

Comparative Genomic Analysis of Three Strains of *Ehrlichia ruminantium* Reveals an Active Process of Genome Size Plasticity†

Roger Frutos,^{1*} Alain Viari,² Conchita Ferraz,³ Anne Morgat,^{2,4} Sophie Eychenié,³ Yane Kandassamy,¹ Isabelle Chantal,¹ Albert Bensaid,^{1,‡} Eric Coissac,² Nathalie Vachieri,¹ Jacques Demaille,³ and Dominique Martinez¹

CIRAD-Emvt, TA30/G, Campus International de Baillarguet, 34398 Montpellier Cedex 5, France¹; Inria Rhône-Alpes—Projet HELIX, 655 Av. de l'Europe, 38330 Montbonnot-Saint Martin, France²; Centre de Séquençage Génomique, IGH-CNRS-UPR 1142, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France³; and Swiss Institute of Bioinformatics, Swiss-Prot Group, 1 rue Michel, Servet, CH-1211 Geneva 4, Switzerland⁴

Received 31 August 2005/Accepted 9 January 2006

Ehrlichia ruminantium is the causative agent of heartwater, a major tick-borne disease of livestock in Africa that has been introduced in the Caribbean and is threatening to emerge and spread on the American mainland. We sequenced the complete genomes of two strains of *E. ruminantium* of differing phenotypes, strains Gardel (Erga; 1,499,920 bp), from the island of Guadeloupe, and Welgevonden (Erwe; 1,512,977 bp), originating in South Africa and maintained in Guadeloupe in a different cell environment. Comparative genomic analysis of these two strains was performed with the recently published parent strain of Erwe (Erwo) and other *Rickettsiales* (*Anaplasma*, *Wolbachia*, and *Rickettsia* spp.). Gene order is highly conserved between the *E. ruminantium* strains and with *A. marginale*. In contrast, there is very little conservation of gene order with members of the *Rickettsiaceae*. However, gene order may be locally conserved, as illustrated by the *tuf* operons. Eighteen truncated protein-encoding sequences (CDSs) differentiate Erga from Erwe/Erwo, whereas four other truncated CDSs differentiate Erwe from Erwo. Moreover, *E. ruminantium* displays the lowest coding ratio observed among bacteria due to unusually long intergenic regions. This is related to an active process of genome expansion/contraction targeted at tandem repeats in noncoding regions and based on the addition or removal of ca. 150-bp tandem units. This process seems to be specific to *E. ruminantium* and is not observed in the other *Rickettsiales*.

Ehrlichia ruminantium is the causative agent of cowdriosis, or heartwater, in domestic and wild ruminants in sub-Saharan Africa; African islands, including Madagascar; and some of the lesser Caribbean islands (56). This tick-borne member of the *Rickettsiales*, an α -proteobacterium, is a small, gram-negative, aerobic, obligate intracellular pathogen of endothelial cells that can cause up to 90% mortality in susceptible animals. Heartwater is responsible for great economic losses in Africa (48) but also represents a threat to the American mainland owing to the presence of potentially transmitting ticks (8, 17). The control strategy is based on vector eradication and immunization, a strategy possible only on islands, where the incoming flow of ticks is highly limited (53). Vaccine development therefore remains critical. The only available commercial vaccine relies on the risky and inappropriate infection of animals with infected blood followed by treatment with antibiotics (12). Attenuated and DNA vaccines were developed (20, 21, 33, 71), but they induce long-lasting protection only against homologous virulent strains while conferring limited protection against heterologous strains in the field. The main difficulty in

developing efficient vaccines is the simultaneous field occurrence of various genotypes in limited geographical areas (39). Similarly, serodiagnosis of heartwater has long been hampered by a lack of specificity and sensitivity (45), although these have been greatly improved recently (34, 36, 65). Additional diagnostic targets, protective proteins, and potential drug targets are therefore still needed. A key step toward this goal is better understanding of how genomes evolve in strains of differing phenotypes and which mechanisms are involved, in order to distinguish modifications linked to adaptation and plasticity from those directly involved in differing traits.

Adaptation to the intracellular lifestyle is accompanied in the early stages by massive gene losses, creation of pseudogenes, and diminution of the genome size (47). The important increase in mobile genetic elements is especially characteristic of the early stage of intracellular parasitism. Following initial reduction, the genome is subjected to opposing evolutionary forces, to generate diversity while eliminating useless sequences in the process of reaching stasis (37, 44, 46, 47, 61). Post-establishment evolution is strongly influenced by the host, leading to specific evolutionary processes in different host-restricted bacterial lineages. The *Rickettsiales*, which specialized in intracellular parasitism 700 million years ago (31), have diverged into several lineages—*Rickettsia*, *Wolbachia*, *Anaplasma*, and *Ehrlichia*—displaying distinctive genomic features (23, 66). While *Rickettsia* spp. and the *Wolbachia pipientis* endosymbiont of *Drosophila melanogaster* (wMel) harbor a large amount of selfish DNA (50, 51, 67), *Anaplasma*, *Ehrlichia*, and the *W. pipientis*

* Corresponding author. Mailing address: CIRAD TA30/G, Campus International de Baillarguet, 34398 Montpellier Cedex 5, France. Phone: 33 4 67 59 39 62. Fax: 33 4 67 59 39 60. E-mail: roger.frutos@cirad.fr.

† Supplemental material for this article may be found at <http://jb.asm.org/>.

‡ Present address: Centre de Recerca en Sanitat Animal, Campus de Bellaterra, Edifici V, 08193 Bellaterra, Barcelona, Spain.

endosymbiont of *Brugia malayi* (wBm) are devoid of insertion sequences (16, 22, 27). *Ehrlichia* spp. display multiple tandem repeats associated with pseudogenes or in intergenic regions, whereas *Anaplasma* spp. do not (16, 22).

We report here the comparative genomic analysis of the complete genomes of three strains of *E. ruminantium*: the Gardel strain (referred to here as Erga), a Welgevonden strain (referred to here as Erwe), and the recently sequenced South African Welgevonden strain (22) (referred to here as Erwo), from which Erwe originates. The Welgevonden genotype is infective and pathogenic to mice, whereas Erga is not. Erga illustrates the separation of the Gardel and Welgevonden genotypes before or at the time of the introduction of *E. ruminantium* in the Caribbean, whereas Erwe represents evolution from the Erwo strain through 14 passages in a different cell environment. Comparative genomic analysis of these three strains therefore allows for the analysis of genome evolution at different time scales. We first report strong gene order conservation between the *Ehrlichia* and *Anaplasma marginale* genomes. We also report the occurrence of differential protein-encoding sequence (CDS) truncations and the presence of several strain-specific genes. A detailed gene-to-gene alignment of the three genomes also allows for assessment of mutational pressure. Finally, we report the description of a process of genome contraction/expansion targeted at tandem repeats in noncoding regions and based on the addition or removal of ca. 150-bp tandem units. This process is specific to *E. ruminantium* and is not observed in the other *Rickettsiales* (including *A. marginale*).

MATERIALS AND METHODS

Bacterial strains and growth conditions. Strain Gardel of *E. ruminantium* (Erga) was isolated on the island of Guadeloupe in 1982 from a goat injected with a homogenate of a female *Amblyomma variegatum* tick collected from cows (64). *E. ruminantium* was multiplied successively in bovine umbilical endothelial cells and in bovine aorta endothelial cells grown in Glasgow minimal essential medium complemented with fetal calf serum, tryptose-phosphate broth, and antibiotics (13) at 37°C, 5% CO₂, with a weekly passage on fresh cells (41). Strain Welgevonden of *E. ruminantium* was isolated in South Africa in 1985 from mice injected with individually homogenized, infected, field-collected *Amblyomma hebraeum* ticks (24). A sample of this Welgevonden strain was received on Guadeloupe island on 25 May 1988 and immediately injected into a naive goat. A blood sample was collected from the infected goat after hyperthermia and confirmation of infection and was used to inoculate a cell culture. A total of 13 additional successive passages were performed over 18 years. Passages 11 to 14 were used to provide DNA for cloning and sequencing. This strain of the Welgevonden genotype was denominated Erwe. The original field-isolated Welgevonden strain was maintained in South Africa, where its genome was recently sequenced (22). This Welgevonden strain is referred to as Erwo. DNA was extracted from Erga and Erwe as described below.

DNA extraction, cloning, and sequencing. Elementary bodies were purified from culture supernatant, as previously described (40), resuspended in 350 µl of phosphate-buffered saline containing 0.36 µg/ml of DNase to remove contaminating host cell DNA, and incubated for 90 min at 37°C prior to the addition of 25 mM EDTA (41). Extraction of DNA from elementary bodies was done as previously described (54). Contamination with host DNA was checked by dot blot hybridization using bovine DNA as a positive control and probe. Purified DNA was broken by sonication to generate fragments of differing sizes. After filling of the ends with Klenow polymerase, DNA fragments ranging from 0.5 kb to 4 kb were separated in a 0.8% agarose gel and collected after Gelase (Epicenter) digestion of a cut-out agarose band. Blunt-end fragments were inserted into pBluescript II KS (Stratagene) digested with EcoRV and dephosphorylated. Ligation was performed with the Fast-Link DNA ligation kit (Epicenter), and competent *Escherichia coli* DH10B cells were transformed prior to colony isolation on LB agar-ampicillin-X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside)-IPTG (isopropyl-β-D-thiogalactopyranoside). About 15,000 clones

were isolated for both Erga and Erwe. Inserts were sequenced on both strands with universal forward and reverse M13 primers and the ET DYEnamic terminator kit (Amersham). Sequences were obtained with ABI 373 and ABI 377 automated sequencers (Applied Biosystems). Data were analyzed and contigs were assembled by the Phred-Phrap and Consed software packages (<http://www.genome.washington.edu>). Gaps were filled in through primer-directed sequencing with custom-made primers. A total of about 20,000 raw sequence runs were generated and analyzed for each *E. ruminantium* strain to generate a full-length consensus sequence with 7× coverage. Sequences of all mutated CDSs were extensively checked to eliminate the possibility of sequencing errors.

Gene prediction and annotation tools. The Erga and Erwe genomes were both annotated as described below. The Erwo genome was independently annotated by Collins et al. (22). The annotation process of Erga and Erwe was mainly conducted with the integrated computer environment GenoStar (25). CDSs were identified by using Markov chain models (15). First, a five-order periodic Markov model for coding sequences was trained on long open reading frames (≥1,000 nucleotides) for each strain. The model was then applied to the genomes, using 120 bp as the cutoff value for CDS length and a probability threshold (*P*) of 0.80 for CDSs of lengths less than 360 bp. To detect frameshifts due to sequencing errors, each CDS of one strain was then checked against all CDSs of the other strain with BlastP (3). A pair of CDSs exhibiting more than 70% amino acid sequence identity and a size difference of less than 20% was considered a pair of homologous genes. A pair of homologous CDSs corresponding to a bidirectional best hit (63) was considered a pair of orthologous CDSs. All other cases (CDSs of different lengths or without any detected homolog in the other strain) were examined manually for possible frameshifts by aligning the corresponding genomic regions using dynamic programming and by close inspection of gel reads. Protein similarity was assessed with BlastP (3) against a database resulting from the nonredundant concatenation of SwissProt (release 41) and the proteomes of two completely sequenced rickettsiae: *R. conorii* (7) and *R. prowazekii* (50). Inferred functions and gene names were checked manually; COG (for “Clusters of Orthologous Groups”) and EC numbers were assigned by a similar procedure on databases of bacterial proteins annotated with COG (62) and EC (14) numbers. tRNAs were located with fastRNA (68), and rRNAs were detected with BlastN (3) in a database of known bacterial RNAs. Genome GC-skew analysis was performed as previously described (28), and an arbitrary origin of coordinates was assigned at the putative origin of replication (no clear DNA box could be located). Codon and amino acid usage were analyzed by correspondence analysis (52, 55). Selection pressure on orthologous genes was assessed by counting synonymous and nonsynonymous nucleotide substitutions according to the methods of Nei and Gojobori (49). The significance of synonymous versus nonsynonymous substitution rates was determined by the Fisher exact test with a Bonferroni correction for multiple tests at an effective *P* value of 0.05 (69). Dispersed and tandem repeats were detected with two complementary programs. Repseek (2) was used for detecting dispersed repeats with a minimal seed length given by the Karlin-Ost formula (35) at a *P* value of 10⁻³. Tandem Repeat Finder (TRF) (10) was used for detecting tandem repeats with default parameters. TRF specifically targets tandem repeats, whereas Repseek can also find separated repeats but is less accurate on tandem repeat boundaries. Therefore, overlapping repeats detected by both programs were considered tandem repeats and only the TRF result was kept. When TRF identified several tandems at the same location, only the longest one was considered. Dispersed repeats were classified into four categories according to the relative orientation and distance between the two copies. The “direct” and “reverse” classes correspond to copies in the same and reverse orientations, respectively. The “close” and “distant” classes correspond to copies less than 1 kb away and more than 1 kb away, respectively.

Computation of size plasticity regions. A pair of orthologous noncoding regions (ONCR) between Erga and Erwe was defined as a pair of intergenic regions flanked by two pairs of orthologous CDSs with no intervening gene (CDS or RNA) in between. ONCR of less than 10 bp were removed. Each ONCR was additionally associated with a pair of orthologous coding regions (OCR), which were arbitrarily chosen as the pair of orthologous CDSs located upstream from the ONCR. The observed expansion/contraction of a region (Δ_{obs}) (i.e., either an OCNCR or an OCR) is defined as follows: $\Delta_{\text{obs}} = \text{length of region (Erwe)} - \text{length of region (Erga)}$. Three classes of expansion were considered: (i) if $-5 < \Delta_{\text{obs}} < 5$, then the expansion class was designated “0,” indicating that the region is stable; (ii) if $\Delta_{\text{obs}} > 5$, then the expansion class was designated “+,” indicating that the genome of Erwe expands with respect to that of Erga; and (iii) if $\Delta_{\text{obs}} < -5$, then the expansion class was designated “-,” indicating that the genome of Erwe contracts with respect to that of Erga.

Expanding/contracting regions (i.e., associated with either the “+” or “-” class) are referred to as “size plasticity regions.” Investigation of the correlation

TABLE 1. *Rickettsiales* genome features

Feature	<i>Rickettsiales</i>								
	<i>Anaplasmataceae</i>				<i>Rickettsiaceae</i>				
	<i>E. ruminantium</i> Gardel (Erga)	<i>E. ruminantium</i> Welgevonden (CIRAD) (Erwe)	<i>E. ruminantium</i> Welgevonden (ARC-OVI) (Erwo)	<i>A. marginale</i> St. Maries (Anmar)	<i>Wolbachia</i> sp. (subsp. <i>B. malayi</i> strain TRS) (wBm)	<i>W. pipientis</i> wMel	<i>R. conorii</i> Malish 7 (Ricon)	<i>R. prowazekii</i> Madrid E (Ripro)	<i>R. typhi</i> Wilmington (Rityp)
Genome size (bp)	1,499,920	1,512,977	1,516,355	1,197,687	1,080,084	1,267,782	1,268,755	1,111,523	1,111,496
GC (%)	27.5	27.5	27.5	49.8	34.2	35.2	32.4	29.0	28.9
No. of CDSs	950	958	920	948	1,194	805	1,372	835	838
tRNAs	36	36	36	37	34	34	33	33	33
rRNAs	3	3	3	3	3	3	3	3	3
% Coding DNA	64.4	63.8	63.1	86.0	67.7	80.8	81.4	76.3	76.2
HAMAP proteome	EHRRG	EHRRW	EHRRW	ANAMM	WOLTR	WOLPM	RICCN	RICPR	RICTY
Accession no.	CR925677	CR925678	CR767821	CP000030	AE017321	AE017196	AE006914	AJ235269	AE017197
Reference or source	This study	This study	22	16	27	67	50	7	42

between tandem repeats and the size plasticity of noncoding regions was conducted by surveying the presence of tandem repeats in each ONCR of Erga/Erwe. A tandem repeat was considered to be present in an ONCR if more than half of its length lies within the ONCR. If more than one repeat was found within the same ONCR, then only the largest one was kept. ONCR were classified into four categories (“Tandem_None,” “None_Tandem,” “Tandem_Tandem,” and “None_None”) depending on whether the ONCR contained a tandem repeat (“Tandem”) or not (“None”) in Erga and Erwe, respectively (for instance, the “Tandem_None” category corresponds to the case in which an ONCR displays a tandem repeat in Erga and no tandem repeat in Erwe). The theoretical expansion/contraction amount due to the tandem repeat (Δ_{theo}) was calculated as follows: $\Delta_{theo} = [(period_Erga + period_Erwe)/2] \times (nb_copy_Erwe - nb_copy_Erga)$, where “period” and “nb_copy” are the period and number of copies of the tandem repeats, respectively (both are set to 0 if no tandem is present).

Accession numbers. The sequences of the complete genomes of Erga and Erwe have been deposited in the EMBL databank under accession numbers CR925677 and CR925678, respectively.

RESULTS

General genome features. The genomes of Erga and Erwe are 1,499,920 and 1,512,977 bp long, respectively. Although originating from Erwo, Erwe does not display exactly the same size as the recently reported Erwo genome (22), which is 1,516,355 bp (Table 1). The G+C content is 27.51% and 27.48% for Erga and Erwe, respectively, which is similar to the ratio reported for Erwo. These values are consistent with the generally low G+C content of *Rickettsiales* with the exception of *A. marginale* (49.8%) (Table 1). The genome of Erga comprises 950 CDSs of an average size of 1,007 bp, representing a coding ratio (number of base pairs involved in CDSs/total number of base pairs) of 64.4% (Table 1). The genome of Erwe bears 958 CDSs of an average size of 998 bp, representing a coding ratio of 63.83% (Table 1). The Erwo genome bears 920 CDSs, of which 32 are considered pseudogenes (22), which leads to a coding ratio of 63.1%. The difference in the numbers of CDSs between Erwe and Erwo (38 CDSs) mostly originates from slightly different annotation strategies (setup of parameters and definition of pseudogenes). However, despite the fact that Erwe and Erwo have been sequenced and annotated completely independently, predictions and annotations are remarkably consistent. Finally, Erga, Erwe, and Erwo contain 36 tRNAs completely scattered across the chromosome (the longest putative operon comprises only 2 tRNAs) and 3 rRNAs, the 16S rRNA being separated from the 5S and 23S rRNAs and displaying the same organization as in other

Rickettsiales. At 63 to 64%, *E. ruminantium* exhibits the lowest coding ratio observed so far among all annotated bacterial and archaeobacterial genomes available from databases (<http://www.ebi.ac.uk/integr8>). The only genome reported to display a lower coding ratio is that of *Mycobacterium leprae* (19). However, this is related to the very large number of pseudogenes in the *M. leprae* genome (19). In the case of *E. ruminantium*, a closer inspection of the gene layout (data not shown) shows that the coding ratio is uniform along the genome and that the low coding ratio is mainly due to unusually long intergenic regions. Figure 1 displays the quantiles (25, 50, and 75) of the length of the intergenic regions for nine completely sequenced *Rickettsiales*, with *E. coli* as a typical bacterial reference. This figure shows that although the tendency to longer intergenic regions is a general feature of *Rickettsiales*, this effect is much

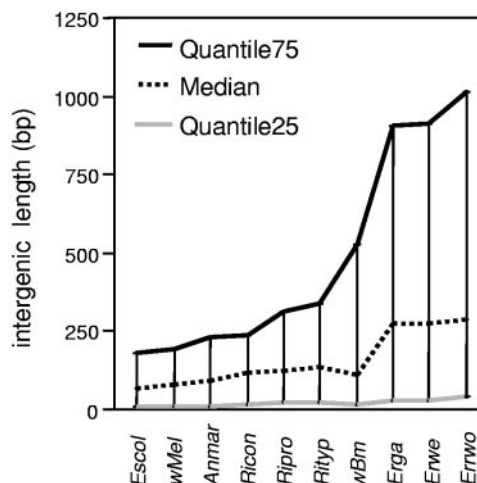


FIG. 1. Quantiles of intergenic length for nine completely sequenced *Rickettsiales* and *E. coli*. Quantile 75 is the value that 75% of the intergenic lengths fall below. The median is quantile 50. *Rickettsiales* tend to exhibit longer intergenic sequences than typical bacteria (represented here by *E. coli*), but this effect is more pronounced in *E. ruminantium*. Abbreviations: Escol, *E. coli*; wMel, *W. pipientis* strain wMel; Anmar, *A. marginale*; Ricon, *R. conorii*; Ripro, *R. prowazekii*; wBm, *Wolbachia* sp. (subsp. *B. malayi* strain TRS); Erga, *E. ruminantium* strain Gardel; Erwe, *E. ruminantium* strain Erwe; Erwo, *E. ruminantium* strain Erwo.

more pronounced in *E. ruminantium* (median, ~270 bp; quantile 75, ~1 kb) (Fig. 1).

Gene order is highly conserved between *E. ruminantium* strains and well conserved with *A. marginale*. Despite slight differences in length, Erga, Erwe, and Erwo exhibit a highly conserved gene order (see Fig. S1a in the supplemental material) and the perfect colinearity is affected by only two small inversions involving 2 and 3 CDSs, respectively, between Erga and Erwe/Erwo. Extending this comparison to other *Rickettsiales* and plotting *E. ruminantium* against *A. marginale* reveals good colinearity between the two species. Gene order is interrupted only by few large inversions (see Fig. S1b in the supplemental material). Gene shuffling is rather limited and mostly present in the central part of the genome, which corresponds to the terminus of replication, a region usually more prone to recombination. Colinearity is not observed between *E. ruminantium* and the *Rickettsiaceae*, i.e., *R. conorii*, *R. prowazekii*, *R. typhi*, wMel, and wBm (see Fig. S1c and d in the supplemental material).

Compositional biases. The cumulative GC-skew profile (28) obtained for *E. ruminantium* (data not shown) exhibits a strong leading/lagging compositional bias, similar to that observed with spirochetes (57). Similarly, the codon usage of *E. ruminantium* significantly differs between the two replication strands. Correspondence analysis of codon usage clearly shows two separated gene clusters (see Fig. S2a in the supplemental material) associated with the leading and lagging strands. This analysis also shows the presence of an additional cluster associated with the second factorial axis (see Fig. S2a in the supplemental material). A second correspondence analysis of amino acid usage (see Fig. S2b in the supplemental material) shows that the previously identified cluster corresponds to a group of genes coding for proteins with biased amino acid composition. Projection of characters (i.e., amino acids) further shows that this bias is directed toward an enrichment of large, hydrophobic amino acids: phenylalanine and tryptophan. Eighty-one CDSs are present in this cluster, of which 42 have been assigned to known membrane proteins (see Table S1 in the supplemental material). The 39 unknown proteins from this cluster might therefore be also associated with membrane proteins.

Comparative analysis of mutational trends. The high colinearity between the three strains of *E. ruminantium* allowed for detailed, gene-by-gene comparison and identification of differences which may explain host range variation. This resulted in an alignment table (see Table S1 in the supplemental material) composed of 986 rows; 888 rows (90%) correspond to triplets of genes (one gene for each strain), and 86 rows (9%) correspond either to singlets (only one gene is observed [6%]) or doublets (one gene is missing in one strain [3%]). Finally, 12 rows (<1%) correspond to fragmented genes (in-frame stop codon), leading to two or more genes per strain in the row. Out of the 888 triplets, 818 (93%) exhibit sequence identity of 95% or more at the amino acid level. Analysis at the nucleotide level (see Table S1 in the supplemental material) indicates that the Welgevonden strains, Erwe and Erwo, are almost identical with respect to their coding sequences. CDS alignments reveal very few substitutions and almost no deletions. Exceptions are almost exclusively associated with recognized pseudogenes and will be presented later. On the other hand, several genes in

Erga and Erwe display a sufficient number of differences to allow for the analysis of selection pressure based on synonymous versus nonsynonymous (S/NS) substitution rates: 181 pairs of orthologs display a significantly larger amount of synonymous substitutions, indicating a strong selection pressure to maintain the protein sequence, whereas only three pairs (i.e., ERGA_CDS_00630/ERWE_CDS_00660, ERGA_CDS_05750/ERWE_CDS_05840, and ERGA_CDS_08580/ERWE_CDS_08680) display a significant amount of nonsynonymous substitutions, indicating putative ongoing pseudogenes or functional changes (see Table S1 in the supplemental material). Two of these CDSs correspond indeed to truncated CDSs (i.e., pseudogenes) in Erwe/Erwo. These three CDSs code for proteins of unknown function. To investigate putative differences in metabolic capabilities between Erga and Erwe/Erwo, genes associated with an EC number were specifically checked (see Table S1 in the supplemental material). Out of a total of 333 rows, 325 (98%) correspond to proteins having more than 95% identity. Six of the eight remaining rows correspond to pseudogenes resulting from duplications, whereas the last two candidates (ERGA_CDS_04370 and ERGA_CDS_05040) display a significant bias toward synonymous substitutions, suggesting that protein function is conserved. Virulence genes and membrane proteins are other candidate targets for investigating host range differences. The *vir* genes are organized in two separate operons (22, 51) with two paralog genes, i.e., *virB4* and *virB8*, located outside the operons (see Table S1 in the supplemental material). The *virB6* and both *virB4* genes display a larger number of synonymous substitutions, suggesting that selective pressure is maintaining their functional capabilities. *virD4*, *virB10*, and the paralogous *virB8* gene also display a large number of substitutions, but their S/NS rates are not sufficiently different to conclude that functional pressure is still acting. The other *vir* genes are highly similar in all strains. With respect to membrane proteins, *cpg1* makes a good candidate, with both high substitution and insertion/deletion rates in Erga and a significant substitution rate in Erwe (see Table S1 in the supplemental material). The other two *cpg*-related genes (i.e., ERGA_CDS_02490 and ERGA_CDS_02500) also display a significant substitution rate in Erga but not between Erwe and Erwo. All these changes are associated with an unbiased S/NS ratio (see Table S1 in the supplemental material). The cluster of paralogous *map1*-related genes (from ERGA_CDS_09000 to ERGA_CDS_09170) is another group of likely candidates since they display a large number of substitutions and insertions/deletions between Erga and Erwe/Erwo (see Table S1 in the supplemental material). Only *map1* and *map1-13* display significant selective pressure toward synonymous substitutions. *map1-2* is truncated (see below) in Erga and is therefore an additional candidate to explain host range differences. Interestingly, the *map1-1* gene from Erwe differs from those of both Erga and Erwo, which are identical (see Table S1 in the supplemental material).

Comparative analysis of unique CDSs. Fifty-seven unique CDSs are found within the three genomes. These unique CDSs are defined as sequences for which no predicted ortholog is found in the other genome. Twenty-eight unique CDSs are annotated only for Erwe, 7 are annotated only for Erwo, and 22 are annotated only for Erga. Careful examination of the differences between Erwo and Erwe shows that they are due

only to different annotation strategies (mostly prediction programs or parameters and definitions of pseudogenes). Therefore, 35 CDSs are specific to the Erwe/Erwo group, whereas 22 are found only in Erga. Only 6 out of these 57 CDSs correspond to major rearrangements in the other genome, i.e., complete or partial gene deletions and extensive mutations (see Table S2 in the supplemental material). The remaining 51 CDSs are unique because of in-frame stop codons in the corresponding sequences in the other strain, making them too short to reach the minimal open reading frame size set for prediction. Out of these 51 unique CDSs, 21 are truncated versions of full-length genes annotated elsewhere in the genome (see Table S2 in the supplemental material), and 5 other CDSs display similarity to known genes not found at full length in *E. ruminantium* (see Table S2 in the supplemental material). They might therefore be remnants of full-length genes eliminated through deletion or have been inserted through ancient horizontal transfer. The remaining 25 CDSs are unknown.

Comparative analysis of partial or fragmented CDSs. Occurrence of a stop codon may not lead to the complete loss of CDS prediction but to the detection of truncated genes, depending on the size of the remaining fragments. Truncated genes resulting in a single CDS are thereafter designated partial CDSs, whereas those resulting in two or more predicted CDS are designated fragmented CDSs; 29 such truncations were observed (see Table S2 in the supplemental material). Sequences were checked to eliminate sequencing errors. Seven genes are affected in all three genomes but differently, depending on the strain. Only one is known and encodes a putative type IV secreted protein (see Table S2 in the supplemental material).

A total of 18 CDS truncations differentiate the genome of Erga from that of Erwe/Erwo; 8 truncations appear in Erga (i.e., full-length CDSs are present in Erwe/Erwo), and 10 truncations appear in Erwe/Erwo (see Table S2 in the supplemental material). Only two genes have a known function. *map1-2* (truncated in Erga) bears a deletion of 48 bases (16 amino acids). This deletion accounts for 80% of the size difference between the *map1* clusters of Erga and Erwe/Erwo. Interestingly, *map1-2* was shown to have recombined with *map1-3* in a subset strain (strain CTVM) of Erga (9). In Erwe/Erwo, the only known gene to be affected is *ftsA*, which is 63 bp shorter than the Erga ortholog. *ftsA* codes for a protein involved in cell division. However, this does not seem to affect the ability of Erwe and Erwo to multiply and develop efficiently. The key-stone protein FtsZ, involved in the recruitment of cell division proteins, and the other Fts proteins (i.e., FtsH, FtsQ, and FtsY) are present and not truncated.

Finally, four CDS truncations (checked on Erwe chromatograms) also occurred between Erwe and Erwo (see Table S2 in the supplemental material). These CDSs are strictly identical in Erwo and Erga and are therefore specific to Erwe. Three of them have a known function. The first is *tufA*, coding for one of the elongation factors Tu. The second *tuf* gene (*tufB*) is present in all strains as a full-length gene. In *E. ruminantium*, the *tuf* operons display the chimeric organization described in *Rickettsia* spp. (6, 22, 59) with respect to *E. coli* (Fig. 2). *tufA* is flanked on one side by *rpsG-fusA* and by tRNA-Trp-*secE-nusG* on the other side (Fig. 2). This organization is intermediate between those of *E. coli* and *Rickettsia*. Like *Rickettsia*, *E.*

ruminantium bears the recombination between the *tufA* and *tufB* operons, but, like *E. coli*, it still bears the *tufA* gene, although fragmented in Erwe (6, 22, 59). The *tufB* operon of *E. ruminantium* is identical to those of *Rickettsia* spp., with *tufB* surrounded by tRNA-Tyr-tRNA-Gly and *rpsJ*. Furthermore, these regions are known to be prone to recombination and inversion (1, 6, 32). Interestingly, *Wolbachia* (*wBm*) displays a novel intermediate configuration. The *tufA* operon is identical to that of *E. coli*, whereas the *tufB* gene is flanked at its 5' end by only two of the *E. coli* tRNA genes (tRNA-Tyr and tRNA-Gly) and at its 3' end by tRNA-Trp-*secE-nusG*. This 3' flanking region is identical to that of the *tufA* operon in the other *Rickettsiales*. The *wBm secE* gene codes for a small protein of 69 amino acids which is part of the prokaryotic protein translocation system. This gene was not detected in the original *wBm* genome annotation (perhaps due to its small size), but the predicted sequence displays high similarity (>60%) to *secE* from *Ehrlichia*, *Anaplasma*, and *Wolbachia* (*wMel*), whereas it does not display similarity to the *E. coli secE* gene. The second truncated known gene in Erwe is *petC*, which is a key element in the cytochrome *bc*₁ complex. Despite interruption of both *tufA* and *petC*, Erwe remains highly virulent. If *tufA* is dispensable, as shown in *Rickettsia* spp. (59), other mechanisms might complement the absence of cytochrome *c*₁. The third known gene to be truncated in Erwe, *ftsK*, is also a key gene. This gene codes for the cell division protein FtsK, which is also a virulence-related gene in *R. prowazekii* (30). It displays a 135-bp deletion in its central part. The corresponding region in Erwo and Erga is a tandem repeat of four in-frame copies of 45 bp translated into the 15-amino-acid sequence LSDQDFEDESFADED. Erwe presents only one copy, and the missing segment corresponds exactly to the three other copies. This region has no sequence similarity with any FtsK protein. This deletion represents one of the very rare occurrences of tandem repeats in coding regions.

Analysis of genome size plasticity. Full genome alignments indicate that the differences in genome size between Erwe and Erwo (3 kb) and between Erga and Erwe (13 kb) are mostly due to intergenic regions. For the latter, the size difference associated with coding regions (CDSs) is only about 200 bp out of 13 kb. Furthermore, this difference is not due to specific large deletions but to small increments/decrements scattered almost uniformly along the genomes, with a slightly more stable region between 250 and 800 kb (data not shown). After removal of ONCR smaller than 10 bp (see Materials and Methods), 591 ONCR were identified between Erga and Erwe (see Materials and Methods), which correspond to a total of 383 kb (67% of the noncoding regions). Associating the pair of orthologous CDSs (OCR) located upstream from each ONCR resulted in 591 OCR covering 630 kb, i.e., 68% of the whole coding region (Table 2). Most OCR (90%) display the same sizes in both genomes ("0" class) reflecting the high conservation of the CDSs. The expanding/contracting regions are mostly noncoding regions with about one third (32%) of ONCR associated either with the "+" or the "-" class (Table 2). To correlate the expanding/contracting ONCR (size plasticity regions) with the presence of repeats, dispersed and tandem repeats were detected in all *Rickettsiales* genomes (Table 3). *E. ruminantium* displays the largest number of tandem repeats (Table 3). Moreover, these tandem repeats are

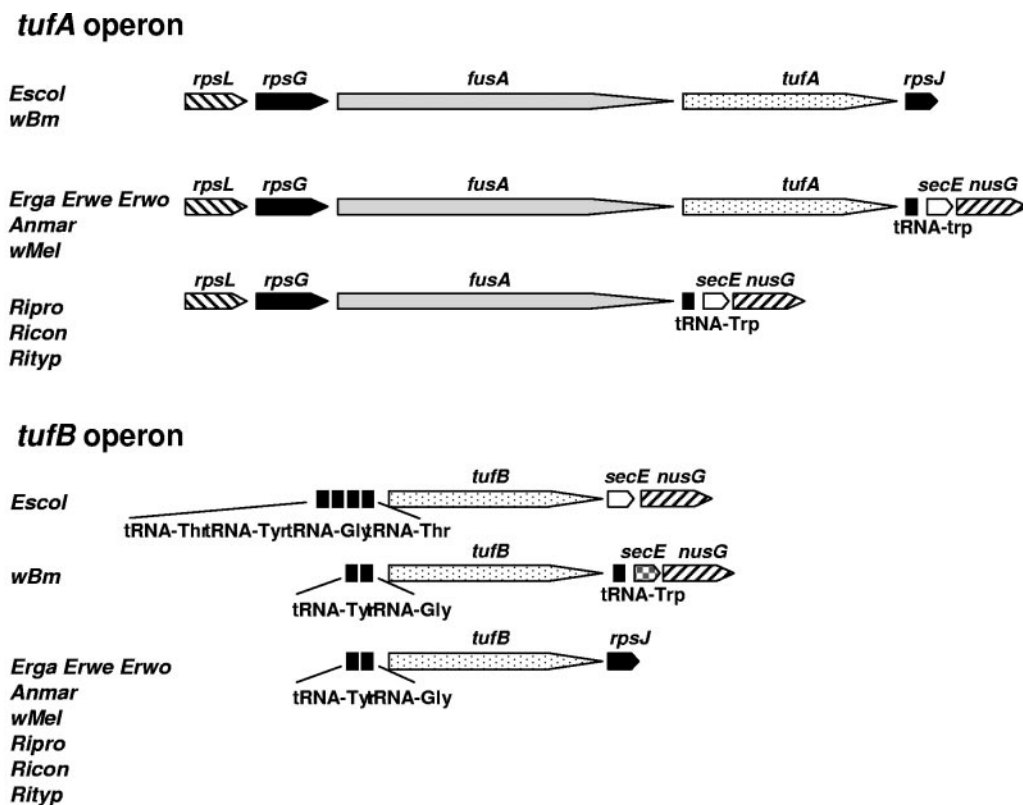


FIG. 2. Organization of the *tufA* and *tufB* operons in the *Rickettsiales* and *E. coli* genomes. The arrangement displayed by *E. coli* (*Escol*) is considered the ancestral organization of the *tuf* genes, whereas an intrachromosomal recombination event led to the shuffled *Rickettsiales* arrangement. Erga, Erwe, Erwo, *A. marginale* (*Anmar*), and *wMel* display the same organization as the *Rickettsia tuf* operons, except that *tufA* is still present (in a split form in Erwe). *Wolbachia* sp. (*wBm*) represents a novel intermediate configuration in which the *tufA* operon is identical to that of *E. coli* and not to other *Rickettsia* spp., whereas the organization of *tufB* is unique (see Discussion). CDSs are represented by arrows (not to scale) textured according to function. tRNA genes are represented by boxes.

mostly located in noncoding regions, i.e., from 59% to 70% (Table 3). By contrast, 60% to 81% of tandem repeats in the other *Rickettsiales* (except *R. typhi*) are located within coding regions (Table 3). The *E. ruminantium* group displays a very low rate of dispersed repeats compared to *A. marginale* and the *Wolbachiae* group. Furthermore, these dispersed repeats are mostly located in coding regions (Table 3). Size plasticity of ONCR is strongly related to the presence of tandem repeats (chi-square test *P* value of $<10^{-3}$) (Table 4). The “Tandem_Tandem” category (i.e., a tandem repeat observed both in Erga and in Erwe) represents the majority of ONCR containing a repeat (52%), and this category is associated with expansion/contraction of the genome (observed for “+” and

–,” 32; expected, 9.8; observed for the “0” class, 8; expected, 30.2). The “None_Tandem” category (i.e., no repeat in Erga and a tandem in Erwe) is associated with the “+” plasticity class (i.e., a longer ONCR in Erwe) (observed, 23; expected, 4.5). This is consistent with a loss of the tandem repeat in Erga (or gain in Erwe). Conversely, the “Tandem_None” class is

TABLE 2. Distribution of size plasticity classes within coding and noncoding regions

Size plasticity class ^a	No. (%)	
	Coding regions	Noncoding regions
+	31 (5.2)	117 (19.8)
–	28 (4.7)	72 (12.2)
0	532 (90.0)	402 (68.0)

^a 0, no size difference; +, the region is larger in Erwe; –, the region is larger in Erga.

TABLE 3. Distribution of repeats in complete genomes of *Rickettsiales*

Organism	Total no. of repeats (% of repeats within coding regions) ^a				
	Close direct	Close reverse	Distant direct	Distant reverse	Tandem
Erga	27 (37)	11 (64)	36 (69)	37 (65)	113 (41)
Erwe	29 (45)	11 (82)	38 (71)	33 (64)	134 (31)
Erwo	26 (42)	11 (82)	40 (70)	33 (56)	138 (30)
Anmar	92 (72)	2 (0)	315 (50)	325 (43)	43 (81)
wBm	3 (100)	1 (0)	226 (2)	229 (1)	9 (78)
wMel	54 (28)	15 (67)	4,480 (22)	4,610 (22)	50 (60)
Ricon	5 (80)	35 (11)	1,324 (21)	1,314 (21)	15 (67)
Ripro	2 (100)	2 (50)	5 (60)	2 (50)	21 (61)
Rityp	1 (100)	2 (50)	4 (25)	0 (0)	30 (43)

^a A repeat is considered to be located within a coding region if more than half its length overlaps the region. Close, copies are separated by less than 1 kb; direct, copies have the same orientation; reverse, copies have the reverse orientation; distant, copies are separated by more than 1 kb.

TABLE 4. Correlation between type of repeat and size plasticity of ONCR

Repeat	No. of repeats observed/expected for size plasticity class ^a			Total no. observed
	+	-	0	
None_None	46/82.6	34/43.5	434/387.9	514
None_Tandem	23/4.5	2/2.4	3/21.1	28
Tandem_None	2/1.4	6/0.8	1/6.8	9
Tandem_Tandem	24/6.4	8/3.4	8/30.2	40
Total no. observed	95	50	446	

^a 0, no size difference; +, the region is larger in Erwe; -, the region is larger in Erga.

associated with the “-” plasticity class that is consistent with a loss of the tandem repeat in Erwe (or gain in Erga).

Expansion/contraction of ONCR occurs by loss of tandem repeat units of ca. 150 bp. A tandem repeat is mainly characterized by two parameters: its period (the length of the repeat unit) and the number of copies. Figure 3a displays the distribution of the period for all tandem repeats found in the “Tandem_Tandem” class. The distribution is bimodal and indicates the presence of two distinct populations. The first population corresponds to small repeats of ca. 12 bp, which represent 15% of the total. The second population corresponds to long-period repeats ranging from 100 to 300 bp. They display a period centered on 150 bp, with the majority ranging between 125 and 175 bp in length. There is no significant difference between the period of the tandem repeats in Erga and in Erwe (correlation coefficient, 0.92 [data not shown]). However, a difference is observed in the number of copies. This suggests that expansion/contraction of ONCR occurs by loss/gain of tandem repeat units, i.e., by varying the number of copies rather than by insertion/deletion of fragments of different sizes. Plotting the theoretical expansion/contraction ratio, Δ_{theo} (see Materials

and Methods), versus the observed one, Δ_{obs} (Fig. 3b), indicates that both quantities correlate perfectly (correlation coefficient, 0.94). This suggests that the Erga and Erwe genomes evolved, with respect to size, by deletion or addition of a variable number of repeat units of similar size centered on 150 bp. Figure S3, in the supplemental material, displays several cases of deletions and expansions in Erga/Erwe/Erwo and *A. marginale*. The tandem repeats between *birA/gst* and *recR/znuA* illustrate the longer intergenic regions in *E. ruminantium* compared to *A. marginale*. These two tandems have periods of 155 bp and 178 bp, respectively, and there is one more copy in Erga than in Erwe/Erwo. The tandem repeat downstream from *ubiB* illustrates a case of complete tandem (four copies of 221 bp) deletion in the Welgevonden strains associated with a shortening of their intergenic regions; conversely, the tandem repeat downstream of ERGA_CDS_02630 (four copies of 187 bp) illustrates a tandem deletion in the Gardel strain. Figure S4a, in the supplemental material, illustrates a case of tandem expansion in Erwe: the period is 7 bases and Erwo exhibits 35 copies, whereas Erwe bears 64 copies, resulting in a 203-bp expansion $[(64 - 35) \times 7]$ of the Erwe intergenic region. Figure S4b, in the supplemental material, illustrates the converse case: the period is here 219 bp and five copies are present in Erwo, whereas Erwe has lost one copy, resulting in a 219-bp reduction. Although the tandem deletion/expansion process mostly affects noncoding regions, few could be observed within genes, as illustrated in the case of the *ftsK* gene.

DISCUSSION

The first feature revealed by this comparative analysis is the highly conserved genome organization between *Ehrlichia* spp. and *Anaplasma* spp. Previous phylogenetic analyses among the *Rickettsiales* have shown that *Ehrlichia* was perceived as being more closely related either to *Wolbachia* (60) or to *Anaplasma*

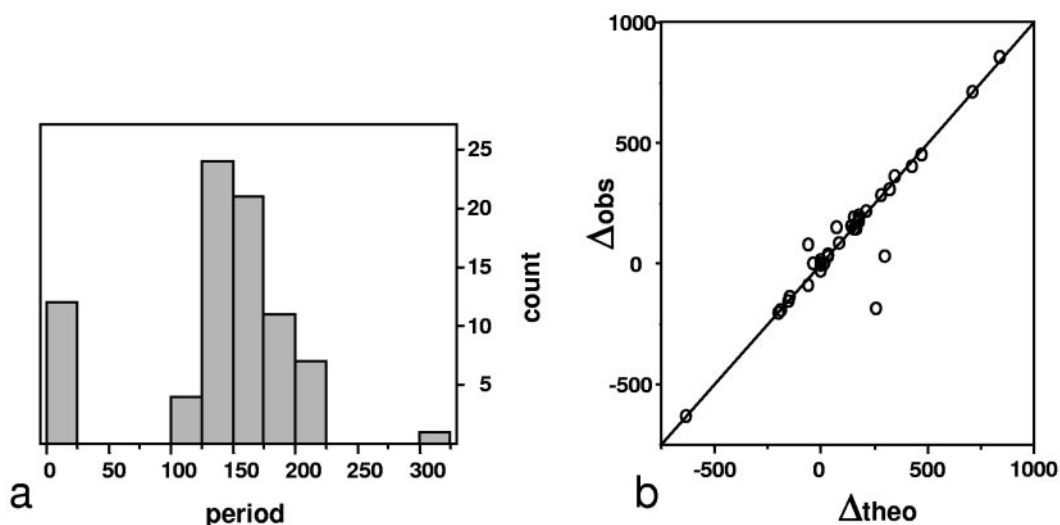


FIG. 3. Distribution of tandem repeats in intergenic regions. (a) Distribution of the periods of the tandem repeats found in expanding/contracting intergenic regions between *E. ruminantium* strains Gardel (Erga) and Welgevonden (Erwe). The distribution is clearly bimodal, with one population of short-period tandem repeats (~12 bp) and a second population of long-period tandem repeats (~150 bp). (b) Correlation between the observed differences in intergenic size (Δ_{obs}) and the values calculated by assuming that the differences are due solely to the different numbers of tandem repeat copies (Δ_{theo}).

(26). The conserved gene organization between *E. ruminantium* and *A. marginale* reported here definitely supports the second hypothesis and strongly suggests that *Ehrlichia* and *Anaplasma* are indeed very closely related bacteria. Moreover, the lack of synteny recorded with *Rickettsia* and *Wolbachia* indicates that *Ehrlichia* was separated long enough from both of these genera for complete gene shuffling to occur and most likely before its separation from *Anaplasma*. The small size of the genome and the lack of several key metabolic pathway genes (22) both indicate that *E. ruminantium* underwent genome shrinkage and massive gene loss in its adaptation to cell parasitism, a common feature in intracellular bacteria (6, 29, 43, 58, 70). This is most likely to have occurred in the common ancestor shared by *Ehrlichia*, *Rickettsia*, *Wolbachia*, and *Anaplasma*, which separated from a free-living α -proteobacterial branch (23, 66). The specific organization in the three genera of the *vir* genes in two separate operons supports this assumption (22, 51). The similar organization of the *tuf* operons in *Rickettsia*, *Wolbachia* (wMel), *E. ruminantium*, and *A. marginale* further indicates that this recombination probably occurred in their common ancestor. The *tufA* gene was further deleted in *Rickettsia* (6), whereas in *Wolbachia* (wMel), *Ehrlichia*, and *Anaplasma*, *tufA* remained, although fragmented in Erwe. The deletion of *tufA* in *Rickettsia* spp. (6), along with the virulence of Erwe, also confirms that *tufA* is dispensable in intracellular parasites, as shown in *R. prowazekii* (59). The conservation of the *vir* and *tuf* operons is relatively surprising, considering that the synteny observed between *Ehrlichia* and *Anaplasma* is globally lost with the other *Rickettsiales*. The evolution of *tuf* genes in *Rickettsia* is associated with palindromic elements named RPE (4, 5), but RPE are not found in *E. ruminantium* or *A. marginale*, indicating that this mechanism was not involved in the recombination and was probably inherited by *Rickettsia* after separation from the ancestral *Ehrlichia/Anaplasma* group.

Another feature to be underlined is the unusually long intergenic regions observed in *E. ruminantium*. Although there is a general trend toward longer intergenic regions in *Rickettsiales*, this trait is very pronounced in *E. ruminantium* and is particularly striking compared to *A. marginale*, which displays the short intergenic sequences usually observed in bacteria. Moreover, these long intergenic regions exhibit important size plasticity related to the presence of tandem repeats. This positive correlation is in good agreement with already-proposed mechanisms of tandem repeat deletion or amplification through DNA slippage (18, 38). Indeed, this RecA-independent mechanism nicely explains why the observed variations affect integral numbers of tandem copies and not the excision of the whole tandem by homologous recombination. Interestingly, although *E. ruminantium* bears the *recA* gene, as well as other genes involved in DNA repair, the RecA-independent mechanism seems to be favored. Large tandem repeats (over several hundred bases) can be produced and maintained under selection (18, 38). However when selection is removed they tend to collapse to single copy by RecA-dependent homologous recombination. Obviously, this phenomenon is here again not observed in *E. ruminantium*. This means either that some selection pressure is still active or that some elements of the excision mechanism are not fully functional. An interesting observation is that the deletion/expansion process is still very

active and occurs in a rather short time frame, since several cases are observable between Erwe and Erwo.

Comparative genomic analysis also provides hints for targeting genes potentially involved in the observed host range diversity. Membrane proteins such as Cpg- and Map1-related proteins are well-studied antigenic proteins (9, 34, 65), and comparative genomics may allow for more accurate targeting. Cpg proteins, and more specifically Cpg1, are clearly good candidates. Several Map1 proteins are also priority targets, as shown by their larger numbers of observed mutations between Erga and Erwe/Erwo. The best candidates are *map1* itself, as well as *map1-2* and *map1-6*. Interestingly, the *map1-1* gene, which was shown to be preferentially expressed in vectoring ticks (9), is the only *map1* gene to be strictly identical between Erga and Erwe while differing between Erwe and Erwo. The *vir* genes provide an additional set of candidates, in particular *virB6* and *virB4*, which display a large number of substitutions between Erga and Erwe/Erwo although they are clearly still subjected to selection pressure. This combination may reflect the need to adapt to a different host and cell environment while preserving key structure-function aspects of the proteins. By contrast, no significant difference appears between genes annotated with EC numbers whose products are involved in intermediate metabolism. It seems therefore unlikely that the metabolic capabilities significantly differ in the three strains.

The truncated genes observed in both Erga and Erwe/Erwo form another group of candidates. The few truncated genes with an identified function cannot explain the host range difference; therefore, candidates must be sought among the 24 unknown truncated genes. Similarly, the new potential membrane proteins of unknown function characterized by high Phe and Trp content comprise a final group of target candidates. However, since nothing is known about their function or localization, transcriptomics and proteomics experiments should thus be further considered to determine which genes are active and whether any differential expression occurs between Erga and Erwe/Erwo. Extending this analysis to the genomes of related bacteria will allow for the identification of genes involved in key mechanisms such as pathogenesis, virulence, host range, and protection and will thus contribute to a better understanding of the biology of the *Rickettsiales*.

The apparent ongoing process of genome plasticity, characterized by permanent deletion-insertion of tandem repeats and occurrence of gene truncations, clearly differentiates *E. ruminantium* from other intracellular bacteria, for which genome stability following initial size reduction is a key feature (11, 37, 44, 46, 61). *E. ruminantium* seems to be capable of rapidly undergoing genomic rearrangements upon exposure to a novel environment, which may thus explain the poor field efficacy of vaccines. The presence of truncated genes in Erwe, while they remained intact in both Erga and Erwo, after a change in cell environment may illustrate this phenomenon. This might be further exemplified both by the recent development of an attenuated phenotype of the Welgevonden strain by propagation in an unusual cell environment, i.e., a canine macrophage-monocyte cell line (71), and by the identification of a recombination event between the *map1-2* and *map1-3* genes of CTVM-Gardel (9) following a modification of the cell environment. Experimental evolution and comparative genomic analysis could help answer this question. Nevertheless, genome

plasticity and the related strain diversity definitely affect both strain-specific diagnostic and vaccine strategies. Therefore, a deeper understanding of this mechanism and its potential for variability, as well as of the minimal genome required for survival, is a prerequisite for development of novel vaccines and diagnostic tools.

ACKNOWLEDGMENTS

The Welgevonden strain was kindly provided by Durr Bezuidenhout (Onderstepoort Veterinary Institute, South Africa) in 1988 in exchange for the Gardel strain from CIRAD.

This work was supported by CIRAD-CNRS grant 751745/00.

REFERENCES

- Abdulkarim, F., and D. Hughes. 1996. Homologous recombination between the *tuf* genes of *Salmonella typhimurium*. *J. Mol. Biol.* **260**:506–522.
- Achaz, G., P. Netter, and E. Coissac. 2001. Study of intrachromosomal duplications among the eukaryote genomes. *Mol. Biol. Evol.* **18**:2280–2288.
- 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.
- Amiri, H., C. M. Alsmark, and S. G. E. Andersson. 2002. Proliferation and deterioration of *Rickettsia* palindromic elements. *Mol. Biol. Evol.* **19**:1234–1243.
- Amiri, H., W. Davids, and S. G. E. Andersson. 2003. Birth and death of orphan genes in *Rickettsia*. *Mol. Biol. Evol.* **20**:1575–1587.
- Andersson, S. G., and C. G. Kurland. 1998. Reductive evolution of resident genomes. *Trends Microbiol.* **6**:263–268.
- 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.
- Barré, N., G. Uilenberg, P. C. Morel, and E. Camus. 1987. Danger of introducing heartwater onto the American mainland: potential role of indigenous and exotic *Amblyomma* ticks. *Onderstepoort J. Vet. Res.* **54**:406–416.
- Bekker, C. P. J., M. Postigo, A. Taoufik, L. Bell-Sakyi, C. Ferraz, D. Martinez, and F. Jongejan. 2005. Transcription analysis of the major antigenic protein 1 multigene family of three in vitro cultured *Ehrlichia ruminantium* isolates. *J. Bacteriol.* **187**:4782–4791.
- Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**:573–580.
- Berg, O. G., and C. G. Kurland. 2002. Evolution of microbial genomes: sequence acquisition and loss. *Mol. Biol. Evol.* **19**:2265–2276.
- Bezuidenhout, J. D. 1989. Cowdria vaccines, p. 31–42. *In* I. G. Wright (ed.), *Veterinary protozoan and hemoparasite vaccines*. CRC Press, Inc., Boca Raton, Fla.
- Bezuidenhout, J. D., C. L. Paterson, and B. J. Barnard. 1985. In vitro cultivation of *Cowdria ruminantium*. *Onderstepoort J. Vet. Res.* **52**:113–120.
- Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. 2003. The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**:365–370.
- Borodovsky, M., and J. McIninch. 1993. GENEMARK: parallel recognition for both DNA strands. *Comput. Chem.* **17**:123–133.
- 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. *Proc. Natl. Acad. Sci. USA* **102**:844–849.
- Burridge, M. J. J., L. A. Simmons, T. F. Peter, and S. M. Mahan. 2002. Increasing risks of introduction of heartwater onto the American mainland associated with animal movements. *Ann. N. Y. Acad. Sci.* **969**:269–274.
- Bzimek, M., and S. T. Lovett. 2001. Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. USA* **98**:8319–8325.
- Cole, S. T., K. Eiglmeier, J. Parkhill, K. D. James, N. R. Thomson, P. R. Wheeler, N. Honore, T. Garnier, C. Churcher, D. Harris, K. Mungall, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. M. Davies, K. Devlin, S. Duthoy, T. Feltwell, A. Fraser, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, C. Lacroix, J. Maclean, S. Moule, L. Murphy, K. Oliver, M. A. Quail, M. A. Rajandream, K. M. Rutherford, S. Rutter, K. Seeger, S. Simon, M. Simmonds, J. Skelton, R. Squares, S. Squares, K. Stevens, K. Taylor, S. Whitehead, J. R. Woodward, and B. G. Barrrell. 2001. Massive gene decay in the leprosy bacillus. *Nature* **409**:1007–1111.
- Collins, N. E., A. Pretorius, M. Van Kleef, K. A. Brayton, E. Zweggarth, and B. A. Allsopp. 2003. Development of improved vaccines for heartwater. *Ann. N. Y. Acad. Sci.* **990**:474–484.
- Collins, N. E., E. P. De Villiers, K. A. Brayton, and B. A. Allsopp. 1998. DNA sequence of a cosmid clone of *Cowdria ruminantium*. *Dev. Biol.* **114**:365–368.
- Collins, N. E., J. Liebenberg, E. P. De Villiers, K. A. Brayton, E. Louw, A. Pretorius, F. E. Faber, H. Van Heerden, A. Josemans, M. Van Kleef, H. C. Steyn, M. F. van Strijp, E. Zweggarth, F. Jongejan, J. C. Maillard, D. Berthier, M. Botha, F. Joubert, C. H. Corton, N. R. Thomson, M. T. Allsopp, and B. A. Allsopp. 2005. The genome of the heartwater agent *Ehrlichia ruminantium* contains multiple tandem repeats of actively variable number. *Proc. Natl. Acad. Sci. USA* **102**:838–843.
- Dumler, J. S., A. F. Barbet, C. P. Bekker, G. A. Dasch, G. H. Palmer, S. C. Ray, Y. Rikihisa, and F. R. Rurangirwa. 2001. Reorganization of genera in the families *Rickettsiaceae* and *Anaplasmataceae* in the order *Rickettsiales*: unification of some species of *Ehrlichia* with *Anaplasma*, *Cowdria* with *Ehrlichia* and *Ehrlichia* with *Neorickettsia*, descriptions of six new species combinations and designation of *Ehrlichia equi* and 'HGE agent' as subjective synonyms of *Ehrlichia phagocytophila*. *Int. J. Syst. Evol. Microbiol.* **51**:2145–2165.
- Du Plessis, J. L. 1985. A method for determining the *Cowdria ruminantium* infection rate of *Amblyomma hebraeum*: effects in mice injected with tick homogenates. *Onderstepoort J. Vet. Res.* **52**:55–61.
- Durand, P., C. Medigue, A. Morgat, Y. Vandenbrouck, A. Viari, and F. Rechenmann. 2003. Integration of data and methods for genome analysis. *Curr. Opin. Drug Discov. Devel.* **6**:346–352.
- Fenollar, F., B. La Scola, H. Inokuma, J. S. Dumler, M. J. Taylor, and D. Raoult. 2003. Culture and phenotypic characterization of a *Wolbachia pipipentis* isolate. *J. Clin. Microbiol.* **41**:5434–5441.
- Foster, J., M. Ganatra, I. Kamal, J. Ware, K. Makarova, N. Ivanova, A. Bhattacharyya, V. Kapatral, S. Kumar, J. Posfai, T. Vincze, J. Ingram, L. Moran, A. Lapidus, M. Omelchenko, N. Kyrpides, E. Ghedin, S. Wang, E. Goldsman, V. Joukov, O. Ostrovskaya, K. Tsukerman, M. Mazur, D. Comb, E. Koonin, and B. Slatko. 2005. The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol.* **3**:599–614.
- Grigoriev, A. 1998. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* **26**:2286–2290.
- Harrison, P. M., and M. Gerstein. 2002. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **318**:1155–1174.
- Hong, G., E. C. Yao-Yu, Z. Shuping, T. Min, T. Mong-Hsun, J. T. Joseph, L. R. Allen, and C. Wei-Mei. 2004. Comparative genomics of *Rickettsia prowazekii* Madrid E and Breinl strains. *J. Bacteriol.* **186**:556–565.
- Horn, M., A. Collingro, S. Schmitz-Esser, C. L. Beier, U. Purkhöld, B. Fartmann, V. Brandt, G. J. Nyakatura, M. Droege, D. Frishman, T. Rattei, H. W. Mewes, and M. Wagner. 2004. Illuminating the evolutionary history of *Chlamydiae*. *Science* **304**:728–730.
- Hughes, D. 2000. Co-evolution of the *tuf* genes links gene conversion with the generation of chromosomal inversions. *J. Mol. Biol.* **297**:355–364.
- Jongejan, F. 1991. Protective immunity to heartwater (*Cowdria ruminantium* infection) is acquired after vaccination with in vitro-attenuated rickettsiae. *Infect. Immun.* **59**:729–731.
- Jongejan, F., N. De Vries, J. Nieuwenhuijs, L. A. Wassink, and A. H. M. Van Vliet. 1993. The immunodominant 32-kilodalton protein of *Cowdria ruminantium* is conserved within the genus *Ehrlichia*. *Rev. Elev. Med. Vet. Trop.* **46**:145–152.
- Karlin, S., G. Ghandour, F. Ost, S. Tavares, and L. J. Korn. 1983. New approaches for computer analysis of nucleic acid sequences. *Proc. Natl. Acad. Sci. USA* **80**:5660–5664.
- Katz, J. B., R. DeWald, J. E. Dawson, E. Camus, D. Martinez, and R. Mondry. 1997. Development and evaluation of a recombinant antigen, monoclonal antibody-based competitive ELISA for heartwater serodiagnosis. *J. Vet. Diagn. Investig.* **9**:130–135.
- Klasson, L., and S. G. E. Andersson. 2004. Evolution of minimal-gene-sets in host-dependent bacteria. *Trends Microbiol.* **12**:37–43.
- Lovett, S. T. 2004. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol. Microbiol.* **52**:1243–1253.
- Martinez, D., N. Vachieri, F. Stachurski, Y. Kandassamy, M. Raliniaina, R. Aprelon, and A. Gueye. 2005. Nested PCR for the detection and genotyping of *Ehrlichia ruminantium*. Use in genetic diversity analysis. *Ann. N. Y. Acad. Sci.* **1026**:106–113.
- Martinez, D., J. C. Maillard, S. Coisne, C. Sheikboudou, and A. Bensaid. 1994. Protection of goats against heartwater acquired by immunisation with inactivated elementary bodies of *Cowdria ruminantium*. *Vet. Immunol. Immunopathol.* **41**:153–163.
- Martinez, D., J. Swinkels, E. Camus, and F. Jongejan. 1990. Comparison between 3 antigens for the serodiagnosis of heartwater disease by indirect immunofluorescence. *Rev. Elev. Med. Vet. Pays Trop.* **43**:159–166.
- McLeod, M. P., X. Qin, S. E. Karpathy, J. Gioia, S. K. Highlander, G. E. Fox, T. Z. McNeill, H. Jiang, D. Muzny, L. S. Jacob, A. C. Hawes, E. Sodergren, R. Gill, J. Hume, M. Morgan, G. Fan, A. G. Amin, R. A. Gibbs, C. Hong, X. J. Yu, D. H. Walker, and G. M. Weinstock. 2004. Complete genome sequence

- of *Rickettsia typhi* and comparison with sequences of other rickettsiae. *J. Bacteriol.* **186**:5842–5855.
43. Mira, A., H. Ochman, and N. A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**:589–596.
 44. Mira, A., L. Klasson, and S. G. E. Andersson. 2002. Microbial genome evolution: sources of variability. *Curr. Opin. Microbiol.* **5**:506–512.
 45. Mondry, R., D. Martinez, E. Camus, A. Liebisch, J. B. Katz, R. Dewald, A. H. van Vliet, and F. Jongejan. 1998. Validation and comparison of three enzyme-linked immunosorbent assays for the detection of antibodies to *Cowdria ruminantium* infection. *Ann. N. Y. Acad. Sci.* **849**:262–272.
 46. Moran, N. A., and A. Mira. 2001. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol.* **2**:RESEARCH0054.1–0054.12. [Online.] <http://genomebiology.com/2001/2/12/RESEARCH/0054>.
 47. Moran, N. A., and G. R. Plague. 2004. Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.* **14**:627–633.
 48. Mukhebi, A. W., T. Chamboko, C. J. O'Callaghan, T. F. Peter, R. L. Kruska, G. F. Medley, S. M. Mahan, and B. D. Perry. 1999. An assessment of the economic impact of heartwater (*Cowdria ruminantium* infection) and its control in Zimbabwe. *Prev. Vet. Med.* **39**:173–189.
 49. Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**:418–426.
 50. Ogata, H., S. Audic, P. Renesto-Audiffren, P. E. Fournier, V. Barbe, D. Samson, V. Roux, P. Cossart, J. Weissenbach, J. M. Claverie, and D. Raoult. 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science* **293**:2093–2098.
 51. Ohashi, N., N. Zhi, Q. Lin, and Y. Rikihisa. 2002. Characterization and transcriptional analysis of gene clusters for a type IV secretion machinery in human granulocytic and monocytic ehrlichiosis agents. *Infect. Immun.* **70**:2128–2138.
 52. Pascal, G., C. Medigue, and A. Danchin. 2005. Universal biases in protein composition of model prokaryotes. *Proteins* **60**:27–35.
 53. Pegram, R. P., E. F. Gersabeck, D. Wilson, and J. W. Hansen. 2002. Eradication of the tropical bont tick in the Caribbean: is the Caribbean *Amblyomma* Program in a crisis? *Ann. N. Y. Acad. Sci.* **969**:297–305.
 54. Perez, J. M., D. Martinez, A. Debus, C. Sheikboudou, and A. Bensaid. 1997. Detection of genomic polymorphisms among isolates of the intracellular bacterium *Cowdria ruminantium* by random amplified polymorphic DNA and Southern blotting. *FEMS Microbiol. Lett.* **154**:73–79.
 55. Perriere, G., and J. Thioulouse. 2003. Use of correspondence discriminant analysis to predict the subcellular location of bacterial proteins. *Comput. Methods Programs Biomed.* **70**:99–110.
 56. Provost, A., and J. D. Bezuidenhout. 1987. The historical background and global importance of heartwater. *Onderstepoort J. Vet. Res.* **54**:165–169.
 57. Rocha, E. P. C., A. Danchin, and A. Viari. 1999. Universal replication biases in bacteria. *Mol. Microbiol.* **32**:11–16.
 58. Stepkowski, T., and B. Legocki. 2001. Reduction of bacterial genome size and expansion resulting from obligate intracellular lifestyle and adaptation to soil habitat. *Acta Biochim. Pol.* **48**:367–381.
 59. Syvanen, A. C., H. Amiri, A. Jamal, S. G. Andersson, and C. G. Kurland. 1996. A chimeric disposition of the elongation factor genes in *Rickettsia prowazekii*. *J. Bacteriol.* **178**:6192–6199.
 60. Taillardat-Bisch, A. V., D. Raoult, and M. Drancourt. 2003. RNA polymerase beta-subunit-based phylogeny of *Ehrlichia* spp., *Anaplasma* spp., *Neorickettsia* spp. and *Wolbachia pipientis*. *Int. J. Syst. Evol. Microbiol.* **53**:455–458.
 61. Tamas, I., L. Klasson, B. Canback, A. K. Naslund, A. S. Eriksson, J. J. Wernegreen, J. P. Sandstrom, N. A. Moran, and S. G. E. Andersson. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**:2376–2379.
 62. Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**:41–54.
 63. Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. *Science* **278**:631–637.
 64. Uilenberg, G., E. Camus, and N. Barré. 1985. Some observations on a stock of *Cowdria ruminantium* from Guadeloupe (French West Indies). *Rev. Elev. Med. Vet. Pays Trop.* **34**:34–42.
 65. Van Vliet, A. H., B. A. van der Zeijst, E. Camus, S. M. Mahan, D. Martinez, and F. Jongejan. 1995. Use of a specific immunogenic region on the *Cowdria ruminantium* MAP1 protein in a serological assay. *J. Clin. Microbiol.* **33**:2405–2410.
 66. Weisburg, W. G., S. M. Barns, D. A. Pelletier, and D. J. Lane. 1991. 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* **173**:697–703.
 67. Wu, M., L. V. Sun, J. J. Vamathevan, M. Riegler, R. T. DeBoy, J. C. Brownlie, E. A. McGraw, W. Martin, C. Esser, N. Ahmadijad, C. Wiegand, R. Madupu, M. J. Beanan, L. M. Brinkac, S. C. Daugherty, A. S. Durkin, J. F. Kolonay, W. C. Nelson, Y. Mohamoud, P. Lee, K. Berry, M. B. Young, T. Utterback, J. Weidman, W. C. Nierman, I. T. Paulsen, K. E. Nelson, H. Tettelin, S. L. O'Neill, and J. A. Eisen. 2004. Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. *PLoS Biol.* **2**:327–341.
 68. Wuyts, J., G. Perrière, and Y. Van de Peer. 2004. The European ribosomal RNA database. *Nucleic Acids Res.* **32**:D101–D103.
 69. Zhang, J., S. Kumar, and M. Nei. 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Mol. Biol. Evol.* **14**:1335–1338.
 70. Zomorodipour, A., and S. G. Andersson. 1999. Obligate intracellular parasites: *Rickettsia prowazekii* and *Chlamydia trachomatis*. *FEBS Lett.* **452**:11–15.
 71. Zwegarth, E., A. I. Josemans, M. F. Van Strijp, L. Lopez-Rebollar, M. Van Kleef, and B. A. Allsopp. 2005. An attenuated *Ehrlichia ruminantium* (Welgevonden stock) vaccine protects small ruminants against virulent heartwater challenge. *Vaccine* **23**:1695–1702.