ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Gene 387 (2007) 31 – 37

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Homologous recombination and the pattern of nucleotide substitution in *Ehrlichia ruminantium*

Austin L. Hughes *, Jeffrey O. French

Department of Biological Sciences, University of South Carolina, Coker Life Sciences Bldg., 700 Sumter Street, Columbia SC 29208, USA

Received 28 June 2006; received in revised form 31 July 2006; accepted 1 August 2006
Available online 23 August 2006
Received by S. Yokoyama

Abstract

Patterns of nucleotide substitution at orthologous loci were examined between three genomes of *Ehrlichia ruminantium*, the causative agent of heartwater disease of ruminants. The most recent common ancestor of two genomes (Erwe and Erwo) belonging to the Welgevonden strain was estimated to have occurred 26,500–57,000 years ago, while the most recent common ancestor of these two genomes and the Erga genome (Gardel strain) was estimated to have occurred 2.1–4.7 million years ago. The search for genes showing extremely high values of the number of synonymous substitutions per site was used to identify genes involved in past homologous recombination. The most striking case involved the *map1* gene, encoding major antigenic protein-1; evidence for homologous recombination is consistent with previous phylogenetic analysis of *map1* alleles. At this and certain other loci, homologous recombination may have contributed to the evolution of host–pathogen interactions. In addition, comparison of the patterns of synonymous and nonsynonymous substitution provided evidence for positive selection favoring a high level of amino acid change between the Welgevonden and Gardel strains at a locus of unknown function (designated Erum4340 in the Erwo genome).

Keywords: Bacterial evolution; *Ehrlichia ruminantium*; Homologous recombination; Nucleotide substitution

1. Introduction

The bacterium *Ehrlichia ruminantium* (Proteobacteria: Alphaproteobacteria: Ricketssiales) is a tick-transmitted obligate intracellular pathogen of endothelial cells that causes heartwater, a disease of wild and domestic ruminants. This bacterium is endemic to sub-Saharan Africa, and has been introduced to several of the Caribbean islands (Uilenberg, 1983). Recently complete genomes of three isolates of *E. ruminantium* have been sequenced: (1) the Gardel strain (abbreviation: Erga) from the island of Guadeloupe; (2) an isolate of the Welgevonden strain (abbreviation: Erwe) isolated in South Africa and maintained in cell culture in Guadeloupe; and (3) the original field-isolated Welgevonden strain (abbreviation: Erwo), maintained in South Africa (Frutos et al., 2006). Differences among these genomes included differential truncation of genes, as well as substantial sequence length variation in non-coding regions, due to expansion or contraction of tandem repeats (Frutos et al., 2006).

The three genomes of *E. ruminantium* were reported to show a high level of conservation of gene order, but relatively little attention has been paid to the extent of nucleotide sequence divergence among these genomes at orthologous loci (i.e., loci related by descent from a common ancestral locus without gene duplication). One interesting finding was an evidence for an apparent interlocus recombination event that took place in a laboratory isolate of the Gardel strain (Bekker et al., 2005). However, the pattern of nucleotide substitution at a genome-wide scale can provide evidence of long-term evolutionary history, including evidence for divergence time and for the roles of natural selection and recombination (Hughes and Friedman, 2004, 2005; Hughes et al., 2002).

Abbreviations: E. ruminantium, *Ehrlichia ruminantium*; d_S, the number of synonymous nucleotide substitutions per synonymous site; d_N, the number of nonsynonymous nucleotide substitutions per nonsynonymous site; AT%, percent A+T.

* Corresponding author. Tel.: +1 803 777 9186; fax: +1 803 777 4002.
E-mail address: austin@biol.sc.edu (A.L. Hughes).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.08.003

Recombinational mechanisms are believed to have played an important role in the evolution of prokaryotes (Andwalla, 2003; Vetsigian and Goldenfeld, 2005). These mechanisms can be placed in two broad categories: (1) non-homologous recombination or horizontal gene transfer (HGT), which involves the acquisition of a gene or genomic region new to a genome (Ochman et al., 2000); and (2) homologous recombination, which involves the introduction to a genome of a new allelic variant of a gene or genomic region already present in that genome (Hanage et al., in press). Homologous recombination among bacterial genomes was the subject of a number of studies because of its effect on phylogenies reconstructed from a selected set of genes (e.g., multi-locus enzyme electrophoresis; Lecointre et al., 1998). Few studies have examined the effect of homologous recombination at the genome-wide level, although the availability of a number of complete genomes from closely related bacterial isolates now makes such studies possible (Hughes and Friedman, 2004, 2005).

Examining the number of synonymous nucleotide substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) in comparison of orthologous protein-coding genes between pairs of closely related genomes can provide reliable indications of past recombination events (Hughes and Friedman, 2005). There is an expected substantial variation among genes with respect to d_N because different proteins are subject to widely different levels of constraint at the amino acid sequence level. Although functional constraints may exist at synonymous sites in at least certain bacterial species (Grantham et al., 1980), the strength of selection at synonymous sites is evidently much less than that at nonsynonymous sites (Bulmer, 1991). Thus, in the comparison of orthologous pairs of genes from two genomes having a common ancestor, it is expected that variance among genes with respect to d_S will be much lower than variance with respect to d_N. On the other hand, an unusually high value of d_S is an evidence of a gene sequence that is derived from a different genetic background than most genes in the genome. Thus, it provides evidence of past homologous recombination.

Here we use simple statistical methods based on linear regression to identify genes with unusually high d_S, thus providing evidence for genome-wide patterns of homologous recombination in the evolutionary history of *E. ruminantium*. In addition, we compare d_S and d_N values in order to identify genes that may have been subjected to positive Darwinian selection, leading to adaptive diversification of *E. ruminantium* genomes. Our analyses thus provide a fuller picture of genome evolution in this species and suggest avenues of empirical research in order to understand the role of key individual genes in the heartwater disease process.

2. Methods

2.1. Sequences analyzed

We downloaded from NCBI the sets of the predicted protein-coding genes from the following genomes of *E. ruminantium*: (1) the Gardel strain from the Caribbean island of Guadeloupe (abbreviation Erga; NC_006831); (2) the Welgevonden strain originating from South Africa but maintained in culture in Guadeloupe (abbreviation Erwe; NC_006832); and a South African strain closely related to Erwe (abbreviation Erwo; NC_005295). In order to assign genes to gene families, we applied the BLASTCLUST program (Altschul et al., 1997), which assembles families by homology search using a single-link method, to the sets of predicted protein sequences from the above-mentioned genomes. In order to identify a set of putative orthologs found in a single copy in each of the three genomes, we used an *E* value of 10^{-6} for the BLASTP homology search and BLAST score density of 1.75 over 90% of the proteins' length. These criteria identified 818 gene families represented by a single member in each genome.

2.2. Nucleotide substitution

Homologous sequences were aligned at the amino acid level using the CLUSTAL W program (Thompson et al., 1994), and the amino acid sequence alignment was imposed on the DNA sequences. The number of synonymous nucleotide substitutions per synonymous site (d_S) and the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) were estimated by the maximum likelihood method of Yang and Nielsen (2000) using the software package PAML (Yang, 1997) (For all values, see Supplementary Table S1).

We also estimated d_S and d_N using Nei and Gojobori's (1986) and Li's (1993) methods (not shown). The results of all three methods were nearly identical. For the comparison between Erwo and Erwe genomes, the correlation between d_N values estimated by Yang and Nielsen's (YN) method and those estimated by Nei and Gojobori's (NG) method was 0.999 ($P<0.001$) and that between d_N values estimated by the YN method and Li's method was 0.996 ($P<0.001$). Similarly, the correlation between d_S values estimated by YN and those estimated by NG was 0.995 ($P<0.001$); and the correlation between d_S values estimated by YN and those estimated by Li's method was 0.945 ($P<0.001$). These results are not unexpected since these sequences were closely related (see Results section below), and thus a simple substitution model is expected to perform as well as a more complicated one (Nei and Kumar, 2000). Divergence time (t) estimates were derived from the formula $d_S=2\lambda t$, using estimations of the substitution rate (λ) obtained from the literature (see below).

In order to examine the possible effects of nucleotide content on the pattern of substitution, we examined nucleotide content at four-fold degenerate sites, since these are not subject to purifying selection on amino acid sequence. We examined the percentage of A+T (AT%) at 4-fold degenerate sites and TA skew, defined as (T−A)/(T+A) (Lobry, 1996). Whether biased nucleotide content is a result of a mutational bias (Chen et al., 2004), of selection on some property such as codon usage (Grantham et al., 1980), or of some combination of these factors, the effects of nucleotide content bias are expected to be more evident at synonymous sites than at nonsynonymous sites, since the latter will be subject to functional constraints on the amino acid sequence (Banerjee et al., 2005; Hughes, 1999). Thus, an expected consequence of biased nucleotide content at

Table 1
Summary of synonymous (d_S) and nonsynonumous (d_N) substitutions per site at 818 orthologous loci among three genomes of *Ehrlichia ruminantium*

Mean±S.E. (median; range) d_S		
	Erwo	Erga
Erwe	0.0005±0.0001 (0.0000; 0.0000–0.0531)	0.0388±0.0013 (0.0332; 0.0000–0.5933)
Erwo		0.0388±0.0013 (0.0330; 0.0000–0.6000)
Mean±S.E. (median; range) d_N		
	Erwo	Erga
Erwe	0.0005±0.0001 (0.0000; 0.0000–0.0285)	0.0049±0.0003 (0.0026; 0.0000–0.0712)
Erwo		0.0049±0.0003 (0.0027; 0.0000–0.0712)

synonymous sites will be a reduction in the observed rate of synonymous substitution (Sharp and Li, 1987; Eyre-Walker and Bulmer, 1995; Zhang et al., 2002).

2.3. Outlier identification

Outliers in linear regression were identified quantitatively by examining studentized residuals. The studentized residual is a *t*-statistic testing the hypothesis that the regression is improved by removing a given data point (Belser et al., 1980). Statistical analyses were conducted using the Minitab statistical package, release 13 (http://www.minitab.com/).

3. Results

3.1. Nucleotide substitution among genomes

The mean and median values of d_S and d_N in pairwise comparison of 818 orthologous genes shared by the Erwe, Erwo, and Erga genomes are shown in Table 1. As expected, Erwe and Erwo were much more similar to each other at both synonymous and nonsynonymous sites than either was to Erga (Table 1). In addition, the mean and median d_S and d_N between Erwe and Erga were identical to those between Erwo and Erga or nearly so (Table 1). In order to estimate the divergence times among these genomes, we used estimates for the rate of synonymous substitution ($4.4–4.7 \times 10^{-9}$ substitutions per site per year) obtained from comparison of *Escherichia coli* and *Salmonella typhimurium*, assumed to have diverged 100 million years ago (Ochman et al., 1999; Sharp, 1991; Smith and Eyre-Walker, 2001). Applying these rates to our data on mean d_S between genomes, we estimate the most recent common ancestor of Erwe and Erwo to have occurred 53,000–57,000 years ago and the most recent common ancestor of Erwe, Erwo, and Erga to have occurred 4.1–4.4 million years ago. However, there is an evidence of the mutation rate and the nucleotide substitution rate at synonymous sites in endosymbiotic members of the Proteobacteria (such as *Buchnera*) than it is in *E. coli* and *Salmonella* (Ochman et al., 1999). Since *E. ruminantium* is obligately intracellular, it may likewise be subject to an enhanced mutation rate. Assuming that, like that of *Buchnera* (Itoh et al., 2002), the mutation rate in *E. ruminantium* is approximately twice that in *E. coli* and *Salmonella*, the most recent common ancestor of Erwe and Erwo would be estimated to have occurred 26,500–28,500 years ago; and the common ancestor of Erwe, Erwo, and Erga would be estimated to have occurred 2.2–2.2 million years ago.

3.2. Homologous recombination in the Erwe and Erwo lineages

Assuming that Erwe and Erwo are more closely related to each other than either is to Erga, events of homologous recombination giving rise to an unusual pattern of synonymous substitution can be classified as follows: (a) recombination in Erwe after its common ancestor with Erwo; (b) recombination in Erwo after its common ancestor with Erwe; and (c) recombination on the branch between Erwe and Erwo, on the one hand, and Erga on the other hand (Fig. 1). If recombination of type (a) has introduced a divergent allele into Erwe, high d_S between Erwe and Erga at that locus should be unusually high in comparison to d_S between Erwo and Erga at the same locus. Conversely, if recombination occurred between Erwe and Erga or a genome closely related to the latter, d_S between Erwe and Erga should be unusually low. Likewise, if recombination of type (b) has introduced a divergent allele into Erwo, d_S between Erwo and Erga should be unusually high in comparison to that between Erwe and Erga. Again, if recombination occurred between Erwo and Erga or a genome closely related to the latter, d_S between Erwo and Erga should be unusually low.

In order to detect loci with these unusual substitution patterns, d_S values in Erwe–Erga comparison were compared to those in Erwo–Erga comparison for the 818 loci. The linear correlation between the two sets of values was very close ($R^2 = 99.0\%$; $P < 0.001$; Fig. 2A). However, there were visible outliers from the overall trend (Fig. 2A). Outliers were identified quantitatively by examining studentized residuals. Plotting studentized residuals against gene location in Erwo revealed that residuals with high absolute value were scattered throughout the genome (Fig. 2B).

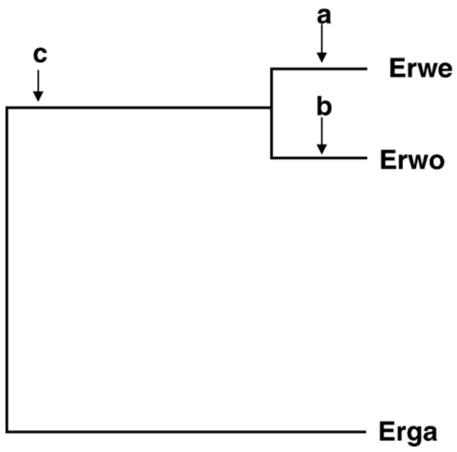

Fig. 1. Schematic representation of the phylogenetic relationships of three *Ehrlichia ruminantium* genomes, illustrating timing of possible events (a, b, and c) of homologous recombination.

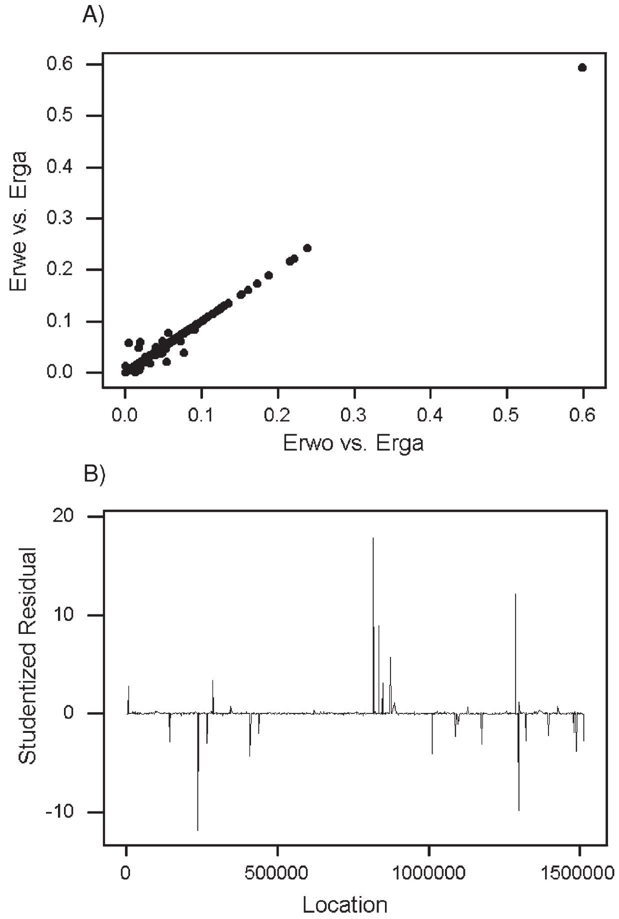

Fig. 2. (A) Plot of d_S in Erwe vs. Erga comparison vs. d_S in Erwo vs. Erga comparison at 818 orthologous loci. (B) Studentized residuals from the regression of d_S in Erwe vs. Erga comparison vs. d_S in Erwo vs. Erga comparison plotted as a function of gene location (start site) in the Erwo genome.

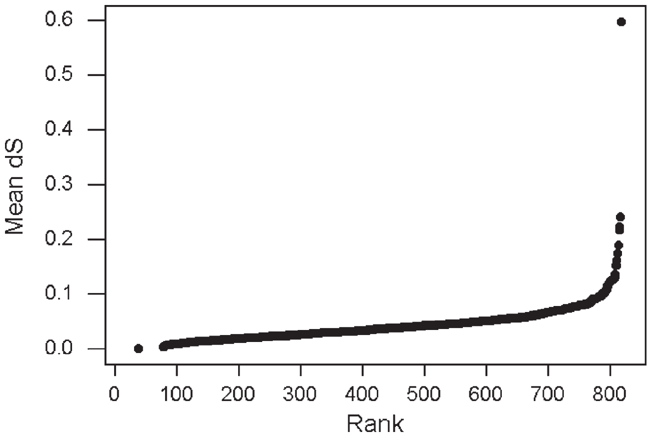

Fig. 3. Plot of mean d_S values in Erwo vs. Erga and Erwe vs. Erga comparison vs. their rank.

We tested the significance of studentized residuals using the Bonferroni procedure to correct for multiple testing; there were eight loci showing statistically significant studentized residuals using this conservative test (Table 2). High positive residuals identified loci at which d_S between Erwe and Erga was unusually high in comparison to that between Erwo and Erga. At the *pdhA* locus, neither d_S between Erwo and Erga (0.0198) nor d_S between Erwe and Erga (0.0593) was unusually low (Table 2). However, the former was much closer to the median value (0.0332; Table 1) than was the latter, suggesting homologous recombination between Erwo and a divergent genome at this locus. On the other hand, at the *nuoM* locus, the d_S value between Erwe and Erga was somewhat high (0.0578), but that between Erwo and Erga (0.0044) was unusually low (Table 2). At this locus, the results support a recombination event in which an allele similar to that of Erga was donated to Erwo after the latter's last common ancestor with Erwe, while Erwe may possibly also have recombined with a divergent genotype at the same locus.

We used negative residuals of high absolute value as a way of identifying loci at which d_S between Erwo and Erga was unusually high in comparison to d_S between Erwe and Erga. For example, at the *proP* locus, d_S between Erwe and Erga (0.0376) was close to the median value (Table 1), whereas that between Erwo and Erga was over twice as high (0.0766), suggesting a recombination between Erwo and a divergent genotype at this locus.

3.3. Ancient events of homologous recombination

In order to identify potential events of homologous recombination of the type designated (c) in Fig. 1, we searched for loci at which both Erwe and Erwo showed unusually high d_S in comparison with Erga. As a simple way of identifying such loci, we first computed the mean of d_S between Erwe and Erga and d_S between Erwo and Erga. Next we ranked the 818 values of mean d_S from least to greatest. Finally, we plotted each mean d_S value against its rank (Fig. 3). There was a significant linear

Table 2
Genes with synonymous divergence from Erga showing strong asymmetry between Erwe and Erwo (Fig. 2A), as indicated by significant studentized residuals

Start site (Erwo)	Gene symbol (Erwo)	Protein function	d_S±S.E. (Erwo–Erga)	d_S±S.E. (Erwe–Erga)	Studentized residual
234,633	Erum1330/*proP*	Proline/betaine transporter	0.0766±0.0210	0.0376±0.0144	−11.89***
410,422	Erum2390/*uvrD*	DNA helicase II	0.0330±0.0101	0.0179±0.0074	−4.32*
817,823	Erum4770/*nuoM*	NADH-quinone oxoreductase chain M	0.0044±0.0044	0.0578±0.0166	18.80***
834,524	Erum4890	Unknown	0.0176±0.0125	0.0478±0.0211	8.95***
873,909	Erum5170/*carA*	Carbamoyl-phosphate synthase small subunit	0.0569±0.0182	0.0764±0.0220	5.69***
1,013,762	Erum5890/*secY*	Preprotein translocase SecY	0.0189±0.0095	0.0047±0.0047	−4.09*
1,287,911	Erum7520/*pdhA*	Pyruvate dehydrogenase E1 component, alpha subunit	0.0198±0.0115	0.0593±0.0207	12.15***
1,298,724	Erum7590	Unknown	0.0541±0.0244	0.0212±0.0151	−9.81***

Significance levels of studentized residuals (Bonferroni-corrected): *P<0.05; ***P<0.001.

Table 3
Genes with unusually high synonymous divergence between Erga and Erwo/Erwe (Fig. 3), as indicated by significant studentized residuals

Start site (Erwo)	Gene symbol (Erwo)	Protein function	d_N±S.E. (Erwo–Erga)	d_S±S.E. (Erwo–Erga)	Studentized residual
280,109	Erum2000/*clpP*	ATP-dependent Clp protease, proteolytic subunit	0.0000±0.0000	0.1735±0.0635	6.14***
620,863	Erum3880	Putative integral membrane protein	0.0360±0.0088	0.2164±0.0678	5.89***
888,358	Erum5220	Putative type IV secretion system protein	0.0475±0.0035	0.2395±0.0266	7.05***
1,365,790	Erum7960	Unknown	0.0712±0.0048	0.1887±0.0223	4.60**
1,490,227	Erum8740/*map1*	Major antigenic protein MAP1	0.0634±0.0097	0.6000±0.1266	37.55***

Significance levels of studentized residuals (Bonferroni-corrected): $^{**}P<0.01$; $^{***}P<0.001$.

relationship between mean d_S and rank $R^2=61.1\%$; $P<0.001$). The greatest outliers corresponded to the highest-ranked values of mean d_S (Fig. 3). There were five outliers with significant (Bonferroni-corrected) studentized ratios (Table 3); and these five genes corresponded to the five highest values of mean d_S. These five genes showed d_S values in the comparison of the Welvegonden strain genotypes with Erga that were between 5.7 and 18 times the median value (Tables 1 and 3). Note, however, that for all of these loci, d_N values were much lower than d_S values (Table 3). Thus, it was clear that the high d_S values could not be attributed to alignment problems.

The most divergent case was the *map1* gene encoding the major antigenic protein-1 (Table 3). At this locus d_S between Erwo and Erga was 0.6000±0.1266, while that between Erwe and Erga was 0.5933±0.1239. The *map1* gene is a member of the MAP multi-gene family, which was represented by 14 other orthologous genes present in each of the three genomes analyzed. Mean d_S between Erwo and Erga at these 14 loci was 0.0374±0.0077. Similarly, the mean d_S between Erwe and Erga at these loci was 0.0354±0.0080. Thus, other members of the MAP family besides *map1* showed synonymous divergence between Erga and the Welgevonden genotypes that was close to the average value for all genes (Table 1), highlighting the strikingly unusual pattern at *map1*.

3.4. Positive selection

We searched for loci subject to positive selection by comparing d_S and d_N using the *Z*-test, with the expectation that d_N significantly greater than d_S is an evidence of positive selection favoring changes at the amino acid level (Hughes and Nei, 1988). There were no cases of d_N significantly greater than d_S in the comparison of Erwe and Erwo. However, there were four loci with d_N significantly greater than d_S at the 5% level in comparison both between Erwe and Erga and between Erwo and Erga (Table 4). In three of these cases, there were no synonymous substitutions between Erga and the Welgevonden genomes (Table 4). By contrast, at the locus identified in the Erwo genome as Erum4340, d_S between Erwo and Erga was close to the median value, but d_N was greatly elevated (Tables 1 and 4). In fact, d_N at this locus (0.0707) was the second-highest value between Erwo and Erga in the 818 loci examined, the highest being at the Erum7960 locus, a locus with an evidence of homologous recombination between Erga and a divergent genome (Table 3).

In order to search for factors that might be responsible for unusually low d_S, we examined the nucleotide content at four-fold degenerate sites. Excluding the 13 loci with strong evidence of homologous recombination (Tables 2 and 3), the Erwo genome showed a high mean AT% (87.3±0.2%) and a substantial TA skew (0.024±0.009). In comparison between Erwo and Erga, 78 of these loci showed no synonymous differences. The mean AT% in the loci without synonymous differences (86.8±0.6%) was not significantly different from that at the remaining loci (87.4±0.2%). However, the mean TA skew at the loci without synonymous substitutions (0.103±0.029) was significantly different from that at other loci (0.015±0.009; *t*-test; two-tailed $P=0.005$). Thus substantial TA skew seemed to be correlated with a reduction in the rate of synonymous substitution between Erwo and Erga. Similar results were found for comparison between Erwe and Erga (not shown).

4. Discussion

Examination of the patterns of nucleotide substitution between three genomes (Erwe, Erwo, and Erga) of *E. ruminantium* revealed substantial sequence polymorphism. Although the Erwe and Erwo genomes are both identified as belonging to the Welgevonden strain, the number of synonymous nucleotide substitutions per site between them suggested that their most recent common ancestor occurred over 25,000 years ago. By contrast, the Erga genotype from the Gardel strain was estimated to have diverged from the two Welgevonden genotypes 2–4 million years ago.

The uncertainty in estimating the age of the most common ancestors of these strains arose from the lack of a well-established molecular clock for *E. ruminantium*. The applicability of rate

Table 4
Genes with d_N significantly greater than d_S in comparison between Erwe/Erwo and Erga

Start site (Erwo)	Gene symbol (Erwo)	Protein function	No. of amino acids	d_N±S.E. (Erwo–Erga)	d_S±S.E. (Erwo–Erga)
665,023	Erum3790	Unknown	235	0.0066±0.0033*	0.0000±0.0000
746,784	Erum4340	Unknown	392	0.0707±0.0088**	0.0281±0.0132
1,041,693	Erum6240	Putative membrane protein	81	0.0289±0.0119*	0.0000±0.0000
1,198,475	Erum7060	Unknown	546	0.0029±0.0014*	0.0000±0.0000

Z-test of the hypothesis that d_N equals d_S: $^{*}P<0.05$; $^{**}P<0.01$.

estimates from *E. coli* and *Salmonella* was uncertain because of the possibility that there is an increased mutation rate in *E. ruminantium*, as is known for certain other obligately intracellular Bacteria (Ochman et al., 1999; Itoh et al., 2002). In spite of this uncertainty, the level of synonymous substitution between Erwe and Erwo (about 5 substitutions per 10,000 synonymous sites) is clearly far too high to have arisen in the two decades (Frutos et al., 2006) since these two isolates have been separately cultured. Thus, the Welgevonden strain must itself be a genetically non-homogeneous mixture of related genotypes.

Examination of synonymous substitution at individual loci revealed unusual patterns suggestive of past events of homologous recombination. Both Erwe and Erwo included genes that showed evidence of homologous recombination with more distantly related genomes, including genomes closely related to Erga. The loci with the strongest evidence of such recombination are known to play roles in important biological processes (Table 2). Since homologous recombination introduces nonsynonymous as well as synonymous differences, these recombinational events may have introduced changes having phenotypic effects on the recipient genotype.

Likewise, the Erga genome showed evidence of homologous recombination with distantly related genomes, and the genes with the strongest evidence of such recombination include several that may play roles in interactions with the host (Table 3). These included a protein putatively involved in the type IV secretion system, which is involved in the secretion of factors involved in pathogenicity of a number of bacterial species (Cascales and Christie, 2003). The most striking evidence for homologous recombination in the Erga genome involved the *map1* gene encoding the major antigenic protein-1 (Table 3). The existence of homologous recombination at this locus is consistent with previous phylogenetic analysis of *E. ruminantium map1* alleles, which revealed a lack of geographical clustering of alleles (Allsopp et al., 2001) and a topology strikingly different in several respects from that of a tree based on housekeeping genes (Allsopp et al., 2003). Together with our results, the phylogenetic results imply that a homologous recombination has been frequent at this locus, implying that relationships among *map1* sequences can be expected to provide no information regarding overall relatedness among genomes.

Evidence of homologous recombination in *E. ruminantium* is interesting because of this species' obligate intracellularity. The available data do not permit quantitative comparison of the extent of homologous recombination in *E. ruminantium* with that in free-living bacteria, but the evidence for recombination in this species supports the hypothesis (Itoh et al., 2002) that an intracellular lifestyle need not subject bacteria to “Muller's ratchet”; that is, to the accumulation of deleterious mutations that cannot be eliminated by recombination (Moran, 1996).

Whether or not the *map1* locus is subject to positive Darwinian selection has been controversial (Allsopp et al., 2001; Jiggins et al., 2002). Evidence for positive selection presented by Jiggins et al. (2002) was based on methods that are known to be non-conservative (Suzuki and Nei, 2004). In the present study, we used a *Z*-test for positive selection by comparing d_N and d_S across entire coding regions in comparison between Erga and the Welgevonden genomes. This approach is expected to be conservative because in many known cases of positive selection at the molecular level, such selection is focused only on certain protein domains (Hughes and Nei, 1988; Hughes, 1999). Given this conservative approach, our strong evidence of positive selection on a locus of unknown function, designated Erum4340 in the Erwo genome, was especially striking (Table 4). Examination of the sequences at this loci showed that nonsynonymous differences scattered across the alignment with no obvious hotspots (not shown). In order to clarify the basis of selection at this locus, it will be important to obtain experimental evidence regarding the function of the protein it encodes. Such knowledge may in turn provide important insights into the biology and pathogenesis of *E. ruminantium*.

Besides Erum4340, there were three other loci with d_N significantly greater than d_S by the *Z*-test, but in all three of these cases there were no synonymous differences between Erga and the Welgevonden genomes (Table 4). Thus, in these three cases, we could not rule out the hypothesis that the pattern of d_N exceeding d_S was produced by some factor acting to reduce synonymous substitution, rather than by positive selection favoring amino acid changes. We obtained evidence that high TA skew is associated with the reduction in synonymous substitution, but there are likely to be other factors at work, including possibly selective constraints on synonymous codon usage (Grantham et al., 1980). Comparison of *Staphylococcus aureus* genome sequences revealed a number of genes with no synonymous differences among a set of genomes far more divergent in sequence than the *E. ruminantium* genomes analyzed here (Hughes and Friedman, 2005). Thus, reduction of synonymous substitution at certain loci maybe a general phenomenon in bacterial genomes and one that remains poorly understood.

Acknowledgment

This research was supported by grant GM43940 from the National Institutes of Health to A.L.H.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2006.08.003.

References

Allsopp, M.T., et al., 2001. *Ehrlichia ruminantium* major antigenic protein gene (map1) variants are not geographically constrained and show no evidence of having evolved under positive selection pressure. J. Clin. Microbiol. 39, 4200–4203.

Allsopp, M.T., van Heerden, H., Steyn, H.C., Allsopp, B.A., 2003. Phylogenetic relationships among *Ehrlichia ruminantium* isolates. Ann. N.Y. Acad. Sci. 990, 685–691.

Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Andwalla, P., 2003. The evolutionary genomics of pathogen recombination. Nat. Rev., Genet. 4, 50–60.

Banerjee, T., Gupta, S.K., Ghosh, T.C., 2005. Role of mutational bias and natural selection on genome-wide nucleotide bias in prokaryotic organisms. BioSystems 81, 11–18.

Bekker, C.P., et al., 2005. Transcription analysis of the major antigenic protein 1 multigene family of three in vitro-cultured *Ehrlichia ruminantium* isolates. J. Bacteriol. 187, 4782–4791.

Belser, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics. John Wiley and Sons, New York.

Bulmer, M., 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897–907.

Cascales, E., Christie, P.J., 2003. The versatile bacterial type IV secretion systems. Nat. Rev. Microbiol. 1, 137–149.

Chen, S.W., Lee, W., Hottes, A.K., Shapiro, L., McAdams, H.H., 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. U. S. A. 101, 3480–3485.

Eyre-Walker, A., Bulmer, M., 1995. Synonymous substitution rates in enterobacteria. Genetics 140, 1407–1412.

Frutos, R., et al., 2006. Comparative genomic analysis of three strains of *Ehrlichia ruminantium* reveals an active process of genome size plasticity. J. Bacteriol. 188, 2533–2542.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pavé, A., 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, r49–r62.

Hanage, W.P., Fraser, C., Spratt, B.G., in press. The impact of homologous recombination on the generation of diversity in bacteria. J. Theor. Biol.

Hughes, A.L., 1999. Adaptive Evolution of Genes and Genomes. Oxford University Press, New York.

Hughes, A.L, Friedman, R., 2004. Patterns of sequence divergence in 5′ intergenic spacers and linked coding regions in 10 species of pathogenic Bacteria reveal distinct recombinational histories. Genetics 168, 1795–1803.

Hughes, A.L, Friedman, R., 2005. Nucleotide substitution and recombination at orthologous loci in *Staphylococcus aureus*. J. Bacteriol. 187, 2698–2704.

Hughes, A.L., Nei, M., 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335, 167–170.

Hughes, A.L., Friedman, R., Murray, M., 2002. Genomewide pattern of synonymous nucleotide substitution in two complete genomes of *Mycobacterium tuberculosis*. Emerg. Infect. Dis. 8, 1342–1346.

Itoh, T., Martin, W., Nei, M., 2002. Acceleration of genome evolution caused by enhanced mutation rate in endocellular symbionts. Proc. Natl. Acad. Sci. U. S. A. 99, 12944–12948.

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

Lecointre, G., Rachdi, L., Darlu, P., Denamur, E., 1998. *Escherichia coli* molecular phylogeny using the incongruence length difference test. Mol. Biol. Evol. 15, 1685–1695.

Lobry, J.R., 1996. Asymmetric substitution patterns in two DNA strands of bacteria. Mol. Biol. Evol. 13, 660–665.

Li, W.-H., 1993. Unbiased estimates of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36, 96–99.

Moran, N.A., 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. U. S. A. 93, 2873–2878.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Ochman, H., Lawrence, J.G., Groisman, E.A., 2000. Lateral transfer and the nature of bacterial innovation. Nature 405, 299–304.

Ochman, H., Elwyn, S., Moran, N.A., 1999. Calibrating bacterial evolution. Proc. Natl. Acad. Sci. U. S. A. 96, 12638–12643.

Sharp, P.M., 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J. Mol. Evol. 33, 23–33.

Sharp, P.M., Li, W.-H., 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4, 222–230.

Smith, N.G., Eyre-Walker, A., 2001. Nucleotide substitution rate estimation in enterobacteria: approximate and maximum-likelihood methods lead to similar conclusions. Mol. Biol. Evol. 18, 2124–2126.

Suzuki, Y., Nei, M., 2004. False-positive selection identified by ML-based methods: examples from the *Sig1* gene of the diatom *Thalassiosira weissflogii* and the *tax* gene of a human T-cell lymphotropic virus. Mol. Biol. Evol. 21, 914–921.

Thompson, J.D., Higgins, D.G., Gibson, T., 1994. CLUSTALW: 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.

Uilenberg, G., 1983. Heartwater (*Cowdria ruminantium* infection): current status. Adv. Vet. Sci. Comp. Med. 27, 428–455.

Vetsigian, K., Goldenfeld, N., 2005. Global divergence of microbial genome sequences mediated by propagating fronts. Proc. Natl. Acad. Sci. U. S. A. 102, 7332–7337.

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

Yang, Z., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17, 32–43.

Zhang, L., Vision, T.J., Gaut, B.S., 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. Mol. Biol. Evol. 19, 1464–1473.