

## Adaptive Evolution and Recombination of *Rickettsia* Antigens

Francis M. Jiggins

Institute of Cell Animal and Population Biology, Ashworth Laboratories, School of Biology, University of Edinburgh, The King's Buildings, West Mains Road, Edinburgh EH9 3JT, UK

Received: 6 April 2005 / Accepted: 8 September 2005 [Reviewing Editor: Dr. Willie J. Swanson]

**Abstract.** The genus *Rickettsia* consists of intracellular bacteria that cause a variety of arthropod vectored human diseases. I have examined the evolutionary processes that are generating variation in antigens that are potential vaccine candidates. The surface proteins rOmpA and rOmpB are subject to intense positive natural selection, causing rapid diversification of their amino acid sequences between species. The positively selected amino acids were mapped and cluster together in regions that may indicate the location of functionally important regions such as epitopes. In contrast to the rOmp antigens, there is no evidence of positive selection on the intracytoplasmic antigen PS120 despite low selective constraints on this gene. All three genes showed evidence of recombination between species, and certain sequences are clear chimeras of two parental sequences. However, recombination has been sufficiently infrequent that the phylogenies of the three genes are similar, although not identical.

**Key words:** *Rickettsia* — rOmpA — rOmpB — PS120 — Positive selection — Recombination

### Introduction

*Rickettsia* is a genus of intracellular parasites of animals and plants that are best known as human

pathogens. The louse-vectored bacterium *Rickettsia prowazekii* causes one of the most severe diseases, epidemic typhus, which kills 10–30% of infected patients (Raoult and Roux 1997). Other human rickettsial diseases include Rocky Mountain spotted fever (*Rickettsia rickettsii*) and Mediterranean spotted fever (*Rickettsia conorii*), which are both vectored by ticks, and murine typhus (*Rickettsia typhi*), which is flea vectored (Raoult and Roux 1997). There are at least 10 other rickettsial diseases of humans, many of which were only recognized in the last 20 years. The genus also includes insect-vectored plant pathogens (Davis et al. 1998) and vertically transmitted insect pathogens (Werren et al. 1994).

Unusually for intracellular bacteria, *Rickettsia* are not enclosed in a host-derived cell membrane vacuole, but are in direct contact with the host's cytoplasm. The outer bacterial membrane is covered by a protein layer arranged in a regular crystalline array (Carl et al. 1990; Ching et al. 1990). One of the components of this protein layer is the rickettsial outer membrane protein B (rOmpB; also called the 120-, 133-, or 135-kDa antigen or SPA), which is attached to the membrane at the carboxyl end of the protein by a hydrophobic membrane spanning region followed by a short hydrophilic anchor (Carl et al. 1990). A second protein, rOmpA, is also an outer membrane protein, and is probably another component of the crystalline layer (Anderson et al. 1990).

Crystalline protein layers in bacteria are known to have a diverse range of roles including cell adhesion, surface recognition, and as a protective coat against a broad range of antagonists (e.g., bacteriophage, lytic

Correspondence to: Francis M. Jiggins; email: francis.jiggins@ed.ac.uk

enzymes, and antibodies (Sleytr and Messner 1983, 1988)). The rOmp proteins have both been implicated in the invasion of host cells (Li and Walker 1998; Uchiyama 2003). They are also both major antigens and are the principal candidates for vaccines against rickettsial infections (Croquet-Valdes et al. 2001). Another rickettsial antigen is the PS120 protein (the gene is referred to as *sca4* or Gene D), which is found within the cytoplasm of the bacteria and is not surface exposed (Schuenke and Walker 1994; Uchiyama 1997).

In this study I have assessed the effects of recombination and natural selection in generating diversity in these antigens. Although many antigens evolve rapidly due to selection to escape the acquired immune response of the host, previous studies have concluded that this is not the case for rOmpA (Fournier et al. 1998). I have revisited this question and detected positive selection acting on both *ompA* and *ompB* but not PS120. I also found that the rate of recombination between different species of *Rickettsia* is low.

## Methods

**Sequence Alignment.** All analyses were performed on previously published sequences of *ompA*, *ompB*, and PS120 (Table 1). The nucleotide sequences were initially aligned using ClustalW and then adjusted by eye to account for the predicted protein sequence and to avoid interrupting the reading frame. I have sequentially numbered the codons and amino acids in each gene according to the predicted protein sequences from the *R. conorii* genome (Ogata et al. 2001). The PS120 alignment includes amino acids 12–1020 of a total of 1026. The *ompA* alignment includes amino acids 953–2013 of a total of 2021. The region of tandem repeats in *ompA* is not included in the alignment. The *ompB* alignment includes amino acids 20–1649 of a total of 1655.

**Tree Reconstruction.** The phylogenetic trees of these genes were reconstructed by maximum likelihood with the program PAUP\* v.4.0b8 (Swofford 1998). The model of sequence evolution for each gene was selected by comparing models by likelihood ratio tests using the program Modeltest v.3.6 (Posada and Crandall 1998). The model selected for both *ompA* and *ompB* was GTR+G, which accounts for the base frequency, gamma distributed rate heterogeneity across sites, and six different rates for the transitions between the different nucleotides. The phylogeny of PS120 was reconstructed using the TVM+G model, which is identical to GTR+G except that the two types of transitions (A-G and C-T) occur at the same rate. The maximum likelihood topology was reconstructed by a heuristic search using nearest-neighbor interchanges.

**Testing for Positively Selected Codons.** The ratio of nonsynonymous-to-synonymous nucleotide substitutions ( $d_N/d_S$  ratio) was estimated to infer the selection pressures acting on the three genes. Neutral evolution will produce a  $d_N/d_S$  value of 1, while  $d_N/d_S < 1$  indicates purifying selection and  $d_N/d_S > 1$  directional selection. Because the selection pressures vary between different regions of the protein, the value of  $d_N/d_S$  averaged across all sites can be uninformative. Therefore, an approach was taken that allows  $d_N/d_S$  to differ across different codons.

The analysis was performed using a maximum likelihood model of codon substitution along the phylogenies of the genes (Nielsen and Yang 1998; Yang et al. 2000) implemented by the codeml program in the PAML v3.13d package (Yang 1997). The phylogenies used were reconstructed as described above and include all the taxa in Table 1. These phylogenies are available from the author on request. Codons with alignment gaps were included during the analysis of *ompA* but not *ompB* or PS120. This was because in the latter two genes including gaps produced problems with convergence during the iteration process (excluding gaps in *ompA* produced qualitatively very similar results to those presented here). The distribution of the  $d_N/d_S$  ratio ( $\omega$ ) across sites was estimated using different models of codon substitution (Yang et al. 2000). Model M0 (one ratio) assumes that all sites have the same value of  $\omega$ . M3 (discrete) assumes three different classes of sites with different  $\omega$  ratios. M7 (beta) allows sites to have 10 different values of  $\omega$ , calculated from the beta distribution with parameters  $p$  and  $q$ . The beta distribution is bounded between 0 and 1 and thus constitutes a null model for testing positive selection. M8 (beta +  $\omega$ ) is similar to M7, but with an additional  $\omega$  category that can exceed 1.

The hypothesis that some amino acid sites are under positive selection can be tested by comparing two nested models with a likelihood ratio test, where these models differ in whether or not they allow some sites to have values of  $\omega$  greater than one. In this case M7 is compared to M8, and M3 to M0. The former comparison (M7 and M8) is the most stringent test of positive selection (Anisimova et al. 2001), while the latter (M0 vs. M3) is primarily a test of variation of  $\omega$  between codons. The null distribution of the likelihood ratio test statistic ( $2\Delta l$ , where  $\Delta l$  is the difference between the log-likelihood scores of the two models) can be approximated using the  $\chi^2$  distribution with the degrees of freedom being the difference in the number of free parameters between the two models (two when comparing M7 and M8 and four when comparing M0 and M3). Analyses of simulated data sets indicate that this test is conservative (Anisimova et al. 2001).

It has recently been noted that the comparison of M7 and M8 may incorrectly detect positive selection if the distribution of  $\omega$  in M7 does not follow a beta distribution and many sites have  $\omega$  near to one (Swanson et al. 2003). Therefore, I also conducted a likelihood ratio test comparing model M8 with M8A (Swanson et al. 2003), which is not sensitive to the distribution of  $\omega$ . The M8A model is identical to M8 except that  $\omega$  is fixed as  $\omega = 1$ . The null distribution of this likelihood ratio test statistic can be approximated using the  $\chi^2$  distribution with 1 degree of freedom (df) (Swanson et al. 2003).

When the likelihood ratio test suggests the presence of sites under positive selection, an empirical Bayes method was used to calculate the posterior probabilities that each site fell into the  $\omega$  classes (Yang et al. 2000). Sites with high probabilities of belonging to the class with  $\omega > 1$  are likely to be under positive selection. The mean  $\omega$  ratio for each site (Figs. 1 and 2) was calculated as the average of the  $\omega$  ratios across all the  $\omega$  classes from model 8, with the posterior probabilities used as weights.

**Recombination: Incongruence of Gene Trees.** Recombination between genes can be detected by comparing the phylogenetic trees of multiple genes; if different genes have had different evolutionary histories, then recombination has occurred. Three pairs of alignments were produced (*ompB*+PS120, *ompA*+PS120, and *ompA*+*ompB*), which only included taxa for which sequences were available for both the genes in question. These alignments are therefore subsets of the larger alignments described above. The phylogeny of each alignment was then reconstructed as described above, and its robustness assessed by 1000 nonparametric bootstrap replicates.

I tested whether the phylogenies reconstructed from the two genes in each pair were significantly different using reciprocal SH tests. In all three comparisons, each gene was forced to take the tree

Table 1. *Rickettsia* sequences

Species	OmpA	OmpB	PS120 ENCODING GENE	Description	Human disease
<i>R. africanae</i>	RAU83436	AF123706	AF151724	Tick vectored vertebrate pathogen	African tick bite fever
<i>R. aeschlimannii</i>	RAU83446	AF123705	AF163006	Tick vectored vertebrate pathogen	
<i>R. akari</i>		AF123707	AF213016	Mite vectored vertebrate pathogen	Rickettsial pox
<i>R. australis</i>		AF123709	AF187982	Tick vectored vertebrate pathogen	Queensland tick typhus
<i>R. conorii</i>	RCU83440, RCU83443, RCU83448, RCU83453	AF123726, AF123721	AF163005, AF163008, U01133	Tick vectored vertebrate pathogen	Mediterranean spotted fever
<i>R. felis</i>		AF182279	AF196973	Flea vectored vertebrate pathogen	Californian flea rickettsiosis
<i>R. heilongjiangii</i>	AF179363	AY260451		Tick vectored vertebrate pathogen	
<i>R. heilongjiangensis</i> <i>v. extremiorientalis</i>	AY280710	AY280712		Tick vectored vertebrate pathogen	
<i>R. helvetica</i>			AF163009	Tick vectored vertebrate pathogen	
<i>R. honei</i>	AF018076	AF123711		Tick vectored vertebrate pathogen	Flinders Island spotted fever
<i>R. honei</i>		AF123724	AF163004	Tick vectored vertebrate pathogen	Thai tick typhus
<i>R. hulimii</i>	AF179366			Tick vectored vertebrate pathogen	
<i>R. hulimensis</i>		AY260452		vertebrate pathogen	
<i>R. japonica</i>	RJU83442	AF123713, RJAB3681	AF155055, AB003696	Isolated from ticks	Japanese fever
<i>R. massiliae</i>	RMU83444, RMU83445	AF123714	AF163003	vertebrate pathogen	
<i>R. mongolotimonae</i>	RMU83439	AF123715	AF151725	Isolated from ticks	
<i>R. montana</i>	RMU83447			Tick vectored vertebrate pathogen	
<i>R. montanensis</i>		AF123716	AF163002	vertebrate pathogen	
<i>R. parkeri</i>	RPU83449	AF123717	AF155059	Isolated from ticks	
<i>R. prowazekii</i>		AF211820, AF211821, AF161079, AF123718	AF200340	Isolated from ticks	Epidemic typhus
<i>R. rhipicephali</i>	RRU83450	AF123719	AF155053	Louse vectored vertebrate pathogen	
<i>R. rickettsii</i>	RRU83451	RRP120	AF163000	Isolated from ticks	
<i>R. sibirica</i>	RSU83455	AF123722	AF155057	Tick vectored vertebrate pathogen	Rocky Mountain spotted fever Siberian tick typhus

(continued)

Table 1. Continued

Species	OmpA	ompB	PSI20 ENCODING GENE	Description	Human disease
<i>R. slovaca</i>	RSU83454	AF123723	AF155054	Tick vectored vertebrate pathogen	
<i>R. typhi</i>			AF188482	Flea vectored vertebrate pathogen	Murine typhus
<i>Rickettsia</i> sp. California 2		AF210695		Isolated from fleas	
<i>Rickettsia</i> sp. DnS28	AF120019			Isolated from ticks	
<i>Rickettsia</i> sp. DnSI4	AF120020			Isolated from ticks	
<i>Rickettsia</i> sp. RpA4	AF120023			Isolated from ticks	
<i>Rickettsia</i> sp. IRS3	AF141910			Isolated from ticks	
<i>Rickettsia</i> sp. IRS4	AF141912		AF163010	Isolated from ticks	
<i>Rickettsia</i> sp. BJ-90	AF179367			Isolated from ticks	
Astrakhan fever <i>Rickettsia</i>	ARU83437	AF123708	AF163007	Tick vectored vertebrate pathogen	
Israeli tick typhus <i>Rickettsia</i>	ITU83441	AF123712	AF155058	Tick vectored vertebrate pathogen	
<i>Rickettsia</i> BAR-29	RSU83438	AF123710	AF155056	vertebrate pathogen	
<i>Rickettsia</i> sp. S	RSU83452	AF123720	AF163001	Isolated from ticks	

topology of the other gene, and the log likelihood calculated. We then tested whether this significantly reduced the likelihood of the topology relative to the maximum likelihood topology using the SH test (Goldman et al. 2000; Shimodaira and Hasegawa 1999), implemented in the program PAUP\* v.4.0b8 (Swofford 1998) using the RELL approximation.

**Recombination Within Genes.** We used three different approaches to detect recombination. The use of multiple methods has been recommended, as analyses of real and simulated data suggest that individual methods sometimes either fail to detect recombination or give false positives (Posada 2002; Posada and Crandall 2001). The methods were selected to reflect a range of different approaches, and because two have been recommended following power analyses (Posada 2002; Posada and Crandall 2001). All these analyses used the same sequence alignments as in the PAML analysis.

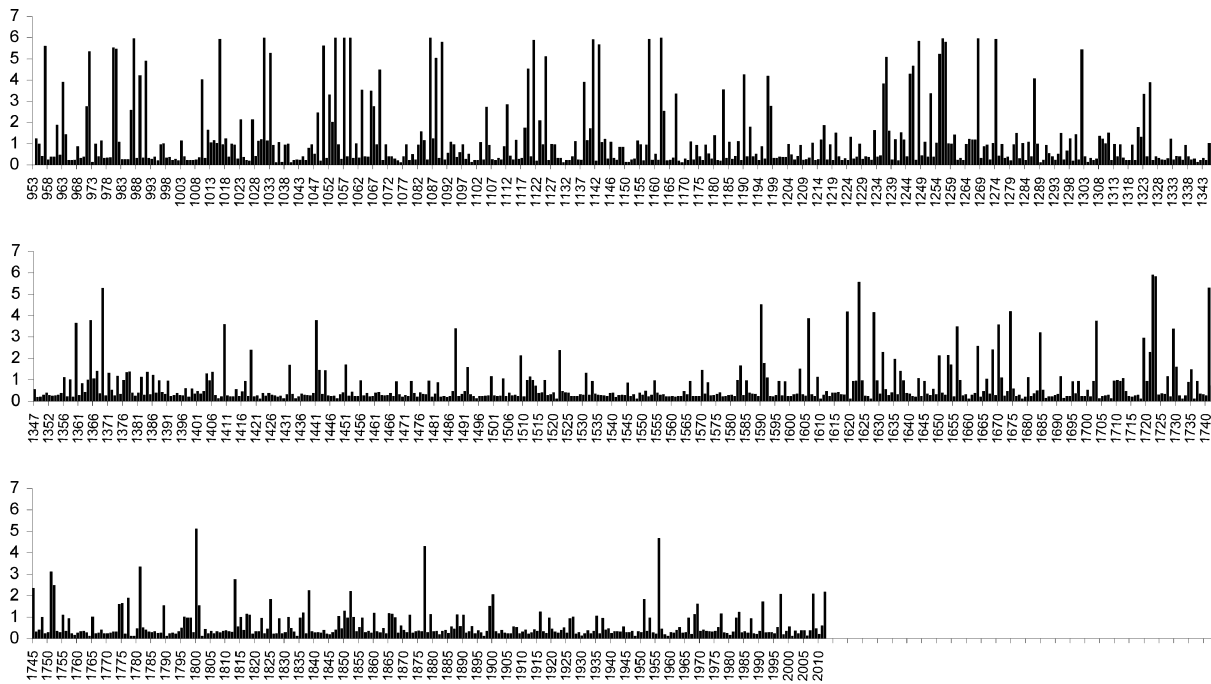
The first method was maximum  $\chi^2$  (Smith 1992). First, alignments were made from all possible triplets of sequences. For each triplet, sites that were monomorphic or contained gaps were discarded. Then, for every pair of sequences in the triplet, a 100-bp window was slid along the alignment in one-nucleotide steps. At each step, the number of variable sites was compared in the left and right halves of the window using a  $\chi^2$  test. Potential breakpoints correspond to peaks in the values of  $\chi^2$ . All *P* values were Bonferroni corrected for multiple tests performed on each alignment. This analysis was implemented using the program RDP (Martin and Rybicki 2000).

The second method used was the Reticulate method (Jakobsen and Easteal 1996). First, a matrix is constructed, each cell of which is a pairwise comparison of two phylogenetically informative sites (excluding sites with more than two states). The cells are classed as compatible if a single tree could be constructed from both sites assuming the minimum number of mutations. The test statistic is the Neighbor Similarity Score (NSS), which is the average proportion of times a cell neighbors a cell of the same type (compatible next to compatible or incompatible next to incompatible). The null distribution of this statistic was generated by recalculating the NSS  $10^4$  times from datasets where the order of sites had been permuted. This analysis was implemented using the program RDP (Martin and Rybicki 2000).

Finally, recombination can be detected by a decline in linkage disequilibrium between pairs of sites with increasing distance. This is because as the distance between two sites decreases, there will have been fewer recombination events to break down linkage disequilibrium. The linkage disequilibrium between pairs of sites was estimated using two different measures,  $r^2$  and  $|D'|$ , and the correlation with distance was calculated using Pearson's coefficient (Awadalla et al. 1999; Jorde and Bamshad 2000). The significance of the negative correlation was calculated by a Mantel test. The position of the sites was randomized, the statistic recalculated a minimum of 1000 times, and the significance taken as the proportion of times the correlation coefficient was the same as or more negative than the observed value. The permutation test included sites with two segregating alleles and was performed using the program LDhat (McVean et al. 2002). The analysis was then repeated including only those sites at which both alleles occurred at frequencies of > 10%. The exclusion of rare alleles is expected to make the test more powerful because common alleles tend to be older and therefore are more likely to show evidence of recombination (Awadalla et al. 1999).

## Results

**Positive Selection on Antigen Genes.** The two outer membrane proteins were both found to be subject to positive selection, causing diversification of their



**Fig. 1.** The  $d_N/d_S$  ( $\omega$ ) ratio along the *ompA* gene estimated as the average of the  $\omega$  ratios across all the  $\omega$  classes weighted by their posterior probabilities. Amino acid sites are numbered according to the *R. conorii* genome sequence. Note that alignment gaps mean that these numbers are not always sequential.

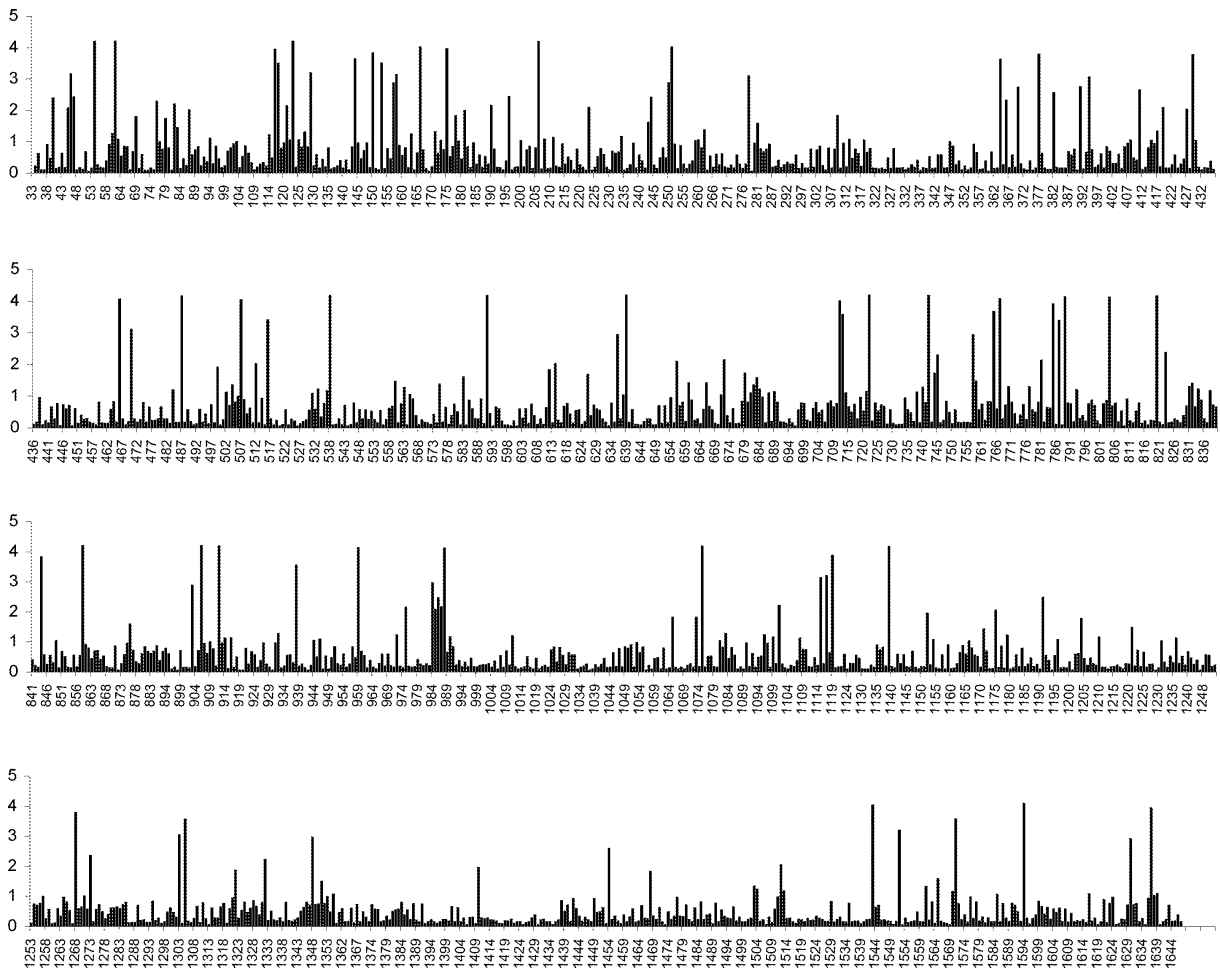
protein sequence. The likelihood scores of the different codon substitution models, the model parameters and sites under selection are shown in Tables 2 and 3, and the results of the likelihood ratio tests in Table 5. In the *ompA* gene, likelihood ratio tests indicated that models that included positive selection significantly increased the likelihood scores compared to models without positive selection (Table 5). In total it was estimated that 7% of the codons in this gene are positively selected, and this number is likely to be conservative because the  $\omega$  ratio of the positively selected class was very high ( $\omega = 8$ ), meaning that codons with  $\omega$  ratios just over 1 will not be included in this category (Table 2). The results from *ompB* are broadly similar to those for *ompA*. Again, models that allow positively selected sites have significantly higher likelihoods than those that do not (Table 5). The proportion of positively selected sites (6%) and estimated  $\omega$  ratio of those sites ( $\omega = 4$ ) are both lower than was the case for *ompA* (Table 3).

In contrast to the surface antigens, there is no evidence of positive selection acting on the intracytoplasmic antigen PS120. The  $\omega$  ratio of the PS120 encoding gene is highly heterogeneous across codons (M1 vs M3; Tables 4 and 5). Furthermore, the comparison of models M7 and M8 suggests that some of the codons are positively selected (Tables 4 and 5). However, this is not a robust result because the  $\omega$  ratio is only slightly greater than 1 (Table 4). This is confirmed as model M8 is not significantly better than M8A, which does not include positively selected

sites (Table 5). Therefore, there is no evidence of positive selection acting on PS120. The significant comparison of M7 and M8 probably results from the distribution of  $\omega$  ratios across codons being a bad fit to the beta distribution combined with a large proportion of codons evolving nearly neutrally ( $\omega = 1$  for 35% of sites in model M8A; Table 4). This suggests that there are many regions of this protein subject to few functional constraints.

*Mapping Sites Under Selection.* The distribution of positively selected sites and conserved sites along the *ompA* and *ompB* genes is shown in Figs. 1 and 2. The location of codons with the strongest statistical support for being positively selected is also shown in Tables 2 and 3. In the *ompA* gene it is notable that most of the positively selected codons are found at the 5' end of the region analyzed. For example, 68% of the codons that had probabilities over 0.5 of being positively selected are found in the first third of this region (data not shown). The positively selected codons were more evenly distributed along *ompB* than was the case for *ompA*, although there was a slight tendency for them to be located in the 5' end of the gene (65% of the codons that had probabilities over 0.5 of being positively selected are found in the first half of the sequence).

*Recombination: Incongruence of Gene Trees.* The phylogenies of the three genes are broadly congruent, suggesting that recombination between these species



**Fig. 2.** The  $d_N/d_S$  ( $\omega$ ) ratio along the *ompB* gene. For details see the legend to Fig. 1.

**Table 2.** Positively selected sites, log-likelihood scores and parameter estimates for the *Rickettsia ompA* gene

Model	Parameters in the $\omega$ distribution	$l$	Positively selected codons
M0, one ratio	$\omega = 0.65$	-9618.40	None
M3, discrete	$\omega_1 = 0.11, p_1 = 0.64$ $\omega_2 = 1.39, p_2 = 0.30$ $\omega_3 = 7.96, (p_3 = 0.06)$	-9381.93	987A, 1015L, <b>1030L</b> , <b>1054D</b> , <b>1056A</b> , <b>1058V</b> , <b>1085T</b> , 1140L, 1157T, <b>1161N</b> , 1267V, 1273Q, 1722A, 1723G
M7, beta	$p = 0.015, q = 0.014$	-9472.53	Not allowed
M8, beta + $\omega$	$p = 0.002, q = 0.003,$ $p_1 = 0.93$ $\omega = 5.96, (p_2 = 0.07)$	-9384.49	<b>987A</b> , 1015L, <b>1030L</b> , <b>1054D</b> , <b>1056A</b> , <b>1058V</b> , <b>1085T</b> , 1089N, 1120L, 1140L, 1157T, <b>1161N</b> , 1247V, <b>1255G</b> , 1256T, <b>1267V</b> , 1273Q, 1722A, 1723G
M8A, beta + $\omega = 1$	$p = 3.32, q = 99.00, p_1 = 0.56$ $(\omega = 1, p_2 = 0.44)$	-9471.31	Not allowed

*Note.* The parameters  $p$  and  $q$  describe the shape of the beta distribution of  $\omega$ , and  $p_1, p_2$ , and  $p_3$  are the proportions of codons belonging to each category. Proportions that are not free parameters are in parentheses. Positively selected codons are those at which the posterior probability that the codon belongs to the positively selected class is  $P > 0.95$ , with those at which  $P > 0.99$  shown in boldface. Only sites in  $\omega_3$  of model 3 are listed as being positively selected.

of *Rickettsia* types is rare. This can be clearly seen from inspection of Figs. 3–5, which compare the phylogenies of the three genes. There are some differences in the tree topology, but these are not sta-

tistically significant in the comparisons of PS120 vs *ompA* or *ompA* vs *ompB* (Table 6). There is, however, evidence of recombination in the comparison of the *ompB* and PS120 tree topologies (Table 6).

**Table 3.** Positively selected sites, log-likelihood scores and parameter estimates for the *Rickettsia ompB* gene

Model	Parameters in the $\omega$ distribution	$l$	Positively selected codons
M0, one ratio	$\omega = 0.48$	-19,832.06	None
M3, discrete	$\omega_1 = 0.09, p_1 = 0.60 \omega_2 = 1.01,$ $p_2 = 0.36 \omega_3 = 4.89, (p_3 = 0.04)$	-19,315.01	<b>53V, 60P, 122H, 205V, 487V,</b> 538G, 591A, <b>639D, 720V, 742I, 789A,</b> 804F, 820G, <b>856G, 906T, 912S,</b> 959I, 988A, 1075F, 1139R
M7, beta	$p = 0.19, q = 0.26$	-19,418.60	Not allowed
M8, beta + $\omega$	$p = 0.26, q = 0.39,$ $p_1 = 0.94 \omega = 4.20, (p_2 = 0.06)$	-19,316.34	<b>53V, 60P, 122H, 205V, 466A, 487V,</b> 507G, <b>538G, 591A, 639D, 720V, 742I,</b> 762A, 789A, 804F, 820G, <b>856G, 906T,</b> <b>912S, 959I, 988A, 1075F, 1139R, 1594V</b>
M8A, beta + $\omega = 1$	$p = 1.31 q = 11.70,$ $p_1 = 0.64 (\omega = 1, p_2 = 0.36)$	-19,399.97	Not allowed

Details in legend of Table 2

**Table 4.** Positively selected sites, likelihood scores and parameter estimates for the PS120 encoding gene

Model	Parameters in the $\omega$ distribution	$l$
M0, one ratio	$\omega = 0.4135$	-11,866.24
M3, discrete	$\omega_1 = 0.00, p_1 = 0.11 \omega_2 = 0.25, p_2 = 0.60 \omega_3 = 1.17, (p_3 = 0.29)$	-11,741.04
M7, beta	$p = 0.51, q = 0.67$	-11,750.90
M8, beta + $\omega$	$p = 1.53, q = 4.75, p_1 = 0.76 \omega = 1.23 (p_2 = 0.24)$	-11,741.11
M8A, beta + $\omega = 1$	$p = 4.55, q = 21.4, p_1 = 0.65 (\omega = 1, p_2 = 0.35)$	-11,742.19

Details in legend of Table 2

**Table 5.** Likelihood ratio tests for positive selection

Gene/protein	Models	$2\Delta l$	d.f.	$P$
<i>OmpA</i>	M0 v M3	472.9	4	<0.001
	M7 v M8	176.1	2	<0.001
	M8A v M8	173.6	1	<0.001
<i>ompB</i>	M0 v M3	1034.1	4	<0.001
	M7 v M8	204.5	2	<0.001
	M8A v M8	167.3	1	<0.001
PS120	M0 v M3	250.4	4	<0.001
	M7 v M8	19.6	2	<0.001
	M8A v M8	2.2	1	Not Sig.

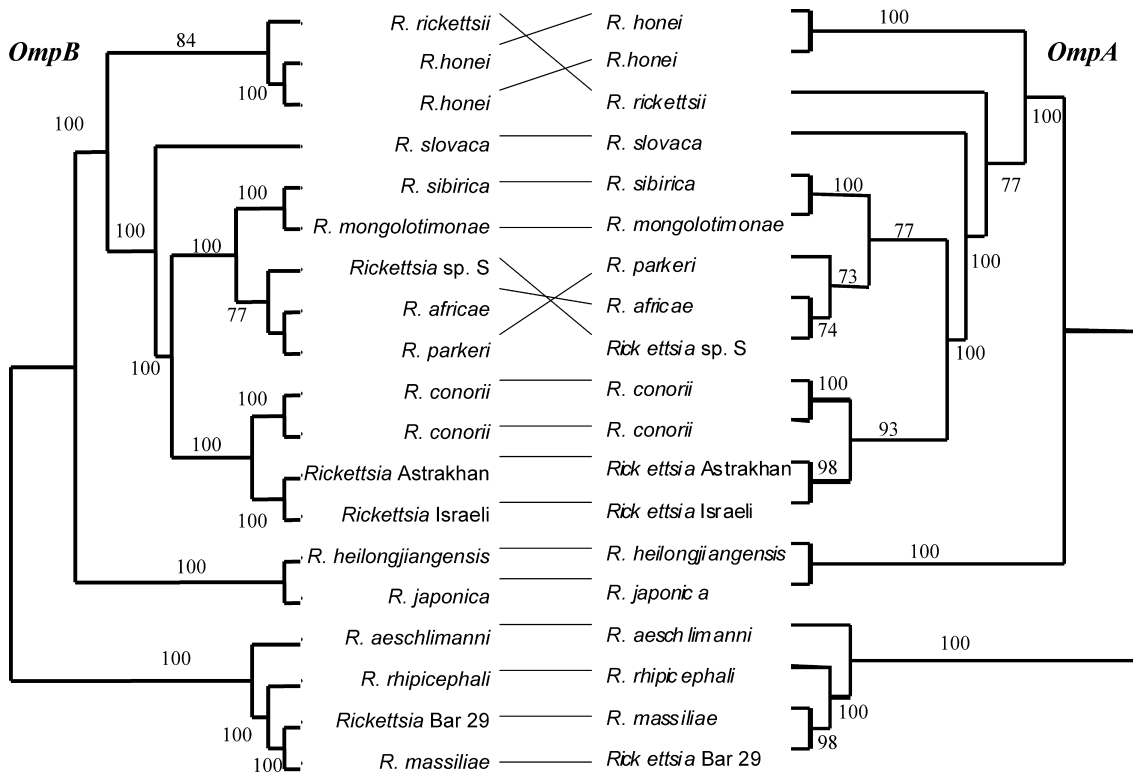
**Recombination Within Genes.** The maximum  $\chi^2$  test detected breakpoints in all three genes. In total there were 2199 significant  $\chi^2$  peaks ( $P < 0.01$ , Bonferroni corrected) in PS120, 1030 in *ompB*, and 187 in *ompA*. These totals are summed across all pairwise comparisons, so they do not reflect the number of crossing over events because each actual crossover event will have been detected in multiple comparisons. All three genes contained breakpoints supported by  $P < 10^{-10}$ . A particularly clear breakpoint that was detected by the maximum  $\chi^2$  test ( $P < 10^{-13}$ ) is shown in Fig. 6. In this case, the *ompB* gene from *R. felis* is a clear chimera between two other sequences. It is noteworthy that this *R. felis* sequence is also partly responsible for the conflict between gene trees described above.

Further support for recombination within the alignments comes from the Reticulate test. In all three cases, the NSS was higher than expected (*ompB*, NSS = 0.83,  $P = 0.001$ ; PS120, NSS = 0.78,  $P < 0.0001$ ; *ompA*, NSS = 0.92,  $P = 0.01$ ). This indicates that neighboring sites tend to have more similar phylogenetic histories than more distant sites, as expected when there is recombination.

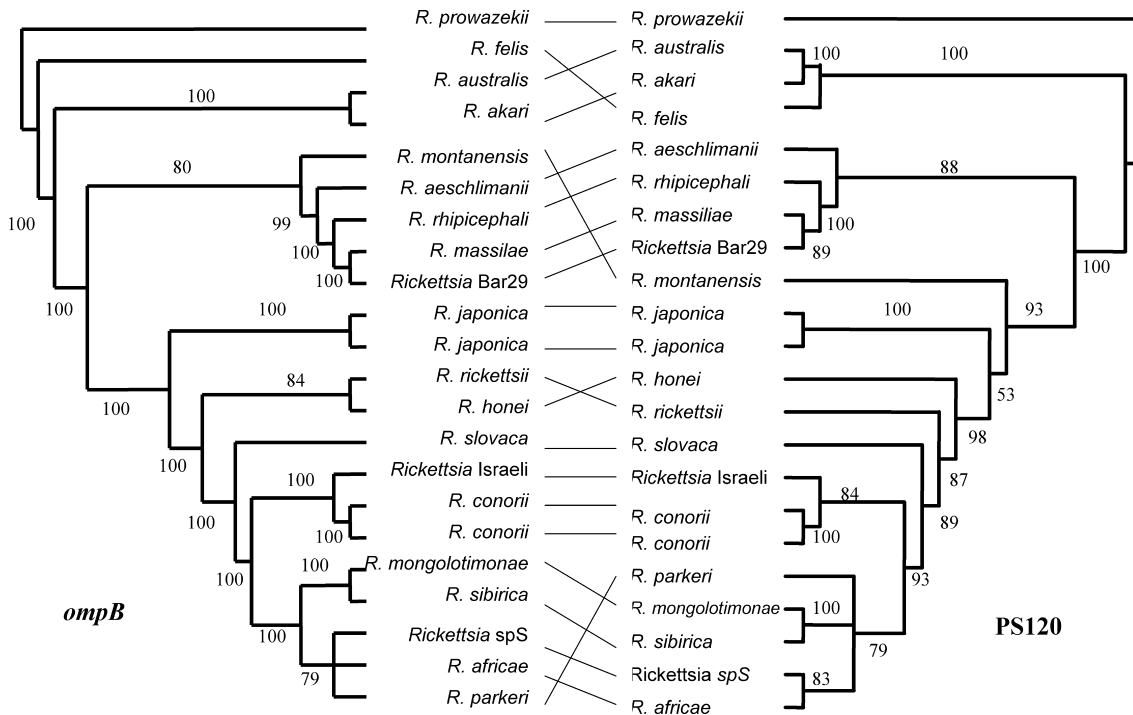
Recombination events within a gene will result in a decline in linkage disequilibrium with increasing genetic distance. This was assessed for all three genes, both based on the entire dataset and using only segregating sites found in more than 10% of the sequences. The exclusion of rare alleles is expected to increase the power of these tests. Although linkage disequilibrium declines with distance in all three genes, there is only consistent statistical support for this relationship in the case of PS120 (Table 7). Furthermore, following the exclusion of rare alleles, the PS120 encoding gene has the most negative correlation coefficients of the three genes (Table 7).

## Discussion

**Positive Selection.** Genes that encode antigens commonly have high rates of nonsynonymous substitutions due to selection favoring antigenic escape



**Fig. 3.** Phylogenies of *ompA* and *ompB*. Percentage bootstrap support is shown by the nodes. The trees only include taxa where sequences of both genes are available from the same strain.



**Fig. 4.** Phylogenies of PS120 and *ompB*. For details see the legend to Fig. 3.

mutants. Consistent with this pattern, a substantial proportion of the amino acids in rOmpA and rOmpB is positively selected. This selection pressure drives

rapid change in the sequence of these proteins and, therefore, will probably generate antigenic diversity within and between these species of *Rickettsia*. The

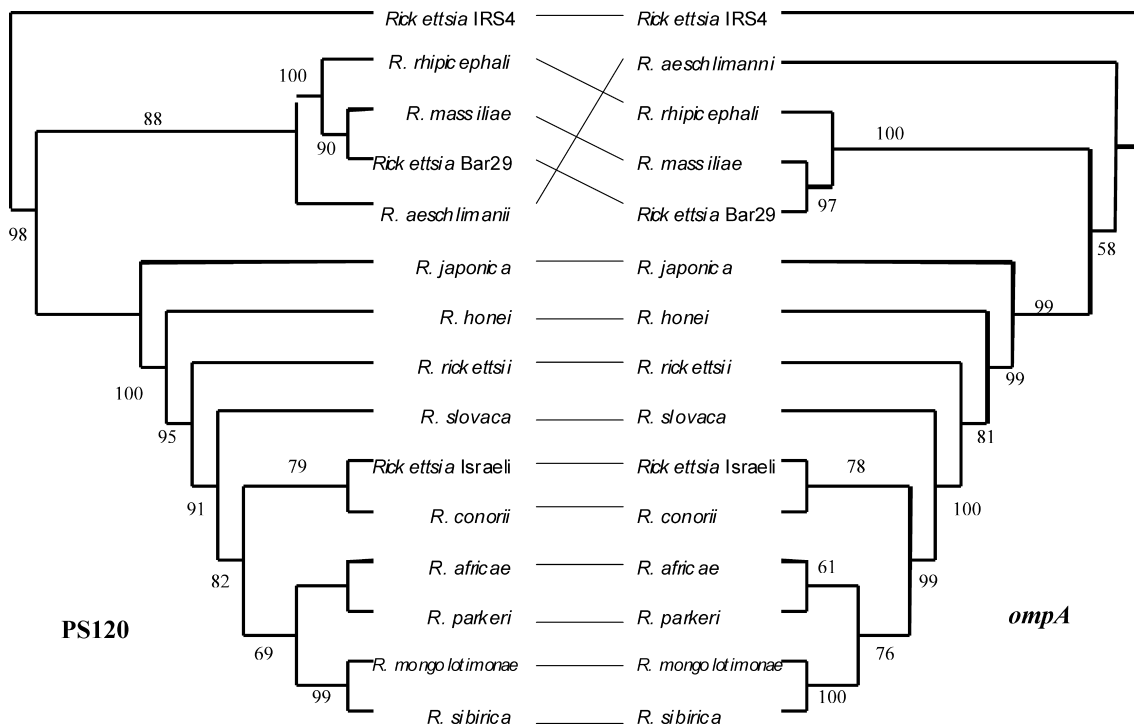


Fig. 5. Phylogenies of *ompA* and PS120. For details see the legend to Fig. 3.

Table 6. SH tests for conflicting phylogenetic signals between different genes

Data	Test tree	$\Delta \ln L$	$P$
PS120	<i>OmpA</i>	8.44	0.065
<i>ompA</i>	PS120	0.421	0.437
<i>ompB</i>	PS120	63.13	0.000
PS120	<i>OmpB</i>	146.92	0.000
<i>ompA</i>	<i>OmpB</i>	11.60	0.046
<i>ompB</i>	<i>OmpA</i>	5.30	0.162

results of this study contrast with a previous analysis of *ompA* that estimated the average  $d_N/d_S$  ratio across the whole gene and failed to detect positive selection (Fournier et al. 1998). The reason for this difference is that while some codons in *ompA* and *ompB* are positively selected ( $d_N/d_S > 1$ ), others are subject to selective constraints ( $d_N/d_S < 1$ ), so the average  $d_N/d_S$  ratio is  $< 1$ .

It is premature to conclude that selection favoring antigenic escape mutants is the cause of positive selection on *ompA* and *ompB*. It is possible that long-term infections of vertebrates may not be an important component in the life cycle of many *Rickettsia* species. Instead, as in most human infections, rickettsemia in other mammals may be short lived. This led Raoult and Roux (1997) to suggest that transmission may often occur between ticks feeding at the same time on a single host. This is particularly likely, as ticks have a tendency to aggregate. If this is the

case, transmission may normally occur before an immune response is mounted against the bacteria, lessening the selection for antigenic change. Furthermore, these genes are also expressed in the arthropod vector, and it may be that the arthropod immune system is the primary selection pressure. They also play a role in the bacteria entering vertebrate cells, and this could also potentially expose them to selection.

I also found that positive selection is not a universal feature of rickettsial antigens, as there is no evidence for positive selection on the PS120 antigen. It is unlikely that functional constraints on PS120 are preventing evolutionary change, because many amino acids evolve nearly neutrally in this protein. It is possible that PS120 is a much less important elicitor of the immune response and is, therefore, under less selection for antigenic change. Alternatively, positive selection on the outer membrane proteins may not be a consequence of their antigenicity, but due to some other function not possessed by PS120 (e.g., adhering to host cells).

The positively selected sites in *ompA* and *ompB* tend to cluster together, which is similar to the pattern observed in similar analyses of other antigens (e.g., Crewther et al. 1996). It is commonly thought that these clusters reflect regions of the protein detected by the host immune system. For example, in the gp120 gene of HIV, the positively selected sites are concentrated at the same regions as epitopes (Seibert et al. 1995). More directly, muta-

**Table 7.** Tests for a decline in linkage disequilibrium with distance

Gene/protein	Sites analyzed <sup>a</sup>	Length of sequences	Number of sequences	Correlation		Mantel test	
				$r^2, d$	$ D' , d$	$P(r^2)$	$P( D' )$
<i>ompA</i>	All	3201	31	-0.028	0.001	<b>0.02</b>	0.53
<i>ompA</i>	> 0.1	3201	31	-0.020	-0.017	0.45	0.55
<i>ompB</i>	All	4983	31	-0.009	-0.020	0.14	<b>0.00</b>
<i>ompB</i>	> 0.1	4983	31	-0.008	0.003	0.29	0.61
PS120	All	3094	27	-0.021	-0.004	<b>0.00</b>	0.33
PS120	> 0.1	3094	27	-0.125	-0.095	<b>0.00</b>	<b>0.00</b>

<sup>a</sup>Analysis includes either all the sites with two alleles segregating or just those with frequencies > 0.1



**Fig. 6.** Polymorphic sites in an alignment of the *ompB* gene from *R. felis* (F), *Rickettsia* sp. strain California 2 (C), and *R. prowazekii* (P; AF123718). Nucleotides in the latter two species that are identical to *R. felis* are indicated by a dot. Note that the *R. felis* is a chimera of the other two sequences.

tions in the HIV Gag polyprotein that cause viral escape from cytotoxic T lymphocyte (CTL) recognition occur at positively selected sites (Leslie et al. 2004).

Can the location of positively selected sites in the *omp* genes predict the location of epitopes? The most comprehensive attempt to locate epitopes in rOmpB from *R. conorii* failed to find any in amino acids 458–692, but found evidence for five epitopes in the region 708–820 (Li et al. 2003). As shown in Fig. 1, although there are many positively selected sites in the latter region, there are also some in the region that does not contain epitopes. In the future, further information

on the location of epitopes from more species, together with the characterization of binding sites and other functional domains, may provide clues as to the causes of positive selection.

In other pathogens it has been suggested that mapping positively selected amino acids within proteins is a useful method of predicting the location of epitopes (Zanotto et al. 1999). This is clearly not the case here, but maps of selection pressures along the gene could still act as a useful guide in the design of vaccines. The most promising regions of the protein to use as a vaccine will be epitopes in the most strongly conserved regions, as these will be the least

variable in natural populations and the least likely to lead to the appearance of antigenic escape mutants.

**Recombination.** Recombination has profound implications for the evolution of pathogenic organisms. For example, it will lead to the accelerated rates of adaptive evolution, which may result in the spread of alleles resistant to drugs or vaccines. Even extremely rare recombination events can be of great importance in creating mosaic progenitors of successful lineages. Recombination can also complicate the naming and classification of organisms if there is genetic exchange between different “species.” In this case the evolutionary history of species may not follow a simple branching phylogenetic tree but, rather, a more complicated network where each species can have multiple ancestors (Smith et al. 1999).

The genome of *R. prowazekii* is very unusual among bacterial pathogens in that it contains no “alien genes” acquired from phylogenetically distant species (Karlin 2001). This suggests that the rate of genetic exchange in this genus may be unusually low. However, the genome sequences of *R. prowazekii* and *R. conorii* contain several genes involved in homologous recombination (Andersson et al. 1998; Ogata et al. 2001), suggesting that recombination can occur. Furthermore, homologous recombination between two strains of *R. prowazekii* has been observed in the laboratory (Rachek et al. 1998).

We found clear evidence that recombination has occurred between different species of *Rickettsia* bacteria. This recombination can be viewed as occurring at an intermediate taxonomic scale to the previous work described above, which looked for genetic exchange between either distantly related taxa (Karlin 2001) or closely related strains of the same species (Andersson et al. 1998; Ogata et al. 2001). Such recombination therefore needs to be considered as a factor during the emergence of new rickettsial diseases or during the spread of drug or vaccine resistance.

I did not attempt to estimate the rate at which recombination is occurring. However, the phylogenies of the three genes, although not identical, were broadly similar. This suggests that the majority of nucleotides in the genome share a common phylogenetic history. There were, however, significant differences in the phylogenetic trees inferred from two of the genes. This may not reflect the exchange of entire genes between bacterial lineages, as one of the key sequences underlying the conflict was itself a recombinant (Fig. 6) Such recombinant sequences can result in the recovery of trees that do not reflect the true phylogeny of either parental sequence (David and Keith 2002).

**Acknowledgments.** This work was funded by a Wellcome Trust Research Career Development Fellowship.

## References

- Anderson BE, McDonald GA, Jones DC, Regnery RL (1990) A protective protein antigen of *Rickettsia-Rickettsii* has tandemly repeated, near-identical sequences. *Infect Immun* 58:2760–2769
- Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UCM, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396:133–140
- Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* 18:1585–1592
- Awadalla P, Eyre-Walker A, Maynard-Smith J (1999) Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science* 286:2524–2525
- Carl M, Dobson ME, Ching WM, Dasch GA (1990) Characterization of the gene encoding the protective paracrystalline-surface-layer protein of *Rickettsia-Prowazekii*—Presence of a truncated identical homolog in *Rickettsia-Typhi*. *Proc Natl Acad Sci USA* 87:8237–8241
- Ching WM, Dasch GA, Carl M, Dobson ME (1990) structural-analyses of the 120-kDa serotype protein antigens of typhus group *Rickettsiae*—Comparison with other S-layer proteins. *Ann NY Acad Sci* 590:334–351
- Crewther PE, Matthew ML, Flegg RH, Anders RF (1996) Protective immune responses to apical membrane antigen 1 of *Plasmodium chabaudi* involve recognition of strain-specific epitopes. *Infect Immun* 64:3310–3317
- Croquet-Valdes PA, Diaz-Montero CM, Feng HM, Li H, Barrett ADT, Walker DH (2001) Immunization with a portion of rickettsial outer membrane protein A stimulates protective immunity against spotted fever rickettsiosis. *Vaccine* 20:979–988
- David P, Keith AC (2002) The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol* 54:396
- Davis MJ, Ying ZT, Brunner BR, Pantoja A, Ferwerda FH (1998) Rickettsial relative associated with papaya bunchy top disease. *Curr Microbiol* 36:80–84
- Fournier PE, Roux V, Raoult D (1998) Phylogenetic analysis of spotted fever group rickettsiae by study of the outer surface protein rOmpA. *Int J Systematic Bacteriol* 48:839–849
- Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood-based tests of topologies in phylogenetics. *Syst Biol* 49:652–670
- Jakobsen IB, Easteal S (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput Appl Biosci* 12:291–295
- Jorde LB, Bamshad M (2000) Questioning evidence for recombination in human mitochondrial DNA. *Science* 288:1931
- Karlin S (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol* 9:335–343
- Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, Feeney M, Tang Y, Holmes EC, Allen T, Prado JG, Altfeld M, Brander C, Dixon C, Ramduth D, Jeena P, Thomas SA, John AS, Roach TA, Kupfer B, Luzzi G, Edwards A, Taylor G, Lyall H, Tudor-Williams G, Novelli V, Martinez-Picado J, Kiepiela P, Walker BD, Goulder PJR (2004) HIV evolution: CTL escape mutation and reversion after transmission. *10:282*

- Li H, Walker DH (1998) rOmpA is a critical protein for the adhesion of *Rickettsia rickettsii* to host cells. *Microbial Pathogen* 24:289–298
- Li Z, Diaz-Montero CM, Valbuena G, Yu XJ, Olano JP, Feng HM, Walker DH (2003) Identification of CD8 T-lymphocyte epitopes in OmpB of *Rickettsia conorii*. *Infect Immun* 71:3920–3926
- Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16:562–563
- McVean GAT, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160:1231–1241
- Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936
- Ogata H, Audic S, Renesto-Audiffren P, Fournier PE, Barbe V, Samson D, Roux V, Cossart P, Weissenbach J, Claverie JM, Raoult D (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science* 293:2093–2098
- Posada D (2002) Evaluation of methods for detecting recombination from DNA sequences: Empirical data. *Mol Biol Evol* 19:708–717
- Posada D, Crandall K (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics* 14:817–818
- Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proc Natl Acad Sci USA* 98:13757–13762
- Rachek LI, Tucker AM, Winkler HH, Wood DO (1998) Transformation of *Rickettsia prowazekii* to rifampin resistance. *J Bacteriol* 180:2118–2124
- Raoult D, Roux V (1997) *Rickettsioses* as paradigms of new or emerging infectious diseases. *Clin Microbiol Rev* 10:694
- Schuenke KW, Walker DH (1994) Cloning, sequencing, and expression of the gene coding for an antigenic 120-kilodalton protein of *Rickettsia-Conorii*. *Infect Immun* 62:904–909
- Seibert SA, Howell CY, Hughes MK, Hughes AL (1995) Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1). *Mol Biol Evol* 12:803–813
- Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* 16:1114–1116
- Sleytr UB, Messner P (1983) Crystalline surface layers on bacteria. *Annu Rev Microbiol* 37:311–339
- Sleytr UB, Messner P (1988) Crystalline surface layers in prokaryotes. *J Bacteriol* 170:2891–2897
- Smith JM (1992) Analyzing the mosaic structure of genes. *J Mol Evol* 34:126–129
- Smith NH, Holmes EC, Donovan GM, Carpenter GA, Spratt BG (1999) Networks and groups within the genus *Neisseria*: Analysis of argF, recA, rho, and 16S rRNA sequences from human *Neisseria* species. *Mol Biol Evol* 16:773–783
- Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol* 20:18–20
- Swofford DL (1998) PAUP\*. Phylogenetic analysis using parsimony (\*and other methods). Version 4. Sinauer Associates, Sunderland, MA
- Uchiyama T (1997) Intracytoplasmic localization of antigenic heat-stable 120- to 130-kilodalton proteins (PS120) common to spotted fever group rickettsiae demonstrated by immunoelectron microscopy. *Microbiol Immunol* 41:815–818
- Uchiyama T (2003) Adherence to and invasion of Vero cells by recombinant *Escherichia coli* expressing the outer membrane protein rOmpB of *Rickettsia japonica*. In: *Rickettsiology: Present and Future Directions*, New York Academy of Science, New York, pp 585–590
- Werren JH, Hurst GDD, Zhang W, Breeuwer JAJ, Stouthamer R, Majerus MEN (1994) *Rickettsial* relative associated with male killing in the ladybird beetle (*Adalia bipunctata*). *J Bacteriol* 176:388–394
- Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556
- Yang ZH, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449
- Zanotto PMdA, Kallas EG, de Souza RF, Holmes EC (1999) Genealogical evidence for positive selection in the nef gene of HIV-1. *Genetics* 153:1077–1089