JOURNAL OF BACTERIOLOGY, Mar. 1997, p. 1985–1991 Vol. 179, No. 6
0021-9193/97/$04.00+0

Comparative Genetics of the *inv-spa* Invasion Gene Complex of *Salmonella enterica*

E. FIDELMA BOYD,[1]* JIA LI,[1] HOWARD OCHMAN,[2] AND ROBERT K. SELANDER[1]

Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, Pennsylvania 16802,[1] and Department of Biology, University of Rochester, Rochester, New York 14627[2]

Received 4 November 1996/Accepted 15 January 1997

The chromosomal region containing the *Salmonella enterica* pathogenic island *inv-spa* was present in the last common ancestor of all the contemporary lineages of salmonellae. For multiple strains of *S. enterica*, representing all eight subspecies, nucleotide sequences were obtained for five genes of the *inv-spa* invasion complex, *invH*, *invE*, *invA*, *spaM*, and *spaN*, all of which encode proteins that are required for entry of the bacteria into cultured epithelial cells. The *invE*, *invA*, *spaM*, and *spaN* genes were present in all eight subspecies of *S. enterica*, and for *invE* and *invA* and their products, levels of sequence variation among strains were within the ranges reported for housekeeping genes. In contrast, the InvH, SpaM, and SpaN proteins were unusually variable in amino acid sequence. Furthermore, *invH* was absent from the subspecies V isolates examined. The SpaM and SpaN proteins provide further evidence of a relationship (first detected by Li et al. [J. Li, H. Ochman, E. A. Groisman, E. F. Boyd, F. Solomon, K. Nelson, and R. K. Selander, Proc. Natl. Acad. Sci. USA 92:7252–7256, 1995]) between the cellular location of the products of the *inv-spa* genes and evolutionary rate, as reflected in the level of polymorphism within *S. enterica*. Invasion proteins that are membrane bound or membrane associated are relatively conserved in amino acid sequence, whereas those that are exported to the extracellular environment are hypervariable, possibly reflecting the action of diversifying selection.

The invasion of host cells by *Salmonella enterica* is mediated by the products of a large number of genes that map to several chromosomal locations (8, 10, 18, 47). Homologs of some of these genes have also been detected in strains of *Escherichia coli* (19), but a pathogenicity island (SPI-1) situated at 63 min on the *S. enterica* Typhimurium LT2 chromosome that is not present in the *E. coli* K-12 chromosome, the *inv-spa* complex, contains genes whose products are required for the invasion of nonphagocytic cells (7, 8, 14, 17, 19, 24, 26, 30, 39). At least 10 genes within this 40-kb region constitute the *inv-spa* complex, which specifies a type III secretion system involved in the export of antigens that promote cell entry (11, 19, 42).

Recent studies have identified a second virulence locus also encoding a type III secretion system in *S. enterica* serovar Typhimurium (35, 45). This second 40-kb region is located at 30.7 min and is presumed to be the same locus previously identified as clone RF333 and characterized as a pathogenicity island (SPI-2) required for survival of *Salmonella* cells in host cells (20, 22, 45).

Homologs of several of the *inv-spa* genes have been identified on the virulence plasmids of *Shigella* and *Yersinia* species and in several other, more distantly related animal and plant pathogens as well as among the genes necessary for the biosynthesis of flagella in several gram-positive and gram-negative bacteria (2, 9, 28, 36), which is perhaps not surprising given that both systems are responsible for transport and localization of proteins across the cell membrane.

While the structure and organization of the gene complexes specifying these secretory pathways, particularly in enteric pathogens, are broadly conserved, their phylogenetic distribution, genomic locations, and base compositions suggest that these sequences arose independently in divergent pathogens (36). The invasion genes of *Shigella* species are plasmid borne and have low GC contents, which is evidence of horizontal transfer from a distantly related organism (21). All species of *Shigella* arose from *E. coli* relatively recently (37, 38) and therefore cannot serve as the source of the *inv-spa* complex, which appears to be ancestral to salmonellae (4, 27). Similarly, the moderately AT-rich *inv-spa* genes of *S. enterica* were not the donors of the *Shigella* sequences. Genes of the *spa* complexes of *S. enterica* and *Shigella* species are more similar to one another than to those of *Yersinia* species, and therefore the latter are excluded as the source of these sequences (19, 36).

The most comprehensive analyses of structure and function of an invasion gene have focused on *invA* (12, 13, 15, 16, 39). Secondary-structure analysis of InvA predicts that there are seven transmembrane domains located in the amino-terminal half, and a largely hydrophilic carboxyl terminus is predicted to be located in the cytoplasm (12, 18). A regulatory role has been suggested for *invE*, and molecular characterization of *invH* by Altmeyer et al. (1) has shown that this gene has features which are commonly involved in transport of proteins beyond the cytoplasmic membrane.

Analysis of *spaO*, *spaP*, and *spaQ* from multiple strains of *S. enterica* representing all subspecies found evidence of a relationship between evolutionary rate and cellular location (27). In the present study, we determined the extent and pattern of sequence diversity in five additional genes of the *inv-spa* complex—*invH*, *invE*, *invA*, *spaM*, and *spaN*—among strains of serovars of the eight subspecies of *S. enterica* to further examine the evolutionary processes underlying divergence in the invasion gene complex. We suggest that *invH* was absent from the most recent common ancestor of subspecies V. Furthermore, we also provide further evidence for a correlation between evolutionary rate and cellular location.

* Corresponding author. Present address: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138.

MATERIALS AND METHODS

Bacterial strains. This study was based on a sample of 19 strains of the eight subspecies of *S. enterica*, including the 16 strains of *Salmonella* reference collec-

TABLE 1. Properties of the 13 isolates representing subspecies V of *S. enterica*

MLEE profile[a]	Strain no.	Antigenic formula	Source	Locality	Year isolated
1	S3050	48:z_{39}:--	Human	Algeria	1986
2	S3048	60:z_{41}:--	Human	Senegal	1985
3	S3039	66:z_{35}:--	Lizard	Italy	1977
4	S3051	61:z_{35}:--	Human	Sudan	1987
5	S3045	44:r:--	Lizard	United Kingdom	1977
6	S3049	1,13,22:i:--	Human	Sweden	1985
7	S3043	48:z_{35}:--	Lizard	Chad	1967
8	S3046	44:z_{39}:--	Food	Ghana	1983
9	S3041	66:z_{41}:--	Frog		1972
10	S3047	1,40:z_{35}:--	Food	Nigeria	1984
11	S3042	48:a:--	Human	California	1983
12	S3044	48:z_{41}:--	Parakeet	United States	1976
13	S3040	66:z_{65}	Fish meal	Malawi	1980

[a] Taken from Fig. 1 in reference 3. MLEE, multilocus enzyme electrophoresis.

tion C (SARC) (3) and three additional subspecies I strains of the host-adapted serovars Dublin (S1518), Gallinarum (S2962), and Choleraesuis (S1280) selected from *Salmonella* reference collection B (6).

Amplification and nucleotide sequencing of *inv-spa* genes. Primers were designed from the published sequences of *invH* (1), *invE* (17), *invA* (12), *spaM*, and *spaN* (19), and internal primers were constructed as sequence information became available. Double-stranded PCR DNA was purified with the Qiaquick PCR purification kit (Qiagen, Chatsworth, Calif.). The sequences were determined in both directions, based on the dideoxy terminator method, using a DNA thermal cycler (Perkin-Elmer Cetus, Norwalk, Conn.) and an automated 373 DNA sequencer (Applied Biosystems) according to the manufacturers' instructions.

DNA hybridization. A DNA fragment for use as a fluorescein-labeled probe in Southern hybridization was prepared from the *S. enterica invH* gene as follows. The *invH* gene was amplified via PCR, with DNA from strain S4194 as template. Following amplification, the PCR product was purified by using the Qiaquick PCR purification kit (Qiagen). The *invH* probe was labeled with fluorescein-conjugated nucleotides and, after filter hybridization, was detected with the enhanced chemiluminescence system of Amersham (Arlington Heights, Ill.). Genomic DNA was extracted from 13 strains of *S. enterica* subspecies V by using G-Nome DNA isolation kits from BIO 101 (Vista, Calif.). The 13 subspecies V isolates were previously analyzed by multilocus enzyme electrophoresis (3). DNAs from the 13 subspecies V strains (Table 1) and 2 subspecies I isolates (S4194 and S3333) were digested with *Eco*RI, and the fragments were separated by agarose gel electrophoresis. The DNA fragments were transferred to Hybond-N nylon membranes for hybridization at 60 and 55°C (high and low stringency, respectively) in 5× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate) 0.1% sodium dodecyl sulfate, and 5% dextran sulfate.

Statistical analysis of *inv-spa* genes from *S. enterica* isolates. Phylogenetic analysis of the data was carried out with the computer program MEGA (25) and computer programs provided by Thomas S. Whittam.

Nucleotide sequence accession numbers. The nucleotide sequences of the *inv-spa* genes of *S. enterica* have been submitted to the GenBank data library. The *inv-spa* genes were assigned the following accession numbers: *spaO*, *spaP*, and *spaQ*, U29345 to U29365; *spaM* and *spaN*, U43300 to U43315; *invA* and *invE*, U43237 to U43274; and *invH*, U84270 to U84286.

RESULTS

Chromosomal gene arrangement. The five invasion genes sequenced are arranged on the *Salmonella* chromosome in the order *invH*, *invE*, *invA*, *spaM*, and *spaN* (Fig. 1). The *invE* gene is located 24 bp upstream of *invA*, and *spaM* and *spaN* are immediately upstream of the previously examined *spaO*, *spaP*, and *spaQ* loci (27). The *invH* gene is located 457 bp upstream of *invF* and is transcribed in the direction opposite that of the other loci in the *inv-spa* cluster.

DNA hybridization. The *invH* gene could not be PCR amplified from subspecies V isolates S3041 and S3044. In order to determine whether *invH* is absent from this subspecies, 13 strains from this group were screened by DNA hybridization (Table 1). A weak hybridization signal was obtained with the *invH* gene probe with one subspecies V strain (S3043) under high-stringency conditions. Moreover, when the stringency of the conditions was reduced, three more subspecies V strains, S3039, S3051, and S3041, gave weak hybridization signals, suggesting that *invH* may be present but highly divergent from the subspecies I *invH* gene.

Rate of synonymous (silent) and nonsynonymous (replacement) substitutions (d_S and d_N). For the 16 strains representing the eight subspecies of *S. enterica*, information on the extent of sequence variation in seven *inv-spa* genes and five housekeeping genes is presented in Table 2. The most notable feature is an unusually high level of amino acid substitutions in the *spaM*, *spaN*, and *spaO* genes, which is reflected in the proportion of amino acids that are polymorphic and by d_N, which is the mean estimated number of nonsynonymous nucleotide substitutions at nonsynonymous sites between pairs of strains. Whereas the percentages of polymorphic amino acids for *invE*, *invA*, *spaP*, and *spaQ* fall within the narrow range of 2.3 to 4.8%, comparable values for *spaM*, *spaN*, and *spaO* are 17.7, 34.4, and 20.4%, respectively. This difference is also reflected in d_N values; d_N is uniformly low for *invE*, *invA*, *spaP*, and *spaQ* (range, 0.25 to 0.68) and very high for *spaM*, *spaN*, and *spaO* (range, 2.84 to 5.99).

In the case of the mean estimated number of synonymous sites between pairs of strains (d_S), levels of sequence diversity in *spaN* and *spaO* (25.41 and 24.30, respectively) are only slightly elevated in comparison with those of *invE*, *invA*, *spaP*, and *spaQ*, and that of *spaM* shows no increase (Table 2).

Data for the *invH* gene are not included in Table 2 because this sequence could not be obtained from strains of subspecies V. To assess the relative level of sequence variation in *invH*, d_N and d_S were calculated for the other genes for 14 strains of seven subspecies (Table 3). The d_S and d_N values, ± standard deviations, for *invH* were 30.96 ± 4.20 and 5.6 ± 0.76, respectively, both of which are elevated. The d_N/d_S ratio for each Inv-Spa protein is given in Table 3. This ratio indicates the

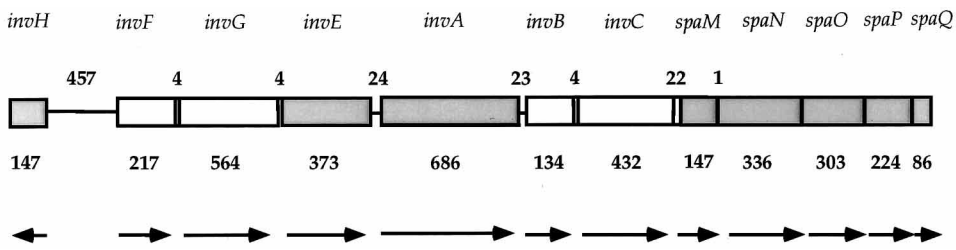

FIG. 1. Organization of the *inv-spa* genes of *S. enterica*. Gene designations are those of Groisman and Ochman (19). Arrows indicate the direction of transcription. The numbers above the gene organization refer to intergenic regions, and those below indicate amino acid numbers.

TABLE 2. Sequence variation in 12 genes among 16 strains of *S. enterica*

Gene	No. (%) of base pairs of gene sequenced	No. (%) of polymorphic:		Mean pairwise value (10^2) for:		d_N/d_S ratio	Reference
		Nucleotides	Amino acids	d_S	d_N		
Invasion							
invE	1,119 (100)	170 (15.2)	18 (4.8)	21.40 ± 1.33	0.50 ± 0.10	0.02	This study
invA	1,950 (95)	307 (15.7)	30 (4.6)	22.87 ± 1.33	0.68 ± 0.12	0.03	This study
spaM	444 (100)	83 (18.8)	26 (17.7)	21.50 ± 2.89[a]	2.84 ± 0.48[a]	0.13	This study
spaN	1,011 (100)	284 (27.6)	118 (34.4)	25.41 ± 2.07	5.99 ± 0.48	0.24	This study
spaO	909 (100)	204 (22.4)	62 (20.4)	24.30 ± 2.04	3.55 ± 0.41	0.15	26
spaP	672 (100)	89 (13.2)	7 (3.1)	19.70 ± 2.08	0.38 ± 0.15	0.02	26
spaQ	258 (100)	28 (10.9)	2 (2.3)	13.78 ± 2.78	0.25 ± 0.17	0.02	26
Housekeeping							
putP	1,467 (97)	216 (14.7)	21 (4.3)	16.70 ± 1.88	0.60 ± 0.23	0.04	33
mdh	849 (90)	133 (15.7)	11 (3.9)	20.13 ± 1.72	0.48 ± 0.16	0.02	5
gapA	924 (93)	118 (12.8)	14 (4.5)	15.15 ± 1.49	0.61 ± 0.15	0.04	34
gnd	1,335 (95)	216 (16.2)	20 (4.5)	21.80 ± 1.40	0.44 ± 0.10	0.02	32
aceK	1,719 (98)	338 (19.7)	47 (8.2)	28.39 ± 1.76	1.05 ± 0.14	0.04	31

[a] Calculated without codon 39, which is a stop codon in strain S3015.

relative degree of functional constraints experienced by a protein if it is evolving in a neutral fashion. For *invH*, *spaM*, *spaN*, and *spaO*, the d_N/d_S ratios are elevated, whereas the ratios for the other *inv-spa* genes are very similar to those of the housekeeping genes. Inasmuch as subspecies V is very strongly differentiated from all other subspecies and, consequently, makes substantial contributions to the values of d_N and d_S shown in Table 2, the fact that very large values for both d_N and d_S were obtained for *invH* in an analysis in which subspecies V was not included indicates that it is even more variable than *spaM*, *spaN*, and *spaO*. As shown in Table 2, the extent of sequence diversity in the conserved invasion genes, *invE*, *invA*, *spaP*, and *spaQ*, is generally similar to that of the five housekeeping genes.

Polymorphism in *invA* and *invE*. The levels of nucleotide sequence diversity in *invE* and *invA* are similar to those of housekeeping genes (Table 2). However the distribution of polymorphic sites in *invA* is nonrandom (Fig. 2). For convenience of reference, four segments of the *invA* gene and its product are designated by the letters A to D in Fig. 2 and segment B is subdivided into B1 and B2. As shown in Fig. 2, region A of InvA (amino acids 1 through 299) is strongly conserved, with only four polymorphic positions, three of which involve substitutions in strains of subspecies V alone. There is a concentration of polymorphic amino acid positions in region B, which extends from amino acid 300 through 499; 22 of the 31 polymorphic amino acids are located in this region.

TABLE 3. Sequence variation in eight invasion genes among 14 strains of *S. enterica* representing seven subspecies[a]

Gene	No. of base pairs	Mean pairwise value (10^2) for:		d_N/d_S ratio
		d_S	d_N	
invH	441	31.0 ± 4.2	5.6 ± 0.8	0.18
invE	1,119	19.5 ± 1.5	0.3 ± 0.1	0.02
invA	1,950	16.7 ± 1.1	0.5 ± 0.1	0.03
spaM	441	17.8 ± 2.7	2.9 ± 0.5	0.16
spaN	999–1,029	19.3 ± 1.9	4.4 ± 0.4	0.23
spaO	909	19.5 ± 1.9	3.0 ± 0.4	0.15
spaP	672	14.6 ± 1.8	0.2 ± 0.1	0.01
spaQ	258	10.8 ± 2.5	0.1 ± 0.1	0.01

[a] Subspecies V not included.

In contrast, the amino terminus of the protein's region C (amino acids 500 to 589) is also strongly conserved, and in region D (amino acids 590 to 650) there is a second, small concentration of polymorphic amino acids, reflecting almost entirely substitutions in strains of subspecies IV, VII, and V, followed by a terminal string of 37 monomorphic amino acids. Within subspecies, there are very few differences in amino acid sequence among strains.

There is marked regional variation in the frequency of silent substitutions along the gene (Fig. 2). A conspicuous feature is the occurrence of two high peaks in the mean proportion of synonymous site differences between pairs of strains in region B, centering on codons 375 in B1 and 455 in B2. These peaks are produced in major part by unusually high concentrations of changes (from the consensus sequence) in strains of subspecies IV and VII, but some contribution is also made by subspecies V (Fig. 2, lower panel). In subspecies IIIb, there is a concentration of changes in B2 but not in B1, and in subspecies IIIa, the frequency of substitutions is not unusually high in either region.

The fact that not all the subspecies exhibit an increased frequency of synonymous substitutions in both B1 and B2 would seem to exclude a localized elevation in mutation rate as an explanation for the overall high incidence of synonymous changes in region B. For subspecies IV and VII, the data provide circumstantial evidence of the importation of a distinctive B region segment from a source not represented by the strains in our sample of *S. enterica*. The basis for this inference is the observation that 19 (58%) of the total of 33 polymorphic nucleotide sites that are uniquely shared in subspecies IV and VII are in region B (11 of them in B1 and 8 in B2), although this region contains only 600 (31%) of the 1,950 nucleotides of the sequenced segment of the gene.

The regional variation in the distribution of synonymous substitution in subspecies IV and VII was detected by Stephens' (46) test for nonrandom clustering of synonymous polymorphic sites as a 648-bp partition (codons 95 to 245) in region A, in which strains of these subspecies share no unique polymorphic sites ($P = 0.001$). Elsewhere in the gene, strains of these subspecies share a total of 38 unique polymorphic sites.

Indicative of a possible intragenic recombination event is the identification, by Stephens' test, of a significant partition that involves a segment of 776 bp (between bp 435 and 1211) in which the strains of subspecies IIIb share no unique polymor-

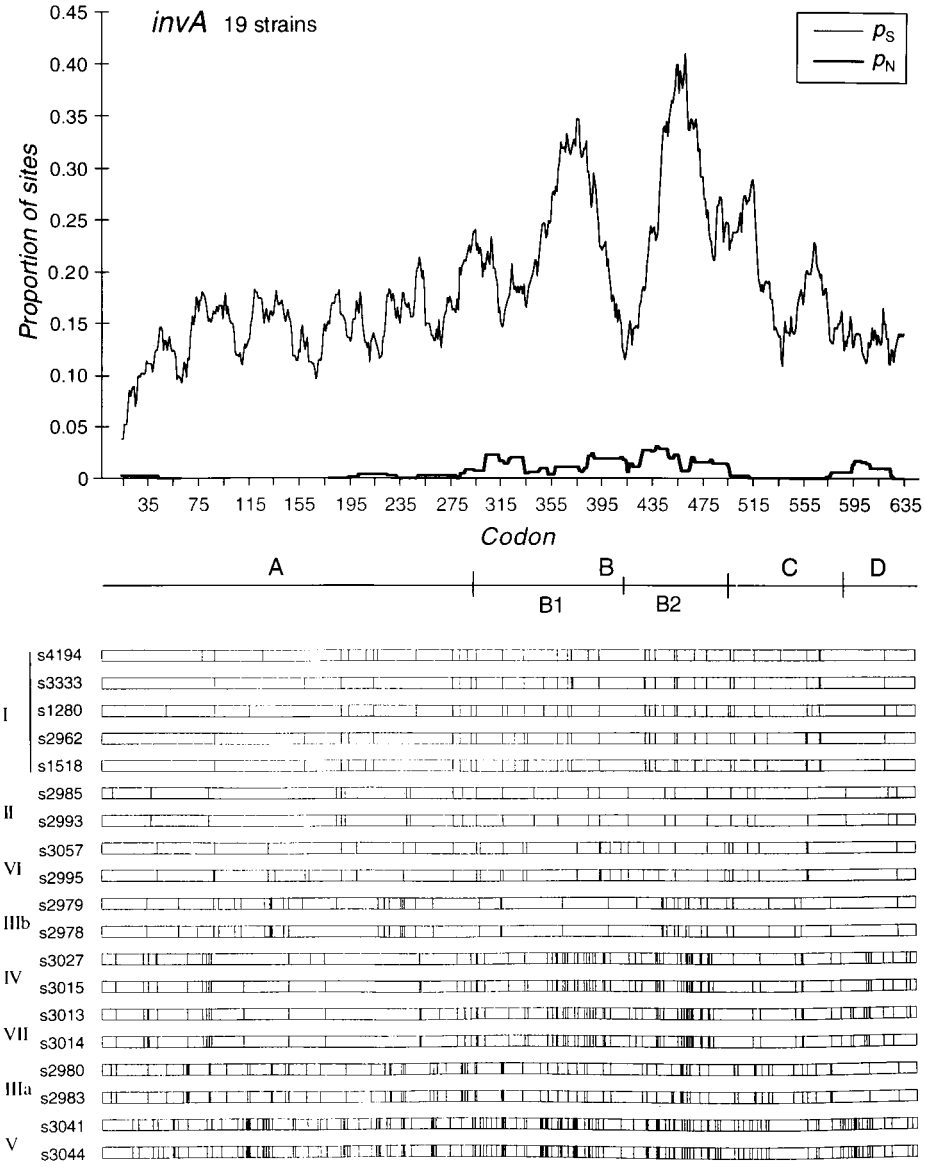

FIG. 2. Distribution of variable sites in the *invA* gene among 19 strains of *S. enterica*. (A) Regional variation in mean proportion of silent site differences between pairs of strains (p_S) and mean proportion of replacement site differences between pairs of strains (p_N) based on a sliding window of 60 nucleotides. (B) The locations of polymorphic sites along the gene from the consensus sequence are indicated by vertical lines. Subspecies are indicated by roman numerals, and strain numbers are indicated.

phic sites ($P = 0.001$). Stephens' test also identified a third partition that involves the sharing by strains of subspecies IIIa, IIIb, and V of five unique polymorphic nucleotide sites in a 384-bp segment ($P = 0.006$). If intragenic recombination events are responsible for the sharing of these sites, they must have occurred early in the evolutionary history of the species because, apart from these few shared sites, there is no particular similarity in the nucleotide sequences of these 384-bp segments of the three strongly differentiated subspecies.

Stephens' test also identified two significant partitions in *invE*. The first identified a unique segment in one strain of subspecies IV (S3015) that had five unique sites. The second partition involves three unique sites in strain S3013 (subspecies VII). Both of these phylogenetic partitions are indicative of intragenic recombination from an unknown source.

Polymorphism in *spaM* and *spaN*. The level of amino acid divergence in SpaM is relatively high. In one strain of subspecies IV (S3015) there is a stop codon at position 59 resulting from a G-T mutation at nucleotide 115.

The *spaN* gene is highly variable in both length (range, 999 to 1029 bp) and amino acid sequence. Length variation involves subspecies IIIa, IV, V, and VII. Subspecies V has a 9-bp deletion beginning at nucleotide 435, and subspecies IIIa has a 3-bp deletion beginning at nucleotide 345. Strains S3027 of subspecies IV and S3013 of subspecies VII both have a 21-bp tandem repeat at position 564; the repeats differ from one another at one nucleotide site.

The amino acid variation in SpaN is the result of highest d_N (5.99 ± 0.48) value calculated thus far in *S. enterica*. The distribution of polymorphic amino acids is nonrandom, with

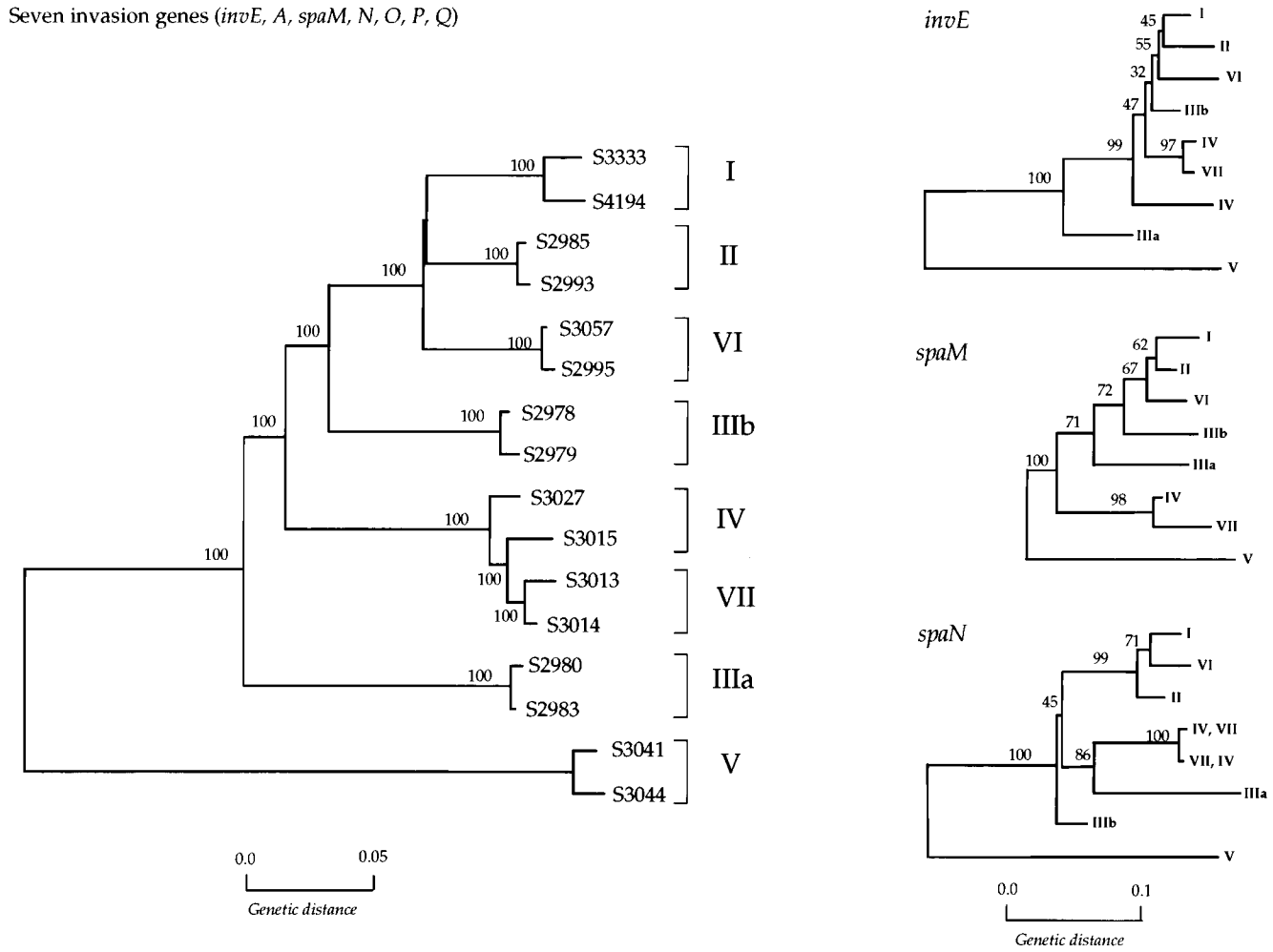

FIG. 3. Evolutionary relationships among 16 strains of the eight subspecies of *S. enterica* based on variation in the combined coding sequences of seven invasion genes. Individual phylogenetic gene trees for *invE*, *spaM*, and *spaN* are based on synonymous sites among 16 strains of *S. enterica*. Subspecies are designated by roman numerals, and bootstrap values, based on 1,000 computer-generated trees, are indicated at the nodes.

only six polymorphic sites in the last 75 positions; the N terminus is similarly conserved, whereas the central region is highly variable and is the site of all the sequence variation. In the *spaM* and *spaN* genes, Stephens' test recognized one recombination event in *spaN* involving subspecies IV (S3027) and VII (S3013).

Polymorphism in *invH*. The *invH* gene was not amplified from all subspecies of *S. enterica*. For the 13 isolates of subspecies V, we were unable to amplify *invH*. However, among the remaining seven subspecies from which *invH* was sequenced, there was a high degree of nucleotide diversity (Table 3), with both d_S and d_N values being elevated.

Genetic relationship among *Salmonella* isolates. Individual neighbor-joining trees (41) were constructed from a pairwise matrix of Jukes' and Cantor's (23) distances for the *invH*, *invE*, *invA*, *spaM*, and *spaN* genes at silent sites. A combined consensus tree for seven invasion genes (*invE*, *invA*, *spaO*, *spaP*, *spaQ*, *spaM*, and *spaN*) is shown in Fig. 3. The pattern (topology) of nucleotide sequence differentiation among the *Salmonella* subspecies for the consensus invasion tree was generally similar to the consensus pattern for five housekeeping genes (3). However, among the housekeeping genes, the position of subspecies II and VI with respect to subspecies I is variable, as reflected in the relatively low bootstrap value (57%) for the common node of the lineages of these three subspecies (3). Among the individual invasion gene trees, some differences in topology are noted. The *invE* gene tree differs from the consensus tree in that strain S3015 of subspecies IV did not cluster with either subspecies VII or subspecies IV isolates (Fig. 3). Stephens' test identified a possible intragenic recombination event in this strain which could account for its placement in the *invE* gene tree. Also of note are the short branch lengths for subspecies I, II, VI, and IIIb in the *invE* tree, which are indicative of recombination among these groups. The *spaM* gene tree differed from the consensus tree in the relative branching order of subspecies IIIa, which clustered with subspecies I, II, and VI (Fig. 3). Similarly, the *spaN* gene tree differed from the consensus tree in the branching order of subspecies IIIa isolates, which cluster with subspecies IV and VII, and isolates of these subspecies clustered with each other (Fig. 3). The *invA* tree and the consensus tree had identical topologies.

DISCUSSION

Relative variability of *invE*, *invA*, *spaM*, and *spaN*. In an earlier study of sequence variation in *spaO*, *spaP*, and *spaQ*, Li

et al. (27) detected a relationship between the evolutionary rate of change in amino acid sequence and the cellular location of the products of these genes. SpaO, which is unusually variable in amino acid sequence and exhibits <25% sequence identity with its *Shigella* and *Yersinia* homologs, was shown to be secreted into the culture medium (27). In contrast, the amino acid sequences of the membrane-associated SpaP and SpaQ proteins are much less variable and exhibit >60% sequence identity with the corresponding proteins of other enteric bacteria.

The findings of the present study confirm this suggested relationship. We have shown that InvA, which spans the cell membrane, is only weakly polymorphic in *S. enterica* and has >60% sequence identity with its homologs in the other enteric bacteria. The amino acid sequence of the InvE protein, which is also membrane associated (17), is similarly conserved within *S. enterica*. However, SpaN, which is exported (7, 48), is hypervariable in amino acid sequence, like SpaO. Similarly, the rather high level of amino acid sequence variation in SpaM suggests that it is secreted, notwithstanding the fact that Collazo et al. (7) failed to detect it in the supernatant.

The casual basis, adaptive or otherwise, for the hypervariability of the secreted invasion proteins remains to be determined. Among the factors that could generate rapid rates of change in amino acid sequence is selection for antigenic diversity to escape the action of host immune systems. A role for *spaO* in differential host adaptation is ruled out by an absence of amino acid sequence variation among several strongly host-adapted serovars (27). Similarly, the possibility that the InvE, InvA, SpaM, and SpaN proteins play any part in host adaptation is remote. The InvA sequences of the strongly host-adapted serovars Gallinarum (fowl), Choleraesuis (swine), and Typhi (humans) were identical, and that of a strain of the cattle-adapted serovar Dublin differed from the other three by only a single amino acid. The SpaM sequences of strains of these four serovars were identical, and that of a strain of Typhi (S3333) differed from them by only one amino acid substitution. The SpaN proteins in the host-adapted strains were more variable, with those of serovars Enteritidis, Gallinarum, Dublin, and Choleraesuis differing from one another by two amino acids and from that of Typhi by four amino acids.

***invH* and subspecies V.** Molecular genetic analysis has confirmed the division of *S. enterica* into seven distinctive groups and, in addition, has identified an eighth group, subspecies VII, composed of several strains that were previously assigned to group IV (3). Furthermore, this and other studies clearly demonstrate that the group V strains are strongly differentiated from all other salmonellae (3, 40, 43, 44). Given the evidence for strong divergence between subspecies V strains and all other groups, it is probable that we were unable to PCR amplify *invH* from any of the subspecies V strains tested due to primer sequence divergence. However, DNA hybridization under reduced-stringency conditions identified only four subspecies V isolates, all of which showed very weak similarity to the *invH* probe. The possibility exists that *invH* was absence from the most recent common ancestor of this group and was subsequently horizontally transferred into this lineage. An alternative hypothesis for the apparent absence of this gene in some strains may be the result of weak selection for protein function, which would allow differences to accumulate and therefore no homology with the *invH* probe from subspecies I. Also, the *invH* gene differs from the other genes in this operon in many respects: it is not part of the *inv-spa* cluster (there is a large, 457-bp intergenic region between *invH* and the other genes of this operon); *invH* is transcribed in the direction opposite all other genes in the *inv-spa* operon; it has a leader sequence unlike those of the other type III secretion system genes; no homologs to *invH* have been identified in either *Shigella* or *Yersinia* species, whereas other genes in the *inv-spa* operon have homologs in these species; and lastly, the GC content of *invH* is considerably lower than average for the *inv-spa* complex (42% compared to 47%). The increased rate of silent and replacement substitutions in *invH* may be due to the circumstance that *invH* is not an essential gene for this organism.

Ancestry of *Salmonella* pathogenic islands. Like *spaO*, *spaP*, and *spaQ*, which were studied by Li et al. (27), *invA*, *invE*, *spaM*, and *spaN* are present in all eight subspecies of *S. enterica* and have evolved in a pattern and at an average rate similar to those of the average housekeeping gene (Fig. 3). The inference is that the chromosomal region containing the *inv-spa* gene cluster was present in the last common ancestor of all contemporary lineages of the salmonellae. However, there is evidence that genes in the *inv-spa* region in subspecies IV and VII, which differ in their housekeeping gene sequences and in chromosomal genotype (as indexed by multilocus enzyme electrophoresis) (3, 43, 44), have highly similar *invE*, *invA*, *spaM*, *spaN*, *spaO*, *spaP*, and *spaQ* sequences (27; this study), which indicates that all or a major part of the *inv-spa* segment has been horizontally transferred between these two lineages. There is evidence that the central (B) region of *invA* in subspecies IV and VII was acquired from a source not represented in the lineages of the extant subspecies (Fig. 2).

Preliminary analysis of a second pathogenic island (SPI-2)—the *spi* locus (35)—at 30.7 min indicated that it is present in all *S. enterica* subspecies except subspecies V. DNA hybridization studies showed that there was no homology between a 3.2-kb *spi* probe and 13 subspecies V strains (2a). Since subspecies V is the most divergent of all the subspecies at all loci thus far examined (3, 40, 43, 44), it is possible that the *spi* region has diverged significantly in this group. On the other hand, the *spi* region may have been introduced into *S. enterica* after the last common ancestor between subspecies V and all the other subspecies.

ACKNOWLEDGMENT

This work was supported by grant AI22144 from the NIH.

REFERENCES

1. **Altmeyer, R. M., J. K. McNern, J. C. Bossio, I. Rossenshine, B. B. Finlay, and J. E. Galán.** 1993. Cloning and molecular characterization of a gene involved in *Salmonella* adherence and invasion of cultured epithelial cells. Mol. Microbiol. **7:**89–98.
2. **Barinaga, M.** 1996. A shared strategy for virulence. Science **272:**1261–1263.
2a. **Boyd, E. F., and D. L. Hartl.** Unpublished data.
3. **Boyd, E. F., F.-S. Wang, T. S. Whittam, and R. K. Selander.** 1996. Molecular genetic relationships of the salmonellae. Appl. Environ. Microbiol. **62:**804–808.
4. **Boyd, E. F., and R. K. Selander.** 1995. Pattern and process in the evolution of *invH* in *Salmonella enterica* subspecies, abstr. H-91, p. 508. *In* Abstracts of the 95th General Meeting of the American Society for Microbiology 1995. American Society for Microbiology, Washington, D.C.
5. **Boyd, E. F., K. Nelson, F.-S. Wang, T. S. Whittam, and R. K. Selander.** 1994. Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) from natural populations of *Escherichia coli* and *Salmonella enterica*. Proc. Natl. Acad. Sci. USA **91:**1280–1284.
6. **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.
7. **Collazo, C. M., M. K. Zierler, and J. E. Galán.** 1995. Functional analysis of the *Salmonella typhimurium* invasion genes *invI* and *invJ* and identification of a target of the protein secretion apparatus encoded in the *inv* locus. Mol. Microbiol. **15:**25–38.
8. **Eichelberg, K., C. C. Ginocchio, and J. E. Galán.** 1994. Molecular and functional characterization of the *Salmonella typhimurium* invasion genes *invB* and *invC*: homology of InvC to the F_0F_1 ATPase family of proteins. J. Bacteriol. **176:**4501–4510.
9. **Falkow, S.** 1996. The evolution of pathogenicity in *Escherichia*, *Shigella*, and

Salmonella, p. 2723–2729. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, D.C.

10. **Finlay, B. B., M. N. Starnbach, C. L. Francis, B. A. Stocker, S. Chatfield, G. Dougan, and S. Falkow.** 1988. Identification and characterization of Tn*phoA* mutants of *Salmonella* that are unable to pass through a polarized MDCK epithelial cell monolayer. Mol. Microbiol. **2:**757–766.
11. **Galán, J. E.** 1996. Molecular genetic bases of *Salmonella* entry into host cells. Mol. Microbiol. **20:**263–271.
12. **Galán, J. E., C. Ginocchio, and P. Costeas.** 1992. Molecular and functional characterization of the *Salmonella* invasion gene *invA*: homology of InvA to members of a new protein family. J. Bacteriol. **174:**4338–4349.
13. **Galán, J. E., and R. Curtiss III.** 1991. Distribution of the *invA*, *-B*, *-C*, and *-D* genes of *Salmonella typhimurium* among other *Salmonella* serovars: *invA* mutants of *Salmonella typhi* are deficient for entry into mammalian cells. J. Bacteriol. **59:**2901–2908.
14. **Galán, J. E., and R. Curtiss III.** 1989. Cloning and molecular characterization of genes whose products allow *Salmonella typhimurium* to penetrate culture cells. Proc. Natl. Acad. Sci. USA **86:**6383–6387.
15. **Ginocchio, C. C., and J. E. Galán.** 1995. Functional conservation among members of the *Salmonella typhimurium* InvA family of proteins. Infect. Immun. **63:**729–732.
16. **Ginocchio, C. C., S. B. Olmsted, C. L. Wells, and J. E. Galán.** 1994. Contact with epithelial cells induces the formation of surface appendages on *Salmonella typhimurium*. Cell **76:**717–724.
17. **Ginocchio, C. C., J. Pace, and J. E. Galán.** 1992. Identification and molecular characterization of a *Salmonella typhimurium* gene involved in triggering the internalization of *Salmonella* into cultured epithelial cells. Proc. Natl. Acad. Sci. USA **89:**5976–5980.
18. **Groisman, E. A., and H. Ochman.** 1994. How to become a pathogen. Trends Microbiol. **2:**289–294.
19. **Groisman, E. A., and H. Ochman.** 1993. Cognate gene clusters govern invasion of host epithelial cells by *Salmonella typhimurium* and *Shigella flexneri*. EMBO J. **12:**3779–3787.
20. **Groisman, E. A., M. A. Sturmoski, F. R. Solomon, R. Lin, and H. Ochman.** 1993. Molecular, functional, and evolutionary analysis of sequences specific to *Salmonella*. Proc. Natl. Acad. Sci. USA **90:**1033–1037.
21. **Hale, T. L.** 1991. Genetic basis of virulence in *Shigella* species. Microbiol. Rev. **55:**206–224.
22. **Hensel, M., J. E. Shea, C. Gleeson, M. D. Jones, E. Dalton, and D. W. Holden.** 1995. Simultaneous identification of bacterial virulence genes by negative selection. Science **269:**400–403.
23. **Jukes, T. H., and C. R. Cantor.** 1969. Evolution of protein molecules, p. 21–132. *In* H. N. Munro (ed.), Mammalian protein metabolism. Academic Press, New York, N.Y.
24. **Kaniga, K., J. C. Bossio, and J. E. Galán.** 1994. The *Salmonella typhimurium* invasion genes *invF* and *invG* encode homologues of the AraC and PulD family of proteins. Mol. Microbiol. **13:**555–568.
25. **Kumar, S., K. Tamura, and M. Nei.** 1993. MEGA: molecular evolutionary genetics analysis, version 1.0. Pennsylvania State University, University Park.
26. **Lee, C. A., B. D. Jones, and S. Falkow.** 1992. Identification of a *Salmonella typhimurium* invasion locus by selection for hyperinvasion mutants. Proc. Natl. Acad. Sci. USA **89:**1847–1851.
27. **Li, J., H. Ochman, E. A. Groisman, E. F. Boyd, F. Solomon, K. Nelson, and R. K. Selander.** 1995. Relationship between evolutionary rate and cellular location among the Inv/Spa invasion proteins of *Salmonella enterica*. Proc. Natl. Acad. Sci. USA **92:**7252–7256.
28. **Maurelli, A. T.** 1994. Virulence protein export systems in *Salmonella* and *Shigella*: a new family or lost relatives? Trends Cell Biol. **4:**240–242.
29. **Miller, S., E. C. Pesci, and C. L. Pickett.** 1993. A *Campylobacter jejuni* homolog of the LcrD/FlbF family of proteins is necessary for flagellar biogenesis. Infect. Immun. **61:**2930–2936.
30. **Mills, D. M., V. Bajaj, and C. A. Lee.** 1994. A 40-kilobase chromosomal fragment encoding *Salmonella typhimurium* invasion genes is absent from the corresponding region of the *Escherichia coli* K-12 chromosome. Mol. Microbiol. **15:**749–759.
31. **Nelson, K., F.-S. Wang, E. F. Boyd, and R. K. Selander.** 1997. Size and sequence polymorphism in the isocitrate dehydrogenase kinase/phosphatase gene (*aceK*) and flanking regions in strains of *Salmonella enterica* and *Escherichia coli*. Genetics, in press.
32. **Nelson, K., and R. K. Selander.** 1994. Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. Proc. Natl. Acad. Sci. USA **91:**10227–10231.
33. **Nelson, K., and R. K. Selander.** 1992. Evolutionary genetics of the proline permease gene (*putP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli*. J. Bacteriol. **174:**6886–6895.
34. **Nelson, K., T. S. Whittam, and R. K. Selander.** 1991. Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli*. Proc. Natl. Acad. Sci. USA **88:**6667–6671.
35. **Ochman, H., F. C. Soncini, F. Solomon, and E. A. Groisman.** 1996. Identification of a pathogenic island required for *Salmonella* survival in host cells. Proc. Natl. Acad. Sci. USA **93:**7800–7804.
36. **Ochman, H., and E. A. Groisman.** 1995. The evolution of invasion in enteric bacteria. Can. J. Microbiol. **41:**555–561.
37. **Ochman, H., and A. C. Wilson.** 1988. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J. Mol. Evol. **26:**74–86.
38. **Ochman, H., T. S. Whittam, D. A. Caugant, and R. K. Selander.** 1983. Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella*. J. Gen. Microbiol. **129:**2715–2726.
39. **Rahn, K., S. A. De Grandis, R. C. Clarke, S. A. McEwen, J. E. Galán, C. Ginocchio, R. Curtiss III, and C. L. Gyles.** 1992. Amplification of an *invA* gene sequence of *Salmonella typhimurium* by polymerase chain reaction as a specific method of detection of *Salmonella*. Mol. Cell Probl. **6:**271–279.
40. **Reeves, M. W., G. M. Evins, A. A. Heiba, B. D. Plikaytis, and J. J. Farmer III.** 1989. Clonal nature of *Salmonella typhi* and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis, and proposal of *Salmonella bongori* comb. nov. J. Clin. Microbiol. **27:**313–320.
41. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
42. **Salmond, G. P. C., and P. J. Reeves.** 1993. Membrane traffic wardens and protein secretion in Gram-negative bacteria. Trends Biochem. Sci. **18:**7–12.
43. **Selander, R. K., J. Li, E. F. Boyd, F.-S. Wang, and K. Nelson.** 1994. DNA sequence analysis of the genetic structure of populations of *Salmonella enterica* and *Escherichia coli*, p. 17–49. *In* F. G. Priest, A. Ramos-Cormenzana, and B. J. Tindall (ed.), Bacterial diversity and systematics. Plenum Press, New York, N.Y.
44. **Selander, R. K., J. Li, and K. Nelson.** 1996. Evolutionary genetics of *Salmonella enterica*, p. 2691–2707. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, D.C.
45. **Shea, J. E., M. Hensel, C. Gleeson, and D. W. Holden.** 1996. Identification of a virulence locus encoding a second type III secretion system in *Salmonella typhimurium*. Proc. Natl. Acad. Sci. USA **93:**2593–2597.
46. **Stephens, J. C.** 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:**539–556.
47. **Stone, B. J., C. M. Garcia, J. L. Badger, T. Hassett, R. I. F. Smith, and V. L. Miller.** 1992. Identification of novel loci affecting entry of *Salmonella enteriditis* into eukaryotic cells. J. Bacteriol. **174:**3945–3952.
48. **Zierler, M. K., and J. E. Galán.** 1995. Contact with cultured epithelial cells stimulates secretion of *Salmonella typhimurium* invasion protein InvJ. Infect. Immun. **63:**4024–4028.