Molecular Evolution of *Rickettsia* Surface Antigens: Evidence of Positive Selection

Guillaume Blanc, Maxime Ngwamidiba,† Hiroyuki Ogata,* Pierre-Edouard Fournier,*† Jean-Michel Claverie,* and Didier Raoult†*

*Information Génomique et Structurale, UPR 2589, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France; and †Unité des rickettsies, IFR 48 CNRS UMR 6020, Faculté de médecine, Université de la Méditerranée, 27 Boulevard Jean Moulin, 13385 Marseille Cedex 05, France

The *Rickettsia* genus is a group of obligate intracellular parasitic α-proteobacteria that includes human pathogens responsible for the typhus disease and various types of spotted fevers. rOmpA and rOmpB are two members of the "surface cell antigen" (Sca) autotransporter (AT) protein family that may play key roles in the adhesion of the *Rickettsia* cells to the host tissue. These molecules are likely determinants for the pathogenicity of the *Rickettsia* and represent good candidates for vaccine development. We identified the 17 members of this family of outer-membrane proteins in nine fully sequenced *Rickettsia* genomes. The typical architecture of the Sca proteins is composed of an N-terminal signal peptide and a C-terminal AT domain that promote the export of the central passenger domain to the outside of the bacteria. A characteristic of this family is the frequent degradation of the genes, which results in different subsets of the *sca* genes being expressed among *Rickettsia* species. Here, we present a detailed analysis of their phylogenetic relationships and evolution. We provide strong evidence that rOmpA and rOmpB as well as three other members of the Sca protein family—Sca1, Sca2, and Sca4—have evolved under positive selection. The exclusive distribution of the predicted positively selected sites within the passenger domains of these proteins argues that these regions are involved in the interaction with the host and may be locked in "arms race" coevolutionary conflicts.

Introduction

The *Rickettsia* are obligate intracellular parasites belonging to the α-protobacteria. They are associated with arthropods (i.e., flea, tick, mite, and lice) and often pathogenic for humans. Members of the *Rickettsia* genus include the etiologic agents of epidemic and murine typhus (*Rickettsia prowazekii* and *Rickettsia typhi*), as well as the etiologic agents of the Rocky Mountain spotted fever (*Rickettsia rickettsii*) and the Mediterranean spotted fever (*Rickettsia conorii*), two tick-borne diseases. Due to their medical importance, this group of bacteria has justified considerable sequencing efforts over the last 7 years. To date, the genome sequences of three *Rickettsia* species have been published (*R. conorii* [Ogata et al. 2001], *R. prowazekii* [Andersson et al. 1998], and *R. typhi* [McLeod et al. 2004]) and at least six other sequencing projects are under way for *Rickettsia belli*, *Rickettsia africae*, *Rickettsia sibirica*, *R. rickettsii*, *Rickettsia akari*, and *Rickettsia felis*. This abundance of data, coupled to a well-preserved colinearity of the genes, make them a very good model to study bacterial evolution.

Rickettsia cells are surrounded by a crystalline proteic layer (Palmer, Martin, and Mallavia 1974), also referred to as S-layer, which represents 10% to 15% of their total protein mass (Ching et al. 1996) and is composed of immunodominant surface protein antigens (SPA) (Dasch 1981; Ching et al. 1990; Ching, Carl, and Dasch 1992). Two SPAs, i.e., rOmpA (Anacker et al. 1987; Vishwanath, McDonald, and Watkins 1990) and rOmpB (Gilmore, Joste, and McDonald 1989; Gilmore et al. 1991), have been identified in several *Rickettsia* species and are the major antigenic determinants eliciting an immune response in patients infected by rickettsioses (Teysseire and Raoult 1992). rOmpA and rOmpB are two large proteins (about 2,000 and 1,600 aa, respectively) that share no significant sequence similarity except that they exhibit a highly conserved ~300-aa C-termini. This region, the autotransporter (AT) domain, folds into a β-barrel structure and inserts into the outer membrane. This structure forms a pore through which the central passenger domain of the proprotein is transported across the membrane (Desvaux, Parham, and Henderson 2004). Experimental studies have demonstrated that following export, the rOmpB proprotein is proteolytically cleaved to release the passenger peptide at the cell surface (Ching et al. 1990; Gilmore et al. 1991; Ching, Carl, and Dasch 1992; Hackstadt et al. 1992). In addition, rOmpA and rOmpB possess an N-terminal signal peptide allowing their passage through the inner membrane presumably by means of the Sec secretion system (Henderson et al. 2004).

Members of the AT protein superfamily are found in α-, β-, γ-, and ε-proteobacteria as well as in *Chlamydia* species (Henderson and Nataro 2001). They all share a homologous C-terminal AT domain. In contrast, the passenger domains are not all homologous and encode for a wide variety of virulence factors that can catalyze proteolysis, serve as adhesins, mediate actin-promoted bacterial motility, or act as cytotoxins to animal cells (Henderson and Nataro 2001). Experimental studies have suggested that rOmpA and rOmpB function as adhesin in *Rickettsia* (Li and Walker 1998; Uchiyama 2003).

Analyses of the genome sequences of *R. conorii* (Ogata et al. 2001) and *R. prowazekii* (Andersson et al. 1998) have revealed the existence of three additional genes encoding a highly conserved C-terminal AT domain. These genes are annotated as "surface cell antigen" (*sca*) genes (i.e., *sca1*, *2*, *3*), together with *rOmpA* (*sca0*) and *rOmpB* (*sca5*). In addition, "*GeneD*" (Sekeyova, Roux, and Raoult 2001) has been renamed *sca4* (Andersson et al. 1998),

Key words: adaptive evolution, *Rickettsia*, autotransporter, adhesin, antigen.

E-mail: guillaume.blanc@igs.cnrs-mrs.fr.

Mol. Biol. Evol. 22(10):2073–2083. 2005
doi:10.1093/molbev/msi199
Advance Access publication June 22, 2005

though the protein product lacks the AT domain. So far, only rOmpA and rOmpB proteins have been detected by sodium dodecyl sulfate–polyacrylamide gel electrophoresis or western blotting in *R. conorii* (Teysseire and Raoult 1992) as well as Sca4 in *R. japonica* (Uchiyama, Zhao, and Uchida 1996).

Given the seemingly prominent role of rOmpA and rOmpB proteins in the virulence of rickettsiae, we further characterized the Sca protein family by using the data from six publicly available *Rickettsia* genomes as well as the complete genome sequences of *R. belli*, *R. felis*, and *R. africae* that have been recently determined in our laboratory. We constructed an updated catalog of 17 subfamilies of *sca* genes found in nine *Rickettsia* spp. and analyzed their phylogenetic relationships. In addition, we have investigated the selective pressure acting on the Sca proteins. Our results suggest that the passenger domains of rOmpA, rOmpB, Sca1, and Sca2 as well as the Sca4 protein have evolved under positive selection.

Methods

Data

The genome sequences of *R. conorii* (NC_003103), *R. typhi* (NC_006142), *R. sibirica* (NZ_AABW00000000), *R. rickettsii* (NZ_AADJ00000000), *R. felis* (CP000053, CP000054) *R. akari* (NZ_AAFE00000000), and *R. prowazekii* (NC_000963) were downloaded from GenBank. For these species, the complete set of proteins was constructed using the GenBank annotations. We used the complete genome sequences of *R. belli* and *R. africae* that have been recently determined in our laboratory and will be published elsewhere. Every open reading frames (ORFs) of more than 100 nonstop codons were collected from the six possible reading frames and virtually translated into proteins. For these organisms, the ORF nucleotide sequences of the *sca* genes were released in GenBank. Locus names and GenBank accession numbers of all analyzed sequences are available as Supplementary Material online. For the study of selective constraints acting on *sca* genes, only the second halves of the gene sequences were available in most organisms for *rOmpA* and *sca1*. In consequence, subsequent analyses were restricted to these regions of proteins, which contain part of the passenger domain and the complete AT domain.

Sequence analysis

ORF products with sequence similarity with the Sca proteins previously identified in *R. conorii* and *R. prowazekii* (i.e., rOmpA, rOmpB, Sca1, Sca2, Sca3, and Sca4) were searched in the nine complete rickettsial proteomes using the BlastP program (E value $< 1 \times 10^{-10}$; Altschul et al. 1997). The AT domains were searched in the Sca protein sequences and the GenBank protein database with the program HMMSEARCH from the HMMER package (R. S. Eddy, unpublished data, http://hmmer.wustl.edu/), using the Pfam AT domain HMM profile (PF03797.6) as query. Hits were considered significant when they matched the Pfam profile with E values $\leq 10^{-5}$. Orthologous relationships between *sca* ORFs were determined by the reciprocal best BlastP match criterion. Except for *R. bellii*, the *Rickettsia* genomes exhibit a nearly perfect colinearity and few genomic rearrangements (Ogata et al. 2001). Hence, we confirmed the orthologous relationships by verifying that the genes surrounding each orthologous *sca* genes were in collinear order. Repeated peptide motifs were identified using the dotter program (Sonnhammer and Durbin 1995). Predictions of N-terminal signal peptides were done using the TMHMM web tool at http://www.cbs.dtu.dk/services/TMHMM/ (Krogh et al. 2001). The significance of sequence similarity between paralogous passenger domains was assessed using the program PRDF from the FASTA2 package (Pearson 1990) with default parameters. The threshold of significance was set to P value (opt delta) < 0.01.

Phylogenetic analysis

Protein and nucleotide alignments were carried out using the programs MUSCLE (Edgar 2004) and ClustalW (Higgins and Sharp 1988), respectively, and corrected manually. We constructed the phylogenetic trees of the *sca5* nucleotide sequences and AT domains (fig. 2) using the Neighbor-Joining (NJ) method implemented in the MEGA software (Kumar, Tamura, and Nei 1994). Phylogenetic distances were calculated using the Tamura-Nei method (Tamura and Nei 1993) for nucleotide sequences and a gamma correction (alpha = 3.34) for the AT domains. The shape parameter alpha for the gamma correction was determined by maximum likelihood using the CODEML program from the PAML package (Yang 1997). For all tree constructions, bootstrap supports for branches were assessed using 100 pseudoreplicates.

Analysis of selective constraints

For *sca1*, *sca2*, *sca4*, *rOmpA*, and *rOmpB*, NJ phylogenetic trees were generated from the nucleotide alignments using the Tamura-Nei distance (trees are provided as supplementary data, Supplementary Material online). To test for positive selection, we fitted four codon-based models of nucleotide substitutions to the data (Yang et al. 2000) using the maximum likelihood method implemented in the CODEML program (Yang 1997). We then compared the models that do not allow for (or do not detect) $\omega > 1$ (M0 and M7 see below) with the ones that identify classes of sites with $\omega > 1$ (M3 and M8) and determined the models that best describe the data. Each analysis was repeated three times with different initial ω values to avoid problems of multiple local optima. We recorded the parameter estimates corresponding to the highest likelihood. The four codon-based models used in this study were selected following recommendations of Yang et al. (2000): Model M0 assumes that all amino acid sites have the same ω estimated freely from the data. Model M3 ($K = 3$) assumes three classes of sites with ω estimated independently from the data (Yang et al. 2000). Model M7 assumes eight classes of sites with independent ω, which are limited to the interval (0, 1) and distributed according to a β distribution. Model M8 (β and ω) adds an extra site class to M7 with a free ω ratio estimated from the data, which may be higher than one. Under model M8, the values of the average (and maximal)

rate of synonymous substitution (d_S) for branches are 0.006 (max. 0.0142), 0.018 (max. 0.0818), 0.021 (max. 0.1588), 0.007 (max. 0.2111), 0.029 (max. 0.3185), and 0.062 (max. 0.7779) for the *sca1*, *sca2*, *sca4*, *rOmpA*, *rOmpB*, and *GltA* phylogenies, respectively. For the rate of nonsynonymous substitutions (d_N), the average (and maximal) values for branches are 0.007 (max. 0.0163), 0.015 (max. 0.0695), 0.011 (max. 0.0814), 0.005 (max. 0.1650), 0.014 (max. 0.1526), and 0.004 (max. 0.0439), respectively. Thus, the levels of substitutions between sequences are sufficiently low to assure that the effects of the saturation of mutation sites are negligible in the analyses ($K < 1$). Nested models (i.e., M0 vs. M3 and M7 vs. M8) were compared using the likelihood ratio test: twice the log-likelihood difference between two models ($2\Delta \ln L$) can be compared to a χ^2 distribution, with the number of degrees of freedom equal to the difference in the number of free parameters between the two models. To assess the sensitivity of the results to the tree topologies, all branches that were supported by less than 50% bootstrap replicates were collapsed into multifurcations. The results obtained with the collapsed trees were consistent with those obtained with the NJ trees, indicating that the method is fairly robust to approximate tree topologies (Yang et al. 2000).

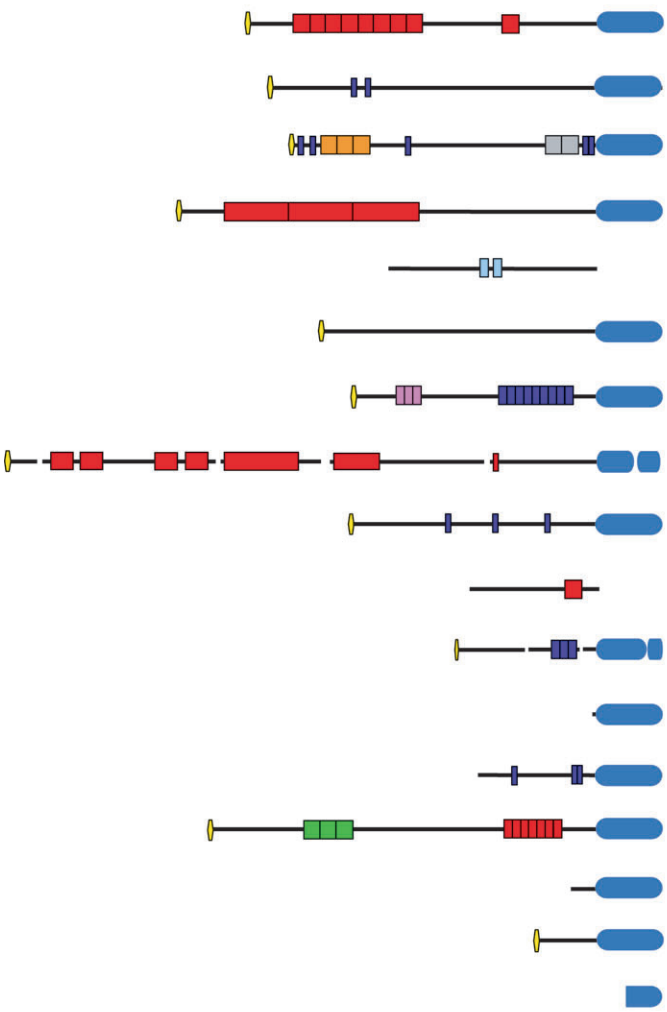
Results

Identification of *sca* genes in the *Rickettsia* genomes

Using BlastP searches against the proteomes of nine fully sequenced *Rickettsia* spp., we identified 145 ORF products with significant sequence similarity (E value $< 1 \times 10^{-10}$) with the six *sca* genes previously identified in the genomes of *R. conorii* and *R. prowazekii* (i.e., *rOmpA*, *rOmpB*, *sca1*, *sca2*, *sca3*, and *sca4*). These Sca Proteins were classified into 17 groups of orthologs (rOmpA, Sca1–Sca4, rOmpB, Sca6–Sca16) on the basis of their levels of sequence similarity and colinearity between genomes (fig. 1). To characterize their functional state, these genes were flagged as complete (flag "C"), split gene (flag "S": the entire protein is encoded by successive ORFs), fragment (flag "F": a single ORF encoding a protein of less than 50% of the longest orthologous protein), or absent (flag "-"). Split genes and gene fragments may be the result of ongoing gene degradation, a common feature in rickettsial genomes that is concomitant with reductive evolution (Andersson et al. 1998; Ogata et al. 2001). Thus, these may be incapable of producing functional proteins and could be pseudogenes. None of the nine *Rickettsia* genomes studied here harbor the 17 intact *sca* genes (fig. 1). *Rickettsia prowazekii* is remarkable as it retained only two intact *sca* genes (*sca3* and *rOmpB*). On the other extreme, the *R. felis* genome contains the largest complement of *sca* genes with 10 complete and 3 split *sca* genes. *rOmpB* is the only *sca* gene found complete in all the nine genomes studied.

The AT domain was identified in all complete *sca* genes except *sca4* and *sca9*. The later gene encodes a protein with significant sequence similarity over its whole length (BlastP E value $< 1 \times 10^{-10}$) with the passenger peptide of Sca3. The *sca4* gene is present in all nine *Rickettsia* genomes (fig. 1), though it is split in two orfs in the *R. prowazekii* genome sequence deposited in GenBank and so, possibly not functional in this species (NB: this gene has been found complete in independent sequencing of *sca4* in *R. prowazekii Madrid E*; unpublished observation). The Sca4 protein was detected in the cytoplasm of spotted fever group (SFG) *Rickettsia* (Uchiyama 1997) and recognized by antirickettsial antibodies (Schuenke and Walker 1994). The longest copy of the *sca9* gene (encoding 557 amino acids) was found in *R. felis* and shorter fragments were identified in the *R. conorii* and *R. sibirica* genomes (118 and 132 amino acids, respectively), as well as a split *sca9* gene in *R. bellii* (fig. 1). Thus, *sca9* sequences are likely degraded genes in *R. conorii*, *R. sibirica*, and *R. bellii*, and experimental studies are needed to determine whether the *R. felis sca9* gene is functional or not. The *sca16* ORF identified in *R. bellii* is also a probable degraded gene because it only encodes a truncated AT domain (fig. 1). The Sca11, Sca12, Sca14, and Sca15 proteins have short passenger domains as compared to the other sca proteins. Although passenger peptides are not always covalently linked to the AT domains as for example in *Escherichia coli* and *Helicobacter pylori* (Yen et al. 2002), the integrity of these proteins can be questioned. The *sca15* gene is uniquely found in the *R. bellii* genome and likely functional because this protein possesses the N-terminal signal peptide needed to be exported across the inner membrane. No signal peptide was predicted in the other three proteins. The passenger domains of rOmpA, Sca1, Sca2, Sca3, Sca6, Sca7, Sca8, Sca10, Sca12, and Sca13 as well as Sca4 and Sca9 contain conserved repeated peptide motifs (fig. 1), which are common in adhesive proteins (Wren 1991; Kehoe 1994). The number of repeated units is often variable among orthologs (not shown) and is as significant a source of sequence variability as already observed for rOmpA (Gilmore and Hackstadt 1991; Gilmore 1993; Fournier, Roux, and Raoult 1998) and other adhesive proteins (Gravekamp et al. 1996). rOmpA, Sca3, Sca7, Sca9, and Sca13 share similar repeated motifs as well as Sca1, Sca2, Sca6, Sca10, and Sca12. The other Sca proteins also present traces of internal sequence duplications, but the weak level of conservation between repeated units make their delineation difficult.

The degree of sequence similarity between the paralogous Sca proteins is contrasted. The AT domain is generally well conserved between paralogs with an average of 31% identical amino acid residues and a maximum of 86% between Sca2 and Sca6. On the other hand, the passenger domain and the signal peptide present a much lower level of sequence conservation. The regions of significant sequence identity were often limited to small segments of the peptides corresponding to the repeated motifs. Nevertheless, paralogous *sca* proteins can be grouped on the basis of the similarity between their passenger domains, as detected by the PRDF program—group 1: rOmpA, Sca3, Sca7, Sca9, and Sca13; group 2: Sca1, Sca2, Sca6, Sca8, Sca11; Sca12 and Sca14. rOmpB, Sca4, Sca10, Sca16, and Sca15 constitute singleton groups as their passenger domains do not exhibit significant sequence similarity with any paralogous Sca proteins. The average levels of amino acid identity between passenger domains are 18% (max. 45% between Sca3 and Sca7) and 9% (max. 14% between Sca2 and Sca8) for groups 1 and 2, respectively. When the



Gene	Length (aa)	Rbe	Rpr	Rty	Rfe	Rak	Rri	Rco	Rsi	Raf
rOmpA (sca0)	2021	F	-	-	S	S	C	C*	C	C
sca1	1902	C	S	C	C	C	C	C*	C	C
sca2	1795	C sca2–sca6 ancestor	-	-	C	C	C	C*	C	C
sca3	2340	S	C*	C	C	S	S	F	F	-
sca4	1026	C	S	C	C	C	C	C*	C	C
rOmpB (sca5)	1655	C	C	C	C	C	C	C*	C	C
sca6	1483	C sca2–sca6 ancestor	S	C*	-	C	-	-	-	-
sca7	3180 (6 ORFs)	-	-	-	S*	S	-	-	-	-
sca8	1502	S	-	-	C*	S	S	S	S	S
sca9	557	S	-	-	C*	-	-	F	F	-
sca10	1011 (3 ORFs)	-	-	-	S	S*	S	F	F	-
sca11	333	-	-	-	C*	-	-	-	-	-
sca12	889	-	-	-	C*	-	-	-	-	F
sca13	2199	S	-	-	C*	-	S	S	S	-
sca14	421	C*	-	-	-	-	-	-	-	-
sca15	637	C*	-	-	-	-	-	-	-	-
sca16	186	F*	-	-	-	-	-	-	-	-

FIG. 1.—The 17 *sca* gene families identified in nine *Rickettsia* genomes. The structural domains of the *sca* proteins are schematized on the left. The yellow diamonds, thick black lines, and the blue ovals represent the predicted peptide signals, passenger domains, and AT domains, respectively. Interruptions in the thick black lines or the blue ovals indicate the occurrence of in-frame stop codons in the coding sequence. Colored boxes show the presence of repeated peptide motifs. Boxes of same color refer to similar peptide motifs. The table on the right summarizes the names, lengths of the peptide products, and the presence or absence of the *sca* genes in the *Rickettsia* genomes. The characters "C," "S," "F," and "-" indicate whether the gene is complete, split, fragment, or absent, respectively, in a given genome. Genes were defined as split if the entire protein is encoded by successive ORFs and as fragment if the gene size was less than 50% of that of the longest ortholog. Asterisks indicate the genes from which the length and the structural domains of the peptide products are shown. For Sca7 and Sca10, the protein lengths were obtained after concatenating the successive ORFs.

Rickettsia passenger domains are used as query in BlastP searches against public databases, no homologous sequence can be found outside of the *Rickettsia* genus. This contrasts with the AT domains for which matches can be found in many proteobacteria and *Chlamydia* (Henderson and Nataro 2001).

Phylogenetic analysis of the Sca sequences

To illustrate the relationships between the nine sequenced *Rickettsia* species, we first reconstructed their phylogeny using the entire *sca5* nucleotide sequences (fig. 2*A*). All branches of the NJ tree are supported by high bootstrap values (≥86%) and the topology is in agreement with previous phylogenetic studies of the *Rickettsia* genus (Stothard, Clark, and Fuerst 1994; Roux et al. 1997; Fournier, Roux, and Raoult 1998; Stenos et al. 1998; Roux and Raoult 2000). *Rickettsia prowazekii* and *R. typhi* are clustered together and form the Typhus Group (TG). The SFG includes all the other species but *R. bellii* and is divided into two subgroups: the rickettsii group, which comprises *R. rickettsii*, *R. conorii*, *R. sibirica*, and *R. africae*, and the akari group, which includes *R. akari* and *R. felis*. *Rickettsia bellii* forms a distinct and distant group, which, according to Stothard, Clark, and Fuerst (1994), diverged before the separation of the genus into the SFG and TG. As shown in superimposition on the phylogeny (fig. 2*A*), the primary hosts differ from species to species (Azad and Beard 1998). This implies

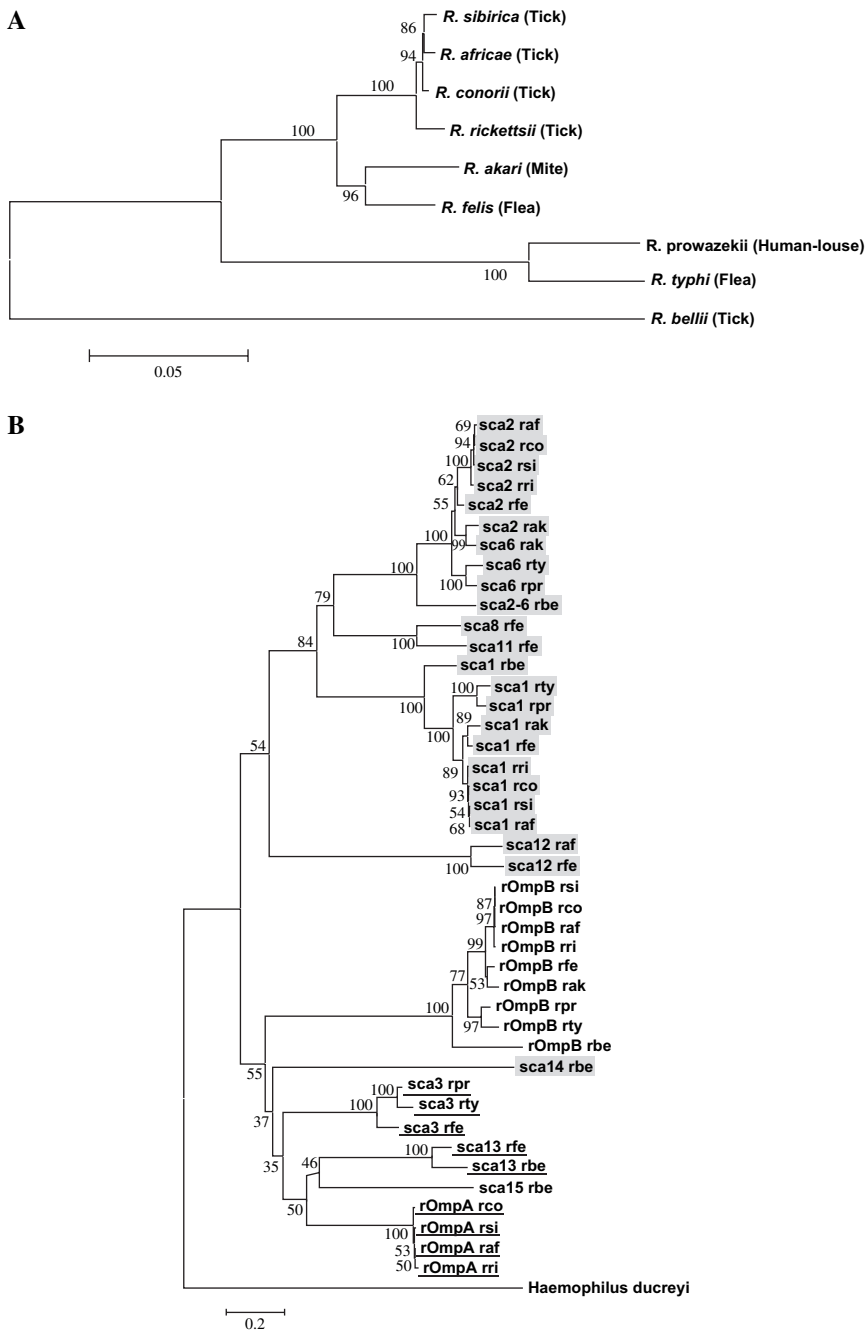


FIG. 2.—NJ phylogenetic trees of the Sca proteins in *Rickettsia*. Bootstrap values based on 100 pseudoreplicates are shown beside nodes. (*A*) Phylogenetic relationships of the nine fully sequenced *Rickettsia* species reconstructed from the analysis of *sca5* gene alignment. Phylogenetic distances between the nucleotide sequences were calculated using the Tamura-Nei method. Note that the tree is unrooted. The choice of *Rickettsia bellii* as out-group species is based on Stothard, Clark, and Fuerst (1994) (*B*) Phylogenetic tree reconstructed from the alignment of the AT domains. Only ORF products with complete AT domains were analyzed. All non-*Rickettsia* AT domains were found distantly related to those of *Rickettsia*, and the AT domain of the *Haemiphilus ducrei* putative serine protease (HD1280) was arbitrarily chosen to root the tree. Phylogenetic distances were corrected with the gamma correction (shape parameter $\alpha = 3.34$). rco, rsi, raf, rri, rfe, rak, rpr, rty, and rbe stand for *Rickettsia conorii*, *Rickettsia sibirica*, *Rickettsia africae*, *Rickettsia rickettsii*, *Rickettsia felis*, *Rickettsia akari*, *Rickettsia prowazekii*, *Rickettsia typhi*, and *R. bellii*, respectively. Similarity analysis of sequences delineated two groups of homologous passenger domains (see text): group 1 sequences are underlined and group 2 sequences are represented on gray background.

that several *Rickettsia* lineages have changed their host range during their evolution.

A total of 283 AT proteins were identified from non-*Rickettsia* species using HMM searches against the GenBank database. Preliminary phylogenetic analyses of the ~300-aa AT domains from these proteins and 43 *Rickettsia* Sca proteins indicated that the *Rickettsia* AT domains are closer to each other than to any other non-*Rickettsia* AT domain (fig. 2*b*). This suggests that the paralogous *sca* genes have appeared relatively late in the evolution of

the *Rickettsia* lineage. However, the phylogenetic distributions of *sca1*, *sca5*, *sca8*, *sca13*, *sca14* and the ancestor of *sca2* and *sca6* imply that these genes arose before the separation of *R. bellii* from the other *Rickettsia* spp. (fig. 2*B*) that occurred early after the emergence of the *Rickettsia* genus (Stothard, Clark, and Fuerst 1994). All orthologous AT domains cluster together in monophyletic groups supported by high bootstrap scores except for the Sca2 and Sca6 AT domains (fig. 2*B*) that form a polyphyletic cluster (see below). We failed to produce a fully resolved phylogeny using either the amino acid or the nucleotide sequence alignment. The branching of the rOmpA, rOmpB, Sca3, Sca12, Sca13, Sca14, and Sca15 subtrees are not supported by high bootstrap scores (fig. 2*B*). This precludes any firm conclusion about the sequential order in which these *sca* genes have been created. Nevertheless, the AT domain tree presented in figure 2*B* is in agreement with the grouping of homologous passenger domains as described above, except Sca14 that falls in the passenger domain group 2, whereas its AT domain presents an affinity to the AT domains of the group 1 Sca proteins.

A notable feature of the evolution of the Sca family is the complex phylogenetic relationships exhibited by the AT domain of Sca2 and Sca6 proteins. The distribution of Sca6 sequences among *Rickettsia* spp. indicates that this gene arose before the divergence of the TG and the SFG (fig. 1). Furthermore, TBlastN alignments revealed that sequence remnants of the *sca2* gene exist in the syntenic orthologous regions of the *R. prowazeki* and *R. typhi* genomes (not shown), indicating that *sca2* also arose before the separation of the TG and SFG *Rickettsia* spp. and was later degraded in the two species. Remarkably, the AT domains of Sca2 and Sca6 in *R. akari* are phylogenetically closer to each other than to any members of their own sub-family (fig. 2*B*). The two genes have a head to tail organization in *R. akari*, and the similarity between their AT domains exceeds 95% at the nucleotide level, whereas only very weak amino acid similarity can be detected elsewhere between these genes. The most likely explanation is that a gene conversion event occurred between the two *R. akari sca* genes that resulted probably in a conversion of the AT domain of Sca6 into a Sca2-like AT domain. Under these circumstances, we cannot exclude that the tandem duplication of *sca2* and *sca6* from a single ancestral gene may have occurred much earlier than shown on the phylogenetic tree (fig. 2*B*) and that recurrent gene conversions between the *sca2* and *sca6* genes may have homogenized the AT domain sequences.

Analysis of the selective pressure acting on the Sca proteins

AT domains are known to form a β-barrel structure into the outer membrane. The complex folding required to form such a structure is likely to constrain the number and the nature of amino acid changes in the peptide sequences. On the other hand, the Sca passenger peptides are presumably exported to the cell surface and are suspected to be involved in host-parasite interactions. For instance, rOmpA and rOmpB are likely responsible for the adhesion of rickettsiae to their host cells (Li and Walker 1998; Uchiyama 2003). Positive selection has been shown to promote the divergence of protein regions involved in host-parasite interactions (Schulenburg et al. 2000; Peek et al. 2001; Jiggins, Hurst, and Yang 2002), but this has not yet been tested for Sca proteins.

Nucleotide substitutions in protein-coding sequences can result either in amino acid change (nonsynonymous substitutions) or not (synonymous substitutions). In *Rickettsia*, selection on synonymous codons is probably ineffective (Andersson and Sharp 1996), and so, synonymous codon positions probably accumulate changes neutrally at a rate similar to the mutation rate. Under such circumstances, the ratio of d_N to d_S, denoted $\omega = d_N/d_S$ herein, measures the magnitude and direction of selective pressure on a protein, with $\omega = 1$, <1, and >1 indicating neutral evolution, purifying selection, and positive diversifying selection, respectively (Li 1997). To examine if the AT and passenger domains evolved under different selection regimes, we estimated the ratio ω by a maximum likelihood approach. In addition, we tested if the Sca passenger domains were subjected to adaptive evolution by performing phylogeny-based statistical tests of positive selection (Yang et al. 2000). We studied five *sca* genes—namely, *sca1*, *sca2*, *sca4*, *rOmpA*, and *rOmpB* (table 1)—because their nucleotide sequences were determined in many *Rickettsia* species for phylogenetic purpose and were included in the analyses (Fournier, Roux, and Raoult 1998; Roux and Raoult 2000; Sekeyova, Roux, and Raoult 2001; M. N. Ngwamidiba, in preparation). The other *sca* genes were not analyzed here because the number of available sequences was not large enough to allow for powerful statistical testing (Anisimova, Bielawski, and Yang 2001).

We first divided each nucleotide alignment into passenger and AT domains (except for *sca4* that has no AT domain) and estimated the average ω_0 by fitting model M0 to the data. The ω_0 estimates for the passenger and AT regions are, respectively, $\omega_{PASSENGER} = 0.58$ versus $\omega_{AT} = 0.44$ for rOmpA, $\omega_{PASSENGER} = 0.41$ versus $\omega_{AT} = 0.23$ for rOmpB, $\omega_{PASSENGER} = 1.20$ versus $\omega_{AT} = 0.35$ for Sca1, and $\omega_{PASSENGER} = 0.77$ versus $\omega_{AT} = 0.39$ for Sca2. Thus, for all four Sca proteins, the ω ratio estimated for the AT domain is consistently lower than for the passenger domain. As expected, this result indicates that the AT domains have evolved under higher selective constraints than their respective passenger domains. Except for the Sca1 passenger domain, the average ω ratios are all smaller than one and therefore averaging ω on all sites failed to detect positive selection. However, adaptive selection typically occurs at a few sites as most amino acids in a protein are under structural and functional constraints with $\omega < 1$. Thus, calculating ω as an average over all amino acid sites is often too conservative to detect positive selection (Yang and Bielawski 2000).

We therefore analyzed the entire alignments of the five *sca* genes using models that allow for several ω categories of amino acid sites. The likelihood scores obtained under models M0 and M3 (table 1) were compared using the likelihood ratio test. The tests were significant for all families after Bonferroni correction ($\alpha = 0.05$; table 2). Thus, model M3, with three ω categories of sites, better fits the data than model M0 that accounts for a single ω category for all sites.

Table 1
Log-likelihood Scores and Parameter Estimates

Gene	n[a]	ls[a]	Tree Length[b]	Model	l	Parameters in the ω Distribution[c]	Predicted Positively Selected Sites[d]
rOmpA (*sca0*)	37	1,032	1.06	M0	−9,563	$\omega = 0.55$	(RC1273), 962, 1,009, 1,015, 1,030, 1,053, 1,056, 1,085, 1,089, 1,120, 1,140, 1,142, 1,157, 1,161, 1,247, 1,255, 1,256, 1,267, 1,273, 1,628, 1,722, 1,723
				M3	−9,331	$\omega_1 = 0.19$, $p_1 = 0.80$; $\omega_2 = 2.50$, $p_2 = 0.17$; $\omega_3 = 11.00$, $p_3 = 0.03$	
				M7	−9,411	B(p = 0.020, q = 0.025)	
				M8	−9,334	B(p = 0.284, q = 0.513), $p_1 = 0.91$, $\omega = 5.29$, $p_2 = 0.09$	
rOmpB (*sca5*)	23	1,591	2.18	M0	−2,0427	$\omega = 0.39$	(AAF34129), 21, 29, 87, 90, 173, 455, 607, 690, 710, 757, 874, 880, 956, 1,043
				M3	−19,959	$\omega_1 = 0.07$, $p_1 = 0.59$; $\omega_2 = 0.86$, $p_2 = 0.37$; $\omega_3 = 4.36$, $p_3 = 0.02$	
				M7	−20,014	B(p = 0.23, q = 0.37)	
				M8	−19,958	B(p = 0.27, q = 0.48); $p_1 = 0.96$, $\omega = 3.9$, $p_2 = 0.04$	
sca1	11	1,172	0.36	M0	−6,950	$\omega = 0.98$	(RC0019), 210, 215, 240, 319, 336, 817, 941, 1,219
				M3	−6,904	$\omega_1 = 0.00$, $p_1 = 0.41$; $\omega_2 = 1.29$, $p_2 = 0.54$; $\omega_3 = 10.29$, $p_3 = 0.05$	
				M7	−6,933	B(p = 0.007, q = 0.005)	
				M8	−6,904	B(p = 0.033, q = 0.018), $p_1 = 0.93$; $\omega = 8.66$, $p_2 = 0.07$	
sca2	14	1,588	1.21	M0	−1,5124	$\omega = 0.71$	(RC0110), 182, 341, 345, 349, 373, 400, 402, 422, 589, 694, 1037, 1053, 1,086, 1,174, 1,278, 1,347, 1,438
				M3	−14,885	$\omega_1 = 0.10$, $p_1 = 0.51$; $\omega_2 = 1.31$, $p_2 = 0.46$; $\omega_3 = 9.19$, $p_3 = 0.03$	
				M7	−14,963	B(p = 0.024, q = 0.020)	
				M8	−14,887	B(p = 0.024, q = 0.014), $p_1 = 0.96$, $\omega = 7.17$, $p_2 = 0.04$	
sca4	17	874	1.17	M0	−8,207	$\omega = 0.44$	(RC0667), 183, 521, 695
				M3	−8,142	$\omega_1 = 0.12$, $p_1 = 0.57$; $\omega_2 = 0.71$, $p_2 = 0.32$; $\omega_3 = 1.97$, $p_3 = 0.11$	
				M7	−8,148	B(p = 0. 24, q = 0.30)	
				M8	−8,141	B(p = 0.74, q = 1.62); $p_1 = 0.87$ $\omega = 1.86$, $p_2 = 0.13$	
Control GltA	20	385	1.75	M0	−3,697	$\omega = 0.05$	None
				M3	−3,674	$\omega_1 = 0.00$, $p_1 = 0.36$; $\omega_2 = 0.00$, $p_2 = 0.34$; $\omega_3 = 0.18$, $p_3 = 0.30$	
				M7	−3,675	B(p = 0.20, q = 3.11)	
				M8	−3,674	B(p = 0.01, q = 0.25); $p_1 = 0.80$, $\omega = 0.18$, $p_2 = 0.20$	

[a] The data have n sequences, each of the ls codons after alignment gaps were removed.
[b] The tree length is the sum of the branch lengths along the tree, measured by the expected number of substitutions per codon.
[c] p_i denotes the proportion of site falling in site class ω_i.
[d] Sites potentially under positive selection identified under model M8 with a posterior probability <5% or <1% (underlined) are listed according to the numbering of the *R. conorii* sequence given in parentheses.

This indicates that the selective pressure varies among the amino acid sites in each protein. In addition, parameter estimates under model M3 suggest the presence of sites under positive selection in each of the five proteins (as indicated by $\omega > 1$; table 1). To confirm this observation, likelihood scores were compared between the more realistic continuous β-distribution models M7 and M8. Again, M8 fits the data significantly better than M7 for all five Sca families after Bonferroni correction ($\alpha = 0.05$; table 2), and the estimated ω ratio for the extra class are all >1 (table 1). These results show that allowing for sites with $\omega > 1$ (model M8) significantly improves the fit to the data. Thus, there is clear statistical evidence that divergence of these five Sca proteins was promoted by positive selection.

Table 2
Likelihood Ratio Tests of Positive Selection

	M0 versus M3		M7 versus M8	
Gene	$2\Delta l$	P value (4 df)	$2\Delta l$	P value (2 df)
rOmpA (*sca0*)	464.32	3.48×10^{-99}	154.66	0.00
rOmpB (*sca5*)	936.44	2.12×10^{-201}	113.32	2.47×10^{-25}
sca1	91.84	5.35×10^{-19}	56.88	4.45×10^{-13}
sca2	477.94	3.95×10^{-102}	152.44	7.91×10^{-34}
sca4	131.49	1.8×10^{-27}	12.74	0.0017
GltA	47.17	1.4×10^{-9}	2.14	0.34

Positively selected amino acid sites were identified under model M8 using the Bayesian method developed by Nielsen and Yang (1998). Consistent with our hypothesis that only the passenger domains are involved in host-parasite interactions, the predicted positively selected sites are exclusively located within this region of Sca1, Sca2, rOmpA, and rOmpB (fig. 3). The distributions of the predicted positively selected sites between passenger and AT domains are highly nonrandom ($P < 3.2 \times 10^{-4}$; binomial distribution), which further confirms that the two domains have evolved under different selection regimes.

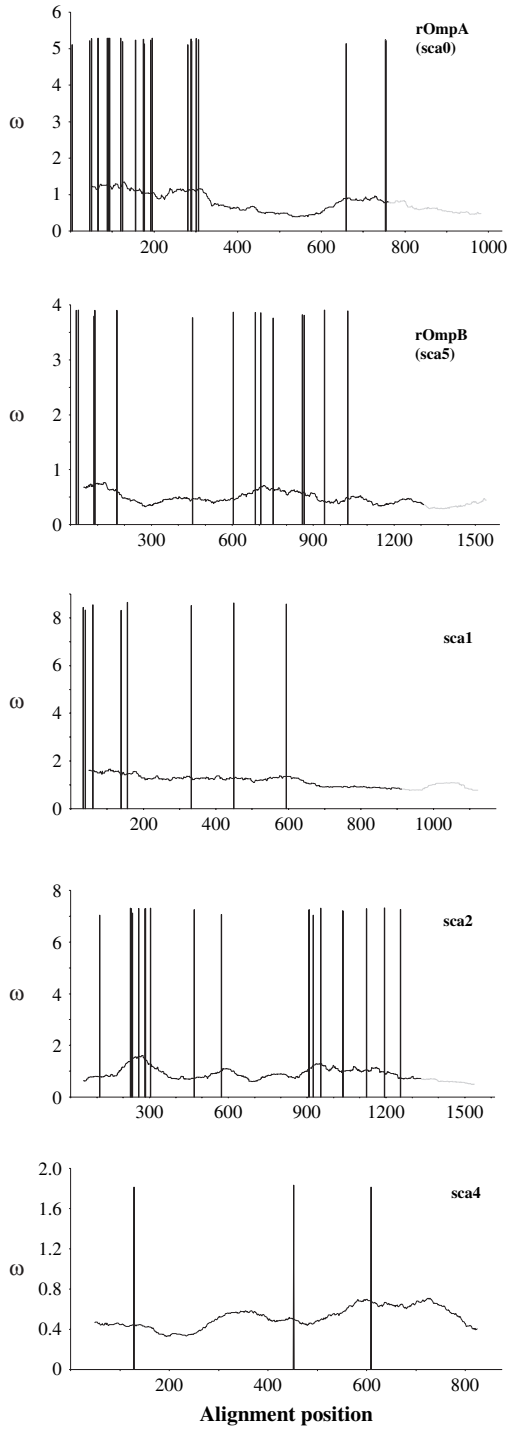


FIG. 3.—Schematic distribution of the site under positive selection along the protein sequences of rOmpA, rOmpB, Sca1, Sca2, and Sca4. Positively selected sites were identified under model M8 using the Bayesian method developed by Nielsen and Yang (1998). Positions of positively selected sites with a posterior probability >95% are indicated by vertical lines whose lengths are proportional to the average (Ω) of the ω ratios over the nine classes from model M8. The black and gray curves represent the means Ω calculated in sliding windows of 100 sites over the passenger and AT domains, respectively. Only the second halves of the rOmpA and Sca1 proteins were analyzed (see *Methods*).

Most *Rickettsia* species are maintained in the host population mainly through transovarial transmission (Azad and Beard 1998). This type of transmission constitutes a strong bottleneck in the *Rickettsia* population and promotes the fixation of slightly deleterious mutations by genetic drift, which tends to increases the rate of nonsynonymous substitutions d_N. To verify that the expected increase of d_N does not mislead likelihood methods into detecting positive selection in Sca proteins, we performed the same likelihood ratio tests for a housekeeping protein that presumably does not evolve under positive selection. We choose the GltA citrate synthase protein because the sequences are available for most species used for the studies of the five *sca* genes (Roux et al. 1997). The GltA sequences analyzed here have an overall level of divergence in the range of the *sca* genes, as measured by the tree length (nucleotide substitutions per codon; table 1). Comparison of model M0 with M3 suggests variation in the ω ratio between sites; however, no signal of positive selection was detected (table 1). Likewise, comparison of model M7 and M8 does not detect positive selection in the GltA sequences. These results suggest that the likelihood ratio tests are reliable for the detection of positive selection, even in organisms that evolve through recurrent bottlenecks.

Discussion

Although only rOmpA, rOmpB, and Sca4 are known to induce immune responses in infected patients (Teysseire and Raoult 1992), our analysis of the nine *Rickettsia* genomes reveals that the diversity of *sca* genes is much greater. With 17 distinct paralogous proteins, the *sca* genes represent the largest family of associated membrane proteins in this genus. Nevertheless, a remarkable feature of the Sca family is the extreme lability of the genes. Except *sca5*, none of the genes are found complete in all nine *Rickettsia* spp. In addition, no intact copy of *sca7*, *sca10*, and *sca16* is found in any genomes. *Rickettsia*, and more generally intracellular parasitic bacteria, are known to undergo genome reduction through degradation of genes presumably dispensable in stable environments (Sakharkar, Dhar, and Chow 2004). If we make the assumption that no lateral gene transfer occurred, the patterns of degradation/conservation of the *rOmpA*, *sca3*, *sca6*, *sca8*, *sca9*, *sca12*, and *sca13* genes among *Rickettsia* imply that they were independently degraded several times in distinct lineages.

There are evidence that rOmpA and rOmpB function as adhesin (Li and Walker 1998; Uchiyama 2003), but the functions of the *sca* genes remain unknown. A high level of sequence similarity is generally a good indicator of shared functionality between proteins. However, the major body of the Sca proteins is composed of the passenger domain that exhibits little or no detectable similarity among paralogous proteins. In addition to the AT domain that allows the mature protein to be exported, the Sca proteins share several structural features that may unify them in a common functional category. First, most Sca proteins have internal repeats within the passenger domain, a feature often found in adhesive proteins (Wren 1991; Kehoe 1994). These repeated motifs are often specific to each Sca protein and may contribute to the specific recognition of different

sets of host receptors. Analyses of the constraints acting on the protein sequences indicate that the selection regimes are similarly distributed along the sequences of Sca1, Sca2, rOmpA, and rOmpB. Remarkably, we found that the Sca passenger peptide may be a hot spot for positive selection. Thus, although the passenger domains of AT proteins in proteobacteria and *Chlamydia* are not always homologous (i.e., do not share a common ancestor), the lack of similarity between Sca passenger domains is likely due to rapid rates of sequence divergence promoted by positive selection. Intriguingly, we found evidence of positive selection in the sequence of Sca4. This protein lacks a recognizable signal peptide and the AT domain required for autonomous export across the outer membrane. Furthermore, Sca4 has been observed into the cytoplasm by immunoelectron microscopy (Uchiyama 1997). These observations argue against an adhesive role for Sca4. This protein is recognized in humoral and cell-mediated immune response (Schuenke and Walker 1994). Although both antibody and cell responses can develop against internal proteins, the fact that Sca4 presents evidence of positive selection as for the passenger domains of Sca1, Sca2, rOmpA, and rOmpB is compatible with the hypothesis that this protein is transported outside the cell at some point. Clearly, more data are needed to understand the function of sca4.

Coevolution between hosts and parasites is expected to take the form of a "Red Queen" scenario. The Red Queen is generally taken to be an ongoing process of reciprocal coadaptation (Lythgoe and Read 1998), whereby parasite and host meet adaptation with counteradaptation through evasion-recognition cycles. The evidence of positive diversifying selection in the passenger domains of five Sca proteins studied here suggests that the host genes and *Rickettsia* Sca proteins may be locked in such coevolutionary conflict. The evasion-recognition competition might take different forms in the case of Sca. Adhesive proteins mediate bacterial attachment to host cell receptors and play a central role in bacterial colonization (Hultgren et al. 1993). Hosts can evade parasite infection by changing the structure of the receptor regions where the recognition/attachment of the Sca adhesin occurs. Under this scenario, positive selection in Sca proteins may result from the selection of the genetic variants that are adapted to the new receptor structure and that are able to restore the interaction. For the presence of a parasite to promote receptor change, the parasite must be both highly prevalent in the host population and exact a severe fitness penalty that manifests itself in the loss of fecundity of the host or survival of the subsequent progeny. Excepted for *R. prowazekii* that kills its host (louse and human), it is unclear whether any of the *Rickettsia* species meet these criteria in either the mammalian or arthropod hosts. It is obvious that host specificity has changed during the evolution of some *Rickettsia* lineages (fig. 2*A*). Such ecological change is likely to involve similar adaptive steps enabling adhesins to interact with new receptor sets. The Sca proteins, as all external structures, may possibly be targets of the recognition of the bacteria by the host defense (Tewari et al. 1993). Positively selected sites in Sca proteins may be involved in sequence variability and, thus, evolve faster to produce novel structures and avoid recognition (Perez et al. 1998; Peek et al. 2001). The identification of those sites may thus provide insights for future efforts of vaccine development (McDonald, Anacker, and Garjian 1987; Wizemann, Adamou, and Langermann 1999; Crocquet-Valdes et al. 2001). The mechanism of sequence variation observed for Sca proteins is markedly different from that reported for outer-membrane proteins in other tick-borne bacterial pathogens such as *Borrelia burgdorferi* (Zhang et al. 1997; Zhang and Norris 1998; Stevenson and Miller 2003), *Anaplasma marginale* (Barbet et al. 1999; Brayton et al. 2002, 2005; Collins et al. 2005), and *Erhlichia ruminantium* (Collins et al. 2005). In these species, gene conversions between expressed functional loci and several variable silent "pseudogene" copies generate surface protein variants. This allows the long persistence of the organism into the host by evading the host immune system (Barbet et al. 1999; Brayton et al. 2002).

The early duplications of the *sca* genes may reflect adaptation of the *Rickettsia* ancestor to its hosts. Later degradation of some of these duplicates may be associated to speciation. It may not be surprising that *R. prowazekii* has only 2 intact *sca* genes (*sca3* and *rOmpB*), while the other species have from 5 to 10 complete *sca* genes (fig. 1). A notable difference between *R. prowazekii* and the other *Rickettsia* is that the primary reservoirs of *R. prowazekii* are mammals, while the others are intracellular parasites of arthropods and probably infect mammals only incidentally (fig. 2*A*). Moreover, *R. prowazekii* is the only known *Rickettsia* able to escape immune control in humans and to cause a late relapse (Brill-Zinsser Disease). Given that the defense system of mammals is more complex than that of arthropods, it is possible that the reduction of the pool of *sca* genes in *R. prowazekii* is also an adaptive response to limit its visibility to the host defense as suggested for the spirochete *Treponema* (Templeton 2004). This scenario is conceivable for *sca1*, *sca4*, and *sca6*, which are degraded in *R. prowazekii* but intact in the *R. typhi* genome and therefore were probably functional in the ancestor of *R. typhi* and *R. prowazekii*.

Conclusion

Although evidence that the Sca proteins function as adhesins is available only for rOmpA and rOmpB (Li and Walker 1998; Uchiyama 2003), our results suggest that Sca1 and Sca2 are paralogs of rOmpA and rOmpB and have evolved under similar selective regimes. This argues that Sca1 and Sca2, and perhaps all *Rickettsia* Sca proteins, may be also involved in adhesion processes. The adhesin (Sca)/host receptor interaction could play a key role in the host recognition. The real host ranges of extant rickettsies have not been exhaustively determined but are likely limited to very few organisms. Thus, the loss of *sca* genes might explain the small host range, and by extrapolation, one may speculate that the *Rickettsia* ancestor had a larger host range because it was encoding for an extended *sca* complement. Meanwhile, the Sca proteins are exposed at the cell surface, and so, represent potential targets for the host defense. Hence, the evolution of the *sca* gene complement has probably been marked by antagonist forces: the need of keeping functional *sca* genes to enter into host cells,

the need to avoid host defenses by loosing *sca* genes and/or sequence variability, the loss of *sca* genes as a consequence of host specialization, and the opportunity to colonize new ecological niche (i.e., a new host) by adaptation (duplication and sequence variation).

Supplementary Material

The NJ phylogenetic trees of the six genes used in the selective constraint analysis (*rOmpA*, *rOmpB*, *sca1*, *sca2*, *sca4*, and *GltA*) and the GenBank accession numbers of the sequences are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Literature Cited

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Anacker, R. L., G. A. McDonald, R. H. List, and R. E. Mann. 1987. Neutralizing activity of monoclonal antibodies to heat-sensitive and heat-resistant epitopes of Rickettsia rickettsii surface proteins. Infect. Immun. **55**:825–827.

Andersson, S. G., and P. M. Sharp. 1996. Codon usage and base composition in *Rickettsia prowazekii*. J. Mol. Evol. **42**: 525–536.

Andersson, S. G., A. Zomorodipour, J. O. Andersson, T. Sicheritz-Ponten, U. C. Alsmark, R. M. Podowski, A. K. Naslund, A. S. Eriksson, H. H. Winkler, and C. G. Kurland. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature **396**:133–140.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. **18**:1585–1592.

Azad, A. F., and C. B. Beard. 1998. Rickettsial pathogens and their arthropod vectors. Emerg. Infect. Dis. **4**:179–186.

Barbet, A. F., R. Blentlinger, J. Yi, A. M. Lundgren, E. F. Blouin, and K. M. Kocan. 1999. Comparison of surface proteins of Anaplasma marginale grown in tick cell culture, tick salivary glands, and cattle. Infect. Immun. **67**:102–107.

Brayton, K. A., L. S. Kappmeyer, D. R. Herndon, M. J. Dark, D. L. Tibbals, G. H. Palmer, T. C. McGuire, and D. P. Knowles Jr. 2005. Complete genome sequencing of Anaplasma marginale reveals that the surface is skewed to two superfamilies of outer membrane proteins. PNAS **102**:844–849.

Brayton, K. A., G. H. Palmer, A. Lundgren, J. Yi, and A. F. Barbet. 2002. Antigenic variation of Anaplasma marginale msp2 occurs by combinatorial gene conversion. Mol. Microbiol. **43**:1151–1159.

Ching, W. M., M. Carl, and G. A. Dasch. 1992. Mapping of monoclonal antibody binding sites on CNBr fragments of the S-layer protein antigens of Rickettsia typhi and Rickettsia prowazekii. Mol. Immunol. **29**:95–105.

Ching, W. M., G. A. Dasch, M. Carl, and M. E. Dobson. 1990. Structural analyses of the 120-kDa serotype protein antigens of typhus group rickettsiae. Comparison with other S-layer proteins. Ann. NY Acad. Sci. **590**:334–351.

Ching, W. M., H. Wang, B. Jan, and G. A. Dasch. 1996. Identification and characterization of epitopes on the 120-kilodalton surface protein antigen of Rickettsia prowazekii with synthetic peptides. Infect. Immun. **64**:1413–1419.

Collins, N. E., J. Liebenberg, E. P. de Villiers et al. (22 coauthors). 2005. The genome of the heartwater agent Ehrlichia ruminantium contains multiple tandem repeats of actively variable copy number. PNAS **102**:838–843.

Crocquet-Valdes, P. A., C. M. Diaz-Montero, H. M. Feng, H. Li, A. D. Barrett, and D. H. Walker. 2001. Immunization with a portion of rickettsial outer membrane protein A stimulates protective immunity against spotted fever rickettsiosis. Vaccine **20**:979–988.

Dasch, G. A. 1981. Isolation of species-specific protein antigens of *Rickettsia typhi* and *Rickettsia prowazekii* for immunodiagnosis and immunoprophylaxis. J. Clin. Microbiol. **14**: 333–341.

Desvaux, M., N. J. Parham, and I. R. Henderson. 2004. The autotransporter secretion system. Res. Microbiol. **155**:53–60.

Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**:113.

Fournier, P. E., V. Roux, and D. Raoult. 1998. Phylogenetic analysis of spotted fever group rickettsiae by study of the outer surface protein rOmpA. Int. J. Syst. Bacteriol. **48**(Pt 3):839–849.

Gilmore, R. D. Jr. 1993. Comparison of the rompA gene repeat regions of rickettsiae reveals species-specific arrangements of individual repeating units. Gene **125**:97–102.

Gilmore, R. D. Jr, W. Cieplak Jr, P. F. Policastro, and T. Hackstadt. 1991. The 120 kilodalton outer membrane protein (rOmp B) of Rickettsia rickettsii is encoded by an unusually long open reading frame: evidence for protein processing from a large precursor. Mol. Microbiol. **5**:2361–2370.

Gilmore, R. D. Jr, and T. Hackstadt. 1991. DNA polymorphism in the conserved 190 kDa antigen gene repeat region among spotted fever group rickettsiae. Biochim. Biophys. Acta. **1097**: 77–80.

Gilmore, R. D. Jr, N. Joste, and G. A. McDonald. 1989. Cloning, expression and sequence analysis of the gene encoding the 120 kD surface-exposed protein of Rickettsia rickettsii. Mol. Microbiol. **3**:1579–1586.

Gravekamp, C., D. S. Horensky, J. L. Michel, and L. C. Madoff. 1996. Variation in repeat number within the alpha C protein of group B streptococci alters antigenicity and protective epitopes. Infect. Immun. **64**:3576–3583.

Hackstadt, T., R. Messer, W. Cieplak, and M. G. Peacock. 1992. Evidence for proteolytic cleavage of the 120-kilodalton outer membrane protein of rickettsiae: identification of an avirulent mutant deficient in processing. Infect. Immun. **60**:159–165.

Henderson, I. R., and J. P. Nataro. 2001. Virulence functions of autotransporter proteins. Infect. Immun. **69**:1231–1243.

Henderson, I. R., F. Navarro-Garcia, M. Desvaux, R. C. Fernandez, and D. Ala'Aldeen. 2004. Type V protein secretion pathway: the autotransporter story. Microbiol. Mol. Biol. Rev. **68**:692–744.

Higgins, D. G., and P. M. Sharp. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene **73**:237–244.

Hultgren, S. J., S. Abraham, M. Caparon, P. Falk, J. W. St Geme III, and S. Normark. 1993. Pilus and nonpilus bacterial adhesins: assembly and function in cell recognition. Cell **73**:887–901.

Jiggins, F. M., G. D. D. Hurst, and Z. Yang. 2002. Host-symbiont conflicts: positive selection on an outer membrane protein of parasitic but not mutualistic Rickettsiaceae. Mol. Biol. Evol. **19**:1341–1349.

Kehoe, M. A. 1994. Cell-wall-associated proteins in gram-positive bacteria. Pp. 217–261 *in* J. M. Ghuysen and R. Hakenbeck, eds. Bacterial cell wall. Elsevier, Amsterdam.

Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305**:567–580.

Kumar, S., K. Tamura, and M. Nei. 1994. MEGA: molecular evolutionary genetics analysis software for microcomputers. Comput. Appl. Biosci. **10**:189–191.

Li, H., and D. H. Walker. 1998. rOmpA is a critical protein for the adhesion of *Rickettsia rickettsii* to host cells. Microb. Pathog. **24**:289–298.

Li, W. H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Lythgoe, K. A., and A. F. Read. 1998. Catching the Red Queen? The advice of the Rose. Trends Ecol. Evol. **13**:473–474.

McDonald, G. A., R. L. Anacker, and K. Garjian. 1987. Cloned gene of *Rickettsia rickettsii* surface antigen: candidate vaccine for Rocky Mountain spotted fever. Science **235**:83–85.

McLeod, M. P., X. Qin, S. E. Karpathy et al. (22 co-authors). 2004. Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other rickettsiae. J. Bacteriol. **186**:5842–5855.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929–936.

Ogata, H., S. Audic, P. Renesto-Audiffren et al. (11 co-authors). 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science **293**:2093–2098.

Palmer, E. L., M. L. Martin, and L. Mallavia. 1974. Ultrastucture of the surface of *Rickettsia prowazeki* and *Rickettsia akari*. Appl. Microbiol. **28**:713–716.

Pearson, W. R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. **183**:63–98.

Peek, A. S., V. Souza, L. E. Eguiarte, and B. S. Gaut. 2001. The interaction of protein structure, selection, and recombination on the evolution of the type-1 fimbrial major subunit (fimA) from *Escherichia coli*. J. Mol. Evol. **52**:193–204.

Perez, J. M., D. Martinez, C. Sheikboudou, F. Jongejan, and A. Bensaid. 1998. Characterization of variable immunodominant antigens of Cowdria ruminantium by ELISA and immunoblots. Parasite Immunol. **20**:613–622.

Roux, V., and D. Raoult. 2000. Phylogenetic analysis of members of the genus *Rickettsia* using the gene encoding the outer-membrane protein rOmpB (ompB). Int. J. Syst. Evol. Microbiol. **50**(Pt 4):1449–1455.

Roux, V., E. Rydkina, M. Eremeeva, and D. Raoult. 1997. Citrate synthase gene comparison, a new tool for phylogenetic analysis, and its application for the rickettsiae. Int. J. Syst. Bacteriol. **47**:252–261.

Sakharkar, K. R., P. K. Dhar, and V. T. K. Chow. 2004. Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. Int. J. Syst. Evol. Microbiol. **54**:1937–1941.

Schuenke, K. W., and D. H. Walker. 1994. Cloning, sequencing, and expression of the gene coding for an antigenic 120-kilodalton protein of *Rickettsia conorii*. Infect. Immun. **62**:904–909.

Schulenburg, J. H., G. D. Hurst, T. M. Huigens, M. M. van Meer, F. M. Jiggins, and M. E. Majerus. 2000. Molecular evolution and phylogenetic utility of Wolbachia ftsZ and wsp gene sequences with special reference to the origin of male-killing. Mol. Biol. Evol. **17**:584–600.

Sekeyova, Z., V. Roux, and D. Raoult. 2001. Phylogeny of Rickettsia spp. inferred by comparing sequences of 'gene D', which encodes an intracytoplasmic protein. Int. J. Syst. Evol. Microbiol. **51**:1353–1360.

Sonnhammer, E. L., and R. Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene **167**:GC1–GC10.

Stenos, J., V. Roux, D. Walker, and D. Raoult. 1998. Rickettsia honei sp. nov., the aetiological agent of Flinders Island spotted fever in Australia. Int. J. Syst. Bacteriol. **48**(Pt 4):1399–1404.

Stevenson, B., and J. C. Miller. 2003. Intra- and interbacterial genetic exchange of Lyme disease spirochete erp genes generates sequence identity amidst diversity. J. Mol. Evol. **57**:309–324.

Stothard, D. R., J. B. Clark, and P. A. Fuerst. 1994. Ancestral divergence of Rickettsia bellii from the spotted fever and typhus groups of Rickettsia and antiquity of the genus Rickettsia. Int. J. Syst. Bacteriol. **44**:798–804.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10**: 512–526.

Templeton, T. J. 2004. Borrelia outer membrane surface proteins and transmission through the tick. J. Exp. Med. **199**:603–606.

Tewari, R., J. I. MacGregor, T. Ikeda, J. R. Little, S. J. Hultgren, and S. N. Abraham. 1993. Neutrophil activation by nascent FimH subunits of type 1 fimbriae purified from the periplasm of *Escherichia coli*. J. Biol. Chem. **268**:3009–3015.

Teysseire, N., and D. Raoult. 1992. Comparison of western immunoblotting and microimmunofluorescence for diagnosis of Mediterranean spotted fever. J. Clin. Microbiol. **30**:455–460.

Uchiyama, T. 1997. Intracytoplasmic localization of antigenic heat-stable 120- to 130-kilodalton proteins (PS120) common to spotted fever group rickettsiae demonstrated by immunoelectron microscopy. Microbiol. Immunol. **41**:815–818.

———. 2003. Adherence to and invasion of Vero cells by recombinant Escherichia coli expressing the outer membrane protein rOmpB of Rickettsia japonica. Ann. NY Acad. Sci. **990**:585–590.

Uchiyama, T., L. Zhao, and T. Uchida. 1996. Demonstration of a heat-stable 120-kilodalton protein of Rickettsia japonica as a spotted fever group-common antigen. Microbiol. Immunol. **40**:133–139.

Vishwanath, S., G. A. McDonald, and N. G. Watkins. 1990. A recombinant Rickettsia conorii vaccine protects guinea pigs from experimental boutonneuse fever and Rocky Mountain spotted fever. Infect. Immun. **58**:646–653.

Wizemann, T. M., J. E. Adamou, and S. Langermann. 1999. Adhesins as targets for vaccine development. Emerg. Infect. Dis. **5**:395–403.

Wren, B. W. 1991. A family of clostridial and streptococcal ligand-binding proteins with conserved C-terminal repeat sequences. Mol. Microbiol. **5**:797–803.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.

Yang, Z., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15**:496–503.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Yen, M. R., C. R. Peabody, S. M. Partovi, Y. Zhai, Y. H. Tseng, and M. H. Saier. 2002. Protein-translocating outer membrane porins of Gram-negative bacteria. Biochim. Biophys. Acta **1562**:6–31.

Zhang, J. R., J. M. Hardham, A. G. Barbour, and S. J. Norris. 1997. Antigenic variation in Lyme disease borreliae by promiscuous recombination of VMP-like sequence cassettes. Cell **89**:275–285.

Zhang, J.-R., and S. J. Norris. 1998. Genetic variation of the Borrelia burgdorferi gene vlsE involves cassette-specific, segmental gene conversion. Infect. Immun. **66**:3698–3704.

Pekka Pamilo, Associate Editor

Accepted June 14, 2005