Evolution and Distribution of the ospC Gene, a Transferable Serotype Determinant of Borrelia burgdorferi

Alan G. Barbour and Bridgit Travinsky
Departments of Microbiology and Molecular Genetics, Medicine, and Ecology and Evolutionary Biology, University of California, Irvine, California, USA

ABSTRACT  Borrelia burgdorferi, an emerging bacterial pathogen, is maintained in nature by transmission from one vertebrate host to another by ticks. One of the few antigens against which mammals develop protective immunity is the highly polymorphic OspC protein, encoded by the ospC gene on the cp26 plasmid. Intragenic recombination among ospC genes is known, but the extent to which recombination extended beyond the ospC locus itself is undefined. We accessed and supplemented collections of DNA sequences of ospC and other loci from ticks in three U.S. regions (the Northeast, the Midwest, and northern California); a total of 839 ospC sequences were analyzed. Three overlapping but distinct populations of B. burgdorferi corresponded to the geographic regions. In addition, we sequenced 99 ospC flanking sequences from different lineages and compared the complete cp26 sequences of 11 strains as well as the cp26 bbb02 loci of 56 samples. Besides recombinations with traces limited to the ospC gene itself, there was evidence of lateral gene transfers that involved (i) part of the ospC gene and one of the two flanks or (ii) the entire ospC gene and different lengths of both flanks. Lateral gene transfers resulted in different linkages between the ospC gene and loci of the chromosome or other plasmids. By acquisition of the complete part or a large part of a novel ospC gene, an otherwise adapted strain would assume a new serotypic identity, thereby being comparatively fitter in an area with a high prevalence of immunity to existing OspC types.

IMPORTANCE  The tick-borne zoonosis Lyme borreliosis is increasing in incidence and spreading geospatially in North America. Further understanding of the evolution and genetics of its cause, Borrelia burgdorferi, in its environments fosters progress toward ecologically based control efforts. By means of DNA sequencing of a large sample collection of the pathogen from across the United States, we studied the gene for the bacterium’s highly diverse OspC protein, protective immunity against which develops in animals. We found that the distributions and frequencies of types of OspC genes differed between populations of B. burgdorferi in the Northeast, the Midwest, and California. Over time, OspC genes were transferred between strains through recombinations involving the whole or parts of the gene and one or both flanks. Acquisitions of OspC genes that are novel for the region confer to recipients unique identities to host immune systems and, presumably, selective advantage when immunity to existing types is widespread among hosts.

Lyme borreliosis is a vector-borne zoonosis with multiple reservoir hosts within sylvatic cycles in temperate regions of the northern hemisphere (1). In North America, the agent is the spirochete Borrelia burgdorferi. A related, often sympatric species is Borrelia bissettii, but this species has not been associated with human disease. Among the reservoirs for B. burgdorferi are rodents, other small mammals, and ground-foraging birds. Its tick vectors in North America are Ixodes scapularis and Ixodes pacificus. Sequential and mixed infections of reservoirs and tick vectors are common (2, 3). In regions where B. burgdorferi is enzootic, sites as small as a few hectares have between 9 and 15 strains (4–6).

Lyme borreliosis is the most frequently reported arthropod-borne infection in the United States and continues to increase in incidence and geographic area of risk (7). Three regions in the United States have enzootic transmission of B. burgdorferi and moderate-to-high risk of infection for humans; these regions comprise several states of the Northeast (Connecticut, Massachusetts, New York, Rhode Island, New Jersey, Maryland, and Pennsylvania), the Midwest (Michigan, Illinois, Minnesota, and Wisconsin), and northern California. The absence or rarity of B. burgdorferi in Ohio and in the Rocky Mountain and Great Basin regions suggests that these three populations of B. burgdorferi are geographically if not ecologically isolated.

Our aim was not to reconstruct the history of the species B. burgdorferi in North America. Other studies have endeavored this (8, 9). Rather, our attention was on a systematically collected set of samples of B. burgdorferi from across the United States and the use of those samples to infer the evolution of the highly polymorphic ospC gene, which encodes OspC, a surface-exposed lipoprotein of B. burgdorferi and other Lyme borreliosis-related species. OspC is homologous to the Vsp protein of the sister taxon of species that cause relapsing fever (10). The OspC and Vsp fam-
ilies of proteins have marked differences in their primary sequences within each family while retaining an α-helical bundle structure (11, 12). Whereas relapsing fever Borrelia genomes each have several different vsp alleles (13, 14), B. burgdorferi genomes have a single ospC gene, which is located on cp26, a circular plasmid of 26 kb (15). There is no antigenic variation of OspC during experimental infection (16), but within a given geographic area, several strains, each expressing a different OspC protein, coexist (4, 6).

A serotype is "a serologically distinguishable strain of a microorganism" (Oxford English Dictionary, 2nd ed., 1989). If any single component of B. burgdorferi determined serotypic identity, OspC could be it, for the following reasons. (i) No other gene of B. burgdorferi approaches ospC in diversity of alleles (17). (ii) The type-specific immunity conferred by immunization with OspC corresponds to the strain-specific immunity to B. burgdorferi in natural infections (18, 19). An animal immunized with one OspC type is protected against infection by a strain with the same OspC type but not a strain with a different one. (iii) The duration of expression of ospC by B. burgdorferi in mice coincides with the 1- to 3-week period in which B. burgdorferi circulates in the blood (20–22).

The population structure of B. burgdorferi reflects the contributions of mutation and recombination and contains both clonal and nonclonal elements (4, 17, 23–25). Traces of lateral gene transfer and recombination were evident in the plasmids of this species (17, 23, 26). While the ospC gene is heavily marked by the recombination process (27–29), this locus was reported to have phylogenetic consistency with chromosomal loci for strain collections largely limited to the Northeast (4, 23, 30). When more strains from the Midwest were included in the analysis, exceptions to strict linkage disequilibrium between ospC alleles and chromosomal markers were observed (24).

In our study of strains from the Midwest and California as well as the Northeast, we noted incongruities between ospC alleles and chromosomal loci across the three geographic areas (25). Some ospC alleles occurred in two different lineages, and there were indications that the same lineage could have two different ospC alleles. Differences between the B. burgdorferi populations of the Northeast and the Midwest were also apparent by multilocus sequence typing (MLST) of the spirochete in I. scapularis ticks collected over different transmission seasons (9).

For the present study, we first completed determination of the ospC genotypes of B. burgdorferi from a survey of infected I. scapularis ticks from the Northeast and the Midwest (8, 31) and then included samples from I. pacificus ticks from northern California (5) and additional ticks from the Midwest. For a fuller accounting of the variety of recombinations involving ospC genes, we extended the sequence analysis to the flanking regions of ospC genes in a large sample of strains and to the entire cp26 plasmid for a subset of these. These results provide insight into the evolution and distribution of ospC of Lyme borreliosis-related Borrelia spp.

RESULTS

Diversity and geographical distribution of ospC alleles. We accessed the most extensive and systematic collection of B. burgdorferi samples yet assembled. Twenty-five ospC types were identified in populations in the Northeast, the Midwest, and California (see Table S1 in the supplemental material). There were two subtypes, A and B, each for types D, H, I, and U, and three subtypes, A, B, and c, for type F. The codon-based nucleotide alignment of the 32 types and subtypes excluded the coding sequence for the signal peptide, the conserved first 11 residues of the mature lipoprotein, and the conserved last 4 C-terminal residues (see data set W1 at http://spiro.mmg.uci.edu/data/ospC). There were 549 sites in the alignment with at least one gap at 30 positions; 294 of the sites were variable. The overall nucleotide diversity per site (\( \pi \)) was 0.189. The 300 pairwise distances between the 25 ospC type sequences of B. burgdorferi were normally distributed, with a mean and median of 0.29 and a standard deviation of 0.05 (see Fig. S1 in the supplemental material). Figure 1 shows the unrooted, network-based phylogram of the 32 type and subtype sequences of B. burgdorferi and an ospC sequence of B. bissettii. With some exceptions, which are considered below, the DNA sequences of the alleles are approximately equidistant from each other.

Table 1 summarizes the frequencies of ospC alleles in I. scapularis nymphs in 3 geographic regions; Data Set S1 in the supplemental material provides the determinations by site and geographic coordinates. Without subtype distinctions, there were 18 alleles in the Northeast, 23 alleles in the Midwest, and 12 alleles at the northern California sites. The tick specimens from the Northeast (\( n = 396 \)) and the Midwest (\( n = 443 \)) were collected by the same protocol and over the same time periods (8). There were 18 types (exclusive of subtypes) common to both regions, but their individual prevalences among 396 sequences from the Northeast and 393 sequences from the Midwest were statistically different (likelihood ratio statistic = 211; \( P = 10^{-9} \)). The coefficient of determination (\( R^2 \)) for rankings paired for the 18 types was only 0.18 (Spearman rank Z = 1.52; \( P > 0.05 \)).

Figure 2 shows the collection sites for I. scapularis nymphs in the Northeast (\( n = 23 \)) and the Midwest (\( n = 28 \)) and the Bayesian
posterior probability contours for a model of two different populations of B. burgdorferi, as defined by ospC alleles (see Data Set S1 in the supplemental material). With a longitude of 83°W, which passes through Detroit, MI, and Columbus, OH, as a dividing line, the probabilities for membership in one or the other cluster were 0.99 for the first cluster for 396 of the 396 samples from the Northeast and 0.99 for the second cluster for 442 of the 443 samples from the Midwest. The probabilities were lower overall when the models specified 3, 4, or 5 populations. When the ospC alleles and their geographic locations for the 839 samples were randomly permuted as a negative control, only one population was defined.

The I. pacificus ticks were collected in California under a different protocol. While data from the two collections are not fully commensurate, the relative frequencies of the different alleles between the three regions can be compared (Table 1). Some ospC types, such as A and H, occurred in all 3 regions, but others were absent from one or two regions. For example, type K was the most prevalent allele in the Northeast sample but was a minor type in the Midwest and was not detected at all in California. Only 9 (50%) of the 18 ospC alleles from the Northeast were found in the California samples. Nevertheless, for each region that was sampled, the pairwise distances in sequence between the prevalent alleles were very similar in their distributions: the means ± standard deviations were 0.29 ± 0.05, 0.29 ± 0.05, and 0.28 ± 0.05 for the indigenous alleles from the Northeast, the Midwest, and California, respectively. This suggests that across the three regions, the balancing selections were of the same magnitude.

**Recombination within ospC genes.** These ospC findings confirmed the results obtained with chromosomal loci (5, 9): the three populations of B. burgdorferi overlap but are genetically distinct. Not only did the ospC alleles differ in frequency between regions, they also differed in their associations with chromosomal and other loci across regions (25). One explanation for the latter finding is lateral gene transfer of ospC genes or their fragments. Before considering recombination events extending beyond the ospC locus, we examined the evidence for recombination within ospC alleles.

The “star” pattern and the tangle at the root of the ospC tree in Fig. 1 indicate a mosaic genetic structure for ospC in these populations. The recombination breakpoints and diversity levels were highest in the middle of the coding sequence (see Fig. S2 in the supplemental material). Most ospC alleles were patchworks, the accumulated effects of multiple recombinations involving donors within the species and genus (27). A possible relic of an event is the 15-bp region (positions 532 to 546), by which subtypes Ia and Ib solely differ. In Ia, the sequence is GAA TCA GTA AAA AAC, which encodes the peptide ESVKN, while in Ib, it is AAA GCA GTA GAG GTC, which encodes KAVEV. The former nucleotide

---

**TABLE 1 Prevalence of ospC alleles of B. burgdorferi in Ixodes ticks in 3 geographic areas**

<table>
<thead>
<tr>
<th>ospC type or subtype</th>
<th>Northeast</th>
<th>Midwest</th>
<th>California</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of ospC alleles</td>
<td>Prevalence</td>
<td>Rank</td>
</tr>
<tr>
<td>A</td>
<td>57</td>
<td>0.144</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>48</td>
<td>0.121</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>15</td>
<td>0.038</td>
<td>11</td>
</tr>
<tr>
<td>Da</td>
<td>15</td>
<td>0.038</td>
<td>12</td>
</tr>
<tr>
<td>Db</td>
<td>0</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>16</td>
<td>0.040</td>
<td>10</td>
</tr>
<tr>
<td>Fa</td>
<td>18</td>
<td>0.045</td>
<td>7</td>
</tr>
<tr>
<td>Fb</td>
<td>2</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>Fc</td>
<td>0</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>0.025</td>
<td>13</td>
</tr>
<tr>
<td>Ha</td>
<td>21</td>
<td>0.053</td>
<td>6</td>
</tr>
<tr>
<td>Hb</td>
<td>1</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>Ib</td>
<td>19</td>
<td>0.048</td>
<td>8</td>
</tr>
<tr>
<td>J</td>
<td>3</td>
<td>0.008</td>
<td>16</td>
</tr>
<tr>
<td>K</td>
<td>77</td>
<td>0.194</td>
<td>1</td>
</tr>
<tr>
<td>L</td>
<td>1</td>
<td>0.003</td>
<td>17</td>
</tr>
<tr>
<td>M</td>
<td>4</td>
<td>0.010</td>
<td>15</td>
</tr>
<tr>
<td>N</td>
<td>37</td>
<td>0.093</td>
<td>4</td>
</tr>
<tr>
<td>O</td>
<td>6</td>
<td>0.015</td>
<td>14</td>
</tr>
<tr>
<td>T</td>
<td>17</td>
<td>0.043</td>
<td>9</td>
</tr>
<tr>
<td>Ua</td>
<td>28</td>
<td>0.071</td>
<td>5</td>
</tr>
<tr>
<td>Ub</td>
<td>0</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>0</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>1</td>
<td>0.003</td>
<td>18</td>
</tr>
<tr>
<td>C3</td>
<td>0</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td>0</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>E3</td>
<td>0</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td>0</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>H3</td>
<td>0</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>I3</td>
<td>0</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>396</td>
<td>1.000</td>
<td>443</td>
</tr>
</tbody>
</table>

*If counts were equal, then the same rank was assigned; subtypes of D, F, H, and U were combined.*

*All type I alleles were subtype Ia.*

---

*mbio.asm.org*
sequence is identical to the corresponding sequence of the type M allele, while the latter is identical to type E3's sequence.

Some pairs of ospC alleles plausibly have a closer evolutionary relationship. In the phylogram in Fig. 1, the pairs comprising H, J, F and I3, N and D3, F3 and D, and C3 and an ospC locus of *B. bissettii* are indicated by significant support for common internal nodes. These pairs featured longer stretches of identity or near-identity between the types than was observed for other pairs (see Fig. S3 in the supplemental material). I3 and A, comprising a type pair, do not stand out as recombinants in Fig. 1, but this is more clearly evident in the alignment of their sequences. As Girard et al. noted (5), type I3 OspC is a chimera of type F for the first two-thirds of the protein and of type A for the last third.

**Recombination of cp26 plasmids.** California strains with a type I3 ospC locus had *rrs-rrlA* and *rrfA-rrlB* intergenic spacers that were identical in sequence to those of type Fa strains but not those of type A strains (25), an indication that a type A-bearing strain was the donor for the chimeric gene. We return to this particular event after first examining the more general case of recombination of cp26 plasmids.

For this analysis, we assembled complete sequences of the 26-kb plasmids bearing ospC in the species. The cp26 sequences of the following 10 strains were publicly available: B31, 64b, ZS7, WI91-23, 29805, 94a, 72a, 118a, CA-11.2a, and 156a. All but 3 strains (ZS7, WI91-23, and CA-11.2a) were from the Northeast. For additional representation from outside this region, we determined the sequence of cp26 of the California isolate CA8.

Since the mosaic character of the ospC locus itself could overwhelm the detection process, we removed its coding sequence from the alignment. The resultant 11 aligned sequences had 25,934 positions, with 749 informative sites (see data set W2 at http://spiro.mmg.uci.edu/data/ospC).

The numbers of recombination events per site (rho) and mutation events per site (theta) were 0.051 and 0.011, respectively, for a rho/theta ratio of 4.7. For the entire 11-sequence alignment with a window size of 100 characters, the PhiTest for recombination gave a mean test value of 0.563, with a variance of $10^{-10}$ and an observed value of 0.203 ($P < 10^{-5}$). Figure 3 shows the inferred recombination breakpoints and their significances along the lengths of the aligned sequences. The highest density of breakpoints surrounded the position which ospC would otherwise occupy. The only other location with a near-comparable density of breakpoints was centered on position 6400, which is in the open

![FIG 2 Population structure of the ospC gene of *B. burgdorferi* in *Ixodes scapularis* nymphs in the Northeast (23 sites) and the Midwest (28 sites) regions of the United States for the years 2004 to 2007. Upper panel, overview map of posterior mode of population membership. Middle and lower panels, contour of posterior probabilities (0.1 to 1.0) for two populations defined by 27 ospC sequence types and subtypes identified in these regions. The y coordinates are latitude, and the x coordinates are longitude. The data for the analysis and the resultant posterior probabilities for cluster 1 or 2 for each site are given in Data Set S1 in the supplemental material.](http://spiro.mmg.uci.edu/data/ospC)
FIG 3  Recombination detection analysis of cp26 plasmids of 11 strains of B. burgdorferi. The ospC coding sequences were removed at the position indicated by the point of the arrowhead. The alignment was subjected to the Sciscan algorithm with a window size of 200, a step size of 20, and 100 permutations. The x axis provides the nucleotide positions of the alignment. The bottom panel (“Hits” on the y axis) indicates the sequence region where recombination was detected. The middle panel gives the number of estimated breakpoints per window. The top panel indicates the log10 of the proportion of overlapping characters, which were linked to different plasmids, and accordingly, we determined the bbb02 sequences of 45 additional B. burgdorferi strains, with an emphasis on strains with different linkages of ospC types to chromosomal loci (see data set W3 at http://spiro.mmg.uci.edu/data/ospC).

Recombination beyond the ospC genes. Examining flanking regions for the ospC gene, we noted that the sequences on each side could be conveniently grouped into 13 sets of oligonucleotide characters, designated A to G, that each included informative polymorphisms (see Fig. S6 in the supplemental material). (Here, “character” is defined in accordance with systematics usage: a variable feature with two or more different states.) For instance, among the 11 cp26 plasmids, there were only 4 variants for the sequence beginning 67 nucleotides 5′ of the ospC start site: CAAATA, CAAAT, CAAAT–, and ATTTGA. There were 5 oligonucleotide characters, designated a, b, c, d, and e, that were upstream of the ospC coding region, and downstream of the stop codon were the characters h, i, j, k, l, and m. We also included in the analysis the characters f and g, from the front and end of the ospC gene, respectively. With the exception of the g character, which typified ospC diversity, there were no more than 6 variants per character.

We amplified and sequenced a cp26 fragment that extended from a position 229 nucleotides upstream of the ospC gene start codon to one 534 nucleotides downstream of the stop codon. This ~1.5-kb fragment corresponded to the plasmid region with the high density of probable breakpoints (Fig. 3). The sequencing was carried out on 99 selected isolates or tick extracts with ospC alleles that were linked to different rrs-rrlA loci.
and were from different geographic regions (see Data Set S2 in the supplemental material). For the alignment, we also included the corresponding sequences from the 11 strains for which the complete cp26 sequences were available. Figure 4 schematically represents the variety of patterns for the 13 oligonucleotide characters before, within, and following ospC genes. Included in the figure are the geographic origins, the rrs-rrlA intergenic spacer and MLST genotypes, and the cp26 classification by bbb02 genotype 1 or 2.

The type I3 ospC gene observed in several samples from California is attributable to a recombination between a recipient strain bearing a cp26 plasmid with a type Fa ospC gene and a donor strain bearing a cp26 plasmid with a type A gene (Fig. 4, group II). The I3 isolates had the same rrs-rrlA and rrFA-rrlB intergenic spacer sequences (25) and the same bbb02 sequences (see data set W3 at http://spiro.mmg.uci.edu/data/ospC) as Fa isolates. The I3 ospC gene characters f and g were the same as the corresponding characters of type F and type A, as befits a chimera. But the 5’ and 3’ flanking regions for the I3 ospC gene were identical to those for subtype Fa ospC and not those for type A-bearing cp26 plasmids. The presumptive proximal crossover was within the sequence TACTGATG, which both type A and subtype Fa share at positions 450 to 457 of the 630-bp-long type A gene. The I3, Fa, and A strains all had variant 1 of oligonucleotide character h but differed over characters i, j, k, and l, suggesting that the distal crossover point was either among the coding sequence’s last 50 nucleotides or among the following 106 nucleotides.

Strains bearing ospC subtype Ha or Hb were an example of another type of recombination (Fig. 5, group II). The three strains had different MLST profiles and different ospA sequences (25). They can also be distinguished by their 5’ and 3’ flanking regions. The Ha-bearing strain in the Northeast has the same 3’ flanking region as Hb-bearing strains of the Midwest, but over their 5’ flanks, the Midwest and California strains with Hb alleles are identical. The Ha and Hb alleles differ by a single synonymous substitution, which is near the gene’s 5’ end, consistent with a recombination involving the 5’ flanking region and the ospC gene itself. All the type H representatives were bbb02 genotype 1.

The other pairs of strains with the same or near-identical ospC genes, different MLST or rrs-rrlA spacer genotypes, and differences in their 5’ and/or 3’ flanking regions involved types B, I, and K. Two strains, exemplified by isolates 64b from the Northeast and ZS7 from Europe, had subtypes Ba and Bb, respectively. There are several differences between 64b and ZS7 in the 5’ flank to the ospC genes (Fig. 4; see also Data Set S2 in the supplemental material). Three of the four polymorphic positions distinguishing Ba and Bb occur in the first third of the sequence and cluster within 15 positions. This is consistent with a lateral transfer of a fragment that included the 5’ end of ospC and the adjacent upstream region. Whereas the strains with subtypes la and lb differed in their 5’ flanking regions, the two polymorphic positions between the ospC alleles occurred in character g at their 3’ ends. The pairs involving the B and I strains had the same bbb02 genotypes. The type K strains from the Northeast and the Midwest had identical ospC sequences but differed in their 5’ and 3’ flanking regions as well as both the rrs-rrlA spacer and the bbb02 genotypes, suggestive of the transfer of the entire ospC gene between lineages.

Another group, group III (Fig. 4), comprised the pairs involving types D, G, N, and T, which had different chromosomal genotypes but the same ospC genes and the same flanking regions to the extent of our sequencing. Also qualifying for this group were the three type A strains in the sampling. The two type A strains from the Northeast or California and the three samples of a Midwest type A strain had bbb02 sequences that were in the two different clades, suggestive of transfer of either an entire plasmid or an extensive length of a plasmid.

**Recombination in the promoter region.** A possible consequence of replacement of all or part of ospC is a collateral effect on adjacent loci or on regulatory regions. We noted that the oligonucleotide character e was located just upstream of the “−35” box of the ospC promoter. Substitutions in this area could affect the inverted repeats implicated in regulation of ospC expression as an operator or through supercoiling (33, 34). Two type H strains, one of subtype Ha and the other of subtype Hb, differed in character e (Fig. 4). Figure 5 shows for these two strains the upstream sequences, numbered in reference to the transcriptional start site (35); oligonucleotide e corresponds to positions −55 to −42. The first inverted repeats, which spanned positions −105 to −54, were the same in sequence and location for the pair, but the second inverted repeats, which included the “−35” σ70 promoter element, were different by an indel and 5 substitutions. Although it was shifted upstream by 4 nucleotides and had a different sequence, the Hb strain still had a predicted stem-loop and a ΔG and a melting temperature (Tm) of −25 kcal/mol and 62.7°C, instead of −22.8 kcal/mol and 61.6°C, respectively, for the Ha strain.

**Transfer of an entire ospC gene.** To this point, we have examined pairs or trios of strains with the same or near-identical ospC alleles and found evidence of lateral gene transfer of all or part of the ospC gene and, in addition, different lengths of sequence on one or both sides of the locus. We next looked at another possible outcome of lateral gene transfer, namely, the occurrence of substantially different ospC genes in members of the same lineage. Notwithstanding the cumulative effects of intra- and interspecies recombination on the chromosome as well as plasmids (17, 36, 37), there was evidence that two strains, 72a and 118a, occupied an internal node of comparatively recent origin. Although strains 72a and 118a have type G and J ospC alleles (24) and different ospA alleles on their lp54 plasmids (25), strains 72a and 118a had the same rrFA-rrlB intergenic spacer (25) and the same dbpA sequences (see Fig. S4 in the supplemental material).

We extended the comparison to include sequences for 8 ribosomal protein genes, which are considered informational genes.
sequences of 11 strains and the \textit{ospC} flanking regions for a larger set of strains, we identified a variety of recombination events that contributed to the nature of North American \textit{B. burgdorferi}.

In our view, strain-specific immunity of reservoir hosts is sufficient to account for the strong balancing selection at the \textit{ospC} locus that is notable in \textit{B. burgdorferi} population structures (4, 6, 40). But \textit{OspC} has also been characterized in functional terms: some strains, defined by their \textit{ospC} alleles, are associated with higher likelihoods of dissemination beyond the skin in humans or experimental animals (41, 42). One study attributed the different \textit{OspC} phenotypes to differential binding of plasminogen (43). Although there are other candidates for host range determinants, such as complement-regulator factor H binding proteins (26, 44), a role for \textit{OspC} in adaptations to different niches cannot be excluded. So, \textit{ospC} diversity could arguably reflect the outcome of niche selection processes (41, 45). Nevertheless, we doubt that the observed antigenic diversity of \textit{OspC} is merely epiphenomenal to functional differences between proteins. The range of pairwise sequence distances among \textit{ospC} alleles nearly matches that of the highly polymorphic family of surface proteins of the relapsing fever agent \textit{Borrelia hermsii}, which employs antigenic variation to evade host immunity (14). Possibly, both immune and niche selective forces are in play, but their relative contributions remain to be determined.

Retention of a gene, like \textit{ospC}, that is necessary for tick-to-vertebrate transmission is more ensured by its location on cp26, apparently the only indispensable plasmid (15, 46). While possession of a cp26 plasmid is required for cell replication, it need not be the original cp26 plasmid. The plasmid encodes compatibility functions, and one cp26 plasmid can be displaced by another if they are incompatible (46). This potentially allows for replacement of entire cp26 plasmids through lateral gene transfer as well as a range of products of recombination between two plasmids that transiently coexist in the same cell. Transfer between \textit{B. burgdorferi} of segments of DNA of \textless 1 kb was noted (27), but recent findings indicate that lateral gene transfer may involve longer lengths of DNA.

We classify recombinations involving \textit{ospC} into 5 patterns. The first is intragenic, that is, effectively limited to the \textit{OspC}-coding sequence itself. Recombination within \textit{ospC} was noted in several reports, beginning with Livey et al. (29), and accounts for its mosaic genetic structure. Possible examples of intragenic recombination include the \textit{ospC} pairs comprising H and I, C and B, and E and H3 (see Fig. S3 in the supplemental material). But our focus here is on recombination outcomes that extend beyond \textit{ospC}'s boundaries. The other four patterns are those that involve (i) part of the \textit{ospC} coding sequence and a sequence extending into the 5′ flanking region, (ii) part of the \textit{ospC} coding sequence and a sequence extending into the 3′ flanking region, (iii) the \textit{ospC} gene and both

**DISCUSSION**

Shakespeare’s late plays were more collaborative in authorship than was previously thought (39), and the same can be said of the origins of existing \textit{B. burgdorferi} strains. Intraspecies recombination had a greater role in shaping the evolution of \textit{B. burgdorferi} than was previously appreciated. With representation from three geographic regions, there were several exceptions to linkage disequilibrium between the plasmid-borne \textit{ospC} gene and chromosomal loci (25). The present study extended the geospatial analysis and demonstrated that the population structures for the \textit{ospC} locus overlapped between the three regions but were distinguishable, thereby confirming the results with MLST and \textit{rrs-rrlA} spacer loci from smaller sample sizes (5, 8, 9). By analyzing whole cp26

![FIG 6](image-url) Nucleotide polymorphisms between the cp26 plasmid sequences, excluding the \textit{ospC} genes, of strains 72a and 118a of \textit{B. burgdorferi}. The position in the sequence where the \textit{ospC} gene sequence was deleted is shown by an arrow.

and, thus, less susceptible to whole or partial replacement than are operational genes, such as that for a metabolic enzyme (38). Among the 11 strains with genome sequences, only 72a and 118a had identical sequences for each of these 8 ribosomal protein genes (see Fig. S7 in the supplemental material). With the MLST set of eight operational housekeeping genes, base substitutions between 72a and 118a were noted, but these were fewer than was observed between other pairs of strains (Fig. S7), and the two strains retained their positions with respect to strain CA-11.2a. Strains 72a and 118a also had the same \textit{bbb02} genotypes and the most closely related cp26 and lp54 plasmids among the 11 strains examined (Fig. S4 and S5). The taxonomic relationship of 72a and 118a with CA-11.2a that was observed with the two sets of chromosomal loci held true for the cp26 sequences. Figure 6 shows the locations of nucleotide differences between the cp26 plasmids of strains 72a and 118a with exclusion of \textit{ospC} coding sequences. The greatest difference, by far, between the two cp26 sequences was at positions on each side of \textit{ospC}, a region extending for \textless 2 kb on the 5′ side and \textless 1 kb on the 3′ side.
flanks but not the entire plasmid, and (iv) replacement of the entire cp26 plasmid.

Inclusion of a sequence upstream of the ospC coding sequence in the recombinant fragment could affect the promoter and inverted repeats that may constitute a regulatory element (33, 34). There resides also the guaA-guaB (bbb17 and bbb18) operon, beginning on the complementary strand 185 nt before the ospC start site. On the 3’ side, there is a highly conserved sequence that would form a stem-loop typical of the rho-independent terminator (12, 35). There are also two short open reading frames for hypothetical peptides of 36 (BBB20) and 31 (BBB21) amino acids (aa) before the stop codon 431 nucleotides downstream on the complementary strand for the open reading frame BBB22, which is homologous to xanthine/uracil permeases of other bacteria. While there may be greater scope for rearrangements without disruptive effects downstream of ospC, a transcription-regulatory element may be changed in sequence without necessarily altering its function (Fig. 5).

When both flanking regions are involved in the recombination, an entire ospC allele may be substituted. The evidence is strongest for the closely related strains 72a and 118a. The presence of the same ospC type in different lineages is exemplified by the type K strains from the Northeast and the Midwest in the collection (Fig. 4). Other possible examples of the latter phenomenon involve the D, G, N, and T strains from different geographic areas. Recombinations that involved part of the ospC gene and either the 5’ or the 3’ flanking region were exemplified by strains of types H, B, and I.

An outcome (outcome 5), i.e., a replacement of one cp26 plasmid by an incompatible plasmid, has been observed in the laboratory (46). We observed discordant tree topologies for the cp26 sequences and the two sets of chromosomal loci (see Fig. S5 and S7 in the supplemental material). Only the strains 118a, 72a, and CA-11.2a maintained the same taxonomic relationship in all 3 phyla genes. But definitive examples of the outcome (outcome 5) were not found in the subset of strains for which the whole-genome sequence was available. The traces of this may be more apparent as more whole-genome sequences are available for comparative analysis.

The mechanisms for lateral gene transfer in B. burgdorferi in nature are unknown. As mixed infections are common (2, 3, 25, 47), there are opportunities for genetic exchange in both reservoir hosts and vectors. There is no evidence of conjugation in the laboratory (48). But membrane vesicles or blebs, which have been shown to contain plasmid DNA (49), could be the vehicles for the higher frequency of transformation events in nature. The cp26 plasmid itself is not a prophage, but transduction of cp26 fragments via another virus, such as the prophage constituting the cp32 replicons (50), is possible.

The ospC gene is clearly transferable, but is it a mobile genetic element? It does not appear to have accessory genes associated with it. The flanking genes, guaA and guaB on one side and bbb22 on the other, encode enzymes for nucleotide metabolism or uptake; these enzymes are not discernibly transposable or integrase. There may be a role for the sets of inverted repeats, which are on each side of ospC and potentially form recombogenic stem-loop structures. These inverted-repeat regions may be included in the transfer, as we have demonstrated. But they need not be, since both recipient and donor have them for their transcription regulation and termination functions. Although we have found evidence of transfer of entire ospC genes in some lineages, a single recombination with incorporation of only part of the gene may be sufficient to confer a new antigenic identity for the recipient cell.

Finally, can we accommodate both the phylogeography and the inferred mechanisms of genetic diversity into a model of the evolution of this pathogen? OspC’s prominence as an abundant, immunogenic surface protein, which is expressed as the spirochete enters the host’s skin and then circulates in the blood, makes this protein an important target for protective immunity. The vertebrate hosts’ immune responses subject the ospC gene to frequency-dependent balancing selection. While recombination does not create polymorphisms at the single-nucleotide level, interstrain exchange of two or more suitably distant sequences can yield novel combinations of substitutions and indels and, as an eventual consequence, a set of antigenically distinctive proteins. One now sees the cumulative effects of intragenic recombination, involving both long and short fragments and occurring in multiple rounds, in the highly polymorphic repositories of ospC genes in three different populations of B. burgdorferi. But we have also seen that intragenic recombination may involve either of its flanking regions. Indeed, it may depend on one of these flanks for a stable heteroduplex if homologous rather than illegitimate recombination is the more common mechanism.

If both flanks were the substrates for a recombination with a heterologous sequence stretch between them, as occurs in the relapsing fever agent B. hermsii during antigenic variation (14), transfer of an entire ospC gene into a different strain would be achieved. This fits the general category of serotype shift and is distinguished from serotype replacement, in which the population structure of a pathogen changes as newly introduced strains gain a foothold in the presence of herd immunity to existing strains. In the case of ospC, a serotype shift would not create a novel allele per se when the greater population of B. burgdorferi is taken into account. But within a partially isolated geographic area, such as the Northeast, with the introduction, e.g., through migratory birds, of a new strain with an ospC locus that is locally unique, acquisition of this one determinant would presumably suffice for enhanced fitness when the beneficiary faces the prospect of reservoir hosts, a large proportion of which are immune to existing strains. The invading bacterial strain itself may not prosper in the new environment, perhaps for lack of other adaptations, e.g., a putative tick midgut adhesin or host-specific complement resistance, suited for parasitism of local ticks and reservoir hosts, such as I. scapularis and the deer mouse (Peromyscus leucopus) at the Northeast sites or I. pacificus and the western gray squirrel (Sciurus griseus) in California (5).

We propose that the ospC phenomenon is analogous to a single gene that upon acquisition and expression provides for a bacterium resistant to an antibiotic in an antibiotic-rich environment, like a hospital or poultry facility. We acknowledge that there may be unrecognized episodic relationships for ospC that operate to constrain the variety of genetic backgrounds of B. burgdorferi in which a particular OspC protein can effectively function. But as long as a newly acquired ospC gene is faithfully positioned next to the promoter, retains the coding sequence for the conserved signal peptide, and is not truncated at its 3’ end, we presume that the novel OspC protein will be expressed and successfully transported.
to the outer membrane and the cell’s surface and function in its new bacterial host (51).

**MATERIALS AND METHODS**

**Strains and culture.** The cultivated *B. burgdorferi* strains were B31 (ATCC 35210), N40, JD 1, Sh2, 2665, and HB19 from the Northeast (52) and CA8, CA11, CA12, CA15, CA16, CA17, CA172, CA337, CA533, and CA534 from northern California (53). Strains VQG, WQR, WQR27, and QQ from the Northeast were provided by Merial Limited, Athens, GA. The strains were cultivated in modified Barbour-Stoenner-Kelly II medium and harvested by centrifugation at 9,500 × g for 20 min at 22°C (52).

**DNA samples from ticks.** The states of the Northeast represented in the study (with the number of ospC sequences per state indicated in parentheses) were Connecticut (28), Massachusetts (8), Maryland (55), Maine (39), New Hampshire (3), New Jersey (16), New York (176), Pennsylvania (47), Rhode Island (11), and Virginia (13); the Midwest states were Iowa (8), Illinois (17), Indiana (4), Michigan (20), Minnesota (218), and Wisconsin (176). The procedures for (i) collection of 7,749 questing *I. scapularis* nymphs during the years 2004 to 2007 at 23 collection sites in the Northeast and 28 in the Midwest, with recording of geographic coordinates, (ii) extraction of DNA from the ticks, (iii) quantitative PCR for identification of ticks with *B. burgdorferi*, and (iv) genotyping of the *B. burgdorferi* isolates were described previously (8, 25). *B. burgdorferi* was identified in 1,540 (20%) of the ticks. PCR amplification of *ospC* genes was carried out blindly with respect to geographic location and was attempted on the 1,522 specimens for which sufficient DNA was available. We reported previously on 741 extracts, in which only a single *ospC* sequence was detected (25). Here, we include the results from an additional 198 extracts, out of a total of 241 with evidence of mixed infections, in which one of the *ospC* types in the mixture could be determined by sequencing, thereby bringing the total number of *ospC* type determinations from this study of the Northeast and the Midwest to 839 (see Data Set S1 in the supplemental material). Girard et al. described the collection of 214 *B. burgdorferi*-infected *I. pacificus* nymphs from 78 woodland sites in Mendocino County, CA, in 2004 (5); *ospC* was amplified and sequenced from 198 (93%) of these nymphs. DNA samples were stored in single-use aliquots at −80°C until use. To confirm the results from the aforementioned collections, we also characterized *ospC* sequences and other loci for 48 infected *I. scapularis* adults from the Midwest (provided by Sarah Hamer and Jean Tsao, Michigan State University) and *B. burgdorferi* isolates VQG, WQR, WQR27, QQ, and 2665.

**Existing sequences.** Table S1 in the supplemental material gives the GenBank accession numbers for existing chromosomes, cp26, and *ospC* sequences. The naming conventions for *B. burgdorferi* sequences was carried out by Bayesian estimation as implemented by MrBayes version 3.1.2 (http://www.mrbayes.csit.fsu.edu/) (57), by maximum likelihood estimation as implemented by PhyML version 3.0 (http://www.atgc-montpellier.fr/phyml/) (58), or by phylogenetic network analysis as implemented by SplitsTree version 4.10 with the NeighborNet protocol (http://www.splitsTREE.org/) (59). The evolutionary model for protein-encoding regions was estimated with ModelTest (60). For Bayesian analysis, there were 10 generations with the first 2,000 sampled trees discarded. For maximum likelihood analysis, there were 1,000 iterations. For whole-plasmid sequences, neighbor-joining and maximum likelihood phylogenograms were generated with Phylo-win (http://pbl.univ lyon1.fr/software/); the observed differences were the distance setting, and the empirical transition-to-transversion ratio was the maximum likelihood setting. Recombination detection and analysis were carried out with the RDP3 suite (http://darwin.uvigo.es/rdp/rdp.html) (61). The SciScan method was used for assessing signals of recombination (62). The PhiTest was also used to assess the likelihood of recombination (63). The R statistical package Geneland was used for stochastic simulation and MCMC (Markov chain Monte Carlo)-based inference of population structure from genetic and geographical data (http://www2.imm.dtu.dk/~gigu/Geneland/) (64). There were 100,000 iterations and thinning by 100; the assumptions were the false-null-allele model and an uncorrelated allele frequency. The postprocess setting was 500 × 150, and the burn-in setting was 200. The goodness-of-fit tests (StatXact version 6.3; Cytel Software, Boston, MA) and hypothesis tests (Statata version 10.1; Stata Corp., College Station, TX) were 2 tailed.

**Nucleotide sequence accession numbers.** The cp26 sequence was assigned GenBank accession number GU569091. The sequence determined in the whole-genome shotgun project has been deposited in GenBank under accession number ADYM00000000. Detailed analysis of strain 3-min denaturation step at 98°C was followed by 40 cycles at 98°C for 30 s, 60°C for 30 s, and 72°C for 90 s and finally a 7-min extension at 72°C. For strain B31, the size of the product was 1,397 bp. The bbb02 gene was amplified by nested PCR with outer forward, outer reverse, inner forward, and inner reverse primers of 5′-TATATATAGCTAGTATGTTTGGTTT3′, 5′-TGAAAATATTTAATTGGAATAGGCAAT3′, 5′-ATTTGGGAATATTAGAATATT3′, and 5′-TGGAATAGAATATTCAACATT3′, respectively. The PCR conditions were the same, except the annealing temperature was 55°C and the extension was for 30 s.

**DNA sequencing.** PCR products were purified using DNA Clean & Concentrator-5 (Zymo Research, Orange, CA) kits and were sequenced directly over both strands by the dideoxy method with a CEQ 8000 DNA sequencer (Beckman-Coulter, Fullerton, CA) or an Applied Biosystems 3730xl DNA analyzer. The sequences of 110 *ospC* genes and their 5′ and 3′ flanking regions are given in Data Set S2 in the supplemental material; sequences of 56 bbb02 genes are given in data set W3 at http://spiro.nmg.ucr.edu/data/ospC.

** Genome sequencing.** DNA was extracted from strain CA8 by use of a DNeasy tissue kit (Qiagen, Valencia, CA). At Ambry Genetics (Aliso Viejo, CA), the DNA was sheared to an average size of 200 bp, the ends were filled in, and adapters were added. The ligated products were size selected by gel purification and then amplified by PCR with primers for the adapters. Library size and fragment concentration were assessed with an Agilent Bioanalyzer (Santa Clara, CA). The paired-end library yielded ~150 × 10^3 plasmids per lane and was sequenced using 39 cycles with an Illumina Genome Analyzer IIx instrument (Hayward, CA). Initial data processing was performed with the Illumina RTA program (SCS version 2.4). Base calling and sequence quality filtering scripts were executed with the Illumina pipeline software program (version 1.4). The assembly was de novo, and the depth of coverage was ≥20X.

**Phylogenetic, recombination detection, and population genetic analyses.** Sequences were aligned using Clustal X (55). DNA distances were determined with DNADIST algorithm, as implemented by Mobyte (http://mobile.pasteur.fr). Nucleotide polymorphism was assessed with DnaSP version 5.10 (56). Phylogenetic inference for coding sequences was carried out by Bayesian estimation as implemented by MrBayes version 3.1.2 (http://mrbayes.csit.fsu.edu/) (57), by maximum likelihood estimation as implemented by PhyML version 3.0 (http://www.atgc-montpellier.fr/phyml/) (58), or by phylogenetic network analysis as implemented by SplitsTree version 4.10 with the NeighborNet protocol (http://www.splitsTREE.org/) (59). The evolutionary model for protein-encoding regions was estimated with ModelTest (60). For Bayesian analysis, there were 10 generations with the first 2,000 sampled trees discarded. For maximum likelihood analysis, there were 1,000 iterations. For whole-plasmid sequences, neighbor-joining and maximum likelihood phylogenograms were generated with Phylo-win (http://pbl.univ lyon1.fr/software/); the observed differences were the distance setting, and the empirical transition-to-transversion ratio was the maximum likelihood setting. Recombination detection and analysis were carried out with the RDP3 suite (http://darwin.uvigo.es/rdp/rdp.html) (61). The SciScan method was used for assessing signals of recombination (62). The PhiTest was also used to assess the likelihood of recombination (63). The R statistical package Geneland was used for stochastic simulation and MCMC (Markov chain Monte Carlo)-based inference of population structure from genetic and geographical data (http://www2.imm.dtu.dk/~gigu/Geneland/) (64). There were 100,000 iterations and thinning by 100; the assumptions were the false-null-allele model and an uncorrelated allele frequency. The postprocess setting was 500 × 150, and the burn-in setting was 200. The goodness-of-fit tests (StatXact version 6.3; Cytel Software, Boston, MA) and hypothesis tests (Statata version 10.1; Stata Corp., College Station, TX) were 2 tailed.

**Nucleotide sequence accession numbers.** The cp26 sequence was assigned GenBank accession number GU569091. The sequence determined in the whole-genome shotgun project has been deposited in GenBank under accession number ADYM00000000. Detailed analysis of strain...
CA8’s chromosome will be presented elsewhere. Alignments not included in the supplemental material are posted as data sets at http://spiro.mmg.asm.org/data/ospC.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org.

Table S1, DOC file, 0.076 MB.
Figure S1, PDF file, 0.038 MB.
Figure S2, PDF file, 0.101 MB.
Figure S3, PDF file, 0.089 MB.
Figure S4, PDF file, 0.107 MB.
Figure S5, PDF file, 0.079 MB.
Figure S6, PDF file, 0.369 MB.
Figure S7, PDF file, 0.080 MB.
Data Set S1, TXT file, 0.055 MB.
Data Set S2, TXT file, 0.255 MB.

ACKNOWLEDGMENTS

We thank Anne Gatewood Hoen, Maria Diuk-Wasser, and Durland Fish of Yale University, Yvette Girard and Robert Lane of University of California Berkeley, Sarah Hamer and Jean Tsao of Michigan State University, and Deborah Grosenbaugh of Merial Limited for providing specimens. We are grateful to Claire Fraser-Liggett, E. F. Mongodin, Sherwood Casjens, John Dunn, Barbara Grosenbaugh of Merial Limited for providing specimens. We are grateful to Claire Fraser-Liggett, E. F. Mongodin, Sherwood Casjens, John Dunn, Ben Luu, Wei-Gang Qiu, and Steve Schutzer for providing public access to the whole-genome shotgun sequences of the B. burgdorferi strains.

This work was supported by Public Health Service grants AI-065359 from the National Institute of Allergy and Infectious Diseases and CI-00171-01 from the Centers for Disease Control and Prevention.

REFERENCES


