthesis (STY2296-2299, detected in 13% of non-Typhi strains) and the fimbrial clusters *tcf* (present in 29% of all non-Typhi strains examined), *sef* (present in 13% of non-Typhi strains), and *ste* (detected in 59% of non-Typhi strains).

Serovars Dublin Du3, Paratyphi C Pc1, and Pc2 contain almost the entire serovar Typhi pathogenicity island SPI7, encompassing 149 genes from STY4521 to STY4680. The only region missing from the entire island in these serovars is the *sopE* moron, a gene cassette of P2-like phage origin that is incorporated into the 3′ end of *samA* within SPI7 (19). However, homologous counterparts of the *sopE* gene itself, likely residing on distinct phage genomes, were detected in approximately one-third of all investigated non-Typhi strains, including Dublin Du3. A well-characterized serovar Typhimurium phage, SopEΦ, contains a gene almost identical to the *sopE* gene identified within SPI7 which is essential for entry of the bacterium into epithelial cells (30). The SPI7 island is completely missing from the serovar Typhi isolate Tp2 (3), and its appearance in a serovar Dublin strain and two serovar Paratyphi C isolates suggests mobility of the cluster as a single insertion. It is also completely missing from a considerable fraction of nosocomial serovar Typhi strains (6) and was, for example, not present in a recent outbreak of serovar Typhi in India (17).

A cluster of genes with an unknown function, STY4412-4415, is present in half of all investigated non-Typhi isolates, and another set of genes likely to encode proteins that are exported (STY2349 to -2350, STY2361, and STY2364) were detected in one-third of all strains. The CT18 prophage present at STY2038-STY2077 was absent from all other isolates, including the serovar Typhi strains investigated. However, approximately 15% of the genes from this phage were also detected in Dublin Du3, likely due to cross-hybridization to homologous genes on a similar phage in that isolate. Another prophage element thought to be specific for CT18, STY1048-STY1071, retains some, but not all, of its genes in serovar Cholerasuis and the other serovar Typhi isolates, Paratyphi B Pb1, PbA7, Paratyphi C Pc1, Pc2, and Saint Paul SpA1, with almost the full complement present in Newport Np15. The phage-like genes at STY2015 to STY2036, just upstream of the CT18-specific prophage, are also found in several of these isolates, suggesting the presence of a single phage in strains that contain these two regions.

Plasmids. The serovar Typhimurium LT2 virulence plasmid pSLT consists of 111 annotated ORFs and contains the *spv* locus, an operon that restores fully virulent behavior to plasmid-cured strains of *Salmonella* (12). This locus is widely distributed within the *S. enterica* subspecies I serovars (2). The entire plasmid is present in all serovar Typhimurium strains included in this study (Table 2). A contiguous region of the plasmid, encompassing pSLT001 to pSLT056 and pSLT103 to pSLT111, is largely present in several other isolates, including serovars Typhisuis, Cholerasuis, Enteritidis, all Dublin strains except for Du2, and Paratyphi C. This set of genes includes *spv* as well as the *sam* locus (involved in DNA repair) and the *par* operon, encoding DNA partition proteins. The fimbrial locus *pef* is also part of this region but has not been detected in the serovar Dublin isolates and in Enteritidis En3. The first part of the *tra* locus (pSLT069 to pSLT084, including *psi*, involved in plasmid SOS response inhibition) has been detected in Enteritidis En1, En3, and EnA1, whereas the second part of the *tra* locus, from pSLT088 to pSLT103, has only been retained in Enteritidis En3 and Dublin Du3. Both of these strains also retained pSLT056 to pSLT067 (including the *ssb* gene for the single-stranded DNA binding protein), rendering En3 with a more or less complete pSLT plasmid except for the *pef* locus and Du3 with homologues of all regions except *pef* and *ssb*. The *srg*/*rck* locus (pSLT008 to pSLT011) involved in resistance to complement killing has not been found in any serovar Choleraesuis isolate. Homologues of *srgAB* have been detected in all serovar Typhi strains and in serovar Paratyphi A, and *srgB* copies have in addition been found in many other isolates. This operon is regulated by the chromosomal *sdiA* gene, encoding a global regulator implicated in the detection of other microbial species (27).

None of the strains investigated retained any genes of the serovar Typhimurium plasmid pKM101, with the notable exception of Dublin Du1, in which nearly all of the pKM101 genes are predicted to be present.

Interserovar divergence. The relationships among strains were analyzed by using a phylogenetic assumption in which the absence or presence of gene clusters was used to determine putative relatedness. While this assumption is not ideal because gene clusters are probably exchanged between serovars, it does provide a convenient, although not exact, indication of relatedness. The tree presented in Fig. 2 was constructed with the data shown in Table 2. The advantage of this data set is that it does not overemphasize big clusters of genes that appear to be acting as a single unit for gene transfer. Trees that were constructed using this type of condensed matrix generally were almost identical to those obtained using single gene predictions (data not shown). In addition, the application of several different algorithms (neighbor joining and the unweighted pair group method with arithmetic mean) did not significantly changed the tree configuration. All of the trees that were constructed with PAUP (Sinauer Associates) were fairly similar to genetic distance trees observed when using MLEE (4). For some serovars, all isolates clustered tightly together, and therefore these can be called monophyletic (serovars Montevideo, Enteritidis, Heidelberg, Munchen, Newport, Paratyphi C, Typhimurium, and Typhi). Other serovars exhibited polyphyletic behavior, i.e., not all isolates of the same serovar clustered (serovars Dublin, Saint Paul, Infantis, Muenster, and Paratyphi B, including the L-+-tartate-+ group formerly classified as Java). While these observations were in general agreement with the clustering behavior of these serovars according to MLEE data, there are some notable exceptions: serovar Newport, which was polyphyletic according to MLEE analysis, appeared to be monophyletic according to its gene content and only interspersed with serovar Munchen isolates. Similarly, serovar Munchen appeared to have a monophyletic distribution of genovars, which was not apparent in the MLEE data. In general, monophylogeny of the serovars was more supported by the genovar tree than by MLEE data, expressing the general trend of serovar isolates to be genetically more closely related to each other than to isolates of a different serovar.

However, the genetic complement of a certain isolate of one *S. enterica* serovar does not always resemble another isolate of the same serovar. Instead, the data indicate associations of a few isolates that have very different serovar assignments. In order to visualize associations of the strains, we created a relationship matrix which displays the numbers of differences in gene presence and absence assignments for all strains against each other in a color-coded fashion. The matrix illustrated in Fig. 3 calculated the number of genetic regions consisting of at least two consecutive genes with differing absence-presence status in all isolates investigated. Obvious phage regions and plasmid genes, i.e., high-mobility regions, were excluded from this calculation. The highest number of differences (117, between SARC14 and Typhi Tp1) is represented as a black square, whereas the lowest number of differences (0) creates a white square. All intermediate values are depicted between these shades on a sliding gray scale. In this matrix, strains of polyphyletic serovars will display markedly different shades than isolates of the same serovar, whereas strains from monophyletic serovars will not. In addition, close relationships between isolates in different serovars will also be easily recognizable as white or light squares in areas that are off the diagonal. With this computation, the close similarity of the serovar Choleraesuis isolates to those of serovar Paratyphi C (only six different regions between Cs11 and Pc1, with no isolate displaying more than 11 differences from another) and of serovar Gallinarum Ga2 to Enteritidis En7 became apparent. Other strains with different serovar assignments that are

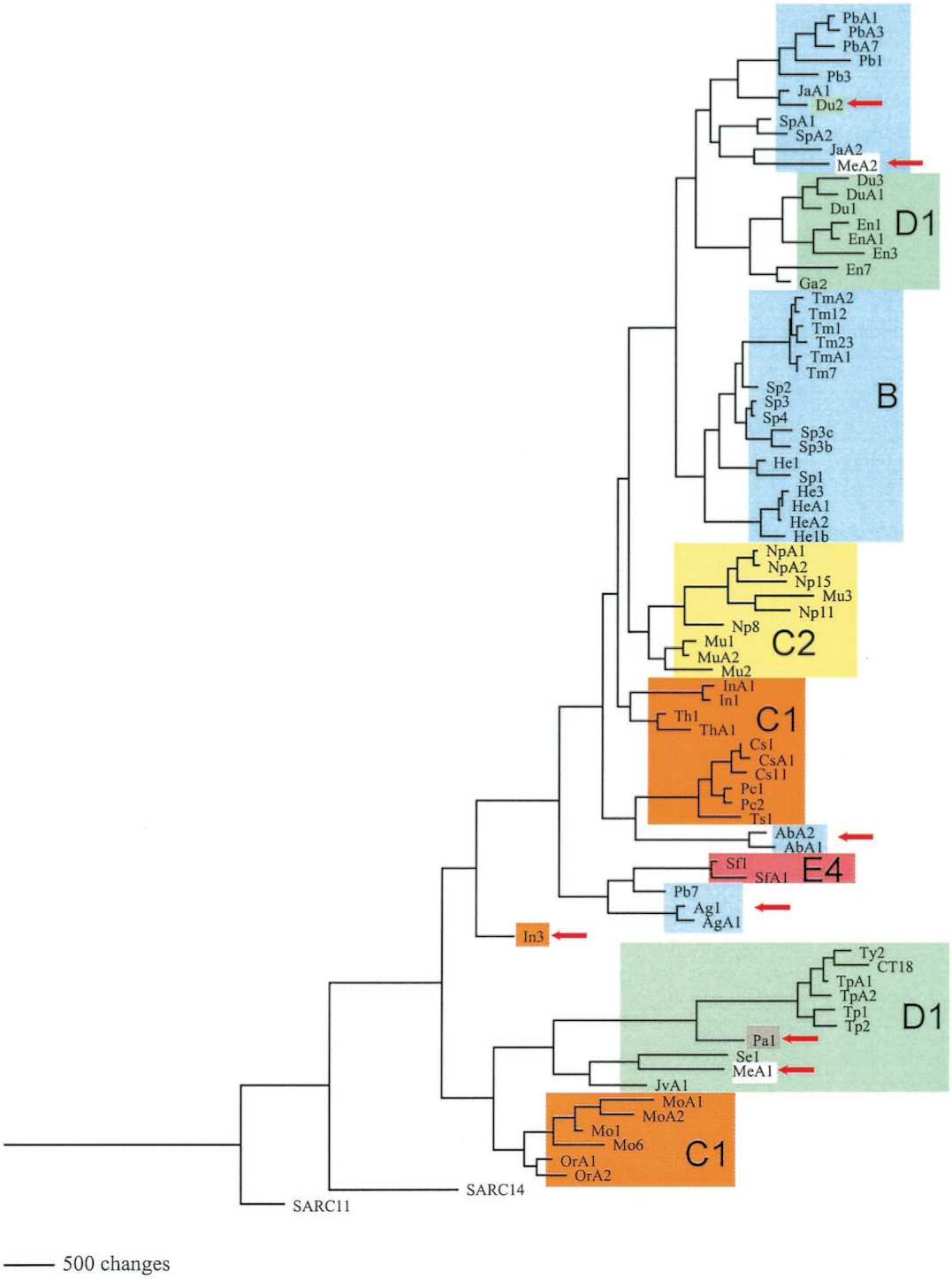


FIG. 2. Phylogenetic tree of *S. enterica* subspecies I isolates. The tree was constructed with PAUP software (Sinauer Assoc. & Co.) by using the presence-absence predictions for the regions as described in Table 2. The following conditions were applied: maximum parsimony, weighting against repeated gains of genes, 10,000 bootstraps. Serogroups are indicated, and notable deviations from the expected clustering by serogroup are depicted with red arrows. Me isolates are serogroup E1, and Pa1 is serogroup A.

genetically quite close (12 or fewer differences) are the serovar Typhimurium isolates and serovar Saint Paul (except the clinical isolates); Typhisuis Ts1 with serovar Paratyphi C and Choleraesuis; Montevideo Mo1 with the serovar Oranienburg isolates; Enteritidis En1 with serovar Dublin (except Du2); and Heidelberg He1 with serovar Saint Paul (except the clinical isolates and Sp3c). All of these similarities are within the same serogroup. However, similarities of isolates of different serogroups can also be found. Dublin Du2, for example, belongs to serogroup D1 and only differs from Java JaA1 (serogroup B) in

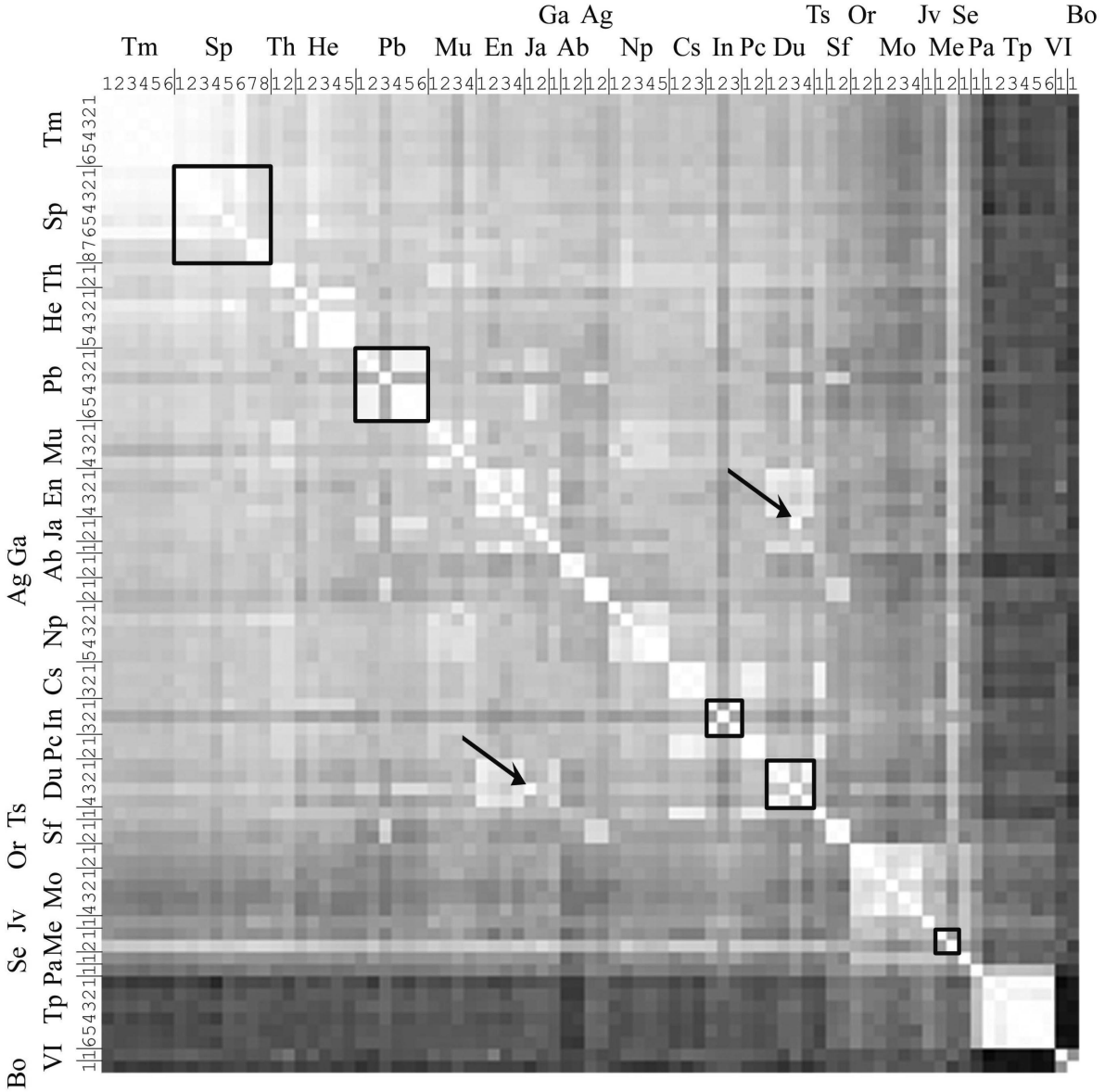


FIG. 3. Relationship matrix depicting the numbers of different absence-presence calls for genomic Typhimurium LT2 and Typhi CT18 regions between strains. Phage regions are excluded. The numbers of differences in gene content are illustrated as shaded squares on a linear scale from white (no differences) to black (maximal number of differences in the matrix [117, for *S. enterica* subsp. *indica*, or VI, versus Typhi Tp1]). Strains are grouped by serovars and in the order of similarity of the most common MLEE type of the respective serovar to Typhimurium Tm1. Within the serovars, strains are ordered as in Table 1. Polyphyletic serovars are marked with black squares. Similarity between Dublin Du2 and Java JaA1 is highlighted with arrows.

five regions. The numerical matrix used for the graphical representation in Fig. 3 can be found as supplementary information (supplement B) at http://bioinformatics.skcc.org/mcclelland/salmonella/subspecies1/.

Intraserovar divergence. The matrix also visualizes the fact that occasionally isolates of the same serovar vary from each other quite substantially when the number of differing regions is considered. The most prominent examples illustrating this effect are Infantis In3, Paratyphi B Pb7, and Muenster MeA1, which differ from the other isolates of their respective serovars that have been tested by at least 38 regions, not counting phages and plasmids. In addition, Dublin Du2, Java JaA1, and Muenchen Mu3 also differ from their respective serovar representatives by more than 20 regions. Nevertheless, Mu3 and JaA1 cluster not far away from the remaining isolates of their respective serovar, whereas polyphyletic behavior is clearly shown for In3, Pb7, MeA2, and Du2 (Fig. 2).

However, most isolates of the same serovars differ by only a few regions. Among the serovars that display close relationships (10 regions or fewer that differ between all isolates investigated) are Abortusovis, Agona, Choleraesuis, Oranienburg, Paratyphi C, Senftenberg, Thompson, Typhimurium, and Typhi. Three of these serovars (Choleraesuis, Typhimurium, and Typhi) were represented by more than two isolates.

Comparison of recent clinical isolates with the SARB strain collection. The SARB collection was established over a decade ago and sampled the reservoir of *Salmonella* serovars at the time. Each different MLEE pattern was associated with a certain frequency of occurrence. In serovars that exhibited more than one MLEE pattern, one of the patterns was usually prevalent and the others were usually rare. When comparing the microarray data for recent clinical isolates sampled within the last 3 to 4 years to the prevalence more than a decade ago, in general those recent clinical isolates clustered tightly with the strain representing the most prevalent MLEE type in the SARB collection. This was seen for serovars Choleraesuis, Infantis, Muenchen, Montevideo, Newport, and Enteritidis. When initially challenged with the unexpected tight clustering of both clinical serovar Heidelberg isolates with the rare He3 isolate of the SARB collection (SARB24), we performed an additional genovar determination of He1 isolates represented in the SARA collection of *S. enterica* strains (1). In fact, three He1 SARA strains investigated (SARA30, SARA32, and SARA33) clustered tightly with the isolate that was assigned as He3 in the SARB collection (SARA32 data are shown as an example). It is very likely that SARB He3 was swapped with the prevalent SARB He1 strain during the establishment of the SARB collection and is now represented by SARB23, whereas He1 is actually represented by SARB24.

We observed one remarkable aberration from the expected clustering of clinical isolates with the prevalent strain in the SARB collection. Whereas the two serovar Saint Paul clinical isolates were almost identical, neither resembled either of the two different Saint Paul MLEE types in the SARB collection. In order to determine whether these clinical isolates resembled any of the serovar Saint Paul ETs that were not represented in the SARB collection, we obtained the genetic profiles of SARA22 and SARA23, representing serovar Saint Paul types Sp1 and Sp2, respectively, as well as the profiles of the Sp3 isolates SARA25 and SARA27 (1). The clinical isolates did not cluster with these strains either. Hence, the two clinical serovar Saint Paul isolates represent a separate lineage within this serovar that was not sampled, and possibly rarer, 15 years ago.

Correlation of MLEE types with genovars. The ETs investigated in this study generally resulted in different genovars by microarray analysis. As an exception, all serovar Typhimurium ETs investigated in this study exhibited very similar genovars, with only minor differences observed (three or fewer regions). Another exception is the SARB Saint Paul Sp4 isolate (SARB56), which displayed a profile identical to that of the Sp3 SARB55 strain.

However, it is possible, if not likely, that isolates of the same MLEE type would belong to different genovars also. In order to test this possibility, we compared the SARB55 (Saint Paul Sp3) genovar with patterns obtained from SARA25 and SARA27, also exhibiting the Sp3 ET. These two SARA strains displayed a very close genetic relationship (two regions were different), but SARA27 differed from the SARB55 genovar in seven regions of at least two consecutive genes. When comparing the Heidelberg He1 profiles in this study, we observed that SARA33 differed from SARB24 in five regions. However, SARA32 and SARA30 only exhibited two differing genomic regions when compared to SARB24 (data not shown).

In conclusion, a separate MLEE type generally results in a different genovar. However, some isolates of closely related MLEE types belonged to the same or a very similar genovar.

DISCUSSION

S. enterica serovars are defined by antigenic variation at lipopolysaccharide moieties (O antigen), flagellar antigens (H antigen), and capsular polysaccharides (Vi antigen). Early indications of significant genetic differences within *S. enterica* serovars were observed by electrophoretic typing using MLEE, which defines groups of strains according to electrophoretic mobility differences in housekeeping proteins caused by amino acid polymorphisms (4). MLEE revealed that, in some serovars, strains could be further subclassified into two or more distinct ETs. Recently, the onset of microarray and genomic sequencing technology has allowed for the differences among strains to be characterized at single-gene resolution. Using microarrays, we have found that separate ETs usually display different gene profiles, defined by the presence and absence of many genes. These differences can be quite substantial: even when one only considers nonphage regions of the serovar Typhimurium LT2 chromosome, within serovar Infantis, for example, In3 differs from In1 in as many as 30 regions. If one looks at all genes present on the microarray, the number rises to 46 differing regions. On the other hand, members of the same ET usually have similar, although not necessarily identical, gene profiles. Recent clinical isolates tend to have a gene profile similar to the one measured for the most common MLEE types from a decade ago in the same serovar.

We postulate that genome structures that arise within any given serovar may be sufficiently stable to define classes of genomes within that serovar. There are examples in our study that support this assumption: virtually identical genome structures were observed, among other examples, for four serovar Heidelberg isolates, the serovar Oranienburg strains, the serovar Agona isolates, and the serovar Senftenberg isolates. We coined the term "genovar" to describe groups of strains that share a similar profile of gene content and to distinguish these groups from the serovars that often contain more than one genovar. While the exact gene profile that defines each genovar and the boundaries between genovars will require further work, it is likely that a practical definition of a genovar will exclude genes carried on highly mobile elements such as phages, plasmids, and transposons, which are expected to regularly leap genotype boundaries. However, the definition of genovars will probably include remnants of phages, transposons, or plasmids which are no longer capable of hyperactive lateral transfer. For a firm definition of genovar boundaries, more data will be required to provide a better overview of the heterogeneity of genotypes among *Salmonella* isolates.

The observation of examples of two or more very different genovars within a serovar represents a quandary. How can two or more significantly different genovars come to exist in the same serovar at the same time? The most obvious explanation involves a mechanism by which the surface antigens that define a serovar are transferred to a different genovar, allowing a new genovar to be recruited to the serovar. There is substantial evidence for the horizontal transfer of genes encoding flagellin and the O antigen within the salmonellae (15, 25, 31). Based on the clustering behavior of the isolates, our data support the

recruitment of Dublin Du2, Infantis In3, and Paratyphi B Pb7 into their genovars by transfer of the surface antigens. However, we cannot exclude the possibility that large numbers of insertion and deletion events (and amino acid changes in housekeeping genes) all took place together in a relatively short time, thereby creating the genovar diversity within the same serovar. If all of the genome changes took place gradually, further sampling should reveal intermediate forms.

One way that may allow a better understanding of the method and rate of formation of genovars will be to sequence DNAs from regions of the genome that can shed light on the phylogenetic histories that define serovars in conjunction with regions that distinguish genovars. Another useful goal will be to monitor the prevalence of genovars over time. The serovar Saint Paul example in this study may suggest that the prevalence of different genovars within *Salmonella* serovars can change within a relatively short time frame, as a Saint Paul genovar that was rare or nonexistent a little more than a decade ago appears to be more common now. However, it is possible that the unusual Saint Paul genovar in the clinical isolates can be explained by a simple serovar typing error, although every effort has been made to prevent misassignment.

As the profiling of hundreds of strains by microarrays is prohibitively expensive, we are currently designing specific PCR methods to perform genovar typing in a high-throughput, relatively inexpensive manner, relying on a knowledge of the binary presence-absence polymorphisms of gene loci scattered throughout the genome that can be used to define each genovar. A PCR approach has already been shown to be useful for the detection of serogroup H (O:6,14) isolates, concentrating on the O antigen gene cluster of these strains (9). We propose that the differences in genovars detected here for the most prevalent serovars of *S. enterica* can form the basis for the detection and distinction of isolates on the potentially disease-relevant level of genovars across the different serovars.

The gene profiles observed to date reveal close relationships that were not necessarily expected between serovars. Despite different host ranges, the serovar Choleraesuis Cs11 serovar (C1 serogroup), for example, was very similar to serovar Paratyphi C isolates (same serogroup). Only 22 single gene differences were observed in the absence-presence patterns of the LT2 chromosomal genes (excluding phage regions) between these two serovars. Among these were only two regions of clustered genes: STM1677 to STM1680 (a thiol peroxidase, an outer membrane protein, a gene similar to the invasin C of *Yersinia*, and a protein kinase) and STM3665 to STM3674, which includes a possible chemotaxis gene as well as several conserved inner membrane proteins. The genovars of serovar Dublin Du2 (D1 serogroup) and serovar Paratyphi B L-+-tartrate-+ Java JaA1 (serogroup B) also differ by 22 LT2 chromosomal genes in only three regions, including the *stc* fimbrial operon (absent from Du2), a region of phage remnants (STM2230 to 2240, absent from Du2), and two *rfb* genes (*rfbX* and *rfbJ*). These relationships between different serovars suggest a close evolutionary relationship. The latter case may in fact be an excellent example of recruitment of an isolate from an exogenous serovar closely related to "Java" into serovar Dublin by the exchange of genes in the *rfb* locus.

The detection of the serovar Typhi CT18 long pathogenicity island SPI7 in serovar Paratyphi C and Dublin isolates has been reported and discussed elsewhere (6, 20). The presence of a plasmid based on pKM101 in Dublin Du1, but not in any other strains examined here, remains unexplained since pKM101 was only deliberately introduced into serovar Typhimurium isolates in the seventies (16).

The bifurcating tree presented in Fig. 2 is an oversimplification of the relationships between strains, because it attempts to build a phylogeny despite the high level of lateral transfer between strains. It is known that transfers between subspecies I isolates of *Salmonella* occur very frequently, possibly surpassing the level of recombination events observed between different subspecies (5). To portray the relationships among strains without imposing a bifurcating tree model, we visualized the genetic distance between strains with a matrix (Fig. 3), using shades of gray, ranging from white, for a perfect match, to black, for the most divergent comparison to the most distant *Salmonella*. While genetic distance is also a crude measure of relatedness, this portrayal allows cases of high divergence within a serovar to be observed as juxtaposed light and dark squares on the diagonal, while indicating similarities among serovars as light boxes off the diagonal. This matrix may be a better indication of relationships than phylogenetic trees in cases where extensive horizontal gene transfer can substantially obscure and override phylogeny.

When comparing genomic contents of the different *S. enterica* subspecies I strains, one would expect the clustering of isolates to be influenced by at least three factors: (i) serogroup, (ii) host specificity, and (iii) disease characteristics. It was unclear to what extent each of these factors would contribute to the clustering behavior. Figure 2 shows that serogroup classification as defined by Kauffmann-White indicates related gene content in most cases. It is therefore a strong expression of genetic relatedness. However, there are exceptions when isolates that do not belong to the same serogroup are similar to each other. Moreover, in our data two serogroups, D1 and C1, formed two different subclusters. For group D1, one of the subclusters consisted primarily of human-restricted serovars (Typhi and Sendai), indicating that host range may also be a determining factor for clustering behavior.

We were unable to detect specific genes or genomic regions that were absent from all host specialists while being present in host generalists or present in human and absent from nonhuman isolates. It is likely that the host range is determined by a combination of genes in different loci, which can be altered by deletion or simple point mutation events. Moreover, additional genes (or phage genomes) and competition between bacterial isolates are also likely to contribute to host range. It can be expected that serovars that have adapted to a narrow host range will lack the functionality of different genes, depending on the host they adapted to. In this context, it is interesting that occasionally a particular host-restricted serovar exclusively lacked certain genes that were otherwise present in all strains investigated. The most dramatic example in our data set is probably the entire region from STM1512 to STM1570, which was absent from both isolates of serovar Abortusovis but from no other serovars. The area contains 35 genes with unknown functions, some of which may have a role in adaptation to a broad host range. The data set provides several observations such as these, which will initiate further investigations.

About 75% of the 803 LT2 chromosomal genes that were absent in some *Salmonella* genomes had no assigned name, compared to 21% of all genes on the array that were not named, indicating that the class of frequently absent genes is enriched in ORFs for which no function has yet been found by genetics. These genes probably contribute to fitness in the wild or they would not be present in groups of strains. Perhaps many of the clusters of genes that distinguish genovars are partly redundant. Genetics would be hard pressed to reveal the function of gene clusters that are partly compensated for by other parts of the genome. The fitness differences that drive this partial redundancy could be subtle or only easily measured in a narrow environmental condition. Another, not mutually exclusive, possibility is that the fitness differences of these variations among genovars are manifested only in environments that have not yet received much attention by researchers. These conditions might include sustained survival in the lumen of the gut, passage into the feces, or survival outside the host. Most research has concentrated on interactions with host cells and has largely ignored survival requirements when unattached to host cells or not in the host. Something as simple as a change in the diet of humans or their domestic animals could change the kind of metabolism that would be optimal for *Salmonella* in the gut or feces. This could put huge selective pressure on the prevalence of genovars.

The fact that genetic differences within a serovar can have profound consequences for the pathogenicity of the isolate has already been noted many times. For example, in serovar Paratyphi B, strains from systemic infections always lack the *avrA* gene but contain *sopE1*, whereas strains from enteric infections generally display different absence-presence patterns for these genes (24). All systemic isolates investigated by Prager et al. belonged to MLEE type 1 (Pb1). In our study, all serovar Paratyphi B isolates except Pb7 lacked the *avrA* gene (*sopE1* was not present on the array). Considering the profound genovar difference between Pb7 and all other serovar Paratyphi B isolates, it is very likely that this difference will also manifest itself in differing, yet to be revealed, characteristics.

Now that the existence of distinct ETs has been expanded to encompass genovars that differ by hundreds of genes, it seems inevitable that these differences will generally be manifested in particular phenotypes that affect various aspects of fitness. It can be expected that genovars will define yet to be determined characteristics in particular groups of strains in the same way as serology proved to be a useful classification because it encompassed differences among strains in host range and in pathogenic mechanism and severity.

ACKNOWLEDGMENTS

We thank K. Sanderson (University of Calgary, Calgary, Alberta, Canada) for supply of the clinical serovar Paratyphi B isolates, the strains from the SARC and the SARB collections, and for critical reading of the manuscript; C. Clark and J. Allen (National Laboratory of Enteric Pathogens, Winnipeg, Manitoba, Canada) for the PFGE plugs of the recent clinical isolates; and S. Uzzau (University of Sassari, Sassari, Italy) for the serovar Abortusovis strains. In addition, we thank F. Long for computational expertise.

This work was supported in part by NIH grant AI34829 (M.M.) and the generosity of Sidney Kimmel.

REFERENCES

1. **Beltran, P., S. A. Plock, N. H. Smith, T. S. Whittam, D. C. Old, and R. K. Selander.** 1991. Reference collection of strains of the Salmonella typhimurium complex from natural populations. J. Gen. Microbiol. **137:**601–606.
2. **Boyd, E. F., and D. L. Hartl.** 1998. *Salmonella* virulence plasmid. Modular acquisition of the *spv* virulence region by an F-plasmid in *Salmonella enterica* subspecies I and insertion into the chromosome of subspecies II, IIIa, IV and VII isolates. Genetics **149:**1183–1190.
3. **Boyd, E. F., S. Porwollik, F. Blackmer, and M. McClelland.** 2003. Differences in gene content among *Salmonella enterica* serovar Typhi isolates. J. Clin. Microbiol. **41:**3823–3828.
4. **Boyd, E. F., F. S. Wang, P. Beltran, S. A. Plock, K. Nelson, and R. K. Selander.** 1993. Salmonella reference collection B (SARB): strains of 37 serovars of subspecies I. J. Gen. Microbiol. **139:**1125–1132.
5. **Brown, E. W., M. K. Mammel, J. E. LeClerc, and T. A. Cebula.** 2003. Limited boundaries for extensive horizontal gene transfer among Salmonella pathogens. Proc. Natl. Acad. Sci. USA **100:**15676–15681.
6. **Bueno, S. M., C. A. Santiviago, A. A. Murillo, J. A. Fuentes, A. N. Trombert, P. I. Rodas, P. Youderian, and G. C. Mora.** 2004. Precise excision of the large pathogenicity island, SPI7, in *Salmonella enterica* serovar Typhi. J. Bacteriol. **186:**3202–3213.
7. **Chan, K., S. Baker, C. C. Kim, C. S. Detweiler, G. Dougan, and S. Falkow.** 2003. Genomic comparison of *Salmonella enterica* serovars and *Salmonella bongori* by use of an *S. enterica* serovar Typhimurium DNA microarray. J. Bacteriol. **185:**553–563.
8. **Deng, W., S. R. Liou, G. Plunkett III, G. F. Mayhew, D. J. Rose, V. Burland, V. Kodoyianni, D. C. Schwartz, and F. R. Blattner.** 2003. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. J. Bacteriol. **185:**2330–2337.
9. **Fitzgerald, C., R. Sherwood, L. L. Gheesling, F. W. Brenner, and P. I. Fields.** 2003. Molecular analysis of the *rfb* O antigen gene cluster of *Salmonella enterica* serogroup O:6,14 and development of a serogroup-specific PCR assay. Appl. Environ. Microbiol. **69:**6099–6105.
10. **Fitzgerald, J. R., and J. M. Musser.** 2001. Evolutionary genomics of pathogenic bacteria. Trends Microbiol. **9:**547–553.
11. **Garaizar, J., S. Porwollik, A. Echeita, A. Rementeria, S. Herrera, R. M. Wong, J. Frye, M. A. Usera, and M. McClelland.** 2002. DNA microarray-based typing of an atypical monophasic *Salmonella enterica* serovar. J. Clin. Microbiol. **40:**2074–2078.
12. **Gulig, P. A., H. Danbara, D. G. Guiney, A. J. Lax, F. Norel, and M. Rhen.** 1993. Molecular analysis of *spv* virulence genes of the *Salmonella* virulence plasmids. Mol. Microbiol. **7:**825–830.
13. **Joyce, E. A., K. Chan, N. R. Salama, and S. Falkow.** 2002. Redefining bacterial populations: a post-genomic reformation. Nat. Rev. Genet. **3:**462–473.
14. **Kingsley, R. A., A. D. Humphries, E. H. Weening, M. R. De Zoete, S. Winter, A. Papaconstantinopoulou, G. Dougan, and A. J. Baumler.** 2003. Molecular and phenotypic analysis of the CS54 island of *Salmonella enterica* serotype Typhimurium: identification of intestinal colonization and persistence determinants. Infect. Immun. **71:**629–640.
15. **Li, J., K. Nelson, A. C. McWhorter, T. S. Whittam, and R. K. Selander.** 1994. Recombinational basis of serovar diversity in Salmonella enterica. Proc. Natl. Acad. Sci. USA **91:**2552–2556.
16. **McCann, J., N. E. Spingarn, J. Kobori, and B. N. Ames.** 1975. Detection of carcinogens as mutagens: bacterial tester strains with R factor plasmids. Proc. Natl. Acad. Sci. USA **72:**979–983.
17. **Mehta, G., and S. C. Arya.** 2002. Capsular Vi polysaccharide antigen in *Salmonella enterica* serovar Typhi isolates. J. Clin. Microbiol. **40:**1127–1128.
18. **Parkhill, J., G. Dougan, K. D. James, N. R. Thomson, D. Pickard, J. Wain, C. Churcher, K. L. Mungall, S. D. Bentley, M. T. Holden, M. Sebaihia, S. Baker, D. Basham, K. Brooks, T. Chillingworth, P. Connerton, A. Cronin, P. Davis, R. M. Davies, L. Dowd, N. White, J. Farrar, T. Feltwell, N. Hamlin, A. Haque, T. T. Hien, S. Holroyd, K. Jagels, A. Krogh, T. S. Larsen, S. Leather, S. Moule, P. O'Gaora, C. Parry, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell.** 2001. Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature **413:**848–852.
19. **Pelludat, C., S. Mirold, and W. D. Hardt.** 2003. The SopEPhi phage integrates into the *ssrA* gene of *Salmonella enterica* serovar Typhimurium A36 and is closely related to the Fels-2 prophage. J. Bacteriol. **185:**5182–5191.
20. **Pickard, D., J. Wain, S. Baker, A. Line, S. Chohan, M. Fookes, A. Barron, P. O. Gaora, J. A. Chabalgoity, N. Thanky, C. Scholes, N. Thomson, M. Quail, J. Parkhill, and G. Dougan.** 2003. Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding *Salmonella enterica* pathogenicity island SPI-7. J. Bacteriol. **185:**5055–5065.
21. **Popoff, M. Y., and L. Le Minor.** 1997. Antigenic formulas of the Salmonella serovars, 7th revision. W.H.O. Collaborating Centre for Reference and Research on Salmonella. Institut Pasteur, Paris, France.
22. **Porwollik, S., J. Frye, L. D. Florea, F. Blackmer, and M. McClelland.** 2003. A non-redundant microarray of genes for two related bacteria. Nucleic Acids Res. **31:**1869–1876.
23. **Porwollik, S., R. M. Wong, and M. McClelland.** 2002. Evolutionary genomics

of Salmonella: gene acquisitions revealed by microarray analysis. Proc. Natl. Acad. Sci. USA **99:**8956–8961.
24. **Prager, R., W. Rabsch, W. Streckel, W. Voigt, E. Tietze, and H. Tschape.** 2003. Molecular properties of *Salmonella enterica* serotype Paratyphi B distinguish between its systemic and its enteric pathovars. J. Clin. Microbiol. **41:**4270–4278.
25. **Reeves, P.** 1993. Evolution of Salmonella O antigen variation by interspecific gene transfer on a large scale. Trends Genet. **9:**17–22.
26. **Selander, R. K., D. A. Caugant, H. Ochman, J. M. Musser, M. N. Gilmour, and T. S. Whittam.** 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl. Environ. Microbiol. **51:**873–884.
27. **Smith, J. N., and B. M. Ahmer.** 2003. Detection of other microbial species by *Salmonella*: expression of the SdiA regulon. J. Bacteriol. **185:**1357–1366.
28. **Swaminathan, B., T. J. Barrett, S. B. Hunter, and R. V. Tauxe.** 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg. Infect. Dis. **7:**382–389.
29. **Uzzau, S., D. J. Brown, T. Wallis, S. Rubino, G. Leori, S. Bernard, J. Casadesus, D. J. Platt, and J. E. Olsen.** 2000. Host adapted serotypes of Salmonella enterica. Epidemiol. Infect. **125:**229–255.
30. **Wood, M. W., R. Rosqvist, P. B. Mullan, M. H. Edwards, and E. E. Galyov.** 1996. SopE, a secreted protein of Salmonella dublin, is translocated into the target eukaryotic cell via a sip-dependent mechanism and promotes bacterial entry. Mol. Microbiol. **22:**327–338.
31. **Xiang, S. H., M. Hobbs, and P. R. Reeves.** 1994. Molecular analysis of the *rfb* gene cluster of a group D2 *Salmonella enterica* strain: evidence for its origin from an insertion sequence-mediated recombination event between group E and D1 strains. J. Bacteriol. **176:**4357–4365.