Genome Degradation is an Ongoing Process in *Rickettsia*

Jan O. Andersson and Siv G. E. Andersson

Department of Molecular Evolution, Uppsala University, Uppsala, Sweden

To study reductive evolutionary processes in bacterial genomes, we examine sequences in the *Rickettsia* genomes which are unconstrained by selection and evolve as pseudogenes, one of which is the *metK* gene, which codes for AdoMet synthetase. Here, we sequenced the *metK* gene and three surrounding genes in eight different species of the genus *Rickettsia.* The *metK* gene was found to contain a high incidence of deletions in six lineages, while the three genes in its surroundings were functionally conserved in all eight lineages. A more drastic example of gene degradation was identified in the *metK* downstream region, which contained an open reading frame in *Rickettsia felis.* Remnants of this open reading frame could be reconstructed in five additional species by eliminating sites of frameshift mutations and termination codons. A detailed examination of the two reconstructed genes revealed that deletions strongly predominate over insertions and that there is a strong transition bias for point mutations which is coupled to an excess of GC-to-AT substitutions. Since the molecular evolution of these inactive genes should reflect the rates and patterns of neutral mutations, our results strongly suggest that there is a high spontaneous rate of deletions as well as a strong mutation bias toward AT pairs in the *Rickettsia* genomes. This may explain the low genomic G+C content (29%), the small genome size (1.1 Mb), and the high noncoding content (24%), as well as the presence of several pseudogenes in the *Rickettsia prowazekii* genome.

Introduction

In order to understand the evolution of genome sizes, it is necessary to determine the rates at which DNA rearrangement and insertion/deletion events occur. In eukaryotes, studies of neutrally evolving sequences, such as pseudogenes and non-LTR retrotransposable elements, have proved to be very valuable for our understanding of the way in which deletions and insertions accumulate (Ophir and Graur 1997; Petrov and Hartl 1997, 1998). Unfortunately, bacterial genomes typically have only 10% noncoding DNA, and until recently, very few nonfunctional genes had been identified (Delorme et al. 1993; Godon et al. 1993). The complete genome sequence of *Rickettsia prowazekii* has revealed an astonishingly high content of noncoding DNA (24%) and has led to the identification of about a dozen pseudogenes (Andersson et al. 1998). It is important to establish the properties of such pseudogenes in order to understand the evolutionary processes whereby bacterial genomes deteriorate.

The *Rickettsia* belong to the α-proteobacteria, and they are obligate intracellular parasites, normally associated with arthropods but often pathogenic for humans (Raoult and Roux 1997). The genus *Rickettsia* can be divided into two major groups; the typhus group (TG), which includes the etiologic agents of epidemic and murine typhus, *R. prowazekii* and *Rickettsia typhi,* and the spotted fever group (SFG), which includes the etiologic agent of Rocky Mountain spotted fever, *Rickettsia rickettsii.*

Epidemic typhus was once a very common infectious disease, widely distributed among both civilians and soldiers. The infection is acquired and transmitted among humans through the body louse *Pediculus humanus corporis* (Brezina et al. 1973; Weiss 1982). Today, its transmission is limited due to the reduced level of body louse infestations, although infections caused by *R. prowazekii* are still considered to have a potential epidemic danger throughout the world. Murine typhus, which is caused by *R. typhi,* is clinically similar to, but milder than, epidemic typhus. Rodents are the principle reservoirs for this species, and the rat flea is the chief vector (Traub, Wisseman, and Farhang-Azad 1978; Azad 1988).

The SFG *Rickettsia* are established in hard body ticks and their animal hosts. Humans are considered only an incidental host, although as many as 10 isolates are known to be human pathogens. The SFG *Rickettsia* normally multiply in the cytoplasm, but some species are also capable of penetrating the host cell nucleus (Wisseman et al. 1976). *Rickettsia felis* has been classified with the SFG *Rickettsia* based on molecular sequence data, although this species causes a murine typhus-like disease in humans and is stable in insect hosts like *R. prowazekii* and *R. typhi* (Adams, Schmidtmann, and Azad 1990; Williams et al. 1992).

The *Rickettsia* are characterized by very small genomes, in the range of 1.1–1.4 Mb (table 2). These are thought to have evolved from an ancestral genome of larger size by reductive evolutionary processes (Andersson and Kurland 1995, 1998). The size of a bacterial genome reflects a balance between the rate at which genes are being lost and the rate at which novel genes are being acquired. The loss of genes in *R. prowazekii* has been shown to be highly biased, such that the proportion of genes responsible for the biosynthesis of compounds that are also present in the host cell cytoplasm is much lower than expected from free-living bacteria (Andersson et al. 1996, 1998). We have speculated that the high fraction of noncoding DNA in the *R. prowazekii* genome may be degraded gene remnants, caught in the final stages of elimination (Andersson et al. 1998).

Genome degradation has previously been described for the chloroplast genome of *Epifagus virigiana,* a plant which lives as a parasite on the roots of beech trees

Key words: deletions, insertions, *metK,* molecular evolution, pseudogenes, *Rickettsia.*

Address for correspondence and reprints: Siv G. E. Andersson, Department of Molecular Evolution, Box 590, Biomedical Center, S-751 24 Uppsala, Sweden. E-mail: siv.andersson@molbio.uu.se.

Mol. Biol. Evol. 16(9):1178–1191. 1999

(Wolfe, Morden, and Palmer 1992). This organism has lost its photosynthetic capacity along with its functional genes for photosynthesis and chlororespiration. The genome contains a multitude of pseudogenes, including six highly truncated photosynthetic genes. However, the genome seems to be functional, since ribosomal RNA genes and several ribosomal protein genes, as well as a few genes coding for other functions, are intact (Wolfe, Morden, and Palmer 1992).

Photosynthesis appears to have been lost many times independently in this group of parasitic plants (dePamphilis, Young, and Wolfe 1997). Among the nonphotosynthetic lineages, the photosynthetic gene *rbcL* has been retained in some lineages, while it has been mutationally degraded or completely eliminated in other lineages (dePamphilis, Young, and Wolfe 1997). In contrast, the translational gene *rps2* has been retained in all lineages, although it has been subjected to accelerated mutation rates in several of these (dePamphilis, Young, and Wolfe 1997). Thus, the transitions to parasitic, nonphotosynthetic lifestyles appear to be associated with drastic increases in the fixation rates for deletion and substitution mutations in these chloroplast genomes.

Accelerated rates of nucleotide substitution have also been observed in *Buchnera,* which are obligate intracellular symbionts of aphids (Baumann et al. 1995; Moran 1996). Species of the genus *Buchnera* are phylogenetically affiliated with the γ-proteobacteria, but they evolve more than 10 times as fast as their free-living relatives *Escherichia coli* and *Salmonella typhimurium* (Brynnel et al. 1998). This substitution rate enhancement appears not to be caused by differences in the intrinsic mutation rate, but to be due to an increased rate of fixation of mutations (Brynnel et al. 1998). This has been explained by Muller's ratchet, which postulates that small populations with low levels of recombination will gradually accumulate deleterious mutations, which may lead to the loss of the most fit class of genomes in the population by genetic drift (Muller 1964; Felsenstein 1974).

We have previously shown that the gene *metK,* which codes for AdoMet synthetase, contains an internal termination codon (TAG) in the Madrid E strain of *R. prowazekii* (Andersson and Andersson 1997). The gene is otherwise very well conserved and exhibits the statistical nucleotide properties expected for a gene that is functional (Andersson and Andersson 1997). AdoMet synthetase, which catalyzes the synthesis of AdoMet from methionine and ATP, is an essential enzyme in bacteria as well as in eukaryotes. This is not surprising, since AdoMet is used as methyl donor in a large number of cellular reactions such as DNA and RNA methylations, protein side-chain modifications, and biosynthesis of amines, creatine, and carnitine (Poirier, Zapisek, and Lyn-Cook 1990; Fujioka 1992). The *metK* gene is a multicopy gene in most eukaryotes, while it is present as a single-copy gene in bacterial genomes. Indeed, there is also only one single copy of the *metK* gene in *R. prowazekii* (Andersson and Andersson 1997).

In order to backtrack through the evolutionary history of the *metK* gene and identify the point at which the termination codon originated, we sequenced the *metK* gene and its surrounding genes in several *Rickettsia* species. Here, we present a detailed analysis of the patterns of deletion, insertion, and substitution mutations in the *metK* gene and in another nearby pseudogene in several members of the TG and SFG *Rickettsia.*

Materials and Methods

DNA Preparation

Propagation of *R. prowazekii* strain Madrid E in the yolk sacs of embryonated hen eggs and preparation of genomic DNA were done according to standard procedures (Winkler 1976). Genomic DNA from *R. prowazekii* strain B, *R. typhi* strain Wilmington, *R. rickettsii* strain 84-21C, *Rickettsia sibirica, Rickettsia montana, Rickettsia rhipicephali, Rickettsia helvetica,* and *R. felis,* which were generous gifts from D. Stothard, A. Azad, and N. Balayeva, were prepared as previously described (Pretzman et al. 1987).

PCR and Sequencing

Genomic DNA was amplified with PCR. Degenerate primers were designed against conserved gene regions in *R. prowazekii* and other rickettsial species (table 1). Each PCR product was first sequenced (see below) with the degenerate primers. The sequence produced allowed the design of species-specific primers, and the whole region was then sequenced by primer walking with the PCR products used as templates. In a typical PCR reaction, the following were used: thermophilic DNA polymerase buffer (Promega), 1.5 mM $MgCl_2$, 0.2 mM each of dATP, dCTP, dGTP, and dTTP (Pharmacia Biotech), 1–2 μM of each primer, 1 U of *Taq* DNA polymerase (Promega), and 1–5 ng of genomic DNA. Amplification was performed in a thermal cycler beginning with 5 min at 95°C, then 30 cycles of 94°C for 1 min, 52°C for 1 min, and 72°C for 2 min. For those degenerate primers not yielding a PCR product when an annealing temperature of 52°C was used, the annealing temperature was gradually lowered until the desired product was obtained.

After thermocycling, aliquots of the PCR reactions were run on an 1% agarose gel. If a single band was seen, the PCR product was purified using the QIAquick PCR Purification Kit (Qiagen). If additional bands were seen, the whole PCR reaction was loaded on a new 1% agarose gel, and the correct bands were excised and purified with the QIAquick Gel Extraction kit (Qiagen).

The purified PCR products were sequenced using the ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction Kit (Perkin-Elmer). The sequencing reactions were purified using Centi-Sep columns (Princeton Separations) and analyzed on an ABI PRISM 377 DNA Sequencer (Perkin-Elmer).

The PCR product from the intergenic region upstream of *dnaE* in *R. helvetica* was sequenced using the Thermo Sequenase Fluorescent Labeled Primer Cycle Sequencing Kit (Amersham) and analyzed on an A.L.F. Sequencer (Pharmacia Biotech) due to failure of the Dye Terminator chemistry, probably caused by a strong secondary structure in the region.

Table 1
Degenerate Primers Used for PCR Amplification of the *metK* Gene and Flanking Sequences in *Rickettsia* Species

Primer	Sequence[a]	Direction[b]	Gene	Position[c]
ftsYfor	GC(A,T) (G,T)GT GAT AC(A,C,T) TTT AGA GC	+	*ftsY*	406–425
polAforB	GGT TAT GAA GC(A,C,T) GAT GAT AT	+	*polA*	325–344
polArevB	ACT TCA CGT AAT TTA TCA G	−	*polA*	512–530
polAfor	AGA CT(A,T) AT(A,T) (C,T)T(A,T) CAA ATT CAT GAT GA	+	*polA*	2440–2465
polArev	ATT TC(A,G,T) GTT ATA AT(A,C,T) GGT AC	−	*polA*	2551–2570
metXfor	TC(G,T) GTT TC(C,T) GAA GG(G,T) CAT CC	+	*metK*	28–47
metXrev	CAA TC(C,G,T) GC(A,T) ACA GGA CC(A,T) CC	−	*metK*	700–719
metXrevB	TTC (A,G)CG ACC (A,G)AA ATG (A,T)CC GTA	−	*metK*	1072–1092
Lcrev71	GCT AAA GCT ACA ATC TAT TTA C	−	*dnaE*	3427–3448
dErevC	TGC A(A,G)(C,T) TCA TCA AA(A,G) CA(A,C,G,T) CC	+	*dnaE*	2761–2780
dErev	GTC ACA CC(C,T) TT(A,G,T) AT(A,T) GC(A,C,G,T) CC	+	*dnaE*	2611–2630
dEforB	TTA GG(A,C,G,T) GCT ATT AAG GG(A,C,G,T) GT	−	*dnaE*	2608–2627
dnaE	GT(A,T) GC(A,T) AAA TTT GC(A,T) GG(A,T) TAT GG	−	*dnaE*	2239–2261

[a] The degenerate nucleotides are given in parentheses.
[b] Primers of the sense orientation are indicated with "+," and primers of the antisense orientation are indicated with "−."
[c] Nucleotide position in the *R. prowazekii* genes.

Sequence Analysis

The analyzed sequenced reactions were collected for each species and assembled using the Staden Package (Staden 1996). Searches for open reading frames (ORFs) were performed using XBBTOOLS (T. Sicheritz-Pontén, personal communication), and homology searches within the data set were performed using the BLAST program, version 2.0a17 (Altschul et al. 1990), within the XBBTOOLS interface. Sequence alignments were performed using CLUSTAL W (Thompson, Higgins, and Gibson 1994). Base frequencies and codon usage statistics were calculated using CODONW (Lloyd and Sharp 1992). Percentages of amino acid identity were calculated using ALIGN (Myers and Miller 1988). Synonymous and nonsynonymous distances were calculated using Li's (1993) method and the program MATDISLI.

In order to facilitate analysis of the *metK* regions in species that contained deletions, insertions, or in-frame stop codons, the alignments were corrected for the species-specific mutations in the following way: The *metK* genes were aligned using CLUSTAL W, with the complete ORF in *R. typhi* used as a template for the alignment. All codons in the alignment that were affected by deletions or in-frame stop codons in one or more species were deleted from all species in the data set using the interface program SEAVIEW (Galtier, Gouy, and Gautier 1996). Insertions were deleted in those species that contained them. The resulting data set was used for calculations of G+C statistics, amino acid identity values, synonymous and nonsynonymous substitution rates and for mutational pattern analysis.

The reconstruction of the second inactivated gene (the *fossil* gene) was performed in a similar manner, with the ORF in *R. felis* used as a representative of the intact gene.

Phylogenetic Analysis

Due to ambiguous alignments, one region of *polA* and one region of *dnaE* were excluded from the phylogenetic analysis. They corresponded to amino acids 430–435 of *polA* and amino acids 828–865 of the full-length *dnaE* of *R. prowazekii*.

The neighbor-joining method (Saitou and Nei 1987) was used with Poisson corrections for amino acid

Table 2
Sizes, Nucleotide Frequencies, and Coding Contents of the Amplified *metK* Gene Fragments in *Rickettsia* Species

	Genomic DNA		Fragment		Coding DNA			Fossil Gene DNA			Noncoding DNA		
Species	L_G[a]	GC_G[b]	L_T[c]	GC_T[d]	L_C[c]	GC_C[d]	F_C[e]	L_F[c]	GC_F[d]	F_F[e]	L_N[c]	GC_N[d]	F_N[e]
R. prowazekii, E ...	1.10	29.3	6,040	30.3	4,101	30.1	0.68	1,143	35.9	0.19	796	23.4	0.13
R. prowazekii, B ...	1.10	29.0	6,040	30.3	5,244	31.3	0.87	—	—	—	796	23.4	0.13
R. typhi	1.13	29.0	7,235	29.8	5,259	31.3	0.73	—	—	—	1,976	25.7	0.27
R. rickettsii	1.27	32.6	6,865	33.4	4,122	33.2	0.60	2,060	33.1	0.30	683	35.4	0.10
R. sibirica	1.24	32.5	6,895	33.7	4,122	33.3	0.60	2,084	33.5	0.30	689	36.1	0.10
R. montana	1.25	ND[f]	8,174	33.1	4,122	33.3	0.50	3,422	31.7	0.42	630	38.7	0.08
R. rhipicephali	1.26	32.2	7,988	32.8	4,125	33.3	0.52	3,306	31.4	0.41	560	37.5	0.07
R. helvetica	1.40	ND[f]	8,033	31.9	4,593	32.9	0.57	2,640	32.1	0.33	800	30.0	0.10
R. felis	ND[f]	ND[f]	9,007	33.7	7,014	32.5	0.78	1,143	38.8	0.12	875	36.6	0.10

[a] L_G = estimated genome size in Mb (Eremeeva, Roux, and Raoult 1993; Roux and Raoult 1993).
[b] GC_G = estimated genomic G+C content (Tyeryar et al. 1973; Schramek 1974).
[c] L_T, L_C, L_F, L_N = sizes (in nucleotides) of whole fragment, coding DNA, fossil gene DNA, and noncoding DNA, respectively.
[d] GC_T, GC_C, GC_F, GC_N = G+C contents of whole fragment, coding DNA, fossil gene DNA, and noncoding DNA, respectively.
[e] F_C, F_F, F_N = fractions of coding DNA, fossil gene DNA, and noncoding DNA, respectively.
[f] Data not available.

distances within the PHYLO_WIN program (Galtier, Gouy, and Gautier 1996). Maximum parsimony analysis was performed within the same program. The level of statistical support for the trees was estimated by performing 500 bootstrap replicates for each method. Maximum likelihood analysis was performed using the PUZZLE program (Strimmer and von Haeseler 1997), with 500 puzzling steps. Global gap removal was used in all methods.

The Replacement Pattern for Universally Conserved Amino Acids

The *ftsY, polA,* and *dnaE* amino acid sequences from *E. coli, Haemophilus influenzae, Helicobacter pylori, Synechocystis* sp., *Bacillus subtilis,* and *Mycobacterium tuberculosis* were aligned and treated as a reference data set. The amino acid sequences of *ftsY, polA,* and *dnaE* from this study were added to this alignment, and the numbers of amino acids that were conserved within the non-*Rickettsia* species and varied in one or more of the *Rickettsia* species were counted and placed in the phylogenetic tree. The same calculation was done for *metK,* for which the disrupted *Rickettsia* sequences were made translatable by removal of all insertions and all nucleotides in codons affected by deletions or in-frame stop codons in the individual species.

Accession Numbers

The SwissProt accession numbers or GenPept sequence ID numbers for the amino acid sequences used in amino acid conservation and phylogenetic analyses are: for *ftsY*— *E. coli,* P10121; *H. influenzae,* P44870; *H. pylori,* 2313887; *Synechocystis* sp., 1653079; *B. subtilis,* P51835; *M. tuberculosis,* Q10969; for *polA*—*E. coli,* P00582; *H. influenzae,* P42741; *H. pylori,* P56105; *Synechocystis* sp., 1006595; *B. subtilis,* 2293272; *M. tuberculosis,* Q07700; for *hicB*—*H. influenzae* strain R3001, 3603326; for *dnaE*—*E. coli,* P10443; *H. influenzae,* P43743; *H. pylori,* P56157; *Synechocystis* sp., 1653960; *B. subtilis,* 2293260; *M. tuberculosis,* Q10779; and for *metK*—*E. coli,* P04384; *H. influenzae,* P43763; *H. pylori,* 2313288; *Synechocystis* sp., 1651961; *B. subtilis,* 2293164; *M. tuberculosis,* P77899; *Staphylococcus aureus,* P50307; *Treponema pallidum,* 3233100; *Pisum sativum,* P49613; *Homo sapiens,* P31153; *Saccharomyces cerevisiae,* P19358.

The EMBL accession numbers for the sequences reported in this paper are: *R. prowazekii* strain Madrid E, AJ238755; *R. prowazekii* strain B, AJ238756; *R. rickettsii,* AJ238758; *R. sibirica,* AJ238759; *R. montana,* AJ238760; *R. rhipicephali,* AJ238761; *R. helvetica,* AJ238762; *R. felis,* AJ23875863; *R. typhi,* AJ238757.

Results

The *metK* Gene and Flanking Sequences in *Rickettsia*

The *metK* gene has been identified in all bacterial genomes sequenced so far, with the exception of *Chlamydia trachomatis* (Stephens et al. 1998). We have previously shown that the *metK* gene in the Madrid E strain of *R. prowazekii* contains a termination codon in a region of the gene that is otherwise very highly conserved among bacterial species (Andersson and Andersson 1997). In order to study the occurrence of mutations in the *metK* genes of other *Rickettsia* species, we sequenced the *metK* gene and its surrounding genes in one other strain of *R. prowazekii* and in seven additional rickettsial species. We have identified the gene *polA,* which codes for DNA polymerase I, immediately upstream of *metK* in the *R. prowazekii* genome (Andersson and Andersson 1997). These two genes are flanked by the genes *ftsY* and *dnaE,* which we used as anchors for PCR reactions designed to amplify the *ftsY-polA-metK-dnaE* segment in other species of the genus *Rickettsia* (table 1).

When genomic DNA from the B and Madrid E strains of *R. prowazekii* was used in the PCR reactions, a region of 6,040 bp was amplified. Successful PCR amplifications were also obtained with genomic DNA from *R. typhi, R. rickettsii, R. sibirica, R. montana, R. rhipicephali, R. helvetica,* and *R. felis.* However, there was a marked size variation of the amplified PCR fragments, ranging from 6,040 bp in *R. prowazekii* to 9,007 bp in *R. felis* (table 2). The overall G+C content values of these regions span from 29.8% in *R. typhi* to 33.7% in *R. sibirica* and *R. felis.* These values are in good agreement with previous estimates of the genomic G+C contents for these species (table 2).

Homologs to *ftsY, polA, metK,* and *dnaE* were identified in all species, and these are arranged in the same order and direction as in *R. prowazekii* (fig. 1). In addition, homologs to *hicB,* which is located within the major pilus gene cluster of *H. influenzae* (Mhlanga-Mutangadura et al. 1998), were observed in *R. felis* and *R. helvetica.* The reading frames for the genes *ftsY, polA, hicB,* and *dnaE* were open, while the *metK* genes were found to be translationally disrupted in all but two species (see below). Finally, an ORF (*orf*) with no similarity to genes in the public databases was observed in *R. felis* (fig. 1). Base composition patterns for *ftsY, polA, hicB, dnaE,* and the ORF in *R. felis* were similar to those of genes that are known to be functionally constrained, i.e., the overall genic G+C content was significantly higher than the G+C content at third synonymous codon positions (table 3) (Andersson and Sharp 1996; Andersson and Andersson 1997). We observed small but marked differences between members of the TG *Rickettsia* and the SFG *Rickettsia* with respect to nucleotide frequencies at synonymous third codon positions ($GC3_S$). Within the TG *Rickettsia,* the genes *ftsY, polA,* and *dnaE* have $GC3_S$ a mean value of 17.3%, whereas the mean $GC3_S$ value of the same genes within the SFG *Rickettsia* is 23.8% (table 3).

Mutational Degradation of the *metK* Genes

Alignments of the *metK* gene sequences showed that *R. typhi* and the B strain of *R. prowazekii* contained complete ORFs with no observable defects, while the *metK* genes in all other species contained in-frame termination codons and/or frameshift mutations (fig. 2*B*). Two species, *R. sibirica* and *R. rhipicephali,* contained termination codons in the middle region of the *metK* gene. For these and other members of the SFG *Rickett-*

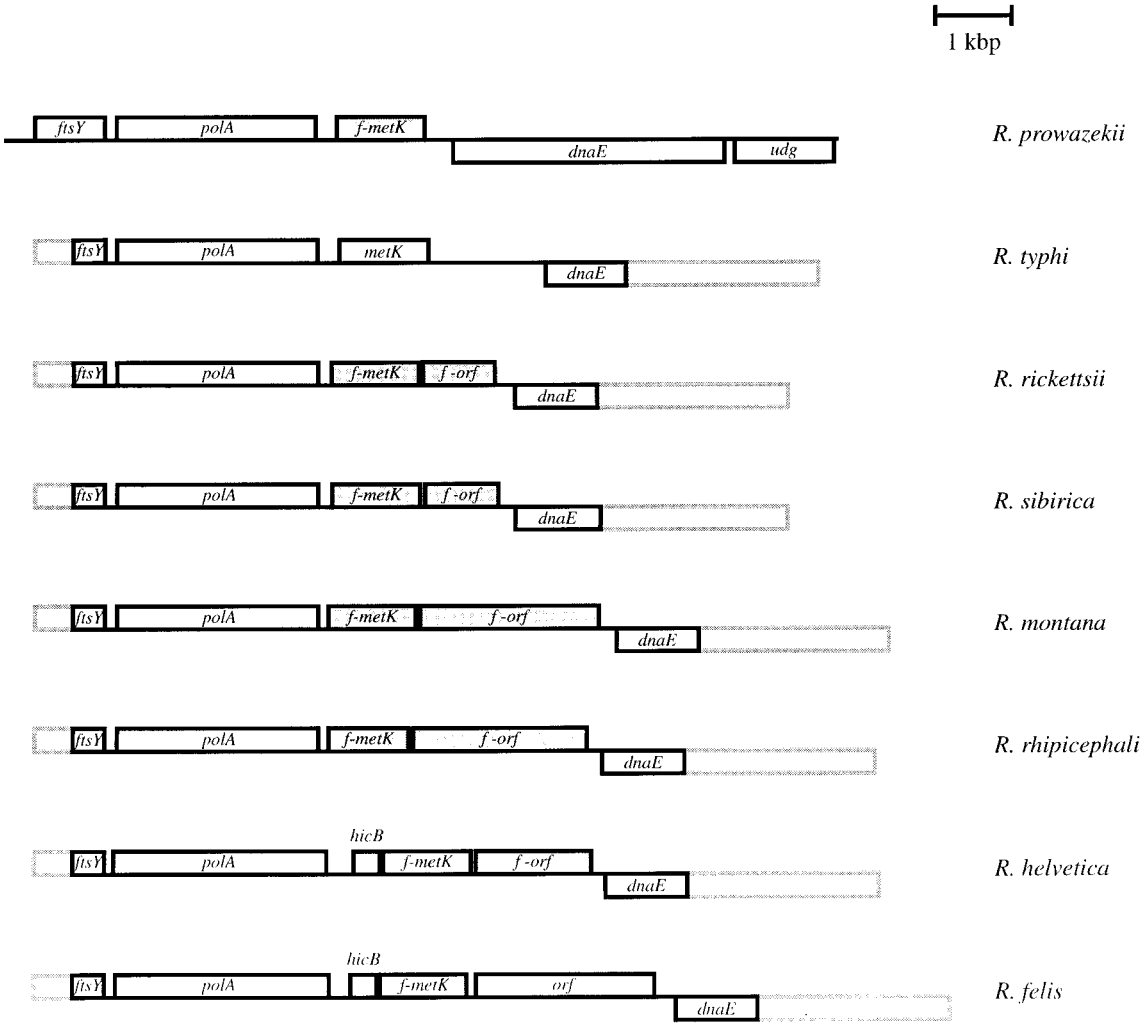

FIG. 1.—Schematic representations of the organization of the *metK* gene region in *Rickettsia*. Thin lines indicate noncoding regions, open boxes indicate coding regions, and shaded boxes indicate pseudogenes. The open boxes with shaded margins represent regions that are inferred to exist but were not sequenced.

sia, numerous short insertion/deletion events were observed, ranging in size from 1 to 96 bp (fig. 2*B*). The *metK* gene sequences of the B and Madrid E strains of *R. prowazekii* were found to be identical, with the exception of a single nucleotide that creates a termination codon in the *metK* gene of the Madrid E strain.

To facilitate further detailed analysis, we reconstructed the original ORFs by combining the regions surrounding the frameshift mutations in the *metK* genes of the SFG *Rickettsia*. These reconstructed *metK* genes were found to have $GC3_S$ values that were indistinguishable from the values of those with no mutational defects (26.0% vs. 23.9% on average) (tables 3 and 4). The reconstructed proteins showed their highest levels of sequence similarity to AdoMet synthetase from *E. coli*, with identity values ranging from 63.2% for *R. rickettsii* to 68.1% for *R. helvetica* (table 4).

Fossil Genes Near Extinction

We observed a surprisingly large size variation of the intergenic sequences between the genes *metK* and *dnaE*, ranging from 366 bp in *R. prowazekii* to 3,018 bp in *R. felis*. The 10-fold longer intergenic sequence in *R. felis* is partially explained by a 2,331-bp-long ORF with no sequence similarity to genes present in the public databases (fig. 1). Furthermore, the other members of the SFG *Rickettsia* were found to contain fragmented remnants of the ORF identified in *R. felis*. However, in these species, numerous insertion, deletion, and point mutations have blurred the traces of the original gene. Nevertheless, translatable ORFs could be restored also in these species by assuming totals of 6, 6, 7, 9, and 12 mutational events for *R. sibirica, R. montana, R. helvetica, R. rickettsii,* and *R. rhipicephali*, respectively (fig. 2*C*). The reconstructed *fossil* genes have $GC3_S$ values similar to those of the complete ORF in *R. felis* (16.8% on the average vs. 19.3% for *R. felis*) (table 4), suggesting that these sequences indeed once represented functional genes. Below, we will refer to these inactivated genes in the SFG *Rickettsia* as the *fossil* genes.

No obvious homology could be seen between the DNA sequences of the *metK-dnaE* intergenic region of *R. felis* and the *metK-dnaE* intergenic regions of *R. typhi* and *R. prowazekii*. However, when the putative gene

Table 3
Nucleotide Frequency Statistics of the *ftsY, polA, hicB,* and *dnaE* Genes in *Rickettsia* Species

Gene	Species	Codons	GC_C[a]	$GC3_S$[b]	%ID[c]
ftsY	*R. prowazekii*	136	0.299	0.141	47.8
	R. typhi	136	0.306	0.163	49.3
	R. rickettsii	136	0.358	0.306	48.6
	R. sibirica	136	0.363	0.321	49.3
	R. montana	136	0.365	0.308	49.3
	R. rhipicephali	136	0.355	0.299	49.3
	R. helvetica	136	0.368	0.328	49.3
	R. felis	136	0.365	0.306	49.3
polA ...	*R. prowazekii*	867	0.298	0.168	35.1
	R. typhi	872	0.299	0.180	35.1
	R. rickettsii	875	0.323	0.215	36.2
	R. sibirica	875	0.325	0.213	35.5
	R. montana	875	0.324	0.217	36.6
	R. rhipicephali	875	0.325	0.218	36.1
	R. helvetica	921	0.329	0.212	35.3
	R. felis	922	0.345	0.250	35.9
hicB ...	*R. helvetica*	113	0.280	0.162	41.7
	R. felis	113	0.277	0.145	42.9
dnaE ...	*R. prowazekii*	361	0.308	0.184	21.4
	R. typhi	359	0.300	0.174	21.3
	R. rickettsii	360	0.343	0.251	23.3
	R. sibirica	360	0.344	0.257	22.5
	R. montana	360	0.344	0.254	22.5
	R. rhipicephali	360	0.346	0.257	22.8
	R. helvetica	357	0.334	0.237	21.6
	R. felis	386	0.351	0.259	22.6

[a] Gene G+C content.
[b] G+C content at silent third codon positions.
[c] Percentage of identity to the homologous genes in *Escherichia coli.*

product of the ORF in *R. felis* was used to search for sequence similarities with the tBLASTx algorithm, several weak homologies to the intergenic regions of *R. typhi* and *R. prowazekii* were observed.

Patterns of Deletions and Insertions

In order to place the mutational degradation of the *metK* gene and the *fossil* genes in an evolutionary context, we performed a phylogenetic analysis based on the combined sequences of the surrounding *ftsY, polA,* and *dnaE* genes. Neighbor-joining, maximum parsimony and maximum likelihood methods produced trees with identical topologies, although with slightly different degrees of statistical support for the internal nodes (fig. 2*A*). This topology was in good agreement with previous phylogenetic reconstructions based on the 16S and 23S rRNA and the citrate synthase gene sequences (Stothard 1995; Raoult and Roux 1997; Roux et al. 1997; Andersson et al. 1999). For example, *R. rickettsii* and *R. sibirica* were found to form a very tight cluster within the SFG *Rickettsia,* as previously observed. Similarly, we identified *R. felis* as an early- diverging species within the SFG *Rickettsia,* in accordance with previous phylogenetic analyses (Stothard 1995).

Insertions can be distinguished from deletions by superimposing the observed mutations onto the reconstructed phylogeny and assuming that the mutational events which correspond to the smallest number of changes represent the actual events. This method has successfully placed all frameshift mutations in the *metK* gene and the *fossil* gene on the individual branches of the phylogenetic tree and unambiguously identified them as either deletions or insertions (fig. 2*B* and *C*).

A closer inspection of the patterns of changes showed that deletions heavily dominate over insertion events. In total, we identified 32 deletion events from 1 to 1,354 bp, and only 9 insertion events of 1–2 bp each (fig. 2*B* and *C*). Two deletions, of 1 and 10 bp, in the *metK* gene were located at identical positions in *R. rickettsii* and *R. sibirica,* suggesting that these deletions occurred prior to the divergence of the two species. All other deletions in the *metK* gene were unique for each individual lineage, suggesting that mutational degradation of the *metK* gene has proceeded independently in most lineages.

A larger fraction of deletions were found to be shared by two or more species in the *fossil* gene sequences, including a 23- bp deletion in *R. rickettsii, R. sibirica,* and *R. rhipicephali,* as well as a termination codon and a deletion of 1,354 bp in *R. rickettsii* and *R. sibirica.* Also for this pseudogene, the patterns of deletions and insertions indicate that gene deterioration is an ongoing process. The shared 1-bp deletion in *R. rickettsii, R. sibirica, R. montana,* and *R. rhipicephali* is likely to represent one of the earliest mutational events in the *fossil* gene.

The unique 24-bp deletion in the *fossil* gene of *R. rhipicephali* was found to be flanked by a perfect 8-bp direct repeat in all other species examined (data not shown), suggesting that the deletion has been generated by replication slippage. For the other 31 deletion events, no obvious patterns in the surrounding sequences could be detected (data not shown).

Synonymous and Nonsynonymous Substitutions

Since insertion/deletions and nucleotide substitutions are expected to start accumulating in a random manner upon gene inactivation, we expected to find a correlation between the fixation rate for deletion/insertion mutations and that for nucleotide substitutions. In order to compare these rates, we estimated the relative fixation rate for mutations at different codon sites by dividing the fraction of substitutions that cause amino acid replacements (K_a) with the fraction of substitutions at synonymous sites (K_s). Provided that a gene has the same function in all species and that there are no selective constraints acting on synonymous sites, this ratio should be a constant factor for each pair of genes that are compared.

Indeed, the estimated K_a/K_s ratios for genes such as *polA* and *dnaE* were approximately similar for all pairwise comparisons within the TG and the SFG *Rickettsia* (0.24 vs. 0.19 on average) (fig. 3). The *metK* gene, on the other hand, yielded a twofold higher ratio for members within the SFG *Rickettsia* (0.18) than for members within the TG *Rickettsia* (0.08) (fig. 3). However, the frequency of nonsynonymous substitutions was very small for closely related pairs, so these ratios may be affected by statistical fluctuations. Nevertheless, the results indicate that the *metK* gene was inactivated more recently in the TG *Rickettsia* than in the SFG *Rickettsia,* which is also consistent with the higher frequency of

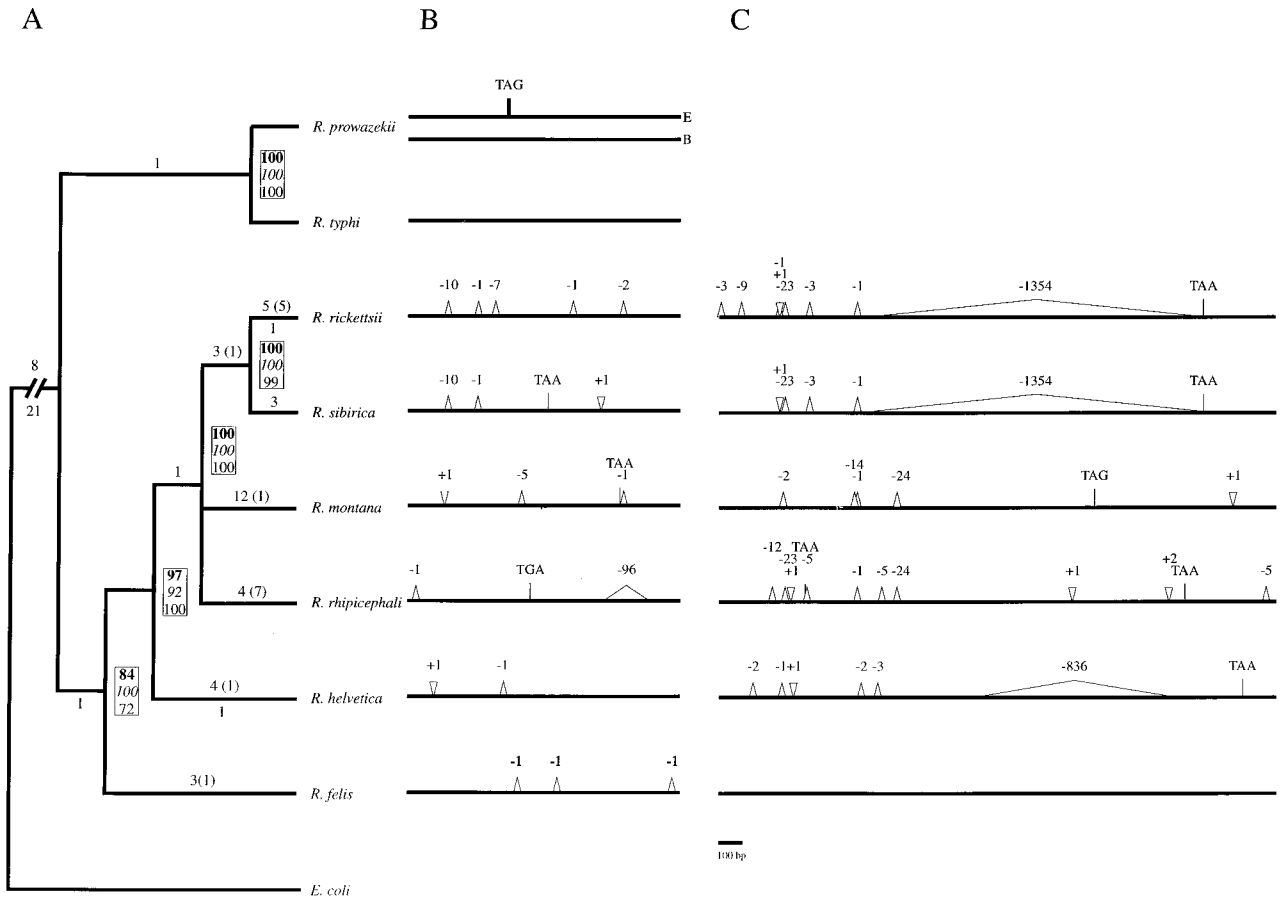

FIG. 2.—*A,* Phylogenetic relationship of *Rickettsia* species based on the combined amino acid sequences of the *ftsY, polA,* and *dnaE* gene products. Neighbor-joining, maximum parsimony and maximum likelihood methods gave identical tree topologies. Values in the boxes indicate the percentages of 500 bootstraps in which the taxa to the right were clustered (neighbor joining in boldface, maximum parsimony in italic, and puzzling support for maximum likelihood in normal text). Numbers above the branches are the numbers of amino acid substitutions at conserved sites in the *metK* gene products. Numbers below the branches are the total numbers of amino acid substitutions at conserved sites in the *ftsY, polA,* and *dnaE* gene products. Numbers in brackets are the numbers of conserved amino acids removed by insertions and/or deletions. *B,* Patterns of deletions and insertions in the *metK* gene. *C,* Patterns of deletions and insertions in the *fossil* gene sequences. Triangles indicate sites of deletion and insertion mutations. Numbers above the triangles are the estimated sizes of the deletions or insertions in base pairs. Vertical lines show the positions of in-frame termination codons, with the identity of the stop codon indicated.

deletion mutations in the *metK* genes of the SFG *Rickettsia.*

For inactivated genes, the fixation rate for nonsynonymous substitution should approach the fixation rate for synonymous substitutions. Indeed, the K_a/K_s ratios were close to one in several pairwise comparisons of substitution rates based on the reconstructed *fossil* gene sequences (fig. 3). For a few closely related taxa, such as, for example, *R. rickettsi* and *R. sibirica,* the total numbers of substitutions were too few to allow accurate measurements of substitution rates, and we therefore combined the data for some of these species. Taken together, our data suggest that the *metK* genes as well as the *fossil* genes have experienced an enhanced rate of nonsynonymous substitution in addition to the accumulation of termination codons and frameshift mutations.

Replacements of Universally Conserved Amino Acids

A limitation of the analysis presented above is that it does not distinguish nonsynonymous substitutions with potentially harmful effects from amino acid replacements that are potentially neutral. To roughly estimate the fraction of potentially deleterious mutations that have accumulated in the *metK* gene since its inactivation, we calculated the fraction of substitutions at sites that are otherwise universally conserved.

Multiple alignment of AdoMet synthetases from a variety of bacterial species identified a total of 141 positions as universally conserved positions (data not shown). Forty-nine of these positions differed in one or more of the *Rickettsia* species. Single substitution or deletion events were observed at 34 positions. Multiple substitution and/or deletion events were observed at 15 positions, 3 of which were excluded from the analysis since they had changed both prior and subsequent to the divergence of the TG and SFG *Rickettsia.* The remaining 16 deletions and 44 substitutions could be assigned to single branches in the phylogenetic tree (fig. 2*A*). Members of the SFG *Rickettsia* have accumulated a total of 3–12 substitutions at universally conserved sites since they diverged from their nearest neighboring species. In

Table 4
Nucleotide Frequency Statistics of the Reconstructed *metK* and *orf* Genes in *Rickettsia* Species

Gene[a]	Species	Codons	GC[b]	GC3[c]	%ID[d]
metK ...	*R. prowazekii*, B	329	0.354	0.199	64.1
f-*metK* ..	*R. prowazekii*, Madrid E	329	0.354	0.199	64.1
metK ...	*R. typhi*	329	0.355	0.193	64.1
f-*metK* ..	*R. rickettsii*	329	0.381	0.259	62.9
	R. sibirica	329	0.382	0.255	63.5
	R. montana	329	0.382	0.268	63.5
	R. rhipicephali	329	0.383	0.268	64.7
	R. helvetica	329	0.384	0.243	67.5
	R. felis	329	0.385	0.265	65.3
f-*orf*	*R. rickettsii*	289	0.270	0.158	—
	R. sibirica	289	0.272	0.161	—
	R. rhipicephali	289	0.283	0.158	—
	R. montana	289	0.287	0.172	—
	R. helvetica	289	0.283	0.193	—
orf	*R. felis*	289	0.287	0.193	—

[a] Fossil genes (indicated by "f-") were reconstructed as described in *Materials and Methods.*

[b] Gene G+C content.

[c] G+C content at silent third codon positions.

[d] Percentage of identity to the homologous genes in *Escherichia coli.*

contrast, only one such substitution was observed in *R. prowazekii* and *R. typhi* since their divergence from the SFG *Rickettsia*.

We performed a similar analysis based on the *ftsY*, *polA*, and *dnaE* genes. For these genes, 32 positions out of 186 conserved sites in the reference alignment were changed in one or more of the *Rickettsia* species (data not shown). All changes were substitutions; five of these were located in regions with ambiguous alignment, and three sites had multiple substitutions. The remaining 24 substitutions were assigned to branches in the phylogenetic tree (fig. 2*A*). A large majority of these changes (21) occurred prior to the divergence of the TG and SFG *Rickettsia*, and only two could be assigned to individual branches within the *Rickettsia* lineage.

In summary, our analysis suggests that there is a significant overrepresentation of amino acid substitutions with potentially deleterious effects in members of the SFG *Rickettsia* as compared with such substitutions in other genes and other species. This is most likely a reflection of relaxed selective constraints acting on the *metK* gene products.

Patterns of Nucleotide Substitutions

We estimated the patterns of point substitutions in the two pseudogenes by first reconstructing the putative ancestral sequences for the SFG *Rickettsia* and then mapping all of the changes to the individual lineages. In total, 158 out of 987 positions were found to be variable in the *metK* gene sequences, of which 7 sites had accumulated 2 independent mutations. Similarly, 138 out of 867 positions were found to be variable in the *fossil* gene sequences, 6 of which were associated with multiple hits. A total of 3 sites were excluded from the analysis because the ancestral states could not be unambiguously resolved due to multiple substitutions. The sequence data for the two pseudogenes were combined, yielding a final data set consisting of 306 mapped mutations, for which no corrections for multiple hits were performed.

The profile of point substitutions at the individual branches is shown in figure 4. Here, it can be seen that transitions are by far the most prevalent type of substitutions and that these are biased toward A and T nucleotides. The ORF in *R. felis* has a significantly higher fraction of substitutions at third codon positions than at the other two positions, as expected for an actively transcribed and translated gene. Codon-specific variations in substitution rates were also observed for the reconstruct-

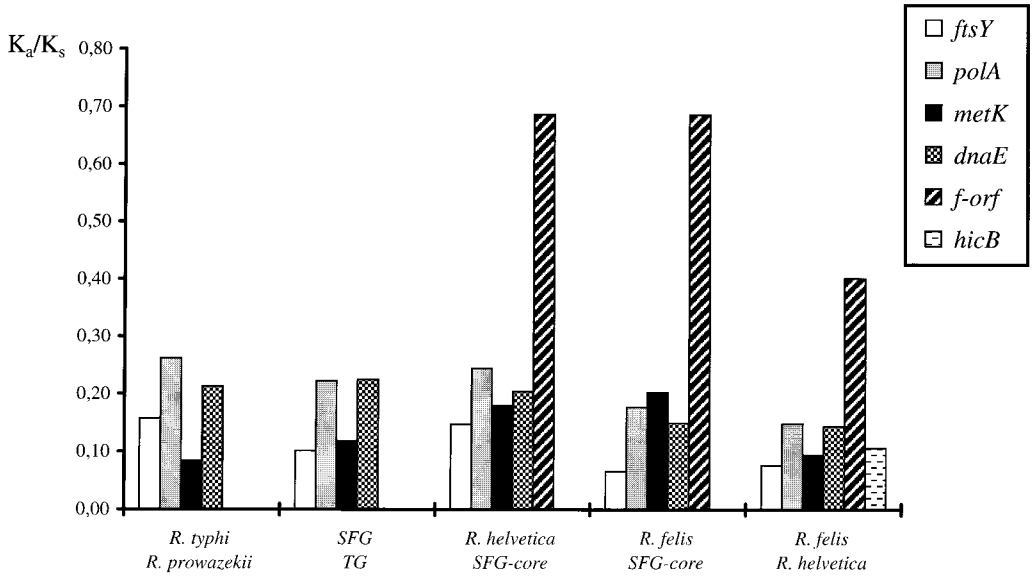

FIG. 3.—Ratios of nonsynonymous and synonymous substitutions (K_a/K_s) for functional genes and reconstructed pseudogenes. Species were in some cases grouped together, and, for these, the average ratios of all pairwise comparisons are shown. The following *Rickettsia* species are included: TG—*R. prowazekii* and *R. typhi*; SFG—*R. rickettsii, R. sibirica, R. montana, R. rhipicephali, R. helvetica,* and *R. felis*; SFG-core—*R. rickettsii, R. sibirica, R. montana,* and *R. rhipicephali.*

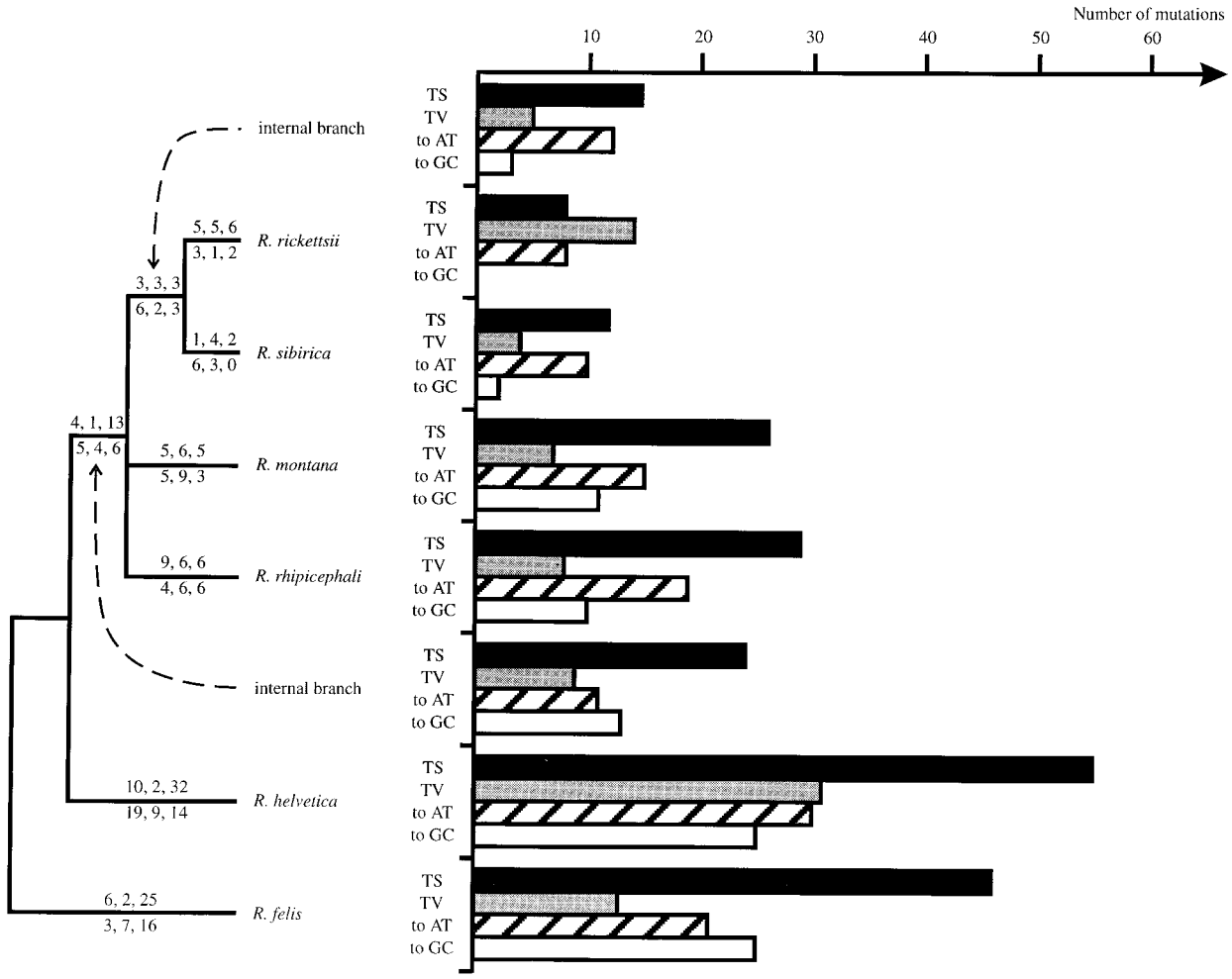

FIG. 4.—Patterns of point mutations in the *metK* and *fossil* gene sequences. The phylogenetic tree has been taken from figure 3*A*. Numbers above the branches show the frequencies of substitutions at the three codon positions in the *metK* gene. Numbers below the branches show the frequencies of substitutions at the three codon positions in the *fossil* gene sequences. The patterns of nucleotide substitutions have been estimated from the total number of changes at all codon positions in the *metK* and *fossil* gene sequences. The frequencies of transitions (black bars) and transversions (gray bars), as well as the fractions of GC-to-AT transitions (hatched bars) and AT-to-GC transitions (open bars), are shown.

ed *metK* genes in *R. felis* and *R. helvetica,* indicating that the *metK* genes in these two species have been under selective constraints for some time since their divergence from the other SFG *Rickettsia.* In contrast, no variations in substitution rates at the three codon positions for the reconstructed *metK* and *fossil* gene sequences could be observed for the branches leading to *R. rickettsii, R. sibirica, R. montana,* and *R. rhipicephali.* This provides evidence to suggest that the *metK* and *fossil* gene sequences in these species have been completely inactive since their divergence. For these lineages, we expect the rates and patterns of substitutions to correspond to the rates and patterns of mutations.

To study these mutational patterns in more detail, we estimated the rates of individual nucleotide changes by normalizing the numbers of mutations for the unequal frequencies of nucleotides in the reconstructed *metK* and *fossil* gene sequences (table 5). Here, it can be seen that the transition from a GC pair to an AT pair is almost 10-fold higher in frequency than the corresponding transversion (0.317 vs. 0.035 on average). The mutation pattern is strongly biased toward A and T nucleotides, since the transition from a GC pair to an AT pair is almost fivefold higher than that from AT to GC (0.317 vs. 0.062 on average). Based on these frequencies, we estimate that the expected G+C content of unconstrained DNA sequences from the SFG *Rickettsia* would be roughly 25% at equilibrium.

Short Repeated Sequences in Spacer Regions

Finally, we examined the remaining spacer regions for nucleotide frequencies and the presence of short repeated sequences. Intergenic sequences have here been defined as regions that are flanked by genes and/or pseudogenes. The typical length of the intergenic regions spanned from 32 bp, for the *metK-fossil* gene spacer in *R. felis,* to 367 bp, for the *polA-orf* spacer region in *R. helvetica* (table 6). In the absence of selective constraints, the G+C content values for spacers in *R. prowazekii* are expected to be lower than the overall genomic G+C content values. Indeed, intergenic regions of *R. prowazekii* have G+C content values in the range of

Table 5
Relative Substitution Frequencies in Pseudogenes in *Rickettsia*

	Relative Substitution Frequencies[a]				
From/To	A	T	C	G	Total[b]
A	—	0.004 (1)	0.080 (19)	0.080 (19)	739
T	0.013 (2)	—	0.044 (7)	0.025 (4)	489
C	0.042 (4)	0.293 (28)	—	0.021 (2)	296
G	0.341 (36)	0.028 (3)	0.028 (3)	—	327

[a] The rate estimates were based on 128 mapped mutations in the *metK* and *fossil* gene sequences in *R. rickettsii, R. sibirica, R. montana,* and *R. rhipicephali.* The rate estimates were normalized for the unequal frequencies of nucleotides in the reconstructed genes. Numbers in parentheses are the numbers of individual nucleotide changes.

[b] The total number of nucleotides in the reconstructed genes.

12%–31%, with a mean of 23% (Andersson and Sharp 1996; Andersson and Andersson 1997; Andersson et al. 1998). We note that the spacer sequences of the *metK* gene region in *R. prowazekii* have G+C content values within the expected range (23.4% on average) (table 2). However, some intergenic regions in the SFG *Rickettsia* showed the expected low G+C content values, while others were associated with G+C content values significantly higher than the genomic G+C content values (tables 2 and 5).

We also observed that the intergenic regions of the SFG *Rickettsia* showed a much higher level of diversity than expected. For example, the spacer sequence downstream of *polA* and upstream of the short ORF have

Table 6
Intergenic Regions: Sizes, Nucleotide Frequencies, and Inverted Repeats in *Rickettsia* Species

Intergenic Regions	Species	*L*[a]	GC[b]	IR[c]	Loop	Homology	Species	Matches
ftsY-polA	*R. prowazekii*	153	30.7	—	—	—	—	—
	R. typhi	166	31.9	—	—	—	—	—
	R. rickettsii	163	36.8	8/8[d]	78	f-*orf–dnaE*	*R. rickettsii*	94/115
	R. sibirica	163	38.0	13/13[d]	78	f-*orf–dnaE*	*R. sibirica*	95/115
	R. montana	162	38.3	11/12[d], 11/11	79, 13	f-*orf–dnaE*	*R. montana*	94/110
	R. rhipicephali	162	37.0	10/10[d]	83	f-*orf–dnaE*	*R. rickettsii*	89/114
	R. helvetica	134	34.3	33/34	45	*orf–dnaE*	*R. felis*	58/77
						dnaE	*R. helvetica*	29/35
	R. felis	145	39.3	—	—	f-*orf-dnaE*	*R. rickettsii*	89/129
polA-metK ...	*R. prowazekii*	277	24.9	—	—	—	—	—
	R. typhi	286	24.5	—	—	—	—	—
	R. rickettsii	175	30.3	8/8[e]	4	*dnaE*	*R. felis*	30/32
	R. sibirica	179	30.7	7/8[e]	4	*dnaE*	*R. felis*	29/32
	R. montana	124	37.1	8/8[e]	4	*dnaE*	*R. felis*	30/32
	R. rhipicephali	120	35.8	15/15[e]	4	*dnaE*	*R. felis*	56/79
polA-orf	*R. helvetica*	367	24.5	—	—	—	—	—
	R. felis	303	36.0	29/29[e]	44	*dnaE*	*R. felis*	77/107
orf-metK	*R. helvetica*	32	31.2	—	—	—	—	—
	R. felis	40	30.0	—	—	—	—	—
metK-polA ...	*R. prowazekii*	366	19.1	—	—	—	—	—
	R. typhi	1,524	25.3	—	—	—	—	—
*metK–*f-*orf* ...	*R. rickettsii*	67	22.4	—	—	—	—	—
	R. sibirica	67	25.4	—	—	—	—	—
	R. montana	71	21.1	—	—	—	—	—
	R. rhipicephali	70	21.4	—	—	—	—	—
	R. helvetica	71	18.3	—	—	—	—	—
metK-orf	*R. felis*	119	24.4	—	—	—	—	—
f-*orf–dnaE* ...	*R. rickettsii*	278	41.0	17/19[f]	144	*ftsY-polA*	*R. rickettsii*	94/115
	R. sibirica	280	41.1	19/21[f]	144	*ftsY-polA*	*R. sibirica*	95/115
	R. montana	273	44.3	19/21[f], 11/12[d]	146, 80	*ftsY-polA*	*R. sibirica*	100/116
	R. rhipicephali	208	44.2	18/21, 18/18[d]	81, 3	—	—	—
	R. helvetica	196	40.8	37/38[d], 16/17	96, 23	*ftsY-polA*	*R. rhipicephali*	42/43
orf-dnaE	*R. felis*	268	42.2	20/20[f]	147	*ftsY-polA*	*R. sibirica*	64/95
						ftsY-polA	*R. helvetica*	58/72
dnaE	*R. felis*	1,161	35.1	12/12, 8/8[e]	103, 4	*polA-orf*	*R. felis*	77/107

[a] Size of intergenic region (in nucleotides).

[b] G+C content of intergenic region.

[c] Inverted repeat.

[d] Inverted repeat contains two copies of CGTCATTGCGAG with at most one mismatch.

[e] Inverted repeat contains a perfect copy of the RS3 sequence 5′-ATTCCC-3′ and 5′-GGGAAT-3′ (Haas and Meyer 1986).

[f] Intergenic region contains at least one copy of the sequence CGTCATTGCGAG with at most one mismatch not associated with the inverted repeat.

G+C content values of 24.5% and 36.0% in *R. helvetica* and *R. felis*, respectively (table 6). Further analysis showed that this striking difference is explained by the presence of short inverted repeats in some intergenic regions but not in others (table 6). Although the positions and sequences of the inverted repeats were not generally conserved among species, some notable similarities were detected.

For example, a core sequence (5′-CGTTCA-TTGCGAG-3′) was part of the inverted repeat located within the *ftsY-polA* spacer regions in *R. rickettsii*, *R. sibirica*, *R. montana*, and *R. rhipicephali* as well as of the inverted repeat in the *fossil*-gene–*dnaE* spacer region in *R. montana* and *R. rhipicephali* (table 6). This core sequence was also detected as a single-copy sequence in the *fossil*-gene–*dnaE* spacer region in *R. rickettsii*, *R. sibirica*, and *R. helvetica* (table 6). However, the overall sequence similarity between the *ftsY-polA* intergenic region and the *fossil*-gene–*dnaE* spacer region could not be explained solely by the presence of inverted repeats, nor was the similarity particularly strong for spacer sequences within the same genome. For example, 58 out of 77 residues in the *ftsY-polA* intergenic region of *R. helvetica* were found to be identical to those in the downstream region of *dnaE* in *R. felis* but not to the corresponding region in *R. helvetica* (table 6).

Another motif identified in several species was the RS3 core sequence (5′-ATTCCC-3′ and 5′-GGGAAT-3′) which has been extensively studied in *Neisseria* species (Haas and Meyer 1986). RS3 repeats are common in noncoding regions of *Neisseria* species, where they may be involved in the shuffling of genetic information within and between genomes so as to create antigenic variation (Haas and Meyer 1986; Seiler et al. 1996). The RS3-like sequence motif was perfectly conserved in the *polA-metK* spacer regions of *R. rickettsii*, *R. sibirica*, *R. montana*, *R. rhipicephali*, and *R. felis* (table 6). Surprisingly, an RS3-like sequence could also be identified within the *dnaE* gene of *R. felis*, in a region in which 77 out of 107 nucleotides were identical to those in the *polA-metK* spacer region of *R. felis* (table 6). This region of the *dnaE* gene is unique to *R. felis*, and no similarities were detected in the *dnaE* genes of the other rickettsial species. Close inspection of the 29/29 hairpin in the *polA-metK* spacer revealed that each arm consisted of an inverted repeat similar to the inverted repeats observed in the *polA-metK* spacer regions of *R. rickettsii*, *R. sibirica*, *R. montana*, and *R. rhipicephali*.

In contrast to the apparent abundance of short repeated sequences in the genomes of the SFG *Rickettsia*, no such repeats were detected in the TG *Rickettsia* in this region. Searches against the entire *R. prowazekii* genome sequence using the inverted repeats failed to reveal any significant similarities, indicating that the absence of short repeats may be a general characteristic of the TG *Rickettsia* genomes.

Discussion

Bacterial transitions to intracellular environments are thought to be associated with massive losses of genetic information (Andersson and Kurland 1995, 1998). This is because many otherwise essential genes become redundant under the influence of the host genome and the enzymatic activities of the host cytoplasm. Provided that the loss of genes such as those encoding biosynthetic functions can be compensated for by an increased repertoire of transport functions, these losses may not be deleterious for the organism. Not surprisingly, epi- or intracellular parasites tend to have much lower fractions of genes coding for biosynthetic functions than do free-living bacteria (Fraser et al. 1995, 1997; Andersson et al. 1998).

In addition, small population sizes in combination with low recombination frequencies and recurrent bottlenecks during transmission between hosts may support the accumulation of irreversible genomic defects by a process commonly referred to as Muller's ratchet (Muller 1964; Felsenstein 1974). Whereas a duplication event can in principle be restored by a matching deletion event, the removal of a sequence may be irreversible in small bacterial populations with low recombination frequencies. Increased fixation rates for deletions may therefore be expected in small, asexual bacterial populations (Andersson and Kurland 1995, 1998).

In short, there may be a variety of explanations for gene loss in intracellular parasites, including (1) a local relaxation of selection for a subset of genes caused by the exploitation of the corresponding gene products from the host; (2) a reduced efficiency of selection caused by small population sizes, bottlenecks, and low recombination frequencies; and (3) direct selection for small genomes and small bacterial cell volumes to ensure rapid replication and large burst sizes (Andersson and Kurland 1995, 1998). In this study, we have presented evidence to suggest that the *metK* gene, which codes for AdoMet synthetase, is gradually decaying in a number of *Rickettsia* species.

Given the variety of possible reasons for gene loss, we may wonder: why is the *metK* gene being degraded? A trivial explanation would be that AdoMet, the product of the reaction catalyzed by AdoMet synthetase is no longer needed by the cell. This seems unlikely, since many genes coding for enzymes that utilize AdoMet have been identified in the *R. prowazekii* genome, such as, for example genes involved in ubiquinone biosynthesis (Andersson et al. 1996, 1998). Another trivial explanation would be that the *metK* gene function is redundant. However, only one single copy of the *metK* gene was identified in the complete genome sequence (Andersson et al. 1998), as also supported by a Southern blot analysis (Andersson and Andersson 1997). A more likely explanation is that the *Rickettsia* have rather recently invented an import system for AdoMet. If so, the *metK* gene would have been free to start accumulating mutations as soon as the intracellular concentration of imported AdoMet reached acceptable concentrations. Uptake of AdoMet has previously been demonstrated in American *Leishmania* promastigotes, as well as in mitochondria (Avila and Polegre 1993; Horne, Holloway, and Eagner 1997). Thus, the observed degradation of the *metK* genes is most likely explained by a local re-

laxation of selection caused by the invention of transport systems for AdoMet, rather than an overall reduced efficiency of selection related to Muller's ratchet.

We detected remnants of a second gene immediately downstream of the *metK* genes in *R. rickettsii, R. sibirica, R. montana, R. rhipicephali,* and *R. helvetica.* However, the same thorough analysis could not be done for this gene because of the lack of protein homologs with known functions. Nevertheless, the codon usage patterns of the reconstructed *fossil* genes suggest that this gene was also once an actively transcribed and translated gene (table 4). However, the ratios of nonsynonymous to synonymous substitutions for the reconstructed *fossil* genes were found to be very high, in the range of 0.6–0.7, suggesting that nonsynonymous substitutions are approaching neutrality. Since there are no similarities to genes in the public databases, we have no clues as to what the function of this gene might have been or why it is currently being eliminated. However, it is conceivable that one or more promoter mutations upstream of the *metK* gene simultaneously inactivated both the *metK* and the *fossil* genes. If so, the silencing of the *fossil* gene may be best viewed as a side effect of the *metK* gene inactivation.

We have shown that the *metK* gene has been destroyed by mutations in at least seven different species of the genus *Rickettsia.* The positions of the observed insertion/deletion mutations were not conserved except for a few deletions that were shared between the two most closely related species, *R. rickettsii* and *R. sibirica.* The extensive decay of the *metK* gene in the SFG *Rickettsia* suggests that the inactivation of the *metK* gene occurred much earlier in this group than in the TG *Rickettsia.* In fact, the sole in-frame termination codon in the *metK* gene in the Madrid E strain of *R. prowazekii* is the first obvious sign of degradation in the TG *Rickettsia.* Since the termination codon is not present in the B strain of *R. prowazekii,* it was probably created as recently as within the last few million years in the Madrid E strain.

The patterns of amino acid substitution along the branches with deletions in the *metK* genes also support an earlier inactivation within the SFG *Rickettsia.* Whereas five universally conserved amino acids on average have been substituted within each member of the SFG *Rickettsia,* only one such replacement mutation could be detected in the TG *Rickettsia.* Furthermore, at least one universally conserved amino acid has been lost by deletion in each lineage within the SFG *Rickettsia.* Again, no such deletions could be detected in the TG *Rickettsia.* Finally, a lower ratio of nonsynonymous to synonymous substitutions also indicates that the inactivation of the *metK* gene occurred more recently in the TG *Rickettsia* than in the SFG *Rickettsia.*

An alternative interpretation is that the *metK* gene was inactivated only once, but that the rates of deletion and substitution mutations have differed among the *Rickettsia* species, perhaps because of differences in lifestyles. One such difference between the two groups of *Rickettsia* is the vector system used for transmission; whereas *R. felis* and the TG *Rickettsia* are established in insects, hard body ticks are the principle vectors for the SFG *Rickettsia.* Unfortunately, parameters such as growth properties, population sizes, and occurrence of bottlenecks in the different vectors and hosts are not very well known, but it seems unlikely that the average generation time would have differed widely for the lineages giving rise to the TG and SFG *Rickettsia.* The most likely scenario may be that an uptake system for AdoMet was acquired or invented only once, presumably prior to the divergence of the TG and the SFG *Rickettsia.* However, the time elapsed from the acquisition of an import system for AdoMet until fixation of the first inactivation mutation in the *metK* gene may have varied among the different lineages.

Nevertheless, the patterns of changes in the inactivated genes were found to be very similar in the different species. In particular, the frequency of deletions was much higher than the frequency of insertions for both pseudogenes in all lineages. One of the deletions in the *fossil* gene sequence in *R. rhipicephali* was flanked by a region containing a direct repeat in all other rickettsial species, suggesting that the repeat may have been the site of slippage replication. Similarly, several sites of deletions/insertions in the spacer region of the *fmt* and 23S rRNA genes were found to be flanked by short direct repeat sequences (Andersson et al. 1999). The strong predominance of deletion mutations implies that there is a general trend toward genomic shrinkage in *Rickettsia* under neutral conditions.

The data presented in this study may also be used to estimate spontaneous mutation frequencies, which are not very well known for bacteria. The analysis of the *metK* and *fossil* genes in the SFG *Rickettsia* indicate that the frequencies of different nucleotide substitutions are far from equal. The greatest bias is in the high frequency of transitions from GC to AT, which are on average twice as frequent as transitions from AT to GC. This is consistent with our expectations that the elevated A+T content of noncoding DNA and the characteristic gradient in G+C content at the three codon positions in *R. prowazekii* genes are primarily caused by a strong mutational bias (Andersson and Sharp 1996; Andersson and Andersson 1997; Andersson et al. 1998). However, it is notable that transversions from A to C occur as frequently as transitions from A to G and twice as often at transitions from T to C. Since the genomic locations of the pseudogenes in the SFG *Rickettsia* are at present unknown, we do not know the extent to which the patterns of substitutions may be accounted for by strand-specific mutation biases.

The pattern of point mutations in *Rickettsia* is similar to that in *Drosophila* and mammals to the extent that transitions occur on average more frequently than do transversions (Li, Wu, and Luo 1984; Petrov and Hartl 1999). Furthermore, the highest proportion of nucleotide changes is for C to T and G to A in *Rickettsia* as well as in *Drosophila* and mammals (Li, Wu, and Luo 1984; Petrov and Hartl 1999). However, this is most likely attributable to sheer coincidence due to the selection of genomes for analysis, rather than to any conserved pattern of point mutations in diverse organisms.

From this perspective, it would be particularly interesting to examine the patterns of point mutations in G+C-biased microbial genomes.

We expect sequence similarities between the TG and SFG *Rickettsia* to be recognizable in the intergenic regions, provided these regions have evolved solely by single-base changes as well as small deletion/insertion events and given that the frequency of synonymous substitutions is typically less than 0.30 in pairwise comparisons across the two groups. Indeed, for a short intergenic region downstream of the *metK* gene, we observed the expected sequence similarity. However, other intergenic regions seem to have evolved in a more complex way. For example, G+C-rich inverted repeats are common in the intergenic regions of the SFG *Rickettsia* but not in the TG *Rickettsia*. We also identified short sequences that are identical to the so-called RS3 repeats of *Neisseria* species, which have been associated with DNA mobility (Haas and Meyer 1986; Seiler et al. 1996). An RS3-like sequence has been inserted into the *dnaE* gene of *R. felis*, suggesting that inverted repeats with RS3-like sequences might be mobile in the *Rickettsia* genome. Transpositions and/or differential loss of longer sequence segments could provide an explanation for the apparent lack of sequence similarity in many intergenic regions of the TG and the SFG *Rickettsia.* Inverted repeats were found to be very rare in the complete *R. prowazekii* genome (Andersson et al. 1998), suggesting that the lack of repeated sequences may be a general feature of the TG *Rickettsia.*

Elimination of gene sequences is likely to be initiated upon transitions to intracellular environments. Differences in genome sizes of intracellular parasites may reflect the time elapsed since the transition to the intracellular lifestyle and the efficiency with which novel transport systems have been invented, as well as differences in rates and patterns of deletion and insertion events. Systematic comparisons of the rates and patterns of gene degradation in obligate intracellular parasites of different phylogenetic affiliations are required in order to fully understand the evolutionary factors responsible for converting large genomes of free-living organisms into small genomes of intracellular parasites. In this way, the rates and patterns of point mutations, deletion as well as other mechanisms leading to genomic decay may be quantified.

Acknowledgments

We thank Charles Kurland for support and many helpful discussions. This work was supported by grants from the Natural Sciences Research Council, the Foundation for Strategic Research, and the Knut and Alice Wallenberg Foundation.

LITERATURE CITED

ADAMS, J. R., E. T. SCHMIDTMANN, and A. F. AZAD. 1990. Infection of colonized cat fleas, *Ctenocephalis felis* (bouche), with a rickettsia-like microorganism. Am. J. Trop. Med. Hyg. **43**:400–409.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

ANDERSSON, J. O., and S. G. E. ANDERSSON. 1997. Genomic rearrangements during evolution of the obligate intracellular parasite *Rickettsia prowazekii* as inferred from an analysis of 52015 bp nucleotide sequence. Microbiology **143**:2783–2795.

ANDERSSON, S. G. E., A.-S. ERIKSSON, A. K. NÄSLUND, M. S. ANDERSEN, and C. G. KURLAND. 1996. The *Rickettsia prowazekii* genome: a random sequence analysis. Microb. Comp. Genomics **1**:293–315.

ANDERSSON, S. G. E., and C. G. KURLAND. 1995. Genomic evolution drives the evolution of the translation system. Biochem. Cell Biol. **73**:775–787.

———. 1998. Reductive evolution of resident genomes. Trends Microbiol. **6**:263–268.

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

ANDERSSON, S. G. E., D. STOTHARD, P. FUERST, and C. G. KURLAND. 1999. Molecular phylogeny and rearrangement of rRNA genes in *Rickettsia* species. Mol. Biol. Evol. (in press).

ANDERSSON, S. G. E., A. ZOMORODIPOUR, J. O. ANDERSSON, T. SICHERITZ-PONTEN, U. C. M. ALSMARK, R. M. PODOWSKI, A. K. NÄSLUND., 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.

AVILA, J., and M. A. POLEGRE. 1993. Uptake and metabolism of S-adenosyl-L-methionine by *Leishmania mexicana* and *Leishmania braziliensis* promastigotes. Mol. Biochem. Parasitol. **58**:123–134.

AZAD, F. A. 1988. Relationship of vector biology and epidemiology of louse- and flea-borne rickettsioses. Pp. 51–61 *in* D. H. WALKER, ed. Biology of rickettsial dieseases. Vol. 1. CRC Press, Boca Raton, Fla.

BAUMANN, P., L. BAUMANN, C. Y. LAI, and D. ROUHBAKSH. 1995. Genetics, physiology, and evolutionary relationships of the genus *Buchnera*: intracellular symbionts of aphids. Annu. Rev. Microbiol. **49**:55–94.

BREZINA, R., E. S. MURRAY, M. L. TARIZZO, and K. BÖGEL. 1973. Rickettsiae and rickettsial diseases. Bull. WHO **49**: 433.

BRYNNEL, E. U., C. G. KURLAND, N. A. MORAN, and S. G. E. ANDERSSON. 1998. Evolutionary rates for *tuf* genes in endosymbionts of aphids. Mol. Biol. Evol. **15**:574–582.

DELORME, C., J.-J. GODON, S. D. EHRLICH, and P. RENAULT. 1993. Gene inactivation in *Lactococcus lactis*: histidine biosynthesis. J. Bacteriol. **175**:4391–4399.

DEPAMPHILIS, C. W., N. D. YOUNG, and A. D. WOLFE. 1997. Evolution of plastid rps2 in a lineage of hemiparasitic and holoparasitic plants: many losses of photosynthesis and complex patterns of rate variation. Proc. Natl. Acad. Sci. USA **94**:7367–7372.

EREMEEVA, M. E., V. ROUX, and D. RAOULT. 1993. Determination of genome size and restriction pattern polymorphism of *Rickettsia prowazekii* and *Rickettsia typhi* by pulsed field gel electrophoresis. FEMS Microbiol. Lett. **112**:105–112.

FELSENSTEIN, J. 1974. The evolutionary advantage of recombination. Genetics **78**:157–159.

FRASER, C. M., S. CASJENS, W. M. HUANG et al. (38 co-authors). 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi.* Nature **390**:580–586.

FRASER, C. M., J. D. GOCAYNE, O. WHITE, et al. (29 co-authors). 1995. The minimal gene complement of *Mycoplasma genitalium.* Science **270**:397–403.

FUJIOKA, M. 1992. Mammalian small molecule methyltransferases: their structural and functional features. Int. J. Biochem. **24**:1917–1924.

GALTIER, N., M. GOUY, and C. GAUTIER. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. **12**:543–548.

GODON, J.-J., C. DELORME, J. BARDOWSKI, M.-C. CHOPIN, S. D. EHRLICH, and P. RENAULT. 1993. Gene inactivation in *Lactococcus lactis*: branched-chain amino acid biosynthesis. J. Bacteriol. **175**:4383–4390.

HAAS, R., and T. F. MEYER. 1986. The repertoire of silent pilus genes in *Neisseria gonorrhoeae*: evidence for gene conversion. Cell **44**:107–115.

HORNE, D. W., R. S. HOLLOWAY, and C. EAGNER. 1997. Transport of *S*-adenosylmethionine in isolated rat liver mitochondria. Arch. Biochem. Biophys. **343**:201–206.

LI, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**: 96–99.

LI, W. H., C.-I. WU, and C.-C. LUO. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21**:58–71.

LLOYD, A. T., and P. M. SHARP. 1992. CODONS: a microcomputer program for codon usage analysis. J. Hered. **83**:239–240.

MHLANGA-MUTANGADURA, T., G. MORLIN, A. L. SMITH, A. EISENSTARK, and M. GOLOMB. 1998. Evolution of the major pilus gene cluster in *Haemophilus influenzae.* J. Bacteriol. **180**:4693–4703.

MORAN, N. A. 1996. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA **93**:2873–2878.

MULLER, J. J. 1964. The relation of recombination to mutational advance. Mutat. Res. **1**:2–9.

MYERS, E., and W. MILLER. 1988. Optimal alignment in linear space. Comput. Appl. Biosci. **4**:11–17.

OPHIR, R., and D. GRAUR. 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene **205**:191–202.

PETROV, F. A., and D. L. HARTL. 1997. Trash DNA is what gets thrown away: high rate of DNA loss in Drosophila. Gene **205**:279–289.

———. 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol. Biol. Evol. **15**:293–302.

———. 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. Proc. Natl. Acad. Sci. USA **96**: 1475–1479.

POIRIER, L. A., W. ZAPISEK, and B. LYN-COOK. 1990. Physiological methylation in carcinogenesis. Prog. Clin. Biol. Res. **340D**:97–112.

PRETZMAN, C. I., Y. RIKIHISA, D. RALPH, J. C. GORDON, and S. BECH-NIELSEN. 1987. Enzyme-linked immunosorbent assay for Potomac horse fever disease. J. Clin. Microbiol. **25**: 31–36.

RAOULT, D., and V. ROUX. 1997. Rickettsioses as paradigms of new or emerging infectious diseases. Clin. Microbiol. Rev. **10**:694–719.

ROUX, V., and D. RAOULT. 1993. Genotypic identification and phylogenetic analysis of the spotted fever group Rickettsiae by pulsed-field gel electrophoresis. J. Bacteriol. **175**:4895–4904.

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.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SCHRAMEK, S. 1974. Deoxyribonucleic acid base composition of Rickettsiae belonging to the Rocky Mountain spotted fever group isolated in Czechoslovakia. Acta Virol. **18**:173–174.

SEILER, A., R. REINHARDT, J. SARKARI, D. A. CAUGANT, and M. ACHTMAN. 1996. Allelic polymorphism and site-specific recombination in the *spc* locus of *Neisseria meningitidis.* Mol. Microbiol. **19**:841–856.

STADEN, R. 1996. The Staden sequence analysis package. Mol. Biotechnol. **5**:233–241.

STEPHENS, R. S., S. KALMAN, C. LAMMEL et al. (12 co-authors). 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis.* Science **282**: 754–759.

STOTHARD, D. R. 1995. The evolutionary history of the genus *Rickettsia* as inferred from 16S and 23S rRNA genes and 17 kDa cell surface antigen gene. Dissertation, Ohio State University, Columbus.

STRIMMER, K., and A. VON HAESELER. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. USA **94**:6815–6819.

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.

TRAUB, R., C. L. WISSEMAN JR., and A. FARHANG-AZAD. 1978. The ecology of murine typhus—critical review. Trop. Dis. Bull. **75**:237–317.

TYERYAR, F. J. J., E. WEISS, D. B. MILLAR, F. M. BOZEMAN, and R. A. ORMSBEE. 1973. DNA base composition of Rickettsiae. Science **180**:415–417.

WEISS, E. 1982. The biology of rickettsiae. Annu. Rev. Microbiol. **36**:345–370.

WILLIAMS, S. G., J. B. SACCI JR., M. E. SCHRIEFER, E. M. ANDERSEN, K. K. FUJIOKA, F. J. SORVILLO, A. R. BARR, and A. F. AZAD. 1992. Typhus and typhus-like rickettsiae associated with opossums and their fleas in Los Angeles County, California. J. Clin. Microbiol. **30**:1758–1762.

WINKLER, H. H. 1976. Rickettsial permeability: and ADP-ATP transport system. J. Biol. Chem. **251**:389–396.

WISSEMAN, C. L. JR., E. A. EDLINGER, A. D. WADDELL, and M. R. JONES. 1976. Infection cycle of *Rickettsia rickettsii* in chick embryo and L-929 cells in culture. Infect. Immun. **14**:1052–1064.

WOLFE, K. H., C. W. MORDEN, and J. D. PALMER. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc. Natl. Acad. Sci. USA **89**:10648–10652.

HOWARD OCHMAN, reviewing editor

Accepted May 19, 1999