INFECTION AND IMMUNITY, Jan. 2006, p. 578–585 Vol. 74, No. 1
0019-9567/06/$08.00+0 doi:10.1128/IAI.74.1.578–585.2006

The *ompA* Gene in *Chlamydia trachomatis* Differs in Phylogeny and Rate of Evolution from Other Regions of the Genome

Brian W. Brunelle† and George F. Sensabaugh*

Graduate Group in Infectious Diseases and Immunity, School of Public Health, University of California, Berkeley, California 94720

Received 3 June 2005/Returned for modification 25 August 2005/Accepted 24 October 2005

Strains of *Chlamydia trachomatis* are classified into serovars based on nucleotide sequence differences in *ompA*, the gene that encodes the major outer membrane protein. Phylogenetic characterization of strains based on *ompA*, however, results in serovar groupings that are inconsistent with the distinguishing features of *C. trachomatis* pathobiology, e.g., tissue tropisms and disease presentation. We have compared nucleotide sequences at multiple sites distributed around the chlamydial genome from 18 strains representing 16 serovars; sampled regions included genes encoding housekeeping enzymes (totaling 2,073 bp), intergenic noncoding segments (1,612 bp), and a gene encoding a second outer membrane protein (*porB*; 1,023 bp), with the *ompA* sequence (1,194 bp) used for reference. These comparative analyses revealed substantial variation in nucleotide substitution patterns among the sampled regions, with average pairwise sequence differences ranging from 0.15% for the housekeeping genes to 12.1% for *ompA*. Phylogenetic characterization of the sampled genomic sequences yielded a strongly supported tree that divides the strains into groupings consistent with *C. trachomatis* biology and which has a topology quite distinct from the *ompA* tree. This phylogenetic incongruity can be accounted for by recombination of the *ompA* gene between different genomic backgrounds. We found, however, no evidence of recombination within or between any of the sampled regions around the *C. trachomatis* genome apart from *ompA*. Parallel analysis of published sequence data on four members of the *pmp* gene family are consistent with the phylogenetic analyses reported here.

Chlamydia trachomatis is a leading cause of infectious blindness and sexually transmitted disease worldwide (13). It is an obligate intracellular organism with a unique biphasic developmental cycle alternating between an extracellular metabolically inert form, the elementary body, and an intracellular metabolically active replicating form, the reticulate body (1). Due to its obligate intracellular nature, chlamydiae have proven refractory to traditional modes of genetic manipulation (43); as a consequence, progress toward understanding its biology has been slow. With full genome sequences for *C. trachomatis* now available, however, it has become possible to study aspects of chlamydial biology that were previously inaccessible, including the evolutionary basis of its pathobiology and epidemiological success.

Strains of *C. trachomatis* infecting humans are subdivided into two biovars, the trachoma biovar, consisting of strains infecting columnar epithelial tissue, and the lymphogranuloma venereum (LGV) biovar, which is made up of strains infecting primarily lymphatic tissue. Strains in each biovar are further subdivided by serological typing using monoclonal antibodies that recognize epitope differences on the surface-exposed major outer membrane protein (MOMP) (46, 51). The trachoma biovar includes serovars A through K, of which serovars A, B, Ba, and C are associated with ocular trachoma and serovars D through K are associated with urogenital infection. The lymphogranuloma venereum biovar consists of three serovars, L1, L2, and L3. Additional strain differentiation within serovars has been achieved by nucleotide sequence analysis of *ompA*, the gene encoding MOMP (45). *ompA* is one of the most polymorphic single-copy genes known in bacteria; sequence variation has been detected at over 25% of its nucleotide sites, resulting in a comparable level of amino acid sequence polymorphism (11). MOMP is the main target of host immune response in humans, and its variability is thought to be due to immune selection (4–6). A large body of *ompA* sequence data currently exists, and these data have been used extensively as the primary point of reference for delineating relationships among strains (11, 15, 28, 29, 47, 48).

It can be questioned, however, whether *ompA* truly reflects the variation among strains of *C. trachomatis*. Phylogenetic analysis of *ompA* subdivides strains into three distinct and well-supported groups, the B-complex (serovars B, Ba, D, E, L1, and L2), the C-complex (serovars A, C, H, I, Ia, J, K, and L3), and the intermediate complex (serovars F and G) (11, 15, 28, 29, 47, 48). As noted over a decade ago by Fitch et al. (11) and subsequently by Stothard and others (47), these divisions are not congruent with groupings based on the tissue tropisms and pathobiological profiles of *C. trachomatis*. More recently, sequence characterization of genes encoding other putative surface-exposed proteins in *C. trachomatis* has yielded strain groupings that are generally consistent with strain pathobiology but which are discordant with the *ompA* phylogeny (15, 48). The apparent discordance between strain relationships based on *ompA* phylogeny and those based on other features of *C. trachomatis* biology has not been explained. One possible explanation is recombination between strains. There is evidence that recombination has occurred within the *ompA* gene (3, 17, 29); recombination involving other genes in the *C. trachomatis* genome might yield gene combinations with different phenotypic profiles. This raises the more general question of the

* Corresponding author. Mailing address: Division of Infectious Diseases, School of Public Health, 140 Earl Warren Hall, University of California, Berkeley, CA 94720. Phone: (510) 642-1271. Fax: (510) 642-6350. E-mail: sensaba@berkeley.edu.

† Present address: Virus and Prion Diseases of Livestock Research Unit, National Animal Disease Center, USDA-ARS, Ames, IA 50010.

extent to which recombination has played a role in shaping the chlamydial genomes and in generating diversity among strains.

To gain a better understanding of the genomic relationships and sequence diversity among strains of *C. trachomatis*, we have performed comparative sequence analysis on representative genome segments from each of 16 serovars. Our genome survey includes loci from six genes encoding housekeeping enzymes, five noncoding regions, and a gene for an outer membrane protein in addition to *ompA*. We augment these analyses with parallel characterizations of recently published sequence data for four genes in the polymorphic membrane protein (*pmp*) family (15, 48). Overall, these analyses yield a picture of the tempo and mode of *C. trachomatis* evolution that is strikingly different from that based on the *ompA* gene alone.

MATERIALS AND METHODS

Source of isolates. Fifteen strains of *C. trachomatis* grown in HeLa cells were obtained through J. Schachter, University of California, San Francisco (A/Har-1, B/Tunis 864, Ba/Apache 2, C/TW-3, D/IC-Cal 8, E/Bour, F/IC-Cal 3, G/392, H/580, Ia/870, J/UW-36, K/UW-31, L1/440, L2/434, and L3/404). Three additional strains were provided by R. S. Stephens, University of California, Berkeley (B/TW-5, D/UW-3, and I/UW-12).

DNA isolation from culture. The DNA was isolated using a standard protocol employing proteinase K digestion, phenol-chloroform-isoamyl extraction, and ethanol precipitation (10).

Loci of interest. Primers were designed to amplify portions of six genes encoding housekeeping enzymes, five noncoding regions along with portions of flanking coding sequence, and the complete *porB* and *ompA* genes. Each PCR product was prepared for sequencing using the exonuclease I and shrimp alkaline phosphatase procedure (Amersham Pharmacia Biotech, Piscataway, NJ) (21). The PCR products were sequenced using ABI Big-Dye Terminator chemistry and an ABI 377 sequencer (Applied Biosystems, Foster City, CA). To ensure accuracy, each locus was amplified twice and sequenced in both directions (fourfold coverage). Any discrepancies in the consensus sequence derived from each 4× sequence were resolved through visual inspection of the electropherogram output. A listing of the sequenced regions and the GenBank accession numbers are shown in Table 1. Orthologous regions from the mouse pneumonitis (MoPn) biovar of *C. trachomatis* strain Nigg were retrieved from the genome sequence (AE002160) (38). The nucleotide sequences for *pmpC* (AF519747 to AF519765), *pmpE* (AY184140 to AY184154), *pmpH* (AY184155 to AY184169), and *pmpI* (AY184170 to AY184184) were retrieved from GenBank.

Alignment and analysis. The freeware sequence tool BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) was used to construct sequence alignments. This was done by translating all coding sequences to their corresponding amino acid sequences, aligning the amino acid sequences using ClustalW (50), and then converting back to the nucleotide sequences. This approach results in a nucleotide sequence that reflects protein functional properties, such as homology and sequence/codon gaps. The noncoding loci were aligned based on their nucleotide sequence. The software packages DnaSP 4.0 (39) and MEGA3 (25) were used to analyze the data and to construct phylogenetic trees. The extent of sequence variation was measured as the p-distance, the proportion of nucleotide sites that differ between pairs of sequences. The average p-distance for a set of sequences can then be used to assess the extent of DNA polymorphism among strains at a particular locus (32). For convenience of notation, the average p-distance is denoted by π. The average nucleotide variation at synonymous (π_s) and nonsynonymous (π_a) sites was calculated using the Nei-Gojobori method (31).

Recombination. Aligned sequences were tested for recombination using the software package RDP (Recombination Detection Program), version 2. This package implements a set of six published methods found to be sensitive for the identification of recombination and to yield the fewest false-positive findings (27, 36, 37). These six methods are RDP (26), GENECONV (35), Bootscan (40), MaxChi (42), Chimaera (37), and SiScan (14). Each method employs a different test for detecting potentially recombinant regions within aligned sequences. The null hypothesis is clonality, i.e., that the pattern of sequence variation among the aligned sequences shows no indication of recombination. Recombination was deemed to occur in a locus if clonality was rejected by three or more tests at a significance level of $P < 0.001$.

TABLE 1. Genome regions sequenced[a]

Group and locus	Gene no.	Size (bp)	GenBank accession nos.
Housekeeping genes			
ssb	CT044	54	DQ064099–DQ064116
pepA	CT045	84	DQ064117–DQ064134
araD	CT121	337	DQ064207–DQ064224
adk	CT128	313	DQ064189–DQ064206
hemZ	CT485	297	DQ064243–DQ064260
gap	CT505	328	DQ064225–DQ064242
thdF	CT698	39	DQ064336–DQ064353
uvrC	CT791	289	DQ064318–DQ064335
mrsA	CT815	334	DQ064261–DQ064278
Noncoding regions			
ssb-pepA	CT044-045	342	DQ064081–DQ064098
rpoB-r17	CT315-316	342	DQ064153–DQ064170
rs4-yceA	CT626-627	266	DQ064135–DQ064152
thdF-psdD	CT698-699	237	DQ064336–DQ064353
glyQ-pgsA	CT796-797	425	DQ064171–DQ064188
porB	CT713	1,023	DQ064300–DQ064317
ompA	CT681	1,194	DQ064279–DQ064296

[a] Loci are specified by the name of the gene containing the sequenced segment and its position referenced on the gene number in the *C. trachomatis* strain D/UW-3 genome (44). The positions of the noncoding regions are indexed on the flanking genes. The length of each sequenced segment is indicated (bp), as are the GenBank accession numbers for the 18 strain sequences from each region.

RESULTS

Patterns of sequence variation among serovars. Sequence variation in each of the surveyed regions of the genome is summarized in Table 2. Genes encoding housekeeping enzymes are believed to be under stabilizing selection, and most variation in these genes is posited to be selectively neutral. To assess variation in genes encoding housekeeping enzymes, sequences from nine different loci were sampled. Six loci were targeted to represent different positions around the *C. trachomatis* genome and to include both leading and lagging strand sequences; short segments of three additional loci were sampled in the course of sequencing noncoding regions. Overall, the nine gene regions yielded 2,073 bp of total sequence (Table 2). Nucleotide sequence polymorphisms were observed at only 12 sites (0.6%); all were single nucleotide substitutions, and 8 resulted in an amino acid replacement. Eight of the 12 substitutions were represented in two or more sequences (parsimony informative); of the 8 replacement substitutions, 6 were parsimony informative. Overall, the average sequence difference between pairs of sequences (π_{total}) was 0.0015; the corresponding values for synonymous sites (π_s) and nonsynonymous sites (π_a) were 0.0022 and 0.0013, respectively. The π_s/π_a ratio was 1.65, a low value compared to other microbes, where ratios are typically in the 2 to 20 range (33).

Noncoding sequence regions were sampled to provide a counterpoint to the coding sequence data; nucleotide substitutions occurring outside of regulatory sites would be expected to be nearly neutral and therefore analogous to synonymous site variation in the coding regions (34). The noncoding regions selected for sequencing were between defined genes to avoid possible ambiguities associated with undefined hypothetical open reading frames. Overall, 47 variable sites were detected in 1,612 bp of sequence (Table 2); of these, 41 were single nucleotide substitutions, 5 were 1-bp insertion/deletions

TABLE 2. Summary of nucleotide sequence variation in sampled regions of the *C. trachomatis* genome[a]

Group and locus	Size (bp)	Δnt	%nt	Δrep	%rep	pars	π	π_s	π_a	π_s/π_a
Housekeeping genes										
ssb	54	0	0.0	0	0.0	0	0.000	0.000	0.000	
pepA	84	2	2.4	1	3.6	2	0.010	0.027	0.005	5.94
araD	337	3	0.9	3	2.7	2	0.002	0.000	0.002	0.00
adk	313	0	0.0	0	0.0	0	0.000	0.000	0.000	
hemZ	297	2	0.7	2	2.0	1	0.001	0.000	0.002	0.00
gap	328	2	0.6	1	0.9	2	0.002	0.005	0.001	3.89
thdF	39	0	0.0	0	0.0	0	0.000	0.000	0.000	
uvrC	289	0	0.0	0	0.0		0.000	0.000	0.000	
mrsA	334	3	0.9	1	0.9	1	0.002	0.003	0.003	1.00
Total	2,073	12	0.6	8	1.2	8	0.002	0.002	0.001	1.65
Intergenic noncoding										
ssb-pepA	342	17	4.9			15	0.014	0.014		
rpoB-r17	342	10	2.9			6	0.009	0.009		
rs4-yceA	266	4	1.5			3	0.004	0.004		
thdF-psdD	237	6	2.5			5	0.007	0.007		
glyQ-pgsA	425	4	0.9			4	0.003	0.003		
Total	1,612	41	2.5			33	0.007	0.007		
porB	1,023	11	1.1	8[c]	2.3	9	0.003	0.003	0.003	0.86
ompA	1,194	334	27.7	99	24.9	301	0.121	0.289	0.069	4.20
pmp genes[b]										
pmpH	2,988	263	8.8	76	7.6	258	0.037	0.101	0.015	6.55
pmpE	2,826	172	6.1	63	6.6	160	0.026	0.061	0.015	3.91
pmpI	2,538	46	1.8	21	2.5	36	0.006	0.012	0.003	3.69
pmpC	5,355	86	1.6	64	3.6	72	0.005	0.005	0.005	1.05

[a] Δnt, number of polymorphic nucleotide sites; %nt, percent nucleotide sites polymorphic; Δrep, number of polymorphic sites resulting in an amino acid replacement; %rep, percent sites with replacement; pars, parsimony informative sites; π, π_s, and π_a, average p-distances at all sites, synonymous sites, and nonsynonymous sites, respectively.

[b] Analyses based on published nucleotide sequence data (15, 48); these sequences represent the same serovars as used in this study, but some sequences come from different strains.

[c] This includes one substitution resulting in a premature stop codon.

(indels), and 1 was an 8-bp indel. All of the indels and 33 of the 41 nucleotide substitutions are parsimony informative. The π_{total} for this region was 0.0073, about three times higher than that of the housekeeping gene loci.

The *porB* gene encodes a surface-exposed protein with porin activity (24); MOMP is also a porin, and the *porB* sequences thus provide a contrast to the *ompA* sequences (Table 2). *porB* is much less variable than *ompA*, exhibiting only 11 variable nucleotide sites over the 1,023-bp gene, 8 of which resulted in amino acid replacement. The π_{total} (0.0032) was in the same range as the corresponding values for the housekeeping gene and noncoding loci. The level of synonymous substitution (π_s = 0.0029) was nearly equal to that observed for the housekeeping gene regions, but the nonsynonymous substitution level was lower (π_a = 0.0034). The π_s/π_a ratio of 0.86 indicates that nucleotide substitutions in *porB* favor amino acid change. Interestingly, strain D/IC-CAL8 contained a substitution at base 977 resulting in a premature stop codon and a predicted protein with a 15-amino-acid truncation.

The full *ompA* gene in each of the surveyed stains was sequenced to verify the serovar and strain designations (Table 2); the data were consistent with sequence data from previous studies. Among the 18 serovar strains sampled, 331 of 1,194 nucleotide sites were polymorphic (27.7%), a level of variability considerably higher than any of the other sampled regions. The π_{total} for *ompA* was 0.1215, a value 15 to 60 times higher than that seen in the housekeeping gene, noncoding, and *porB* regions. More remarkable was the π_s of 0.2893, indicating a synonymous substitution rate 40 to 145 times greater than that detected in the other regions.

Table 2 summarizes published sequence data for four *pmp* genes, *pmpC*, *pmpE*, *pmpH*, and *pmpI* (15, 48); though the strains used in those studies did not fully overlap the strains used in this study, representative sequences are provided for each serovar. The *pmp*'s are part of the nine-member paralogous gene family that are thought to encode outer membrane proteins (44), though only three, PmpE, PmpG, and PmpH, have been demonstrated to be surface exposed (30, 49). Each of the four *pmp* gene sequences exhibited more differences than the housekeeping gene, noncoding, and *porB* regions, but none was as variable as *ompA*. Interestingly, the two that encode proteins known to be surface exposed, *pmpE* and *pmpH*, exhibited much higher nucleotide variability (π_{total} = 0.0262 to 0.0367) than *pmpC* and *pmpI* (π_{total} = 0.0051 to 0.0055). The range for synonymous substitutions varied considerably (π_s = 0.0053 to 0.1009) but, again, was well below that for *ompA*.

Sequence divergence of human strains from the mouse strain of *C. trachomatis*. To assess whether differences in mutation rates might account for the observed variability in nucleotide polymorphism patterns shown in Table 2, the nucleo-

TABLE 3. Sequence divergence between human and mouse strains of *C. trachomatis*[a]

Sequence region	Size (bp)	K_{total}	K_s	K_a	K_s/K_a	π_s/π_a
Housekeeping (9)	2,073	0.165	0.528	0.050	10.56	1.65
Noncoding (5)	1,612	0.211	0.211			
ompA	1,194	0.207	0.545	0.104	5.24	4.20
porB	1,023	0.178	0.585	0.051	11.44	0.85
pmpE	2,826	0.262	0.580	0.161	3.60	3.91
pmpH	2,988	0.234	0.550	0.130	4.23	6.55
pmpI	2,538	0.231	0.597	0.115	5.19	3.69

[a] K_{total}, K_s, and K_a, human-mouse strain divergence at all sites, synonymous sites, and nonsynonymous sites, respectively; π_s/π_a, ratio of p-distances at synonymous and nonsynonymous sites within human strains (from Table 2).

tide divergence (K) at each gene region was measured relative to the orthologous gene region of the outgroup MoPn strain; the extent to which the divergence values parallel the corresponding variation in π_{total}, π_s, and π_a provides an indication of locus-specific differences in mutation rates. As shown in Table 3, the nucleotide divergence over all sites (K_{total}) varied less than twofold among the sampled gene regions, with the *pmp* genes exhibiting the most divergence. The divergence values for synonymous sites in coding regions were even more similar in magnitude (K_s = 0.528 to 0.597). Sequence divergence at nonsynonymous sites exhibited a somewhat greater range, with the housekeeping gene and *porB* sequences registering divergence values (K_a) of about 0.05, whereas the values for the *ompA* and *pmp* genes ranged two- to threefold larger.

Comparison of the ratios of synonymous-to-replacement substitutions within (π_s/π_a) and between (K_s/K_a) populations provides an additional perspective. The within and between ratios for *ompA* and the *pmp* genes are similar in magnitude (3.6 to 6.6), whereas the ratios for *porB* and the housekeeping genes are distinctly different (1.7 versus 10.6 and 0.7 versus 11.4, respectively). This variation is likely due to differences in selection pressures operating at the disparate loci.

Phylogenetic reconstruction. Gene trees were constructed for *ompA*, *porB*, and each individual housekeeping gene and noncoding region (data not shown) using the MoPn strain of *C. trachomatis* as the outgroup. The gene tree for *ompA* (Fig. 1) is consistent with *ompA* trees reported previously (11, 15, 28, 29, 47, 48). The *ompA* sequence alignments include a large number of parsimony informative sites and yield a gene tree

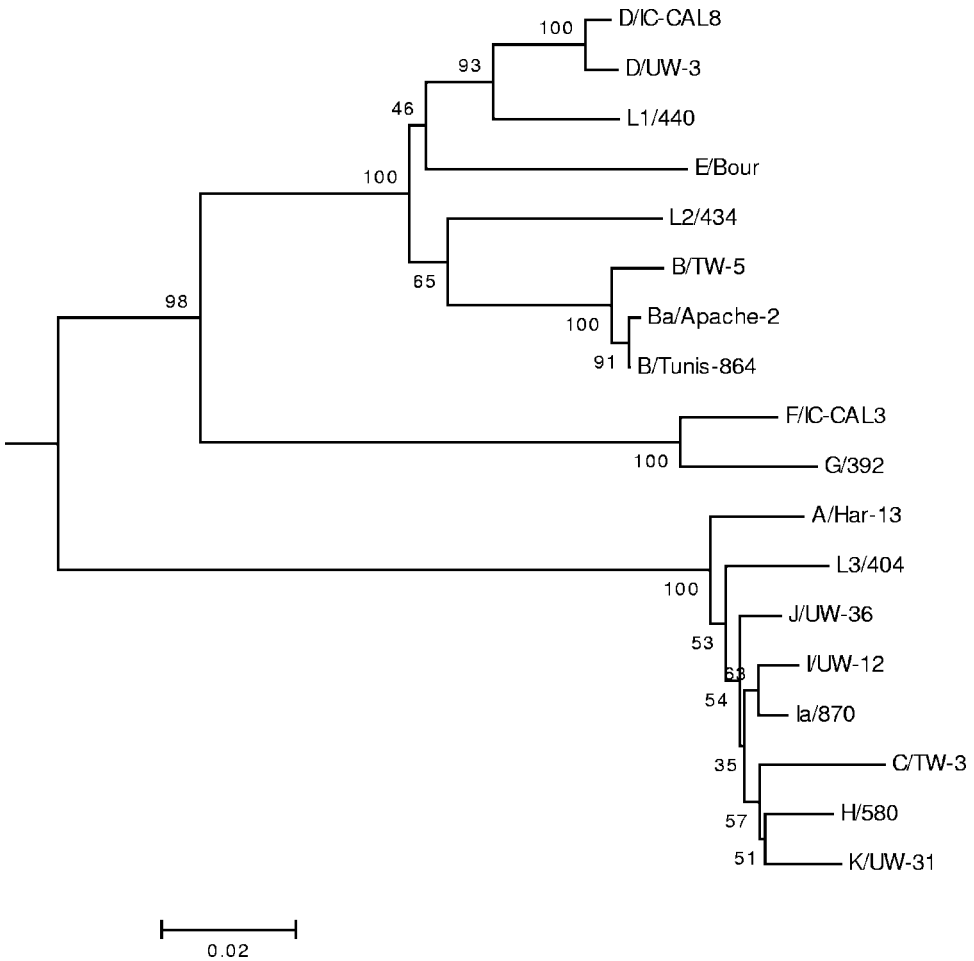

FIG. 1. Gene tree for *ompA* nucleotide sequences from 18 different human-specific strains of *C. trachomatis*. The tree was based on uncorrected p-distances and was generated using the neighbor-joining method with the MoPn strain *ompA* sequence as the root. Branch lengths are proportional to the number of substitutions per nucleotide site. The numbers at the nodes are percent bootstrap values for 1,000 replications.

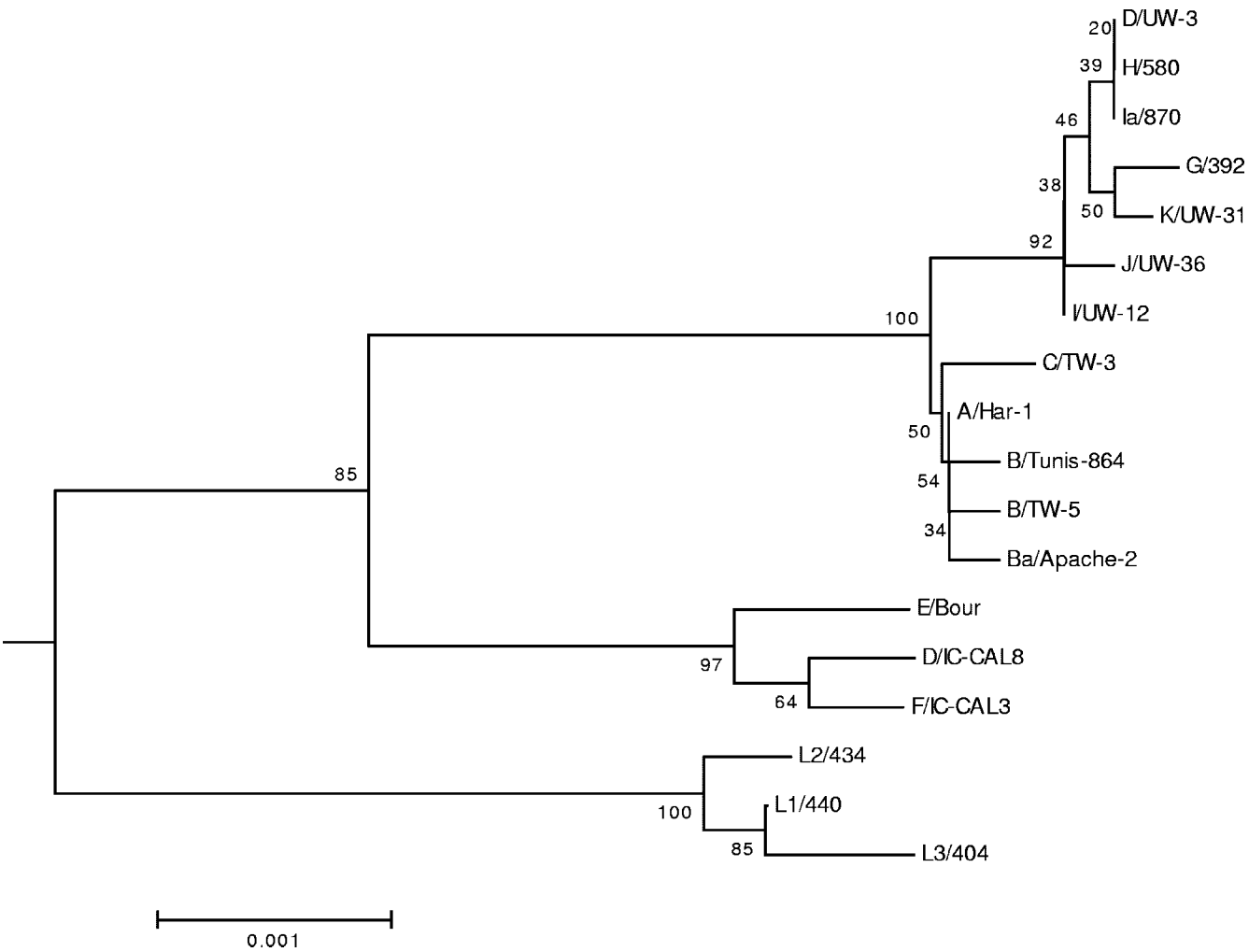

FIG. 2. Phylogenetic relationships of 18 different human-specific strains of *C. trachomatis* based on concatenated nucleotide sequences from segments of nine genes encoding housekeeping enzymes, six intergenic noncoding segments, and the *porB* gene. The tree was constructed as described for Fig. 1.

with strong bootstrap support. As previously noted, this tree subdivides the serovars into three distinct groups, the B-complex (serovars B, Ba, D, E, L1, and L2), the C-complex (serovars A, C, H, I, Ia, J, K, and L3), and the intermediate complex (serovars F and G).

The gene trees for each of the *porB*, housekeeping gene, and noncoding regions had few branch points, keeping with the relatively low level of sequence variation within each region. However, none showed a branching pattern consistent with the *ompA* tree. Since the relationships indicated by the individual trees for the housekeeping, noncoding, and *porB* gene regions were not in conflict, the sequence data for these regions were concatenated and a single phylogenetic tree was constructed from the concatenated data (Fig. 2). Concatenation of sequence data allows all informative sites from the different loci to be combined to create a more comprehensive data set, permitting more robust phylogenetic inferences to be made (12). To test whether the phylogeny resulting from the concatenated sequence data was biased by a single locus, a subset of trees was built using the concatenated data with each region omitted; this resulted in no perturbation of the tree topology.

The phylogenetic tree constructed from the concatenated sequence data distinguished four distinct clades. This clustering was highly supported by bootstrap analysis and by a large fraction of the parsimony informative sites; the tree was also consistent with the distribution of five of the six indels present in the intergenic regions. The most basal division in the concatenated sequence tree was between the three strains that belong to the LGV biovar (L1/440, L2/434, and L3/404) and the 15 strains that belong to the trachoma biovar. The trachoma biovar in turn is separated into three distinct clades: one containing the strains associated with ocular disease (A/Har-1, B/Tunis 864, B/TW-5, Ba/Apache 2, and C/TW-3), and two containing strains found in urogenital infections, the strains D/IC-Cal8, E/Bour, and F/IC-Cal 3 on one branch and the remainder of the urogenital strains (D/UW-3, G/392, H/580, I/UW-12, Ia/870, J/UW-36, and K/UW-31) on the other. The topology of the trachoma branches suggests an early division of

urogenital strains into two groups with one subsequently undergoing a split, giving rise to the ocular serovars. The phylogenetic tree indicated by the concatenated sequence data is supported by the gene trees for the polymorphic membrane proteins *pmpC*, *pmpH*, and *pmpI*; the gene tree for *pmpE* differs only in moving two urogenital strains from one urogenital branch to the other (15, 48). Overall, the general consistency of the serovar divisions exhibited in these trees supports an evolutionary basis for the biological differences between the *C. trachomatis* strains.

It is to be noted that the two D serovar strains were nearly identical in *ompA* sequence but were differentiated in the concatenated sequence tree (Fig. 2). In the latter tree, the D/IC-Cal8 strain is in the clade with the E/Bour and F/IC-CAL3 strains, whereas the D/UW-3 strain is in the clade containing the other urogenital strains. The sequence analysis differentiating these two strains was repeated to verify that the assessments were not a consequence of sample mix-up or sequencing error. It is notable that these two strains would be considered clinically identical based on serological and *ompA* classification but are clearly different elsewhere in their genomes.

Recombination. Each of the sampled gene regions was tested for evidence of recombination using the six test algorithms included in the RDP. No trace of recombination was detected in any of the housekeeping gene, noncoding, or *porB* gene regions. In contrast, the *ompA* sequences were found to deviate from clonality by all six recombination tests ($P < 0.001$); this finding is consistent with previous reports of recombination in *ompA* (3, 17, 29). To test for recombination between gene regions, the non-*ompA* sequence data were concatenated and analyzed by RDP. Again, no indication of recombination was detected. Analysis of the published *pmp* gene sequence data provided evidence for recombination within *pmpE* and *pmpH*, the two most diverse members of the family, but not for *pmpC* or *pmpI*. In contrast to *ompA*, however, the *pmpE* and *pmpH* gene trees are similar in branching topology to the concatenated genome sequence tree.

DISCUSSION

The comparative sequence analyses described in this study reveal three striking features of genomic relationships among the serovars of *C. trachomatis*: (i) the rate of nucleotide substitution for different regions of the genome varies by as much as 100-fold, (ii) the *ompA* gene has a phylogenetic history distinct from the remainder of the genome, and (iii) there is no evidence of recombination occurring in the *C. trachomatis* genome apart from genes encoding surface-exposed proteins. These observations raise important questions about the course of *C. trachomatis* evolution and have implications for the clinical classification of chlamydial infections.

Variation in nucleotide substitution rate. This study demonstrates a spectrum of nucleotide substitution patterns among different loci in *C. trachomatis*: the rates of substitution at synonymous sites vary over 100-fold, and the rates of replacement substitution vary over 50-fold. Although substitution rate differences approaching this magnitude have been noted in comparisons of homologous sequences between species (18, 33), to the best of our knowledge the extent of rate variation detected within the human *C. trachomatis* serovars is unprecedented. There is no simple apparent explanation for the variation in substitution rates, i.e., there are no striking differences in GC content, nor is there evidence of significant codon usage bias among the coding loci. Although only small segments of the genome were sampled, there is no indication that the rate differences can be attributed to chromosomal position or to location of the coding sequence on a leading or lagging strand. There is experimental evidence suggesting that in some microbial species spontaneous mutation rates increase in highly transcribed genes (18, 52) and in genomes under environmental stress (2); the former mechanism might account for the high rate of substitution in *ompA* given that it is one of the most highly expressed genes in the *C. trachomatis* genome. Neither of these mechanisms can be excluded based on the data in hand; whether either is actually in play in *C. trachomatis* is an open question.

Variation in the rates of nonsynonymous substitution can be linked to predicted variation in selection pressures. The very high level of replacement substitution in the *ompA* gene can be attributed to immune selection pressure on the protein it encodes, MOMP, as it is known that MOMP elicits a strong immune response (4, 20). The high level of replacement substitution in *pmpE* and *pmpH* may also be associated with immune selection pressure, since both encode surface-exposed proteins (49). The surface exposure of the less variable PmpC and PmpI proteins is not known, and nothing is known of their immunogenic potential. The PorB protein is surface exposed but appears not to be a natural target for the immune response (22); this may account for its relatively low level of replacement substitution, a level only marginally higher than that seen for the genes encoding the housekeeping enzyme. The pattern of nucleotide substitution in the noncoding regions suggests that they may also be under selection pressure, presumably to conserve regulatory element binding sites. Two lines of evidence support this idea, both derived from comparison of the noncoding sequences from the human and mouse strains. First, the sequence divergence in the noncoding regions is substantially lower than what would be expected if the substitutions in the noncoding regions were neutral, assuming that neutral substitution rates are reflected in the divergence values for synonymous sites in the coding regions (noncoding $K = 0.211$, versus $K_s = 0.528$ to 0.597). This rationale has been applied to account for differences between K and K_s in other bacteria (19). Second, the noncoding region sequence alignments contain multiple runs of invariant sequence greater than 15 bp in length; such runs would not be expected, given a random mutation model.

The substantial differences seen with synonymous site substitution rates at the different loci in *C. trachomatis* contrast with the relative uniformity of divergence rates at synonymous sites between the human *C. trachomatis* strains and the mouse strain MoPn. This pattern might be accounted for under an evolutionary scenario in which it is assumed that synonymous substitutions are essentially neutral and serve as a molecular clock for events in *C. trachomatis* evolution (23). Under this scenario, the uniformity of divergence rates indicates a common point in time for the split between the mouse and the human strain genomes. Along the human strain lineage, however, different genes diverged at different points in time, with *ompA* at the deepest remove in time, then the *pmp* genes and, most recently, the housekeeping and *porB* genes. A possible

implication of this is that the divergence of the *ompA* gene and possibly the *pmp* genes began before the division of the genomic lineages leading to the contemporary serovars.

The alternative to this scenario is that the variation in substitution rates among the genes within the human *C. trachomatis* strains developed after the division of the contemporary serovar lineages. This scenario would entail extraordinary locus-specific mechanisms involving acceleration of substitution rates at some gene loci (notably *ompA*), intense selection constraints extending to the codon level within the genes exhibiting very low substitution rates, or some combination of these. Obviously, explication of such mechanisms, should they exist, would be of considerable interest for advancing understanding of the biology of *C. trachomatis*.

Phylogenetic relationships among serovars within *C. trachomatis*. The comparative sequence analyses presented here provide evidence that the *C. trachomatis* genome has evolved along at least two distinct phylogenetic trajectories. The phylogenetic branching pattern represented by the *ompA* gene represents one of the trajectories. The *ompA* tree distinguishes three very strongly supported groups (Fig. 1); *ompA* variants within a group differ in sequence by 0 to 8%, whereas variants in different groups differ by 12 to 20%. The other phylogenetic trajectory is represented, at least to the extent indicated by the data described here, by the remainder of the *C. trachomatis* genome. Tree building based on the multiple sites sampled from around the genome as well as the from four *pmp* genes yields a coherent and internally consistent tree with a different topology than the *ompA* tree (Fig. 2). Significantly, the four main branches of this genome tree coincide with the tissue tropisms and patterns of disease presentation associated with *C. trachomatis* infection in the human host. This differentiation of serovars is supported by recent microarray-based genome surveys (8). It is worth speculating that the divergent phylogenetic trajectories observed here may reflect different selection pressures associated with the biphasic life cycle of *C. trachomatis*: the general genomic trajectory tied to the organism's functioning in the intracellular milieu and the *ompA* trajectory determined by immune and possibly other host niche selection pressures during its extracellular phase.

Recombination. The discordant phylogeny of *ompA* compared to other genes in the *C. trachomatis* genome is prima facie evidence of recombination involving *ompA* (36). Mosaic gene structures of *ompA* have been identified in previous studies and attributed to recombination involving segments within the *ompA* gene (3, 17, 29). This study provides two additional lines of evidence for recombination involving *ompA*. First, all six computational approaches used by the Recombination Detection Program indicated a high probability of recombination events among the aligned *ompA* sequences. Second, we detected two strains that unambiguously fall into different genome groups based on sequence differences in the sampled gene set but which would be classified in serovar D based on *ompA* sequence typing; the most parsimonious explanation is that an entire *ompA* gene has moved from one background genome type to another. Thus, it appears possible the entire *ompA* gene as well as *ompA* gene segments can undergo recombinational transfer between strains.

Recombination provides a ready explanation for the dissimilarity of *ompA* genes among otherwise biologically related strains, for example, the LGV strains. Recombination can also account for two recently reported observations of incongruent associations between genomic markers and serovar types. In the first case, clinical isolates of ocular origin are differentiated from urogenital strains by carriage in the former of defective *trp* operon genes; urogenital strains have functional *trp* operons. Although isolates in serovar B are typically of ocular origin and have defective *trp* operons, some urogenital tract isolates carrying intact *trp* operons have been identified as belonging to serovar B (7); it is possible that these isolates are the result of a transfer of an *ompA* gene from an ocular strain into a urogenital strain genomic background. The second apparent anomaly is the detection in some clinical isolates of serovar incongruent *pmpC* sequences (15); though this was attributed to recombination involving the *pmpC* gene, given the findings of this study it is more likely that the recombinational transfer involves the *ompA* gene.

The likelihood that recombination involving *ompA* has occurred in *C. trachomatis* prompts the question of whether it occurs elsewhere in the genome as well. The RDP provides statistical evidence of recombination in the *pmpE* and *pmpH* genes but finds no significant deviation from clonality across the remainder of the genome. To the extent recombination may occur at the two *pmp* loci, it appears not to have resulted in substantial departures from the general genome phylogeny as defined by the concatenated sequence data. There are a sufficient number of polymorphic sites in the segments of genome sampled here to have allowed detection of a recombination event had one occurred. Thus, the available evidence argues against genetic fluidity in the *C. trachomatis* genome. Rather, it suggests that recombination is probably uncommon and is localized to a few sites, most prominently at *ompA*. Although diverse recombination mechanisms (including gene conversion) have been invoked to account for localized genetic variation in other organisms (9, 16, 41), the details of recombination in *C. trachomatis* remain to be identified and are questions for future study.

Finally, a clinical and epidemiological consequence of *ompA* recombination is that serovar classification based on *ompA* sequence variation does not necessarily reflect the genetic content of the remainder of the *C. trachomatis* genome. To the extent that the content of the genome determines the pathobiology and the epidemiological success of the organism, strain typing based on *ompA* alone may paint an incomplete picture. The same consideration applies for studies looking at possible associations between chlamydia infection and other maladies, such as cervical cancer. A classification system for *C. trachomatis* that incorporates more extensive genomic characterization would be beneficial.

ACKNOWLEDGMENTS

We thank Julius Schachter for the gift of isolates, Malcolm McGinnis for use of his sequencing facilities, and Richard S. Stephens for the gift of isolates and for critical input to the development of this project.

This work was supported in part by a Faculty Bridging Grant to G.F.S.

REFERENCES

1. **Bedson, S. P., and J. O. W. Bland.** 1932. A morphological study of psitticosis virus, with the description of a developmental cycle. Br. J. Exp. Pathol. **13:**461–466.

2. **Bjedov, I., O. Tenaillon, B. Gerard, V. Souza, E. Denamur, M. Radman, F. Taddei, and I. Matic.** 2003. Stress-induced mutagenesis in bacteria. Science **300:**1404–1409.
3. **Brunham, R., C. Yang, I. Maclean, J. Kimani, G. Maitha, and F. Plummer.** 1994. Chlamydia trachomatis from individuals in a sexually transmitted disease core group exhibit frequent sequence variation in the major outer membrane protein (omp1) gene. J. Clin. Investig. **94:**458–463.
4. **Brunham, R. C., F. A. Plummer, and R. S. Stephens.** 1993. Bacterial antigenic variation, host immune response, and pathogen-host coevolution. Infect. Immun. **61:**2273–2276.
5. **Caldwell, H. D., and J. Schachter.** 1982. Antigenic analysis of the major outer membrane protein of *Chlamydia* spp. Infect. Immun. **35:**1024–1031.
6. **Caldwell, H. D., and J. Schachter.** 1983. Immunoassay for detecting *Chlamydia trachomatis* major outer membrane protein. J. Clin. Microbiol. **18:** 539–545.
7. **Caldwell, H. D., H. Wood, D. Crane, R. Bailey, R. B. Jones, D. Mabey, I. Maclean, Z. Mohammed, R. Peeling, C. Roshick, J. Schachter, A. W. Solomon, W. E. Stamm, R. J. Suchland, L. Taylor, S. K. West, T. C. Quinn, R. J. Belland, and G. McClarty.** 2003. Polymorphisms in *Chlamydia trachomatis* tryptophan synthase genes differentiate between genital and ocular isolates. J. Clin. Investig. **111:**1757–1769.
8. **Carlson, J. H., S. Hughes, D. Hogan, G. Cieplak, D. E. Sturdevant, G. McClarty, H. D. Caldwell, and R. J. Belland.** 2004. Polymorphisms in the *Chlamydia trachomatis* cytotoxin locus associated with ocular and genital isolates. Infect. Immun. **72:**7063–7072.
9. **Centurion-Lara, A., R. E. LaFond, K. Hevner, C. Godornes, B. J. Molini, W. C. Van Voorhis, and S. A. Lukehart.** 2004. Gene conversion: a mechanism for generation of heterogeneity in the tprK gene of Treponema pallidum during infection. Mol. Microbiol. **52:**1579–1596.
10. **Davis, L. G., W. M. Kuehl, and J. F. Battey.** 1994. Basic methods in molecular biology, 2nd ed., p. 16–21. Appleton and Lange, Norwalk, Conn.
11. **Fitch, W. M., E. M. Peterson, and L. M. de la Maza.** 1993. Phylogenetic analysis of the outer-membrane-protein genes of Chlamydiae, and its implication for vaccine development. Mol. Biol. Evol. **10:**892–913.
12. **Gadagkar, S. R., M. S. Rosenberg, and S. Kumar.** 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J. Exp. Zool. B **304:**64–74.
13. **Gerbase, A. C., J. T. Rowley, D. H. Heymann, S. F. Berkley, and P. Piot.** 1998. Global prevalence and incidence estimates of selected curable STDs. Sex. Transm. Infect. **74**(Suppl. 1)**:**S12–S16.
14. **Gibbs, M. J., J. S. Armstrong, and A. J. Gibbs.** 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics **16:**573–582.
15. **Gomes, J. P., W. J. Bruno, M. J. Borrego, and D. Dean.** 2004. Recombination in the genome of *Chlamydia trachomatis* involving the polymorphic membrane protein C gene relative to *ompA* and evidence for horizontal gene transfer. J. Bacteriol. **186:**4295–4306.
16. **Haake, D. A., M. A. Suchard, M. M. Kelley, M. Dundoo, D. P. Alt, and R. L. Zuerner.** 2004. Molecular evolution and mosaicism of leptospiral outer membrane proteins involves horizontal DNA transfer. J. Bacteriol. **186:** 2818–2828.
17. **Hayes, L. J., P. Yearsley, J. D. Treharne, R. A. Ballard, G. H. Fehler, and M. E. Ward.** 1994. Evidence for naturally occurring recombination in the gene encoding the major outer membrane protein of lymphogranuloma venereum isolates of *Chlamydia trachomatis*. Infect. Immun. **62:**5659–5663.
18. **Hudson, R. E., U. Bergthorsson, and H. Ochman.** 2003. Transcription increases multiple spontaneous point mutations in Salmonella enterica. Nucleic Acids Res. **31:**4517–4522.
19. **Hughes, A. L., and R. Friedman.** 2004. Patterns of sequence divergence in 5′ intergenic spacers and linked coding regions in 10 species of pathogenic bacteria reveal distinct recombinational histories. Genetics **168:**1795–1803.
20. **Hughes, A. L., T. Ota, and M. Nei.** 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol. Biol. Evol. **7:**515–524.
21. **Innis, M., D. Gelfand, and J. Sninsky.** 1999. PCR applications: protocols for functional genomics. Academic Press, San Diego, Calif.
22. **Kawa, D. E., J. Schachter, and R. S. Stephens.** 2004. Immune response to the Chlamydia trachomatis outer membrane protein PorB. Vaccine **22:**4282–4286.
23. **Kimura, M.** 1968. Evolutionary rate at the molecular level. Nature **217:**624–626.
24. **Kubo, A., and R. S. Stephens.** 2000. Characterization and functional analysis of PorB, a Chlamydia porin and neutralizing target. Mol. Microbiol. **38:**772–780.
25. **Kumar, S., K. Tamura, and M. Nei.** 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. **5:**150–163.
26. **Martin, D., and E. Rybicki.** 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics **16:**562–563.
27. **Martin, D. P., C. Williamson, and D. Posada.** 2005. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics **21:**260–262.
28. **Millman, K., C. M. Black, R. E. Johnson, W. E. Stamm, R. B. Jones, E. W. Hook, D. H. Martin, G. Bolan, S. Tavare, and D. Dean.** 2004. Population-based genetic and evolutionary analysis of *Chlamydia trachomatis* urogenital strain variation in the United States. J. Bacteriol. **186:**2457–2465.
29. **Millman, K. L., S. Tavare, and D. Dean.** 2001. Recombination in the *ompA* gene but not the *omcB* gene of *Chlamydia* contributes to serovar-specific differences in tissue tropism, immune surveillance, and persistence of the organism. J. Bacteriol. **183:**5997–6008.
30. **Mygind, P. H., G. Christiansen, P. Roepstorff, and S. Birkelund.** 2000. Membrane proteins PmpG and PmpH are major constituents of Chlamydia trachomatis L2 outer membrane complex. FEMS Microbiol. Lett. **186:**163–169.
31. **Nei, M., and T. Gojobori.** 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:**418–426.
32. **Nei, M., and S. Kumar.** 2000. Molecular evolution and phylogenetics, p. 33–34. Oxford University Press, New York, N.Y.
33. **Ochman, H., S. Elwyn, and N. A. Moran.** 1999. Calibrating bacterial evolution. Proc. Natl. Acad. Sci. USA **96:**12638–12643.
34. **Ohta, T.** 1997. The meaning of near-neutrality at coding and non-coding regions. Gene **205:**261–267.
35. **Padidam, M., S. Sawyer, and C. M. Fauquet.** 1999. Possible emergence of new geminiviruses by frequent recombination. Virology **265:**218–225.
36. **Posada, D.** 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol. Biol. Evol. **19:**708–717.
37. **Posada, D., and K. A. Crandall.** 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. USA **98:**13757–13762.
38. **Read, T. D., R. C. Brunham, C. Shen, S. R. Gill, J. F. Heidelberg, O. White, E. K. Hickey, J. Peterson, T. Utterback, K. Berry, S. Bass, K. Linher, J. Weidman, H. Khouri, B. Craven, C. Bowman, R. Dodson, M. Gwinn, W. Nelson, R. DeBoy, J. Kolonay, G. McClarty, S. L. Salzberg, J. Eisen, and C. M. Fraser.** 2000. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res. **28:**1397–1406.
39. **Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas.** 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:**2496–2497.
40. **Salminen, M. O., J. K. Carr, D. S. Burke, and F. E. McCutchan.** 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res. Hum. Retrovir. **11:**1423–1425.
41. **Sibley, M. H., and E. A. Raleigh.** 2004. Cassette-like variation of restriction enzyme genes in Escherichia coli C and relatives. Nucleic Acids Res. **32:** 522–534.
42. **Smith, J. M.** 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. **34:**126–129.
43. **Stephens, R. S.** 1992. Challenge of Chlamydia research. Infect. Agents Dis. **1:**279–293.
44. **Stephens, R. S., S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R. L. Tatusov, Q. Zhao, E. V. Koonin, and R. W. Davis.** 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science **282:**754–759.
45. **Stephens, R. S., R. Sanchez-Pescador, E. A. Wagar, C. Inouye, and M. S. Urdea.** 1987. Diversity of *Chlamydia trachomatis* major outer membrane protein genes. J. Bacteriol. **169:**3879–3885.
46. **Stephens, R. S., M. R. Tam, C. C. Kuo, and R. C. Nowinski.** 1982. Monoclonal antibodies to Chlamydia trachomatis: antibody specificities and antigen characterization. J. Immunol. **128:**1083–1089.
47. **Stothard, D. R., G. Boguslawski, and R. B. Jones.** 1998. Phylogenetic analysis of the *Chlamydia trachomatis* major outer membrane protein and examination of potential pathogenic determinants. Infect. Immun. **66:**3618–3625.
48. **Stothard, D. R., G. A. Toth, and B. E. Batteiger.** 2003. Polymorphic membrane protein H has evolved in parallel with the three disease-causing groups of *Chlamydia trachomatis*. Infect. Immun. **71:**1200–1208.
49. **Tanzer, R. J., and T. P. Hatch.** 2001. Characterization of outer membrane proteins in *Chlamydia trachomatis* LGV serovar L2. J. Bacteriol. **183:**2686–2690.
50. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
51. **Wang, S. P., C. C. Kuo, R. C. Barnes, R. S. Stephens, and J. T. Grayston.** 1985. Immunotyping of Chlamydia trachomatis with monoclonal antibodies. J. Infect. Dis. **152:**791–800.
52. **Yoshiyama, K., and H. Maki.** 2003. Spontaneous hotspot mutations resistant to mismatch correction in Escherichia coli: transcription-dependent mutagenesis involving template-switching mechanisms. J. Mol. Biol. **327:**7–18.

Editor: A. D. O'Brien